# Engineering Fundamentals

Major Contributors
Jean-Francois Chamberland and Henry D. Pfister

September 28, 2016

# Contents

# 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 the Zermelo-Fraenkel set theory, together with the axiom of choice. 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 statements 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. It is connected to the notion of logical implication. The **conditional** from $P$ to $Q$, denoted $P \to Q$, is mathematically true if it is not the case that $P$ is true and $Q$ is false. The precise definition of $P \to Q$ is given in the truth table

| $P$ | $Q$ | $P \to Q$ |
|---|---|---|
| T | T | T |
| T | F | F |
| F | T | T |
| F | F | T |

.

This table should match your intuition when $P$ is true. When $P$ is false, students often think the resulting truth value should be undefined. Although it may seem

strange at first glance, this truth table is universally accepted by mathematicians. To motivate this definition, one can think of $P \to Q$ as a promise that $Q$ is true whenever $P$ is true. When $P$ is false, the promise is kept by default. 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 $P \to Q$ 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,

$$(P \to R) \wedge (Q \vee \neg R). \tag{1.1}$$

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 | T | T | T | T | F |
| T | T | F | T | F | F | F | T | T | T |
| T | F | T | T | T | T | F | F | F | F |
| T | F | F | T | F | F | F | F | T | T |
| F | T | T | F | T | T | T | T | T | F |
| F | T | F | F | T | F | T | T | T | T |
| F | F | T | F | T | T | F | F | F | F |
| F | F | F | F | T | F | T | F | T | T |
| | | | 1 | 5 | 2 | 7 | 3 | 6 | 4 |

.

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 | T | F |
| F | F | T | T |
| | 1 | 3 | 2 |

.

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 | T | F | F |
| F | F | F | T |
| | 1 | 3 | 2 |

.

Of course, most statement 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.

A **logical implication** can be intuitively described as "$P$ implies $Q$" if $Q$ must be true whenever $P$ is true. That is, $Q$ cannot be false if $P$ is true. Necessity is the key aspect of this sentence, the fact that $P$ and $Q$ both happen to be true cannot be coincidental. To have $P$ implies $Q$, we need the conditional $P \to Q$ to be true under all possible circumstances.

Meta-statements, such as "$P$ implies Q", can be defined formally when $P$ and $Q$ are both logical functions of other propositions. For example, consider $P = R \wedge (R \to S)$ and $Q = S$. Then, the truth of the statement $P \to Q$ 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 $P \to Q$ is a tautology. We abbreviate $P$ implies $Q$ by writing $P \Rightarrow Q$. It is important to understand the difference between "$P \to Q$" and "$P \Rightarrow Q$." The former, $P \to Q$, 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 $P \to Q$ 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. $(P \to Q) \wedge P \Rightarrow Q$.

2. $(P \to Q) \wedge \neg Q \Rightarrow \neg P$.

3. $P \wedge Q \Rightarrow P$.

4. $(P \vee Q) \wedge \neg P \Rightarrow Q$.

5. $P \leftrightarrow Q \Rightarrow P \to Q$.

6. $(P \to Q) \wedge (Q \to P) \Rightarrow P \to Q$.

7. $(P \to Q) \wedge (Q \to R) \Rightarrow P \to R$

8. $(P \to Q) \wedge (R \to S) \wedge (P \vee R) \Rightarrow Q \vee S$.

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

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

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

| $P$ | $Q$ | $R$ | $((P$ | $\to$ | $Q)$ | $\wedge$ | $(Q$ | $\to$ | $R))$ | $\to$ | $(P$ | $\to$ | $R)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | T | T | T | T | T | T | T | T | T | T | T | T | T |
| T | T | F | T | T | T | F | T | F | F | T | T | F | F |
| T | F | T | T | F | F | F | F | T | T | T | T | T | T |
| T | F | F | T | F | F | F | F | T | F | T | T | F | F |
| F | T | T | F | T | T | T | T | T | T | T | F | T | T |
| F | T | F | F | T | T | F | T | F | F | T | F | T | F |
| F | F | T | F | T | F | T | F | T | T | T | F | T | T |
| F | F | F | F | T | F | T | F | T | F | T | 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 $Q \to P$ is the converse of $P \to Q$. This reversal may not preserve the truth of the statement though and therefore logical implications are not always reversible. For instance, although $(P \to Q) \wedge (Q \to R)$ implies

$P \to R$, the converse is not always true. It can easily be seen from columns 9 & 10 above that

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

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

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. $P \to Q \Leftrightarrow \neg P \vee Q$.
9. $P \to Q \Leftrightarrow \neg Q \to \neg P$ *(Contrapositive).*
10. $P \leftrightarrow Q \Leftrightarrow (P \to Q) \wedge (Q \to P)$.
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).*

The first equivalence in this list, $\neg(\neg P) \Leftrightarrow P$, may appear trivial. However, from the point of view of constructing mathematical proofs, this equivalence is frequently employed. Indeed, one method to prove that statement $P$ is true is to hypothesize that $\neg P$ is true and then derive a contradiction. It then follows that, if $\neg P$ is false, then $P$ is true. This popular technique is called **proof by contradiction**. It is illustrated below through a classic example.

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 $P \rightarrow Q \Leftrightarrow \neg Q \rightarrow \neg P$ is also used extensively in constructing mathematical proofs.

### 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

Quantifiers are of paramount importance in rigorous proofs. They are employed to make statements about collections of elements. Universal quantification is used to formalize the notion that a statement is true for all possible values of a collection. The **universal quantifier** is typically denoted by $\forall$ and it is informally read "for

all." Let $U$ be a specific collection of elements, and let $P(x)$ be a statement that applies to $x$. Then the statement $\forall x \in U, P(x)$ is true if $P(x)$ is true for all values of $x$ in $U$. The other type of quantifier often encountered 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 $U$. Based on the meanings of the quantifiers, we have the following equivalences

$$\neg\left((\forall x \in U)\, P(x)\right) \Leftrightarrow (\exists x \in U)\,(\neg P(x))$$
$$\neg\left((\exists x \in U)\, P(x)\right) \Leftrightarrow (\forall x \in U)\,(\neg P(x))\,.$$

In mathematics, a **free variable** is a notation for a place or places in an expression into which some definite substitution may take place, or with respect to which some operation (e.g. quantification) may take place. A **bound variable** is a variable for which we have no ability to choose the value.

## 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 that 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 $P \to Q$ 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 $P \to Q \Leftrightarrow \neg Q \to \neg P$. 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 $P \to Q$.

**Proof by Contradiction:** A proof by contradiction is based on the mathematical equivalence $\neg(P \to Q) \Leftrightarrow P \wedge \neg Q$. 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 $P \to Q$.

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

*First, assume that $\sqrt{2}$ is a rational number. This assumption implies 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 it can be expressed as an irreducible fraction." The contrapositive of this statement is "If $\sqrt{2}$ cannot be expressed as an irreducible fraction, then it 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.*

## 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 any collection of objects, mathematical or otherwise. For example, we 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.

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

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

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

$$\begin{aligned}\mathbb{N} &= \{n \in \mathbb{Z} | n \geq 1\} \\ \mathbb{Q} &= \{q \in \mathbb{R} | q = a/b, a \in \mathbb{Z}, b \in \mathbb{N}\}.\end{aligned}$$

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

$$\begin{aligned}\text{Open interval: } (a, b) &\triangleq \{x \in \mathbb{R} | a < x < b\} \\ \text{Closed interval: } [a, b] &\triangleq \{x \in \mathbb{R} | a \leq x \leq b\} \\ \text{Half-open intervals: } (a, b] &\triangleq \{x \in \mathbb{R} | a < x \leq b\} \\ [a, b) &\triangleq \{x \in \mathbb{R} | a \leq x < b\}\end{aligned}$$

**Example 1.4.1 (Russell's Paradox).** *Let $R$ be the set of all sets that do not contain themselves or $R = \{S | S \notin S\}$. The paradox arises from the fact that the definition leads to the logical contradiction $R \in R \Leftrightarrow R \notin R$. Axiomatic set theory eliminates this paradox by disallowing self-referencing constructions like this.*

There are also some standard relations defined between any two sets $A, B$.

**Definition 1.4.2.** *We say that $A$ **equals** $B$ (denoted $A = B$) if, for all $x$, $x \in A$ iff $x \in B$. This means that*

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

**Definition 1.4.3.** *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*

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

*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 1.4.4.** *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 1.4.5.** *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 1.4.6.** *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,

$$\bigcup_{\alpha \in I} S_\alpha = \{x | x \in S_\alpha \text{ for some } \alpha \in I\}$$
$$\bigcap_{\alpha \in I} S_\alpha = \{x | x \in S_\alpha \text{ for all } \alpha \in I\}.$$

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 1.4.7.** *The **Cartesian Product**, denoted $A \times B$, of two sets is the set of ordered pairs $\{(a,b)|a \in A, b \in B\}$. 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 1.4.8.** *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 $X^\omega$, is the set of infinite sequences $(x_1, x_2, x_3, \ldots)$ 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 $X^\infty$.

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 $E = \{(x, y) \in A \times A | P(x, y)\}$. 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 $[a] = \{x \in A | x \sim a\}$. 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 $A \backslash \sim = \{[a] | a \in A\}$.

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 1.4.9.** *Consider the set $\mathbb{N} \times \mathbb{N} = \{(a, b)|a, b \in \mathbb{N}\}$ 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 $\mathbb{N} \times \mathbb{N} \backslash \sim$ can therefore be associated with the set of reduced fractions.*

## 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** $f\colon X \to Y$ is a rule that assigns a single value $f(x) \in Y$ to each element $x \in X$. The notation $f\colon X \to Y$ 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** $f\colon X \to Y$ 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 $f\colon X \to Y$ has a unique **inverse function** $f^{-1}\colon Y \to X$ 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 $f\colon X \to Y$ can be transformed into a bijective function $g\colon X \to R$ 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 $f\colon X \to Y$ 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 $f\colon \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2$. Let $A = [1, 2]$ and $B = f(A) = [1, 4]$. Then,*

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

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

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

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

$(a)$ $x \in A \Rightarrow f(x) \in f(A)$ $\quad$ $(b)$ $y \in f(A) \Rightarrow \exists x \in A$ *s.t.* $f(x) = y$

$(c)$ $x \in f^{-1}(B) \Rightarrow f(x) \in B$ $\quad$ $(d)$ $f(x) \in B \Rightarrow x \in f^{-1}(B)$.

*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

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

and the definition of subset. The second result follows from

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

and the definition of subset.

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

$$(1) \quad f\left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} f\left(A_i\right) \qquad (2) \quad f\left(\bigcap_{i\in I} A_i\right) \subseteq \bigcap_{i\in I} f\left(A_i\right)$$

$$(3)\ f^{-1}\left(\bigcup_{i\in I} B_i\right) = \bigcup_{i\in I} f^{-1}\left(B_i\right) \qquad (4)\ f^{-1}\left(\bigcap_{i\in I} B_i\right) = \bigcap_{i\in I} f^{-1}\left(B_i\right).$$

# Chapter 2

# Metric Spaces and Topology

From an engineering perspective, the most important way to construct a topology on a set is to define the topology in terms of a metric on the set. 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 engineering problems. To gain better insight into metric spaces, we need to review the notion of a metric and to introduce a definition for topology.

## 2.1 Metric Spaces

A **metric space** is a set that has a well-defined "distance" between any two elements of the set. Mathematically, the notion of a metric space abstracts a few basic properties of Euclidean space. Formally, a metric space $(X, d)$ is a set $X$ and a function $d$ that is a metric on $X$.

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

$$d\colon X \times X \to \mathbb{R}$$

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

**Example 2.1.3.** *Given $\underline{x} = (x_1, \ldots, x_n), \underline{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, the **Euclidean metric** $d$ on $\mathbb{R}^n$ is defined by the equation*

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

*As implied by its name, the function $d$ defined above is a metric.*

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

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

*Show that $\rho$ is a metric.*

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

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

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

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 $B_d(x,\epsilon) = \{y \in X | d(x,y) < \epsilon\}$. This set is called the **$d$-open ball** (or open ball) of radius $\epsilon$ 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 $\delta$, 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 $x_1, x_2, \ldots$ where $x_i \in X$ for all $i \in \mathbb{N}$. Formally, a sequence is equivalent to a function $f\colon \mathbb{N} \to X$ where $x_i = f(i)$ for all $i \in \mathbb{N}$.*

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

**Definition 2.1.9.** *A sequence $x_1, x_2, \ldots$ in $(X, d)$ is a **Cauchy sequence** if, for any $\epsilon > 0$, there is a natural number $N$ (depending on $\epsilon$) such that, for all $m, n > N$,*

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

**Theorem 2.1.10.** *Every convergent sequence is a Cauchy sequence.*

*Proof.* Since $x_1, x_2, \ldots$ 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, $x_1, x_2, \ldots$ is a Cauchy sequence. □

### 2.1.1 Metric Topology

**Definition 2.1.11.** *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 2.1.12.** *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 if open*

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

One might be curious why only finite intersections are allowed in Theorem 2.1.12. The following example highlights the problem with allowing infinite intersections.

**Example 2.1.13.** *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*

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

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

**Definition 2.1.14.** *A subset $W$ of a metric space $(X, d)$ is closed if its complement $W^c = X - W$ is open.*

**Corollary 2.1.15.** *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 if closed*

*Sketch of proof.* Using the definition of closed, one can apply De Morgan's Laws to Theorem 2.1.12 verify this result. □

**Definition 2.1.16.** *For any metric space $(X, d)$ and subset $W \subseteq X$, a point $w \in W$ is in the **interior** of $W$ if there is a $\delta > 0$ such that $x \in W$ if $d(x, w) < \delta$.*

**Definition 2.1.17.** *For any metric space $(X, d)$ and subset $W \subseteq X$, a point $w \in W$ is a **limit point** of $W$ if there is a sequence $w_1, w_2, \ldots \in W$ of distinct elements that converges to $w$.*

**Definition 2.1.18.** *For any metric space $(X, d)$ and subset $W \subseteq X$, a point $x \in X$ is in the **closure** of $W$ if, for every $\delta > 0$, there is a $w \in W$ such that $d(x, w) < \delta$.*

The interior of $A$ is denoted by $A^\circ$ and the closure of $A$ is denoted by $\overline{A}$. Using Definition 2.1.11, it is easy to verify that $A^\circ$ is open. One can show that closure of $W$ is equal to the union of $W$ and its limit points. Thus, $\overline{A}$ is closed because a subset of a metric space is closed if and only if it contains all of its limit points.

### 2.1.2 Continuity

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

**Definition 2.1.19.** *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 $d_X(x_0, x) < \delta$,*

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

*In precise mathematical notation, one has*

$$(\forall\, \epsilon > 0)(\exists\, \delta > 0),$$
$$((x \in X) \wedge (d_X(x_0, x) < \delta)) \to d_Y\left(f(x_0), f(x)\right) < \epsilon.$$

**Theorem 2.1.20.** *If $f$ is continuous at $x_0$, then $f(x_n) \to f(x_0)$ for all sequences $x_1, x_2, \ldots \in X$ such that $x_n \to x_0$. Conversely, if $f(x_n) \to f(x_0)$ for all sequences $x_1, x_2, \ldots \in X$ such that $x_n \to x_0$, 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 $x_n \to x_0$, 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 $f(x_n) \to f(x_0)$.

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 $\epsilon$ and any positive sequence $\delta_n \to 0$, let $x_n$ be the promised $x$. Then, $x_n \to x_0$ because $d_X(x_0, x_n) < \delta_n \to 0$ 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 $x_n \to x_0$. □

**Definition 2.1.21.** *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*

$$(\forall\, x_0 \in X)(\forall\, \epsilon > 0)(\exists\, \delta > 0),$$
$$((x \in X) \wedge (d_X(x_0, x) < \delta)) \to d_Y\,(f(x_0), f(x)) < \epsilon.$$

**Definition 2.1.22.** *The function $f$ is called **uniformly continuous** if it continuous and, for all $\epsilon > 0$, the $\delta > 0$ can be chosen independently of $x_0$. In precise mathematical notation, one has*

$$(\forall\, \epsilon > 0)(\exists\, \delta > 0)(\forall\, x_0 \in X),$$
$$((x \in X) \wedge (d_X(x_0, x) < \delta)) \to d_Y\,(f(x_0), f(x)) < \epsilon.$$

**Definition 2.1.23.** *A function $f\colon X \to Y$ is called **Lipschitz continuous** on $A \subseteq X$ if there is a constant $L \in \mathbb{R}$ such that $d_Y(f(x), f(y)) \leq L d_X(x, y)$ for all $x, y \in A$.*

Let $f_A$ denote the **restriction** of $f$ to $A \subseteq X$ defined by $f_A\colon A \to Y$ 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.

### 2.1.3 Completeness

Suppose $(X, d)$ is a metric space. From Definition 2.1.8, we know that a sequence $x_1, x_2, \ldots$ 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 $X$.

**Example 2.1.24.** *Let $C[-1,1]$ be the vector space of continuous functions on the interval $[-1,1]$ and consider the $L^2$ norm*

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

*Define the sequence of functions $f_n(t)$ given by*

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

*Assuming that $m \geq n$, we get*

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

*This sequence satisfies the Cauchy criterion, but it does not converge to a continuous function in $C[-1,1]$.*

**Definition 2.1.25.** *A metric space $(X,d)$ is said to be **complete** if every Cauchy sequence in $X$ converges to a limit $x \in X$.*

**Example 2.1.26.** *Any closed subset of $\mathbb{R}^n$ (or $\mathbb{C}^n$) is complete.*

**Example 2.1.27.** *Consider the sequence $x_n \in \mathbb{Q}$ defined by $x_n = \left(1 + \frac{1}{n}\right)^n$. It is well-known that this sequence converges to $e \in \mathbb{R}$, but this number is not rational. Therefore, the rational numbers $\mathbb{Q}$ are not complete.*

**Theorem 2.1.28.** *A closed subset $A$ of a complete metric space $X$ is itself a complete metric space.*

**Definition 2.1.29.** *An **isometry** is a mapping $\phi\colon X \to Y$ 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 2.1.30.** *A subset $A$ of a 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 2.1.31.** *The **completion** of a metric space $X$ consists of a complete metric space $X'$ and an isometry $\phi\colon X \to X'$ such that $\phi(X)$ is a dense subset of $X'$. Moreover, the completion is unique up to isometry.*

**Example 2.1.32.** *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 2.1.33.** *Two Cauchy sequences $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ 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 2.1.34.** *Let $\mathcal{C}(\mathbb{Q})$ denote the set of all Cauchy sequences $q_1, q_2, \ldots$ of rational numbers where $\sim$ represents the equivalence relation on this set defined above. Then, the set of equivalence classes (or quotient set) $\mathcal{C}(\mathbb{Q})\backslash\sim$ 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 2.1.35.** *Let $A$ be a subset of a metric space $(X, d)$ and $f\colon X \to X$ be a function such that $f(A) \subseteq A$. The function $f$ is a **contraction** on $A$ if there exists a constant $\gamma < 1$ such that $d\left(f(x), f(y)\right) \leq \gamma d(x, y)$ for all $x, y \in A$.*

Consider the following important results in applied mathematics: Picard's theorem for differential equations, the implicit function theorem, and Bellman's principle of optimality 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 2.1.36** (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, $d(y,z) = d\left(f(y), f(z)\right) \leq \gamma d(y,z)$ 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.

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

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

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 $x_n \to x^*$ 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 $A$.

Arguments similar to the above can be used to prove the stated error bounds. □

**Example 2.1.37.** *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* $x_{n+1} = \cos(x_n)$ *converges to the unique fixed point* $x^* = \cos(x^*)$ *for all* $x_1 \in [0,1]$.

### 2.1.4 Compactness

**Definition 2.1.38.** *A metric space* $X$ *is* ***totally bounded*** *if, for any* $\epsilon > 0$, *there exists a finite set of* $B(x, \epsilon)$ *balls that cover (i.e., whose union equals)* $X$.

**Definition 2.1.39.** *A metric space* $X$ *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 2.1.40.** *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 2.1.41.** *Let $X$ be a compact space and $x_1, x_2, \ldots \in X$ be a sequence. Then, there is a subsequence $x_{n_1}, x_{n_2}, \ldots$, defined by some increasing sequence $n_1, n_2, \ldots \in \mathbb{N}$, that converges.*

*Proof.* Suppose for some $x \in X$, it holds that, for any $\epsilon > 0$, the set $B(x, \epsilon)$ contains infinitely many elements in the sequence $x_1, x_2, \ldots$. In this case, we can choose $n_1, n_2, \ldots$ so that $|x - x_{n_j}| < \frac{1}{j}$. Therefore, the subsequence converges to $x$.

On the other hand, suppose the sequence has no convergent subsequence. Then, for every $x \in X$, there is an $\epsilon > 0$ such that $B(x, \epsilon)$ does not contain infinitely many elements in the sequence. Since $X$ is totally bounded, we also find that, for any $\epsilon > 0$, $X$ is covered by a finite set of balls of radius $\epsilon$. But, for each $\epsilon > 0$, this gives a contradiction because all the elements in the infinite sequence must fall into a finite number balls each containing finitely many elements. By contradiction, this implies that there is a subsequence that converges. □

Functions from compact sets to the real numbers are very important in practice. To keep the discussion self-contained, we first review the extreme values of sets of real numbers. First, we must define the **extended real numbers** $\overline{\mathbb{R}}$ by augmenting the real numbers to include limit points for unbounded sequences $\overline{\mathbb{R}} \triangleq \mathbb{R} \cup \{\infty, -\infty\}$. Surprisingly, it turns out that $\overline{\mathbb{R}}$ is a compact metrizable space.

**Definition 2.1.42.** *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 2.1.43.** *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 2.1.44.** *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 $\infty$ if $X = \emptyset$.*

**Definition 2.1.45.** *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 2.1.46.** *Let $X$ be a metric space and $f\colon X \to \mathbb{R}$ 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 $x_1, x_2, \ldots \in A$ such that $\lim_n f(x_n) = M$.*

*Proof.* If $M = \infty$, then $A$ 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 $f(x_{n+1}) > f(x_n) + 1$.

If $M < \infty$, then $A$ 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 $x_1, x_2, \ldots$ by choosing $x_n \in A$ to be any point that satisfies $M - f(x_n) \leq \frac{1}{n}$. □

**Theorem 2.1.47.** *Let $X$ be a metric space and $f\colon X \to \mathbb{R}$ 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 2.1.46, one finds that there is a sequence $x_1, x_2, \ldots \in A$ such that $\lim_n f(x_n) = \sup f(A)$. Since $A$ is compact, there must also be a subsequence $x_{n_1}, x_{n_2}, \ldots$ that converges. As $A$ is closed, this subsequence must converge to some $x^* \in A$. Finally, the continuity of $f$ shows that

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

□

**Corollary 2.1.48.** *A continuous function from a compact subset $A$, of a metric space $X$, to the real numbers achieves a minimum on $A$.*

**Theorem 2.1.49.** *Any bounded non-decreasing sequence of real numbers converges to its supremum.*

*Proof.* Let $x_1, x_2, \ldots \in \mathbb{R}$ be a sequence satisfying $x_{n+1} \geq x_n$ and $x_n \leq M < \infty$ for all $n \in \mathbb{N}$. Without loss of generality, we can take the upper bound $M$ to be the supremum $\sup\{x_1, x_2, \ldots\}$. This sequence converges to $M$ if, for any $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $M - x_n < \epsilon$ for all $n > N$.

To prove this by contradiction, we suppose that $x_n$ does not converge to $M$. Since $x_n$ is non-decreasing, this implies that there is an $\epsilon > 0$ such that $M - x_n \geq \epsilon$ for all $n \in \mathbb{N}$. But, this result contradicts $\sup f(A) = M$. Therefore, the sequence converges to $M$. □

### 2.1.5 Sequences of Functions

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

**Definition 2.1.50.** *The sequence $f_n$ **converges pointwise** to $f \colon X \to Y$ if*

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

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

$$\forall x \in X, \forall \epsilon > 0, \exists N \in \mathbb{N},\ (n > N) \to \left(d_Y\left(f_n(x), f(x)\right) < \epsilon\right).$$

**Definition 2.1.51.** *The sequence $f_n$ **converges uniformly** to $f \colon X \to Y$ if*

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall x \in X,\ (n > N) \to \left(d_Y\left(f_n(x), f(x)\right) < \epsilon\right).$$

*This condition is also equivalent to*

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

**Theorem 2.1.52.** *If each $f_n$ is continuous and $f_n$ converges uniformly to $f \colon X \to Y$, 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 $f_n \to f$ 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

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

□

## 2.2 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 $\epsilon$-balls*

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

*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 2.1.11.*

**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 2.1.3 and 2.1.4 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 $f : [0, 1] \to \mathbb{R}$ 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*

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

*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 $A^\circ$ and the closure of $A$ is denoted by $\overline{A}$. We note that $A^\circ$ is open and $\overline{A}$ is closed. Furthermore, $A^\circ \subseteq A \subseteq \overline{A}$.

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

*Proof.* We prove instead the equivalent contrapositive statement: $x \notin \overline{A}$ if and only if there is an open set $B$ containing $x$ that does not intersect $A$. Clearly, if $x \notin \overline{A}$, then $\overline{A}^c = X - \overline{A}$ 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 $\overline{A}$. But $x \notin B^c$, so $x \notin \overline{A}$. □

**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 $f\colon X \to Y$ 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 $\{x \in X | f(x) \in B\}$. 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 $f\colon X \to Y$. The following are equivalent:*

1. *$f$ is continuous*
2. *for every subset $A \subseteq X$, $f\left(\overline{A}\right) \subseteq \overline{f(A)}$*
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. Suppose $x \in \overline{A}$, where $A$ is a subset of $X$. Let $O$ be a neighborhood of $f(x)$. Since $f^{-1}(O)$ is an open set of $X$ containing $x \in \overline{A}$, it must intersect with $A$ in some point $x'$. It follows that $O$ intersects $f(A)$ in the point $f(x')$. By Theorem 2.2.9, we find that $f(x) \in \overline{f(A)}$.

$(2 \Rightarrow 3)$. Suppose that for every subset $A \subseteq X$, $f\left(\overline{A}\right) \subseteq \overline{f(A)}$. Let $C \subseteq Y$ be a closed set and let $A = f^{-1}(C)$. By P 1.5.4, we have $f(A) \subseteq C$. If $x \in \overline{A}$, then

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

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 $f\colon X \to Y$. 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

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

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 $d_Y\left(f(x_1), f(x_2)\right) < \epsilon$ 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 exits a $d$-open ball $B_{d_X}(x, \delta)$ of radius $\delta > 0$ centered at $x$ such that $f\left(B_{d_X}(x, \delta)\right) \subset B_{d_Y}\left(f(x), \epsilon\right)$. Therefore, every $x \in f^{-1}(O_y)$ has a neighborhood in the same set, and that implies $f^{-1}(O_y)$ is open. □

**Definition 2.2.19.** *A sequence $x_1, x_2, \ldots$ 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 $x_n \to x$, 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 $B_d\left(x, \frac{1}{n}\right)$. Since $x \in \overline{A}$, the set $A \cap B_d\left(x, \frac{1}{n}\right)$ is not empty and we choose $x_n$ to be any point in this set. It follows that the sequence $x_1, x_2, \ldots$ converges to $x$. Notice that the "only if" proof holds for any topological space, while "if" requires a metric. □

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

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 $x_1, x_2, \ldots$ of points of $A$ converging to $x$. By assumption, $f(x_n) \to f(x)$. Since $f(x_n) \in f(A)$, Theorem 2.2.21 implies that $f(x) \in \overline{f(A)}$. Hence $f\left(\overline{A}\right) \subseteq \overline{f(A)}$ and $f$ is continuous. □

# Chapter 3

# Linear Algebra

## 3.1 Fields

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 $(s+t) \in F$. For every pair of elements $s, t \in F$ then $st \in F$. 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 $F = \left\{s + t\sqrt{2} : s, t \in \mathbb{Q}\right\}$ together with the standard addition and multiplication is a field. Let $s, t, u, v \in \mathbb{Q}$,*

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

*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 $x_1, \ldots, x_n$ which satisfy the conditions

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

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

$$A\underline{x} = \underline{y},$$

where $A$ is the matrix representation given by

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

and $\underline{x}$, $\underline{y}$ denote

$$\underline{x} = (x_1, \ldots, x_n)^T$$
$$\underline{y} = (y_1, \ldots, y_m)^T.$$

**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*

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

**Definition 3.2.2.** *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 $A^{-1}$.*

**Lemma 3.2.3.** *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.* This result follows easily from row reduction and the proof is left as an exercise for the reader. □

## 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* $\underline{v} = (v_1, \ldots, v_n)$ *of scalar* $v_i \in F$*. If* $\underline{w} = (w_1, \ldots, w_n)$ *with* $w_i \in F$*, the sum of* $\underline{v}$ *and* $\underline{w}$ *is defined by*

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

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

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

*The set of* $n$*-tuples, denoted by* $F^n$*, together with the vector addition and scalar product defined above forms a vector space.*

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

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

*The product of scalar* $s \in F$ *and the function* $f \in V$ *is the function* $sf$ *defined by*

$$(sf)(x) = sf(x)\ \forall x \in X.$$

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

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

**Definition 3.3.5.** *Consider a complex* $n \times n$ *matrix* $A$ *with elements* $a_{ij}$*. The* ***Hermitian transpose*** $B = A^H$ *of* $A$ *has elements defined by* $b_{ij} = \overline{a_{ji}}$ *where* $\overline{a}$ *denotes the complex conjugate of* $a$*.*

.

### 3.3.1 Subspaces

**Definition 3.3.6.** *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 3.3.7.** *A non-empty subset $W \subset V$ is a subspace of $V$ if and only if for every pair of vector $\underline{w}_1, \underline{w}_2 \in W$ 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 3.3.8.** *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 3.3.9.** *Let $U$ be a set of vectors in a vector space $V$. The **span** of $U$, denoted $\mathrm{span}(U)$, is defined to be the set of all finite linear combinations of vectors in $U$.*

The subspace spanned by $U$ can also be defined equivalently as the intersection of all subspaces of $V$ which contain $U$. It is easy to see that all finite linear combinations of vectors in $U$ must be in all subspaces containing $U$. One can show the converse by considering its contrapositive: There exists a finite linear combination of vectors in $U$ and a subspace containing $U$ such that the vector is not in the subspace. Of course, this contradicts the definition of a vector space.

**Definition 3.3.10.** *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*

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

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 Dimension

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

**Definition 3.3.11.** *Let $V$ be a vector space over $F$. A subset $U \subset V$ is **linearly dependent** if there exist distinct vectors $\underline{u}_1, \ldots, \underline{u}_n \in U$ and scalars $s_1, \ldots, s_n \in F$, not all of which are $0$, such that*

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

*A set which is not linearly dependent 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 3.3.12.** *Let $V$ be a vector space over $F$. Let $\mathcal{B} = \{\underline{v}_\alpha | \alpha \in A\}$ be a subset of linearly independent vectors from $V$, such that every $\underline{v} \in V$ has a unique decomposition 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.*

**Theorem 3.3.13.** *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 $V$. □

**Example 3.3.14.** *Let $F$ be a field and let $U \subset F^n$ be the subset consisting of the vectors $\underline{e}_1, \ldots, \underline{e}_n$ defined by*

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

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

$$\underline{v} = \sum_{i=1}^{n} v_i \underline{e}_i. \tag{3.2}$$

*Thus, the collection $U = \{\underline{e}_1, \ldots, \underline{e}_n\}$ spans $F^n$. Since $\underline{v} = \underline{0}$ in (3.2) if and only if $v_1 = \cdots = v_n = 0$, $U$ is linearly independent. Accordingly, the set $U$ is a basis for $F^{n\times 1}$. This basis is termed the **standard basis** of $F^n$.*

**Example 3.3.15.** *Let $A$ be an invertible matrix over $F$. The columns of $A$, denoted by $A_1, \ldots, A_n$, form a basis for the space of column vectors $F^{n\times 1}$. This can be seen as follows. If $\underline{v} = (v_1, \ldots, v_n)^T$ is a column vector, then*

$$A\underline{v} = \sum_{i=1}^{n} v_i A_i.$$

*Since $A$ is invertible,*

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

*That is, $\{A_1, \ldots, A_n\}$ is a linearly independent set. 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, as a consequence, $\{A_1, \ldots, A_n\}$ is a basis for $F^n$.*

**Theorem 3.3.16.** *Let $V$ be a vector space which is spanned by a finite set of vectors $W = \{\underline{w}_1, \ldots, \underline{w}_n\}$. Then, any linearly independent set of vectors in $V$ is finite and contains no more than $n$ elements.*

*Proof.* Assume that $U = \{\underline{u}_1, \ldots, \underline{u}_m\} \subset V$ with $m > n$. 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 $s_1, \ldots, s_m$ 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 $n < m$, Lemma 3.2.3 implies that there exist scalars $s_1, \ldots, s_n$, not all $0$, such that $t_1 = t_2 = \cdots = t_m = 0$. For these scalars, $\sum_{j=1}^{m} s_j \underline{u}_j = 0$. That is, the set $U$ is linearly dependent. □

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. While this may seem intuitively obvious to many readers, the previous theorem shows that this intuition from $\mathbb{R}^n$ does not break down for other fields and vector spaces.

**Definition 3.3.17.** *The **dimension** of a finite-dimensional vector space is defined as 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 argue that the empty set $\emptyset$ spans $\{\underline{0}\}$ because the intersection of all the subspaces containing the empty set is $\{0\}$. Though this is only a minor point.

**Theorem 3.3.18.** *Let $A$ be an $n \times n$ matrix over $F$ and suppose that the columns of $A$, denoted by $A_1, \ldots, A_n$, form a linearly independent set of vectors in $F^{n\times 1}$. Then $A$ is invertible.*

*Proof.* Suppose that $W$ is the subspace of $F^{n\times 1}$ spanned by $A_1, \ldots, A_n$. Since $A_1, \ldots, A_n$ are linearly independent, $\dim(W) = n = \dim(F^{n\times 1})$. It follows that $W = V$ and, as a consequence, there exist scalars $b_{ij} \in F$ such that

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

where $\{\underline{e}_1, \ldots, \underline{e}_n\}$ is the standard basis for $F^{n\times 1}$. Then, for the matrix $B$ with entries $b_{ij}$, we have

$$AB = I.$$

Note also that if the rows of $A$ form a linearly independent set of vectors in $F^{1\times n}$ then $A$ is invertible. □

### 3.3.3 Coordinate System

Let $\{\underline{v}_1, \ldots, \underline{v}_n\}$ be a basis for the $n$-dimensional vector space $V$. 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., $\underline{v}_1, \ldots, \underline{v}_n$) allows the first element in the coordinate vector to be associated with the first vector in our basis and so on.

**Definition 3.3.19.** *If $V$ is a finite-dimensional vector space, an **ordered basis** for $V$ is a finite sequence of vectors which is linearly independent and spans $V$.*

In particular, if the sequence $\underline{v}_1, \ldots, \underline{v}_n$ is an ordered basis for $V$, then the set $\{\underline{v}_1, \ldots, \underline{v}_n\}$ is a basis for $V$. The ordered basis $\mathcal{B}$ is the set $\{\underline{v}_1, \ldots, \underline{v}_n\}$, together with the specific ordering of the vectors. Based on this ordered basis, a vector $\underline{w} \in V$ can be unambiguously represented as an $n$-tuple,

$$\underline{w} = (s_1, \ldots, s_n) = \sum_{i=1}^{n} s_i \underline{v}_i.$$

Equivalently, vector $\underline{w}$ can be described using the *coordinate matrix of $\underline{w}$ relative to the ordered basis $\mathcal{B}$*:

$$\underline{w} = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix}.$$

The dependence of this coordinate matrix on $\mathcal{B}$ can be specified explicitly using the notation $[\underline{w}]_{\mathcal{B}}$. This will be particularly important when multiple coordinates 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: $\underline{e}_1, \ldots, \underline{e}_n$. Vectors in $F^n$ can therefore be unambiguously expressed as $n$-tuples.*

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

$$[\underline{u}]_{\mathcal{A}} = P\,[\underline{u}]_{\mathcal{B}}$$
$$[\underline{u}]_{\mathcal{B}} = P^{-1}\,[\underline{u}]_{\mathcal{A}}$$

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

**S 3.3.21**. Consider the ordered basis $\mathcal{A} = \underline{v}_1, \ldots, \underline{v}_n$ 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

$$\underline{w}_i = \sum_{k=1}^{n} p_{ki} \underline{v}_k,$$

and

$$t_i = \sum_{i=1} q_{ij} s_j,$$

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

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

*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 $T\left(\underline{v}\right) = A\underline{v}$ 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,*

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

**Example 3.4.5.** *Let $V$ be the space of continuous functions from $\mathbb{R}$ 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 $T\left(\underline{0}\right) = \underline{0}$. This is essential since

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

**Definition 3.4.6.** *A linear transformation $T\colon V \to W$ 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

**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 $G\colon \mathcal{B} \to W$, there is a unique linear transformation $T\colon V \to W$ such that $T\underline{v}_\alpha = G(\underline{v}_\alpha)$.*

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

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

where $s_\alpha \neq 0$ for only a finite subset of $A$. Using this expansion and the mapping $G$ for basis vectors, we define the mapping $T$, for each $\underline{v}$, as

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

Using the unique decomposition of vectors, it is easy to verify that this mapping is linear.

To show that it is unique, we assume that there is another linear mapping $U\colon V \to W$ such that $U\underline{v}_\alpha = G(\underline{v}_\alpha)$. In this case, the linearity guarantees that

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

This implies that $U\underline{v} = T\underline{v}$ for all $\underline{v} \in V$ and therefore that $U = T$. □

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

**Definition 3.4.8.** *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*

$$\mathcal{R}(T) \triangleq \{\underline{w} \in W | \exists \underline{v} \in V \text{ s.t. } T\underline{v} = \underline{w}\} = \{T\underline{v} | \underline{v} \in V\}.$$

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 3.4.9.** *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*

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

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

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

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

**Definition 3.4.10.** *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 3.4.11.** *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 $\underline{v}_1, \ldots, \underline{v}_k$ be a basis for $\mathcal{N}(T)$, the nullspace of $T$. There are vectors $\underline{v}_{k+1}, \ldots, \underline{v}_n \in V$ such that $\underline{v}_1, \ldots, \underline{v}_n$ is a basis for $V$. We want to show that $T\underline{v}_{k+1}, \ldots, T\underline{v}_n$ is a basis for the range of $T$. The vectors $T\underline{v}_1, \ldots, T\underline{v}_n$ certainly span $\mathcal{R}(T)$ and, since $T\underline{v}_j = \underline{0}$ for $j = 1, \ldots, k$, it follows that $T\underline{v}_{k+1}, \ldots, \underline{v}_n$ span $\mathcal{R}(T)$. Suppose that there exist scalars $s_{k+1}, \ldots, s_n$ 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 $\underline{v}_1, \ldots, \underline{v}_k$ 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 $\underline{v}_1, \ldots, \underline{v}_n$ are linearly independent, this implies that

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

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

**Corollary 3.4.12.** *If $A$ is an $m \times n$ matrix with entries in the field $F$, then*

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

*Proof.* First, we note that the range $\mathcal{R}(A)$ of $A$ equals the column space of $A$ and therefore $\text{rank}(A) = \dim(\mathcal{R}(A))$. Next, we use the definition of the orthogonal complement and $\mathcal{R}(A^H) = \{A^H \underline{y} | \underline{y} \in F^m\}$ to see that

$$\mathcal{R}(A^H)^\perp = \left\{\underline{x} \in F^n \,|\, \underline{x}^H \underline{z} = \underline{0} \;\forall\, \underline{z} \in \mathcal{R}(A^H\right\} = \left\{\underline{x} \in F^n \,|\, \underline{x}^H A^H \underline{y} = 0 \;\forall\, \underline{y} \in F^m\right\}.$$

This implies that $\mathcal{R}(A^H)^\perp = \left\{\underline{x} \in F^n \,|\, \underline{x}^H A^H = 0\right\} = \mathcal{N}(A)$ and therefore that

$$\dim\left(\mathcal{R}(A^H)\right) = n - \dim\left(\mathcal{R}(A^H)\right)^\perp = n - \dim\left(\mathcal{N}(A)\right) = \dim\left(\mathcal{R}(A)\right).$$

□

## 3.5 Norms

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

A **norm** on vector space $V$ is a real-valued function $\|\cdot\| : V \to \mathbb{R}$ 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 the concept of a metric. For instance, a metric can be defined in terms of a 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.

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 3.5.1.** *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 $f : \mathbb{R}^n \to \mathbb{R}$ 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 $f : \mathbb{R}^n \to \mathbb{R}$ defined by*

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

*is not a norm.*

**Example 3.5.2.** *The following functions are examples of norms for $F^n$,*

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

3. *the $l^\infty$ norm:* $\|\underline{v}\|_\infty = \max_{1,\ldots,n}\{|v_i|\}$.

**Example 3.5.3.** *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:* $\|f(t)\|_p = \left(\int_a^b |f(t)|^p dt\right)^{\frac{1}{p}}, \quad p \in (1,\infty)$
3. *the $L^\infty$ norm:* $\|f(t)\|_\infty = \text{esssup}_{[a,b]}\{|f(t)|\}$.

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

## What is the Lebesgue Integral?

Many important spaces include functions that are not Riemann integrable. The Lebesgue integral 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 norm of a function is zero if and only if it is 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*.

Consider the set of all functions $f\colon X \to \mathbb{R}$ from $X$ to the real numbers. The normed vector space $L^p(X)$ (with $1 \leq p < \infty$) is the subset 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?". For this course, we will be satisfied with the following definition:

**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 $L^p(X)$ by considering sequences $f_1, f_2, \ldots$ of "simple" functions that are constructed by rounding the function values down to a finite set of possibilities. 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 can define the Lebesgue integral of $f(x)$ to be the limit of this sequence. In fact, the elements of $L^p(X)$ are in one-to-one correspondence with the limits of non-decreasing sequences of simple functions that satisfy $\|f_n\|_{L^p} \to M < \infty$.*

**Definition 3.5.5.** *The **Lebesgue measure** of a set is equal to the Lebesgue integral of its indicator function when both quantites exist. In particular, the set is measurable iff the Lebesgue integral of its indicator function exists.*

**Definition 3.5.6.** *A vector $\underline{v} \in V$ is said to be **normalized** if $\|\underline{v}\| = 1$. Any vector can be normalized, except the zero vector:*

$$\underline{u} = \frac{\underline{v}}{\|\underline{v}\|} \tag{3.3}$$

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

**Definition 3.5.7.** *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 3.5.8.** *The vector spaces $\mathbb{R}^n$ (or $\mathbb{C}^n$) with any well-defined norm are Banach spaces.*

**Example 3.5.9.** *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 3.5.10.** *A Banach space $V$ has a **Schauder basis**, $\underline{v}_1, \underline{v}_2, \ldots$, if every $\underline{v} \in V$ can be written uniquely as*

$$\underline{v} = \sum_{i\in\mathbb{N}} s_i \underline{v}_i.$$

**Example 3.5.11.** *Let $V = \mathbb{R}^\omega$ be the vector space of semi-infinite real sequences. The **standard Schauder basis** is the countably infinite extension $\{\underline{e}_1, \underline{e}_2, \ldots\}$ of the standard basis.*

**Definition 3.5.12.** *A **closed subspace** of a Banach space is a subspace that is a closed set in the topology generated by the norm.*

**Theorem 3.5.13.** *All finite dimensional subspaces of a Banach space are closed.*

*Proof.* This proof requires material from later in the notes, but is given here for completeness. Let $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n$ be a basis for a finite dimensional subspace $W$ of a Banach space $V$ over $F$. Let $U = F^n$ be the standard Banach space, which is closed by definition, and consider the mapping $f : U \to W$ defined by

$$f(\underline{s}) = \sum_{i=1}^{n} s_i \underline{w}_i.$$

It is easy to verify that this linear mapping is non-singular and onto. Therefore, it has a linear inverse mapping $g = f^{-1}$ that must be continuous (i.e., bounded) because $U, W$ are finite dimensional. Since $g$ is continuous, we find that $W = g^{-1}(U) = f(U)$ is closed because $U$ is closed. □

**Example 3.5.14.** *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. More generally, the span of any infinite set of linearly independent vectors only includes finite linear combinations and is therefore not closed.*

## 3.6 Inner Products

**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* $\underline{v} = \underline{0}$.

Note that the conditions of Definition 3.6.1 imply that

$$\langle \underline{u} | s\underline{v} + \underline{w} \rangle = \overline{s} \langle \underline{u} | \underline{v} \rangle + \langle \underline{u} | \underline{w} \rangle .$$

**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*

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

*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*

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

*is an inner product.*

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

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

Also, we have

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

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

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

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 $\langle \cdot | \cdot \rangle$ is an inner product on $W$. If $T$ is a non-singular linear transformation from $V$ into $W$, then the equation*

$$p_T (\underline{v}_1, \underline{v}_2) = \langle T\underline{v}_1 | T\underline{v}_2 \rangle$$

*defines an inner product $p_T$ on $V$.*

**Theorem 3.6.7.** *Let $V$ be a finite-dimensional space, and suppose that*

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

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

$$h_{ji} = \langle \underline{w}_i | \underline{w}_j \rangle$$

*it assumes on pairs of vectors in $\mathcal{B}$.*

*Proof.* If $\underline{u} = \sum_i s_i \underline{w}_i$ and $\underline{v} = \sum_j t_j \underline{w}_j$, then

$$\begin{aligned}
\langle \underline{u} | \underline{v} \rangle &= \left\langle \sum_i s_i \underline{w}_i \middle| \underline{v} \right\rangle = \sum_i s_i \langle \underline{w}_i | \underline{v} \rangle \\
&= \sum_i s_i \left\langle \underline{w}_i \middle| \sum_j t_j \underline{w}_j \right\rangle = \sum_i \sum_j s_i \bar{t}_j \langle \underline{w}_i | \underline{w}_j \rangle \\
&= \sum_i \sum_j \bar{t}_j h_{ji} s_i = [\underline{v}]_{\mathcal{B}}^H \, H \, [\underline{u}]_{\mathcal{B}}
\end{aligned}$$

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 $H$ is called the *matrix of the inner product in the ordered basis $\mathcal{B}$.* □

It is easily verified that $H$ is a Hermitian matrix, i.e., $H = H^H$. Furthermore, $H$ must satisfy the additional condition

$$\underline{w}^H H \underline{w} > 0, \quad \forall \underline{w} \neq \underline{0}. \tag{3.4}$$

In particular, $H$ must be invertible.

Conversely if $H$ is an $n \times n$ Hermitian matrix over $F$ which satisfies (3.4), then $H$ is the matrix in the ordered basis $\mathcal{B}$ of an inner product on $V$. This inner product is given by

$$\langle \underline{u} | \underline{v} \rangle = [\underline{v}]_{\mathcal{B}}^H \, H \, [\underline{u}]_{\mathcal{B}} \, .$$

**Problem 3.6.8.** *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 3.6.9.** *Consider any set $W$ of real-valued random variables, defined on a common probability space, that have finite 2nd moments. It turns out that $V = \mathrm{span}(W)$ is a vector space over $\mathbb{R}$. In fact, one can define the inner product*

$$\langle X|Y\rangle = \mathrm{E}\left[XY\right],$$

*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 $\|X - Y\|^2 = E\left[|X - Y|^2\right] = 0$ (or equivalently $\Pr(X \neq Y) = 0$).*

### 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 3.6.10.** *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**,*

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

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

**Theorem 3.6.11.** *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 items follow immediately from the definitions involved. The third inequality, $|\langle \underline{v}|\underline{w}\rangle| \leq \|\underline{v}\|\,\|\underline{w}\|$, 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

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

Then $\langle \underline{u}|\underline{v}\rangle = 0$ and

$$\begin{aligned} 0 \leq \|\underline{u}\|^2 &= \left\langle \underline{w} - \frac{\langle \underline{w}|\underline{v}\rangle}{\|\underline{v}\|^2}\underline{v} \middle| \underline{w} - \frac{\langle \underline{w}|\underline{v}\rangle}{\|\underline{v}\|^2}\underline{v} \right\rangle \\ &= \langle \underline{w}|\underline{w}\rangle - \frac{\langle \underline{w}|\underline{v}\rangle \langle \underline{v}|\underline{w}\rangle}{\|\underline{v}\|^2} = \|\underline{w}\|^2 - \frac{|\langle \underline{v}|\underline{w}\rangle|^2}{\|\underline{v}\|^2}. \end{aligned}$$

Hence $|\langle \underline{v}|\underline{w}\rangle|^2 \leq \|\underline{v}\|^2 \|\underline{w}\|^2$. Notice that equality occurs iff $\underline{u} = \underline{0}$, or equivalently iff $\underline{w} = \underline{0}$ or $\underline{v} = s\underline{w}$.

Using this result, we get

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

with equality iff 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 $\underline{v} = s\underline{w}$ (i.e., $\underline{v}$ and $\underline{w}$ are linearly dependent). □

**Theorem 3.6.12.** *Consider the vector space $\mathbb{R}^n$ with the standard inner product of Example 3.6.3. The function $f : V \to F$ defined by $f(\underline{w}) = \langle \underline{w}|\underline{v}\rangle$ is continuous.*

*Proof.* Let $\underline{w}_1, \underline{w}_2, \ldots$ 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 $\|\underline{w}_n - \underline{w}\| \to 0$, the convergence of $\langle \underline{w}_n, \underline{v}\rangle$ is established. □

## 3.7 Orthogonal Vectors and Subspaces

**Definition 3.7.1.** *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.*

**Definition 3.7.2.** *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 3.7.3.** *A collection $W$ of vectors in $V$ is an **orthogonal set** if all pairs of distinct vectors in $W$ are orthogonal.*

**Example 3.7.4.** *The standard basis of $\mathbb{R}^n$ is an orthonormal set with respect to the standard inner product.*

**Example 3.7.5.** *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 $\{1, f_1, g_1, f_2, g_2, \ldots\}$ is a countably infinite orthonormal set that is a Schauder basis for this vector space.*

**Theorem 3.7.6.** *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 $\underline{w}_1, \ldots, \underline{w}_n$ are distinct vectors in $W$ and consider

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

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. □

**Corollary 3.7.7.** *If $\underline{v} \in V$ is a linear combination of an orthogonal sequence of distinct, non-zero vectors $\underline{w}_1, \ldots, \underline{w}_n$, then $\underline{v}$ is the particular linear combination*

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

**Theorem 3.7.8.** *Let $V$ be an inner-product space and assume $\underline{v}_1, \ldots, \underline{v}_n$ are linearly independent vectors in $V$. Then it is possible to construct an orthogonal sequence of vectors $\underline{w}_1, \ldots, \underline{w}_n \in V$ such that for each $k = 1, \ldots, n$ the set*

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

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

*Proof.* First let $\underline{w}_1 = \underline{v}_1$. Define the remaining vectors inductively as follows. Suppose the vectors

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

have been chosen so that for every $k$

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

is an orthogonal basis for the subspace spanned by $\underline{v}_1, \ldots, \underline{v}_k$. 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 $\underline{w}_1, \ldots, \underline{w}_m$ and hence a linear combination of $\underline{v}_1, \ldots, \underline{v}_m$. Furthermore, for $j \in 1, \ldots, m$

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

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

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

**Corollary 3.7.9.** *Every finite-dimensional inner-product space has a basis of orthonormal vectors.*

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

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

□

**Example 3.7.10.** *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*

$$\begin{aligned}
\underline{w}_1 &= (2, 2, 1) \\
\underline{w}_2 &= (3, 6, 0) - \frac{\langle (3, 6, 0) | (2, 2, 1) \rangle}{9} (2, 2, 1) \\
&= (3, 6, 0) - 2(2, 2, 1) = (-1, 2, -2) \\
\underline{w}_3 &= (6, 3, 9) - \frac{\langle (6, 3, 9) | (2, 2, 1) \rangle}{9} (2, 2, 1) - \frac{\langle (6, 3, 9) | (-1, 2, -2) \rangle}{9} (-1, 2, -2) \\
&= (6, 3, 9) - 3(2, 2, 1) + 2(-1, 2, -2) = (-2, 1, 2).
\end{aligned}$$

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

**Definition 3.7.11.** *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*

$$W^\perp = \left\{ \underline{v} \in V \middle| \langle \underline{v} | \underline{w} \rangle = 0 \ \forall \ \underline{w} \in W \right\}.$$

**Problem 3.7.12.** *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 3.7.12**. 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 $\underline{m}_1, \underline{m}_2, \ldots \in W^\perp$ be a sequence that converges to $\underline{m}$. The continuity of the inner product, from Theorem 3.6.12, implies that, for all $\underline{w} \in W$,

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

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

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

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

### 3.7.1 Hilbert Spaces

**Definition 3.7.13.** *A complete inner-product space is called a **Hilbert space**.*

**Definition 3.7.14.** *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 $V$. Each element $\underline{v} \in V$ has the form

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

where the sum converges in the (2)-norm. The **Parseval identity**

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

is obtained by computing $\langle \underline{v} | \underline{v} \rangle$.

**Theorem 3.7.15.** *Every orthogonal set in a Hilbert space $V$ can be enlarged to an orthonormal basis for $V$.*

*Proof.* Let $X$ be the set of orthonormal subsets of $V$. Furthermore, for $x, y \in X$ consider the strict partial order defined by proper inclusion. If $x = \{\underline{v}_\alpha | \alpha \in A_0\}$ is an element of $X$, then by the Hausdorff maximal principle 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 an orthonormal basis. □

**Theorem 3.7.16.** *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. Using the well-ordering theorem, this subset can be ordered into a sequence $\underline{v}_1, \underline{v}_2, \ldots$ such that, for every vector $\underline{v} \in V$ and any $\epsilon > 0$, there exists an $n$ such that $\|\underline{v} - \underline{v}_n\| < \epsilon$. Therefore, applying Gram-Schmidt orthogonalization, to this ordered sequence of vectors, generates a countable orthonormal basis. Conversely, if $V$ has a countable orthonormal basis, then linear combinations with rational coefficients can be used to construct a countable dense subset. □

## 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

$$f\left(s\underline{v}_1 + \underline{v}_2\right) = sf\left(\underline{v}_1\right) + f\left(\underline{v}_2\right)$$

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 $s_1, \ldots, s_n$ be scalars in $F$. Then the functional $f$ on $F^n$ defined by*

$$f(v_1, \ldots, v_n) = s_1 v_1 + \cdots + s_n v_n$$

*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 $s_1, \ldots, s_n$.*

**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 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.* If we choose an orthonormal basis $\mathcal{B} = \underline{v}_1, \ldots, \underline{v}_n$ 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

$$f(\underline{w}) = f(t_1\underline{v}_1 + \cdots + t_n\underline{v}_n) = t_1 f\left(\underline{v}_j\right) + \cdots + t_n f\left(\underline{v}_j\right).$$

Thus, we can choose $\bar{s}_j = f\left(\underline{v}_j\right)$ to get $\langle \underline{w} | \underline{v} \rangle = f\left(\underline{w}\right)$ and this gives

$$\underline{v} = \overline{f\left(\underline{v}_1\right)}\underline{v}_1 + \cdots + \overline{f\left(\underline{v}_n\right)}\underline{v}_n.$$

Let $\underline{v}'$ be any vector that satisfies $f\left(\underline{w}\right) = \langle \underline{w} | \underline{v}' \rangle$ 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 $\underline{w}_1, \ldots, \underline{w}_n \in V$, then we can write

$$\begin{aligned} \underline{v} &= \underline{w} + \underline{e} \\ &= s_1\underline{w}_1 + \cdots + s_n\underline{w}_n + \underline{e}, \end{aligned}$$

where $\underline{e}$ is the approximation error.

This problem is, in general, rather difficult. 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.

**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 the 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 $\underline{w}_1, \underline{w}_2, \ldots$ and is closed, then*

$$\underline{w} = \sum_{i=1}^{\infty} \frac{\langle \underline{v} | \underline{w}_i \rangle}{\|\underline{w}_i\|^2} \underline{w}_i \tag{4.1}$$

*is 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 $W$. Let $\underline{w}' \in W$ such that $\underline{w}' \neq \underline{w}$. Then $\underline{v} - \underline{w}' = (\underline{v} - \underline{w}) + (\underline{w} - \underline{w}')$ and

$$\begin{aligned} \|\underline{v} - \underline{w}'\|^2 &= \|\underline{v} - \underline{w}\|^2 + 2\mathrm{Re}\,\langle \underline{v} - \underline{w} | \underline{w} - \underline{w}' \rangle + \|\underline{w} - \underline{w}'\|^2 \\ &= \|\underline{v} - \underline{w}\|^2 + \|\underline{w} - \underline{w}'\|^2 \\ &\geq \|\underline{v} - \underline{w}\|^2 . \end{aligned} \tag{4.2}$$

Conversely, suppose that $\|\underline{v} - \underline{w}'\| \geq \|\underline{v} - \underline{w}\|$ for every $\underline{w}' \in W$. From (4.2), we get

$$2\mathrm{Re}\,\langle \underline{v} - \underline{w} | \underline{w} - \underline{w}' \rangle + \|\underline{w} - \underline{w}'\|^2 \geq 0$$

for all $\underline{w}' \in W$. Note that every vector in $W$ can be expressed as $\underline{w} - \underline{w}'$ where $\underline{w}' \in W$, it follows that

$$2\mathrm{Re}\,\langle \underline{v} - \underline{w} | \underline{w}'' \rangle + \|\underline{w}''\|^2 \geq 0 \tag{4.3}$$

for every $\underline{w}'' \in W$. If $\underline{w}'$ is in $W$ and $\underline{w}' \neq \underline{w}$ then we may take

$$\underline{w}'' = -\frac{\langle \underline{w} - \underline{w}' | \underline{v} - \underline{w} \rangle}{\|\underline{w} - \underline{w}'\|^2} (\underline{w} - \underline{w}') .$$

Inequality (4.3) then reduces to the statement

$$-2\frac{|\langle \underline{v} - \underline{w} | \underline{w} - \underline{w}' \rangle|^2}{\|\underline{w} - \underline{w}'\|^2} + \frac{|\langle \underline{v} - \underline{w} | \underline{w} - \underline{w}' \rangle|^2}{\|\underline{w} - \underline{w}'\|^2} \geq 0.$$

This inequality holds if and only if $\langle \underline{v} - \underline{w} | \underline{w} - \underline{w}' \rangle = 0$. Therefore, $\underline{v} - \underline{w}$ is orthogonal to every vector in $W$. Hence 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 $W$.

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.

Assume that $W$ is closed and has a countable orthogonal basis $\underline{w}_1, \underline{w}_2, \ldots$ and let $\underline{w}$ be defined by (4.1). Then $\underline{v} - \underline{w}$ is orthogonal to $\underline{w}_j$ for $j \in \mathbb{N}$, i.e.,

$$\begin{aligned}\langle \underline{v} - \underline{w} | \underline{w}_j \rangle &= \langle \underline{v} | \underline{w}_j \rangle - \left\langle \sum_{i=1}^{\infty} \frac{\langle \underline{v} | \underline{w}_i \rangle}{\|\underline{w}_i\|^2} \underline{w}_i \middle| \underline{w}_j \right\rangle \\ &= \langle \underline{v} | \underline{w}_j \rangle - \frac{\langle \underline{v} | \underline{w}_j \rangle}{\|\underline{w}_i\|^2} \langle \underline{w}_j | \underline{w}_j \rangle = 0.\end{aligned}$$

That is, $\underline{v} - \underline{w}$ is orthogonal to every vector in $W$ and therefore $\underline{w}$ is the best approximation of $\underline{v}$ by vectors in $W$. □

**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 $E\colon V \to W$, 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 4.1.13 to verify that this is consistent with the concept of orthogonal projection from Definition 4.1.10.

**Problem 4.1.4.** *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 4.1.5.** *A function $F\colon X \to Y$ 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 4.1.6.** *Let $V$ be a vector space and $T\colon V \to V$ be a linear transformation. Then, $T$ is a projection if $T$ is idempotent.*

**Example 4.1.7.** *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 4.1.8.** *Let $V$ be a vector space and $T\colon V \to V$ be a projection operator. Then, the range $\mathcal{R}(T)$ and the $\mathcal{N}(T)$ 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. □

**Example 4.1.9.** *Consider the linear transform $T\colon V \to V$ 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 4.1.10.** *Let $V$ be an inner-product space and $P\colon V \to V$ be a projection operator. Then, $P$ is an **orthogonal projection** if $\mathcal{R}(P) \perp \mathcal{N}(P)$.*

**Example 4.1.11.** *Let $V$ be an inner-product space and $P\colon V \to V$ be an orthogonal projection. Then, $\underline{v} = P\underline{v} + (I - P)\underline{v}$ defines a natural orthogonal decomposition of $\underline{v}$ because $P\underline{v} \in \mathcal{R}(P)$ and $(I - P)\underline{v} \in \mathcal{N}(P)$. Therefore, $\mathcal{N}(P) = \mathcal{R}(P)^{\perp}$ and $P\underline{v} = \underline{v}$ if and only if $\underline{v} \in \mathcal{R}(P)$.*

**Theorem 4.1.12.** *Any idempotent Hermitian matrix $P$ defines an orthogonal projection operator.*

*Proof.* We simply must verify that the range a null space are orthogonal. Since $P\underline{u} \in \mathcal{R}(P)$ and $(I - P)\underline{v} \in \mathcal{N}(P)$, we can simply point out that

$$\langle P\underline{u} | (I - P)\underline{v} \rangle = \underline{v}^H (I - P)^H P\underline{u} = \underline{v}^H (P - P^H P)\underline{u} = \underline{v}^H (P - P^2)\underline{u} = 0.$$

□

**Theorem 4.1.13.** *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, $E\left(E\underline{v}\right) = E\underline{v}$ 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 $\underline{w}_1, \underline{w}_2, \ldots$ be a countable orthonormal basis for $W$ (whose existence follows from Theorem 3.7.16). Using part 3 of Theorem 4.1.2, we find that

$$\begin{aligned} E\left(s_1\underline{v}_1 + \underline{v}_2\right) &= \sum_{i=1}^{\infty} \langle s_1\underline{v}_1 + \underline{v}_2 | \underline{w}_i \rangle \underline{w}_i \\ &= s_1 \sum_{i=1}^{\infty} \langle \underline{v}_1 | \underline{w}_i \rangle \underline{w}_i + \sum_{i=1}^{\infty} \langle \underline{v}_2 | \underline{w}_i \rangle \underline{w}_i \\ &= s_1 E\underline{v}_1 + E\underline{v}_2. \end{aligned}$$

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

$$W \cap W^{\perp} = \{\underline{u} \in W | \langle \underline{u} | \underline{w} \rangle = 0 \; \forall \; \underline{w} \in W\} \subseteq \{\underline{u} \in W | \langle \underline{u} | \underline{u} \rangle = 0\} = \{\underline{0}\}.$$

□

**Corollary 4.1.14.** *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 4.1.13. One can also verify that $I - E$ is an idempotent linear transformation of $V$ with range $W^{\perp}$ and nullspace $W$. From Definition 4.1.10, we see that $I - E$ is an orthogonal projection. □

Theorem 4.1.13 also implies the following result, known as Bessel's inequality.

**Corollary 4.1.15.** *Let $\underline{v}_1, \underline{v}_2, \ldots$ be a countable orthogonal set of distinct non-zero vectors in an inner-product space $V$. If $\underline{v} \in V$ then*

$$\sum_{i=1}^{\infty} \frac{|\langle \underline{v} | \underline{v}_i \rangle|^2}{\|\underline{v}_i\|^2} \leq \|\underline{v}\|^2 .$$

*Moreover, equality holds if and only if*

$$\underline{v} = \sum_{i=1}^{\infty} \frac{\langle \underline{v} | \underline{v}_i \rangle}{\|\underline{v}_i\|^2} \underline{v}_i .$$

*Proof.* Define

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

Then $\underline{v} = \underline{w} + \underline{u}$, where $\langle \underline{w} | \underline{u} \rangle = 0$ and $\|\underline{v}\|^2 = \|\underline{w}\|^2 + \|\underline{u}\|^2$. Noting that

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

we obtain the desired result, with equality iff $\underline{u} = \underline{0}$. □

If $\underline{v}_1, \underline{v}_2, \ldots$ is an orthonormal set, Bessel's inequality states that

$$\sum_{i=1}^{\infty} |\langle \underline{v} | \underline{v}_i \rangle|^2 \leq \|\underline{v}\|^2 .$$

It follows that the vector $\underline{v}$ is in the closure of the subspace spanned by $\underline{v}_1, \underline{v}_2, \ldots$ if an only if

$$\underline{v} = \sum_{i=1}^{\infty} \langle \underline{v} | \underline{v}_i \rangle \, \underline{v}_i .$$

## 4.2 Computing Approximations in Hilbert Spaces

### 4.2.1 Normal Equations

Suppose $V$ is a Hilbert space the subspace $W$ is spanned by $\underline{w}_1, \ldots, \underline{w}_n \in V$. Consider the situation where the sequence $\underline{w}_1, \ldots, \underline{w}_n$ 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

$$\left\langle \underline{v} - \hat{\underline{v}} | \underline{w}_j \right\rangle = \left\langle \underline{v} - \sum_{i=1}^{n} s_i \underline{w}_i \middle| \underline{w}_j \right\rangle = 0$$

or, equivalently,

$$\sum_{i=1}^{n} s_i \left\langle \underline{w}_i | \underline{w}_j \right\rangle = \left\langle \underline{v} | \underline{w}_j \right\rangle$$

for $j = 1, \ldots, n$. 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

$$\underline{t}^T = \left( \langle \underline{v} | \underline{w}_1 \rangle, \langle \underline{v} | \underline{w}_2 \rangle, \ldots, \langle \underline{v} | \underline{w}_n \rangle \right)$$

is the **cross-correlation vector**, and

$$\underline{s}^T = (s_1, s_2, \ldots, s_n)$$

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*

$$G = \begin{bmatrix} \langle \underline{w}_1 | \underline{w}_1 \rangle & \langle \underline{w}_2 | \underline{w}_1 \rangle & \cdots & \langle \underline{w}_n | \underline{w}_1 \rangle \\ \langle \underline{w}_1 | \underline{w}_2 \rangle & \langle \underline{w}_2 | \underline{w}_2 \rangle & \cdots & \langle \underline{w}_n | \underline{w}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{w}_1 | \underline{w}_n \rangle & \langle \underline{w}_2 | \underline{w}_n \rangle & \cdots & \langle \underline{w}_n | \underline{w}_n \rangle \end{bmatrix} \tag{4.4}$$

*is called the **Grammian** matrix. Since $G_{ji} = \left\langle \underline{w}_i | \underline{w}_j \right\rangle$, it follows that the Grammian 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 $\underline{v}^H M\underline{v} \geq 0$ for all $\underline{v} \in F^n$. A matrix $M \in F^{n\times n}$ is **positive-definite** if $\underline{v}^H M\underline{v} > 0$ for all $\underline{v} \in F^n - \{\underline{0}\}$.*

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 $\underline{v}^H M\underline{v} = \underline{0}$ and contradicts the definition of positive definite.

**Theorem 4.2.3.** *A Grammian matrix $G$ is always positive-semidefinite. Furthermore, it is positive-definite if and only if the sequence of vectors $\underline{w}_1, \ldots, \underline{w}_n$ is linearly independent.*

*Proof.* Let $\underline{v} = (v_1, \ldots, v_n)^T \in F^n$. Then,

$$\begin{aligned}\underline{v}^H G\underline{v} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{v}_j G_{ji} v_i = \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{v}_j \left\langle \underline{w}_i | \underline{w}_j \right\rangle v_i \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \left\langle v_i\underline{w}_i | v_j\underline{w}_j \right\rangle = \left\langle \sum_{i=1}^{n} v_i\underline{w}_i \middle| \sum_{j=1}^{n} v_j\underline{w}_j \right\rangle \\ &= \left\| \sum_{i=1}^{n} v_i\underline{w}_i \right\|^2 \geq 0.\end{aligned} \tag{4.5}$$

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 (4.5), this implies that

$$\sum_{i=1}^{n} v_i\underline{w}_i = 0$$

and hence the sequence of vectors $\underline{w}_1, \ldots, \underline{w}_n$ 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 $\underline{v} \in F^n - \{\underline{0}\}$. Therefore, the sequence of vectors $\underline{w}_1, \ldots, \underline{w}_n$ is linearly independent. □

### 4.2.2 Orthogonality Principle

**Theorem 4.2.4.** *Let $\underline{w}_1, \ldots, \underline{w}_n$ be vectors in an inner-product space $V$ and denote the span of $\underline{w}_1, \ldots, \underline{w}_n$ by $W$. For any vector $\underline{v} \in V$, the norm of the error vector*

$$\underline{e} = \underline{v} - \sum_{i=1}^{n} s_i \underline{w}_i \tag{4.6}$$

*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*

$$\left\langle \underline{v} - \hat{\underline{v}} | \underline{w}_j \right\rangle = 0$$

*for $j = 1, \ldots, n$.*

*Proof.* Minimizing $\|\underline{e}\|^2$ over $\underline{s}$, where $\underline{e}$ is given by (4.6) requires minimizing

$$\begin{aligned} J(\underline{s}) &= \left\langle \underline{v} - \sum_{i=1}^{n} s_i \underline{w}_i \middle| \underline{v} - \sum_{j=1}^{n} s_j \underline{w}_j \right\rangle \\ &= \langle \underline{v} | \underline{v} \rangle - \sum_{i=1}^{n} \langle s_i \underline{w}_i | \underline{v} \rangle - \sum_{j=1}^{n} \left\langle \underline{v} | s_j \underline{w}_j \right\rangle + \sum_{i=1}^{n} \sum_{j=1}^{n} \left\langle s_i \underline{w}_i | s_j \underline{w}_j \right\rangle \\ &= \langle \underline{v} | \underline{v} \rangle - \sum_{i=1}^{n} s_i \langle \underline{w}_i | \underline{v} \rangle - \sum_{j=1}^{n} \bar{s}_j \left\langle \underline{v} | \underline{w}_j \right\rangle + \sum_{i=1}^{n} \sum_{j=1}^{n} s_i \bar{s}_j \left\langle \underline{w}_i | \underline{w}_j \right\rangle . \end{aligned}$$

Taking the gradient of $J(\underline{s})$, we get

$$\begin{aligned} \nabla J(\underline{s}) &= - \begin{bmatrix} \langle \underline{v} | \underline{w}_1 \rangle \\ \langle \underline{v} | \underline{w}_2 \rangle \\ \vdots \\ \langle \underline{v} | \underline{w}_n \rangle \end{bmatrix} + \begin{bmatrix} \langle \underline{w}_1 | \underline{w}_1 \rangle & \langle \underline{w}_2 | \underline{w}_1 \rangle & \ldots & \langle \underline{w}_n | \underline{w}_1 \rangle \\ \langle \underline{w}_1 | \underline{w}_2 \rangle & \langle \underline{w}_2 | \underline{w}_2 \rangle & \ldots & \langle \underline{w}_n | \underline{w}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{w}_1 | \underline{w}_n \rangle & \langle \underline{w}_2 | \underline{w}_n \rangle & \ldots & \langle \underline{w}_n | \underline{w}_n \rangle \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} \\ &= \underline{0}. \end{aligned}$$

In matrix form, this yields the familiar equation

$$G\underline{s} = \underline{t}.$$

To ensure that this extremum is in fact a minimum, we compute the Hessian matrix

$$\nabla^2 J(\underline{s}) = G.$$

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 $W$. □

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 $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n \in V$ 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 $j = 1, \ldots, n$.

### 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 Grammian $G$ defined in (4.4). The vector $A^H \underline{v}$ is the cross correlation vector $\underline{t}$.

When the vectors $\underline{w}_1, \ldots, \underline{w}_n$ are linearly independent, the Grammian matrix is positive definite and hence invertible. The optimal solution for the least-squares problem is therefore given by

$$\underline{s} = \left(A^H A\right)^{-1} A^H \underline{v} = G^{-1} \underline{t}.$$

The matrix $\left(A^H A\right)^{-1} A^H$ is often called the **pseudoinverse**.

The best approximation of $\underline{v} \in V$ by vectors in $W$ is equal to

$$\hat{\underline{v}} = A\underline{s} = A\left(A^H A\right)^{-1} A^H \underline{v}.$$

The matrix $P = A\left(A^H A\right)^{-1} A^H$ 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

$$h_{ji} = \left\langle \underline{e}_i | \underline{e}_j \right\rangle,$$

and that this inner product can be expressed 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 $\underline{w}_1, \ldots, \underline{w}_n$ are linearly independent, the optimal solution is given by

$$\underline{s} = \left(A^H H A\right)^{-1} A^H H \underline{v}.$$

### 4.3.4 Minimum Error

Let $\underline{\hat{v}} \in W$ be the best approximation of $\underline{v}$ by vectors in $W$. Again, we can write

$$\underline{v} = \underline{\hat{v}} + \underline{e}$$

where $\underline{e} \in W^{\perp}$ is the minimum achievable error. The squared norm of the minimum error is given implicitly by

$$\|\underline{v}\|^2 = \|\underline{\hat{v}} + \underline{e}\|^2 = \langle \underline{\hat{v}} + \underline{e} | \underline{\hat{v}} + \underline{e} \rangle = \langle \underline{\hat{v}} | \underline{\hat{v}} \rangle + \langle \underline{e} | \underline{e} \rangle = \|\underline{\hat{v}}\|^2 + \|\underline{e}\|^2 .$$

We can then find an explicit expression for the approximation error,

$$\begin{aligned} \|\underline{e}\|^2 &= \|\underline{v}\|^2 - \|\underline{\hat{v}}\|^2 = \underline{v}^H H \underline{v} - \underline{\hat{v}}^H H \underline{\hat{v}} \\ &= \underline{v}^H H \underline{v} - \underline{s}^H A^H H A \underline{s} \\ &= \underline{v}^H H \underline{v} - \underline{v}^H H A \left( A^H H A \right)^{-1} A^H H \underline{v} \\ &= \underline{v}^H \left( H - H A \left( A^H H A \right)^{-1} A^H H \right) \underline{v}. \end{aligned}$$

## 4.4 Applications and Examples in Signal Processing

### 4.4.1 Linear Regression

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a collection of points in $\mathbb{R}^2$. A **linear regression** problem consists in finding scalars $a$ and $b$ such that

$$y_i \approx a x_i + b$$

for $i = 1, \ldots, n$. Definite the error component $e_i$ by $e_i = y_i - a x_i - b$, 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 $\underline{y} = (y_1, \ldots, y_n)^T$, $\underline{s} = (a, b)^T$, $\underline{e} = (e_1, \ldots, e_n)^T$, 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 problems. 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

$$\underline{s} = \left(A^H A\right)^{-1} A^H \underline{y}.$$

### 4.4.2 Linear Minimum Mean-Squared Estimation

Let $Y, X_1, \ldots, X_n$ be a set of zero-mean random variables. The goal of the linear minimum mean-squared estimation (LMMSE) problem is to find coefficients $s_1, \ldots, s_n$ such that

$$\hat{Y} = s_1 X_1 + \cdots + s_n X_n.$$

minimizes the MSE $\mathrm{E}[|Y - \hat{Y}|^2]$. Using the inner product defined by

$$\langle X|Y\rangle = \mathrm{E}\left[X\overline{Y}\right], \tag{4.7}$$

we can compute the linear minimum mean-squared estimate $\hat{Y}$ using

$$G\underline{s} = \underline{t},$$

where

$$G = \begin{bmatrix} \mathrm{E}\left[X_1\overline{X}_1\right] & \mathrm{E}\left[X_2\overline{X}_1\right] & \cdots & \mathrm{E}\left[X_n\overline{X}_1\right] \\ \mathrm{E}\left[X_1\overline{X}_2\right] & \mathrm{E}\left[X_2\overline{X}_2\right] & \cdots & \mathrm{E}\left[X_n\overline{X}_2\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}\left[X_1\overline{X}_n\right] & \mathrm{E}\left[X_2\overline{X}_n\right] & \cdots & \mathrm{E}\left[X_n\overline{X}_n\right] \end{bmatrix}$$

and

$$\underline{t} = \begin{bmatrix} \mathrm{E}\left[Y\overline{X}_1\right] \\ \mathrm{E}\left[Y\overline{X}_2\right] \\ \vdots \\ \mathrm{E}\left[Y\overline{X}_n\right] \end{bmatrix}.$$

If the matrix $G$ is invertible, the minimum mean-squared error is given by

$$\left\|Y - \hat{Y}\right\|^2 = \mathrm{E}\left[Y\overline{Y}\right] - \underline{t}^H G^{-1} \underline{t}.$$

### 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

$$\begin{aligned} Y[t] &= \sum_{k=0}^{K-1} h[k]X[t-k] \\ &= \begin{bmatrix} X[t] & \dots & X[t-K+1] \end{bmatrix} \begin{bmatrix} h[0] \\ \vdots \\ h[K-1] \end{bmatrix} = (\underline{X}[t])^T \, \underline{h}. \end{aligned}$$

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

$$\|Z[t] - Y[t]\|^2 = \mathrm{E}\left[|Z[t] - Y[t]|^2\right].$$

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

$$\left\langle Z[t] - \sum_{k=0}^{K-1} h[k]X[t-k] \middle| X[t-j] \right\rangle = 0,$$

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 (4.7), we obtain

$$\mathrm{E}\left[Z[t]\overline{X}[t-j]\right] = \sum_{k=0}^{K-1} h[k]\mathrm{E}\left[X[t-k]\overline{X}[t-j]\right]. \tag{4.8}$$

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 Grammian 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

$$R_{xx}(j-k) = R_{xx}(j,k) = \mathrm{E}\left[X[t-k]\overline{X}[t-j]\right] = \langle X[t-k] | X[t-j] \rangle .$$

Also define

$$R_{zx}(j) = \mathrm{E}\left[Z[t]\overline{X}[t-j]\right] = \langle Z[t]|X[t-j]\rangle\,.$$

Using this notation, we can rewrite (4.8) 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

$$\begin{aligned} \|Z - Y\|^2 &= \|Z\|^2 - \|Y\|^2 \\ &= \mathrm{E}[Z\overline{Z}] - \mathrm{E}\left[\underline{h}^H \overline{\underline{X}}\,\underline{X}^T \underline{h}\right] \\ &= \mathrm{E}[Z\overline{Z}] - \underline{h}^H R_{xx} \underline{h} \\ &= \mathrm{E}[Z\overline{Z}] - R_{zx}^H \underline{h}, \end{aligned}$$

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 $x[1], x[2], \ldots_1, x[N]$ and $z[1], z[2], \ldots, z[N]$ 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 $h[0], h[1], \ldots, h[K-1]$ 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 $\lambda$ emphasizes the most recently received symbols because, in reality, the channel conditions are changing with time.

Using the vector $\underline{z} = \begin{bmatrix} z[K] & z[K+1] & \cdots & z[N] \end{bmatrix}$ 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 $\Lambda$ is a diagonal matrix whose diagonal contains $\begin{bmatrix} \lambda^{N-K} & \lambda^{N-K+1} & \cdots & \lambda^1 & \lambda^0 \end{bmatrix}$. 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 $\lambda$ close to 1, we have

$$\left[A^H \Lambda A\right]_{ij} = \sum_{t=K}^{N} \lambda^{N-t} x[t-j+1]\overline{x}[t-i+1] \approx \frac{R_{xx}(i-j)}{1-\lambda}$$

and

$$\left[A^H \Lambda \underline{z}\right]_i = \sum_{t=K}^{N} \lambda^{N-t} z[t]\overline{x}[t-i+1] \approx \frac{R_{zx}(i)}{1-\lambda}.$$

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 $\underline{t}_{N+1} = \lambda \underline{b}_N + z[N+1]\underline{u}^H$, where

$$\underline{u} = \begin{bmatrix} x[N+1] & x[N] & \cdots & x[N-K+2] \end{bmatrix}.$$

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 $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n$ be a set of linearly independent vectors in $W$. For any $\underline{v} \in V$, consider finding the scalars $s_1, s_2, \ldots, s_n$ 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 $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n$. The orthogonality principle tells us that $s_1, s_2, \ldots, s_n$ must satisfy

$$\begin{bmatrix} \langle \underline{w}_1 | \underline{w}_1 \rangle & \langle \underline{w}_2 | \underline{w}_1 \rangle & \cdots & \langle \underline{w}_n | \underline{w}_1 \rangle \\ \langle \underline{w}_1 | \underline{w}_2 \rangle & \langle \underline{w}_2 | \underline{w}_2 \rangle & \cdots & \langle \underline{w}_n | \underline{w}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{w}_1 | \underline{w}_n \rangle & \langle \underline{w}_2 | \underline{w}_n \rangle & \cdots & \langle \underline{w}_n | \underline{w}_n \rangle \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} = \begin{bmatrix} \langle \underline{v} | \underline{w}_1 \rangle \\ \langle \underline{v} | \underline{w}_2 \rangle \\ \vdots \\ \langle \underline{v} | \underline{w}_n \rangle \end{bmatrix}. \tag{4.9}$$

The same problem can also be posed in a different manner.

**Theorem 4.5.1.** *Let $V$ be a Hilbert space and $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n$ 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 $i = 1, \ldots, n$. This vector is given by*

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

*where the coefficients $s_1, s_2, \ldots, s_n$ can be found by solving* (4.9) *with $\langle \underline{v} | \underline{w}_i \rangle = c_i$.*

*Proof.* Let $W = \operatorname{span}(\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n)$ and notice that the subset

$$A = \{\underline{u} \in V | \, \langle \underline{u} | \underline{w}_i \rangle = c_i \; \forall \; i = 1, \ldots, n\}$$

is simply the orthogonal complement $W^{\perp}$ translated by any vector $\underline{v} \in A$. Therefore, the vector achieving $\min_{\underline{u} \in A} \|\underline{u}\|$ is the error vector in the best approximation of some $\underline{v} \in A$ by vectors in $W^{\perp}$. Using the unique decomposition $\underline{v} = \hat{\underline{v}} + \underline{w}$ implied by the orthogonal decomposition $V = W^{\perp} \oplus W$, one finds that the error vector $\underline{w}$ must lie in $W$. Moreover, the normal equations, given by modifying (4.9), show that the error vector $\underline{w}$ is the unique vector in $W$ that satisfies $\langle \underline{w} | \underline{w}_i \rangle = c_i$ for $i = 1, \ldots, n$. □

### 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 linear 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 $(A^H)^H \underline{s} = \underline{v}$. Then, the theorem concludes that the minimum norm solution lies in $\mathcal{R}(A^H)$ (i.e., the column space of $A^H$). Using this, one can define $\hat{\underline{s}} = A^H \underline{t}$ and see that $A(A^H \underline{t}) = \underline{v}$. 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

$$\hat{\underline{s}} = A^H \left(AA^H\right)^{-1} \underline{v}.$$

## 4.6 Projection onto Convex Sets

So far, we have focused on the projection of vectors on to subspaces. Convex sets are generalize the closure property of subspaces to subsets of points.

**Definition 4.6.1.** *Let $V$ be a vector space over $\mathbb{R}$. 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 $\lambda\underline{a}_1 + (1-\lambda)\underline{a}_2 \in A^\circ$.*

Let $V$ be a Hilbert space and $A \subseteq V$ be a closed convex set. For any $\underline{v}$, let

$$f(\underline{v}) = \inf_{\underline{u} \in A} \|\underline{u} - \underline{v}\|$$

be the minimum distance between $\underline{v}$ and the set $A$. In this section, we will show that this infimum is achieved as a minimum by a unique $\underline{u}^* \in A$. Moreover, the mapping $P_A : V \to A$ is called the orthogonal projection of $\underline{v}$ onto $A$ and is defined by $\underline{v} \mapsto \underline{u}^*$.

**Theorem 4.6.2.** *The orthogonal projection of $\underline{v} \in V$ onto a closed convex set $A \subseteq V$ exists and is unique.*

*Proof.* Let $d = f(\underline{v})$ be the infimal distance and consider any sequence $\underline{u}_1, \underline{u}_2, \ldots \in A$ that achieves the infimum so that

$$\lim_{n\to\infty} \|\underline{u}_n - \underline{v}\| = d.$$

Since $A$ is complete, the next step is showing that this sequence is Cauchy. The paralleogram law states that $\|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2 - \|x+y\|^2$ and applying this to $\underline{u}_n + \underline{u}_m$ 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 $m, n \to \infty$) 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. □

**Remark 4.6.3.** *The same result holds for norm projections in many other Banach spaces including $L^p$ and $\ell^p$ 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.

**Lemma 4.6.4.** *For any $\underline{v} \notin A$, a necessary and sufficient condition for $\underline{u}^* = P_A(\underline{v})$ is that $\langle \underline{v} - \underline{u}^*, \underline{u} - \underline{u}^* \rangle \leq 0$ 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

$$\begin{aligned}\|\underline{v} - \underline{u}^*\|^2 &\leq \|\underline{v} - \underline{u}'\| \\ &\leq \|\underline{v} - \underline{u}^* - \alpha(\underline{u} - \underline{u}^*)\|^2 \\ &= \|\underline{v} - \underline{u}^*\|^2 + \alpha^2\|\underline{u} - \underline{u}^*\|^2 - 2\alpha\langle v - \underline{u}^*, \underline{u} - \underline{u}^*\rangle.\end{aligned}$$

Thus, $\langle v - \underline{u}^*, \underline{u} - \underline{u}^*\rangle \leq \frac{\alpha}{2}\|\underline{u} - \underline{u}^*\|^2$. One can establish necessity by taking the limit as $\alpha \to 0$. For sufficiency, we assume $\langle \underline{v} - \underline{u}^*, \underline{u} - \underline{u}^* \rangle \leq 0$ and we write

$$\begin{aligned}\|\underline{v} - \underline{u}\|^2 - \|\underline{v} - \underline{u}^*\|^2 &= \|(\underline{v} - \underline{u}^*) - (\underline{u} - \underline{u}^*)\|^2 - \|\underline{v} - \underline{u}^*\|^2 \\ &= \|\underline{v} - \underline{u}^*\|^2 + \|\underline{u} - \underline{u}^*\|^2 - 2\langle v - \underline{u}^*, \underline{u} - \underline{u}^*\rangle - \|\underline{v} - \underline{u}^*\|^2 \\ &\geq 0.\end{aligned}$$

Thus, $\|\underline{v} - \underline{u}\|^2 \geq \|\underline{v} - \underline{u}^*\|^2$ for all $\underline{u} \in A$ and $\underline{u}^* = P_A(\underline{v})$. □

### 4.6.1 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 number of 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 $A \times B$, one finds that any local minimum distance is a global minimum distance.we have the following results.

**Theorem 4.6.5.** *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 the set $B$ is compact, then the infimum is achieved. If the set $A$ is strictly convex and the infimum is achieved, then the minimizing points $\underline{u}^*, \underline{v}^*$ are unique.*

*Proof.* Consider any sequence $(\underline{u}_1, \underline{v}_1), (\underline{u}_2, \underline{v}_2), \ldots \in A \times B$ which satisfies

$$\lim_{n \to \infty} \|\underline{u}_n - \underline{v}_n\| = d.$$

If $B$ is compact, then there is a subsequence $\underline{v}_{n_j}$ that converges to some $\underline{v}^* \in B$. Therefore, we can set $B = \{v^*\}$ and use Theorem 4.6.2 to see that there exists a unique $\underline{u}^* \in A$ which achieves the smallest distance. Notice that the $\underline{v}^*$ is not unique in general and that $\underline{u}^*$ is only unique for a given $\underline{v}^*$.

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

$$M = \left\{ (\underline{u}, \underline{v}) \in A \times B \;\middle|\; \|\underline{u} - \underline{v}\| = d \right\}.$$

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. □

**Remark 4.6.6.** *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*

$$\underline{u}_{n+1} = \arg\min_{\underline{u}\in A} \|\underline{u} - \underline{v}_n\|$$
$$\underline{v}_{n+1} = \arg\min_{\underline{v}\in B} \left\|\underline{u}_{n+1} - \underline{v}\right\|.$$

*Notice that the sequence $d_n = \|\underline{u}_n - \underline{v}_n\|$ is non-increasing and must therefore have a limit. One can show that, if either set is compact (resp. bounded), then the sequence $(\underline{u}_n, \underline{v}_n)$ converges strongly (resp. weakly) to vectors that minimize the distance. Variations of this approach can even be useful when $A \cap B \neq \emptyset$. In this case, sequence $\underline{v}_n$ converges weakly to some point $\underline{v} \in A \cap B$.*

# 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 optimization theory.

## 5.1 Derivatives in Banach Spaces

**Definition 5.1.1.** *Let $f\colon X \to Y$ be a mapping from a vector space $X$ to a Banach space $(Y, \|\cdot\|)$. Then, if it exists, the **Gâteuax differential** of $f$ at $\underline{x}$ in direction $\underline{h}$ is given by*

$$\delta f(\underline{x}; \underline{h}) \triangleq \lim_{t \to 0} \frac{f(\underline{x} + t\underline{h}) - f(\underline{x})}{t},$$

*where the limit is with respect to the implied mapping from $\mathbb{R}$ to $Y$.*

**Lemma 5.1.2.** *Let $Y = (\mathbb{R}, |\cdot|)$ and suppose that $\delta f(\underline{x}; \underline{h}) < 0$ exists for some $f$, $\underline{x}$, and $\underline{h}$. Then, there exists $t_0 > 0$ such that, for all $t \in (0, t_0)$, one has*

$$f(\underline{x} + t\underline{h}) < f(\underline{x}).$$

*Proof.* This follows from applying the definition of the limit. □

**Problem 5.1.3.** *Suppose $X = Y = L^1([0,1])$ is the Banach space of Lebesgue absolutely integrable functions mapping $[0,1]$ to $\mathbb{R}$ and $f(\underline{x}) = \|\underline{x}\| = \int_0^1 |x(s)| \mathrm{d}s$ is the norm of $\underline{x}$. Compute $\delta f(\underline{x}; \underline{h})$ for this case by interchaging the limit and integration in*

$$\lim_{t \to 0} \int_0^1 \frac{1}{t} \left( |x(s) + th(s)| - |x(s)| \right) \mathrm{d}s.$$

**Definition 5.1.4.** *Let $f\colon X \to Y$ be a mapping from a vector space $X$ to a Banach space $(Y, \|\cdot\|)$. Then, $f$ is **Gâteuax differentiable** at $\underline{x}$ if the Gâteuax differential $\delta f(\underline{x}; \underline{h})$ exists for all $\underline{h} \in X$ and is a continuous linear function of $\underline{h}$.*

**Definition 5.1.5.** *Let $f\colon X \to Y$ be a mapping from a Banach space $(X, \|\cdot\|_X)$ to a Banach space $(Y, \|\cdot\|_Y)$. Then, $f$ is **Fréchet differentiable** at $\underline{x}_0$ if there is a continuous linear transformation $T\colon X \to Y$ satisfying*

$$\lim_{\underline{h}\to\underline{0}} \frac{\|f(\underline{x}+\underline{h}) - f(\underline{x}) - T(\underline{h})\|_Y}{\|\underline{h}\|_X} = 0,$$

*where the limit is with respect to the implied Banach space mapping $X \to \mathbb{R}$.*

**Example 5.1.6.** *A function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ with $f = (f_1, f_2, \ldots, f_m)^T$ 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 also called the **gradient** of the function*

$$f'(\underline{x}) = \nabla f(\underline{x}) \triangleq \begin{bmatrix} \frac{\partial f}{\partial x_1}(\underline{x}) & \frac{\partial f}{\partial x_2}(\underline{x}) & \cdots & \frac{\partial f}{\partial x_n}(\underline{x}) \end{bmatrix}.$$

## 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 $f\colon V \to F$ 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 $f\colon X \to \mathbb{R}$ 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 $x \in X$, then the local minimum is also a **global minimum value**.*

**Theorem 5.2.3.** *Let $(X, \|\cdot\|)$ be a normed vector space and $f\colon X \to \mathbb{R}$ 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 5.1.2 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 no be no $\epsilon > 0$ satisfying the definition of a local minimum value in Definition 5.2.2. □

## 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.** *A Banach space $X$ is called **strictly convex** if the unit ball, given by $\{x \in X | \|x\| \leq 1\}$, 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 $s \in F$.*

**Definition 5.3.2.** *Let $V$ be a vector space, $A \subseteq V$ be a convex set, and $f\colon V \to \mathbb{R}$ be a functional. Then, the 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 $\underline{a}_1 = \underline{a}_2$.*

**Example 5.3.3.** *Let $(X, \|\cdot\|)$ be a normed vector space. Then, the norm $\|\cdot\|\colon X \to \mathbb{R}$ is a convex functional on $X$. Try proving this as an exercise.*

**Theorem 5.3.4.** *Let $(X, \|\cdot\|)$ be a normed vector space, $A \subseteq X$ be a convex set, and $f\colon X \to \mathbb{R}$ 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

$$f(\underline{x}) = f\left((1-\lambda)\underline{x}_0 + \lambda\underline{x}_1\right) \leq (1-\lambda)f(\underline{x}_0) + \lambda f(\underline{x}_1) < f(\underline{x}_0).$$

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

$$f\left((1-\lambda)\underline{x}_0 + \lambda\underline{x}_1\right) < (1-\lambda)f(\underline{x}_0) + \lambda f(\underline{x}_1) = f(\underline{x}_0).$$

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. □

**Theorem 5.3.5.** *Let $(X, \|\cdot\|)$ be a normed vector space and $f\colon X \to \mathbb{R}$ be a convex functional on a convex set $A \subseteq X$. If $f$ is Gâteuax 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

$$f\left(\underline{x}_0 + \lambda(\underline{x} - \underline{x}_0)\right) \leq f(\underline{x}_0) + \lambda\left(f(\underline{x}) - f(\underline{x}_0)\right) \tag{5.1}$$

for all $\lambda \in (0,1)$. Also, if $f$ is strictly convex, then (5.1) strict for $\underline{x} \neq \underline{x}_0$. Thus,

$$f(\underline{x}) \geq f(\underline{x}_0) + \frac{f\left(\underline{x} + \lambda(\underline{x} - \underline{x}_0)\right) - f(\underline{x}_0)}{\lambda}$$

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 (5.1) is a strict inequality for $\underline{x} \neq \underline{x}_0$. □

**Corollary 5.3.6.** *Let $(X, \|\cdot\|)$ be a normed vector space and $f\colon X \to \mathbb{R}$ be a convex functional on a convex set $A \subseteq X$. If $f$ is Gâteuax 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 $\mathbb{R}^n$ can be put into the following **standard form**. Let $f_i\colon \mathcal{D} \to \mathbb{R}$ and $h_j\colon \mathcal{D} \to \mathbb{R}$ be a real functionals on $\mathcal{D} \subseteq \mathbb{R}^n$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, p$. Then, the standard form is

$$\begin{aligned} \text{minimize} \quad & f_0(\underline{x}) \\ \text{subject to} \quad & f_i(\underline{x}) \leq 0, \quad i = 1, 2, \ldots, m \\ & h_j(\underline{x}) = 0, \quad j = 1, 2, \ldots, p. \end{aligned}$$

The function $f_0$ is called the **objective function** while the functions $f_1, \ldots, f_m$ are called inequality constraints and the functions $h_1, \ldots, h_p$ are called equality constraints.

**Definition 5.4.1.** *A vector $\underline{x} \in \mathcal{D}$ is **feasible** if it satisfies the constraints. Let $\mathcal{F} = \{\underline{x} \,|\, f_i(\underline{x}) \leq 0, i = 1, 2, \ldots, m\,, h_j(\underline{x}) = 0, j = 1, \ldots, p\}$ be the set of feasible vectors. Then, the problem is feasible if $\mathcal{F} \neq \emptyset$.*

**Definition 5.4.2.** *The **optimal value** is*

$$p^* = \inf \{ f_0(\underline{x}) \,|\, \underline{x} \in \mathcal{F} \}.$$

*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 gives the trivial upper bound

$$p^* \leq f_0(\underline{x}) \; \forall \underline{x} \in \mathcal{F}.$$

The optimization of linear function with arbitrary affine equality and inequality constraints is called a **linear program**. Any linear program can be reduced to the following standard form by adding additional variables.

**Definition 5.4.3.** *The standard form of a linear program (LP) is given by*

$$\begin{aligned} \text{minimize} \quad & \underline{c}^T \underline{x} \\ \text{subject to} \quad & A\underline{x} = \underline{b} \\ & \underline{x} \succeq \underline{0}. \end{aligned}$$

### 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 non-negative cost $\lambda_i$ (resp. $\nu_j$) associated with violating the $i$-th inequality (resp. $j$-th equality) constraint.

**Definition 5.4.4.** *The **Lagrangian** $L \colon \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ 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 $\lambda_i$ is the **Lagrange multiplier** associated with the $i$-th inequality constraint and $\nu_j$ 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.*

**Theorem 5.4.6** (Karush-Kuhn-Tucker)**.** *Assume the functions $f_i$ and $h_j$ are continuously differentiable and let $A = \{i \in [m] \,|\, f_i(\underline{x}^*) = 0\}$ be the set of active constraints at $\underline{x}^*$. Then, $\underline{x}^*$ is locally optimal only if $\underline{\lambda}^* \geq 0$ and $\underline{\nu}^*$ exist such that*

$$\nabla f_0(\underline{x}^*) + \sum_{i \in A} \lambda_i^* \nabla f_i(\underline{x}^*) + \sum_{j=1}^{p} \nu_j^* \nabla h_j(\underline{x}^*) = \underline{0} \tag{5.2}$$

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 upon which it is based. First, we note that the negative gradient $-\nabla f_0(\underline{x}^*)$ gives the direction of steepest descent for the objective function.

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

$$f(\underline{x}(t)) = f(\underline{x}^*) + t\underline{y}^H \nabla f(\underline{x}^*) + o(t),$$

where $o(t) \to 0$ as $t \to 0$. If the problem is unconstrained (e.g., $m = p = 0$), then $\nabla f_0(\underline{x}^*)$ must be $\underline{0}$. Otherwise, one is guaranteed to reduce the function by choosing $\underline{y} = -\nabla f_0(\underline{x}^*)$ (e.g., see Lemma 5.1.2). 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 $\underline{x}(t)$ will be infeasible if $|\underline{y}^H \nabla h_j(\underline{x}^*)| > 0$. Thus, we certainly need $\underline{y}^H \nabla h_j(\underline{x}^*) = 0$ 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 $\underline{x}(t)$ will be infeasible if $\underline{y}^H \nabla f_i(\underline{x}^*) > 0$. Thus, we certainly need $\underline{y}^H \nabla f_i(\underline{x}^*) \leq 0$ 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 small perturbations of $\underline{x}^*$.

The geometric picture implied by Theorem 5.4.6 is that of a game where one would like to decrease the objective $f_0(\underline{x}^*)$ choosing $\underline{y}$ such that $\underline{y}^H \nabla f_0(\underline{x}^*) < 0$ 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

$$F = \left\{ \sum_{i \in A} \lambda_i \nabla f_i(\underline{x}^*) \,\middle|\, \lambda_i \geq 0, i \in A \right\}.$$

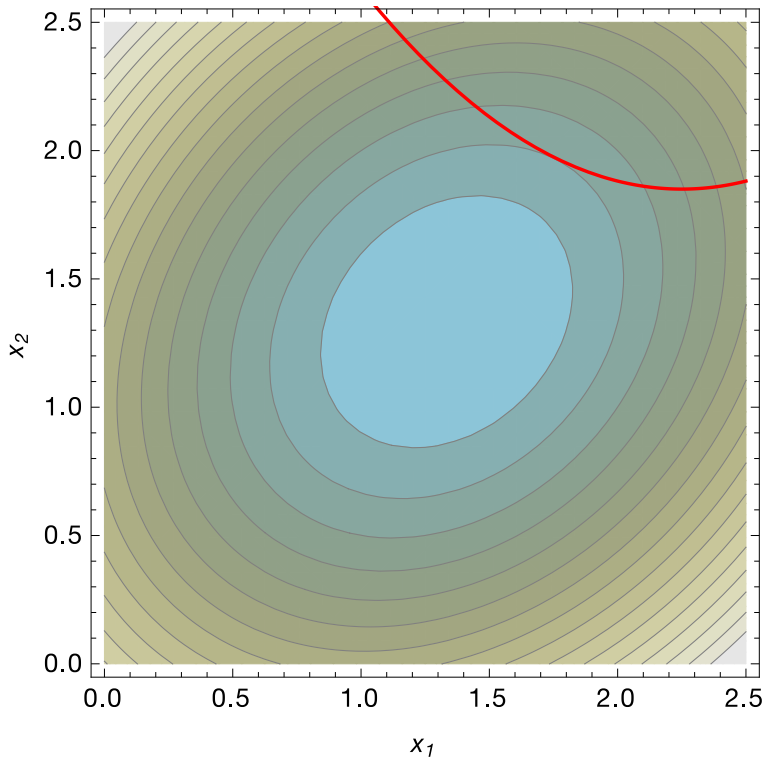


Figure 5.1: A contour plot of the function $f_0(x_1, x_2) = (x_1-1)^2+(x_2-1)^2-x_1x_2/2$ whose minimum occurs at $(4/3, 4/3)$ (i.e., the center of the blue ellipse). The red line indicates the inequality constraint $f_1(x_1, x_2) = 1.85+(x_1-2.25)^2/2-x_2 \leq 0$. The picture shows that the constrained minimum occurs at the intersection of the contour tangent line and the active constraint line.

Thus, one must at least pick a direction $\underline{y}$ that is orthogonal to all vectors in $H$ and has 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 (5.2) into the statement

$$-P\nabla f_0(\underline{x}^*) \in PF$$

or "the projection of the descent direction lies in the projection of the cone of directions that violate the inequality constraints". The reason for this is that we can absorb the $\nabla h_j$ terms into the $\nabla f_i$ 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 $i = 0, 1, \ldots, m$. Then, the cone $PF$ is defined by

$$PF = \left\{ \sum_{i \in A} \lambda_i \underline{f}^{(i)} \;\middle|\; \lambda_i \geq 0, i \in A \right\}.$$

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 and the constraints remain *almost* satisfied. The challenge in making this proof precise is that, unless the equality constraints are affine, the constraints may not be exactly satisfied for $t > 0$. In standard profos of this result, this difficulty is overcome by using the implicit function theorem to construct an $\underline{x}(t)$ that starts in the direction on $\underline{y}$ but is perturbed slightly to remain feasible.

*Proof.* For simplicity, we prove only the case where $h_j(\underline{x}) = \underline{a}_j^H \underline{x} - \underline{b}$ is affine and $PF$ does contain a line (i.e., $\{\alpha \underline{z} \,|\, \alpha \in \mathbb{R}\}$ for some $\underline{z}$). 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 $\underline{\nu}^*(\underline{\lambda}) = \arg\min_{\underline{\nu} \in \mathbb{R}^p} \|\underline{y}(\underline{\lambda}, \underline{\nu})\|_2$ and apply the best approximation theorem to see that

$$\underline{y}(\underline{\lambda}, \underline{\nu}^*(\underline{\lambda})) = P\underline{y}(\underline{\lambda}, \underline{\nu}),$$

where $P$ is orthogonal projection onto $H^\perp$ and $H = \operatorname{span}(\{\underline{a}_j\})$. 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 define $\underline{y}^* = \arg\min_{\underline{\lambda} \in \mathbb{R}^m, \underline{\lambda} \geq \underline{0}} \|\underline{y}(\underline{\lambda}, \underline{\nu}^*(\underline{\lambda}))\|_2$. This implies that $\underline{y}^*$ is the error vector for the projection of $-P\nabla f_0(\underline{x}^*)$ onto the convex set $PF$. The projection itself is given by $\underline{z} = -P\nabla f_0 - \underline{y}^*$ and Lemma 4.6.4 shows that $\left(\underline{y}^*\right)^H (\underline{z}) \geq 0$. Using this, we see that

$$\left(\underline{y}^*\right)^H \left(-P\nabla f_0(\underline{x}^*)\right) = \left(\underline{y}^*\right)^H \left(\underline{z} + \underline{y}^*\right) \geq \|\underline{y}^*\|_2^2.$$

If (5.2) cannot be satisfied by some $\underline{\lambda} \geq \underline{0}$ and $\underline{\nu}$, then $\underline{y}^* \neq \underline{0}$ and $\|\underline{y}^*\|_2 > 0$. This shows that $\underline{y}^*$ points in a direction that decreases the value of the objective function.

But, the $\underline{y}^*$ direction is only guaranteed to preserve feasibility to first order (i.e., $(\underline{y}^*)^H P\nabla f_i(\underline{x}^*) \leq 0$). To fix this, one can add to $\underline{y}^*$ a sufficiently small vector $\underline{w}$

satisfying $\underline{w}^H P \nabla f_i(\underline{x}^*) < 0$ for all $i = 1, 2, \ldots, m$. Such a $\underline{w}$ lies in the "interior of the polar cone of $PF$" and will exist as long as $PF$ does not contain a line. With this modification, the definition of the derivative implies that, for sufficiently small $t$, $\underline{x}(t) = \underline{x}^* + t(\underline{y}^* + \underline{w})$ will be a feasible vector satisfying $f_0(\underline{x}(t)) < f_0(\underline{x}^*)$. □

### 5.4.2 Lagrangian Duality

**Definition 5.4.7.** *The **Lagrangian dual** function is defined to be*

$$g(\underline{\lambda}, \underline{\nu}) = \inf_{\underline{x} \in \mathcal{D}} L(\underline{x}, \underline{\lambda}, \underline{\nu}).$$

**Lemma 5.4.8.** *The Lagrangian dual problem*

$$\begin{aligned} &\text{maximize} && g(\underline{\lambda}, \underline{\nu}) \\ &\text{subject to} && \underline{\lambda} \geq 0 \end{aligned}$$

*has a unique maximum value $d^* \geq p^*$. This property is known as **weak duality**.*

*Proof.* Since the Lagrangian dual function is the pointwise infimum of affine functions, it is always concave. Therefore, the function has a unique maximum value $d^*$. The upper bound on $d^*$ follows from

$$\begin{aligned} g(\underline{\lambda}, \underline{\nu}) &= \inf_{\underline{x} \in \mathcal{D}} L(\underline{x}, \underline{\lambda}, \underline{\nu}) \\ &\leq \inf_{\underline{x} \in \mathcal{F}} L(\underline{x}, \underline{\lambda}, \underline{\nu}) \\ &= p^* + \sum_{i=1}^{m} \lambda_i f_i(\underline{x}) \\ &\leq p^*. \end{aligned}$$

□

The Lagrangian dual function can be $-\infty$ for a wide range of $(\underline{\lambda}, \underline{\nu})$. In this case, it makes sense to define the set of **implicit constraints** given by

$$\mathcal{C} = \{(\underline{\lambda}, \underline{\nu}) | g(\underline{\lambda}, \underline{\nu}) > -\infty\}.$$

The points $(\underline{\lambda}, \underline{\nu}) \in \mathcal{C}$ are called **dual feasible**.

**Definition 5.4.9.** *If $d^* = p^*$, then one says that **strong duality** holds for the problem.*

**Example 5.4.10.** *The Lagrangian of a linear program (LP) in standard form is*

$$L(\underline{x}, \underline{\lambda}, \underline{\nu}) = \underline{c}^T \underline{x} + \underline{\nu}^T (A\underline{x} - \underline{b}) - \underline{\lambda}^T \underline{x}$$

*and the Lagrangian dual function is given by*

$$g(\underline{\lambda}, \underline{\nu}) = \inf_{\underline{x} \in \mathcal{D}} L(\underline{x}, \underline{\lambda}, \underline{\nu}) = \begin{cases} -\underline{b}^T \underline{\nu} & \text{if } A^T \underline{\nu} - \underline{\lambda} + \underline{c} = \underline{0} \\ -\infty & \text{otherwise.} \end{cases}$$

*Solving the implicit constraint for $\underline{\lambda}$ and using the fact that $\underline{\lambda} \succeq \underline{0}$ gives the standard form of the LP dual problem*

$$\begin{aligned} \text{maximize} \quad & -\underline{b}^T \underline{\nu} \\ \text{subject to} \quad & A^T \underline{\nu} + c \succeq \underline{0}. \end{aligned}$$

*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 5.4.11.** *An optimization problem in standard form is called **convex** if the function $f_i$ is convex for $i = 0, 1, \ldots, m$, the function $h_j$ is affine (i.e., $h_j(\underline{x}) = \underline{a}_j^T \underline{x} - b_j$) for $j = 1, 2, \ldots, p$, and $\mathcal{D} = \mathbb{R}^n$.*

Applying Theorem 5.3.4 to this setup shows that a standard convex optimization problem has a unique minimum value. Also, if the function $f_0$ is strictly convex, then the minimum value 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 [BV04, Sec. 5.3.2].

**Theorem 5.4.12** (Slater's Condition)**.** *If a convex optimization problem has a point $\underline{x}_0$ where $f_i(\underline{x}_0) < 0$ for $i = 1, \ldots, m$ and $h_j(\underline{x}_0) = 0$ for $j = 1, \ldots, p$, then strong duality holds for the problem.*

**Example 5.4.13.** *For a channel with colored noise, the input distribution that maximizes the achievable information rate can be found by solving the convex optimiza-*

*tion problem, known as water-filling, given by*

$$\begin{aligned} \text{minimize} \quad & -\sum_{i=1}^{n} \log\left(x_i + \alpha_i\right) \\ \text{subject to} \quad & \sum_{i=1}^{n} x_i = P \\ & \underline{x} \succeq 0. \end{aligned}$$

*Choosing $x_i = \frac{P}{n}$ for $i = 1, \ldots, n$ gives a point that satisfies Slater's condition, so strong duality holds for this problem.*

**Example 5.4.14.** *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 $x_i \to -\infty$. Thus, the system is implicitly constrained to have $\lambda_i \geq 0$. The first-order optimality conditions, for $i = 1, 2, \ldots, n$, 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 $\lambda_i$ (for $\lambda_i \geq 0$) and this implies that $g(\underline{\lambda}, \nu)$ is decreasing in $\lambda_i$ (for $\lambda_i \geq 0$ and $x_i \geq 0$).*

*Thus, the expression $\max_{\underline{\lambda} \geq 0} g(\underline{\lambda}, \nu)$ is given by choosing the smallest non-negative $\lambda_i$'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 $\nu$ 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 $\nu$ values and computing $P$ in terms of $\nu$.*

# 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. □

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. □

**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.** *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 6.1.6.** *Consider the vector space $V$ of semi-infinite real sequences $\mathbb{R}^\omega$ where $\underline{v} = (v_1, v_2, v_3, \ldots) \in V$ with $v_n \in \mathbb{R}$ for $n \in \mathbb{N}$. Let $L\colon V \to V$ be the left-shift linear transformation defined by*

$$L\underline{v} = (v_2, v_3, v_4, \ldots)$$

*and $R\colon V \to V$ be the right-shift linear transformation defined by*

$$R\underline{v} = (0, v_1, v_2, \ldots).$$

*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 6.1.7.** *Consider the normed vector space $V$ of semi-infinite real sequences $\mathbb{R}^\omega$ with the standard Schauder basis $\{\underline{e}_1, \underline{e}_2, \ldots\}$. Let $T\colon V \to V$ be the linear transformation that satisfies $T\underline{e}_i = i^{-1}\underline{e}_i$ for $i = 1, 2, \ldots$. Let the linear transformation $U\colon V \to V$ satisfy $U\underline{e}_i = i\underline{e}_i$ for $i = 1, 2, \ldots$. 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 6.1.8.** *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

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

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

$$T^{-1}\left(s\underline{w}_1 + \underline{w}_2\right) = s\underline{v}_1 + \underline{v}_2 = s\left(T^{-1}\underline{w}_1\right) + T^{-1}\underline{w}_2$$

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 operation, 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 $\mathcal{B} = \underline{v}_1, \ldots, \underline{v}_n$ be a basis for $V$. There is a unique dual basis $\mathcal{B}^* = f_1, \ldots, f_n$ for $V^*$ such that $f_j\left(\underline{v}_i\right) = \delta_{ij}$. For each linear functional on $V$, we have*

$$f = \sum_{i=1}^{n} f\left(\underline{v}_i\right) f_i$$

*and for each vector $\underline{v}$ in $V$, we have*

$$\underline{v} = \sum_{i=1}^{n} f_i\left(\underline{v}\right) \underline{v}_i.$$

*Proof.* Let $\mathcal{B} = \underline{v}_1, \ldots, \underline{v}_n$ be a basis for $V$. According to Theorem 3.4.7, there is a unique linear functional $f_i$ on $V$ such that

$$f_i\left(\underline{v}_j\right) = \delta_{ij}.$$

Thus, we obtain from $\mathcal{B}$ a set of $n$ distinct linear functionals $f_1, \ldots, f_n$ on $V$. These functionals are linearly independent; suppose that

$$f = \sum_{i=1}^{n} s_i f_i,$$

then

$$f\left(\underline{v}_j\right) = \sum_{i=1}^{n} s_i f_i\left(\underline{v}_j\right) = \sum_{i=1}^{n} s_i \delta_{ij} = s_j.$$

In particular, if $f$ is the zero functional, $f\left(\underline{v}_j\right) = 0$ for $j = 1, \ldots, n$ and hence the scalars $\{s_j\}$ must all equal $0$. It follows that the functionals $f_1, \ldots, f_n$ are linearly independent. Since $\dim V^* = n$, we conclude that $\mathcal{B}^* = f_1, \ldots, f_n$ 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 $f_1, \ldots, f_n$ with

$$f = \sum_{i=1}^{n} s_i f_i.$$

Furthermore, by construction, we must have $s_j = f\left(\underline{v}_j\right)$ for $j = 1, \ldots, n$. Similarly, if

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

is a vector in $V$, then

$$f_j\left(\underline{v}\right) = \sum_{i=1}^{n} t_i f_j\left(\underline{v}_i\right) = \sum_{i=1}^{n} t_i \delta_{ij} = t_j.$$

That is, the unique expression for $\underline{v}$ as a linear combination of $\underline{v}_1, \ldots, \underline{v}_n$ is

$$\underline{v} = \sum_{i=1}^{n} f_i\left(\underline{v}\right) \underline{v}_i.$$

□

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 $T\colon V \to W$ be a linear transform. If $g \in W^*$ is a linear functional on W (i.e., $g\colon W \to F$), then $g(T\underline{v}) \in V^*$ is a linear functional on $V$. The **transpose** of $T$ is the mapping $U\colon W^* \to V^*$ 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 $U\colon W \to V$ 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 identify $V = V^*$ via isomorphism. The interesting case that does not simplify is that of linear transforms between infinite dimensional Banach spaces.

## 6.3 Operator Norms

Intuitively, the operator norm is the largest factor by which a linear transform can increase the length of a vector. This provides a simple "worst-case" characterization of any linear transform.

**Definition 6.3.1.** *Let $V$ and $W$ be two normed vector spaces and let $T\colon V \to W$ be a linear transformation. The induced **operator norm** of $T$ is defined to*

$$\|T\| = \sup_{\underline{v} \in V - \{\underline{0}\}} \frac{\|T\underline{v}\|}{\|\underline{v}\|} = \sup_{\underline{v} \in V, \|\underline{v}\| = 1} \|T\underline{v}\| .$$

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 standard conditions required by a norm.

This norm also has a new property that follows easily from the definition. The induced operator norm is called **submultiplicative** because $\|T\underline{v}\| \leq \|T\|\|\underline{v}\|$. From this, it is easy to see that it also provides a submultiplicative norm for the algebra of linear operators in that

$$\|UT\underline{v}\| \leq \|U\| \, \|T\underline{v}\| \leq \|U\| \, \|T\| \, \|\underline{v}\| .$$

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

$$\sup_{\underline{v}\in V-\{\underline{0}\}} \frac{\|T\underline{v}\|}{\|\underline{v}\|} = \sup_{\underline{v}\in V-\{\underline{0}\}} \left\| T\frac{\underline{v}}{\|\underline{v}\|} \right\| = \sup_{\underline{u}\in V, \|\underline{u}\|=1} \|T\underline{u}\| .$$

### 6.3.1 Bounded Transformations

**Definition 6.3.2.** *If the norm of a linear transformation is finite, then the transformation is said to be* ***bounded****.*

**Theorem 6.3.3.** *A linear transformation $T\colon V \to W$ 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 $\underline{v}_1, \underline{v}_2, \ldots$ be a convergent sequence in $V$, then

$$\left\|T\underline{v}_i - T\underline{v}_j\right\| = \left\|T\left(\underline{v}_i - \underline{v}_j\right)\right\| \leq M \left\|\underline{v}_i - \underline{v}_j\right\| .$$

This implies that $T\underline{v}_1, T\underline{v}_2, \ldots$ 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 $T\colon V \to W$ be a linear transformation. If $V$ is finite dimensional, then $T$ is continuous and bounded.*

**Lemma 6.3.5.** *Let $V$ be a finite-dimensional normed vector space, and let*

$$\mathcal{B} = \underline{v}_1, \ldots, \underline{v}_n$$

*be a basis for $V$. Then, for $\underline{v} \in V$, each coefficient $s_i$ in the expansion*

$$\underline{v} = s_1\underline{v}_1 + \cdots + s_n\underline{v}_n$$

*is a continuous linear function of $\underline{v}$. Being continuous, it is also bounded, so there exists a constant $M$ such that $|s_i| \leq M \|\underline{v}\|$.*

*Proof of Lemma.* The linearity property is straightforward, its proof is omitted. It will suffice to show that there is an $m > 0$ such that

$$m|s_i| \leq m\left(|s_1| + \cdots + |s_n|\right) \leq \|\underline{v}\|, \tag{6.1}$$

since (6.1) implies that $|s_i| \leq m^{-1}\|\underline{v}\|$. We first show that this holds for coefficients $\{s_1, \ldots, s_n\}$ satisfying the condition $|s_1| + \cdots + |s_n| = 1$. Let

$$S = \left\{ (s_1, \ldots, s_n) \middle| \sum_{i=1}^{n} |s_i| = 1 \right\}.$$

This set is closed and bounded; it is therefore compact. Define the function $f\colon S \to \mathbb{R}$ by

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

It can be shown that $f$ is continuous, and it is clear that $f > 0$ over $S$. Let

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

Since $f$ is continuous and $S$ is compact, this minimum exists and is attained by some point $(s_1', \ldots, s_n') \in S$. Note that $m > 0$ for otherwise $\underline{v}_1, \ldots, \underline{v}_n$ are linearly dependent, contradicting the fact that $\mathcal{B}$ is a basis. Thus $m$ so defined satisfies (6.1).

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

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

This is the desired result. □

We are now ready to prove the theorem.

*Proof of Theorem.* Let $\mathcal{B} = \underline{v}_1, \ldots, \underline{v}_n$ be a basis for $V$. Let $\underline{v} \in V$ be expressed in terms of this basis as

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

Let $C = \max_{1 \leq i \leq n} \|T\underline{v}_i\|$. Then,

$$\begin{aligned} \|T\underline{v}\| &= \|T\left(s_1\underline{v}_1 + \cdots + s_n\underline{v}_n\right)\| \\ &\leq |s_1| \, \|T\underline{v}_1\| + \cdots + |s_n| \, \|T\underline{v}_n\| \\ &\leq C\left(|s_1| + \cdots + |s_n|\right). \end{aligned}$$

By the previous lemma, this implies that there exists an $M$ such that $|s_1| + \cdots + |s_n| \leq M \|\underline{v}\|$, so that

$$\|T\underline{v}\| \leq CM \|\underline{v}\|.$$

□

### 6.3.2 The Neumann Expansion

**Theorem 6.3.6.** *Let $\|\cdot\|$ be a submultiplicative operator norm and $T\colon V \to V$ 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 $m \to \infty$, we see that the limit $\lim_{n\to\infty} A_n$ exists.

Next, we observe that

$$(I - T)\left(I + T + T^2 + \cdots + T^{n-1}\right) = I - T^n.$$

Since $\|T\| < 1$, we have $\lim_{k\to\infty} T^k = 0$ because $\left\|T^k\right\| \leq \|T\|^k \to 0$. Taking the limit $n \to \infty$ of both sides gives

$$(I - T)\sum_{i=0}^{\infty} T^i = \lim_{n\to\infty}(I - T^n) = I.$$

Likewise, reversing the order multiplication results in the same result. This shows that $\sum_{i=0}^{\infty} T^i$ must be the inverse of $I - T$. □

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, $\lambda$ must be the largest eigenvalue of $A^H A$. For an $n\times n$ matrix $B$ with eigenvalues $\lambda_1, \ldots, \lambda_n$, 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 $\|A\|_F^2 = \mathrm{tr}(A^H A)$.

## 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 $\underline{v} - \underline{u} = \underline{0}$. □

**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 $\underline{v}_1, \underline{v}_2, \ldots$ 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 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

$$\begin{aligned} M\|\underline{x}_n\| \geq |f(\underline{x}_n)| &= \left|\sum_{i=1}^{n} \overline{f(\underline{v}_i)} f(\underline{v}_i)\right| = \sum_{i=1}^{n} |f(\underline{v}_i)|^2 = \sum_{i=1}^{n} f(\underline{v}_i)\,\overline{f(\underline{v}_i)} \\ &= \sum_{i=1}^{n} \left\langle \overline{f(\underline{v}_i)}\underline{v}_i \middle| \overline{f(\underline{v}_i)}\underline{v}_i \right\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} \left\langle \overline{f(\underline{v}_j)}\underline{v}_j \middle| \overline{f(\underline{v}_i)}\underline{v}_i \right\rangle \\ &= \left\langle \sum_{j=1}^{n} \overline{f(\underline{v}_j)}\underline{v}_j \middle| \sum_{i=1}^{n} \overline{f(\underline{v}_i)}\underline{v}_i \right\rangle = \langle \underline{x}_n | \underline{x}_n \rangle = \|\underline{x}_n\|^2. \end{aligned}$$

This implies that $\|\underline{x}_n\| \leq M$ for all $n$. Hence, $\lim_{n\to\infty} \sum_{i=1}^{n} |f(\underline{v}_i)|^2$ 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

$$\|f_{\underline{v}}\| \triangleq \sup_{\underline{u} \in V - \{\underline{0}\}} \frac{f_{\underline{v}}(\underline{u})}{\|\underline{u}\|} = \|\underline{v}\|.$$

Since $f$ is continuous, it follows that $\|f\| < \infty$ and $\|\underline{v}\| < \infty$. Then,

$$f_{\underline{v}}(\underline{v}_j) = \left\langle \underline{v}_j \middle| \sum_{i=1}^{\infty} \overline{f(\underline{v}_i)}\underline{v}_i \right\rangle = f(\underline{v}_j).$$

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. □

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 $R\colon V^* \to V$ 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 $R(sf_1 + f_2) = \overline{s}R(f_1) + R(f_2)$ 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 $T\colon V \to W$ 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 $T^*\colon W \to V$ 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 $\underline{v} \in V$,

$$\begin{aligned}
\langle \underline{v}|T^*\left(s\underline{w}_1 + \underline{w}_2\right)\rangle &= \langle T\underline{v}|\left(s\underline{w}_1 + \underline{w}_2\right)\rangle \\
&= \bar{s}\,\langle T\underline{v}|\underline{w}_1\rangle + \langle T\underline{v}|\underline{w}_2\rangle \\
&= \bar{s}\,\langle \underline{v}|T^*\underline{w}_1\rangle + \langle \underline{v}|T^*\underline{w}_2\rangle \\
&= \langle \underline{v}|sT^*\underline{w}_1\rangle + \langle \underline{v}|T^*\underline{w}_2\rangle \\
&= \langle \underline{v}|sT^*\underline{w}_1 + T^*\underline{w}_2\rangle\,.
\end{aligned}$$

Since this holds for all $\underline{v} \in V$, we gather from Lemma 6.4.1 that $T^*\left(s\underline{v}_1 + \underline{v}_2\right) = sT^*\underline{v}_1 + T^*\underline{v}_2$. 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 $\underline{v} \in V$. □

**Theorem 6.4.4.** *Let $V$ be a finite-dimensional inner-product space and let*

$$\mathcal{B} = \underline{v}_1, \ldots, \underline{v}_n$$

*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 $A_{kj} = \left\langle T\underline{v}_j|\underline{v}_k\right\rangle$.*

*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 $A_{kj} = \langle T\underline{v}_j | \underline{v}_k \rangle$. □

**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 $\mathcal{B} = \underline{v}_1, \ldots, \underline{v}_n$ be an orthonormal basis for $V$, let $A = [T]_{\mathcal{B}}$ and $B = [T^*]_{\mathcal{B}}$. According to the previous theorem,

$$A_{kj} = \langle T\underline{v}_j | \underline{v}_k \rangle$$
$$B_{kj} = \langle T^*\underline{v}_j | \underline{v}_k \rangle$$

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 $T\colon V \to W$. 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 $T\colon V \to W$ can be summarized as follows,

$$R_T \subseteq W$$
$$N_T \subseteq V$$
$$R_{T^*} \subseteq V$$
$$N_{T^*} \subseteq W.$$

**Theorem 6.5.1.** *Let $T\colon V \to W$ be a bounded linear transformation between two Hilbert spaces $V$ and $W$, and let $R_T$ and $R_{T^*}$ be 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*

$$\overline{R_T} = R_T = [N_{T^*}]^\perp$$
$$\overline{R_{T^*}} = R_{T^*} = [N_T]^\perp .$$

*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 $[R_{T^*}]^\perp = N_T$. □

## 6.6 Pseudoinverses

**Theorem 6.6.1.** *Let $T$ be a bounded linear transformation from $V$ to $W$. 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 $\underline{u} \in N_{T^*}$. Then

$$\langle \underline{w} | \underline{u} \rangle = \langle T\underline{v} | \underline{u} \rangle = \langle \underline{v} | T^* \underline{u} \rangle = \langle \underline{v} | \underline{0} \rangle = 0.$$

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 $\underline{w} \notin R_T$, then

$$\underline{w}_o = \underline{w} - P_{R_T} \underline{w} = \underline{w} - \underline{w}_r \neq \underline{0}.$$

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 contradict 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. □

**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 $T \colon V \to W$ 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 $\|\underline{w} - \underline{\hat{w}}\|$, where $\underline{\hat{w}} = T\underline{v} \in R_T$. By the projection theorem, we must have

$$\underline{w} - \underline{\hat{w}} \in [R_T]^{\perp}.$$

But this is equivalent to

$$\underline{w} - \underline{\hat{w}} \in N_{T^*}.$$

That is, $T^*\left(\underline{w} - \underline{\hat{w}}\right) = \underline{0}$, or equivalently $T^*\underline{w} = T^*\underline{\hat{w}}$. Conversely, if $T^*T\underline{v} = T^*\underline{w}$, then

$$T^*\left(T\underline{v} - \underline{w}\right) = \underline{0},$$

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. □

**Corollary 6.6.4.** *If $A$ is a matrix such that $A^HA$ is invertible, then the least-squares solution to $A\underline{v} = \underline{w}$ is*

$$\underline{v} = \left(A^HA\right)^{-1}A^H\underline{w}.$$

The matrix $\left(A^HA\right)^{-1}A^H$ is the left inverse of $A$ and is an example of a Moore-Penrose **pseudoinverse**.

**Theorem 6.6.5.** *Suppose the vector $\underline{\hat{v}} \in V$ minimizes $\|\underline{v}\|$ over all $\underline{v} \in V$ satisfying $T\underline{v} = \underline{w}$. Then, $\underline{\hat{v}} \in [N_T]^\perp$ and, if $R_{T^*}$ is closed, $\underline{\hat{v}} = T^*\underline{u}$ for some $\underline{u} \in W$.*

*Proof.* Suppose $\underline{\hat{v}} \notin [N_T]^\perp$, then the orthogonal decomposition $V = [N_T]^\perp + N_T$ shows that the projection of $\underline{\hat{v}}$ onto $[N_T]^\perp$ has smaller norm but still satisfies $T\underline{\hat{v}} = \underline{w}$. This gives a contradiction and shows that $\underline{\hat{v}} \in [N_T]^\perp$. If $R_{T^*}$ is closed, then $R_{T^*} = [N_T]^\perp$ and $\underline{\hat{v}} = T^*\underline{u}$ for some $\underline{u} \in W$. □

**Corollary 6.6.6.** *If $A$ is a matrix such that $AA^H$ is invertible, then the minimum-norm solution to $A\underline{v} = \underline{w}$ is*

$$\underline{v} = A^H\left(AA^H\right)^{-1}\underline{w}.$$

*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. □

The matrix $A^H\left(AA^H\right)^{-1}$ is the right inverse of $A$ and is another example of a Moore-Penrose **pseudoinverse**.

**Definition 6.6.7.** *Let $T\colon V \to W$ 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 $\underline{\hat{v}}$ 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 $\underline{\hat{v}}$.*

# 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

$$y_j = \frac{1}{l_{jj}} \left( b_j - \sum_{i=1}^{j-1} l_{ji} y_i \right), \quad j = 1, 2, \ldots, n.$$

**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

$$x_j = \frac{1}{u_{jj}} \left( y_j - \sum_{i=j+1}^{n} u_{ji} x_i \right), \quad j = n, n-1, \ldots, 1.$$

**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. □

## 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.

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 2 & 8 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix}$$

This allows one to write

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix}$$

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 $i, j \in \{1, 2, \ldots, n\}$. Then, adding $\alpha$ 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:*

$$\begin{aligned} &(i) \;\; E_{ij}E_{kl} = \delta_{j,k}E_{il} \\ &(ii) \;\; (I + \alpha E_{ij})(I + \beta E_{kl}) = I + \alpha E_{ij} + \beta E_{kl} \quad \textit{if } j \neq k \\ &(iii) \;\; (I + \alpha E_{ij})^{-1} = (I - \alpha E_{ij}) \quad \textit{if } i \neq j. \end{aligned}$$

*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 $L = L_1 L_2 \cdots L_{n-1}$ 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 $j = 1, 2, \ldots, n-1$. 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. □

**Theorem 7.2.5.** *This process generates one column of $L$ per round because*

$$[L]_{ij} = \begin{cases} \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}} \textit{ if } 1 \leq i < j \\ 1 \textit{ if } i = j \\ 0 \textit{ otherwise}. \end{cases}$$

*Proof.* First, we note that

$$\begin{aligned} L &= L_1 L_2 \cdots L_{n-1} \\ &= \prod_{j=1}^{n-1} \left( \prod_{i=j+1}^{n} \left( I - \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}} E_{i,j} \right) \right)^{-1} \\ &\overset{(a)}{=} \prod_{i=1}^{n-1} \prod_{i=j+1}^{n} \left( I - \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}} E_{i,j} \right)^{-1} \\ &\overset{(b)}{=} \prod_{i=1}^{n-1} \prod_{i=j+1}^{n} \left( I + \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}} E_{i,j} \right) \\ &= I + \sum_{i=1}^{n-1} \sum_{i=j+1}^{n} \frac{a_{i,j}^{(j)}}{a_{j,j}^{(j)}} E_{i,j}, \end{aligned}$$

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. □

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

$$p_j = \arg \max_{i=j,j+1,\ldots,n} |a_{i,j}^{(j)}|.$$

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 $A = \left(LD^{1/2}\right)\left(LD^{1/2}\right)^H$. 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 $\widetilde{L} = LD^{1/2}$. This gives the Cholesky decomposition $\widetilde{L}\widetilde{L}^H = LDL^H = A$.

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 $\underline{Y} = \widetilde{L}^{-1}\underline{X}$ and see that

$$E[\underline{Y}\underline{Y}^H] = E[\widetilde{L}^{-1}\underline{X}\underline{X}^H\widetilde{L}^{-H}] = \widetilde{L}^{-1}R\widetilde{L}^{-H} = I.$$

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 $Q^HQ = QQ^H = I$. If all elements of the matrix are real, then it is called **orthogonal** and $Q^TQ = QQ^T = I$.

**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, $QQ^H = I$, 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 $\underline{a}_1, \ldots, \underline{a}_n$ of $A$. This results in orthonormal vectors

$\{\underline{q}_1, \ldots, \underline{q}_l\}$, where $l = \min(m, n)$, such that

$$\underline{a}_j = \sum_{i=1}^{\min(j,l)} r_{i,j}\underline{q}_i \quad \text{for } j = 1, 2, \ldots, n.$$

This gives an $m \times l$ matrix $Q = [\underline{q}_1 \ \ldots \ \underline{q}_l]$ 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 $\{\underline{q}_1, \ldots, \underline{q}_l\}$ to an orthonormal basis $\{\underline{q}_1, \ldots, \underline{q}_m\}$ of $\mathbb{C}^m$. This gives an $m \times m$ unitary matrix $Q' = [\underline{q}_1 \ \ldots \ \underline{q}_m]$. Adding $m - n$ rows of zeros to the previous $R$ matrix gives an $m \times n$ matrix $R'$ such that $A = Q'R'$. □

## 7.4 Hermitian Matrices and Complex Numbers

**Definition 7.4.1.** *A square matrix $Q \in \mathbb{R}^{n\times n}$ is **orthogonal** if $Q^TQ = QQ^T = I$.*

**Definition 7.4.2.** *A square matrix $U \in \mathbb{C}^{n\times n}$ is **unitary** if $U^HU = UU^H = I$.*

It is worth noting that, for unitary (resp. orthogonal) matrices, it suffices to check only that $U^HU = I$ (resp. $Q^TQ = I$) because $U$ is invertible (e.g., it has linearly independent columns) and

$$U^HU = I \implies I = UU^{-1} = U(U^HU)U^{-1} = UU^H.$$

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 $U^HU = I$ are analogous to complex numbers on the unit circle, satisfying $\bar{z}z = 1$.
- *Orthogonal matrices* satisfying $Q^TQ = I$ 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 $\lambda$.*

**Definition 8.1.2.** *The **spectrum** $\sigma(T)$ of a linear operator $T\colon V \to V$ 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 $T : V \to V$ be the right-shift operator defined by*

$$T(v_1, v_2, \ldots) = (0, v_1, v_2, \ldots).$$

*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 $\underline{v} = 0$ 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 $\lambda$ be a scalar. The following are equivalent:*

1. *$\lambda$ 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 third are equivalent. If $\lambda$ is an eigenvalue of $A$, then there exists a 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 $\underline{v} \in V$ 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. □

The last criterion is important. It implies that every eigenvalue $\lambda$ 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 $\lambda$ 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 $\lambda$ be an eigenvalue of the matrix $A$. The **eigenspace** associated with $\lambda$ is the set $E_\lambda = \{\underline{v} \in V | A\underline{v} = \lambda\underline{v}\}$. The **algebraic multiplicity** of $\lambda$ is the multiplicity of the zero at $t = \lambda$ in the characteristic polynomial $\chi_A(t)$. The **geometric multiplicity** of an eigenvalue $\lambda$ is equal to dimension of the eigenspace $E_\lambda$ 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 $\lambda_1, \lambda_2, \ldots, \lambda_k$ are distinct eigenvalues with eigenvectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_k$, then the eigenvectors are linearly independent. Suppose that

$$\sum_{i=1}^{k} c_i \underline{v}_i = \underline{0}$$

for scalars $c_1, c_2, \ldots, c_k$. 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

$$\prod_{j=1,j\neq m}^{k}(A-\lambda_j I)\sum_{i=1}^{k}c_j\underline{v}_i = \left(\prod_{j=1,j\neq m}^{k}(A-\lambda_j I)\right)c_m\underline{v}_m$$
$$= c_m\prod_{j=1,j\neq m}^{k}(\lambda_m-\lambda_j) = \underline{0}.$$

Since all eigenvalues are distinct, we must conclude that $c_m = 0$. Since the choice of $m$ was arbitrary, it follows that $c_1, c_2, \ldots, c_k$ are all zero. Therefore, the vectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_k$ are linearly independent. □

**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 $\Lambda$ 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 $\underline{v}_1, \ldots, \underline{v}_n$. Let the eigenvalue of $\underline{v}_i$ be denoted by $\lambda_i$ so that

$$A\underline{v}_j = \lambda_j\underline{v}_j, \quad j = 1, \ldots, n.$$

In matrix form, we have

$$A\begin{bmatrix} \underline{v}_1 & \cdots & \underline{v}_n \end{bmatrix} = \begin{bmatrix} A\underline{v}_1 & \cdots & A\underline{v}_n \end{bmatrix}$$
$$= \begin{bmatrix} \lambda_1\underline{v}_1 & \cdots & \lambda_n\underline{v}_n \end{bmatrix}.$$

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

$$A = S\Lambda S^{-1}$$
$$\Lambda = S^{-1}AS.$$

That is, the matrix $A$ is diagonalizable. □

The type of the transformation from $A$ to $\Lambda$ 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 association with distinct eigenvalues are orthogonal.*

*Proof.* First, we notice that $A = A^H$ implies $\underline{v}^H A \underline{v}$ is real because

$$\overline{s} = \left(\underline{v}^H A \underline{v}\right)^H = \underline{v}^H A^H \underline{v} = \underline{v}^H A \underline{v} = s.$$

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, $\lambda_1$ is real. Next, assume that $A\underline{w} = \lambda_2 \underline{w}$ and $\lambda_2 \neq \lambda_1$. Then, we have

$$\lambda_1 \lambda_2 \underline{w}^H \underline{v} = \underline{w}^H A^H A\underline{v} = \underline{w} A^2 \underline{v} = \lambda_1^2 \underline{w}^H \underline{v}.$$

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. □

# 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 $A = S\Lambda S^{-1}$ 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

$$\underline{x}(t) = Se^{\Lambda t}S^{-1}\underline{x}(0).$$

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 $\Lambda^n$ 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 $e^A$,

$$e^A = \sum_{i=0}^{\infty}\frac{A^i}{i!} = S\left(\sum_{i=0}^{\infty}\frac{\Lambda^i}{i!}\right)S^{-1} = Se^{\Lambda}S^{-1},$$

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 $\lambda$ 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 $\underline{v}_1, \ldots, \underline{v}_n$, then let $S$ be an invertible matrix whose columns are there $n$ 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 $\Lambda$ 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 $\Lambda$. 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 $\lambda_1, \ldots, \lambda_l$ 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

$$J_m(0)\underline{e}_{j+1} = \underline{e}_j \text{ for } j = 1, 2, \ldots, m-1.$$

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 $\lambda$, the **generalized eigenspace**

$$G_\lambda = \left\{ \underline{v} \in \mathbb{C}^n \middle| (A - \lambda I)^n \underline{v} = \underline{0} \right\}.$$

Let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of $A$ ordered by decreasing magnitude. Let $d_j$ be the dimension of $G_{\lambda_j}$, which is equal to the sum of the sizes of the Jordan

blocks associated with $\lambda$, then $\sum_{j=1}^{k} d_j = n$. Let $T$ be a matrix whose first $d_1$ columns for a basis for $G_{\lambda_1}$, next $d_2$ columns form a basis for $G_{\lambda_2}$, 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 $\lambda_j$.

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., $\mathcal{N}\left((B-\lambda I)^{j-1})^{\perp}\right)$. This gives the sequence $V_1, \ldots, V_J$ of non-empty subspaces (e.g., $V_j$ is empty for $j > J$). Now, we can form a sequence of bases $W_J, W_{J-1}, \ldots, W_1$ recursively starting from $W_J$ with

$$W_j = W_{j+1} \cup \{(B-\lambda I)\underline{w} | \underline{w} \in W_{j+1}\} \cup \text{basis}(V_j - V_{j-1}),$$

where 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 $\underline{v}_1^{(i)}, \underline{v}_2^{(i)}, \ldots$ 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 $\lambda_1$ 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 $A = TJT^{-1}$ 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 $\|A^k\| \to 0$ 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 $i = 1, \ldots, m-1$ 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 $k^{m-1}|\lambda|^{k-m+1} \to 0$ as $k \to \infty$ 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 $|\left[(\lambda I + N)^k\right]_{1,1}| \geq 1$ for all $k \geq 0$. □

**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 $J_1, \ldots, J_k$. 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 $\lambda$ associated with each Jordan block satisfies $\|\lambda\| < 1$. In this case, the lemma shows that $\|J_i^k\| \to 0$ which implies that $\|J^k\| \to 0$. Therefore, $\|A^k\| \to 0$ 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 $k \geq 0$. □

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 $1 \times 1$ 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 $\lambda_1$ of maximum modulus, then*

$$\lim_{k\to\infty} \frac{1}{\lambda_1^k} A^k = \underline{u}\underline{v}^H,$$

*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 $B = TJT^{-1}$ 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 $n \to \infty$. This implies that $B^n = TJ^nT^{-1}$ 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 $\lambda_1$. Therefore, $\underline{v}^H A = \lambda_1 \underline{v}^H$. Finally, the fact that $\underline{u} = B^n \underline{u} \to \underline{u}\underline{v}^H \underline{u}$ implies that $\underline{v}^H \underline{u} = 1$. □

# 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 $k = 1, 2, \ldots, n-1$ 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 $\mathcal{B} = \underline{x}_1, \ldots, \underline{x}_n$ 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 $AU_n$ can be expressed as a linear combination of vectors in $\mathcal{B}$, say,

$$A\underline{x}_i = \sum_{j=1}^{n} s_{j,i}\underline{x}_j \quad i = 1, \ldots, n.$$

Note that $A\underline{x}_1 = \lambda_1 \underline{x}_1$ for some $\lambda_1$ since $\underline{x}_1 = \underline{u}$, an eigenvector of $A$. We can then

write

$$AU_n = \begin{bmatrix} A\underline{x}_1 & \cdots & A\underline{x}_n \end{bmatrix} = U_n \begin{bmatrix} \lambda_1 & s_{1,2} & \dots & s_{1,n} \\ 0 & s_{2,2} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & s_{n,2} & \cdots & s_{n,n} \end{bmatrix} = U_n \begin{bmatrix} \lambda_1 & \underline{s}^T \\ \underline{0} & A_{n-1} \end{bmatrix},$$

where we have used the convenient notation

$$A_{n-1} = \begin{bmatrix} s_{2,2} & \cdots & \underline{s}_{2,n} \\ \vdots & \ddots & \vdots \\ s_{n,2} & \cdots & s_{n,n} \end{bmatrix}$$

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. □

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 $\Lambda$ is a diagonal matrix.*

*Proof.* Note that $A^H = A$ and $T = U^H AU$. Consider the matrix $T^H$ given by

$$T^H = (U^H AU)^H = U^H A^H U = U^H AU = T.$$

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. □

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*

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p),$$

*where $p = \min(m, n)$.*

The diagonal elements of $\Sigma$ are called the *singular values* of $A$ and are typically ordered so that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0.$$

*Proof.* Let

$$A^H AV = V\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

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 $\lambda_1, \lambda_2, \cdots, \lambda_r > 0$ and $\lambda_{r+1} = \cdots = \lambda_n = 0$, where $r \leq p$. For $i \leq r$, let

$$\underline{u}_i = \frac{A\underline{v}_i}{\sqrt{\lambda_i}},$$

and observe that

$$\langle \underline{u}_i | \underline{u}_j \rangle = \frac{\underline{v}_j^H A^H A \underline{v}_i}{\sqrt{\lambda_i \lambda_j}} = \frac{\underline{v}_j^H \underline{v}_i \lambda_i}{\sqrt{\lambda_i \lambda_j}} = \delta_{ij}.$$

Also note that $\{\underline{u}_i\}$ are eigenvectors of $AA^H$ since

$$AA^H \underline{u}_i = AA^H A \frac{\underline{v}_i}{\sqrt{\lambda_i}} = \sqrt{\lambda_i} A \underline{v}_i = \lambda_i \underline{u}_i.$$

The set $\{\underline{u}_i : i = 1, \ldots, r\}$ 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

$$\underline{u}_i^H AV = \frac{1}{\sqrt{\lambda_i}} \underline{v}_i^H A^H AV = \sqrt{\lambda_i} \underline{e}_i^H.$$

On the other hand, if $i > r$ then $\underline{u}_i^H AV = \underline{0}$. Hence,

$$U^H AV = \operatorname{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\right) = \Sigma,$$

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

$$A^H A = \begin{bmatrix} 27 & -9 \\ -9 & 27 \end{bmatrix} = V \Lambda V^H = \left( \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right) \begin{bmatrix} 18 & 0 \\ 0 & 36 \end{bmatrix} \left( \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right).$$

This implies that $\Sigma_1 = \Lambda^{1/2}$ and $V_1 = V$. Therefore, we can compute $U_1 = AV_1\Sigma_1^{-1}$ with

$$U_1 = \begin{bmatrix} 1 & 1 \\ 5 & -1 \\ -1 & 5 \end{bmatrix} \left( \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right) \begin{bmatrix} \frac{1}{\sqrt{18}} & 0 \\ 0 & \frac{1}{\sqrt{36}} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 0 \\ \frac{2}{3} & \frac{1}{\sqrt{2}} \\ \frac{2}{3} & -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

Putting this all together, we have the compressed SVD

$$A = U_1\Sigma_1V_1 = \begin{bmatrix} \frac{1}{3} & 0 \\ \frac{2}{3} & \frac{1}{\sqrt{2}} \\ \frac{2}{3} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{18} & 0 \\ 0 & \sqrt{36} \end{bmatrix} \left( \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right).$$

## 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

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

The factorization $A = U\Sigma V^H$ is called the **full SVD** of the matrix $A$ while the factorization $A = U_1\Sigma_1V_1$ 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 a orthonormal basis for the vectors in $Y$ that lie in the range of $A$ while the vectors in $U_2$ form 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

$$P_W = W(W^H W)^{-1} W^H = WW^H.$$

Therefore, the projection matrices for the fundamental subspaces are given by

$$\begin{aligned}P_{\mathcal{R}(A)} &= U_1 U_1^H\\ P_{\mathcal{R}(A^H)} &= V_1 V_1^H\\ P_{\mathcal{N}(A)} &= V_2 V_2^H\\ P_{\mathcal{N}(A^H)} &= U_2 U_2^H\end{aligned}$$

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 $\underline{v}$ is the $i$th column of $V$. This shows $A$ as the sum of $r$ rank-1 matrices. It also allows one to compute

$$\begin{aligned}\|A\|_F &= \sum_{i=1}^{r} \sigma_i^2\\ \|A\|_2 &= \sigma_1\end{aligned}$$

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 $\sigma$ is $\sigma^{-1}$ 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*

$$c : \mathcal{Y} \to \bigcup_{Y \in \mathcal{Y}} Y$$

*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

$$Y' = \{(Y, y) \,|\, y \in Y\} .$$

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

$$\mathcal{Z} = \{Y' | Y \in \mathcal{Y}\}.$$

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

$$c : \mathcal{Z} \to \mathcal{Y} \times \bigcup_{Y \in \mathcal{Y}} Y$$

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. □

### 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*

$$Y_x = \{y \in Y | y < x\}$$

*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

[BV04] S. P. Boyd, L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

# Index