LEHIGH UNIVERSITY

Mathematical Background

Prof. Frank E. Curtis

August 20, 2024

INDUSTRIAL AND SYSTEMS ENGINEERING

ISE

LEHIGH UNIVERSITY

* Please send comments and corrections to frank.e.curtis@gmail.com.

Table of Contents

Chapter 1

Mathematical Logic

Logic is the formal science of reason, where one of the primary goals is to identify true (or false) inferences. Mathematical logic involves the use of both symbolic and natural languages to construct formal mathematical statements, mostly in the forms of theorems and their corresponding proofs. In the following sections, we cover a few basic definitions and concepts in mathematical logic.

1.1 Operations and operators

An operation is a procedure that produces a new value from one or more inputs. Operations are indicated by operators. A unary operator indicates an operation that acts on a single object, known as an operand, whereas a binary operator indicates an operation involving two operands. A common unary operator is the "not" operator, which is denoted by "$\neg$." Consider a statement, call it A. The statement $\neg A$, i.e., "not A," is true if and only if A is false, meaning $\neg A$ is false if and only if A is true. Most other logical operators are binary operators. Given statements A and B, below is a list of some common binary operators.

- $A \wedge B$, i.e., "A and B." This statement is true if and only if A and B are both true.
- $A \vee B$, i.e., "A or B." This statement is true if and only if A or B (or both) is true.
- $A \Rightarrow B$, i.e., "A implies B." This statement is true if A being true means B is also true.
- $A \Leftrightarrow B$, i.e., "A if and only if B." This statement is true if A being true means B is also true, and A being false means B is also false. It is the logical equivalent of $(A \Rightarrow B) \wedge (B \Rightarrow A)$.

1.2 Quantification

One may have a statement involving a variable, call it x, and the truth of the statement may depend on x, in which case we may refer to the statement as $A(x)$. Below are examples of notation used in such cases.

- $\forall x : A(x)$, i.e., "$A(x)$ for all x." This statement is true if and only if $A(x)$ is true for all x.
- $\exists x : A(x)$, i.e., "$A(x)$ for some x." This statement is true if and only if $A(x)$ is true for at least one x.
- $\exists! x : A(x)$, i.e., "$A(x)$ for a unique x." This statement is true if and only if $A(x)$ is true for unique x.

For reasons evident in the definitions above, the symbol $\forall$ is known as the universal quantifier, $\exists$ is known as the existential quantifier, and $\exists!$ is known as the uniqueness quantifier.

1.3 Operator precedence

One may have a collection of statements combined with logical operators. In such cases, the order in which the operations are to be understood (i.e., their precedence) is critically important. Standard practice is to have the following order of precedence for the operators defined in §1.1: $\neg$, $\wedge$, $\vee$, $\Rightarrow$, and $\Leftrightarrow$. In order to avoid confusion and/or impose precedence in a series of operations, parentheses should be used. All operations within a set of parentheses have higher precedence than operations with the parenthetical statement. For example, recall the example provided in §1.1 that $A \Leftrightarrow B$ is logically equivalent to $(A \Rightarrow B) \wedge (B \Rightarrow A)$; the parentheses in this latter statement are important since $\wedge$ normally takes precedence over $\Rightarrow$.

1.4 Binary operator properties

There are a variety of properties that binary operators may or may not possess. The following are examples.

- *Associativity.* A binary operator $\star$ is associative if and only if it satisfies the associative law:

$$(a \star b) \star c \quad \text{is equivalent to} \quad a \star (b \star c).$$

  That is, within an expression containing two or more occurrences of an associative operator, the order in which the operators are applied does not matter as long as the operand sequence does not change.

- *Commutativity.* A binary operator is commutative if and only if changing the order of the operands does not change the result; i.e., if $\star$ is commutative, then

$$a \star b \quad \text{is equivalent to} \quad b \star a.$$

- *Distributivity.* Consider two binary operators, $\star$ and $\dagger$. The operator $\star$ is left-distributive over the operator $\dagger$ if and only if

$$a \star (b \dagger c) \quad \text{is equivalent to} \quad (a \star b) \dagger (a \star c),$$

  it is right-distributive over $\dagger$ if and only if

$$(a \dagger b) \star c \quad \text{is equivalent to} \quad (a \star c) \dagger (b \star c),$$

  and it is distributive over $\dagger$ if it is both left- and right-distributive over $\dagger$.

For example, in simple arithmetic with numbers, addition and multiplication are both associative and commutative, whereas subtraction and division are both neither associative nor commutative. Similarly, in logic, $\wedge$ and $\vee$ are both associative and commutative, whereas $\Rightarrow$ is neither associative nor commutative.

1.5 Implications and related statements

Given two statements, A and B, an implication such as $A \Rightarrow B$ is extremely important, especially when considering mathematical proof techniques. Below is a list of important related statements, where in each definition we refer to the implication $A \Rightarrow B$ simply as "the implication."

- $\neg(A \Rightarrow B)$ is the negation of the implication. By the definition of negation, if the implication is true, then its negation is false, and if the implication is false, then its negation is true.

- $\neg A \Rightarrow \neg B$ is the inverse of the implication, which may or may not be true, regardless of whether or not the implication is true.

- $\neg B \Rightarrow \neg A$ is the contrapositive of the implication, which is true if and only if the implication is true.

- $B \Rightarrow A$ is the converse of the implication, which may or may not be true, regardless of whether or not the implication is true. That being said, the converse is the contrapositive of the inverse, meaning that it is true if and only if the inverse of the implication is true.

1.6 Logical equivalences

The definitions and concepts discussed in the previous sections lead to a variety of useful logical equivalences. Consider two statements, A and B. If A is logically equivalent to B, then A can be proved true if and only if B can be proved true. Such an equivalence leads to the possibility of replacing one statement for another (i.e., A for B, or vice versa), which can be extremely useful when writing mathematical proofs. The table below contains a summary of some useful logical equivalences; in each case, a statement A is given along with a logically equivalent statement B. Each statement involves a combination of other statements (e.g., P, Q, etc.). We use $\top$ ($\bot$) to represent an inherently true (false) statement.

Name	A	B
Identity laws	$P \wedge \top$	P
	$P \vee \bot$	P
Domination laws	$P \vee \top$	$\top$
	$P \wedge \bot$	$\bot$
Idempotent laws	$P \vee P$	P
	$P \wedge P$	P
Double negation law	$\neg(\neg P)$	P
Negation laws	$P \vee \neg P$	$\top$
	$P \wedge \neg P$	$\bot$
Associative laws	$(P \vee Q) \vee R$	$P \vee (Q \vee R)$
	$(P \wedge Q) \wedge R$	$P \wedge (Q \wedge R)$
Commutative laws	$P \vee Q$	$Q \vee P$
	$P \wedge Q$	$Q \wedge P$
Distributive laws	$P \vee (Q \wedge R)$	$(P \vee Q) \wedge (P \vee R)$
	$P \wedge (Q \vee R)$	$(P \wedge Q) \vee (P \wedge R)$
De Morgan's laws	$\neg(P \wedge Q)$	$\neg P \vee \neg Q$
	$\neg(P \vee Q)$	$\neg P \wedge \neg Q$
Absorption laws	$P \vee (P \wedge Q)$	P
	$P \wedge (P \vee Q)$	P
Conditional laws	$P \Rightarrow Q$	$\neg P \vee Q$
	$P \Rightarrow Q$	$\neg Q \Rightarrow \neg P$
	$P \vee Q$	$\neg P \Rightarrow Q$
	$P \wedge Q$	$\neg(P \Rightarrow \neg Q)$
	$\neg(P \Rightarrow Q)$	$P \wedge \neg Q$
	$(P \Rightarrow Q) \wedge (P \Rightarrow R)$	$P \Rightarrow (Q \wedge R)$
	$(P \Rightarrow Q) \vee (P \Rightarrow R)$	$P \Rightarrow (Q \vee R)$
	$(P \Rightarrow R) \wedge (Q \Rightarrow R)$	$(P \wedge Q) \Rightarrow R$
	$(P \Rightarrow R) \vee (Q \Rightarrow R)$	$(P \vee Q) \Rightarrow R$
Biconditional laws	$P \Leftrightarrow Q$	$(P \Rightarrow Q) \wedge (Q \Rightarrow P)$
	$P \Leftrightarrow Q$	$\neg P \Leftrightarrow \neg Q$
	$P \Leftrightarrow Q$	$(P \wedge Q) \vee (\neg P \wedge \neg Q)$
	$\neg(P \Leftrightarrow Q)$	$(P \vee Q) \wedge \neg(P \wedge Q)$

Table 1.1: Examples of logical equivalences

It is also useful to consider statements that are locally equivalent to the negation of quantified statements; e.g., the statement $\neg(\forall x : A(x))$ is equivalent to $\exists x : \neg A(x)$, the statement $\neg(\exists x : A(x))$ is equivalent to $\forall x : \neg A(x)$, and the statement $\neg(\exists! x : A(x))$ is equivalent to $(\forall x : \neg A(x)) \vee (\exists (x, y) : A(x) \wedge A(y))$.

1.7 Proof techniques

Mathematical logic is often used to prove the truth (or falsehood) of certain mathematical statements. Below is a list of typical techniques for proving (or disproving) mathematical statements.

- *Direct proof.* In a direct proof, a statement is proved by logically combining axioms, definitions, or previously proved statements in a direct manner to prove the statement. For example, to prove that 4 is an even number, one can provide the direct argument that $4 = 2 \times 2$, which by the definition of an even number (being an integer multiple of 2) shows that 4 is even.

- *Proof by construction.* Similar to a direct proof, this type of proof involves the construction of an example possessing a particular property to show that something having that property exists. For example, to prove that there exists an integer multiple of π between 6 and 7, one can construct the example $2\pi \approx 6.28$ which is simultaneously an integer multiple of π and between 6 and 7.

- *Proof by counterexample.* A common example of a proof by construction is a proof by counterexample, where one creates an object having (respectively, lacking) a certain property to disprove a proposition that all objects lack (respectively, have) the property. For example, to disprove the incorrect proposition that the square root of any integer is an irrational number, one can cite the counterexample 4, which has $\sqrt{4} = 2$, a rational number.

- *Proof by exhaustion.* In such a proof, the realm of possibilities is divided into a finite number of groups, where in each group the proposition is proved while ignoring the other groups. For example, to prove that 2 is the only even prime number, one can exhaustively consider the realm of natural numbers: (*i*) 2 is both even and prime, so it is an even prime; (*ii*) any odd number is not an even number, so no odd number can be an even prime; (*iii*) all even numbers greater than 2 are not prime since they can be expressed as integer multiples of 2, so no even number greater than 2 can be an even prime. All possibilities being exhausted, it has been shown that 2 is the only even prime.

- *Proof by contraposition.* If the statement to be proved is an implication, then the result follows if the contrapositive of the statement is proved; such is a proof by contraposition. For example, to prove the implication that if x^2 is even, then x is even, one can instead prove that if x is not even, then x^2 is not even. This fact follows from the fact that the product of an odd number with itself (or with any odd number, for that matter) is an odd number.

- *Proof by contradiction.* Since the negation of a statement is false if and only if the statement is true, one may prove a statement to be true by proving that its negation leads to a logical falsehood; such is a proof by contradiction. A well-known example of a proof of this type is Euclid's proof of the existence of an infinite number of primes. The proof begins by supposing that the negation of the statement is true, i.e., that there are only a finite number of primes (with one of them being the largest), and then proceeds by showing (by construction) that this supposition leads to the existence of another (larger) prime. Since this conclusion contradicts the initial supposition, one has arrived at a logical falsehood, thus showing that the initial supposition of a finite number of primes must be false, which is logically equivalent to the statement that there are an infinite number of primes.

- *Proof by induction.* In a proof by induction, a "base case" is established first. Then, an "induction rule" is proved to show that one statement (e.g., the "base case") implies another case to be true. The rule being applied repeatedly, one has proved that a large (or even infinite) number of statements are true without having had to prove each statement individually. Such is a proof by induction. For example, one can use such a proof to show that the sum of all natural numbers from 0 to n is equal to $n(n+1)/2$ for all any natural number n. The base case of $n = 0$ clearly holds. Then, supposing that the statement is true up to some natural number n, one can observe that

$$0 + 1 + \cdots + n + (n+1) = \frac{n(n+1)}{2} + n + 1 = \frac{(n+1)(n+2)}{2},$$

  which shows that it also holds up to $n + 1$. By induction, the statement holds true for all natural n.

Chapter 2

Elements of Real Analysis, Geometry, and Linear Algebra

The three topics in this chapter represent three major branches of mathematics. Evolved from calculus (i.e., the study of change), mathematical analysis is the study of functions. In particular, real analysis is the study of real numbers, sequences of real numbers, real-valued functions, and so on. Geometry, on the other hand, is the study of shape, size, the relative position of figures, and the properties of space. Finally, algebra is the study of arithmetic operations, often involving non-numerical objects representing numbers that are either unknown (i.e., variables) or unspecified (i.e., parameters). Each of these branches, while distinct in focus, are intimately related, and the purpose of this chapter is to provide background on relevant topics in each of them. As opposed to the previous chapter where the objects of interest were statements (being true or false), the objects of interest in this chapter are numbers, variables, parameters, sets, functions, etc. The first few sections introduce notation that will be used throughout the remainder of the chapter; see §2.11 for a discussion of this notation and exceptions that will be made.

2.1 Operations

A unary operation is an operation involving a single quantity. This document presumes that the reader is familiar with common operations, such as $-$ (negation), a^x (where x is an exponent that raises a to the power of x), log (logarithm, presumed to be the natural logarithm unless otherwise indicated by a subscript), and so forth. This notation is standard. A binary operation is an operation involving two quantities. This document presumes that the reader is familiar with the most common binary operations between numbers: $+$ (addition), $-$ (subtraction), $\times$ (multiplication, for which we use $\cdot$ or simply write the numbers next to each other), and $\div$ (division, for which we write a fraction or use /). Some of these operations (e.g., negation, addition, and subtraction) apply elementwise for vectors and matrices as well, at least as far as the rules of linear algebra allow, which are discussed later in the chapter.

2.2 Relationships

A binary relationship is a relationship between two, often numerical, quantities. We write $x = y$ to indicate that x is equal to y, and write $x \neq y$ to indicate that x is not equal to y. If x and y are numbers and x is greater than or equal to y, then we write $x \geq y$; if x is strictly greater than y, then we write $x > y$. Similarly, $x \leq y$ indicates that x is less than or equal to y, and $x < y$ indicates that x is strictly less than y. In some contexts, less precise statements are appropriate for illustrating the nature of a binary relationship. For example, if x is approximately equal to y, then we write $x \approx y$. Similarly, if x is substantially greater than y, then we write $x \gg y$, and if x is substantially less than y, then we write $x \ll y$.

If x is an object that is *defined* to be equal to y, then we write $x := y$. This notation is used when the value of y has already been introduced, and we would like to introduce x; here, x derives its definition from being equal to y. Similarly, we may write $x =: y$ to indicate that y is defined as being equal to x. If we

intend to set the value of x to the value of y, such as in the context of an algorithm within which x may be set to any value, then we write $x \leftarrow y$. However, $x \to y$ does not indicate that y is being set to the value of x; the operator $\to$ (introduced later) is reserved for notation for functions and limits of sequences.

2.3 Sets

A collection of elements in which no element can be repeated is known as a set. (Various names, such as multiset, are used when a collection can contain repeated elements) A rigorous definition of a set from first principles is beyond our scope, but suffice it to say that there are certain fundamental sets, such as the set of real numbers, that are defined *ab initio*, i.e., not with respect to other sets. Otherwise, a set is defined with respect to a reference set, call it $\mathcal{S}$, with a statement such as

$$\mathcal{X} := \{x \text{ is an element of } \mathcal{S} : x \text{ satisfies } P\},$$

where P is a given property or collection of properties. (Sometimes, the reference set $\mathcal{S}$ is previously and/or universally defined, so it is omitted.) In this notation, the value x is used as a placeholder for an element of $\mathcal{X}$, and the phrase after the colon within the brackets discusses the property (or properties) that x possesses to be included in $\mathcal{X}$. (In some cases, for brevity and/or clarity, properties may be included before the colon as well.) If $\mathcal{X}$ is a set and x is an element of $\mathcal{X}$, then we write $x \in \mathcal{X}$. Otherwise, if x is not an element of $\mathcal{X}$, then we write $x \notin \mathcal{X}$. The complement of a set $\mathcal{X}$ is the set of elements in the reference set $\mathcal{S}$ that are not included in $\mathcal{X}$; with $\mathcal{S}$ presumed to be known, it is denoted as $\mathcal{X}^c := \{x \in \mathcal{S} : x \notin \mathcal{X}\}$.

If all elements of a set $\mathcal{X}$ are also elements of a set $\mathcal{Y}$, then $\mathcal{X}$ is called a subset of $\mathcal{Y}$ and we write $\mathcal{X} \subseteq \mathcal{Y}$. Similarly, if all elements of $\mathcal{Y}$ are also elements of $\mathcal{X}$, then $\mathcal{X}$ is called a superset of $\mathcal{Y}$ and we write $\mathcal{X} \supseteq \mathcal{Y}$. If $\mathcal{X}$ is both a subset and a superset of $\mathcal{Y}$ (meaning that both sets contain exactly the same collection of elements), then $\mathcal{X}$ and $\mathcal{Y}$ are equal and we write $\mathcal{X} = \mathcal{Y}$. If $\mathcal{X}$ is a subset of $\mathcal{Y}$, but the sets are not equal (i.e., $\mathcal{X} \neq \mathcal{Y}$), then $\mathcal{X}$ is known as a proper subset of $\mathcal{Y}$ and we write $\mathcal{X} \subset \mathcal{Y}$. A proper superset $\mathcal{X}$ of $\mathcal{Y}$ is defined similarly, and this relationship is indicated with the notation $\mathcal{X} \supset \mathcal{Y}$. In certain cases, indicating whether a subset (or superset) is proper or not is necessary, but in others it is not.

The empty set, which contains no elements, is denoted as $\emptyset$.

2.4 Sets of numbers

The set of real numbers is denoted as $\mathbb{R}$. Infinity, denoted by ∞, is a value defined to be greater in value than all real numbers. Negative infinity, denoted by $-\infty$, is a value defined to be lesser in value than all real numbers. The set of extended real numbers, defined as $\mathbb{R}$ augmented with the set $\{-\infty, \infty\}$, is denoted as $\overline{\mathbb{R}}$. It follows that $-\infty < x < \infty$ for all $x \in \mathbb{R}$ and $-\infty \leq x \leq \infty$ for all $x \in \overline{\mathbb{R}}$. We write $[a, b]$ to denote the set of all (possibly extended) real x such that $a \leq x \leq b$. A round (instead of square) bracket indicates that the inequality in the definition is strict; e.g., the set $(a, b]$ denotes the set of all (possibly extended) real x such that $a < x \leq b$. The rules of arithmetic extend in certain cases when infinite values are involved: $x \cdot \infty = \infty$ for any real $x > 0$; $x \cdot \infty = -\infty$ for any real $x < 0$; and $x + \infty = \infty$ and $x - \infty = -\infty$ for any $x \in \mathbb{R}$. However, $\infty \cdot 0$, $-\infty \cdot 0$, $\infty - \infty$ and ∞/∞ are meaningless.

The set of real numbers greater than or equal to $a \in \mathbb{R}$ is denoted as $\mathbb{R}_{\geq a} := \{x \in \mathbb{R} : x \geq a\}$, and the set of real numbers greater than a is denoted as $\mathbb{R}_{>a} := \{x \in \mathbb{R} : x > a\}$. The sets $\mathbb{R}_{\leq a}$ (real numbers less than or equal to a) and $\mathbb{R}_{<a}$ (real numbers less than a) are defined similarly.

The set of integers is denoted as $\mathbb{Z} := \{\dots, -2, -1, 0, 1, 2, \dots\}$. Defined similarly as for real numbers, we define the sets of nonnegative integers $\mathbb{Z}_{\geq 0}$, positive integers $\mathbb{Z}_{>0}$, nonpositive integers $\mathbb{Z}_{\leq 0}$, and negative integers $\mathbb{Z}_{<0}$. Moreover, we adopt the convention that the set of nonnegative integers can be referred to as the set of natural numbers, denoted as $\mathbb{N} := \mathbb{Z}_{\geq 0}$. We also define $\mathbb{N}_{>0} := \mathbb{Z}_{>0}$. The set of rational numbers (i.e., those that can be represented as quotients of integers) is denoted as $\mathbb{Q}$, and the corresponding sets $\mathbb{Q}_{\geq 0}$, $\mathbb{Q}_{>0}$, $\mathbb{Q}_{\leq 0}$, and $\mathbb{Q}_{<0}$ are defined using the same scheme as with the sets of real numbers and integers. Finally, the set of complex numbers is denoted as $\mathbb{C}$.

2.5 Countable sets, uncountable sets, and cardinality

Let $\mathcal{X}$ be a nonempty set. If the elements of $\mathcal{X}$ can be counted using a finite subset of $\mathbb{N}_{>0}$ (i.e., if one can derive a one-to-one correspondence between elements of $\mathcal{X}$ and a finite subset of $\mathbb{N}_{>0}$), then $\mathcal{X}$ is said to be a finite set; otherwise, it is said to be an infinite set. If the elements of $\mathcal{X}$ can be counted using a finite subset of $\mathbb{N}_{>0}$ or all of $\mathbb{N}_{>0}$, then $\mathcal{X}$ is said to be countable; otherwise, it is said to be uncountable. The number of elements of a set $\mathcal{X}$, known as the cardinality of $\mathcal{X}$, is written as $|\mathcal{X}|$. For example, $|\{1, 2, 3, 4\}| = 4$. If $\mathcal{X}$ is a nonempty finite set, then we write $|\mathcal{X}| < \infty$. The empty set is said to have cardinality of zero, i.e., $|\emptyset| = 0$. The cardinality of a countably infinite set is a so-called infinite cardinal known as aleph-zero, written as $\aleph_0$. However, as is common and since it does not lead to confusion for our purposes, we overload the symbol ∞ and write that if $\mathcal{X}$ is a countably infinite set, then $|\mathcal{X}| = \infty$.

It is a remarkable fact that there indeed exist sets whose elements cannot be counted using the natural numbers, i.e., uncountable sets. For example, the set of real numbers $\mathbb{R}$ is uncountable since one cannot derive a one-to-one correspondence between the elements of $\mathbb{R}$ and those of $\mathbb{N}$. For this reason, despite the fact that both sets have an infinite number of elements, one may say that the set $\mathbb{R}$ has more elements than the set $\mathbb{N}$. This fact has subtle, but important consequences in analysis. The notion of cardinality of an uncountable set, such as $\mathbb{R}$, requires a notion of cardinality of a continuum; this is outside of our scope.

2.6 Set operations

The following is a list of (binary) operations that may be performed with two sets $\mathcal{X}_1$ and $\mathcal{X}_2$.

- Their union is denoted as $\mathcal{X}_1 \cup \mathcal{X}_2 := \{x : x \in \mathcal{X}_1 \text{ or } x \in \mathcal{X}_2\}$.
- Their intersection is denoted as $\mathcal{X}_1 \cap \mathcal{X}_2 := \{x : x \in \mathcal{X}_1 \text{ and } x \in \mathcal{X}_2\}$.
- Their set difference is denoted as $\mathcal{X}_1 \setminus \mathcal{X}_2 := \{x : x \in \mathcal{X}_1 \text{ and } x \notin \mathcal{X}_2\}$.
- Their (Cartesian) product is denoted as $\mathcal{X}_1 \times \mathcal{X}_2 := \{(x_1, x_2) : x_1 \in \mathcal{X}_1 \text{ and } x_2 \in \mathcal{X}_2\}$.

We have the following additional operations when the elements of the sets can be added or subtracted.

- Their vector (Minkowski) sum is denoted as $\mathcal{X}_1 + \mathcal{X}_2 := \{x_1 + x_2 : x_1 \in \mathcal{X}_1 \text{ and } x_2 \in \mathcal{X}_2\}$.
- Their vector (Minkowski) difference is denoted as $\mathcal{X}_1 - \mathcal{X}_2 := \{x_1 - x_2 : x_1 \in \mathcal{X}_1 \text{ and } x_2 \in \mathcal{X}_2\}$.

Unions, intersections, vector sums, and products extend naturally to cases when more than two sets are involved using the operation prefixes $\bigcup$, $\bigcap$, $\sum$, and $\prod$, respectively. For two examples, the sets

$$\bigcup_{\alpha \in \mathcal{Y}} \mathcal{X}_\alpha \quad \text{and} \quad \sum_{i=a}^{b} \mathcal{X}_i$$

are, respectively, the union of all $\mathcal{X}_\alpha$ indexed over α in some (finite, countable, or uncountable) set $\mathcal{Y}$ and the vector sum of all $\mathcal{X}_i$ indexed over i in some ordered (finite or countable) set $\{a, \ldots, b\}$.

2.7 Function notation

A function defines a mapping from a set of inputs to a set of outputs with the property that each input is mapped to exactly one output. If f is a function, then we use the notation $f : \mathcal{X} \to \mathcal{Y}$ to indicate that f maps inputs in $\mathcal{X}$ (its domain) to outputs in $\mathcal{Y}$ (its codomain). If $f : \mathcal{X} \to \mathcal{Y}$ is a function and $\mathcal{U}$ is a subset of $\mathcal{X}$, then $\{f(x) : x \in \mathcal{U}\}$ is the (forward) image of $\mathcal{U}$ under f, where, if $\mathcal{U} = \mathcal{X}$, then this set is simply called the image of f. Similarly, if $f : \mathcal{X} \to \mathcal{Y}$ is a function and $\mathcal{V}$ is a subset of $\mathcal{Y}$, then $\{x \in \mathcal{X} : f(x) \in \mathcal{V}\}$ is the preimage of $\mathcal{V}$ under f, where, if the image of f is $\mathcal{V}$, then this set is simply called the preimage of f.

A multifunction (or set-valued function) is a mapping in which each input is mapped to at least one output, i.e., a mapping for which a given input may be mapped to one, or more than one, output. If $\hat{f}$ is a multifunction, then we write $\hat{f} : \mathcal{X} \rightrightarrows \mathcal{Y}$ to indicate that $\hat{f}$ maps inputs in $\mathcal{X}$ (its domain) to outputs

in $\mathcal{Y}$ (its codomain). For example, if f is a function with domain $\mathcal{X}$ and the image of f is $\mathcal{V}$, then one may consider the multifunction $\hat{f} : \mathcal{V} \rightrightarrows \mathcal{X}$ defined by $\hat{f}(v) = \{x \in \mathcal{X} : f(x) = v\}$.

A typical convention, adopted in this document, is that the domain of a function (multifunction) may include values at which the function (multifunction) may be undefined and/or yield an infinite value (values). In such cases, it is implicitly assumed that one should ignore points in the domain at which the function is undefined, though the same should not necessarily be assumed about points yielding infinite values since they may be of special interest. For example, the domain of a function $f : \mathcal{X} \to \mathcal{Y}$, where $\{-\infty, \infty\} \subseteq \mathcal{Y}$, should be distinguished from the effective domain of f, which we denote and define as

$$\mathrm{dom}(f) := \{x \in \mathcal{X} : f(x) < \infty\}.$$

Such a function is said to be proper if $\mathrm{dom}(f) \neq \varnothing$ and $f(x) > -\infty$ for all $x \in \mathcal{X}$; otherwise, it is improper.

Another important set related to a function is its support, defined as

$$\mathrm{supp}(f) := \{x \in \mathcal{X} : f(x) \neq 0\}.$$

2.8 Infima and suprema

Let $\mathcal{X}$ be a nonempty set of real numbers. The infimum of $\mathcal{X}$, denoted as $\inf \mathcal{X}$, is the largest real number y such that $y \leq x$ for all $x \in \mathcal{X}$. The infimum is equal to $-\infty$ if no such real number exists. Similarly, the supremum of $\mathcal{X}$, denoted as $\sup \mathcal{X}$, is the smallest real number y such that $y \geq x$ for all $x \in \mathcal{X}$, and is equal to ∞ if no such real number exists. For the empty set, we use the convention that $\inf \varnothing = \infty$ and $\sup \varnothing = -\infty$. If $\inf \mathcal{X} = x_*$ for some $x_* \in \mathcal{X}$, then x_* is a minimum point of $\mathcal{X}$ and we write $x_* = \min \mathcal{X}$. Similarly, if $\sup \mathcal{X} = x_*$ for some $x_* \in \mathcal{X}$, then x_* is a maximum point of $\mathcal{X}$ and we write $x_* = \max \mathcal{X}$. In this manner, whenever "$\min \mathcal{X}$" is written in place of "$\inf \mathcal{X}$" (or "$\max \mathcal{X}$" in place of "$\sup \mathcal{X}$"), it is done so only to emphasize when it is known that the infimum (or supremum) of $\mathcal{X}$ is an element of $\mathcal{X}$.

We use similar, but slightly enhanced notation when a set is defined as a set of function values as they are evaluated over a given set of inputs. For example, we may write $\inf\{f(x) : x \in \mathcal{X}\}$ as

$$\inf_{x \in \mathcal{X}} f(x).$$

As in the case of sets, in this notation we replace inf (sup) with min (max) when the infimum (supremum) of $\{f(x) : x \in \mathcal{X}\}$ is known to be attained at some element $x_* \in \mathcal{X}$.

2.9 Vectors and matrices

A vector is a multidimensional object composed of a set of numbers. We adopt the convention that, by default, a vector is organized into a column, i.e., it is a column vector. If x is a vector, then its ith element (otherwise known as ith component or ith entry) is denoted as x_i. If each element of a vector belongs to the same set of numbers, then the dimension of the vector is indicated by an exponent on the set; e.g., the set of n-dimensional real vectors (i.e., real n-vectors) is denoted as $\mathbb{R}^n$. Overall, if $x \in \mathbb{R}^n$, then

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{where} \quad x_i \in \mathbb{R} \quad \text{for all} \quad i \in \{1, \ldots, n\}.$$

A matrix is a multidimensional object composed of a set of vectors of the same dimension. If A is a matrix composed of n vectors of dimension m placed side-by-side, then we say that A is an $m \times n$ matrix. For example, if A is composed of n vectors in $\mathbb{R}^m$, then we write $A \in \mathbb{R}^{m \times n}$ and have

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ldots & a_{mn} \end{bmatrix}, \quad \text{where} \quad a_{ij} \in \mathbb{R} \quad \text{for all} \quad i \in \{1, \ldots, m\} \quad \text{and} \quad j \in \{1, \ldots, n\}.$$

Here, a_{ij} denotes the (i,j)-element of the matrix A, where i is the row index and j is the column index. An n-dimensional vector can be viewed as an $n \times 1$ matrix, so all matrix operations apply to vectors as well. We also remark that a matrix such as A above can be viewed as a collection of m row vectors in $\mathbb{R}^n$.

If all pairs of corresponding elements in a pair of matrices satisfy a binary relationship, then, for brevity, it is written that the pair of matrices satisfies the relationship (which is meant to be understood to hold elementwise). For example, for $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$, we write $x > y$ to indicate that $x_i > y_i$ for all $i \in \{1, \ldots, n\}$. Correspondingly, if at least one pair of elements of a pair of matrices does not satisfy a binary relationship, then it is written that the pair of matrices does not satisfy the relationship. For example, for $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$, we write $x \neq y$ to indicate that $x_i \neq y_i$ for at least one $i \in \{1, \ldots, n\}$. (This is consistent with a logical negation of the statement that $x_i = y_i$ for all $i \in \{1, \ldots, n\}$.)

2.10 Transposition

Given $A \in \mathbb{R}^{m \times n}$, the transpose of A, denoted by $A^T \in \mathbb{R}^{n \times m}$, is the matrix whose (j,i)-element is equal to the (i,j)-element of A for all $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$. (We remark that if $A \in \mathbb{C}^{m \times n}$, then we also have the notion of a conjugate transpose, denoted A^* which has its (j,i)-element equal to the complex conjugate of the (i,j)-element of A for all $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$.)

2.11 Notational conventions

Throughout the remainder of this chapter, we maintain the conventions that have been employed thus far. In particular, we generally use lowercase letters to indicate numbers, vectors, or functions; uppercase letters to indicate matrices; and calligraphic uppercase letters to indicate sets. Subscripts are used to indicate the number of an element of a vector or matrix, where in the case of a matrix we switch from an uppercase to the corresponding lowercase letter when indicating an element; e.g., x_i is the ith element of the vector x and a_{ij} (or $a_{i,j}$) is the (i,j)-element of the matrix A. An exception occurs when subscripts are used for the index of a vector (or matrix) in a collection of vectors (or matrices). For example, we may write

$$A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \quad \text{where} \quad a_j \in \mathbb{R}^m \quad \text{for all} \quad j \in \{1, \ldots, n\},$$

or we may write

$$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} \quad \text{where} \quad a_i \in \mathbb{R}^n \quad \text{for all} \quad i \in \{1, \ldots, m\}.$$

In such cases, the meaning of the subscript will be made clear in the context.

We also assume for the remainder of the chapter that all quantities are defined in a real space. However, it should be clear that certain concepts also apply for integer quantities, rational quantities, etc.

2.12 Matrix addition and multiplication

Given $(A, B) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n}$, the sum of A and B is $C = A + B \in \mathbb{R}^{m \times n}$, where

$$c_{ij} = a_{ij} + b_{ij} \quad \text{for all } i \in \{1, \ldots, m\} \text{ and } j \in \{1, \ldots, n\}.$$

The difference between A and B, i.e., $A - B$, is defined in a similar fashion. If two matrices have different row or column dimension, then neither addition nor subtraction is defined between the matrices.

Given $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$, the inner product product of x and y is defined as

$$x^T y = \sum_{i=1}^{n} x_i y_i \in \mathbb{R}.$$

Similarly, the outer product of x and y is defined as $Z = xy^T \in \mathbb{R}^{n \times n}$, where

$$z_{ij} = x_i y_j \in \mathbb{R} \quad \text{for all } i \in \{1, \ldots, n\} \text{ and } j \in \{1, \ldots, n\}.$$

More generally, given $A \in \mathbb{R}^{m\times p}$ and $B \in \mathbb{R}^{p\times n}$, their product is $C = AB \in \mathbb{R}^{m\times n}$ where

$$c_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj} \in \mathbb{R} \quad \text{for all } i \in \{1,\dots,m\} \text{ and } j \in \{1,\dots,n\}.$$

If $A \in \mathbb{R}^{m\times p}$ and $B \in \mathbb{R}^{q\times n}$ where $p \neq q$, then AB is undefined and has no meaning. Moreover, there is no such thing as matrix division; the most relevant concept is matrix inversion, which is defined in §2.18.

We remark that if A and B are matrices such that AB is well-defined, then $(AB)^T = B^T A^T$.

2.13 Subspaces, affine sets, convex sets, and cones

A nonempty subset $\mathcal{S}$ of $\mathbb{R}^n$ is a subspace if $\alpha_1 s_1 + \alpha_2 s_2 \in \mathcal{S}$ for every $(s_1, s_2) \in \mathcal{S}\times\mathcal{S}$ and $(\alpha_1,\alpha_2) \in \mathbb{R}\times\mathbb{R}$. It is easily shown that if $\mathcal{S}$ is a subspace, then it includes the origin and contains every line that passes through two distinct points in $\mathcal{S}$. In addition, if $s \in \mathcal{S}$, then $\mathcal{S} = s + \mathcal{S}$ and $\mathcal{S} = s - \mathcal{S}$. (Here, given a set $\{x\}$ composed of a single vector x, an operation with $\{x\}$ is simply written as an operation with x.)

An affine set $\mathcal{X}$ in $\mathbb{R}^n$ is a translated subspace, in the sense that there exists a vector $x \in \mathbb{R}^n$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^n$ such that $\mathcal{X} = x + \mathcal{S} = \{x + s : s \in \mathcal{S}\}$. Here, $\mathcal{S}$ is known as the subspace parallel to $\mathcal{X}$.

Theorem 2.13.1. *Given an affine set $\mathcal{X}$, there is a unique subspace $\mathcal{S}$ parallel to $\mathcal{X}$.*

Proof. Let $\mathcal{X} = x + \mathcal{S}$ and $\mathcal{X} = \bar{x} + \overline{\mathcal{S}}$. Then, since $x \in \mathcal{X}$, we must have $x = \bar{x} + \bar{s}$ for some $\bar{s} \in \overline{\mathcal{S}}$ so that $\mathcal{X} = \bar{x} + \bar{s} + \mathcal{S}$. Since we also have $\mathcal{X} = \bar{x} + \overline{\mathcal{S}}$, it follows that $\mathcal{S} = \overline{\mathcal{S}} - \bar{s} = \overline{\mathcal{S}}$. □

As for a subspace, an affine set includes every line that passes through two distinct points in the set, though an affine set might not include the origin (and if it does, then it is subspace). We also remark that an alternative definition is that a nonempty subset $\mathcal{X}$ of $\mathbb{R}^n$ is an affine set if $\alpha_1 x_1 + \alpha_2 x_2 \in \mathcal{X}$ for every $(x_1, x_2) \in \mathcal{X}\times\mathcal{X}$ and $(\alpha_1,\alpha_2) \in \mathbb{R}\times\mathbb{R}$ such that $\alpha_1 + \alpha_2 = 1$.

A nonempty subset $\mathcal{X}$ of $\mathbb{R}^n$ is convex if $\alpha_1 x_1 + \alpha_2 x_2 \in \mathcal{X}$ for every $(x_1, x_2) \in \mathcal{X}\times\mathcal{X}$ and $(\alpha_1,\alpha_2) \in \mathbb{R}_{\geq 0}\times\mathbb{R}_{\geq 0}$ such that $\alpha_1 + \alpha_2 = 1$. Clearly, all subspaces and affine sets are convex.

A nonempty subset $\mathcal{X}$ of $\mathbb{R}^n$ is a cone if $\alpha x \in \mathcal{X}$ for every $x \in \mathcal{X}$ and $\alpha \geq 0$. A related definition is that of a cone generated by a set $\mathcal{X} \subseteq \mathbb{R}^n$ (not necessarily a cone), which is denoted and defined by

$$\text{cone}(\mathcal{X}) := \{\alpha x \in \mathbb{R}^n : x \in \mathcal{X} \text{ and } \alpha \geq 0\}.$$

(We remark that in some settings it is preferred to replace $\alpha \geq 0$ with $\alpha > 0$ in these definitions, in which case a cone does not necessarily contain the origin.)

2.14 Combinations of vectors and hulls

A linear combination of a set of vectors $(x_1,\dots,x_m) \in \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ is any $x \in \mathbb{R}^n$ that can be written as

$$x = \sum_{i=1}^{m} \alpha_i x_i, \quad \text{where} \quad \alpha_i \in \mathbb{R} \quad \text{for all} \quad i \in \{1,\dots,m\}. \tag{2.14.1}$$

The (linear) span of $\{x_1,\dots,x_m\}$, denoted as $\text{span}(\{x_1,\dots,x_m\})$, is the set composed of all linear combinations of the elements in $\{x_1,\dots,x_m\}$. Similarly, viewing $A \in \mathbb{R}^{m\times n}$ as being composed of the set of column vectors $(a_1,\dots,a_n) \in \mathbb{R}^m \times \cdots \times \mathbb{R}^m$, we denote the span of these vectors simply as $\text{span}(A)$. If $\text{span}(\{x_1,\dots,x_m\})$ is equal to a subspace $\mathcal{S}$, then we call $\{x_1,\dots,x_m\}$ a spanning set for the subspace $\mathcal{S}$.

We also define three other types of combinations of a set of vectors. An affine combination of the vectors in $\{x_1,\dots,x_m\}$ is any $x \in \mathbb{R}^n$ that can be written as (2.14.1) with the added restriction that $\sum_{i=1}^m \alpha_i = 1$, a conical combination is any $x \in \mathbb{R}^n$ that can be written as (2.14.1) with the added restriction that $\alpha_i \geq 0$ for all $i \in \{1,\dots,m\}$, and a convex combination is any $x \in \mathbb{R}^n$ that can be written as (2.14.1) with the added restrictions that $\sum_{i=1}^m \alpha_i = 1$ and $\alpha_i \geq 0$ for all $i \in \{1,\dots,m\}$.

Each of the definitions in the two preceding paragraphs leads to the concept of a hull (of a set) when the set of vectors are drawn from a given set. In particular, the linear (respectively, affine, conical, or convex)

hull of a set $\mathcal{X} \subseteq \mathbb{R}^n$ is the set of all linear (respectively, affine, conical, or convex) combinations of all elements of $\mathcal{X}$. To denote such a hull of $\mathcal{X}$, we use the notation $\text{span}(\mathcal{X})$ (respectively, $\text{aff}(\mathcal{X})$, $\text{coni}(\mathcal{X})$, or $\text{conv}(\mathcal{X})$). In the case of a linear (respectively, affine or convex) hull, the hull is equal to the smallest subspace (respectively, affine set or convex set) containing the given set. This is not the case for the conical hull; the smallest cone containing the given set is the cone generated by the set, but this is not necessarily equal to the conical hull of the set.

2.15 Linear and affine independence

The vectors $(x_1, \ldots, x_m) \in \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ are called linearly independent if there are no real numbers $(\alpha_1, \ldots, \alpha_m) \in \mathbb{R} \times \cdots \times \mathbb{R}$ (at least one being nonzero) such that

$$0 = \sum_{i=1}^{m} \alpha_i x_i.$$

Otherwise, if the above equality holds for some set of real numbers (at least one being nonzero), then the set of vectors are linearly dependent. Clearly, in any linearly independent set of vectors, we must have $x_i \neq 0$ for all $i \in \{1, \ldots, n\}$. Then, an equivalent definition of linear independence is that $x_1 \neq 0$ and for all $k \in \{2, \ldots, n\}$ the vector x_k does not belong to $\text{span}(\{x_1, \ldots, x_{k-1}\})$.

The vectors $(x_1, \ldots, x_m) \in \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ are affinely independent if there are no real numbers $(\alpha_1, \ldots, \alpha_m) \in \mathbb{R} \times \cdots \times \mathbb{R}$ (at least one being nonzero) such that

$$\sum_{i=1}^{m} \alpha_i = 0 \quad \text{and} \quad 0 = \sum_{i=1}^{m} \alpha_i x_i.$$

Clearly, linear independence of a set of vectors implies affine independence, but not vice versa. That being said, the m vectors $(x_1, \ldots, x_m) \in \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ are affinely independent if and only if the $m - 1$ vectors $(x_2 - x_1, \ldots, x_m - x_1) \in \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ are linearly independent.

2.16 Bases and dimensions

If $\mathcal{S}$ is a subspace containing at least one nonzero vector, then a basis for $\mathcal{S}$ is a collection of linearly independent vectors whose span is equal to $\mathcal{S}$. Every basis for a given subspace $\mathcal{S}$ has the same number of vectors, and this number is called the dimension of $\mathcal{S}$. (By convention, the subspace $\{0\}$ has zero dimension.) The dimension of an affine set $\mathcal{X}$ is the dimension of the subspace parallel to $\mathcal{X}$.

2.17 Orthogonality and orthonormality of vectors

Two vectors x and $\bar{x}$ in $\mathbb{R}^n$ are called orthogonal if their inner product is zero, i.e., $x^T \bar{x} = 0$. Moreover, they are called orthonormal if they are orthogonal, $x^T x = 1$, and $\bar{x}^T \bar{x} = 1$. (It can be seen along with the definitions in §2.21 that these latter requirements are that x and $\bar{x}$ each have Euclidean norm equal to 1.) Given $\mathcal{X} \subseteq \mathbb{R}^n$, the set of vectors that are orthogonal to all elements of $\mathcal{X}$ is denoted as

$$\mathcal{X}^\perp := \{\bar{x} : x^T \bar{x} = 0 \quad \text{for all} \quad x \in \mathcal{X}\}.$$

If $\mathcal{S} \subseteq \mathbb{R}^n$ is a subspace, then $\mathcal{S}^\perp$ is called the orthogonal complement of $\mathcal{S}$, where it follows that $(\mathcal{S}^\perp)^\perp = \mathcal{S}$.

Theorem 2.17.1. *Given a subspace $\mathcal{S} \subseteq \mathbb{R}^n$, any vector $x \in \mathbb{R}^n$ can be uniquely decomposed as*

$$x = s + \bar{s}, \quad \textit{where} \quad s \in \mathcal{S} \quad \textit{and} \quad \bar{s} \in \mathcal{S}^\perp.$$

2.18 Matrix inversion, definiteness, and special forms

A matrix $A \in \mathbb{R}^{m\times n}$ is square if $m = n$. The square $n \times n$ matrix with its (i,i)-element equal to 1 for all $i \in \{1,\ldots,n\}$ and each (i,j)-element equal to 0 for all $(i,j) \in \{1,\ldots,n\}\times\{1,\ldots,n\}$ with $i \neq j$ is known as an identity matrix and is denoted by $I \in \mathbb{R}^{n\times n}$. (We use the notation I to denote the identity matrix of any size, where the size is determined by the context in which it appears.) Given a square matrix $A \in \mathbb{R}^{n\times n}$, if there exists a unique matrix, which we denote by A^{-1}, such that $AA^{-1} = I$, then A^{-1} is called the inverse of A; in such cases, it can be shown that we also have $A^{-1}A = I$. When the inverse of a given matrix A exists, we say that A is invertible or nonsingular; otherwise, we say that A is non-invertible or singular. If A and B are invertible matrices in $\mathbb{R}^{n\times n}$, then AB is invertible and $(AB)^{-1} = B^{-1}A^{-1}$.

Let $A \in \mathbb{R}^{n\times n}$. We say A is diagonal if its (i,j)-element is equal to 0 for all $(i,j) \in \{1,\ldots,n\}\times\{1,\ldots,n\}$ with $i \neq j$; e.g., the identity matrix I is diagonal. We say that A is symmetric if $A^T = A$; e.g., any diagonal matrix is symmetric. A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is called positive semidefinite if for all $x \in \mathbb{R}^n$ we have $x^TAx \geq 0$, and it is called positive definite if this inequality is strict for all nonzero x. Similarly, such a matrix is negative semidefinite if for all $x \in \mathbb{R}^n$ we have $x^TAx \leq 0$, and it is negative definite if this inequality is strict for all nonzero x. All of these notions apply exclusively to symmetric matrices, so, e.g., when one says that a matrix is positive definite, then it is implicitly assumed that the matrix is symmetric, though the adjective "symmetric" may be added for clarity.

Let $A \in \mathbb{R}^{n\times n}$. We say that A is lower triangular if its (i,j)-element is equal to 0 for all $(i,j) \in \{1,\ldots,n\}\times\{1,\ldots,n\}$ with $i < j$. Similarly, we say that A is upper triangular if its (i,j)-element is equal to 0 for all $(i,j) \in \{1,\ldots,n\}\times\{1,\ldots,n\}$ with $i > j$. Clearly, if A is lower triangular, then A^T is upper triangular, and if A is upper triangular, then A^T is lower triangular. Moreover, A is lower triangular and upper triangular if and only if it is diagonal, in which case it is also symmetric.

2.19 Images and inverse images of matrix transformations

Given $\mathcal{X} \subseteq \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$, the image of $\mathcal{X}$ under A is denoted by

$$A\mathcal{X} := \{Ax : x \in \mathcal{X}\}.$$

Similarly, given $\mathcal{Y} \subseteq \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times n}$, the inverse image of $\mathcal{Y}$ under A is denoted by

$$A^{-1}\mathcal{Y} := \{x : Ax \in \mathcal{Y}\}.$$

Note that this notation may be used even if A is not invertible.

2.20 Range and null spaces

Let $A \in \mathbb{R}^{m\times n}$. The range space of A, denoted by range(A), is the set of all $\bar{x} \in \mathbb{R}^m$ such that $\bar{x} = Ax$ for some $x \in \mathbb{R}^n$. The null space of A, denoted by null(A), is the set of all $x \in \mathbb{R}^n$ such that $Ax = 0$. The range space and null space of any matrix are subspaces. The rank of A, denoted by rank(A), is the dimension of the range space of A, which is equal to the maximal number of linearly independent columns of A, and is also equal to the maximal number of linearly independent rows of A. In this manner, given any matrix A, the transpose A^T has the same rank as A. We say that $A \in \mathbb{R}^{m\times n}$ has full row rank if rank$(A) = m$, and that it has full column rank if rank$(A) = n$. Overall, A has full rank if rank$(A) = \min\{m,n\}$, which can only be true if either all of its rows or all of its columns are linearly independent.

Theorem 2.20.1 (Fundamental Theorem of Linear Algebra)**.** *Given any $A \in \mathbb{R}^{m\times n}$, we have*

$$\text{range}(A^T) = \text{null}(A)^\perp \quad \textit{and} \quad \text{range}(A) = \text{null}(A^T)^\perp.$$

Observing Theorems 2.17.1 and 2.20.1, we may conclude that, given any $A \in \mathbb{R}^{m\times n}$, we have

$$\text{range}(A^T) + \text{null}(A) = \mathbb{R}^n \quad \text{and} \quad \text{range}(A) + \text{null}(A^T) = \mathbb{R}^m.$$

2.21 Vector norms

A norm on $\mathbb{R}^n$ is a function that assigns a real number $\|x\|$ to $x \in \mathbb{R}^n$ and has the following properties:

- $\|x\| \geq 0$ for all $x \in \mathbb{R}^n$.
- $\|\alpha x\| = |\alpha| \|x\|$ for all $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^n$. (Here, $|\alpha|$ denotes the absolute value of α.)
- $\|x\| = 0$ if and only if $x = 0$.
- $\|x + \bar{x}\| \leq \|x\| + \|\bar{x}\|$ for all $(x, \bar{x}) \in \mathbb{R}^n \times \mathbb{R}^n$. (This is known as the triangle inequality.)

The most common norms on $\mathbb{R}^n$ are ℓ_p-norms (with $p \geq 1$), where for $x \in \mathbb{R}^n$ we define

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}. \tag{2.21.2}$$

In particular, the well-known ℓ_1-, ℓ_2-, and ℓ_∞-norms have this form, but also have the simpler forms

$$\|x\|_1 := \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 := \sqrt{x^T x}, \quad \text{and} \quad \|x\|_\infty := \max_{i \in \{1,\ldots,n\}} |x_i|.$$

Any norm $\|\cdot\|$ on $\mathbb{R}^n$ has a dual norm, which is denoted and defined for $x \in \mathbb{R}^n$ by

$$\|x\|_* := \max_{\|\bar{x}\| \leq 1} x^T \bar{x}.$$

The dual of the ℓ_p-norm is the ℓ_q-norm where $\frac{1}{p} + \frac{1}{q} = 1$; e.g., the ℓ_2-norm is self-dual. This fact also extends to the extreme cases, i.e., the dual of the ℓ_1-norm is the ℓ_∞-norm, and vice versa.

Two norms, say $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, are equivalent if there exist real numbers a and b such that

$$a\|x\|_\alpha \leq \|x\|_\beta \leq b\|x\|_\alpha \quad \text{for all} \quad x \in \mathbb{R}^n.$$

We have the following result for norms on finite-dimensional vector spaces (such as $\mathbb{R}^n$), though we remark that this result is not necessarily true in infinite-dimensional spaces.

Theorem 2.21.1. *All norms on $\mathbb{R}^n$ are equivalent.*

This result can be verified in that, in $\mathbb{R}^n$, for $\alpha > \beta \geq 1$ we have

$$\|x\|_\alpha \leq \|x\|_\beta \leq n^{\left(\frac{1}{\beta} - \frac{1}{\alpha}\right)} \|x\|_\alpha.$$

This leads to the following useful relationships for the commonly employed norms mentioned previously:

$$\begin{aligned} \|x\|_2 &\leq \|x\|_1 \leq \sqrt{n}\, \|x\|_2 \\ \|x\|_\infty \leq \|x\|_1 &\leq n\, \|x\|_\infty \\ \|x\|_\infty \leq \|x\|_2 &\leq \sqrt{n}\, \|x\|_\infty \end{aligned}$$

Since all norms on $\mathbb{R}^n$ are equivalent, it is common to use the ℓ_2-norm, known as the Euclidean norm; i.e., we often define $\|\cdot\| := \|\cdot\|_2$. The ℓ_2-norm satisfies useful properties, as shown in the following theorems.

Theorem 2.21.2 (Cauchy-Schwarz Inequality)**.** *Given $\{x, \bar{x}\} \in \mathbb{R}^n$ and $\|\cdot\| := \|\cdot\|_2$, we have*

$$|x^T \bar{x}| \leq \|x\| \|\bar{x}\|.$$

Theorem 2.21.3 (Pythagorean Theorem)**.** *Given $\{x, \bar{x}\} \in \mathbb{R}^n$ and $\|\cdot\| := \|\cdot\|_2$, we have*

$$\|x + \bar{x}\|^2 = \|x\|^2 + 2x^T \bar{x} + \|\bar{x}\|^2,$$

meaning that if x and $\bar{x}$ are orthogonal (i.e., $x^T \bar{x} = 0$), then we have

$$\|x + \bar{x}\|^2 = \|x\|^2 + \|\bar{x}\|^2.$$

If $x \in \mathbb{R}^n$ has $\|x\| = 1$, then x is called a unit vector (with respect to the norm $\|\cdot\|$).

2.22 Norms and condition numbers of matrices

The notion of a norm of a vector can be extended to that of a matrix. Commonly used norms are vector-induced norms, where for a norm $\|\cdot\|$ on $\mathbb{R}^n$ we define the matrix norm of $A \in \mathbb{R}^{m\times n}$ by

$$\|A\| := \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

In the case of the common ℓ_1- and ℓ_∞-norms, this leads to the following explicit formulas for $A \in \mathbb{R}^{m\times n}$:

$$\|A\|_1 = \max_{j\in\{1,\ldots,n\}} \sum_{i=1}^{m} |a_{ij}| \quad \text{and} \quad \|A\|_\infty = \max_{i\in\{1,\ldots,m\}} \sum_{j=1}^{n} |a_{ij}|.$$

(For a simpler formula for the ℓ_2-norm of a matrix, see the discussion of eigenvalues of a matrix in §2.25.) As in the case of vectors, the ℓ_2-norm (i.e., Euclidean norm) satisfies useful properties, such as

$$\|AB\|_2 \leq \|A\|_2\|B\|_2 \quad \text{whenever } AB \text{ is well-defined.}$$

Another common matrix norm is the Frobenius norm, denoted $\|\cdot\|_F$, which is defined for $A \in \mathbb{R}^{m\times n}$ by

$$\|A\|_F := \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}.$$

If $A \in \mathbb{R}^{n\times n}$ is invertible, then for a given norm $\|\cdot\|$ we define the condition number of A as

$$\kappa(A) := \|A\|\|A^{-1}\|.$$

Different norms lead to different condition numbers; e.g., for $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$, one typically defines the corresponding condition numbers $\kappa_1(\cdot)$, $\kappa_2(\cdot)$, and $\kappa_\infty(\cdot)$, respectively.

2.23 Orthogonality of matrices

The concept of orthogonality of vectors is often discussed as a property of the vectors composing a matrix. For example, each pair of columns of a matrix $A \in \mathbb{R}^{m\times n}$ is orthogonal (orthonormal) if and only if A^TA is diagonal (the identity matrix). Similarly, each pair of rows of a matrix $A \in \mathbb{R}^{m\times n}$ is orthogonal (orthonormal) if and only if AA^T is diagonal (the identity matrix). However, oddly enough, the convention is that the term orthogonal matrix is reserved for any square matrix $A \in \mathbb{R}^{n\times n}$ such that each pair of columns (or rows) is orthonormal, which is to say that $A^TA = I$ and $AA^T = I$, i.e., $A^T = A^{-1}$. An important type of orthonormal matrix is a permutation matrix, which is any square matrix obtained by taking an identity matrix and permuting the orders of the rows and/or columns.

The complex analog of an orthogonal matrix is a unitary matrix. Specifically, a matrix $U \in \mathbb{C}^{n\times n}$ is unitary if $UU^* = I$ and $U^*U = I$, where U^* is the conjugate transpose of U; recall 2.10.

2.24 Matrix factorizations

Numerical algorithms make extensive use of matrix computations, which in turn may be made more efficient by performing matrix factorizations, also known as decompositions. If $A \in \mathbb{R}^{n\times n}$ is a matrix, then factoring (or decomposing) A involves computing other matrices—known as factors—whose product is A. This may be of interest if computations involving its factors are more efficient or can be performed more accurately (see Chapter 5) than computations directly with A or some function of A. In addition, properties of the factors of A often reveal important properties about A itself.

In the following section, two fundamental factorizations of a matrix A are discussed, commonly known as the singular value decomposition and, for square A, the eigenvalue decomposition. Other factorizations of importance in computational mathematics are presented in §2.27–2.29.

2.25 Singular value and eigenvalue decompositions

Any matrix $A \in \mathbb{R}^{m \times n}$ can be written as the product of three matrices with certain properties, as shown in this section. The decomposition of A into the product of these three matrices is known as the singular value decomposition (SVD) of A. If $m > n$, then the SVD of A has the form

$$A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V,$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{n \times n}$ is diagonal with entries

$$0 \leq \sigma_1 \leq \cdots \leq \sigma_n.$$

These entries are known as the singular values of A. Similarly, when $m \leq n$, the SVD of A has the form

$$A = U \begin{bmatrix} \Sigma & 0 \end{bmatrix} V,$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{m \times m}$ is diagonal with nonnegative entries being the singular values of A. For any (possibly nonsquare) matrix $A \in \mathbb{R}^{m \times n}$ with $\sigma_1 > 0$, we may define the condition number of A as the ratio of its largest to its smallest singular value, i.e., σ_n / σ_1. In cases when A is square and nonsingular, this corresponds to the value $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2$ (recall §2.22).

Given $A \in \mathbb{R}^{n \times n}$, consider any real or complex number λ such that for some nonzero $v \in \mathbb{R}^n$ we have

$$Av = \lambda v.$$

Such a real number is known as an eigenvalue of the matrix A, and the vector v is the corresponding eigenvector. A fundamental fact is that any matrix $A \in \mathbb{R}^{n \times n}$ has exactly n eigenvalues, all of which are either real or can be grouped into complex conjugate pairs. It also well-known that A is invertible if all of its eigenvalues are nonzero. The eigenvalues of a symmetric matrix are all real. In addition, a matrix is positive (negative) definite if and only if all of its eigenvalues are positive (negative). Similarly, a matrix is positive (negative) semidefinite if and only if all of its eigenvalues are nonnegative (nonpositive).

If $A \in \mathbb{R}^{n \times n}$ is symmetric, then its n real eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ and associated (real) unit eigenvectors $\{v_1, \ldots, v_n\}$ can be used to write the spectral decomposition of A, which has the form

$$A = \sum_{j=1}^{n} \lambda_j v_j v_j^T.$$

Defining the diagonal matrix $\Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, i.e., the diagonal matrix with (i,i)-entry equal to λ_i for all $i \in \{1, \ldots, n\}$, and $V = [v_1 \ \cdots \ v_n]$, the decomposition can also be written as

$$A = V \Lambda V^T.$$

If A is positive semidefinite, then its spectral decomposition coincides with its singular value decomposition.

There are important relationships between singular values, eigenvalues, and norms of matrices. For example, we have for any (possibly nonsquare) matrix $A \in \mathbb{R}^{m \times n}$ that

$$\|A\|_2^2 = \text{largest eigenvalue of } A^T A.$$

In particular, if $A \in \mathbb{R}^{n \times n}$ is positive definite, then

$$\|A\|_2 = \text{largest eigenvalue of } A,$$
$$\text{and} \quad \|A^{-1}\|_2 = \text{inverse of smallest eigenvalue of } A,$$

where, since the spectral and singular value decompositions coincide for a positive definite matrix, we may replace "eigenvalue" with "singular value" in these statements. Finally, for an orthogonal matrix $V \in \mathbb{R}^{n \times n}$,

$$\|Vx\|_2 = \|x\|_2 \quad \text{for all } x \in \mathbb{R}^n,$$

and all singular values of V are equal to 1.

2.26 Spectra, determinants, and traces

The spectrum of a matrix $A \in \mathbb{R}^{n\times n}$ is the set of its n eigenvalues. However, we may also be interested in information about such a square matrix that can be captured in a single number, and typically this number is related to the eigenvalues of the matrix. For example, the trace of A is defined as

$$\text{trace}(A) := \sum_{j=1}^{n} a_{jj},$$

and it can be shown that if A has eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$, then

$$\text{trace}(A) = \sum_{j=1}^{n} \lambda_j.$$

Another important quantity related to the matrix A is its determinant, which is defined as

$$\det(A) = \prod_{j=1}^{n} \lambda_j.$$

The determinant of A has many revealing properties. For example, $\det(A) = 0$ if and only if A is singular. Also, $\det(AB) = \det(A)\det(B)$ for any $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$, and if A is invertible, then $\det(A^{-1}) = 1/\det(A)$. Finally, if A is orthogonal, then $A^{-1} = A^T$ and $\det(A) = \det(A^{-1}) = \det(A^T) \in \{-1, 1\}$.

2.27 *LU* factorization

An *LU* factorization (with row partial pivoting) of a matrix $A \in \mathbb{R}^{n\times n}$ is defined as a factorization into a permutation matrix $P \in \mathbb{R}^{n\times n}$, a unit lower triangular matrix $L \in \mathbb{R}^{n\times n}$ (i.e., a lower triangular matrix with all diagonal elements equal to 1), and an upper triangular matrix $U \in \mathbb{R}^{n\times n}$ such that

$$PA = LU. \tag{2.27.3}$$

Any square matrix $A \in \mathbb{R}^{n\times n}$ admits such a factorization. The phrase *row partial pivoting* refers to the fact that multiplying A on the left by a permutation matrix P has the effect of rearranging the rows of A. Similarly, an *LU* factorization with column partial pivoting involves a permutation matrix Q such that $AQ = LU$, and an *LU* factorization with full pivoting involves permutation matrices P and Q such that $PAQ = LU$. The matrices L and U in these alternatives are not necessarily the same as those in (2.27.3), but they are still unit lower and upper triangular, respectively, as in (2.27.3).

A common use of such a factorization is solving systems of linear equations. If one aims to solve $Ax = b$ where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, then one can (i) compute the factors of A satisfying (2.27.3), (ii) compute the permuted right-hand side $\tilde{b} = Pb$, (iii) solve $L\tilde{x} = \tilde{b}$ for $\tilde{x} \in \mathbb{R}^n$, then (iv) solve $Ux = \tilde{x}$ for $x \in \mathbb{R}^n$. The idea here is that since L and U are triangular matrices, steps (iii) and (iv) can be performed efficiently via forward and backward substitution, respectively. For example, forward substitution for $L\tilde{x} = \tilde{b}$ involves the straightforward computation of the elements of $\tilde{x}$ in forward order:

$$\tilde{x}_1 \leftarrow \frac{\tilde{b}_1}{l_{1,1}}, \quad \tilde{x}_2 \leftarrow \frac{\tilde{b}_2 - l_{2,1}\tilde{x}_1}{l_{2,2}}, \quad \ldots, \quad \text{then} \quad \tilde{x}_n \leftarrow \frac{\tilde{b}_n - l_{n,1}\tilde{x}_1 - \cdots - l_{n,n-1}\tilde{x}_{n-1}}{l_{n,n}}.$$

This use of an *LU* factorization is equivalent to the well-known procedure known as Gaussian elimination.

Algorithm 2.27.1 computes the factors in (2.27.3) for a given input matrix $A \in \mathbb{R}^{n\times n}$.

When A is square and dense, a sophisticated implementation of Algorithm 2.27.1—that, for one thing, does not involve explicit computation and storage of a set of n permutation matrices—requires at most $2n^3/3$ arithmetic operations. This can be reduced if A has special structure; e.g., see §2.29.

Such a procedure can also be applied when $A \in \mathbb{R}^{m\times n}$ is not square. When $m > n$, then an algorithm similar to Algorithm 2.27.1 yields a permutation matrix $P \in \mathbb{R}^{m\times m}$, a unit lower triangular $L \in \mathbb{R}^{m\times n}$ (which is to say a matrix L with $L^T = [L_1^T \; L_2^T]$ with $L_1 \in \mathbb{R}^{n\times n}$ unit lower triangular), and an upper triangular $U \in \mathbb{R}^{n\times n}$ such that (2.27.3) holds. When $m < n$, then this modified algorithm can be used to compute such an *LU* factorization of A^T, i.e., such matrices P, L, and U such that $PA^T = LU$.

Algorithm 2.27.1 *LU* factorization with row partial pivoting

initialize $L \leftarrow 0$, $U \leftarrow A$, and $P_i^T \leftarrow I$ for all $i \in \{1, \ldots, n\}$
for $i = 1, \ldots, n$ **do**
  find $j \in \{i, \ldots, n\}$ such that $|u_{ji}| = \max_{k \in \{i,\ldots,n\}} |u_{ki}|$
  if $u_{ij} = 0$ **then** break
  if $i \neq j$ **then** replace P_i^T by the permutation matrix that swap rows i and j
  set $L \leftarrow P_i^T L$ and $U \leftarrow P_i^T U$
  set $l_{ii} \leftarrow 1$
  for $j = i + 1, \ldots, n$ **do**
    set $l_{ji} \leftarrow u_{ji}/u_{ii}$
    for $k = i + 1, \ldots, n$ **do**
      set $u_{jk} \leftarrow u_{jk} - l_{ji}u_{ik}$
    end for
  end for
end for
replace U by its upper triangular part (setting all other components to zero)
set $P \leftarrow P_n \cdots P_1$

2.28 QR factorization

A QR factorization of a matrix $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ is defined as a factorization into an orthogonal matrix $Q \in \mathbb{R}^{m \times m}$ and an upper triangular matrix $R \in \mathbb{R}^{m \times n}$ (which is to say a matrix R with $R^T = [R_1^T \; 0]$ with $R_1 \in \mathbb{R}^{n \times n}$ upper triangular) such that

$$A = QR. \tag{2.28.4}$$

Any such matrix admits such a factorization. A common use of such a factorization is solving overdetermined systems of linear equations such as $Ax = b$ when $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ has full column rank and $b \in \mathbb{R}^m$. In such a situation, one can (i) compute the factors of A satisfying (2.28.4), (ii) compute the permuted right-hand side $\tilde{b} = Q^T b$, and (iii) solve $Rx = \tilde{b}$ using backward substitution. Another use of such a factorization comes from the fact that when A has full column rank, the matrix Q_1 composed of the first n columns of Q satisfies $\text{range}(A) = \text{range}(Q_1)$ and the matrix Q_2 composed of the remaining $m - n$ columns of Q satisfies $\text{null}(A) = \text{range}(Q_2)$. That is, one can use a QR factorization to compute orthogonal bases for the range and null spaces of A. Unfortunately, when $\text{rank}(A) < n$, the factorization (2.28.4) might not be as useful in such situations; in particular, when solving $Ax = b$, the aforementioned procedure may break down in the back substitution phase, and one does not necessarily have $\text{range}(A) = \text{range}(Q_1)$. One can overcome these issues when A is column rank deficient by using a column pivoting procedure, but for simplicity in these notes we consider such situations out of our scope.

There are a variety of methods for computing the factors in (2.28.4) when A has full column rank. The most popular approaches involve Householder reflections, Givens rotations, or a Gram-Schmidt process. Here, we present a technique using Givens rotations. The use of Givens rotations can be understood by observing the effect of a 2-dimensional orthogonal rotation matrix, which for any $\theta \in \mathbb{R}$ has the form

$$G(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

One can confirm that premultiplying $x \in \mathbb{R}^2$ by $G(\theta)^T$ amounts to a counterclockwise rotation of x about the origin by θ radians. This means that, with an appropriate choice of θ, one can rotate x to a point $y = G(\theta)^T x$ with $y_2 = 0$. In particular, one can confirm that this is achieved by

$$G(x) = \begin{bmatrix} c(x) & s(x) \\ -s(x) & c(x) \end{bmatrix} \quad \text{with} \quad c(x) = \frac{x_1}{\sqrt{x_1^2 + x_2^2}} \quad \text{and} \quad s(x) = \frac{-x_2}{\sqrt{x_1^2 + x_2^2}}.$$

Generalizing this approach, one can take $x \in \mathbb{R}^m$ and define an orthogonal matrix $G(x_i, x_j) \in \mathbb{R}^{m \times m}$ so as to rotate the subvector $[x_i \; x_j]^T$ such that $y = G(x_i, x_j)^T x$ has $y_j = 0$ and $y_k = x_k$ for all $k \notin \{i, j\}$.

Algorithm 2.28.2 QR factorization using Givens rotations

initialize $Q \leftarrow I$ and $R \leftarrow A$
for $j = 1, \ldots, n$ **do**
  for $i = m, \ldots, j+1$ (decreasing) **do**
    set $Q \leftarrow QG(r_{i-1,j}, r_{i,j})$
    set $R \leftarrow G(r_{i-1,j}, r_{i,j})^T R$
  end for
end for

If $A \in \mathbb{R}^{m \times n}$ has full column rank, Algorithm 2.28.2 computes (Q, R) in (2.28.4) using Givens rotations.

When A is square and dense, a sophisticated implementation of Algorithm 2.28.2—that, for one thing, does not involve explicit computation and storage of each orthogonal rotation matrix—requires approximately $4m^2n/3$ arithmetic operations. This means that, when A is square, a QR factorization is roughly twice as expensive as an LU factorization. Consequently, if the goal is to solve a square system of equations, then the LU factorization is preferred, while if the goal is to solve a nonsquare system or explicitly to form orthogonal bases for the range and/or null space of A, then the QR factorization is preferred. If $A \in \mathbb{R}^{m \times n}$ with $m < n$, then one can apply a similar procedure to compute a QR factorization of A^T.

2.29 Cholesky and LBL^T factorizations

A Cholesky factorization of a symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$ is defined as a factorization into a lower triangular matrix $L \in \mathbb{R}^{n \times n}$ (or, equivalently, an upper triangular matrix $R \in \mathbb{R}^{n \times n}$) such that

$$A = LL^T \quad (\text{or } A = R^T R). \tag{2.29.5}$$

This can be seen as a special case of an LU factorization for cases when A is symmetric positive definite. The following algorithm computes the Cholesky factorization of such a matrix $A \in \mathbb{R}^{n \times n}$. By exploiting the symmetry of A—in particular, by only needing to access the lower triangular elements of A—it requires approximately $n^3/3$ arithmetic operations, making it roughly half as expensive as Algorithm 2.27.1.

Algorithm 2.29.3 Cholesky factorization

initialize $L \leftarrow 0$ and $\tilde{A} \leftarrow A$
for $i = 1, \ldots, n$ **do**
  set $l_{i,i} \leftarrow \sqrt{\tilde{a}_{i,i}}$
  for $j = i+1, \ldots, n$ **do**
    set $l_{j,i} \leftarrow \tilde{a}_{j,i}/l_{i,i}$
    for $k = i+1, \ldots, j$ **do**
      set $\tilde{a}_{j,k} \leftarrow a_{j,k} - l_{j,i}l_{k,i}$
    end for
  end for
end for

As for an LU factorization, a Cholesky factorization can be used when solving linear systems with a symmetric positive definite matrix. It can also be used to verify that a symmetric matrix is positive definite: If Algorithm 2.29.3 runs to completion with all $\{l_{i,i}\}$ elements positive, then A is positive definite.

When a matrix $A \in \mathbb{R}^{n \times n}$ is symmetric but not positive definite, then Algorithm 2.29.3 will break down when it tries to compute the square root of a negative number. However, all symmetric matrices admit a similar factorization, known as a symmetric indefinite factorization or LBL^T factorization, of the form

$$PAP^T = LBL^T, \tag{2.29.6}$$

where $P \in \mathbb{R}^{n \times n}$ is a permutation matrix, $L \in \mathbb{R}^{n \times n}$ is unit lower triangular, and B is a block-diagonal matrix with blocks of dimension at most 2. That is, B is symmetric and diagonal except one may have

$b_{i-1,i} = b_{i,i-1} \neq 0$ if and only if $b_{i-2,i-1} = b_{i-1,i-2} = 0$ and $b_{i,i+1} = b_{i+1,i} = 0$. Letting $\tilde{A}$ represent the matrix A after its rows and columns have been symmetrically permuted, an iteration toward computing the factors in (2.29.6) involves three steps: (i) Identifying a nonsingular submatrix of the form

$$E = \begin{bmatrix} \tilde{a}_{i,i} \end{bmatrix} \quad \text{or} \quad E = \begin{bmatrix} \tilde{a}_{i,i} & \tilde{a}_{i,j} \\ \tilde{a}_{j,i} & \tilde{a}_{j,j} \end{bmatrix}$$

to be used as a pivot block, (ii) finding a permutation matrix P so that

$$P\tilde{A}P^T = \begin{bmatrix} E & F^T \\ F & G \end{bmatrix}, \tag{2.29.7}$$

and (iii) performing a block factorization on this permuted matrix to obtain

$$P\tilde{A}P^T = \begin{bmatrix} I & 0 \\ FE^{-1} & I \end{bmatrix} \begin{bmatrix} E & 0 \\ 0 & G - FE^{-1}F^T \end{bmatrix} \begin{bmatrix} I & E^{-1}F^T \\ 0 & I \end{bmatrix}.$$

The process is then repeated by applying the same steps to the submatrix $G - FE^{-1}F^T$, otherwise known as the Schur complement of the matrix G with respect to the right-hand side of (2.29.7). Overall, this process involves approximately $n^3/3$ arithmetic operations—as in a Cholesky factorization—in addition to the work that must be performed to identify suitable pivot blocks.

2.30 Sherman-Morrison-Woodbury formula

A matrix $A \in \mathbb{R}^{n \times n}$ is said to undergo a rank-one update to become another matrix, call it $\overline{A} \in \mathbb{R}^{n \times n}$, if

$$\overline{A} \leftarrow A + cuv^T,$$

where $c \in \mathbb{R}$, $u \in \mathbb{R}^n$, and $v \in \mathbb{R}^n$ are all nonzero. (If any of c, u, or v is zero, then $cuv^T = 0$ and we have the trivial update $\overline{A} = A$.) This terminology comes from the fact that $\text{rank}(cuv^T) = 1$. A particular case of interest is when both $\overline{A}$ and A are invertible, in which case we have

$$\overline{A}^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}.$$

This is known as the Sherman-Morrison formula. Generalizing this result to higher rank updates of form

$$\overline{A} \leftarrow A + UCV^T,$$

where $\overline{A} \in \mathbb{R}^{n \times n}$, $A \in \mathbb{R}^{n \times n}$, and $C \in \mathbb{R}^{m \times m}$ are invertible, $U \in \mathbb{R}^{n \times m}$, and $V \in \mathbb{R}^{n \times m}$, we have that

$$\overline{A}^{-1} = A^{-1} - A^{-1}U(C^{-1} + V^TA^{-1}U)^{-1}V^TA^{-1}.$$

This is known as the Woodbury formula, or the Sherman-Morrison-Woodbury formula. The formula is useful when solving linear systems of the form $\overline{A}x = b$. In particular, we have that

$$x = \overline{A}^{-1}b = A^{-1}b - A^{-1}U(C^{-1} + V^TA^{-1}U)^{-1}V^TA^{-1}b,$$

so x can be found by solving $m + 1$ systems with A (to obtain $A^{-1}b$ and $A^{-1}U$), inverting the $m \times m$ matrices C and $C^{-1} + V^TA^{-1}U$, and performing elementary matrix algebra. If solving systems with A can be done cheaply and $m \ll n$, then this may be more efficient than solving the system with $\overline{A}$ directly.

2.31 Interlacing eigenvalue theorem

Another interesting result related to rank-one updates is the following.

Theorem 2.31.1 (Interlacing Eigenvalue Theorem)**.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric with eigenvalues (indexed in increasing order) $\{\lambda_1, \ldots, \lambda_n\}$, $v \in \mathbb{R}^n$ be a unit vector, and $c \in \mathbb{R}$. Then, denoting the eigenvalues of the matrix $A + cvv^T$ as those in the set (indexed in increasing order) $\{\xi_1, \ldots, \xi_n\}$, we have for $c > 0$ that*

$$\xi_n \geq \lambda_n \geq \xi_{n-1} \geq \lambda_{n-1} \geq \cdots \geq \xi_2 \geq \lambda_2 \geq \xi_1 \geq \lambda_1,$$

and we have for $c < 0$ that

$$\lambda_n \geq \xi_n \geq \lambda_{n-1} \geq \xi_{n-1} \geq \cdots \geq \lambda_2 \geq \xi_2 \geq \lambda_1 \geq \xi_1.$$

Moreover, in any case (i.e., for all $c \in \mathbb{R}$) we have

$$\sum_{j=1}^{n}(\xi_j - \lambda_j) = c.$$

An important consequence of this theorem is that, with $c > 0$, a rank-one update to a positive (semi)definite matrix results in a positive (semi)definite matrix. Another consequence is that, with $c < 0$, a rank-one update to a positive (semi)definite matrix results in one with at most one negative eigenvalue.

2.32 Sequences of real numbers

A sequence is an ordered set, i.e., a set in which the order of the elements matters. (In fact, sequences are typically defined as multisets so as to allow elements to appear more than once in the sequence.) With respect to the order of the sequence, each element of a sequence is typically numbered with an index, where for our purposes we use subscripts to denote the index; e.g., we may denote a sequence as $\{x_k : k \in \mathbb{N}\}$, which for brevity may be written as $\{x_k\}_{k\in\mathbb{N}}$, $\{x_k\}_{k=0}^{\infty}$, or simply $\{x_k\}$. (By $\{x_k\}_{k=a}^{b}$, we are referring to the ordered set composed of the elements of $\{x_k\}$ from index a to index b, inclusive.) Sequences can be finite or infinite. By choosing a subset of the elements of a sequence, we obtain a subsequence. Typically, a subsequence is defined through a definition or choice of an ordered subset $\mathcal{K}$ of $\mathbb{N}$, i.e., we may be interested in the subsequence $\{x_k : k \in \mathcal{K}\}$, which may also be written as $\{x_k\}_{k\in\mathcal{K}}$.

A sequence of real numbers $\{x_k\}$ is said to converge if there exists a real number x such that for every real number $\epsilon > 0$ we have $|x_k - x| < \epsilon$ for every k greater than some index K (that may depend on ϵ). In such a case, the sequence $\{x_k\}$ is said to converge to x, which is known as the limit of $\{x_k\}$. This may be indicated in various ways, including

$$\lim_{k\to\infty} x_k = x, \quad \{x_k\} \to x, \quad \text{or} \quad x_k \to x. \tag{2.32.8}$$

If for every real number α there exists some index K such that $x_k \geq \alpha$ for all $k \geq K$, then we say that $\{x_k\}$ diverges and we write $\{x_k\} \to \infty$; similarly, if for every real number α there exists some index K such that $x_k \leq \alpha$ for all $k \geq K$, then $\{x_k\}$ diverges and we write $\{x_k\} \to -\infty$. When one says that "$\{x_k\}$ converges," it is implicitly assumed that the limit point of the sequence is finite, but when one says "the limit of $\{x_k\}$ exists" or "$\{x_k\}$ has a limit," then the limit point might be finite or infinite.

The real number sequence $\{x_k\}$ is said to be bounded above (below) if there exists a real number α such that $x_k \leq \alpha$ ($x_k \geq \alpha$) for all $k \in \mathbb{N}$. Moreover, $\{x_k\}$ is said to be bounded if it is bounded above and below. The real number sequence $\{x_k\}$ is said to be monotonically nondecreasing (nonincreasing) if $x_{k+1} \geq x_k$ ($x_{k+1} \leq x_k$) for all k. If the inequality in this definition is replaced with a strict inequality, then the sequence is monotonically increasing (decreasing). In many cases, the monotonicity and convergence of a sequence are both important, so we have special notation to capture both concepts simultaneously. In particular, if $\{x_k\} \to x$ and $\{x_k\}$ is monotonically nondecreasing (nonincreasing), then we may simply write $\{x_k\} \nearrow x$ ($\{x_k\} \searrow x$).

Theorem 2.32.1. *Let $\{x_k\}$ be a real number sequence. Then, the following hold true.*

(a) If $\{x_k\}$ is bounded above and nondecreasing, then it converges.

(b) If $\{x_k\}$ is bounded below and nonincreasing, then it converges.

Consequently, if $\{x_k\}$ is bounded and nondecreasing or nonincreasing, then it converges.

Theorem 2.32.2. *If a real number sequence $\{x_k\}$ is monotonically nondecreasing, then it either converges or $\{x_k\} \to \infty$. Similarly, if $\{x_k\}$ is monotonically nonincreasing, then it either converges or $\{x_k\} \to -\infty$.*

As in the case of a monotonically nondecreasing (nonincreasing) sequence, we remark that a monotonically increasing (decreasing) sequence may converge or diverge.

Given a real number sequence $\{x_k\}$, consider the corresponding sequences (indexed by j) defined by

$$\overline{x}_j := \sup\{x_k : k \geq j\} \quad \text{and} \quad \hat{x}_j := \inf\{x_k : k \geq j\} \quad \text{for all } j \in \mathbb{N}.$$

The sequences $\{\overline{x}_j\}$ and $\{\hat{x}_j\}$ are nonincreasing and nondecreasing, respectively, and so, by Theorem 2.32.1, they have limits whenever $\{x_k\}$ is bounded above or below, respectively. The limit of $\{\overline{x}_j\}$ is denoted $\limsup_{k\to\infty} x_k$ and is referred to as the upper limit of $\{x_k\}$. Similarly, the limit of $\{\hat{x}_j\}$ is denoted $\liminf_{k\to\infty} x_k$ and is referred to as the lower limit of $\{x_k\}$. If $\{x_k\}$ is unbounded above, then we write $\limsup_{k\to\infty} x_k = \infty$, and if it is unbounded below, then we write $\liminf_{k\to\infty} x_k = -\infty$.

Theorem 2.32.3. *Let $\{x_k\}_{k=0}^{\infty}$ and $\{z_k\}_{k=0}^{\infty}$ be real number sequences. Then, the following hold true.*

(a) We have

$$\inf\{x_k : k \geq 0\} \leq \liminf_{k\to\infty} x_k \leq \limsup_{k\to\infty} x_k \leq \sup\{x_k : k \geq 0\}.$$

(b) The sequence $\{x_k\}$ converges if and only if

$$-\infty < \liminf_{k\to\infty} x_k = \limsup_{k\to\infty} x_k < \infty.$$

Furthermore, if $\{x_k\}$ converges, then its limit is equal to $\liminf_{k\to\infty} x_k$ and $\limsup_{k\to\infty} x_k$.

(c) If $x_k \geq z_k$ for all k, then

$$\liminf_{k\to\infty} x_k \geq \liminf_{k\to\infty} z_k \quad \text{and} \quad \limsup_{k\to\infty} x_k \geq \limsup_{k\to\infty} z_k$$

(d) We have

$$\liminf_{k\to\infty} x_k + \liminf_{k\to\infty} z_k \leq \liminf_{k\to\infty}(x_k + z_k)$$
$$\text{and} \quad \limsup_{k\to\infty} x_k + \limsup_{k\to\infty} z_k \geq \limsup_{k\to\infty}(x_k + z_k).$$

2.33 Sequences of real vectors

A sequence $\{x_k\}$ in $\mathbb{R}^n$ is said to converge to some $x \in \mathbb{R}^n$ if, for all $i \in \{1, \ldots, n\}$, the ith component of $\{x_k\}$ converges to the ith component of x. As for real number sequences, we use the notation in (2.32.8) to indicate that a vector sequence $\{x_k\}$ converges to x. The vector sequence $\{x_k\}$ is bounded above (below) if each of its corresponding component sequences is bounded above (below). We also have the following result in which, due to Theorem 2.21.1, the choice of norm $\|\cdot\|$ on $\mathbb{R}^n$ is arbitrary.

Theorem 2.33.1. *A sequence $\{x_k\}$ in $\mathbb{R}^n$ is bounded if and only if $\|x_k\| \leq \alpha$ for some $\alpha \in \mathbb{R}$ for all k.*

A vector $x \in \mathbb{R}^n$ is said to be a limit point of a sequence $\{x_k\}$ if there exists a subsequence of $\{x_k\}$ that converges to x. The following is a classical result related to bounded sequences and limit points.

Theorem 2.33.2 (Bolzano-Weierstrauss Theorem)**.** *A bounded sequence in $\mathbb{R}^n$ has at least one limit point.*

2.34 Rates of convergence

In this section, we describe two types of rates of convergence of sequences, one prefixed by Q (for "quotient") and the other prefixed by R (for "root"). When no prefix is indicated, the prefix Q is assumed.

Let $\{x_k\}$ be a sequence in $\mathbb{R}^n$ that converges to $x \in \mathbb{R}^n$ and let $\|\cdot\|$ be a vector norm on $\mathbb{R}^n$. The convergence of $\{x_k\}$ to x is Q-sublinear if

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x\|}{\|x_k - x\|} = 1.$$

The convergence of $\{x_k\}$ to x is Q-linear if there exists $c \in (0, 1)$ such that

$$\frac{\|x_{k+1} - x\|}{\|x_k - x\|} \leq c \quad \text{for all sufficiently large } k.$$

The constant c indicates the rate of linear convergence, i.e., if the above holds with $c = 1/2$, then the sequence is said to converge Q-linearly with rate $c = 1/2$. The convergence of $\{x_k\}$ to x is Q-superlinear if

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x\|}{\|x_k - x\|} = 0.$$

It is easily seen that any sequence that converges Q-superlinearly also converges Q-linearly.

We now distinguish between various types of Q-superlinear convergence. First, for the most common type, the convergence of $\{x_k\}$ to x is Q-quadratic if there exists $c > 0$ such that

$$\frac{\|x_{k+1} - x\|}{\|x_k - x\|^2} \leq c \quad \text{for all sufficiently large } k.$$

More generally, we say that the Q-order of convergence of $\{x_k\}$ to x is $p > 1$ if there exists $c > 0$ such that

$$\frac{\|x_{k+1} - x\|}{\|x_k - x\|^p} \leq c \quad \text{for all sufficiently large } k.$$

For example, this leads to the definitions of Q-cubic ($p = 3$) and Q-quartic ($p = 4$) convergence. Clearly, the rate of convergence depends on the constant c, but the dependence on $p > 1$ is more important; i.e., if the Q-order of convergence of $\{\bar{x}_k\}$ to $\bar{x}$ is p_1 and that of $\{\hat{x}_k\}$ to $\hat{x}$ is $p_2 > p_1$, then for all sufficiently large k we have $\|\bar{x}_k - \bar{x}\| > \|\hat{x}_k - \hat{x}\|$, regardless of the corresponding constants, call them c_1 and c_2, respectively.

A drawback of the above definitions is that they focus on sequences in which the decrease in norm to the limit point is consistent for all k. Thus, they do not apply for certain sequences that converge reasonably quickly, but whose elements do not approach the limit point in a straighforward manner. For such cases, we define a slightly weaker form of convergence known as R-order convergence. If $\{\epsilon_k\}$ converges to 0 and

$$\|x_k - x\| \leq \epsilon_k \quad \text{for all } k, \tag{2.34.9}$$

then the R-order of convergence of $\{x_k\}$ to x is said to be that of the Q-order of convergence of $\{\epsilon_k\}$ to 0. If $\{\epsilon_k\}$ converges Q-sublinearly to 0 and (2.34.9) holds, then $\{x_k\}$ converges R-sublinearly to x; if $\{\epsilon_k\}$ converges Q-linearly to 0 and (2.34.9) holds, then $\{x_k\}$ converges R-linearly to x; and so on.

2.35 Open, closed, and/or bounded sets

A set $\mathcal{X} \subseteq \mathbb{R}^n$ is open if for every $x \in \mathcal{X}$ there exists a real number $\epsilon > 0$ such that

$$\{\bar{x} : \|\bar{x} - x\| \leq \epsilon\} \subset \mathcal{X}.$$

On the other hand, a set $\mathcal{X} \subseteq \mathbb{R}^n$ is closed if for any sequence $\{x_k\}$ in $\mathcal{X}$, any limit point of $\{x_k\}$ (also called a closure point of the set $\mathcal{X}$) is an element of $\mathcal{X}$. A set is open if and only if its complement is closed,

though, by convention, one typically says that the empty set $\varnothing$ and the real space $\mathbb{R}^n$ are each both open and closed. The set of all closure points of a set $\mathcal{X}$ is called the closure of $\mathcal{X}$, and is denoted $\text{cl}(\mathcal{X})$.

A set $\mathcal{X} \subseteq \mathbb{R}^n$ is bounded if for a norm $\|\cdot\|$ on $\mathbb{R}^n$ there exists a constant α such that

$$\|x\| \leq \alpha \quad \text{for all } x \in \mathcal{X}.$$

A special case is when $\mathcal{X}$ is both bounded and closed, in which case it is called compact.

Given $x \in \mathbb{R}^n$, a real number $\epsilon > 0$, and a norm $\|\cdot\|$ on $\mathbb{R}^n$, the sets

$$\{\bar{x} : \|\bar{x} - x\| < \epsilon\} \quad \text{and} \quad \{\bar{x} : \|\bar{x} - x\| \leq \epsilon\}$$

are known as an open ball and an closed ball, respectively, about x. (Instead of ball, one may also say sphere.) An open (closed) neighborhood of a vector x is any open (closed) ball containing x.

The following theorem highlights the fact that, in the previous definitions, the choice of norm is arbitrary.

Theorem 2.35.1. *If a subset of $\mathbb{R}^n$ is open (closed, bounded, or compact) with respect to some norm $\|\cdot\|$ on $\mathbb{R}^n$, then it is open (closed, bounded, or compact) with respect to all other norms on $\mathbb{R}^n$.*

The notion of a neighborhood of a point allows us to define notions of interiors and boundaries of a set. In particular, a point x is an interior point of $\mathcal{X} \subseteq \mathbb{R}^n$ if there exists a neighborhood of x that is contained in $\mathcal{X}$; the set of all interior points of $\mathcal{X}$ is known as the interior of $\mathcal{X}$, and is denoted $\text{int}(\mathcal{X})$. On the other hand, any $x \in \text{cl}(\mathcal{X})$ that is not an interior point of $\mathcal{X}$ (i.e., $x \notin \text{int}(\mathcal{X})$) is said to be a boundary point of $\mathcal{X}$; the set of all boundary points of $\mathcal{X}$ is called the boundary of $\mathcal{X}$.

Theorem 2.35.2. *The following hold true.*

(a) The union of a finite collection of closed sets is closed.

(b) The intersection of any collection of closed sets is closed.

(c) The union of any collection of open sets is open.

(d) The intersection of a finite collection of open sets is open.

(e) A set is open if and only if all of its elements are interior points.

(f) Every subspace of $\mathbb{R}^n$ (with dimension less than n) is closed.

(g) A set $\mathcal{X}$ is compact if and only if every sequence of elements of $\mathcal{X}$ has a subsequence that converges to an element of $\mathcal{X}$.

(h) If $\{\mathcal{X}_k\}$ is a sequence of nonempty and compact sets such that $\mathcal{X}_{k+1} \subseteq \mathcal{X}_k$ for all k, then the intersection $\cap_{k=0}^{\infty} \mathcal{X}_k$ is nonempty and compact.

2.36 Order notation

Let $\{x_k\}$ be a real number sequence and let $\{\bar{x}_k\}$ be a nonnegative real number sequence. We write

$$x_k = \mathcal{O}(\bar{x}_k)$$

to indicate that there exists a positive constant c such that

$$|x_k| \leq c\bar{x}_k \quad \text{for all sufficiently large } k.$$

A stronger statement is

$$x_k = o(\bar{x}_k),$$

which (assuming $\bar{x}_k > 0$ for all k) indicates that the sequence of ratios $\{|x_k|/\bar{x}_k\}$ approaches zero, i.e.,

$$\lim_{k \to \infty} \frac{|x_k|}{\bar{x}_k} = 0,$$

Similarly, we write

$$x_k = \Omega(\overline{x}_k)$$

to indicate that there exists a positive constant c such that

$$|x_k| \geq c\overline{x}_k \quad \text{for all sufficiently large } k,$$

and we write

$$x_k = \omega(\overline{x}_k)$$

(assuming $|x_k| > 0$ for all k) to indicate that the sequence of ratios $\{\overline{x}_k/|x_k|\}$ approaches zero, i.e.,

$$\lim_{k\to\infty} \frac{\overline{x}_k}{|x_k|} = 0.$$

Finally, we write

$$x_k = \Theta(\overline{x}_k)$$

to indicate that there exist positive constants c_1 and c_2 such that

$$c_1\overline{x}_k \leq |x_k| \leq c_2\overline{x}_k,$$

which is equivalent to saying that $x_k = \mathcal{O}(\overline{x}_k)$ and $x_k = \Omega(\overline{x}_k)$.

All of the notation in the previous paragraph extends in obvious ways to the case of having a function $f : \mathbb{R}^n \to \mathbb{R}$ and a nonnegative function $\overline{f} : \mathbb{R}^n \to \mathbb{R}$. For example,

$$f(x) = \mathcal{O}(\overline{f}(x)) \quad \text{as } x \to \infty$$

indicates that there exists a positive constant c such that

$$|f(x)| \leq c\overline{f}(x) \quad \text{for all } x \text{ with sufficiently large norm},$$

whereas we write

$$f(x) = o(\overline{f}(x)) \quad \text{as } x \to \infty$$

to indicate that the ratio $|f(x)|/\overline{f}(x)$ converges to zero as $x \to \infty$. By default, the assumption is that we are interested in the behavior of the functions as $x \to \infty$ (and so this part of the statement is often unwritten), but we may also be interested in their behavior as x approaches other values. For example, we may write

$$f(x) = \mathcal{O}(\overline{f}(x)) \quad \text{as } x \to \hat{x}$$

to indicate that there exists a positive constant c such that

$$|f(x)| \leq c\overline{f}(x) \quad \text{for all } x \text{ with sufficiently small } \|x - \hat{x}\|.$$

2.37 Continuity of functions

Consider a set $\mathcal{X} \subset \mathbb{R}^n$, function $f : \mathcal{X} \to \mathbb{R}^m$, and point $x \in \mathcal{X}$. If there exists a point $y \in \mathbb{R}^m$ such that $\{f(x_k)\} \to y$ for every sequence $\{x_k\}$ in $\mathcal{X}$ such that $\{x_k\} \to x$, then we write

$$\lim_{\overline{x}\to x} f(\overline{x}) = y.$$

Similarly, if there exists $y \in \mathbb{R}^m$ such that $\{f(x_k)\} \to y$ for every sequence $\{x_k\}$ in $\mathcal{X}$ such that $\{x_k\} \to x$ and $x_k \leq x$ ($x_k \geq x$), then we write

$$\lim_{\overline{x}\nearrow x} f(\overline{x}) = y \quad \left(\lim_{\overline{x}\searrow x} f(\overline{x}) = y\right).$$

A function $f : \mathcal{X} \to \mathbb{R}^m$ is called continuous at $x \in \mathcal{X}$ if

$$\lim_{\overline{x} \to x} f(\overline{x}) = f(x).$$

Similarly, a function $f : \mathcal{X} \to \mathbb{R}^m$ is left-continuous (right-continuous) at $x \in \mathcal{X}$ if

$$\lim_{\overline{x} \nearrow x} f(\overline{x}) = f(x) \quad \left(\lim_{\overline{x} \searrow x} f(\overline{x}) = f(x) \right).$$

Finally, a (real number-valued) function $f : \mathcal{X} \to \mathbb{R}$ is lower semicontinuous (upper semicontinuous) at $x \in \mathcal{X}$ if for every sequence $\{x_k\}$ in $\mathcal{X}$ converging to x we have

$$f(x) \leq \liminf_{k \to \infty} f(x_k) \quad \left(f(x) \geq \limsup_{k \to \infty} f(x_k) \right).$$

We say that f is continuous over a subset of $\mathcal{X}$ if it is continuous at all points in the subset, and we say that f itself is continuous (without qualification) if it is continuous over the entire set $\mathcal{X}$. Similarly terminology is used for left-continuous, right-continuous, upper semicontinuous, and lower semicontinuous functions.

Theorem 2.37.1. *The following hold true.*

(a) Any vector norm $\|\cdot\|$ on $\mathbb{R}^n$ is a continuous function.

(b) If $\overline{f} : \mathbb{R}^p \to \mathbb{R}^m$ and $f : \mathbb{R}^n \to \mathbb{R}^p$ are continuous, then the composition $(\overline{f} \cdot f) : \mathbb{R}^n \to \mathbb{R}^m$ defined by $(\overline{f} \cdot f)(x) = \overline{f}(f(x))$ is a continuous function.

(c) If $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous and $\mathcal{X}$ is an open (closed) subset of $\mathbb{R}^m$, then the inverse image of $\mathcal{X}$, namely $\{x \in \mathbb{R}^n : f(x) \in \mathcal{X}\}$, is an open (closed) set.

(d) If $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous and $\mathcal{X}$ is a compact subset of $\mathbb{R}^n$, then the image of $\mathcal{X}$, namely $\{f(x) : x \in \mathcal{X}\}$, is a compact set.

Theorem 2.37.2 (Weierstrauss' Extreme Value Theorem)**.** *A continuous function $f : \mathbb{R}^n \to \mathbb{R}$ attains a minimum over any compact subset of $\mathbb{R}^n$.*

A function $f : \mathcal{X} \to \mathbb{R}^m$ is Lipschitz continuous (on its domain $\mathcal{X}$) if there exists $L \in \mathbb{R}_{>0}$ such that

$$\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\| \quad \text{for all} \quad (x_1, x_2) \in \mathcal{X} \times \mathcal{X}.$$

Any such positive constant L satisfying this definition is called *a* Lipschitz constant for f (on $\mathcal{X}$) while the smallest such positive constant is known as *the* Lipschitz constant for f (on $\mathcal{X}$). If the Lipschitz constant for a function is strictly less than one, then the function is called a contraction. A function $f : \mathcal{X} \to \mathbb{R}^m$ is locally Lipschitz continuous if for every $x \in \mathcal{X}$ there exists a neighborhood $\mathcal{N}$ of x such that f restricted to $\mathcal{N}$ (i.e., the function $f_{\mathcal{N}} : \mathcal{N} \to \mathbb{R}^m$ defined by $f_{\mathcal{N}}(x) = f(x)$ for all $x \in \mathcal{N}$) is Lipschitz continuous (on $\mathcal{N}$).

Lipschitz continuity is a special case of Hölder continuity. In particular, a function $f : \mathcal{X} \to \mathbb{R}^m$ is Hölder continuous of order α (on its domain $\mathcal{X}$) if there exists a real number $L \in \mathbb{R}_{>0}$ such that

$$\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\|^\alpha \quad \text{for all} \quad (x_1, x_2) \in \mathcal{X} \times \mathcal{X}.$$

By the equivalence of norms on $\mathbb{R}^n$, the norm $\|\cdot\|$ in the preceding paragraphs can be any norm on $\mathbb{R}^n$.

2.38 Differentiability of functions

A univariate function $f : \mathbb{R} \to \mathbb{R}$ is said to be differentiable at $x \in \mathbb{R}$ if and only if its first derivative, i.e.,

$$f'(x) := \lim_{\epsilon \to 0} \frac{f(x + \epsilon) - f(x)}{\epsilon}$$

exists and is finite at x. This definition extends to differentiability of a function over a set or even the entire domain; that is, such a function is differentiable over $\mathcal{X} \subseteq \mathbb{R}$ if and only if it is differentiable at all $x \in \mathcal{X}$. A univariate function f is said to be continuously differentiable over $\mathcal{X} \subseteq \mathbb{R}$ if and only if f' exists and is finite at all $x \in \mathcal{X}$ and f' itself is a continuous function over $\mathcal{X}$. Similarly, a univariate function f is said to be twice differentiable at $x \in \mathbb{R}$ if and only if its first derivative and its second derivative, i.e.,

$$f''(x) := \lim_{\epsilon \to 0} \frac{f'(x+\epsilon) - f'(x)}{\epsilon},$$

exist and are finite at x. This definition extends to twice differentiability over a set or even the entire domain. A univariate function f is said to be twice continuously differentiable over $\mathcal{X} \subseteq \mathbb{R}$ if and only if f' and f'' exist and are finite at all $x \in \mathcal{X}$ and f'' is continuous over $\mathcal{X}$. Higher order differentiation and derivatives are defined in a similar manner. The notation $f^{(p)}$ is often used for the pth derivative of f.

Now consider a multivariate function $f : \mathbb{R}^n \to \mathbb{R}$. We say that f is Fréchet differentiable (or simply differentiable) at x if and only if there exists a linear function $g_x : \mathbb{R}^n \to \mathbb{R}$ such that, for any vector norm $\|\cdot\|$ on $\mathbb{R}^n$, one has that

$$\lim_{d \to 0} \frac{|f(x+d) - f(x) - g_x(d)|}{\|d\|} = 0. \tag{2.38.10}$$

If such a function g_x exists, then it has the form $g_x(d) = \nabla f(x)^T d$, where

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$

is the gradient of f at x. In this componentwise form, the jth element (for each $j \in \{1, \ldots, n\}$) is known as the partial derivative of f with respect to x_j evaluated at x, and it can be shown that

$$\frac{\partial f}{\partial x_j}(x) = \lim_{\epsilon \to 0} \frac{f(x + \epsilon e_j) - f(x)}{\epsilon},$$

where e_j represents the n-vector of all zeros except for a one in the jth position (called the jth unit vector). A gradient with respect to only a subset of variables can be expressed via a subscript on the gradient symbol, namely ∇. For example, the gradient of $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ at $(x, \overline{x}) \in \mathcal{X} \times \mathcal{X}$ with respect only to the first set of variables (i.e., those corresponding to x) can be written as $\nabla_x f(x, \overline{x})$.

For a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$, we define $\nabla f(x)$ to be the $n \times m$ matrix whose ith column (for all $i \in \{1, \ldots, m\}$) is $\nabla f_i(x) \in \mathbb{R}^n$, i.e., the gradient of the real number-valued function f_i evaluated at x. It is common to work with the transpose of this matrix, which is called the Jacobian of f at x:

$$\nabla f(x)^T = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}.$$

Returning to the case of a real number-valued multivariate function $f : \mathbb{R}^n \to \mathbb{R}$, the matrix of second partial derivatives of f at x is known as the Hessian of f at x. It has the form

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{bmatrix}.$$

If f is twice-continuously differentiable, then the Hessian is always symmetric since, for any $x \in \mathbb{R}^n$,

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x) \quad \text{for all} \quad (i,j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}.$$

In summary, for a multivariate function $f : \mathbb{R}^n \to \mathbb{R}$, we say that it is differentiable over $\mathcal{X} \subseteq \mathbb{R}^n$ if $\nabla f(x)$ exists for all $x \in \mathcal{X}$, and we say that it is continuously differentiable over $\mathcal{X}$ if, in addition, ∇f is continuous over $\mathcal{X}$. Similarly, we say that f is twice differentiable over $\mathcal{X} \subseteq \mathbb{R}^n$ if $\nabla f(x)$ and $\nabla^2 f(x)$ exist for all $x \in \mathcal{X}$, and we say that it is twice continuously differentiable over $\mathcal{X}$ if, in addition, $\nabla^2 f$ is continuous over $\mathcal{X}$. If f is continuously differentiable over $\mathcal{X}$, then for any $x \in \mathcal{X}$ we have

$$f(x+d) = f(x) + \nabla f(x)^T d + \mathcal{O}(\|d\|_2^2) \quad \text{for all } d \in \mathbb{R}^n \text{ such that } x + d \in \mathcal{X}.$$

2.39 Directional derivatives

The existence of a function g_x such that (2.38.10) holds is a rather strong statement in that the limit implies convergence of the quotient to zero for any sequence $\{d_k\} \to 0$. We obtain a weaker statement by saying that f is directionally differentiable at $x \in \text{int}(\mathcal{X})$ along a given direction $d \in \mathbb{R}^n$, which is to say that the (one-sided) directional derivative of f at x along d, namely

$$f'(d; x) = \lim_{\alpha \searrow 0} \frac{f(x + \alpha d) - f(x)}{\alpha}$$

exists and is finite. If a function $f : \mathcal{X} \to \mathbb{R}$ is directionally differentiable at $x \in \text{int}(\mathcal{X})$ along any direction $d \in \mathbb{R}^n$, then f is said to be Gâteaux differentiable at x.

Theorem 2.39.1. *If $f : \mathcal{X} \to \mathbb{R}$ is Fréchet differentiable at x, then it is Gâteaux differentiable at $x \in \mathcal{X}$. However, if f is Gâteaux differentiable at x, then it is not necessarily Fréchet differentiable at x.*

Relating the concepts of differentiability and directional differentiability, we have that f is differentiable at x if and only if the gradient $\nabla f(x)$ exists and satisfies $\nabla f(x)^T d = f'(d; x)$ for all $d \in \mathbb{R}^n$, i.e.,

$$f(x + \alpha d) = f(x) + \alpha \nabla f(x)^T d + \mathcal{O}(|\alpha|^2) \quad \text{for all } \alpha \in \mathbb{R}.$$

2.40 Mean Value Theorem

The following is a useful result about continuously differentiable functions.

Theorem 2.40.1 (Mean Value Theorem)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be continuously differentiable over an open convex set $\mathcal{X} \subseteq \mathbb{R}^n$ and consider $x \in \mathcal{X}$ and $d \in \mathbb{R}^n$ such that $x + d \in \mathcal{X}$. Then, there exists $\alpha \in [0, 1]$ with*

$$f(x + d) = f(x) + \nabla f(x + \alpha d)^T d.$$

A generalization of this fact leads to the following well-known theorem attributed to Taylor. For this theorem, we require multi-index notation. For $\beta \in \mathbb{N}^n$ and $d \in \mathbb{R}^n$, we define

$$\begin{aligned} |\beta| &:= \beta_1 + \cdots + \beta_n, \\ \beta! &:= \beta_1! \cdots \beta_n!, \\ d^\beta &:= d_1^{\beta_1} \cdots d_n^{\beta_n}, \\ \text{and} \quad \partial^\beta f &:= \partial_1^{\beta_1} \cdots \partial_n^{\beta_n} f. \end{aligned}$$

Theorem 2.40.2 (Taylor's Theorem)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be $(k+1)$-times differentiable over an open convex set $\mathcal{X} \subseteq \mathbb{R}^n$ and consider $x \in \mathcal{X}$ and $d \in \mathbb{R}^n$ such that $x + d \in \mathcal{X}$. Then, there exists $\alpha \in [0, 1]$ such that*

$$f(x + d) = \sum_{|\beta| \leq k} \tfrac{1}{\beta!} \partial^\beta f(x) d^\beta + \sum_{|\beta| = k+1} \tfrac{1}{\beta!} \partial^\beta f(x + \alpha d) d^\beta.$$

In Theorem 2.40.2, the first term on the right-hand side is known as the kth order Taylor series approximation of f at x with the second term being the remainder. The form of the remainder written here is known as the Lagrange form; the remainder may also be written in integral form, leading to the equation

$$f(x + d) = \sum_{|\beta| \leq k} \tfrac{1}{\beta!} \partial^\beta f(x) d^\beta + (k+1) \sum_{|\beta| = k+1} \tfrac{1}{\beta!} d^\beta \int_0^1 (1 - \alpha)^k \partial^\beta f(x + \alpha d) d\beta.$$

As a consequence of Theorem 2.40.2, one obtains a natural extension of Theorem 2.40.1 to second-order derivatives. In particular, if, under the conditions of Theorem 2.40.1, the function f is twice continuously differentiable over $\mathcal{X}$, then there exists $\alpha \in [0,1]$ such that

$$f(x+d) = f(x) + \nabla f(x)^T d + \tfrac{1}{2} d^T \nabla^2 f(x+\alpha d) d.$$

2.41 Implicit Function Theorem

The following is another useful result about continuously differentiable functions.

Theorem 2.41.1 (Implicit Function Theorem)**.** *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^n$ be such that*

(i) $f(\bar{x}, 0) = 0$ for some $(\bar{x}, 0) \in \text{int}(\mathcal{X}) \times \text{int}(\mathcal{Y})$,

(ii) f is continuously differentiable in a neighborhood of $(\bar{x}, 0)$, and

(iii) $\nabla_x f(x, y)$ is nonsingular at $(x, y) = (\bar{x}, 0)$.

Then, there exist open sets $\mathcal{N}_\mathcal{X} \subset \mathcal{X}$ and $\mathcal{N}_\mathcal{Y} \subset \mathcal{Y}$ containing $\bar{x}$ and 0, respectively, and a continuous function $\hat{x} : \mathcal{N}_\mathcal{Y} \to \mathcal{N}_\mathcal{X}$ such that $\bar{x} = \hat{x}(0)$ and $f(\hat{x}(y), y) = 0$ for all $y \in \mathcal{N}_\mathcal{Y}$. In addition, the function $\hat{x}$ satisfying these properties is unique. Finally, if f is p-times continuously differentiable (for some $p > 0$) with respect to both of its arguments, then $\hat{x}$ is also p-times continuously differentiable and for all $y \in \mathcal{N}_\mathcal{Y}$ we have

$$\nabla \hat{x}(y) = -\nabla_y f(\hat{x}(y), y)(\nabla_x f(\hat{x}(y), y))^{-1}.$$

2.42 Norms of functions

Though this chapter has focused on quantities in $\mathbb{R}^n$, it is useful to understand the concept of a norm of a function, even in finite dimensional settings. There are a variety of types of function spaces considered by mathematicians, such as Hilbert, Banach, and Sobolev spaces. However, for simplicity, we focus on the function space $\mathcal{C}^k(\mathcal{X}, \mathbb{R}^n)$, which denotes the set of functions that map $\mathcal{X} \subseteq \mathbb{R}$ to $\mathbb{R}^n$ and are k-times differentiable on $\mathcal{X}$. Let $f : \mathcal{X} \to \mathbb{R}^n$. For $f \in \mathcal{C}^0(\mathcal{X}, \mathbb{R}^n)$, a common norm is the 0-norm defined by

$$\|f\|_0 := \max_{x \in \mathcal{X}} |f(x)|.$$

Similarly, for $f \in \mathcal{C}^1(\mathcal{X}, \mathbb{R}^n)$, we may use the 0-norm or the 1-norm defined by

$$\|f\|_1 := \max_{x \in \mathcal{X}} |f(x)| + \max_{x \in \mathcal{X}} |f'(x)|.$$

These definitions extend in the obvious way to define the p-norm on $\mathcal{C}^q(\mathcal{X}, \mathbb{R}^n)$ for any $q \geq p$. Another type of norm on such a function space is the L_p-norm defined by

$$\|f\|_{L_p} := \left(\int_{x \in \mathcal{X}} |f(x)|^p \, \mathrm{d}x \right)^{1/p},$$

which is reminiscient of (2.21.2). Observe that it makes sense to refer to the 0-norm as the L_∞-norm.

Chapter 3

Combinatorics

Combinatorics is a branch of mathematics encompassing the study and characterization of finite or countably infinite structures (recall §2.5). There are various subfields of combinatorics, but in these notes we restrict attention to a brief introduction of enumerative combinatorics, which concentrates on counting techniques. These concepts are important in their own right, and are especially important in the study of probability (see Chapter 4) as they can be used to calculate the relative likelihoods of different experimental outcomes.

3.1 Basic counting principles

Counting techniques are built upon a few basic principles. These principles appear to be straightforward, but enumerating and remembering them is extremely useful when considering complicated counting problems. The following are a few basic counting principles. In the first two, the same rule applies of one replaces "ways of doing thing i" with "outcomes of event i" and the remaining language is modified accordingly.

Theorem 3.1.1 (Summation principle)**.** *If there are m ways of doing one thing and n ways of doing another thing and both things cannot be done, then there are $m + n$ ways of doing one of the two things. Similarly, if there are n_1 ways of doing thing 1, n_2 ways of doing thing 2, and so on up to n_p ways of doing thing p, and only one thing can be done, then there are $n_1 + n_2 + \cdots + n_p$ ways of doing one of the p things.*

Theorem 3.1.2 (Multiplication principle)**.** *If there are m ways of doing one thing and n ways of doing another thing, then there are $m \cdot n$ ways of doing both things. Similarly, if there are n_1 ways of doing thing 1, n_2 ways of doing thing 2, and so on up to n_p ways of doing thing p, then there are $n_1 \cdot n_2 \cdots n_p$ ways of doing all p things.*

Theorem 3.1.3 (Inclusion-exclusion principle)**.** *If $\mathcal{X}$ and $\mathcal{Y}$ are finite sets, then the number of elements in their union is the number of elements in each minus the number of elements in their intersection; i.e.,*

$$|\mathcal{X} \cup \mathcal{Y}| = |\mathcal{X}| + |\mathcal{Y}| - |\mathcal{X} \cap \mathcal{Y}|.$$

More generally, if $\mathcal{X}_i$ is a finite set for all $i \in \{1, \ldots, n\}$, then

$$\left|\bigcup_{i=1}^{n} \mathcal{X}_i\right| = \sum_{i=1}^{n} |\mathcal{X}_i| - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |\mathcal{X}_i \cap \mathcal{X}_j| + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} |\mathcal{X}_i \cap \mathcal{X}_j \cap \mathcal{X}_k| - \cdots + (-1)^{n-1} \left|\bigcap_{i=1}^{n} \mathcal{X}_i\right|;$$

i.e., the number of elements in their union is found by including the number of elements in each, minus the number of elements in each pairwise intersection, plus the number of elements in each triple-wise intersection, minus the number of elements in each quadruple-wise intersection, and so on in an alternating fashion until one adds or subtracts the number of elements in the intersection of all of the sets.

Theorem 3.1.4 (Pigeonhole principle)**.** *If m items are placed in n bins where $m > n$, then at least one of the bins will end up with more than one item in it.*

This last principle may be obvious, but it can often be called upon to prove results that are somewhat unexpected. As a frivolous example, it can be used to argue that in any major city with more than 1,000,000 people, there are at least two people with the exact same numbers of hairs on their heads. This can be argued since a human typically has many fewer than 1,000,000 hairs on their head—so this can be considered an upper bound on the number of hairs on any person's head. Therefore, even if one first considers people with numbers of hairs on their head equal to 0, 1, 2, ..., 1,000,000, the next person considered would have the same number of hairs as one of these previously considered people.

3.2 Permutations

Numerous situations in which sophisticated counting techniques are called upon involve counting the number of ways in which one can place a set of objects into a uniquely ordered sequence. If you have n objects, then the number of ways that this can be done is referred to as the number of permutations of these objects. For example, if the objects are the elements of $\{1, 2, 3\}$, then the permutations are

$$\{1, 2, 3\},\ \{1, 3, 2\},\ \{2, 1, 3\},\ \{2, 3, 1\},\ \{3, 1, 2\},\ \text{and } \{3, 2, 1\}.$$

In general, the number of permutations of a set of n distinct objects is

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1.$$

This can be seen by noting that there are n possible choices for the first element in the sequence, then $n-1$ choices for the second element, and so on. The fact then follows from the multiplication principle. Similar logic can be used to state that the number of k-permutations of n objects—i.e., the number of ways to construct a sequence of length k from a set of distinct objects of size n—is equal to

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

An even simpler concept than the number of permutations of a set is the number of ways that you can fill k spots, where for each spot you are able to choose any from a set of n objects, even if an object has already been selected to fill another spot. Since there are n choices for the first spot, n choices for the second spot, etc., it follows that the number of ways that this can be done is n^k. These are sometimes referred to as permutations with repetition, but this is somewhat of a misnomer as they are not true permutations.

On the other hand, a more complicated concept than the number of permutations of a set of distinct objects is the number of permutations of a set of objects in which certain objects are identical to other objects in the set. (Such sets are better referred to as multisets; recall §2.3.) For example, the permutations of the elements of the set $\{1, 2, 2\}$ are only

$$\{1, 2, 2\},\ \{2, 1, 2\},\ \text{and } \{2, 2, 1\}.$$

Even though the numbers of elements in $\{1, 2, 3\}$ and $\{1, 2, 2\}$ are the same, there are fewer permutations of the elements of the latter set as certain rearrangements of the objects lead to the same result. In general, the number of permutations of a set of n objects in which object a_1 appears n_1 times, object a_2 appears n_2 times, and so on up to object a_p appearing n_p times is equal to

$$\frac{n!}{n_1! n_2! \cdots n_p!}.$$

For example, the number of permutations of the letters in MATHEMATICS is $11!/(2!2!2!) = 6{,}652{,}800$. This follows as there are 11 letters in the word with the letters M, A, and T each appearing twice.

3.3 Combinations

A signifying feature of a k-permutation from a set of n objects is that the order in which the k objects appear in the sequence matters when counting the number of possibilities. However, one is often interested

in the number of ways of choosing k objects from a set of n objects when the order in which the objects are chosen does not matter. This leads to the notion of the number of combinations of k objects chosen from a set of size n, which is equal to

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

This fact can be understood by noting that each k-permutation from a set of n objects is duplicated $k!$ times in the total number of k-permutations; thus, one obtains the number of combinations by taking the number of k-permutations and dividing by $k!$. For example, there are 60 different 3-permutations from $\{1, 2, 3, 4, 5\}$, but the combinations of 3 objects from this set are only

$$\{1,2,3\},\ \{1,2,4\},\ \{1,2,5\},\ \{1,3,4\},\ \{1,3,5\},\ \{1,4,5\},\ \{2,3,4\},\ \{2,3,5\},\ \{2,4,5\},\ \text{and } \{3,4,5\}.$$

It is worthwhile to note that one obtains a simple expression for the total number of combinations of any length from a set of n objects. In particular, it can be verified that

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n,$$

which can be viewed as the total number of subsets of any n-element set. In fact, one can arrive at this value in another way: Corresponding to each element of an n-element set, consider the two possible choices of "yes" and "no" as to whether the element will be included in a particular subset. The total number of "permutations with repetition" of these choices is equal to 2^n.

3.4 Partitions

Suppose that you have a set of n objects. A useful quantity is the number of ways of partitioning the set such that subset 1 has n_1 objects, subset 2 has n_2 objects, and so on until subset p has n_p objects, where in all subsets the order of the objects does not matter. Clearly, for this quantity to be well-defined, one must have $n_1 + n_2 + \cdots + n_p = n$. The number of possible partitions of this type can be determined using combinations and the multiplication principle. To start, notice that a combination is a special case of a partition in which one partitions a set of n objects into a "chosen" set of k objects and a "not chosen" set of $n - k$ objects; thus, $\binom{n}{k}$ is the number of ways to partition a set of n objects into subsets of sizes k and $n - k$. Generalizing this idea, one can determine the number of partitions by considering a p-stage process in which one choose n_1 elements in stage 1 (for subset 1), then n_2 elements in stage 2 (for subset 2), and so on. The multiplication principle then leads to the number of partitions being equal to

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-n_2-\cdots-n_{p-1}}{n_p}.$$

By expanding and canceling equal terms, one finds that this is equal to

$$\frac{n!}{n_1!n_2!\cdots n_p!} =: \binom{n}{n_1, n_2, \ldots, n_p}.$$

Chapter 4

Probability and Statistics

Although often discussed together and intertwined in their terminology, the subjects of probability and statistics have distinct scopes. Probability theory is a branch of mathematics that has been built to analyze and make predictions about random and/or uncertain phenomena. Rooted with concrete axiomatic principles, probability theory allows for the use of standard mathematical methods to concretely answer probabilistic questions. That is, even though they relate to uncertain phenomenon and their answers may only involve the probability that an event occurs, typical questions in probability theory have definite mathematical answers. Statistics, on the other hand, is the process of analyzing and extracting information about an unknown variable or model from a given set of data (e.g., observations), a process that almost undoubtedly involves an element of art along with knowledge about sound mathematical principles. The purpose of this chapter is to provide basic definitions of concepts related to probabilistic models along with a few fundamental topics related to statistics and statistical inference.

It is worthwhile to note at the outset that there are two major prevailing—and, some would say, competing—interpretations of the fundamental nature of probability. The first interpretation involves defining probabilities in terms of frequency of occurrence; e.g., in this interpretation, one would say that the probability of success for an experiment is the fraction of times that one would obtain a successful result if the experiment were repeated many times (or perhaps indefinitely). Those that submit to this view are called frequentists or objectivists. The other prevailing interpretation is that probabilities are more appropriately considered as a representation of an expert's subjective belief; e.g., in this interpretation, an expert has some (subjective) prior belief about the probability of a given event, which, whenever new information in obtained, would be updated to produce a posterior belief. The idea here is that, even if they started with different initial beliefs, a set of experts given enough evidence would eventually have similar assessments of the likelihoods of the outcomes of uncertain events. Those that submit to this view are called subjectivists or Bayesians. While a given person may consider themselves as submitting more to one or the other of these interpretations, one often finds that both are useful in modern statistical inference.

4.1 Probabilistic models

A probabilistic model is a mechanism for characterizing knowledge (or belief) about the outcomes of a particular process, typically referred to as an experiment. The idea is that while the outcome of an experiment may be uncertain, each (uncertain) outcome has a certain likelihood of occurring relative to the likelihoods of the other possible outcomes. The set of all possible outcomes of an experiment is known as its sample space. Any subset of a sample space (involving one or possibly many outcomes) is known as an event. Along with the sample space, the other essential element of a probabilistic model for a given experiment is a probability law, which assigns to any given event $\mathcal{X}$ a nonnegative number $\mathbb{P}(\mathcal{X})$ that encodes our knowledge (or belief) about the collective likelihood of the outcomes in $\mathcal{X}$.

For a valid probabilitistic model, one must place a few restrictions on the choices of sample space and probability law. The elements of the sample space should be distinct, mutually exclusive, and collectively exhaustive. That is, each outcome should be different and the sample space should include all possible

outcomes of the experiment. As for the probability law, it should assign to any event $\mathcal{X}$ a number $\mathbb{P}(\mathcal{X})$, called the probability of $\mathcal{X}$, such that the following axioms are satisfied.

- *Nonnegativity.* $\mathbb{P}(\mathcal{X}) \geq 0$ for any event $\mathcal{X}$.
- *Additivity.* If $\mathcal{X}$ and $\mathcal{Y}$ are disjoint events, then $\mathbb{P}(\mathcal{X} \cup \mathcal{Y}) = \mathbb{P}(\mathcal{X}) + \mathbb{P}(\mathcal{Y})$.
- *Normalization.* If Ω includes the entire sample space, then $\mathbb{P}(\Omega) = 1$.

The following theorem follows as a direct result of the above axioms.

Theorem 4.1.1 (Discrete Probability Law)**.** *If the sample space of a probabilistic model consists of a finite number of outcomes, e.g., $\Omega = \{x_1, \ldots, x_n\}$, then the probability law can be completely specified by the probabilities of the events that consist of a single element. In particular,*

$$\mathbb{P}(\{x_1, \ldots, x_n\}) = \mathbb{P}(x_1) + \cdots + \mathbb{P}(x_n).$$

A variety of properties of probabilistic models can be derived from the axioms above. For a few examples, letting $\mathcal{X}$ and $\mathcal{Y}$ be events in a given model, the following hold true.

- If $\mathcal{X} \subseteq \mathcal{Y}$, then $\mathbb{P}(\mathcal{X}) \leq \mathbb{P}(\mathcal{Y})$.
- $\mathbb{P}(\mathcal{X} \cup \mathcal{Y}) = \mathbb{P}(\mathcal{X}) + \mathbb{P}(\mathcal{Y}) - \mathbb{P}(\mathcal{X} \cap \mathcal{Y})$.
- $\mathbb{P}(\mathcal{X} \cup \mathcal{Y}) = \mathbb{P}(\mathcal{X}) + \mathbb{P}(\mathcal{X}^c \cap \mathcal{Y})$.
- $\mathbb{P}(\mathcal{X} \cup \mathcal{Y}) \leq \mathbb{P}(\mathcal{X}) + \mathbb{P}(\mathcal{Y})$.

The second property can be seen as a consequence of the inclusion-exclusion principle, Theorem 3.1.3.

4.2 Conditional probability

In certain situations, one is interested in defining a probabilistic model for an experiment in which partial information about the outcome is known or at least supposed. In particular, supposing that the outcome is within a given event $\mathcal{X}$, one may seek a new probability law that takes this supposition into account. The resulting law is said to specify, say, the conditional probability of an event $\mathcal{Y}$ given $\mathcal{X}$, denoted $\mathbb{P}(\mathcal{Y}|\mathcal{X})$. One can easily argue that, as long as $\mathbb{P}(\mathcal{X}) > 0$, the conditional probability of $\mathcal{Y}$ given $\mathcal{X}$ satisfies

$$\mathbb{P}(\mathcal{Y}|\mathcal{X}) = \frac{\mathbb{P}(\mathcal{X} \cap \mathcal{Y})}{\mathbb{P}(\mathcal{X})}.$$

Here, we refer to $\mathcal{X}$ as the conditioning event. A generalization of the above equation leads to the following.

Theorem 4.2.1 (Multiplication Rule)**.** *If all of the conditioning events have positive probability, then*

$$\mathbb{P}\left(\bigcap_{i=1}^{n} \mathcal{X}_i\right) = \mathbb{P}(\mathcal{X}_1)\mathbb{P}(\mathcal{X}_2|\mathcal{X}_1)\mathbb{P}(\mathcal{X}_3|\mathcal{X}_1 \cap \mathcal{X}_2) \cdots \mathbb{P}\left(\mathcal{X}_n \middle| \bigcap_{i=1}^{n-1} \mathcal{X}_i\right).$$

4.3 Independence

If the occurrence of an event $\mathcal{X}$ provides no information about the likelihood of an event $\mathcal{Y}$ in that knowledge of $\mathcal{X}$ does not affect the probability of $\mathcal{Y}$ (i.e., $\mathbb{P}(\mathcal{Y}|\mathcal{X}) = \mathbb{P}(\mathcal{Y})$), then one says that $\mathcal{Y}$ is independent of $\mathcal{X}$. When $\mathbb{P}(\mathcal{X}) > 0$, one obtains from the definition of conditional probability that such independence is equivalently stated as $\mathbb{P}(\mathcal{X} \cap \mathcal{Y}) = \mathbb{P}(\mathcal{X})\mathbb{P}(\mathcal{Y})$. Note that this latter equation holds even if $\mathbb{P}(\mathcal{X}) = 0$, and it shows that if $\mathcal{Y}$ is independent of $\mathcal{X}$, then $\mathcal{X}$ is independent of $\mathcal{Y}$, i.e., the relationship is symmetric. Generally, one says that the events $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ are independent if, for every $\mathcal{S} \subseteq \{1, \ldots, n\}$, one has

$$\mathbb{P}\left(\bigcap_{i \in \mathcal{S}} \mathcal{X}_i\right) = \prod_{i \in \mathcal{S}} \mathbb{P}(\mathcal{X}_i).$$

Given a pair of events $(\mathcal{X}, \mathcal{Y})$ and a third event $\mathcal{Z}$ with $\mathbb{P}(\mathcal{Z}) > 0$, one says that $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent given $\mathcal{Z}$ if, similar to the above, the following relationship holds true:

$$\mathbb{P}(\mathcal{X} \cap \mathcal{Y}|\mathcal{Z}) = \mathbb{P}(\mathcal{X}|\mathcal{Z})\mathbb{P}(\mathcal{Y}|\mathcal{Z}).$$

Moreover, if $\mathbb{P}(\mathcal{Y}|\mathcal{Z}) > 0$, then the condition above is equivalent to

$$\mathbb{P}(\mathcal{X}|\mathcal{Y} \cap \mathcal{Z}) = \mathbb{P}(\mathcal{X}|\mathcal{Z}).$$

It is important to note that independence of events does not imply conditional independence of the events. Similarly, conditional independence does not imply independence.

4.4 Total probability and Bayes' theorem

Applying the concept of conditional probability, we obtain the following two useful theorems. For the theorems, we define a partition of the sample space of an experiment as any set of events such that each possible outcome of the experiment is included in exactly one element of the set.

Theorem 4.4.1 (Total Probability Theorem)**.** *Let $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ be a partition of the sample space such that $\mathbb{P}(\mathcal{X}_i) > 0$ for all $i \in \{1, \ldots, n\}$. Then, for any event $\mathcal{Y}$, the probability of $\mathcal{Y}$ is given by*

$$\begin{aligned}\mathbb{P}(\mathcal{Y}) &= \mathbb{P}(\mathcal{X}_1 \cap \mathcal{Y}) + \cdots + \mathbb{P}(\mathcal{X}_n \cap \mathcal{Y}) \\ &= \mathbb{P}(\mathcal{X}_1)\mathbb{P}(\mathcal{Y}|\mathcal{X}_1) + \cdots + \mathbb{P}(\mathcal{X}_n)\mathbb{P}(\mathcal{Y}|\mathcal{X}_n).\end{aligned}$$

Theorem 4.4.2 (Bayes' Theorem)**.** *Let $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ be a partition of the sample space such that $\mathbb{P}(\mathcal{X}_i) > 0$ for all $i \in \{1, \ldots, n\}$. Then, for any $i \in \{1, \ldots, n\}$ and any event $\mathcal{Y}$ such that $\mathbb{P}(\mathcal{Y}) > 0$, the conditional probability of $\mathcal{X}_i$ given $\mathcal{Y}$ is given by*

$$\begin{aligned}\mathbb{P}(\mathcal{X}_i|\mathcal{Y}) &= \frac{\mathbb{P}(\mathcal{X}_i)\mathbb{P}(\mathcal{Y}|\mathcal{X}_i)}{\mathbb{P}(\mathcal{Y})} \\ &= \frac{\mathbb{P}(\mathcal{X}_i)\mathbb{P}(\mathcal{Y}|\mathcal{X}_i)}{\mathbb{P}(\mathcal{X}_1)\mathbb{P}(\mathcal{Y}|\mathcal{X}_1) + \cdots + \mathbb{P}(\mathcal{X}_n)\mathbb{P}(\mathcal{Y}|\mathcal{X}_n)}.\end{aligned}$$

Bayes' Theorem implies that for any events $\mathcal{X}$ and $\mathcal{Y}$ such that $\mathbb{P}(\mathcal{Y}) > 0$ one has

$$\mathbb{P}(\mathcal{X}|\mathcal{Y}) = \frac{\mathbb{P}(\mathcal{X})\mathbb{P}(\mathcal{Y}|\mathcal{X})}{\mathbb{P}(\mathcal{Y})} = \frac{\mathbb{P}(\mathcal{X})\mathbb{P}(\mathcal{Y}|\mathcal{X})}{\mathbb{P}(\mathcal{X})\mathbb{P}(\mathcal{Y}|\mathcal{X}) + \mathbb{P}(\mathcal{X}^c)\mathbb{P}(\mathcal{Y}|\mathcal{X}^c)}.$$

This fact is typically used for statistical inference. That is, given a certain number of causes that may result in a certain number of effects, one may aim to infer the cause when a particular effect is observed. In the notation above, the event $\mathcal{X}$ corresponds to a potential cause while $\mathcal{Y}$ corresponds to the observed effect. (Do not confuse the meaning of an effect with that of an outcome. In the present setting, one should imagine the "outcome" of the "experiment" being a cause-and-effect event that has occurred. We assume that we know the effect part of the pair, but not the cause part of it.) Given that the effect $\mathcal{Y}$ has been observed, we are interested in the probability that the cause was $\mathcal{X}$; i.e., we are interested in $\mathbb{P}(\mathcal{X}|\mathcal{Y})$. In this setting, one typically refers to this conditional probability as the posterior probability, which is computed from the equation based on some prior probability $\mathbb{P}(\mathcal{X})$ (which is assumed to be known).

For example, consider the following "False-Positive Puzzle" for which the intuition of many people is faulty: A test for a certain disease is assumed to be correct 95% of the time; i.e., if a person has the disease, then the test returns positive with probability 0.95, and if a person does not have the disease, then the test returns negative with probability 0.95. Suppose that a person drawn randomly from the population has the disease with probability 0.001. If a person drawn randomly from the population tests positive for the disease, then what is the probability that they actually have it? In fact, the probability is quite small. Indeed, if $\mathcal{X}$ is the event that the person has the disease and $\mathcal{Y}$ is the event that they test positive, then the probability of interest (i.e., the probability they have the disease given that they tested positive) is

$$\mathbb{P}(\mathcal{X}|\mathcal{Y}) = \frac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot 0.05} \approx 0.0187.$$

In hindsight, one can see that this probability greatly depends on the likelihood of having a "false-positive." If a randomly drawn person tests positive for the disease, then they are much more likely *not* to have the disease and receive an incorrect diagnosis than they are to actually have the disease.

4.5 Random variables

The outcome of an experiment is not always a numerical value. For example, the outcome of a flip of a coin is either "heads" or "tails." However, if the outcome of an experiment is a numerical value, or if one has a strategy of associating numerical values to the outcomes of an experiment, then it is often useful to assign probabilities to these numerical values. This is done through the concept of a random variable. Specifically, supposing without loss of generality that the numerical values obtained from an experiment are real, a random variable is a real-valued function of the outcome of an experiment. For example, the outcome of a flip of a coin is not a random variable, but the number of heads obtained in a flip (or a sequence of flips) of a coin is a random variable. A discrete random variable is one whose range is a finite or countably infinite set, whereas a general random variable may take values in an uncountable set. A particular type of general random variable is a continuous random variable, the definition of which is given in §4.6.

Throughout the remainder of this chapter, we assume without loss of generality that any random variable X is real-valued. Here, we follow the widespread convention of using capital letters to denote random variables. We also note that many of the concepts discussed in the remainder of the chapter are easily generalized to the case of having a random vector $X \in \mathbb{R}^n$.

4.6 Probability distribution functions

Probabilities corresponding to the values that a random variable X can take are characterized by a probability distribution function, commonly referred to as the distribution function (or simply distribution) of X. If X is a discrete random variable, then one can ascribe nonzero probabilities to each of the finite or countably infinite values that X can take. In such cases, the distribution of X is characterized by a probability mass function (or PMF); see §4.7. However, if X is a general random variable, then one may not be able to assign nonzero probabililities to all values that X can take. (Intuitively, if one assigns a nonzero probability to all of the uncountable number of values that X can take, then the accumulation of these probabilities would be infinite, violating the *normalization* axiom for valid probabilistic models; recall §4.1.) In such cases, one can only assign nonzero probabilities to events that X falls into particular intervals of values. A continuous random variable X is one for which the probabilities of such events can be characterized by a probability density function (or PDF); see §4.8.

Another important function related to a random variable X is its cumulative distribution function (or CDF), which, for any real value, captures the probability that X takes a value less than or equal to that real value; see §4.9. This definition of a CDF applies for both discrete and general random variables.

4.7 Probability mass functions

The probability mass function corresponding to a discrete random variable X is a function $f : \mathbb{R} \to [0, 1]$ that characterizes the probabilities of the values that X can take. In particular, if X is a discrete random variable with such a probability mass function f, then, for any real number x,

$$f(x) = \mathbb{P}(X = x).$$

Here, the right-hand side would be written more precisely using set notation as $\mathbb{P}(\{X = x\})$, but, for brevity here and throughout the remainder of this chapter, we use the abbreviated notation above. Similarly, later on, probabilities such as $\mathbb{P}(\{X \leq x\})$ and $\mathbb{P}(\{X = x\} \cap \{Y = y\})$ will be written simply as $\mathbb{P}(X \leq x)$ and $\mathbb{P}(X = x, Y = y)$, respectively. In any case, the relationships in the parentheses describe an event.

Some examples of discrete random variables and their probability mass functions are as follows. For brevity, we only indicate the real values for which each probability mass function takes a nonzero value.

- *Discrete uniform.* If the outcome of an experiment is an integer value in the interval $[\underline{x}, \overline{x}]$ with $(\underline{x}, \overline{x}) \in \mathbb{Z} \times \mathbb{Z}$ and all such outcomes are equally likely, then the outcome is represented by a discrete uniform random variable X with probability mass function

$$f(x) = \frac{1}{\overline{x} - \underline{x} + 1} \quad \text{for all} \quad x \in [\underline{x}, \overline{x}] \cap \mathbb{Z}.$$

- *Bernoulli.* Suppose that an experiment has a probability $p \in [0, 1]$ of "success" and a probability $1 - p$ of "failure", and that one assigns the real value 1 to "success" and 0 to "failure." Then, the outcome of the experiment is represented by a Bernoulli random variable X with probability mass function

$$f(x) = \begin{cases} p & \text{if } x = 1; \\ 1 - p & \text{if } x = 0. \end{cases}$$

- *Geometric.* If an experiment whose outcome can be represented by a Bernoulli random variable with parameter p (as above) is run repeatedly until the first success, then the random variable corresponding to the number of trials needed until the first success is represented by a geometric random variable with parameter p. Such a random variable has the probability mass function

$$f(x) = p(1 - p)^{x-1} \quad \text{for all} \quad x \in \{1, 2, \dots\}.$$

- *Binomial.* If an experiment whose outcome can be represented by a Bernoulli random variable with parameter p (as above) is run n times, then the random variable corresponding to the number of successes is represented by a binomial random variable with parameters n and p. Such a random variable has the probability mass function

$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad \text{for all} \quad x \in \{0, 1, \dots, n\}.$$

- *Poisson.* If events occur at a rate λ (in number of events per unit time) independent of the time of the last event, then the number of events that occur within a unit of time is represented by a Poisson random variable with parameter λ. The probability mass function for such a random variable is

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{for all} \quad x \in \{0, 1, 2, \dots\}.$$

A Poisson random variable can also be used to approximate a binomial random variable when n is relatively large and p is relatively small. Indeed, the number of events that occur within a unit of time is related to a total number of successes if one breaks the time unit into small subintervals and defines "success" to mean that an event occurs within a given subinterval. (The idea is that, if the subintervals into which the time unit is split are small enough that the probability of two events occurring within one subinterval is negligible, then the subintervals can be thought of as separate trials of an experiment.) This relationship is formalized in the following theorem, also known as the law of rare events.

Theorem 4.7.1 (Poisson Limit Theorem)**.** *If $n \to \infty$ and $p \to 0$ such that $np \to \lambda$, then*

$$\binom{n}{x} p^x (1 - p)^{n-x} \to \frac{\lambda^x e^{-\lambda}}{x!}.$$

4.8 Probability density functions

A random variable X is called continuous if there exists a function $f : \mathbb{R} \to \mathbb{R}_+$, called the probability density function of X, such that, for any $\{\underline{x}, \overline{x}\} \in \mathbb{R} \times \mathbb{R}$,

$$\mathbb{P}(X \in [\underline{x}, \overline{x}]) = \int_{\underline{x}}^{\overline{x}} f(x)\, dx.$$

By this definition, it is clear that for a continuous random variable X one has $\mathbb{P}(X = x) = 0$ for any $x \in \mathbb{R}$, from which it follows that including or excluding endpoints of an interval have no effect on the probability in the displayed equation above; i.e., $\mathbb{P}(X \in [\underline{x}, \overline{x}]) = \mathbb{P}(X \in (\underline{x}, \overline{x}]) = \mathbb{P}(X \in [\underline{x}, \overline{x})) = \mathbb{P}(X \in (\underline{x}, \overline{x}))$. It is also worthwhile to note that a probability density function can take arbitrarily large values, but to qualify as a PDF the function f must be nonnegative and satisfy the normalization property

$$P(X \in (-\infty, \infty)) = \int_{-\infty}^{\infty} f(x)\, dx = 1.$$

Some examples of continuous random variables and their probability density functions are as follows.

- *Uniform.* A random variable X is uniform over an interval $[\underline{x}, \overline{x}]$ if it has the PDF

$$f(x) = \begin{cases} (\overline{x} - \underline{x})^{-1} & \text{if } x \in [\underline{x}, \overline{x}], \\ 0 & \text{otherwise.} \end{cases}$$

- *Exponential.* A random variable X is exponential with parameter $\lambda > 0$ if it has the PDF

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

  This type of random variable is named for the fact that the probability that X exceeds a given real value decreases exponentially. That is, for any $x \geq 0$, it follows from this definition that

$$\mathbb{P}(X \geq x) = \int_x^{\infty} \lambda e^{-\lambda \tilde{x}}\, d\tilde{x} = -e^{-\lambda x}\big|_x^{\infty} = e^{-\lambda x}.$$

- *Normal (or Gaussian).* A random variable X is normal with parameters μ and $\sigma^2 \geq 0$ if its PDF is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

  In particular, a standard normal random variable is so-defined when $\mu = 0$ and $\sigma = 1$.

4.9 Cumulative distribution functions

The past two sections have shown that one needs to describe discrete and continuous random variables differently when it comes to their probability distribution functions. The notion of a cumulative distribution function, on the other hand, represents a single entity through which one can describe the probabilities associated with any random variable, either discrete or general (e.g., continuous). In particular, if X is a random variable, then its cumulative distribution function $F : \mathbb{R} \to [0, 1]$ is defined as

$$F(x) = \mathbb{P}(X \leq x) = \begin{cases} \displaystyle\sum_{\tilde{x} \leq x} f(\tilde{x}) & \text{if } X \text{ is discrete,} \\ \\ \displaystyle\int_{-\infty}^{x} f(\tilde{x})\, d\tilde{x} & \text{if } X \text{ is continuous.} \end{cases}$$

The CDF also provides a convenient way to characterize another type of general random variable, known as a mixed random variable, defined as a mixture between a discrete random variable, say Y, and a continuous random variable, say Z. For example, if X is a random variable that follows the probability law of Y with probability p and that of Z with probability $1 - p$, then X is a mixed random variable with CDF

$$F(x) = \mathbb{P}(X \leq x) = p\mathbb{P}(Y \leq x) + (1 - p)\mathbb{P}(Z \leq x),$$

the right-hand side of which is well-defined by the CDFs of Y and Z.

For any random variable X, it follows that its cumulative distribution function F is a monotonically nondecreasing function that tends to 0 when $x \to -\infty$ and tends to 1 when $x \to \infty$. If X is discrete, then F is a piecewise constant function of x, while if X is continuous, then F is also continuous.

4.10 Expected value, variance, and moments

The probability or cumulative distribution function of a random variable provides numbers corresponding to probabilities that the random variable takes on certain values. Often, however, it is convenient to encapsulate a relevant property (or properties) of a random variable in a single (or a few) numbers. In particular, a common property of a random variable is its expected value (or mean), which can be understood as the value one would expect to get if one were to average numerous "realized values" or "realizations" of the random variable. For a discrete or continuous random variable X with probability mass or density function f, this expected value is denoted and defined as

$$\mathbb{E}(X) = \begin{cases} \sum_x f(x)x & \text{if } X \text{ is discrete,} \\ \\ \int_{-\infty}^{\infty} f(x)x\,dx & \text{if } X \text{ is continuous.} \end{cases}$$

(In the right-hand side above, it is possible that the sum is not convergent or the integral is infinite or not well-defined, in which case the expected value is infinite or not well-defined. For simplicity throughout this document, we assume that such a sum is always convergent and such an integral is always finite and well-defined. Issues related to when they are not can be found in more advanced treatments of these topics.) More generally, if $g : \mathbb{R} \to \mathbb{R}$ is a function, then the expected value of $g(X)$ is given by

$$\mathbb{E}(g(X)) = \begin{cases} \sum_x f(x)g(x) & \text{if } X \text{ is discrete,} \\ \\ \int_{-\infty}^{\infty} f(x)g(x)\,dx & \text{if } X \text{ is continuous.} \end{cases}$$

The expected value of a random variable X is also known as the first moment of X. More generally, the nth moment of X is $\mathbb{E}(X^n)$, i.e., the expected value of $g(X) = X^n$. Related to the second moment of X is an important quantity known as the variance of X, which is denoted and defined as

$$\text{var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2);$$

i.e., the variance of X is the expected squared difference between X and its expected value. This quantity measures the extent to which one expects X to vary from its mean; the larger the value for $\text{var}(X)$, the more one expects that a particular realization of X may vary from its mean. Since $(X - \mathbb{E}(X))^2$ is nonnegative, it follows that $\text{var}(X)$ is always nonnegative. One also finds that

$$\text{var}(X) = \mathbb{E}(X^2 - 2X\mathbb{E}(X) + \mathbb{E}(X)^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2,$$

which reveals the relationship between $\text{var}(X)$ and the first and second moments of X.

If the random variable X takes values measured with certain units (e.g., number of horses), then $\mathbb{E}(X)$ has a value measured in the same units while $\text{var}(X)$ takes values of those units squared (e.g., horses^2). This makes it somewhat unnatural to compare the magnitudes of $\mathbb{E}(X)$ and $\text{var}(X)$. Thus, an important related measure of dispersion of X from its mean is the standard deviation of X, denoted and defined as

$$\text{std}(X) = \sqrt{\text{var}(X)}.$$

The standard deviation is measured in the same units as X and $\mathbb{E}(X)$.

4.11 Joint and marginal distribution functions

It is common for a probability model to involve multiple quantities, each of which can be represented by its own random variable. This motivates the introduction of probability distribution functions and related concepts for events involving multiple random variables. For simplicity, we restrict attention to the situation

of having two random variables, say X and Y, with the idea that the following concepts can be similarly defined when one has a finite or countably infinite set of random variables.

First, suppose that X and Y are discrete random variables. The joint probability mass function (or joint PMF) of X and Y is a function $f : \mathbb{R} \times \mathbb{R} \to [0, 1]$ defined for all $x \in \mathbb{R}$ and $y \in \mathbb{R}$ as

$$f(x, y) = \mathbb{P}(X = x, Y = x).$$

Such a joint PMF encapsulates the individual PMFs of X and Y, referred to in such a context as the marginal PMFs of X and Y. These can be recovered, respectively, using the formulas

$$f_X(x) = \sum_y f(x, y) \quad \text{and} \quad f_Y(y) = \sum_x f(x, y).$$

Similarly, supposing now that X and Y are continuous random variables, the joint probability density function (or joint PDF) of X and Y can be represented as a function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ defined for all pairs $(\underline{x}, \overline{x}) \in \mathbb{R} \times \mathbb{R}$ and $(\underline{y}, \overline{y}) \in \mathbb{R} \times \mathbb{R}$ as

$$\mathbb{P}(X \in [\underline{x}, \overline{x}], Y \in [\underline{y}, \overline{y}]) = \int_{\underline{x}}^{\overline{x}} \int_{\underline{y}}^{\overline{y}} f(x, y)\, dy\, dx.$$

From such a function, the marginal PDFs of X and Y can be recovered, respectively, by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx.$$

One can also similarly define joint and marginal distribution functions when, say, X is discrete and Y is continuous, or vice versa. Whether X and Y are discrete, continuous, or general, their joint cumulative distribution function (or joint CDF) is a function $F : \mathbb{R} \times \mathbb{R} \to [0, 1]$ such that, for all $(x, y) \in \mathbb{R} \times \mathbb{R}$,

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

4.12 Conditional distribution functions

As in the discussion in §4.2, one may be interested in characterizing the probability or cumulative distribution function for a random variable conditioning on an event $\mathcal{X}$ having occurred. This is easily done with appropriate normalization. For example, the conditional PMF of a discrete random variable X conditioning on an event $\mathcal{X}$ with $\mathbb{P}(\mathcal{X}) > 0$ is the function $f_{X|\mathcal{X}}(\cdot|\mathcal{X}) : \mathbb{R} \to [0, 1]$ such that, for all $x \in \mathbb{R}$,

$$f_{X|\mathcal{X}}(x|\mathcal{X}) = \mathbb{P}(X = x|\mathcal{X}) = \frac{\mathbb{P}(\{X = x\} \cap \mathcal{X})}{P(\mathcal{X})}.$$

If the event is characterized by another discrete random variable Y taking the value y, then the conditional PMF of X given $Y = y$ (or simply "given Y") is the function $f_{X|Y} : \mathbb{R} \to [0, 1]$ with

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)},$$

where $f_{X,Y}$ and f_Y are the joint PMF of X and Y and the corresponding marginal PMF of Y, respectively. From this definition, one obtains a variety of useful relationships between the joint PMF of X and Y, the conditional PMF of X given Y (or Y given X), and the marginal PMFs of X and Y. In particular, the joint PMF of X and Y can be represented via the product of a marginal and conditional PMF as in

$$f_{X,Y}(x, y) = f_Y(y) f_{X|Y}(x|y) \quad \text{and} \quad f_{X,Y}(x, y) = f_X(x) f_{Y|X}(y|x),$$

from which it follows that each marginal PMF can be represented via a sum of such quantities as in

$$f_X(x) = \sum_y f_Y(y) f_{X|Y}(x|y) \quad \text{and} \quad f_Y(y) = \sum_x f_X(x) f_{Y|X}(y|x).$$

The definitions in the previous paragraph can be extended in natural ways to continuous random variables. In particular, the conditional PDF of a continuous random variable X conditioning on an event $\mathcal{X}$ with $P(\mathcal{X}) > 0$ is the function $f_{X|\mathcal{X}}(\cdot|\mathcal{X}) : \mathbb{R} \to \mathbb{R}_+$ such that, for all $(\underline{x}, \overline{x}) \in \mathbb{R} \times \mathbb{R}$,

$$\mathbb{P}(X \in [\underline{x}, \overline{x}]|\mathcal{X}) = \int_{\underline{x}}^{\overline{x}} f_{X|\mathcal{X}}(x|\mathcal{X})\, dx.$$

If the event is characterized by another continuous random variable Y taking the value y, then the conditional PDF of X given Y is the function $f_{X|Y} : \mathbb{R} \to \mathbb{R}_+$ such that, for all $x \in \mathbb{R}$,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

While this formula may appear natural, its derivation is not trivial; after all, since Y is a continuous random variable, the event $Y = y$ occurs with probability zero! Still, the formula applies and, for all $(\underline{x}, \overline{x}) \in \mathbb{R} \times \mathbb{R}$,

$$\mathbb{P}(X \in [\underline{x}, \overline{x}]|Y = y) = \int_{\underline{x}}^{\overline{x}} f_{X|Y}(x|y)\, dx.$$

4.13 Expected values (continued)

As in the discussion in §4.10, when multiple random variables X and Y are present, one may desire to encapsulate information about the pair (X, Y) in one or a few values. For example, the expected value of a function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of X and Y can be obtained via their joint distribution function by

$$\mathbb{E}(g(X, Y)) = \begin{cases} \displaystyle\sum_{x,y} f(x, y)g(x, y) & \text{if } X \text{ and } Y \text{ are discrete,} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)g(x, y)\, dy\, dx & \text{if } X \text{ and } Y \text{ are continuous.} \end{cases}$$

Similarly, given a function $g : \mathbb{R} \to \mathbb{R}$, one can use conditional distribution functions to compute, say, the expected value of $g(X)$ given that Y takes on the value $y \in \mathbb{R}$:

$$\mathbb{E}(g(X)|Y = y) = \begin{cases} \displaystyle\sum_{x} f_{X|Y}(x|y)g(x) & \text{if } X \text{ is discrete,} \\ \displaystyle\int_{-\infty}^{\infty} f_{X|Y}(x|y)g(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

Using the these formulas, we obtain the following theorem, which can be seen as an extension of the Total Probability Theorem (i.e., Theorem 4.4.1) applied to expected values.

Theorem 4.13.1 (Total Expectation Theorem)**.** *If Y is a discrete random variable with probability mass function $f_Y : \mathbb{R} \to [0, 1]$, then, for any random variable X,*

$$\mathbb{E}(X) = \sum_{y} f_Y(y)\mathbb{E}[X|Y = y].$$

Similarly, if Y is continuous with probability density function $f_Y : \mathbb{R} \to \mathbb{R}_+$, then

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} f_Y(y)\mathbb{E}[X|Y = y]\, dy.$$

4.14 Independence (continued)

The notion of independence of two random variables is similar to the notion of independence of events introduced in §4.3. In particular, discrete random variables X and Y are said to be independent if, for all $(x, y) \in \mathbb{R} \times \mathbb{R}$, the events $\{X = x\}$ and $\{Y = y\}$ are independent. In such a case, it follows that the joint and marginal PMFs of X and Y are such that, for all $(x, y) \in \mathbb{R} \times \mathbb{R}$,

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

Analogously, continuous random variables X and Y are said to be independent if the above relationship holds when $f_{X,Y}$, f_X, and f_Y represent the joint and marginal PDFs of X and Y.

Expressions involving independent random variables often simplify nicely. If X and Y are independent, then for any $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ the random variables $g(X)$ and $h(Y)$ are independent and

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)).$$

This implies, for example, that if X and Y are independent, then

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

One can use this relationship to show that if X and Y are independent, then

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y), \tag{4.14.1}$$

a property that does not generally hold when X and Y are not independent, as seen in the next section.

4.15 Covariance and correlation

If random variables X and Y are independent, then the likelihood that X takes a particular value is the same regardless of the value that Y takes, and vice versa. In certain situations, however, the likelihood that a random variable X takes a value may depend on the value that another random variable Y takes. Such a relationship between X and Y can be quantified by their covariance, which is denoted and defined as

$$\operatorname{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))).$$

(From this definition, it is clear that the variance of a random variable is merely the covariance of the random variable with itself.) By manipulating this formula, one derives an equivalent definition of covariance:

$$\begin{aligned}\operatorname{cov}(X, Y) &= \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, from which the above implies that $\operatorname{cov}(X, Y) = 0$. More generally, if $\operatorname{cov}(X, Y) = 0$, then it is said that X and Y are uncorrelated. If X and Y are independent, then they are clearly uncorrelated, but uncorrelated random variables are not necessarily independent.

In fact, the term uncorrelated is derived from the notion of the correlation between random variables X and Y, which is defined, as long as X and Y both have nonzero variances, as

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \operatorname{var}(Y)}}.$$

Clearly, the correlation between X and Y is merely a scaling of the covariance between X and Y.

Generally speaking, X and Y have a positive covariance if $X - \mathbb{E}(X)$ and $Y - \mathbb{E}(Y)$ tend to have the same sign, whereas a negative covariance indicates that these quantities tend to have opposite signs. That is, positive covariance suggests that when X takes a value larger (or smaller) than its mean, then Y tends to also take a value larger (or smaller) than its mean, and this tendency is more pronounced the

more positive the covariance. Similarly, negative covariance suggests that when X takes a value larger (or smaller) than its mean, then Y tends to take a value smaller (or larger) than its mean, and this tendency is more pronounced the more negative the covariance. The scaling employed to produce the corresponding correlation between X and Y merely ensures that $\text{corr}(X,Y) \in [-1,1]$. If $\text{corr}(X,Y) = 1$ (or -1), then X and Y are said to be perfectly positively (or negatively) correlated, which occurs if and only if there exists a positive (or negative) constant c such that

$$Y - \mathbb{E}(Y) = c(X - \mathbb{E}(X)).$$

For example, a situation in which this occurs (with $c = 1$) is when $X = Y$; i.e., a random variable X always has $\text{corr}(X,X) = 1$, which is to say that it is always perfectly positively correlated with itself.

Using the definition of covariance, one can show that

$$\text{var}(X+Y) = \text{var}(X) + 2\,\text{cov}(X,Y) + \text{var}(Y),$$

from which one obtains (4.14.1) if and only if X and Y are uncorrelated.

4.16 Estimator basics

If the event that a random variable Y takes a certain value y provides information about another random variable X, then we have discussed that one may be interested in the expected value of X given $Y = y$, namely $\mathbb{E}(X|Y=y)$. This value, in turn, can be seen as a function of y; different values, say y_1 and y_2, lead to the different values $\mathbb{E}(X|Y=y_1)$ and $\mathbb{E}(X|Y=y_2)$. One can then define the quantity $\mathbb{E}(X|Y)$, which is itself a random variable—a function of the random variable Y. The properties of such a random variable $\mathbb{E}(X|Y)$ are worthy of note, particularly in the context of estimation and statistical inference. For one thing, viewing Y as an observation that provides information about X, one can view $\mathbb{E}(X|Y)$ as an estimator of X given Y. A first important property of this estimator is revealed in the following theorem.

Theorem 4.16.1 (Law of Iterated Expectations)**.** *If X and Y are random variables, then*

$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X).$$

The proof of this theorem follows by viewing $\mathbb{E}(X|Y)$ as a function of Y, whose expected value can be computed using the formulas from §4.13, then applying the Total Expectation Theorem 4.13.1.

One important property of the estimator $\mathbb{E}(X|Y)$ for X is that it is unbiased; i.e., the estimator does not have tendency to underestimate or overestimate $\mathbb{E}(X)$. Indeed, the estimation error $\tilde{X} = \mathbb{E}(X|Y) - X$ is a random variable satisfying $\mathbb{E}(\tilde{X}|Y) = \mathbb{E}((\mathbb{E}(X|Y) - X)|Y) = \mathbb{E}(X|Y) - \mathbb{E}(X|Y) = 0$, meaning that $\mathbb{E}(\tilde{X}|Y=y) = 0$ for any y. Consequently, by Theorem 4.16.1, one finds that

$$\mathbb{E}(\tilde{X}) = \mathbb{E}(\mathbb{E}(\tilde{X}|Y)) = 0.$$

Another important property of $\mathbb{E}(X|Y)$ is that the estimation error $\tilde{X}$ is uncorrelated with it. This is reassuring, since it ensures that the estimation error is expected to be consistent regardless of the particular realization of Y. To see that this is true, note that the expected values of $\mathbb{E}(X|Y)$ and $\tilde{X}$ are both zero, so their covariance is merely given by the expected value of their product, namely

$$\begin{aligned}\mathbb{E}(\mathbb{E}(X|Y)\tilde{X}) &= \mathbb{E}(\mathbb{E}(X|Y)\tilde{X}|Y) && \text{(by Theorem 4.16.1)}\\ &= \mathbb{E}(\mathbb{E}(X|Y)\mathbb{E}(\tilde{X}|Y)) = 0. && \text{(since } \mathbb{E}(X|Y) \text{ is determined by } Y)\end{aligned}$$

Due to the fact that these random variables are uncorrelated, it follows that the variance of $X = \mathbb{E}(X|Y) - \tilde{X}$ is simply the sum of the variances of $\mathbb{E}(X|Y)$ and $\tilde{X}$. This can be used to prove the following theorem.

Theorem 4.16.2 (Law of Total Variance)**.** *If X and Y are random variables, then, defining*

$$\text{var}(X|Y) = \mathbb{E}((X - \mathbb{E}(X|Y))^2|Y) = \mathbb{E}(\tilde{X}^2|Y),$$

it follows that

$$\text{var}(X) = \mathbb{E}(\text{var}(X|Y)) + \text{var}(\mathbb{E}(X|Y)).$$

That is, the variance of X can be expressed, using any random variable Y, as the variance of the estimator $\mathbb{E}(X|Y)$ plus the variance of the estimation error of $\mathbb{E}(X|Y)$ for X.

4.17 Markov and Chebyshev inequalities

The following theorem quantifies the intuitive fact that if a nonnegative random variable has a small expected value, then the probability that the random variable takes a large value is small.

Theorem 4.17.1 (Markov's inequality)**.** *If X is a nonnegative random variable with mean μ, then*

$$\mathbb{P}(X \geq x) \leq \frac{\mu}{x} \quad \textit{for any} \quad x \in \mathbb{R}_{++}.$$

A useful trick can be used to prove this theorem. In particular, defining a so-called indicator random variable X_x that takes on the value 0 if $X < x$ and the value x if $X \geq x$, one can easily show that $\mathbb{E}(X_x) = x\mathbb{P}(X \geq x)$ and $\mathbb{E}(X_x) \leq \mathbb{E}(X) = \mu$, which together give the result.

Markov's inequality can be used to prove the following, which quantifies the intuitive fact that if a random variable has a small variance, then the probability that it takes a value far from its mean is small.

Theorem 4.17.2 (Chebyshev's inequality)**.** *If X is a random variable with mean μ and variance σ^2, then*

$$\mathbb{P}(|X - \mu| \geq x) \leq \frac{\sigma^2}{x^2} \quad \textit{for any} \quad x \in \mathbb{R}_{++}.$$

In particular, by considering $x = z\sigma$ for some $z \in \mathbb{R}_{++}$, this theorem implies that

$$\mathbb{P}(|X - \mu| \geq z\sigma) \leq \frac{\sigma^2}{(z\sigma)^2} = \frac{1}{z^2},$$

which is to say that the probability that a random variable takes a value more than z standard deviations away from its mean is at most $1/z^2$.

4.18 Laws of large numbers

One of the most common questions in statistics is: How much data do I need? In order to provide an answer for such a question, probabilists have derived theorems that each provide an answer for a related question: What is the effect of increasing the size of a set of random variables? In this section, we present two fundamental theorems of this type, known as laws of large numbers.

Consider a sequence of random variables $\{X_1, \ldots, X_n\}$ that are all independent and identically distributed with mean μ and variance σ^2. The laws in which we are interested relate to the random variable

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \quad \text{for} \quad n \in \mathbb{N}_+,$$

which can be seen as an estimator for the mean μ. In particular,

$$\mathbb{E}(\overline{X}_n) = \frac{\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)}{n} = \frac{\mu n}{n} = \mu,$$

and from independence of the random variables in the sequence it follows that

$$\text{var}(\overline{X}_n) = \frac{\text{var}(X_1) + \cdots + \text{var}(X_n)}{n^2} = \frac{\sigma^2 n}{n^2} = \frac{\sigma^2}{n}.$$

Using Chebyshev's inequality, Theorem 4.17.2, it follows that

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \quad \text{for any} \quad \epsilon \in \mathbb{R}_{++},$$

along with which we obtain the proof of the following theorem.

Theorem 4.18.1 (Weak law of large numbers)**.** *If $\{X_1, X_2, \ldots\}$ is a sequence of random variables that are independent and identically distributed with mean μ and variance σ^2, then for every $\epsilon \in \mathbb{R}_{++}$ it follows that*

$$\lim_{n \to \infty} \mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \to 0.$$

This theorem confirms the intuitive idea that, for large n, the distribution of the sample mean $\overline{X}_n$ is concentrated around the true mean μ. That is, for any interval of the form $[\mu - \epsilon, \mu + \epsilon]$, there is a high probability that the sample mean will fall into this interval as n tends to ∞. One may have to accumulate more random variables into the sample mean in order to achieve this when ϵ is relatively small, but eventually the probability of $\overline{X}_n$ falling into such an interval is high for any positive value of ϵ.

The following theorem also states a property of the sample mean as $n \to \infty$.

Theorem 4.18.2 (Strong law of large numbers)**.** *If $\{X_1, X_2, \dots\}$ is a sequence of random variables that are independent and identically distributed with mean μ, then*

$$\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

Intuitively, this theorem states that even though $\overline{X}_n$ is a random variable with variance σ^2/n (which is strictly positive whenever $\sigma^2 > 0$), the limiting random variable obtained when $n \to \infty$ has all of its distribution concentrated at a single point, namely μ. That is, the probability of having an infinite sequence of such random variables yielding $\lim_{n\to\infty} \overline{X}_n \neq \mu$ is zero.

The difference between the weak and strong laws of large numbers is subtle. The former result is referred to as "weak" as it merely ensures that the probability of finding $\overline{X}_n$ far from its mean μ becomes less likely as $n \to \infty$. However, observing the sequence $\{\overline{X}_1, \overline{X}_2, \dots\}$, the law does not provide concrete information about how often $\overline{X}_n$ will deviate significantly from μ. On the other hand, the "strong" law, as its name suggests, guarantees a stronger property: For any $\epsilon > 0$, the probability that the difference $|\overline{X}_n - \mu|$ will exceed ϵ an infinite number of times is zero. The two laws are in fact representative of two different types of convergence that one can define for sequences of random variables, as defined in the next section.

4.19 Convergence of sequences of random variables

As random variables are not numbers, one cannot discuss convergence of random variables using the same terminology and definitions as in §2.32. However, notions of convergence of random variables do exist. In this section, we provide definitions of two types of convergence of random variables, each of which can be seen as a generic statement similar to those seen in the weak and strong laws of large numbers.

If $\{Y_n\}$ is a sequence of random variables and y is a scalar such that, for any $\epsilon \in \mathbb{R}_{++}$,

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - y| \geq \epsilon) = 0,$$

then $\{Y_n\}$ is said to converge in probability to y. According to this definition, the weak law of large numbers states that the sample mean of a sequence of independent and identically distributed random variables converges in probability to the mean of the random variables. This definition can be equivalently stated in a manner that makes it reminiscent of the definition of convergence for a sequence of real numbers: If for every $\epsilon \in \mathbb{R}_{++}$ and $\delta \in \mathbb{R}_{++}$ there exists some $k \in \mathbb{N}$ such that

$$\mathbb{P}(|Y_n - y| \geq \epsilon) \leq \delta \quad \text{for all} \quad k \geq n,$$

then $\{Y_n\}$ converges in probability to y.

A stronger type of convergence occurs when the sequence $\{Y_n\}$ and scalar y satisfy

$$\mathbb{P}\left(\lim_{n\to\infty} Y_n = y\right) = 1;$$

if such a property holds, then $\{Y_n\}$ is said to converge with probability 1 (or almost surely) to y. The strong law of large numbers states that the sample mean of a sequence of independent and identically distributed random variables converges almost surely to the mean of the random variables.

If a sequence of random variables converges almost surely, then it can be shown that it converges in probability. However, the converse is not always true.

4.20 Central Limit Theorem

In the previous sections, we have seen various important properties of the sample mean of a sequence of random variables as the size of the sequence increases indefinitely. In this section, we investigate the properties of a related sequence defined by a series of random variables.

If $\{X_1, \dots, X_n\}$ is a sequence of independent and identically distributed random variables, then the sum $X_1 + \cdots + X_n = n\overline{X}_n$ has a variance that increases indefinitely as $n \to \infty$. However, with appropriate centering and normalization, we obtain a sequence with interesting asymptotic properties. Let

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \quad \text{for all} \quad n \in \mathbb{N}_+.$$

For any $n \in \mathbb{N}_+$, the expected value and variance of Z_n are easily found to be

$$\mathbb{E}(Z_n) = \frac{\mathbb{E}(X_1 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}} = 0$$

and, by independence of the random variables,

$$\text{var}(Z_n) = \frac{\text{var}(X_1 + \cdots + X_n)}{\sigma^2 n} = \frac{\sigma^2 n}{\sigma^2 n} = 1.$$

The following theorem shows that for any such sequence of random variables that follow any distribution, the random variable Z_n tends to behave as a standard normal random variable as $n \to \infty$.

Theorem 4.20.1 (Central Limit Theorem)**.** *If $\{X_1, X_2, \dots\}$ is a sequence of random variables that are independent and identically distributed with mean μ and variance σ^2, then the cumulative distribution function of Z_n converges to that of a standard normal random variable, namely*

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx,$$

in the sense that

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \Phi(z) \quad \textit{for all} \quad z \in \mathbb{R}.$$

Chapter 5

Computational Mathematics

In modern engineering and applied mathematics, computing plays an essential role. There exist computer programs that can perform symbolic computation as one does with pencil and paper, but these types of computations can be expensive in terms of both computing time and data storage. As this limits the size and complexity of problems that can be solved using symbolic computation, most computations performed in modern computers are numerical computations involving floating point numbers that only approximate exact computations with real numbers. This makes it important to understand how numerical computations are performed, how these computations can lead to calculation errors, and how these errors can be avoided or at least minimized. The fields known as computational mathematics, scientific computing, and numerical analysis focus on the study of algorithms involving numerical computations with the goal of designing, analyzing, and implementing efficient software for solving large and complex mathematical problems.

5.1 Floating point numbers

Most modern computers employ the IEEE 754 standard for binary floating point numbers. The standard involves a few formats with similar structure that differ only in a few details. For example, one of the most widely known and used formats is that for double-precision floating point numbers wherein a number is represented by 64 binary digits. In such a representation, each digit falls into one of three segments:[1]

- Digit 1 is the sign bit (also known as the sign indicator), call it s. A value $s = 0$ indicates that the number is positive, whereas a value $s = 1$ indicates that the number is negative.
- Digits 2–12 compose the exponent (also known as the characteristic or scale), call its decimal value e. These 11 binary digits correspond to a decimal integer value for e in the interval $[0, 2047]$. The number being represented is considered a product involving 2^{e-1023}—where the fixed number 1023 is called the exponent bias—meaning that this term can range from 2^{-1023} to 2^{1024}. This makes it possible to represent relatively small and relatively large numbers.
- Digits 13–64 compose the significand (also known as the mantissa or coefficient), call its decimal value f. These 52 binary digits represent a binary fraction with value less than 1. The number being represented is considered a product involving $(1 + f)$.

Overall, a representation with sign s, exponent e, and significand f corresponds to the decimal number

$$(-1)^s 2^{e-1023}(1 + f). \tag{5.1.1}$$

For example, consider the following string of 64 digits:

$$\underbrace{0}_{s}\ \underbrace{10000000011}_{e}\ \underbrace{1011100100010000000000000000000000000000000000000000000000}_{f}$$

[1] The situation is in fact slightly more complicated than is presented here as recent versions of the standard involve various sophistications, such as the reservation of certain strings of digits for special quantities (such as infinity) and a distinction between so-called normal and subnormal floating point numbers. However, for simplicity in these notes, we ignore such sophistications and instead discuss a simplified version of the standard.

Converting from binary to decimal, we obtain the values

$$\begin{aligned} s &= 0, \\ e &= 2^{10} + 2^1 + 2^0, \\ \text{and} \quad f &= \left(\tfrac{1}{2}\right)^1 + \left(\tfrac{1}{2}\right)^3 + \left(\tfrac{1}{2}\right)^4 + \left(\tfrac{1}{2}\right)^5 + \left(\tfrac{1}{2}\right)^8 + \left(\tfrac{1}{2}\right)^{12}, \end{aligned}$$

which, plugging into (5.1.1), means that the digits correspond to the decimal number 27.56640625.

Other formats in the IEEE 754 standard are similar, but differ in a few ways such as the number of binary digits used to represent a given number. There are also other schemes that have been proposed and used, such as that of fixed point numbers. However, since the IEEE 754 standard is so widely used, we do not bother considering alternatives in these notes. Moreover, to understand issues related to other formats in the IEEE 754 standard, it is sufficient to consider the double precision format described above.

5.2 Overflow, underflow, and NaN

The fact that only a finite number of digits are used to represent a floating point number automatically implies that not all real numbers can be represented exactly as floating point numbers. Indeed, not even all rational numbers can be represented! For one thing, there is a limit to the magnitude of any number that can be represented in the form (5.1.1); it is essentially equal to the number that results when all digits in the exponent and significand are equal to 1. (It may not be exactly this value as this precise string of digits may be reserved as a representation for infinity.) In general, the largest number that can be represented in any floating point format represents an upper bound on the real number that can be provided from any calculation. If a calculation (say, a sum) would otherwise lead to result larger than this largest floating point number, then overflow is said to occur. When a computer is requested to make such a calculation, the typical response is for it to say that an error has occurred or for it to return a reserved quantity representing infinity. This can lead to catastrophic failure of computational software if this value is allowed to propogate through subsequent calculations. One manner in which this can be avoided is to perform a sequence of computations in a (valid) order such that the likelihood of overflow is minimized. For example, rather than computing an average of two large numbers by attempting to add them (possibly leading to overflow) and then dividing by two, one can first divide each number by two and then add the results.

There is also a limit to the absolute value of the smallest nonzero number that can be represented in any floating point format. In terms of (5.1.1), the smallest positive number is essentially equal to the number that results when all digits in the exponent and significand are equal to 0. (It may not be exactly this value due to the use of so-called subnormal floating point numbers.) If a calculation would otherwise lead to a result between this number and zero, then underflow is said to occur. When a computer is requested to make such a calculation, the typical response is for it to say that an error has occurred or—if it is determined that no significant loss of accuracy will occur—for it to return a number, possibly even zero! As in the case of overflow, this can lead to catastrophic failure if this value is allowed to propogate.

Modern computers also make special consideration of calculations that are not considered well-defined. For certain such computations, the IEEE 754 standard has a reserved quanity known as NaN ("not a number"). For example, this is the result that is returned when one attempts to divide a number by 0, multiply or add infinity, or perform any calculation with a quantity that is equal to NaN. It is also typically given as the result when one attempts to perform a calculation whose result would otherwise be a complex number, such as taking a square root of a negative number. (Complex numbers are typically handled specially as a pair of floating point numbers corresponding to the complex number's real and imaginary parts.) In some situations a result of NaN causes the computer to say that an error has occurred, or the result can propogate through subsequent calculations leading to potentially serious consequences.

5.3 Round-off error

Numerical errors do not only occur at the extremes of a particular floating point number format. There also exists a gap between any two consecutive numbers that can be represented. When a computer attempts to perform a calculation involving a real number that lies in one of these gaps (such as any calculation

that should involve the irrational number π), or when it is requested to perform a calculation that in exact arithmetic would lead to a number that cannot be represented exactly as a floating point number, then the result can only be an approximation of the true result. There are two prevailing methods for handling such situations—chopping and rounding—either of which is said to result in round-off error. For example, consider a model floating point system wherein a number is represented by one binary digit, call it b, and five decimal digits, call them d_i for $i \in \{1, \ldots, 5\}$, such that a decimal number being represented is

$$(-1)^b 10^{d_1 - 4}(1.d_2 d_3 d_4 d_5).$$

One can verify that the decimal number 1.00005 cannot be represented exactly in this floating point system. If a calculation that should lead to this result is performed and the computer chops off the last digit to store 040000 (corresponding to 1.0000), then the computer has performed chopping. On the other hand, if the computer rounds the last digit to store 040001 (corresponding to 1.0001), then it has performed rounding.

5.4 Absolute error, relative error, and significant digits

One of the goals in the field of numerical analysis is to estimate the round-off errors that may occur in any numerical computation. There are two main types of errors considered in such analyses: absolute and relative. If y is a real number and $\text{fl}(y)$ is its representation in a floating point format, then the absolute error is defined as $|y - \text{fl}(y)|$ and the relative error (provided that $|y| > 0$) is $|y - \text{fl}(y)|/|y|$. For various reasons, relative error is more often considered the measure of interest as it weighs the error versus the magnitude of the value being considered. However, since relative error is undefined when $y = 0$ and can be sensitive to measurement units, absolute error is also often considered. This sensitivity to measurement units can be seen through various examples; e.g., the relative error between 1° and 2° in Celsius is very large, but the relative error between these temperatures on the Kelvin scale is very small.

Rather than convey the inaccuracy of a numerical computation in terms of absolute or relative error, it is often convenient to convey its accuracy in terms of a property of the computed value itself. The most common such property is the number of significant digits that the value possesses. A floating point representation $\text{fl}(y)$ of a real number y has k significant digits if k is the largest nonnegative integer with

$$5 \times 10^{-k} > \frac{|y - \text{fl}(y)|}{|y|} \quad \text{(i.e., the relative error)}.$$

5.5 Machine epsilon

Users of numerical software typically do not have the time or interest to perform an extensive analysis of the errors in their computations! However, in order to manage expectations about the accuracy to which one can expect a computation to be performed, a value about which one should be aware is known as machine epsilon. This value, which differs from one hardware-software combination to the next, has various related definitions, but generally speaking it represents an upper bound on the relative error of any computation due to round-off error. A more precise definition is that machine epsilon is the difference between 1 and the smallest number larger than 1 that can be represented in a particular floating point format. For example, for a standard double precision floating point format (recall (5.1.1)), machine epsilon is approximately $2^{-52} \approx 2 \times 10^{-16}$. When employing a computer that uses this format, one should not expect the result from any numerical software to have a relative error smaller than this value, or, in many situations, the square root of this value (i.e., approximately 10^{-8}). Many software languages provide a simple command with which a user can determine the value of machine epsilon for their computer when using that language.

5.6 Error bound theorems for linear systems

Numerical analysts have derived various theorems for estimating numerical errors at a more macroscopic level than by considering the round-off error of every arithmetic operation. In this section, we provide a few example theorems of this type in the widely applicable context of solving linear systems of equations.

Suppose that one aims to solve $Ax = b$ for $x \in \mathbb{R}^n$ where $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. If numerical software provides an approximate solution $\tilde{x} \in \mathbb{R}^n$ for this system, then one may want to have a general impression about the magnitude of $\|x - \tilde{x}\|$ for some norm $\|\cdot\|$. As it has been requested to attempt to solve $Ax = b$, the software has performed its job well if the resulting residual from the calculation, namely $r := \|A\tilde{x} - b\|$, is small (at least relative to the magnitudes of $\|A\|$ and/or $\|b\|$). Unfortunately, however, a small residual does not necessarily guarantee a small $\|x - \tilde{x}\|$. This can be seen in the following theorem.

Theorem 5.6.1. *If $b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ is nonsingular, then $x = A^{-1}b$ and $r = A\tilde{x} - b$ satisfy*

$$\|x - \tilde{x}\| \leq \|A^{-1}\|\|r\|.$$

If, in addition, $\|x\| \neq 0$ and $\|b\| \neq 0$, then

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|r\|}{\|b\|} = \kappa(A)\frac{\|r\|}{\|b\|}.$$

This theorem highlights the importance of the condition number for a matrix in terms of the accuracy that one can expect from numerical computations involving the matrix.

The previous theorem considers a simplified setting in which one is able to store the matrix A and right-hand side vector b exactly. In general, however, only the floating point representations of these quantities can be stored, leading to other sources of error when attempting to solve $Ax = b$. The following theorem provides a glimpse of the effect of these errors; in particular, it provides an upper bound on the relative error between x and $\tilde{x}$ when the former is an exact solution of $Ax = b$ and the latter is an exact solution of $\text{fl}(A)\tilde{x} = \text{fl}(b)$ when $\text{fl}(A) = A + \delta_A$ and $\text{fl}(b) = b + \delta_b$.

Theorem 5.6.2. *If $(b, \delta_b) \in \mathbb{R}^n \times \mathbb{R}^N$ and $(A, \delta_A) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n}$ are such that A is nonsingular and $\|A^{-1}\|\|\delta_A\| < 1$, then x and $\tilde{x}$ satisfying $Ax = b$ and $(A + \delta_A)\tilde{x} = b + \delta_b$, respectively, yield*

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \|A^{-1}\|\|\delta_A\|}\left(\frac{\|\delta_b\|}{\|b\|} + \frac{\|\delta_A\|}{\|A\|}\right).$$

Of course, the situation in reality is plagued by a combination of the above issues. In particular, one obtains a true upper bound on the relative error between the true and an approximate solution of the linear system $Ax = b$ by accounting for the errors in representing A and b as well as the error when attempting to solve the linear system involving the floating point representations of these quantities. Formulas for such bounds can be derived, and while the precise form of these bounds may not be particularly important, one should appreciate the important factors involved: the norms of A and b, the round-off error from storing A and b, the condition number of A, and the residual in the approximate solution of the system.

5.7 Rules of thumb

When implementing numerical software, one should often follow a few key rules of thumb. First, even if one would make a decision in exact arithmetic based on whether or not two numbers are equal, this should be avoided in numerical software. For example, while 2 and $(\sqrt{2})^2$ are equal in exact arithmetic, they would be determined to be unequal by a computer due to the round-off error that would result from computing $\sqrt{2}$. Instead, it may be appropriate to make the decision based on the relative difference between the two values being small with respect to some prescribed tolerance (perhaps related to machine epsilon). One should also avoid subtracting nearly identical numbers and the explicit computation of an inverse of a matrix as these are classic examples of computations that are known to result in large round-off errors. (Rather than explicitly computing the inverse of a matrix, sophisticated numerical software would perform some sort of Gaussian elimination or other factorization technique in which the computations would be ordered intelligently to reduce round-off errors.) Overall, unless one is experienced at implementing numerical software, it is recommended to rely on previously implemented software for performing low-level computations, such as the software library known as BLAS (Basic Linear Algebra Subprograms).