Introduction to Machine Learning
Fall 2025
University of Science and Technology of China

Lecturer: Zhihui Li, Xiaojun Chang — Homework 1
Posted: Sep. 28th, 2025 — Due: Oct. 16th, 2025

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 $\mathbf{x}$.

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.

Solution 1: Limit and Limit Points

1. ①Necessity:

   $\lim\limits_{n\to\infty} x_n = x \Longrightarrow \forall\, \epsilon > 0,\, \exists\, N > 0$, s.t. when $n \geq N$, $\|x_n - x\| < \epsilon \Longrightarrow \|x_n\| < \|x\| + \epsilon$.

   Set
   $$M = \max\{\|x_1\|, \|x_2\|, \cdots, \|x_{N-1}\|, \|x\| + \epsilon\}.$$

   Clearly, $\|x_n\| \leq M$, $\forall\, n > 0$, i.e., sequence $\{x_n\}$ is bounded.

   Besides, $\lim\limits_{n\to\infty} x_n = x \Longrightarrow$ all subsequences $\{x_{n_k}\}$ also converge to $x \Longrightarrow$ limit point x is unique.

   ②Sufficiency:

   Suppose, for the sake of contradiction, that $\lim\limits_{n\to\infty} x_n \neq x$. Then $\exists\, \epsilon_0 > 0$ and a subsequence $\{x_{n_k}\}$ such that
   $$\|x_{n_k} - x\| \geq \epsilon_0,\ \forall\, k > 0$$

   Accoring to Bolzano-Weierstrass Theorem, $\{x_{n_k}\}$ is bounded $\Longrightarrow$ there is a further convergent subsequence $\{x_{n_{k_j}}\}$ with $\lim\limits_{j\to\infty} x_{n_{k_j}} = y \in \mathbb{R}^n$. By the uniqueness of limit points, $y = x$. Therefore, $\|x_{n_{k_j}} - x\| \to 0$, which contradicts $\|x_{n_k} - x\| \geq \epsilon_0$, $\forall\, k > 0$.

   Hence our supposition was false and $x_n \to x$.

2. (a) For any $x \in [0,1]$, take $x_n = x + \frac{1}{n}$ if $x < 1$; $x_n = x - \frac{1}{n}$ if $x > 0$. Then it is clear that $x_n \neq x$, $\forall n > 0$ and $x_n \to x$. Thus, x is a limit point of C.
   If $x \notin [0,1]$, $dist(x, (0,1)) > 0 \Longrightarrow \exists\, \epsilon > 0$, s.t. $B(x,\epsilon)\backslash\{x\} \cap C = \emptyset \Longrightarrow x$ is not a limit point of C.
   Thus $C' = [0,1]$. Correspondingly, $x = 2$ is an isolated point of C.

   (b) To prove C' is closed $\Longleftrightarrow$ to prove $C' = \overline{C'} \Longleftrightarrow$ to prove all of the limit points of C' belong to C'
   Set x as any limit point of $C' \Longrightarrow$ there exists a sequence $\{x_n\} \subset C'\backslash\{x\}$ with $x_n \to x$.
   For each n, since $x_n \in C'$, there exists a sequence $\{x_{n,m}\} \subset C\backslash\{x_n\}$ with $x_{n,m} \to x_n$.
   $\forall\, \epsilon > 0$, $\exists N > 0$, s.t. when $n \geq N$, $\|x_n - x\| < \frac{\epsilon}{2}$. Let m be large enough, then $\|x_{n,m} - x_n\| < \frac{\epsilon}{2}$. By triangle inequality,
   $$\|x_{n,m} - x\| \leq \|x_{n,m} - x_n\| + \|x_n - x\| \leq \epsilon$$
   Therefore, $x_{n,m} \in C$ implies $x \in C'$.
   In conclusion, the set C' is closed.

■

Exercise 2: Norms

In this exercise, we will give some examples of norms and a useful theorem related to norms in **finite** dimensional vector space.

1. l_p **norm:** The l_p norm is defined by

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

where $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $p \geq 1$.

   (a) Please show that the l_p norm is a norm.

   (b) Please show that the following equality.

$$\lim_{p\to\infty} \|\mathbf{x}\|_p = \|\mathbf{x}\|_\infty = \max_{1\leq i\leq n} |x_i|.$$

   The l_∞ norm is defined as above.

2. **Operator norms:** Suppose that $\mathbf{A} \in \mathbb{R}^{m\times n}$, which can be viewed as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Please show the following operator norms' equality.

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

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

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

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

3. **(Optional) Dual norm:** Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. The dual norm of $\|\cdot\|$ is defined by

$$\|\mathbf{x}\|_* = \sup_{\mathbf{y}\in\mathbb{R}^n, \|\mathbf{y}\|\leq 1} \mathbf{y}^\top \mathbf{x}.$$

   (a) Please show that the dual of the Euclidean norm is the Euclidean norm itself. i.e.,

$$\sup_{\mathbf{y}\in\mathbb{R}^n, \|\mathbf{y}\|_2\leq 1} \mathbf{y}^\top \mathbf{x} = \|\mathbf{x}\|_2.$$

   (b) Please show that the dual of the l_1 norm is the l_∞ norm. i.e.,

$$\sup_{\mathbf{y}\in\mathbb{R}^n, \|\mathbf{y}\|_1\leq 1} \mathbf{y}^\top \mathbf{x} = \|\mathbf{x}\|_\infty.$$

Solution 2: Norms

1. (a) ① Positive Definiteness:

$$|x_i| \geq 0 \Longrightarrow \|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \geq 0$$

where the equality holds if and only if $x_i = 0,\ \forall\ 1 \leq i \leq n \Longleftrightarrow \mathbf{x} = \mathbf{0}$.
② Homogeneity: $\forall\ \alpha \in \mathbb{R}$,

$$\|\alpha\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |\alpha x_i|^p\right)^{1/p} = |\alpha| \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} = |\alpha|\|\mathbf{x}\|_p$$

③ Triangle Inequality: Let $q = \frac{p}{p-1}$ to satisfy $\frac{1}{p} + \frac{1}{q} = 1$.
$\forall\ \mathbf{x},\ \mathbf{y} \in \mathbb{R}^n$, according to Hölder inequality,

$$\begin{aligned}
\|\mathbf{x}+\mathbf{y}\|_p^p &= \sum_{i=1}^{n} |x_i + y_i|^p \\
&\leq \sum_{i=1}^{n} |x_i| \cdot |x_i + y_i|^{p-1} + \sum_{i=1}^{n} |y_i| \cdot |x_i + y_i|^{p-1} \\
&\leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \cdot \left(\sum_{i=1}^{n} |x_i + y_i|^{(p-1)q}\right)^{1/q} \\
&\quad + \left(\sum_{i=1}^{n} |y_i|^p\right)^{1/p} \cdot \left(\sum_{i=1}^{n} |x_i + y_i|^{(p-1)q}\right)^{1/q} \\
&= (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p) \cdot \left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{1/q}
\end{aligned}$$

$$\Longrightarrow \|\mathbf{x}+\mathbf{y}\|_p = \left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{1/p} = \left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{1-1/q} \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$$

(b) It is clear that

$$\|\mathbf{x}\|_\infty^p \leq \|\mathbf{x}\|_p^p \leq n\|\mathbf{x}\|_\infty^p \Longleftrightarrow \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_p \leq n^{1/p}\|\mathbf{x}\|_\infty$$

By the Squeeze Theorem,

$$\lim_{p\to\infty} n^{1/p} = 1 \Longrightarrow \lim_{p\to\infty} \|\mathbf{x}\|_p = \|\mathbf{x}\|_\infty$$

2. When $\mathbf{A} = \mathbf{0}$, the conclusion is trivial. Thus, the following discussion is based on $\mathbf{A} \neq \mathbf{0}$.

Divide matrix $\mathbf{A}$ into blocks by columns as $(a_1, a_2, \cdots, a_n)$ and let $\|a_{j_0}\|_1 = \max\limits_{1\le j\le n} \|a_j\|_1$.

Then $\forall\ \mathbf{x} \in \mathbb{R}^n$ that satisfies $\|\mathbf{x}\|_1 = \sum\limits_{i=1}^{n} |x_i| = 1$,

$$\|\mathbf{A}\mathbf{x}\|_1 = \|\sum_{j=1}^{n} x_j a_j\|_1 \le \sum_{j=1}^{n} |x_j|\|a_j\|_1 \le \left(\sum_{i=1}^{n} |x_i|\right) \max_{1\le j\le n} \|a_j\|_1 = \|a_{j_0}\|_1$$

Besides, $\|\mathbf{A}\mathbf{e_{j_0}}\|_1 = \|a_{j_0}\|_1$. In conclusion,

$$\|\mathbf{A}\|_1 = \max_{\|\mathbf{x}\|_1=1} \|\mathbf{A}\mathbf{x}\|_1 = \max_{1\le j\le n} \|a_j\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|.$$

(b) $\forall\ \mathbf{x} \in \mathbb{R}^n$ that satisfies $\|\mathbf{x}\|_\infty = 1$,

$$\|\mathbf{A}\mathbf{x}\|_\infty = \max_{1\le i\le n} \left|\sum_{j=1}^{n} a_{ij} x_j\right| \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}||x_j| \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$$

Set $\max\limits_{1\le i\le n} \sum\limits_{j=1}^{n} |a_{ij}| = \sum\limits_{j=1}^{n} |a_{kj}|$. Then let $\tilde{\mathbf{x}} = (\operatorname{sgn}(a_{k1}), \cdots, \operatorname{sgn}(a_{kn}))^T$.

$\mathbf{A} \ne \mathbf{0} \Longrightarrow \|\tilde{\mathbf{x}}\|_\infty = 1$ and it is clear that $\|\mathbf{A}\tilde{\mathbf{x}}\|_\infty = \max\limits_{1\le i\le n} \sum\limits_{j=1}^{n} |a_{ij}|$.

3. (a) By Cauchy-Schwarz Inequality,

$$\mathbf{y^T x} \le \|\mathbf{y}\|_2 \|\mathbf{x}\|_2 \le \|\mathbf{x}\|_2 \Longrightarrow \sup_{\mathbf{y}\in\mathbb{R}^n, \|\mathbf{y}\|_2\le 1} \mathbf{y}^\top \mathbf{x} \le \|\mathbf{x}\|_2.$$

If $\mathbf{x} = \mathbf{0}$, the target equality is trivial; if not, choose $\mathbf{y_0} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$, which satisfies $\|\mathbf{y_0}\|_2 = 1$. Then $\mathbf{y_0^T x} = \frac{\mathbf{x^T x}}{\|\mathbf{x}\|_2} = \|\mathbf{x}\|_2$.
In conclusion, $\sup_{\mathbf{y}\in\mathbb{R}^n, \|\mathbf{y}\|_2\le 1} \mathbf{y}^\top \mathbf{x} = \|\mathbf{x}\|_2$.

(b) For any $\mathbf{y} = (y_1, \cdots, y_n)^T$ with $\|y\|_1 \le 1$,

$$\mathbf{y^T x} = \sum_{i=1}^{n} x_i y_i \le \sum_{i=1}^{n} |x_i||y_i| \le \max_{1\le i\le n} |x_i| \sum_{i=1}^{n} |y_i| \le \|\mathbf{x}\|_\infty$$

$$\Longrightarrow \sup_{\mathbf{y}\in\mathbb{R}^n, \|\mathbf{y}\|_1\le 1} \mathbf{y}^\top \mathbf{x} \le \|\mathbf{x}\|_\infty.$$

If $\mathbf{x} = \mathbf{0}$, the target equality is trivial; if not, let $k = \arg\max\limits_{1\le i\le n} |x_i|$ and choose $\mathbf{y_0} = \operatorname{sgn}(x_k)\mathbf{e_k}$ so that $\|y_0\|_1 = 1$. Then $\mathbf{y_0^T x} = \operatorname{sgn}(\mathbf{x_k})\mathbf{x_k} = \|\mathbf{x}\|_\infty$.
In conclusion, $\sup_{\mathbf{y}\in\mathbb{R}^n, \|\mathbf{y}\|_1\le 1} \mathbf{y}^\top \mathbf{x} = \|\mathbf{x}\|_\infty$.

■

Exercise 3: 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 $\mathbf{x} \in C$.

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 3: Open and Closed Sets

1. (a) $\Longrightarrow$ (b):

   Suppose, for the sake of contradiction, that C^c is not open. Then $\exists\ \mathbf{x} \in C$, s.t. $\forall\ \epsilon > 0$, $B_\epsilon(\mathbf{x}) \not\subset C^c \Longrightarrow$ take $\mathbf{x_n} \subset B_\epsilon(\mathbf{x}) \backslash C^c \subset C$ for $\epsilon_n > 0$. Without generality, set $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_n \to 0$, then it is clear that sequence $\{\mathbf{x_n}\}$ satisfies $\lim\limits_{n\to\infty} \mathbf{x_n} = \mathbf{x}$. Therefore, $\mathbf{x} \in \mathbf{cl}\ C = C$, which contradicts $\mathbf{x} \in C^c$.

   (b) $\Longrightarrow$ (c):

   Suppose, for the sake of contradiction, that $\exists\ \mathbf{x} \in C^c$, s.t. $B_\epsilon(\mathbf{x}) \cap C \neq \emptyset$ for every $\epsilon > 0$. Because C^c is open, $\exists\ \delta > 0$, s.t. $B_\delta(\mathbf{x}) \subset C^c \Longrightarrow B_\delta(\mathbf{x}) \subset C \cap C^c = \emptyset$. So this is a contradiction.

   (c) $\Longrightarrow$ (a):

   Suppose, for the sake of contradiction, that $\exists$ a sequence $\{\mathbf{x_n}\} \subset C$, which satisfies $\mathbf{x_n} \to \mathbf{x} \in C^c$, i.e., $\forall\ \epsilon > 0$, $\exists\ N > 0$, s.t. when $n \geq N$, $\|\mathbf{x_n} - \mathbf{x}\| < \epsilon$. Therefore, $B_\epsilon(\mathbf{x}) \cap C \supset \{x_n\} \neq \emptyset \Longrightarrow \mathbf{x} \in C$, which contradicts $\mathbf{x} \in C^c$.

   $$\begin{tikzcd} & (b) \arrow[dr, Rightarrow] & \\ (a) \arrow[ur, Rightarrow] & & (c) \arrow[ll, Rightarrow] \end{tikzcd}$$

   In summary, these three hypothesis are equivalent.

2. (a) $\forall\ \epsilon > 0$, $-\frac{\epsilon}{2} \notin [0,1] \Longrightarrow B_\epsilon(0) = (-\epsilon, \epsilon) \not\subset [0,1] \Longrightarrow [0,1]$ is not an open set in $\mathbb{R}$.

   If $\mathbf{x} \in (0,1)$, take $\epsilon = \min\{\mathbf{x}, \mathbf{1} - \mathbf{x}\} > 0$. Then $B_\epsilon(\mathbf{x}) \subset [0,1] \Longrightarrow B_\epsilon(\mathbf{x}) \cap B = B_\epsilon(\mathbf{x}) \subset [0,1]$.

   If $\mathbf{x} = 0$, take $\epsilon = \frac{1}{2} > 0$. Then $B_\epsilon(0) = (-\frac{1}{2}, \frac{1}{2}) \Longrightarrow B_\epsilon(0) \cap B = [0, \frac{1}{2}] \subset [0,1]$. In the same way, if $\mathbf{x} = 1$, take $\epsilon = \frac{1}{2} > 0$. Then $B_\epsilon(1) = (\frac{1}{2}, \frac{3}{2}) \Longrightarrow B_\epsilon(1) \cap B = [\frac{1}{2}, 1] \subset [0,1]$.

   In conclusion, $[0,1]$ is open in B.

   Beside, if $\mathbf{x} = 2$, take $\epsilon = \frac{1}{2} > 0$. Then $B_\epsilon(2) = (\frac{3}{2}, \frac{5}{2}) \Longrightarrow B_\epsilon(2) \cap B = \{\frac{1}{2}\} \subset \{\frac{1}{2}\} \Longrightarrow \{2\} = B \backslash [0,1]$ is open in B.

   In summary, $[0,1]$ is both open and closed in B.

   (b) ① Necessity:

   $C \subset A$ is open in $A \Longrightarrow \forall\ \mathbf{x} \in C$, $\exists\ \epsilon_x > 0$, s.t. $B_{\epsilon_x}(\mathbf{x}) \cap A \subset C \Longrightarrow$

   $$\bigcup_{\mathbf{x}\in C} (B_{\epsilon_x}(\mathbf{x}) \cap A) = \left(\bigcup_{\mathbf{x}\in C} B_{\epsilon_x}(\mathbf{x})\right) \cap A \subset C.$$

   Set $U = \bigcup\limits_{\mathbf{x}\in C} B_{\epsilon_x}(\mathbf{x})$. Then $U \cap A \subset C$ and it is clear that U is open in $\mathbb{R}^n$.

$\forall\, \mathbf{x} \in C$, $\mathbf{x} \in B_{\epsilon_x}(\mathbf{x})$ and $\mathbf{x} \in C \subset A \implies$

$$\mathbf{x} \in B_{\epsilon_x}(\mathbf{x}) \cap A \subset \bigcup_{\mathbf{x} \in C} (B_{\epsilon_x}(\mathbf{x}) \cap A) = \left(\bigcup_{\mathbf{x} \in C} B_{\epsilon_x}(\mathbf{x}) \right) \cap A = U \cap A.$$

$\implies C \subset U \cap A$.

In conclusion, $C = U \cap A$ and U is open in $\mathbb{R}^n$.

② Sufficiency:

U is open in $\mathbb{R}^n \implies \forall\, \mathbf{x} \in U \supset C$, $\exists\, \epsilon_x > 0$, s.t. $B_{\epsilon_x}(\mathbf{x}) \subset U \implies B_{\epsilon_x}(\mathbf{x}) \cap A \subset U \cap A = C$. Thus $C \subset A$ is open in A.

■

Exercise 4: Projection

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

$$\mathbf{P_A}(\mathbf{x}) = \underset{\mathbf{z}\in\mathbb{R}^m}{\mathbf{argmin}}\,\{\|\mathbf{x}-\mathbf{z}\|_2 : \mathbf{z} \in \mathcal{C}(\mathbf{A})\}.$$

We call $\mathbf{P_A}(\mathbf{x})$ the projection of the point $\mathbf{x}$ onto the column space of $\mathbf{A}$.

1. Please show that $\mathbf{P_A}(\mathbf{x})$ is unique for any $\mathbf{x} \in \mathbb{R}^m$.
2. Let $\mathbf{v}_i \in \mathbb{R}^n$, $i = 1, \ldots, d$ with $d \leq n$, which are linearly independent.
   (a) For any $\mathbf{w} \in \mathbb{R}^n$, please find $\mathbf{P}_{\mathbf{v}_1}(\mathbf{w})$, which is the projection of $\mathbf{w}$ onto the subspace spanned by $\mathbf{v}_1$.
   (b) Please show $\mathbf{P}_{\mathbf{v}_1}(\cdot)$ is a linear map, i.e.,
   $$\mathbf{P}_{\mathbf{v}_1}(\alpha\mathbf{u} + \beta\mathbf{w}) = \alpha\mathbf{P}_{\mathbf{v}_1}(\mathbf{u}) + \beta\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}),$$
   where $\alpha, \beta \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^n$.
   (c) Please find the projection matrix corresponding to the linear map $\mathbf{P}_{\mathbf{v}_1}(\cdot)$, i.e., find the matrix $\mathbf{H}_1 \in \mathbb{R}^{n\times n}$ such that
   $$\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \mathbf{H}_1\mathbf{w}.$$
   (d) Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$.
      i. For any $\mathbf{w} \in \mathbb{R}^n$, please find $\mathbf{P_V}(\mathbf{w})$ and the corresponding projection matrix $\mathbf{H}$.
      ii. Please find $\mathbf{H}$ if we further assume that $\mathbf{v}_i^\top\mathbf{v}_j = 0$, $\forall\, i \neq j$.
3. (a) Suppose that
   $$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$
   What are the coordinates of $\mathbf{P_A}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ for any $\mathbf{x} \in \mathbb{R}^2$? Are the coordinates unique?
   (b) Suppose that
   $$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}.$$
   What are the coordinates of $\mathbf{P_A}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ for any $\mathbf{x} \in \mathbb{R}^2$? Are the coordinates unique?
4. A matrix $\mathbf{P}$ is called a projection matrix if $\mathbf{Px}$ is the projection of $\mathbf{x}$ onto $\mathcal{C}(\mathbf{P})$ for any $\mathbf{x}$.

(a) Let λ be the eigenvalue of $\mathbf{P}$. Show that λ is either 1 or 0. (*Hint: you may want to figure out what the eigenspaces corresponding to* $\lambda = 1$ *and* $\lambda = 0$ *are, respectively.*)

(b) Show that $\mathbf{P}$ is a projection matrix if and only if $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P}$ is symmetric.

5. Let $\mathbf{B} \in \mathbb{R}^{m \times s}$ and $\mathcal{C}(\mathbf{B})$ be its column space. Suppose that $\mathcal{C}(\mathbf{B})$ is a proper subspace of $\mathcal{C}(\mathbf{A})$. Is $\mathbf{P_B}(\mathbf{x})$ the same as $\mathbf{P_B}(\mathbf{P_A}(\mathbf{x}))$? Please show your claim rigorously.

Solution 4: Projection

1. ① Strict Convexity:

   Define
   $$f(\mathbf{z}) = \|\mathbf{x}-\mathbf{z}\|_2^2 = \|\mathbf{z}\|_2^2 + 2\langle \mathbf{x}, \mathbf{z}\rangle + \|\mathbf{x}\|_2^2$$

   Then $\nabla^2 f(\mathbf{z}) = 2\mathbf{I_m} > 0 \Longrightarrow f$ is strictly convex.

   The column space $\mathcal{C}(\mathbf{A})$ is convex $\Longrightarrow \forall\, \mathbf{a}, \mathbf{b} \in \mathcal{C}(\mathbf{A}), \forall\, t \in (0,1), t\mathbf{a}+(1-t)\mathbf{b} \in \mathcal{C}(\mathbf{A})$.

   Suppose $\mathbf{z_1}, \mathbf{z_2} \in \mathcal{C}(\mathbf{A})$ are both minimizers of f over $\mathcal{C}(\mathbf{A})$. Then according to the convexity of f,

   $$f(t\mathbf{z_1} + (1-t)\mathbf{z_2}) < tf(\mathbf{z_1}) + (1-t)f(\mathbf{z_2}) = f(\mathbf{z_1}) = f(\mathbf{z_2})$$

   contradicting the minimality of $\mathbf{z_1}$ and $\mathbf{z_2}$. Therefore, the minimizer $\mathbf{P_A}(\mathbf{x})$ is unique.

   ② Orthogonality:

   Suppose $\mathbf{z_0}$ is a minimizer of f over $\mathcal{C}(\mathbf{A})$. For any $\mathbf{y} \in \mathcal{C}(\mathbf{A})$ and $t \in \mathbb{R}$, define

   $$\Phi(t) = \|\mathbf{x} - (\mathbf{z_0} + t\mathbf{y})\|_2^2 = \|\mathbf{x} - \mathbf{z_0}\|_2^2 - 2t\langle \mathbf{x} - \mathbf{z_0}, \mathbf{y}\rangle + t^2\|\mathbf{y}\|_2^2$$

   Notice that $\mathbf{z_0} + t\mathbf{y} \in \mathcal{C}(\mathbf{A})$.

   Since $\mathbf{z_0}$ minimizes f over $\mathcal{C}(\mathbf{A})$, $\Phi(t)$ achieves its minimum at $t = 0 \Longrightarrow \Phi'(0) = -2\langle \mathbf{x} - \mathbf{z_0}, \mathbf{y}\rangle = 0 \Longrightarrow \mathbf{x} - \mathbf{z_0} \perp \mathcal{C}(\mathbf{A})$, i.e., $\mathbf{x} - \mathbf{z_0} \in \mathcal{C}(\mathbf{A})^\perp$.

   Fruthermore, if $\mathbf{x} - \mathbf{z_0} \perp \mathcal{C}(\mathbf{A})$, then for any $\mathbf{y} \in \mathcal{C}(\mathbf{A})$,

   $$\|\mathbf{x} - \mathbf{y}\|_2^2 = \|\mathbf{x} - \mathbf{z_0} + \mathbf{z_0} - \mathbf{y}\|_2^2 = \|\mathbf{x} - \mathbf{z_0}\|_2^2 + \|\mathbf{z_0} - \mathbf{y}\|_2^2 \geq \|\mathbf{x} - \mathbf{z_0}\|_2^2$$

   $\Longrightarrow \mathbf{z_0}$ is a minimizer of f over $\mathcal{C}(\mathbf{A})$.

   If $\mathbf{z_1}$, $\mathbf{z_2}$ both satisfy $\mathbf{x} - \mathbf{z_i} \perp \mathcal{C}(\mathbf{A})$ $(i = 1, 2)$, then $\mathbf{z_1} - \mathbf{z_2} \in \mathcal{C}(\mathbf{A})$ and $\mathbf{z_1} - \mathbf{z_2} = (\mathbf{x} - \mathbf{z_2}) - (\mathbf{x} - \mathbf{z_1}) \perp \mathcal{C}(\mathbf{A}) \Longrightarrow \mathbf{z_1} - \mathbf{z_2} = \mathbf{0} \Longrightarrow \mathbf{z_1} = \mathbf{z_2}$, i.e. the minimizer $\mathbf{P_A}(\mathbf{x})$ is unique.

2. (a)
   $$\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \underset{\substack{\alpha\mathbf{v_1} \\ \alpha\in\mathbb{R}, \mathbf{v_1}\in\mathbb{R}^n}}{\mathbf{argmin}} \|\mathbf{w} - \alpha\mathbf{v}_1\|_2$$

   $$\frac{\mathrm{d}}{\mathrm{d}\alpha}\|\mathbf{w} - \alpha\mathbf{v}_1\|_2^2 = \frac{\mathrm{d}}{\mathrm{d}\alpha}(\mathbf{w} - \alpha\mathbf{v}_1)^\top(\mathbf{w} - \alpha\mathbf{v}_1) = -2\mathbf{v}_1^\top(\mathbf{w} - \alpha\mathbf{v}_1) = 0$$

   $$\Longrightarrow \alpha^* = \frac{\mathbf{v_1^T w}}{\mathbf{v_1^T v_1}} \Longrightarrow \mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \frac{\mathbf{v_1^T w}}{\mathbf{v_1^T v_1}}\mathbf{v}_1$$

   (b) According to the result in (a),

   $$\begin{aligned}\mathbf{P}_{\mathbf{v}_1}(\alpha\mathbf{u} + \beta\mathbf{w}) = \frac{\mathbf{v_1^T}(\alpha\mathbf{u} + \beta\mathbf{w})}{\mathbf{v_1^T v_1}}\mathbf{v}_1 &= \alpha\frac{\mathbf{v_1^T u}}{\mathbf{v_1^T v_1}}\mathbf{v}_1 + \beta\frac{\mathbf{v_1^T w}}{\mathbf{v_1^T v_1}}\mathbf{v}_1 \\ &= \alpha\mathbf{P}_{\mathbf{v}_1}(\mathbf{u}) + \beta\mathbf{P}_{\mathbf{v}_1}(\mathbf{w})\end{aligned}$$

(c)

$$\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \frac{\mathbf{v_1^T w}}{\mathbf{v_1^T v_1}}\mathbf{v}_1 = \left(\frac{\mathbf{v_1 v_1^T}}{\mathbf{v_1^T v_1}}\right)\mathbf{w} := \mathbf{H_1 w}$$

(d) i.

$$\mathbf{P_V}(\mathbf{w}) = \underset{\substack{\mathbf{Vy} \\ \mathbf{y}\in\mathbb{R}^d, \mathbf{V}\in\mathbb{R}^{n\times d}}}{\mathbf{argmin}} \|\mathbf{w} - \mathbf{Vy}\|_2$$

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{y}}\|\mathbf{w} - \mathbf{Vy}\|_2^2 = \frac{\mathrm{d}}{\mathrm{d}\mathbf{y}}(\mathbf{w} - \mathbf{Vy})^\top(\mathbf{w} - \mathbf{Vy}) = -2\mathbf{V}^\top(\mathbf{w} - \mathbf{Vy}) = 0$$

$\mathbf{v_i}$, $i = 1, \ldots, d$ are linearly independent $\Longrightarrow$ $\mathbf{V}$ has full column rank $\Longrightarrow$ $\mathbf{V}^\top\mathbf{V}$ is invertible.

$$\Longrightarrow \mathbf{y}^* = (\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top\mathbf{w} \Longrightarrow \mathbf{P_V}(\mathbf{w}) = \mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top\mathbf{w}$$

$$\Longrightarrow \mathbf{H} = \mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top$$

ii.

$$\mathbf{v_i^T v_j} = 0, \forall i \neq j \Longrightarrow \mathbf{V}^\top\mathbf{V} = \begin{bmatrix} \mathbf{v_1^T v_1} & 0 & \cdots & 0 \\ 0 & \mathbf{v_2^T v_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{v_d^T v_d} \end{bmatrix}$$

$$\Longrightarrow (\mathbf{V}^\top\mathbf{V})^{-1} = \begin{bmatrix} \frac{1}{\mathbf{v_1^T v_1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\mathbf{v_2^T v_2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\mathbf{v_d^T v_d}} \end{bmatrix}$$

$$\Longrightarrow \mathbf{H} = \sum_{i=1}^{d} \frac{\mathbf{v_i v_i^T}}{\mathbf{v_i^T v_i}}$$

3. (a)

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \Longrightarrow \mathbf{x} \in \mathcal{C}(\mathbf{A}) = \mathbb{R}^2 \Longrightarrow \mathbf{P_A}(\mathbf{x}) = \mathbf{x} = \mathbf{A}\cdot\mathbf{x}$$

The coordinates of $\mathbf{P_A}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ are unique and equal to $\mathbf{x}$ itself.

(b)

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} \Longrightarrow \mathcal{C}(\mathbf{A}) = \{\alpha(1,1)^T : \alpha \in \mathbb{R}\}$$

$$\mathbf{P_A}(\mathbf{x}) = \underset{\substack{\alpha(1,1)^T \\ \alpha\in\mathbb{R}}}{\mathbf{argmin}} \|\mathbf{x} - \alpha(1,1)^T\|_2$$

$$\overset{2.(a)}{\Longrightarrow} \mathbf{P_A}(\mathbf{x}) = \frac{(1,1)\mathbf{x}}{(1,1)(1,1)^T}(1,1)^T = \frac{x_1 + x_2}{2}(1,1)^T = \mathbf{A}\cdot(c_1, c_2)^T$$

$\Longrightarrow$ The set of all coordinates of $\mathbf{P_A}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ is the affine line:

$$\left\{(c_1, c_2)^T \in \mathbb{R}^2 : c_1 + 2c_2 = \frac{x_1 + x_2}{2}\right\}$$

Thus the coordinates are not unique.

4. (a) Let λ be an eigenvalue of $\mathbf{P}$ and $\mathbf{v}$ be the corresponding eigenvector. Then $\mathbf{Pv} = \lambda\mathbf{v}$. Since $\mathbf{P}$ is a projection matrix, $\mathbf{Pv}$ is the projection of $\mathbf{v}$ onto $\mathcal{C}(\mathbf{P})$ $\Longrightarrow \mathbf{v} = \mathbf{Pv} + (\mathbf{v} - \mathbf{Pv})$, where $\mathbf{Pv} \in \mathcal{C}(\mathbf{P})$ and $\mathbf{v} - \mathbf{Pv} \in \mathcal{C}(\mathbf{P})^\perp$.

$$\Longrightarrow \|\mathbf{v}\|_2^2 = \|\mathbf{Pv}\|_2^2 + \|\mathbf{v} - \mathbf{Pv}\|_2^2 \xRightarrow{\mathbf{Pv}=\lambda\mathbf{v}} \left(\lambda^2 - \lambda\right)\|\mathbf{v}\|_2^2 = 0 \Longrightarrow \lambda \in \{0, 1\}$$

(b) ① Necessity:
$\forall\, \mathbf{x} \in \mathbb{R}^m$, $\mathbf{Px} \in \mathcal{C}(\mathbf{P})$. We need to prove that $\mathbf{x} - \mathbf{Px} \in \mathcal{C}(\mathbf{P})^\perp$.
$\forall\, \mathbf{y} \in \mathcal{C}(\mathbf{P})$, $\exists\, \mathbf{z} \in \mathbb{R}^m$, s.t. $\mathbf{y} = \mathbf{Pz}$. Then

$$\begin{aligned}\langle \mathbf{x} - \mathbf{Px}, \mathbf{y}\rangle = \langle \mathbf{x} - \mathbf{Px}, \mathbf{Pz}\rangle &= \langle \mathbf{P^T}(\mathbf{x} - \mathbf{Px}), \mathbf{z}\rangle \\ &\xlongequal{\mathbf{P^T}=\mathbf{P}} \langle \mathbf{P}(\mathbf{x} - \mathbf{Px}), \mathbf{z}\rangle \\ &\xlongequal{\mathbf{P^2}=\mathbf{P}} \langle \mathbf{Px} - \mathbf{P}^2\mathbf{x}, \mathbf{z}\rangle = 0\end{aligned}$$

$\Longrightarrow \mathbf{x} - \mathbf{Px} \in \mathcal{C}(\mathbf{P})^\perp \Longrightarrow \mathbf{P}$ is a projection matrix.
② Sufficiency:
$\mathbf{P}$ is a projection matrix $\Longrightarrow \forall\, \mathbf{x} \in \mathbb{R}^m$, $\mathbf{Px} \in \mathcal{C}(\mathbf{P})$
$\Longrightarrow \mathbf{P}(\mathbf{Px}) = \mathbf{argmin}_{\mathbf{z}\in\mathcal{C}(\mathbf{P})} \|\mathbf{Px} - \mathbf{z}\|_2 = \mathbf{Px} \Longrightarrow \mathbf{P^2} = \mathbf{P}$.
$\mathbf{P}$ is a projection matrix $\Longrightarrow \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, $\mathbf{x} = \mathbf{Px} + (\mathbf{x} - \mathbf{Px})$, where $\mathbf{Px} \in \mathcal{C}(\mathbf{P})$ and $\mathbf{x} - \mathbf{Px} \in \mathcal{C}(\mathbf{P})^\perp$. Since $\mathbf{Py} \in \mathcal{C}(\mathbf{P})$, we have

$$\langle \mathbf{Px}, \mathbf{y} - \mathbf{Py}\rangle = 0 \Longrightarrow \langle \mathbf{Px}, \mathbf{y}\rangle = \langle \mathbf{Px}, \mathbf{Py}\rangle = \langle \mathbf{x}, \mathbf{Py}\rangle = \langle \mathbf{P^T x}, \mathbf{y}\rangle$$

$\Longrightarrow \mathbf{P^T} = \mathbf{P}$.

5. $\forall\, \mathbf{x} \in \mathbb{R}^m$,

$$\mathcal{C}(\mathbf{B}) \text{ is a proper subspace of } \mathcal{C}(\mathbf{A}) \Longrightarrow \forall \mathbf{z} \in \mathcal{C}(\mathbf{P}), z \in \mathcal{C}(\mathbf{A}) \Longrightarrow \mathbf{P_A}(\mathbf{x}) - \mathbf{z} \in \mathcal{C}(\mathbf{A})$$

$$\begin{aligned}\mathbf{x} - \mathbf{P_A}(\mathbf{x}) \in \mathcal{C}(\mathbf{A})^\perp \Longrightarrow \|\mathbf{x} - z\|_2^2 &= \|(\mathbf{P_A}(\mathbf{x}) - \mathbf{z}) + (\mathbf{x} - \mathbf{P_A}(\mathbf{x}))\|_2^2 \\ &= \|\mathbf{P_A}(\mathbf{x}) - \mathbf{z}\|_2^2 + \|\mathbf{x} - \mathbf{P_A}(\mathbf{x})\|_2^2\end{aligned}$$

$$\Longrightarrow \underset{\mathbf{z}\in\mathcal{C}(\mathbf{B})}{\mathbf{argmin}} \|\mathbf{x} - \mathbf{z}\|_2 = \underset{\mathbf{z}\in\mathcal{C}(\mathbf{B})}{\mathbf{argmin}} \|\mathbf{P_A}(\mathbf{x}) - \mathbf{z}\|_2 \Longrightarrow \mathbf{P_B}(\mathbf{x}) = \mathbf{P_B}(\mathbf{P_A}(\mathbf{x}))$$

$\blacksquare$

Exercise 5: Derivatives with matrices

Definition 1 (Differentiability)**.** [?] 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) $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{x}$.

2. Consider a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$. The **Jacobian Matrix with denominator layout** is defined by:

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_1} \\ \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \frac{\partial f_2(\mathbf{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}.$$

   Please show that

$$L(\mathbf{x}-\mathbf{x}_0) = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^\top (\mathbf{x}-\mathbf{x}_0),$$

   where $L : \mathbb{R}^n \to \mathbb{R}^m$ is the derivative in Definition 1.

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

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

5. (Optional) 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 $f'(\mathbf{X})$.

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

Solution 5: Derivatives with matrices

1. (a) Let $L(\mathbf{h}) = \mathbf{a}^\top \mathbf{h}$, $\forall \mathbf{h} \in \mathbb{R}^n$, then L is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}$. For any $\mathbf{x}_0 \in \mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{x}_0$,

$$\begin{aligned}\frac{|f(\mathbf{x}) - f(\mathbf{x}_0) - L(\mathbf{x}-\mathbf{x}_0)|}{\|\mathbf{x}-\mathbf{x}_0\|_2} &= \frac{|\mathbf{a}^\top \mathbf{x} - \mathbf{a}^\top \mathbf{x}_0 - \mathbf{a}^\top(\mathbf{x}-\mathbf{x}_0)|}{\|\mathbf{x}-\mathbf{x}_0\|_2} \\ &= \frac{0}{\|\mathbf{x}-\mathbf{x}_0\|_2} = 0\end{aligned}$$

$\Longrightarrow$ f is differentiable at $\mathbf{x}_0$ with derivative L and $f'(\mathbf{x}) = \mathbf{a}^\top$.

(b) Let $L(\mathbf{h}) = 2\mathbf{x}_0^\top \mathbf{h}$, $\forall \mathbf{h} \in \mathbb{R}^n$, then L is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}$. For any $\mathbf{x}_0 \in \mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{x}_0$,

$$\begin{aligned}\frac{|f(\mathbf{x}) - f(\mathbf{x}_0) - L(\mathbf{x}-\mathbf{x}_0)|}{\|\mathbf{x}-\mathbf{x}_0\|_2} &= \frac{|\mathbf{x}^\top \mathbf{x} - \mathbf{x}_0^\top \mathbf{x}_0 - 2\mathbf{x}_0^\top(\mathbf{x}-\mathbf{x}_0)|}{\|\mathbf{x}-\mathbf{x}_0\|_2} \\ &= \frac{|(\mathbf{x}-\mathbf{x}_0)^\top(\mathbf{x}-\mathbf{x}_0)|}{\|\mathbf{x}-\mathbf{x}_0\|_2} \\ &= \frac{\|\mathbf{x}-\mathbf{x}_0\|_2^2}{\|\mathbf{x}-\mathbf{x}_0\|_2} \\ &= \|\mathbf{x}-\mathbf{x}_0\|_2\end{aligned}$$

$\Longrightarrow \lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \frac{|f(\mathbf{x})-f(\mathbf{x}_0)-L(\mathbf{x}-\mathbf{x}_0)|}{\|\mathbf{x}-\mathbf{x}_0\|_2} = 0 \Longrightarrow f$ is differentiable at $\mathbf{x}_0$ with derivative L and $f'(\mathbf{x}) = 2\mathbf{x}^\top$.

Key point: Observe the linear term of $\mathbf{x} - \mathbf{x_0}$ in $f(\mathbf{x}) - f(\mathbf{x_0})$.

2. Denote $\{\mathbf{e_1}, \cdots, \mathbf{e_n}\}$ as the standard basis of $\mathbb{R}^n$.

For a fixed $i \in \{1, \cdots, n\}$, we have

$$\lim_{t\to 0} \frac{f(\mathbf{x_0} + t\mathbf{e_i}) - f(\mathbf{x_0})}{t} = \left(\frac{\partial f_1}{\partial x_i}(\mathbf{x_0}), \frac{\partial f_2}{\partial x_i}(\mathbf{x_0}), \cdots, \frac{\partial f_m}{\partial x_i}(\mathbf{x_0})\right)^\top = L\mathbf{e_i}$$

which is the i-th column of

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_1} \\ \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \frac{\partial f_2(\mathbf{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}.$$

For any $\mathbf{h} = \sum_{i=1}^{n} h_i \mathbf{e_i}$, we have

$$L\mathbf{h} = L\left(\sum_{i=1}^{n} h_i \mathbf{e_i}\right) = \sum_{i=1}^{n} h_i L\mathbf{e_i} = \sum_{i=1}^{n} h_i \left[\left(\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x_0})\right)^{\top}\right]_{\cdot,i} = \left(\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x_0})\right)^{\top} \mathbf{h}$$

Let $\mathbf{h} = \mathbf{x} - \mathbf{x_0}$, we conclude

$$L(\mathbf{x} - \mathbf{x}_0) = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^{\top} (\mathbf{x} - \mathbf{x}_0).$$

3. Let $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ be a function, $\mathbf{X}_0 \in \mathbb{R}^{n\times n}$ be a point, 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

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

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

4. $\forall\ \alpha, \beta > 0$ and $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times m}$,

$$f(\alpha\mathbf{X} + \beta\mathbf{Y}) = \operatorname{tr}\left(\mathbf{A}^{\top}(\alpha\mathbf{X} + \beta\mathbf{Y})\right) = \alpha \operatorname{tr}\left(\mathbf{A}^{\top}\mathbf{X}\right) + \beta \operatorname{tr}\left(\mathbf{A}^{\top}\mathbf{Y}\right) = \alpha f(\mathbf{X}) + \beta f(\mathbf{Y})$$

$\Longrightarrow f$ is linear $\Longrightarrow$ let $L(\mathbf{H}) = \operatorname{tr}(\mathbf{A}^{\top}\mathbf{H})$. It is clear that $f(\mathbf{X}) - f(\mathbf{X_0}) - L(\mathbf{X} - \mathbf{X_0}) = 0$ $\Longrightarrow f$ is differentiable and $f'(\mathbf{X}) = \operatorname{tr}(\mathbf{A}^{\top}\mathbf{X})$.

5.

6.

■

Exercise 6: Linear Space

1. Let $P_n[x]$ be the set of all polynomials on $\mathbb{R}$ with degree at most n. Show that $P_n[x]$ is a linear space.

2. A real symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is called *positive definite*, written $\mathbf{A} \succ \mathbf{0}$, if for all $\mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0}$,
$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0.$$
Let the set of all positive definite matrices be
$$\mathbb{S}^n_{++} := \Big\{ \mathbf{A} \in \mathbb{R}^{n \times n} : \mathbf{A} = \mathbf{A}^\top,\ \mathbf{x}^\top \mathbf{A}\mathbf{x} > 0 \text{ for all } \mathbf{x} \neq \mathbf{0} \Big\}.$$
Is $\mathbb{S}^n_{++}$ a linear subspace of $\mathbb{R}^{n \times n}$? Please show your conclusion in detail.

Solution 6: Linear Space

1. $\forall\ p(x) = \sum\limits_{k=0}^{m} a_k x^k,\ q(x) = \sum\limits_{k=0}^{m} b_k x^k,\ r(x) = \sum\limits_{k=0}^{m} c_k x^k,\ a_k,\ b_k,\ c_k \in \mathbb{R},\ 0 \leq k \leq n,\ \forall\ \alpha,$ $\beta \in \mathbb{R}$:

   (a) Closure under addition:

   $$p(x) + q(x) = \sum_{k=0}^{m} (a_k + b_k) x^k \in P_n[x]$$

   (b) Associativity of addition:

   $$(p+q) + r = \sum_{k=0}^{m} [(a_k + b_k) + c_k] x^k = \sum_{k=0}^{m} [a_k + (b_k + c_k)] x^k = p + (q + r)$$

   (c) Commutativity of addition:

   $$p(x) + q(x) = \sum_{k=0}^{m} (a_k + b_k) x^k = \sum_{k=0}^{m} (b_k + a_k) x^k = q(x) + p(x)$$

   (d) Additive identity:

   $$0 \in P_n[x] \text{ and } p + 0 = 0 + p$$

   (e) Additive Inverse:

   $$-p = \sum_{k=0}^{m} (-a_k) x^k \in P_n[x] \text{ and } p + (-p) = 0$$

   (f) Closure under scalar multiplication:

   $$\alpha p = \sum_{k=0}^{m} (\alpha a_k) x^k \in P_n[x]$$

   (g) Distributivity of scalar over vector addition:

   $$\alpha(p+q) = \sum_{k=0}^{m} [\alpha(a_k + b_k)] x^k = \sum_{k=0}^{m} (\alpha a_k + \alpha b_k) x^k = \alpha p + \alpha q$$

   (h) Distributivity of scalar addition over vector:

   $$(\alpha + \beta) p = \sum_{k=0}^{m} [(\alpha + \beta) a_k] x^k = \sum_{k=0}^{m} (\alpha a_k + \alpha b_k) x^k = \alpha p + \beta p$$

(i) Compatibility of scalar multiplication:

$$\alpha(\beta p) = \sum_{k=0}^{m} [\alpha(\beta a_k)] x^k = \sum_{k=0}^{m} [(\alpha\beta) a_k] x^k = (\alpha\beta) p$$

(j) Unit scalar:

$$1 \in P_n[x] \text{ and } 1 \cdot p = p \cdot 1 = p$$

2. No.

(a)

$$\forall \mathbf{x} \in \mathbb{R}^n, \mathbf{x}^\top \mathbf{0} \mathbf{x} = 0 \Longrightarrow \mathbf{0} \notin \mathbb{S}^n_{++}$$

(b)

$$\forall \alpha < 0, \mathbf{A} \in \mathbb{S}^n_{++}, \mathbf{x}^\top (\alpha A) \mathbf{x} < 0 \Longrightarrow \alpha \mathbf{A} \notin \mathbb{S}^n_{++}$$

■

Exercise 7: 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 $\lambda_1, \lambda_2, \ldots, \lambda_n$.

2. Let $V = \mathbb{R}^n$, $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) \in \mathbb{R}^{n\times n}$ and $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n) \in \mathbb{R}^{n\times n}$. $(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n) = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)\mathbf{P}$, 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 $\mathbf{P}$.

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 $i \in \{1, \ldots, n\}$.

4. Suppose $\mathbf{a} = (1, 0)$, $\mathbf{b} = (0, 1)$ and $\mathbf{c} = (-1, 0)$ are three unit vectors in two-dimensional space. $\mathbf{v} = (x, y)$ is a vector in two-dimensional space.

   (a) Please find the coordinate of $\mathbf{v}$ under basis $\{\mathbf{c}, \mathbf{b}\}$? Is the coordinate unique?

   (b) Please find all the possible combination coefficients of $\mathbf{v}$ under vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$, i.e., $\mathbf{v} = x'\mathbf{a} + y'\mathbf{b} + z'\mathbf{c}$.

   (c) (**Bonus**) Each set of combination coefficients (x', y', z') in (b) forms a vector in $\mathbb{R}^3$. Please find the combination coefficients with minimum ℓ_1-norm.

Solution 7: Basis and Coordinates

1. Suppose $\sum\limits_{i=1}^{n} c_i(\lambda_i \mathbf{a_i}) = \sum\limits_{i=1}^{n} (c_i\lambda_i)\mathbf{a_i} = 0$. Because $\{\mathbf{a_i}\}_{i=1}^{n}$ is linear independent, we must have $c_i\lambda_i = 0 \xrightarrow{\lambda_i \neq 0} c_i = 0$ for all i, i.e., $\{\lambda_i \mathbf{a_i}\}$ is linear independent.

   $\forall$ $\mathbf{v} \in \mathbf{V}$, $\{\mathbf{a_i}\}$ is a basis $\Longrightarrow$ there exists unique scalars $\{\alpha_i\}$ such that

   $$\mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{a_i} = \sum_{i=1}^{n} \left(\frac{\alpha_i}{\lambda_i}\right)(\lambda_i \alpha_i).$$

   Thus $\mathbf{v}$ is a linear combination fo the vectors in $\{\lambda_i \mathbf{a_i}\}$.

   In conclusion, $\{\lambda_i \mathbf{a_i}\}$ is also a basis of an n-dimensional vector space V.

2. ① $|\mathbf{B}| = |\mathbf{AP}| = |\mathbf{A}| \cdot |\mathbf{P}| \neq 0 \Longrightarrow \mathbf{B}$ is invertible $\Longrightarrow$ the columns $\mathbf{b_i}$, $1 \leq i \leq n$ are linearly independent. Being n independent vectors in $\mathbb{R}^n$, they span $\mathbb{R}^n$, i.e., $\{\mathbf{b_i}\}$, $1 \leq i \leq n$ is also a basis of V for any invertible matrix $\mathbf{P}$.

   ② $\forall$ $\mathbf{x} \in \mathbb{R}^n$, there exists a unique $\mathbf{c} \in \mathbb{R}^n$ with $\mathbf{x} = \mathbf{Ac} \xrightarrow{\mathbf{P} \text{ is invertible}} \mathbf{x} = \mathbf{Ac} = \mathbf{APP}^{-1}\mathbf{c} = \mathbf{B}\left(\mathbf{P}^{-1}\mathbf{c}\right) \Longrightarrow \{\mathbf{b_i}\}$, $1 \leq i \leq n$ is also a basis of V for any invertible matrix $\mathbf{P}$.

3. (a) Set $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ as the coordinate of $\mathbf{v}$ under $\{\lambda_i \mathbf{a_i}\}$, $1 \leq i \leq n$.

   $$\mathbf{v} = \sum_{i=1}^{n} y_i(\lambda_i \mathbf{a_i}) = \sum_{i=1}^{n} x_i \mathbf{a_i} \Longrightarrow y_i = \frac{x_i}{\lambda_i}$$

   Therefore, the coordinate of $\mathbf{v}$ under $\{\lambda_i \mathbf{a_i}\}$, $1 \leq i \leq n$ is $\mathbf{y} = \left(\frac{x_1}{\lambda_1}, \frac{x_2}{\lambda_2}, \cdots, \frac{x_n}{\lambda_n}\right)$.

   (b) It is clear that $\mathbf{x} = (1, 1, \cdots, 1)^\top$ and $\mathbf{y} = \left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \cdots, \frac{1}{\lambda_n}\right)$.

4. (a) It is clear that the coordinate of $\mathbf{v}$ under basis $\{\mathbf{c}, \mathbf{b}\}$ is $(-x, y)$. And the coordinate is unique because $\{\mathbf{c}, \mathbf{b}\}$ is a basis of $\mathbb{R}^2$.

   (b) $\mathbf{v} = x'\mathbf{a} + y'\mathbf{b} + z'\mathbf{c} = (x' - z', y') = (x, y) \Longrightarrow (x', y', z') = (x + t, y, t)$, $\forall$ $t \in \mathbb{R}$.

   (c)

   $$\min \|(x', y', z')\|_1 = \min |x'| + |y'| + |z'| = \min |x'| + |z'| = \min |x + t| + |t|$$

   $$\begin{cases} |x + t| + |t| \geq |x + t - t| = |x| \\ |x + t| + |t| = |x| \text{ when } t \in [-x, 0] \text{ and } x \geq 0 \text{ or } t \in [0, -x] \text{ and } x < 0 \end{cases}$$

   In conclusion, $\min \|(x', y', z')\|_1 = |x| + |y|$.

■

Exercise 8: 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) $\mathbf{rank}(\mathbf{A}) = \mathbf{rank}(\mathbf{A}^\top) = \mathbf{rank}(\mathbf{A}^\top\mathbf{A}) = \mathbf{rank}(\mathbf{A}\mathbf{A}^\top)$;

   (b) $\mathbf{rank}(\mathbf{AB}) \leq \mathbf{rank}(\mathbf{A})$; (please give an example when the equality holds)

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

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

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

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

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

   Please show that

   (a) $\mathbf{rank}(\mathbf{A}) = \dim(\mathcal{C}(\mathbf{A}))$;

   (b) $\mathbf{rank}(\mathbf{A}) + \dim(\mathcal{N}(\mathbf{A})) = n$.

3. Given that

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

   Please show the results in 1.(b) by Eq. (1).

Solution 8: Rank of matrices

1. (a) The column rank of a matrix equals its row rank $\Longrightarrow \mathbf{rank}(\mathbf{A}) = \mathbf{rank}(\mathbf{A})$.
   For $\mathbf{x} \in \mathbb{R}^n$,

$$\begin{cases}(\mathbf{Ax} = \mathbf{0} \Longrightarrow \mathbf{A^T A x} = \mathbf{0}) \Longrightarrow \mathcal{N}(\mathbf{A^T A}) \supset \mathcal{N}(\mathbf{A}) \\ (\mathbf{A^T A x} = \mathbf{0} \Longrightarrow \|\mathbf{Ax}\|_2^2 = \mathbf{x^T A^T A x} = \mathbf{0} \Longrightarrow \mathbf{Ax} = \mathbf{0}) \Longrightarrow \mathcal{N}(\mathbf{A^T A}) \subset \mathcal{N}(\mathbf{A})\end{cases}$$

$$\Longrightarrow \mathcal{N}(\mathbf{A^T A}) = \mathcal{N}(\mathbf{A})$$

$$\Longrightarrow \mathbf{rank}(\mathbf{A^T A}) = n - \dim \mathbf{A^T A} = n - \dim \mathbf{A} = \mathbf{rank}(\mathbf{A})$$

   In the same way, $\mathbf{rank}(\mathbf{AA^T}) = \mathbf{rank}(\mathbf{A})$.
   In conclusion, $\mathbf{rank}(\mathbf{A}) = \mathbf{rank}(\mathbf{A}^\top) = \mathbf{rank}(\mathbf{A}^\top\mathbf{A}) = \mathbf{rank}(\mathbf{A}\mathbf{A}^\top)$.

   (b) $\mathrm{Im}(\mathbf{AB}) = \mathbf{A}(\mathrm{Im}(\mathbf{B})) \subset \mathrm{Im}(\mathbf{A}) \Longrightarrow \mathbf{rank}(\mathbf{AB}) = \dim(\mathrm{Im}(\mathbf{AB})) \leq \dim(\mathrm{Im}(\mathbf{A})) = \mathbf{rank}(\mathbf{A})$.
   E.g.:
   Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} = I_n$. Then $\mathbf{AB} = \mathbf{A} \Longrightarrow \mathbf{rank}(\mathbf{AB}) = \mathbf{rank}(\mathbf{A})$.

2. (a) By definition, $\mathbf{rank}(\mathbf{A})$ is the maximal number of linearly independent columns of $\mathbf{A}$ and the dimension of the span of a finite set of vectors equals the maximal size of a linearly independent subset of that set. Therefore, $\mathbf{rank}(\mathbf{A}) = \dim(\mathcal{C}(\mathbf{A}))$.

   (b) Let $\{\mathbf{v_1}, \cdots, \mathbf{v_k}\}$ be a basis of $\mathcal{N}(\mathbf{A})$, where $k = \dim(\mathcal{N}(\mathbf{A}))$. We can extend this basis to a basis of $\mathbb{R}^n$, i.e., there exist vectors $\{\mathbf{w_1}, \cdots, \mathbf{w_{n-k}}\}$ such that $\{\mathbf{v_1}, \cdots, \mathbf{v_k}, \mathbf{w_1}, \cdots, \mathbf{w_{n-k}}\}$ is a basis of $\mathbb{R}^n$.
   We claim that $\{\mathbf{Aw_1}, \cdots, \mathbf{Aw_{n-k}}\}$ is a basis of $\mathcal{C}(\mathbf{A})$.
   Spanning: $\forall\ \mathbf{y} \in \mathcal{C}(\mathbf{A})$, there exists $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{y} = \mathbf{Ax}$.
   Since $\{\mathbf{v_1}, \cdots, \mathbf{v_k}, \mathbf{w_1}, \cdots, \mathbf{w_{n-k}}\}$ is a basis of $\mathbb{R}^n$, there exist scalars $\{\alpha_i\}_{i=1}^k$ and $\{\beta_j\}_{j=1}^{n-k}$ such that

$$\mathbf{x} = \sum_{i=1}^k \alpha_i \mathbf{v_i} + \sum_{j=1}^{n-k} \beta_j \mathbf{w_j}$$

$$\Longrightarrow \mathbf{y} = \mathbf{Ax} = \sum_{i=1}^k \alpha_i \mathbf{Av_i} + \sum_{j=1}^{n-k} \beta_j \mathbf{Aw_j} = \sum_{j=1}^{n-k} \beta_j \mathbf{Aw_j}$$

   $\Longrightarrow \mathbf{y}$ is a linear combination of $\{\mathbf{Aw_1}, \cdots, \mathbf{Aw_{n-k}}\}$.
   Linear independence: Suppose $\sum\limits_{j=1}^{n-k} \gamma_j \mathbf{Aw_j} = \mathbf{0}$. Then $\mathbf{A}(\sum\limits_{j=1}^{n-k} \gamma_j \mathbf{w_j}) = \mathbf{0} \Longrightarrow \sum\limits_{j=1}^{n-k} \gamma_j \mathbf{w_j} \in \mathcal{N}(\mathbf{A})$. Since $\{\mathbf{v_1}, \cdots, \mathbf{v_k}, \mathbf{w_1}, \cdots, \mathbf{w_{n-k}}\}$ is a basis of $\mathbb{R}^n$, the vectors $\{\mathbf{w_1}, \cdots, \mathbf{w_{n-k}}\}$ are linearly independent and none of them can be expressed as a linear combination of $\{\mathbf{v_1}, \cdots, \mathbf{v_k}\}$. Therefore, $\sum\limits_{j=1}^{n-k} \gamma_j \mathbf{w_j} = \mathbf{0} \Longrightarrow \gamma_j = 0$ for all j.

In conclusion, $\{\mathbf{Aw_1}, \cdots, \mathbf{Aw_{n-k}}\}$ is a basis of $\mathcal{C}(\mathbf{A}) \Longrightarrow \dim(\mathcal{C}(\mathbf{A})) = n - k = n - \dim(\mathcal{N}(\mathbf{A})) \Longrightarrow \mathbf{rank}(\mathbf{A}) + \dim(\mathcal{N}(\mathbf{A})) = n$.

3.

$$\begin{aligned}\mathbf{rank}(\mathbf{AB}) \overset{1.(a)}{=\!=\!=} \mathbf{rank}(\mathbf{B^T A^T}) \overset{1.(b)}{=\!=\!=} \mathbf{rank}(\mathbf{A^T}) - \dim(\mathcal{C}(\mathbf{A^T}) \cap \mathcal{N}(\mathbf{B^T}))\\ \leq \mathbf{rank}(\mathbf{A^T}) \overset{1.(a)}{=\!=\!=} \mathbf{rank}(\mathbf{A})\end{aligned}$$

■

Exercise 9: Properties of Eigenvalues and Singular Values

1. Suppose the maximum eigenvalue, minimum eigenvalue and maximum singular value 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

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

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

   (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

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

Solution 9: Properties of Eigenvalues and Singular Values

1. Define Rayleigh quotient:
$$R(x) = \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$
Note that $R(\alpha \mathbf{x}) = R(\mathbf{x})$ for any nonzero scalar α. Hence
$$\sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq 0} R(x) = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x}, \qquad \inf_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq 0} R(x) = \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x}$$
Because $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{A}\mathbf{x}$ is continuous and the unit sphere is compact, both extrema are attained.

   By the spectral theorem, there exists an orthogonal matrix $\mathbf{Q}$ and a diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ with real eigenvalues λ_i such that $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$.

   Without generality, order the eigenvalues so that $\lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = \lambda_{\max}$.

   For any $\mathbf{x} \in \mathbb{R}^n$,
$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top \mathbf{x} := \mathbf{y}^\top \mathbf{\Lambda}\mathbf{y} = \sum_{i=1}^n \lambda_i y_i^2, \qquad \mathbf{x}^\top \mathbf{x} = (\mathbf{Q}\mathbf{y})^\top (\mathbf{Q}\mathbf{y}) = \mathbf{y}^\top \mathbf{y}$$
$$\implies R(x) = \frac{\sum_{i=1}^n \lambda_i y_i^2}{\sum_{i=1}^n y_i^2} \in [\lambda_{\min}, \lambda_{\max}]$$
Let $\mathbf{v}_{\max}$ be a unit eigenvector of $\mathbf{A}$ associated with $\lambda_{\max}$, then
$$R(\mathbf{v}_{\max}) = \frac{\mathbf{v}_{\max}{}^\top \mathbf{A}\mathbf{v}_{\max}}{\mathbf{v}_{\max}{}^\top \mathbf{v}_{\max}} = \lambda_{\max}.$$
Similarly, let $\mathbf{v}_{\min}$ be a unit eigenvector associated with $\lambda_{\min}$, then $R(\mathbf{v}_{\min}) = \lambda_{\min}$.

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

2. (a)
$$\|\mathbf{B}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{B}\mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2=1} [(\mathbf{B}\mathbf{x})^\top \mathbf{B}\mathbf{x}]^{\frac{1}{2}} = \max_{\|\mathbf{x}\|_2=1} [\mathbf{x}^\top (\mathbf{B}^\mathbf{T}\mathbf{B})\mathbf{x}]^{\frac{1}{2}}$$
Notice that $\mathbf{B}^\mathbf{T}\mathbf{B}$ is symmetric and positive semi-definite. Without generality, let its eigenvalues be
$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0.$$

and let their corresponding orthogonal normalized eigenvectors be $\mathbf{v_1}, \cdots, \mathbf{v_n} \in \mathbb{R}^n$. Then for any $\mathbf{x} \in \mathbb{R}^n$ with $\|\mathbf{x}\|_2 = 1$, there exist scalars $\alpha_1, \cdots, \alpha_n$ such that

$$\mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{v_i}, \qquad \sum_{i=1}^{n} \alpha_i^2 = 1$$

$$\Longrightarrow \mathbf{x}^\top (\mathbf{B^T B}) \mathbf{x} = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \le \lambda_1 \sum_{i=1}^{n} \alpha_i^2 = \lambda_1$$

Besides, let $\mathbf{x} = \mathbf{v_1}$, then $\mathbf{x}^\top (\mathbf{B^T B}) \mathbf{x} = \lambda_1$.
In conclusion, $\|\mathbf{B}\|_2 = \sqrt{\lambda_1} = \sigma_{\max}(\mathbf{B})$.

(b) By Cauchy-Schwarz inequality,

$$\mathbf{x}^\top \mathbf{B} \mathbf{y} \le \|\mathbf{x}\|_2 \|\mathbf{B}\mathbf{y}\|_2 \le \|\mathbf{x}\|_2 \|\mathbf{B}\|_2 \|\mathbf{y}\|_2 = \|\mathbf{x}\|_2 \sigma_{\max}(\mathbf{B}) \|\mathbf{y}\|_2$$

Let $\mathbf{x} = \mathbf{u_1}$ and $\mathbf{y} = \mathbf{v_1}$, where $\mathbf{u_1}$ and $\mathbf{v_1}$ are the left and right singular vectors of $\mathbf{B}$ associated with $\sigma_{\max}(\mathbf{B})$. Then

$$\mathbf{x}^\top \mathbf{B} \mathbf{y} = \mathbf{u_1}^\top \mathbf{B} \mathbf{v_1} = \sigma_{\max}(\mathbf{B}) = \|\mathbf{u_1}\|_2 \sigma_{\max}(\mathbf{B}) \|\mathbf{v_1}\|_2$$

In conclusion,

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

$\blacksquare$

Exercise 10: Matrix SVD Decomposition and Pseudoinverse

1. For any real matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, the **Moore-Penrose generalized inverse** (or pseudoinverse) of $\mathbf{A}$, denoted by $\mathbf{A}^+ \in \mathbb{R}^{m \times n}$, is a matrix that satisfies the following four conditions:

   (a) $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$ (Consistency condition)

   (b) $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$ (Reflexivity condition)

   (c) $(\mathbf{A}\mathbf{A}^+)^\top = \mathbf{A}\mathbf{A}^+$ (Symmetry condition 1)

   (d) $(\mathbf{A}^+\mathbf{A})^\top = \mathbf{A}^+\mathbf{A}$ (Symmetry condition 2)

   Suppose that the matrix $\mathbf{A}$ can be decomposed via Singular Value Decomposition (SVD) as $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, Please show that $\mathbf{A}^+ = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^\top$, where $\mathbf{\Sigma}^+ \in \mathbb{R}^{m \times n}$ is defined by:

$$\mathbf{\Sigma}^+_{ij} = \begin{cases} \frac{1}{\mathbf{\Sigma}_{ii}} & \text{if } i = j \text{ and } \mathbf{\Sigma}_{ii} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

2. **(Optional)** Please show that $\mathbf{A}^+$ is unique for any matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$.

3. Consider the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{n \times m}$, $\mathbf{x} \in \mathbb{R}^m$, and $\mathbf{b} \in \mathbb{R}^n$. Please show that if the system has no solution (i.e., $\mathbf{b}$ is not in the column space of $\mathbf{A}$), the least squares solution to the system

$$\arg\min_{\mathbf{x} \in \mathbb{R}^m} \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2,$$

   is given by $\mathbf{x} = \mathbf{A}^+\mathbf{b}$, where $\mathbf{A}^+ \in \mathbb{R}^{m \times n}$ is the Moore-Penrose generalized inverse of matrix $\mathbf{A}$ defined above.

   (**Hint**: For any orthogonal matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$ and vector $\mathbf{x} \in \mathbb{R}^n$, then $\|\mathbf{U}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$)

Solution 10: Matrix SVD Decomposition and Pseudoinverse

1. (a)
$$\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}\mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{\top}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^{+}\mathbf{\Sigma}\mathbf{V}^{\top} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top} = \mathbf{A}$$
   (b)
$$\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{\top}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}\mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{\top} = \mathbf{V}\mathbf{\Sigma}^{+}\mathbf{\Sigma}\mathbf{\Sigma}^{+}\mathbf{U}^{\top} = \mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{\top} = \mathbf{A}^{+}$$
   (c)
$$(\mathbf{A}\mathbf{A}^{+})^{\top} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}\mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{\top})^{\top} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^{+}\mathbf{U}^{\top} = \mathbf{A}\mathbf{A}^{+}$$
   (d)
$$(\mathbf{A}^{+}\mathbf{A})^{\top} = (\mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{\top}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top})^{\top} = \mathbf{V}\mathbf{\Sigma}^{+}\mathbf{\Sigma}\mathbf{V}^{\top} = \mathbf{A}^{+}\mathbf{A}$$

2. Assume that there exist two pseudoinverses $\mathbf{B}$ and $\mathbf{C}$ of $\mathbf{A}$. Set $\mathbf{X} = \mathbf{B} - \mathbf{C}$.

   Given that $\mathbf{B}$ and $\mathbf{C}$ satisfy the four Moore-Penrose conditions:
$$\mathbf{ABA} = \mathbf{A}, \qquad \mathbf{BAB} = \mathbf{B}, \qquad (\mathbf{AB})^{\top} = \mathbf{AB}, \qquad (\mathbf{BA})^{\top} = \mathbf{BA}$$
$$\mathbf{ACA} = \mathbf{A}, \qquad \mathbf{CAC} = \mathbf{C}, \qquad (\mathbf{AC})^{\top} = \mathbf{AC}, \qquad (\mathbf{CA})^{\top} = \mathbf{CA}$$
   Then
$$\mathbf{AXA} = \mathbf{ABA} - \mathbf{ACA} = \mathbf{A} - \mathbf{A} = \mathbf{0}$$
$$\mathbf{AX} = \mathbf{AB} - \mathbf{AC} = (\mathbf{AB})^{\top} - (\mathbf{AC})^{\top} = (\mathbf{AB} - \mathbf{AC})^{\top} = (\mathbf{AX})^{\top}$$
$$\mathbf{XA} = \mathbf{BA} - \mathbf{CA} = (\mathbf{BA})^{\top} - (\mathbf{CA})^{\top} = (\mathbf{BA} - \mathbf{CA})^{\top} = (\mathbf{XA})^{\top}$$
$$\Longrightarrow \begin{cases} \|\mathbf{AX}\|_2^2 = \operatorname{tr}\left((\mathbf{AX})^{\top}\mathbf{AX}\right) = \operatorname{tr}(\mathbf{AXAX}) = \operatorname{tr}((\mathbf{AXA})\mathbf{X}) = \operatorname{tr}(\mathbf{0X}) = 0 \\ \|\mathbf{XA}\|_2^2 = \operatorname{tr}\left((\mathbf{XA})^{\top}\mathbf{XA}\right) = \operatorname{tr}(\mathbf{XAXA}) = \operatorname{tr}(\mathbf{X}(\mathbf{AXA})) = \operatorname{tr}(\mathbf{X0}) = 0 \end{cases}$$
$$\Longrightarrow \begin{cases} \mathbf{AX} = \mathbf{0} \\ \mathbf{XA} = \mathbf{0} \end{cases}$$
$$\begin{aligned} \Longrightarrow \mathbf{X} = \mathbf{B} - \mathbf{C} &= \mathbf{BAB} - \mathbf{CAC} \\ &= \mathbf{BAB} - \mathbf{CAB} + \mathbf{CAB} - \mathbf{CAC} \\ &= (\mathbf{B} - \mathbf{C})\mathbf{AB} + \mathbf{CA}(\mathbf{B} - \mathbf{C}) \\ &= \mathbf{XAB} + \mathbf{CAX} = \mathbf{0B} + \mathbf{C0} = \mathbf{0} \end{aligned}$$
$$\Longrightarrow \mathbf{B} = \mathbf{C}$$
   In conclusion, the pseudoinverse $\mathbf{A}^{+}$ is unique.

3. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$ be the SVD of $\mathbf{A}$, where $\mathbf{U} \in \mathbb{R}^{n\times n}$ and $\mathbf{V} \in \mathbb{R}^{m\times m}$ are orthogonal matrices, and $\mathbf{\Sigma} \in \mathbb{R}^{n\times m}$ is a diagonal matrix with singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$ on the diagonal (where $r = \operatorname{rank}(\mathbf{A})$).

   Then
$$\|\mathbf{Ax} - \mathbf{b}\|_2^2 = \|\mathbf{U}^{\top}(\mathbf{Ax} - \mathbf{b})\|_2^2 = \|\mathbf{\Sigma}\mathbf{V}^{\top}\mathbf{x} - \mathbf{U}^{\top}\mathbf{b}\|_2^2$$

Let $\mathbf{y} = \mathbf{V}^\top \mathbf{x}$ and $\mathbf{c} = \mathbf{U}^\top \mathbf{b}$. Then the problem reduces to

$$\min_{\mathbf{y}\in\mathbb{R}^m} \|\mathbf{\Sigma y} - \mathbf{c}\|_2^2 = \min_{\mathbf{y}\in\mathbb{R}^m} \sum_{i=1}^{r} (\sigma_i y_i - c_i)^2 + \sum_{i=r+1}^{n} c_i^2$$

Note that the second term $\sum_{i=r+1}^{n} c_i^2$ is constant with respect to $\mathbf{y}$. Therefore, we only need to minimize the first term, which is minimized when $y_i = \frac{c_i}{\sigma_i}$ for $i = 1, 2, \dots, r$ and y_i can be any value for $i = r+1, r+2, \dots, m$.

Hence, the set of least squares minimizers is

$$\mathbf{x} = \mathbf{V}\mathbf{y}, \qquad \mathbf{y} = (\frac{c_1}{\sigma_1}, \frac{c_2}{\sigma_2}, \dots, \frac{c_r}{\sigma_r}, y_{r+1}, \dots, y_m)^\top, y_{r+1}, \dots, y_m \in \mathbb{R}$$

Let $y_i = 0$, $i = r+1, r+2, \dots, m$. Then the least-norm solution is

$$\mathbf{x} = \mathbf{V}(\frac{c_1}{\sigma_1}, \frac{c_2}{\sigma_2}, \dots, \frac{c_r}{\sigma_r}, 0, \dots, 0)^\top = \mathbf{V\Sigma^+U^\top b} = \mathbf{A^+ b}$$

■