FRTN50: Optimization for learning
Mathematical prerequisites

Table of contents

Introduction 1

Sets 2

Logic 4

Functions 7

Sets revisited 8

Basic facts about $\mathbb{R}$ 10

Inner-product spaces 13

Real-valued vectors and matrices 17

Some analysis 26

Probability theory 33

Methods of proof 34

References 34

Introduction

The course *optimization for learning* uses quite a lot of mathematics from basic analysis and linear algebra. We have noticed that some students struggle with the material in this course since they have forgotten much from earlier courses in mathematics.

This short document covers some elementary facts from set theory, logic, analysis, linear algebra, probability theory, and methods of proof that we expect students taking the course to know. If you are unsure of anything here, consider reading up on these subjects. Parts written in small font size can be skipped in a first reading but are included for completeness and for the more interested students.

The reference section at the end contains some resources that inspired parts of the contents in this document.

Written by: Manu Upadhyaya, `manu.upadhyaya@control.lth.se`

Latest update: August 29, 2023

Sets

Basic notation. Let A and B be *sets*. We say that a is an *object* or *element* of A if a is a member of A and write this as

$$a \in A.$$

If a is not a member of A we write

$$a \notin A.$$

The set containing no elements is called the *empty set* and is denoted as $\emptyset$. The set A is said to be *nonempty* if $A \neq \emptyset$.

A set that only contains a single element, e.g., the set $\{a\}$ that only contains the element a, is called called a *singleton*. A singleton is an example of a nonempty set.

We say that A is a *subset* of B if every element of A is also a member of B and indicate this by writing

$$A \subseteq B.$$

In particular, for any set A, the empty set $\emptyset$ and the set itself are always subsets of A. I.e.,

$$\emptyset \subseteq A \quad \text{and} \quad A \subseteq A,$$

always hold. If $A \subseteq B$, we also write $B \supseteq A$, which is read "B *contains* A".

Suppose we are given some *property* that elements of A may or may not possess. How do we form the set consisting of all elements of A having that property? Consider for example the set of positive integers, i.e., the *natural numbers*, which is written as

$$\mathbb{N} = \{1, 2, 3, \ldots\},$$

and wish to select the even positive integers. We write the set of even positive integers as

$$\{x \in \mathbb{N} \mid x \text{ is even}\}.$$

The symbol "|" is read as "such that". Sometimes the colon ":" is used in place of "|".

Union and "or". The set consisting of all elements of A together with all the elements of B is called the *union* of A and B, and is denoted as $A \cup B$. Formally, we define

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}.$$

The word "or" used in this context is always given the unambiguous interpretation as an *inclusive or*, i.e., the statement "P or Q" means "P or Q, or both". In particular, the set $A \cup B$ consists of all the elements that belong to A or to B or to both. Clearly, the union operation is *commutative*, i.e., $A \cup B = B \cup A$.

Intersection. The set that consists of all elements common to A and B is called the *intersection* of A and B, and is denoted as $A \cap B$. Formally, we define

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

The intersection operation is also commutative, i.e., $A \cap B = B \cap A$.

Disjoint sets. Using the intersection operation, we may express the statement that A and B have no elements is common by the equation

$$A \cap B = \emptyset.$$

In such a case, we say that A and B are *disjoint*.

Set difference. The *(set) difference* of two sets, denoted by $A \setminus B$, is defined as the set consisting of those elements of A that not in B. Formally, we define

$$A \setminus B = \{x \mid x \in A \text{ and } x \notin B\}.$$

Note that the set difference operation does not commute.

Complement. Suppose that there is some base set X such that $A \subseteq X$. Then *complement* of A in X is defined as the set difference of X and A, i.e., $X \setminus A$.

Cartesian product. We define the *Cartesian product* of the sets A and B, denoted by $A \times B$, to be the set of all *ordered pairs* or 2-*tuples* (a,b) for which a is an element of A and b is an element of B. Formally, we define

$$A \times B = \{(a,b) \mid a \in A \text{ and } b \in B\}.$$

Note that this definition assumes that the concept of an ordered pair is already given, and we simply take it as a primitive concept. Similarly, recall that the concept of a set was assumed as given. In general, $A \times B$ is not equal to $B \times A$. Unfortunately, the notation (a,b) has another well-established meaning, i.e., denoting an open interval of the real numbers $\mathbb{R}$. However, this conflict in notation will cause no difficulty and the meaning will be clear from context.

Collection of sets. We will sometimes refer to sets whose elements are sets as *collections*. For instance, the collection whose elements are the subsets of A is called the *power set* of A and is denoted by $\mathcal{P}(A)$. Formally, we define

$$\mathcal{P}(A) = \left\{ \tilde{A} \;\middle|\; \tilde{A} \subseteq A \right\}.$$

Arbitrary unions and intersections. Given a collection $\mathcal{A}$ of sets, the *union* of the elements of $\mathcal{A}$ is defined by the equation

$$\bigcup \mathcal{A} = \bigcup_{A \in \mathcal{A}} A = \{x \mid x \in A \text{ for at least one } A \in \mathcal{A}\}.$$

The *intersection* of the elements of $\mathcal{A}$ is defined by the equation

$$\bigcap \mathcal{A} = \bigcap_{A \in \mathcal{A}} A = \{x \mid x \in A \text{ for every } A \in \mathcal{A}\}.$$

The for the particular case $\mathcal{A} = \emptyset$, it is clear that

$$\bigcup_{A \in \mathcal{A}} A = \emptyset.$$

However, we leave $\bigcap_{A \in \mathcal{A}} A$ undefined in this case for technical reasons.

Logic

Statements. Defining precisely what is meant by a mathematical statement is surprisingly difficult. Therefore, we will only provide a practical working definition. A *statement* is a sentence that is either true or false, but not both.

Note that not all sentences are statements. E.g. consider the sentence

"This statement is false".

Why is this not a statement?

Negation. The *negation* of statement P is the statement that is false when P is true, and true when P is false. The negation of P is denoted as

$$\text{not } P.$$

The negation not P is sometimes written as $\neg P$. We may indicate the effect of the negation operation with the following truth table:

P	not P
True	False
False	True

We give here a simple example of a negation: Consider the statement $a \in A$. The negation is simply $a \notin A$.

"If ..., then ..." statements. Many statements have the following structure:

$$\text{If } P\text{, then } Q \tag{1}$$

where

- P is a statement called the *hypothesis* or *assumption*, and
- Q is a statement called the *conclusion*.

A statement of the form (1) is called an *implication*. We say P implies Q and sometimes write $P \Rightarrow Q$. What is the logical value of the statement "if P, then Q"? This will depend on the logical values of P and Q. We indicate it by the following truth table:

P	Q	If P, then Q
True	True	True
True	False	False
False	True	True
False	False	True

Vacuous truth. Pay special attention to the last two rows in the truth table above. E.g. consider the following implication about real numbers x:

$$\text{If } x^2 < 0\text{, then } x = 1.$$

Note that the hypothesis can never hold, but the statement is true, independent of the conclusion. Such implications are said to be *vacuously true*.

Why do we care about vacuously true implications? Vacuously true implications are generally not of interest by themselves, but they frequently arise in proofs as special cases.

Let us see another example of a vacuous truth. We claimed previously that $\emptyset \subseteq A$ for any set A. But how do we formally prove this? Writing $\emptyset \subseteq A$ is a shorthand way of writing the statement, "For every object x, if $x \in \emptyset$, then $x \in A$". However, it is obvious that the statement $x \in \emptyset$ is not satisfied for any object x. Thus, $\emptyset \subseteq A$ is vacuously true.

Converse. Consider the implication

$$P \Rightarrow Q.$$

The *converse* of the above implication is the implication

$$Q \Rightarrow P.$$

Equivalence. If $P \Rightarrow Q$ and its converse $Q \Rightarrow P$ are both true, we say that P and Q are *(logically) equivalent* statements and write this statement as $P \Leftrightarrow Q$. The sign $\Leftrightarrow$ is read as "if and only if" (abbreviated iff sometimes) and the statement is called a *biconditional statement*.

If the equivalence $P \Leftrightarrow Q$ holds, we also say that P is a *necessary* and *sufficient* condition for Q. In particular:

- A necessary condition is one which must hold for a conclusion to be true. It does not guarantee that the result is true. Here, the $Q \Rightarrow P$ part is responsible for the necessity.
- A sufficient condition is one which guarantees the conclusion is true. The conclusion may be true even if the condition is not satisfied. Here, the $P \Rightarrow Q$ part is responsible for the sufficiency.

Many theorems and exercises in the course will require that you prove that some particular equivalence $P \Leftrightarrow Q$ holds. i.e., you will need to establish the validity of $P \Rightarrow Q$ and $Q \Rightarrow P$.

Contrapositive. Consider the implication

$$P \Rightarrow Q.$$

The *contrapositive* of the above implication is the implication

$$(\text{not } Q) \Rightarrow (\text{not } P).$$

One can show that an implication and its contrapositive are logically equivalent (prove it). Why is this useful? In certain cases, it turns out easier to prove the contrapositive than the original implication.

Quantifiers. Many statements include the phrases "for all" and "there exists". Such phrases are called *logical quantifiers*.

- The phrases "for all", denoted by $\forall$, is known as the *universal quantifier*. Other common equivalent ways of writing the phrase "for all" is "for every" and "for each'.
- The phrases "there exists", denoted by $\exists$, is known as the *existential quantifier*. Another common equivalent way of writing the phrase "there exists" is "for at least one".

Let us see some examples of their usage. Let $\mathbb{Z}$ denote the integers, i.e.,

$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}.$$

- The statement "For all x in $\mathbb{Z}$, x^2 is greater or equal to 0" can be written symbolically as $\forall x \in \mathbb{Z},\ x^2 \geq 0$.
- The statement "There exists an x in $\mathbb{Z}$ such that x^2 is equal to 4" can be written symbolically as $\exists x \in \mathbb{Z},\ x^2 = 4$.
- For consecutive quantifiers of the same type we introduce a symbolic shorthand. E.g. for the sentence "For all x in $\mathbb{Z}$ and for all y in $\mathbb{Z}$, the sum $x+y$ is in $\mathbb{Z}$", instead of writing $\forall x \in \mathbb{Z}, \forall y \in \mathbb{Z}, x+y \in \mathbb{Z}$, we write $\forall x, y \in \mathbb{Z}, x+y \in \mathbb{Z}$. The corresponding rule for the existential quantifier and extensions to more than two quantifiers is straightforward.
- It is possible to combine different quantifiers. E.g. consider the statement "For each x in $\mathbb{Z}$ there exists a y in $\mathbb{Z}$ such that $x+y$ is equal to 0". This can be written symbolically as $\forall x \in \mathbb{Z}, \exists y \in \mathbb{Z}, x+y = 0$. Note that the order of the quantifiers matters in general.

Negation of quantifiers. Suppose that A is some set and that it is possible for each x in A to construct a statement $P(x)$, usually a condition on x. It is not hard to show that (prove it)

- not $(\forall x \in A, P(x)) \Leftrightarrow \exists x \in A, (\text{not } P(x))$, and
- not $(\exists x \in A, P(x)) \Leftrightarrow \forall x \in A, (\text{not } P(x))$.

For convenience, we state these facts using sentences:

- The negation of the statement

  "For all $x \in A$, the statement $P(x)$ holds"

  is

  "There exists an $x \in A$ such that the statement $P(x)$ does not hold".

- The negation of the statement

  "There exists an $x \in A$ such that the statement $P(x)$ holds"

  is

  "For all $x \in A$, the statement $P(x)$ does not hold".

Why are these facts important, especially the first fact? In many problems, you will be asked to determine if a statement $\forall x \in A, P(x)$ is true or false. Suppose that the statement is in fact false. It is then sometimes convenient or easier to show that the negation is true instead. I.e., to show that there exists an $x \in A$ such that $P(x)$ does not hold. In particular, this x is then called a *counterexample*, i.e., an example that shows that the statement is indeed false.

Next, we give the general rule to negate statements with multiple quantifiers. To negate a statement of the form

$$Q_1 x_1 \in A_1, \ldots, Q_n x_n \in A_n, P(x_1, \ldots, x_n)$$

where Q_i is $\forall$ or $\exists$, and A_i is a set, for each $i \in \{1, \ldots, n\}$, and $P(x_1, \ldots, x_n)$ is a statement, do the following:

(i) Change every $\forall$ to $\exists$ and every $\exists$ to $\forall$.

(ii) Replace $P(x_1,\ldots,x_n)$ by its negation.

Functions

Definition. Given two sets X and Y, a *function* f from X to Y is a rule of assignment between these two sets, denoted by $f : X \to Y$. In particular, for each element x of the set X, the function assigns a unique element, denoted by $f(x)$, of the set Y. Here, x is called the *argument* or *input* of the function, and $f(x)$ is the *value* of the function at x.

Note that a function $f : X \to Y$ is only well-defined if the sets X and Y are explicitly specified. The set X is called the *domain* of the function[1] and Y is called the *codomain* of the function.

Sometimes, we will define a function $f : X \to Y$ by using an arrow $\mapsto$ to show the action of f on an element of X. E.g., the function "$f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = x + 2$" can be written as "$x \mapsto x + 2$ from $\mathbb{R}$ to $\mathbb{R}$".

Example: The simplest function is probably the *identity function*; given a set X, the identity function on X, denoted Id, is defined as the mapping $x \mapsto x$ from X to X. Often, the identity function Id is used without specifying the set X, but it is usually easily inferred from the context.

Image. Let $f : X \to Y$ be a function and let $A \subseteq X$. The *image* of A under f, written $f(A)$, is defined by

$$f(A) = \{y \in Y \mid y = f(x) \text{ for at least one } x \text{ in } A\}.$$

Pre-image. Let $f : X \to Y$ be a function and let $B \subseteq Y$. The *pre-image* or *inverse image* of B under f, written $f^{-1}(B)$, is defined by

$$f^{-1}(B) = \{x \in X \mid f(x) \in B\}.$$

Composition. Let $g : X \to Y$ and $f : Z \to E$ such that $Y \subseteq Z$. The function $f \circ g : X \to E$, called the *composition* of f and g, and is defined by

$$(f \circ g)(x) = f(g(x))$$

for each $x \in X$.

Restriction. Let $f : X \to Y$ be a function and let $A \subseteq X$. The *restriction* of f to A, written $f|_A$, is the function $f|_A : A \to Y$ defined by

$$f|_A(x) = f(x)$$

for each $x \in A$.

Injective, surjective and bijective. Let $f : X \to Y$ be a function. We have the following definitions:

- The function f is said to be *injective* or *one-to-one* if

  $$x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$$

  for each $x_1, x_2 \in X$, or equivalently,

  $$f(x_1) = f(x_2) \Rightarrow x_1 = x_2$$

  for each $x_1, x_2 \in X$. An injective function is sometimes called an *injection*.

[1] Do not confuse the domain of a function with the effective domain defined later in the course.

- The function f is said to be *surjective* or *onto* if

$$\forall y \in Y, \exists x \in X, f(x) = y$$

or equivalently,

$$y \in Y \Rightarrow \exists x \in X, f(x) = y.$$

A surjective function is sometimes called a *surjection*.

- If the function f is injective and surjective, we say that f is *bijective*. A bijective function is sometimes called a *bijection*.

Binary operators. A *binary operator* on a set X is a function with domain $X \times X$ and codomain X.

Binary operators are usually written with infix notation, i.e., the placement of the operator is between the operands. E.g. the standard addition operator $+$ on $\mathbb{R}$ is a binary operator on $\mathbb{R}$.

Sets revisited

Finite sets. A set A is said to be *finite* if it is empty or if there exists a bijective function

$$f : A \to \{1, \ldots, n\}$$

for some $n \in \mathbb{N}$. In the former case, we say that A has *cardinality* 0; in the latter case, we say that A has *cardinality* n. We sometimes denote the cardinality of A by the symbol $|A|$.

We have the following fact: The cardinality of a finite set A is uniquely determined by A.

Infinite sets. A set A is said to be *infinite* if it is not finite. If there exists a bijective function

$$f : A \to \mathbb{N}$$

we say that A is *countably infinite*.

Countable and uncountable sets. A set is said to be *countable* if it is either finite or countably infinite. A set that is not countable is said to be *uncountable* or *uncountably infinite*.

Indexed family of sets. Let $\mathcal{A}$ be a nonempty collection of sets. An *indexing function* for $\mathcal{A}$ is a surjective function

$$f : I \to \mathcal{A}$$

for some nonempty set I called the *index set*. The collection $\mathcal{A}$, together with the indexing function f, is called an *indexed family of sets*.

Given $i \in I$, the set $f(i)$ is typically denoted by some symbol A_i. Then, the indexed family of sets is typically denoted by the symbol

$$\{A_i\}_{i \in I}$$

dropping the reference to $\mathcal{A}$ and f altogether.

How do collections of sets and indexed families of sets differ? Roughly speaking, collections of sets do not allow duplicate members, while indexed families of sets do.

Unions and intersections again. Let $\{A_i\}_{i \in I}$ be an indexed family of sets. We define

$$\bigcup_{i \in I} A_i = \{a \mid \exists i \in I, a \in A_i\}$$

and

$$\bigcap_{i \in I} A_i = \{a \mid \forall i \in I, a \in A_i\}$$

as the union and intersections of the members of this family, respectively.

There are two especially useful cases for the index set I. In the case of $I = \{1, \ldots, n\}$ for some positive integer n, we write the indexed family of sets as $\{A_i\}_{i=1}^n$. Sometimes, we simply write the indexed family of sets as $A_1, \ldots, A_n$ and call it a *(finite) sequence* of sets. We write

$$\bigcup_{i=1}^{n} A_i \quad \text{and} \quad \bigcap_{i=1}^{n} A_i$$

as the union and intersections of the members of this family, respectively.

Similarly, in the case of $I = \mathbb{N}$, we write the indexed family of sets as $\{A_i\}_{i=1}^\infty$. Sometimes, we simply write the indexed family of sets as $A_1, A_2, \ldots$ and call it an *(infinite) sequence* of sets. We write

$$\bigcup_{i=1}^{\infty} A_i \quad \text{and} \quad \bigcap_{i=1}^{\infty} A_i$$

as the union and intersections of the members of this family, respectively.

Cartesian product. We wish to generalize the concept of a Cartesian product beyond two sets.

First, we start with the finite case. Let $n \in \mathbb{N}$. Given a set X, we define an *n-tuple* of elements of X to be a function

$$x : \{1, \ldots, n\} \to X.$$

Given an n-tuple x, for each $i \in \{1, \ldots, n\}$,[2] we often denote the value of x at i by x_i rather than $x(i)$, and call it the ith coordinate of x. We often denote the function x itself by the symbol

$$(x_1, \ldots, x_n).$$

Next, let $\{A_i\}_{i=1}^n$ be an indexed family of sets. Let $X = \bigcup_{i=1}^n A_i$. The *Cartesian product* of this indexed family is defined to be the set of all n-tuples of elements of X such that $x_i \in A_i$ for each $i \in \{1, \ldots, n\}$, and is denoted by

$$\prod_{i=1}^{n} A_i \quad \text{or} \quad A_1 \times \cdots \times A_n.$$

In the particular case that all of the members of the family are identical, i.e., $A_i = A$ for each $i \in \{1, \ldots, n\}$, for some set A, we denote the Cartesian product by

$$A^n.$$

Example: The set of all n-tuples of elements of $\mathbb{R}$ is denoted as $\mathbb{R}^n$ and is often called *n-dimensional real space.*

Second, we present the countably infinite case. Given a set X, we define an *(infinite) sequence* of elements of X to be a function

$$x : \mathbb{N} \to X.$$

Given a sequence x, for each $i \in \mathbb{N}$, we often denote the value of x at i by x_i rather than $x(i)$, and call it the ith coordinate of x. We often denote the function x itself by the symbol

$$(x_i)_{i=1}^\infty \quad \text{or} \quad (x_i)_{i \in \mathbb{N}}.$$

Next, let $\{A_i\}_{i=1}^\infty$ be an indexed family of sets. Let $X = \bigcup_{i=1}^\infty A_i$. The *Cartesian product* of this indexed family is defined to be the set of all sequences of elements of X such that $x_i \in A_i$ for each $i \in \mathbb{N}$, and is denoted by

$$\prod_{i=1}^{\infty} A_i.$$

Example: You have probably seen sequences of elements of $\mathbb{R}$ or even $\mathbb{R}^n$ before. In optimization, we often look at algorithms that produce sequences $(x_i)_{i=1}^\infty$ of elements of $\mathbb{R}^n$.

Third, it is possible to define the Cartesian product for the general case, which includes the uncountably infinite case. However, this structure is not explicitly used in the course and is therefore excluded from this presentation.

[2] We will sometimes write the set membership $i \in \{1, \ldots, n\}$ as $i = 1, \ldots, n$.

Basic facts about $\mathbb{R}$

We will simply assume that the reader is familiar with the real numbers $\mathbb{R}$ and the regular operations and relations on them. However, we would like to highlight some properties of $\mathbb{R}$ relevant for optimization.

Upper/lower bound. Let $A \subseteq \mathbb{R}$ and $b \in \mathbb{R}$.

- The element b is called an *upper bound* of A if $a \leq b$ for every $a \in A$. In such case, A is said to be *bound from above (by b)* or that it has an *upper bound*. If A does not have an upper bound, we say that A is *unbounded above*.
- The element b is called a *lower bound* of A if $b \leq a$ for every $a \in A$. In such case, A is said to be *bound from below (by b)* or that it has a *lower bound*. If A does not have a lower bound, we say that A is *unbounded below*.

If A is unbounded above or unbounded below, we say that A is *unbounded*.

Best bounds. Let $A \subseteq \mathbb{R}$ and $b \in \mathbb{R}$.

- The element b is called a *least upper bound* of A if b is an upper bound of A and $b \leq c$ for every upper bound c of A.
- The element b is called a *greatest lower bound* of A if b is a lower bound of A and $c \leq b$ for every lower bound c of A.

A least upper bound does not necessarily exists. However, if a least upper bound does exist, it is unique (show it). A similar remark holds for the greatest lower bound.

$\mathbb{R}$ **is complete.** Every nonempty subset of $\mathbb{R}$ that has an upper bound has a least upper bound. This property is called *completeness* or the *least-upper-bound property*.

One can show that this is equivalent to that every nonempty subset of $\mathbb{R}$ that has a lower bound has a greatest lower bound. This equivalent property is called the *greatest-lower-bound property*.

Supremum/infimum. Let $A \subseteq \mathbb{R}$.

- The *supremum* of A, denoted $\sup A$, is defined as

$$\sup A = \begin{cases} \text{the least upper bound of } A & \text{if } A \text{ has an upper bound and } A \neq \emptyset, \\ \infty & \text{if } A \text{ is unbounded above,} \\ -\infty & \text{if } A = \emptyset. \end{cases}$$

- The *infimum* of A, denoted $\inf A$, is defined as

$$\inf A = \begin{cases} \text{the greatest lower bound of } A & \text{if } A \text{ has a lower bound and } A \neq \emptyset, \\ -\infty & \text{if } A \text{ is unbounded below,} \\ \infty & \text{if } A = \emptyset. \end{cases}$$

Min/max. Let $A \subseteq \mathbb{R}$.

- If $\sup A \in A$, we say that the supremum of A is *attained* or *achieved*, and we sometimes write $\sup A$ as $\max A$.

- If $\inf A \in A$, we say that the infimum of A is *attained* or *achieved*, and we sometimes write $\inf A$ as $\min A$.

If the set A is finite, the supremum and infimum of A are always attained.

Optimizing real-valued functions. Let $f : X \to \mathbb{R}$ be a function for some set X. Let $Y \subseteq X$. We define the *supremum* of f over Y as

$$\sup_{y \in Y} f(y) = \sup f(Y)$$

and the *infimum* of f over Y as

$$\inf_{y \in Y} f(y) = \inf f(Y).$$

If the supremum or infimum above is attained, we sometimes write

$$\max_{y \in Y} f(y) = \max f(Y)$$

and

$$\min_{y \in Y} f(y) = \min f(Y),$$

respectively. Moreover, if the supremum is attained, we denote the set of elements in Y at which the function f is maximized by $\operatorname{Arg\,max}_{y \in Y} f(y)$, also known as *arguments of the maxima*, i.e.,

$$\operatorname*{Arg\,max}_{y \in Y} f(y) = \{y \in Y \mid f(y) = \max f(Y)\}$$

and if the infimum is attained, we denote the set of elements in Y at which the function f is minimized by $\operatorname{Arg\,min}_{y \in Y} f(y)$, also known as *arguments of the minima*, i.e.,

$$\operatorname*{Arg\,min}_{y \in Y} f(y) = \{y \in Y \mid f(y) = \min f(Y)\}.$$

Note that, given that they exists, $\operatorname{Arg\,max}_{y \in Y} f(y)$ and $\operatorname{Arg\,min}_{y \in Y} f(y)$ are subsets of Y. If we know that they are singletons, we write the unique element in $\operatorname{Arg\,max}_{y \in Y} f(y)$ as

$$\operatorname*{arg\,max}_{y \in Y} f(y)$$

and unique element in $\operatorname{Arg\,min}_{y \in Y} f(y)$ as

$$\operatorname*{arg\,min}_{y \in Y} f(y).$$

In particular, given that they exists, $\operatorname{arg\,max}_{y \in Y} f(y)$ and $\operatorname{arg\,min}_{y \in Y} f(y)$ are elements of Y.

Absolute value. The *absolute value* of a real number $x \in \mathbb{R}$, denoted $|x|$, is defined by

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

If $x, y \in \mathbb{R}$, then

$$|x| \leq y \quad (|x| < y)$$

if and only if

$$-y \leq x \leq y \quad (-y < x < y).$$

Intervals and rays.

Let $a, b \in \mathbb{R}$. There are four subsets of $\mathbb{R}$ that are called the *intervals* determined by a and b. They are the following:

$$\begin{aligned}(a,b) &= \{x \in \mathbb{R} \mid a < x < b\} \\ (a,b] &= \{x \in \mathbb{R} \mid a < x \leq b\} \\ [a,b) &= \{x \in \mathbb{R} \mid a \leq x < b\} \\ [a,b] &= \{x \in \mathbb{R} \mid a \leq x \leq b\}.\end{aligned}$$

The first interval is called an *open interval* in $\mathbb{R}$, the middle two are called *half-open intervals* in $\mathbb{R}$ and the last one is called a *closed interval* in $\mathbb{R}$.

Moreover, there are four subsets of $\mathbb{R}$ that are called the *rays* determined by a. They are the following:

$$\begin{aligned}(a,\infty) &= \{x \in \mathbb{R} \mid a < x\} \\ (-\infty,a) &= \{x \in \mathbb{R} \mid x < a\} \\ [a,\infty) &= \{x \in \mathbb{R} \mid a \leq x\} \\ (-\infty,a] &= \{x \in \mathbb{R} \mid x \leq a\}.\end{aligned}$$

The first two rays are called *open rays* in $\mathbb{R}$ and the last two are called *closed rays* in $\mathbb{R}$. Sometimes, we do not distinguish between intervals and rays, and simply call both of them intervals.

Extending the real line. In this course, we will sometimes work with the extended real numbers[3], i.e., $\mathbb{R}$ together with the infinite values ∞ and $-\infty$. In this section, we extend some arithmetic operations and relations in $\mathbb{R}$ to all of $\mathbb{R} \cup \{\infty, -\infty\}$.

The order between ∞ and $-\infty$ is $-\infty < \infty$. We use the following order between elements in $\mathbb{R}$ and $\{\infty, -\infty\}$:

$$-\infty < x < \infty$$

[3]To be precise, we will consider functions that take values in $\mathbb{R} \cup \{\infty\}$. However, since $\mathbb{R} \cup \{\infty\}$ is a subset of $\mathbb{R} \cup \{\infty, -\infty\}$, our discussion here applies.

for each $x \in \mathbb{R}$. We define

$$\begin{aligned}
\infty + x = x + \infty = \infty, && \forall x \in \mathbb{R} \cup \{\infty\},\\
x - \infty = (-\infty) + x = -\infty, && \forall x \in \mathbb{R} \cup \{-\infty\},\\
\lambda\infty = \infty\lambda = \infty, && \forall \lambda \in (0, \infty),\\
\lambda(-\infty) = (-\infty)\lambda = -\infty, && \forall \lambda \in (0, \infty),\\
\lambda\infty = \infty\lambda = -\infty, && \forall \lambda \in (-\infty, 0),\\
\lambda(-\infty) = (-\infty)\lambda = \infty, && \forall \lambda \in (-\infty, 0),\\
0\infty = \infty 0 = 0,\\
0(-\infty) = (-\infty)0 = 0,\\
\infty + (-\infty) = (-\infty) + \infty = \infty.
\end{aligned}$$

We also need to extend the definition of the supermum and infimum. Let $A \subseteq \mathbb{R} \cup \{\infty, -\infty\}$ and $b \in \mathbb{R} \cup \{\infty, -\infty\}$.

- The element b is called an *upper bound* of A if $a \leq b$ for every $a \in A$.
- The element b is called a *lower bound* of A if $b \leq a$ for every $a \in A$.

Note that every subset of $\mathbb{R} \cup \{\infty, -\infty\}$ has an upper and lower bound in $\mathbb{R} \cup \{\infty, -\infty\}$.

- The element b is called a *least upper bound* of A if b is an upper bound of A and $b \leq c$ for every upper bound c of A. This number is denoted as $\sup A$.
- The element b is called a *greatest lower bound* of A if b is a lower bound of A and $c \leq b$ for every lower bound c of A. This number is denoted as $\inf A$.

Note that every subset of $\mathbb{R} \cup \{\infty, -\infty\}$ has a least upper bound and greatest lower bound in $\mathbb{R} \cup \{\infty, -\infty\}$ and they are both necessarily unique.

- If $\sup A \in A$, we say that the supremum of A is *attained* or *achieved*, and we sometimes write $\sup A$ as $\max A$.
- If $\inf A \in A$, we say that the infimum of A is *attained* or *achieved*, and we sometimes write $\inf A$ as $\min A$.

Let $f : X \to \mathbb{R} \cup \{\infty, -\infty\}$ be a function for some set X. Let $Y \subseteq X$. We define the *supremum* and *infimum* of f over Y completely analogous to the case of $f : X \to \mathbb{R}$, but with our extended definitions of supremum and infimum, respectively. We extend correspondingly the discussion if the supremum or infimum is attained.

Inner-product spaces

Vector spaces. A *vector space* over $\mathbb{R}$, also known as a *real vector space*, is a set $\mathbb{V}$ of elements called *vectors*, together with operations $(x, y) \mapsto x + y$ from $\mathbb{V} \times \mathbb{V}$ to $\mathbb{V}$ and $(\alpha, x) \mapsto \alpha \cdot x$ from $\mathbb{R} \times \mathbb{V}$ to $\mathbb{V}$ such that the following holds:

1. $x + y = y + x$ for each $x, y \in \mathbb{V}$,
2. $x + (y + z) = (x + y) + x$ for each $x, y, z \in \mathbb{V}$,

3. There exists in $\mathbb{V}$ a unique vector 0 (called the *zeros vector*) such that $x+0=x$ for each $x \in \mathbb{V}$,
4. For each $x \in \mathbb{V}$, there exists a vector $-x \in \mathbb{V}$ such that $x+(-x)=0$,
5. $(\alpha\beta)\cdot x = \alpha\cdot(\beta\cdot x)$ for each $\alpha,\beta \in \mathbb{R}$ and for each $x \in \mathbb{V}$,
6. $1\cdot x = x$ for each $x \in \mathbb{V}$,
7. $\alpha\cdot(x+y) = \alpha\cdot x+\alpha\cdot y$ for each $\alpha \in \mathbb{R}$ and for each $x,y \in \mathbb{V}$,
8. $(\alpha+\beta)\cdot x = \alpha\cdot x+\beta\cdot x$ for each $\alpha,\beta \in \mathbb{R}$ and for each $x \in \mathbb{V}$.

We will, of course, usually write αx in place of $\alpha\cdot x$. Moreover, the elements in $\mathbb{R}$ are usually called *scalars*. Thus, $+$ is usually referred to as *vector-vector addition* and $\cdot$ as *scalar-vector multiplication*.

Operations involving scalars, vectors, and sets. Let $\mathbb{V}$ be a real vector space, $X,Y \subseteq \mathbb{V}$, $a \in X$ and $\alpha \in \mathbb{R}$. We define:

$$\begin{aligned}
\alpha X &= \{\alpha x \mid x \in X\}, && \textit{scalar-set multiplication}\\
a+X = X+a &= \{a+x \mid x \in X\}, && \textit{vector-set addition}\\
X+Y &= \{x+y \mid x \in X, y \in Y\}, && \textit{Minkowski sum}\\
X-Y &= \{x-y \mid x \in X, y \in Y\}. && \textit{Minkowski difference}
\end{aligned}$$

Linear subspace. Let $\mathbb{V}$ be a real vector space. A nonempty subset $\mathbb{W}$ of $\mathbb{V}$ is called a *linear subspace* of $\mathbb{V}$ if

$$\alpha x+\beta y \in \mathbb{W}$$

for each $\alpha,\beta \in \mathbb{R}$ and for each $x,y \in \mathbb{W}$.

Note that all linear subspaces $\mathbb{W}$ of $\mathbb{V}$ are vector spaces when the operators $+$ and $\cdot$ are restricted to $\mathbb{W}\times\mathbb{W}$ and $\mathbb{R}\times\mathbb{W}$, respectively.

Linear functions. Let $\mathbb{V}$ and $\mathbb{W}$ be two real vector spaces. A function $L:\mathbb{V}\to\mathbb{W}$ is called *linear* if

$$L(\alpha x+\beta y) = \alpha L(x)+\beta L(y)$$

for each $\alpha,\beta \in \mathbb{R}$ and for each $x,y \in \mathbb{V}$.

Affine functions. Let $\mathbb{V}$ and $\mathbb{W}$ be two real vector spaces. A function $T:\mathbb{V}\to\mathbb{W}$ is called *affine* if the function $x \mapsto T(x)-T(0)$ from $\mathbb{V}$ to $\mathbb{W}$ is linear.

Independence. Let $\mathbb{V}$ be a real vector space. A finite set of vectors $\{x_1,\ldots,x_n\} \subseteq \mathbb{V}$ is said to be *linearly dependent* if there exist scalars $\alpha_1,\ldots,\alpha_n \in \mathbb{R}$ not all zero, such that

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

If the finite set of vectors is not linearly dependent, it is said to be *linearly independent*. Thus, the finite set of vectors is linearly independent if the equation

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

implies that

$$\alpha_i = 0$$

for each $i = 1, \ldots, n$.

Linear combination. Let $\mathbb{V}$ be a real vector space and let $\{x_1, \ldots, x_n\} \subseteq \mathbb{V}$ be a finite set of vectors. A vector $y \in \mathbb{V}$ is said to be a *linear combination* of the finite set of vectors if there exist scalars $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that

$$y = \sum_{i=1}^{n} \alpha_i x_i.$$

Basis and dimension. Let $\mathbb{V}$ be a real vector space and let $A \subseteq \mathbb{V}$. The family of vectors A is said to be a *(Hamel or algebraic) basis* of $\mathbb{V}$ if

1. Every finite subset of vectors of A is linearly independent.
2. For every vector $y \in \mathbb{V}$, there exists a finite subset of vectors from A such that y is a linear combination of the finite subset of vectors.

The existence of a basis for every real vector space can be established using a result known as Zorn's lemma, which, although beyond the scope of this course, ensures the existence of such bases. Moreover, it is well known that the number of vectors in all the bases of a real vector space $\mathbb{V}$ is the same — this number is called the *(Hamel or algebraic) dimension* of the space and is denoted by $\dim \mathbb{V}$. We will only work with finite-dimensional vector spaces in this course. However, it is possible to study optimization in infinite dimensional vector spaces.

Normed spaces. Let $\mathbb{V}$ be a real vector space. A *norm* on $\mathbb{V}$ is a function $\|\cdot\| : \mathbb{V} \to \mathbb{R}$ such that the following holds:

1. Positive definiteness: $\|x\| \geq 0$ for each $x \in \mathbb{V}$ and $\|x\| = 0$ if and only if $x = 0$.
2. Absolute homogeneity: $\|\alpha x\| = |\alpha| \, \|x\|$ for each $\alpha \in \mathbb{R}$ and each $x \in \mathbb{V}$.
3. Subadditivity/triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$ for each $x, y \in \mathbb{V}$.

A real vector space $\mathbb{V}$ together with a norm $\|\cdot\|$ on $\mathbb{V}$ is called a *normed vector space* or *normed space*. Such a normed vector space is sometimes written as $(\mathbb{V}, \|\cdot\|)$. Sometimes, the norm of a space $\mathbb{V}$ is denoted by $\|\cdot\|_{\mathbb{V}}$ to emphasize the identity of the vector space and to distinguish it from other norms.

A vector $x \in \mathbb{V}$ is said to be an *unit vector* or *normalized* if $\|x\| = 1$.

Inner-product spaces. Let $\mathbb{V}$ be a real vector space. A *inner product* on $\mathbb{V}$ is a function $\langle \cdot, \cdot \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ such that the following holds:

1. Linearity: $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$ for each $\alpha, \beta \in \mathbb{R}$ and for each $x, y, z \in \mathbb{V}$.
2. Commutativity/symmetry: $\langle x, y \rangle = \langle y, x \rangle$ for each $x, y \in \mathbb{V}$.
3. Positive definiteness: $\langle x, x \rangle \geq 0$ for each $x \in \mathbb{V}$ and $\langle x, x \rangle = 0$ if and only if $x = 0$.

A real vector space $\mathbb{V}$ together with an inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{V}$ is called a *inner-product space*. Such an inner-product space is sometimes written as $(\mathbb{V}, \langle \cdot, \cdot \rangle)$. Sometimes, the inner

product of a space $\mathbb{V}$ is denoted by $\langle \cdot, \cdot \rangle_{\mathbb{V}}$ to emphasize the identity of the vector space and to distinguish it from other inner products.

Two vectors $x, y \in \mathbb{V}$ are said to be *orthogonal* if

$$\langle x, y \rangle = 0.$$

Note that every inner product on $\mathbb{V}$ induces a norm on $\mathbb{V}$, called the *canonical norm*, that is defined by

$$\|x\| = \sqrt{\langle x, x \rangle}$$

for each $x \in \mathbb{V}$. Prove that the canonical norm actually is a norm.

A finite sequence of vectors $x_1, \ldots, x_n$ in $\mathbb{V}$ is said to be *orthonormal* if

$$\langle x_i, x_j \rangle = \delta_{ij}$$

for each $i, j = 1, \ldots, n$, where δ_{ij} is the *Kronecker delta* function defined by

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases}$$

where the range of i and j is clear from context. I.e., the vectors are pairwise orthogonal and normalized.

Cauchy-Schwarz inequality. Let $(\mathbb{V}, \langle \cdot, \cdot \rangle)$ be an inner-product space and let $\|\cdot\|$ be the canonical norm. Then

$$|\langle x, y \rangle| \leq \|x\| \, \|y\| \tag{2}$$

for each $x, y \in \mathbb{V}$, with equality holding in (2) if and only if x and y are linearly dependent. Moreover, if (2) holds with equality and $y \neq 0$, then

$$x = \frac{\langle x, y \rangle}{\|y\|^2} y.$$

Inequality (2) is called the *Cauchy-Schwarz* inequality.

Orthogonal complement. Let $\mathbb{V}$ be a real vector space and $\mathbb{X} \subseteq \mathbb{V}$. The *orthogonal complement* of $\mathbb{X}$, denoted $\mathbb{X}^{\perp}$, is defined as

$$\mathbb{X}^{\perp} = \{v \in \mathbb{V} \mid \forall x \in \mathbb{X}, \langle v, x \rangle = 0\}.$$

Notice that $\mathbb{X}^{\perp}$ is a linear subspace of $\mathbb{V}$.

Complementary subspaces. Let $\mathbb{V}$ be a real vector space and $\mathbb{X}, \mathbb{Y} \subseteq \mathbb{V}$ be two linear subspaces of $\mathbb{V}$. The subspaces $\mathbb{X}$ and $\mathbb{Y}$ are said to be *complementary* whenever

$$\mathbb{V} = \mathbb{X} + \mathbb{Y} \quad \text{and} \quad \mathbb{X} \cap \mathbb{Y} = \{0\},$$

in which case $\mathbb{V}$ is said to be the *direct sum* of $\mathbb{X}$ and $\mathbb{Y}$, and this is denoted by writing $\mathbb{V} = \mathbb{X} \oplus \mathbb{Y}$. This is equivalent to that for each vector $v \in \mathbb{V}$ there are unique vectors $x \in \mathbb{X}$ and $y \in \mathbb{Y}$ such that $v = x + y$.

Real-valued vectors and matrices

n-dimensional real space $\mathbb{R}^n$. The set $\mathbb{R}^n$ is the set of all n-tuples $(x_1, \ldots, x_n)$ of elements of $\mathbb{R}$. We can write this as

$$\mathbb{R}^n = \{(x_1, \ldots, x_n) \mid \forall i \in \{1, \ldots, n\}, x_i \in \mathbb{R}\}.$$

We use the convention that the n-tuple $(x_1, \ldots, x_n)$ of elements of $\mathbb{R}$ can also be written as a column vector, i.e.,

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

using standard matrix notation. This means that we identify vectors in $\mathbb{R}^n$ with matrices in $\mathbb{R}^{n \times 1}$. $\mathbb{R}^n$ is also a real vector space with vector-vector addition defined by

$$\underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}_{\in \mathbb{R}^n} + \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\in \mathbb{R}^n} = \underbrace{\begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_1 + y_n \end{bmatrix}}_{\in \mathbb{R}^n}$$

and scalar-vector multiplication defined by

$$\underbrace{\alpha}_{\in \mathbb{R}} \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}_{\in \mathbb{R}^n} = \underbrace{\begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}}_{\in \mathbb{R}^n}.$$

The most common inner product on $\mathbb{R}^n$ is the *dot product* defined by

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i \tag{3}$$

for each $x, y \in \mathbb{R}^n$, making $\mathbb{R}^n$ an inner-product space. The associated canonical norm is

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{4}$$

for each $x \in \mathbb{R}^n$, called the 2-*norm*. Thus, the Cauchy-Schwarz inequality (2) implies that

$$\left(\sum_{i=1}^{n} x_i y_i \right)^2 \leq \left(\sum_{i=1}^{n} x_i^2 \right) \left(\sum_{i=1}^{n} y_i^2 \right)$$

for each $(x_1, \ldots, x_n), (y_1, \ldots, y_n) \in \mathbb{R}^n$.

$\mathbb{R}^n$ comes with a standard basis, denoted by $\{e_1, \ldots, e_n\}$, where e_i is the vector whose ith coordinate is one while all the others are zeros. The vector of all ones will be denoted by $\mathbf{1}$.

Subsets of $\mathbb{R}^n$. Given a positive integer $n \in \mathbb{N}$, we define

- the *nonnegative orthant* as $\mathbb{R}^n_+ = \{(x_1,\ldots,x_n) \mid \forall i \in \{1,\ldots,n\}, x_i \in [0,\infty)\}$,
- the *positive orthant* as $\mathbb{R}^n_{++} = \{(x_1,\ldots,x_n) \mid \forall i \in \{1,\ldots,n\}, x_i \in (0,\infty)\}$,
- the *nonpositive orthant* as $\mathbb{R}^n_- = \{(x_1,\ldots,x_n) \mid \forall i \in \{1,\ldots,n\}, x_i \in (-\infty,0]\}$, and
- the *negative orthant* as $\mathbb{R}^n_{--} = \{(x_1,\ldots,x_n) \mid \forall i \in \{1,\ldots,n\}, x_i \in (-\infty,0)\}$.

Real-valued $m \times n$ matrices $\mathbb{R}^{m\times n}$. The set of all real-valued $m \times n$ matrices is denoted by $\mathbb{R}^{m\times n}$. We will sometimes use the shorthand notation $X = (X_{ij})$ for matrices in $\mathbb{R}^{m\times n}$, where X_{ij} is entry i,j of the matrix X. The range of i and j will always be clear from context. We will also sometimes use the notation $(X)_{ij}$ to represent entry i,j of the matrix X.

The set $\mathbb{R}^{m\times n}$ is a real vector space with the component-wise addition as the vector-vector addition and the component-wise scalar multiplication as the scalar-vector multiplication. The most common inner product in $\mathbb{R}^{m\times n}$ is defined by

$$\langle X,Y\rangle = \operatorname{Tr}\left(X^T Y\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}Y_{ij}$$

for each $X,Y \in \mathbb{R}^{m\times n}$, making $\mathbb{R}^{m\times n}$ an inner-product space. Here Tr is the trace operation on matrices. The associated canonical norm is

$$\|X\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2} \tag{5}$$

for each $X \in \mathbb{R}^{m\times n}$, called the *Frobenius norm*. Thus, the Cauchy-Schwarz inequality (2) implies that

$$\left(\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}Y_{ij}\right)^2 \le \left(\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2\right)\left(\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}^2\right)$$

for each $X,Y \in \mathbb{R}^{m\times n}$.

Trace. Let us for completeness define the trace operation and present some properties of it. Suppose that $A \in \mathbb{R}^{n\times n}$. The *trace* of the square matrix A, denoted $\operatorname{Tr}(A)$, is defined as

$$\operatorname{Tr}(A) = \sum_{i=1}^{n} A_{ii}$$

The following properties hold:

- $\operatorname{Tr}(\alpha A + \beta B) = \alpha \operatorname{Tr}(A) + \beta \operatorname{Tr}(B)$ for each $\alpha,\beta \in \mathbb{R}$ and each $A,B \in \mathbb{R}^{n\times n}$.
- $\operatorname{Tr}(A) = \operatorname{Tr}(A^T)$ for each $A \in \mathbb{R}^{n\times n}$.
- $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$ for each $A \in \mathbb{R}^{m\times n}$ and each $B \in \mathbb{R}^{n\times m}$.
- In particular, the last point implies that the trace operation is invariant under *cyclic permutations* i.e.,

  $$\operatorname{Tr}(ABC) = \operatorname{Tr}(BCA) = \operatorname{Tr}(CAB)$$

  for each $A \in \mathbb{R}^{m\times n}$, for each $B \in \mathbb{R}^{n\times l}$ and for each $C \in \mathbb{R}^{l\times m}$.

Hadamard product. Suppose that $A, B \in \mathbb{R}^{m \times n}$. The Hadamard product of A and B, denoted $A \odot B$, is defined as the element-wise product

$$(A \odot B)_{ij} = A_{ij} B_{ij}.$$

Kronecker product. Suppose that $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$. The Kronecker product of A and B, denoted $A \otimes B$, is defined by

$$A \otimes B = \begin{bmatrix} A_{11}B & \cdots & A_{1n}B \\ \vdots & \ddots & \vdots \\ A_{m1}B & \cdots & A_{mn}B \end{bmatrix},$$

using block matrix notation.

Range. Let $A \in \mathbb{R}^{m \times n}$. The *range* of A, denoted by $\mathcal{R}(A)$, is defined by

$$\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\}.$$

Note that $\mathcal{R}(A)$ is a linear subspace of $\mathbb{R}^m$.

Nullspace. Let $A \in \mathbb{R}^{m \times n}$. The *nullspace* or *kernel* of A, denoted by $\mathcal{N}(A)$, is defined by

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}.$$

Note that $\mathcal{N}(A)$ is a linear subspace of $\mathbb{R}^n$.

Orthogonal decomposition theorem. Let $A \in \mathbb{R}^{m \times n}$. Then

$$\mathcal{R}(A)^\perp = \mathcal{N}\left(A^T\right) \quad \text{and} \quad \mathcal{N}(A)^\perp = \mathcal{R}\left(A^T\right).$$

Moreover,

$$\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{N}\left(A^T\right) \quad \text{and} \quad \mathbb{R}^n = \mathcal{N}(A) \oplus \mathcal{R}\left(A^T\right).$$

Rank. Let $A \in \mathbb{R}^{m \times n}$. The *rank* of A, denoted by $\operatorname{rank}(A)$, is defined by

$$\operatorname{rank}(A) = \dim \mathcal{R}(A).$$

The rank of A is equal to

- the maximum number of linearly independent columns of A, and
- the maximum number of linearly independent rows of A.

The matrix A is said to have *full rank* if $\operatorname{rank}(A) = \min(m, n)$. The matrix A is said to be *rank-deficient* if it does not have full rank.

We have the following properties:

- $\operatorname{rank}(A) = 0$ if and only if $A = 0$, for each $A \in \mathbb{R}^{m \times n}$.
- $\operatorname{rank}(A) \leq \min(m, n)$, for each $A \in \mathbb{R}^{m \times n}$.
- $\operatorname{rank}(cA) = \operatorname{rank}(A)$, for each $c \in \mathbb{R} \setminus \{0\}$ and for each $A \in \mathbb{R}^{m \times n}$.

- If $\operatorname{rank}(A) = n$, then $\operatorname{rank}(AB) = \operatorname{rank}(B)$ for each $A \in \mathbb{R}^{m\times n}$ and each $B \in \mathbb{R}^{n\times p}$.
- If $\operatorname{rank}(B) = n$, then $\operatorname{rank}(AB) = \operatorname{rank}(A)$ for each $A \in \mathbb{R}^{m\times n}$ and each $B \in \mathbb{R}^{n\times p}$.
- The rank function is subadditive, i.e., $\operatorname{rank}(A+B) \leq \operatorname{rank}(A) + \operatorname{rank}(B)$, for each $A, B \in \mathbb{R}^{m\times n}$.
- $\operatorname{rank}(A) = \operatorname{rank}(A^T) = \operatorname{rank}(A^T A) = \operatorname{rank}(AA^T)$, for each $A \in \mathbb{R}^{m\times n}$.

Rank–nullity theorem. Let $A \in \mathbb{R}^{m\times n}$. Then

$$\operatorname{rank}(A) + \dim \mathcal{N}(A) = n \quad \text{and} \quad \operatorname{rank}(A) + \dim \mathcal{N}\left(A^T\right) = m.$$

Invertibility. Let $A \in \mathbb{R}^{n\times n}$. The square matrix A is called *invertible* or *nonsingular* or *nondegenerate* if there exists an matrix $B \in \mathbb{R}^{n\times n}$ such that

$$AB = BA = I,$$

where I denotes the *identity matrix* of comfortable size. If this is the case, the matrix B is uniquely determined by A, and is called the *(multiplicative) inverse* of A, denoted by A^{-1}. A square matrix that is not invertible is called *singular* or *degenerate*.

Next follows a list of equivalent statements that is useful to determine if the matrix A is invertible or not.

- A is invertible.
- The *determinant* of A is nonzero, i.e., $\det A \neq 0$.
- A has full rank, i.e., $\operatorname{rank}(A) = n$.
- The nullspace of A is trivial, i.e., $\mathcal{N}(A) = \{0\}$.
- The columns of A are linearly independent.
- The rows of A are linearly independent.
- The transpose A^T is an invertible matrix.

Furthermore, if A is invertible, the following properties hold:

- $(A^{-1})^{-1} = A$.
- $(\alpha A)^{-1} = \alpha^{-1} A^{-1}$ for each $\alpha \in \mathbb{R} \setminus \{0\}$.
- $(A^T)^{-1} = (A^{-1})^T$.
- $\det A^{-1} = (\det A)^{-1}$.
- $(AB)^{-1} = B^{-1} A^{-1}$ for each invertible $B \in \mathbb{R}^{n\times n}$.

Symmetric matrices. A square matrix $A \in \mathbb{R}^{n\times n}$ is said to be *symmetric* if $A = A^T$.

Orthogonal matrices. A matrix $U \in \mathbb{R}^{m\times n}$ is said to be *orthogonal* if $U^T U = I$. Note that we must have $m \geq n$. If U is square, i.e., $m = n$, then

$$U^T U = UU^T = I$$

and consequently $U^{-1} = U^T$.

Eigenvalues and eigenvectors. Let $A \in \mathbb{R}^{n \times n}$ be a square matrix and let $x \in \mathbb{C}^n \setminus \{0\}$ be a nonzero vector (the set $\mathbb{C}$ denotes the set of complex numbers, which we assume the reader is familiar with). Suppose that

$$Ax = \lambda x,$$

for some $\lambda \in \mathbb{C}$. Then λ is said to be an *eigenvalue* of A and x is said to be an *eigenvector* of A corresponding to λ.

The *spectrum* of the matrix A is the set of its eigenvalues, i.e.,

$$\{\lambda \in \mathbb{C} \mid \exists x \in \mathbb{C}^n \setminus \{0\}, Ax = \lambda x\}.$$

The spectrum of A can be found by solving the *characteristic equation* of A, i.e.,

$$\det(A - \lambda I) = 0$$

for $\lambda \in \mathbb{C}$. The characteristic equation will always have n solutions, counting multiplicity. Denote these by $\lambda_1, \ldots, \lambda_n$. Then the following holds:

- $\operatorname{Tr} A = \sum_{i=1}^n \lambda_i$.
- $\det A = \prod_{i=1}^n \lambda_i$.
- A is invertible if and only if $\lambda_i \neq 0$ for each $i = 1, \ldots, n$. If A is invertible, the eigenvalues of A^{-1} are $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$, counting multiplicity.

The *spectral radius* of A, denoted $\rho(A)$, is defined as the maximum modulus of the eigenvalues of A, i.e.,

$$\rho(A) = \max_{i=1,\ldots,n} |\lambda_i|.$$

Eigendecomposition. Let $A \in \mathbb{R}^{n \times n}$ and suppose that $q_1, \ldots, q_n$ are linearly independent eigenvectors of A with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$. Then we can decompose A as

$$A = Q \Lambda Q^{-1} \tag{6}$$

where

$$Q = \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix}$$

and

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

The factorization in (6) is called an *eigendecomposition* or *spectral decomposition* of A. Note that the eigendecomposition is not unique.

If in addition A is invertible, the inverse of A can be calculated as

$$A^{-1} = Q \Lambda^{-1} Q^{-1}.$$

Spectral theorem for real symmetric matrices. Let $A \in \mathbb{R}^{n \times n}$. When can we guarantee the existence of n linearly independent eigenvectors of A in the eigendecomposition above? Here we give a condition:

If A is symmetric, then A has n orthonormal (and thus linearly independent) real eigenvectors $q_1, \ldots, q_n \in \mathbb{R}^n \setminus \{0\}$ with corresponding real eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$, and can be decomposed as

$$A = Q\Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T, \tag{7}$$

where

$$Q = \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix}$$

is orthogonal, i.e., $QQ^T = Q^TQ = I$, and

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

We will order the eigenvalues as $\lambda_1 \geq \ldots \geq \lambda_n$. Moreover, sometimes we denote the ith eigenvalue of A by $\lambda_i(A)$, the largest eigenvalue of A by $\lambda_{\max}(A)$, and the smallest eigenvalue of A by $\lambda_{\min}(A)$.

Variational characterization of eigenvalues. Let $A \in \mathbb{R}^{n \times n}$ by symmetric. Then

$$\lambda_{\min}(A)x^Tx \leq x^TAx \leq \lambda_{\max}(A)x^Tx$$

for each $x \in \mathbb{R}^n$. Moreover,

$$\lambda_{\max}(A) = \max\left\{x^TAx \;\middle|\; x \in \mathbb{R}^n, \|x\|_2 = 1\right\},$$
$$\lambda_{\min}(A) = \min\left\{x^TAx \;\middle|\; x \in \mathbb{R}^n, \|x\|_2 = 1\right\}.$$

Comparing eigenvalues. Suppose that $A, B \in \mathbb{R}^{n \times n}$ are symmetric. Then

$$\lambda_i(A) + \lambda_{\min}(B) \leq \lambda_i(A+B) \leq \lambda_i(A) + \lambda_{\max}(B)$$

for each $i = 1, \ldots, n$.

Definiteness. Let $A \in \mathbb{R}^{n \times n}$ be symmetric. We make the following definitions:

- A is *positive definite* if $x^TAx > 0$ for each $x \in \mathbb{R}^n \setminus \{0\}$. We denote this by $A \succ 0$.
- A is *positive semidefinite* if $x^TAx \geq 0$ for each $x \in \mathbb{R}^n$. We denote this by $A \succeq 0$.
- A is *negative definite* if $x^TAx < 0$ for each $x \in \mathbb{R}^n \setminus \{0\}$. We denote this by $A \prec 0$.
- A is *negative semidefinite* if $x^TAx \leq 0$ for each $x \in \mathbb{R}^n$. We denote this by $A \preceq 0$.
- A is *indefinite* is A is neither positive semidefinite nor negative semidefinite.

One can show that:

- A is positive definite if and only if all of its eigenvalues are positive,
- A is positive semidefinite if and only if all of its eigenvalues are non-negative,
- A is negative definite if and only if all of its eigenvalues are negative,
- A is negative semidefinite if and only if all of its eigenvalues are non-positive, and

- A is indefinite if and only if it has both positive and negative eigenvalues.

The set $\mathbb{S}^n$ and some subsets. We denote the set of $n \times n$ real symmetric matrices as $\mathbb{S}^n$, i.e.,

$$\mathbb{S}^n = \left\{ A \in \mathbb{R}^{n \times n} \;\middle|\; A = A^T \right\}.$$

Note that $\mathbb{S}^n$ is a linear subspace of $\mathbb{R}^{n \times n}$. We denote

- the set of $n \times n$ positive definite matrices as $\mathbb{S}^n_{++} = \{A \in \mathbb{S}^n \mid A \succ 0\}$,
- the set of $n \times n$ positive semidefinite matrices as $\mathbb{S}^n_{+} = \{A \in \mathbb{S}^n \mid A \succeq 0\}$,
- the set of $n \times n$ negative definite matrices as $\mathbb{S}^n_{--} = \{A \in \mathbb{S}^n \mid A \prec 0\}$,
- the set of $n \times n$ negative semidefinite matrices as $\mathbb{S}^n_{-} = \{A \in \mathbb{S}^n \mid A \preceq 0\}$,

Matrix inequalities. Let $A, B \in \mathbb{S}^n$. We use the notation $A \prec B$ to mean $B - A \succ 0$, and so on. These are called *matrix inequalities*.

Symmetric square root. Let $A \in \mathbb{S}^n_+$ and suppose that $A = Q \operatorname{diag}(\lambda_1, \ldots, \lambda_n) Q^T$ is an eigendecomposition as in (7). We define the *symmetric square* root of A as

$$A^{1/2} = Q \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) Q^T.$$

Note that $A^{1/2}$ is a symmetric matrix that solves the equation $A = X^2$ for X.

Singular value decomposition. Let $A \in \mathbb{R}^{m \times n}$ such that $\operatorname{rank}(A) = r$. Then A can be expressed as

$$A = U \Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T \tag{8}$$

where

$$U = \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix} \in \mathbb{R}^{m \times r}$$

such that U is orthogonal,

$$V = \begin{bmatrix} v_1 & \cdots & v_r \end{bmatrix} \in \mathbb{R}^{n \times r}$$

such that V is orthogonal,

$$\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$$

such that $\sigma_1 \geq \ldots \geq \sigma_r > 0$. The factorization is called a *(compact-form) singular value decomposition* of A, the positive numbers σ_i are called the *singular values* of A, vectors u_i are called the *left singular vectors* of A, and v_i the *right singular vectors* of A. These satisfy

$$A v_i = \sigma_i u_i \quad \text{and} \quad u_i^T A = \sigma_i v_i^T$$

for each $i = 1, \ldots, r$. Moreover,

$$\sigma_i^2 = \lambda_i \left(A A^T \right) = \lambda_i \left(A^T A \right)$$

for each $i = 1, \ldots, r$, and u_i, v_i are eigenvectors of $A^T A$ and of AA^T, respectively.

Sometimes, we denote the ith singular value of A by $\sigma_i(A)$, the largest singular value of A by $\sigma_{\max}(A)$ and the smallest singular value of A by $\sigma_{\min}(A)$.

Remark. Given a singular value decomposition (8) of A, we see that $\operatorname{rank}(A) = r$, i.e., the rank of A is equal the number of singular values.

Norms on $\mathbb{R}^n$. In this section we introduce some standard norms on $\mathbb{R}^n$. We have previously seen the 2-norm in (4). We generalize this to the *p-norm*, defined by

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, & \text{if } p \in [1, \infty), \\ \max\{|x_i| \mid i = 1, \ldots, n\}, & \text{if } p = \infty \end{cases}$$

for any $x \in \mathbb{R}^n$. Particular important cases, besides the 2-norm, are the 1-*norm* or *taxicab norm* or *Manhattan norm* or *sum-absolute-value norm* (i.e., $p = 1$)

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

and the ∞-*norm* or *Chebyshev norm* (i.e., $p = \infty$)

$$\|x\|_\infty = \max\{|x_i| \mid i = 1, \ldots, n\}.$$

Let $H \in \mathbb{S}^n_{++}$. We have previously seen the dot product in (3) as an example of an inner product on $\mathbb{R}^n$. Another useful option is the *H-inner product*, defined by

$$\langle x, y \rangle_H = x^T H y$$

for each $x, y \in \mathbb{R}^n$. The associated canonical norm is the *H-norm*, given by

$$\|x\|_H = \sqrt{x^T H x}$$

for each $x \in \mathbb{R}^n$.

Equivalence of norms. It is well-known in analysis that all norms on a finite-dimensional vector space are *equivalent*. E.g., suppose that $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbb{R}^n$. Then $\|\cdot\|_a$ and $\|\cdot\|_b$ are said to be *equivalent* if there exists positive constants α and β such that

$$\alpha \|x\|_a \leq \|x\|_b \leq \beta \|x\|_a$$

for each $x \in \mathbb{R}^n$.

In particular, all norms on a finite-dimensional vector space generate the same *topology*, e.g., they define the same open set, the same set of convergent sequences, the same set of continuous functions, etc. Some of these concepts will be defined later in the document.

Norms on $\mathbb{R}^{m \times n}$. Let $A \in \mathbb{R}^{m \times n}$ such that $\operatorname{rank}(A) = r$.

- We have already seen the Frobenius norm in (5). However, since we introduced eigenvalues and singular values, we have an alternative expression for the Frobenius norm of A, i.e.,

$$\|A\|_F = \sqrt{\operatorname{Tr}(A^T A)} = \sqrt{\sum_{i=1}^n \lambda_i (A^T A)} = \sqrt{\sum_{i=1}^r (\sigma_i (A))^2}.$$

- Suppose that $\|\cdot\|_a$ is a norm on $\mathbb{R}^m$ and $\|\cdot\|_b$ is a norm on $\mathbb{R}^n$. The *induced matrix norm* or *operator norm*, denoted $\|\cdot\|_{a,b}$, on $\mathbb{R}^{m\times n}$ is defined by

$$\|A\|_{a,b} = \max\{\|Ax\|_a \mid x \in \mathbb{R}^n, \|x\|_b \leq 1\}.$$

This definition implies that

$$\|Ax\|_a \leq \|A\|_{a,b}\|x\|_b$$

for each $x \in \mathbb{R}^n$. We refer to the matrix norm $\|\cdot\|_{a,b}$ as the (a,b)-*norm*. When $a = b$, we will simply refer to it as an a-*norm* and omit one of the subscripts in its notation, that is, use the notation $\|\cdot\|_a$ instead of $\|\cdot\|_{a,a}$. Here follows some important examples of induced matrix norms:

  - ◇ The 2-norm on $\mathbb{R}^{m\times n}$ (i.e., $a = 2$) is given by

$$\|A\|_2 = \|A\|_{2,2} = \sqrt{\lambda_{\max}(A^TA)} = \sigma_{\max}(A).$$

    The 2-norm on $\mathbb{R}^{m\times n}$ is sometimes also called the *spectral norm*.

  - ◇ The 1-norm on $\mathbb{R}^{m\times n}$ (i.e., $a = 1$) is given by

$$\|A\|_1 = \|A\|_{1,1} = \max_{j=1,\ldots,n}\sum_{i=1}^{m}|A_{ij}|.$$

    The 1-norm on $\mathbb{R}^{m\times n}$ is sometimes also called the *maximum absolute column sum norm*.

  - ◇ The ∞-norm on $\mathbb{R}^{m\times n}$ (i.e., $a = \infty$) is given by

$$\|A\|_\infty = \|A\|_{\infty,\infty} = \max_{i=1,\ldots,m}\sum_{j=1}^{n}|A_{ij}|.$$

    The ∞-norm on $\mathbb{R}^{m\times n}$ is sometimes also called the *maximum absolute row sum norm*.

In general, for each (a,b)-norm on $\mathbb{R}^{m\times n}$, we have that

$$\rho(A) \leq \left\|A^k\right\|_{a,b}^{1/k}$$

for each $k \in \mathbb{N}$, where $\rho(A)$ is the spectral radius of A. Moreover, we have the *spectral radius formula* or *Gelfand's formula*, i.e.,

$$\rho(A) = \lim_{r\to\infty}\|A^r\|_{a,b}^{1/r}.$$

- Let

$$\boldsymbol{\sigma}(A) = \begin{bmatrix}\sigma_1(A) & \cdots & \sigma_r(A)\end{bmatrix}^T \in \mathbb{R}^r_{++}$$

and let $p \in [1,\infty)\cup\{\infty\}$. The *Schatten p-norm* on $\mathbb{R}^{m\times n}$, denoted as $\|\cdot\|_{S_p}$, is defined as

$$\|A\|_{S_p} = \|\boldsymbol{\sigma}(A)\|_p.$$

Here follows some important examples of Schatten p-norms:

◇ The Schatten 2-norm on $\mathbb{R}^{m\times n}$ (i.e., $p=2$) is given by

$$\|A\|_{S_2} = \sqrt{\sum_{i=1}^{r} (\sigma_i(A))^2} = \|A\|_F$$

which coincides with the Frobenius norm.

◇ The Schatten 1-norm on $\mathbb{R}^{m\times n}$ (i.e., $p=1$) is given by

$$\|A\|_{S_1} = \sum_{i=1}^{r} \sigma_i(A).$$

The Schatten 1-norm on $\mathbb{R}^{m\times n}$ is sometimes also called the *nuclear norm* and is sometimes also denoted as $\|\cdot\|_*$.

◇ The Schatten ∞-norm on $\mathbb{R}^{m\times n}$ (i.e., $p=\infty$) is given by

$$\|A\|_{S_\infty} = \sigma_{\max}(A) = \|A\|_2$$

which coincides with the spectral norm.

- Let $k \in \{1,\dots,r\}$. The *Ky Fan k-norm* on $\mathbb{R}^{m\times n}$, denoted as $\|\cdot\|_{\langle k\rangle}$, is defined as

$$\|A\|_{\langle k\rangle} = \sum_{i=1}^{k} \sigma_i(A).$$

Note that the Ky Fan r-norm and nuclear norm coincide, and so do the Ky Fan 1-norm and the spectral norm.

Some analysis

Open balls. For each $x \in \mathbb{R}^n$ and each real number $r > 0$, the *open ball* centered at x with radius r, denoted by $B(x,r)$, is defined as

$$B(x,r) = \{y \in \mathbb{R}^n \mid \|x-y\|_2 < r\}.$$

Open sets. A subset X of $\mathbb{R}^n$ is called *open* if for each $x \in X$, there exists a real number $r > 0$ such that

$$B(x,r) \subseteq X.$$

The empty set $\emptyset$ is (vacuously) open and the entire space $\mathbb{R}^n$ is open. One can show that arbitrary unions of open set are open, and that finite intersections of open sets are open.

Boundedness. A subset X of $\mathbb{R}^n$ is called *bounded* if there exists an $x \in \mathbb{R}^n$ and a real number $R > 0$ such that

$$X \subseteq B(x,R),$$

i.e., if the set is contained in some open ball. Equivalently, the set X is bounded if there exists a real number $M > 0$ such that

$$\|x\|_2 \le M$$

for each $x \in X$. If X is not bounded, we say that X is *unbounded*.

A sequence $(x_k)_{k=1}^{\infty}$ of elements of $\mathbb{R}^n$ is called *bounded* if the set $\{x_k \mid k \in \mathbb{N}\}$ is bounded.

Closed sets. A subset of $\mathbb{R}^n$ is called *closed* if the complement in $\mathbb{R}^n$ is open.

Hence, both the set $\emptyset$ and $\mathbb{R}^n$ are closed[4]. One can show that arbitrary intersections of closed sets are closed and that finite unions of closed sets are closed.

Limit of sequences. Let $(x_k)_{k=1}^{\infty}$ be a sequence of elements of $\mathbb{R}^n$ and $\bar{x} \in \mathbb{R}^n$. We say that $\bar{x}$ is a *limit* of the sequence $(x_k)_{k=1}^{\infty}$, denote by

$$\lim_{k\to\infty} x_k = \bar{x} \quad \text{or} \quad x_k \to \bar{x} \text{ as } k \to \infty \quad \text{or} \quad x_k \xrightarrow[k\to\infty]{} \bar{x},$$

if for each $\epsilon > 0$, there exists $K \in \mathbb{N}$ such that

$$\|x_k - \bar{x}\|_2 < \epsilon$$

for each integer $k \geq K$.

A sequence of elements of $\mathbb{R}^n$ is said to *converge* and to be a *convergent sequence* if it has a limit. In such a case, the limit is always unique.

Squeeze theorem. Suppose that $(a_k)_{k=1}^{\infty}, (b_k)_{k=1}^{\infty}$ and $(c_k)_{k=1}^{\infty}$ are sequences of elements of $\mathbb{R}$ and

$$\lim_{k\to\infty} a_k = \lim_{k\to\infty} c_k = l$$

for some real number $l \in \mathbb{R}$. If there exists $K \in \mathbb{N}$ such that

$$a_k \leq b_k \leq c_k$$

for each integer $k \geq K$, then

$$\lim_{k\to\infty} b_k = l.$$

This result is known as the *squeeze theorem*. It is also known as the *pinching theorem*, the *sandwich rule*, the *police theorem* and the *between theorem*.

Characterization of closed sets. A subset X of $\mathbb{R}^n$ is closed if and only if the limit of every convergent sequence of elements of X is contained in X.

Compact sets. A subset of $\mathbb{R}^n$ is called *compact* if it is closed and bounded.

Interior of a set. Let $X \subseteq \mathbb{R}^n$ and $x \in X$. The element x is called an *interior point* of X if there exists an $\epsilon > 0$ such that

$$B(x, \epsilon) \subseteq X$$

i.e., there exists an open ball centered at x that is contained in X. The set of all interior point of X, denoted $\operatorname{int} X$, is called the *interior* of X, i.e.,

$$\operatorname{int} X = \{x \in X \mid \exists \epsilon > 0, B(x, \epsilon) \subseteq X\}.$$

[4]Thus, the sets $\emptyset$ and $\mathbb{R}^n$ are both open and close. Sets that are both open and closed are sometimes called *clopen*.

Closure of a set. Let $X \subseteq \mathbb{R}^n$. The *closure* of X, denoted as $\operatorname{cl} X$, is defined as

$$\operatorname{cl} X = \{x \in \mathbb{R}^n \mid \forall \epsilon > 0, \exists y \in X, \|x - y\|_2 < \epsilon\}.$$

The closure of X is also given by $\operatorname{cl} X = \mathbb{R}^n \setminus \operatorname{int}(\mathbb{R}^n \setminus X)$.

Boundary of a set. Let $X \subseteq \mathbb{R}^n$. The *boundary* of X, denoted as $\operatorname{bd} X$, is defined as

$$\operatorname{bd} X = \operatorname{cl} X \setminus \operatorname{int} X.$$

Elements of $\operatorname{bd} X$ are called *boundary points* of X.

We can characterize closed and open sets using the boundary: X is closed if it contains its boundary, i.e., $\operatorname{bd} X \subseteq X$. It is open if it contains no boundary points, i.e., $X \cap \operatorname{bd} X = \emptyset$.

Affine sets. Let $V \subseteq \mathbb{R}^n$.

- The set V is said to be *affine* if

  $$\alpha x + (1-\alpha) y \in V$$

  for each $x, y \in V$ and each $\alpha \in \mathbb{R}$.
- If $x_1, \ldots, x_m \in \mathbb{R}^n$ and $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ such that $\sum_{i=1}^m \alpha_m = 1$, we say that the point

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

  is an *affine combination* of the points $x_1, \ldots, x_m$.
- Using induction, one can show that every affine set V contains every affine combination of the points of V.
- One can show that arbitrary intersections of affine sets are affine.
- If V is affine and $x_0 \in V$, then the set

  $$S = V - x_0$$

  is a linear subspace of $\mathbb{R}^n$. Thus, the affine set V can be written as

  $$V = S + x_0,$$

  i.e., as a linear subspace plus an offset. Note that S does not depend on the choice of x_0. The set S is sometimes called the *linear subspace parallel to* V.

Affine hull. Let $X \subseteq \mathbb{R}^n$. The *affine hull* of X, denoted $\operatorname{aff} X$, is defined by

$$\operatorname{aff} X = \left\{ \sum_{i=1}^m \alpha_i x_i \;\middle|\; m \in \mathbb{N}, (x_1, \ldots, x_m) \in X^m, (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m, \sum_{i=1}^m \alpha_i = 1 \right\}.$$

One can show that $\operatorname{aff} X$ is equal to the smallest affine set containing X, i.e.,

$$\operatorname{aff} X = \bigcap \{A \subseteq \mathbb{R}^n \mid X \subseteq A, A \text{ affine}\}.$$

Relative interior. Let $X \subseteq \mathbb{R}^n$. The *relative interior* of X, denoted $\operatorname{relint} X$, is defined by

$$\operatorname{relint} X = \{x \in X \mid \exists \epsilon > 0, B(x, \epsilon) \cap \operatorname{aff} X \subseteq X\}.$$

I.e, the relative interior of X is its interior relative to the affine hull of X.

Relative boundary. Let $X \subseteq \mathbb{R}^n$. The *relative boundary* of X, denoted $\operatorname{relbd} X$, is defined by

$$\operatorname{relbd} X = \operatorname{cl} X \setminus \operatorname{relint} X.$$

Monotone sequences. A sequence $(x_k)_{k=1}^{\infty}$ of elements of $\mathbb{R}$ is called

- *nondecreasing* if $x_k \leq x_{k+1}$ for each $k \in \mathbb{N}$,
- *increasing* if $x_k < x_{k+1}$ for each $k \in \mathbb{N}$,
- *nonincreasing* if $x_k \geq x_{k+1}$ for each $k \in \mathbb{N}$,
- *decreasing* if $x_k > x_{k+1}$ for each $k \in \mathbb{N}$, and
- *monotone* if it is either nondecreasing or nonincreasing.

Monotone convergence theorem. Every monotone sequence of real numbers converges if and only if the sequence is bounded.

Continuity. Let $X \subseteq \mathbb{R}^n$ and $f : X \to \mathbb{R}^m$. Let $x_0 \in X$. We say that f is *continuous* at x_0 if for each $\epsilon > 0$, there exists $\delta > 0$ such that

$$\|f(x) - f(x_0)\|_2 < \epsilon$$

whenever $x \in X$ and satisfies $\|x - x_0\|_2 < \delta$.

We have an alternative characterization of continuity. The function f is continuous at x_0 if and only if

$$f(x_k) \to f(x_0) \quad \text{as} \quad k \to \infty$$

for each sequence $(x_k)_{k=1}^{\infty}$ of elements of X that converges and has x_0 as its limit.

We say that the function f is *continuous* if f is continuous at x_0 for each $x_0 \in X$.

Limit of functions. Let $X \subseteq \mathbb{R}^n$, $f : X \to \mathbb{R}^m$, $x_0 \in \operatorname{cl} X$ and $\bar{f} \in \mathbb{R}^m$. We say that f has *limit* $\bar{f}$ as x approaches x_0 and denote this by

$$\lim_{x \to x_0} f(x) = \bar{f} \quad \text{or} \quad f(x) \to \bar{f} \text{ as } x \to x_0 \quad \text{or} \quad f(x) \xrightarrow[x \to x_0]{} \bar{f}$$

if for each $\epsilon > 0$, there exists $\delta > 0$ such that

$$\left\| f(x) - \bar{f} \right\|_2 < \epsilon$$

whenever $x \in X$ and satisfies $0 < \|x - x_0\|_2 < \delta$. Note that if $\bar{f}$ is the limit of f as x approaches x_0, then $\bar{f}$ is the unique limit of f as x approaches x_0.

One can check that f has limit $\bar{f}$ as x approaches x_0 if and only if

$$\lim_{x_k \to x_0} f(x_k) = \bar{f}$$

for each sequence $(x_k)_{k=1}^{\infty}$ of elements of X, all different from x_0, such that $\lim_{k\to\infty} x_k = x_0$.

Differentiability. Let $X \subseteq \mathbb{R}^n$, $f : X \to \mathbb{R}^m$ and $x \in \operatorname{int} X$. We say that f is *differentiable* at x if there exists a matrix $B \in \mathbb{R}^{m\times n}$ such that

$$\frac{f(x+h) - f(x) - Bh}{\|h\|_2} \to 0 \quad \text{as} \quad h \to 0. \tag{9}$$

In such a case, the matrix B is unique and is called the *derivative* of f at x, and is denoted $D_f(x)$. It is not hard to show that if f is differentiable at x, then f is continuous at x.

We say that f is *differentiable* if X is open and f is differentiable every point in X. In such a case, the derivative of f can be seen as a function $D_f : X \to \mathbb{R}^{m\times n}$ with value $D_f(x)$ at $x \in X$.

Linearity of derivative. Let $X \subseteq \mathbb{R}^n$ and $f, g : X \to \mathbb{R}^m$. Let $x \in \operatorname{int} X$ and suppose that f and g are both differentiable at x. Let $\alpha, \beta \in \mathbb{R}$. Then the function $\alpha f + \beta g$ is differentiable at x and

$$D_{\alpha f+\beta g}(x) = \alpha D_f(x) + \beta D_g(x).$$

Partial derivatives. Let $X \subseteq \mathbb{R}^n$ and let $f : X \to \mathbb{R}$ be a real-valued function. Let $x \in \operatorname{int} X$. We define the jth *partial derivative* of f at x, denoted by $\frac{\partial}{\partial x_j} f(x)$, to be

$$\frac{\partial}{\partial x_j} f(x) = \lim_{t\to 0} \frac{f(x+te_j) - f(x)}{t}$$

provided the limit exists.

Calculating the derivative. Let $X \subseteq \mathbb{R}^n$ and $f : X \to \mathbb{R}^m$. Let $f_i : X \to \mathbb{R}$ be the ith *component function* of f, i.e.,

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}$$

for each $x \in X$. Let $x \in \operatorname{int} X$. Then:

- The function f is differentiable at x if and only if each component function f_i is differentiable at x.
- If the function f is differentiable at x, then

  $$(D_f(x))_{ij} = \frac{\partial}{\partial x_j} f_i(x)$$

  for each $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

Example: Let $f : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$f(x) = Ax + b$$

for each $x \in \mathbb{R}^n$, where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Then

$$D_f(x) = A$$

for each $x \in \mathbb{R}^n$.

The Jacobian. Let $X \subseteq \mathbb{R}^n$ and $f : X \to \mathbb{R}^m$. Let $x \in \operatorname{int} X$. Suppose that the jth partial derivative of the component function f_i at x exists, for each $i = 1 \ldots, m$ and each $j = 1, \ldots, n$. Then the *Jacobian* of f at x exists, is denoted as $J_f(x)$, and is the matrix in $\mathbb{R}^{m \times n}$ defined by

$$(J_f(x))_{ij} = \frac{\partial}{\partial x_j} f_i(x)$$

for each $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

Note that if the derivative of f at x exists, it is equal to the Jacobian of f at x. However, it is possible for the partial derivatives, and hence the Jacobian, to exist, without it following that f is differentiable at x. This fact leaves us in something of a quandary. We have no convenient way to determine whether or not a function is differentiable (other than going back to the definition). However, the following result provides a useful criterion for differentiability:

Let $X \subseteq \mathbb{R}^n$, X open and $f : X \to \mathbb{R}^m$. Suppose that the jth partial derivative of the component function f_i exists at each point x of X and are continuous on X, for each $i = 1 \ldots, m$ and for each $j = 1, \ldots, n$. Then f is differentiable at each point of X.

The gradient. Let $X \subseteq \mathbb{R}^n$, $f : X \to \mathbb{R}$, $x \in \operatorname{int} X$ and suppose that f is differentiable at x. Since f is real-valued, the derivative $D_f(x)$ at x is a $1 \times n$ matrix, i.e., a row vector. The transpose of the derivative $D_f(x)$ at x is called the *gradient* of f at x and is denoted by $\nabla f(x)$, i.e.,

$$\nabla f(x) = D_f(x)^T.$$

From the above, we know that we can calculate the gradient by

$$(\nabla f(x))_i = \frac{\partial}{\partial x_i} f(x)$$

for each $i = 1, \ldots, n$.

If f is differentiable, the *gradient mapping* is the function $\nabla f : X \to \mathbb{R}^n$ with value $\nabla f(x)$ at x, for each $x \in X$.

Example: Let $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x) = \frac{1}{2} x^T P x + q^T x + r$$

for each $x \in \mathbb{R}^n$, where $P \in \mathbb{S}^n$, $q \in \mathbb{R}^n$ and $r \in \mathbb{R}$. Then

$$\nabla f(x) = Px + q$$

for each $x \in \mathbb{R}^n$.

The chain rule. Let $X \subseteq \mathbb{R}^n$ and $f : X \to \mathbb{R}^m$. Let $Y \subseteq \mathbb{R}^m$ such that $f(X) \subseteq Y$ and let $g : Y \to \mathbb{R}^p$. Let $x \in \operatorname{int} X$ and $f(x) \in \operatorname{int} Y$. Suppose that f is differentiable at x and that g is differentiable at $f(x)$. Then the composition $g \circ f$ is differentiable at x and

$$D_{g \circ f}(x) = D_g(f(x)) D_f(x).$$

Example: Suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable, $A \in \mathbb{R}^{n \times p}$ and $b \in \mathbb{R}^n$. Define the function $g : \mathbb{R}^p \to \mathbb{R}^m$ such that

$$g(x) = f(Ax + b)$$

for each $x \in \mathbb{R}^p$. Then g is differentiable with derivative

$$D_g(x) = D_f(Ax+b)A$$

for each $x \in \mathbb{R}^p$, and therefore has the gradient

$$\nabla g(x) = A^T \nabla f(Ax+b)$$

for each $x \in \mathbb{R}^p$.

Higher-order partial derivatives. Note that partial derivatives may have partial derivatives, which themselves might have partial derivatives, and so on. Let us give this a notation. Let $X \subseteq \mathbb{R}^n$ and $f : X \to \mathbb{R}^m$. Let $x \in \operatorname{int} X$. Given $i \in \{1,\ldots,m\}$, the *kth-order partial derivatives* of component function f_i at x, each one denoted by $\frac{\partial^k}{\partial x_{j_1}^{\alpha_1} \ldots \partial x_{j_l}^{\alpha_l}} f_i(x)$, is defined by

$$\frac{\partial^k}{\partial x_{j_1}^{\alpha_1} \ldots \partial x_{j_l}^{\alpha_l}} f_i(x) = \underbrace{\frac{\partial}{\partial x_{j_1}} \cdots \frac{\partial}{\partial x_{j_1}}}_{\alpha_1 \text{ times}} \cdots \underbrace{\frac{\partial}{\partial x_{j_l}} \cdots \frac{\partial}{\partial x_{j_l}}}_{\alpha_l \text{ times}} f_i(x)$$

provided that it exists, for each $l \in \{1,\ldots,n\}$, for each $j_1,\ldots,j_l \in \{1,\ldots,n\}$ and for each $\alpha_1,\ldots,\alpha_l \in \{1,\ldots,k\}$ such that $\sum_{o=1}^{l} \alpha_o = k$.

Let $k \in \mathbb{N}$. If all the partial derivatives of order less than or equal to k, of all the component functions of f exist for each point in X and they are continuous on X, we say that f is *k-times continuously differentiable* on X and sometimes denote the class of functions with this property by $\mathcal{C}^k(X,\mathbb{R}^m)$. This is in fact equivalent to that the kth-order derivative of f exists and is continuous.

We have the following result, as the reader may recall: Let $X \subseteq \mathbb{R}^n$ be open and let $f : X \to \mathbb{R}$ be *twice continuously differentiable* (i.e., 2-times continuously differentiable) on X. Then

$$\frac{\partial^2}{\partial x_i \partial x_j} f(x) = \frac{\partial^2}{\partial x_j \partial x_i} f(x)$$

for each $x \in X$ and for each $i,j \in \{1,\ldots,n\}$.

The Hessian. Let $X \subseteq \mathbb{R}^n$ be open and let $f : X \to \mathbb{R}$ be a real-valued function. Suppose that the gradient mapping $\nabla f : X \to \mathbb{R}^n$ exists and is differentiable at $x \in X$. Then the *Hessian* of f at x is the symmetric matrix in $\mathbb{R}^{n\times n}$, denoted by $\nabla^2 f(x)$, defined by

$$\nabla^2 f(x) = D_{\nabla f}(x)$$

and is equal to

$$(\nabla^2 f(x))_{i,j} = \frac{\partial^2}{\partial x_j \partial x_i} f(x) \tag{10}$$

for each $i,j \in \{1,\ldots,n\}$.

Example: Let $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x) = \frac{1}{2} x^T P x + q^T x + r$$

for each $x \in \mathbb{R}^n$, where $P \in \mathbb{S}^n$, $q \in \mathbb{R}^n$ and $r \in \mathbb{R}$. Then

$$\nabla^2 f(x) = P$$

for each $x \in \mathbb{R}^n$.

Example: Suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ is twice differentiable, $A \in \mathbb{R}^{n \times p}$ and $b \in \mathbb{R}^n$. Define the function $g : \mathbb{R}^p \to \mathbb{R}^m$ such that

$$g(x) = f(Ax + b)$$

for each $x \in \mathbb{R}^p$. Then g is twice differentiable with Hessian

$$\nabla^2 g(x) = A^T \nabla^2 f(Ax + b) A$$

for each $x \in \mathbb{R}^p$.

Example: Let $X \subseteq \mathbb{R}^n$ open, $f : X \to \mathbb{R}$ twice differentiable and $g : \mathbb{R} \to \mathbb{R}$ twice differentiable. Define the function $h : X \to \mathbb{R}$ such that

$$h(x) = g(f(x))$$

for each $x \in X$. Then h is twice differentiable with Hessian

$$\nabla^2 h(x) = g'(f(x)) \nabla^2 f(x) + g''(f(x)) \nabla f(x) \nabla f(x)^T$$

for each $x \in X$.

Big $\mathcal{O}$ notation. Here we introduce a type of standard asymptotic notation, restricted to the case used in the course. Let $f, g : \mathbb{N} \to \mathbb{R}$ and $g(k)$ positive for $k \in \mathbb{N}$ large enough. We write

$$f(k) = \mathcal{O}(g(k))$$

if there exists a positive integer $K \in \mathbb{N}$ and positive constant $c > 0$ such that

$$|f(k)| \leq c g(k)$$

for each integer $k \geq K$.

Probability theory

The course requires familiarity with basic definitions and results in probability theory. We recommended any one of the following basic books if you need a refresher:

- (Bertsekas and Tsitsiklis, 2008)
- (Blitzstein and Hwang, 2019)
- (Grimmett and Stirzaker, 2020)
- (Gut, 2009)
- (Wasserman, 2010)

Important keywords are:

Probability spaces.

Distributions.

Random variables.

Expected value.

Variance.

Conditional expectation/variance.

Modes of convergence.

Methods of proof

You will have to write proofs in the course. If you are unfamiliar with sound mathematical arguments we recommended any one of the following basic books:

- (Bond and Keane, 2007)
- (Chartrand et al., 2018)
- (Gerstein, 2012)
- (Hammack, 2018)
- (Houston, 2009)
- (Velleman, 2006)

Important keywords are:

Direct method.

Proof by cases.

Contradiction.

Induction.

Contrapositive method.

References

Axler, S. (2020), *Measure, integration & real analysis*, Graduate texts in mathematics: 282, Springer Open.

Bauschke, H. H. and Combettes, P. L. (2017), *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics, Springer International Publishing.

Beck, A. (2017), *First-order methods in optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Bertsekas, D. P. and Tsitsiklis, J. N. (2008), *Introduction to Probability*, 2 edn, Athena Scientific.

Blitzstein, J. K. and Hwang, J. (2019), *Introduction to probability.*, CRC Press.

Bond, R. J. and Keane, W. J. (2007), *An introduction to abstract mathematics*, Waveland Pr Inc.

Boyd, S. P. and Vandenberghe, L. (2004), *Convex optimization*, Cambridge University Press.

Calafiore, G. C. and El Ghaoui, L. (2014), *Optimization models*, Cambridge University Press.

Cartan, H. (1971), *Differential calculus*, Kershaw.

Chartrand, G., Polimeni, A. D. and Zhang, P. (2018), *Mathematical proofs: A transition to advanced mathematics*, Pearson.

Cohn, D. L. (2013), *Measure theory*, Birkhäuser Advanced Texts Basler Lehrbücher, Springer New York.

Duistermaat, J. J. and Kolk, J. A. C. (2004*a*), *Multidimensional Real Analysis I: Differentiation*, Cambridge Studies in Advanced Mathematics, Cambridge University Press.

Duistermaat, J. J. and Kolk, J. A. C. (2004*b*), *Multidimensional Real Analysis II: Integration*, Cambridge Studies in Advanced Mathematics, Cambridge University Press.

Gerstein, L. J. (2012), *Introduction to mathematical structures and proofs*, Springer New York, NY.

Grimmett, G. and Stirzaker, D. (2020), *Probability and random processes*, Oxford University Press.

Gut, A. (2009), *An Intermediate Course in Probability*, 2 edn, Springer Publishing Company, Incorporated.

Güler, O. (2010), *Foundations of optimization*, Graduate texts in mathematics: 258, Springer.

Halmos, P. R. (1974), *Naive set theory*, Springer.

Hammack, R. (2018), *Book of proof*.

Holst, A. and Ufnarovski, V. (2014), *Matrix theory*, Studentlitteratur.

Houston, K. (2009), *How to think like a mathematician: A companion to undergraduate mathematics*, Cambridge University Press.

Meyer, C. D. (2000), *Matrix analysis and applied linear algebra*, Society for Industrial and Applied Mathematics.

Munkres, J. R. (1991), *Analysis on manifolds*, Perseus Books Advanced Book Program.

Munkres, J. R. (2014), *Topology*, Pearson.

Rockafellar, R. T. (1970), *Convex analysis*, Princeton mathematical series: 28, Princeton Univ. Press.

Rudin, W. (1976), *Principles of mathematical analysis*, International series in pure and applied mathematics (New York), McGraw-Hill.

Velleman, D. J. (2006), *How to prove it: A structured approach*, 2 edn, Cambridge University Press.

Wasserman, L. (2010), *All of statistics : a concise course in statistical inference*, Springer.

Zalinescu, C. (2002), *Convex Analysis in General Vector Spaces*, World Scientific.