

Engineering Fundamentals

Major Contributors

Jean-Francois Chamberland and Henry D. Pfister

October 30, 2024

Contents

1	Logic and Set Theory	1
1.1	Statements	2
1.2	Relations between Statements	6
1.2.1	Fallacious Arguments	9
1.2.2	Quantifiers	10
1.3	Strategies for Proofs	12
1.4	Set Theory	14
1.5	Functions	19
2	Metric Spaces and Topology	23
2.1	Metric Spaces	24
2.1.1	Convergence	26
2.1.2	Metric Topology	27
2.1.3	Continuity	29
2.1.4	Properties of the Real Numbers	30
2.1.5	Sequences of Functions	33
2.1.6	Completeness	33
2.1.7	Compactness	37
2.2	General Topology*	39
2.2.1	Closed Sets and Limit Points	40
2.2.2	Continuity	42
3	Linear Algebra	47
3.1	Fields	47
3.2	Matrices	48
3.3	Vector Spaces	51

3.3.1	Subspaces	52
3.3.2	Bases and Dimensions	53
3.3.3	Coordinate System	56
3.4	Linear Transformations	58
3.4.1	Definitions	58
3.4.2	Properties	59
3.5	Norms	62
3.6	Inner Products	70
3.6.1	Induced Norms	74
3.7	Sets of Orthogonal Vectors	76
3.7.1	Hilbert Spaces	81
3.8	Linear Functionals	82
4	Representation and Approximation	85
4.1	Best Approximation	85
4.1.1	Projection Operators	88
4.2	Computing Approximations in Hilbert Spaces	91
4.2.1	Normal Equations	91
4.2.2	Orthogonality Principle	93
4.3	Approximation for Systems of Linear Equations	94
4.3.1	Matrix Representation	94
4.3.2	Standard Inner Products	95
4.3.3	Generalized Inner Products	96
4.3.4	Minimum Error	96
4.4	Applications and Examples in Signal Processing	97
4.4.1	Linear Regression	97
4.4.2	Linear Minimum Mean-Squared Error Estimation	97
4.4.3	The Wiener Filter	98
4.4.4	LMMSE Filtering in Practice	100
4.5	Dual Approximation	102
4.5.1	Minimum-Norm Solutions	102
4.5.2	Underdetermined Linear Systems	103
4.6	Projection onto Convex Sets	103
4.6.1	Projection Properties and Examples	106

4.6.2	Minimum Distance Between Two Convex Sets	109
5	Optimization	111
5.1	Derivatives in Banach Spaces	111
5.2	Unconstrained Optimization	117
5.3	Convex Functionals	118
5.4	Constrained Optimization	120
5.4.1	The Lagrangian	121
5.4.2	Lagrangian Duality	126
5.4.3	Convex Optimization	128
6	Linear Transformations and Operators	131
6.1	The Algebra of Linear Transformations	131
6.2	The Dual Space	134
6.3	Operator Norms	136
6.3.1	Bounded Transformations	138
6.3.2	The Neumann Expansion	139
6.3.3	Matrix Norms	139
6.4	Linear Functionals on Hilbert Spaces	141
6.5	Fundamental Subspaces	144
6.6	Pseudoinverses	145
6.6.1	Least Squares	146
7	Matrix Factorization and Analysis	149
7.1	Triangular Systems	149
7.1.1	Solution by Substitution	149
7.1.2	The Determinant	151
7.2	LU Decomposition	151
7.2.1	Introduction	151
7.2.2	Formal Approach	153
7.2.3	Partial Pivoting	155
7.3	LDLT and Cholesky Decomposition	156
7.3.1	Cholesky Decomposition	157
7.3.2	QR decomposition	157
7.4	Hermitian Matrices and Complex Numbers	158

8	Canonical Forms	159
8.1	Eigenvalues and Eigenvectors	159
8.2	Applications of Eigenvalues	163
8.2.1	Differential Equations	163
8.2.2	Functions of a Matrix	164
8.3	The Jordan Form	165
8.4	Applications of Jordan Normal Form	168
8.4.1	Convergent Matrices	168
9	Singular Value Decomposition	171
9.1	Diagonalization of Hermitian Matrices	171
9.2	Singular Value Decomposition	173
9.3	Properties of the SVD	175
A	Optional Topics	177
A.1	Dealing with Infinity*	177
A.1.1	The Axiom of Choice	177
A.1.2	Well-Ordered Sets	178
A.1.3	The Maximum Principle	179

Chapter 1

Logic and Set Theory

To criticize mathematics for its abstraction is to miss the point entirely. Abstraction is what makes mathematics work. If you concentrate too closely on too limited an application of a mathematical idea, you rob the mathematician of his most important tools: analogy, generality, and simplicity.

– *Ian Stewart*

Does God play dice? The mathematics of chaos

In mathematics, a **proof** is a demonstration that, assuming certain axioms, some statement is necessarily true. That is, a proof is a logical argument, not an empirical one. One must demonstrate that a proposition is true in all cases before it is considered a theorem of mathematics. An unproven proposition for which there is some sort of empirical evidence is known as a **conjecture**. Mathematical logic is the framework upon which rigorous proofs are built. It is the study of the principles and criteria of valid inference and demonstrations.

Logicians have analyzed set theory in great details, formulating a collection of axioms that affords a broad enough and strong enough foundation to mathematical reasoning. The standard form of axiomatic set theory is denoted ZFC and it consists of the Zermelo-Fraenkel (ZF) axioms combined with the axiom of choice (C). Each of the axioms included in this theory expresses a property of sets that is widely accepted by mathematicians. It is unfortunately true that careless use of set theory can lead to contradictions. Avoiding such contradictions was one of the original motivations for the axiomatization of set theory.

A rigorous analysis of set theory belongs to the foundations of mathematics and mathematical logic. The study of these topics is, in itself, a formidable task. For our purposes, it will suffice to approach basic logical concepts informally. That is, we adopt a naive point of view regarding set theory and assume that the meaning of a set as a collection of objects is intuitively clear. While informal logic is not itself rigorous, it provides the underpinning for rigorous proofs. The rules we follow in dealing with sets are derived from established axioms. At some point of your academic career, you may wish to study set theory and logic in greater detail. Our main purpose here is to learn how to state mathematical results clearly and how to prove them.

1.1 Statements

A proof in mathematics demonstrates the truth of certain **statement**. It is therefore natural to begin with a brief discussion of statements. A statement, or **proposition**, is the content of an assertion. It is either true or false, but cannot be both true and false at the same time. For example, the expression “There are no classes at Texas A&M University today” is a statement since it is either true or false. The expression “Do not cheat and do not tolerate those who do” is not a statement. Note that an expression being a statement does not depend on whether we personally can verify its validity. The expression “The base of the natural logarithm, denoted e , is an irrational number” is a statement that most of us cannot prove.

Statements on their own are fairly uninteresting. What brings value to logic is the fact that there are a number of ways to form new statements from old ones. In this section, we present five ways to form new statements from old ones. They correspond to the English expressions: and; or; not; if, then; if and only if. In the discussion below, P and Q represent two abstract statements.

A logical **conjunction** is an operation on two logical propositions that produces a value of true if both statements are true, and is false otherwise. The conjunction (or logical AND) of P and Q , denoted by $P \wedge Q$, is precisely defined by

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

Similarly, a logical **disjunction** is an operator on two logical propositions that is true if either statement is true or both are true, and is false otherwise. The disjunction (or logical OR) of P and Q , denoted $P \vee Q$, is defined by

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

In mathematics, a **negation** is an operator on the logical value of a proposition that sends true to false and false to true. The negation (or logical NOT) of P , denoted $\neg P$, is given by

P	$\neg P$
T	F
F	T

The next method of combining mathematical statements is slightly more subtle than the preceding ones. The **conditional connective** $P \rightarrow Q$ is a logical statement that is read “if P then Q ” and defined by the truth table

P	Q	$P \rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

In this statement, P is called the **antecedent** and Q is called the **consequent**. The truth table should match your intuition when P is true. When P is false, students often think the resulting truth value should be undefined. Although the given definition may seem strange at first glance, this truth table is universally accepted by mathematicians.

To motivate this definition, one can think of $P \rightarrow Q$ as a promise that Q is true whenever P is true. When P is false, the promise is kept by default. For example, suppose your friend promises “if it is sunny tomorrow, I will ride my bike”. We will call this a true statement if they keep their promise. If it rains and they don’t ride their bike, most people would agree that they have still kept their promise. Therefore, this definition allows one to combine many statements together and detect broken promises without being distracted by uninformative statements.

Logicians draw a firm distinction between the **conditional connective** and the **implication relation**. They use the phrase “if P then Q ” for the conditional connective and the phrase “ P implies Q ” for the implication relation. They explain the difference between these two forms by saying that the conditional is the contemplated relation, while the implication is the asserted relation. We will discuss this distinction in the Section 1.2, where we formally study relations between statements. The importance and soundness of the conditional form $P \rightarrow Q$ will become clearer then.

The logical **biconditional** is an operator connecting two logical propositions that is true if the statements are both true or both false, and it is false otherwise. The biconditional from P to Q , denoted $P \leftrightarrow Q$, is precisely defined by

P	Q	$P \leftrightarrow Q$
T	T	T
T	F	F
F	T	F
F	F	T

We read $P \leftrightarrow Q$ as “ P if and only if Q .” The phrase “if and only if” is often abbreviated as “iff”.

Using the five basic operations defined above, it is possible to form more complicated compound statements. We sometimes need parentheses to avoid ambiguity in writing compound statements. We use the convention that \neg takes precedence over the other four operations, but none of these operations takes precedence over the others. For example, let P , Q and R be three propositions. We wish to make a truth table for the following statement,

$$(P \rightarrow R) \wedge (Q \vee \neg R). \quad (1.1)$$

We can form the true table for this statement, using simple steps, as follows

P	Q	R	$(P \rightarrow R)$	\wedge	$(Q \vee \neg R)$
T	T	T	T	T	F
T	T	F	F	F	T
T	F	T	T	F	F
T	F	F	F	F	T
F	T	T	T	T	F
F	T	F	F	T	T
F	F	T	T	F	F
F	F	F	F	T	T
			1	5	2

We conclude this section with a brief mention of two important concepts. A **tautology** is a statement that is true in every valuation of its propositional variables, independent of the truth values assigned to these variables. The proverbial tautology is $P \vee \neg P$,

P	$P \vee \neg P$
T	T
F	T
	1

For instance, the statement “The Aggies won their last football game or the Aggies did not win their last football game” is true regardless of whether the Aggies actually defeated their latest opponent.

The negation of a tautology is a **contradiction**, a statement that is necessarily false regardless of the truth values of its propositional variables. The statement $P \wedge \neg P$ is a contradiction, and its truth table is

P	$P \wedge \neg P$
T	F
F	F
	1

Of course, most statements we encounter are neither tautologies nor contradictions. For example, (1.1) is not necessarily either true or false. Its truth value depends on the values of P , Q and R . Try to see whether the statement

$$((P \wedge Q) \rightarrow R) \rightarrow (P \rightarrow (Q \rightarrow R))$$

is a tautology, a contradiction, or neither.

1.2 Relations between Statements

Strictly speaking, relations between statements are not formal statements themselves. They are *meta-statements* about some propositions. We study two types of relations between statements, *implication* and *equivalence*. An example of an implication meta-statement is the observation that “if the statement ‘Robert graduated from Texas A&M University’ is true, then it implies that the statement ‘Robert is an Aggie’ is also true.” Another example of a meta-statement is “the statement ‘Fred is an Aggie and Fred is honest’ being true is equivalent to the statement ‘Fred is honest and Fred is an Aggie’ being true.” These two examples illustrate how meta-statements describe the relationship between statements. It is also instructive to note that implications and equivalences are the meta-statement analogs of conditionals and biconditionals.

Consider two compound statements P and Q that depend on other logical statements (e.g., $P = (R \rightarrow S) \wedge (S \rightarrow T)$ and $Q = R \rightarrow T$). A **logical implication** from P to Q , read as “ P implies Q ”, asserts that Q must be true whenever P is true (i.e., for all possible truth values of the dependent statements R, S, T). Necessity is the key aspect of this sentence; the fact that P and Q both happen to be true cannot be coincidental. To state that P implies Q , denoted by $P \Rightarrow Q$, one needs the conditional $P \rightarrow Q$ to be true under all possible circumstances.

Meta-statements, such as “ P implies Q ”, can be defined formally only when P and Q are both logical functions of other propositions. For example, consider $P = R \wedge (R \rightarrow S)$ and $Q = S$. Then, the truth of the statement $P \rightarrow Q$ depends only on the truth of external propositions R and S .

The notion of implication can be rigorously defined as follows, P implies Q if the statement $P \rightarrow Q$ is a tautology. We abbreviate P implies Q by writing $P \Rightarrow Q$. It is important to understand the difference between “ $P \rightarrow Q$ ” and “ $P \Rightarrow Q$.” The former, $P \rightarrow Q$, is a compound statement that may or may not be true. On the other hand, $P \Rightarrow Q$ is a relation stating that the compound statement $P \rightarrow Q$ is true under all instances of the external propositions.

While the distinction between implication and conditional may seem extraneous, we will soon see that meta-statements become extremely useful in building valid arguments. In particular, the following implications are used extensively in constructing proofs.

Fact 1.2.1. Let P, Q, R and S be statements.

1. $(P \rightarrow Q) \wedge P \Rightarrow Q$.
2. $(P \rightarrow Q) \wedge \neg Q \Rightarrow \neg P$.
3. $P \wedge Q \Rightarrow P$.
4. $(P \vee Q) \wedge \neg P \Rightarrow Q$.
5. $P \leftrightarrow Q \Rightarrow P \rightarrow Q$.
6. $(P \rightarrow Q) \wedge (Q \rightarrow P) \Rightarrow P \rightarrow Q$.
7. $(P \rightarrow Q) \wedge (Q \rightarrow R) \Rightarrow P \rightarrow R$
8. $(P \rightarrow Q) \wedge (R \rightarrow S) \wedge (P \vee R) \Rightarrow Q \vee S$.

As an illustrative example, we show that $(P \rightarrow Q) \wedge (Q \rightarrow R)$ implies $P \rightarrow R$. To demonstrate this assertion, we need to show that

$$((P \rightarrow Q) \wedge (Q \rightarrow R)) \rightarrow (P \rightarrow R) \tag{1.2}$$

is a tautology. This is accomplished in the truth table below

P	Q	R	$((P \rightarrow Q) \wedge (Q \rightarrow R))$	\rightarrow	$(P \rightarrow R)$
T	T	T	T	T	T
T	T	F	F	T	F
T	F	T	F	T	T
T	F	F	F	T	F
F	T	T	T	T	T
F	T	F	F	T	F
F	F	T	F	T	T
F	F	F	F	T	F
			1	7	2
			10	3	8
			4	11	5
			9	6	

Column 11 has the truth values for statement (1.2). Since (1.2) is true under all circumstances, it is a tautology and the implication holds. Showing that the other relations are valid is left to the reader as an exercise.

Reversing the arrow in a conditional statement gives the **converse** of that statement. For example, the statement $Q \rightarrow P$ is the converse of $P \rightarrow Q$. This reversal

may not preserve the truth of the statement though and therefore logical implications are not always reversible. For instance, although $(P \rightarrow Q) \wedge (Q \rightarrow R)$ implies $P \rightarrow R$, the converse is not always true. It can easily be seen from columns 9 & 10 above that

$$(P \rightarrow R) \rightarrow ((P \rightarrow Q) \wedge (Q \rightarrow R))$$

is not a tautology. That is, $P \rightarrow R$ certainly does not imply $(P \rightarrow Q) \wedge (Q \rightarrow R)$.

A logical implication that is reversible is called a **logical equivalence**. More precisely, P is equivalent to Q if the statement $P \leftrightarrow Q$ is a tautology. We denote the sentence “ P is equivalent to Q ” by simply writing “ $P \Leftrightarrow Q$.” The meta-statement $P \Leftrightarrow Q$ holds if and only if $P \Rightarrow Q$ and $Q \Rightarrow P$ are both true. Being able to recognize that two statements are equivalent will become handy. It is sometime possible to demonstrate a result by finding an alternative, equivalent form of the statement that is easier to prove than the original form. A list of important equivalences appears below.

Fact 1.2.2. *Let P, Q and R be statements.*

1. $\neg(\neg P) \Leftrightarrow P$.
2. $P \vee Q \Leftrightarrow Q \vee P$.
3. $P \wedge Q \Leftrightarrow Q \wedge P$.
4. $(P \vee Q) \vee R \Leftrightarrow P \vee (Q \vee R)$.
5. $(P \wedge Q) \wedge R \Leftrightarrow P \wedge (Q \wedge R)$.
6. $P \wedge (Q \vee R) \Leftrightarrow (P \wedge Q) \vee (P \wedge R)$.
7. $P \vee (Q \wedge R) \Leftrightarrow (P \vee Q) \wedge (P \vee R)$.
8. $P \rightarrow Q \Leftrightarrow \neg P \vee Q$.
9. $P \rightarrow Q \Leftrightarrow \neg Q \rightarrow \neg P$ (*Contrapositive*).
10. $P \leftrightarrow Q \Leftrightarrow (P \rightarrow Q) \wedge (Q \rightarrow P)$.
11. $\neg(P \wedge Q) \Leftrightarrow \neg P \vee \neg Q$ (*De Morgan's Law*).
12. $\neg(P \vee Q) \Leftrightarrow \neg P \wedge \neg Q$ (*De Morgan's Law*).

Given a conditional statement of the form $P \rightarrow Q$, we call $\neg Q \rightarrow \neg P$ the **contrapositive** of the original statement. The equivalence $P \rightarrow Q \Leftrightarrow \neg Q \rightarrow \neg P$ noted above is used extensively in constructing mathematical proofs.

One must be careful not to allow contradictions in logical arguments because, starting from a contradiction, anything can be proven true. For example, one can verify that $P \wedge \neg P \Rightarrow Q$ is a valid logical equivalence. But, Q doesn't appear on the LHS. Thus, a contradiction in your assumptions can lead to a "correct" proof for an arbitrary statement.

Fortunately, propositional logic has an axiomatic formulation that is consistent, complete, and decidable. In this context, the term **consistent** means that the logical implications generated by the axioms do not contain a contradiction, the term **complete** means that any valid logical implication can be generated by applying the axioms, and the term **decidable** means there is a terminating method that always determines whether a postulated implication is valid or invalid.

1.2.1 Fallacious Arguments

A **fallacy** is a component of an argument that is demonstrably flawed in its logic or form, thus rendering the argument invalid. Recognizing fallacies in mathematical proofs may be difficult since arguments are often structured using convoluted patterns that obscure the logical connections between assertions. We give below examples for three types of fallacies that are often found in attempted mathematical proofs.

Affirming the Consequent: If the Indian cricket team wins a test match, then all the players will drink tea together. All the players drank tea together. Therefore the Indian cricket team won a test match.

Denying the Antecedent: If Diego Maradona drinks coffee, then he will be fidgety. Diego Maradona did not drink coffee. Therefore, he is not fidgety.

Unwarranted Assumptions: If Yao Ming gets close to the basket, then he scores a lot of points. Therefore, Yao Ming scores a lot of points.

1.2.2 Quantifiers

Consider the statements “Socrates is a person” and “Every person is mortal”. In propositional logic, there is no formal way to combine these statements to deduce that “Socrates is mortal”. In the first statement, the noun “Socrates” is called the subject and the phrase “is a person” is called the **predicate**. Likewise, in predicate logic, the statement $P(x) = “x \text{ is a person}”$ is called a predicate and x is called a **free variable** because its value is not fixed in the statement $P(x)$.

Let U be a specific collection of elements and let $P(x)$ be a statement that can be applied to any $x \in U$. In first-order predicate logic, quantifiers are applied to predicates in order to make statements about collections of elements. Later, we will see that quantifiers are of paramount importance in rigorous proofs.

The **universal quantifier** is typically denoted by \forall and it is informally read “for all.” It follows that the statement “ $\forall x \in U, P(x)$ ” is true if $P(x)$ is true for all values of x in U . It can be seen as shorthand for an iterated conjunction because

$$\forall x \in U, P(x) \Leftrightarrow \bigwedge_{x \in U} P(x),$$

where \Leftrightarrow indicates that these statements are equivalent for all sets U and predicates P . If $U = \emptyset$ is the empty set, then $\forall x \in U, P(x)$ is vacuously true by convention because there are no elements in U to test with $P(x)$.

Returning to the motivating example, let us also define $Q(x) = “x \text{ is mortal}”$. With these definitions, we can write the statement “Every person is mortal” as $\forall x, (P(x) \rightarrow Q(x))$. In logic, this usage implies that x ranges over the universal set. In engineering mathematics, however, the range of free variables is typically stated explicitly.

The other type of quantifier often seen in mathematical proofs is the **existential quantifier**, denoted \exists . The statement “ $\exists x \in U, P(x)$ ” is true if $P(x)$ is true for at least one value of x in U . It can be seen as shorthand for an iterated disjunction because

$$\exists x \in U, P(x) \Leftrightarrow \bigvee_{x \in U} P(x),$$

If $U = \emptyset$ is the empty set, then $\exists x \in U, P(x)$ is false by convention because there are no elements in U .

The idea of **universal instantiation** is that, if a statement $P(x)$ is true for all $x \in U$, then there must exist some $x_0 \in U$ such that $P(x_0)$ is true. However, this

implicitly assumes that U is not empty. In fact, universal instantiation does not hold when U is empty. If U is not empty though, then universal instantiation implies that $\forall x \in U, P(x) \Rightarrow \exists x \in U, P(x)$.

Based on the meaning of these quantifiers, one can infer the logical implications

$$\begin{aligned}\neg(\forall x \in U, P(x)) &\Leftrightarrow \exists x \in U, \neg P(x) \\ \neg(\exists x \in U, P(x)) &\Leftrightarrow \forall x \in U, \neg P(x).\end{aligned}$$

Using the connection to conjunction and disjunction, these rules are actually equivalent to De Morgan's law for iterated conjunctions and disjunctions.

One can also define predicates with multiple free variables such as $P(x, y)$ = “ x contains y ”. Once again, these statements are assumed to be true or false for every choice of x, y . There are 8 possible quantifiers for a 2-variable predicate and they can be arranged according based on some natural implications. Assuming that x, y are taken from non-empty sets, one finds that

$$\begin{array}{ccccccc}\forall x, \forall y, P(x, y) & \Rightarrow & \exists x, \forall y, P(x, y) & \Rightarrow & \forall y, \exists x, P(x, y) & \Rightarrow & \exists y, \exists x, P(x, y) \\ & & \Downarrow & & & & \Downarrow \\ \forall y, \forall x, P(x, y) & \Rightarrow & \exists y, \forall x, P(x, y) & \Rightarrow & \forall x, \exists y, P(x, y) & \Rightarrow & \exists x, \exists y, P(x, y)\end{array}$$

All of these implications follow from $\forall x \forall y = \forall y \forall x$, $\exists x \exists y = \exists y \exists x$, and the single variable inference rule $\forall x, P(x) \Rightarrow \exists x, P(x)$ except for two: $\exists x, \forall y, P(x, y) \Rightarrow \forall y, \exists x, P(x, y)$ and its symmetric pair.

To understand this last implication, consider an example where x is in a set I of images and y is in a set C of colors. Then, $\exists x, \forall y, P(x, y)$ means “there is an image that contains all the colors” (e.g., an image of a rainbow) and $\forall y, \exists x, P(x, y)$ means “for each color there is an image containing that color”. The first statement implies the second because, in the second, the rainbow image satisfies the $\exists x$ quantifier for all y . To see that the implication is not an equivalence, consider a set of pictures where each image contains exactly one color and there is one such image for each color. In this case, it is true that “for each color there is an image containing that color” but it is not true that “there is an image that contains all the colors”.

In quantified statements, such as $\exists x \in U, P(x)$, the variable x is called a **bound variable** because its value cannot be chosen freely. Similarly, in the statement $\exists y \in U, P(x, y)$, x is a free variable and y is a bound variable.

Finally, we note that first-order predicate logic has an axiomatic formulation that is consistent, complete, and semidecidable. In this context, **semidecidable** means that there is an algorithm that, if it terminates, correctly determines the truth of any postulated implication. But, it is only guaranteed to terminate for true postulates.

1.3 Strategies for Proofs

The relation between intuition and formal rigor is not a trivial matter. Intuition tells us what is important, what might be true, and what mathematical tools may be used to prove it. Rigorous proofs are used to verify that a given statement which appears intuitively true is indeed true. Ultimately, a mathematical proof is a convincing argument that starts from some premises, and logically deduces the desired conclusion. Most proofs do not mention the logical rules of inference used in the derivation. Rather, they focus on the mathematical justification of each step, leaving to the reader the task of filling the logical gaps. The mathematics is the major issue. Yet, it is essential that you understand the underlying logic behind the derivation as to not get confused while reading or writing a proof.

True statements in mathematics have different names. They can be called theorems, propositions, lemmas, corollaries and exercises. A **theorem** is a statement that can be proved on the basis of explicitly stated or previously agreed assumptions. A **proposition** is a statement not associated with any particular theorem; this term sometimes connotes a statement with a simple proof. A **lemma** is a proven proposition which is used as a stepping stone to a larger result rather than an independent statement in itself. A **corollary** is a mathematical statement which follows easily from a previously proven statement, typically a mathematical theorem. The distinction between these names and their definitions is somewhat arbitrary. Ultimately, they are all synonymous to a true statement.

A proof should be written in grammatically correct English. Complete sentences should be used, with full punctuation. In particular, every sentence should end with a period, even if the sentence ends in a displayed equation. Mathematical formulas and symbols are parts of sentences, and are treated no differently than words. One way to learn to construct proofs is to read a lot of well written proofs, to write progressively more difficult proofs, and to get detailed feedback on the proofs you write.

Direct Proof: The simplest form of proof for a statement of the form $P \rightarrow Q$ is the **direct proof**. First assume that P is true. Produce a series of steps, each one following from the previous ones, that eventually leads to conclusion Q . It warrants the name “direct proof” only to distinguish it from other, more intricate, methods of proof.

Proof by Contrapositive: A proof by contrapositive takes advantage of the mathematical equivalence $P \rightarrow Q \Leftrightarrow \neg Q \rightarrow \neg P$. That is, a proof by contrapositive begins by assuming that Q is false (i.e., $\neg Q$ is true). It then produces a series of direct implications leading to the conclusion that P is false (i.e., $\neg P$ is true). It follows that Q cannot be false when P is true, so $P \rightarrow Q$.

Proof by Contradiction: A proof by contradiction is based on the mathematical equivalence $\neg(P \rightarrow Q) \Leftrightarrow P \wedge \neg Q$. In a proof by contradiction, one starts by assuming that both P and $\neg Q$ are true. Then, a series of direct implications are given that lead to a logical contradiction. Hence, $P \wedge \neg Q$ cannot be true and $P \rightarrow Q$.

Example 1.3.1. *We wish to show that $\sqrt{2}$ is an irrational number.*

First, suppose that $\sqrt{2}$ is a rational number. This would imply that there exist integers p and q with $q \neq 0$ such that $p/q = \sqrt{2}$. In fact, we can further assume that the fraction p/q is irreducible. That is, p and q are coprime integers (they have no common factor greater than 1). From $p/q = \sqrt{2}$, it follows that $p = \sqrt{2}q$, and so $p^2 = 2q^2$. Thus p^2 is an even number, which implies that p itself is even (only even numbers have even squares). Because p is even, there exists an integer r satisfying $p = 2r$. We then obtain the equation $(2r)^2 = 2q^2$, which is equivalent to $2r^2 = q^2$ after simplification. Because $2r^2$ is even, it follows that q^2 is even, which means that q is also even. We conclude that p and q are both even. This contradicts the fact that p/q is irreducible. Hence, the initial assumption that $\sqrt{2}$ is a rational number must be false. That is to say, $\sqrt{2}$ is irrational.

Example 1.3.2. *Consider the following statement, which is related to Example 1.3.1. “If $\sqrt{2}$ is rational, then $\sqrt{2}$ can be expressed as an irreducible fraction.” The contrapositive of this statement is “If $\sqrt{2}$ cannot be expressed as an irreducible fraction, then $\sqrt{2}$ is not rational.” Above, we proved that $\sqrt{2}$ cannot be expressed as an irreducible fraction and therefore $\sqrt{2}$ is not a rational number.*

The final proof strategy we discuss is finite induction.

Definition 1.3.3. Let $P(n)$ be a logical statement for each $n \in \mathbb{N}$. The principle of **mathematical induction** states that $P(n)$ is true all $n \in \mathbb{N}$ if:

1. $P(1)$ is true, and
2. $P(n) \rightarrow P(n + 1)$ for all $n \in \mathbb{N}$.

From a foundational perspective, this statement is essentially equivalent to the existence and uniqueness of the natural numbers. It is taken as an axiom in the Peano axiomatic formulation of arithmetic. In contrast, the ZF axiomatic formulation of set theory defines the natural numbers as the smallest inductive set and the existence of an inductive set is taken as an axiom.

Example 1.3.4. Let $S_n = \sum_{i=1}^n i$. We wish to show that the statement $P(n) = "S_n = \frac{n^2+n}{2}"$ is true for all $n \in \mathbb{N}$. For $n = 1$, this is true because both expressions equal 1. For $P(n + 1)$, we are given $P(n)$ and can write

$$S_{n+1} = S_n + (n + 1) = \frac{n^2 + n}{2} + n + 1 = \frac{n^2 + 3n + 2}{2} = \frac{(n + 1)^2 + (n + 1)}{2}.$$

Thus, the result follows from mathematical induction.

More general forms of finite induction are also quite common but they can be reduced to the original form. For example, let $Q(m)$ be a predicate for $m \in \mathbb{N}$ and define $P(n) = "\forall m \in S_n, Q(m)"$ for a sequence of nested finite sets $S_1 \subset S_2 \subset \dots \subseteq \mathbb{N}$. Defining $S_\infty = \cup_{n \in \mathbb{N}} S_n$, we see that $"\forall n \in \mathbb{N}, P(n)" \Leftrightarrow "\forall m \in S_\infty, Q(m)"$ follows from $P(1) = "\forall m \in S_1, Q(m)"$ and $"P(n) \rightarrow P(n + 1)" \Leftrightarrow "\forall m \in S_n, Q(m) \rightarrow \forall m \in S_{n+1}, Q(m)"$.

1.4 Set Theory

Set theory is generally considered to be the foundation of all modern mathematics. This means that most mathematical objects (numbers, relations, functions, etc.) are defined in terms of sets. Unfortunately for engineers, set theory is not quite as simple as it seems. It turns out that simple approaches to set theory include paradoxes (e.g., statements which are both true and false). These paradoxes can

be resolved by putting set theory in a firm axiomatic framework, but that exercise is rather unproductive for engineers. Instead, we adopt what is called **naive set theory** which rigorously defines the operations of set theory without worrying about possible contradictions. This approach is sufficient for most of mathematics and also acts as a stepping-stone to more formal treatments.

A **set** is taken to be any collection of objects, mathematical or otherwise. For example, one can think of “the set of all books published in 2007”. The objects in a set are referred to as **elements** or members of the set. The logical statement “ a is a member of the set A ” is written

$$a \in A.$$

Likewise, its logical negation “ a is not a member of the set A ” is written $a \notin A$. Therefore, exactly one of these two statements is true. In naive set theory, one assumes the existence of any set that can be described in words. Later, we will see that this can be problematic when one considers objects like the “set of all sets”.

One may present a set by listing its elements. For example, $A = \{a, e, i, o, u\}$ is the set of standard English vowels. It is important to note that the order elements are presented is irrelevant and the set $\{i, o, u, a, e\}$ is the same as A . Likewise, repeated elements have no effect and the set $\{a, e, i, o, u, e, o\}$ is the same as A . A **singleton** set is a set containing exactly one element such as $\{a\}$.

There are a number of standard sets worth mentioning: the **integers** \mathbb{Z} , the **real numbers** \mathbb{R} , and the **complex numbers** \mathbb{C} . It is possible to construct these sets in a rigorous manner, but instead we will assume their meaning is intuitively clear. New sets can be defined in terms of old sets using **set-builder notation**. Let $P(x)$ be a logical statement about objects x in the set X , then the “set of elements in X such that $P(x)$ is true” is denoted by

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

For example, the set of even integers is given by

$$\{x \in \mathbb{Z} | “x \text{ is even}”\} = \{\dots, -4, -2, 0, 2, 4, \dots\}.$$

If no element $x \in X$ satisfies the condition, then the result is the **empty set** which is denoted \emptyset . Using set-builder notation, we can also recreate the **natural numbers**

\mathbb{N} and the **rational numbers** \mathbb{Q} with

$$\mathbb{N} = \{n \in \mathbb{Z} | n \geq 1\}$$

$$\mathbb{Q} = \{q \in \mathbb{R} | q = a/b, a \in \mathbb{Z}, b \in \mathbb{N}\}.$$

The following standard notation is used for interval subsets of the real numbers:

$$\text{Open interval: } (a, b) \triangleq \{x \in \mathbb{R} | a < x < b\}$$

$$\text{Closed interval: } [a, b] \triangleq \{x \in \mathbb{R} | a \leq x \leq b\}$$

$$\text{Half-open intervals: } (a, b] \triangleq \{x \in \mathbb{R} | a < x \leq b\}$$

$$[a, b) \triangleq \{x \in \mathbb{R} | a \leq x < b\}$$

Definition 1.4.1. For a finite set A , the **cardinality** $|A|$ equals the number of elements in A . If there is a bijective mapping between the set A and the natural numbers \mathbb{N} , then $|A| = \infty$ and the set is called **countably infinite**. If $|A| = \infty$ and the set is not countably infinite, then A is called **uncountably infinite**.

Example 1.4.2. The set of rational numbers is countably infinite while the set of real numbers is uncountably infinite.

Example 1.4.3 (Russell's Paradox). Let R be the set of all sets that do not contain themselves or $R = \{S | S \notin S\}$. Such a set is said to exist in naive set theory (though it may empty) simply because it can be described in words. The paradox arises from the fact that the definition leads to the logical contradiction $R \in R \leftrightarrow R \notin R$.

What this proves is that *naive set theory is not consistent* because it allows constructions that lead to contradictions. Axiomatic set theory eliminates this paradox by disallowing self-referential and other problematic constructions. Thus, another reasonable conclusion is that Russell's paradox shows that the set R cannot exist in any consistent theory of sets.

Another common question is whether there are sets that contains themselves. In naive set theory, the answer is yes and some examples are the "set of all sets" and the "set of all abstract ideas". On the other hand, in the ZF axiomatic formulation of set theory, it is a theorem that no set contains itself.

There are a few standard relationships defined between any two sets A, B .

Definition 1.4.4. We say that A **equals** B (denoted $A = B$) if, for all x , $x \in A$ iff $x \in B$. This means that

$$A = B \Leftrightarrow \forall x ((x \in A) \leftrightarrow (x \in B)).$$

Definition 1.4.5. We say that A is a **subset** of B (denoted $A \subseteq B$) if, for all x , if $x \in A$ then $x \in B$. This means that

$$A \subseteq B \Leftrightarrow \forall x ((x \in A) \rightarrow (x \in B)).$$

It is a **proper subset** (denoted $A \subset B$) if $A \subseteq B$ and $A \neq B$.

There are also a number of operations between sets. Let A, B be any two sets.

Definition 1.4.6. The **union** of A and B (denoted $A \cup B$) is the set of elements in either A or B . This means that $A \cup B = \{x \in A \text{ or } x \in B\}$ is also defined by

$$x \in A \cup B \Leftrightarrow (x \in A) \vee (x \in B).$$

Definition 1.4.7. The **intersection** of A and B (denoted $A \cap B$) is the set of elements in both A and B . This means that $A \cap B = \{x \in A | x \in B\}$ is also defined by

$$x \in A \cap B \Leftrightarrow (x \in A) \wedge (x \in B).$$

Two sets are said to be **disjoint** if $A \cap B = \emptyset$.

Definition 1.4.8. The **set difference** between A and B (denoted $A - B$ or $A \setminus B$) is the set of elements in A but not in B . This means that

$$x \in A - B \Leftrightarrow (x \in A) \wedge (x \notin B).$$

If there is some implied universal set U , then the **complement** (denoted A^c) is defined by $A^c = U - A$

One can apply De Morgan's Law in set theory to verify that

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup B^c,$$

which allows us to interchange union or intersection with set difference.

We can also form the union or the intersection of arbitrarily many sets. This is defined in a straightforward way,

$$\bigcup_{\alpha \in I} S_\alpha = \{x \mid x \in S_\alpha \text{ for some } \alpha \in I\}$$

$$\bigcap_{\alpha \in I} S_\alpha = \{x \mid x \in S_\alpha \text{ for all } \alpha \in I\}.$$

It is worth noting that the definitions apply whether the index set is finite, countably infinite, or even uncountably infinite.

Another way to build sets is by grouping elements into pairs, triples, and vectors.

Definition 1.4.9. *The **Cartesian Product**, denoted $A \times B$, of two sets is the set of ordered pairs $\{(a, b) \mid a \in A, b \in B\}$. For n -tuples taken from the same set, the notation A^n denotes the n -fold product $A \times A \times \cdots \times A$.*

Example 1.4.10. *If $A = \{a, b\}$, then the set of all 3-tuples from A is given by*

$$A^3 = \{(a, a, a), (a, a, b), (a, b, a), (a, b, b), (b, a, a), (b, a, b), (b, b, a), (b, b, b)\}.$$

The countably infinite product of X , denoted X^ω , is the set of infinite sequences (x_1, x_2, x_3, \dots) where $x_n \in X$ is arbitrary for $n \in \mathbb{N}$. If the sequences are restricted to have only a finite number of non-zero terms, then the set is usually denoted X^∞ .

One can also formalize relationships between elements of a set. A **relation** \sim between elements of the set A is defined by the pairs $(x, y) \in A \times A$ for which the relation holds. Specifically, the relation is defined by the subset of ordered pairs $E \subseteq A \times A$ where the relation $a \sim b$ holds; so $x \sim y$ if and only if $(x, y) \in E$. A relation on A is said to be:

1. Reflexive if $x \sim x$ holds for all $x \in A$
2. Symmetric if $x \sim y$ implies $y \sim x$ for all $x, y \in A$
3. Transitive if $x \sim y$ and $y \sim z$, then $x \sim z$ for all $x, y, z \in A$

A relation is called an **equivalence relation** if it is reflexive, symmetric, and transitive. For example, let A be a set of people and $P(x, y)$ be the statement “ x has the same birthday (month and day) as y .” Then, we can define \sim such that $a \sim b$ holds if and only if $P(x, y)$ is true. In this case, the set E is given by

$E = \{(x, y) \in A \times A \mid P(x, y)\}$. One can verify that this is an equivalence relation by checking that it is reflexive, symmetric, and transitive.

One important characteristic of an equivalence relation is that it partitions the entire set A into disjoint **equivalence classes**. The equivalence class associated with $a \in A$ is given by $[a] = \{x \in A \mid x \sim a\}$. In the birthday example, there is a natural equivalence class associated with each day of the year. The set of all equivalence classes is called the **quotient set** and is denoted $A/\sim \triangleq \{[a] \mid a \in A\}$.

In fact, there is a natural equivalence relation defined by any disjoint partition of a set. For example, let $A_{i,j}$ be the set of people in A whose birthday was on the j -th day of the i -th month. It follows that $x \sim y$ if and only if there exists a unique pair i, j such that $x, y \in A_{i,j}$. In this case, the days of year are used as equivalence classes to define the equivalence relation.

Example 1.4.11. Consider the set $\mathbb{N}^2 = \{(a, b) \mid a, b \in \mathbb{N}\}$ of ordered pairs of natural numbers. If one associates the element (a, b) with the fraction a/b , then the entire set is associated with the set of (possibly reducible) fractions. Now, consider the equivalence relation $(a, b) \sim (c, d)$ if $ad = bc$. In this case, two ordered pairs are equivalent if their associated fractions evaluate to the same real number. The quotient set \mathbb{N}^2/\sim can therefore be associated with the set of reduced fractions.

Unfortunately, this section will not end on a happy note by saying that the ZFC axiomatic formulation of set theory is consistent. Instead, we observe that Kurt Gödel's Incompleteness Theorems imply that, if ZFC is consistent, then this cannot be proven using statements in ZFC and, moreover, it cannot be complete. On the other hand, if ZFC is inconsistent, then it contains a paradox and one can prove anything using statements in ZFC. Since ZFC manages to avoid all known paradoxes and no contradictions have been so far, it is still the most popular formal system in which to define mathematics.

1.5 Functions

In elementary mathematics, functions are typically described in terms of graphs and formulas. The drawback of this approach is that one tends to picture only “nice” functions. In fact, Cauchy himself published in 1821 an incorrect proof of the false assertion that “a sequence of continuous functions that converges everywhere has a

continuous limit function.” Nowadays, every teacher warns their students that one must be careful because the world is filled with “not so nice” functions.

The modern approach to defining functions is based on set theory. A **function** $f: X \rightarrow Y$ is a rule that assigns a single value $f(x) \in Y$ to each element $x \in X$. The notation $f: X \rightarrow Y$ is used to emphasize the role of the **domain** X and the **codomain** Y . The **range** of f is the subset of Y which is actually achieved by f , $\{f(x) \in Y | x \in X\}$. Since the term codomain is somewhat uncommon, people often use the term range instead of codomain either intentionally (for simplicity) or unintentionally (due to confusion).

Definition 1.5.1. *Formally, a **function** $f: X \rightarrow Y$ from X to Y is defined by a subset $F \subset X \times Y$ such that $A_x = \{y \in Y | (x, y) \in F\}$ has exactly one element for each $x \in X$. The **value** of f at $x \in X$, denoted $f(x)$, is the unique element of Y contained in A_x .*

Two functions are said to be equal if they have the same domain, codomain, and value for all elements of the domain. A function f is called:

1. **one-to-one** or **injective** if, for all $x, x' \in X$, if $f(x) = f(x')$ then $x = x'$;
2. **onto** or **surjective** if its range $\{f(x) | x \in X\}$ equals Y ;
3. a **one-to-one correspondence** or **bijective** if it is both one-to-one and onto.

A bijective function $f: X \rightarrow Y$ has a unique **inverse function** $f^{-1}: Y \rightarrow X$ such that $f^{-1}(f(x)) = x$ for all $x \in X$ and $f(f^{-1}(y)) = y$ for all $y \in Y$. In fact, any one-to-one function $f: X \rightarrow Y$ can be transformed into a bijective function $g: X \rightarrow R$ with $g(x) = f(x)$ by restricting its codomain Y to its range R .

Functions can also be applied to sets in a natural way. For a function $f: X \rightarrow Y$ and subset $A \subseteq X$, the **image** of A under f is

$$f(A) \triangleq \{y \in Y | \exists x \in A \text{ s.t. } f(x) = y\} = \{f(x) | x \in A\}.$$

Using this definition, we see that the range of f is simply $f(X)$. One benefit of allowing functions to have set-valued images is that a set-valued inverse function always exists. The **inverse image** or **preimage** of a subset $B \subseteq Y$ is

$$f^{-1}(B) \triangleq \{x \in X | f(x) \in B\}.$$

For a one-to-one function f , the inverse image of any singleton set $\{f(x)\}$ is the singleton set $\{x\}$. It is worth noting that the notation $f^{-1}(B)$ for the preimage of B can be somewhat misleading because, in some cases, $f^{-1}(f(A)) \neq A$. In general, a function gives rise to the following property, $f(f^{-1}(B)) \subseteq B$ and $f^{-1}(f(A)) \supseteq A$.

Example 1.5.2. Let the function $f: \mathbb{R} \rightarrow \mathbb{R}$ be defined by $f(x) = x^2$. Let $A = [1, 2]$ and notice that $B = f(A) = [1, 4]$. Then,

$$f^{-1}(B) = f^{-1}([1, 4]) = [-2, -1] \cup [1, 2] \supseteq A.$$

Example 1.5.3. Let the function $f: \mathbb{R} \rightarrow \mathbb{R}$ be defined by $f(x) = x^2 + 1$. Let $B = [0, 2]$ and notice that $A = f^{-1}(B) = [-1, 1]$. Then,

$$f(A) = f([-1, 1]) = [1, 2] \subseteq B.$$

Problem 1.5.4. For all $f: X \rightarrow Y$, $A \subseteq X$, and $B \subseteq Y$, we have the rules:

$$\begin{aligned} (a) \quad x \in A &\Rightarrow f(x) \in f(A) & (b) \quad y \in f(A) &\Rightarrow \exists x \in A \text{ s.t. } f(x) = y \\ (c) \quad x \in f^{-1}(B) &\Rightarrow f(x) \in B & (d) \quad f(x) \in B &\Rightarrow x \in f^{-1}(B). \end{aligned}$$

Use these rules to show that $f^{-1}(f(A)) \supseteq A$ and $f(f^{-1}(B)) \subseteq B$.

Solution 1.5.4. The first result follows from

$$x \in A \stackrel{(a)}{\Rightarrow} f(x) \in f(A) \stackrel{(d)}{\Rightarrow} x \in f^{-1}(f(A)),$$

and the definition of subset. The second result follows from

$$y \in f(f^{-1}(B)) \stackrel{(b)}{\Rightarrow} \exists x \in f^{-1}(B) \text{ s.t. } f(x) = y \stackrel{(c)}{\Rightarrow} y \in B,$$

and the definition of subset.

Problem 1.5.5. Let $f: X \rightarrow Y$, $A_i \subseteq X$ for all $i \in I$, and $B_i \subseteq Y$ for all $i \in I$. Show that the following expressions hold:

$$\begin{aligned} (1) \quad f\left(\bigcup_{i \in I} A_i\right) &= \bigcup_{i \in I} f(A_i) & (2) \quad f\left(\bigcap_{i \in I} A_i\right) &\subseteq \bigcap_{i \in I} f(A_i) \\ (3) \quad f^{-1}\left(\bigcup_{i \in I} B_i\right) &= \bigcup_{i \in I} f^{-1}(B_i) & (4) \quad f^{-1}\left(\bigcap_{i \in I} B_i\right) &= \bigcap_{i \in I} f^{-1}(B_i). \end{aligned}$$

Chapter 2

Metric Spaces and Topology

We initiate this chapter with a brief discussion of key mathematical concepts and motivate how they can be generalized to address important challenges. We assume that the intended reader has experience with the real numbers, convergent sequences, and continuous functions. Specifically, the reader should be familiar with the following notions.

- **Convergence:** Suppose x_1, x_2, \dots is a sequence of real numbers. This sequence converges to a point x if, for any $\epsilon > 0$, there exists a number N such that $|x_n - x| < \epsilon$ for all $n > N$.
- **Continuity:** Suppose $f : \mathbb{R} \mapsto \mathbb{R}$ is a real-valued function over the real numbers. This function is continuous at x_0 if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that $|f(x) - f(x_0)| < \epsilon$ for all $x \in \mathbb{R}$ satisfying $|x - x_0| < \delta$.

One scenario where convergence plays an important role is in the design of iterative methods applied to optimization. Therein, convergence properties can ensure that the iterative process leads to a well-defined answer. Likewise, continuity is a cornerstone of calculus, differential equations, and probability. These related concepts give rise to many practical mathematical tools like gradient descent and Newton's method.

The definitions above are empowering when dealing with the real numbers. Yet, we are interested in exploring spaces that contain objects such as vectors, time series, images, arrays of data, polynomials, matrices, and functions. This brings up an important question: How can we extend the notions of convergence and continuity

to more general spaces? Answering this question is key in better understanding the structure of abstract spaces and, also, in designing algorithms and iterative methods attuned to a rich class of problems.

A first step in developing a more powerful theory is to identify what makes these definitions work for the real numbers. Based on the insights we gain from familiar examples, we can then extract key attributes and build intuition on how we can define similar concepts for abstract spaces. We begin with pertinent questions. First off, what are the real numbers and why do we use them all the time, as opposed to, say, the rational numbers? Second, in the definition of convergence above, we need to know the limit of the sequence to show convergence. However, when we design an iterative algorithm, we most likely do not know its limit; else we would not need to use an iteration process to get there. Is there a notion of convergent sequence for which we do not need to possess an explicit characterization of its limit beforehand? Interestingly, both definitions rely on proximity, respectively $|x_n - x|$ and $|f(x) - f(x_0)|$. We have a pretty good grasp of the distance between two points on the real line. How can we define the distance between two points in more general setting? These are the questions we seek to address below, in the hope of leveraging our familiarity with real numbers into understanding more general spaces.

2.1 Metric Spaces

From an applied perspective, the quintessential way to construct a topology on a space is to define the open sets in terms of a metric. This approach underlies our intuitive understanding of open and closed sets on the real line. Generally speaking, a metric captures the notion of a distance between two elements of a set. Topologies that are defined through metrics possess a number of properties that make them suitable for analysis. Identifying these common properties permits the unified treatment of different spaces that are useful in solving practical problems. To gain better insight into metric spaces, we review the notion of a metric and we introduce a formal definition for topology.

A **metric space** is a set with a well-defined *distance* between any two elements. In some sense, such a space abstracts a few basic properties of Euclidean space. Formally, a metric space (X, d) is a set X and a function d called a metric. The function $d(\cdot, \cdot)$ must fulfill the following properties.

Definition 2.1.1. A *metric* on a set X is a function

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

that satisfies the following properties,

1. $d(x, y) \geq 0 \quad \forall x, y \in X$; equality holds if and only if $x = y$
2. $d(x, y) = d(y, x) \quad \forall x, y \in X$
3. $d(x, y) + d(y, z) \geq d(x, z) \quad \forall x, y, z \in X$.

Example 2.1.2. The collection of real numbers equipped with the metric of absolute distance, $d(x, y) = |x - y|$, defines the standard metric space for the real numbers \mathbb{R} .

Example 2.1.3. The set of vectors in \mathbb{R}^n can also be endowed with a metric. Suppose $\underline{x} = (x_1, \dots, x_n)$ and $\underline{y} = (y_1, \dots, y_n)$ are elements in \mathbb{R}^n . The **Euclidean metric** d on \mathbb{R}^n is defined by

$$d(\underline{x}, \underline{y}) = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2}.$$

As implied by its name, the function $d(\cdot, \cdot)$ given above possesses all the properties of a metric.

It may be beneficial to also look at metrics that are less common. The following two problems introduce alternate distances between points.

Problem 2.1.4. Let $\underline{x} = (x_1, \dots, x_n), \underline{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ and consider the function ρ given by

$$\rho(\underline{x}, \underline{y}) = \max \{|x_1 - y_1|, \dots, |x_n - y_n|\}.$$

Show that ρ is a metric.

Problem 2.1.5. Let X be a metric space with metric d . Define $\bar{d}: X \times X \rightarrow \mathbb{R}$ by

$$\bar{d}(x, y) = \min \{d(x, y), 1\}.$$

Show that \bar{d} is also a metric.

2.1.1 Convergence

Let (X, d) be a metric space. Then, elements of X are called **points** and the number $d(x, y)$ is called the **distance** between x and y . Let $\epsilon > 0$ and consider the set $B_d(x, \epsilon) = \{y \in X \mid d(x, y) < \epsilon\}$. This set is called the **d -open ball** (or open ball) of radius ϵ centered at x .

Problem 2.1.6. Suppose $a \in B_d(x, \epsilon)$ with $\epsilon > 0$. Show that there exists a d -open ball centered at a of radius δ , say $B_d(a, \delta)$, that is contained in $B_d(x, \epsilon)$.

One of the main benefits of having a metric is that it provides some notion of “closeness” between points in a set. This allows one to discuss limits, convergence, open sets, and closed sets.

Definition 2.1.7. A **sequence** of elements from a set X is an infinite list x_1, x_2, \dots where $x_i \in X$ for all $i \in \mathbb{N}$. Formally, a sequence is equivalent to a function $f: \mathbb{N} \rightarrow X$ where $x_i = f(i)$ for all $i \in \mathbb{N}$.

Definition 2.1.8. Consider a sequence x_1, x_2, \dots of points in a metric space (X, d) . We say that x_n **converges** to $x \in X$ (denoted by $x_n \rightarrow x$) if, for any $\epsilon > 0$, there is natural number N such that $d(x, x_n) < \epsilon$ for all $n > N$.

Problem 2.1.9. For a sequence x_n , show that $x_n \rightarrow a$ and $x_n \rightarrow b$ implies $a = b$.

Definition 2.1.10. A sequence x_1, x_2, \dots in (X, d) is a **Cauchy sequence** if, for any $\epsilon > 0$, there is a natural number N (depending on ϵ) such that, for all $m, n > N$,

$$d(x_m, x_n) < \epsilon.$$

Theorem 2.1.11. Every convergent sequence is a Cauchy sequence.

Proof. Since x_1, x_2, \dots converges to some x , there is an N , for any $\epsilon > 0$, such that $d(x, x_n) < \epsilon/2$ for all $n > N$. The triangle inequality for $d(x_m, x_n)$ shows that, for all $m, n > N$,

$$d(x_m, x_n) \leq d(x_m, x) + d(x, x_n) \leq \epsilon/2 + \epsilon/2 = \epsilon.$$

Therefore, x_1, x_2, \dots is a Cauchy sequence. □

Example 2.1.12. Let (X, d) be the metric space of rational numbers defined by $X = \mathbb{Q}$ and $d(x, y) = |x - y|$. The sequence $x_1 = 2$, $x_{n+1} = \frac{1}{2}x_n + \frac{1}{x_n}$ satisfies $x_n \in \mathbb{Q}$ and, using $x_{n+1} - \sqrt{2} = \frac{1}{2x_n}(x_n - \sqrt{2})^2$, one can show it is Cauchy. But, it does not converge in (X, d) because its limit point is the irrational number $\sqrt{2} \notin \mathbb{Q}$.

2.1.2 Metric Topology

Definition 2.1.13. Let W be a subset of a metric space (X, d) . The set W is called **open** if, for every $w \in W$, there is an $\epsilon > 0$ such that $B_d(w, \epsilon) \subseteq W$.

Theorem 2.1.14. For any metric space (X, d) ,

1. \emptyset and X are open
2. any union of open sets is open
3. any finite intersection of open sets is open

Proof. This proof is left as an exercise for the reader. □

One might be curious why only finite intersections are allowed in Theorem 2.1.14. The following example highlights the problem with allowing infinite intersections.

Example 2.1.15. Let $I_n = \left(-\frac{1}{n}, \frac{1}{n}\right) \subset \mathbb{R}$, for $n \in \mathbb{N}$, be a sequence of open real intervals. The infinite intersection

$$\bigcap_{n \in \mathbb{N}} I_n = \{x \in \mathbb{R} \mid \forall n \in \mathbb{N}, x \in I_n\} = \{0\}.$$

But, it is easy to verify that $\{0\}$ is not an open set.

Definition 2.1.16. A subset W of a metric space (X, d) is **closed** if its complement $W^c = X - W$ is open.

Corollary 2.1.17. For any metric space (X, d) ,

1. \emptyset and X are closed
2. any intersection of closed sets is closed
3. any finite union of closed sets is closed

Sketch of proof. Using the definition of closed, one can apply De Morgan's Laws to Theorem 2.1.14 verify this result. □

Actually, the sets \emptyset and X are both open and closed. Such sets are called **clopen**. For a non-trivial example, consider the standard metric space of rational numbers and choose $W = \{x \in \mathbb{Q} \mid x < \sqrt{2}\}$. This set is open because, for all $x \in W$, we have $B(x, \sqrt{2} - x) \subseteq W$. Since $\sqrt{2} \notin \mathbb{Q}$, it follows that $U = \{x \in \mathbb{Q} \mid x \geq \sqrt{2}\} = \{x \in \mathbb{Q} \mid x > \sqrt{2}\}$ which is open by the same argument. But $U^c = W$, so W is also closed.

Definition 2.1.18. For any metric space (X, d) and subset $W \subseteq X$, a point $x \in X$ is a **limit point** of W if there is a sequence $w_1, w_2, \dots \in W \setminus \{x\}$ of distinct elements that converges to x . Equivalently, x is a limit point of W if, for all $\delta > 0$, the set $\{w \in W \mid d(x, w) < \delta\}$ contains some point besides x .

Problem 2.1.19. Prove that the two definitions of a limit point are equivalent.

By this definition (which is standard), an **isolated point** $w \in W$ is not a limit point because there is no sequence of distinct elements that converges to it. Instead, for any sequence that converges to an isolated point w , there must be an $N \in \mathbb{N}$ such that $w_n = w$ for all $n > N$.

Theorem 2.1.20. For any metric space (X, d) , a subset $W \subseteq X$ is closed if and only if it contains all of its limit points.

Proof. Assume W is closed and suppose $x \notin W$ (i.e., $x \in W^c$) is a limit point of W . Since W^c is open, there is a $\delta > 0$ such that $B(x, \delta) \subseteq W^c$ and no sequence in W can approach x . This contradicts the supposition that $x \notin W$ is a limit point and implies all limit points must be in W . On the other hand, if W contains all its limit points, then no $x \notin W$ (i.e., $x \in W^c$) can be a limit point. Negating the second definition of a limit point and applying it to any $x \in W^c$, we see that there is a $\delta > 0$ such that $\{w \in W \mid d(x, w) < \delta\}$ is empty. Thus, $B_d(x, \delta) \subseteq W^c$ and this implies that W^c is open. Thus, W is closed. \square

It follows that, for a closed subset $W \subseteq X$, any sequence $w_n \in W$ that converges must converge to a point $w \in W$.

Definition 2.1.21. For any metric space (X, d) and subset $W \subseteq X$, a point $w \in W$ is in the **interior** of W if and only if there is a $\delta > 0$ such that, for all $x \in X$ with $d(x, w) < \delta$, it follows that $x \in W$.

Problem 2.1.22. Prove that the interior of a set is open.

Definition 2.1.23. For any metric space (X, d) and subset $W \subseteq X$, a point $x \in X$ is in the **closure** of W if and only if, for all $\delta > 0$, there is a $w \in W$ such that $d(x, w) < \delta$.

Problem 2.1.24. Prove that the closure of a set contains all its limit points and, thus, is closed.

The interior of A is denoted by A° and the closure of A is denoted by \bar{A} . The **boundary** ∂A of a set A is defined by $\partial A \triangleq \bar{A} \setminus A^\circ$.

2.1.3 Continuity

Let $f: X \rightarrow Y$ be a function between the metric spaces (X, d_X) and (Y, d_Y) .

Definition 2.1.25. The function f is **continuous** at x_0 if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that, for all $x \in X$ satisfying $d_X(x_0, x) < \delta$, we have

$$d_Y(f(x_0), f(x)) < \epsilon.$$

In precise mathematical notation, one has

$$(\forall \epsilon > 0)(\exists \delta > 0)(\forall x \in \{x' \in X \mid d_X(x_0, x') < \delta\}), d_Y(f(x_0), f(x)) < \epsilon.$$

Theorem 2.1.26. If f is continuous at x_0 , then $f(x_n) \rightarrow f(x_0)$ for all sequences $x_1, x_2, \dots \in X$ such that $x_n \rightarrow x_0$. Conversely, if $f(x_n) \rightarrow f(x_0)$ for all sequences $x_1, x_2, \dots \in X$ such that $x_n \rightarrow x_0$, then f is continuous at x_0 .

Proof. If f is continuous at x_0 , then, for any $\epsilon > 0$, there is a $\delta > 0$ such that $d_Y(f(x_0), f(x)) < \epsilon$ if $d_X(x_0, x) < \delta$. If $x_n \rightarrow x_0$, then there is an $N \in \mathbb{N}$ such that $d_X(x_n, x_0) < \delta$ for all $n > N$. Thus, $d_Y(f(x_0), f(x_n)) < \epsilon$ for all $n > N$ and $f(x_n) \rightarrow f(x_0)$.

For the converse, we show the contrapositive. If f is not continuous at x_0 , then there exists an $\epsilon > 0$ such that, for all $\delta > 0$, there is an $x \in X$ with $d_X(x_0, x) < \delta$ and $d_Y(f(x_0), f(x)) \geq \epsilon$. For this ϵ and any positive sequence $\delta_n \rightarrow 0$, let x_n be the promised x . Then, $x_n \rightarrow x_0$ because $d_X(x_0, x_n) < \delta_n \rightarrow 0$ but $d_Y(f(x_0), f(x_n)) \geq \epsilon$. Thus, $f(x_n)$ does not converge to $f(x_0)$ for some sequence where $x_n \rightarrow x_0$. \square

Definition 2.1.27. The **limit of f at x_0** , $\lim_{x \rightarrow x_0} f(x)$, exists and equals $y \in Y$ if $f(x_n) \rightarrow y$ for all sequences $x_n \in X$ such that $x_n \rightarrow x_0$. Thus, Theorem 2.1.26 implies that $\lim_{x \rightarrow x_0} f(x) = f(x_0)$ if and only if f is continuous at x_0 .

Remark 2.1.28. Comparing this with the definition of continuous, we see that $\lim_{x \rightarrow x_0} f(x) = y$ can be defined equivalently as: for any $\epsilon > 0$, there exists a $\delta > 0$ such that, for all $x \in X$ satisfying $d_X(x_0, x) < \delta$, we have

$$d_Y(y, f(x)) < \epsilon.$$

Definition 2.1.29. The function f is called **continuous** if, for all $x_0 \in X$, it is continuous at x_0 . In precise mathematical notation, one has

$$(\forall x_0 \in X)(\forall \epsilon > 0)(\exists \delta > 0) \\ (\forall x \in \{x' \in X \mid d_X(x_0, x') < \delta\}), d_Y(f(x_0), f(x)) < \epsilon.$$

Definition 2.1.30. The function f is called **uniformly continuous** if it is continuous and, for all $\epsilon > 0$, the $\delta > 0$ can be chosen independently of x_0 . In precise mathematical notation, one has

$$(\forall \epsilon > 0)(\exists \delta > 0)(\forall x_0 \in X) \\ (\forall x \in \{x' \in X \mid d_X(x_0, x') < \delta\}), d_Y(f(x_0), f(x)) < \epsilon.$$

Definition 2.1.31. A function $f: X \rightarrow Y$ is called **Lipschitz continuous** on $A \subseteq X$ if there is a constant $L \in \mathbb{R}$ such that $d_Y(f(x), f(y)) \leq Ld_X(x, y)$ for all $x, y \in A$.

Let f_A denote the **restriction** of f to $A \subseteq X$ defined by $f_A: A \rightarrow Y$ with $f_A(x) = f(x)$ for all $x \in A$. It is easy to verify that, if f is Lipschitz continuous on A , then f_A is uniformly continuous.

Problem 2.1.32. Let (X, d) be a metric space and define $f: X \rightarrow \mathbb{R}$ by $f(x) = d(x, x_0)$ for some fixed $x_0 \in X$. Show that f is Lipschitz continuous with $L = 1$.

2.1.4 Properties of the Real Numbers

In this section, we review some important properties of the real numbers. To proceed, we will use some facts now that are only proven later. For example, we will use the fact that “every Cauchy sequence of real numbers converges to a real number”. This property, known as completeness, is the focus on Section 2.1.6 and can be

established by formally constructing the real numbers as limit points of sequences of rational numbers.

To discuss notions of extreme values for sets of real numbers, we also define the **extended real numbers** $\overline{\mathbb{R}}$ by augmenting the real numbers to include limit points for unbounded sequences $\overline{\mathbb{R}} \triangleq \mathbb{R} \cup \{\infty, -\infty\}$. Using the metric $d_{\overline{\mathbb{R}}}(x, y) \triangleq \left| \frac{x}{1+|x|} - \frac{y}{1+|y|} \right|$ with $\frac{x}{1+|x|} \Big|_{x=\pm\infty} \triangleq \pm 1$, this defines the standard metric space of extended real numbers. Later, we will see that this is actually a compact metric space. The main difference from \mathbb{R} is that, for $x_n \in \overline{\mathbb{R}}$, the statement $x_n \rightarrow \infty$ is well defined and equivalent to $\forall M > 0, \exists N \in \mathbb{N}, \forall n > N, x_n > M$.

Definition 2.1.33. The *supremum* (or *least upper bound*) of $X \subseteq \mathbb{R}$, denoted $\sup X$, is the smallest extended real number $M \in \overline{\mathbb{R}}$ such that $x \leq M$ for all $x \in X$. It is always well-defined and equals $-\infty$ if $X = \emptyset$.

Definition 2.1.34. The *maximum* of $X \subseteq \mathbb{R}$, denoted $\max X$, is the largest value achieved by the set. It equals $\sup X$ if $\sup X \in X$ and is undefined otherwise.

Definition 2.1.35. The *infimum* (or *greatest lower bound*) of $X \subseteq \mathbb{R}$, denoted $\inf X$, is the largest extended real number $m \in \overline{\mathbb{R}}$ such that $x \geq m$ for all $x \in X$. It is always well-defined and equals ∞ if $X = \emptyset$.

Definition 2.1.36. The *minimum* of $X \subseteq \mathbb{R}$, denoted $\min X$, is the smallest value achieved by the set. It equals $\inf X$ if $\inf X \in X$ and is undefined otherwise.

Lemma 2.1.37. Let X be a set and $f: X \rightarrow \mathbb{R}$ be a function from X to the real numbers. Let $M = \sup f(A)$ for some non-empty $A \subseteq X$. Then, there exists a sequence $x_1, x_2, \dots \in A$ such that $\lim_n f(x_n) = M$.

Proof. If $M = \infty$, then $f(A)$ has no finite upper bound and, for any $y \in \mathbb{R}$, there exists an $x \in A$ such that $f(x) > y$. In this case, we can let x_1 be any element of A and x_{n+1} be any element of A such that $f(x_{n+1}) > f(x_n) + 1$. In the metric space $(\overline{\mathbb{R}}, d_{\overline{\mathbb{R}}})$, this implies that $d_{\overline{\mathbb{R}}}(f(x_n), \infty) = \left| \frac{f(x_n)}{1+|f(x_n)|} - 1 \right| \rightarrow 0$ and thus $f(x_n) \rightarrow \infty$.

If $M < \infty$, then $f(A)$ has a finite upper bound and, for any $\epsilon > 0$, there is an x such that $M - f(x) < \epsilon$. Otherwise, one arrives at the contradiction $\sup f(A) < M$. Therefore, we can construct the sequence x_1, x_2, \dots by choosing $x_n \in A$ to be any point that satisfies $M - f(x_n) \leq \frac{1}{n}$. \square

Theorem 2.1.38. *In the standard metric space of real numbers, any non-decreasing sequence with a finite upper bound converges to its supremum.*

Proof. Let $x_1, x_2, \dots \in \mathbb{R}$ be a sequence satisfying $x_{n+1} \geq x_n$ (i.e., non-decreasing) and $x_n \leq M < \infty$ for all $n \in \mathbb{N}$ (i.e., with a finite upper bound). Without loss of generality, we can choose the upper bound M to be the supremum $\sup\{x_1, x_2, \dots\}$. Now, we will prove directly that $x_n \rightarrow M$.

First, we note that the definition of the supremum implies that $x_n \leq M$ for all $n \in \mathbb{N}$ and, for any $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $x_N > M - \epsilon$. Since x_n is non-decreasing, this implies that $x_n > M - \epsilon$ for all $n > N$. Next, it follows from $x_n \leq M$ that $|M - x_n| = M - x_n < \epsilon$ for all $n > N$. Thus, the constructed N satisfies all elements in the definition of $x_n \rightarrow M$. \square

Lemma 2.1.39. *Let $y_n \in \mathbb{R}$ be a real sequence. If $\sum_{i=1}^{\infty} |y_i| = M < \infty$, then $x_n = \sum_{i=1}^n y_i$ satisfies $x_n \rightarrow x$ with $|x| < M$.*

Proof. Let $w_n = \sum_{i=1}^n |y_i|$ and observe that the following inequality holds,

$$|x_m - x_n| = \left| \sum_{i=n+1}^m y_i \right| \leq \sum_{i=n+1}^m |y_i| = |w_m - w_n|.$$

Since w_n converges, it is Cauchy and the inequality implies that x_m is Cauchy. Thus, x_n converges to some $x \in \mathbb{R}$ and $|x| < M$ follows from $|x_n| \leq w_n \leq M$. \square

Lemma 2.1.40. *Let $y_n \in \mathbb{R}$ be a real sequence and $x_n = \sum_{i=1}^n y_i$ be its sequence of partial sums. Then, $\sum_{i=1}^{\infty} y_i \triangleq \lim_{n \rightarrow \infty} x_n$ exists if and only if the tail of the sum is negligible:*

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall m, n > N, \left| \sum_{i=n+1}^m y_i \right| < \epsilon.$$

Proof. In a metric space, convergent sequences must be Cauchy. For the real numbers, all Cauchy sequences converge. Thus, $\lim_{n \rightarrow \infty} x_n$ exists if and only if x_n is a Cauchy sequence. Thus, x_n converges if and only if " $\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall m, n > N, |x_m - x_n| < \epsilon$ ". Thus, the result follows from the fact that $|x_m - x_n| = \left| \sum_{i=n+1}^m y_i \right|$. \square

2.1.5 Sequences of Functions

Let (X, d_X) and (Y, d_Y) be metric spaces and $f_n: X \rightarrow Y$ for $n \in \mathbb{N}$ be a sequence of functions mapping X to Y .

Definition 2.1.41. *The sequence f_n converges pointwise to $f: X \rightarrow Y$ if*

$$\lim_{n \rightarrow \infty} f_n(x) = f(x)$$

for all $x \in X$. Using mathematical symbols, we can write

$$\forall x \in X, \forall \epsilon > 0, \exists N \in \mathbb{N}, \forall n \in \{n' \in \mathbb{N} \mid n' > N\}, d_Y(f_{n'}(x), f(x)) < \epsilon.$$

Definition 2.1.42. *The sequence f_n converges uniformly to $f: X \rightarrow Y$ if*

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall n \in \{n' \in \mathbb{N} \mid n' > N\}, \forall x \in X, d_Y(f_{n'}(x), f(x)) < \epsilon.$$

This condition is also equivalent to

$$\lim_{n \rightarrow \infty} \sup_{x \in X} d_Y(f_n(x), f(x)) = 0.$$

Theorem 2.1.43. *If each f_n is continuous and f_n converges uniformly to $f: X \rightarrow Y$, then f is continuous.*

Proof. The goal is to show that, for all $x \in X$ and any $\epsilon > 0$, there is a $\delta > 0$ such that $d_Y(f(x), f(y)) < \epsilon$ if $d_X(x, y) < \delta$. Since $f_n \rightarrow f$ uniformly, for any $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $d_Y(f_n(x), f(x)) < \epsilon/3$ for all $n > N$ and all $x \in X$. Now, we can fix $\epsilon > 0$ use the N promised above. Then, for any $n > N$, the continuity of f_n implies that, for all $x \in X$ and any $\epsilon > 0$, there is a $\delta > 0$ such that $d_Y(f_n(x), f_n(y)) < \epsilon/3$ if $d_X(x, y) < \delta$. Thus, if $d_X(x, y) < \delta$, then

$$\begin{aligned} d_Y(f(x), f(y)) &\leq d_Y(f(x), f_n(x)) + d_Y(f_n(x), f_n(y)) + d_Y(f_n(y), f(y)) \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \end{aligned}$$

□

2.1.6 Completeness

Suppose (X, d) is a metric space. From Definition 2.1.8, we know that a sequence x_1, x_2, \dots of points in X converges to $x \in X$ if, for every $\delta > 0$, there exists an integer N such that $d(x_i, x) < \delta$ for all $i \geq N$.

It is possible for a sequence in a metric space X to satisfy the Cauchy criterion, but not to converge in X .

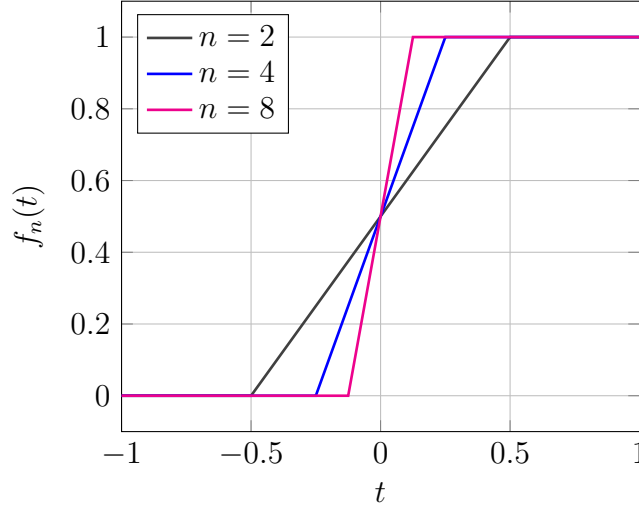


Figure 2.1: The sequence of continuous functions in Example 2.1.44 satisfies the Cauchy criterion. But, it does not converge to a continuous function in $C[-1, 1]$.

Example 2.1.44. Let $X = C[-1, 1]$ be the space of continuous functions that map $[-1, 1]$ to \mathbb{R} and satisfy $\|f\|_2 < \infty$, where $\|f\|_2$ denotes the L^2 norm

$$\|f\|_2 \triangleq \left(\int_{-1}^1 |f(t)|^2 dt \right)^{\frac{1}{2}}.$$

This set forms a metric space (X, d) when equipped with the distance

$$d(f, g) \triangleq \|f - g\|_2 = \left(\int_{-1}^1 |f(t) - g(t)|^2 dt \right)^{\frac{1}{2}}.$$

Consider the sequence of functions $f_n(t)$ given by

$$f_n(t) \triangleq \begin{cases} 0 & t \in [-1, -\frac{1}{n}] \\ \frac{nt}{2} + \frac{1}{2} & t \in (-\frac{1}{n}, \frac{1}{n}) \\ 1 & t \in [\frac{1}{n}, 1]. \end{cases}$$

Assuming that $m \geq n$, a bit of calculus shows that

$$d(f_n, f_m) = \|f_n(t) - f_m(t)\|_2 = \left(\int_{-1}^1 |f_n(t) - f_m(t)|^2 dt \right)^{\frac{1}{2}} = \frac{(m-n)^2}{6m^2n}.$$

Since $m \geq n$, this distance is upper bounded by $\frac{1}{6n}$ and the sequence satisfies the Cauchy criterion. But, it does not converge to a continuous function in $C[-1, 1]$.

Definition 2.1.45. A metric space (X, d) is said to be **complete** if every Cauchy sequence in (X, d) converges to a limit $x \in X$.

The standard metric space of real numbers with absolute distance is a complete metric space. This fact and other foundational properties of the real numbers can be derived formally using the techniques described below. However, a formal construction of the real numbers will not be provided in these notes.

Example 2.1.46. Consider the sequence $x_1 = 2$, $x_{n+1} = \frac{1}{2}x_n + \frac{1}{x_n}$ and observe that $x_n \in \mathbb{Q}$ for all $n \in \mathbb{N}$. We saw earlier that x_n is a Cauchy sequence with limit point $\sqrt{2} \in \mathbb{R}$. But, $\sqrt{2}$ is irrational and thus the rational numbers \mathbb{Q} are not complete.

Theorem 2.1.47. A closed subset A of a complete metric space X is itself a complete metric space.

Proof. Any Cauchy sequence $x_1, x_2, \dots \in A$ is also a Cauchy sequence in X . This implies that $x_n \rightarrow x \in X$ and it follows that $x \in \bar{A}$. Since A is closed, $x \in A$. \square

Definition 2.1.48. An **isometry** is a mapping $\phi: X \rightarrow Y$ between two metric spaces (X, d_X) and (Y, d_Y) that is distance preserving (i.e., it satisfies $d_X(x, x') = d_Y(\phi(x), \phi(x'))$ for all $x, x' \in X$).

Definition 2.1.49. A subset A of a metric space (X, d) is **dense** in X if every $x \in X$ is a limit point of the set A . This is equivalent to its closure \bar{A} being equal to X .

Definition 2.1.50. The **completion** of a metric space (X, d_X) consists of a complete metric space (Y, d_Y) and an isometry $\phi: X \rightarrow Y$ such that $\phi(X)$ is a dense subset of Y . Moreover, the completion is unique up to isometry.

Example 2.1.51. Consider the metric space \mathbb{Q} of rational numbers equipped with the metric of absolute distance. The completion of this metric space is \mathbb{R} because the isometry is given by the identity mapping and \mathbb{Q} is a dense subset of \mathbb{R} .

Cauchy sequences have many applications in analysis and signal processing. For example, they can be used to construct the real numbers from the rational numbers. In fact, the same approach is used to construct the completion of any metric space.

Definition 2.1.52. Two Cauchy sequences x_1, x_2, \dots and y_1, y_2, \dots are equivalent if, for every $\epsilon > 0$, there exists an integer N such that $d(x_k, y_k) \leq \epsilon$ for all $k \geq N$.

Example 2.1.53. Let $\mathcal{C}(\mathbb{Q})$ denote the set of all Cauchy sequences q_1, q_2, \dots of rational numbers where \sim represents the equivalence relation on this set defined above. Then, the set of equivalence classes (or quotient set) $\mathcal{C}(\mathbb{Q})/\sim$ is in one-to-one correspondence with the real numbers. This construction is the standard completion of \mathbb{Q} . Since every Cauchy sequence of rationals converges to a real number, the isometry is given by mapping each equivalence class to its limit point in \mathbb{R} .

Definition 2.1.54. Let A be a subset of a metric space (X, d) and $f: X \rightarrow X$ be a function. Then, f is a **contraction** on A if $f(A) \subseteq A$ and there exists a constant $\gamma < 1$ such that $d(f(x), f(y)) \leq \gamma d(x, y)$ for all $x, y \in A$.

Consider the following important results in applied mathematics: Picard's uniqueness theorem for differential equations, the implicit function theorem, and the existence of stationary optimal policies for Markov decision processes. What do they have in common? They each establish the existence and uniqueness of a function and have relatively simple proofs based on the contraction mapping theorem.

Theorem 2.1.55 (Contraction Mapping Theorem). Let (X, d) be a complete metric space and f be contraction on a closed subset $A \subseteq X$. Then, f has a unique fixed point x^* in A such that $f(x^*) = x^*$ and the sequence $x_{n+1} = f(x_n)$ converges to x^* for any point $x_1 \in A$. Moreover, x_n satisfies the error bounds $d(x^*, x_n) \leq \gamma^{n-1}d(x^*, x_1)$ and $d(x^*, x_{n+1}) \leq d(x_n, x_{n+1})\gamma/(1 - \gamma)$.

Proof. Suppose f has two fixed points $y, z \in A$. Then, $d(y, z) = d(f(y), f(z)) \leq \gamma d(y, z)$ and $d(y, z) = 0$ because $\gamma \in [0, 1)$. This shows that $y = z$ and any two fixed points in A must be identical. If x^* is a fixed point, then $d(x^*, x_{n+1}) = d(f(x^*), f(x_n)) \leq \gamma d(x^*, x_n)$ and, by induction, one gets the first error bound.

Since $d(f(x_n), f(x_{n+1})) \leq \gamma d(x_n, x_{n+1})$, induction shows that $d(x_n, x_{n+1}) \leq \gamma^{n-1}d(x_1, x_2)$. Using this, we can bound the distance $d(x_m, x_n)$ (for $m < n$) with

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x_{m+1}) + d(x_{m+1}, x_n) \\ &\leq \sum_{i=m}^{n-1} d(x_i, x_{i+1}) \leq \sum_{i=m}^{n-1} \gamma^{i-1} d(x_1, x_2) \\ &\leq \sum_{i=m}^{\infty} \gamma^{i-1} d(x_1, x_2) \leq \frac{\gamma^{m-1}}{1 - \gamma} d(x_1, x_2). \end{aligned} \quad (2.1)$$

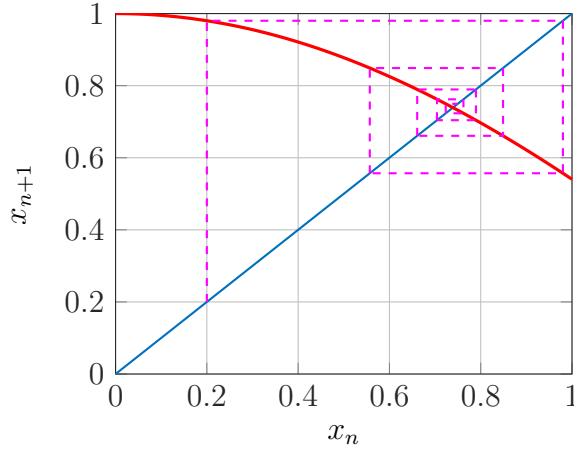


Figure 2.2: Starting from $x_1 = 0.2$, the iteration in Example 2.1.56 maps x_n to $x_{n+1} = \cos(x_n)$. The points are also connected to the slope-1 line to show the path.

The sequence x_n is Cauchy because $d(x_m, x_n)$ can be made arbitrarily small (for all $n > m$) by increasing m . As (X, d) is complete, it follows that $x_n \rightarrow x^*$ for some $x^* \in X$. Since f is Lipschitz continuous, this implies that $x^* = \lim_n x_n = \lim_n f(x_n) = f(x^*)$ the unique fixed point of f in A . Since $x_n \rightarrow x^*$, we can use essentially the same argument to get the second error bound

$$d(x_{n+1}, x^*) \leq \sum_{i=n+1}^{\infty} d(x_i, x_{i+1}) \leq \sum_{i=n+1}^{\infty} \gamma^{i-n} d(x_n, x_{n+1}) = \frac{\gamma}{1-\gamma} d(x_n, x_{n+1}). \quad \square$$

Example 2.1.56. Consider the cosine function restricted to the subset $[0, 1] \subseteq \mathbb{R}$. Since $\cos(x)$ is decreasing for $0 \leq x < \pi$, we have $\cos([0, 1]) = [\cos(1), 1]$ with $\cos(1) \approx 0.54$. The mean value theorem of calculus also tells us that $\cos(y) - \cos(x) = \cos'(t)(y - x)$ for some $t \in [x, y]$. Since $\cos'(t) = -\sin(t)$ and $\sin(t)$ is increasing on $[0, 1]$, we find that $\sin([0, 1]) = [0, \sin(1)]$ with $\sin(1) \approx 0.84$.

Taking the absolute value, shows that $|\cos(y) - \cos(x)| \leq 0.85|y - x|$. Therefore, $\cos(t)$ is a contraction on $[0, 1]$ and the sequence $x_{n+1} = \cos(x_n)$ (e.g., see Figure 2.2) converges to the unique fixed point $x^* = \cos(x^*)$ for all $x_1 \in [0, 1]$.

2.1.7 Compactness

Definition 2.1.57. A metric space (X, d) is **totally bounded** if, for any $\epsilon > 0$, there exists a finite set of $B_d(x, \epsilon)$ balls that cover (i.e., whose union equals) X .

Definition 2.1.58. A metric space is **compact** if it is complete and totally bounded.

The closed interval $[0, 1] \subset \mathbb{R}$ is compact. In fact, a subset of \mathbb{R}^n is compact if and only if it is closed and bounded. On the other hand, the standard metric space of real numbers is not compact because it is not totally bounded.

Theorem 2.1.59. A closed subset A of a compact space X is itself a compact space.

The following theorem highlights one of the main reasons that compact spaces are desirable in practice.

Theorem 2.1.60. Let (X, d) be a compact metric space and $x_1, x_2, \dots \in X$ be a sequence. Then, there is a subsequence x_{n_1}, x_{n_2}, \dots , defined by some increasing sequence $n_1, n_2, \dots \in \mathbb{N}$, that converges.

Proof. We proceed by recursively constructing subsequences $z_n^{(k)}$ starting from $z_n^{(0)} = x_n$. Since X is totally bounded, let $C_k \subset X$ be the centers of a finite set of balls with radius 2^{-k} that cover X (i.e., $\cup_{x \in C_k} B(x, 2^{-k}) = X$). Then, one of these balls (say centered at x') must contain infinitely many elements in $z_n^{(k-1)}$ (i.e., $\exists x' \in C_k, |\{n \in \mathbb{N} \mid z_n^{(k-1)} \in B(x', 2^{-k})\}| = \infty$). Next, we extract the subsequence contained in this ball by choosing $z_n^{(k)}$ to be the subsequence of $z_n^{(k-1)}$ contained in $B(x', 2^{-k})$. From the triangle inequality, it follows that $d(y, y') < 2(2^{-k})$ for all $y, y' \in B(x', 2^{-k})$. Thus, $d(z_m^{(k')}, z_n^{(k)}) < 2^{-k+1}$ for all $m > n \geq 1$ and $k' \geq k \geq 1$.

Let $I(k, n)$ be the index in the original sequence associated with $z_n^{(k)}$. Since each stage only removes elements from the previous subsequence and relabels, it follows that $I(k+1, k+1) \geq I(k, k+1) > I(k, k)$. This implies that the sequence $y_k = z_k^{(k)} = x_{I(k,k)}$ is a subsequence of x_n and $d(y_m, y_k) \leq 2^{-k+1}$ for all $m > k$ and $k \geq 1$. Thus, for any $\epsilon > 0$, choosing $N = \lceil \log_2 \frac{1}{\epsilon} \rceil + 1$ shows that y_k is a Cauchy sequence. Since X is complete, it follows that y_k converges to some $y \in X$. \square

Functions from compact sets to the real numbers are very important in practice.

Theorem 2.1.61. Let X be a metric space and $f: X \rightarrow \mathbb{R}$ be a continuous function from X to the real numbers. If A is a compact subset of X , then there exists $x \in A$ such that $f(x) = \sup f(A)$ (i.e., f achieves a maximum on A).

Proof. Using Lemma 2.1.37, one finds that there is a sequence $x_1, x_2, \dots \in A$ such that $\lim_n f(x_n) = \sup f(A)$. Since A is compact, there must also be a subsequence x_{n_1}, x_{n_2}, \dots that converges. As A is closed, this subsequence must converge to some $x^* \in A$. Finally, the continuity of f shows that

$$\sup f(A) = \lim_n f(x_n) = \lim_k f(x_{n_k}) = f(\lim_k x_{n_k}) = f(x^*). \quad \square$$

Corollary 2.1.62. *Let (X, d) be a metric space. Then, a continuous function from a compact subset $A \subseteq X$ to the real numbers achieves a minimum on A .*

2.2 General Topology*

While topology originated with the study of sets of finite-dimensional real vectors, its mathematical abstraction can also be useful. We note that some of the terms used above, for metric spaces, are redefined below. Fortunately, these new definitions are compatible with the old ones when the topology is generated by a metric.

Definition 2.2.1. *A **topology** on a set X is a collection \mathcal{J} of subsets of X that satisfies the following properties,*

1. \emptyset and X are in \mathcal{J}
2. the union of the elements of any subcollection of \mathcal{J} is in \mathcal{J}
3. the intersection of the elements of any finite subcollection of \mathcal{J} is in \mathcal{J} .

A subset $A \subseteq X$ is called an **open set** of X if $A \in \mathcal{J}$. Using this terminology, a topological space is a set X together with a collection of subsets of X , called *open sets*, such that \emptyset and X are both open and such that arbitrary unions and finite intersections of open sets are open.

Definition 2.2.2. *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 exists a basis element B containing x .
2. if $x \in B_1$ and $x \in B_2$ where $B_1, B_2 \in \mathcal{B}$, then there exists a basis element B_3 containing x such that $B_3 \subseteq B_1 \cap B_2$.

3. a subset $A \subseteq X$ is open in the topology on X generated by \mathcal{B} if and only if, for every $x \in A$, there exists a basis element $B \in \mathcal{B}$ such that $x \in B$ and $B \subseteq A$.

Probably the most important and frequently used way of imposing a topology on a set is to define the topology in terms of a metric.

Example 2.2.3. If d is a metric on the set X , then the collection of all ϵ -balls

$$\{B_d(x, \epsilon) \mid x \in X, \epsilon > 0\}$$

is a basis for a topology on X . This topology is called the **metric topology** induced by d .

Applying the meaning of open set from Definition 2.2.2 to this basis, one finds that a set A is open if and only if, for each $x \in A$, there exists a $\delta > 0$ such that $B_d(x, \delta) \subset A$. Clearly, this condition agrees with the definition of d -open from Definition 2.1.13.

Definition 2.2.4. Let X be a topological space. This space is said to be **metrizable** if there exists a metric d on the set X that induces the topology of X .

We note that definitions and results in Sections 2.1.6 and 2.1.7 for metric spaces actually apply to any metrizable space. For example, a metrizable space is complete if and only if there the metric that induces its topology also defines a complete metric space.

Example 2.2.5. While most of the spaces discussed in these notes are metrizable, there is a very common notion of convergence that is not metrizable. The topology on the set of functions $f: [0, 1] \rightarrow \mathbb{R}$ where the open sets are defined by pointwise convergence is not metrizable.

2.2.1 Closed Sets and Limit Points

Definition 2.2.6. A subset A of a topological space X is **closed** if the set

$$A^c = X - A = \{x \in X \mid x \notin A\}$$

is open.

Note that a set can be open, closed, both, or neither! It can be shown that the collection of closed subsets of a space X has properties similar to those satisfied by the collection of open subsets of X .

Fact 2.2.7. *Let X be a topological space. The following conditions hold,*

1. \emptyset and X are closed
2. arbitrary intersections of closed sets are closed
3. finite unions of closed sets are closed.

Definition 2.2.8. *Given a subset A of a topological space X , the **interior** of A is defined as the union of all open sets contained in A . The **closure** of A is defined as the intersection of all closed sets containing A .*

The interior of A is denoted by A° and the closure of A is denoted by \bar{A} . We note that A° is open and \bar{A} is closed. Furthermore, $A^\circ \subseteq A \subseteq \bar{A}$.

Theorem 2.2.9. *Let A be a subset of the topological space X . The element x is in \bar{A} if and only if every open set B containing x intersects A .*

Proof. We prove instead the equivalent contrapositive statement: $x \notin \bar{A}$ if and only if there is an open set B containing x that does not intersect A . Clearly, if $x \notin \bar{A}$, then $\bar{A}^c = X - \bar{A}$ is an open set containing x that does not intersect A . Conversely, if there is an open set B containing x that does not intersect A , then $B^c = X - B$ is a closed set containing A . The definition of closure implies that B^c must also contain \bar{A} . But $x \notin B^c$, so $x \notin \bar{A}$. \square

Definition 2.2.10. *An open set O containing x is called a **neighborhood** of x .*

Definition 2.2.11. *Suppose A is a subset of the topological space X and let x be an element of X . Then x is a **limit point** of A if every neighborhood of x intersects A in some point other than x itself.*

In other words, $x \in X$ is a limit point of $A \subset X$ if $x \in \overline{A - \{x\}}$, the closure of $A - \{x\}$. The point x may or may not be in A .

Theorem 2.2.12. *A subset of a topological space is closed if and only if it contains all its limit points.*

Definition 2.2.13. A subset A of a topological space X is **dense** in X if every $x \in X$ is a limit point of the set A . This is equivalent to its closure \overline{A} being equal to X .

Definition 2.2.14. A topological space X is **separable** if it contains a countable subset that is dense in X .

Example 2.2.15. Since every real number is a limit point of rational numbers, it follows that \mathbb{Q} is a dense subset of \mathbb{R} . This also implies that \mathbb{R} , the standard metric space of real numbers, is separable.

2.2.2 Continuity

Definition 2.2.16. Let X and Y be topological spaces. A function $f: X \rightarrow Y$ is **continuous** if for each open subset $O \subseteq Y$, the set $f^{-1}(O)$ is an open subset of X .

Recall that $f^{-1}(B)$ is the set $\{x \in X \mid f(x) \in B\}$. Continuity of a function depends not only upon the function f itself, but also on the topologies specified for its domain and range!

Theorem 2.2.17. Let X and Y be topological spaces and consider a function $f: X \rightarrow Y$. The following are equivalent:

1. f is continuous
2. for every subset $A \subseteq X$, $f(\overline{A}) \subseteq \overline{f(A)}$
3. for every closed set $C \subseteq Y$, the set $f^{-1}(C)$ is closed in X .

Proof. (1 \Rightarrow 2). Assume f is a continuous function. We wish to show $f(\overline{A}) \subseteq \overline{f(A)}$ for every subset $A \subseteq X$. To begin, suppose A is fixed and let $y \in f(\overline{A})$. Then, there exists $x \in \overline{A}$ such that $f(x) = y$. Let $O \subseteq Y$ be a neighborhood of $f(x)$. Preimage $f^{-1}(O)$ is an open set containing x because f is continuous. Since $x \in \overline{A} \cap f^{-1}(O)$, we gather that $f^{-1}(O)$ must intersect with A in some point x' . Moreover, $f(x') \in f(f^{-1}(O)) \subseteq O$ and $f(x') \in f(A)$. Thus, O intersects with $f(A)$ in the point $f(x')$. Since O is an arbitrary neighborhood of $f(x)$, we deduce that $f(x) \in \overline{f(A)}$ by Theorem 2.2.9. Collecting these results, we get that any $y \in f(\overline{A})$ is also in $\overline{f(A)}$.

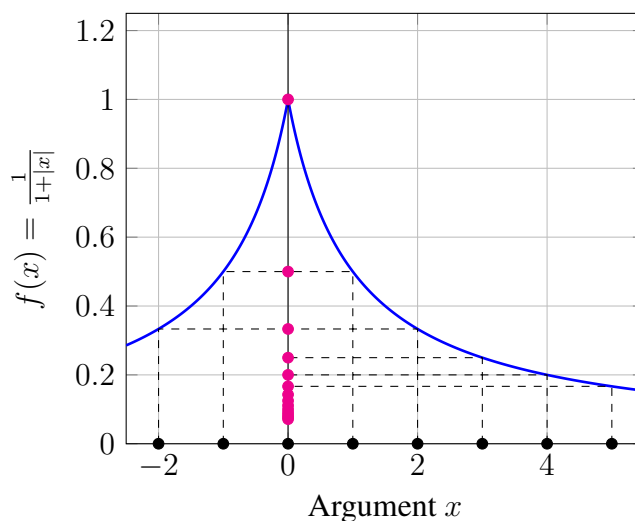


Figure 2.3: The function $f(x) = \frac{1}{1+|x|}$ is continuous. The set of integers \mathbb{Z} is closed. Yet, the image of this set, $f(\mathbb{Z}) = \{1/n : n \in \mathbb{N}\}$, is not closed. Thus, this is an example of a continuous function along with a set for which $f(\overline{\mathbb{Z}}) \subsetneq \overline{f(\mathbb{Z})}$.

(2 \Rightarrow 3). For this step, we assume that $f(\overline{A}) \subseteq \overline{f(A)}$ for every subset $A \subseteq X$. Let $C \subseteq Y$ be a closed set and let $A = f^{-1}(C)$. Then, $f(A) = f(f^{-1}(C)) \subseteq C$. If $x \in \overline{A}$, we get

$$f(x) \in f(\overline{A}) \subseteq \overline{f(A)} \subseteq \overline{C} = C.$$

So that $x \in f^{-1}(C) = A$ and, as a consequence, $\overline{A} \subseteq A$. Thus, $A = \overline{A}$ is closed.

(3 \Rightarrow 1). Let O be an open set in Y . Let $O^c = Y - O$; then O^c is closed in Y . By assumption, $f^{-1}(O^c)$ is closed in X . Using elementary set theory, we have

$$X - f^{-1}(O^c) = \{x \in X | f(x) \notin O^c\} = \{x \in X | f(x) \in O\} = f^{-1}(O).$$

That is, $f^{-1}(O)$ is open. □

Theorem 2.2.18. *Suppose X and Y are two metrizable spaces with metrics d_X and d_Y . Consider a function $f: X \rightarrow Y$. The function f is continuous if and only if it is d -continuous with respect to these metrics.*

Proof. Suppose that f is continuous. For any $x_1 \in X$ and $\epsilon > 0$, let $O_y = B_{d_Y}(f(x_1), \epsilon)$ and consider the set

$$O_x = f^{-1}(O_y)$$

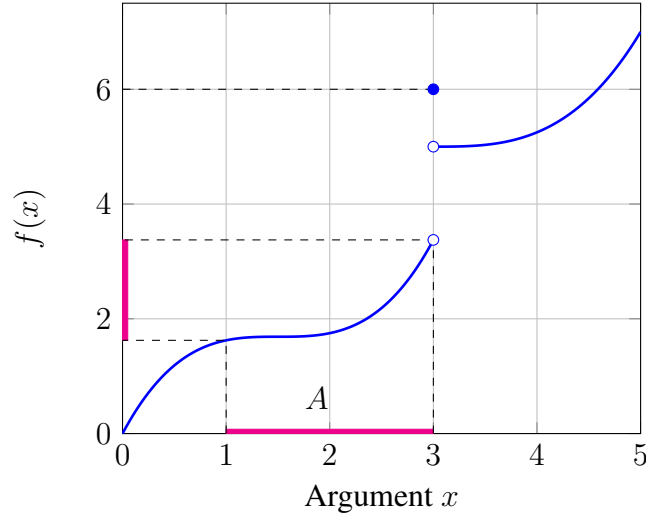


Figure 2.4: Given a function with a discontinuity and a set A , the image of the closure, $f(\overline{A})$, need not be a subset of the closure of the image, $\overline{f(A)}$, as seen in the example above.

which is open in X and contains the point x_1 . Since O_x is open and $x_1 \in O_x$, there exists a d -open ball $B_{d_X}(x_1, \delta)$ of radius $\delta > 0$ centered at x_1 such that $B_{d_X}(x_1, \delta) \subset O_x$. We also see that $f(x_2) \in O_y$ for any $x_2 \in B_{d_X}(x_1, \delta)$ because $A \subseteq O_x$ implies $f(A) \subseteq O_y$. It follows that $d_Y(f(x_1), f(x_2)) < \epsilon$ for all $x_2 \in B_{d_X}(x_1, \delta)$.

Conversely, let O_y be an open set in Y and suppose that the function f is d -continuous with respect to d_X and d_Y . For any $x \in f^{-1}(O_y)$, there exists a d -open ball $B_{d_Y}(f(x), \epsilon)$ of radius $\epsilon > 0$ centered at $f(x)$ that is entirely contained in O_y . By the definition of d -continuous, there exists a d -open ball $B_{d_X}(x, \delta)$ of radius $\delta > 0$ centered at x such that $f(B_{d_X}(x, \delta)) \subset B_{d_Y}(f(x), \epsilon)$. Therefore, every $x \in f^{-1}(O_y)$ has a neighborhood in the same set, and that implies $f^{-1}(O_y)$ is open. \square

Definition 2.2.19. A sequence x_1, x_2, \dots of points in X is said to **converge** to $x \in X$ if for every neighborhood O of x there exists a positive integer N such that $x_i \in O$ for all $i \geq N$.

A sequence need not converge at all. However, if it converges in a metrizable space, then it converges to only one element.

Theorem 2.2.20. Suppose that X is a metrizable space, and let $A \subseteq X$. There

exists a sequence of points of A converging to x if and only if $x \in \overline{A}$.

Proof. Suppose $x_n \rightarrow x$, where $x_n \in A$. Then, for every open set O containing x , there is an N , such that $x_n \in O$ for all $n > N$. By Theorem 2.2.9, this implies that $x \in \overline{A}$. Let d be a metric for the topology of X and x be a point in \overline{A} . For each positive integer n , consider the neighborhood $B_d(x, \frac{1}{n})$. Since $x \in \overline{A}$, the set $A \cap B_d(x, \frac{1}{n})$ is not empty and we choose x_n to be any point in this set. It follows that the sequence x_1, x_2, \dots converges to x . Notice that the “only if” proof holds for any topological space, while “if” requires a metric. \square

Theorem 2.2.21. *Let $f: X \rightarrow Y$ where X is a metrizable space. The function f is continuous if and only if for every convergent sequence $x_n \rightarrow x$ in X , the sequence $f(x_n)$ converges to $f(x)$.*

Proof. Suppose that f is continuous. Let O be a neighborhood of $f(x)$. Then $f^{-1}(O)$ is a neighborhood of x , and so there exists an integer N such that $x_n \in f^{-1}(O)$ for $n \geq N$. Thus, $f(x_n) \in O$ for all $n \geq N$ and $f(x_n) \rightarrow f(x)$.

To prove the converse, assume that the convergent sequence condition is true. Let $A \subseteq X$. Since X is metrizable, one finds that $x \in \overline{A}$ implies that there exists a sequence x_1, x_2, \dots of points of A converging to x . By assumption, $f(x_n) \rightarrow f(x)$. Since $f(x_n) \in f(A)$, Theorem 2.2.17 implies that $f(x) \in \overline{f(A)}$. Hence $f(\overline{A}) \subseteq \overline{f(A)}$ and f is continuous. \square

Chapter 3

Linear Algebra

3.1 Fields

This section focuses on key properties of the real and complex numbers that make them useful for linear algebra. Consider a set F of objects and two operations on the elements of F , addition and multiplication. For every pair of elements $s, t \in F$ then their sum $(s + t) \in F$. For every pair of elements $s, t \in F$ then their product $st \in F$. Suppose that these two operations satisfy

1. addition is commutative: $s + t = t + s \forall s, t \in F$
2. addition is associative: $r + (s + t) = (r + s) + t \forall r, s, t \in F$
3. to each $s \in F$ there exists a unique element $(-s) \in F$ such that $s + (-s) = 0$
4. multiplication is commutative: $st = ts \forall s, t \in F$
5. multiplication is associative: $r(st) = (rs)t \forall r, s, t \in F$
6. there is a unique non-zero element $1 \in F$ such that $s1 = s \forall s \in F$
7. to each $s \in F - 0$ there exists a unique element $s^{-1} \in F$ such that $ss^{-1} = 1$
8. multiplication distributes over addition: $r(s + t) = rs + rt \forall r, s, t \in F$.

Then, the set F together with these two operations is a **field**.

Example 3.1.1. *The real numbers with the usual operations of addition and multiplication form a field. The complex numbers with these two operations also form a field.*

Example 3.1.2. *The set of integers with addition and multiplication is not a field.*

Problem 3.1.3. *Is the set of rational numbers a subfield of the real numbers?*

Example 3.1.4. *Is the set of all real numbers of the form $s + t\sqrt{2}$, where s and t are rational, a subfield of the complex numbers?*

The set $F = \{s + t\sqrt{2} : s, t \in \mathbb{Q}\}$ together with the standard addition and multiplication is a field. Let $s, t, u, v \in \mathbb{Q}$,

$$\begin{aligned} s + t\sqrt{2} + u + v\sqrt{2} &= (s + u) + (t + v)\sqrt{2} \in F \\ (s + t\sqrt{2})(u + v\sqrt{2}) &= (su + 2tv) + (sv + tu)\sqrt{2} \in F \\ (s + t\sqrt{2})^{-1} &= \frac{s - t\sqrt{2}}{s^2 + 2t^2} = \frac{s}{s^2 + 2t^2} - \frac{t}{s^2 + 2t^2}\sqrt{2} \in F \end{aligned}$$

Again, the remaining properties are straightforward to prove. The field $s + t\sqrt{2}$, where s and t are rational, is a subfield of the complex numbers.

3.2 Matrices

Let F be a field and consider the problem of finding n scalars x_1, \dots, x_n which satisfy the conditions

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2 \\ \vdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= y_m \end{aligned} \tag{3.1}$$

where $y_1, \dots, y_m \in F$ and $a_{ij} \in F$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. These conditions form a system of m linear equations in n unknowns. A shorthand notation for (3.1) is the matrix equation

$$Ax = y,$$

where $\underline{x} = (x_1, \dots, x_n)^T$, $\underline{y} = (y_1, \dots, y_m)^T$, and A is the matrix given by

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

We also use $[A]_{i,j}$ to denote the entry of A in the i -th row and j -th column (i.e., a_{ij}) and $[\underline{x}]_i$ to denote the i -th entry in \underline{x} (i.e., x_i).

Definition 3.2.1. Let A be an $m \times n$ matrix over F and let B be an $n \times p$ matrix over F . The **matrix product** AB is the $m \times p$ matrix C whose i, j entry is

$$c_{ij} = \sum_{r=1}^n a_{ir} b_{rj}. \quad (3.2)$$

Remark 3.2.2. Consider (3.2) when j is fixed and i is eliminated by grouping the elements of C and A into column vectors $\underline{c}_1, \dots, \underline{c}_p$ and $\underline{a}_1, \dots, \underline{a}_n$. For this case, (3.2) shows that the j -th column of C is a linear combination of the columns of A ,

$$\underline{c}_j = \sum_{r=1}^n \underline{a}_r b_{rj},$$

Similarly, one can fix i and eliminate the index j by grouping the elements of C and B into row vectors $\underline{c}_1, \dots, \underline{c}_m$ and $\underline{b}_1, \dots, \underline{b}_n$. Then, (3.2) shows that the i -th row of C is a linear combination of the rows of B ,

$$\underline{c}_i = \sum_{r=1}^n a_{ir} \underline{b}_r.$$

Definition 3.2.3. Consider an $m \times n$ matrix A with elements $a_{ij} \in F$. The **transpose** of A is the $n \times m$ matrix $B = A^T$ with elements defined by $b_{ij} = a_{ji}$.

Definition 3.2.4. Consider a complex $m \times n$ matrix A with elements $a_{ij} \in \mathbb{C}$. Its **Hermitian transpose** $B = A^H$ is the $n \times m$ matrix with elements defined $b_{ij} = \overline{a_{ji}}$, where \bar{a} denotes the complex conjugate of a .

Problem 3.2.5. For matrices $A \in \mathbb{C}^{m \times p}$ and $B \in \mathbb{C}^{p \times n}$, show $(AB)^H = B^H A^H$.

Definition 3.2.6. An $m \times n$ matrix A over F is in **row echelon form** if

1. all rows containing only zeros, if they exist, are the bottom of the matrix, and
2. For rows with non-zero entries, the leading coefficient (i.e., the first non-zero element from the left) is strictly to the right of the leading coefficient of the row above it.

These two conditions imply that entries below the leading coefficient in a column are zero. A matrix is **column echelon form** if its transpose is in row echelon form.

Definition 3.2.7. An $m \times n$ matrix A over F is in **reduced row echelon form** if it is in row echelon form and

1. every leading coefficient is 1, and
2. every leading coefficient is the only non-zero element in its column.

Definition 3.2.8. Let A be an $n \times n$ matrix over F . An $n \times n$ matrix B is called the **inverse** of A if

$$AB = BA = I.$$

In this case, A is called **invertible** and its inverse is denoted by A^{-1} .

Problem 3.2.9. For a matrix $A \in \mathbb{C}^{n \times n}$, show that $(A^H)^{-1} = (A^{-1})^H$ if A^{-1} exists.

Definition 3.2.10. An **elementary row operation** on an $m \times n$ matrix consists of

1. multiplying a row by a non-zero scalar;
2. swapping two rows, or
3. adding a scalar multiple of one row to another row.

An **elementary column operation** is the same but applied to the columns.

Lemma 3.2.11. For any $m \times n$ matrix A over F , there is an invertible $m \times m$ matrix P over F such that $R = PA$ is in reduced row echelon form.

Sketch of Proof. This follows from the fact that elementary row operations (i.e., Gaussian elimination) can be used to reduce any matrix to reduced row echelon form. To construct the P matrix, one applies Gaussian elimination to the augmented matrix $A' = [A \ I]$. This results in an augmented matrix $R' = [R \ P]$ in reduced row echelon form. It follows that R is also in reduced row echelon form. Since elementary row operations can be implemented by (invertible) matrix multiplies on the left side, one also finds that $R' = PA'$, $R = PA$, and P is invertible. \square

Lemma 3.2.12. *Let A be an $m \times n$ matrix over F with $m < n$. Then, there exists a length- n column vector $\underline{x} \neq \underline{0}$ (over F) such that $A\underline{x} = \underline{0}$.*

Proof. First, we use row reduction to compute the reduced row echelon form $R = PA$ of A , where P is invertible. Then, we observe that the columns of R containing leading elements can be combined in a linear combination to cancel any other column of R . This allows us to construct a vector \underline{x} satisfying $R\underline{x} = \underline{0}$ and thus $A\underline{x} = P^{-1}R\underline{x} = \underline{0}$. \square

3.3 Vector Spaces

Definition 3.3.1. *A vector space consists of the following,*

1. *a field F of scalars*
2. *a set V of objects, called vectors*
3. *an operation called vector addition, which associates with each pair of vectors $\underline{v}, \underline{w} \in V$ a vector $\underline{v} + \underline{w} \in V$ such that*
 - (a) *addition is commutative: $\underline{v} + \underline{w} = \underline{w} + \underline{v}$*
 - (b) *addition is associative: $\underline{u} + (\underline{v} + \underline{w}) = (\underline{u} + \underline{v}) + \underline{w}$*
 - (c) *there is a unique vector $\underline{0} \in V$ such that $\underline{v} + \underline{0} = \underline{v}, \forall \underline{v} \in V$*
 - (d) *to each $\underline{v} \in V$ there is a unique vector $-\underline{v} \in V$ such that $\underline{v} + (-\underline{v}) = \underline{0}$*
4. *an operation called scalar multiplication, which associates with each $s \in F$ and $\underline{v} \in V$ a vector $s\underline{v} \in V$ such that*
 - (a) $1\underline{v} = \underline{v}, \forall \underline{v} \in V$
 - (b) $(s_1 s_2)\underline{v} = s_1(s_2\underline{v})$
 - (c) $s(\underline{v} + \underline{w}) = s\underline{v} + s\underline{w}$
 - (d) $(s_1 + s_2)\underline{v} = s_1\underline{v} + s_2\underline{v}$.

Example 3.3.2. *Let F be a field, and let V be the set of all n -tuples $\underline{v} = (v_1, \dots, v_n)$ of scalar $v_i \in F$. If $\underline{w} = (w_1, \dots, w_n)$ with $w_i \in F$, the sum of \underline{v} and \underline{w} is defined by*

$$\underline{v} + \underline{w} = (v_1 + w_1, \dots, v_n + w_n).$$

The product of a scalar $s \in F$ and vector \underline{v} is defined by

$$s\underline{v} = (sv_1, \dots, sv_n).$$

The set of n -tuples, denoted by F^n , with the vector addition and scalar product defined above forms a vector space. This is the standard vector space for F^n .

Example 3.3.3. Let X be a non-empty set and let Y be a vector space over F . Consider the set V of all functions from X into Y . The sum of two vectors $f, g \in V$ is the function from X into Y defined by

$$(f + g)(x) = f(x) + g(x) \quad \forall x \in X,$$

where the RHS uses vector addition from Y . The product of scalar $s \in F$ and the function $f \in V$ is the function sf defined by

$$(sf)(x) = sf(x) \quad \forall x \in X,$$

where the RHS uses scalar multiplication from Y . This is the standard vector space of functions from a set X to a vector space Y .

Definition 3.3.4. A vector $\underline{w} \in V$ is said to be a **linear combination** of the vectors $\underline{v}_1, \dots, \underline{v}_n \in V$ provided that there exist scalars $s_1, \dots, s_n \in F$ such that

$$\underline{w} = \sum_{i=1}^n s_i \underline{v}_i.$$

3.3.1 Subspaces

Definition 3.3.5. Let V be a vector space over F . A **subspace** of V is a subset $W \subset V$ which is itself a vector space over F .

Fact 3.3.6. A non-empty subset $W \subset V$ is a subspace of V if and only if for every pair $\underline{w}_1, \underline{w}_2 \in W$ and every scalar $s \in F$ the vector $s\underline{w}_1 + \underline{w}_2$ is again in W .

If V is a vector space then the intersection of any collection of subspaces of V is a subspace of V .

Example 3.3.7. Let A be an $m \times n$ matrix over F . The set of all $n \times 1$ column vectors V such that

$$\underline{v} \in V \implies A\underline{v} = \underline{0}$$

is a subspace of $F^{n \times 1}$.

Definition 3.3.8. Let U be a set (or list) of vectors in V . The **span** of U , denoted $\text{span}(U)$, is defined to be the set of all finite linear combinations of vectors in U .

The subspace spanned by U can be defined equivalently as the intersection of all subspaces of V that contain U . To see this, we note that the intersection of all subspaces containing U is a subspace containing U because the intersection of subspaces is also a subspace. The intersection cannot be larger than $\text{span}(U)$, however, because $\text{span}(U)$ is a subspace containing U .

Definition 3.3.9. Let V be a vector space and U, W be subspaces. If U, W are disjoint (i.e., $U \cap W = \{\underline{0}\}$), their **direct sum** $U \oplus W$ is defined by

$$U \oplus W \triangleq \{\underline{u} + \underline{w} \mid \underline{u} \in U, \underline{w} \in W\}.$$

An important property of a direct sum is that any vector $\underline{v} \in U \oplus W$ has a unique decomposition $\underline{v} = \underline{u} + \underline{w}$ where $\underline{u} \in U$ and $\underline{w} \in W$.

3.3.2 Bases and Dimensions

The dimension of a vector space is defined using the concept of a basis.

Definition 3.3.10. Let V be a vector space over F . A list of vectors $\underline{u}_1, \dots, \underline{u}_n \in V$ is called **linearly dependent** if there are scalars $s_1, \dots, s_n \in F$, not all of which are 0, such that

$$\sum_{i=1}^n s_i \underline{u}_i = \underline{0}.$$

A list that is not linearly dependent is called **linearly independent**. Similarly, a subset $U \subset V$ is called linearly dependent if there is a finite list $\underline{u}_1, \dots, \underline{u}_n \in U$ of distinct vectors that is linearly dependent. Otherwise, it is called linearly independent.

A few important consequences follow immediately from this definition. Any subset of a linearly independent set is also linearly independent. Any set which contains the $\underline{0}$ vector is linearly dependent. A set $U \subset V$ is linearly independent if and only if each finite subset of U is linearly independent.

Definition 3.3.11. Let V be a vector space over F . Let $\mathcal{B} = \{\underline{v}_\alpha \mid \alpha \in A\}$ be a subset of linearly independent vectors from V such that every $\underline{v} \in V$ can be written

as a finite linear combination of vectors from \mathcal{B} . Then, the set \mathcal{B} is a **Hamel basis** for V . The space V is **finite-dimensional** if it has a finite basis.

Using this definition, we note that a basis decomposition $\underline{v} = \sum_{i=1}^n s_i \underline{v}_{\alpha_i}$ must be unique because the difference between any two distinct decompositions produces a finite linear dependency in the basis and, hence, a contradiction.

Theorem 3.3.12. *Every vector space has a Hamel basis.*

Proof. Let X be the set of linearly independent subsets of V . Furthermore, for $x, y \in X$ consider the strict partial order defined by proper inclusion. By the maximum principle, if x is an element of X , then there exists a maximal simply ordered subset Z of X containing x . This element is a Hamel basis for V . \square

Example 3.3.13. *Let F be a field and let $U \subset F^n$ be the subset consisting of the vectors $\underline{e}_1, \dots, \underline{e}_n$ defined by*

$$\begin{aligned}\underline{e}_1 &= (1, 0, \dots, 0) \\ \underline{e}_2 &= (0, 1, \dots, 0) \\ &\vdots \\ \underline{e}_n &= (0, 0, \dots, 1).\end{aligned}$$

For any $\underline{v} = (v_1, \dots, v_n) \in F^n$, we have

$$\underline{v} = \sum_{i=1}^n v_i \underline{e}_i. \quad (3.3)$$

Thus, the collection $U = \{\underline{e}_1, \dots, \underline{e}_n\}$ spans F^n . Since $\underline{v} = \underline{0}$ in (3.3) if and only if $v_1 = \dots = v_n = 0$, U is linearly independent. Accordingly, the set U is a basis for $F^{n \times 1}$. This basis is termed the **standard basis** of F^n .

Lemma 3.3.14. *Let $A \in F^{n \times n}$ be an invertible matrix. Then, the columns of A form a basis for F^n . Similarly, the rows of A will also form a basis for F^n*

Proof. If $\underline{v} = (v_1, \dots, v_n)^T$ is a column vector, then

$$A\underline{v} = \sum_{i=1}^n v_i \underline{a}_i,$$

where the columns of A are denoted by $\underline{a}_1, \dots, \underline{a}_n$. Since A is invertible,

$$A\underline{v} = \underline{0} \implies I\underline{v} = A^{-1}\underline{0} \implies \underline{v} = \underline{0}.$$

Thus, $\{\underline{a}_1, \dots, \underline{a}_n\}$ is a linearly independent set. Next, for any column vector $\underline{w} \in F^n$, let $\underline{v} = A^{-1}\underline{w}$. It follows that $\underline{w} = A\underline{v}$ and, thus, $\{\underline{a}_1, \dots, \underline{a}_n\}$ is a basis for F^n . If A is invertible, then $(A^T)^{-1}$ exists. Thus, the same holds for the rows. \square

Theorem 3.3.15. *Let V be a finite-dimensional vector space that is spanned by a finite set of vectors $W = \{\underline{w}_1, \dots, \underline{w}_n\}$. If $U = \{\underline{u}_1, \dots, \underline{u}_m\} \subset V$ is a linearly independent set of vectors, then $m \leq n$.*

Proof. Suppose that $U = \{\underline{u}_1, \dots, \underline{u}_m\} \subset V$ is linearly independent and $m > n$. Since W spans V , there exists scalars a_{ij} such that

$$\underline{u}_j = \sum_{i=1}^n a_{ij}\underline{w}_i.$$

For any m scalars s_1, \dots, s_m we have

$$\sum_{j=1}^m s_j \underline{u}_j = \sum_{j=1}^m s_j \sum_{i=1}^n a_{ij} \underline{w}_i = \sum_{j=1}^m \sum_{i=1}^n (a_{ij} s_j) \underline{w}_i = \sum_{i=1}^n \left(\sum_{j=1}^m a_{ij} s_j \right) \underline{w}_i.$$

Collecting the a_{ij} coefficients into an n by m matrix A shows that

$$\begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix} = A \begin{bmatrix} s_1 \\ \vdots \\ s_m \end{bmatrix}.$$

Since $A \in F^{n \times m}$ with $n < m$, Lemma 3.2.12 implies there are scalars s_1, \dots, s_n , not all 0, such that $t_1 = t_2 = \dots = t_m = 0$. For these scalars, $\sum_{j=1}^m s_j \underline{u}_j = \underline{0}$. Thus, the set U is linearly dependent. and the contradiction implies $m \leq n$. \square

Now, suppose that V is a finite-dimensional vector space with bases $U = \{\underline{u}_1, \dots, \underline{u}_n\}$ and $W = \{\underline{w}_1, \dots, \underline{w}_m\}$ where $m \neq n$. Then, without loss of generality, we can assume $m > n$ and apply Theorem 3.3.15 to see that W must be linearly dependent. Since a basis must be linearly independent, this gives a contradiction and implies that $m = n$. Hence, if V is a finite-dimensional vector space, then any two bases of V have the same number of elements. Therefore, the dimension of a finite-dimensional vector space is uniquely defined. Thus, our intuition about dimension from \mathbb{R}^n does not break down for other vector spaces and fields.

Definition 3.3.16. The *dimension* of a finite-dimensional vector space is the number of elements in any basis for V . We denote the dimension of a finite-dimensional vector space V by $\dim(V)$.

The zero subspace of a vector space V is the subspace spanned by the vector $\underline{0}$. Since the set $\{\underline{0}\}$ is linearly dependent and not a basis, we assign a dimension 0 to the zero subspace. Alternatively, it can be argued that the empty set \emptyset spans $\{\underline{0}\}$ because the intersection of all the subspaces containing the empty set is $\{\underline{0}\}$. Though this is only a minor point.

Theorem 3.3.17. Let A be an $n \times n$ matrix over F whose columns, denoted by $\underline{a}_1, \dots, \underline{a}_n$, form a linearly independent set of vectors in F^n . Then A is invertible.

Proof. Let W be the subspace of $V = F^n$ spanned by $\underline{a}_1, \dots, \underline{a}_n$. Since $\underline{a}_1, \dots, \underline{a}_n$ are linearly independent, $\dim(W) = n = \dim(V)$. Now, suppose $W \neq V$. Since $W \subseteq V$, that implies there is a vector $\underline{v} \in V$ such that $\underline{v} \notin W$. It would follow that $\dim(V) > \dim(W)$ but this contradicts $\dim(V) = \dim(W)$. Thus, $W = V$.

Since $W = V$, one can write the standard basis vectors $\underline{e}_1, \dots, \underline{e}_n \in F^n$ in terms of the columns of A . In particular, there exist scalars $b_{ij} \in F$ such that

$$\underline{e}_j = \sum_{i=1}^n b_{ij} \underline{a}_i, \quad 1 \leq j \leq n.$$

Then, for the matrix B with entries b_{ij} , we have $AB = I$

Next, suppose that the columns of B are linearly dependent. Then, there is a non-zero $\underline{v} \in F^n$ such that $B\underline{v} = \underline{0}$. But, that gives the contradiction that $A(B\underline{v}) = \underline{0}$ and $(AB)\underline{v} = I\underline{v} = \underline{v}$. Thus, the columns of B are linearly independent.

Using the first argument again, one finds that there is a matrix C such that $BC = I$. This also implies that $A = AI = A(BC) = (AB)C = IC = C$. Thus, A^{-1} exists and equals B . \square

3.3.3 Coordinate System

Let $\{\underline{v}_1, \dots, \underline{v}_n\}$ be a basis for the n -dimensional vector space V and recall that every vector $\underline{w} \in V$ can be expressed uniquely as

$$\underline{w} = \sum_{i=1}^n s_i \underline{v}_i.$$

While standard vector and matrix notation requires that the basis elements be ordered, a set is an unordered collection of objects. Ordering this set (e.g., $\underline{v}_1, \dots, \underline{v}_n$) allows the first element in the coordinate vector to be associated with the first vector in our basis and so on.

Definition 3.3.18. *If V is a finite-dimensional vector space, an **ordered basis** for V is a finite list of vectors that is linearly independent and spans V .*

In particular, if the sequence $\underline{v}_1, \dots, \underline{v}_n$ is an ordered basis for V , then the set $\{\underline{v}_1, \dots, \underline{v}_n\}$ is a basis for V . The ordered basis \mathcal{B} , denoted by $(\underline{v}_1, \dots, \underline{v}_n)$, defines the set and a specific ordering of the vectors. Based on this ordered basis, a vector $\underline{v} \in V$ can be unambiguously represented as an n -tuple $(s_1, \dots, s_n) \in F^n$ such that

$$\underline{v} = \sum_{i=1}^n s_i \underline{v}_i.$$

Definition 3.3.19. *For a finite-dimensional vector space V with ordered basis $\mathcal{B} = (\underline{v}_1, \dots, \underline{v}_n)$, the **coordinate vector** of $\underline{v} \in V$ is denoted by $[\underline{v}]_{\mathcal{B}}$ and equals the unique vector $\underline{s} \in F^n$ such that*

$$\underline{v} = \sum_{i=1}^n s_i \underline{v}_i.$$

The dependence of the coordinate vector $[\underline{v}]_{\mathcal{B}}$ on the basis is explicitly specified using the subscript. This can be particularly useful when multiple coordinate systems are involved.

Example 3.3.20. *The canonical example of an ordered basis is the standard basis for F^n introduced in Section 3.3.2. Note that the standard basis contains a natural ordering: $\underline{e}_1, \dots, \underline{e}_n$. Vectors in F^n can therefore be unambiguously expressed as n -tuples.*

Problem 3.3.21. *Suppose that $\mathcal{A} = \underline{v}_1, \dots, \underline{v}_n$ is an ordered basis for V . Let P be an $n \times n$ invertible matrix. Show that there exists an ordered basis $\mathcal{B} = \underline{w}_1, \dots, \underline{w}_n$ for V such that*

$$\begin{aligned} [\underline{u}]_{\mathcal{A}} &= P [\underline{u}]_{\mathcal{B}} \\ [\underline{u}]_{\mathcal{B}} &= P^{-1} [\underline{u}]_{\mathcal{A}} \end{aligned}$$

for every $\underline{u} \in V$.

S 3.3.21. Consider the ordered basis $\mathcal{A} = \underline{v}_1, \dots, \underline{v}_n$ and let $Q = P^{-1}$. For all $\underline{u} \in V$, we have $\underline{u} = \sum_{i=1}^n s_i \underline{v}_i$, where

$$[\underline{u}]_{\mathcal{A}} = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix}.$$

If we define

$$\underline{w}_i = \sum_{k=1}^n p_{ki} \underline{v}_k \quad \text{and} \quad t_i = \sum_{j=1}^n q_{ij} s_j,$$

then we find that

$$\begin{aligned} \sum_{i=1}^n t_i \underline{w}_i &= \sum_{i=1}^n \sum_{j=1}^n q_{ij} s_j \underline{w}_i = \sum_{i=1}^n \sum_{j=1}^n q_{ij} s_j \sum_{k=1}^n p_{ki} \underline{v}_k \\ &= \sum_{j=1}^n s_j \sum_{k=1}^n \underline{v}_k \sum_{i=1}^n p_{ki} q_{ij} = \sum_{j=1}^n s_j \sum_{k=1}^n \underline{v}_k \delta_{jk} \\ &= \sum_{j=1}^n s_j \underline{v}_j = \underline{u}. \end{aligned}$$

This shows that $\mathcal{B} = \underline{w}_1, \dots, \underline{w}_n$ is an ordered basis for V and

$$[\underline{u}]_{\mathcal{B}} = \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix}.$$

The definition of t_i also shows that $[\underline{u}]_{\mathcal{B}} = P^{-1} [\underline{u}]_{\mathcal{A}}$ and therefore $[\underline{u}]_{\mathcal{A}} = P [\underline{u}]_{\mathcal{B}}$.

3.4 Linear Transformations

3.4.1 Definitions

Definition 3.4.1. Let V and W be vector spaces over a field F . A **linear transform** from V to W is a function T from V into W such that

$$T(s\underline{v}_1 + \underline{v}_2) = sT\underline{v}_1 + T\underline{v}_2$$

for all \underline{v}_1 and \underline{v}_2 in V and all scalars s in F .

Definition 3.4.2. Let $L(V, W)$ denote the **set of all linear transforms** from V into W , where V and W are vector spaces over a field F .

Example 3.4.3. Let A be a fixed $m \times n$ matrix over F . The function T defined by $T(\underline{v}) = A\underline{v}$ is a linear transformation from $F^{n \times 1}$ into $F^{m \times 1}$.

Example 3.4.4. Let $P \in F^{m \times m}$ and $Q \in F^{n \times n}$ be fixed matrices. Define the function T from $F^{m \times n}$ into itself by $T(A) = PAQ$. Then T is a linear transformation from $F^{m \times n}$ into $F^{m \times n}$. In particular,

$$\begin{aligned} T(sA + B) &= P(sA + B)Q \\ &= sPAQ + PBQ \\ &= sT(A) + T(B). \end{aligned}$$

Example 3.4.5. Let V be the space of continuous functions from $[0, 1]$ to \mathbb{R} , and define T by

$$(Tf)(x) = \int_0^x f(t)dt.$$

Then T is a linear transformation from V into V . The function Tf is continuous and differentiable.

It is important to note that if T is a linear transformation from V to W , then $T(\underline{0}) = \underline{0}$. This is essential since

$$T(\underline{0}) = T(\underline{0} + \underline{0}) = T(\underline{0}) + T(\underline{0}).$$

Definition 3.4.6. A linear transformation $T: V \rightarrow W$ is **singular** if there is a non-zero vector $\underline{v} \in V$ such that $T\underline{v} = \underline{0}$. Otherwise, it is called **non-singular**.

3.4.2 Properties

The following theorem illuminates a very important structural element of linear transformations: they are uniquely defined by where they map a set of basis vectors for their domain.

Theorem 3.4.7. Let V, W be vector spaces over F and $\mathcal{B} = \{\underline{v}_\alpha | \alpha \in A\}$ be a Hamel basis for V . For each mapping $G: \mathcal{B} \rightarrow W$, there is a unique linear transformation $T: V \rightarrow W$ such that $T\underline{v}_\alpha = G(\underline{v}_\alpha)$ for all $\alpha \in A$.

Proof. Since \mathcal{B} is a Hamel basis for V , every vector $\underline{w} \in V$ has a unique expansion

$$\underline{w} = \sum_{\alpha \in A} s_{\alpha}(\underline{w}) \underline{v}_{\alpha},$$

where $s_{\alpha}(\underline{w})$ is the unique α coefficient for \underline{w} and $s_{\alpha}(\underline{w}) \neq 0$ only for a finite subset of A . Using the unique expansion and vector space properties, one can show that

$$s_{\alpha}(t\underline{w}_1 + \underline{w}_2) = ts_{\alpha}(\underline{w}_1) + s_{\alpha}(\underline{w}_2).$$

Next, we define the mapping $T: V \rightarrow W$ in terms of $s_{\alpha}(\cdot)$ and $G(\cdot)$ with

$$T\underline{w} = \sum_{\alpha \in A} s_{\alpha}(\underline{w}) G(\underline{v}_{\alpha}).$$

Using the linearity of $s_{\alpha}(\cdot)$, it is easy to verify that T is a linear transform.

To show that T is unique, we let $U: V \rightarrow W$ be any other linear mapping satisfying $U\underline{v}_{\alpha} = G(\underline{v}_{\alpha})$ for all $\alpha \in A$. In this case, the linearity of U guarantees that

$$U\underline{w} = U\left(\sum_{\alpha \in A} s_{\alpha}(\underline{w}) \underline{v}_{\alpha}\right) = \sum_{\alpha \in A} s_{\alpha}(\underline{w}) U(\underline{v}_{\alpha}) = \sum_{\alpha \in A} s_{\alpha}(\underline{w}) G(\underline{v}_{\alpha}).$$

From this, we see that $U\underline{w} = T\underline{w}$ for all $\underline{w} \in V$ and therefore that $U = T$. \square

Definition 3.4.8. Let V and W be vector spaces with ordered bases \mathcal{A} and \mathcal{B} . Then, the **coordinate matrix** for the linear transform $T: V \rightarrow W$ with respect to \mathcal{A} and \mathcal{B} is denoted $[T]_{\mathcal{A}, \mathcal{B}}$ and, for all $\underline{v} \in V$, satisfies

$$[T\underline{v}]_{\mathcal{B}} = [T]_{\mathcal{A}, \mathcal{B}}[\underline{v}]_{\mathcal{A}}.$$

If $V = W$ and $\mathcal{A} = \mathcal{B}$, then the coordinate matrix $[T]_{\mathcal{A}, \mathcal{A}}$ is denoted by $[T]_{\mathcal{A}}$.

Definition 3.4.9. If T is a linear transformation from V into W , the **range** of T is the set of all vectors $\underline{w} \in W$ such that $\underline{w} = T\underline{v}$ for some $\underline{v} \in V$. We denote the range of T by

$$\mathcal{R}(T) \triangleq \{\underline{w} \in W \mid \exists \underline{v} \in V, T\underline{v} = \underline{w}\} = \{T\underline{v} \mid \underline{v} \in V\}.$$

The set $\mathcal{R}(T)$ is a subspace of W . Let $\underline{w}_1, \underline{w}_2 \in \mathcal{R}(T)$ and let s be a scalar. By definition, there exist vectors \underline{v}_1 and \underline{v}_2 in V such that $T\underline{v}_1 = \underline{w}_1$ and $T\underline{v}_2 = \underline{w}_2$. Since T is a linear transformation, we have

$$T(s\underline{v}_1 + \underline{v}_2) = sT\underline{v}_1 + T\underline{v}_2 = s\underline{w}_1 + \underline{w}_2,$$

which shows that $s\underline{w}_1 + \underline{w}_2$ is also in $\mathcal{R}(T)$.

Definition 3.4.10. If T is a linear transformation from V into W , the **nullspace** of T is the set of all vectors $\underline{v} \in V$ such that $T\underline{v} = \underline{0}$. We denote the nullspace of T by

$$\mathcal{N}(T) \triangleq \{\underline{v} \in V \mid T\underline{v} = \underline{0}\}.$$

It can easily be verified that $\mathcal{N}(T)$ is a subspace of V .

$$T(\underline{0}) = \underline{0} \implies \underline{0} \in \mathcal{N}(T).$$

Furthermore, if $T\underline{v}_1 = T\underline{v}_2 = \underline{0}$ then

$$T(s\underline{v}_1 + \underline{v}_2) = sT(\underline{v}_1) + T(\underline{v}_2) = s\underline{0} + \underline{0} = \underline{0},$$

so that $s\underline{v}_1 + \underline{v}_2 \in \mathcal{N}(T)$.

Definition 3.4.11. Let V and W be vector spaces over a field F , and let T be a linear transformation from V into W . If V is finite-dimensional, the **rank** of T is the dimension of the range of T and the **nullity** of T is the dimension of the nullspace of T .

Theorem 3.4.12. Let V and W be vector spaces over the field F and let T be a linear transformation from V into W . If V is finite-dimensional, then

$$\text{rank}(T) + \text{nullity}(T) = \dim(V)$$

Proof. Let $\underline{v}_1, \dots, \underline{v}_k$ be a basis for $\mathcal{N}(T)$, the nullspace of T . There are vectors $\underline{v}_{k+1}, \dots, \underline{v}_n \in V$ such that $\underline{v}_1, \dots, \underline{v}_n$ is a basis for V . We want to show that $T\underline{v}_{k+1}, \dots, T\underline{v}_n$ is a basis for the range of T . The vectors $T\underline{v}_1, \dots, T\underline{v}_n$ certainly span $\mathcal{R}(T)$ and, since $T\underline{v}_j = \underline{0}$ for $j = 1, \dots, k$, it follows that $T\underline{v}_{k+1}, \dots, T\underline{v}_n$ span $\mathcal{R}(T)$. Suppose that there exist scalars s_{k+1}, \dots, s_n such that

$$\sum_{j=k+1}^n s_j T\underline{v}_j = \underline{0}.$$

This implies that

$$T\left(\sum_{j=k+1}^n s_j \underline{v}_j\right) = \underline{0}.$$

and accordingly the vector $\underline{v} = \sum_{j=k+1}^n s_j \underline{v}_j$ is in the nullspace of T . Since $\underline{v}_1, \dots, \underline{v}_k$ form a basis for $\mathcal{N}(T)$, there must be a linear combination such that

$$\underline{v} = \sum_{j=1}^k t_j \underline{v}_j.$$

But then,

$$\sum_{j=1}^k t_j \underline{v}_j - \sum_{j=k+1}^n s_j \underline{v}_j = \underline{0}.$$

Since the vectors $\underline{v}_1, \dots, \underline{v}_n$ are linearly independent, this implies that

$$t_1 = \dots = t_k = s_{k+1} = \dots = s_n = 0.$$

That is, the set $T\underline{v}_{k+1}, \dots, T\underline{v}_n$ is linearly independent in W and therefore forms a basis for $\mathcal{R}(T)$. In turn, this implies that $n = \text{rank}(T) + \text{nullity}(T)$. \square

Theorem 3.4.13. *If A is an $m \times n$ matrix with entries in the field F , then*

$$\text{row rank}(A) \triangleq \dim(\mathcal{R}(A^T)) = \dim(\mathcal{R}(A)) \triangleq \text{rank}(A).$$

Proof. Let $R = PA$ be the reduced row echelon form of A , where P is invertible. Let r be the number of non-zero rows in R and observe that $\text{row rank}(A) = r$ because the rows of R form a basis for the row space of A . Next, we write $A = P^{-1}R$ and observe that each column of R has non-zero entries only in the first r rows. Thus, each column of A is a linear combination of the first r columns in P^{-1} . Thus, the column space of A is spanned by r vectors and $\text{rank}(A) \leq \text{row rank}(A)$.

The proof is completed by applying the above bound to both A and A^T to get

$$\text{rank}(A) \leq \text{row rank}(A) = \text{rank}(A^T) \leq \text{row rank}(A^T) = \text{rank}(A). \quad \square$$

When $F = \mathbb{C}$, the space $\mathcal{R}(A^H)$ has many nice properties and can also be called the row space of A . Regardless, it holds that $\text{rank}(A) = \text{rank}(A^T) = \text{rank}(A^H)$.

3.5 Norms

Let V be a vector space over the real numbers or the complex numbers.

Definition 3.5.1. *A **norm** on vector space V is a real-valued function $\|\cdot\| : V \rightarrow \mathbb{R}$ that satisfies the following properties.*

1. $\|\underline{v}\| \geq 0 \quad \forall \underline{v} \in V$; equality holds if and only if $\underline{v} = \underline{0}$
2. $\|s\underline{v}\| = |s| \|\underline{v}\| \quad \forall \underline{v} \in V, s \in F$
3. $\|\underline{v} + \underline{w}\| \leq \|\underline{v}\| + \|\underline{w}\| \quad \forall \underline{v}, \underline{w} \in V$.

The concept of a norm is closely related to that of a metric. For instance, a metric can be defined from any norm. Let $\|\underline{v}\|$ be a norm on vector space V , then

$$d(\underline{v}, \underline{w}) = \|\underline{v} - \underline{w}\|$$

is the metric induced by the norm.

Definition 3.5.2. A vector $\underline{v} \in V$ is said to be **normalized** if $\|\underline{v}\| = 1$. Any vector can be normalized, except the zero vector:

$$\underline{u} = \frac{\underline{v}}{\|\underline{v}\|} \quad (3.4)$$

has norm $\|\underline{u}\| = 1$. A normalized vector is also referred to as a **unit vector**.

Normed vector spaces are very useful because they have all the properties of a vector space and all the benefits of a topology generated by the norm. Therefore, one can discuss limits and convergence in a meaningful way.

Example 3.5.3. Consider vectors in \mathbb{R}^n with the euclidean metric

$$d(\underline{v}, \underline{w}) = \sqrt{(v_1 - w_1)^2 + \cdots + (v_n - w_n)^2}.$$

Recall that the standard bounded metric introduced in Problem 2.1.5 is given by

$$\bar{d}(\underline{v}, \underline{w}) = \min \{d(\underline{v}, \underline{w}), 1\}.$$

Define the function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ by $f(\underline{v}) = \bar{d}(\underline{v}, \underline{0})$. Is the function f a norm?

By the properties of a metric, we have

1. $\bar{d}(\underline{v}, \underline{0}) \geq 0 \quad \forall \underline{v} \in V$; equality holds if and only if $\underline{v} = \underline{0}$
2. $\bar{d}(\underline{v}, \underline{0}) + \bar{d}(\underline{w}, \underline{0}) = \bar{d}(\underline{v}, \underline{0}) + \bar{d}(\underline{0}, \underline{w}) \geq \bar{d}(\underline{v}, \underline{w}) \quad \forall \underline{v}, \underline{w} \in V$.

However, $\bar{d}(s\underline{v}, \underline{0})$ is not necessarily equal to $s\bar{d}(\underline{v}, \underline{0})$. For instance, $\bar{d}(2\underline{e}_1, \underline{0}) = 1 < 2\bar{d}(\underline{e}_1, \underline{0})$. Thus, the function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ defined by

$$f(\underline{v}) = \bar{d}(\underline{v}, \underline{0}).$$

is not a norm.

Example 3.5.6. The following functions are examples of norms for \mathbb{R}^n and \mathbb{C}^n :

What are L^p spaces? What is the Lebesgue Integral?

Many important spaces include functions that are not Riemann integrable. For $X = [a, b]^d$, the Lebesgue integral of $f: X \rightarrow \mathbb{R}$ is defined using measure theory and is often used in advanced probability courses. Since there are many non-zero Lebesgue-integrable functions whose integral is zero, this definition has a subtlety. The Lebesgue integral of a non-negative function is zero if and only if the function equals zero **almost everywhere** (abbreviated a.e.). Therefore, two functions are *equal almost everywhere* if the norm of their difference is zero. Strictly speaking, a vector space of “functions” with the L^p norm actually has elements that are equivalence classes of functions defined by *equality almost everywhere*.

The normed vector space $L^p(X)$ (with $1 \leq p < \infty$) is the subset of all functions mapping X to the real numbers (i.e., $f: X \rightarrow \mathbb{R}$) where the Lebesgue integral

$$\|f\|_{L^p} \triangleq \left(\int_X |f(x)|^p dx \right)^{1/p}$$

exists and is finite. Of course, this definition begs the question, “What is the Lebesgue integral?”. The following definition is sufficient for these notes:

Definition 3.5.4. *The **Lebesgue integral** is a generalization of the Riemann integral that applies to wider class of functions. The values of these two integrals coincide on the set of Riemann integrable functions. Loosely speaking, one can construct any non-negative function $f \in L^p(X)$ by considering sequences f_1, f_2, \dots of “simple” functions formed by rounding values of f down to values in a finite set $S_i \subset [0, \infty)$ where $\{0\} \subset S_1 \subset S_2 \subset \dots \subset [0, \infty)$. By construction, the sequence of functions is non-decreasing (i.e., $f_{n+1}(x) \geq f_n(x)$ for all $x \in X$) and, therefore, it converges pointwise to a limit function $f(x)$. Moreover, the Lebesgue integral of each simple function is easy to define. Thus, this sequence of simple functions gives rise to a non-decreasing sequence of Lebesgue integrals and one defines the Lebesgue integral of $f(x)$ to be the limit of this sequence. In fact, the non-negative functions in $L^p(X)$ are in one-to-one correspondence with the limits of non-decreasing sequences of simple functions that satisfy $\|f_n\|_{L^p} \rightarrow M < \infty$, up to a.e. equivalence.*

Definition 3.5.5. *The **Lebesgue measure** of a set $A \subseteq X$ is equal to the Lebesgue integral of its indicator function when both quantities exist. In particular, the set is measurable if and only if the Lebesgue integral of its indicator function exists.*

1. the l^1 norm: $\|\underline{v}\|_1 = \sum_{i=1}^n |v_i|$
2. the l^p norm: $\|\underline{v}\|_p = \left(\sum_{i=1}^n |v_i|^p\right)^{\frac{1}{p}}$, $p \in (1, \infty)$
3. the l^∞ norm: $\|\underline{v}\|_\infty = \max_{1, \dots, n} \{|v_i|\}$.

Example 3.5.7. Similarly, norms can be defined for the vector space of functions from $[a, b]$ to \mathbb{R} (or \mathbb{C}) with

1. the L^1 norm: $\|f(t)\|_1 = \int_a^b |f(t)| dt$
2. the L^p norm: $\|f(t)\|_p = \left(\int_a^b |f(t)|^p dt\right)^{\frac{1}{p}}$, $p \in (1, \infty)$
3. the L^∞ norm: $\|f(t)\|_\infty = \text{ess sup}_{[a,b]} \{|f(t)|\}$.

In this example, the integral notation refers to the **Lebesgue integral** (rather than the **Riemann integral**).

Example 3.5.8. Consider any set W of real-valued random variables, defined on a common probability space, such that $\|X\|_p \triangleq E[|X|^p]^{1/p} < \infty$ for all $X \in W$ and some fixed $p \in [1, \infty)$. Then, $V = \text{span}(W)$ is a normed vector space over \mathbb{R} and $X, Y \in V$ are considered to be equal if $\|X - Y\|^p = E[|X - Y|^p] = 0$ (or equivalently $\Pr(X \neq Y) = 0$). In addition, the closure of V is a Banach space.

Remark 3.5.9. We have not shown that the ℓ^p and L^p norm definitions above satisfy all the required properties. In particular, to prove the triangle inequality, one requires the Minkowski inequality which is deferred until Theorem 3.5.24.

Problem 3.5.10. For a normed vector space V , show that $\|\underline{v} - \underline{w}\| \geq | \|\underline{v}\| - \|\underline{w}\| |$ (i.e., the **reverse triangle inequality**) for all $\underline{v}, \underline{w} \in V$.

Definition 3.5.11. A complete normed vector space is called a **Banach space**.

Banach spaces are the standard setting for many problems because completeness is a powerful tool for solving problems.

Example 3.5.12. The vector spaces \mathbb{R}^n (or \mathbb{C}^n) with any well-defined norm are Banach spaces.

Example 3.5.13. The vector space of all continuous functions from $[a, b]$ to \mathbb{R} is a Banach space under the supremum norm

$$\|f(t)\| = \sup_{t \in [a, b]} f(t).$$

Definition 3.5.14. A Banach space V has a **Schauder basis**, $\underline{v}_1, \underline{v}_2, \dots$, if every $\underline{v} \in V$ can be written uniquely as

$$\underline{v} = \sum_{i=1}^{\infty} s_i \underline{v}_i.$$

Remark 3.5.15. While every vector space has a Hamel basis, not all Banach spaces have a Schauder basis. In particular, there are Banach spaces which are separable (with respect to the induced metric) but which do not contain a Schauder basis. Fortunately, these atypical spaces are essentially never encountered in practice.

Lemma 3.5.16. If $\sum_{i=1}^{\infty} \|\underline{v}_i\| = a < \infty$, then $\underline{u}_n = \sum_{i=1}^n \underline{v}_i$ satisfies $\underline{u}_n \rightarrow \underline{u}$.

Proof. This is left as an exercise for the reader because it is a straightforward generalization of the proof of Lemma 2.1.39. \square

Example 3.5.17. Let $V = \mathbb{R}^{\omega}$ be the vector space of semi-infinite real sequences. The **standard Schauder basis** is the countably infinite extension $\{\underline{e}_1, \underline{e}_2, \dots\}$ of the standard basis.

Definition 3.5.18. Let V be a vector space over \mathbb{R} or \mathbb{C} . Two norms on V , denoted $\|\cdot\|_V$ and $\|\cdot\|_{V'}$, are called **equivalent norms** if, for all $\underline{v} \in V$, there are positive real numbers m, M such that

$$m\|\underline{v}\|_{V'} \leq \|\underline{v}\|_V \leq M\|\underline{v}\|_{V'}. \quad (3.5)$$

Lemma 3.5.19. For finite-dimensional normed spaces, all norms are equivalent.

Proof. Let V be a finite-dimensional normed space over F with norm $\|\cdot\|_V$. Then, V has an ordered basis $\mathcal{B} = \underline{v}_1, \dots, \underline{v}_n$. We can also define a new norm

$$\|\underline{v}\|_{V'} = \|\underline{v}\|_{\mathcal{B}} = \sum_{i=1}^n |s_i|,$$

where $\underline{v}\|_{\mathcal{B}} = (s_1, \dots, s_n)$ is the unique coordinate vector such that $\underline{v} = s_1 \underline{v}_1 + \dots + s_n \underline{v}_n$ and $\|\underline{s}\|_1$ is the 1-norm of $\underline{s} \in F^n$. It is easy to verify that this is a valid norm.

Since equivalence is a transitive property of norms and $\|\cdot\|_V$ is arbitrary, it is sufficient to show that $\|\cdot\|_V$ is equivalent to $\|\cdot\|_{V'}$. First, we observe that

$$\begin{aligned}\|\underline{v}\|_V &= \|s_1\underline{v}_1 + \cdots + s_n\underline{v}_n\|_V \\ &\leq \|s_1\underline{v}_1\|_V + \cdots + \|s_n\underline{v}_n\|_V \\ &= \sum_{i=1}^n |s_i| \|\underline{v}_i\|_V \\ &\leq M \sum_{i=1}^n |s_i|,\end{aligned}$$

where $M = \max_{i \in \{1, \dots, n\}} \|\underline{v}_i\|_V$.

To show the lower bound, we first consider the case where the coefficients $\{s_1, \dots, s_n\}$ satisfy the condition $|s_1| + \cdots + |s_n| = 1$. Let

$$S = \left\{ (s_1, \dots, s_n) \mid \sum_{j=1}^n |s_j| = 1 \right\}.$$

This set is compact because it is a closed and totally bounded subset of the complete space F^n . Define the function $f: S \rightarrow \mathbb{R}$ by

$$f(s_1, \dots, s_n) = \|s_1\underline{v}_1 + \cdots + s_n\underline{v}_n\|_V.$$

The function f is continuous because

$$\begin{aligned}|f(s_1, \dots, s_n) - f(t_1, \dots, t_n)| &= \left| \left\| \sum_{j=1}^n s_j \underline{v}_j \right\|_V - \left\| \sum_{j=1}^n t_j \underline{v}_j \right\|_V \right| \\ &\leq \left\| \sum_{j=1}^n s_j \underline{v}_j - \sum_{j=1}^n t_j \underline{v}_j \right\|_V \\ &\leq M \sum_{j=1}^n |s_j - t_j|.\end{aligned}$$

Now, let the minimum of f over S be the non-negative real number

$$m = \min_{(s_1, \dots, s_n) \in S} f(s_1, \dots, s_n).$$

Since f is continuous and S is compact, this minimum exists and is attained by some point $(s'_1, \dots, s'_n) \in S$. Moreover, $\|\underline{v}\|_V \geq m$ for all $\underline{v} \in f(S)$. Note

that we must have $m > 0$ because otherwise $\underline{v}_1, \dots, \underline{v}_n$ are linearly dependent, contradicting the fact that \mathcal{B} is a basis. Since $|s_1| + \dots + |s_n| = 1$, this value of m satisfies $\|\underline{v}\|_V \geq m\|\underline{v}\|_{V'}$ for all $(s_1, \dots, s_n) \in S$.

For general sets of coefficients $\{s_i\}$, let $c = |s_1| + \dots + |s_n|$. If $c = 0$, the result is trivial. If $c > 0$, then write

$$\begin{aligned} \|s_1\underline{v}_1 + \dots + s_n\underline{v}_n\|_V &= c \left\| \frac{s_1}{c}\underline{v}_1 + \dots + \frac{s_n}{c}\underline{v}_n \right\|_V \\ &= cf \left(\frac{s_1}{c}, \dots, \frac{s_n}{c} \right) \\ &\geq cm \\ &= m(|s_1| + \dots + |s_n|). \end{aligned}$$

This completes the proof of (3.5) for the constructed values of m and M . \square

Definition 3.5.20. A *closed subspace* of a Banach space is a subspace that is a closed set in the topology generated by the norm.

Theorem 3.5.21. All finite-dimensional subspaces of a Banach space are closed and all finite-dimensional normed spaces are complete.

Proof. Consider a Banach space V with norm $\|\underline{v}\|_V$ and a subspace W satisfying $\dim(W) = n < \infty$. Then, W has an ordered basis $\mathcal{B} = \underline{v}_1, \dots, \underline{v}_n$. Let \underline{w}_k be a sequence in W and observe that there are n sequences $s_k^{(i)}$ such that

$$\underline{w}_k = \sum_{i=1}^n s_k^{(i)} \underline{v}_i.$$

By Lemma 3.5.19, the norm $\|\underline{v}\|_{V'} = \|[\cdot]_{\mathcal{B}}\|_1$ is equivalent to $\|\cdot\|_V$. It follows that there is a positive real constant m such that

$$|s_k^{(i)} - s_l^{(i)}| \leq \sum_{j=1}^n |s_k^{(j)} - s_l^{(j)}| \leq \frac{1}{m} \|\underline{w}_k - \underline{w}_l\|_V$$

for all i, k, l .

To show W is closed, we assume the sequence \underline{w}_k converges in V which implies that it is Cauchy in V . Then, we combine the fact that \underline{w}_k is Cauchy in V with the previous inequality to see that, for all i , the sequence $s_k^{(i)}$ is Cauchy. Since the

scalar field (either \mathbb{R} or \mathbb{C}) is complete, it follows that $s_k^{(i)}$ converges and we denote its limit by t_i . Using the continuity of norm, we observe that

$$0 = \lim_{k \rightarrow \infty} \left\| \underline{w}_k - \sum_{i=1}^n s_k^{(i)} \underline{v}_i \right\|_V = \left\| \underline{w} - \sum_{i=1}^n t_i \underline{v}_i \right\|_V.$$

Thus, $\underline{w} = \sum_{i=1}^n t_i \underline{v}_i \in W$ and we see that W is closed.

For the completeness result, we assume that V is finite dimensional and choose $W = V$. Then, instead of assuming that \underline{w}_k converges in V , we assume that \underline{w}_k is Cauchy in V by hypothesis. The argument above shows that \underline{w}_k converges in V and it follows that V is complete. □

Example 3.5.22. Let $V = L^p([a, b])$, for $1 \leq p < \infty$, be the set of real Lebesgue-integrable functions on $[a, b]$. We say that $f \in V$ is continuous if the equivalence class generated by equality almost everywhere contains a continuous function. It is easy to verify that the subset $W \subset V$ of continuous functions is a subspace. It is not closed, however, because sequences in W may converge to discontinuous functions. In fact, the set of continuous functions is dense in $L^p([a, b])$ for $p \in [1, \infty)$.

Example 3.5.23. Let $W = \{\underline{w}_1, \underline{w}_2, \dots\}$ be a linearly independent sequence of normalized vectors in a Banach space. The span of W only includes finite linear combinations. However, a sequence of finite linear combinations, like

$$\underline{u}_n = \sum_{i=1}^n \frac{1}{i^2} \underline{w}_i,$$

converges to the infinite linear combination $\underline{u} = \lim_{n \rightarrow \infty} \underline{u}_n$ if the limit exists. Applying Lemma 3.5.16 to $\underline{v}_i = \frac{1}{i^2} \underline{w}_i$ shows that the limit exists if $\sum_{i=1}^{\infty} i^{-2} < \infty$ and this can be shown using the integral test. Thus, the span of any infinite set of linearly independent vectors is not closed.

Theorem 3.5.24 (Hölder and Minkowski Inequalities). Consider the following weighted versions of the ℓ^p and L^p norms defined by

$$\|\underline{v}\|_{\ell^p(\underline{w})} = \left(\sum_{i=1}^n w_i |v_i|^p \right)^{\frac{1}{p}}$$

$$\|f\|_{L^p(X, w)} = \left(\int_X w(x) |f(x)|^p dx \right)^{\frac{1}{p}},$$

where the vector \underline{w} and function $w(x)$ define real positive weights and X is chosen so that the Lebesgue integral is well-defined. For $p \in [1, \infty)$, the Minkowski inequality states that

$$\begin{aligned}\|\underline{u} + \underline{v}\|_{\ell^p(\underline{w})} &\leq \|\underline{u}\|_{\ell^p(\underline{w})} + \|\underline{v}\|_{\ell^p(\underline{w})} \\ \|f + g\|_{L^p(X,w)} &\leq \|f\|_{L^p(X,w)} + \|g\|_{L^p(X,w)}.\end{aligned}$$

Choose $p, q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$, where $1/\infty = 0$. For the ℓ^p/ℓ^q case, assume $\underline{u} \in \ell^p(\underline{w})$, $\underline{v} \in \ell^q(\underline{w})$, and define the product vector $\underline{t} = (u_1v_1, \dots, u_nv_n)$. For the L^p/L^q case, assume that $f \in L^p(X, w)$ (i.e., $\|f\|_{L^p(X,w)} < \infty$), $g \in L^q(X, w)$ (i.e., $\|g\|_{L^q(X,w)} < \infty$), and define the product function $h(x) = f(x)g(x)$. Then, the Hölder inequality states that

$$\begin{aligned}\|\underline{t}\|_{\ell^1(\underline{w})} &\leq \|\underline{u}\|_{\ell^p(\underline{w})} \|\underline{v}\|_{\ell^q(\underline{w})} \\ \|h\|_{L^1(X,w)} &\leq \|f\|_{L^p(X,w)} \|g\|_{L^q(X,w)}.\end{aligned}$$

3.6 Inner Products

An *inner product space* is a vector space with an *inner product*, defined below. Such spaces have an inherent structure that admits the rigorous development of geometrical notions such as the length of a vector and the angle between two vectors. They offer a means to define the projection of a vector onto another vector, and the orthogonality of a pair of vectors. In some sense, inner product spaces generalize Euclidean spaces with their dot products. Importantly, an inner product naturally induces a norm and, hence, a metric. Consequently, they are necessarily equipped with a metric topology, open sets, and a definition of convergence.

In practice, one often encounters inner product space that are finite dimensional. Therein, vectors can be expressed as linear combinations of elements within a Hamel basis. Yet, an inner product space can also be of infinite dimension. Hilbert spaces are complete inner product spaces, which makes them suitable to discuss Cauchy sequences of vectors and their limits. Furthermore, when an inner product space is separable (in the induced metric topology), then it will have a countable Schauder basis. Intuitively, a separable Hilbert space with its Schauder basis can be handled mathematically in a way similar to a finite-dimensional inner product space. For this reason, our treatment of infinite-dimensional inner product spaces is

largely restricted to *separable Hilbert spaces*. More general spaces require a higher level of mathematical sophistication, beyond this treatment.

Definition 3.6.1. Let F be the field of real numbers or the field of complex numbers, and assume V is a vector space over F . An **inner product** on V is a function which assigns to each ordered pair of vectors $\underline{v}, \underline{w} \in V$ a scalar $\langle \underline{v}, \underline{w} \rangle \in F$ in such a way that for all $\underline{u}, \underline{v}, \underline{w} \in V$ and any scalar $s \in F$

1. $\langle \underline{u} + \underline{v}, \underline{w} \rangle = \langle \underline{u}, \underline{w} \rangle + \langle \underline{v}, \underline{w} \rangle$
2. $\langle s\underline{v}, \underline{w} \rangle = s\langle \underline{v}, \underline{w} \rangle$
3. $\langle \underline{v}, \underline{w} \rangle = \overline{\langle \underline{w}, \underline{v} \rangle}$, where the overbar denotes complex conjugation;
4. $\langle \underline{v}, \underline{v} \rangle \geq 0$ with equality iff $\underline{v} = \underline{0}$.

We emphasize that the conditions of Definition 3.6.1 imply

$$\langle \underline{u}, s\underline{v} + \underline{w} \rangle = \bar{s}\langle \underline{u}, \underline{v} \rangle + \langle \underline{u}, \underline{w} \rangle.$$

Thus, an inner product is linear in the first argument and conjugate linear in the second argument.

Definition 3.6.2. A real or complex vector space equipped with an inner product is called an **inner-product space**.

Example 3.6.3. Consider the inner product on F^n defined by

$$\langle \underline{v}, \underline{w} \rangle = \langle (v_1, \dots, v_n), (w_1, \dots, w_n) \rangle = \sum_{j=1}^n v_j \bar{w}_j.$$

This inner product is called the **standard inner product**. When $F = \mathbb{R}$, the standard inner product can also be written as

$$\langle \underline{v}, \underline{w} \rangle = \sum_{j=1}^n v_j w_j.$$

In this context it is often called the **dot product**, denoted by $\underline{v} \cdot \underline{w}$. In either case, it can also be written in terms of the Hermitian transpose as $\langle \underline{v}, \underline{w} \rangle = \underline{w}^H \underline{v}$.

Problem 3.6.4. For $\underline{v} = (v_1, v_2)$ and $\underline{w} = (w_1, w_2)$ in \mathbb{R}^2 , show that

$$\langle \underline{v}, \underline{w} \rangle = v_1 w_1 - v_2 w_1 - v_1 w_2 + 4v_2 w_2$$

is an inner product.

S 3.6.4. For all $\underline{u}, \underline{v}, \underline{w} \in V$ and all scalars s

$$\begin{aligned} \langle \underline{u} + \underline{v}, \underline{w} \rangle &= (u_1 + v_1)w_1 - (u_2 + v_2)w_1 - (u_1 + v_1)w_2 + 4(u_2 + v_2)w_2 \\ &= u_1 w_1 - u_2 w_1 - u_1 w_2 + 4u_2 w_2 + v_1 w_1 - v_2 w_1 - v_1 w_2 + 4v_2 w_2 \\ &= \langle \underline{u}, \underline{w} \rangle + \langle \underline{v}, \underline{w} \rangle. \end{aligned}$$

Also, we have

$$\langle s\underline{v}, \underline{w} \rangle = sv_1 w_1 - sv_2 w_1 - sv_1 w_2 + 4sv_2 w_2 = s\langle \underline{v}, \underline{w} \rangle.$$

Since $V = \mathbb{R}^2$, we have $\langle \underline{v}, \underline{w} \rangle = \overline{\langle \underline{w}, \underline{v} \rangle}$. Furthermore,

$$\langle \underline{v}, \underline{v} \rangle = v_1^2 - 2v_1 v_2 + 4v_2^2 = (v_1 - v_2)^2 + 3v_2^2 \geq 0 \quad \text{with equality iff } \underline{v} = \underline{0}.$$

That is, $\langle \underline{v}, \underline{v} \rangle$ is an inner product.

Example 3.6.5. Let V be the vector space of all continuous complex-valued functions on the unit interval $[0, 1]$. Then,

$$\langle f, g \rangle = \int_0^1 f(t) \overline{g(t)} dt$$

is an inner product.

Example 3.6.6. Let V and W be two vector spaces over F and suppose that $\langle \cdot, \cdot \rangle_W$ is an inner product on W . If T is a non-singular linear transformation from V into W , then the equation

$$\langle \underline{v}_1, \underline{v}_2 \rangle_V = \langle T\underline{v}_1, T\underline{v}_2 \rangle_W$$

defines an inner product on V .

Example 3.6.7. Let $V = F^{m \times n}$ be the space of $m \times n$ matrices over F and define the inner product for matrices $A, B \in V$ to be

$$\langle A, B \rangle \triangleq \text{tr}(B^H A) = \sum_{i=1}^n \sum_{j=1}^m \bar{b}_{j,i} a_{j,i}.$$

This also equals $\text{tr}(AB^H)$ and both are identical to writing the entries of the matrices as length- mn vectors and then applying the standard inner product.

Theorem 3.6.8. *Let V be a finite-dimensional space, and suppose that*

$$\mathcal{B} = \underline{w}_1, \dots, \underline{w}_n$$

is an ordered basis for V . Any inner product on V is determined by the values

$$g_{ij} = \langle \underline{w}_j, \underline{w}_i \rangle$$

that it takes on pairs of vectors in \mathcal{B} .

Proof. If $\underline{u} = \sum_j s_j \underline{w}_j$ and $\underline{v} = \sum_i t_i \underline{w}_i$, then

$$\begin{aligned} \langle \underline{u}, \underline{v} \rangle &= \left\langle \sum_j s_j \underline{w}_j, \underline{v} \right\rangle = \sum_j s_j \langle \underline{w}_j, \underline{v} \rangle \\ &= \sum_j s_j \left\langle \underline{w}_j, \sum_i t_i \underline{w}_i \right\rangle = \sum_j \sum_i s_j \bar{t}_i \langle \underline{w}_j, \underline{w}_i \rangle \\ &= \sum_j \sum_i \bar{t}_i g_{ij} s_j = [\underline{v}]_{\mathcal{B}}^H G [\underline{u}]_{\mathcal{B}} \end{aligned}$$

where $[\underline{u}]_{\mathcal{B}}$ and $[\underline{v}]_{\mathcal{B}}$ are the coordinate matrices of $\underline{u}, \underline{v}$ in the ordered basis \mathcal{B} . The matrix G is called the *weight matrix* of the inner product in the ordered basis \mathcal{B} . \square

Since $g_{ij} = \langle \underline{w}_j, \underline{w}_i \rangle = \overline{\langle \underline{w}_i, \underline{w}_j \rangle} = \bar{g}_{ji}$, we see that G is a Hermitian matrix, i.e., $G = G^H$. Furthermore, G must satisfy the additional condition

$$\underline{w}^H G \underline{w} > 0, \quad \forall \underline{w} \neq \underline{0} \quad (3.6)$$

so that the induced norm is non-negative and zero only for the zero vector. A Hermitian matrix that satisfies this condition is called positive definite and this also implies that G is invertible.

Conversely, if G is an $n \times n$ Hermitian matrix over F which satisfies (3.6), then

$$\langle \underline{u}, \underline{v} \rangle_G = [\underline{v}]_{\mathcal{B}}^H G [\underline{u}]_{\mathcal{B}}$$

is a well-defined inner product on V .

Problem 3.6.9. *Let V be a vector space over F . Show that the sum of two inner products on V is an inner product on V . Show that a positive multiple of an inner product is also an inner product.*

Example 3.6.10. Consider any set W of real-valued random variables, defined on a common probability space, that have finite 2nd moments. It turns out that $V = \text{span}(W)$ is a vector space over \mathbb{R} . In fact, one can define the inner product

$$\langle X, Y \rangle = E[XY],$$

for any $X, Y \in V$. Using the induced norm, this inner product provides the topology of mean-square convergence and two random variables $X, Y \in V$ are considered equal if $\|X - Y\|^2 = E[|X - Y|^2] = 0$ (or equivalently $\Pr(X \neq Y) = 0$).

In terms of abstract mathematics, the introduction of an inner product allows one to introduce the key concept of orthogonality.

Definition 3.6.11. Let \underline{v} and \underline{w} be vectors in an inner-product space V . Then \underline{v} is **orthogonal** to \underline{w} (denoted $\underline{v} \perp \underline{w}$) if $\langle \underline{v}, \underline{w} \rangle = 0$. Since this relation is reflexive and \underline{w} is also orthogonal to \underline{v} , we simply say that \underline{v} and \underline{w} are orthogonal.

3.6.1 Induced Norms

A finite-dimensional real inner-product space is often referred to as a **Euclidean space**. A complex inner-product space is sometimes called a unitary space.

Definition 3.6.12. Let V be an inner-product space with inner product $\langle \cdot, \cdot \rangle$. This inner product can be used to define a norm, called the **induced norm**, where

$$\|\underline{v}\| = \langle \underline{v}, \underline{v} \rangle^{\frac{1}{2}}$$

for every $\underline{v} \in V$.

Definition 3.6.13. Let $\underline{w}, \underline{v}$ be vectors in an inner-product space V with inner product $\langle \cdot, \cdot \rangle$. As shown in Figure 3.1, the **projection** of \underline{w} onto \underline{v} is defined to be

$$\underline{u} = \frac{\langle \underline{w}, \underline{v} \rangle}{\|\underline{v}\|^2} \underline{v}$$

Lemma 3.6.14. Let \underline{u} be the projection of \underline{w} onto \underline{v} . Then, $\langle \underline{w} - \underline{u}, \underline{u} \rangle = 0$ and

$$\|\underline{w} - \underline{u}\|^2 = \|\underline{w}\|^2 - \|\underline{u}\|^2 = \|\underline{w}\|^2 - \frac{|\langle \underline{w}, \underline{v} \rangle|^2}{\|\underline{v}\|^2}.$$

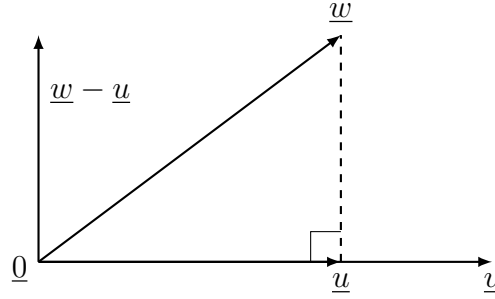


Figure 3.1: The projection of \underline{w} onto \underline{v} is given by \underline{u} and $\underline{w} - \underline{u}$ is orthogonal to \underline{v} .

Proof. First, we observe that

$$\langle \underline{w} - \underline{u}, \underline{v} \rangle = \langle \underline{w}, \underline{v} \rangle - \langle \underline{u}, \underline{v} \rangle = \langle \underline{w}, \underline{v} \rangle - \frac{\langle \underline{w}, \underline{v} \rangle}{\|\underline{v}\|^2} \langle \underline{v}, \underline{v} \rangle = 0.$$

Since $\underline{u} = s\underline{v}$ for some scalar s , we gather that $\langle \underline{w} - \underline{u}, \underline{u} \rangle = \langle \underline{w} - \underline{u}, s\underline{v} \rangle = s\langle \underline{w} - \underline{u}, \underline{v} \rangle = 0$. Using $\langle \underline{w} - \underline{u}, \underline{u} \rangle = 0$, we can write

$$\begin{aligned} \|\underline{w}\|^2 &= \|(\underline{w} - \underline{u}) + \underline{u}\|^2 = \langle (\underline{w} - \underline{u}) + \underline{u}, (\underline{w} - \underline{u}) + \underline{u} \rangle \\ &= \|\underline{w} - \underline{u}\|^2 + 2\operatorname{Re}\langle \underline{w} - \underline{u}, \underline{u} \rangle + \|\underline{u}\|^2 = \|\underline{w} - \underline{u}\|^2 + \|\underline{u}\|^2. \end{aligned}$$

The proof is completed by noting that $\|\underline{u}\|^2 = \frac{|\langle \underline{w}, \underline{v} \rangle|^2}{\|\underline{v}\|^4} \langle \underline{v}, \underline{v} \rangle = \frac{|\langle \underline{w}, \underline{v} \rangle|^2}{\|\underline{v}\|^2}$. \square

Theorem 3.6.15. *If V is an inner-product space and $\|\cdot\|$ is its associated induced norm, then for any $\underline{v}, \underline{w} \in V$ and any scalar s*

1. $\|s\underline{v}\| = |s| \|\underline{v}\|$
2. $\|\underline{v}\| > 0$ for $\underline{v} \neq \underline{0}$
3. $|\langle \underline{v}, \underline{w} \rangle| \leq \|\underline{v}\| \|\underline{w}\|$ with equality iff $\underline{v} = \underline{0}$, $\underline{w} = \underline{0}$, or $\underline{v} = s\underline{w}$
4. $\|\underline{v} + \underline{w}\| \leq \|\underline{v}\| + \|\underline{w}\|$ with equality iff $\underline{v} = \underline{0}$, $\underline{w} = \underline{0}$, or $\underline{v} = s\underline{w}$.

Proof. The first two properties follow immediately from the definitions involved. The third property, $|\langle \underline{v}, \underline{w} \rangle| \leq \|\underline{v}\| \|\underline{w}\|$, is called the **Cauchy-Schwarz inequality**. When $\underline{v} = \underline{0}$, then clearly $|\langle \underline{v}, \underline{w} \rangle| = \|\underline{v}\| \|\underline{w}\| = 0$. Assume $\underline{v} \neq \underline{0}$ and let

$$\underline{u} = \frac{\langle \underline{w}, \underline{v} \rangle}{\|\underline{v}\|^2} \underline{v}$$

be the projection \underline{w} onto \underline{v} . By Lemma 3.6.14, we have

$$0 \leq \|\underline{w} - \underline{u}\|^2 = \|\underline{w}\|^2 - \frac{|\langle \underline{w}, \underline{v} \rangle|^2}{\|\underline{v}\|^2}, \quad (3.7)$$

where equality holds iff $\underline{w} - \underline{u} = \underline{0}$, or equivalently iff $\underline{w} = \underline{0}$ or $\underline{v} = s\underline{w}$. Rearranging (3.7) by isolating the inner product component and cross-multiplying, we find that

$$|\langle \underline{v}, \underline{w} \rangle|^2 = |\langle \underline{w}, \underline{v} \rangle|^2 \leq \|\underline{v}\|^2 \|\underline{w}\|^2,$$

with equality iff $\underline{v} = \underline{0}$, $\underline{w} = \underline{0}$, or $\underline{v} = s\underline{w}$. Using this, we get the fourth property

$$\begin{aligned} \|\underline{v} + \underline{w}\|^2 &= \|\underline{v}\|^2 + \langle \underline{v}, \underline{w} \rangle + \langle \underline{w}, \underline{v} \rangle + \|\underline{w}\|^2 \\ &= \|\underline{v}\|^2 + 2\operatorname{Re}\langle \underline{v}, \underline{w} \rangle + \|\underline{w}\|^2 \\ &\leq \|\underline{v}\|^2 + 2|\langle \underline{v}, \underline{w} \rangle| + \|\underline{w}\|^2 \\ &\leq \|\underline{v}\|^2 + 2\|\underline{v}\|\|\underline{w}\| + \|\underline{w}\|^2, \end{aligned}$$

where the first inequality holds with equality iff $\operatorname{Re}\langle \underline{v}, \underline{w} \rangle = |\langle \underline{v}, \underline{w} \rangle|$ (i.e., $\langle \underline{v}, \underline{w} \rangle$ is real and non-negative) and the second inequality (i.e., Cauchy-Schwarz) holds with equality. Thus, $\|\underline{v} + \underline{w}\| \leq \|\underline{v}\| + \|\underline{w}\|$ with equality iff $\underline{v} = \underline{0}$, $\underline{w} = \underline{0}$, or $\underline{v} = s\underline{w}$ for real $s \geq 0$. \square

Theorem 3.6.16. *Consider the vector space \mathbb{R}^n with the standard inner product. Then, the function $f: V \rightarrow F$ defined by $f(\underline{w}) = \langle \underline{w}, \underline{v} \rangle$ is continuous.*

Proof. Let $\underline{w}_1, \underline{w}_2, \dots$ be a sequence in V converging to \underline{w} . Then,

$$|\langle \underline{w}_n, \underline{v} \rangle - \langle \underline{w}, \underline{v} \rangle| = |\langle \underline{w}_n - \underline{w}, \underline{v} \rangle| \leq \|\underline{w}_n - \underline{w}\| \|\underline{v}\|.$$

Since $\|\underline{w}_n - \underline{w}\| \rightarrow 0$, the convergence of $\langle \underline{w}_n, \underline{v} \rangle$ is established. \square

3.7 Sets of Orthogonal Vectors

Definition 3.7.1. *Let V be an inner-product space and U, W be subspaces. Then, the subspace U is an **orthogonal** to the subspace W (denoted $U \perp W$) if $\underline{u} \perp \underline{w}$ for all $\underline{u} \in U$ and $\underline{w} \in W$.*

Definition 3.7.2. *A collection W of vectors in V is an **orthogonal set** if all pairs of distinct vectors in W are orthogonal.*

Example 3.7.3. *The standard basis of \mathbb{R}^n is an orthonormal set with respect to the standard inner product.*

Example 3.7.4. *Let V be the vector space (over \mathbb{C}) of continuous complex-valued functions on the interval $0 \leq x \leq 1$ with the inner product*

$$\langle f, g \rangle = \int_0^1 f(x)\overline{g(x)}dx.$$

Let $f_n(x) = \sqrt{2} \cos 2\pi nx$ and $g_n(x) = \sqrt{2} \sin 2\pi nx$. Then $\{1, f_1, g_1, f_2, g_2, \dots\}$ is a countably infinite orthonormal set that is a Schauder basis for this vector space.

Theorem 3.7.5. *An orthogonal set of non-zero vectors is linearly independent.*

Proof. Let W be an orthogonal set of non-zero vectors in a given inner-product space V . Suppose $\underline{w}_1, \dots, \underline{w}_n$ are distinct vectors in W and consider

$$\underline{v} = s_1\underline{w}_1 + \dots + s_n\underline{w}_n.$$

The inner product $\langle \underline{v}, \underline{w}_i \rangle$ is given by

$$\langle \underline{v}, \underline{w}_i \rangle = \left\langle \sum_j s_j \underline{w}_j, \underline{w}_i \right\rangle = \sum_j s_j \langle \underline{w}_j, \underline{w}_i \rangle = s_i \langle \underline{w}_i, \underline{w}_i \rangle.$$

Since $\langle \underline{w}_i, \underline{w}_i \rangle \neq 0$, it follows that

$$s_i = \frac{\langle \underline{v}, \underline{w}_i \rangle}{\|\underline{w}_i\|^2} \quad 1 \leq i \leq n.$$

In particular, if $\underline{v} = 0$ then $s_j = 0$ for $1 \leq j \leq n$ and the vectors in W are linearly independent. \square

Corollary 3.7.6. *If $\underline{v} \in V$ is a linear combination of an orthogonal sequence of distinct, non-zero vectors $\underline{w}_1, \dots, \underline{w}_n$, then \underline{v} satisfies the identity*

$$\underline{v} = \sum_{i=1}^n \frac{\langle \underline{v}, \underline{w}_i \rangle}{\|\underline{w}_i\|^2} \underline{w}_i,$$

and equals the sum of the projections of \underline{v} onto $\underline{w}_1, \dots, \underline{w}_n$.

Theorem 3.7.7. *Let V be an inner-product space and assume $\underline{v}_1, \dots, \underline{v}_n$ are linearly independent vectors in V . Then it is possible to construct an orthogonal sequence of vectors $\underline{w}_1, \dots, \underline{w}_n \in V$ such that for each $k = 1, \dots, n$ the set*

$$\{\underline{w}_1, \dots, \underline{w}_k\}$$

is a basis for the subspace spanned by $\underline{v}_1, \dots, \underline{v}_k$.

Proof. First, let $\underline{w}_1 = \underline{v}_1$. The remaining vectors are defined inductively as part during the proof. Suppose the vectors

$$\underline{w}_1, \dots, \underline{w}_m \quad (1 \leq m < n)$$

have been chosen so that for every k

$$\{\underline{w}_1, \dots, \underline{w}_k\} \quad 1 \leq k \leq m$$

is an orthogonal basis for the subspace spanned by $\underline{v}_1, \dots, \underline{v}_k$. Let

$$\underline{w}_{m+1} = \underline{v}_{m+1} - \sum_{i=1}^m \frac{\langle \underline{v}_{m+1}, \underline{w}_i \rangle}{\|\underline{w}_i\|^2} \underline{w}_i.$$

Then $\underline{w}_{m+1} \neq 0$, for otherwise \underline{v}_{m+1} is a linear combination of $\underline{w}_1, \dots, \underline{w}_m$ and hence a linear combination of $\underline{v}_1, \dots, \underline{v}_m$. For $j \in \{1, \dots, m\}$, we also have

$$\begin{aligned} \langle \underline{w}_{m+1}, \underline{w}_j \rangle &= \langle \underline{v}_{m+1}, \underline{w}_j \rangle - \sum_{i=1}^m \frac{\langle \underline{v}_{m+1}, \underline{w}_i \rangle}{\|\underline{w}_i\|^2} \langle \underline{w}_i, \underline{w}_j \rangle \\ &= \langle \underline{v}_{m+1}, \underline{w}_j \rangle - \frac{\langle \underline{v}_{m+1}, \underline{w}_j \rangle}{\|\underline{w}_j\|^2} \langle \underline{w}_j, \underline{w}_j \rangle \\ &= 0. \end{aligned}$$

Clearly, $\{\underline{w}_1, \dots, \underline{w}_{m+1}\}$ is an orthogonal set consisting of $m + 1$ non-zero vectors in the subspace spanned by $\underline{v}_1, \dots, \underline{v}_{m+1}$. Since the dimension of the latter subspace is $m + 1$, this set is a basis for the subspace. \square

The inductive construction of the vectors $\underline{w}_1, \dots, \underline{w}_n$ is known as the **Gram-Schmidt orthogonalization** process.

Corollary 3.7.8. *Every finite-dimensional inner-product space has a basis of orthonormal vectors.*

Proof. Let V be a finite-dimensional inner-product space. Suppose that $\underline{v}_1, \dots, \underline{v}_n$ is a basis for V . Apply the Gram-Schmidt process to obtain a basis of orthogonal vectors $\underline{w}_1, \dots, \underline{w}_n$. Then, a basis of orthonormal vectors is given by

$$\underline{u}_1 = \frac{\underline{w}_1}{\|\underline{w}_1\|}, \dots, \underline{u}_n = \frac{\underline{w}_n}{\|\underline{w}_n\|}.$$

□

Example 3.7.9. Consider the vectors

$$\underline{v}_1 = (2, 2, 1)$$

$$\underline{v}_2 = (3, 6, 0)$$

$$\underline{v}_3 = (6, 3, 9)$$

in \mathbb{R}^3 equipped with the standard inner product. Apply the Gram-Schmidt process to $\underline{v}_1, \underline{v}_2, \underline{v}_3$ to obtain an orthogonal basis.

Applying the Gram-Schmidt process to $\underline{v}_1, \underline{v}_2, \underline{v}_3$, we get

$$\underline{w}_1 = (2, 2, 1)$$

$$\underline{w}_2 = (3, 6, 0) - \frac{\langle (3, 6, 0), (2, 2, 1) \rangle}{9} (2, 2, 1)$$

$$= (3, 6, 0) - 2(2, 2, 1) = (-1, 2, -2)$$

$$\underline{w}_3 = (6, 3, 9) - \frac{\langle (6, 3, 9), (2, 2, 1) \rangle}{9} (2, 2, 1) - \frac{\langle (6, 3, 9), (-1, 2, -2) \rangle}{9} (-1, 2, -2)$$

$$= (6, 3, 9) - 3(2, 2, 1) + 2(-1, 2, -2) = (-2, 1, 2).$$

It is easily verified that $\underline{w}_1, \underline{w}_2, \underline{w}_3$ is an orthogonal set of vectors.

Definition 3.7.10. Let V be an inner-product space and W be any set of vectors in V . The **orthogonal complement** of W denoted by W^\perp is the set of all vectors in V that are orthogonal to every vector in W or

$$W^\perp = \{ \underline{v} \in V \mid \langle \underline{v}, \underline{w} \rangle = 0 \forall \underline{w} \in W \}.$$

Problem 3.7.11. Let W be any subset of vector space V . Show that W^\perp is a closed subspace of V and that any vector in the subspace spanned by W is orthogonal to any vector in W^\perp .

S 3.7.11. Let $\underline{m}_1, \underline{m}_2 \in W^\perp$ and $s \in F$. For any vector $\underline{w} \in W$, we have

$$\langle \underline{m}_1, \underline{w} \rangle = \langle \underline{m}_2, \underline{w} \rangle = 0.$$

This implies

$$\langle s\underline{m}_1 + \underline{m}_2, \underline{w} \rangle = s\langle \underline{m}_1, \underline{w} \rangle + \langle \underline{m}_2, \underline{w} \rangle = 0.$$

That is, $s\underline{m}_1 + \underline{m}_2 \in W^\perp$. Hence, W^\perp is a subspace of V .

To see that W^\perp is closed, we let \underline{m} be any point in the closure of W^\perp and $\underline{m}_1, \underline{m}_2, \dots \in W^\perp$ be a sequence that converges to \underline{m} . The continuity of the inner product, from Theorem 3.6.16, implies that, for all $\underline{w} \in W$,

$$\langle \underline{m}, \underline{w} \rangle = \left\langle \lim_{n \rightarrow \infty} \underline{m}_n, \underline{w} \right\rangle = \lim_{n \rightarrow \infty} \langle \underline{m}_n, \underline{w} \rangle = 0.$$

Therefore, $\underline{m} \in W^\perp$ and the orthogonal complement contains all of its limit points.

Notice also that any vector \underline{w} in the subspace spanned by W can be written as $\underline{w} = \sum_i s_i \underline{w}_i$ with $\underline{w}_i \in W$ and $s_i \in F$. Therefore, the inner product of \underline{w} with any $\underline{w}' \in W^\perp$ is given by

$$\langle \underline{w}, \underline{w}' \rangle = \left\langle \sum_i s_i \underline{w}_i, \underline{w}' \right\rangle = \sum_i s_i \langle \underline{w}_i, \underline{w}' \rangle = 0.$$

It follows that the subspace spanned by W is orthogonal to the subspace W^\perp .

Definition 3.7.12. A complex matrix $U \in \mathbb{C}^{n \times n}$ is called **unitary** if $U^H U = I$. Similarly, a real matrix $Q \in \mathbb{R}^{n \times n}$ is called **orthogonal** if $Q^T Q = I$.

Theorem 3.7.13. Let $V = \mathbb{C}^n$ be the standard inner product space and let $U \in \mathbb{C}^{n \times n}$ define a linear operator on V . Then, the following conditions are equivalent:

- (i) The columns of U form an orthonormal basis (i.e., $U^H U = I$),
- (ii) the rows of U form an orthonormal basis (i.e., $U U^H = I$),
- (iii) U preserves inner products (i.e., $\langle U \underline{v}, U \underline{w} \rangle = \langle \underline{v}, \underline{w} \rangle$ for all $\underline{u}, \underline{v} \in V$), and
- (iv) U is an isometry (i.e., $\|U \underline{v}\| = \|\underline{v}\|$ for all $\underline{v} \in V$).

Proof. If (i) holds, then U is invertible because its columns are linearly independent. Thus, $U^H U = I$ implies $U^H = U^{-1}$ and (ii) follows. Likewise, (iii) holds because $\langle U \underline{v}, U \underline{w} \rangle = \underline{w}^H U^H U \underline{v} = \underline{w}^H \underline{v} = \langle \underline{v}, \underline{w} \rangle$ for all $\underline{u}, \underline{v} \in V$. Choosing $\underline{w} = \underline{v}$ gives (iv). Lastly, if $\|U \underline{v}\| = \|\underline{v}\|$ for all $\underline{v} \in V$, then $\underline{v}^H (U^H U - I) \underline{v} = \|U \underline{v}\|^2 - \|\underline{v}\|^2 = 0$ for all $\underline{v} \in V$. By Theorem 9.1.2, $U^H U - I$ is diagonalizable because it is Hermitian. Thus, all its eigenvalues must be 0 and $U^H U - I = 0$. \square

3.7.1 Hilbert Spaces

Definition 3.7.14. A complete inner-product space is called a **Hilbert space**.

Definition 3.7.15. Recall that a subset $\{\underline{v}_\alpha | \alpha \in A\}$ of a Hilbert space V is said to be orthonormal if $\|\underline{v}_\alpha\| = 1$ for every $\alpha \in A$ and $\langle \underline{v}_\alpha, \underline{v}_\beta \rangle = 0$ for all $\alpha \neq \beta$. If the subspace spanned by the family $\{\underline{v}_\alpha | \alpha \in A\}$ is dense in V , we call this set an **orthonormal basis**.

Note that, according to this definition, an orthonormal basis for a Hilbert space V is not necessarily a Hamel basis for V . However, it can be shown that any orthogonal basis is a subset of a Hamel basis. In practice it is the orthonormal basis, not the Hamel basis itself, which is of most use. None of these issues arise in finite-dimensional spaces, where an orthogonal basis is always a Hamel basis.

Let $\mathcal{B} = \{\underline{v}_\alpha | \alpha \in A\}$ be an orthonormal basis for Hilbert space V . Then, each element $\underline{v} \in V$ has a unique representation as

$$\underline{v} = \sum_{\alpha \in A} s_\alpha \underline{v}_\alpha.$$

Using orthogonality to compute $\langle \underline{v}, \underline{v} \rangle$, one gets the **Parseval identity**

$$\|\underline{v}\|^2 = \sum_{\alpha \in A} |s_\alpha|^2.$$

Since $\|\underline{v}\|^2 < \infty$ for all $\underline{v} \in V$, the RHS also exists and is finite for all $\underline{v} \in V$.

Theorem 3.7.16. Every orthonormal set in a Hilbert space V can be enlarged to an orthonormal basis for V .

Proof. Let $x = \{\underline{v}_\alpha | \alpha \in A_0\}$ be the initial orthonormal set. Let X be the set of orthonormal subsets of V and, for $x, y \in X$, consider the strict partial order defined by proper inclusion. Since x is an element of X , the Hausdorff maximal principle implies that there exists a maximal simply ordered subset Z of X containing x . This shows the existence of a maximal orthonormal set $\{\underline{v}_\alpha | \alpha \in A\}$, where $A_0 \subset A$.

Let W be the closed subspace of V generated by $\{\underline{v}_\alpha | \alpha \in A\}$. If $W \neq V$, there is a unit vector $\underline{u} \in W^\perp$, contradicting the maximality of the system $\{\underline{v}_\alpha | \alpha \in A\}$. Thus, $W = V$ and we have enlarged the orthonormal set x to a basis. \square

Theorem 3.7.17. A Hilbert space V has a countable orthonormal basis if and only if V is separable.

Sketch of proof. If V is separable, then it contains a countable dense subset. Since this set is countable, it can be ordered into a sequence $\underline{v}_1, \underline{v}_2, \dots$ such that, for every vector $\underline{v} \in V$ and any $\epsilon > 0$, there exists an n such that $\|\underline{v} - \underline{v}_n\| < \epsilon$. By removing all vectors that are linear combinations of previous vectors, this sequence can be pruned into a countable linearly independent set. Then, a countable orthonormal basis is generated by applying Gram-Schmidt orthogonalization to the pruned sequence of vectors. Conversely, if V has a countable orthonormal basis, then linear combinations with rational coefficients can be used to construct a countable dense subset. \square

Lemma 3.7.18. *Let V be a Hilbert space and $\underline{v}_1, \underline{v}_2, \dots$ be a countable orthogonal set. Then, $\underline{v} = \sum_{i=1}^{\infty} \underline{v}_i$ exists if and only if $\sum_{i=1}^{\infty} \|\underline{v}_i\|^2 = M < \infty$.*

Proof. For $\underline{u}_m = \sum_{i=1}^m \underline{v}_i$ and $w_m = \sum_{i=1}^m \|\underline{v}_i\|^2$, orthogonality implies that

$$\|\underline{u}_m - \underline{u}_n\|^2 = \left\| \sum_{i=n+1}^m \underline{v}_i \right\|^2 = \sum_{i=n+1}^m \|\underline{v}_i\|^2 = |w_m - w_n|.$$

Thus, the sequence \underline{u}_n is Cauchy in V if and only if w_n is Cauchy in \mathbb{R} . \square

3.8 Linear Functionals

Definition 3.8.1. *Let V be a vector space over a field F . A linear transformation f from V into the scalar field F is called a **linear functional** on V .*

That is, f is a functional on V such that

$$f(s\underline{v}_1 + \underline{v}_2) = sf(\underline{v}_1) + f(\underline{v}_2)$$

for all $\underline{v}_1, \underline{v}_2 \in V$ and $s \in F$.

Example 3.8.2. *Let F be a field and let s_1, \dots, s_n be scalars in F . Then the functional f on F^n defined by*

$$f(v_1, \dots, v_n) = s_1v_1 + \dots + s_nv_n$$

is a linear functional. It is the linear functional which is represented by the matrix

$$\begin{bmatrix} s_1 & s_2 & \cdots & s_n \end{bmatrix}$$

relative to the standard ordered basis for F^n . Every linear functional on F^n is of this form, for some scalars s_1, \dots, s_n .

Definition 3.8.3. Let n be a positive integer and F a field. If A is an $n \times n$ matrix with entries in F , the **trace** of A is the scalar

$$\operatorname{tr}(A) = A_{11} + A_{22} + \cdots + A_{nn}.$$

Example 3.8.4. The trace function is a linear functional on the matrix space $F^{n \times n}$ since

$$\begin{aligned} \operatorname{tr}(sA + B) &= \sum_{i=1}^n (sA_{ii} + B_{ii}) \\ &= s \sum_{i=1}^n A_{ii} + \sum_{i=1}^n B_{ii} \\ &= s \operatorname{tr}(A) + \operatorname{tr}(B). \end{aligned}$$

Example 3.8.5. Let $[a, b]$ be a closed interval on the real line and let $C([a, b])$ be the space of continuous real-valued functions on $[a, b]$. Then

$$L(g) = \int_a^b g(t) dt$$

defines a linear functional L on $C([a, b])$.

Theorem 3.8.6 (Riesz). Let V be a finite-dimensional Hilbert space and f be a linear functional on V . Then, there exists a unique vector $\underline{v} \in V$ such that $f(\underline{w}) = \langle \underline{w}, \underline{v} \rangle$ for all $\underline{w} \in V$.

Proof. If we choose an orthonormal basis $\mathcal{B} = \underline{v}_1, \dots, \underline{v}_n$ for V , then the inner product of $\underline{w} = t_1 \underline{v}_1 + \cdots + t_n \underline{v}_n$ and $\underline{v} = s_1 \underline{v}_1 + \cdots + s_n \underline{v}_n$ will be

$$\langle \underline{w}, \underline{v} \rangle = t_1 \bar{s}_1 + \cdots + t_n \bar{s}_n.$$

If f is a linear functional on V , then f has the form

$$f(\underline{w}) = f(t_1 \underline{v}_1 + \cdots + t_n \underline{v}_n) = t_1 f(\underline{v}_1) + \cdots + t_n f(\underline{v}_n).$$

Thus, we can choose $\bar{s}_j = f(\underline{v}_j)$ to get $\langle \underline{w}, \underline{v} \rangle = f(\underline{w})$ and this gives

$$\underline{v} = \overline{f(\underline{v}_1)} \underline{v}_1 + \cdots + \overline{f(\underline{v}_n)} \underline{v}_n.$$

Let \underline{v}' be any vector that satisfies $f(\underline{w}) = \langle \underline{w}, \underline{v}' \rangle$ for all $\underline{w} \in V$. Then, we see that $\langle \underline{w}, \underline{v} - \underline{v}' \rangle = 0$ for all $\underline{w} \in V$. This implies that $\underline{v} - \underline{v}' = \underline{0}$.

□

Chapter 4

Representation and Approximation

4.1 Best Approximation

Suppose W is a subspace of a Banach space V . For any $\underline{v} \in V$, consider the problem of finding a vector $\underline{w} \in W$ such that $\|\underline{v} - \underline{w}\|$ is as small as possible.

Definition 4.1.1. *The vector $\underline{w} \in W$ is a **best approximation** of $\underline{v} \in V$ by vectors in W if*

$$\|\underline{v} - \underline{w}\| \leq \|\underline{v} - \underline{w}'\|$$

for all $\underline{w}' \in W$.

If W is spanned by the vectors $\underline{w}_1, \dots, \underline{w}_n \in V$, then we can write

$$\begin{aligned}\underline{v} &= \underline{w} + \underline{e} \\ &= s_1 \underline{w}_1 + \dots + s_n \underline{w}_n + \underline{e},\end{aligned}$$

where \underline{e} is the approximation error.

Finding a best approximation is, in general, rather difficult¹. However, if the norm $\|\cdot\|$ corresponds to the induced norm of an inner product, then one can use orthogonal projection and the problem is greatly simplified. This chapter focuses mainly on computing the best approximation of arbitrary vectors in a Hilbert space.

¹A best approximation exists if W is closed and the Banach space V is reflexive (i.e., it equals its double dual). In addition, it is unique if the Banach space V is strictly convex (i.e., $\|\underline{v} + \underline{w}\| < 2$ for all distinct $\underline{v}, \underline{w} \in V$ such that $\|\underline{v}\| = \|\underline{w}\| = 1$).

Theorem 4.1.2. *Suppose W is a subspace of a Hilbert space V and \underline{v} is a vector in V . Then, we have the following:*

1. *The vector $\underline{w} \in W$ is a best approximation of $\underline{v} \in V$ by vectors in W if and only if $\underline{v} - \underline{w}$ is orthogonal to every vector in W .*
2. *If a best approximation of $\underline{v} \in V$ by vectors in W exists, it is unique.*
3. *If W has a countable orthogonal basis $\underline{w}_1, \underline{w}_2, \dots$ and is closed, then*

$$\underline{w} = \sum_{i=1}^{\dim(W)} \frac{\langle \underline{v}, \underline{w}_i \rangle}{\|\underline{w}_i\|^2} \underline{w}_i \quad (4.1)$$

exists and equals the best approximation of \underline{v} by vectors in W .

Proof. Let $\underline{w} \in W$ and suppose $\underline{v} - \underline{w}$ is orthogonal to every vector in W . For any $\underline{w}' \in W$, we have $\underline{v} - \underline{w}' = (\underline{v} - \underline{w}) + (\underline{w} - \underline{w}')$ and

$$\begin{aligned} \|\underline{v} - \underline{w}'\|^2 &= \|\underline{v} - \underline{w}\|^2 + 2\operatorname{Re}\langle \underline{v} - \underline{w}, \underline{w} - \underline{w}' \rangle + \|\underline{w} - \underline{w}'\|^2 \\ &= \|\underline{v} - \underline{w}\|^2 + \|\underline{w} - \underline{w}'\|^2 \\ &\geq \|\underline{v} - \underline{w}\|^2. \end{aligned} \quad (4.2)$$

For the converse, we note that, if $\underline{v} - \underline{w}$ is not orthogonal to all vectors in W , then there must be some $\underline{u} \in W$ such that $\langle \underline{v} - \underline{w}, \underline{u} \rangle \neq 0$. Then, we let \underline{w}'' be the projection of $\underline{v} - \underline{w}$ onto \underline{u} . Next, we define $\underline{w}' = \underline{w} + \underline{w}''$ and observe that $\underline{w}' \in W$. Thus, Lemma 3.6.14 implies

$$\|\underline{v} - \underline{w}'\|^2 = \|\underline{v} - \underline{w} - \underline{w}''\|^2 = \|\underline{v} - \underline{w}\|^2 - \frac{|\langle \underline{v} - \underline{w}, \underline{u} \rangle|^2}{\|\underline{u}\|^2} < \|\underline{v} - \underline{w}\|^2.$$

Thus, \underline{w} is not a best approximation of \underline{v} by vectors in W .

For uniqueness, suppose $\underline{w}, \underline{w}' \in W$ are best approximations of \underline{v} by vectors in W . Then $\|\underline{v} - \underline{w}\| = \|\underline{v} - \underline{w}'\|$ and (4.2) implies that $\|\underline{w} - \underline{w}'\| = 0$. That is, if a best approximation exists then it is unique.

Finally, assume W is closed and $\underline{w}_1, \underline{w}_2, \dots$ is a countable orthogonal basis. Then, for (4.1), let the sequence of partial sums be $\underline{u}_n \triangleq \sum_{i=1}^n \underline{w}_i \langle \underline{v}, \underline{w}_i \rangle / \|\underline{w}_i\|^2$. Next, observe that $\underline{v} - \underline{u}_n$ is orthogonal to \underline{w}_j for $j \in \{1, \dots, n\}$, i.e.,

$$\begin{aligned} \langle \underline{v} - \underline{u}_n, \underline{w}_j \rangle &= \langle \underline{v}, \underline{w}_j \rangle - \left\langle \sum_{i=1}^n \frac{\langle \underline{v}, \underline{w}_i \rangle}{\|\underline{w}_i\|^2} \underline{w}_i, \underline{w}_j \right\rangle \\ &= \langle \underline{v}, \underline{w}_j \rangle - \frac{\langle \underline{v}, \underline{w}_j \rangle}{\|\underline{w}_j\|^2} \langle \underline{w}_j, \underline{w}_j \rangle = 0. \end{aligned}$$

Since $\underline{v} - \underline{u}_n$ is orthogonal to every vector in $W_n = \text{span}\{\underline{w}_1, \dots, \underline{w}_n\}$, we see that \underline{u}_n is the best approximation of \underline{v} by vectors in W_n . If $\dim(W) < \infty$, then this completes the proof.

If $\dim(W) = \infty$, then we must argue that the limit of \underline{u}_n exists and lies in W . The orthogonality of $\underline{w}_1, \dots, \underline{w}_n$ implies $\|\underline{v}\|^2 = \|\underline{v} - \underline{u}_n\|^2 + \|\underline{u}_n\|^2$. From this, we see that $\|\underline{u}_n\|^2 = \sum_{i=1}^n |\langle \underline{v}, \underline{w}_i \rangle|^2 / \|\underline{w}_i\|^2$ is an increasing real sequence upper bounded by $\|\underline{v}\|^2$. It follows that the RHS converges to a finite limit. Thus, we can apply Lemma 3.7.18 to show convergence $\underline{u}_n \rightarrow \underline{w}$. Since W is closed, it follows that $\underline{w} \in W$. By construction, $\underline{v} - \underline{w}$ is orthogonal to \underline{w}_j for $j \in \mathbb{N}$ and, thus, every vector in W . Hence, \underline{w} is the best approximation of \underline{v} by vectors in W . \square

Definition 4.1.3. *Whenever the vector \underline{w} in Theorem 4.1.2 exists, it is called the **orthogonal projection** of \underline{v} onto W . If every vector in V has an orthogonal projection onto W , then the mapping $E: V \rightarrow W$, which assigns to each vector in V its orthogonal projection onto W , is called the **orthogonal projection of V onto W** .*

One can use Theorem 4.1.14 to verify that this is consistent with the concept of orthogonal projection from Definition 4.1.11. Theorem 4.1.2 also implies the following result, known as Bessel's inequality.

Corollary 4.1.4. *Let $\underline{v}_1, \underline{v}_2, \dots$ be a countable orthogonal set of distinct non-zero vectors in an inner-product space V and let $W = \text{span}(\{\underline{v}_1, \underline{v}_2, \dots\})$. If $\underline{v} \in V$, then*

$$\sum_{i=1}^{\dim(W)} \frac{|\langle \underline{v}, \underline{v}_i \rangle|^2}{\|\underline{v}_i\|^2} \leq \|\underline{v}\|^2.$$

Moreover, equality holds if and only if

$$\underline{v} = \sum_{i=1}^{\dim(W)} \frac{\langle \underline{v}, \underline{v}_i \rangle}{\|\underline{v}_i\|^2} \underline{v}_i.$$

Proof. Let the projection of \underline{v} onto the closure of the span of $\underline{v}_1, \underline{v}_2, \dots$ (i.e., \overline{W}) be

$$\underline{w} = \sum_{i=1}^{\dim(W)} \frac{\langle \underline{v}, \underline{v}_i \rangle}{\|\underline{v}_i\|^2} \underline{v}_i.$$

Then, the error $\underline{u} = \underline{v} - \underline{w}$ satisfies $\langle \underline{u}, \underline{w} \rangle = 0$ and $\|\underline{u}\|^2 = \|\underline{v}\|^2 - \|\underline{w}\|^2$. Noting that $\|\underline{u}\|^2 \geq 0$ and

$$\|\underline{w}\|^2 = \sum_{i=1}^{\dim(W)} \frac{|\langle \underline{v}, \underline{v}_i \rangle|^2}{\|\underline{v}_i\|^2},$$

we see that $\|\underline{w}\|^2 \leq \|\underline{v}\|^2$ with equality iff $\underline{u} = \underline{0}$. \square

Problem 4.1.5. Let W be the subspace of \mathbb{R}^2 spanned by the vector $(1, 2)$. Using the standard inner product, let E be the orthogonal projection of \mathbb{R}^2 onto W . Find

1. a formula for $E(x_1, x_2)$
2. the matrix of E in the standard ordered basis, i.e., $E(x_1, x_2) = E\underline{x}$
3. W^\perp
4. an orthonormal basis in which E is represented by the matrix

$$E = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

4.1.1 Projection Operators

Definition 4.1.6. A function $F: X \rightarrow Y$ with $Y \subseteq X$ is **idempotent** if $F(F(x)) = F(x)$. When F is a linear transformation, this reduces to $F^2 = F \cdot F = F$.

Definition 4.1.7. Let V be a vector space and $T: V \rightarrow V$ be a linear transformation. If T is idempotent, then T is called a **projection**.

Example 4.1.8. The idempotent matrix A is a projection onto the first two coordinates.

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

Theorem 4.1.9. Let V be a vector space and $T: V \rightarrow V$ be a projection operator. Then, the range $\mathcal{R}(T)$ and the nullspace $\mathcal{N}(T)$ are disjoint subspaces of V .

Proof. For all $\underline{v} \in V - \{0\}$, we need to prove that \underline{v} is not in both the range and nullspace. Let $\underline{v} \in V$ be in the range of T so that there is a $\underline{w} \in V$ such that $T\underline{w} = \underline{v}$. Then, $T\underline{v} = T^2\underline{w} = T\underline{w} = \underline{v}$ and \underline{v} is not in the null space unless $\underline{v} = \underline{0}$.

Let \underline{v} be in the null space of T , then $T\underline{v} = \underline{0}$. But, $T\underline{v} = \underline{v}$ for all \underline{v} in the range. Therefore, \underline{v} is not in the range unless $\underline{v} = \underline{0}$. From this, we see that only $\underline{0} \in V$ is in both the range and nullspace. Therefore, they are disjoint subspaces. \square

Example 4.1.10. Consider the linear transform $T: V \rightarrow V$ defined by $T = I - P$, where P is a projection. It is easy to verify that T is a projection operator because

$$T^2 = (I - P)(I - P) = I - P - P + P^2 = I - P = T.$$

Notice also that $P(I - P)\underline{v} = \underline{0}$ implies that $\mathcal{R}(T) \subseteq \mathcal{N}(P)$ and $T\underline{v} = \underline{v}$ for $\underline{v} \in \mathcal{N}(P)$ implies $\mathcal{N}(P) \subseteq \mathcal{R}(T)$. Therefore, $\mathcal{R}(T) = \mathcal{N}(P)$ and $I - P$ is a projection onto $\mathcal{N}(P)$.

Definition 4.1.11. Let V be an inner-product space and $P: V \rightarrow V$ be a projection operator. If $\mathcal{R}(P) \perp \mathcal{N}(P)$, then P is called a **orthogonal projection**.

Example 4.1.12. Let V be an inner-product space and $P: V \rightarrow V$ be an orthogonal projection. Since $P\underline{v} = \underline{0}$ iff $(I - P)\underline{v} = \underline{v}$ and $I - P$ is a projection, we have $\mathcal{N}(P) = \mathcal{R}(I - P)$. Then, $\underline{v} = P\underline{v} + (I - P)\underline{v}$ defines an orthogonal decomposition of \underline{v} because $P\underline{v} \in \mathcal{R}(P)$ is orthogonal to $(I - P)\underline{v} \in \mathcal{N}(P)$. In addition, $V = \mathcal{R}(P) \oplus \mathcal{N}(P)$ and hence $\mathcal{N}(P) = \mathcal{R}(P)^\perp$.

Theorem 4.1.13. For $V = F^n$ with the standard inner product, a projection matrix P is an orthogonal projection matrix if and only if it is Hermitian (i.e. $P^H = P$).

Proof. First, we observe that

$$\langle P\underline{u}, (I - P)\underline{v} \rangle = \underline{v}^H (I - P)^H P\underline{u} = \underline{v}^H (P - P^H P)\underline{u}.$$

Then, if $P = P^H$, we have $P = P^2 = P^H P$ and the RHS above is zero. Thus, the LHS is zero and we have $\mathcal{R}(P) \perp \mathcal{N}(P)$. Conversely, if $\mathcal{R}(P) \perp \mathcal{N}(P)$, then

$$\langle P\underline{u}, \underline{v} \rangle = \langle P\underline{u}, P\underline{v} + (I - P)\underline{v} \rangle = \langle P\underline{u}, P\underline{v} \rangle + \langle P\underline{u}, (I - P)\underline{v} \rangle = \langle P\underline{u}, P\underline{v} \rangle.$$

By symmetry, we also have $\langle \underline{u}, P\underline{v} \rangle = \langle P\underline{u}, P\underline{v} \rangle$ and together these imply that

$$\underline{v}^H P\underline{u} = \langle P\underline{u}, \underline{v} \rangle = \langle \underline{u}, P\underline{v} \rangle = \underline{v}^H P^H \underline{u}$$

for all $\underline{u}, \underline{v} \in V$. Thus, $P = P^H$. □

Theorem 4.1.14. Suppose W is a closed subspace of a separable Hilbert space V and let E denote the orthogonal projection of V on W . Then, E is an idempotent linear transformation of V onto W , $E\underline{w}' = \underline{0}$ iff $\underline{w}' \in W^\perp$, and

$$V = W \oplus W^\perp.$$

Proof. Let \underline{v} be any vector in V . Since $E\underline{v}$ is the best approximation of \underline{v} by vectors in W , it follows that $\underline{v} \in W$ implies $E\underline{v} = \underline{v}$. Therefore, $E(E\underline{v}) = E\underline{v}$ for any $\underline{v} \in V$ since $E\underline{v} \in W$. That is, $E^2 = E$ and E is idempotent.

To show that E is a linear transformation, let $\underline{w}_1, \underline{w}_2, \dots$ be a countable orthonormal basis for W (whose existence follows from Theorem 3.7.17). Using part 3 of Theorem 4.1.2, we find that

$$\begin{aligned} E(s_1\underline{v}_1 + \underline{v}_2) &= \sum_{i=1}^{\dim(W)} \langle s_1\underline{v}_1 + \underline{v}_2, \underline{w}_i \rangle \underline{w}_i \\ &= s_1 \sum_{i=1}^{\dim(W)} \langle \underline{v}_1, \underline{w}_i \rangle \underline{w}_i + \sum_{i=1}^{\dim(W)} \langle \underline{v}_2, \underline{w}_i \rangle \underline{w}_i \\ &= s_1 E\underline{v}_1 + E\underline{v}_2. \end{aligned}$$

Therefore, E is a linear transformation. It also follows that $E\underline{w}' = \underline{0}$ iff $\underline{w}' \in W^\perp$ because W^\perp can be defined by the fact that $\langle \underline{w}', \underline{w}_i \rangle = 0$ for $i \in \mathbb{N}$.

Again, let $\underline{v} \in V$ and recall that (by Theorem 4.1.2) $E\underline{v}$ is the unique vector in W such that $\underline{v} - E\underline{v}$ is in W^\perp . Therefore, the equation $\underline{v} = E\underline{v} + (\underline{v} - E\underline{v})$ gives a unique decomposition of \underline{v} into $E\underline{v} \in W$ and $\underline{v} - E\underline{v} \in W^\perp$. This unique decomposition implies that V is the direct sum of W and W^\perp . Lastly, one finds from the definition of W^\perp that

$$W \cap W^\perp = \{\underline{u} \in W \mid \langle \underline{u}, \underline{w} \rangle = 0 \forall \underline{w} \in W\} \subseteq \{\underline{u} \in W \mid \langle \underline{u}, \underline{u} \rangle = 0\} = \{\underline{0}\}. \quad \square$$

Corollary 4.1.15. *Let W be a closed subspace of a separable Hilbert space V and E be the orthogonal projection of V on W . Then $I - E$ is the orthogonal projection of V on W^\perp .*

Proof. This follows directly from the orthogonal decomposition in Theorem 4.1.14. One can also verify that $I - E$ is an idempotent linear transformation of V with range W^\perp and nullspace W . From Definition 4.1.11, we see that $I - E$ is an orthogonal projection. \square

Example 4.1.16. *Let $V = \mathbb{C}^n$ be the standard n -dimensional complex Hilbert space. Let $U \in \mathbb{C}^{n \times m}$ be a matrix whose columns $\underline{u}_1, \dots, \underline{u}_m$ form an orthonormal set in V . Then, the best approximation of $\underline{v} \in V$ by vectors in $\mathcal{R}(U)$ (as defined by (4.1)) can also be written as*

$$\underline{w} = UU^H \underline{v} = \sum_{i=1}^m \underline{u}_i (\underline{u}_i^H \underline{v}).$$

4.2 Computing Approximations in Hilbert Spaces

4.2.1 Normal Equations

Suppose V is a Hilbert space the subspace W is spanned by $\underline{w}_1, \dots, \underline{w}_n \in V$. Consider the situation where the sequence $\underline{w}_1, \dots, \underline{w}_n$ is linearly independent, but not orthogonal. In this case, it is not possible to apply (4.1) directly. It is nevertheless possible to obtain a similar expression for the best approximation of \underline{v} by vectors in W . Theorem 4.1.2 asserts that $\hat{\underline{v}} \in W$ is a best approximation of $\underline{v} \in V$ by vectors in W if and only if $\underline{v} - \hat{\underline{v}}$ is orthogonal to every vector in W . This implies that

$$\langle \underline{v} - \hat{\underline{v}}, \underline{w}_j \rangle = \left\langle \underline{v} - \sum_{i=1}^n s_i \underline{w}_i, \underline{w}_j \right\rangle = 0$$

or, equivalently,

$$\sum_{i=1}^n s_i \langle \underline{w}_i, \underline{w}_j \rangle = \langle \underline{v}, \underline{w}_j \rangle$$

for $j = 1, \dots, n$. These conditions yield a system of n linear equations in n unknowns, which can be written in the matrix form

$$\begin{bmatrix} \langle \underline{w}_1, \underline{w}_1 \rangle & \langle \underline{w}_2, \underline{w}_1 \rangle & \cdots & \langle \underline{w}_n, \underline{w}_1 \rangle \\ \langle \underline{w}_1, \underline{w}_2 \rangle & \langle \underline{w}_2, \underline{w}_2 \rangle & \cdots & \langle \underline{w}_n, \underline{w}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{w}_1, \underline{w}_n \rangle & \langle \underline{w}_2, \underline{w}_n \rangle & \cdots & \langle \underline{w}_n, \underline{w}_n \rangle \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} = \begin{bmatrix} \langle \underline{v}, \underline{w}_1 \rangle \\ \langle \underline{v}, \underline{w}_2 \rangle \\ \vdots \\ \langle \underline{v}, \underline{w}_n \rangle \end{bmatrix}.$$

We can rewrite this matrix equation as

$$G\underline{s} = \underline{t}$$

where

$$\underline{t}^T = (\langle \underline{v}, \underline{w}_1 \rangle, \langle \underline{v}, \underline{w}_2 \rangle, \dots, \langle \underline{v}, \underline{w}_n \rangle)$$

is the **cross-correlation vector**, and

$$\underline{s}^T = (s_1, s_2, \dots, s_n)$$

is the vector of coefficients. Equations of this form are collectively known as the **normal equations**.

Definition 4.2.1. The $n \times n$ matrix

$$G = \begin{bmatrix} \langle \underline{w}_1, \underline{w}_1 \rangle & \langle \underline{w}_2, \underline{w}_1 \rangle & \cdots & \langle \underline{w}_n, \underline{w}_1 \rangle \\ \langle \underline{w}_1, \underline{w}_2 \rangle & \langle \underline{w}_2, \underline{w}_2 \rangle & \cdots & \langle \underline{w}_n, \underline{w}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{w}_1, \underline{w}_n \rangle & \langle \underline{w}_2, \underline{w}_n \rangle & \cdots & \langle \underline{w}_n, \underline{w}_n \rangle \end{bmatrix} \quad (4.3)$$

is called the **Gramian matrix**. Since $g_{ij} = \langle \underline{w}_j, \underline{w}_i \rangle$, it follows that the Gramian is a Hermitian symmetric matrix, i.e., $G^H = G$.

Definition 4.2.2. A matrix $M \in F^{n \times n}$ is **positive-semidefinite** if $M^H = M$ and $\underline{v}^H M \underline{v} \geq 0$ for all $\underline{v} \in F^n - \{\underline{0}\}$. If the inequality is strict, M is **positive-definite**.

An important aspect of positive-definite matrices is that they are always invertible. This follows from noting that $M\underline{v} = \underline{0}$ for $\underline{v} \neq \underline{0}$ implies that $\underline{v}^H M \underline{v} = 0$ and contradicts the definition of positive definite.

Theorem 4.2.3. A Gramian matrix G is always positive-semidefinite. It is positive-definite if and only if the vectors $\underline{w}_1, \dots, \underline{w}_n$ are linearly independent.

Proof. Since $g_{ij} = \langle \underline{w}_j, \underline{w}_i \rangle$, the conjugation property of the inner product implies $G^H = G$. Using $\underline{v} = (v_1, \dots, v_n)^T \in F^n$, we can write

$$\begin{aligned} \underline{v}^H G \underline{v} &= \sum_{i=1}^n \sum_{j=1}^n \bar{v}_i g_{ij} v_j = \sum_{i=1}^n \sum_{j=1}^n \bar{v}_i \langle \underline{w}_j, \underline{w}_i \rangle v_j \\ &= \sum_{i=1}^n \sum_{j=1}^n \langle v_j \underline{w}_j, v_i \underline{w}_i \rangle = \left\langle \sum_{j=1}^n v_j \underline{w}_j, \sum_{i=1}^n v_i \underline{w}_i \right\rangle \\ &= \left\| \sum_{i=1}^n v_i \underline{w}_i \right\|^2 \geq 0. \end{aligned} \quad (4.4)$$

That is, $\underline{v}^H G \underline{v} \geq 0$ for all $\underline{v} \in F^n$.

Suppose that G is not positive-definite. Then, there exists $\underline{v} \in F^n - \{\underline{0}\}$ such that $\underline{v}^H G \underline{v} = 0$. By (4.4), this implies that

$$\sum_{i=1}^n v_i \underline{w}_i = \underline{0}$$

and hence the sequence of vectors $\underline{w}_1, \dots, \underline{w}_n$ is not linearly independent.

Conversely, if G is positive-definite then $\underline{v}^H G \underline{v} > 0$ and

$$\left\| \sum_{i=1}^n v_i \underline{w}_i \right\| > 0$$

for all $\underline{v} \in F^n - \{0\}$. Thus, the vectors $\underline{w}_1, \dots, \underline{w}_n$ are linearly independent. \square

4.2.2 Orthogonality Principle

Theorem 4.2.4. *Let $\underline{w}_1, \dots, \underline{w}_n$ be vectors in an inner-product space V and denote the span of $\underline{w}_1, \dots, \underline{w}_n$ by W . For any vector $\underline{v} \in V$, the norm of the error vector*

$$\underline{e} = \underline{v} - \sum_{i=1}^n s_i \underline{w}_i \quad (4.5)$$

*is minimized when the error vector \underline{e} is orthogonal to every vector in W . If $\hat{\underline{v}}$ denotes the **least-squares** approximation of \underline{v} then*

$$\langle \underline{v} - \hat{\underline{v}}, \underline{w}_j \rangle = 0$$

for $j = 1, \dots, n$.

Proof. Minimizing $\|\underline{e}\|^2$ over \underline{s} , where \underline{e} is given by (4.5) requires minimizing

$$\begin{aligned} J(\underline{s}) &= \left\langle \underline{v} - \sum_{i=1}^n s_i \underline{w}_i, \underline{v} - \sum_{j=1}^n s_j \underline{w}_j \right\rangle \\ &= \langle \underline{v}, \underline{v} \rangle - \sum_{i=1}^n \langle s_i \underline{w}_i, \underline{v} \rangle - \sum_{j=1}^n \langle \underline{v}, s_j \underline{w}_j \rangle + \sum_{i=1}^n \sum_{j=1}^n \langle s_i \underline{w}_i, s_j \underline{w}_j \rangle \\ &= \langle \underline{v}, \underline{v} \rangle - \sum_{i=1}^n s_i \langle \underline{w}_i, \underline{v} \rangle - \sum_{j=1}^n \bar{s}_j \langle \underline{v}, \underline{w}_j \rangle + \sum_{i=1}^n \sum_{j=1}^n s_i \bar{s}_j \langle \underline{w}_i, \underline{w}_j \rangle. \end{aligned}$$

To take the derivative with respect to $\underline{s} \in \mathbb{C}^n$, we use the decomposition $\underline{s} = \underline{a} + j\underline{b}$, with $\underline{a}, \underline{b} \in \mathbb{R}^n$, and define the differential operators

$$\frac{\partial}{\partial \underline{s}} \triangleq \frac{1}{2} \left(\frac{\partial}{\partial \underline{a}} - j \frac{\partial}{\partial \underline{b}} \right) \quad \frac{\partial}{\partial \bar{\underline{s}}} \triangleq \frac{1}{2} \left(\frac{\partial}{\partial \underline{a}} + j \frac{\partial}{\partial \underline{b}} \right),$$

where $\partial/\partial \underline{a} = (\partial/\partial a_1, \dots, \partial/\partial a_n)^T$. Since $J(\underline{s})$ is real function of \underline{s} , a stationary

point of J can be found by setting either derivative to 0. Choosing \underline{s} , gives

$$\begin{aligned} \frac{\partial}{\partial \underline{s}} J(\underline{s}) &= - \begin{bmatrix} \langle \underline{v}, \underline{w}_1 \rangle \\ \langle \underline{v}, \underline{w}_2 \rangle \\ \vdots \\ \langle \underline{v}, \underline{w}_n \rangle \end{bmatrix} + \begin{bmatrix} \langle \underline{w}_1, \underline{w}_1 \rangle & \langle \underline{w}_2, \underline{w}_1 \rangle & \cdots & \langle \underline{w}_n, \underline{w}_1 \rangle \\ \langle \underline{w}_1, \underline{w}_2 \rangle & \langle \underline{w}_2, \underline{w}_2 \rangle & \cdots & \langle \underline{w}_n, \underline{w}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{w}_1, \underline{w}_n \rangle & \langle \underline{w}_2, \underline{w}_n \rangle & \cdots & \langle \underline{w}_n, \underline{w}_n \rangle \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} \\ &= \underline{0}. \end{aligned}$$

In matrix form, this yields the familiar equation

$$G\underline{s} = \underline{t}.$$

To ensure that this extremum is in fact a minimum, one can compute the 2nd derivative to show that the Hessian is G . Since G is a positive-semidefinite matrix, the extremum is indeed a minimum.

This implies that $\|\underline{e}\|^2$ is minimized if and only if $G\underline{s} = \underline{t}$. That is, $\|\underline{e}\|^2$ is minimized if and only if $\underline{v} - \hat{\underline{v}}$ is orthogonal to every vector in W . \square

Note that it is also possible to prove this theorem using the Cauchy-Schwarz inequality or the projection theorem.

4.3 Approximation for Systems of Linear Equations

4.3.1 Matrix Representation

For finite-dimensional vector spaces, least-squares (i.e., best approximation) problems have natural matrix representations. Suppose $V = F^m$ and $\underline{w}_1, \underline{w}_2, \dots, \underline{w}_n \in V$ are column vectors. Then, the approximation vector is given by

$$\hat{\underline{v}} = \sum_{i=1}^n s_i \underline{w}_i$$

In matrix form, we have

$$\hat{\underline{v}} = A\underline{s},$$

where $A = [\underline{w}_1 \cdots \underline{w}_n]$. The optimization problem can then be reformulated as follows. Determine $\underline{s} \in F^n$ such that

$$\|\underline{e}\|^2 = \|\underline{v} - \hat{\underline{v}}\|^2 = \|\underline{v} - A\underline{s}\|^2$$

is minimized. Note that this occurs when the error vector is orthogonal to every vector in W , i.e.,

$$\langle \underline{e}, \underline{w}_j \rangle = \langle \underline{v} - \hat{\underline{v}}, \underline{w}_j \rangle = \langle \underline{v} - A\underline{s}, \underline{w}_j \rangle = 0$$

for $j = 1, \dots, n$.

4.3.2 Standard Inner Products

When $\|\cdot\|$ is the norm induced by the standard inner product, these conditions can be expressed as

$$\begin{bmatrix} \underline{w}_1^H \\ \vdots \\ \underline{w}_n^H \end{bmatrix} (\underline{v} - A\underline{s}) = \underline{0}.$$

Using the definition of A , we obtain

$$A^H A\underline{s} = A^H \underline{v}.$$

The matrix $A^H A$ is the Gramian G defined in (4.3). The vector $A^H \underline{v}$ is the cross correlation vector \underline{t} .

When the vectors $\underline{w}_1, \dots, \underline{w}_n$ are linearly independent, the Gramian matrix is positive definite and hence invertible. The optimal solution for the least-squares problem is therefore given by

$$\underline{s} = (A^H A)^{-1} A^H \underline{v} = G^{-1} \underline{t}.$$

The matrix $(A^H A)^{-1} A^H$ is often called the **pseudoinverse**.

The best approximation of $\underline{v} \in V$ by vectors in W is equal to

$$\hat{\underline{v}} = A\underline{s} = A (A^H A)^{-1} A^H \underline{v}.$$

The matrix $P = A (A^H A)^{-1} A^H$ is called the **projection matrix** for the range of A . It defines an orthogonal projection onto the range of A (i.e., the subspace spanned by the columns of A).

4.3.3 Generalized Inner Products

We can also consider the case of a general inner product. Recall that an inner product on V is completely determined by the values

$$h_{ji} = \langle \underline{e}_i, \underline{e}_j \rangle$$

and that it can be expressed in terms of the matrix H (where $[H]_{j,i} = h_{j,i}$) as

$$\langle \underline{v}, \underline{w} \rangle = \underline{w}^H H \underline{v}.$$

Minimizing $\|\underline{e}\|^2 = \|\underline{v} - A\underline{s}\|^2$ and using the orthogonality principle lead to the matrix equation

$$A^H H A \underline{s} = A^H H \underline{v}.$$

When the vectors $\underline{w}_1, \dots, \underline{w}_n$ are linearly independent, the optimal solution is given by

$$\underline{s} = (A^H H A)^{-1} A^H H \underline{v}.$$

4.3.4 Minimum Error

Let $\hat{\underline{v}} \in W$ be the best approximation of \underline{v} by vectors in W . Again, we can write

$$\underline{v} = \hat{\underline{v}} + \underline{e},$$

where $\underline{e} \in W^\perp$ is the minimum achievable error. The squared norm of the minimum error is given implicitly by

$$\|\underline{v}\|^2 = \|\hat{\underline{v}} + \underline{e}\|^2 = \langle \hat{\underline{v}} + \underline{e}, \hat{\underline{v}} + \underline{e} \rangle = \langle \hat{\underline{v}}, \hat{\underline{v}} \rangle + \langle \underline{e}, \underline{e} \rangle = \|\hat{\underline{v}}\|^2 + \|\underline{e}\|^2.$$

We can then find an explicit expression for the approximation error,

$$\begin{aligned} \|\underline{e}\|^2 &= \|\underline{v}\|^2 - \|\hat{\underline{v}}\|^2 = \underline{v}^H H \underline{v} - \hat{\underline{v}}^H H \hat{\underline{v}} \\ &= \underline{v}^H H \underline{v} - \underline{s}^H A^H H A \underline{s} \\ &= \underline{v}^H H \underline{v} - \underline{v}^H H A (A^H H A)^{-1} A^H H \underline{v} \\ &= \underline{v}^H \left(H - H A (A^H H A)^{-1} A^H H \right) \underline{v}. \end{aligned}$$

4.4 Applications and Examples in Signal Processing

4.4.1 Linear Regression

Let $(x_1, y_1), \dots, (x_n, y_n)$ be a collection of points in \mathbb{R}^2 . A **linear regression** problem consists in finding scalars a and b such that

$$y_i \approx ax_i + b$$

for $i = 1, \dots, n$. Define the error component e_i by $e_i = y_i - ax_i - b$, then

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = a \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + b \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}.$$

In vector form, we can rewrite this equation as

$$\underline{y} = A\underline{s} + \underline{e},$$

where $\underline{y} = (y_1, \dots, y_n)^T$, $\underline{s} = (a, b)^T$, $\underline{e} = (e_1, \dots, e_n)^T$, and

$$A = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}.$$

This equation has a form analog to the matrix representation of a least-squares problem. Consider the goal of minimizing $\|\underline{e}\|^2$. The line that minimizes the sums of the squares of the *vertical* distances between the data abscissas and the line is then given by

$$\underline{s} = (A^H A)^{-1} A^H \underline{y}.$$

4.4.2 Linear Minimum Mean-Squared Error Estimation

Let Y, X_1, \dots, X_n be a set of zero-mean random variables. The goal of the linear minimum mean-squared error (LMMSE) estimation problem is to find coefficients s_1, \dots, s_n such that

$$\hat{Y} = s_1 X_1 + \dots + s_n X_n$$

minimizes the MSE $E[|Y - \hat{Y}|^2]$. Using the inner product defined by

$$\langle X, Y \rangle = E[X\bar{Y}], \quad (4.6)$$

we can compute the linear minimum mean-squared estimate \hat{Y} using

$$G\underline{s} = \underline{t},$$

where

$$G = \begin{bmatrix} E[X_1\bar{X}_1] & E[X_2\bar{X}_1] & \cdots & E[X_n\bar{X}_1] \\ E[X_1\bar{X}_2] & E[X_2\bar{X}_2] & \cdots & E[X_n\bar{X}_2] \\ \vdots & \vdots & \ddots & \vdots \\ E[X_1\bar{X}_n] & E[X_2\bar{X}_n] & \cdots & E[X_n\bar{X}_n] \end{bmatrix}$$

and

$$\underline{t} = \begin{bmatrix} E[Y\bar{X}_1] \\ E[Y\bar{X}_2] \\ \vdots \\ E[Y\bar{X}_n] \end{bmatrix}.$$

If the matrix G is invertible, the minimum mean-squared error is given by

$$\|Y - \hat{Y}\|^2 = E[Y\bar{Y}] - \underline{t}^H G^{-1} \underline{t}.$$

4.4.3 The Wiener Filter

Suppose that the sequence of zero-mean random variables $\{X[t]\}$ is wide-sense stationary, and consider the FIR filter

$$\begin{aligned} Y[t] &= \sum_{k=0}^{K-1} h[k]X[t-k] \\ &= \begin{bmatrix} X[t] & \cdots & X[t-K+1] \end{bmatrix} \begin{bmatrix} h[0] \\ \vdots \\ h[K-1] \end{bmatrix} = (\underline{X}[t])^T \underline{h}. \end{aligned}$$

The goal is to design this filter in such a way that its output is as close as possible to a desired sequence $\{Z[t]\}$. In particular, we want to minimize the mean-squared error

$$\|Z[t] - Y[t]\|^2 = E[|Z[t] - Y[t]|^2].$$

By the orthogonality principle, the mean-squared error is minimized when the error is orthogonal to the data; that is, for $j = 0, 1, \dots, K - 1$, we have

$$\left\langle Z[t] - \sum_{k=0}^{K-1} h[k]X[t-k], X[t-j] \right\rangle = 0,$$

or, equivalently, we can write

$$\langle Z[t], X[t-j] \rangle = \sum_{k=0}^{K-1} h[k] \langle X[t-k], X[t-j] \rangle.$$

Using (4.6), we obtain

$$E [Z[t]\overline{X}[t-j]] = \sum_{k=0}^{K-1} h[k] E [X[t-k]\overline{X}[t-j]]. \quad (4.7)$$

where $j = 1, \dots, K - 1$.

For this specific case where the normal equations are defined in terms of the expectation operator, these equations are called the **Wiener-Hopf** equations. The Gramian of the Wiener-Hopf equations can be expressed in a more familiar form using the autocorrelation matrix. Recall that $\{X[t]\}$ is a wide-sense stationary process. As such, we have

$$R_{xx}(j-k) = R_{xx}(j,k) = E [X[t-k]\overline{X}[t-j]] = \langle X[t-k], X[t-j] \rangle.$$

Also define

$$R_{zx}(j) = E [Z[t]\overline{X}[t-j]] = \langle Z[t], X[t-j] \rangle.$$

Using this notation, we can rewrite (4.7) as

$$R_{zx} = \begin{bmatrix} R_{zx}(0) \\ R_{zx}(1) \\ \vdots \\ R_{zx}(K-1) \end{bmatrix} = R_{xx} \begin{bmatrix} h[0] \\ h[1] \\ \vdots \\ h[K-1] \end{bmatrix}$$

where the $K \times K$ autocorrelation matrix is given by

$$R_{xx} = \begin{bmatrix} R_{xx}[0] & \overline{R}_{xx}[1] & \cdots & \overline{R}_{xx}[K-1] \\ R_{xx}[1] & R_{xx}[0] & \cdots & \overline{R}_{xx}[K-2] \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}[K-1] & R_{xx}[K-2] & \cdots & R_{xx}[0] \end{bmatrix}.$$

Note that the matrix R_{xx} is Toeplitz, i.e., all the elements on a diagonal are equal. Assuming that R_{xx} is invertible, the optimal filter taps are then given by

$$\underline{h} = R_{xx}^{-1} R_{zx}.$$

The minimum mean-squared error is given by

$$\begin{aligned} \|Z - Y\|^2 &= \|Z\|^2 - \|Y\|^2 \\ &= \text{E}[Z\bar{Z}] - \text{E}[\underline{h}^H \bar{X} X^T \underline{h}] \\ &= \text{E}[Z\bar{Z}] - \underline{h}^H R_{xx} \underline{h} \\ &= \text{E}[Z\bar{Z}] - R_{zx}^H \underline{h}, \end{aligned}$$

where t can be ignored because the processes are WSS.

4.4.4 LMMSE Filtering in Practice

While theoretical treatments of optimal filtering often assume one has well-defined random variables with known statistics, this is rarely the case in practice. Yet, there is a very close connection between Wiener filtering and natural data driven approaches. Consider the problem from the previous section and let $x[1], x[2], \dots, x[N]$ and $z[1], z[2], \dots, z[N]$ be realizations of the random processes.

As an application, one can think of the $x[t]$ sequence as the received samples in a wireless communication system and the $z[t]$ sequence as a *pilot sequence* (i.e., known to both the transmitter and receiver). It is assumed the transmitted sequence has been convolved with an unknown LTI system. This type of degradation is known as intersymbol interference (ISI) and the goal is to find a linear filter $h[0], h[1], \dots, h[K-1]$ that removes as much ISI as possible. A suitable cost function for this goal is

$$J(\underline{h}) = \sum_{t=K}^N \lambda^{N-t} \left| z[t] - \sum_{k=0}^{K-1} h[k] x[t-k] \right|^2,$$

where the exponential weighting factor λ emphasizes the most recently received symbols because, in reality, the channel conditions are changing with time.

Using the vector $\underline{z} = [z[K] \ z[K+1] \ \cdots \ z[N]]$ and the matrix

$$A = \begin{bmatrix} x[K] & x[K-1] & \cdots & x[1] \\ x[K+1] & x[K] & \cdots & x[2] \\ \vdots & \vdots & \ddots & \vdots \\ x[N] & x[N-1] & \cdots & x[N-K+1] \end{bmatrix},$$

we can rewrite this cost function as

$$J(\underline{h}) = (A\underline{h} - \underline{z})^H \Lambda (A\underline{h} - \underline{z}),$$

where Λ is a diagonal matrix with diagonal entries $[\lambda^{N-K} \ \lambda^{N-K+1} \ \cdots \ \lambda^0]$. Using the orthogonality principle, one finds that the optimal solution is given by the normal equation

$$A^H \Lambda A \underline{h} = A^H \Lambda \underline{z}.$$

To see the connection with Wiener filtering, the key observation is that the matrix $A^H \Lambda A$ and the vector $A^H \Lambda \underline{z}$ are sample-average estimates of the correlation matrix and cross-correlation vector. This is because, for large N and λ close to 1, we have

$$[A^H \Lambda A]_{ij} = \sum_{t=K}^N \lambda^{N-t} x[t-j+1] \bar{x}[t-i+1] \approx \frac{R_{xx}(i-j)}{1-\lambda}$$

and

$$[A^H \Lambda \underline{z}]_i = \sum_{t=K}^N \lambda^{N-t} z[t] \bar{x}[t-i+1] \approx \frac{R_{zx}(i)}{1-\lambda}.$$

Another benefit of this approach is that, as each new sample arrives, the solution \underline{h} can be updated with low complexity. Consider the matrix $G_N = A^H \Lambda A$ and vector $\underline{b}_N = A^H \Lambda \underline{z}$ as a function of N . Then, $G_{N+1} = \lambda G_N + \underline{u}^H \underline{u}$ and $\underline{b}_{N+1} = \lambda \underline{b}_N + z[N+1] \underline{u}^H$, where

$$\underline{u} = [x[N+1] \ x[N] \ \cdots \ x[N-K+2]]. \quad (4.8)$$

The updated solution vector $\underline{h}_{N+1} = G_{N+1}^{-1} \underline{b}_{N+1}$ can be computed efficiently using the Sherman-Morrison matrix inversion formula.

4.5 Dual Approximation

4.5.1 Minimum-Norm Solutions

In many cases, one is interested in finding the minimum-norm vector that satisfies some feasibility constraints. For example, an underdetermined system of linear equations has an infinite number of solutions. But, in practice, it often makes sense to prefer the minimum-norm solution over other solutions. Finding this solution is very similar to finding the best approximation.

Let V be a Hilbert space and $\underline{w}_1, \underline{w}_2, \dots, \underline{w}_n$ be a set of linearly independent vectors in V . For any $\underline{v} \in V$, consider finding the scalars s_1, s_2, \dots, s_n that minimize

$$\left\| \underline{v} - \sum_{i=1}^n s_i \underline{w}_i \right\|.$$

The answer is clearly given by the best approximation of \underline{v} by vectors in the span of $\underline{w}_1, \underline{w}_2, \dots, \underline{w}_n$. The orthogonality principle tells us that s_1, s_2, \dots, s_n must satisfy

$$\begin{bmatrix} \langle \underline{w}_1, \underline{w}_1 \rangle & \langle \underline{w}_2, \underline{w}_1 \rangle & \cdots & \langle \underline{w}_n, \underline{w}_1 \rangle \\ \langle \underline{w}_1, \underline{w}_2 \rangle & \langle \underline{w}_2, \underline{w}_2 \rangle & \cdots & \langle \underline{w}_n, \underline{w}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{w}_1, \underline{w}_n \rangle & \langle \underline{w}_2, \underline{w}_n \rangle & \cdots & \langle \underline{w}_n, \underline{w}_n \rangle \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} = \begin{bmatrix} \langle \underline{v}, \underline{w}_1 \rangle \\ \langle \underline{v}, \underline{w}_2 \rangle \\ \vdots \\ \langle \underline{v}, \underline{w}_n \rangle \end{bmatrix}. \quad (4.9)$$

The same problem can also be posed in a different manner.

Theorem 4.5.1. *Let V be a Hilbert space and $\underline{w}_1, \underline{w}_2, \dots, \underline{w}_n$ be a set of linearly independent vectors in V . The **dual approximation** problem is to find the vector $\underline{w} \in V$ of minimum-norm that satisfies $\langle \underline{w}, \underline{w}_i \rangle = c_i$ for $i = 1, \dots, n$. This vector is given by*

$$\underline{w} = \sum_{i=1}^n s_i \underline{w}_i, \quad (4.10)$$

where the coefficients s_1, s_2, \dots, s_n can be found by solving (4.9) with $\langle \underline{v}, \underline{w}_i \rangle = c_i$.

Proof. Let $W = \text{span}(\underline{w}_1, \underline{w}_2, \dots, \underline{w}_n)$ and notice that the subset

$$A = \{ \underline{u} \in V \mid \langle \underline{u}, \underline{w}_i \rangle = c_i \ \forall i = 1, \dots, n \}$$

is simply the orthogonal complement W^\perp translated by some fixed vector $\underline{v} \in A$. This is because, for all $\underline{x} \in W^\perp$, we have $\langle \underline{x}, \underline{w}_i \rangle = 0$ and $\langle \underline{v} - \underline{x}, \underline{w}_i \rangle = c_i$.

Thus, there is a one-to-one correspondence between $\underline{u} \in A$ and $\underline{x} \in W^\perp$ given by $\underline{u} = \underline{v} - \underline{x}$ and this implies

$$\min_{\underline{u} \in A} \|\underline{u}\| = \min_{\underline{x} \in W^\perp} \|\underline{v} - \underline{x}\|.$$

Since the optimizer \underline{x}^* of the RHS expression is the orthogonal projection of \underline{v} onto W^\perp , the resulting error vector $\underline{v} - \underline{x}^*$ must be contained in $(W^\perp)^\perp = W$. With this knowledge, we let $\underline{w} = \underline{v} - \underline{x}^*$ and use (4.9) to solve for s_1, \dots, s_n because we know the RHS satisfies $\langle \underline{v}, \underline{w}_i \rangle = c_i$. Since $\underline{w}_1, \dots, \underline{w}_n$ are linearly independent, the Gramian is full rank and this linear system has a unique solution. \square

4.5.2 Underdetermined Linear Systems

Let $A \in \mathbb{C}^{m \times n}$ with $m < n$ be the matrix representation of an underdetermined system of linearly independent equations and $\underline{v} \in \mathbb{C}^m$ be any column vector. Then, the dual approximation theorem can be applied to solve the problem

$$\min_{\underline{s}: A\underline{s} = \underline{v}} \|\underline{s}\|.$$

To see this as a dual approximation, we can rewrite the constraint as $(A^H)^H \underline{s} = \underline{v}$ so that the columns of A^H define the vectors $\underline{w}_1, \dots, \underline{w}_n$ in Theorem 4.5.1. Then, the theorem concludes that the minimum norm solution lies in $\mathcal{R}(A^H)$ (i.e., the column space of A^H). Using (4.10), one gathers that $\hat{\underline{s}} = A^H \underline{t}$ for some \underline{t} with $A(A^H \underline{t}) = \underline{v}$. If the rows of A are linearly independent, then the columns of A^H are linearly independent and $(AA^H)^{-1}$ exists. In this case, the solution $\hat{\underline{s}}$ can be obtained in closed form and is given by

$$\hat{\underline{s}} = A^H (AA^H)^{-1} \underline{v}.$$

4.6 Projection onto Convex Sets

So far, we have focused on the projection of vectors onto subspaces. In this section, similar results are obtained for the projection of vectors onto convex sets.

Definition 4.6.1. *Let V be a vector space. The subset $A \subseteq V$ is called a **convex set** if, for all $\underline{a}_1, \underline{a}_2 \in A$ and $\lambda \in (0, 1)$, we have $\lambda \underline{a}_1 + (1 - \lambda) \underline{a}_2 \in A$. The set is **strictly convex** if, for all $\underline{a}_1, \underline{a}_2 \in \overline{A}$ and $\lambda \in (0, 1)$, we have $\lambda \underline{a}_1 + (1 - \lambda) \underline{a}_2 \in A^\circ$.*

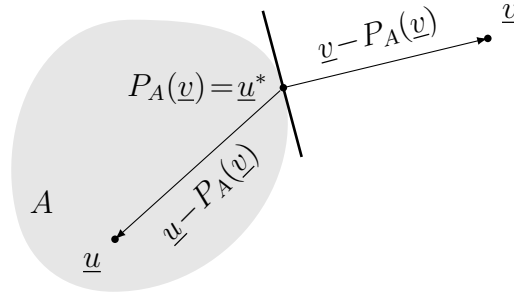


Figure 4.1: Orthogonal projection of \underline{v} onto closed convex set A .

Problem 4.6.2. Show that the intersection of convex sets is convex.

Definition 4.6.3. Let V be a Hilbert space and $A \subseteq V$ be a closed convex set. The **orthogonal projection** of $\underline{v} \in V$ onto A is the mapping $P_A: V \rightarrow A$ defined by

$$P_A(\underline{v}) \triangleq \arg \min_{\underline{u} \in A} \|\underline{u} - \underline{v}\|.$$

Remark 4.6.4. If A is compact, then the existence of the minimum (for any norm) is given by topology because $\|\underline{v} - \underline{u}\|$ is continuous in \underline{u} . Similarly, if both \underline{u} and \underline{u}' achieve the minimum distance d , then the convexity of the norm implies that the line segment between them also achieves the minimum distance. This implies that the closed ball of radius d in V contains a line segment on its boundary but one can use the Cauchy-Schwarz inequality to show this is impossible.

The following theorem instead uses vector space methods to establish the same result for all closed convex A .

Theorem 4.6.5. For Hilbert space V , the orthogonal projection of $\underline{v} \in V$ onto a closed convex set $A \subseteq V$ exists and is unique.

Proof. Let $d = \inf_{\underline{u} \in A} \|\underline{u} - \underline{v}\|$ be the infimal distance between \underline{v} and the set A . Next, consider any sequence $\underline{u}_1, \underline{u}_2, \dots \in A$ that achieves the infimum so that

$$\lim_{n \rightarrow \infty} \|\underline{u}_n - \underline{v}\| = d.$$

Since A is complete, the next step is showing that this sequence is Cauchy. The parallelogram law states that $\|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2 - \|x + y\|^2$ and applying

this to $\underline{x} = \underline{v} - \underline{u}_n$ and $\underline{y} = \underline{v} - \underline{u}_m$ gives

$$\begin{aligned} \|\underline{u}_m - \underline{u}_n\|^2 &= \|(\underline{v} - \underline{u}_n) - (\underline{v} - \underline{u}_m)\|^2 \\ &= 2\|\underline{v} - \underline{u}_n\|^2 + 2\|\underline{v} - \underline{u}_m\|^2 - \|(\underline{v} - \underline{u}_n) + (\underline{v} - \underline{u}_m)\|^2 \\ &= 2\|\underline{v} - \underline{u}_n\|^2 + 2\|\underline{v} - \underline{u}_m\|^2 - 4\left\|\underline{v} - \frac{\underline{u}_n + \underline{u}_m}{2}\right\|^2 \\ &\leq 2\|\underline{v} - \underline{u}_n\|^2 + 2\|\underline{v} - \underline{u}_m\|^2 - 4d^2 \end{aligned}$$

because the convexity of A implies $\frac{\underline{u}_n + \underline{u}_m}{2} \in A$ and therefore $\left\|\underline{v} - \frac{\underline{u}_n + \underline{u}_m}{2}\right\|^2 \geq d^2$. Since the limit of the RHS (as $m, n \rightarrow \infty$) equals 0, we find that the sequence \underline{u}_n is Cauchy and therefore the limit \underline{u}^* must exist. Since $\underline{u}_n \in A$ and A is closed, we also see that $\underline{u}^* \in A$. Therefore, the infimum is achieved as a minimum.

Uniqueness can be seen by assuming instead that $\underline{u}_m, \underline{u}_n$ are two elements in A which are both at a distance d from \underline{v} . Then, the above derivation shows that $\|\underline{u}_m - \underline{u}_n\|^2 \leq 0$. Therefore, they are the same point. \square

Remark 4.6.6. *The same result holds for norm projections in many other Banach spaces including L^p and ℓ^p for $1 < p < \infty$. In general, it is required that the Banach space be strictly convex (for uniqueness) and reflexive (for existence).*

Earlier in this chapter, we studied the equivalence between the orthogonality and Hilbert-space projections onto subspaces. The following result can be seen as a generalization of that result to Hilbert-space projections onto convex sets.

Theorem 4.6.7. *For any $\underline{v} \notin A$, a necessary and sufficient condition for $\underline{u}^* = P_A(\underline{v})$ is that $\operatorname{Re}\langle \underline{v} - \underline{u}^*, \underline{u} - \underline{u}^* \rangle \leq 0$ for all $\underline{u} \in A$.*

Proof. Let $\underline{u}^* = P_A(\underline{v})$ be the unique projection of \underline{v} onto A . For all $\underline{u} \in A$ and any $\alpha \in (0, 1)$, observe that $\underline{u}' = (1 - \alpha)\underline{u}^* + \alpha\underline{u} = \underline{u}^* + \alpha(\underline{u} - \underline{u}^*) \in A$ due to convexity. The optimality of \underline{u}^* implies that

$$\begin{aligned} \|\underline{v} - \underline{u}^*\|^2 &\leq \|\underline{v} - \underline{u}'\|^2 \\ &\leq \|\underline{v} - \underline{u}^* - \alpha(\underline{u} - \underline{u}^*)\|^2 \\ &= \|\underline{v} - \underline{u}^*\|^2 + \alpha^2\|\underline{u} - \underline{u}^*\|^2 - 2\alpha\operatorname{Re}\langle \underline{v} - \underline{u}^*, \underline{u} - \underline{u}^* \rangle. \end{aligned}$$

Thus, $\operatorname{Re}\langle \underline{v} - \underline{u}^*, \underline{u} - \underline{u}^* \rangle \leq \frac{\alpha}{2}\|\underline{u} - \underline{u}^*\|^2$. One can establish necessity by taking the limit as $\alpha \rightarrow 0$. For sufficiency, we assume $\operatorname{Re}\langle \underline{v} - \underline{u}^*, \underline{u} - \underline{u}^* \rangle \leq 0$ and we

write

$$\begin{aligned} \|\underline{v} - \underline{u}\|^2 - \|\underline{v} - \underline{u}^*\|^2 &= \|(\underline{v} - \underline{u}^*) - (\underline{u} - \underline{u}^*)\|^2 - \|\underline{v} - \underline{u}^*\|^2 \\ &= \|\underline{v} - \underline{u}^*\|^2 + \|\underline{u} - \underline{u}^*\|^2 - 2\operatorname{Re}\langle \underline{v} - \underline{u}^*, \underline{u} - \underline{u}^* \rangle - \|\underline{v} - \underline{u}^*\|^2 \\ &\geq 0. \end{aligned}$$

Thus, $\|\underline{v} - \underline{u}\|^2 \geq \|\underline{v} - \underline{u}^*\|^2$ for all $\underline{u} \in A$ and $\underline{u}^* = P_A(\underline{v})$. \square

4.6.1 Projection Properties and Examples

Let A be a closed convex subset of a Hilbert space V over \mathbb{R} . By drawing a simple picture (e.g., see Figure 4.1), one can see that projecting \underline{v} onto A is an operation that is translation invariant. Specifically, this means that translating the set A and the vector \underline{v} by the same vector \underline{v}_0 results in an output that is also translated by \underline{v}_0 . Mathematically, this means that, for all $\underline{v}, \underline{v}_0 \in V$, the projection onto V satisfies

$$\begin{aligned} P_{A+\underline{v}_0}(\underline{v} + \underline{v}_0) &= \arg \min_{\underline{u} \in A+\underline{v}_0} \|\underline{u} - \underline{v} - \underline{v}_0\| \\ &= \underline{v}_0 + \arg \min_{\underline{u}' \in A} \|(\underline{u}' + \underline{v}_0) - \underline{v} - \underline{v}_0\| \\ &= \underline{v}_0 + \arg \min_{\underline{u}' \in A} \|\underline{u}' - \underline{v}\| \\ &= \underline{v}_0 + P_A(\underline{v}). \end{aligned}$$

This also leads to the following trick. If a projection is easy when the set is centered, then one can: (i) translate the problem so that the set is centered, (ii) project onto the centered set, and (iii) translate back.

Using the best approximation theorem, it is easy to verify that the orthogonal projection of $\underline{v} \in V$ onto a one-dimensional subspace $W = \operatorname{span}(\underline{w})$ is given by

$$P_W(\underline{v}) = \frac{\langle \underline{v}, \underline{w} \rangle}{\|\underline{w}\|^2} \underline{w}.$$

A **hyperplane** is a closed subspace of $U \subset V$ that satisfies a single linear equality of the form $\langle \underline{v}, \underline{w} \rangle = 0$ for all $\underline{v} \in U$. Such a subspace is said to have co-dimension one (e.g., if $\dim(V) = n$, then $\dim(U) = n - 1$). Equivalently, U can be seen as the orthogonal complement of a one-dimensional subspace (e.g., $U = W^\perp$). Thus, we can write

$$P_U(\underline{v}) = P_{W^\perp}(\underline{v}) = \underline{v} - \frac{\langle \underline{v}, \underline{w} \rangle}{\|\underline{w}\|^2} \underline{w}.$$

Similarly, a linear equality such as $\langle \underline{v}, \underline{w} \rangle = c$, where \underline{v}_0 is any vector in V satisfying $\langle \underline{v}_0, \underline{w} \rangle = c$, defines an **affine hyperplane**. This is the shifted subspace $U + \underline{v}_0$ of co-dimension one because

$$\langle \underline{v}, \underline{w} \rangle = \langle \underline{u} + \underline{v}_0, \underline{w} \rangle = \langle \underline{u}, \underline{w} \rangle + \langle \underline{v}_0, \underline{w} \rangle = 0 + c = c.$$

Thus, we can project onto $U + \underline{v}_0$ by translating, projecting, and then translating back. This gives

$$P_{U+\underline{v}_0}(\underline{v}) = \left((\underline{v} - \underline{v}_0) - \frac{\langle \underline{v} - \underline{v}_0, \underline{w} \rangle}{\|\underline{w}\|^2} \underline{w} \right) + \underline{v}_0 = \underline{v} - \frac{\langle \underline{v}, \underline{w} \rangle - c}{\|\underline{w}\|^2} \underline{w},$$

which does not depend on the choice of \underline{v}_0 .

Next, let H be the subset of $\underline{v} \in V$ satisfying the linear inequality $\langle \underline{v}, \underline{w} \rangle \geq c$. Then, H is a closed convex set known as a **half space**. For any $\underline{v} \in H$, we have $P_H(\underline{v}) = \underline{v}$ and, for any $\underline{v} \notin H$, we have $P_H(\underline{v}) = P_{U+\underline{v}_0}(\underline{v})$ because the closest point must lie on the separating hyperplane and achieve the inequality with equality. For any $\underline{v} \in H$, one can put these together to see that

$$P_H(\underline{v}) = \begin{cases} \underline{v} & \text{if } \langle \underline{v}, \underline{w} \rangle \geq c \\ \underline{v} - \frac{\langle \underline{v}, \underline{w} \rangle - c}{\|\underline{w}\|^2} \underline{w} & \text{if } \langle \underline{v}, \underline{w} \rangle < c. \end{cases} \quad (4.11)$$

Theorem 4.6.8. *Let V be a Hilbert space over \mathbb{R} and $A \subset V$ be a closed convex set. For any $\underline{v} \notin A$, there is an affine hyperplane $U' = \{\underline{u} \in V \mid \langle \underline{u}, \underline{w} \rangle = c\}$ (defined by $\underline{w} \in V$ and $c \in \mathbb{R}$) such that $\langle \underline{v}, \underline{w} \rangle < c$ and $\langle \underline{u}, \underline{w} \rangle \geq c$ for all $\underline{u} \in A$.*

Proof. Let $\underline{u}^* = P_A(\underline{v})$ be the orthogonal projection of \underline{v} onto A and define $\underline{w} = \underline{u}^* - \underline{v}$ and $c = \langle \underline{u}^*, \underline{w} \rangle$. From Theorem 4.6.7, we see that $\langle \underline{v} - \underline{u}^*, \underline{u} - \underline{u}^* \rangle \leq 0$ for all $\underline{u} \in A$. Thus, for all $\underline{u} \in A$, we have

$$\begin{aligned} \langle \underline{u}, \underline{w} \rangle &= \langle \underline{w}, \underline{u} \rangle = \langle \underline{u}^* - \underline{v}, \underline{u} \rangle = -\langle \underline{v} - \underline{u}^*, \underline{u} \rangle \\ &\stackrel{(a)}{\geq} -\langle \underline{v} - \underline{u}^*, \underline{u}^* \rangle = \langle \underline{u}^* - \underline{v}, \underline{u}^* \rangle \\ &= \langle \underline{w}, \underline{u}^* \rangle = \langle \underline{u}^*, \underline{w} \rangle = c, \end{aligned}$$

where (a) follows from $\langle \underline{v} - \underline{u}^*, \underline{u} - \underline{u}^* \rangle \leq 0$ for all $\underline{u} \in A$. For $\langle \underline{v}, \underline{w} \rangle$, we observe that together, $\underline{u}^* \in A$ and $\underline{v} \notin A$, imply that

$$\begin{aligned} 0 &< \|\underline{u}^* - \underline{v}\|^2 = \langle \underline{u}^* - \underline{v}, \underline{u}^* - \underline{v} \rangle \\ &= \langle \underline{u}^*, \underline{u}^* - \underline{v} \rangle - \langle \underline{v}, \underline{u}^* - \underline{v} \rangle = c - \langle \underline{v}, \underline{u}^* - \underline{v} \rangle, \end{aligned}$$

which shows that $\langle \underline{v}, \underline{w} \rangle < c$ and completes the proof. \square

Corollary 4.6.9 (Farkas' Lemma). *Let $B \in \mathbb{R}^{m \times n}$ be a matrix and $\underline{v} \in \mathbb{R}^m$ be a vector. Then, exactly one of the following two conditions hold:*

1. *There exists $\underline{s} \in \mathbb{R}^n$ such that $B\underline{s} = \underline{v}$ and $\underline{s} \geq \underline{0}$.*
2. *There exists $\underline{w} \in \mathbb{R}^m$ such that $\underline{w}^T B \geq \underline{0}$ and $\underline{w}^T \underline{v} < 0$.*

Proof. Consider the closed convex set $A = \{B\underline{s} \mid \underline{s} \geq \underline{0}\}$, where $\underline{s} \geq \underline{0}$ denotes $s_i \geq 0$ for $i = 1, \dots, n$. If $\underline{v} \in A$, then we see that the first condition holds.

If $\underline{v} \notin A$, then we apply Theorem 4.6.8 using the standard Hilbert space \mathbb{R}^m to see that there exists $\underline{w} \in \mathbb{R}^m$ and $c \in \mathbb{R}$ such that $\underline{w}^T \underline{v} < c$ and $\underline{w}^T \underline{u} \geq c$ for all $\underline{u} \in A$. Next, we observe that A is a cone (i.e., for all $\alpha \geq 0$ and $\underline{u} \in A$, we have $\alpha \underline{u} \in A$). Thus, if there is any $\underline{u} \in A$ such that $\underline{w}^T \underline{u} < 0$, then $\underline{w}^T \underline{u}$ can take any non-positive value for some $\underline{u} \in A$. This gives a contradiction and shows that $\underline{w}^T \underline{u} \geq 0$ for all $\underline{u} \in A$. Since $\underline{0} \in A$, $c > 0$ would also give a contradiction and it follows that $c \leq 0$ and $\underline{w}^T \underline{v} < 0$.

This establishes that either the first or second condition must hold. But, if they both hold, then we get the contradiction

$$0 > \underline{w}^T \underline{v} = \underline{w}^T (B\underline{s}) = (\underline{w}^T B)\underline{s} \geq 0,$$

where the last step holds because $\underline{w}^T B \geq \underline{0}$ and $\underline{s} \geq \underline{0}$. Thus, exactly one holds. \square

Theorem 4.6.10. *Let V be Hilbert space over \mathbb{R} and $A \subset V$ be a closed convex set. Then, A equals the intersection of a set of half spaces.*

Proof. Let \mathcal{H} be the set of all half spaces in V and let $\mathcal{G} = \{H \in \mathcal{H} \mid A \subseteq H\}$ be the subset of half spaces containing A . For example, consider the half spaces defined by tangent planes passing through points on the boundary of A . Let $B = \bigcap_{H \in \mathcal{G}} H$ be the intersection of all the half spaces in \mathcal{G} . Since each half space contains A , it is clear that $A \subseteq B$. To show that $A = B$, we will show that $(x \notin A) \rightarrow (x \notin B)$. If $x \notin A$, then Theorem 4.6.8 shows that there is an affine hyperplane that separates x from A and an associated half space G that contains A but not x . Since G contains A , it follows that $G \in \mathcal{G}$. But, $x \notin G$ implies $x \notin B$ because B is the intersection of the half spaces in \mathcal{G} . This completes the proof. \square

4.6.2 Minimum Distance Between Two Convex Sets

Now, consider the smallest distance between two disjoint closed convex sets $A, B \subseteq V$. In this case, a unique solution may exist but a some things can go wrong. If the two sets are not strictly convex (e.g., consider two squares), then it is clearly possible for their to multiple pairs of points that achieve the minimum distance. Even if the two sets are strictly convex, one may find that the infimum is achieved as the points wander off to infinity. For example, consider the strictly convex hyperbolic sets $A = \{(x, y) | x^2 - y^2 \geq 1, x > 0\}$ and $B = \{(x, y) | y^2 - x^2 \geq 1, y \geq 0\}$. These two sets share the line $x = y > 0$ as an asymptote, so their infimal distance is 0.

To understand this behavior, we first note that the distance $f(\underline{u}, \underline{v}) = \|\underline{u} - \underline{v}\|$ is a convex function on the convex product set $A \times B$. It follows that any local minimum value is a global minimum value distance.

Theorem 4.6.11. *Let V be a Hilbert space and consider the infimal distance*

$$d = \inf_{\underline{u} \in A, \underline{v} \in B} \|\underline{u} - \underline{v}\|$$

between two disjoint closed convex sets $A, B \subseteq V$. If either set is compact, then the infimum is achieved. If the infimum is achieved and either set is strictly convex, then the minimizing points $\underline{u}^, \underline{v}^*$ are unique.*

Proof. Consider any sequence $(\underline{u}_1, \underline{v}_1), (\underline{u}_2, \underline{v}_2), \dots \in A \times B$ that satisfies

$$\lim_{n \rightarrow \infty} \|\underline{u}_n - \underline{v}_n\| = d.$$

If B is compact, then there is a subsequence \underline{v}_{n_j} that converges to some $\underline{v}^* \in B$. Since $\|P_A(\underline{v}_n) - \underline{v}_n\| \leq \|\underline{u}_n - \underline{v}_n\|$, we can replace \underline{u}_{n_j} by $P_A(\underline{v}_{n_j})$ and still achieve the infimum. The continuity of P_A also shows that $\underline{u}_{n_j} \rightarrow \underline{u}^* = P_A(\underline{v}^*)$ and this implies $\|\underline{u}^* - \underline{v}^*\| = d$. Also, $P_B(\underline{u}^*) = \underline{v}^*$ because \underline{v}^* is the unique closest point in B to \underline{u}^* . Notice that \underline{v}^* may not be unique (due to the subsequence construction) and, thus, the pair $(\underline{u}^*, \underline{v}^*)$ is not unique in general.

Since $\|\underline{u} - \underline{v}\|$ is a convex function on the convex product set $A \times B$, there is a (possibly empty) convex set of minimizers

$$M = \{(\underline{u}, \underline{v}) \in A \times B \mid \|\underline{u} - \underline{v}\| = d\}.$$

Also, each component of the $(\underline{u}, \underline{v})$ points in M must lie on the boundary of its set because otherwise one could reduce the smallest distance by moving one point along the minimum distance line towards the boundary. Now, suppose that (i) A is strictly convex and (ii) M contains more than one pair of minimizers. Then, condition (ii) implies that there must be two boundary points $\underline{u}_1, \underline{u}_2 \in \partial A$ such that $\alpha \underline{u}_1 + (1 - \alpha) \underline{u}_2 \in \partial A$ for $\alpha \in [0, 1]$. But this contradicts condition (i) and shows that, if A is strictly convex, then there is at most one pair $(\underline{u}^*, \underline{v}^*) \in M$ of minimizing points. \square

Remark 4.6.12. *Finding the minimum distance between two disjoint closed convex sets $A, B \subseteq V$ is a classic problem that is solved nicely by the idea of alternating minimization. Let $\underline{v}_0 \in B$ be an arbitrary initial point and define*

$$\begin{aligned}\underline{u}_{n+1} &= \arg \min_{\underline{u} \in A} \|\underline{u} - \underline{v}_n\| \\ \underline{v}_{n+1} &= \arg \min_{\underline{v} \in B} \|\underline{u}_{n+1} - \underline{v}\|.\end{aligned}$$

Notice that the sequence $d_n = \|\underline{u}_n - \underline{v}_n\|$ is non-increasing and must therefore have a limit. By adapting the previous proof, one can show that, if either set is compact, then the sequence $(\underline{u}_n, \underline{v}_n)$ converges to a pair of vectors that minimize the distance.

Theorem 4.6.13. *Let V be a Hilbert space over \mathbb{R} and A, B be disjoint closed convex subsets of V . If either set is compact, then there is an affine hyperplane $\{\underline{a} \in V \mid \langle \underline{a}, \underline{w} \rangle = c\}$ (defined by $\underline{w} \in V$ and $c \in \mathbb{R}$) such that $\langle \underline{u}, \underline{w} \rangle > c$ for all $\underline{u} \in A$ and $\langle \underline{u}, \underline{w} \rangle < c$ for all $\underline{u} \in B$.*

Proof. Applying Theorem 4.6.11 gives a pair of points $(\underline{u}^*, \underline{v}^*) \in A \times B$ that minimize the distance and satisfy $\underline{u}^* = P_A(\underline{v}^*)$ and $\underline{v}^* = P_B(\underline{u}^*)$. Applying Theorem 4.6.8 to $P_A(\underline{v}^*)$ shows that $\langle \underline{u}, \underline{u}^* - \underline{v}^* \rangle \geq \langle \underline{u}^*, \underline{u}^* - \underline{v}^* \rangle$ for all $\underline{u} \in A$. Similarly, applying Theorem 4.6.8 to $P_B(\underline{u}^*)$ shows that $\langle \underline{v}, \underline{v}^* - \underline{u}^* \rangle \geq \langle \underline{v}^*, \underline{v}^* - \underline{u}^* \rangle$ for all $\underline{v} \in B$. Negating this gives $\langle \underline{v}, \underline{u}^* - \underline{v}^* \rangle \leq \langle \underline{v}^*, \underline{u}^* - \underline{v}^* \rangle$. Now, we observe that $\langle \underline{u}^*, \underline{u}^* - \underline{v}^* \rangle - \langle \underline{v}^*, \underline{u}^* - \underline{v}^* \rangle = \|\underline{u}^* - \underline{v}^*\|^2 > 0$ because $\underline{u}^* \neq \underline{v}^*$. Thus, we can choose $\underline{w} = \underline{u}^* - \underline{v}^*$ and $c = \frac{1}{2}(\langle \underline{u}^*, \underline{u}^* - \underline{v}^* \rangle + \langle \underline{v}^*, \underline{u}^* - \underline{v}^* \rangle)$ to guarantee that $\langle \underline{u}, \underline{w} \rangle > c$ for all $\underline{u} \in A$ and $\langle \underline{u}, \underline{w} \rangle < c$ for all $\underline{u} \in B$. \square

Chapter 5

Optimization

The foundation of engineering is the ability to use math and physics to design and optimize complex systems. The advent of computers has made this possible on an unprecedented scale. This chapter provides a brief introduction to mathematical optimization theory.

5.1 Derivatives in Banach Spaces

In this chapter, we assume that readers are familiar with derivatives as defined in undergraduate multivariable calculus. To gain insight, we first recall the standard interpretation of the derivative as a local linear approximation of a function. For a function $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$, this interpretation gives

$$f(\underline{x} + \underline{h}) = f(\underline{x}) + J(\underline{x}) \underline{h} + \text{higher order terms},$$

where $J(\underline{x}) \in \mathbb{R}^{m \times n}$ is the Jacobian matrix of f at \underline{x} .

Instead of interpreting a multivariate derivative as a matrix, we will view the derivative $f'(x)$ as a linear transform T from the domain to codomain. This transform maps the input perturbation \underline{h} to a local approximation of the output perturbation. Since both are finite dimensional in our example, the linear transform T is represented by the Jacobian matrix and we have

$$f'(\underline{x})(\underline{h}) = T\underline{h} = J(\underline{x}) \underline{h}.$$

Mathematically, such definitions require the structure of a Banach space because

one needs the linear structure to compute differences, and the norm topology to define limits. Completeness guarantees that limits exist under mild conditions.

In Chapter 2, limits are discussed primarily for sequences but Definition 2.1.27 does extend this to cases where points approach other points. By treating Banach spaces as metric spaces, we arrive at the following induced definition of such limits.

Definition 5.1.1. For Banach spaces X, Y with $f: X \rightarrow Y$ and $x, z \in X$, we define $\lim_{x \rightarrow z} f(x)$ as follows. For $y \in Y$, we say that $\lim_{x \rightarrow z} f(x) = y$ (i.e., the limit as $x \rightarrow z$ of $f(x)$ equals y) if $f(x_n) \rightarrow y$ for all sequences $x_n \in X$ such that $x_n \rightarrow z$. If there is no such $y \in Y$, then we say the limit does not exist.

The proof of Theorem 2.1.26 shows that this is equivalent to: for any $\epsilon > 0$, there is a $\delta > 0$ such that, for all $x \in X$ satisfying $\|x - z\| < \delta$, we have $\|y - f(x)\| < \epsilon$.

Definition 5.1.2. Let $f: X \rightarrow Y$ be a mapping from a vector space X over \mathbb{R} to a Banach space $(Y, \|\cdot\|)$. Then, if it exists, the **Gâteaux differential** (or directional derivative) of f at \underline{x} in direction \underline{h} is given by

$$\delta f(\underline{x}; \underline{h}) \triangleq \lim_{t \rightarrow 0} \frac{f(\underline{x} + t\underline{h}) - f(\underline{x})}{t},$$

where the limit is with respect to the implied mapping from $t \in \mathbb{R}$ to Y .

When this directional derivative exists, we can write the approximation

$$f(\underline{x} + t\underline{h}) \approx f(\underline{x}) + t\delta f(\underline{x}; \underline{h}).$$

In fact, we can get a tighter characterization that is especially meaningful in the context of optimization.

Lemma 5.1.3. Let $Y = (\mathbb{R}, |\cdot|)$ and suppose that $\delta f(\underline{x}; \underline{h})$ exists and is negative for some f , \underline{x} , and \underline{h} . Then, there exists $t_0 > 0$ such that $f(\underline{x} + t\underline{h}) < f(\underline{x})$ for all $t \in (0, t_0)$.

Proof. The $\delta f(\underline{x}; \underline{h})$ limit implies that, for any $\epsilon > 0$, there is a $t_0 > 0$ such that

$$f(\underline{x} + t\underline{h}) - f(\underline{x}) \leq (\delta f(\underline{x}; \underline{h}) + \epsilon)t$$

for all $t \in (0, t_0)$. If $\delta f(\underline{x}; \underline{h}) < 0$, then one can choose $\epsilon = -\frac{1}{2}\delta f(\underline{x}; \underline{h})$ to see that the RHS is negative for all $t \in (0, t_0)$. The stated result follows. \square

Example 5.1.4. For the standard Banach space $X = Y = \mathbb{R}^2$, let $f(\underline{x}) = (x_1 x_2, x_1 + x_2^2)$. Then, for $\underline{x} = (1, 1)$, $\underline{h} = (1, 2)$, we have

$$\delta f(\underline{x}, \underline{h}) = \frac{d}{dt}((1+t)(1+2t), (1+t) + (1+2t)^2) \Big|_{t=0} = (3, 5).$$

Problem 5.1.5. Suppose $X = Y = L^1([0, 1])$ is the Banach space of Lebesgue absolutely integrable functions mapping $[0, 1]$ to \mathbb{R} and $f(\underline{x}) = \|\underline{x}\| = \int_0^1 |x(s)| ds$ is the norm of \underline{x} . Assuming the set $\{s \in [0, 1] | x(s) = 0\}$ has measure 0, show that

$$\delta f(\underline{x}; \underline{h}) \triangleq \lim_{t \rightarrow 0} \int_0^1 \frac{1}{t} (|x(s) + th(s)| - |x(s)|) ds = \int_0^1 \text{sgn}(x(s)) h(s) ds.$$

Definition 5.1.6. Let $f: X \rightarrow Y$ be a mapping from a vector space X over \mathbb{R} to a Banach space $(Y, \|\cdot\|)$. Then, f is **Gâteaux differentiable** at \underline{x} if the Gâteaux differential $\delta f(\underline{x}; \underline{h})$ exists for all $\underline{h} \in X$ and is a linear function of \underline{h} . If, in addition, X is a Banach space, then $\delta f(\underline{x}; \underline{h})$ must be a continuous linear function of \underline{h} .

Remark 5.1.7. For simplicity, our treatment of Gâteaux derivatives assumes X is a vector space over \mathbb{R} but similar results are possible over \mathbb{C} as well.

Definition 5.1.8. Let $f: X \rightarrow Y$ be a mapping from a Banach space $(X, \|\cdot\|_X)$ to a Banach space $(Y, \|\cdot\|_Y)$. Then, f is **Fréchet differentiable** at \underline{x} if there is a linear transformation $T: X \rightarrow Y$ with $\|T\| < \infty$ that satisfies

$$\lim_{\underline{h} \rightarrow 0} \frac{\|f(\underline{x} + \underline{h}) - f(\underline{x}) - T(\underline{h})\|_Y}{\|\underline{h}\|_X} = 0, \quad (5.1)$$

where the limit is with respect to the implied Banach space mapping $X \rightarrow \mathbb{R}$. In this case, the **Fréchet derivative** at \underline{x} equals T and is denoted by $f'(\underline{x})$ in general.

Example 5.1.9. A function $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ with $f = (f_1, f_2, \dots, f_m)^T$ is (Fréchet) differentiable at \underline{x}_0 if the mapping J from \mathbb{R}^n to the **Jacobian matrix**,

$$J(\underline{x}) = f'(\underline{x}) \triangleq \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\underline{x}) & \frac{\partial f_1}{\partial x_2}(\underline{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\underline{x}) \\ \frac{\partial f_2}{\partial x_1}(\underline{x}) & \frac{\partial f_2}{\partial x_2}(\underline{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\underline{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\underline{x}) & \frac{\partial f_m}{\partial x_2}(\underline{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\underline{x}) \end{bmatrix},$$

exists and is continuous in \underline{x} at $\underline{x} = \underline{x}_0$. A necessary and sufficient condition for this is that each partial derivative is continuous in \underline{x} at $\underline{x} = \underline{x}_0$.

If $m = 1$, then the Jacobian is closely related to the **gradient** of the function

$$\nabla f(\underline{x}) \triangleq f'(\underline{x})^H = \left[\frac{\partial f}{\partial x_1}(\underline{x}) \quad \frac{\partial f}{\partial x_2}(\underline{x}) \quad \cdots \quad \frac{\partial f}{\partial x_n}(\underline{x}) \right]^H.$$

It is worth noting that the orientation of the gradient vector (i.e., row versus column vector) is sometimes defined differently. This is because derivatives can be understood as linear transforms and either orientation can be used to define the correct linear transform.

Example 5.1.10. Let X be a Hilbert space over \mathbb{R} and $f: X \rightarrow \mathbb{R}$ be a real functional. If the Fréchet derivative $f'(\underline{x})$ exists, then it is a continuous linear functional on X . Thus, the Riesz representation theorem guarantees that there is a vector $\underline{u} \in X$ such that $f'(\underline{x})(\underline{h}) = \langle \underline{h}, \underline{u} \rangle$ for all $\underline{h} \in X$. This vector is called the gradient $\nabla f(\underline{x})$ and it follows that

$$f'(\underline{x})(\underline{h}) = \langle \underline{h}, \nabla f(\underline{x}) \rangle \text{ for all } \underline{h} \in X.$$

Problem 5.1.11. In the setting of the previous example, show that, if $\nabla f(\underline{x}) \neq \underline{0}$, then $f(\underline{x} - \delta \nabla f(\underline{x})) < f(\underline{x})$ for some $\delta > 0$.

Theorem 5.1.12. Let $f: X \rightarrow Y$ be a mapping from a Banach space $(X, \|\cdot\|_X)$ to a Banach space $(Y, \|\cdot\|_Y)$. If f is Fréchet differentiable at \underline{x} with derivative f' , then f is Gâteaux differentiable at \underline{x} with Gâteaux differential $\delta f(\underline{x}; \underline{h}) = f'(\underline{x})(\underline{h})$.

Proof. For $\underline{h} = \underline{0}$, the statement is trivial. For $\underline{h} \neq \underline{0}$, we first observe that $t\underline{h} \rightarrow \underline{0}$ as $t \rightarrow 0$. Letting $T = f'(\underline{x})$, we can combine this with (5.1) to see that

$$\begin{aligned} 0 &= \lim_{t \rightarrow 0} \frac{\|f(\underline{x} + t\underline{h}) - f(\underline{x}) - T(t\underline{h})\|_Y}{\|t\underline{h}\|_X} \\ &= \lim_{t \rightarrow 0} \left\| \frac{f(\underline{x} + t\underline{h}) - f(\underline{x})}{t\|\underline{h}\|_X} - \frac{tT(\underline{h})}{t\|\underline{h}\|_X} \right\|_Y \\ &= \frac{1}{\|\underline{h}\|_X} \lim_{t \rightarrow 0} \left\| \frac{f(\underline{x} + t\underline{h}) - f(\underline{x})}{t} - T(\underline{h}) \right\|_Y. \end{aligned}$$

Thus, the Gâteaux differential exists and satisfies $\delta f(\underline{x}; \underline{h}) = T(\underline{h}) = f'(\underline{x})(\underline{h})$. \square

Theorem 5.1.13. Let X, Y, Z be Banach spaces and let $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ be functions. If f is Fréchet differentiable at \underline{x} and g is Fréchet differentiable at $\underline{y} = f(\underline{x})$, then $(g \circ f)(\underline{x}) = g(f(\underline{x}))$ is Fréchet differentiable at \underline{x} with derivative $g'(f(\underline{x})) \circ f'(\underline{x})$.

Proof. For the stated derivatives, the errors in the implied linear approximations are

$$\begin{aligned}\phi(\underline{v}) &= f(\underline{x} + \underline{v}) - f(\underline{x}) - f'(\underline{x})(\underline{v}) \\ \psi(\underline{u}) &= g(\underline{y} + \underline{u}) - g(\underline{y}) - g'(\underline{y})(\underline{u}) \\ \rho(\underline{h}) &= g(f(\underline{x} + \underline{h})) - g(f(\underline{x})) - (g'(\underline{y}) \circ f'(\underline{x}))(\underline{h}).\end{aligned}$$

From the assumptions of differentiability, we know that the first two approximations become tight for small perturbations. In other words,

$$\lim_{\underline{v} \rightarrow \underline{0}} \frac{\|\phi(\underline{v})\|_Y}{\|\underline{v}\|_X} = 0, \quad \lim_{\underline{u} \rightarrow \underline{0}} \frac{\|\psi(\underline{u})\|_Z}{\|\underline{u}\|_Y} = 0.$$

Next, we observe that the definition of ϕ implies

$$g(f(\underline{x} + \underline{h})) - g(f(\underline{x})) = g(f(\underline{x}) + f'(\underline{x})(\underline{h}) + \phi(\underline{h})) - g(\underline{y}).$$

Combining this with the definition of ρ shows that

$$\begin{aligned}\rho(\underline{h}) &= g(f(\underline{x}) + f'(\underline{x})(\underline{h}) + \phi(\underline{h})) - g(\underline{y}) - (g'(\underline{y}) \circ f'(\underline{x}))(\underline{h}) \\ &= \psi(f'(\underline{x})(\underline{h}) + \phi(\underline{h})) + g'(\underline{y})(f'(\underline{x})(\underline{h}) + \phi(\underline{h})) - (g'(\underline{y}) \circ f'(\underline{x}))(\underline{h}) \\ &= \psi(f'(\underline{x})(\underline{h}) + \phi(\underline{h})) + g'(\underline{y})(\phi(\underline{h})).\end{aligned}$$

We take this opportunity to note that $\|g'(f(\underline{x})) \circ f'(\underline{x})\| \leq \|g'(f(\underline{x}))\| \|f'(\underline{x})\| \leq \infty$ because $\|f'(\underline{x})\| \leq \infty$ and $\|g'(f(\underline{x}))\| < \infty$. Since $\lim_{\underline{h} \rightarrow \underline{0}} \|\phi(\underline{h})\|_Y / \|\underline{h}\|_X = 0$, there is a $t > 0$ such that $\|\phi(\underline{h})\|_Y \leq \|f'(\underline{x})\| \|\underline{h}\|_X$ if $\|\underline{h}\|_X < t$. Under the same condition, it follows that $2\|f'(\underline{x})\| \|\underline{h}\|_X \geq \|f'(\underline{x})\| \|\underline{h}\|_X + \|\phi(\underline{h})\|_Y$. Using this, we can write

$$\begin{aligned}\frac{\|\rho(\underline{h})\|_Z}{\|\underline{h}\|_X} &= \frac{\|\psi(f'(\underline{x})(\underline{h}) + \phi(\underline{h})) + g'(\underline{y})(\phi(\underline{h}))\|_Z}{\|\underline{h}\|_X} \\ &\leq 2\|f'(\underline{x})\| \frac{\|\psi(f'(\underline{x})(\underline{h}) + \phi(\underline{h}))\|_Z}{2\|f'(\underline{x})\| \|\underline{h}\|_X} + \frac{\|g'(\underline{y})(\phi(\underline{h}))\|_Z}{\|\underline{h}\|_X} \\ &\leq 2\|f'(\underline{x})\| \frac{\|\psi(f'(\underline{x})(\underline{h}) + \phi(\underline{h}))\|_Z}{\|f'(\underline{x})\| \|\underline{h}\|_X + \|\phi(\underline{h})\|_Y} + \frac{\|g'(\underline{y})\| \|\phi(\underline{h})\|_Y}{\|\underline{h}\|_X} \\ &\leq 2\|f'(\underline{x})\| \frac{\|\psi(f'(\underline{x})(\underline{h}) + \phi(\underline{h}))\|_Z}{\|f'(\underline{x})(\underline{h}) + \phi(\underline{h})\|_Y} + \frac{\|g'(\underline{y})\| \|\phi(\underline{h})\|_Y}{\|\underline{h}\|_X}.\end{aligned}$$

Since $(f'(\underline{x})(\underline{h}) + \phi(\underline{h})) \rightarrow \underline{0}$ as $\underline{h} \rightarrow \underline{0}$, it follows that the limit of the RHS, as $\underline{h} \rightarrow \underline{0}$, also exists and equals 0. Thus, $\lim_{\underline{h} \rightarrow \underline{0}} \|\rho(\underline{h})\|_Z / \|\underline{h}\|_X = 0$ and the Fréchet derivative of $g(f(\underline{x}))$ exists and satisfies the chain rule. \square

Theorem 5.1.14. *Let X, Y be Banach spaces and $f: X \rightarrow Y$ be a function. For $\underline{x}_1, \underline{x}_2 \in X$, let $\underline{h} = \underline{x}_2 - \underline{x}_1$ and assume the Gâteaux differential $\delta f((1-s)\underline{x}_1 + s\underline{x}_2; \underline{h})$ exists for all $s \in [0, 1]$. Then, $\|f(\underline{x}_2) - f(\underline{x}_1)\| \leq M\|\underline{x}_2 - \underline{x}_1\|$, where*

$$M = \frac{\sup_{s \in [0,1]} \|\delta f((1-s)\underline{x}_1 + s\underline{x}_2; \underline{h})\|}{\|\underline{x}_2 - \underline{x}_1\|}.$$

Proof. For $\underline{w}_1 = \frac{1}{2}(\underline{x}_1 + \underline{x}_2)$, observe that

$$\begin{aligned} \frac{\|f(\underline{x}_2) - f(\underline{x}_1)\|}{\|\underline{x}_2 - \underline{x}_1\|} &= \frac{\|f(\underline{x}_2) - f(\underline{w}_1) + f(\underline{w}_1) - f(\underline{x}_1)\|}{\|\underline{x}_2 - \underline{x}_1\|} \\ &\leq \frac{\|f(\underline{x}_2) - f(\underline{w}_1)\| + \|f(\underline{w}_1) - f(\underline{x}_1)\|}{\|\underline{x}_2 - \underline{x}_1\|} \\ &= \frac{\|f(\underline{x}_2) - f(\underline{w}_1)\|}{2\|\underline{x}_2 - \underline{w}_1\|} + \frac{\|f(\underline{w}_1) - f(\underline{x}_1)\|}{2\|\underline{w}_1 - \underline{x}_1\|}. \end{aligned}$$

Suppose that $\|f(\underline{x}_2) - f(\underline{x}_1)\| > M\|\underline{x}_2 - \underline{x}_1\|$. Then, there is an $\epsilon > 0$ such that one or both of the following conditions must hold:

$$\frac{\|f(\underline{x}_2) - f(\underline{w}_1)\|}{\|\underline{x}_2 - \underline{w}_1\|} \geq M + \epsilon \quad \text{and} \quad \frac{\|f(\underline{w}_1) - f(\underline{x}_1)\|}{\|\underline{w}_1 - \underline{x}_1\|} \geq M + \epsilon.$$

Repeating indefinitely and choosing a satisfying subinterval at each step, one gets a sequence \underline{w}_n of midpoints that converges to $\underline{x} = (1-s)\underline{x}_1 + s\underline{x}_2$ for some $s \in [0, 1]$. Since the Gâteaux differential $\delta f(\underline{x}; \underline{h})$ exists by assumption, it follows that

$$M + \epsilon \leq \frac{\|f(\underline{w}_n) - f(\underline{x})\|}{\|\underline{w}_n - \underline{x}\|} = \left\| \frac{f(\underline{x} \pm 2^{-n}\underline{h}) - f(\underline{x})}{2^n\|\underline{x}_2 - \underline{x}_1\|} \right\| \rightarrow \frac{\|\delta f(\underline{x}; \underline{h})\|}{\|\underline{x}_2 - \underline{x}_1\|}.$$

This contradicts the definition of M and, thus, $\|f(\underline{x}_2) - f(\underline{x}_1)\| \leq M\|\underline{x}_2 - \underline{x}_1\|$. \square

Lemma 5.1.15. *Let X, Y be Banach spaces and $f: X \rightarrow Y$ be a function. If the Fréchet derivative $f'(\underline{x})$ exists and satisfies $\|f'(\underline{x})\| \leq L$ for all \underline{x} in a convex set $A \subseteq X$, then f is Lipschitz continuous on A with Lipschitz constant L .*

Proof. Assume $\|f'(\underline{x})\| \leq L$ for all \underline{x} in a convex set $A \subseteq X$. Then, for any $\underline{x}_1, \underline{x}_2 \in A$, let $\underline{h} = \underline{x}_2 - \underline{x}_1$ and notice that Theorem 5.1.12 implies that

$$\|\delta f(\underline{x}_1 + s\underline{h}; \underline{h})\| = \|f'(\underline{x}_1 + s\underline{h})(\underline{h})\| \leq \|f'(\underline{x}_1 + s\underline{h})\|\|\underline{h}\|,$$

for all $s \in [0, 1]$. Applying Theorem 5.1.14, we see that $\|f(\underline{x}_2) - f(\underline{x}_1)\| \leq M\|\underline{x}_2 - \underline{x}_1\|$ with $M \leq \|f'(\underline{x})\| \leq L$. This completes the proof. \square

Lemma 5.1.16. *Let $f: X \rightarrow \mathbb{R}$ map the Hilbert space X to the real numbers. If $\nabla f(\underline{x})$ exists and satisfies $\|\nabla f(\underline{y}) - \nabla f(\underline{x})\| \leq L\|\underline{y} - \underline{x}\|$, then*

$$|f(\underline{y}) - f(\underline{x}) - \langle \underline{y} - \underline{x}, \nabla f(\underline{x}) \rangle| \leq \frac{1}{2}L\|\underline{y} - \underline{x}\|^2.$$

Proof. Let $\underline{h} = \underline{y} - \underline{x}$ and $\phi(t) = f(\underline{x} + t\underline{h})$. Then, $\phi'(t) = \langle \underline{h}, \nabla f(\underline{x} + t\underline{h}) \rangle$ and

$$\begin{aligned} |f(\underline{y}) - f(\underline{x}) - \langle \underline{h}, \nabla f(\underline{x}) \rangle| &= \left| \int_0^1 (\phi'(t) - \phi'(0)) dt \right| \\ &= \left| \int_0^1 \langle \underline{h}, \nabla f(\underline{x} + t\underline{h}) - \nabla f(\underline{x}) \rangle dt \right| \\ &\leq \left| \int_0^1 \|\underline{h}\| \|\nabla f(\underline{x} + t\underline{h}) - \nabla f(\underline{x})\| dt \right| \\ &\leq \int_0^1 \|\underline{h}\| L \|t\underline{h}\| dt \\ &= \frac{1}{2}L\|\underline{h}\|^2. \end{aligned} \quad \square$$

5.2 Unconstrained Optimization

Functions mapping elements of a vector space (over F) down to the scalar field F play a very special role in the analysis of vector spaces.

Definition 5.2.1. *Let V be a vector space over F . Then, a **functional** on V is a function $f: V \rightarrow F$ that maps V to F .*

Linear functionals (i.e., functionals that are linear) are used to define many important concepts in abstract vector spaces. For unconstrained optimization, however, linear functionals are not interesting because they are either zero or they achieve all values in F .

Definition 5.2.2. *Let $(X, \|\cdot\|)$ be a normed vector space. Then, a real functional $f: X \rightarrow \mathbb{R}$ achieves a **local minimum value** at $\underline{x}_0 \in X$ if there is an $\epsilon > 0$ such that, for all $\underline{x} \in X$ satisfying $\|\underline{x} - \underline{x}_0\| < \epsilon$, we have $f(\underline{x}) \geq f(\underline{x}_0)$. If the bound holds for all $\underline{x} \in X$, then the local minimum is also a **global minimum value**.*

Theorem 5.2.3. *Let $(X, \|\cdot\|)$ be a normed vector space and $f: X \rightarrow \mathbb{R}$ be a real functional. If $\delta f(\underline{x}_0, \underline{h})$ exists and is negative for any $\underline{h} \in X$, then \underline{x}_0 is not a local minimum value.*

Proof. First, we apply Lemma 5.1.3 with the \underline{x} and \underline{h} for which $\delta f(\underline{x}_0, \underline{h}) < 0$. This gives a $t_0 > 0$ such that $f(\underline{x}_0 + t\underline{h}) < f(\underline{x}_0)$ for all $t \in (0, t_0)$. Thus, there can be no $\epsilon > 0$ satisfying the definition of a local minimum value in Definition 5.2.2. \square

5.3 Convex Functionals

Convexity is a particularly nice property of spaces and functionals that leads to well-defined minimum values.

Definition 5.3.1. Let V be a vector space, $A \subseteq V$ be a convex set, and $f: V \rightarrow \mathbb{R}$ be a functional. Then, a functional f is called **convex** on A if, for all $\underline{a}_1, \underline{a}_2 \in A$ and $\lambda \in (0, 1)$, we have

$$f(\lambda \underline{a}_1 + (1 - \lambda) \underline{a}_2) \leq \lambda f(\underline{a}_1) + (1 - \lambda) f(\underline{a}_2).$$

The functional is **strictly convex** if equality occurs only when $\underline{a}_1 = \underline{a}_2$. A functional f is called (strictly) **concave** if $-f$ is (strictly) convex.

Definition 5.3.2. A Banach space X is called **strictly convex** if the unit ball, given by $\{x \in X \mid \|x\| \leq 1\}$, is a strictly convex set. An equivalent condition is that equality in the triangle inequality (i.e., $\|\underline{x} + \underline{y}\| = \|\underline{x}\| + \|\underline{y}\|$) for non-zero vectors implies that $\underline{x} = s\underline{y}$ for some $s \in F$.

Example 5.3.3. Let $(X, \|\cdot\|)$ be a normed vector space. Then, the norm $\|\cdot\|: X \rightarrow \mathbb{R}$ is a convex functional on X . Proving this is a good introductory exercise.

Example 5.3.4. Let X be an inner-product space. For $\underline{x}, \underline{y} \in X$ and $\lambda \in (0, 1)$,

$$\begin{aligned} \|\lambda \underline{x} + (1 - \lambda) \underline{y}\|^2 &= \lambda^2 \|\underline{x}\|^2 + 2\lambda(1 - \lambda) \operatorname{Re}\langle \underline{x}, \underline{y} \rangle + (1 - \lambda)^2 \|\underline{y}\|^2 \\ &= \lambda \|\underline{x}\|^2 + (1 - \lambda) \|\underline{y}\|^2 - \lambda(1 - \lambda) (\|\underline{x}\|^2 + \|\underline{y}\|^2 - 2\operatorname{Re}\langle \underline{x}, \underline{y} \rangle) \\ &= \lambda \|\underline{x}\|^2 + (1 - \lambda) \|\underline{y}\|^2 - \lambda(1 - \lambda) \|\underline{x} - \underline{y}\|^2 \\ &\leq \lambda \|\underline{x}\|^2 + (1 - \lambda) \|\underline{y}\|^2, \end{aligned}$$

with equality iff $\underline{x} = \underline{y}$. Thus, the square of the induced norm $\|\cdot\|^2$ is a strictly convex functional on X .

Theorem 5.3.5. *Let $(X, \|\cdot\|)$ be a normed vector space, $A \subseteq X$ be a convex set, and $f: X \rightarrow \mathbb{R}$ be a convex functional on A . Then, any local minimum value of f on A is a global minimum value on A . If the functional is strictly convex on A and achieves a local minimum value on A , then there is a unique point $\underline{x}_0 \in A$ that achieves the global minimum value on A .*

Proof. Let $\underline{x}_0 \in A$ a point where the functional achieves a local minimum value. Proving by contradiction, we suppose that there is another point $\underline{x}_1 \in A$ such that $f(\underline{x}_1) < f(\underline{x}_0)$. From the definition of a local minimum value, we find an $\epsilon > 0$ such that $f(\underline{x}) \geq f(\underline{x}_0)$ for all $\underline{x} \in A$ satisfying $\|\underline{x} - \underline{x}_0\| < \epsilon$. Choosing $\lambda < \frac{\epsilon}{\|\underline{x}_0 - \underline{x}_1\|}$ in $(0, 1)$ and $\underline{x} = (1 - \lambda)\underline{x}_0 + \lambda\underline{x}_1$ implies that $\|\underline{x} - \underline{x}_0\| < \epsilon$ while the convexity of f implies that

$$f(\underline{x}) = f((1 - \lambda)\underline{x}_0 + \lambda\underline{x}_1) \leq (1 - \lambda)f(\underline{x}_0) + \lambda f(\underline{x}_1) < f(\underline{x}_0).$$

This contradicts the definition of a local minimum value and implies that $f(\underline{x}_0)$ is a global minimum value on A . If f is strictly convex and $f(\underline{x}_1) = f(\underline{x}_0)$, then we suppose that $\underline{x}_0 \neq \underline{x}_1$. In this case, strict convexity implies that

$$f((1 - \lambda)\underline{x}_0 + \lambda\underline{x}_1) < (1 - \lambda)f(\underline{x}_0) + \lambda f(\underline{x}_1) = f(\underline{x}_0).$$

This contradicts the fact that $f(\underline{x}_0)$ is a global minimum value on A and implies that $\underline{x}_0 = \underline{x}_1$ is unique. \square

Theorem 5.3.6. *Let $(X, \|\cdot\|)$ be a normed vector space and $f: X \rightarrow \mathbb{R}$ be a convex functional on a convex set $A \subseteq X$. If f is Gâteaux differentiable at $\underline{x}_0 \in A$, then*

$$f(\underline{x}) \geq f(\underline{x}_0) + \delta f(\underline{x}_0; \underline{x} - \underline{x}_0)$$

for all $\underline{x} \in A$. If f is strictly convex then the inequality is strict for $\underline{x} \neq \underline{x}_0$.

Proof. By the convexity of A and f , we have $\underline{x}_0 + \lambda(\underline{x} - \underline{x}_0) \in A$ and

$$f(\underline{x}_0 + \lambda(\underline{x} - \underline{x}_0)) \leq f(\underline{x}_0) + \lambda(f(\underline{x}) - f(\underline{x}_0)) \quad (5.2)$$

for all $\lambda \in (0, 1)$. Also, if f is strictly convex, then (5.2) strict for $\underline{x} \neq \underline{x}_0$. Thus,

$$f(\underline{x}) \geq f(\underline{x}_0) + \frac{f(\underline{x}_0 + \lambda(\underline{x} - \underline{x}_0)) - f(\underline{x}_0)}{\lambda}$$

and taking the limit at $\lambda \downarrow 0$ completes the proof for a convex functional.

For the case where f is strictly convex, we first apply the convex result to see

$$f(\underline{x}_0 + \lambda(\underline{x} - \underline{x}_0)) \geq f(\underline{x}_0) + \delta f(\underline{x}_0; \lambda(\underline{x} - \underline{x}_0)) = f(\underline{x}_0) + \lambda \delta f(\underline{x}_0; \underline{x} - \underline{x}_0),$$

where the second step holds because $\delta f(\underline{x}; \underline{h})$ is linear in \underline{h} . This gives

$$\delta f(\underline{x}_0; \underline{x} - \underline{x}_0) \leq \frac{f(\underline{x}_0 + \lambda(\underline{x} - \underline{x}_0)) - f(\underline{x}_0)}{\lambda} < f(\underline{x}) - f(\underline{x}_0),$$

where the second inequality holds because (5.2) is a strict inequality for $\underline{x} \neq \underline{x}_0$. \square

Corollary 5.3.7. *Let $(X, \|\cdot\|)$ be a normed vector space and $f: X \rightarrow \mathbb{R}$ be a convex functional on a convex set $A \subseteq X$. If f is Gâteaux differentiable at $\underline{x}_0 \in A$ and $\delta f(\underline{x}_0; \underline{x} - \underline{x}_0) = 0$ for all $\underline{x} \in A$, then*

$$f(\underline{x}_0) = \min_{\underline{x} \in A} f(\underline{x}).$$

If f is strictly convex, \underline{x}_0 is the unique minimizer over A .

5.4 Constrained Optimization

Lagrangian optimization is an indispensable tool in engineering and physics that allows one to solve constrained non-linear optimization problems. For convex problems, there are now efficient algorithms that can handle thousands of variables and constraints. In some cases, there are also analytical techniques that allow one to derive tight bounds on optimum value. These approaches have become so common that convex Lagrangian optimization problems are now taught as a fundamental part of the graduate engineering curriculum. For simplicity, we focus on the case where the domain \mathcal{D} is a subset of the finite-dimensional real space \mathbb{R}^n .

Constrained non-linear optimization problems over $\mathcal{D} \subseteq \mathbb{R}^n$ can be put into the following **standard form**. Let $f_i: \mathcal{D} \rightarrow \mathbb{R}$ and $h_j: \mathcal{D} \rightarrow \mathbb{R}$ be a real functionals on \mathcal{D} for $i = 0, 1, \dots, m$ and $j = 1, 2, \dots, p$. Then, the standard form is

$$\begin{aligned} & \text{minimize} && f_0(\underline{x}) \\ & \text{subject to} && f_i(\underline{x}) \leq 0, \quad i = 1, 2, \dots, m \\ & && h_j(\underline{x}) = 0, \quad j = 1, 2, \dots, p \\ & && \underline{x} \in \mathcal{D}. \end{aligned}$$

The function f_0 is called the **objective function** while the functions f_1, \dots, f_m are called inequality constraints and the functions h_1, \dots, h_p are called equality constraints.

Definition 5.4.1. A vector $\underline{x} \in \mathcal{D}$ is **feasible** if it satisfies the constraints. Let $\mathcal{F} = \{\underline{x} \in \mathcal{D} \mid f_i(\underline{x}) \leq 0, i = 1, 2, \dots, m, h_j(\underline{x}) = 0, j = 1, \dots, p\}$ be the set of feasible vectors. Then, the problem is feasible if $\mathcal{F} \neq \emptyset$.

Definition 5.4.2. The **optimal value** is

$$p^* = \inf_{\underline{x} \in \mathcal{F}} f_0(\underline{x}).$$

By convention, p^* is allowed to take infinite values and $p^* = \infty$ if the problem is not feasible.

Evaluating the function at any feasible point automatically an upper bound because

$$p^* \leq f_0(\underline{x}) \quad \forall \underline{x} \in \mathcal{F}.$$

The optimization of a linear function with arbitrary affine equality and inequality constraints is called a **linear program**. Linear programs (LPs) have many equivalent forms and any linear program can be transformed into any standard form.

Definition 5.4.3. Two standard minimization forms of an LP are given by:

$$\begin{array}{ll} \text{minimize} & \underline{c}^T \underline{x} \\ \text{subject to} & A\underline{x} = \underline{b} \\ & \underline{x} \succeq \underline{0} \end{array} \qquad \begin{array}{ll} \text{minimize} & \underline{c}^T \underline{x} \\ \text{subject to} & A\underline{x} \succeq \underline{b}. \\ & \underline{x} \succeq \underline{0}. \end{array}$$

5.4.1 The Lagrangian

The Lagrangian is used to transform constrained optimization problems into unconstrained optimization problems. One can think of it as introducing a cost $\lambda_i \geq 0$ associated with violating the i -th inequality constraint and a variable ν_j used to enforce the j -th equality constraint.

Definition 5.4.4. The **Lagrangian** $L: \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ associated with optimization problem is

$$L(\underline{x}, \underline{\lambda}, \underline{\nu}) = f_0(\underline{x}) + \sum_{i=1}^m \lambda_i f_i(\underline{x}) + \sum_{j=1}^p \nu_j h_j(\underline{x}),$$

where λ_i is the **Lagrange multiplier** associated with the i -th inequality constraint and ν_j is the Lagrange multiplier associated with the j -th equality constraint.

Definition 5.4.5. A point \underline{x}^* is called **locally optimal** if there is an $\epsilon_0 > 0$ such that, for all $\epsilon < \epsilon_0$, it holds that $f_0(\underline{x}) \geq f_0(\underline{x}^*)$ for all $\underline{x} \in \mathcal{F}$ satisfying $\|\underline{x} - \underline{x}^*\| < \epsilon$. The i -th inequality constraint is **active** at \underline{x}^* if $f_i(\underline{x}^*) = 0$. Otherwise, it is **inactive**. Let $A = \{i \in \{1, \dots, m\} \mid f_i(\underline{x}^*) = 0\}$ be the set of active constraints at \underline{x}^* .

Definition 5.4.6 (Mangasarian-Fromovitz). A standard constrained optimization problem satisfies the MF constraint qualification at \underline{x}^* if the functions f_i and h_j are all continuously differentiable at \underline{x}^* and there exists a vector $\underline{w} \in \mathbb{R}^n$ satisfying $\nabla f_i(\underline{x}^*)^T \underline{w} < 0$ for $i \in A$ and $\nabla h_j(\underline{x}^*)^T \underline{w} = 0$ for $j = 1, \dots, p$. If the constraints h_j are not all affine, then one additionally needs that the vectors $\nabla h_1(\underline{x}^*), \dots, \nabla h_p(\underline{x}^*) \in \mathbb{R}^n$ form a linearly independent set.

Theorem 5.4.7 (Karush-Kuhn-Tucker). If \underline{x}^* is a constrained local optimum that satisfies the MF constraint qualification, then there exist $\underline{\lambda}^* \geq 0$ and $\underline{\nu}^*$ such that

$$\nabla f_0(\underline{x}^*) + \sum_{i \in A} \lambda_i^* \nabla f_i(\underline{x}^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(\underline{x}^*) = \underline{0}. \quad (5.3)$$

This theorem provides a necessary condition for a point \underline{x}^* to be locally optimal for a constrained optimization problem. Before considering its proof, it is useful to discuss the geometric picture underlying its contrapositive statement: if (5.3) does not hold for all $\underline{\lambda}^* \geq 0$ and $\underline{\nu}^* \in \mathbb{R}^p$, then \underline{x}^* is not locally optimal.

Now, consider what happens if we evaluate the function at $\underline{x}(t) = \underline{x}^* + t\underline{y}$ for some direction \underline{y} and a sufficiently small $t > 0$. For any continuously differentiable function f , the definition of the derivative implies that

$$f(\underline{x}(t)) = f(\underline{x}^*) + t \nabla f(\underline{x}^*)^T \underline{y} + o(t),$$

where $o(t) \rightarrow 0$ as $t \rightarrow 0$. If the problem is unconstrained (e.g., $m = p = 0$), then $\nabla f_0(\underline{x}^*)$ must be $\underline{0}$. This is because the negative gradient $-\nabla f_0(\underline{x}^*)$ gives the direction of steepest descent for the objective function and one is guaranteed to reduce the function by choosing $\underline{y} = -\nabla f_0(\underline{x}^*)$ (e.g., see Lemma 5.1.3). If there are constraints, however, then $\underline{x}(t)$ may be infeasible. For the j -th equality constraint, the definition of the derivative implies that, for sufficiently small t , $\underline{x}(t)$

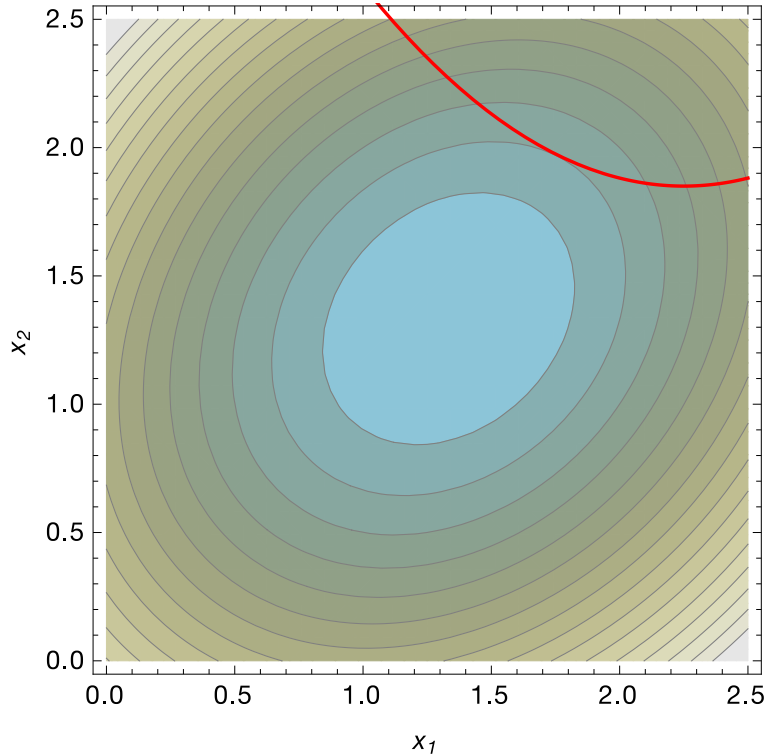


Figure 5.1: A contour plot of the function $f_0(x_1, x_2) = (x_1 - 1)^2 + (x_2 - 1)^2 - x_1 x_2 / 2$ whose minimum occurs at $(4/3, 4/3)$ (i.e., the center of the blue ellipse). The red line indicates the inequality constraint $f_1(x_1, x_2) = 1.85 + (x_1 - 2.25)^2 / 2 - x_2 \leq 0$. The picture shows that the constrained minimum occurs where the objective contour line is tangent to the active constraint line.

will be infeasible if $|\nabla h_j(\underline{x}^*)^T \underline{y}| > 0$. Thus, we certainly need $\nabla h_j(\underline{x}^*)^T \underline{y} = 0$ for all j .

If the i -th inequality constraint is active (i.e., $f_i(\underline{x}^*) = 0$), then the definition of the derivative implies that, for sufficiently small t , $\underline{x}(t)$ will be infeasible if $\nabla f_i(\underline{x}^*)^T \underline{y} > 0$. Thus, we certainly need $\nabla f_i(\underline{x}^*)^T \underline{y} \leq 0$ for all $i \in A$. If the constraint is inactive (i.e., $f_i(\underline{x}^*) < 0$), then due to continuity it will remain satisfied for sufficiently small t .

The geometric picture implied by Theorem 5.4.7 is that of a game where one would like to decrease the objective $f_0(\underline{x}^*)$ by choosing \underline{y} such that $\nabla f_0(\underline{x}^*)^T \underline{y} < 0$ but there are constraints on the set of allowable \underline{y} 's. Let $H = \text{span}(\{\nabla h_j(\underline{x}^*)\})$ be the subspace of directions that violate the equality constraints at \underline{x}^* . Similarly, let

the cone of directions that violate the active inequality constraints is given by

$$F = \left\{ \sum_{i \in A} \lambda_i \nabla f_i(\underline{x}^*) \mid \lambda_i \geq 0, i \in A \right\}.$$

Thus, one can only pick directions \underline{y} that are orthogonal to all vectors in H and also have a non-positive inner product with all vectors in F .

Let the matrix P define the orthogonal projection of \mathbb{R}^n onto H^\perp . Using this, we can translate the equation (5.3) into the statement

$$-P\nabla f_0(\underline{x}^*) \in PF$$

or “the projection, onto H^\perp , of the descent direction lies in the projection, onto H^\perp , of the cone of directions that violate the inequality constraints”. The reason for this is that we can absorb the ∇h_j terms into the ∇f_i terms by defining

$$\underline{f}^{(i)} = \nabla f_i(\underline{x}^*) + \sum_{j=1}^p \nu_{j,i} \nabla h_j(\underline{x}^*) = P\nabla f_i(\underline{x}^*)$$

so that $\underline{f}^{(i)} \in H^\perp$ for $i = 0, 1, \dots, m$. Then, the cone PF is defined by

$$PF = \left\{ \sum_{i \in A} \lambda_i \underline{f}^{(i)} \mid \lambda_i \geq 0, i \in A \right\}.$$

If $-P\nabla f_0(\underline{x}^*) \notin PF$, then we project $-P\nabla f_0(\underline{x}^*)$ onto PF to get a non-zero residual \underline{y} . The resulting vector gives a direction where the objective function decreases linearly in t and the constraint violations are $o(t)$. The challenge in making this proof precise is that, unless the equality constraints are affine, they may not be exactly satisfied for $t > 0$. In standard proofs of this result, this difficulty is overcome by using the implicit function theorem to construct an $\underline{x}(t)$ that starts in the direction of \underline{y} but is perturbed slightly to remain feasible.

Proof. For simplicity, we prove only the case where $h_j(\underline{x}) = \underline{a}_j^T \underline{x} - \underline{b}$ is affine. First, we define

$$\underline{y}(\underline{\lambda}, \underline{\nu}) = -\nabla f_0(\underline{x}^*) - \sum_{i=1}^m \lambda_i \nabla f_i(\underline{x}^*) - \sum_{j=1}^p \nu_j \underbrace{\nabla h_j(\underline{x}^*)}_{\underline{a}_j}.$$

The vector $\underline{y}(\underline{\lambda}, \underline{\nu})$ can be seen as the residual of the descent direction for the objective function after the constraint gradients have been used to cancel some parts.

Next, we let $\underline{\nu}^*(\underline{\lambda}) = \arg \min_{\underline{\nu} \in \mathbb{R}^p} \|\underline{y}(\underline{\lambda}, \underline{\nu})\|$ and apply the best approximation theorem (for the standard inner product space) to see that

$$\underline{y}(\underline{\lambda}, \underline{\nu}^*(\underline{\lambda})) = P\underline{y}(\underline{\lambda}, \underline{0}),$$

where the matrix P defines an orthogonal projection onto H^\perp and $H = \text{span}(\{\underline{a}_j\})$. This ensures that each $h_j(\underline{x}^* + t\underline{y}(\underline{\lambda}, \underline{\nu}^*(\underline{\lambda}))) = 0$ for all $\underline{\lambda} \in \mathbb{R}^m$ and $t \in \mathbb{R}$.

Continuing, we let $S = \{\underline{\lambda} \in \mathbb{R}^m \mid \underline{\lambda} \geq \underline{0}, \lambda_i = 0, i \notin A\}$ and compute

$$\underline{y}^* = \arg \min_{\underline{\lambda} \in S} \|\underline{y}(\underline{\lambda}, \underline{\nu}^*(\underline{\lambda}))\|.$$

This optimization uses the gradients of the active constraints to cancel as much of the residual descent direction as possible. Thus, $\underline{y}^* \neq \underline{0}$ implies there is a descent direction that does not violate the constraints. Looking at the formulas for $\underline{y}(\underline{\lambda}, \underline{\nu})$ and $\underline{y}(\underline{\lambda}, \underline{\nu}^*(\underline{\lambda}))$, we can also interpret \underline{y}^* as the error vector for the projection of $\underline{v} = -P\nabla f_0(\underline{x}^*)$ onto the convex set PF , which is the closed convex cone of perturbations that preserve the equality constraints but locally violate the inequality constraints. Then, $\underline{u}^* = \underline{v} - \underline{y}^*$ equals the projection itself and we observe that Theorem 4.6.7 implies $(\underline{u} - \underline{u}^*)^T(\underline{v} - \underline{u}^*) \leq 0$ for all $\underline{u} \in PF$. Since $\underline{0} \in PF$, we can choose $\underline{u} = \underline{0}$ to see that $(\underline{u}^*)^T(\underline{v} - \underline{u}^*) \geq 0$. Using this, we can write

$$\begin{aligned} -(P\nabla f_0(\underline{x}^*))^T \underline{y}^* &= \underline{v}^T(\underline{v} - \underline{u}^*) \\ &= (\underline{v} - \underline{u}^*)^T(\underline{v} - \underline{u}^*) + (\underline{u}^*)^T(\underline{v} - \underline{u}^*) \\ &\geq \|\underline{y}^*\|^2. \end{aligned}$$

If (5.3) cannot be satisfied by some $\underline{\lambda} \in S$ and $\underline{\nu} \in \mathbb{R}^p$, then $\underline{y}^* \neq \underline{0}$ and $\|\underline{y}^*\| > 0$. Thus, a perturbation in the \underline{y}^* direction will decrease the value of the objective function while essentially preserving feasibility.

But, the \underline{y}^* direction is only guaranteed to preserve feasibility to first order (i.e., $(P\nabla f_i(\underline{x}^*))^T \underline{y}^* \leq 0$) for $i = 1, \dots, m$. To fix this, first one needs to augment \underline{y}^* with a small amount of some vector \underline{w} satisfying $(P\nabla f_i(\underline{x}^*))^T \underline{w} < 0$ for all $i = 1, \dots, m$. Such a vector is guaranteed by the constraint qualification. Then, we can choose some $\delta > 0$ such that $(P\nabla f_0(\underline{x}^*))^T(\underline{y}^* + \delta \underline{w}) \leq -\frac{1}{2}\|\underline{y}^*\|$. With this modification, the definition of the derivative implies that, for sufficiently small t , $\underline{x}(t) = \underline{x}^* + t(\underline{y}^* + \delta \underline{w})$ will be a feasible vector satisfying $f_0(\underline{x}(t)) < f_0(\underline{x}^*)$. \square

5.4.2 Lagrangian Duality

Definition 5.4.8. The *Lagrangian dual function* is defined to be

$$g(\underline{\lambda}, \underline{\nu}) \triangleq \inf_{\underline{x} \in \mathcal{D}} L(\underline{x}, \underline{\lambda}, \underline{\nu}).$$

Lemma 5.4.9. The *Lagrangian dual problem*

$$\begin{aligned} & \text{maximize} && g(\underline{\lambda}, \underline{\nu}) \\ & \text{subject to} && \underline{\lambda} \succeq 0 \end{aligned}$$

has a unique maximum value $d^* \leq p^*$. This property is known as **weak duality**.

Proof. The Lagrangian dual function is concave because it is the pointwise infimum of affine functions

$$\begin{aligned} & g(\alpha \underline{\lambda} + (1 - \alpha) \underline{\lambda}', \alpha \underline{\nu} + (1 - \alpha) \underline{\nu}') \\ &= \inf_{\underline{x} \in \mathcal{D}} L(\underline{x}, \alpha \underline{\lambda} + (1 - \alpha) \underline{\lambda}', \alpha \underline{\nu} + (1 - \alpha) \underline{\nu}') \\ &= \inf_{\underline{x} \in \mathcal{D}} (\alpha L(\underline{x}, \underline{\lambda}, \underline{\nu}) + (1 - \alpha) L(\underline{x}, \underline{\lambda}', \underline{\nu}')) \\ &\geq \inf_{\underline{x} \in \mathcal{D}} \alpha L(\underline{x}, \underline{\lambda}, \underline{\nu}) + \inf_{\underline{x}' \in \mathcal{D}} (1 - \alpha) L(\underline{x}', \underline{\lambda}', \underline{\nu}') \\ &= \alpha g(\underline{\lambda}, \underline{\nu}) + (1 - \alpha) g(\underline{\lambda}', \underline{\nu}'). \end{aligned}$$

Thus, it follows from Theorem 5.3.5 that g has a unique maximum value d^* which can be upper bounded by

$$\begin{aligned} g(\underline{\lambda}, \underline{\nu}) &= \inf_{\underline{x} \in \mathcal{D}} L(\underline{x}, \underline{\lambda}, \underline{\nu}) \stackrel{(a)}{\leq} \inf_{\underline{x} \in \mathcal{F}} L(\underline{x}, \underline{\lambda}, \underline{\nu}) \\ &\stackrel{(b)}{=} p^* + \sum_{i=1}^m \lambda_i f_i(\underline{x}) \stackrel{(c)}{\leq} p^*, \end{aligned}$$

where (a) is implied by $\mathcal{F} \subseteq \mathcal{D}$, (b) follows from $h_j(\underline{x}) = 0$ for $\underline{x} \in \mathcal{F}$, and (c) holds by combining $f_i(\underline{x}) \leq 0$ for $\underline{x} \in \mathcal{F}$ and $\lambda_i \geq 0$. \square

The Lagrangian dual function can be $-\infty$ for a wide range of $(\underline{\lambda}, \underline{\nu})$. In this case, it makes sense to eliminate these points by defining the implicit constraint set

$$\mathcal{C} \triangleq \{(\underline{\lambda}, \underline{\nu}) \in \mathbb{R}^m \times \mathbb{R}^p \mid \underline{\lambda} \succeq 0, g(\underline{\lambda}, \underline{\nu}) > -\infty\}.$$

The points $(\underline{\lambda}, \underline{\nu}) \in \mathcal{C}$ are called **dual feasible** and it follows that

$$d^* = \sup_{(\underline{\lambda}, \underline{\nu}) \in \mathcal{C}} g(\underline{\lambda}, \underline{\nu}).$$

By convention, $d^* = -\infty$ if the dual problem is not feasible (i.e., $\mathcal{C} = \emptyset$).

Definition 5.4.10. *If $d^* = p^*$, then one says **strong duality** holds for the problem.*

Theorem 5.4.11. *Let \underline{x}^* be a primal optimal point and $(\underline{\lambda}^*, \underline{\nu}^*)$ be a dual optimal point. If strong duality holds, $\underline{x}^* \in \mathcal{D}^\circ$, and all f_i and h_j functions are differentiable at \underline{x}^* , then we get the KKT conditions of complementary slackness, $\lambda_i^* f_i(\underline{x}^*) = 0$ for $i = 1, \dots, m$, and stationarity (5.3).*

Proof. By weak duality, we have

$$d^* = g(\underline{\lambda}^*, \underline{\nu}^*) \leq f_0(\underline{x}^*) = p^*.$$

Since \underline{x}^* is feasible, combining $d^* = p^*$ with the proof of weak duality shows that

$$\inf_{\underline{x} \in \mathcal{D}} L(\underline{x}, \underline{\lambda}^*, \underline{\nu}^*) = L(\underline{x}^*, \underline{\lambda}^*, \underline{\nu}^*)$$

and $\lambda_i^* = 0$ if the i -th inequality constraint is inactive (i.e., $f_i(\underline{x}^*) < 0$). Thus, we also observe that complementary slackness condition $\lambda_i^* f_i(\underline{x}^*) = 0$ holds for $i = 1, \dots, m$. Since $\underline{x}^* \in \mathcal{D}^\circ$, it follows that \underline{x}^* is a locally optimal point of $L(\underline{x}, \underline{\lambda}^*, \underline{\nu}^*)$. Thus, \underline{x}^* must be a stationary point of $L(\underline{x}, \underline{\lambda}^*, \underline{\nu}^*)$ and taking the \underline{x} -derivative gives (5.3). \square

Example 5.4.12. *For the first LP in Definition 5.4.3, the Lagrangian is given by*

$$L(\underline{x}, \underline{\lambda}, \underline{\nu}) = \underline{c}^T \underline{x} + \underline{\nu}^T (\underline{b} - A\underline{x}) - \underline{\lambda}^T \underline{x},$$

where the $\underline{\lambda}$ term is negative because the constraint is $\underline{x} \succeq \underline{0}$. Thus, the Lagrangian dual function is given by

$$g(\underline{\lambda}, \underline{\nu}) = \inf_{\underline{x} \in \mathcal{D}} L(\underline{x}, \underline{\lambda}, \underline{\nu}) = \begin{cases} \underline{b}^T \underline{\nu} & \text{if } \underline{c} - A^T \underline{\nu} - \underline{\lambda} = \underline{0} \\ -\infty & \text{otherwise.} \end{cases}$$

Solving the implicit constraint and using the fact that $\underline{\lambda} \succeq \underline{0}$, one gets the dual LP problem

$$\begin{aligned} &\text{maximize} && \underline{b}^T \underline{\nu} \\ &\text{subject to} && A^T \underline{\nu} \preceq \underline{c}. \end{aligned}$$

Strong duality for linear programs says that, if the original LP has an optimal solution (i.e., it is neither unbounded nor infeasible), then the dual LP has an optimal solution of the same value.

5.4.3 Convex Optimization

Definition 5.4.13. *An optimization problem in standard form is called **convex** if all f_i functions are convex, all the h_j functions are affine (i.e., $h_j(\underline{x}) = \underline{a}_j^T \underline{x} - b_j$), and $\mathcal{D} = \mathbb{R}^n$.*

Problem 5.4.14. *For a convex standard-form optimization problem (i.e., satisfying Definition 5.4.13), show that the feasible set is a convex set.*

Applying Theorem 5.3.5 to this setup shows that a convex standard-form optimization problem has a unique minimum value. Also, if the function f_0 is strictly convex, then the minimum value is achieved uniquely. There are a number of stronger conditions that also imply strong duality for convex optimization problems. **Slater's condition** is stated below as a theorem and its proof can be found in [?, Sec. 5.3.2].

Theorem 5.4.15 (Slater's Condition). *If a convex optimization problem has a point \underline{x}_0 where $f_i(\underline{x}_0) < 0$ for $i = 1, \dots, m$ and $h_j(\underline{x}_0) = 0$ for $j = 1, \dots, p$, then the MF constraint qualification and strong duality both hold for the problem. In addition, if all f_i functions are differentiable, then the KKT conditions are necessary and sufficient for optimality.*

Example 5.4.16. *For a channel with colored noise, the input distribution that maximizes the achievable information rate can be found by solving the convex optimization problem, known as water-filling, given by*

$$\begin{aligned} & \text{minimize} && - \sum_{i=1}^n \log(x_i + \alpha_i) \\ & \text{subject to} && \sum_{i=1}^n x_i = P \\ & && \underline{x} \succeq 0. \end{aligned}$$

Choosing $x_i = \frac{P}{n}$ for $i = 1, \dots, n$ gives a point that satisfies Slater's condition, so strong duality holds for this problem.

Example 5.4.17. For the water-filling problem, the Lagrangian can be written as

$$L(\underline{x}, \underline{\lambda}, \nu) = - \sum_{i=1}^n \log(x_i + \alpha_i) - \sum_{i=1}^m \lambda_i x_i + \nu \left(-P + \sum_{i=1}^n x_i \right)$$

and the Lagrangian dual is given by $g(\underline{\lambda}, \nu) = \inf_{\underline{x} \in \mathbb{R}^n} L(\underline{x}, \underline{\lambda}, \nu)$.

If $\lambda_i < 0$, then the Lagrangian tends to $-\infty$ as $x_i \rightarrow -\infty$. Thus, the system is implicitly constrained to have $\lambda_i \geq 0$. The first-order optimality conditions, for $i = 1, 2, \dots, n$, are given by

$$-\frac{1}{x_i + \alpha_i} - \lambda_i + \nu = 0.$$

Solving this for x_i shows that x_i is increasing in λ_i (for $\lambda_i \geq 0$) and this implies that $g(\underline{\lambda}, \nu)$ is decreasing in λ_i (for $\lambda_i \geq 0$ and $x_i \geq 0$).

Thus, the expression $\max_{\lambda \geq 0} g(\underline{\lambda}, \nu)$ is given by choosing the smallest non-negative λ_i 's for which $x_i \geq 0$. This implies that

$$(x_i, \lambda_i) = \begin{cases} \left(\frac{1}{\nu} - \alpha_i, 0 \right) & \text{if } \nu < \frac{1}{\alpha_i} \\ \left(0, \nu - \frac{1}{\alpha_i} \right) & \text{if } \nu \geq \frac{1}{\alpha_i}. \end{cases}$$

From this, the value of ν can be determined by solving

$$\sum_{i=1}^n x_i = \sum_{i=1}^n \max \left\{ 0, \frac{1}{\nu} - \alpha_i \right\} = P.$$

By strong duality, the optimal value of the dual problem equals the optimal value of the original problem. Finally, the problem can be easily solved for a range of P values by sweeping through a range of ν values and computing P in terms of ν .

Chapter 6

Linear Transformations and Operators

6.1 The Algebra of Linear Transformations

Theorem 6.1.1. *Let V and W be vector spaces over the field F . Let T and U be two linear transformations from V into W . The function $(T + U)$ defined pointwise by*

$$(T + U)(\underline{v}) = T\underline{v} + U\underline{v}$$

is a linear transformation from V into W . Furthermore, if $s \in F$, the function (sT) defined by

$$(sT)(\underline{v}) = s(T\underline{v})$$

is also a linear transformation from V into W . The set of all linear transformation from V into W , together with the addition and scalar multiplication defined above, is a vector space over the field F .

Proof. Suppose that T and U are linear transformation from V into W . For $(T + U)$ defined above, we have

$$\begin{aligned}(T + U)(s\underline{v} + \underline{w}) &= T(s\underline{v} + \underline{w}) + U(s\underline{v} + \underline{w}) \\ &= s(T\underline{v}) + T\underline{w} + s(U\underline{v}) + U\underline{w} \\ &= s(T\underline{v} + U\underline{v}) + (T\underline{w} + U\underline{w}) \\ &= s(T + U)\underline{v} + (T + U)\underline{w},\end{aligned}$$

which shows that $(T + U)$ is a linear transformation. Similarly, we have

$$\begin{aligned}
 (rT)(s\underline{v} + \underline{w}) &= r(T(s\underline{v} + \underline{w})) \\
 &= r(s(T\underline{v}) + (T\underline{w})) \\
 &= rs(T\underline{v}) + r(T\underline{w}) \\
 &= s(r(T\underline{v})) + rT(\underline{w}) \\
 &= s((rT)\underline{v}) + (rT)\underline{w}
 \end{aligned}$$

which shows that (rT) is a linear transformation.

To verify that the set of linear transformations from V into W together with the operations defined above is a vector space, one must directly check the conditions of Definition 3.3.1. These are straightforward to verify, and we leave this exercise to the reader. \square

We denote the space of linear transformations from V into W by $L(V, W)$. Note that $L(V, W)$ is defined only when V and W are vector spaces over the same field.

Fact 6.1.2. *Let V be an n -dimensional vector space over the field F , and let W be an m -dimensional vector space over F . Then the space $L(V, W)$ is finite-dimensional and has dimension mn .*

Theorem 6.1.3. *Let V , W , and Z be vector spaces over a field F . Let $T \in L(V, W)$ and $U \in L(W, Z)$. Then the composed function UT defined by $(UT)(\underline{v}) = U(T(\underline{v}))$ is a linear transformation from V into Z .*

Proof. Let $\underline{v}_1, \underline{v}_2 \in V$ and $s \in F$. Then, we have

$$\begin{aligned}
 (UT)(s\underline{v}_1 + \underline{v}_2) &= U(T(s\underline{v}_1 + \underline{v}_2)) \\
 &= U(sT\underline{v}_1 + T\underline{v}_2) \\
 &= sU(T\underline{v}_1) + U(T\underline{v}_2) \\
 &= s(UT)(\underline{v}_1) + (UT)(\underline{v}_2),
 \end{aligned}$$

as desired. \square

Definition 6.1.4. *If V is a vector space over the field F , a **linear operator** on V is a linear transformation from V into V .*

Definition 6.1.5. An **algebra** over a field F is a vector space V over F that has a bilinear vector product “ \cdot ”: $V \times V \rightarrow V$ satisfying $(s\underline{u}) \cdot (t\underline{v}) = (st)(\underline{u} \cdot \underline{v})$ and

$$(s\underline{u} + \underline{v}) \cdot (t\underline{w} + \underline{x}) = st(\underline{u} \cdot \underline{w}) + s(\underline{u} \cdot \underline{x}) + t(\underline{v} \cdot \underline{w}) + (\underline{v} \cdot \underline{x}),$$

for all $s, t \in F$ and $\underline{u}, \underline{v}, \underline{w}, \underline{x} \in V$. If V is a Banach space and the norm of the vector product satisfies $\|\underline{u} \cdot \underline{v}\| \leq \|\underline{u}\| \|\underline{v}\|$, then it is called a **Banach algebra**.

Example 6.1.6. The set $L(V, V)$ of linear operators on V forms an algebra when the vector product is defined by functional composition $UT(\underline{v}) = U(T(\underline{v}))$. If V is a Banach space and $L(V, V)$ is equipped with the induced operator norm, then it forms a Banach algebra.

Definition 6.1.7. A linear transformation T from V into W is called **invertible** if there exists a function U from W to V such that UT is the identity function on V and TU is the identity function on W . If T is invertible, the function U is unique and is denoted by T^{-1} . Furthermore, T is invertible if and only if

1. T is one-to-one: $T\underline{v}_1 = T\underline{v}_2 \implies \underline{v}_1 = \underline{v}_2$
2. T is onto: the range of T is W .

Example 6.1.8. Consider the vector space V of semi-infinite real sequences \mathbb{R}^ω where $\underline{v} = (v_1, v_2, v_3, \dots) \in V$ with $v_n \in \mathbb{R}$ for $n \in \mathbb{N}$. Let $L: V \rightarrow V$ be the left-shift linear transformation defined by

$$L\underline{v} = (v_2, v_3, v_4, \dots)$$

and $R: V \rightarrow V$ be the right-shift linear transformation defined by

$$R\underline{v} = (0, v_1, v_2, \dots).$$

Notice that L is onto but not one-to-one and R is one-to-one but not onto. Therefore, neither transformation is invertible.

Example 6.1.9. Consider the normed vector space V of semi-infinite real sequences \mathbb{R}^ω with the standard Schauder basis $\{\underline{e}_1, \underline{e}_2, \dots\}$. Let $T: V \rightarrow V$ be the linear transformation that satisfies $T\underline{e}_i = i^{-1}\underline{e}_i$ for $i = 1, 2, \dots$. Let the linear transformation $U: V \rightarrow V$ satisfy $U\underline{e}_i = i\underline{e}_i$ for $i = 1, 2, \dots$. It is easy to verify that $U = T^{-1}$ and $UT = TU = I$.

This example should actually bother you somewhat. Since T reduces vector components arbitrarily, its inverse must enlarge them arbitrarily. Clearly, this is not a desirable property. Later, we will introduce a norm for linear transforms which quantifies this problem.

Theorem 6.1.10. *Let V and W be vector spaces over the field F and let T be a linear transformation from V into W . If T is invertible, then the inverse function T^{-1} is a linear transformation from W onto V .*

Proof. Let \underline{w}_1 and \underline{w}_2 be vectors in W and let $s \in F$. Define $\underline{v}_j = T^{-1}\underline{w}_j$, for $j = 1, 2$. Since T is a linear transformation, we have

$$T(s\underline{v}_1 + \underline{v}_2) = sT(\underline{v}_1) + T(\underline{v}_2) = s\underline{w}_1 + \underline{w}_2.$$

That is, $s\underline{v}_1 + \underline{v}_2$ is the unique vector in V that maps to $s\underline{w}_1 + \underline{w}_2$ under T . It follows that

$$T^{-1}(s\underline{w}_1 + \underline{w}_2) = s\underline{v}_1 + \underline{v}_2 = s(T^{-1}\underline{w}_1) + T^{-1}\underline{w}_2$$

and T^{-1} is a linear transformation. □

A **homomorphism** is a mapping between algebraic structures which preserves all relevant structure. An **isomorphism** is a homomorphism which is also invertible. For vector spaces, the relevant structure is given by vector addition and scalar multiplication. Since a linear transformation preserves both of these operations, it is also a *vector space homomorphism*. Likewise, an invertible linear transformation is a *vector space isomorphism*.

6.2 The Dual Space

Definition 6.2.1. *Let V be a vector space. The collection of all linear functionals on V , denoted $L(V, F)$, forms a vector space. We also denote this space by V^* and call it the **dual space** of V .*

The following theorem shows that, if V is finite dimensional, then

$$\dim V^* = \dim V.$$

In this case, one actually finds that V is isomorphic to V^* . Therefore, the two spaces can be identified with each other so that $V = V^*$ for finite dimensional V .

Theorem 6.2.2. *Let V be a finite-dimensional vector space over the field F , and let $\mathcal{B} = \underline{v}_1, \dots, \underline{v}_n$ be a basis for V . There is a unique dual basis $\mathcal{B}^* = f_1, \dots, f_n$ for V^* such that $f_j(\underline{v}_i) = \delta_{ij}$. For each linear functional on V , we have*

$$f = \sum_{i=1}^n f(\underline{v}_i) f_i$$

and for each vector \underline{v} in V , we have

$$\underline{v} = \sum_{i=1}^n f_i(\underline{v}) \underline{v}_i.$$

Proof. Let $\mathcal{B} = \underline{v}_1, \dots, \underline{v}_n$ be a basis for V . According to Theorem 3.4.7, there is a unique linear functional f_i on V such that

$$f_i(\underline{v}_j) = \delta_{ij}.$$

Thus, we obtain from \mathcal{B} a set of n distinct linear functionals f_1, \dots, f_n on V . These functionals are linearly independent; suppose that

$$f = \sum_{i=1}^n s_i f_i,$$

then

$$f(\underline{v}_j) = \sum_{i=1}^n s_i f_i(\underline{v}_j) = \sum_{i=1}^n s_i \delta_{ij} = s_j.$$

In particular, if f is the zero functional, $f(\underline{v}_j) = 0$ for $j = 1, \dots, n$ and hence the scalars $\{s_j\}$ must all equal 0. It follows that the functionals f_1, \dots, f_n are linearly independent. Since $\dim V^* = n$, we conclude that $\mathcal{B}^* = f_1, \dots, f_n$ forms a basis for V^* , the **dual basis** of \mathcal{B} .

Next, we want to show that there is a unique basis which is dual to \mathcal{B} . If f is a linear functional on V , then f is some linear combination of f_1, \dots, f_n with

$$f = \sum_{i=1}^n s_i f_i.$$

Furthermore, by construction, we must have $s_j = f(\underline{v}_j)$ for $j = 1, \dots, n$. Similarly, if

$$\underline{v} = \sum_{i=1}^n t_i \underline{v}_i.$$

is a vector in V , then

$$f_j(\underline{v}) = \sum_{i=1}^n t_i f_j(\underline{v}_i) = \sum_{i=1}^n t_i \delta_{ij} = t_j.$$

That is, the unique expression for \underline{v} as a linear combination of $\underline{v}_1, \dots, \underline{v}_n$ is

$$\underline{v} = \sum_{i=1}^n f_i(\underline{v}) \underline{v}_i.$$

□

One important use of the dual space is to define the transpose of a linear transform in a way that generalizes to infinite dimensional vector spaces. Let V, W be vector spaces over F and $T: V \rightarrow W$ be a linear transform. If $g \in W^*$ is a linear functional on W (i.e., $g: W \rightarrow F$), then $g(T\underline{v}) \in V^*$ is a linear functional on V . The **transpose** of T is the mapping $U: W^* \rightarrow V^*$ defined by $f(\underline{v}) = g(T\underline{v}) \in V^*$ for all $g \in W^*$. If V, W are finite-dimensional, then one can identify $V = V^*$ and $W = W^*$ via isomorphism and recover the standard transpose mapping $U: W \rightarrow V$ implied by the matrix transpose.

The details of this definition are not used in the remainder of these notes, but can be useful in understanding the subtleties of infinite dimensional spaces. For infinite dimensional Hilbert spaces, we will see later that the definition again simplifies because one identifies $V = V^*$ via isomorphism. The interesting case that does not simplify is that of linear transforms between infinite dimensional Banach spaces.

6.3 Operator Norms

For any vector space of linear transforms, one can define a norm to get a normed vector space of linear transforms (e.g., consider the Frobenius norm of a matrix). In contrast, an operator norm is defined for linear transforms between normed spaces and it is induced by the vector norms of the underlying spaces. Intuitively, the induced operator norm is the largest factor by which a linear transform can increase the length of a vector. This defines a simple “worst-case” expansion for any linear transform.

Definition 6.3.1. Let V and W be two normed vector spaces and let $T: V \rightarrow W$ be a linear transformation. The induced **operator norm** of T is defined to

$$\|T\| = \sup_{\underline{v} \in V - \{0\}} \frac{\|T\underline{v}\|}{\|\underline{v}\|} = \sup_{\underline{v} \in V, \|\underline{v}\|=1} \|T\underline{v}\|.$$

A common question about the operator norm is, “How do I know the two expressions give the same result?”. To see this, we can write

$$\sup_{\underline{v} \in V - \{0\}} \frac{\|T\underline{v}\|}{\|\underline{v}\|} = \sup_{\underline{v} \in V - \{0\}} \left\| T \frac{\underline{v}}{\|\underline{v}\|} \right\| = \sup_{\underline{u} \in V, \|\underline{u}\|=1} \|T\underline{u}\|.$$

Previously, we have seen that the set $L(V, W)$ of linear transformations from V into W , with the standard addition and scalar multiplication, satisfies the conditions required to be a vector space. Now, we have a norm for that vector space. Interested readers should verify that the above definition satisfies the first two standard conditions required by a norm. To verify the triangle inequality, we can write

$$\begin{aligned} \|T + U\| &= \sup_{\underline{v} \in V, \|\underline{v}\|=1} \|(T + U)\underline{v}\| \\ &\leq \sup_{\underline{v} \in V, \|\underline{v}\|=1} (\|T\underline{v}\| + \|U\underline{v}\|) \\ &\leq \sup_{\underline{v} \in V, \|\underline{v}\|=1} \|T\underline{v}\| + \sup_{\underline{v} \in V, \|\underline{v}\|=1} \|U\underline{v}\| \\ &= \|T\| + \|U\|. \end{aligned}$$

The induced operator norm also has another property that follows naturally from its definition. Notice that

$$\|T\| = \sup_{\underline{v} \in V - \{0\}} \frac{\|T\underline{v}\|}{\|\underline{v}\|} \geq \frac{\|T\underline{u}\|}{\|\underline{u}\|}$$

for all non-zero $\underline{u} \in V$. Checking the special case of $\underline{u} = \underline{0}$ separately, one can show the **induced operator-norm inequality** $\|T\underline{u}\| \leq \|T\| \|\underline{u}\|$ for all $\underline{u} \in V$.

For the space $L(V, V)$ of linear operators on V , a norm is called **submultiplicative** if $\|TU\| \leq \|T\| \|U\|$ for all $T, U \in L(V, V)$. The induced operator-norm inequality shows that all induced operator norms are submultiplicative because

$$\|UT\underline{v}\| \leq \|U\| \|T\underline{v}\| \leq \|U\| \|T\| \|\underline{v}\|.$$

This also defines a submultiplicative norm for the algebra of linear operators on V .

6.3.1 Bounded Transformations

Definition 6.3.2. A linear transformation is called **bounded** if its induced operator norm is finite.

Theorem 6.3.3. A linear transformation $T: V \rightarrow W$ is bounded if and only if it is continuous.

Proof. Suppose that T is bounded; that is, there exists M such that $\|T\underline{v}\| \leq M \|\underline{v}\|$ for all $\underline{v} \in V$. Let $\underline{v}_1, \underline{v}_2, \dots$ be a convergent sequence in V , then

$$\|T\underline{v}_i - T\underline{v}_j\| = \|T(\underline{v}_i - \underline{v}_j)\| \leq M \|\underline{v}_i - \underline{v}_j\|.$$

This implies that $T\underline{v}_1, T\underline{v}_2, \dots$ is a convergent sequence in W , and T is continuous.

Conversely, assume T is continuous and notice that $T\underline{0} = \underline{0}$. Therefore, for any $\epsilon > 0$, there is a $\delta > 0$ such that $\|T\underline{v}\| < \epsilon$ for all $\|\underline{v}\| < \delta$. Since the norm of $\underline{u} = \frac{\delta\underline{v}}{2\|\underline{v}\|}$ is equal to $\delta/2$, we get

$$\|T\underline{v}\| = \left\| T \frac{\delta\underline{v}}{2\|\underline{v}\|} \right\| \frac{2\|\underline{v}\|}{\delta} < \frac{2\epsilon}{\delta} \|\underline{v}\|.$$

The value $M = \frac{2\epsilon}{\delta}$ serves as an upper bound on $\|T\|$. □

Then, by showing that linear transformations over finite-dimensional spaces are continuous, one concludes that they are also bounded. This is accomplished in the following theorem.

Theorem 6.3.4. Let V and W be normed vector spaces and let $T: V \rightarrow W$ be a linear transformation. If V is finite dimensional, then T is continuous and bounded.

Proof. Let $\mathcal{B} = \underline{v}_1, \dots, \underline{v}_n$ be a basis for V . Let $\underline{v} \in V$ be expressed in terms of this basis as

$$\underline{v} = s_1\underline{v}_1 + \dots + s_n\underline{v}_n.$$

Let $C = \max_{1 \leq i \leq n} \|T\underline{v}_i\|_W$. Then,

$$\begin{aligned} \|T\underline{v}\|_W &= \|T(s_1\underline{v}_1 + \dots + s_n\underline{v}_n)\|_W \\ &\leq |s_1| \|T\underline{v}_1\|_W + \dots + |s_n| \|T\underline{v}_n\|_W \\ &\leq C(|s_1| + \dots + |s_n|). \end{aligned}$$

By Lemma 3.5.19, there is a real m such that $|s_1| + \dots + |s_n| \leq \frac{1}{m} \|\underline{v}\|_V$ and

$$\|T\underline{v}\|_W \leq \frac{C}{m} \|\underline{v}\|_V. \quad \square$$

6.3.2 The Neumann Expansion

Theorem 6.3.5. *Let $\|\cdot\|$ be a submultiplicative operator norm and $T: V \rightarrow V$ be a linear operator with $\|T\| < 1$. Then, $(I - T)^{-1}$ exists and*

$$(I - T)^{-1} = \sum_{i=0}^{\infty} T^i.$$

Proof. First, we observe that the sequence

$$A_n = \sum_{i=0}^{n-1} T^i.$$

is Cauchy. This follows from the fact that, for $m < n$, we have

$$\|A_n - A_m\| = \left\| \sum_{i=m}^{n-1} T^i \right\| \leq \sum_{i=m}^{n-1} \|T\|^i = \frac{\|T\|^m - \|T\|^n}{1 - \|T\|} \leq \frac{\|T\|^m}{1 - \|T\|}.$$

Since this goes to zero as $m \rightarrow \infty$, we see that the limit $\lim_{n \rightarrow \infty} A_n$ exists.

Next, we observe that

$$(I - T)(I + T + T^2 + \cdots + T^{n-1}) = I - T^n.$$

Since $\|T\| < 1$, we have $\lim_{k \rightarrow \infty} T^k = 0$ because $\|T^k\| \leq \|T\|^k \rightarrow 0$. Taking the limit $n \rightarrow \infty$ of both sides gives

$$(I - T) \sum_{i=0}^{\infty} T^i = \lim_{n \rightarrow \infty} (I - T^n) = I.$$

Likewise, reversing the order multiplication results in the same result. This shows that $\sum_{i=0}^{\infty} T^i$ must be the inverse of $I - T$. \square

If one only needs to show that $I - T$ is non-singular, then proof by contradiction is somewhat simpler. Suppose $I - T$ is singular, then there exists a non-zero vector \underline{v} such that $(I - T)\underline{v} = \underline{0}$. But, this implies that $\|\underline{v}\| = \|T\underline{v}\| \leq \|T\| \|\underline{v}\|$. Since $\|\underline{v}\| \neq 0$, this gives the contradiction $\|T\| \geq 1$ and implies that $I - T$ is non-singular.

6.3.3 Matrix Norms

$$\|A\|_{\infty} = \max_{\|\underline{v}\|_{\infty}=1} \|A\underline{v}\|_{\infty} = \max_i \sum_j |a_{ij}|$$

$$\|A\|_1 = \max_{\|\underline{v}\|_1=1} \|A\underline{v}\|_1 = \max_j \sum_i |a_{ij}|$$

The 2-norm of a matrix can be found by solving

$$\max_{\underline{v}^H \underline{v}=1} \|A\underline{v}\|_2^2 = \underline{v}^H A^H A \underline{v}.$$

Using the Lagrange multiplier technique, one seeks to minimize

$$J = \underline{v}^H A^H A \underline{v} - \lambda \underline{v}^H \underline{v}.$$

Taking the gradient with respect to \underline{v} and equating the result to zero, we get

$$A^H A \underline{v} = \lambda \underline{v}.$$

The corresponding \underline{v} must be an eigenvector of the matrix $A^H A$. Left multiplying this equation by \underline{v}^H and using the fact that $\underline{v}^H \underline{v} = 1$, we obtain

$$\underline{v}^H A^H A \underline{v} = \lambda \underline{v}^H \underline{v} = \lambda.$$

Since we are maximizing the left hand side of this equation, λ must be the largest eigenvalue of $A^H A$. For an $n \times n$ matrix B with eigenvalues $\lambda_1, \dots, \lambda_n$, the **spectral radius** $\rho(B)$ is defined by

$$\rho(B) = \max_i |\lambda_i|.$$

The spectral radius of B is the smallest radius of a circle centered at the origin that contains all the eigenvalues of B . It follows that

$$\|A\|_2 = \sqrt{\rho(A^H A)}.$$

When A is Hermitian, $\|A\|_2 = \rho(A)$. The 2-norm is also called the **spectral norm**.

The **Frobenius norm** is given by

$$\|A\|_F = \left(\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2 \right)^{\frac{1}{2}}.$$

This norm is also called the **Euclidean norm**. Note that $\|A\|_F^2 = \text{tr}(A^H A)$.

6.4 Linear Functionals on Hilbert Spaces

Let V be an inner-product space, and let \underline{v} be some fixed vector in V . Define the function $f_{\underline{v}}$ from V into F by

$$f_{\underline{v}}(\underline{w}) = \langle \underline{w}, \underline{v} \rangle.$$

Clearly, $f_{\underline{v}}$ is a linear functional on V . If V is a Hilbert space, then every continuous linear functional on V arises in this way from some vector \underline{v} . This result is known as the **Riesz representation theorem**.

Lemma 6.4.1. *If $\langle \underline{v}, \underline{w} \rangle = \langle \underline{u}, \underline{w} \rangle$ for all $\underline{w} \in V$, then $\underline{v} = \underline{u}$.*

Proof. Then, $\langle \underline{v} - \underline{u}, \underline{w} \rangle = 0$ for all $\underline{w} \in V$. Therefore, $\langle \underline{v} - \underline{u}, \underline{v} - \underline{u} \rangle = 0$ and this implies $\underline{v} - \underline{u} = \underline{0}$. \square

Theorem 6.4.2 (Riesz). *Let V be a Hilbert space and f be a continuous linear functional on V . Then, there exists a unique vector $\underline{v} \in V$ such that $f(\underline{w}) = \langle \underline{w}, \underline{v} \rangle$ for all $\underline{w} \in V$.*

Proof. While the result holds in any Hilbert space, this proof assumes V is separable for simplicity. Therefore, we let $\underline{v}_1, \underline{v}_2, \dots$ be a countable orthonormal basis for V . We wish to find a candidate vector \underline{v} for the inner product.

First, we note that f is bounded because it is continuous and, as such, there exists M such that $|f(\underline{x})| \leq M\|\underline{x}\|$ for all $\underline{x} \in V$. Let $\underline{x}_n = \sum_{i=1}^n \overline{f(\underline{v}_i)} \underline{v}_i$. For any n , we have

$$\begin{aligned} M \|\underline{x}_n\| &\geq |f(\underline{x}_n)| = \left| \sum_{i=1}^n \overline{f(\underline{v}_i)} f(\underline{v}_i) \right| = \sum_{i=1}^n |f(\underline{v}_i)|^2 = \sum_{i=1}^n f(\underline{v}_i) \overline{f(\underline{v}_i)} \\ &= \sum_{i=1}^n \langle \overline{f(\underline{v}_i)} \underline{v}_i, \overline{f(\underline{v}_i)} \underline{v}_i \rangle = \sum_{i=1}^n \sum_{j=1}^n \langle \overline{f(\underline{v}_j)} \underline{v}_j, \overline{f(\underline{v}_i)} \underline{v}_i \rangle \\ &= \left\langle \sum_{j=1}^n \overline{f(\underline{v}_j)} \underline{v}_j, \sum_{i=1}^n \overline{f(\underline{v}_i)} \underline{v}_i \right\rangle = \langle \underline{x}_n, \underline{x}_n \rangle = \|\underline{x}_n\|^2. \end{aligned}$$

This implies that $\|\underline{x}_n\| \leq M$ for all n . Hence, $\lim_{n \rightarrow \infty} \sum_{i=1}^n |f(\underline{v}_i)|^2$ is bounded and the vector

$$\underline{v} = \sum_{i=1}^{\infty} \overline{f(\underline{v}_i)} \underline{v}_i,$$

is in V because it is the limit point of a Cauchy sequence. Let $f_{\underline{v}}$ be the functional defined by

$$f_{\underline{v}}(\underline{w}) = \langle \underline{w}, \underline{v} \rangle.$$

By the Cauchy-Schwarz, we can verify that

$$\|f_{\underline{v}}\| \triangleq \sup_{\underline{u} \in V - \{0\}} \frac{f_{\underline{v}}(\underline{u})}{\|\underline{u}\|} = \|\underline{v}\|.$$

Since f is continuous, it follows that $\|f\| < \infty$ and $\|\underline{v}\| < \infty$. Then,

$$f_{\underline{v}}(\underline{v}_j) = \left\langle \underline{v}_j, \sum_{i=1}^{\infty} \overline{f(\underline{v}_i)} \underline{v}_i \right\rangle = f(\underline{v}_j).$$

Since this is true for each \underline{v}_j , it follows that $f = f_{\underline{v}}$. Now, consider any $\underline{v}' \in V$ such that $\langle \underline{w}, \underline{v} \rangle = \langle \underline{w}, \underline{v}' \rangle$ for all $\underline{w} \in W$. Applying Lemma 6.4.1 shows that $\underline{v} = \underline{v}'$ and we conclude that \underline{v} is unique. \square

An important consequence of this theorem is that the continuous dual space V^* of a Hilbert space V is isometrically isomorphic to the original space V . Let $R: V^* \rightarrow V$ be the implied Riesz mapping from continuous linear functionals on V (i.e., V^*) to elements of V . Then, $f(\underline{v}) = \langle \underline{v}, R(f) \rangle$ for all $f \in V^*$. The isomorphism can be shown by verifying that $R(sf_1 + f_2) = \bar{s}R(f_1) + R(f_2)$ and one finds that the mapping R is conjugate linear. The mapping is isometric because $\|f\| = \|R(f)\|$. Based on this isomorphism, one can treat a Hilbert space as self-dual and assume without confusion that $V = V^*$.

Theorem 6.4.3. *Let V and W be Hilbert spaces, and assume $T: V \rightarrow W$ is a continuous linear transformation. Then, the **adjoint** is the unique linear transformation T^* on W such that*

$$\langle T\underline{v}, \underline{w} \rangle = \langle \underline{v}, T^*\underline{w} \rangle$$

for all vectors $\underline{v} \in V, \underline{w} \in W$.

Proof. Let \underline{w} be any vector in W . Then $f(\underline{v}) = \langle T\underline{v}, \underline{w} \rangle$ is a continuous linear functional on V . It follows from the Riesz representation theorem (Theorem 6.4.2) that there exists a unique vector $\underline{v}' \in V$ such that $f(\underline{v}) = \langle T\underline{v}, \underline{w} \rangle = \langle \underline{v}, \underline{v}' \rangle$. Of course, the vector \underline{v}' depends on the choice of \underline{w} . So, we define the adjoint mapping $T^*: W \rightarrow V$ to give the required \underline{v}' for each \underline{w} . In other words,

$$\underline{v}' = T^*\underline{w}.$$

Next, we must verify that T^* is a linear transformation. Let $\underline{w}_1, \underline{w}_2$ be in W and s be a scalar. For all $\underline{v} \in V$,

$$\begin{aligned} \langle \underline{v}, T^*(s\underline{w}_1 + \underline{w}_2) \rangle &= \langle T\underline{v}, (s\underline{w}_1 + \underline{w}_2) \rangle \\ &= s\langle T\underline{v}, \underline{w}_1 \rangle + \langle T\underline{v}, \underline{w}_2 \rangle \\ &= s\langle \underline{v}, T^*\underline{w}_1 \rangle + \langle \underline{v}, T^*\underline{w}_2 \rangle \\ &= \langle \underline{v}, sT^*\underline{w}_1 \rangle + \langle \underline{v}, T^*\underline{w}_2 \rangle \\ &= \langle \underline{v}, sT^*\underline{w}_1 + T^*\underline{w}_2 \rangle. \end{aligned}$$

Since this holds for all $\underline{v} \in V$, we gather from Lemma 6.4.1 that $T^*(s\underline{v}_1 + \underline{v}_2) = sT^*\underline{v}_1 + T^*\underline{v}_2$. Therefore, T^* is linear. The uniqueness of T^* is inherited from Theorem 6.4.2 because, for each $\underline{w} \in W$, the vector $T^*\underline{w}$ is determined uniquely as the vector \underline{v}' such that $\langle T\underline{v}, \underline{w} \rangle = \langle \underline{v}, \underline{v}' \rangle$ for all $\underline{v} \in V$. \square

Theorem 6.4.4. *Let V be a finite-dimensional inner-product space and let*

$$\mathcal{B} = \underline{v}_1, \dots, \underline{v}_n$$

be an orthonormal basis for V . Let T be a linear operator on V and let A be the matrix representation of T in the ordered basis \mathcal{B} . Then $A_{kj} = \langle T\underline{v}_j, \underline{v}_k \rangle$.

Proof. Since \mathcal{B} is an orthonormal basis, we have

$$\underline{v} = \sum_{k=1}^n \langle \underline{v}, \underline{v}_k \rangle \underline{v}_k.$$

The matrix A is defined by

$$T\underline{v}_j = \sum_{k=1}^n A_{kj} \underline{v}_k$$

and since

$$T\underline{v}_j = \sum_{k=1}^n \langle T\underline{v}_j, \underline{v}_k \rangle \underline{v}_k,$$

we conclude that $A_{kj} = \langle T\underline{v}_j, \underline{v}_k \rangle$. \square

Corollary 6.4.5. *Let V be a finite-dimensional inner-product space, and let T be a linear operator on V . In any orthonormal basis for V , the matrix for T^* is the conjugate transpose of the matrix of T .*

Proof. Let $\mathcal{B} = \underline{v}_1, \dots, \underline{v}_n$ be an orthonormal basis for V , let $A = [T]_{\mathcal{B}}$ and $B = [T^*]_{\mathcal{B}}$. According to the previous theorem,

$$\begin{aligned} A_{kj} &= \langle T\underline{v}_j, \underline{v}_k \rangle \\ B_{kj} &= \langle T^*\underline{v}_j, \underline{v}_k \rangle \end{aligned}$$

By the definition of T^* , we then have

$$B_{kj} = \langle T^*\underline{v}_j, \underline{v}_k \rangle = \overline{\langle \underline{v}_k, T^*\underline{v}_j \rangle} = \overline{\langle T\underline{v}_k, \underline{v}_j \rangle} = \overline{A_{jk}}.$$

□

We note here that every linear operator on a finite-dimensional inner-product space V has an adjoint on V . However, in the infinite-dimensional case this is not necessarily true. In any case, there exists at most one such operator T^* .

6.5 Fundamental Subspaces

There are four fundamental subspaces of a linear transformation $T: V \rightarrow W$ when V and W are Hilbert spaces. We have already encountered two such spaces: The range of T and the nullspace of T . Recall that the range of a linear transformation T is the set of all vectors $\underline{w} \in W$ such that $\underline{w} = T\underline{v}$ for some $\underline{v} \in V$. The nullspace of T consists of all vectors $\underline{v} \in V$ such that $T\underline{v} = \underline{0}$.

The other two fundamental subspaces of T are the **range of the adjoint** T^* , denoted R_{T^*} and the **nullspace of the adjoint** T^* , denoted N_{T^*} . The various subspaces of the transformation $T: V \rightarrow W$ can be summarized as follows,

$$\begin{aligned} R_T &\subseteq W \\ N_T &\subseteq V \\ R_{T^*} &\subseteq V \\ N_{T^*} &\subseteq W. \end{aligned}$$

Theorem 6.5.1. *Let V and W be Hilbert spaces and $T: V \rightarrow W$ be a bounded linear transformation from V to W such that R_T and R_{T^*} are both closed. Then,*

1. *the range R_T is the orthogonal complement of N_{T^*} , i.e., $[R_T]^\perp = N_{T^*}$;*

2. the nullspace N_T is the orthogonal complement of R_{T^*} , i.e., $[R_{T^*}]^\perp = N_T$.

Complementing these equalities, we get

$$\begin{aligned}\overline{R_T} &= R_T = [N_{T^*}]^\perp \\ \overline{R_{T^*}} &= R_{T^*} = [N_T]^\perp.\end{aligned}$$

Proof. Let $\underline{w} \in R_T$, then there exists $\underline{v} \in V$ such that $T\underline{v} = \underline{w}$. Assume that $\underline{n} \in N_{T^*}$, then

$$\langle \underline{w}, \underline{n} \rangle = \langle T\underline{v}, \underline{n} \rangle = \langle \underline{v}, T^*\underline{n} \rangle = 0.$$

That is, \underline{w} and \underline{n} are orthogonal vectors. It follows that $N_{T^*} \subseteq [R_T]^\perp$. Now, let $\underline{w} \in [R_T]^\perp$. Then, for every $\underline{v} \in V$, we have

$$\langle T\underline{v}, \underline{w} \rangle = 0.$$

This implies that $\langle \underline{v}, T^*\underline{w} \rangle = 0$, by the definition of the adjoint. Since this is true for every $\underline{v} \in V$, we get $T^*\underline{w} = \underline{0}$, so $\underline{w} \in N_{T^*}$. Then $[R_T]^\perp \subseteq N_{T^*}$, which combined with our previous result yields $[R_T]^\perp = N_{T^*}$. Using a similar argument, one can show that $[R_{T^*}]^\perp = N_T$. \square

6.6 Pseudoinverses

Theorem 6.6.1. *Let V and W be Hilbert spaces and T be a bounded linear transformation from V to W where R_T is closed. The equation $T\underline{v} = \underline{w}$ has a solution if and only if $\langle \underline{w}, \underline{u} \rangle = 0$ for every vector $\underline{u} \in N_{T^*}$, i.e.,*

$$\underline{w} \in R_T \Leftrightarrow \underline{w} \perp N_{T^*}.$$

In matrix notation, $A\underline{v} = \underline{w}$ has a solution if and only if $\underline{u}^H \underline{w} = 0$ for every vector \underline{u} such that $A^H \underline{u} = \underline{0}$.

Proof. Assume that $T\underline{v} = \underline{w}$, and let $\underline{u} \in N_{T^*}$. Since T is bounded, the adjoint T^* exists and

$$\langle \underline{w}, \underline{u} \rangle = \langle T\underline{v}, \underline{u} \rangle = \langle \underline{v}, T^*\underline{u} \rangle = \langle \underline{v}, \underline{0} \rangle = 0.$$

To prove the reverse implication, suppose that $\langle \underline{w}, \underline{u} \rangle = 0$ when $\underline{u} \in N_{T^*}$ and $T\underline{v} = \underline{w}$ has no solution. Since $\underline{w} \notin R_T$ and R_T is closed, it follows that

$$\underline{w}_o = \underline{w} - P_{R_T} \underline{w} = \underline{w} - \underline{w}_r \neq \underline{0}.$$

But

$$\langle \underline{w}, \underline{w}_o \rangle = \langle \underline{w}_r + \underline{w}_o, \underline{w}_o \rangle = \langle \underline{w}_o, \underline{w}_o \rangle > 0,$$

which contradicts the assumption that $\langle \underline{w}, \underline{u} \rangle = 0$ when $\underline{u} \in N_{T^*}$. We must conclude that $T\underline{v} = \underline{w}$ has a solution. \square

Fact 6.6.2. *The solution to $T\underline{v} = \underline{w}$ (if it exists) is unique if and only if the only solution to $T\underline{v} = \underline{0}$ is $\underline{v} = \underline{0}$. That is, if $N_T = \{\underline{0}\}$.*

6.6.1 Least Squares

Let $T: V \rightarrow W$ be a bounded linear transformation. If the equation $T\underline{v} = \underline{w}$ has no solution, then we can find a vector \underline{v} that minimizes

$$\|T\underline{v} - \underline{w}\|^2.$$

Theorem 6.6.3. *The vector $\underline{v} \in V$ minimizes $\|T\underline{v} - \underline{w}\|$ if and only if*

$$T^*T\underline{v} = T^*\underline{w}.$$

Proof. Minimizing $\|\underline{w} - T\underline{v}\|$ is equivalent to minimizing $\|\underline{w} - \hat{\underline{w}}\|$, where $\hat{\underline{w}} = T\underline{v} \in R_T$. By the projection theorem, we must have

$$\underline{w} - \hat{\underline{w}} \in [R_T]^\perp.$$

But this is equivalent to

$$\underline{w} - \hat{\underline{w}} \in N_{T^*}.$$

That is, $T^*(\underline{w} - \hat{\underline{w}}) = \underline{0}$, or equivalently $T^*\underline{w} = T^*\hat{\underline{w}}$. Conversely, if $T^*T\underline{v} = T^*\underline{w}$, then

$$T^*(T\underline{v} - \underline{w}) = \underline{0},$$

so that $T\underline{v} - \underline{w} \in N_{T^*}$. Hence, the error is orthogonal to the subspace R_T and has minimal length by the projection theorem. \square

Corollary 6.6.4. *If A is a matrix such that $A^H A$ is invertible, then the least-squares solution to $A\underline{v} = \underline{w}$ is*

$$\underline{v} = (A^H A)^{-1} A^H \underline{w}.$$

The matrix $(A^H A)^{-1} A^H$ is the left inverse of A and is an example of a Moore-Penrose **pseudoinverse**.

Theorem 6.6.5. *Suppose the vector $\hat{v} \in V$ minimizes $\|v\|$ over all $v \in V$ satisfying $Tv = \underline{w}$. Then, $\hat{v} \in [N_T]^\perp$ and, if R_{T^*} is closed, $\hat{v} = T^*\underline{u}$ for some $\underline{u} \in W$.*

Proof. Suppose $\hat{v} \notin [N_T]^\perp$, then the orthogonal decomposition $V = [N_T]^\perp + N_T$ shows that the projection of \hat{v} onto $[N_T]^\perp$ has smaller norm but still satisfies $T\hat{v} = \underline{w}$. This gives a contradiction and shows that $\hat{v} \in [N_T]^\perp$. If R_{T^*} is closed, then $R_{T^*} = [N_T]^\perp$ and $\hat{v} = T^*\underline{u}$ for some $\underline{u} \in W$. \square

Corollary 6.6.6. *If A is a matrix such that AA^H is invertible, then the minimum-norm solution to $Av = \underline{w}$ is*

$$\underline{v} = A^H (AA^H)^{-1} \underline{w}.$$

Proof. The theorem shows that $\underline{v} = A^H \underline{u}$ and $A\underline{v} = AA^H \underline{u} = \underline{w}$. Since AA^H is invertible, this gives $\underline{u} = (AA^H)^{-1} \underline{w}$ and computing \underline{v} gives the desired result. \square

The matrix $A^H (AA^H)^{-1}$ is the right inverse of A and is another example of a Moore-Penrose **pseudoinverse**.

Definition 6.6.7. *Let $T: V \rightarrow W$ be a bounded linear transformation, where V and W are Hilbert spaces, and R_T is closed. For each $\underline{w} \in W$, there is a unique vector \hat{v} of minimum norm in the set of vectors that minimize $\|T\underline{v} - \underline{w}\|$. The **pseudoinverse** T^\dagger is the transformation mapping each $\underline{w} \in W$ to its unique \hat{v} .*

Chapter 7

Matrix Factorization and Analysis

Matrix factorizations are an important part of the practice and analysis of signal processing. They are at the heart of many signal-processing algorithms. Their applications include solving linear equations (LU), decorrelating random variables (LDLT, Cholesky), orthogonalizing sets of vectors (QR), and finding low-rank matrix approximations (SVD). Their usefulness is often two-fold: they allow efficient computation of important quantities and they are (often) designed to minimize round-off error due to finite-precision calculation. An algorithm is called *numerically stable*, for a particular set of inputs, if the error in the final solution is proportional to the round-off error in the elementary field operations.

7.1 Triangular Systems

A square matrix $L \in F^{n \times n}$ is called **lower triangular** (or **upper triangular**) if all elements above (or below) the main diagonal are zero. Likewise, a triangular matrix (lower or upper) is a **unit triangular** if it has all ones on the main diagonal. A system of linear equations is called *triangular* if it can be represented by the matrix equation $A\underline{x} = \underline{b}$ where A is either upper or lower triangular.

7.1.1 Solution by Substitution

Let $L \in F^{n \times n}$ be a lower triangular matrix with entries $l_{ij} = [L]_{ij}$. The matrix equation $L\underline{y} = \underline{b}$ can be solved efficiently using **forward substitution**, which is

defined by the recursion

$$y_j = \frac{1}{l_{jj}} \left(b_j - \sum_{i=1}^{j-1} l_{ji} y_i \right), \quad j = 1, 2, \dots, n.$$

Example 7.1.1. Consider the system

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}.$$

Applying the above recursion gives

$$\begin{aligned} y_1 &= \frac{1}{1} = 1 \\ y_2 &= \frac{1}{1}(2 - 1 \cdot 1) = 1 \\ y_3 &= \frac{1}{1}(9 - 1 \cdot 1 - 2 \cdot 1) = 6. \end{aligned}$$

Let $U \in F^{n \times n}$ be an upper triangular matrix with entries $u_{ij} = [U]_{ij}$. The matrix equation $U\underline{x} = \underline{y}$ can be solved efficiently using **backward substitution**, which is defined by the recursion

$$x_j = \frac{1}{u_{jj}} \left(y_j - \sum_{i=j+1}^n u_{ji} x_i \right), \quad j = n, n-1, \dots, 1.$$

Example 7.1.2. Consider the system

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 6 \end{bmatrix}.$$

Applying the above recursion gives

$$\begin{aligned} x_3 &= \frac{6}{2} = 3 \\ x_2 &= \frac{1}{1}(1 - 3 \cdot 3) = -8 \\ x_1 &= \frac{1}{1}(1 - 1 \cdot 6 - 1 \cdot (-8)) = 3. \end{aligned}$$

The computational complexity of each substitution is roughly $\frac{1}{2}n^2$ operations.

Problem 7.1.3. Show that set of upper triangular matrices is a subalgebra of the set of all matrices. Since it is clearly a subspace, only two properties must be verified:

1. that the product of two upper triangular matrices is upper triangular
2. that the inverse of an upper triangular matrix is upper triangular

7.1.2 The Determinant

The determinant $\det(A)$ of a square matrix $A \in F^{n \times n}$ is a scalar which captures a number of important properties of that matrix. For example, A is invertible iff $\det(A) \neq 0$ and the determinant satisfies $\det(AB) = \det(A) \det(B)$ for square matrices A, B . Mathematically, it is the unique function mapping matrices to scalars that is (i) linear in each column, (ii) negated by column transposition, and (iii) satisfies $\det(I) = 1$.

The determinant of a square matrix can be defined recursively using the fact that $\det([a]) = a$. Let $A \in F^{n \times n}$ be an arbitrary square matrix with entries $a_{ij} = [A]_{ij}$. The (i, j) -minor of A is the determinant of the $(n - 1) \times (n - 1)$ matrix formed by deleting the i -th row and j -th column of A .

Fact 7.1.4 (Laplace's Formula). *The determinant of A is given by*

$$\det(A) = \sum_{j=1}^n a_{ij}(-1)^{i+j} M_{ij} = \sum_{i=1}^n a_{ij}(-1)^{i+j} M_{ij},$$

where M_{ij} is the (i, j) -minor of A .

Theorem 7.1.5. *The determinant of a triangular matrix is the product of its diagonal elements.*

Proof. For upper (lower) triangular matrices, this can be shown by expanding the determinant along the first column (row) to compute each minor. \square

7.2 LU Decomposition

7.2.1 Introduction

LU decomposition is a generalization of Gaussian elimination which allows one to efficiently solve a system of linear equations $A\underline{x} = \underline{b}$ multiple times with different

right-hand sides. In its basic form, it is numerically stable only if the matrix is positive definite or diagonally dominant. A slight modification, known as *partial pivoting*, makes it stable for a very large class of matrices.

Any square matrix $A \in F^{n \times n}$ can be factored as $A = LU$, where L is a unit lower-triangular matrix and U is an upper-triangular matrix. The following example uses elementary row operations to cancel, in each column, all elements below the main diagonal. These elementary row operations are represented using left multiplication by a unit lower-triangular matrix.

$$\begin{aligned} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} &= \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} &= \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 2 & 8 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} &= \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix} \end{aligned}$$

This allows one to write

$$\begin{aligned} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix} \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix} \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix} \end{aligned}$$

LU decomposition can also be used to efficiently compute the determinant of A . Since $\det(A) = \det(LU) = \det(L)\det(U)$, the problem is reduced to computing the determinant of triangular matrices. Using Theorem 7.1.5, it is easy to see that $\det(L) = 1$ and $\det(U) = \prod_{i=1}^n u_{ii}$.

7.2.2 Formal Approach

To describe LU decomposition formally, we first need to describe the individual operations that are used to zero out matrix elements.

Definition 7.2.1. Let $A \in F^{n \times n}$ be an arbitrary matrix, $\alpha \in F$ be a scalar, and $i, j \in \{1, 2, \dots, n\}$. Then, adding α times the j -th row to the i -th row an **elementary row-addition operation**. Moreover, $I + \alpha E_{ij}$, where $E_{ij} \triangleq \underline{e}_i \underline{e}_j^T$ and \underline{e}_k is the k -th standard basis vector, is the **elementary row-addition matrix** which effects this operation via left multiplication.

Example 7.2.2. For example, elementary row operations are used to cancel the $(2, 1)$ matrix entry in

$$(I - E_{2,1})A = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 1 & 3 & 9 \end{bmatrix}.$$

Lemma 7.2.3. The following identities capture the important properties of elementary row-operation matrices:

- (i) $E_{ij}E_{kl} = \delta_{j,k}E_{il}$
- (ii) $(I + \alpha E_{ij})(I + \beta E_{kl}) = I + \alpha E_{ij} + \beta E_{kl}$ if $j \neq k$
- (iii) $(I + \alpha E_{ij})^{-1} = (I - \alpha E_{ij})$ if $i \neq j$.

Proof. This proof is left as an exercise. □

Now, consider the process for computing the LU decomposition of A . To initialize the process, we let $A^{(1)} = A$. In each round, we let

$$L_j^{-1} = \prod_{i=j+1}^n \left(I - \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}} E_{i,j} \right)$$

be the product of elementary row operation matrices which cancel the subdiagonal elements of the j -th column. The process proceeds by defining $A^{(j+1)} = L_j^{-1}A^{(j)}$ so that $A^{(j+1)}$ has all zeros below the diagonal in the first j columns. After $n - 1$ rounds, the process terminates with

$$U = A^{(n)} = L_{n-1}^{-1}L_{n-2}^{-1} \cdots L_1^{-1}A$$

where $L = L_1L_2 \cdots L_{n-1}$ is unit lower triangular.

Lemma 7.2.4. *From the structure of elementary row operation matrices, we see*

$$\prod_{j=1}^{n-1} \prod_{i=j+1}^n (I + \alpha_{ij} E_{i,j}) = I + \sum_{j=1}^{n-1} \sum_{i=j+1}^n \alpha_{ij} E_{i,j}.$$

Proof. First, we notice that

$$\prod_{i=j+1}^n (I + \alpha_{ij} E_{i,j}) = I + \sum_{i=j+1}^n \alpha_{ij} E_{i,j}$$

for $j = 1, 2, \dots, n-1$. Expanding the product shows that any term with two E matrices must contain a product $E_{i,j} E_{l,j}$ with $l > i > j$. By Lemma 7.2.3i, we see that this term must be zero because $j \neq l$.

Now, we can prove the main result via induction. First, we assume that

$$\prod_{j=1}^k \prod_{i=j+1}^n (I + \alpha_{ij} E_{i,j}) = I + \sum_{j=1}^k \sum_{i=j+1}^n \alpha_{ij} E_{i,j}.$$

Next, we find that if $k \leq n-2$, then

$$\begin{aligned} \prod_{j=1}^{k+1} \prod_{i=j+1}^n (I + \alpha_{ij} E_{i,j}) &= \left(\prod_{j=1}^k \prod_{i=j+1}^n (I + \alpha_{ij} E_{i,j}) \right) \left(\prod_{l=k+2}^n (I + \alpha_{l,k+1} E_{l,k+1}) \right) \\ &= \left(I + \sum_{j=1}^k \sum_{i=j+1}^n \alpha_{ij} E_{i,j} \right) \left(I + \sum_{l=k+2}^n \alpha_{l,k+1} E_{l,k+1} \right) \\ &= I + \sum_{j=1}^{k+1} \sum_{i=j+1}^n \alpha_{ij} E_{i,j} + \sum_{j=1}^k \sum_{i=j+1}^n \sum_{l=k+2}^n \alpha_{ij} \alpha_{l,k+1} E_{i,j} E_{l,k+1} \\ &= I + \sum_{j=1}^{k+1} \sum_{i=j+1}^n \alpha_{ij} E_{i,j} + \sum_{j=1}^k \sum_{i=j+1}^n \sum_{l=k+2}^n \alpha_{ij} \alpha_{l,k+1} E_{i,k+1} \delta_{j,l} \\ &= I + \sum_{j=1}^{k+1} \sum_{i=j+1}^n \alpha_{ij} E_{i,j}. \end{aligned}$$

Finally, we point out that the base case $k = 1$ is given by the initial observation. \square

Theorem 7.2.5. *This process generates one column of L per round because*

$$[L]_{ij} = \begin{cases} \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}} & \text{if } 1 \leq i < j \\ 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Proof. First, we note that

$$\begin{aligned}
L &= L_1 L_2 \cdots L_{n-1} \\
&= \prod_{j=1}^{n-1} \left(\prod_{i=j+1}^n \left(I - \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}} E_{i,j} \right) \right)^{-1} \\
&\stackrel{(a)}{=} \prod_{i=1}^{n-1} \prod_{i=j+1}^n \left(I - \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}} E_{i,j} \right)^{-1} \\
&\stackrel{(b)}{=} \prod_{i=1}^{n-1} \prod_{i=j+1}^n \left(I + \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}} E_{i,j} \right) \\
&= I + \sum_{i=1}^{n-1} \sum_{i=j+1}^n \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}} E_{i,j},
\end{aligned}$$

where (a) follows from Lemma 7.2.3ii (i.e., all matrices in the inside product commute) and (b) follows from Lemma 7.2.3iii. Picking off the (i, j) entry of L (e.g., with $\underline{e}_i^T L \underline{e}_j$) gives the stated result. \square

Finally, we note that the LU decomposition can be computed in roughly $\frac{2}{3}n^3$ field operations.

7.2.3 Partial Pivoting

Sometimes the *pivot element* $a_{j,j}^{(j)}$ can be very small or zero. In this case, the algorithm will either fail (e.g., divide by zero) or return a very unreliable result. The algorithm can be easily modified to avoid this problem by swapping rows of $A^{(j)}$ to increase the magnitude of the pivot element before each cancellation phase. This results in a decomposition of the form $PA = LU$, where P is a permutation matrix.

In this section, we will describe LU decomposition with partial pivoting using the notation from the previous section. The main difference is that, in each round, we will define $A^{(j+1)} = M_j^{-1} P_j A^{(j)}$ where P_j is a permutation matrix. In particular, left multiplication by P_j swaps row j with row p_j , where

$$p_j = \arg \max_{i=j,j+1,\dots,n} |a_{i,j}^{(j)}|.$$

The matrix M_j^{-1} is now chosen to cancel the subdiagonal elements in j -th column of $P_j A^{(j)}$. After $n - 1$ rounds, the resulting decomposition has the form

$$A^{(n)} = M_{n-1}^{-1} P_{n-1} M_{n-2}^{-1} P_{n-2} \cdots M_1^{-1} P_1 A = U.$$

To show this can also be written in the desired form, we need to understand some properties of the permutations. First, we point that swapping two rows is a transposition and therefore $P_j^2 = I$. Next, we will show that the permutations can be moved to the right.

Lemma 7.2.6. *Let $M = I + \sum_{j=1}^k \sum_{i=j+1}^n \alpha_{ij} E_{ij}$ and Q be a permutation matrix which swaps row $l \geq k + 1$ and row $m > l$. Then, $QM = \widetilde{M}Q$ where*

$$\widetilde{M} = I + \sum_{j=1}^k \sum_{i=j+1}^n \alpha_{ij} Q E_{ij}.$$

Therefore, we can write

$$A^{(n)} = \underbrace{\widetilde{M}_{n-1}^{-1} \widetilde{M}_{n-2}^{-1} \cdots \widetilde{M}_1^{-1}}_{L^{-1}} \underbrace{P_{n-1} \cdots P_2 P_1}_P A = U$$

and $PA = LU$.

Proof. The proof is left as an exercise. □

7.3 LDLT and Cholesky Decomposition

If the matrix $A \in \mathbb{C}^{n \times n}$ is Hermitian, then the LU decomposition allows the factorization $A = LDL^H$, where L is unit lower triangular and D is diagonal. Since this factorization is typically applied to real matrices, it is referred to as **LDLT decomposition**. If A is also positive definite, then the diagonal elements of D are positive and we can write $A = (LD^{1/2})(LD^{1/2})^H$. The form $A = \widetilde{L}\widetilde{L}^H$, where \widetilde{L} is lower triangular, is known as **Cholesky factorization**.

To see this, we will describe the LDLT decomposition using the notation from LU decomposition starting from $A^{(1)} = A$. In the j -th round, define L_j^{-1} to be the product of elementary row-operation matrices which cancel the subdiagonal elements of the j -th column $A^{(j)}$. Then, define $A^{(j+1)} = L_j^{-1} A^{(j)} L_j^{-H}$ and notice that $A^{(j+1)}$ is Hermitian because $A^{(j)}$ is Hermitian. Next, notice that $A^{(j+1)}$ has zeros below the diagonal in the first j columns and zeros to the right of diagonal in the first j rows. This follows from the fact that the first j rows of $A^{(j)}$ are not affected by applying L_j^{-1} on left. Therefore, applying L_j^{-H} on the right also cancels

the elements to the right of the diagonal in the j -th row. After $n - 1$ rounds, we find that $D = A^{(n)}$ is a diagonal matrix.

There are a number of redundancies in the computation described above. First off, the L matrix computed by LU decomposition is identical to the L matrix computed by LDLT decomposition. Therefore, one can save operations by defining $A^{(j+1)} = L_j^{-1}A^{(j)}$. Moreover, the elements to the right of the diagonal in $A^{(j)}$ do not affect the computation at all. So, one can roughly half the number of additions and multiplies by only updating the lower triangular part of $A^{(j)}$. The resulting computational complexity is roughly $\frac{1}{3}n^3$ field operations.

7.3.1 Cholesky Decomposition

For a positive-definite matrix A , we can first apply the LDLT decomposition and then define $\tilde{L} = LD^{1/2}$. This gives the Cholesky decomposition $\tilde{L}\tilde{L}^H = LDL^H = A$.

The Cholesky decomposition is typically used to compute whitening filters for random variables. For example, one can apply it to the correlation matrix $R = E[\underline{X}\underline{X}^H]$ of a random vector \underline{X} . Then, one can define $\underline{Y} = \tilde{L}^{-1}\underline{X}$ and see that

$$E[\underline{Y}\underline{Y}^H] = E[\tilde{L}^{-1}\underline{X}\underline{X}^H\tilde{L}^{-H}] = \tilde{L}^{-1}R\tilde{L}^{-H} = I.$$

From this, one sees that \underline{Y} is a vector of uncorrelated (or white) random variables.

7.3.2 QR decomposition

A complex matrix $Q \in \mathbb{C}^{n \times n}$ is called **unitary** if $Q^H Q = Q Q^H = I$. If all elements of the matrix are real, then it is called **orthogonal** and $Q^T Q = Q Q^T = I$.

Theorem 7.3.1. Any matrix $A \in \mathbb{C}^{m \times n}$ can be factored as

$$A = QR,$$

where Q is an $m \times m$ unitary matrix, $Q Q^H = I$, and R is an $m \times n$ upper-triangular matrix.

Proof. To show this decomposition, we start by applying Gram-Schmidt Orthogonalization to the columns $\underline{a}_1, \dots, \underline{a}_n$ of A . This results in orthonormal vectors

$\{\underline{q}_1, \dots, \underline{q}_l\}$, where $l = \min(m, n)$, such that

$$\underline{a}_j = \sum_{i=1}^{\min(j,l)} r_{i,j} \underline{q}_i \quad \text{for } j = 1, 2, \dots, n.$$

This gives an $m \times l$ matrix $Q = [\underline{q}_1 \ \dots \ \underline{q}_l]$ and an $l \times n$ upper-triangular matrix R , with entries $[R]_{i,j} = r_{i,j}$, such that $A = QR$. If $m \leq n$, then $l = m$, Q is unitary, and the decomposition is complete. Otherwise, we must extend the orthonormal set $\{\underline{q}_1, \dots, \underline{q}_l\}$ to an orthonormal basis $\{\underline{q}_1, \dots, \underline{q}_m\}$ of \mathbb{C}^m . This gives an $m \times m$ unitary matrix $Q' = [\underline{q}_1 \ \dots \ \underline{q}_m]$. Adding $m - n$ rows of zeros to the previous R matrix gives an $m \times n$ matrix R' such that $A = Q'R'$. \square

7.4 Hermitian Matrices and Complex Numbers

Definition 7.4.1. A square matrix $Q \in \mathbb{R}^{n \times n}$ is **orthogonal** if $Q^T Q = Q Q^T = I$.

Definition 7.4.2. A square matrix $U \in \mathbb{C}^{n \times n}$ is **unitary** if $U^H U = U U^H = I$.

It is worth noting that, for unitary (resp. orthogonal) matrices, it suffices to check only that $U^H U = I$ (resp. $Q^T Q = I$) because U is invertible (e.g., it has linearly independent columns) and

$$U^H U = I \implies I = U U^{-1} = U (U^H U) U^{-1} = U U^H.$$

A useful analogy between matrices and complex numbers is as follows.

- *Hermitian matrices* satisfying $A^H = A$ are analogous to real numbers, whose complex conjugates are equal to themselves.
- *Unitary matrices* satisfying $U^H U = I$ are analogous to complex numbers on the unit circle, satisfying $\bar{z}z = 1$.
- *Orthogonal matrices* satisfying $Q^T Q = I$ are analogous to the real numbers $z = \pm 1$, such that $z^2 = 1$.

The transformation

$$z = \frac{1 + jr}{1 - jr}$$

maps real number r into the unit circle $|z| = 1$. Analogously, by *Cayley's formula*,

$$U = (I + jR)(I - jR)^{-1},$$

a Hermitian matrix R is mapped to a unitary matrix.

Chapter 8

Canonical Forms

8.1 Eigenvalues and Eigenvectors

Definition 8.1.1. Let V be a vector space over the field F and let T be a linear operator on V . An **eigenvalue** of T is a scalar $\lambda \in F$ such that there exists a non-zero vector $\underline{v} \in V$ with $T\underline{v} = \lambda\underline{v}$. Any vector \underline{v} such that $T\underline{v} = \lambda\underline{v}$ is called an **eigenvector** of T associated with the eigenvalue value λ .

Definition 8.1.2. The **spectrum** $\sigma(T)$ of a linear operator $T: V \rightarrow V$ is the set of all scalars such that the operator $(T - \lambda I)$ is not invertible.

Example 8.1.3. Let $V = \ell_2$ be the Hilbert space of infinite square-summable sequences and $T: V \rightarrow V$ be the right-shift operator defined by

$$T(v_1, v_2, \dots) = (0, v_1, v_2, \dots).$$

Since T is not invertible, it follows that the scalar 0 is in the spectrum of T . But, it is not an eigenvalue because $T\underline{v} = \underline{0}$ implies $\underline{v} = \underline{0}$ and an eigenvector must be a non-zero vector. In fact, this operator does not have any eigenvalues.

For finite-dimensional spaces, things are quite a bit simpler.

Theorem 8.1.4. Let A be the matrix representation of a linear operator on a finite-dimensional vector space V , and let λ be a scalar. The following are equivalent:

1. λ is an eigenvalue of A
2. the operator $(A - \lambda I)$ is singular

$$3. \det(A - \lambda I) = 0.$$

Proof. First, we show the first and second are equivalent. If λ is an eigenvalue of A , then there exists a non-zero vector $\underline{v} \in V$ such that $A\underline{v} = \lambda\underline{v}$. Therefore, $(A - \lambda I)\underline{v} = 0$ and $(A - \lambda I)$ is singular. Likewise, if $(A - \lambda I)\underline{v} = 0$ for some non-zero $\underline{v} \in V$ and $\lambda \in F$, then $A\underline{v} = \lambda\underline{v}$. To show the second and third are equivalent, we note that the determinant of a matrix is zero iff it is singular. \square

The last criterion is important. It implies that every eigenvalue λ is a root of the polynomial

$$\chi_A(\lambda) \triangleq \det(\lambda I - A)$$

called the **characteristic polynomial** of A . The equation $\det(A - \lambda I) = 0$ is called the characteristic equation of A . The spectrum $\sigma(A)$ is given by the roots of the characteristic polynomial $\chi_A(\lambda)$.

Let A be a matrix over the field of real or complex numbers. A nonzero vector \underline{v} is called a **right eigenvector** for the eigenvalue λ if $A\underline{v} = \lambda\underline{v}$. It is called a **left eigenvector** if $\underline{v}^H A = \lambda\underline{v}^H$.

Definition 8.1.5. Let λ be an eigenvalue of the matrix A . The **eigenspace** associated with λ is the set $E_\lambda = \{\underline{v} \in V \mid A\underline{v} = \lambda\underline{v}\}$. The **algebraic multiplicity** of λ is the multiplicity of the zero at $t = \lambda$ in the characteristic polynomial $\chi_A(t)$. The **geometric multiplicity** of an eigenvalue λ is equal to dimension of the eigenspace E_λ or nullity($A - tI$).

Theorem 8.1.6. If the eigenvalues of an $n \times n$ matrix are all distinct, then the eigenvectors of A are linearly independent.

Proof. We will prove the slightly stronger statement: if $\lambda_1, \lambda_2, \dots, \lambda_k$ are distinct eigenvalues with eigenvectors $\underline{v}_1, \underline{v}_2, \dots, \underline{v}_k$, then the eigenvectors are linearly independent. Suppose that

$$\sum_{i=1}^k c_i \underline{v}_i = \underline{0}$$

for scalars c_1, c_2, \dots, c_k . Notice that one can annihilate \underline{v}_j from this equation by multiplying both sides by $(A - \lambda_j I)$. So, multiplying both sides by a product of

these matrices gives

$$\begin{aligned} \prod_{j=1, j \neq m}^k (A - \lambda_j I) \sum_{i=1}^k c_j \underline{v}_i &= \left(\prod_{j=1, j \neq m}^k (A - \lambda_j I) \right) c_m \underline{v}_m \\ &= c_m \underline{v}_m \prod_{j=1, j \neq m}^k (\lambda_m - \lambda_j) = \underline{0}. \end{aligned}$$

Since all eigenvalues are distinct, we must conclude that $c_m = 0$. Since the choice of m was arbitrary, it follows that c_1, c_2, \dots, c_k are all zero. Therefore, the vectors $\underline{v}_1, \underline{v}_2, \dots, \underline{v}_k$ are linearly independent. \square

Definition 8.1.7. Let T be a linear operator on a finite-dimensional vector space V . The operator T is **diagonalizable** if there exists a basis \mathcal{B} for V such that each basis vector is an eigenvector of T ,

$$[T]_{\mathcal{B}} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

Similarly, a matrix A is diagonalizable if there exists an invertible matrix S such that

$$A = S\Lambda S^{-1}$$

where Λ is a diagonal matrix.

Theorem 8.1.8. If an $n \times n$ matrix has n linearly independent eigenvectors, then it is diagonalizable.

Proof. Suppose that the $n \times n$ matrix A has n linearly independent eigenvectors, which we denote by $\underline{v}_1, \dots, \underline{v}_n$. Let the eigenvalue of \underline{v}_i be denoted by λ_i so that

$$A\underline{v}_j = \lambda_j \underline{v}_j, \quad j = 1, \dots, n.$$

In matrix form, we have

$$\begin{aligned} A \begin{bmatrix} \underline{v}_1 & \cdots & \underline{v}_n \end{bmatrix} &= \begin{bmatrix} A\underline{v}_1 & \cdots & A\underline{v}_n \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 \underline{v}_1 & \cdots & \lambda_n \underline{v}_n \end{bmatrix}. \end{aligned}$$

We can rewrite the last matrix on the right as

$$\begin{bmatrix} \lambda_1 \underline{v}_1 & \cdots & \lambda_n \underline{v}_n \end{bmatrix} = \begin{bmatrix} \underline{v}_1 & \cdots & \underline{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} = S\Lambda.$$

where

$$S = \begin{bmatrix} \underline{v}_1 & \cdots & \underline{v}_n \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix},$$

Combining these two equations, we obtain the equality

$$AS = S\Lambda.$$

Since the eigenvectors are linearly independent, the matrix S is full rank and hence invertible. We can therefore write

$$\begin{aligned} A &= S\Lambda S^{-1} \\ \Lambda &= S^{-1}AS. \end{aligned}$$

That is, the matrix A is diagonalizable. □

The type of the transformation from A to Λ arises in a variety of contexts.

Definition 8.1.9. *If there exists an invertible matrix T such that*

$$A = TBT^{-1},$$

*then matrices A and B are said to be **similar**.*

If A and B are similar, then they have the same eigenvalues. Similar matrices can be considered representations of the same linear operator using different bases.

Lemma 8.1.10. *Let A be an $n \times n$ Hermitian matrix (i.e., $A^H = A$). Then, the eigenvalues of A are real and the eigenvectors associated with distinct eigenvalues are orthogonal.*

Proof. First, we notice that $A = A^H$ implies $\underline{v}^H A \underline{v}$ is real because

$$\bar{s} = (\underline{v}^H A \underline{v})^H = \underline{v}^H A^H \underline{v} = \underline{v}^H A \underline{v} = s.$$

If $A\underline{v} = \lambda_1\underline{v}$, left multiplication by \underline{v}^H shows that

$$\underline{v}^H A\underline{v} = \lambda_1 \underline{v}^H \underline{v} = \lambda_1 \|\underline{v}\|.$$

Therefore, λ_1 is real. Next, assume that $A\underline{w} = \lambda_2\underline{w}$ and $\lambda_2 \neq \lambda_1$. Then, we have

$$\lambda_1 \lambda_2 \underline{w}^H \underline{v} = \underline{w}^H A^H A \underline{v} = \underline{w}^H A^2 \underline{v} = \lambda_1^2 \underline{w}^H \underline{v}.$$

We also assume, without loss of generality, that $\lambda_1 \neq 0$. Therefore, if $\lambda_2 \neq \lambda_1$, then $\underline{w}^H \underline{v} = 0$ and the eigenvectors are orthogonal. \square

8.2 Applications of Eigenvalues

8.2.1 Differential Equations

It is well known that the solution of the 1st-order linear differential equation

$$\frac{d}{dt}x(t) = ax(t)$$

is given by

$$x(t) = e^{at}x(0).$$

It turns out that this formula can be extended to coupled differential equations. Let A be a diagonalizable matrix and consider the the set of 1st order linear differential equations defined by

$$\frac{d}{dt}\underline{x}(t) = A\underline{x}(t).$$

Using the decomposition $A = SAS^{-1}$ and the substitution $\underline{x}(t) = S\underline{y}(t)$, we find that

$$\begin{aligned} \frac{d}{dt}\underline{x}(t) &= \frac{d}{dt}S\underline{y}(t) \\ &= S\frac{d}{dt}\underline{y}(t). \end{aligned}$$

and

$$\begin{aligned} \frac{d}{dt}\underline{x}(t) &= A\underline{x}(t) \\ &= AS\underline{y}(t). \end{aligned}$$

This implies that

$$\frac{d}{dt}\underline{y}(t) = S^{-1}AS\underline{y}(t) = \Lambda\underline{y}(t).$$

Solving each individual equation gives

$$y_j(t) = e^{\lambda_j t} y_j(0)$$

and we can group them together in matrix form with

$$\underline{y}(t) = e^{\Lambda t} \underline{y}(0).$$

In terms of $\underline{x}(t)$, this gives

$$\underline{x}(t) = S e^{\Lambda t} S^{-1} \underline{x}(0).$$

In the next section, we will see this is equal to $\underline{x}(t) = e^{At} \underline{x}(0)$.

8.2.2 Functions of a Matrix

The diagonal form of a diagonalizable matrix can be used in a number of applications. One such application is the computation of matrix exponentials. If $A = S\Lambda S^{-1}$ then

$$A^2 = S\Lambda S^{-1}S\Lambda S^{-1} = S\Lambda^2 S^{-1}$$

and, more generally,

$$A^n = S\Lambda^n S^{-1}.$$

Note that Λ^n is obtained in a straightforward manner as

$$\Lambda^n = \begin{bmatrix} \lambda_1^n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^n \end{bmatrix}.$$

This observation drastically simplifies the computation of the matrix exponential e^A ,

$$e^A = \sum_{i=0}^{\infty} \frac{A^i}{i!} = S \left(\sum_{i=0}^{\infty} \frac{\Lambda^i}{i!} \right) S^{-1} = S e^{\Lambda} S^{-1},$$

where

$$e^{\Lambda} = \begin{bmatrix} e^{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n} \end{bmatrix}.$$

Theorem 8.2.1. *Let $p(\cdot)$ be a given polynomial. If λ is an eigenvalue of A , while \underline{v} is an associated eigenvector, then $p(\lambda)$ is an eigenvalue of the matrix $p(A)$ and \underline{v} is an eigenvector of $p(A)$ associated with $p(\lambda)$.*

Proof. Consider $p(A)\underline{v}$. Then,

$$p(A)\underline{v} = \sum_{k=0}^l p_k A^k \underline{v} = \sum_{k=0}^l p_k \lambda^k \underline{v} = p(\lambda)\underline{v}.$$

That is $p(A)\underline{v} = p(\lambda)\underline{v}$. □

A matrix A is singular if and only if 0 is an eigenvalue of A .

8.3 The Jordan Form

Not all matrices are diagonalizable. In particular, if A has an eigenvalue whose algebraic multiplicity is larger than its geometric multiplicity, then that eigenvalue is called **defective**. A matrix with a defective eigenvalue is not diagonalizable.

Theorem 8.3.1. *Let A be an $n \times n$ matrix. Then A is diagonalizable if and only if there is a set of n linearly independent vectors, each of which is an eigenvector of A .*

Proof. If A has n linearly independent eigenvectors $\underline{v}_1, \dots, \underline{v}_n$, then let S be an invertible matrix whose columns are these n vectors. Consider

$$\begin{aligned} S^{-1}AS &= S^{-1} \begin{bmatrix} A\underline{v}_1 & \cdots & A\underline{v}_n \end{bmatrix} \\ &= S^{-1} \begin{bmatrix} \lambda_1 \underline{v}_1 & \cdots & \lambda_n \underline{v}_n \end{bmatrix} \\ &= S^{-1}S\Lambda = \Lambda. \end{aligned}$$

Conversely, suppose that there is a similarity matrix S such that $S^{-1}AS = \Lambda$ is a diagonal matrix. Then $AS = S\Lambda$. This implies that A times the i th column of S is the i th diagonal entry of Λ times the i th column of S . That is, the i th column of S is an eigenvector of A associated with the i th diagonal entry of Λ . Since S is nonsingular, there are exactly n linearly independent eigenvectors. □

Definition 8.3.2. The **Jordan normal form** of any matrix $A \in \mathbb{C}^{n \times n}$ with $l \leq n$ linearly independent eigenvectors can be written as

$$A = TJT^{-1},$$

where T is an invertible matrix and J is the block-diagonal matrix

$$J = \begin{bmatrix} J_{m_1}(\lambda_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & J_{m_l}(\lambda_l) \end{bmatrix}.$$

The $J_m(\lambda)$ are $m \times m$ matrices called **Jordan blocks**, and they have the form

$$J_m(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{bmatrix}.$$

It is important to note that the eigenvalues $\lambda_1, \dots, \lambda_l$ are not necessarily distinct (i.e., multiple Jordan blocks may have the same eigenvalue). The Jordan matrix J associated with any matrix A is unique up to the order of the Jordan blocks. Moreover, two matrices are similar iff they are both similar to the same Jordan matrix J .

Since every matrix is similar to a Jordan block matrix, one can gain some insight by studying Jordan blocks. In fact, Jordan blocks exemplify the way that matrices can be degenerate. For example, $J_m(\lambda)$ has the single eigenvector \underline{e}_1 (i.e., the standard basis vector) and satisfies

$$J_m(\lambda)\underline{e}_{j+1} = \underline{e}_j \quad \text{for } j = 1, 2, \dots, m-1.$$

So, the reason this matrix has only one eigenvector is that left-multiplication by this matrix shifts all elements in a vector up element.

Computing the Jordan normal form of a matrix can be broken into two parts. First, one can identify, for each distinct eigenvalue λ , the **generalized eigenspace**

$$G_\lambda = \{ \underline{v} \in \mathbb{C}^n \mid (A - \lambda I)^n \underline{v} = \underline{0} \}.$$

Let $\lambda_1, \dots, \lambda_k$ be the distinct eigenvalues of A ordered by decreasing magnitude. Let d_j be the dimension of G_{λ_j} , which is equal to the sum of the sizes of the Jordan

blocks associated with λ , then $\sum_{j=1}^k d_j = n$. Let T be a matrix whose first d_1 columns form a basis for G_{λ_1} , next d_2 columns form a basis for G_{λ_2} , and so on. In this case, the matrix $T^{-1}AT$ is block diagonal and the j -th block B_j is associated with the eigenvalue λ_j .

To put A in Jordan normal form, we now need to transform each block matrix B into Jordan normal form. One can do this by identifying the subspace V_j that is not mapped to $\underline{0}$ by $(B - \lambda I)^{j-1}$ (i.e., $\mathcal{N}((B - \lambda I)^{j-1})^\perp$). This gives the sequence V_1, \dots, V_J of non-empty subspaces (e.g., V_j is empty for $j > J$). Now, we can form a sequence of bases W_J, W_{J-1}, \dots, W_1 recursively starting from W_J with

$$W_j = W_{j+1} \cup \{(B - \lambda I)\underline{w} \mid \underline{w} \in W_{j+1}\} \cup \text{basis}(V_j - V_{j-1}),$$

where $\text{basis}(V_j - V_{j-1})$ is some set basis vectors that extends V_{j-1} to V_j . Each vector in W_j gives rise to a length j **Jordan chain** of vectors $\underline{v}_{i-1} = (B - \lambda I)\underline{v}_i \in W_{i-1}$ starting from any $\underline{v}_j \in W_j$. Each vector \underline{v}_j defined in this way is called a **generalized eigenvector** of order j . By correctly ordering the basis W_1 as columns of T , one finds that $T^{-1}BT$ is a Jordan matrix.

Example 8.3.3. Consider the matrix

$$\begin{bmatrix} 4 & 0 & 1 & 0 \\ 2 & 2 & 3 & 0 \\ -1 & 0 & 2 & 0 \\ 4 & 0 & 1 & 2 \end{bmatrix}.$$

First, we find the characteristic polynomial

$$\chi_A(t) = \det(tI - A) = t^4 - 10t^3 + 37t^2 - 60t + 36 = (t - 2)^2(t - 3)^2.$$

Next, we find the eigenvectors associated with the eigenvalues $\lambda_1 = 3$ and $\lambda_2 = 2$. This is done by finding a basis $\underline{v}_1^{(i)}, \underline{v}_2^{(i)}, \dots$ for the nullspace of $A - \lambda_i I$ and gives

$$\begin{aligned} \underline{v}_1^{(1)} &= [1 \ -1 \ -1 \ 3]^T \\ \underline{v}_1^{(2)} &= [0 \ 1 \ 0 \ 0]^T \\ \underline{v}_2^{(2)} &= [0 \ 0 \ 0 \ 1]^T. \end{aligned}$$

Since the eigenvalue λ_1 has algebraic multiplicity 2 and geometric multiplicity 1, we still need to find another generalized eigenvector associated with this eigenspace.

In particular, we need a vector \underline{w} which satisfies $(A - \lambda_1 I)\underline{w} = \underline{v}_1^{(1)}$. This gives

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 2 & -1 & 3 & 0 \\ -1 & 0 & -1 & 0 \\ 4 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 3 \end{bmatrix}.$$

Using the pseudoinverse of $(A - \lambda_1 I)$, one finds that $\underline{w} = \left[\frac{11}{12} \frac{37}{12} \frac{1}{12} \frac{9}{12} \right]$. Using this, we construct the Jordan normal form by noting that

$$\begin{bmatrix} 4 & 0 & 1 & 0 \\ 2 & 2 & 3 & 0 \\ -1 & 0 & 2 & 0 \\ 4 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} \underline{v}_1^{(1)} & \underline{w} & \underline{v}_1^{(2)} & \underline{v}_2^{(2)} \end{bmatrix} = \begin{bmatrix} 3\underline{v}_1^{(1)} & \underline{v}_1^{(1)} + 3\underline{w} & 2\underline{v}_1^{(2)} & 2\underline{v}_2^{(2)} \end{bmatrix}$$

$$= \begin{bmatrix} \underline{v}_1^{(1)} & \underline{w} & \underline{v}_1^{(2)} & \underline{v}_2^{(2)} \end{bmatrix} \begin{bmatrix} 3 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

This implies that $A = T J T^{-1}$ with

$$T = \begin{bmatrix} \underline{v}_1^{(1)} & \underline{w} & \underline{v}_1^{(2)} & \underline{v}_2^{(2)} \end{bmatrix} = \begin{bmatrix} 1 & \frac{11}{12} & 0 & 0 \\ -1 & \frac{37}{12} & 1 & 0 \\ -1 & \frac{1}{12} & 0 & 0 \\ 3 & \frac{9}{12} & 0 & 1 \end{bmatrix}.$$

8.4 Applications of Jordan Normal Form

Jordan normal form often allows one to extend to all matrices results that are easy to prove for diagonalizable matrices.

8.4.1 Convergent Matrices

Definition 8.4.1. An $n \times n$ matrix A is **convergent** if $\|A^k\| \rightarrow 0$ for any norm.

Of course, this is equivalent to the statement “ A^k converges to the all zero matrix”. Since all finite-dimensional vector norms are equivalent, it also follows that this condition does not depend on the norm chosen.

Recall that the spectral radius $\rho(A)$ of a matrix A is the magnitude of the largest eigenvalue. If A is diagonalizable, then $A^k = T\Lambda^k T^{-1}$ and it is easy to see that

$$\|A^k\| \leq \|T\| \|\Lambda^k\| \|T^{-1}\|.$$

Since all finite-dimensional vector norms are equivalent, we know that $\|\Lambda^k\| \leq M\|\Lambda^k\|_1 = M\rho(A)^k$. Therefore, A is convergent if $\rho(A) < 1$. If $\rho(A) \geq 1$, then it is easy to show that $\|\Lambda^k\| > 0$ and therefore that $\|A^k\| > 0$. For general matrices, we can instead use the Jordan normal form and the following lemma.

Lemma 8.4.2. *The Jordan block $J_m(\lambda)$ is convergent iff $|\lambda| < 1$.*

Proof. This follows from the fact that $J_m(\lambda) = \lambda I + N$, where $[N]_{i,j} = \delta_{i+1,j}$. Using the Binomial formula, we write

$$\begin{aligned} \|(\lambda I + N)^k\| &= \left\| \sum_{i=0}^k \binom{k}{i} N^i \lambda^{k-i} \right\| \\ &\leq \sum_{i=0}^{m-1} \binom{k}{i} |\lambda|^{k-i}, \end{aligned}$$

where the second step follows from the fact that $\|N^i\|$ is 1 for $i = 1, \dots, m-1$ and zero for $i \geq m$. Notice that $\binom{k}{i} |\lambda|^{k-i} \leq k^{m-1} |\lambda|^{k-m+1}$ for $0 \leq i \leq m-1$. Since $k^{m-1} |\lambda|^{k-m+1} \rightarrow 0$ as $k \rightarrow \infty$ iff $|\lambda| < 1$, we see that each term in the sum converges to zero under the same condition. On the other hand, if $|\lambda| \geq 1$, then $|[(\lambda I + N)^k]_{1,1}| \geq 1$ for all $k \geq 0$. \square

Theorem 8.4.3. *A matrix $A \in \mathbb{C}^{n \times n}$ is convergent iff $\rho(A) < 1$.*

Proof. Using the Jordan normal form, we can write $A = TJT^{-1}$, where J is a block diagonal with k Jordan blocks J_1, \dots, J_k . Since J is block diagonal, we also have that $\|J^k\| \leq \sum_{i=1}^k \|J_i^k\|$. If $\rho(A) < 1$, then the eigenvalue λ associated with each Jordan block satisfies $|\lambda| < 1$. In this case, the lemma shows that $\|J_i^k\| \rightarrow 0$ which implies that $\|J^k\| \rightarrow 0$. Therefore, $\|A^k\| \rightarrow 0$ and A is convergent. On the other hand, if $\rho(A) \geq 1$, then there is a Jordan block J_i with $|\lambda| \geq 1$ and $|[J_i^k]_{1,1}| \geq 1$ for all $k \geq 0$. \square

In some cases, one can make stronger statements about large powers of a matrix.

Definition 8.4.4. A matrix A has a **unique eigenvalue of maximum modulus** if the Jordan block associated with that eigenvalue is 1×1 and all other Jordan blocks are associated with eigenvalues of smaller magnitude.

The following theorem shows that a properly normalized matrix of this type converges to a non-zero limit.

Theorem 8.4.5. If A has a unique eigenvalue λ_1 of maximum modulus, then

$$\lim_{k \rightarrow \infty} \frac{1}{\lambda_1^k} A^k = \underline{u} \underline{v}^H,$$

where $A \underline{u} = \lambda_1 \underline{u}$, $\underline{v}^H A = \lambda_1 \underline{v}^H$, and $\underline{v}^H \underline{u} = 1$.

Proof. Let $B = \frac{1}{\lambda_1} A$ so that maximum modulus eigenvalue is now 1. Next, choose the Jordan normal form $B = T J T^{-1}$ so that the Jordan block associated with the eigenvalue 1 is in the top left corner of J . In this case, it follows from the lemma that J^n converges to $\underline{e}_1 \underline{e}_1^H$ as $n \rightarrow \infty$. This implies that $B^n = T J^n T^{-1}$ converges to $T \underline{e}_1 \underline{e}_1^H T^{-1} = \underline{u} \underline{v}^H$ where \underline{u} is the first column of T and \underline{v}^H is the first row of T^{-1} .

By construction, the first column of T is the right eigenvector \underline{u} and satisfies $A \underline{u} = \lambda_1 \underline{u}$. Likewise, the first row of T^{-1} is the left eigenvector \underline{v}^H associated with the eigenvalue 1 because $B^H = T^{-H} J^H T^H$ and the first column of T^{-H} (i.e., Hermitian conjugate of first row of T^{-1}) is the right eigenvector of A^H associated with λ_1 . Therefore, $\underline{v}^H A = \lambda_1 \underline{v}^H$. Finally, the fact that $\underline{u} = B^n \underline{u} \rightarrow \underline{u} \underline{v}^H \underline{u}$ implies that $\underline{v}^H \underline{u} = 1$. \square

Chapter 9

Singular Value Decomposition

9.1 Diagonalization of Hermitian Matrices

Lemma 9.1.1 (Schur Decomposition). *For any square matrix A , there exists a unitary matrix U such that*

$$U^H A U = T$$

where T is upper triangular. That is, every square matrix is similar to an upper-triangular matrix.

Proof. We prove this lemma by induction on the size n of the matrix. Since it is clearly true for scalars (i.e., matrices of size $n = 1$), the base case is trivial. Now, suppose that the result holds for all $k = 1, 2, \dots, n - 1$ and let $A \in \mathbb{C}^{n \times n}$. Since every matrix has at least one eigenvector, we let \underline{u} be an eigenvector of A normalized so that $\|\underline{u}\|_2 = 1$. Using the Gram-Schmidt procedure, it is possible to construct an orthonormal basis $\mathcal{B} = \{\underline{x}_1, \dots, \underline{x}_n\}$ for \mathbb{C}^n , with $\underline{x}_1 = \underline{u}$. Define the matrix U_n by

$$U_n = \begin{bmatrix} \underline{x}_1 & \cdots & \underline{x}_n \end{bmatrix}.$$

Since \mathcal{B} is a basis for \mathbb{C}^n , every column of the matrix $A U_n$ can be expressed as a linear combination of vectors in \mathcal{B} , say,

$$A \underline{x}_i = \sum_{j=1}^n s_{j,i} \underline{x}_j \quad i = 1, \dots, n.$$

Note that $A \underline{x}_1 = \lambda_1 \underline{x}_1$ for some λ_1 since $\underline{x}_1 = \underline{u}$, an eigenvector of A . We can then

write

$$AU_n = \begin{bmatrix} A\underline{x}_1 & \cdots & A\underline{x}_n \end{bmatrix} = U_n \begin{bmatrix} \lambda_1 & s_{1,2} & \cdots & s_{1,n} \\ 0 & s_{2,2} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & s_{n,2} & \cdots & s_{n,n} \end{bmatrix} = U_n \begin{bmatrix} \lambda_1 & \underline{s}^T \\ \underline{0} & A_{n-1} \end{bmatrix},$$

where we have used the convenient notation

$$A_{n-1} = \begin{bmatrix} s_{2,2} & \cdots & s_{2,n} \\ \vdots & \ddots & \vdots \\ s_{n,2} & \cdots & s_{n,n} \end{bmatrix}$$

and $\underline{s}^T = (s_{1,2}, \dots, s_{1,n})$. By the inductive hypothesis, we can write $A_{n-1} = U_{n-1}T_{n-1}U_{n-1}^H$ where T_{n-1} is upper triangular and U_{n-1} is unitary. It follows that

$$\begin{aligned} AU_n &= U_n \begin{bmatrix} \lambda_1 & \underline{s}^T \\ \underline{0} & A_{n-1} \end{bmatrix} = U_n \begin{bmatrix} \lambda_1 & \underline{s}^T \\ \underline{0} & U_{n-1}T_{n-1}U_{n-1}^H \end{bmatrix} \\ &= U_n \begin{bmatrix} 1 & \underline{0}^T \\ \underline{0} & U_{n-1} \end{bmatrix} \begin{bmatrix} \lambda_1 & \underline{s}^T U_{n-1} \\ \underline{0} & T_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \underline{0}^T \\ \underline{0} & U_{n-1}^H \end{bmatrix}. \end{aligned}$$

Let U be the matrix given by

$$U = U_n \begin{bmatrix} 1 & \underline{0}^T \\ \underline{0} & U_{n-1} \end{bmatrix},$$

and note that U is unitary. It follows that

$$U^H AU = \begin{bmatrix} \lambda_1 & \underline{s}^T U_{n-1} \\ \underline{0} & T_{n-1} \end{bmatrix}.$$

That is, U is a unitary matrix such that $U^H AU$ is upper-triangular. \square

We use this lemma to prove the following theorem.

Theorem 9.1.2. *Every Hermitian $n \times n$ matrix A can be diagonalized by a unitary matrix,*

$$U^H AU = \Lambda,$$

where U is unitary and Λ is a diagonal matrix.

Proof. Note that $A^H = A$ and $T = U^H A U$. Consider the matrix T^H given by

$$T^H = (U^H A U)^H = U^H A^H U = U^H A U = T.$$

That is, T is also Hermitian. Since T is upper triangular, this implies that T is a diagonal matrix. We must conclude that every Hermitian matrix is diagonalized by a unitary matrix. \square

This proves every Hermitian matrix has a complete set of orthonormal eigenvectors.

9.2 Singular Value Decomposition

The singular value decomposition (SVD) provides a matrix factorization related to the eigenvalue decomposition that works for all matrices. In general, any matrix $A \in \mathbb{C}^{m \times n}$ can be factored into a product of unitary matrices and a diagonal matrix, as explained below.

Theorem 9.2.1. *Let A be a matrix in $\mathbb{C}^{m \times n}$. Then A can be factored as*

$$A = U \Sigma V^H$$

where $U \in \mathbb{C}^{m \times m}$ is unitary, $V \in \mathbb{C}^{n \times n}$ is unitary, and $\Sigma \in \mathbb{R}^{m \times n}$ has the form

$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_p),$$

where $p = \min(m, n)$.

The diagonal elements of Σ are called the *singular values* of A and are typically ordered so that

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0.$$

Proof. Let

$$A^H A V = V \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$$

be the spectral decomposition of $A^H A$, where the columns of V are orthonormal eigenvectors

$$V = \begin{bmatrix} \underline{v}_1 & \underline{v}_2 & \cdots & \underline{v}_n \end{bmatrix},$$

with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$ and $r \leq p$. For $i \leq r$, let

$$\underline{u}_i = \frac{Av_i}{\sqrt{\lambda_i}},$$

and observe that

$$\langle \underline{u}_i, \underline{u}_j \rangle = \frac{v_j^H A^H Av_i}{\sqrt{\lambda_i \lambda_j}} = \frac{v_j^H v_i \lambda_i}{\sqrt{\lambda_i \lambda_j}} = \delta_{ij}.$$

Also note that $\{\underline{u}_i\}$ are eigenvectors of AA^H since

$$AA^H \underline{u}_i = AA^H A \frac{v_i}{\sqrt{\lambda_i}} = \sqrt{\lambda_i} Av_i = \lambda_i \underline{u}_i.$$

The set $\{\underline{u}_i : i = 1, \dots, r\}$ can be extended using the Gram-Schmidt procedure to form an orthonormal basis for \mathbb{C}^m . Let

$$U = \begin{bmatrix} \underline{u}_1 & \cdots & \underline{u}_m \end{bmatrix}.$$

For the zero eigenvalues, the eigenvectors must come from the nullspace of AA^H since the eigenvectors with zero eigenvalues are, by construction, orthogonal to the eigenvectors with nonzero eigenvalues that are in the range of AA^H .

For \underline{u}_i where $i \leq r$, we get

$$\underline{u}_i^H AV = \frac{1}{\sqrt{\lambda_i}} v_i^H A^H AV = \sqrt{\lambda_i} e_i^H.$$

On the other hand, if $i > r$ then $\underline{u}_i^H AV = \underline{0}$. Hence,

$$U^H AV = \text{diag} \left(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n} \right) = \Sigma,$$

as desired. □

This proof gives a recipe for computing the SVD of an arbitrary matrix. Consider the matrix

$$A = \begin{bmatrix} 1 & 1 \\ 5 & -1 \\ -1 & 5 \end{bmatrix}.$$

The eigenvalue decomposition of $A^H A$ is given by

$$A^H A = \begin{bmatrix} 27 & -9 \\ -9 & 27 \end{bmatrix} = V \Lambda V^H = \left(\frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \right) \begin{bmatrix} 36 & 0 \\ 0 & 18 \end{bmatrix} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \right).$$

This implies that $\Sigma_1 = \Lambda^{1/2}$ and $V_1 = V$. Therefore, we can compute $U_1 = AV_1\Sigma_1^{-1}$ with

$$U_1 = \begin{bmatrix} 1 & 1 \\ 5 & -1 \\ -1 & 5 \end{bmatrix} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \right) \begin{bmatrix} \frac{1}{\sqrt{36}} & 0 \\ 0 & \frac{1}{\sqrt{18}} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{3} \\ \frac{1}{\sqrt{2}} & \frac{2}{3} \\ -\frac{1}{\sqrt{2}} & \frac{2}{3} \end{bmatrix}.$$

Putting this all together, we have the compact SVD

$$A = U_1\Sigma_1V_1^H = \begin{bmatrix} 0 & \frac{1}{3} \\ \frac{1}{\sqrt{2}} & \frac{2}{3} \\ -\frac{1}{\sqrt{2}} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} \sqrt{36} & 0 \\ 0 & \sqrt{18} \end{bmatrix} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \right).$$

9.3 Properties of the SVD

Many of the important properties of the SVD can be understood better by separating the non-zero singular values from the zero singular values. To do this, we note that every rank r matrix $A \in \mathbb{C}^{m \times n}$ has a singular value decomposition

$$A = U\Sigma V^H = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^H \\ V_2^H \end{bmatrix} = U_1\Sigma_1V_1^H,$$

where $U \in \mathbb{C}^{m \times m}$ and $V \in \mathbb{C}^{n \times n}$ are unitary and $U_1 \in \mathbb{C}^{m \times r}$, $U_2 \in \mathbb{C}^{m \times m-r}$, $V_1 \in \mathbb{C}^{n \times r}$, and $V_2 \in \mathbb{C}^{n \times n-r}$ have orthonormal columns. The diagonal matrix $\Sigma_1 \in \mathbb{R}^{r \times r}$ contains the non-zero singular values

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0.$$

The factorization $A = U\Sigma V^H$ is called the **full SVD** of the matrix A while the factorization $A = U_1\Sigma_1V_1^H$ is called the **compact SVD** of A . The compact SVD of a rank- r matrix retains only the r columns of U, V associated with non-zero singular values.

Let X, Y be inner product spaces and let A define a mapping from X to Y . Then, the columns of V_1 form an orthonormal basis for the vectors in X that are mapped to non-zero vectors (i.e., $\mathcal{N}(A)^\perp$) while the columns of V_2 form an orthonormal basis of $\mathcal{N}(A)$. Likewise, the columns of U_1 form an orthonormal basis for the vectors in Y that lie in the range of A while the vectors in U_2 form an orthonormal basis for $\mathcal{R}(A)^\perp$. It follows that the full SVD computes orthonormal bases for

all of the four fundamental subspaces of the matrix A . For example, it is easy to show that

$$\begin{aligned}\mathcal{R}(A) &= \text{span}(U_1) \\ \mathcal{R}(A^H) &= \text{span}(V_1) \\ \mathcal{N}(A) &= \text{span}(V_2) \\ \mathcal{N}(A^H) &= \text{span}(U_2)\end{aligned}$$

To see this, notice that $A \sum_{i=1}^t c_i \underline{v}_i = \sum_{i=1}^t c_i \sigma_i \underline{u}_i$.

From this, we can compute easily any projection onto a fundamental subspace. First, we point out that the projection onto the column space of any matrix $W \in \mathbb{C}^{m \times n}$ with orthonormal columns (i.e., $W^H W = I$) is given by

$$P_W = W(W^H W)^{-1} W^H = W W^H.$$

Therefore, the projection matrices for the fundamental subspaces are given by

$$\begin{aligned}P_{\mathcal{R}(A)} &= U_1 U_1^H \\ P_{\mathcal{R}(A^H)} &= V_1 V_1^H \\ P_{\mathcal{N}(A)} &= V_2 V_2^H = I - V_1 V_1^H \\ P_{\mathcal{N}(A^H)} &= U_2 U_2^H = I - U_1 U_1^H.\end{aligned}$$

This decomposition also provides a rank revealing decomposition of a rank- r matrix

$$A = \sum_{i=1}^r \sigma_i \underline{u}_i \underline{v}_i^H,$$

where \underline{u}_i is the i th column of U and \underline{v}_i is the i th column of V . This shows A as the sum of r rank-1 matrices. It also allows one to compute

$$\begin{aligned}\|A\|_F^2 &= \sum_{i=1}^r \sigma_i^2 \\ \|A\|_2 &= \sigma_1\end{aligned}$$

The pseudoinverse of A is also very easy to compute from the SVD. In particular, one finds that

$$A^\dagger = V \Sigma^\dagger U^H = V_1 \Sigma_1^{-1} U_1^H.$$

One can verify this by computing $A^\dagger A$ and AA^\dagger . It also follows from the fact that the pseudoinverse of a scalar σ is σ^{-1} if $\sigma \neq 0$ and zero otherwise.

Appendix A

Optional Topics

A.1 Dealing with Infinity*

A.1.1 The Axiom of Choice

The **axiom of choice**, formulated by Zermelo in 1904, is innocent-looking. However, one can prove theorems with its aid that some mathematicians were originally reluctant to accept in the past.

Definition A.1.1 (The Axiom of Choice). *Given a collection \mathcal{X} of disjoint nonempty sets, there exists a set C having exactly one element in common with each element of \mathcal{X} . That is, for each $X \in \mathcal{X}$ the set $C \cap X$ contains a single element.*

Most mathematicians today accept the axiom of choice as part of the set theory on which they base their mathematics. A straightforward consequence of the axiom of choice is the existence of a choice function.

Lemma A.1.2 (Existence of a Choice Function). *Given a collection \mathcal{Y} of non-empty sets, there exists a function*

$$c : \mathcal{Y} \rightarrow \bigcup_{Y \in \mathcal{Y}} Y$$

satisfying $c(Y) \in Y$ for every $Y \in \mathcal{Y}$.

Proof. The difference between the axiom of choice and the lemma is that in the latter statement the sets of the collection \mathcal{Y} need not be disjoint. Given an element $Y \in \mathcal{Y}$, define the set Y' by

$$Y' = \{(Y, y) \mid y \in Y\}.$$

That is, Y' is the collection of all ordered pairs where the first coordinate of the ordered pair is the set Y , and the second coordinate is an element of Y . Because Y contains at least one element, the set Y' is nonempty. Furthermore, Y' is a subset of the cartesian product

$$\mathcal{Y} \times \bigcup_{Y \in \mathcal{Y}} Y.$$

If Y_1 and Y_2 are two different sets in \mathcal{Y} , then the sets Y'_1 and Y'_2 are disjoint; specifically, the elements of Y'_1 and Y'_2 differ at least in their first coordinates.

Consider the collection

$$\mathcal{Z} = \{Y' \mid Y \in \mathcal{Y}\}.$$

This is a collection of disjoint nonempty subsets of

$$\mathcal{Y} \times \bigcup_{Y \in \mathcal{Y}} Y.$$

By the axiom of choice, there exists a set Z having exactly one element in common with each element of \mathcal{Z} . Define the function

$$c : \mathcal{Z} \rightarrow \mathcal{Y} \times \bigcup_{Y \in \mathcal{Y}} Y$$

by $c(Y') = Y' \cap Z$. This function c implicitly provides the rule for a function from \mathcal{Y} to the set $\bigcup_{Y \in \mathcal{Y}} Y$ such that y belongs to Y whenever $(Y, y) \in Z$. This rule is the desired choice function. \square

A.1.2 Well-Ordered Sets

A **simple order** $<$ on a set X is a relation such that, for all $x, y, z \in X$,

1. if $x \neq y$ then either $x < y$ or $y < x$
2. if $x < y$ then $x \neq y$
3. if $x < y$ and $y < z$ then $x < z$.

Definition A.1.3. A set X with an order relation $<$ is said to be **well-ordered** if every nonempty subset of X has a smallest element.

The set of natural numbers, for example, is well-ordered. On the other hand, the set of integers is not well-ordered.

Fact A.1.4 (Well-ordering theorem). *If X is a set, there exists an order relation on X that is a well-ordering.*

This theorem was proved by Zermelo using the axiom of choice. It startled the mathematical community in 1904 and spurred much controversy about the axiom of choice. It is given here without a proof.

Corollary A.1.5. *There exists an uncountable well-ordered set.*

Definition A.1.6. *Let X be an ordered set. Given $x \in X$, the set*

$$Y_x = \{y \in Y \mid y < x\}$$

*is called the **section** of X by x .*

Corollary A.1.7. *There exists an uncountable well-ordered set, every section of which is countable.*

The well-ordering principle is a necessary tool in proofs by induction when the set over which the induction process is applied is not a segment of the natural numbers; this is the so-called transfinite induction.

A.1.3 The Maximum Principle

A **strict partial order** \prec on a set X is a relation such that for all $x, y, z \in X$

1. if $x \prec y$ then $x \neq y$
2. if $x \prec y$ and $y \prec z$ then $x \prec z$.

A strict partial order is similar to a simple order, except that it need not be true that for every distinct $x, y \in X$, either $x \prec y$ or $y \prec x$.

Fact A.1.8 (The maximum principle). *Let X be a set and suppose that \prec is a strict partial order on X . If Y is a subset of X that is simply ordered by \prec , then there exists a maximal simply ordered subset Z of X containing Y .*

The maximum principle is given here without a proof. It is interesting to note that the well-ordering theorem and the maximum principle are equivalent; either of

them implies the other. Furthermore, each of them is equivalent to the axiom of choice.

Let \prec be a strict partial order on X . For $x, y \in X$, the relation $x \preceq y$ holds if $x \prec y$ or $x = y$. The relation \preceq so defined is called a **partial order** on X . For example, the inclusion relation \subset on a collection of sets is a partial order, whereas proper inclusion is a strict partial order.

Bibliography

Index

- applications, 97
 - linear regression, 97
 - Wiener-Hopf, 99
- Banach space
 - strictly convex, 118
- eigenvalues, 159
 - algebraic multiplicity, 160
 - characteristic polynomial, 160
 - defective, 165
 - diagonalizable, 161
 - eigenspace, 160
 - eigenvalue, 159
 - eigenvector, 159
 - generalized eigenspace, 166
 - generalized eigenvector, 167
 - geometric multiplicity, 160
 - Jordan chain, 167
 - Jordan normal form, 166
 - similar, 162
 - spectrum, 159
- field, 47
- functions, 20
 - bijjective, 20
 - codomain, 20
 - concave, 118
 - convex, 118
 - domain, 20
 - global minimum value, 117
 - image, 20
 - injective, 20
 - inverse function, 20
 - inverse image, 20
 - one-to-one, 20
 - one-to-one correspondence, 20
 - onto, 20
 - preimage, 20
 - surjective, 20
- inner-product space, 71
 - adjoint, 142
 - best approximation, 85
 - Cauchy-Schwarz inequality, 75
 - dual approximation, 102
 - Euclidean space, 74
 - Gram-Schmidt orthogonalization, 78
 - half space, 107
 - Hilbert space, 81
 - induced norm, 74
 - inner product, 71
 - least-squares, 93
 - normal equations, 91
 - orthogonal, 74, 76, 80
 - orthogonal complement, 79
 - orthogonal set, 76

- orthonormal basis, 81
- Parseval identity, 81
- projection, 74
- Riesz representation theorem, 141
- standard inner product, 71
- unitary, 80
- integral, 65
 - almost everywhere, 64
 - Lebesgue integral, 64, 65
 - Lebesgue measure, 64
 - Riemann integral, 65
- linear transform, 58
 - algebra, 133
 - Banach algebra, 133
 - bounded, 138
 - coordinate matrix, 60
 - idempotent, 88
 - invertible, 133
 - linear operator, 132
 - non-singular, 59
 - nullity, 61
 - nullspace, 61
 - operator norm, 137
 - orthogonal projection, 89
 - projection, 88
 - pseudoinverse, 147
 - range, 60
 - rank, 61
 - singular, 59
 - transpose, 136
 - vector product, 133
- logic, 1
 - biconditional, 4
 - complete, 9
 - conditional connective, 3, 4
 - conjecture, 1
 - conjunction, 2
 - consistent, 9
 - contradiction, 5
 - contrapositive, 9
 - converse, 7
 - corollary, 12
 - decidable, 9
 - disjunction, 3
 - existential quantifier, 10
 - fallacy, 9
 - free variable, 10
 - implication relation, 4
 - lemma, 12
 - logical equivalence, 8
 - logical implication, 6
 - mathematical induction, 14
 - negation, 3
 - predicate, 10
 - proof, 1
 - proposition, 12
 - semidecidable, 12
 - tautology, 5
 - theorem, 12
 - universal instantiation, 10
 - universal quantifier, 10
- matrix
 - compact SVD, 175
 - convergent, 168
 - elementary column operation, 50
 - elementary row operation, 50

- Frobenius norm, 140
- Gramian, 92
- Hermitian transpose, 49
- inverse, 50
- invertible, 50
- matrix product, 49
- orthogonal, 158
- positive-definite, 92
- positive-semidefinite, 92
- projection matrix, 95
- pseudoinverse, 95, 146, 147
- reduced row echelon form, 50
- row echelon form, 49
- spectral radius, 140
- trace, 83
- transpose, 49
- unitary, 158
- matrix factorization, 149
 - backward substitution, 150
 - Cholesky factorization, 156
 - forward substitution, 149
 - LDLT decomposition, 156
 - lower triangular, 149
 - LU decomposition, 151
 - orthogonal, 157
 - unit triangular, 149
 - unitary, 157
 - upper triangular, 149
- metric space, 24
 - d -open ball, 26
 - boundary, 29
 - Cauchy sequence, 26
 - closed, 27
 - closure, 29
 - compact, 38
 - complete, 35
 - completion, 35
 - continuous, 29, 30
 - contraction, 36
 - converges, 26
 - dense, 35
 - distance, 26
 - Euclidean metric, 25
 - interior, 28
 - isolated point, 28
 - isometry, 35
 - limit, 30
 - limit point, 28
 - Lipschitz continuous, 30
 - metric, 25
 - open, 27
 - points, 26
 - pointwise convergence, 33
 - sequence, 26
 - totally bounded, 37
 - uniform convergence, 33
 - uniformly continuous, 30
- optimization
 - active, 122
 - convex, 128
 - feasible, 121
 - Fréchet derivative, 113
 - Fréchet differentiable, 113
 - Gâteaux differentiable, 113
 - Gâteaux differential, 112
 - gradient, 114
 - Jacobian matrix, 113

- Lagrange multiplier, 122
 - Lagrangian, 121
 - Lagrangian dual, 126
 - linear program, 121
 - local minimum value, 117
 - locally optimal, 122
 - objective function, 121
 - optimal value, 121
 - Slater's condition, 128
 - standard form, 120
 - strong duality, 127
 - weak duality, 126
- set theory
- axiom of choice, 177
 - cardinality, 16
 - Cartesian Product, 18
 - complement, 17
 - complex numbers, 15
 - countably infinite, 16
 - disjoint, 17
 - elements, 15
 - empty set, 15
 - equivalence classes, 19
 - equivalence relation, 18
 - infimum, 31
 - integers, 15
 - intersection, 17
 - maximum, 31
 - minimum, 31
 - naive set theory, 15
 - natural numbers, 15
 - partial order, 180
 - quotient set, 19
 - rational numbers, 16
 - real numbers, 15
 - Russell's Paradox, 16
 - set, 15
 - set difference, 17
 - set-builder notation, 15
 - simple order $<$, 178
 - singleton, 15
 - strict partial order, 179
 - subset, 17
 - supremum, 31
 - uncountably infinite, 16
 - union, 17
 - well-ordered, 178
- topology, 39
- basis, 39
 - closed, 40
 - closure, 41
 - continuous, 42
 - converge, 44
 - dense, 42
 - extended real numbers, 31
 - interior, 41
 - limit point, 41
 - metric topology, 40
 - metrizable, 40
 - neighborhood, 41
 - open set, 39
 - separable, 42, 70
- vector space, 51, 70
- affine hyperplane, 107
 - Banach space, 65
 - best approximation, 85

closed subspace, 68
convex set, 103
coordinate vector, 57
dimension, 56
direct sum, 53
dual basis, 135
dual space, 134
equivalent norms, 66
finite-dimensional, 54
functional, 117
Hamel basis, 54
Hilbert spaces, 70
homomorphism, 134
hyperplane, 106
isomorphism, 134
linear combination, 52
linear functional, 82
linear transform, 58
linearly dependent, 53
linearly independent, 53
norm, 62
normalized, 63
ordered basis, 57
reverse triangle inequality, 65
Schauder basis, 66
span, 53
standard basis, 54
standard Schauder basis, 66
subspace, 52
unit vector, 63