Lecture 01. Review of Mathematics

Lecturer: Jie Wang Date: Sep 7, 2023

In machine learning area, we model many problems as optimization problems, i.e., finding the maxima and minima of functions. We also model the input space as a linear space, i.e., we use vectors to represent a data point and use matrices to represent a image. To this end, we need some mathematical tools to analyze the properties of the functions and linear space. In this lecture, we introduce a suite of powerful tools from **mathematical analysis** and **linear algebra** that are widely used in machine learning. The major references of this lecture are [1, 2, 3, 4, 6].

1 Mathematical Analysis

We start by recalling some basic concept in mathematical analysis.

1.1 Supremum and Infimum

We begin from some basic definitions, which characterize the properties of real numbers.

Definition 1. A nonempty set $S \subseteq \mathbb{R}$ is **bounded above** if there exists a number $u \in \mathbb{R}$ such that $x \leq u$ for all $x \in S$. The number u is called an **upper bound** for S.

Similarly, the set S is **bounded below** if there exists a number $l \in \mathbb{R}$ such that $l \leq x$ for all $x \in S$. The number l is called a **lower bound** for S.

Definition 2. The real number u is the **least upper bound** for a nonempty set $S \subseteq \mathbb{R}$ if

1. u is an upper bound for S;
2. if u' is any upper bound for S, then $u \leq u'$.

The least upper bound is called the **supremum** of the set S, which is denoted by

$$u = \sup S.$$

If $u \in S$, then u is called the **maximum** point of S, i.e.,

$$u = \max S.$$

Question 1. For any nonempty subset of real numbers that is bounded above, can we always find it a least upper bound?

The Completeness Axiom. Suppose that S is a nonempty subset of real numbers that is bounded above. Then, the set S has a least upper bound.

Definition 3. The real number l is the **greatest lower bound** for a set $S \subseteq \mathbb{R}$ if

1. l is a lower bound for S;
2. if l' is any lower bound for S, then $l \geq l'$.

The greatest lower bound is called the **infimum** of the set S, which is denoted by

$$l = \inf S.$$

If $l \in S$, then l is called the **minimum** point of S, i.e.,

$$l = \min S.$$

1.2 Norms and Inner Products

1.2.1 Norms

In a vector space, norm measures the "length" of a vector, and thus the "distance" between two vectors. Once we have a distance function defined, we can discuss limits, followed by many important concepts and tools in mathematical analysis, especially differentiation and integration (we can of course discuss these concepts and tools without a distance function defined under a topological space setting, which is out of the scope of this class).

Definition 4. A function $f : \mathbb{R}^n \to \mathbb{R}$ with **dom** $f = \mathbb{R}^n$ is called a **norm** if

- f is nonnegative: $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$;
- f is definite: $f(\mathbf{x}) = 0$ only if $\mathbf{x} = 0$;
- f is homogeneous: $f(t\mathbf{x}) = |t|f(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$;
- f satisfies the triangle inequality: $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

We often use the notation $f(\mathbf{x}) = \|\mathbf{x}\|$ to denote the norm function.

Definition 5. The **unit ball** of a given norm $\|\cdot\|$ is the set of vectors with norm less than or equal to one, that is,

$$\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq 1\}.$$

Example 1. For $\mathbf{x} \in \mathbb{R}^n$, the commonly seen ℓ_p norm, $p \geq 1$, is defined by

$$\|\mathbf{x}\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}.$$

The ℓ_1-norm and ℓ_2-norm (the Euclidean norm) are commonly-used regularization terms. Moreover, the Chebyshev or ℓ_∞-norm is given by

$$\|\mathbf{x}\|_\infty = \max\{|x_1|, \ldots, |x_n|\}.$$

Moreover, for any $\mathbf{P} \in \mathbb{S}^n_{++}$—which is the set of $n \times n$ positive definite matrices—we define the $\mathbf{P}$-quadratic norm as

$$\|\mathbf{x}\|_\mathbf{P} = (\langle \mathbf{x}, \mathbf{x} \rangle_\mathbf{P})^{1/2} = (\langle \mathbf{x}, \mathbf{P}\mathbf{x} \rangle)^{1/2} = (\mathbf{x}^\top \mathbf{P}\mathbf{x})^{1/2} = \|\mathbf{P}^{1/2}\mathbf{x}\|_2.$$

Example 2. Norm can also be defined on an other space such as a matrix space. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$.

- The Frobenius norm is

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{i,j}^2}.$$

- The matrix p-norms are

$$\|\mathbf{A}\|_p = \max_{\|\mathbf{x}\|_p = 1} \|\mathbf{A}\mathbf{x}\|_p.$$

Specifically,

$$\begin{aligned}\|\mathbf{A}\|_1 =& \max_j \sum_i |a_{ij}|,\\ \|\mathbf{A}\|_2 =& \sigma_{\max}(\mathbf{A}) = (\lambda_{\max}(\mathbf{A}^\top \mathbf{A}))^{1/2},\\ \|\mathbf{A}\|_\infty =& \max_i \sum_j |a_{ij}|.\end{aligned}$$

- The trace (nuclear/spectral) norm is

$$\|\mathbf{A}\|_* = \sum_i \sigma_i(\mathbf{A}).$$

1.2.2 Inner Products

Definition 6. A function $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with **dom** $f = \mathbb{R}^n \times \mathbb{R}^n$ is called an **inner product** if

- f is nonnegative: $f(\mathbf{x}, \mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$;
- f is definite: $f(\mathbf{x}, \mathbf{x}) = 0$ if and only if $\mathbf{x} = 0$;
- f is symmetric: $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}, \mathbf{x})$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$;
- f is bilinear: $f(a\mathbf{x} + b\mathbf{y}, \mathbf{z}) = af(\mathbf{x}, \mathbf{z}) + bf(\mathbf{y}, \mathbf{z})$ and $f(\mathbf{x}, a\mathbf{y} + b\mathbf{z}) = af(\mathbf{x}, \mathbf{y}) + bf(\mathbf{x}, \mathbf{z})$, for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ and $a, b \in \mathbb{R}$.

We often use the notation $f(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$ to denote the inner product function.

Example 3. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Then

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = x_1 y_1 + \cdots + x_n y_n$$

is an inner product. For any positive definite matrix $\mathbf{P}$, we can also define an inner product as

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{P}} = \langle \mathbf{x}, \mathbf{P}\mathbf{y} \rangle = \mathbf{x}^\top \mathbf{P}\mathbf{y}.$$

Example 4. Inner product can also be defined as above on a general linear space.

- Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$. Then

$$\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^\top \mathbf{B}) = \sum_{i,j} a_{ij} b_{ij}.$$

  is an inner product.

- Let $l^2(\mathbb{R}) = \{\mathbf{x} = (x_1, x_2, \cdots, x_n, \cdots), x_i \in \mathbb{R}, \sum_{i=1}^\infty |x_i|^2 < \infty\}$. Then

$$\langle \mathbf{x}, \mathbf{y} \rangle_{l^2} = \sum_{i=1}^\infty x_i y_i$$

  is an inner product.

- Let $L^2(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R}, \int_{\mathbb{R}} |f(x)|^2 \mathrm{d}x < \infty\}$. Then

$$\langle f, g \rangle_{L^2} = \int_{\mathbb{R}} f(x)g(x)\mathrm{d}x$$

  is an inner product. (Here we view two functions that equals almost everywhere as the same in $L^2(\mathbb{R})$).

Note that for any inner product, we can naturally define a norm, that is, a norm induced by the inner product

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

Question 2. Can any norm be induced by an inner product?

Proposition 1 (**Cauchy-Schwarz Inequality**). *For all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$*, there holds*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|.$$

The equality holds when and only when $\mathbf{x}$ *and* $\mathbf{y}$ *are linearly dependent, i.e.,* $\lambda \mathbf{x} + \mu \mathbf{y} = 0$ *for some* $\lambda, \mu \in \mathbb{R}$.

Definition 7. For any non-zero vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the **included angle** is defined as

$$\Theta(\mathbf{x}, \mathbf{y}) = \arccos \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}.$$

Definition 8. Two vectors $\mathbf{x}$ and $\mathbf{y}$ are **orthogonal** if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.

1.3 Basic Topology of $\mathbb{R}^n$

1.3.1 Open Sets

Definition 9. Given $\epsilon > 0$, the **ϵ-neighborhood** of a point $x \in \mathbb{R}^n$ is

$$N_\epsilon(\mathbf{x}) = \{\mathbf{y} : \mathbf{y} \in \mathbb{R}^n, \|\mathbf{y} - \mathbf{x}\| < \epsilon\}.$$

The number ϵ is called the radius of $N_\epsilon(\mathbf{x})$.

Definition 10. An element $\mathbf{x} \in S \subseteq \mathbb{R}^n$ is called an **interior point** of S if there exists an $\epsilon > 0$ such that $N_\epsilon(\mathbf{x}) \subseteq S$.

Definition 11. The set of interior points of S is called the **interior** of S, which is denoted by S° or **int** S.

Definition 12. A set $O \subseteq \mathbb{R}^n$ is **open** if every point in O is an interior point of O, i.e., $O =$ **int** O.

Question 3.

3.1. Is the ϵ-neighborhood an open set?

3.2. Is $(0, 1) \subset \mathbb{R}$ an open set?

3.3. Is $(0, 1) \subset \mathbb{R}^2$ an open set?

1.3.2 Closed Sets

Another type of sets that is closely related to the open sets is the so-called **closed sets**. We can easily define the closed sets by using open sets.

Definition 13. A set $F \subseteq \mathbb{R}^n$ is **closed** if its complement set, that is,

$$\mathbb{R}^n \setminus F = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \notin F\},$$

is open.

Definition 13 implies that, if $F \subseteq$ is closed in $\mathbb{R}^n$, we can find for each $\mathbf{x} \notin F$ a neighborhood $N_\epsilon(\mathbf{x}) \subset \mathbb{R}^n \setminus F$, where ϵ may depend on $\mathbf{x}$. Another useful approach to characterize the topological properties of closed sets is by **convergent sequences**.

Definition 14. A **sequence** $(\mathbf{x}_k)$ of vectors in $\mathbb{R}^n$ is said to **converge** to $\mathbf{x} \in \mathbb{R}^n$ if for any $\epsilon > 0$, there exists a positive integer N such that

$$\|\mathbf{x}_k - \mathbf{x}\|_2 < \epsilon, \forall\, k \geq N.$$

Symbolically, $\mathbf{x}_k \to \mathbf{x}$ or $\lim_{k\to\infty} \mathbf{x}_k = \mathbf{x}$.

Theorem 1. *If the limit of a sequence exists, it must be unique.*

Definition 15. A vector $\mathbf{x} \in \mathbb{R}^n$ is a **limit (cluster/accumulation) point** of a set $S \subseteq \mathbb{R}^n$ if there exists a sequence $(\mathbf{x}_k) \subseteq S$ and $\mathbf{x}_k \neq \mathbf{x}$ for $k = 1, 2, \ldots$ such that $\mathbf{x}_k \to \mathbf{x}$.

Definition 16. A set $F \subseteq \mathbb{R}^n$ is **closed** if it contains all of its limit point.

Question 4. Let S' be the set of all limit points of S. How to characterize the points left in S after we remove all the points in S'? In other words, what can we say about the points in $S \setminus S'$?

To Question 4, we give an equivalent definition of limit point by neighborhood.

Definition 17. A vector $\mathbf{x} \in \mathbb{R}^n$ is a **limit (cluster/accumulation) point** of a set $S \subseteq \mathbb{R}^n$ if every neighborhood of $\mathbf{x}$ contains a point $\mathbf{x}' \neq \mathbf{x}$ such that $\mathbf{x}' \in S$.

In view of Definition 17, we can easily characterize the points in $S \setminus S'$, which is known as the isolated points.

Definition 18. A vector $\mathbf{x} \in S \subseteq \mathbb{R}^n$ is an **isolated point** of S if it is not a limit point of S, that is, there exists a neighbor of $\mathbf{x}$ that contains no other points in S other than $\mathbf{x}$.

***Remark* 1.** Notice that, for a nonempty set $S \subseteq \mathbb{R}^n$, its limit points may not belong to S, while its isolated points must be one of the elements in S. However, either S' or $S \setminus S'$ can be empty (when?), but not both under the nonempty assumption of S.

When a set S does not contain all of its limit points, we may say that S is not closed to the limit operations of the sequences in S. This may lead to practical difficulties. For example, what is the length of the diagonal of the unit square if you only know rational numbers? Thus, expanding the set such that it contains all its limit points becomes desirable.

Definition 19. The **closure** of the set S, denoted by **cl** S or $\bar{S}$, is the set $S \cup S'$.

In view of Definitions 16 and 19, we immediately have the result as follows.

Theorem 2. *Let $S \subseteq \mathbb{R}^n$.*

1. *The set $\bar{S}$ is closed.*
2. *The set S is closed if and only if $S = \bar{S}$.*

Question 5.

1. Is $(0, 1]$ closed in $\mathbb{R}$?
2. Is $(0, 1]$ closed in $(0, \infty)$?

Remark 2. When we discuss the openness or closedness of a given set S, we always refers to another set Ω that includes S. Specifically, even for the same set S, it can be open with respect to a set Ω with $S \subseteq \Omega$, and it can also be closed with respect to another set Ω' with $S \subseteq \Omega'$ as well (please see Questions 3 and 5). Thus, a rigorious way to claim that "the set S is open or closed" is to say that "the set S is open or closed in Ω (with $S \subseteq \Omega$)".

Question 6.

1. Can you find a set that is **open-and-closed**?
2. Can you find a set that is neither open nor closed?

1.3.3 The Boundary of A Set

Enlightened by the definition of open sets introduced in Section 1.3.1, we can characterize the inside of a set $S \subseteq \mathbb{R}^n$ by its interior. This naturally raises two questions.

Question 7.

1. How to characterize the outside of a set $S \subseteq \mathbb{R}^n$?
2. How to characterize the boundary of a set $S \subseteq \mathbb{R}^n$?

As long as we know how to characterize the inside of a set, we can easily characterize its outside (how?). Thus, given a set S, the points left by removing the inside and outside of S naturally belong to the boundary of S. We formalize this idea by the definition as follows.

Definition 20. A point $\mathbf{x}$ is a **boundary point** of a set $S \subseteq \mathbb{R}^n$ if every ϵ-neighborhood of $\mathbf{x}$ contains both points belonging to S and points not belonging to S.

We can further characterize the boundary points by the results as follows.

Theorem 3. *Let ∂S (also denoted by* **bd** *S) be the boundary of a set $S \subseteq \mathbb{R}^n$. Then,*

$$\partial S = \bar{S} \setminus S^\circ.$$

Question 8.

1. Can we claim that **bd** $S \subseteq S$?
2. Is that possible **bd** $S = S$?

1.3.4 Compact Sets

Definition 21. A set $S \subseteq \mathbb{R}^n$ is **bounded** if there exists a scalar M such that

$$\|\mathbf{x}\|_2 \leq M, \forall \mathbf{x} \in S.$$

Definition 22. A set $S \subseteq \mathbb{R}^n$ is **compact** if every sequence in S has a subsequence that converges to a point in S.

Theorem 4. *A set $S \subseteq \mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Remark 3. Different from the "openness" and "closedness", the property of "compactness" is intrinsic [5]—that is, if $A \subseteq B \subseteq C$, then A is compact in B if and only if A is compact in C, while the property of being closed (or open) is not intrinsic (see Question 5).

1.4 Continuous Functions

Definition 23. Let $f : S \to \mathbb{R}$ where $S \subseteq \mathbb{R}^n$. We say f is **continuous** at $\mathbf{x}_0 \in S$ if for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|f(\mathbf{x}) - f(\mathbf{x}_0)| < \epsilon,$$

for all $\mathbf{x} \in S$ and $\|\mathbf{x} - \mathbf{x}_0\| < \delta$. A function is continuous if it is continuous at every point in its domain.

Question 9. Let $f : \mathbb{N} \to \mathbb{R}$, where $\mathbb{N}$ is the set of all integers. Is f a continous function?

A handy property of continuous functions f is that, for any sequence $(\mathbf{x}_k) \subset \mathbf{dom}\ f$ that converges to $\mathbf{x} \in \mathbf{dom}\ f$, we have

$$\lim_{k\to\infty} f(\mathbf{x}_k) = f\left(\lim_{k\to\infty} \mathbf{x}_k\right) = f(\mathbf{x}).$$

Proposition 2 (Bolzano-Weierstrass Theorem). *Every bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.*

Theorem 5 (Extreme Value Theorem). *Let C be a compact subset of $\mathbb{R}^n$ and $f : C \to \mathbb{R}$ be continuous. Then, there exist $a, b \in C$ such that*

$$f(a) \leq f(\mathbf{x}) \leq f(b), \forall \mathbf{x} \in C.$$

In other words, f attains maximum and minimum values in C.

Theorem 5 is one of the most important results in calculus, as it provides a method to show the existence of the optimum of optimization problems.

Question 10.

Which property will be preserved by a continuous function, openness, closedness, or compactness? Specifically, let $S \subset \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function, and

$$I = \{\mathbf{y} : \mathbf{y} \in \mathbb{R}, \exists\, \mathbf{x} \in S, \text{such that } f(\mathbf{x}) = y\}.$$

If S is open/closed/compact, will I be open/closed/compact?

If you can show that the compactness is preserved by continuous functions, you can immediately obtain the result in Theorem 5.

2 Linear Algebra

Linear algebra is central to almost all areas of mathematics and is also powerful in most sciences and fields of engineering, including machine learning. In this section, we review some of the basics of linear algebra.

2.1 Linear Space

We start with linear space, which is one of the most important concepts in linear algebra. A critical property of linear spaces is that each vector can be linearly represented by a finite number of other vectors. To understand this statement, we need to answer the following questions.

1. What is a linear space?
2. What is the linear combination?
3. What is the basis of a linear space?
4. Do all linear spaces have a basis? Are their bases always finite or countable?

We can find the answers to the first three questions in this lecture. However, to answer the last question, we need Axiom of Choice or Zorn's Lemma, which are beyond our scope. In this course, we admit that every linear space has a set of basis and know that some linear spaces have non-countable basis.

Definition 24. Let $\mathcal{V}$ be a nonempty set and F be a number field (e.g., $\mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$). We say that $\mathcal{V}$ is the **linear space** (or **vector space**) **over** F if the following conditions hold.

1. We have defined two binary operations in $\mathcal{V}$.
   (a) The first operation, called **vector addition** or simply **addition**, assigns to any two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathcal{V}$ a third vector in $\mathcal{V}$ which is commonly written as $\mathbf{u}+\mathbf{v}$, and called the sum of these two vectors.
   (b) The second operation, called **scalar multiplication**, assigns to any scalar $a \in F$ and any vector $\mathbf{v} \in \mathcal{V}$ another vector in $\mathcal{V}$, which is denoted $a\mathbf{v}$.
2. The addition and the scalar multiplication defined in $\mathcal{V}$ satisfy the following eight axioms.
   (a) The addition is **commutative**, i.e., $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}, \forall \mathbf{u}, \mathbf{v} \in \mathcal{V}$.
   (b) The addition is **associative**, i.e., $(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w}), \forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$.
   (c) There exists a **zero vector** in $\mathcal{V}$, i.e., there exists $\theta \in \mathcal{V}$ such that $\theta+\mathbf{v}=\mathbf{v}+\theta=\mathbf{v}$, $\forall \mathbf{v} \in \mathcal{V}$. The zero vector is also denoted by $\mathbf{0}$.
   (d) There exists an **additive inverse** for each vector in $\mathcal{V}$, i.e., for every $\mathbf{v} \in \mathcal{V}$, there exists a vector $\mathbf{v}' \in \mathcal{V}$ such that $\mathbf{v}+\mathbf{v}'=\mathbf{v}'+\mathbf{v}=\mathbf{0}$. We also denote $\mathbf{v}'$ by $-\mathbf{v}$.
   (e) The scalar multiplication is **compatible** with field multiplication, i.e., $(ab)\mathbf{v}=a(b\mathbf{v})$, $\forall \mathbf{v} \in \mathcal{V}, a, b \in F$.
   (f) The **multiplicative identity** in F is the **identity element** of scalar multiplication, i.e., $\mathbf{1v}=\mathbf{v}, \forall \mathbf{v} \in \mathcal{V}$.
   (g) The scalar multiplication is **distributive** with respect to vector addition, i.e., $a(\mathbf{u}+\mathbf{v})=a\mathbf{u}+a\mathbf{v}, \forall a \in F, \mathbf{u}, \mathbf{v} \in \mathcal{V}$.

(h) The scalar multiplication is **distributive** with respect to field addition, i.e., $(a+b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}, \forall a, b \in F, \mathbf{v} \in \mathcal{V}$.

Remark 4. When we talk about a linear space, compared to what the elements in it (i.e., vectors) are, we care more about the operations (i.e., the vector addition and scalar multiplication) defined on it and its linear structure.

Example 5. 1. $\mathbb{R}[x]$ is the linear space consisting of all the polynomials with real coefficients.

2. $C[a, b]$ is the linear space consisting of all the continuous functions on $[a, b]$.

Definition 25. Let $\mathcal{V}$ be a linear space over F and S be a subset of $\mathcal{V}$. For any finite subset $S_1 = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset S$ and $a_1, \ldots, a_k \in F$, we call $a_1\mathbf{v}_1 + \ldots + a_k\mathbf{v}_k$ a **linear combination** of S. If a vector $\mathbf{u} \in \mathcal{V}$ is a linear combination of S, we say that $\mathbf{u}$ can be **linearly represented** by S. The set of all linear combinations of S is denoted $V(S)$.

Note that for every subset $S \subset \mathcal{V}, V(S)$ is a subspace of $\mathcal{V}$. It is easy to show that any subspace $\mathcal{W}$ that contains S also contains $V(S)$. Hence, $V(S)$ is the smallest subspace of $\mathcal{V}$ containing S.

Definition 26. We say that S **spans** or **generates** $V(S), V(S)$ is the **linear span** of S, and S is a **spanning set** or a generating set of $V(S)$.

Definition 27. Let $\mathcal{V}$ be a linear space over F and S be a subset of $\mathcal{V}$. If there exists some difnite subset $S_1 = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset S$ and scalars $a_1, \ldots, a_k$ that are not all 0 such that

$$a_1\mathbf{v}_1 + \ldots + a_k\mathbf{v}_k = \mathbf{0},$$

we say that S is **linearly dependent**. Otherwise, S is **linearly independent**.

Remark 5. Please note the geometric meaning of linear dependencies.

Theorem 6. *Let $\mathcal{V}$ be a linear space over F and S be a subset of $\mathcal{V}$, then S is linearly dependent if and only if there exists some vector $\mathbf{v} \in S$, which is the linear combination of other vectors.*

Definition 28. Let $\mathcal{V}$ be a linear space over F.

1. If there exist n vectors in $\mathcal{V}$ that are linearly independent and every $n+1$ vectors in $\mathcal{V}$ are linearly dependent, we say that the **dimension** of $\mathcal{V}$ is n, denoted $\dim \mathcal{V} = n$.

2. If there exists a set of vectors $M = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ such that every vector $v \in \mathcal{V}$ is the linear combination of M, i.e.,
$$\mathbf{v} = a_1\mathbf{v}_1 + \ldots + a_n\mathbf{v}_n,$$
and the coefficients are uniquely determined by $\mathbf{v}$, we say that M is a **basis** of $\mathcal{V}$. The ordered array $(a_1, \ldots, a_n)$ is called the **coordinate** of $\mathbf{v}$ under the basis M.

2.2 Range and Nullspace

We define the range and the nullsapce of $\mathbf{A} \in \mathbb{R}^{m\times n}$ as follows.

Definition 29. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$. The **range** of $\mathbf{A}$, denoted $\mathcal{R}(\mathbf{A})$, is the set of all vectors in $\mathbb{R}^m$ that can be written as linear combinations of the columns of $\mathbf{A}$, i.e.,

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

The **nullsapce** (or **kernel**) of $\mathbf{A}$, denoted $\mathcal{N}(\mathbf{A})$, is the set of all vectors $\mathbf{x}$ mapped into $\mathbf{0}$ by $\mathbf{A}$, i.e.,

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

Theorem 7. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$.

1. $\mathcal{R}(\mathbf{A})$ *is a subspace of* $\mathbb{R}^m$ *and* $\mathcal{N}(\mathbf{A})$ *is a subspace of* $\mathbb{R}^n$.
2. *The dimension of* $\mathcal{R}(\mathbf{A})$ *is the rank of* $\mathbf{A}$, *i.e.,*
$$\dim \mathcal{R}(\mathbf{A}) = \operatorname{rank} \mathbf{A}.$$
3. *The dimension of* $\mathcal{N}(\mathbf{A})$ *is that of the solution space of* $\mathbf{Ax} = \mathbf{0}$, *i.e.,*
$$\dim \mathcal{N}(\mathbf{A}) = \dim V_{\mathbf{A}}.$$
In fact, we have $\mathcal{N}(\mathbf{A}) = V_{\mathbf{A}}$.
4. *The sum of dimensions of* $\mathcal{R}(\mathbf{A})$ *and* $\mathcal{N}(\mathbf{A})$ *is the dimension of* $\mathbb{R}^n$, *i.e.,*
$$\dim \mathcal{R}(\mathbf{A}) + \dim \mathcal{N}(\mathbf{A}) = n,$$
which is equivalent to what we know about $\mathbf{A}$ *:*
$$\operatorname{rank} \mathbf{A} + \dim V_{\mathbf{A}} = n.$$

Question 11. How many ways can you think of to explain the rank of a matrix?

2.3 Symmetric Eigenvalue Decomposition

Symmetric matrix is import in machine learning with some useful properties.

Theorem 8. *Let* $\mathbf{A} \in S^n$, *i.e.,* $\mathbf{A}$ *is a real symmetric* $n \times n$ *matrix. Then* $\mathbf{A}$ *can be factored as*
$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top,$$
where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ *is orthogonal and* $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$. *Such a decomposition is called the symmetric eigenvalue decomposition or spectral decomposition of* $\mathbf{A}$.

Suppose that $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_n)$, then we have
$$\mathbf{A}(\mathbf{q}_1, \ldots, \mathbf{q}_n) = (\mathbf{q}_1, \ldots, \mathbf{q}_n) \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix},$$
which leads to $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i, 1 \le i \le n$. Note that $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$, which means that $\mathbf{q}_i^\top \mathbf{q}_j = 0$, $1 \le i \ne j \le n$. Hence $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is an orthonormal set of eigenvectors of $\mathbf{A}$.

2.3.1 Definiteness and Matrix Inequalities

First we have a simple observation. Suppose $\mathbf{A} \in S^n$, we let $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ be the maximum and minimum eigenvalue of $\mathbf{A}$, respectively. Then we have
$$\lambda_{\max}(\mathbf{A}) = \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\ne 0} \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}, \quad \lambda_{\min}(\mathbf{A}) = \inf_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\ne 0} \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}.$$

To see this, let $\mathbf{A} = \mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q}$ be the orthogonal decomposition of $\mathbf{A}$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_n)$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\mathbf{A}$. Then we have

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} = \mathbf{x}^\top \mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q} \mathbf{x} \leq \lambda_{\max}(\mathbf{A}) \mathbf{x}^\top \mathbf{Q}^\top \mathbf{Q} \mathbf{x} = \lambda_{\max}(\mathbf{A}) \mathbf{x}^\top \mathbf{x}.$$

The equality holds if and only if $\mathbf{x}$ is an eigenvector of $\lambda_{\max}(\mathbf{A})$. The equation of $\lambda_{\min}(\mathbf{A})$ holds similarly, and we leave it for an exercise.

In the following context, we are going to investigate how the sign of $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ influence the properties of $\mathbf{A}$. First, we introduce the concept of definiteness.

Definition 30. Let $\mathbf{A} \in S^n$.

1. We say $\mathbf{A}$ is **positive definite** or $\mathbf{A} > 0$, if $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$. We denote the set of all positive definite matrices by S^n_{++}.
2. We say $\mathbf{A}$ is **positive semidefinite** or $\mathbf{A} \geq 0$, if $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$. We denote the set of all positive definite matrices by S^n_{+}.
3. We say $\mathbf{A}$ is **negative definite** or $\mathbf{A} < 0$, if $\mathbf{x}^\top \mathbf{A} \mathbf{x} < 0$ for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$.
4. We say $\mathbf{A}$ is **negative semidefinite** or $\mathbf{A} \leq 0$, if $\mathbf{x}^\top \mathbf{A} \mathbf{x} \leq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$.

Notice that by definition, $\mathbf{A}$ is positive definite is equivalent to $\lambda_{\min}(\mathbf{A}) > 0$; $\mathbf{A}$ is positive semidefinite is equivalent to $\lambda_{\min}(\mathbf{A}) \geq 0$. For negative and negative semidefinite cases, we also have similar results. Another observation is that $\mathbf{A} < 0$ is equivalent to $-\mathbf{A} > 0$, $\mathbf{A} \leq 0$ is equivalent to $-\mathbf{A} \geq 0$.

2.4 Singular Value Decomposition (SVD)

Singular value decomposition separates any matrix into simple pieces and is widely used in numerical linear algebra field.

Theorem 9. *Suppose $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $\text{rank}(\mathbf{A}) = r$, then $\mathbf{A}$ can be factorized as*

$$\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top.$$

Here $\mathbf{U} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are both orthogonal matrices, $\mathbf{\Sigma} = \text{diag}(\sigma_1, \cdots, \sigma_r, 0, \cdots, 0)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_r > 0$.

We call σ_i $(i = 1, \cdots, r)$ the singular values of $\mathbf{A}$, the columns of $\mathbf{U}$ left singular vectors and the columns of $\mathbf{V}$ right singular vectors. Then we have

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top,$$

where $\mathbf{u}_i$ is the ith column of $\mathbf{U}$ $(i = 1, \cdots, m)$ and $\mathbf{v}_j$ is the jth column of $\mathbf{V}$ $(j = 1, \cdots, n)$.

Further, one can show that the set of singular values of $\mathbf{A}$ is equal to the set of the arithmetic square root of non-zero eigenvalues of $\mathbf{A}^\top \mathbf{A}$ or $\mathbf{A} \mathbf{A}^\top$.

We denote the largest singular value of $\mathbf{A}$ by $\sigma_{\max}(\mathbf{A})$. Then we can prove that

$$\sigma_{\max}(\mathbf{A}) = \sup_{\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n, \mathbf{x}, \mathbf{y} \neq 0} \frac{\mathbf{x}^\top \mathbf{A} \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} = \sup_{\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \neq 0} \frac{\|\mathbf{A}\mathbf{y}\|_2}{\|\mathbf{y}\|_2} = \|\mathbf{A}\|_2.$$

References

[1] S. Abbott. *Understanding Analysis, 2ed.* Springer, 2015.

[2] D. Bertsekas. *Convex Optimization Theory.* Athena Scientific, 2009.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[4] R. Courant and F. John. *Introduction to Calculus and Analysis.* Springer, 1989.

[5] J. M. Erdman. *A Problems Based Course in Advanced Calculus.* AMS, 2018.

[6] W. Rudin. *Principles of Mathematical Analysis.* McGraw-Hill Education, 1976.