Introduction to Machine Learning
Fall 2022
University of Science and Technology of China

Lecturer: Jie Wang — Homework 1
Name: Yunqin Zhu — ID: PB20061372

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: Limit and Limit Points

1. Show that $\{\mathbf{x}_n\}$ in $\mathbb{R}^n$ converges to $\mathbf{x} \in \mathbb{R}^n$ if and only if $\{\mathbf{x}_n\}$ is bounded and has a unique limit point $\mathbf{x}$.

   Solution:

   ($\Rightarrow$) Suppose that $\mathbf{x}_n \to \mathbf{x}$. For any $\epsilon > 0$, there exists a positive integer N such that $\|\mathbf{x}_k - \mathbf{x}\|_2 < \epsilon$ whenever $k \geq N$. Hence $\|\mathbf{x}_k\|_2 \leq \|\mathbf{x}_k - \mathbf{x}\|_2 + \|\mathbf{x}\|_2 < \epsilon + \|\mathbf{x}\|_2$. Let $M = \max\{\|\mathbf{x}_1\|, \|\mathbf{x}_2\|, \ldots, \|\mathbf{x}_{N-1}\|, \epsilon + \|\mathbf{x}\|_2\}$. Then $\|\mathbf{x}_n\|_2 \leq M$, which implies that $\{\mathbf{x}_n\}$ is bounded.
   Let $\{\mathbf{x}_{n_k}\}_k$ be an arbitrary subsequence of $\{\mathbf{x}_n\}$. Since $n_k \geq k$, we have $\|\mathbf{x}_{n_k} - \mathbf{x}\|_2 < \epsilon$ whenever $k \geq N$, i.e. $\lim_{k\to\infty} \mathbf{x}_{n_k} = \mathbf{x}$. Clearly $\mathbf{x}_{n_k} \neq \mathbf{x}$, so $\mathbf{x}$ is a unique limit point of $\{\mathbf{x}_n\}$.

   ($\Leftarrow$) Suppose that $\{\mathbf{x}_n\}$ is bounded and has a unique limit point $\mathbf{x}$. Assume that $\mathbf{x}_n \not\to \mathbf{x}$. Then there exists $\delta > 0$ such that given any positive integer N, we can find $k \geq N$ satisfying $\|\mathbf{x}_k - \mathbf{x}\|_2 \geq \delta$. Let $n_1 = 1$ and, for any n_k, find $n_{k+1} \geq n_k + 1$ such that $\|\mathbf{x}_{n_{k+1}} - \mathbf{x}\|_2 \geq \delta$. By Bolzano-Weierstrass theorem, the bounded sequence $\{\mathbf{x}_{n_k}\}$ has a convergent subsequence $\{\mathbf{x}_{m_k}\}$. Since $\mathbf{x}$ is the unique limit point, $\{\mathbf{x}_{m_k}\}$ must converge to $\mathbf{x}$, which contradicts $\|\mathbf{x}_{m_k} - \mathbf{x}\|_2 \geq \delta$. Therefore, $\mathbf{x}_n \to \mathbf{x}$. ■

2. (**Limit Points of a Set**). Let C be a subset of $\mathbb{R}^n$. A point $\mathbf{x} \in \mathbb{R}^n$ is called a limit point of C if there is a sequence $\{\mathbf{x}_n\}$ in C such that $\mathbf{x}_n \to \mathbf{x}$ and $\mathbf{x}_n \neq \mathbf{x}$ for all positive integers n. If $\mathbf{x} \in C$ and $\mathbf{x}$ is not a limit point of C, then $\mathbf{x}$ is called an isolated point of C. Let C' be the set of limit points of the set C. Please show the following statements.

   (a) If $C = (0,1) \cup \{2\} \subset \mathbb{R}$, then $C' = [0,1]$ and $x = 2$ is an isolated point of C.

   (b) The set C' is closed.

   (c) The closure of C is the union of C' and C; that is $\mathbf{cl}\ C = C' \cup C$. Moreover, $C' \subset C$ if and only if C is closed.

   Solution:

   (a) For $x \in (0,1]$, the sequence $\{(1-2^{-n})x\} \subset C$ converges to x. For $x = 0$, $\{2^{-n}\} \subset C$ converges to x. Hence $C' \supset [0,1]$.
   Suppose $\{x_n\}$ converges to some $x \notin [0,1]$. For $x \in (-\infty, 0)$, the neighborhood $(-2x, 0)$ intersects $\{x_n\}$ but does not intersects C. For $x \in (1, \infty)$, the neighborhood $(1,x) \cap (x, 2x-1)$ intersects $\{x_n\}$ but does not intersects C. Hence $\{x_n\} \not\subset C \implies C' \subset [0,1] \implies C' = [0,1]$.
   Specifically, $x = 2 \in C \setminus C'$ is an isolated point.

   (b) Let $\mathbf{x}$ be a limit point of C'. From a sequence $\{\mathbf{x}_n\} \subset C'$ converging to $\mathbf{x}$, pick $\mathbf{x}_{n_k}$ satisfying $\|\mathbf{x}_{n_k} - \mathbf{x}\|_2 < \frac{1}{k}$, where k is a positive integer. Since $\mathbf{x}_{n_k}$ is a limit point of C, we can similarly find some $\mathbf{y}_k \in C$ satisfying $\|\mathbf{y}_k - \mathbf{x}_{n_k}\|_2 < \frac{1}{k}$ and $\mathbf{y}_k \neq \mathbf{x}$. Then we have $\|\mathbf{y}_k - \mathbf{x}\|_2 \le \|\mathbf{y}_k - \mathbf{x}_{n_k}\|_2 + \|\mathbf{x}_{n_k} - \mathbf{x}\|_2 < \frac{2}{k}$, which implies that $\lim_{k\to\infty} \mathbf{y}_k = \mathbf{x}$. Hence $\mathbf{x}$ is a limit point of $C \implies \mathbf{x} \in C'$. Since $\mathbf{x}$ is arbitrary, we know C' contains all of its limit points. According to (c), C' is closed.

(c) ($\Rightarrow$) Assume that $C' \subset C$ but C is not closed. Let $\mathbf{x}$ be a point in $\mathbb{R}^n \setminus C$. Since $\mathbb{R}^n \setminus C$ is not open, for any $\epsilon > 0$, the ϵ-neighborhood $N_\epsilon(\mathbf{x})$ must intersect C. For each positive integer n, we can let $\epsilon = \frac{1}{n}$ and find some $\mathbf{x}_n \in N_\epsilon(\mathbf{x}) \cap C$. Clearly $\mathbf{x}_n \neq \mathbf{x}$ and $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$, so $\mathbf{x} \in C'$. But $\mathbf{x} \notin C$, which is a contradiction.

($\Leftarrow$) If $\mathbf{x}$ is a limit point of C, then there exists $\{\mathbf{x}_n\} \subset C$ converging to $\mathbf{x}$. Given any $N_\epsilon(\mathbf{x})$ with $\epsilon > 0$, we can always find some $\mathbf{x}_n$ in $N_\epsilon(\mathbf{x})$, which implies that $N_\epsilon(\mathbf{x})$ intersects C, i.e. $N_\epsilon(\mathbf{x}) \not\subset \mathbb{R}^n \setminus C$. Since $\mathbb{R}^n \setminus C$ is open, $\mathbf{x}$ must be in C. Hence $C' \subset C$. ■

Exercise 2: Open and Closed Sets

The norm ball $\{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y}-\mathbf{x}\|_2 < r, \mathbf{x} \in \mathbb{R}^n\}$ is denoted by $B_r(\mathbf{x})$.

1. Given a set $C \subset \mathbb{R}^n$, please show the following are equivalent.

   (a) The set C is closed; that is $\mathbf{cl}\ C = C$.

   (b) The complement of C is open.

   (c) If $B_\epsilon(\mathbf{x}) \cap C \neq \emptyset$ for every $\epsilon > 0$, then $\mathbf{x} \in C$.

   Solution:

(a) $\Rightarrow$ (b) Assume that $\mathbb{R}^n \setminus C$ is not open. Let $\mathbf{x}$ be a point in $\mathbb{R}^n \setminus C$. For each positive integer n, we have $B_{\frac{1}{n}}(\mathbf{x}) \not\subset \mathbb{R}^n \setminus C$, so there exists $\mathbf{x}_n \in B_{\frac{1}{n}}(\mathbf{x}) \cap C$. Clearly $\mathbf{x}_n \neq \mathbf{x}$ and $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$, so $\mathbf{x}$ is a limit point of C, i.e. $\mathbf{x} \in \mathbf{cl}\ C = C$. But $\mathbf{x} \in \mathbb{R}^n \setminus C$, which is a contradiction.

(b) $\Rightarrow$ (c) If $B_\epsilon(\mathbf{x}) \cap C \neq \emptyset \implies B_\epsilon(\mathbf{x}) \not\subset \mathbb{R}^n \setminus C$ for every $\epsilon > 0$, then $\mathbf{x}$ is not an interior point of $\mathbb{R}^n \setminus C$. Since $\mathbb{R}^n \setminus C$ is open, $\mathbf{x}$ must be in C.

(c) $\Rightarrow$ (a) Suppose $\mathbf{x}$ is a limit point of C. Then there exists $\{\mathbf{x}_n\} \subset C\setminus\{\mathbf{x}\}$ converging to $\mathbf{x}$. Given any $\epsilon > 0$, we can always find some $\mathbf{x}_n \in B_\epsilon(\mathbf{x}) \implies B_\epsilon(\mathbf{x}) \cap C \neq \emptyset \implies \mathbf{x} \in C$. Since $\mathbf{x}$ is arbitrary, it follows that C contains all of its limit points, i.e. $\mathbf{cl}\ C = C$. ■

2. Given $A \subset \mathbb{R}^n$, a set $C \subset A$ is called open in A if

$$C = \{\mathbf{x} \in C : B_\epsilon(\mathbf{x}) \cap A \subset C \text{ for some } \epsilon > 0\}.$$

   A set C is said to be closed in A if $A \setminus C$ is open in A.

   (a) Let $B = [0,1] \cup \{2\}$. Please show that $[0,1]$ is not an open set in $\mathbb{R}$, while it is both open and closed in B.

   (b) Please show that a set $C \subset A$ is open in A if and only if $C = A \cap U$, where U is open in $\mathbb{R}^n$.

   Solution:

   (a) For every $\epsilon > 0$, we have $(\epsilon, 0] \subset B_\epsilon(0)$ and $(\epsilon, 0] \not\subset [0,1]$. Hence $B_\epsilon(0) \cap \mathbb{R} \not\subset [0,1]$, i.e. $x = 0$ is not an interior point of $[0,1]$. Therefore, $[0,1]$ is not an open set in $\mathbb{R}$.
   For every $x \in [0,1]$, $B_1(x) \cap B \subset (-1,2) \cap B = [0,1]$, so $[0,1]$ is open in B.
   For $x \in B \setminus [0,1] = \{2\}$, $B_1(x) \cap B = \{2\}$, so $\{2\}$ is open in B, i.e. $[0,1]$ is closed in B.

   (b) ($\Rightarrow$) For every $\mathbf{x} \in C$, there exists $\epsilon_\mathbf{x} > 0$ such that $B_{\epsilon_\mathbf{x}}(\mathbf{x}) \cap A \subset C$. Let $U = \bigcup_{\mathbf{x}\in C} B_{\epsilon_\mathbf{x}}(\mathbf{x})$. On the one hand, since $\mathbf{x} \in B_{\epsilon_\mathbf{x}}(\mathbf{x})$, it follows that $C \subset U \implies C \subset U \cap A$. On the other hand, $U \cap A = \bigcup_{\mathbf{x}\in C}(B_{\epsilon_\mathbf{x}}(\mathbf{x}) \cap A) \subset C$. Hence $U \cap A = C$. Since every $B_{\epsilon_\mathbf{x}}(\mathbf{x})$ is open in $\mathbb{R}^n$, U is also open in $\mathbb{R}^n$.

   ($\Leftarrow$) For every $\mathbf{x} \in C \subset U$, there exists $\epsilon > 0$ such that $B_\epsilon(\mathbf{x}) \subset U \implies B_\epsilon(\mathbf{x}) \cap A \subset (U \cap A) = C$. Hence C is open in A. ■

Exercise 3: Bolzano-Weierstrass Theorem

The Least Upper Bound Axiom

Any nonempty set of real numbers with an upper bound has a least upper bound. That is, $\sup C$ *always exists for a nonempty bounded above set* $C \subset \mathbb{R}$.

Please show the following statements from **the least upper bound axiom**.

1. Let C be a nonempty subset of $\mathbb{R}$ that is bounded above. Prove that $u = \sup C$ if and only if u is an upper bound of C and

$$\forall\, \epsilon > 0, \exists\, a \in C \text{ such that } a > u - \epsilon.$$

Solution:

($\Rightarrow$) Since u is the least upper bound, for any $\epsilon > 0$, $u - \epsilon$ cannot be an upper bound of C. Hence, there must be some $a \in C$ such that $a > u - \epsilon$.

($\Leftarrow$) For any $u' < u$, there exists some $a \in C$ such that $a > u'$, implying that u' is not an upper bound of C. Hence, u is the least upper bound. ■

2. Every bounded sequence in $\mathbb{R}$ has at least one limit point.

Solution:
Suppose $\{x_n\}$ to be a bounded sequence in $\mathbb{R}$. Let $C = \{x : x < x_n$ for an infinite number of $x_n\}$. C is nonempty because it contains all lower bounds of x_n. Moreover, C is bounded above by every upper bound of x_n. By the least upper bound axiom, $u = \sup C$ must exist. Then we can find $a \in C$ satisfying $a > u - \frac{1}{k}$ for each positive integer k. That is, there exist infinitely many x_n greater than $u - \frac{1}{k}$. Furthermore, since $u + \frac{1}{k} \notin C$, only a finite number of x_n is greater than $u + \frac{1}{k}$. From those infinitely many x_n who satisfy $|x_n - u| < \frac{1}{k}$, pick x_{n_k} such that $x_{n_k} \neq u$. Since $\lim_{k\to\infty} x_{n_k} = u$, we know u is a limit point of $\{x_n\}$. ■

3. Every bounded sequence in $\mathbb{R}^n$ has at least one limit point.

Solution:
Suppose $\{\mathbf{x}^m\}$ to be a bounded sequence in $\mathbb{R}^n$ and $\mathbf{x}^m = (x_1^m, x_2^m, \ldots, x_n^m)$. Then $\{x_1^m\}$ is a bounded sequence in $\mathbb{R}$ with at least one limit point x_1, which implies that it has a subsequence $\{x_1^{m_{1,k}}\}_k$ converging to x_1. For each $d = 2, 3, \ldots, n$, let x_d be a limit point of the bounded sequence $\{x_d^{m_{d-1,k}}\}_k$. Then there exists a subsequence $\{x_d^{m_{d,k}}\}_k \subset \{x_d^{m_{d-1,k}}\}_k$ converging to x_d. Now, consider the components of $\{\mathbf{x}^{m_{n,k}}\}_k$. Since $\{x_1^{m_{n,k}}\}_k \subset \{x_1^{m_{1,k}}\}_k$, $\{x_1^{m_{n,k}}\}_k$ must converge to x_1. Analogously, $\{x_d^{m_{n,k}}\}_k$ must converge to x_d for each $d = 2, 3, \ldots, n-1$. If we denote $(x_1, x_2, \ldots, x_n)$ by $\mathbf{x}$, then $\lim_{k\to\infty} \mathbf{x}^{m_{n,k}} = \mathbf{x}$. Clearly $\mathbf{x}^{m_{n,k}} \neq \mathbf{x}$, so $\{\mathbf{x}^m\}$ has a limit point $\mathbf{x}$. ■

Exercise 4: Extreme Value Theorem

1. Show that a set $C \subset \mathbb{R}^n$ is compact if and only if C is closed and bounded.

   Solution:

   ($\Rightarrow$) Let C be compact. Assume that C is not bounded. Then, given any positive integer n, there must exist some $\mathbf{x}_n \in C$ such that $\|\mathbf{x}_n\|_2 > n$. Since C is compact, $\{\mathbf{x}_n\}$ has a convergent subsequence $\{\mathbf{x}_{n_k}\}$. According to Exercise 1.1, $\{\mathbf{x}_{n_k}\}$ must be bounded, but $\|\mathbf{x}_{n_k}\|_2 > n_k$, which is a contradiction. Therefore, C is bounded.
   Let $\mathbf{x}$ be a limit point of C. Then there is a sequence $\{\mathbf{x}_n\} \subset C \setminus \{\mathbf{x}\}$ that converges to $\mathbf{x}$. Since C is compact, $\{\mathbf{x}_n\}$ has a convergent subsequence $\{\mathbf{x}_{n_k}\}$ and $\lim_{k\to\infty} \mathbf{x}_{n_k} \in C$. According to Exercise 1.1, we have $\lim_{k\to\infty} \mathbf{x}_{n_k} = \mathbf{x} \implies \mathbf{x} \in C$. Since $\mathbf{x}$ is arbitrary, we know C contains all of its limit points. Therefore, C is closed.

   ($\Leftarrow$) Let $C \subset \mathbb{R}^n$ be closed and bounded. Then any sequence $\{\mathbf{x}_n\} \subset C$ is also bounded. By Bolzano-Weierstrass theorem, we can find a convergent subsequence $\{\mathbf{x}_{n_k}\}$ whose limit $\mathbf{x}$ is also a limit point of C. Because C is closed, it follows that $\mathbf{x} \in C$. Hence C is compact. ■

2. Let C be a compact subset of $\mathbb{R}^n$ and $f : C \to \mathbb{R}$ be continuous. Please show that there exist $\mathbf{a}, \mathbf{b} \in C$ such that

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

   (**Hint:** first prove that $f(C)$ is compact, in $\mathbb{R}$.)

   Solution:
   Consider an arbitrary sequence $\{y_n\} \subset f(C)$. For each positive integer n, there exists $\mathbf{x}_n \in C$ such that $f(\mathbf{x}_n) = y_n$. Since C is compact, $\{\mathbf{x}_n\} \subset C$ has a convergent subsequence $\{\mathbf{x}_{n_k}\}$ satisfying $\lim_{k\to\infty} \mathbf{x}_{n_k} \in C$. Denote $\lim_{k\to\infty} \mathbf{x}_{n_k}$ by $\mathbf{x}$. Because f is continuous, $\mathbf{x}_{n_k} \to \mathbf{x} \implies \mathbf{y}_{n_k} \to f(\mathbf{x})$. Moreover, $\mathbf{x} \in C \implies f(\mathbf{x}) \in f(C)$. Therefore, $f(C)$ is compact in $\mathbb{R}$, i.e. bounded and closed.
   Since $f(C)$ is bounded, we can find $u = \sup f(C)$ and $l = \inf f(C)$. For each positive n, there exists $u_n \in f(C)$ such that $u - \frac{1}{n} < u_n \leq u$, which implies that $\lim_{n\to\infty} u_n = u$. If $u_n = u$ for some n, then $u \in f(C)$. Otherwise u is a limit point of $f(C)$, and since $f(C)$ is closed, it still follows that $u \in f(C)$. Analogously, we can conclude that $l \in f(C)$. Hence, there exists $\mathbf{a}, \mathbf{b} \in C$ such that $f(\mathbf{a}) = l$ and $f(\mathbf{b}) = u$, which leads to the desired statement. ■

3. Let $f : [a, b] \to \mathbb{R}$ be continuous. Show that the range of f is a compact interval $[c, d]$ for some $c, d \in \mathbb{R}$.

   Solution:
   Clearly, $[a, b]$ is a compact in $\mathbb{R}$. By Exercise 4.2, we know that $f([a, b])$ is also compact in $\mathbb{R}$ and $c = \min f([a, b])$, $d = \max f([a, b])$ both exist. For any $y \in (c, d)$, we define $C = \{x \in [a, b] : f(x) \leq y\}$. Since C is nonempty and bounded above, $u = \sup C$ must exist.
   Assume that $f(u) > y$. Since f is continuous, there exists $\epsilon > 0$ such that $f(x) > y$ for all $x \in (u - \epsilon, u + \epsilon)$. Then $u - \epsilon$ is an upper bound of C less than u, which is a contradiction.
   Assume that $f(u) < y$. Since f is continuous, there exists $\epsilon > 0$ such that $f(x) < y$ for all $x \in (u - \epsilon, u + \epsilon)$. Then $u + \frac{\epsilon}{2}$ is an element of C greater than u, which is a contradiction.
   Hence $f(u) = y$, from which we conclude that $f([a, b]) \supset [c, d]$, i.e. $f([a, b]) = [c, d]$. ■

Exercise 5: Basis and Coordinates

Suppose that $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ is a basis of an n-dimensional vector space V.

1. Show that $\{\lambda_1\mathbf{a}_1, \lambda_2\mathbf{a}_2, \ldots, \lambda_n\mathbf{a}_n\}$ is also a basis of V for nonzero scalars $\lambda_1, \lambda_2, \ldots, \lambda_n$.

   Solution:
   Because $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ is a basis of V, for every $\mathbf{v} \in V$, there exists $x_1, x_2, \ldots, x_n$ such that $\mathbf{v} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n$. Let $y_i = \frac{x_i}{\lambda_i}$, then $\mathbf{v} = y_1\lambda_1\mathbf{a}_1 + y_2\lambda_2\mathbf{a}_2 + \cdots + y_n\lambda_n\mathbf{a}_n$. Therefore, $\{\lambda_1\mathbf{a}_1, \lambda_2\mathbf{a}_2, \ldots, \lambda_n\mathbf{a}_n\}$ is a basis of V. ■

2. Let $V = \mathbb{R}^n$ and $(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n) = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)\mathbf{P}$, where $\mathbf{P} \in \mathbb{R}^{n\times n}$ and $\mathbf{b}_i \in \mathbb{R}^n$, for any $i \in \{1, \ldots, n\}$. Show that $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ is also a basis of V for any invertible matrix $\mathbf{P}$.

   Solution:
   Let $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ and $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n)$, then $\mathbf{B} = \mathbf{AP}$. Because $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ is a basis of V, for every $\mathbf{v} \in V$, there exists $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$ such that $\mathbf{v} = \mathbf{Ax}$. Let $\mathbf{y} = \mathbf{P}^{-1}\mathbf{x}$, then $\mathbf{v} = \mathbf{By}$. Therefore, $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ is a basis of V. ■

3. Suppose that the coordinate of a vector $\mathbf{v}$ under the basis $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ is $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

   (a) What is the coordinate of $\mathbf{v}$ under $\{\lambda_1\mathbf{a}_1, \lambda_2\mathbf{a}_2, \ldots, \lambda_n\mathbf{a}_n\}$?

   (b) What are the coordinates of $\mathbf{w} = \mathbf{a}_1 + \cdots + \mathbf{a}_n$ under $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ and $\{\lambda_1\mathbf{a}_1, \lambda_2\mathbf{a}_2, \ldots, \lambda_n\mathbf{a}_n\}$? Note that $\lambda_i \neq 0$ for any $i \in \{1, \ldots, n\}$.

   Solution:

   (a) By Exercise 5.1, the coordinate under $\{\lambda_1\mathbf{a}_1, \lambda_2\mathbf{a}_2, \ldots, \lambda_n\mathbf{a}_n\}$ is $(\frac{x_1}{\lambda_1}, \frac{x_2}{\lambda_2}, \ldots, \frac{x_n}{\lambda_n})$.

   (b) The coordinate under $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ is $(1, 1, \ldots, 1)$.
   The coordinate under $\{\lambda_1\mathbf{a}_1, \lambda_2\mathbf{a}_2, \ldots, \lambda_n\mathbf{a}_n\}$ is $(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_n})$. ■

Exercise 6: Derivatives with Matrices

Definition 1 (Differentiability)**.** [1] Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a function, $\mathbf{x}_0 \in \mathbb{R}^n$ be a point, and let $L : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. We say that f is *differentiable at* $\mathbf{x}_0$ *with derivative* L if we have

$$\lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \frac{\|f(\mathbf{x}) - f(\mathbf{x}_0) - L(\mathbf{x}-\mathbf{x}_0)\|_2}{\|\mathbf{x}-\mathbf{x}_0\|_2} = 0.$$

We denote this derivative by $f'(\mathbf{x}_0)$.

1. Let $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Consider the functions as follows. Please show that they are differentiable and find $f'(\mathbf{x})$.

   (a) $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$.

   (b) $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{x}$.

   (c) $f(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$.

   Solution:

   (a)
$$\begin{aligned}
& \lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \frac{\mathbf{a}^\top\mathbf{x} - \mathbf{a}^\top\mathbf{x}_0 - \mathbf{a}^\top(\mathbf{x}-\mathbf{x}_0)}{\|\mathbf{x}-\mathbf{x}_0\|_2} \\
= & \lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \frac{0}{\|\mathbf{x}-\mathbf{x}_0\|_2} \quad = \quad 0 \quad \Longrightarrow f'(\mathbf{x}) \equiv \mathbf{a}^\top.
\end{aligned}$$

   (b)
$$\begin{aligned}
& \lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \frac{\mathbf{x}^\top\mathbf{x} - \mathbf{x}_0^\top\mathbf{x}_0 - 2\mathbf{x}_0^\top(\mathbf{x}-\mathbf{x}_0)}{\|\mathbf{x}-\mathbf{x}_0\|_2} \\
= & \lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \frac{(\mathbf{x}-\mathbf{x}_0)^\top(\mathbf{x}-\mathbf{x}_0)}{\|\mathbf{x}-\mathbf{x}_0\|_2} \\
= & \lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \|\mathbf{x}-\mathbf{x}_0\|_2 \quad = \quad 0 \quad \Longrightarrow f'(\mathbf{x}) = 2\mathbf{x}^\top.
\end{aligned}$$

   (c)
$$\begin{aligned}
f(\mathbf{x}) &= \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \\
&= (\mathbf{y} - \mathbf{A}\mathbf{x})^\top(\mathbf{y} - \mathbf{A}\mathbf{x}) \\
&= \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x}.
\end{aligned}$$
$$\begin{aligned}
& \lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \frac{f(\mathbf{x}) - f(\mathbf{x}_0) - (-2\mathbf{y}^\top\mathbf{A} + 2\mathbf{x}_0^\top\mathbf{A}^\top\mathbf{A})(\mathbf{x}-\mathbf{x}_0)}{\|\mathbf{x}-\mathbf{x}_0\|_2} \\
= & \lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \frac{\mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{x}_0^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} + \mathbf{x}_0^\top\mathbf{A}^\top\mathbf{A}\mathbf{x}_0^\top}{\|\mathbf{x}-\mathbf{x}_0\|_2} \\
= & \lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \frac{(\mathbf{x}-\mathbf{x}_0)^\top\mathbf{A}^\top\mathbf{A}(\mathbf{x}-\mathbf{x}_0)}{\|\mathbf{x}-\mathbf{x}_0\|_2} \\
\leq & \lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \|\mathbf{A}^\top\mathbf{A}\|_2 \cdot \|\mathbf{x}-\mathbf{x}_0\|_2 \ = \ 0 \ \Longrightarrow f'(\mathbf{x}) = -2\mathbf{y}^\top\mathbf{A} + 2\mathbf{x}^\top\mathbf{A}^\top\mathbf{A}.
\end{aligned}$$
■

2. Please follow Definition 1 and give the definition of the differentiability of the functions $f : \mathbb{R}^{n\times n} \to \mathbb{R}$.

Solution:
Let $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ be a function, $\mathbf{X}_0 \in \mathbb{R}^{n\times n}$ be a matrix, and let $L : \mathbb{R}^{n\times n} \to \mathbb{R}$ be a linear transformation. We say that f is *differentiable at* $\mathbf{X}_0$ *with derivative* L if we have

$$\lim_{\mathbf{X}\to\mathbf{X}_0;\mathbf{X}\neq\mathbf{X}_0} \frac{f(\mathbf{X}) - f(\mathbf{X}_0) - L(\mathbf{X}-\mathbf{X}_0)}{\|\mathbf{X}-\mathbf{X}_0\|_F} = 0.$$

We denote this derivative by $f'(\mathbf{X}_0)$. ■

3. Let $f(\mathbf{X}) = \det(\mathbf{X})$, where $\det(\mathbf{X})$ is the determinant of $\mathbf{X} \in \mathbb{R}^{n\times n}$. Please discuss the differentiability of f rigorously according to your definition in the last part. If f is differentiable, please find $f'(\mathbf{X})$.

Solution:

$$\begin{aligned}
& \det(\mathbf{X}+\Delta\mathbf{X}) \\
= \; & \sum_{\sigma\in S_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} (x_{i,\sigma_i} + \Delta x_{i,\sigma_i}) \right) \\
= \; & \sum_{\sigma\in S_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} x_{i,\sigma_i} \right) + \sum_{i=1}^{n} \sum_{\sigma\in S_n} \left(\operatorname{sgn}(\sigma) \Delta x_{i,\sigma_i} \prod_{j\neq i} x_{j,\sigma_j} \right) + O(\|\Delta\mathbf{X}\|_F^2) \\
= \; & \det(\mathbf{X}+\Delta\mathbf{X}) + \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{cof}(\mathbf{X})_{ij} \Delta x_{ij} + O(\|\Delta\mathbf{X}\|_F^2) \\
= \; & \det(\mathbf{X}+\Delta\mathbf{X}) + \operatorname{tr}(\operatorname{adj}(\mathbf{X})\Delta\mathbf{X}) + o(\|\Delta\mathbf{X}\|_F).
\end{aligned}$$

$$\begin{aligned}
\therefore \; & \lim_{\Delta\mathbf{X}\to\mathbf{O}} \frac{\det(\mathbf{X}+\Delta\mathbf{X}) - \det(\mathbf{X}) - \operatorname{tr}(\operatorname{adj}(\mathbf{X})\Delta\mathbf{X})}{\|\Delta\mathbf{X}\|_F} \\
= \; & \lim_{\Delta\mathbf{X}\to\mathbf{O}} \frac{o(\|\Delta\mathbf{X}\|_F)}{\|\Delta\mathbf{X}\|_F} \quad = \quad 0 \quad \Longrightarrow \; f'(\mathbf{X}) : \boldsymbol{\Xi} \mapsto \operatorname{tr}(\operatorname{adj}(\mathbf{X})\boldsymbol{\Xi}).
\end{aligned}$$ ■

4. Let $f(\mathbf{X}) = \operatorname{tr}(\mathbf{A}^\top\mathbf{X})$, where $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{n\times m}$, and $\operatorname{tr}(\cdot)$ denotes the trace of a matrix. Please discuss the differentiability of f and find f' if it is differentiable.

Solution:

$$\begin{aligned}
& \lim_{\mathbf{X}\to\mathbf{X}_0;\mathbf{X}\neq\mathbf{X}_0} \frac{\operatorname{tr}(\mathbf{A}^\top\mathbf{X}) - \operatorname{tr}(\mathbf{A}^\top\mathbf{X}_0) - \operatorname{tr}(\mathbf{A}^\top(\mathbf{X}-\mathbf{X}_0))}{\|\mathbf{X}-\mathbf{X}_0\|_F} \\
= \; & \lim_{\mathbf{X}\to\mathbf{X}_0;\mathbf{X}\neq\mathbf{X}_0} \frac{0}{\|\mathbf{X}-\mathbf{X}_0\|_F} \quad = \quad 0 \quad \Longrightarrow \; f'(\mathbf{X}) : \boldsymbol{\Xi} \mapsto \operatorname{tr}(\mathbf{A}^\top\boldsymbol{\Xi}).
\end{aligned}$$ ■

5. Let $\mathbf{S}_{++}^n$ be the space of all positive definite $n \times n$ matrices. Prove the function $f : \mathbf{S}_{++}^n \to \mathbb{R}$ defined by $f(\mathbf{X}) = \operatorname{tr}\mathbf{X}^{-1}$ is differentiable on $\mathbf{S}_{++}^n$. (Hint: Expand the expression $(\mathbf{X}+t\mathbf{Y})^{-1}$ as a power series.)

Solution:

$$\frac{\mathrm{d}\mathbf{X}^{-1}}{\mathrm{d}t}\mathbf{X} + \mathbf{X}^{-1}\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = \frac{\mathrm{d}(\mathbf{X}^{-1}\mathbf{X})}{\mathrm{d}t} = \frac{\mathrm{d}\mathbf{I}}{\mathrm{d}t} = \mathbf{O} \implies \frac{\mathrm{d}\mathbf{X}^{-1}}{\mathrm{d}t} = -\mathbf{X}^{-1}\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t}\mathbf{X}^{-1}.$$

$$\therefore \quad (\mathbf{X}+t\mathbf{Y})^{-1} = \mathbf{X}^{-1} - \mathbf{X}^{-1}(t\mathbf{Y})\mathbf{X}^{-1} + O(t^2).$$

$$\therefore \quad \lim_{t\to 0} \frac{\operatorname{tr}(\mathbf{X}+t\mathbf{Y}) - \operatorname{tr}(\mathbf{X}) + \operatorname{tr}(\mathbf{X}^{-1}(t\mathbf{Y})\mathbf{X}^{-1})}{\|t\mathbf{Y}\|_F}$$

$$= \quad \lim_{t\to 0} \frac{o(t)}{O(t)} \quad = \quad 0 \quad \Longrightarrow f'(\mathbf{X}) : \mathbf{\Xi} \mapsto \operatorname{tr}(-\mathbf{X}^{-1}\mathbf{\Xi}\mathbf{X}^{-1}).$$ ■

6. Define a function $f : \mathbf{S}_{++}^n \to \mathbb{R}$ by $f(\mathbf{X}) = \log\det \mathbf{X}$. Prove $\nabla f(\mathbf{I}) = \mathbf{I}$. Deduce $\nabla f(\mathbf{X}) = \mathbf{X}^{-1}$ for any $\mathbf{X}$ in $\mathbf{S}_{++}^n$.

   Solution:

$$\frac{\partial \det(\mathbf{X})}{\partial x_{ij}} = \sum_{k=1}^{n} \left(\frac{\partial x_{ik}}{\partial x_{ij}} \operatorname{cof}(\mathbf{X})_{ik} + x_{ik} \frac{\partial \operatorname{cof}(\mathbf{X})_{ik}}{\partial x_{ij}} \right) = \operatorname{cof}(\mathbf{X})_{ij}$$

$$\Longrightarrow \frac{\partial f(\mathbf{X})}{\partial x_{ij}} = \frac{\mathrm{d} f(\mathbf{X})}{\mathrm{d}\det(\mathbf{X})} \frac{\partial \det(\mathbf{X})}{\partial x_{ij}} = \frac{\operatorname{cof}(\mathbf{X})_{ij}}{\det(\mathbf{X})}$$

$$\Longrightarrow \nabla f(\mathbf{X}) = \frac{\operatorname{cof}(\mathbf{X})}{\det(\mathbf{X})} = \mathbf{X}^{-\top}.$$

   Let $\mathbf{X} = \mathbf{I} \implies \nabla f(\mathbf{I}) = \mathbf{I}$. For any $\mathbf{X} \in \mathbf{S}_{++}^n$, $\mathbf{X} = \mathbf{X}^\top \implies \nabla f(\mathbf{X}) = \mathbf{X}^{-1}$. ■

Exercise 7: Rank of Matrices

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$.

1. Please show that

   (a) $\mathbf{rank}(\mathbf{A}) = \mathbf{rank}\left(\mathbf{A}^\top\right)$;

   (b) $\mathbf{rank}(\mathbf{AB}) \le \mathbf{rank}(\mathbf{A})$;

   (c) $\mathbf{rank}(\mathbf{AB}) \le \mathbf{rank}(\mathbf{B})$;

   (d) $\mathbf{rank}(\mathbf{A}) = \mathbf{rank}\left(\mathbf{A}^\top\mathbf{A}\right)$.

   Solution:

   (a) Suppose $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_r\}$ is a basis of $\mathcal{C}(\mathbf{A})$, where $r = \mathbf{rank}(\mathbf{A})$. Let $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_r) \in \mathbb{R}^{m\times r}$. Since all n columns of $\mathbf{A}$ are linear combinations of $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_r\}$, there exists a matrix $\mathbf{R} \in \mathbb{R}^{r\times n}$ such that $\mathbf{A} = \mathbf{CR}$. Then each row of $\mathbf{A}$ can be written as a linear combination of rows of $\mathbf{R}$, implying that $\mathbf{rank}\left(\mathbf{A}^\top\right) \le \mathbf{rank}(\mathbf{R}) \le r$. Moreover, $\mathbf{rank}(\mathbf{A}) = \mathbf{rank}\left((\mathbf{A}^\top)^\top\right) \le \mathbf{rank}\left(\mathbf{A}^\top\right)$. Therefore, $\mathbf{rank}(\mathbf{A}) = \mathbf{rank}\left(\mathbf{A}^\top\right)$.

   (b) Columns of $\mathbf{AB}$ are linear combinations of columns of $\mathbf{A}$ $\implies$ $\mathcal{C}(\mathbf{AB}) \subset \mathcal{C}(\mathbf{A})$ $\implies$ $\mathbf{rank}(\mathbf{AB}) = \dim(\mathcal{C}(\mathbf{AB})) \le \dim(\mathcal{C}(\mathbf{A})) = \mathbf{rank}(\mathbf{A})$.

   (c) Rows of $\mathbf{AB}$ are linear combinations of rows of $\mathbf{B}$ $\implies$ $\mathcal{C}(\mathbf{B}^\top\mathbf{A}^\top) \subset \mathcal{C}(\mathbf{B}^\top)$ $\implies$ $\mathbf{rank}(\mathbf{AB}) = \dim(\mathcal{C}(\mathbf{B}^\top\mathbf{A}^\top)) \le \dim(\mathcal{C}(\mathbf{B}^\top)) = \mathbf{rank}(\mathbf{B})$.

   (d) According to (a), $\mathbf{A} = \mathbf{CR}$, where $\mathbf{C}$ has full column rank and $\mathbf{R}$ has full row rank. If $\mathbf{C}^\top\mathbf{Cx} = \mathbf{0}$ for some $\mathbf{x} \in \mathbb{R}^r$, then $\mathbf{x}^\top\mathbf{C}^\top\mathbf{Cx} = \mathbf{0} \implies \mathbf{Cx} = \mathbf{0} \implies \mathbf{x} = \mathbf{0}$. Hence $\mathbf{C}^\top\mathbf{C}$ also has full rank. Analogously, we show that $\mathbf{RR}^\top$ has full rank. Then $\mathbf{A} = \mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1}(\mathbf{RR}^\top)^{-1}\mathbf{RA}^\top\mathbf{A}$, which implies that $\mathbf{rank}(\mathbf{A}) \le \mathbf{rank}\left(\mathbf{A}^\top\mathbf{A}\right)$. On the other hand, $\mathbf{rank}\left(\mathbf{A}^\top\mathbf{A}\right) \le \mathbf{rank}(\mathbf{A})$, so $\mathbf{rank}(\mathbf{A}) = \mathbf{rank}\left(\mathbf{A}^\top\mathbf{A}\right)$. ■

2. The *column space* of $\mathbf{A}$ is defined by

   $$\mathcal{C}(\mathbf{A}) = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = \mathbf{Ax},\, \mathbf{x} \in \mathbb{R}^n\}.$$

   The *null space* of $\mathbf{A}$ is defined by

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

   Notice that, the rank of $\mathbf{A}$ is the dimension of the column space of $\mathbf{A}$.

   Please show that

   (a) $\mathbf{rank}(\mathbf{A}) + \dim(\mathcal{N}(\mathbf{A})) = n$;

   (b) $\mathbf{y} = \mathbf{0}$ if and only if $\mathbf{a}_i^\top\mathbf{y} = 0$ for $i = 1, \ldots, m$, where $\mathbf{y} \in \mathbb{R}^m$ and $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}$ is a basis of $\mathbb{R}^m$.

   Solution:

   (a) Let $r = \mathbf{rank}(\mathbf{A})$. There exists a permutation matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ such that $\mathbf{AP} = (\mathbf{C} \quad \mathbf{CR})$, where $\mathbf{C} \in \mathbb{R}^{m\times r}$ is r linearly independent columns of $\mathbf{A}$ and $\mathbf{CR} \in \mathbb{R}^{m\times(n-r)}$ is the other columns.

Let $\mathbf{X} = \mathbf{P}\begin{pmatrix} -\mathbf{R} \\ \mathbf{I}_{n-r} \end{pmatrix} \in \mathbb{R}^{n\times(n-r)}$. Then $\mathbf{AX} = (\mathbf{C} \quad \mathbf{CR})\begin{pmatrix} -\mathbf{R} \\ \mathbf{I}_{n-r} \end{pmatrix} = \mathbf{0}$, implying that all columns of $\mathbf{X}$ are in $\mathcal{N}(\mathbf{A})$.

If $\mathbf{Xu} = \mathbf{0}$ for some $\mathbf{u} \in \mathbb{R}^{n-r}$, then $\mathbf{P}\begin{pmatrix} -\mathbf{Ru} \\ \mathbf{u} \end{pmatrix} = \mathbf{0} \implies \mathbf{u} = \mathbf{0}$. Hence columns of $\mathbf{X}$ are linearly independent.

If $\mathbf{Ax} = \mathbf{0}$ for some $\mathbf{x} = \mathbf{P}\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix}$ where $\mathbf{u}_1 \in \mathbb{R}^r$ and $\mathbf{u}_2 \in \mathbb{R}^{n-r}$, then $\mathbf{C}(\mathbf{I}_r \quad \mathbf{R})\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = \mathbf{0} \implies (\mathbf{I}_r \quad \mathbf{R})\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = \mathbf{0} \implies \mathbf{u}_1 = -\mathbf{Ru}_2 \implies \mathbf{x} = \mathbf{P}\begin{pmatrix} -\mathbf{Ru}_2 \\ \mathbf{u}_2 \end{pmatrix} = \mathbf{Xu}_2$, i.e. any $\mathbf{x} \in \mathcal{N}(\mathbf{A})$ is a linear combination of columns of $\mathbf{X}$.

To conclude, columns of $\mathbf{X}$ form a basis of $\mathcal{N}(\mathbf{A})$. Therefore, $\mathbf{rank}\,(\mathbf{A}) + \dim(\mathcal{N}(\mathbf{A})) = \mathbf{rank}\,(\mathbf{A}) + \mathbf{rank}\,(\mathbf{X}) = r + (n-r) = n$.

(b) ($\Rightarrow$) $\mathbf{y} = \mathbf{0} \implies \mathbf{a}_i^\top \mathbf{y} = \mathbf{a}_i^\top \mathbf{0} = 0, \forall i = 1, 2, \ldots, m$.

($\Leftarrow$) $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}$ is a basis of $\mathbb{R}^m \implies \mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m) \in \mathbb{R}^{m\times m}$ is invertible. Then $\mathbf{Ay} = \mathbf{0} \implies \mathbf{y} = \mathbf{A}^{-1}\mathbf{0} = \mathbf{0}$. ■

3. Show that

$$\mathbf{rank}\,(\mathbf{AB}) = \mathbf{rank}\,(\mathbf{B}) - \dim(\mathcal{C}(\mathbf{B}) \cap \mathcal{N}(\mathbf{A})). \tag{1}$$

Solution:

Suppose $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$. Let $r = \mathbf{rank}\,(\mathbf{AB})$ and $s = \mathbf{rank}\,(\mathbf{B})$. Factor $\mathbf{B}$ as $\mathbf{C_B R_B}$, where $\mathbf{C_B} \in \mathbb{R}^{n\times s}$ has full column rank and $\mathbf{R_B} \in \mathbb{R}^{s\times p}$ has full row rank. $\mathbf{AB} = \mathbf{AC_B R_B} \implies \mathbf{rank}\,(\mathbf{AB}) \le \mathbf{rank}\,(\mathbf{AC_B})$. $\mathbf{AC_B} = \mathbf{ABR_B^\top}(\mathbf{R_B R_B^\top})^{-1} \implies \mathbf{rank}\,(\mathbf{AC_B}) \le \mathbf{rank}\,(\mathbf{AB})$. Hence $\mathbf{rank}\,(\mathbf{AC_B}) = r$. Find a permutation matrix $\mathbf{P} \in \mathbb{R}^{s\times s}$ such that $\mathbf{AC_B P} = (\mathbf{C} \quad \mathbf{CR})$, where $\mathbf{C} \in \mathbb{R}^{m\times r}$ is r linearly independent columns of $\mathbf{AC_B}$ and $\mathbf{CR} \in \mathbb{R}^{m\times(s-r)}$ is the other columns.

Let $\mathbf{X} = \mathbf{C_B P}\begin{pmatrix} -\mathbf{R} \\ \mathbf{I}_{s-r} \end{pmatrix} \in \mathbb{R}^{n\times(s-r)}$. Clearly, columns of $\mathbf{X}$ are in $\mathcal{C}(\mathbf{B})$. Since $\mathbf{AX} = (\mathbf{C} \quad \mathbf{CR})\begin{pmatrix} -\mathbf{R} \\ \mathbf{I}_{s-r} \end{pmatrix} = \mathbf{0}$, columns of $\mathbf{X}$ are also in $\mathcal{N}(\mathbf{A})$, and thus in $\mathcal{C}(\mathbf{B}) \cap \mathcal{N}(\mathbf{A})$.

If $\mathbf{Xu} = \mathbf{0}$ for some $\mathbf{u} \in \mathbb{R}^{s-r}$, then $(\mathbf{C_B^\top C_B})^{-1}\mathbf{C_B^\top Xu} = \mathbf{0} \implies \mathbf{P}\begin{pmatrix} -\mathbf{Ru} \\ \mathbf{u} \end{pmatrix} = \mathbf{0} \implies \mathbf{u} = \mathbf{0}$. Hence columns of $\mathbf{X}$ are linearly independent.

If $\mathbf{Ax} = \mathbf{0}$ for some $\mathbf{x} = \mathbf{C_B P}\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix}$ where $\mathbf{u}_1 \in \mathbb{R}^r$ and $\mathbf{u}_2 \in \mathbb{R}^{s-r}$, then $\mathbf{C}(\mathbf{I}_r \quad \mathbf{R})\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = \mathbf{0}$ $\implies (\mathbf{I}_r \quad \mathbf{R})\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = \mathbf{0} \implies \mathbf{u}_1 = -\mathbf{Ru}_2 \implies \mathbf{x} = \mathbf{C_B P}\begin{pmatrix} -\mathbf{Ru}_2 \\ \mathbf{u}_2 \end{pmatrix} = \mathbf{Xu}_2$, i.e. any $\mathbf{x} \in \mathcal{C}(\mathbf{B}) \cap \mathcal{N}(\mathbf{A})$ is a linear combination of columns of $\mathbf{X}$.

To conclude, columns of $\mathbf{X}$ form a basis of $\mathcal{C}(\mathbf{B}) \cap \mathcal{N}(\mathbf{A})$. Therefore, $\mathbf{rank}\,(\mathbf{AB}) = r = s - (s-r) = \mathbf{rank}\,(\mathbf{B}) - \mathbf{rank}\,(\mathbf{X}) = \mathbf{rank}\,(\mathbf{B}) - \dim(\mathcal{C}(\mathbf{B}) \cap \mathcal{N}(\mathbf{A}))$. ■

4. Suppose that the first term on the right-hand side (RHS) of Eq. (1) changes to $\mathbf{rank}\,(\mathbf{A})$. Please find the second term on the RHS of Eq. (1) such that it still holds.

Solution:

$\mathbf{rank}\left(\mathbf{B}^\top \mathbf{A}^\top\right) = \mathbf{rank}\left(\mathbf{A}^\top\right) - \dim(\mathcal{C}(\mathbf{A}^\top) \cap \mathcal{N}(\mathbf{B}^\top))$

$\implies$ $\textbf{rank}\,(\mathbf{AB}) = \textbf{rank}\,(\mathbf{A}) - \dim(\mathcal{C}(\mathbf{A}^\top) \cap \mathcal{N}(\mathbf{B}^\top))$.
The second term on the RHS changes to $\dim(\mathcal{C}(\mathbf{A}^\top) \cap \mathcal{N}(\mathbf{B}^\top))$. ■

5. Show the results in 1. by Eq. (1) or the one you established in 4.

   Solution:

   (a) $\textbf{rank}\,(\mathbf{IA}) = \textbf{rank}\,(\mathbf{I}) - \dim(\mathcal{C}(\mathbf{I})^\top \cap \mathcal{N}(\mathbf{A}^\top)) \implies \textbf{rank}\,(\mathbf{A}) = n - \dim(\mathcal{N}(\mathbf{A}^\top))$.
   $\textbf{rank}\,\left(\mathbf{A}^\top\mathbf{I}\right) = \textbf{rank}\,(\mathbf{I}) - \dim(\mathcal{C}(\mathbf{I}) \cap \mathcal{N}(\mathbf{A}^\top)) \implies \textbf{rank}\,\left(\mathbf{A}^\top\right) = n - \dim(\mathcal{N}(\mathbf{A}^\top))$.
   Hence $\textbf{rank}\,(\mathbf{A}) = \textbf{rank}\,\left(\mathbf{A}^\top\right)$.

   (b) $\textbf{rank}\,(\mathbf{AB}) = \textbf{rank}\,(\mathbf{A}) - \dim(\mathcal{C}(\mathbf{A}^\top) \cap \mathcal{N}(\mathbf{B}^\top)) \le \textbf{rank}\,(\mathbf{A})$.

   (c) $\textbf{rank}\,(\mathbf{AB}) = \textbf{rank}\,(\mathbf{B}) - \dim(\mathcal{C}(\mathbf{B}) \cap \mathcal{N}(\mathbf{A})) \le \textbf{rank}\,(\mathbf{B})$.

   (d) $\textbf{rank}\,\left(\mathbf{A}^\top\mathbf{A}\right) = \textbf{rank}\,(\mathbf{A}) - \dim(\mathcal{C}(\mathbf{A}) \cap \mathcal{N}(\mathbf{A}^\top))$. For any $\mathbf{x} = \mathbf{Au} \in \mathcal{C}(\mathbf{A}) \cap \mathcal{N}(\mathbf{A}^\top)$, we have $\mathbf{x}^\top\mathbf{x} = \mathbf{u}^\top\mathbf{A}^\top\mathbf{x} = \mathbf{u}^\top\mathbf{0} = 0 \implies \mathbf{x} = \mathbf{0}$, so $\dim(\mathcal{C}(\mathbf{A}) \cap \mathcal{N}(\mathbf{A}^\top)) = 0$, and hence $\textbf{rank}\,\left(\mathbf{A}^\top\mathbf{A}\right) = \textbf{rank}\,(\mathbf{A})$. ■

Exercise 8: Linear Equations

Consider the system of linear equations in $\mathbf{w}$

$$\mathbf{y} = \mathbf{X}\mathbf{w}, \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{w} \in \mathbb{R}^d$, and $\mathbf{X} \in \mathbb{R}^{n\times d}$.

1. Give an example for "$\mathbf{X}$" and "$\mathbf{y}$" to satisfy the following three situations respectively:
   (a) there exists one unique solution;
   (b) there does not exist any solution;
   (c) there exists more than one solution.

   Solution:
   (a) $\mathbf{X} = \begin{pmatrix} \mathbf{I} \\ \mathbf{O} \end{pmatrix} (n > d)$, $\mathbf{y} = \mathbf{0}$. The unique solution is $\mathbf{w} = \mathbf{0}$.
   (b) $\mathbf{X} = \mathbf{O}$, $\mathbf{y} \neq \mathbf{0}$. No solution because $\mathbf{X}\mathbf{w} \equiv \mathbf{0}$.
   (c) $\mathbf{X} = \mathbf{O}$, $\mathbf{y} = \mathbf{0}$. The solution can be any $\mathbf{w} \in \mathbb{R}^d$. ■

2. Suppose that $\mathbf{X}$ has full column rank and $\mathbf{rank}\,((\mathbf{X}, \mathbf{y})) = \mathbf{rank}\,(\mathbf{X})$. Show that the system of linear equations (2) always admits a unique solution.

   Solution:
   $\mathbf{rank}\,((\mathbf{X}, \mathbf{y})) = \mathbf{rank}\,(\mathbf{X}) \implies \dim \mathcal{C}((\mathbf{X}, \mathbf{y})) = \dim \mathcal{C}(\mathbf{X}) = d$. Since $\mathbf{X}$ has full column rank, the d columns of $\mathbf{X}$ are linearly independent and thus form a basis of $\mathcal{C}((\mathbf{X}, \mathbf{y}))$. Then $\mathbf{y}$ can be represented as a linearly combination of columns of $\mathbf{X}$, i.e. $\mathbf{y} = \mathbf{X}\mathbf{w}$ admits a solution. Because $\mathbf{X}^\top\mathbf{X}$ is invertible, we have $\mathbf{w} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, which must be unique. ■

3. (**Normal equations**) Consider another system of linear equations in $\mathbf{w}$

$$\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X}\mathbf{w}. \tag{3}$$

   Please show that the system (3) always admits a solution. Moreover, does it always admit a unique solution?

   Solution:
   $\mathbf{rank}\left(\mathbf{X}^\top\right) = \mathbf{rank}\left(\mathbf{X}^\top\mathbf{X}\right) \leq \mathbf{rank}\left((\mathbf{X}^\top\mathbf{X} \quad \mathbf{X}^\top\mathbf{y})\right) \leq \mathbf{rank}\left(\mathbf{X}^\top\right) \implies \mathbf{rank}\left(\mathbf{X}^\top\mathbf{X}\right) = \mathbf{rank}\left((\mathbf{X}^\top\mathbf{X} \quad \mathbf{X}^\top\mathbf{y})\right)$. According to Exercise 8.2, $\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X}\mathbf{w}$ admits a solution.
   If $\mathbf{X}$ has full column rank, then $\mathbf{X}^\top\mathbf{X}$ is invertible and $\mathbf{w} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is the unique solution.
   If $\mathbf{X} \neq \mathbf{O}$ is rank-deficient, it can be factored as $\mathbf{C}\mathbf{R}$, where $\mathbf{C}$ has full column rank and $\mathbf{R}$ has full row rank. Then $\mathbf{w}_0 = \mathbf{R}^\top(\mathbf{R}\mathbf{R}^\top)^{-1}(\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top\mathbf{y}$ is a solution to $\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X}\mathbf{w}$. For any $\mathbf{u} \in \mathbb{R}^d$, $\mathbf{v} = (\mathbf{I} - \mathbf{R}^\top(\mathbf{R}\mathbf{R}^\top)^{-1}\mathbf{R})\mathbf{u} \neq \mathbf{0}$ is a solution to $\mathbf{0} = \mathbf{X}^\top\mathbf{X}\mathbf{w}$, and hence $\mathbf{w} = \mathbf{w}_0 + \mathbf{v}$ is a solution to $\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X}\mathbf{w}$, which is not unique. ■

Exercise 9: Properties of Eigenvalues and Singular Values

1. Suppose the maximum eigenvalue, minimum eigenvalue of a given symmetric matrix $\mathbf{A} \in S^n$ are denoted by $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$, respectively. Please show that

$$\lambda_{\max}(\mathbf{A}) = \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}, \quad \lambda_{\min}(\mathbf{A}) = \inf_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}.$$

Solution:
Consider the optimization problem max / min $R_{\mathbf{A}} = \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}$ s.t. $\mathbf{x}^\top\mathbf{x} = 1$. We have the Lagrangian $L(\mathbf{x}, \lambda) = \mathbf{x}^\top\mathbf{A}\mathbf{x} - \lambda(\mathbf{x}^\top\mathbf{x} - 1)$. Then the first order conditions become $\nabla_{\mathbf{x}}L(\mathbf{x}, \lambda) = 2\mathbf{A}\mathbf{x} - 2\lambda\mathbf{x} = 0 \implies (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ and $\mathbf{x}^\top\mathbf{x} = 1$, implying that λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ is a unit eigenvector corresponding to λ. For the maximization problem, the second order condition is $\nabla^2_{\mathbf{x}\mathbf{x}}L(\mathbf{x}, \lambda) = 2\mathbf{A} - 2\lambda\mathbf{I} \leq 0 \implies \mathbf{A} - \lambda\mathbf{I} \leq 0$, which is satified if and only if $\lambda = \lambda_{\max}(\mathbf{A})$. For the minimization problem, the second order condition is $\nabla^2_{\mathbf{x}\mathbf{x}}L(\mathbf{x}, \lambda) = 2\mathbf{A} - 2\lambda\mathbf{I} \geq 0 \implies \mathbf{A} - \lambda\mathbf{I} \geq 0$, which is satified if and only if $\lambda = \lambda_{\min}(\mathbf{A})$. Therefore, the global maximum is $R_{\mathbf{A}} = \mathbf{x}^\top\lambda_{\max}(\mathbf{A})\mathbf{x} = \lambda_{\max}(\mathbf{A})$ and the global minimum is $R_{\mathbf{A}} = \mathbf{x}^\top\lambda_{\min}(\mathbf{A})\mathbf{x} = \lambda_{\min}(\mathbf{A})$. In other words, $\lambda_{\max}(\mathbf{A}) = \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}$ and $\lambda_{\min}(\mathbf{A}) = \inf_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}$. ■

2. Suppose $\mathbf{B} = (b_{ij}) \in \mathbb{R}^{m\times n}$ with maximum singular value $\sigma_{\max}(\mathbf{B})$.

   (a) Let $\|\mathbf{B}\|_2 := \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{B}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$. Please show that

$$\sigma_{\max}(\mathbf{B}) = \|\mathbf{B}\|_2.$$

   (b) Please show that

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

   (c) Let $\|\mathbf{B}\|_1 := \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{B}\mathbf{x}\|_1}{\|\mathbf{x}\|_1}$. Please show that

$$\|\mathbf{B}\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^m |b_{ij}|.$$

   (d) Let $\|\mathbf{B}\|_\infty := \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{B}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty}$. Please show that

$$\|\mathbf{B}\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^n |b_{ij}|.$$

Solution:

   (a) $\|\mathbf{B}\|_2 = \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \sqrt{\frac{\mathbf{x}^\top\mathbf{B}^\top\mathbf{B}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}}$. Since $\mathbf{B}^\top\mathbf{B}$ is symmetric, by Exercise 9.1, we have $\|\mathbf{B}\|_2 = \sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})} = \sigma_{\max}(\mathbf{B})$.

(b) Consider maximizing $R_{\mathbf{B}} = \frac{\mathbf{x}^\top \mathbf{B}\mathbf{y}}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}$ subject to $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$. Let $\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$ and the Lagrangian $L(\mathbf{z}, \boldsymbol{\lambda}) = \mathbf{x}^\top \mathbf{B}\mathbf{y} - \frac{1}{2}\lambda_1(\|\mathbf{x}\|_2^2 - 1) - \frac{1}{2}\lambda_2(\|\mathbf{y}\|_2^2 - 1)$. The necessary conditions $\nabla_{\mathbf{z}} L(\mathbf{z}, \boldsymbol{\lambda}) = \begin{pmatrix} \mathbf{B}\mathbf{y} - \lambda_1 \mathbf{x} \\ \mathbf{B}^\top \mathbf{x} - \lambda_2 \mathbf{y} \end{pmatrix} = 0$ and $\mathbf{x}^\top \mathbf{x} = \mathbf{y}^\top \mathbf{y} = 1$ imply that $\sigma := \lambda_1 = \lambda_2$ is a singular value of $\mathbf{B}$ with unit left and right singular vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. Let $\mathbf{A} = \begin{pmatrix} & \mathbf{B} \\ \mathbf{B}^\top & \end{pmatrix}$. Then $\nabla^2_{\mathbf{zz}} L(\mathbf{z}, \boldsymbol{\lambda}) = \begin{pmatrix} -\lambda_1 \mathbf{I}_m & \mathbf{B} \\ \mathbf{B}^\top & -\lambda_2 \mathbf{I}_n \end{pmatrix} = \mathbf{A} - \sigma \mathbf{I}$. Clearly, every eigenvalue of $\mathbf{A}$ is also a singular value of $\mathbf{B}$, and $\sigma_{\max}(\mathbf{B})$ is the largest eigenvalue of $\mathbf{A}$. So $\nabla^2_{\mathbf{zz}} L(\mathbf{z}, \boldsymbol{\lambda})$ is negative semidefinite if and only if $\sigma = \sigma_{\max}(\mathbf{B})$, and hence the maximum of $R_{\mathbf{B}} = \mathbf{x}^\top \mathbf{B}\mathbf{y} = \mathbf{x}^\top \sigma \mathbf{x} = \sigma_{\max}(\mathbf{B})$. ' In other words, $\sigma_{\max}(\mathbf{B}) = \sup_{\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n, \mathbf{x}, \mathbf{y} \neq 0} \frac{\mathbf{x}^\top \mathbf{B}\mathbf{y}}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}$.

(c)
$$\begin{aligned}\frac{\|\mathbf{B}\mathbf{x}\|_1}{\|\mathbf{x}\|_1} &= \frac{\sum\limits_{i=1}^m |\sum\limits_{j=1}^n b_{ij} x_j|}{\sum\limits_{j=1}^n |x_j|} \leq \frac{\sum\limits_{i=1}^m \sum\limits_{j=1}^n |b_{ij}||x_j|}{\sum\limits_{j=1}^n |x_j|} = \frac{\sum\limits_{j=1}^n (\sum\limits_{i=1}^m |b_{ij}|)|x_j|}{\sum\limits_{j=1}^n |x_j|} \\ &\leq \frac{\sum\limits_{j=1}^n |x_j|}{\sum\limits_{j=1}^n |x_j|} \max_{1 \leq j \leq n} \sum_{i=1}^m |b_{ij}| = \max_{1 \leq j \leq n} \sum_{i=1}^m |b_{ij}|.\end{aligned}$$

The equality holds if $x_j = \begin{cases} 1, & \text{if } j = \mathop{\mathbf{argmax}}\limits_{1 \leq j \leq n} \sum\limits_{i=1}^m |b_{ij}| \\ 0, & \text{otherwise} \end{cases}$, so $\|\mathbf{B}\|_1 = \max\limits_{1 \leq j \leq n} \sum\limits_{i=1}^m |b_{ij}|$.

(d)
$$\begin{aligned}\frac{\|\mathbf{B}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} &= \frac{\max\limits_{1 \leq i \leq m} |\sum\limits_{j=1}^n b_{ij} x_j|}{\max\limits_{1 \leq j \leq n} |x_j|} \leq \frac{\max\limits_{1 \leq i \leq m} \sum\limits_{j=1}^n |b_{ij}||x_j|}{\max\limits_{1 \leq j \leq n} |x_j|} \\ &\leq \frac{\max\limits_{1 \leq j \leq n} |x_j|}{\max\limits_{1 \leq j \leq n} |x_j|} \max_{1 \leq i \leq m} \sum_{j=1}^n |b_{ij}| = \max_{1 \leq i \leq m} \sum_{j=1}^n |b_{ij}|.\end{aligned}$$

Let $k = \mathop{\mathbf{argmax}}\limits_{1 \leq i \leq m} \sum\limits_{j=1}^n |b_{ij}|$. The equality holds if $x_j = \begin{cases} 1, & \text{if } b_{kj} \geq 0 \\ -1, & \text{if } b_{kj} < 0 \end{cases}$, so $\|\mathbf{B}\|_\infty = \max\limits_{1 \leq i \leq m} \sum\limits_{j=1}^n |b_{ij}|$. ■

Exercise 10: Projection to a Linear Subspace

1. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ with rank d and $\mathbf{y} \in \mathbb{R}^n$. Consider the optimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2,$$

   (a) We denote the column space of $\mathbf{X}$ by $\mathcal{C}(\mathbf{X})$. Please show that $\hat{\mathbf{y}} := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is the projection of $\mathbf{y}$ on $\mathcal{C}(\mathbf{X})$, i.e. $\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{x}\rangle = 0$ for any $\mathbf{x} \in \mathcal{C}(\mathbf{X})$.

   (b) Please solve the above optimization problem by completing the square.

   (c) Please show that $\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2 \le \|\mathbf{y}\|_2$. Then find the necessary and sufficient condition where the equality holds and give it a geometric interpretation.

   Solution:

   (a) For any $\mathbf{x} = \mathbf{X}\mathbf{u} \in \mathcal{C}(\mathbf{X})$, $\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{x}\rangle = (\mathbf{y} - \hat{\mathbf{y}})^\top\mathbf{x} = \mathbf{y}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{x} = \mathbf{y}^\top(\mathbf{X}\mathbf{u} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X})\mathbf{u}) = \mathbf{y}^\top\mathbf{0} = 0$.

   (b) $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 = \|\mathbf{y} - \hat{\mathbf{y}} + \hat{\mathbf{y}} - \mathbf{X}\mathbf{w}\|_2^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 + 2\langle \mathbf{y} - \hat{\mathbf{y}}, \hat{\mathbf{y}} - \mathbf{X}\mathbf{w}\rangle + \|\hat{\mathbf{y}} - \mathbf{X}\mathbf{w}\|_2^2$. Since $\hat{\mathbf{y}} - \mathbf{X}\mathbf{w} = \mathbf{X}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} - \mathbf{w}) \in \mathcal{C}(\mathbf{X})$, it follows that $\langle \mathbf{y} - \hat{\mathbf{y}}, \hat{\mathbf{y}} - \mathbf{X}\mathbf{w}\rangle = 0$. So $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 + \|\hat{\mathbf{y}} - \mathbf{X}\mathbf{w}\|_2^2 \ge \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$, where the equality holds if and only if $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$, i.e. $\mathbf{w} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. And hence $\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$.

   (c) $\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 = \langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{y} + \hat{\mathbf{y}} - 2\hat{\mathbf{y}}\rangle = \|\mathbf{y}\|_2^2 - \|\hat{\mathbf{y}}\|_2^2 - 2\langle \mathbf{y} - \hat{\mathbf{y}}, \hat{\mathbf{y}}\rangle = \|\mathbf{y}\|_2^2 - \|\hat{\mathbf{y}}\|_2^2 \le \|\mathbf{y}\|_2^2 \implies \min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2 \le \|\mathbf{y}\|_2$. The equality holds if and only if $\|\hat{\mathbf{y}}\|_2 = 0 \iff \hat{\mathbf{y}} = \mathbf{0} \iff \langle \mathbf{y}, \mathbf{x}\rangle = 0$ for any $\mathbf{x} \in \mathcal{C}(\mathbf{X})$. In other words, the projection of $\mathbf{y}$ on $\mathcal{C}(\mathbf{X})$ is zero, and thus $\mathbf{y}$ is orthogonal to $\mathcal{C}(\mathbf{X})$, i.e. $\mathbf{y} \in \mathcal{N}(\mathbf{X}^\top)$. ∎

2. Suppose X and Y are both random variables defined in the same sample space Ω with finite second-order moment, i.e. $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$.

   (a) Let $L^2(\Omega) = \{Z : \Omega \to \mathbb{R} \mid \mathbb{E}[Z^2] < \infty\}$ be the set of random variables with finite second-order moment. Please show that $L^2(\Omega)$ is a linear space, and $\langle X, Y\rangle := \mathbb{E}[XY]$ defines an inner product in $L^2(\Omega)$. Then find the projection of Y on the subspace of $L^2(\Omega)$ consisting of all constant variables.

   (b) Please find a real constant $\hat{c}$, such that

$$\hat{c} = \mathop{\mathbf{argmin}}_{c\in\mathbb{R}} \mathbb{E}[(Y - c)^2].$$

   [Hint: you can solve it by completing the square.]

   (c) Please find the necessary and sufficient condition where $\min_{c\in\mathbb{R}} \mathbb{E}[(Y - c)^2] = \mathbb{E}[Y^2]$. Then give it a geometric interpretation using inner product and projection.

   Solution:

   (a) Suppose $X, Y, Z \in L^2(\Omega)$ and $a, b \in \mathbb{R}$. Then $\mathbb{E}[(X + Y)^2] = \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY] \le \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]} < \infty$ and $\mathbb{E}[aX^2] = a\mathbb{E}[X^2] < \infty$, implying that $X + Y, aX \in L^2(\Omega)$. Therefore, we can define vector addition in $L^2(\Omega)$ as the usual addition of random variables, and scalar multiplication as the usual multiplication of a random variable by a real constant. Clearly, the addition is associative and commutative, with 0 as the identity element, i.e. $X + (Y + Z) = (X + Y) + Z$, $X + Y = Y + X$ and

$X + 0 = X$. The scalar multiplication, with 1 as the identity element, is compatible with field multiplication and distributive with respect to both vector addition and field addition, i.e. $1X = X$, $a(bX) = (ab)X$, $a(X+Y) = aX + aY$ and $(a+b)X = aX + bX$. To conclude, $L^2(\Omega)$ is a linear space.
Morever, the given definition of inner product satisfies symmetry, linearity and positive definiteness, i.e. $\mathbb{E}[XY] = \mathbb{E}[YX]$, $\mathbb{E}[(aX + bY)Z] = a\mathbb{E}[XZ] + b\mathbb{E}[YZ]$ and $\mathbb{E}[X^2] \geq 0$, where the equality holds if and only if $X \equiv 0$.
For any real constant variable C in the subspace, $\langle Y - \mathbb{E}[Y], C\rangle = \mathbb{E}[C](\mathbb{E}[Y] - \mathbb{E}[\mathbb{E}[Y]]) = C(\mathbb{E}[Y] - \mathbb{E}[Y] = 0$. Hence the projection of Y on the subspace is $\mathbb{E}[Y]$.

(b) $\mathbb{E}[(Y-c)^2] = \mathbb{E}[(Y - \mathbb{E}[Y] + \mathbb{E}[Y] - c)^2] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] + 2\langle Y - \mathbb{E}[Y], \mathbb{E}[Y] - c\rangle + \mathbb{E}[(\mathbb{E}[Y] - c)^2] = \mathrm{Var}\,(Y) + (\mathbb{E}[Y] - c)^2 \geq \mathrm{Var}\,(Y)$, where the equality holds if and only if $\mathbb{E}[Y] - c = 0$. Hence $\hat{c} = \mathbb{E}[Y]$.

(c) $\min_{c\in\mathbb{R}} \mathbb{E}[(Y-c)^2] = \mathrm{Var}\,(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \leq \mathbb{E}[Y^2]$. The equality holds if and only if $\mathbb{E}[Y] = 0 \iff \langle Y, C\rangle = 0$ for any C in the subspace consisting of constant variables. In other words, the projection of Y on the subspace is zero, and thus Y is orthogonal to the subspace. ■

3. Suppose X and Y are both random variables defined in the same sample space Ω and all the expectations exist in this problem. Consider the problem

$$\min_{f:\mathbb{R}\to\mathbb{R}} \mathbb{E}[(f(X) - Y)^2].$$

(a) Please solve the above problem by completing the square.

(b) We let $\mathcal{C}(X)$ denote the subspace $\{f(X) \mid f(\cdot) : \mathbb{R} \to \mathbb{R}, \mathbb{E}[f(X)^2] < \infty\}$ of $L^2(\Omega)$. Please show that the solution of the above problem is the projection of Y on $\mathcal{C}(X)$.

(c) Please show that question 2 is a special case of question 3.

Solution:

(a)
$$\begin{aligned}\mathbb{E}[(f(X) - Y)^2] &= \mathbb{E}[(f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y)^2] \\ &= \mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] \\ &\quad + 2\,\mathbb{E}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)].\end{aligned}$$

$$\begin{aligned}\mathbb{E}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)] &= \mathbb{E}\{(f(X) - \mathbb{E}[Y|X])\,\mathbb{E}[\mathbb{E}[Y|X] - Y|X]\} \\ &= \mathbb{E}\{(f(X) - \mathbb{E}[Y|X])\,(\mathbb{E}[Y|X] - \mathbb{E}[Y|X])\} \\ &= \mathbb{E}\{(f(X) - \mathbb{E}[Y|X]) \cdot 0\} = 0.\end{aligned}$$

$$\therefore \mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] \geq \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2].$$

The equality holds if and only if $f(X) = \mathbb{E}[Y|X]$. Hence $\min\limits_{f:\mathbb{R}\to\mathbb{R}} \mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] = \mathrm{Var}\,(Y|X)$.

(b) For any $f(X) \in \mathcal{C}(X)$, $\langle Y - \mathbb{E}[Y|X], f(X)\rangle = \mathbb{E}[f(X)(Y - \mathbb{E}[Y|X])] = \mathbb{E}\{f(X)\,\mathbb{E}[Y - \mathbb{E}[Y|X]|X]\} = \mathbb{E}\{f(X) \cdot 0\} = 0$. Hence the projection of Y on $\mathcal{C}(X)$ is $\mathbb{E}[Y|X]$.

(c) If X, Y are independent, then $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ and $\min\limits_{f:\mathbb{R}\to\mathbb{R}} \mathbb{E}[(f(X) - Y)^2] = \min\limits_{c\in\mathbb{R}} \mathbb{E}[(Y - c)^2] = \mathrm{Var}\,(Y)$. More specifically, if X is a constant variable, then $\mathcal{C}(X)$ consists of all constant variables, and hence question 3 becomes question 2. ■

References

[1] T. Tao. *Analysis II*. Springer, 2015.