# A Complete Solution Guide to Introduction to Nonlinear Optimization Theory, Algorithms, and Applications with MATLAB

Youming Zhao
Email: youming0.zhao@gmail.com

First draft: May 24, 2022 Last update: October 14, 2023

## Contents

# 1 Chapter 1 Mathematical Preliminaries

## 1.1 Some important concepts

### 1.1.1 Induced matrix norm and several equivalent definitions

Here we introduce the definition of the induced matrix norm from the textbook. That is, the induced matrix norm $\|A\|_{a,b}$ is defined by

$$\|\mathbf{A}\|_{a,b} = \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_b \colon \|\mathbf{x}\|_a \leq 1\}. \tag{1}$$

$\|\mathbf{A}\|_{a,b}$ can also be computed in the following alternative ways (Horn and Johnson, 2013, p. 343, Definition 5.6.1):

$$\|\mathbf{A}\|_{a,b} = \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_b \colon \|\mathbf{x}\|_a = 1\} = \max_{\|\mathbf{x}\|_a \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a}. \tag{2}$$

Now we show that they are valid alternatives of (1) by proving two lemmas. The first alternative is exactly the following lemma.

**Lemma 1.1.** *The maximum points* $\mathbf{x}^*$ *of the RHS of* (1) *must satisfy* $\|\mathbf{x}^*\|_a = 1$.

*Proof.* We will prove it by contradiction. Given $\mathbf{A} \neq \mathbf{0}$, it is obvious that $\mathbf{x}^* \neq \mathbf{0}$ must hold, otherwise $\|\mathbf{A}\mathbf{x}^*\|_b = 0$ which is the minimum value and it is easy to find an $\mathbf{x}$ such that $\|\mathbf{A}\mathbf{x}\|_b > 0$. Suppose that the maximum points satisfy $\|\mathbf{x}^*\|_a < 1$, then there exists real numbers $k$ such that $\|k\mathbf{x}^*\|_a = 1$ in which $|k| = 1/\|\mathbf{x}^*\|_a > 1$. Let $\mathbf{y} = k\mathbf{x}^*$, then we get

$$\|\mathbf{A}\mathbf{y}\|_b = \|\mathbf{A}(k\mathbf{x}^*)\|_b = |k|\|\mathbf{A}\mathbf{x}^*\|_b > \|\mathbf{A}\mathbf{x}^*\|_b \tag{3}$$

which contradicts that $\mathbf{x}^*$ are the maximum points. Thus, $\|\mathbf{x}^*\|_a = 1$ holds. □

We directly present the second alternative as a lemma as follows and prove it through Lemma 1.1.

**Lemma 1.2.** *For any* $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{A}\|_{a,b} = \max_{\|\mathbf{x}\|_a \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a}. \tag{4}$$

*Proof.* An equivalent form of Lemma 1.1 is

$$\|\mathbf{A}\|_{a,b} = \max_{\mathbf{y}}\left\{\frac{\|\mathbf{A}\mathbf{y}\|_b}{\|\mathbf{y}\|_a} : \|\mathbf{y}\|_a = 1\right\} = \max_{\|\mathbf{y}\|_a = 1} \frac{\|\mathbf{A}\mathbf{y}\|_b}{\|\mathbf{y}\|_a}. \tag{5}$$

By letting $\mathbf{y} = k\mathbf{x}$ where $k \in \mathbb{R}\backslash\{0\}$, we have

$$\|\mathbf{A}\|_{a,b} = \max_{|k|\|\mathbf{x}\|_a = 1} \frac{|k|\|\mathbf{A}\mathbf{x}\|_b}{|k|\|\mathbf{x}\|_a} = \max_{\|\mathbf{x}\|_a = 1/|k|} \frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a} = \max_{\|\mathbf{x}\|_a \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a} \tag{6}$$

where the last equality follows from that $k$ is an *arbitrary* nonnegative real number. This completes our proof. □

The textbook gives a result about the induced matrix norm without a proof right after its definition. Here, we will present it as a proposition with a proof. The proof is an immediate result of Lemma 4.

**Proposition 1.3.** *For any* $\mathbf{x} \in \mathbb{R}^n$ *the inequality*

$$\|\mathbf{A}\mathbf{x}\|_b \leq \|\mathbf{A}\|_{a,b}\|\mathbf{x}\|_a \tag{7}$$

*holds.*

*Proof.* According to Lemma 4, for any $\mathbf{x} \neq \mathbf{0}$, it follows that

$$\frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a} \leq \|\mathbf{A}\|_{a,b} \Longleftrightarrow \|\mathbf{A}\mathbf{x}\|_b \leq \|\mathbf{A}\|_{a,b}\|\mathbf{x}\|_a \tag{8}$$

completing the proof. □

### 1.1.2 Accumulation point

**Definition 1.4 (accumulation points).** *If any open ball of a point $x$ contains infinitely many points of a set $S$, then $x$ is called an accumulation point of $S$. The set of all accumulation points of $S$ is denoted by $S'$.*

### 1.1.3 Closed set

We describe the definition of closed sets in a slightly different way than the textbook. However, in essence, they are the same thing.

**Definition 1.5 (closed sets).** *If a set $S$ contains all of its accumulation points, then we call $S$ a closed set.*

### 1.1.4 Boundary point

**Definition 1.6 (boundary points).** *Given a set $U \subseteq \mathbb{R}^n$, a* ***boundary point*** *of $U$ is a point $\mathbf{x} \in \mathbb{R}^n$ satisfying the following: any neighborhood of $\mathbf{x}$ contains at least one point in $U$ and at least one point in its complement $U^c$. The set of all boundary points of a set is denoted by $\mathrm{bd}(U)$ or $\partial U$ and is called the boundary of $U$.*

### 1.1.5 Closure

**Definition 1.7 (closure of a set).** *The closure of a set $U \subseteq \mathbb{R}^n$ is the smallest closed set containing $U$:*

$$\mathrm{cl}(U) = \bigcap \{T : U \subseteq T,\ T \text{ is closed}\}. \tag{9}$$

*Another equivalent definition of $\mathrm{cl}(U)$ is given by*

$$\mathrm{cl}(U) = U \cup \mathrm{bd}(U). \tag{10}$$

The closure set is indeed a closed set as an intersection of closed sets (see Exercise 1.16(ii)).

### 1.1.6 Interior point and interior of a set

**Definition 1.8 (interior points).** *Given a set $U \subseteq \mathbb{R}^n$, a point $\mathbf{c} \in U$ is an interior point of $U$ if there exists $r > 0$ for which $B(\mathbf{c}, r) \subseteq U$.*

**Definition 1.9 (interior of a set).** *The set of all interior points of a given set $U$ is called the interior of a set and is denoted by* $\text{int}(U)$*:*

$$\text{int}(U) = \{\mathbf{x} \in U : B(\mathbf{x}, r) \subseteq U \textit{ for some } r > 0\}. \tag{11}$$

### 1.1.7 De Morgan's Law/Theorem

Here we present a generalized form of De Morgan's Law which is also known as De Morgan's Theorem from Wikipedia[1].

**Theorem 1.10 (De Morgan's Law/Theorem).**

$$\left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c \tag{12}$$

$$\left(\bigcap_{i\in I} A_i\right)^c = \bigcup_{i\in I} A_i^c \tag{13}$$

*where $I$ is some, possibly countably or uncountably infinite, indexing set.*

## 1.2 Exercises

**Exercise 1.1**

Show that $\|\cdot\|_{1/2}$ is not a norm.

*Proof.* To show that a function is not a norm, it suffices to find a counterexample which does not satisfy at least one of the three properties of a norm. For $\|\cdot\|_{1/2}$, we let

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Then we have

$$\begin{aligned}
\|\mathbf{x}+\mathbf{y}\|_{1/2} &= \left\| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\|_{1/2} = (\sqrt{1}+\sqrt{1})^2 = 4 \\
\|\mathbf{x}\|_{1/2} &= (\sqrt{1}+\sqrt{0})^2 = 1 \\
\|\mathbf{y}\|_{1/2} &= (\sqrt{0}+\sqrt{1})^2 = 1
\end{aligned}$$

However,

$$\|\mathbf{x}+\mathbf{y}\|_{1/2} = 4 > \|\mathbf{x}\|_{1/2} + \|\mathbf{y}\|_{1/2} = 1 + 1 = 2.$$

Hence, $\|\cdot\|_{1/2}$ does not satisfy the triangle inequality. This completes the proof. □

In fact, when $0 < p < 1$, $\|\cdot\|_p$ satisfies the reverse of Minkowski's inequality within the domain of $\mathbb{R}_+^n$. Formally, we have the following theorem.

[1] https://en.wikipedia.org/wiki/De_Morgan%27s_laws

**Theorem 1.11 (reversed Minkowski's inequality).** *For any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_+$ *and* $0 < p < 1$, *the following inequality*

$$\|\mathbf{x} + \mathbf{y}\|_p \geq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$$

*holds.*

64

65 The following proof largely follows Jax (2016) but in greater detail.

*Proof.* Obviously, the claim holds when either $\mathbf{x} = 0$ or $\mathbf{y} = 0$. We only need to consider the case when $\mathbf{x} \neq 0$ and $\mathbf{y} \neq 0$, which guarantees $\|\mathbf{x} + \mathbf{y}\|_p \neq 0$. Let $f(x) = x^p$ with $x > 0$ and $0 < p < 1$. Since $f''(x) = p(p-1)x^{p-2} < 0$ for any $x > 0$, $f(x)$ is concave. Thus, we have

$$\begin{aligned}
(x_i + y_i)^p &= \left(t \cdot \frac{x_i}{t} + (1-t) \cdot \frac{y_i}{1-t}\right)^p, \quad 0 < t < 1, i \in \{1, 2, \dots, n\} \\
&\geq t \cdot \frac{x_i^p}{t^p} + (1-t) \cdot \frac{y_i^p}{(1-t)^p}.
\end{aligned}$$

Taking summation over $i$ gives

$$\begin{aligned}
\sum_{i=1}^n (x_i + y_i)^p &\geq t \sum_{i=1}^n \frac{x_i^p}{t^p} + \frac{y_i^p}{(1-t)^p} \\
\|\mathbf{x} + \mathbf{y}\|_p^p &\geq t \frac{\|\mathbf{x}\|_p^p}{t^p} + (1-t) \frac{\|\mathbf{y}\|_p^p}{(1-t)^p}
\end{aligned}$$

66 Letting $t = \frac{\|\mathbf{x}\|_p}{\|\mathbf{x}\|_p + \|\mathbf{y}\|_p}$ yields

$$\begin{aligned}
\|\mathbf{x} + \mathbf{y}\|_p^p &\geq t \frac{\|\mathbf{x}\|_p^p}{\frac{\|\mathbf{x}\|_p^p}{(\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)^p}} + (1-t) \frac{\|\mathbf{y}\|_p^p}{\frac{\|\mathbf{y}\|_p^p}{(\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)^p}} \\
&= t(\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)^p + (1-t)(\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)^p \\
&= (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)^p \\
\implies \quad \|\mathbf{x} + \mathbf{y}\|_p &\geq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p,
\end{aligned}$$

67 as desired. □

*Remark* 1.12. You may observe that the reversed Minkowski's inequality does not hold when $\mathbf{x} = -\mathbf{y} \neq 0$. The reason is that in the above proof, the condition $x_i, y_i \geq 0, \forall i$ is required to ensure that $f(x)$ is concave and well defined. Concretely speaking, $\sqrt[3]{x}$ is convex on $\mathbb{R}_-$ and $\sqrt[4]{x}$ is not well defined on $\mathbb{R}_-$. Hence, the reversed Minkowski's inequality only works for both vectors with nonnegative entries. Note that Minkowski's inequality works not only for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ but also for $\mathbf{x}, \mathbf{y} \in \mathbf{C}^n$.

68

**Extensions**

Since $\|\cdot\|_0$ does not satisfy the positive homogeneity, it is not a true norm.

69

**Exercise 1.2**

Prove that for any $\mathbf{x} \in \mathbb{R}^n$ one has

$$\|\mathbf{x}\|_\infty = \lim_{p \to \infty} \|\mathbf{x}\|_p.$$

70

*Proof.* Since the definitions $\|\mathbf{x}\|_\infty \equiv \max_{i=1,2,\ldots,n}|x_i|$ and $\|\mathbf{x}\|_p \equiv \sqrt[p]{\sum_{i=1}^n |x_i|^p}$, we only need to show $\lim_{p\to\infty}\|\mathbf{x}\|_p = \max_{i=1,2,\ldots,n}|x_i|$. Given any $\mathbf{x}\in\mathbb{R}^n$ where $n$ is a finite positive integer, we have

$$\lim_{p\to\infty}\sqrt[p]{\left(\max_{i=1,2,\ldots,n}|x_i|\right)^p} \le \lim_{p\to\infty}\sqrt[p]{\sum_{i=1}^n |x_i|^p} \le \lim_{p\to\infty}\sqrt[p]{\left(n\cdot\max_{i=1,2,\ldots,n}|x_i|\right)^p}$$

$$\Downarrow$$

$$\max_{i=1,2,\ldots,n}|x_i| \le \lim_{p\to\infty}\sqrt[p]{\sum_{i=1}^n |x_i|^p} \le \underbrace{\lim_{p\to\infty}\sqrt[p]{n}}_{=1}\cdot\max_{i=1,2,\ldots,n}|x_i|$$

$$\Downarrow$$

$$\max_{i=1,2,\ldots,n}|x_i| \le \lim_{p\to\infty}\sqrt[p]{\sum_{i=1}^n |x_i|^p} \le \max_{i=1,2,\ldots,n}|x_i|$$

$$\Downarrow$$

$$\lim_{p\to\infty}\sqrt[p]{\sum_{i=1}^n |x_i|^p} = \max_{i=1,2,\ldots,n}|x_i|.$$

$\square$

This completes our proof.

**Exercise 1.3**

Show that for any $\mathbf{x},\mathbf{y},\mathbf{z}\in\mathbb{R}^n$

$$\|\mathbf{x}-\mathbf{z}\| \le \|\mathbf{x}-\mathbf{y}\| + \|\mathbf{y}-\mathbf{z}\|.$$

*Proof.* Here, $\|\cdot\|$ refers to the vector norm $\|\cdot\|_2$ whose subscript is frequently omitted for brevity. By the definition of the vector norm, $\|\cdot\|_2$ satisfies the triangle inequality as follows.

$$\begin{aligned}\|\mathbf{x}-\mathbf{z}\|_2 &= \|\mathbf{x}-\mathbf{y}+\mathbf{y}-\mathbf{z}\|_2\\ &\le \|\mathbf{x}-\mathbf{y}\| + \|\mathbf{y}-\mathbf{z}\|_2\end{aligned}$$

as desired. $\square$

**Exercise 1.4**

Prove the Cauchy-Schwarz inequality (Lemma 1.5)

$$|\mathbf{x}^T\mathbf{y}| \le \|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2, \quad \forall \mathbf{x},\mathbf{y}\in\mathbb{R}^n. \tag{14}$$

Show that equality holds if and only if the vectors $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent.

*Proof.* This lemma can be concisely proved via the following formula from geometry.

$$\mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2\cdot\cos\theta \tag{15}$$

where $\theta$ denotes the angle between $\mathbf{x}$ and $\mathbf{y}$. Since $|\cos\theta| \le 1$, it follows that

$$-\|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2 \le \mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2\cdot\cos\theta \le \|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2 \tag{16}$$

where the equality holds if and only if $|\cos\theta|= 1$ which geometrically implies that $\mathbf{x}$ and $\mathbf{y}$ are parallel to each other, in other words, $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent. If we express (16) in a compact way, then we get

$$|\mathbf{x}^T\mathbf{y}|\leq \|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2, \quad \forall \mathbf{x},\mathbf{y}\in\mathbb{R}^n. \tag{17}$$

This completes the proof. □

**Exercise 1.5**

Suppose that $\mathbb{R}^m$ and $\mathbb{R}^n$ are equipped with norms $\|\cdot\|_b$ and $\|\cdot\|_a$, respectively. Show that the induced matrix norm $\|\cdot\|_{a,b}$ satisfies the triangle inequality. That is, for any $\mathbf{A},\mathbf{B}\in\mathbb{R}^{m\times n}$ the inequality

$$\|\mathbf{A}+\mathbf{B}\|_{a,b}\leq \|\mathbf{A}\|_{a,b}+\|\mathbf{B}\|_{a,b} \tag{18}$$

holds.

*Proof.* By the definition of the induced norm, namely (1),

$$\begin{aligned}
\|\mathbf{A}+\mathbf{B}\|_{a,b} &= \max_{\mathbf{x}}\{\|(\mathbf{A}+\mathbf{B})\mathbf{x}\|_b\colon \|\mathbf{x}\|_a\leq 1\} && (19)\\
&= \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{x}\|_b\colon \|\mathbf{x}\|_a\leq 1\} && (20)\\
&\leq \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_b+\|\mathbf{B}\mathbf{x}\|_b\colon \|\mathbf{x}\|_a\leq 1\} && (21)\\
&\leq \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_b\colon \|\mathbf{x}\|_a\leq 1\} + \max_{\mathbf{x}}\{\|\mathbf{B}\mathbf{x}\|_b\colon \|\mathbf{x}\|_a\leq 1\} && (22)\\
&= \|\mathbf{A}\|_{a,b}+\|\mathbf{B}\|_{a,b} && (23)
\end{aligned}$$

where the first inequality follows from the triangle inequality. This completes the proof. □

**Exercise 1.6**

Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Show that the norm function $f(\mathbf{x}) = \|\mathbf{x}\|$ is a continuous function over $\mathbb{R}^n$.

*Proof.* As we know, the continuity of $f(\mathbf{x})$ at a point $\mathbf{x}_0$ requires that, for any $\epsilon > 0$ and the point $\mathbf{x}_0$ in the domain $\mathcal{D}$ of $f$, there always exists a $\delta$ such that $|f(\mathbf{x}) - f(\mathbf{x}_0)|< \epsilon$ whenever $\mathbf{x}\in\mathcal{D}$ and $\|\mathbf{x}-\mathbf{x}_0\|< \delta$. Here, any nonnegative $\delta < \epsilon$ will satisfy this requirement. To see this, we need to analyze two cases. For the case when $\|\mathbf{x}\|> \|\mathbf{x}_0\|$,

$$\begin{aligned}
|f(\mathbf{x}) - f(\mathbf{x}_0)| &= \|\mathbf{x}\|-\|\mathbf{x}_0\| && (24)\\
&= \|\mathbf{x}-\mathbf{x}_0+\mathbf{x}_0\|-\|\mathbf{x}_0\| && (25)\\
&\leq \|\mathbf{x}-\mathbf{x}_0\|+\|\mathbf{x}_0\|-\|\mathbf{x}_0\| && (26)\\
&= \|\mathbf{x}-\mathbf{x}_0\|< \delta < \epsilon. && (27)
\end{aligned}$$

The case of $\|\mathbf{x}\|= \|\mathbf{x}_0\|$ is trivial. For the case when $\|\mathbf{x}\|< \|\mathbf{x}_0\|$,

$$\begin{aligned}
|f(\mathbf{x}) - f(\mathbf{x}_0)| &= \|\mathbf{x}_0\|-\|\mathbf{x}\| && (28)\\
&= \|\mathbf{x}_0-\mathbf{x}+\mathbf{x}\|-\|\mathbf{x}\| && (29)\\
&\leq \|\mathbf{x}-\mathbf{x}_0\|+\|\mathbf{x}\|-\|\mathbf{x}\| && (30)\\
&= \|\mathbf{x}-\mathbf{x}_0\|< \delta < \epsilon. && (31)
\end{aligned}$$

Since the above argument holds for any $\mathbf{x}_0\in\mathbb{R}^n$, it follows that $f(\mathbf{x}) = \|\mathbf{x}\|$ is continuous over $\mathbb{R}^n$. This completes the proof. □

**Exercise 1.7**

(**attainment of the maximum in the induced norm definition**) Suppose that $\mathbb{R}^m$ and $\mathbb{R}^n$ are equipped with norms $\|\cdot\|_b$ and $\|\cdot\|_a$, respectively, and let $\mathbf{A} \in \mathbb{R}^{m\times n}$. Show that there exists $\mathbf{x} \in \mathbb{R}^n$ such that $\|\mathbf{x}\|_a \leq 1$ and $\|\mathbf{A}\mathbf{x}\|_b = \|\mathbf{A}\|_{a,b}$.

*Proof.* Define the set $C = \{\mathbf{x} \in \mathbb{R}^n | \|\mathbf{x}\|_a \leq 1\}$. It is easy to see that $C$ contains all the limits of convergent sequences of points in $C$, so $C$ is closed. We can find a positive number $M$, say 2, such that $C \subset B(\mathbf{0}, M)$, so $C$ is bounded. Since $\mathbf{0} \in C$, $C$ is nonempty. Thus, $C$ is a nonempty and compact set. From Exercise 1.6, since $\|\cdot\|_b$ is a norm, $\|\mathbf{A}\mathbf{x}\|_b$ is continuous. According to Weierstrass theorem (see Theorem 2.30 in the textbook), there exists a global minimum of $f$ and a global maximum of $f$ over $C$. By the definition of the induced norm, the maximum is denoted $\|\mathbf{A}\|_{a,b}$. This completes our proof. □

**Exercise 1.8**

Suppose that $\mathbb{R}^m$ and $\mathbb{R}^n$ are equipped with norms $\|\cdot\|_b$ and $\|\cdot\|_a$, respectively. Show that the induced matrix norm $\|\cdot\|_{a,b}$ can be computed by the formula

$$\|\mathbf{A}\|_{a,b} = \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_b : \|\mathbf{x}\|_a = 1\}. \tag{32}$$

*Proof.* By the definition of the induced norm, the claim is equivalent to proving that the maxima are achieved at $\mathbf{x}^*$ satisfying $\|\mathbf{x}^*\| = 1$, which has been shown in Lemma 1.1. □

**Exercise 1.9**

Suppose that $\mathbb{R}^m$ and $\mathbb{R}^n$ are equipped with norms $\|\cdot\|_b$ and $\|\cdot\|_a$, respectively. Show that the induced matrix norm $\|\cdot\|_{a,b}$ can be computed by the formula

$$\|\mathbf{A}\|_{a,b} = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a}. \tag{33}$$

*Proof.* This is exactly Lemma 2 which includes a proof. □

**Exercise 1.10**

Let $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times k}$ and assume that $\mathbb{R}^m, \mathbb{R}^n, \mathbb{R}^k$ are equipped with the norms $\|\cdot\|_c$, $\|\cdot\|_b$, and $\|\cdot\|_a$, respectively. Prove that

$$\|\mathbf{A}\mathbf{B}\|_{a,c} \leq \|\mathbf{A}\|_{b,c}\|\mathbf{B}\|_{a,b}. \tag{34}$$

*Proof.* From Exercise 1.9, we have

$$\|\mathbf{A}\mathbf{B}\|_{a,c} \leq \frac{\|\mathbf{A}\mathbf{B}\mathbf{x}\|_c}{\|\mathbf{x}\|_a} \tag{35}$$

where $\mathbf{x} \neq \mathbf{0}$. For every $\mathbf{x} \neq \mathbf{0}$, if $\mathbf{B}\mathbf{x} = \mathbf{0}$, then $\mathbf{B} = \mathbf{0}$ must hold, in which case the claim is obviously true. When $\mathbf{B}\mathbf{x} \neq \mathbf{0}$, let $\mathbf{y} = \mathbf{B}\mathbf{x}$ and then,

$$\|\mathbf{A}\mathbf{B}\|_{a,c} \leq \frac{\|\mathbf{A}\mathbf{y}\|_c}{\|\mathbf{y}\|_b}\frac{\|\mathbf{B}\mathbf{x}\|_b}{\|\mathbf{x}\|_a} \leq \|\mathbf{A}\|_{b,c}\|\mathbf{B}\|_{a,b}. \tag{36}$$

This completes the proof. □

**Exercise 1.11**

Prove the formula of the $\infty$-matrix norm given in Example 1.9 of the textbook. Specifically, given $\mathbf{A} \in \mathbb{R}^{m\times n}$,

$$\|\mathbf{A}\|_\infty = \max_{i=1,2,\ldots,m} \sum_{j=1}^{n} |A_{i,j}|. \tag{37}$$

*Proof.* From Exercise 1.8, the induced norm $\|\mathbf{A}\|_\infty$ can also be computed by

$$\begin{aligned}
\|\mathbf{A}\|_\infty &= \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_\infty : \|\mathbf{x}\|_\infty = 1\} && (38)\\
&= \max_{\mathbf{x}}\left\{\max_{i=1,\ldots,m} \Big|\sum_{j=1}^{n} A_{ij}x_j\Big| : \max_{j=1,\ldots,n}|x_j| = 1\right\} && (39)\\
&= \max_{\mathbf{x}}\left\{\max_{i=1,\ldots,m} \sum_{j=1}^{n} |A_{ij}x_j| : \max_{j=1,\ldots,n}|x_j| = 1\right\} && (40)\\
&= \max_{i=1,\ldots,m} \sum_{j=1}^{n} |A_{ij}\mathrm{sign}(A_{ij})| = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |A_{ij}| && (41)
\end{aligned}$$

where $\mathrm{sign}(A_{ij}) = 1$ if $A_{ij} \geq 0$ otherwise $\mathrm{sign}(A_{ij}) = -1$. Note that, besides the last line, (40) also makes use of the constraint $|x_j| \leq 1$ for every $j \in \{1,\ldots,n\}$. $\square$

**Exercise 1.12**

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$. Prove that

(i) $\frac{1}{\sqrt{n}}\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\|_2 \leq \sqrt{m}\|\mathbf{A}\|_\infty$,

(ii) $\frac{1}{\sqrt{m}}\|\mathbf{A}\|_1 \leq \|\mathbf{A}\|_2 \leq \sqrt{n}\|\mathbf{A}\|_1$.

*Proof.* Before we prove the claimed 4 inequalities, we have

$$\begin{aligned}
\|\mathbf{A}\|_2 &= \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 && \text{(Definition of } \|\mathbf{A}\|_2) && (42)\\
&= \max_{\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n} A_{ij}x_j\right)^2} && \text{(Definition of } \|\mathbf{A}\|_2) && (43)\\
&= \max_{\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n} |A_{ij}||x_j|\right)^2} && (\forall j,\ \mathrm{sgn}(x_j) \text{ does not change } \|\mathbf{x}\|_2) && (44)
\end{aligned}$$

Given this, for Part (i), we first show the second inequality.

$$\max_{\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n} |A_{ij}||x_j|\right)^2} \leq \max_{\|\mathbf{x}\|_\infty=1} \sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n} |A_{ij}||x_j|\right)^2} \quad (\{\mathbf{x} \mid \|x\|_2 = 1\} \subset \{\mathbf{x} \mid \|x\|_\infty = 1\}) \tag{45}$$

$$= \sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|A_{ij}|\right)^2} \qquad \text{(Maximum is attained at } |x_i|= 1\ \forall i) \tag{46}$$

$$\leq \sqrt{\sum_{i=1}^{m}\left(\max_{i=1,\dots,m}\sum_{j=1}^{n}|A_{ij}|\right)^2} \qquad (u_i \leq \max_i |u_i|,\ \forall i) \tag{47}$$

$$= \sqrt{\sum_{i=1}^{m}(\|\mathbf{A}\|_\infty)^2} = \sqrt{m}\|\mathbf{A}\|_\infty \qquad \text{(Definition of } \|\mathbf{A}\|_\infty) \tag{48}$$

as desired. Now we prove the first inequality of Part (i).

$$\max_{\|\mathbf{x}\|_2=1}\sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|A_{ij}||x_j|\right)^2} \geq \sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|A_{ij}|\cdot\frac{1}{\sqrt{n}}\right)^2} \qquad \left(\sum_{j=1}^{n}\left(\frac{1}{\sqrt{n}}\right)^2 = 1\right) \tag{49}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|A_{ij}|\right)^2} \qquad \left(\left(\frac{1}{\sqrt{n}}\right)^2 = \frac{1}{n}\right) \tag{50}$$

$$\geq \sqrt{\max_{i=1,\dots,m}\frac{1}{n}\left(\sum_{j=1}^{n}|A_{ij}|\right)^2} \qquad \left(\sum_i|u_i|\geq \max_i |u_i|\ \forall i\right) \tag{51}$$

$$= \frac{1}{\sqrt{n}}\max_{i=1,\dots,m}\sum_{j=1}^{n}|A_{ij}| = \frac{1}{\sqrt{n}}\|\mathbf{A}\|_\infty \qquad \text{(Definition of } \|\mathbf{A}\|_\infty) \tag{52}$$

For part (ii), we first consider the left inequality.

$$\max_{\|\mathbf{x}\|_2=1}\sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|A_{ij}||x_j|\right)^2} = \sqrt{m}\cdot\max_{\|\mathbf{x}\|_2=1}\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|A_{ij}||x_j|}{m} \qquad \text{(AM-QM inequality)} \tag{53}$$

$$= \frac{1}{\sqrt{m}}\cdot\max_{\|\mathbf{x}\|_2=1}\sum_{j=1}^{n}|x_j|\left(\sum_{i=1}^{m}|A_{ij}|\right) \qquad \left(\forall m, n < \infty, \sum_{i=1}^{m}\sum_{j=1}^{n} = \sum_{j=1}^{n}\sum_{i=1}^{m}\right) \tag{54}$$

$$= \frac{1}{\sqrt{m}}\cdot\max_{\|\mathbf{x}\|_2=1}\sqrt{\sum_{j=1}^{n}|x_j|^2}\sqrt{\sum_{j=1}^{n}\left(\sum_{i=1}^{m}|A_{ij}|\right)^2} \qquad \text{(Cauchy-Schwarz inequality)} \tag{55}$$

$$= \frac{1}{\sqrt{m}}\sqrt{\sum_{j=1}^{n}\left(\sum_{i=1}^{m}|A_{ij}|\right)^2} \qquad (\|\mathbf{A}\|_2= 1) \tag{56}$$

$$\geq \frac{1}{\sqrt{m}}\sqrt{\max_{j=1,\dots,n}\left(\sum_{i=1}^{m}|A_{ij}|\right)^2} \qquad \left(\sum_i|u_i| \geq \max_i |u_i|\ \forall i\right) \tag{57}$$

$$= \frac{1}{\sqrt{m}} \max_{j=1,\dots,n} \sum_{i=1}^{m} |A_{ij}| = \frac{1}{\sqrt{m}} \|\mathbf{A}\|_1 \qquad \text{(Definition of } \|\mathbf{A}\|_1) \tag{58}$$

When applying the AM-GM inequality, the equality holds if and only if $\sum_{j=1}^{n} |A_{1j}x_j| = \cdots = \sum_{j=1}^{n} |A_{mj}x_j|$, which is attainable. For Cauchy-Schwarz inequality, the equality holds if and only if $\sum_{i=1}^{m} |A_{ij}| = k|x_j|$ for all $j = 1, \dots, n$ where $k$ is a constant, which is attainable too.

Now we show the inequality on the right hand side.

$$\max_{\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^{m} \left( \sum_{j=1}^{n} |A_{ij}||x_j| \right)^2} \le \max_{\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2 \sum_{j=1}^{n} x_j^2} \qquad \text{(Cauchy-Schwarz inequality)} \tag{59}$$

$$= \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2} \qquad (\|\mathbf{x}\|_2 = 1) \tag{60}$$

$$= \sqrt{\sum_{j=1}^{n} \sum_{i=1}^{m} |A_{ij}|^2} \qquad \left( \forall m, n < \infty, \sum_{i=1}^{m} \sum_{j=1}^{n} = \sum_{j=1}^{n} \sum_{i=1}^{m} \right) \tag{61}$$

$$\le \sqrt{\sum_{j=1}^{n} \left( \sum_{i=1}^{m} |A_{ij}| \right)^2} \qquad \left( \forall a_i \ge 0, \sum_{i=1}^{m} a_i^2 \le \left( \sum_{i=1}^{m} a_i \right)^2 \right) \tag{62}$$

$$\le \sqrt{\sum_{j=1}^{n} \left( \max_{j=1,\dots,n} \sum_{i=1}^{m} |A_{ij}| \right)^2} \qquad (u_i \le \max_i |u_i|, \ \forall i) \tag{63}$$

$$= \sqrt{n} \cdot \max_{j=1,\dots,n} \sum_{i=1}^{m} |A_{ij}| \qquad \left( \sum_{j=1}^{n} c = nc \right) \tag{64}$$

$$= \sqrt{n} \|\mathbf{A}\|_1 \qquad \text{(Definition of } \|\mathbf{A}\|_1) \tag{65}$$

where in the first line the equality holds if and only if $|A_{ij}| = k_i|x_j|$ for all $i = 1, \dots, m$ and $j = 1, \dots, n$, and $k_i$ is a constant, which is not necessarily attainable. This completes the proof. $\square$

**Exercise 1.13**

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$. Show that

(i) $\|\mathbf{A}\| = \|\mathbf{A}^T\|$ (here $\|\cdot\|$ is the spectral norm),

(ii) $\|\mathbf{A}\|_F^2 = \sum_{i=1}^{n} \lambda_i(\mathbf{A}^T\mathbf{A})$.

*Proof.* For part (i), the spectral norm is defined by

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})} = \sigma_{\max}(\mathbf{A}) \tag{66}$$

where $\lambda_{\max}(\mathbf{A}^T\mathbf{A})$ is the maximum eigenvalue of $\mathbf{A}^T\mathbf{A}$, and $\sigma_{\max}(\mathbf{A})$ is the largest singular values of $\mathbf{A}$. Similarly,

$$\|\mathbf{A}^T\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}\mathbf{A}^T)} = \sigma_{\max}(\mathbf{A}^T) \tag{67}$$

By the Theorem 2.6.3(a) in Horn and Johnson (2013), the singular values are supposed to be nonnegative. And by the Theorem 2.6.3(b) in Horn and Johnson (2013), the nonzero eigenvalues of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ are identical. Thus,

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})} = \sqrt{\lambda_{\max}(\mathbf{A}\mathbf{A}^T)} = \|\mathbf{A}^T\|_2 \tag{68}$$

107 as desired.

Now we consider part (ii).

$$\begin{aligned}
\|\mathbf{A}\|_F^2 =& \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2 && \text{(Definition of Frobenius norm)} && (69)\\
=& \mathrm{Tr}(\mathbf{A}^T\mathbf{A}) && \text{(Definition of trace)} && (70)\\
=& \sum_{n=1}^{n} \lambda_i(\mathbf{A}^T\mathbf{A}) && && (71)
\end{aligned}$$

where the last line follows from the following argument[2]. By definition, the characteristic polynomial of $\mathbf{A}^T\mathbf{A}$ is given by

$$\begin{aligned}
p(t) =& \det(t\mathbf{I} - \mathbf{A}^T\mathbf{A}) && && (72)\\
=& t^n - \mathrm{Tr}(\mathbf{A}^T\mathbf{A})t^{n-1} + \cdots + (-1)^n \det(\mathbf{A}^T\mathbf{A}) && \text{(Definition of determinant)} && (73)
\end{aligned}$$

Also, by the definition, eigenvalues are the roots of $p(t)$. Hence,

$$p(t) = (t - \lambda_1)(t - \lambda_2)\cdots(t - \lambda_n) \tag{74}$$

By comparing coefficients, we have

$$\mathrm{Tr}(\mathbf{A}^T\mathbf{A}) = \sum_{i=1}^{n} \lambda_i(\mathbf{A}^T\mathbf{A}) \tag{75}$$

108 which completes the proof. □

**Exercise 1.14**

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Show that

$$\max_{\mathbf{x}}\{\mathbf{x}^T\mathbf{A}\mathbf{x} : \|\mathbf{x}\|^2 = 1\} = \lambda_{\max}(\mathbf{A}). \tag{76}$$

109

110 The inspiration of the following proof is from the proof of Lemma 1.11 in the textbook.

*Proof.* According to the spectral decomposition theorem there exists an orthogonal matrix $\mathbf{U} \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ such that $\mathbf{U}^T\mathbf{A}\mathbf{U} = \mathbf{D}$. Without the loss of generality, we can assume that the diagonal elements of $\mathbf{D}$, which are the eigenvalues of $\mathbf{A}$, are ordered nonincreasingly: $d_1 \geq d_2 \geq \cdots \geq d_n$, where $d_1 = \lambda_{\max}(\mathbf{A})$. Since $\mathbf{U}$ is an orthogonal matrix, we can make the change of variables $\mathbf{x} = \mathbf{U}\mathbf{y}$.

$$\begin{aligned}
\max_{\|\mathbf{x}\|_2^2=1} \mathbf{x}^T\mathbf{A}\mathbf{x} =& \max_{\|\mathbf{U}\mathbf{y}\|_2^2=1} (\mathbf{U}\mathbf{y})^T\mathbf{A}\mathbf{U}\mathbf{y} && && (77)\\
=& \max_{\|\mathbf{y}\|_2^2=1} \mathbf{y}^T\mathbf{U}^T\mathbf{A}\mathbf{U}\mathbf{y} && (\|\mathbf{U}\mathbf{y}\|_2^2 = \|\mathbf{y}\|_2^2) && (78)
\end{aligned}$$

[2] https://math.stackexchange.com/questions/546155/proof-that-the-trace-of-a-matrix-is-the-sum-of-its-eigenvalues

$$
\begin{aligned}
&= \max_{\|\mathbf{y}\|_2^2=1} \mathbf{y}^T\mathbf{D}\mathbf{y} && (\mathbf{U}^T\mathbf{A}\mathbf{U}=\mathbf{D}) && (79)\\
&= \max_{\|\mathbf{y}\|_2^2=1} \sum_{i=1}^n d_i y_i^2 \le d_1 \max_{\|\mathbf{y}\|_2^2=1} \sum_{i=1}^n y_i^2 && (d_1 \ge d_2 \ge \cdots \ge d_n) && (80)\\
&= d_1 = \lambda_{\max}(\mathbf{A}) && && (81)
\end{aligned}
$$

□

**Exercise 1.15**

Prove that a set $U \subseteq R^n$ is closed if and only if its complement $U^c$ is open.

*Proof.* We first prove the sufficiency. Given $U^c$ is open, we suppose that $U$ is not closed. Then there must exist at least one accumulation point of $U$, say $x$, such that $x \notin U$, i.e., $x \in U^c$. Since $U^c$ is open, then there exists an open ball $B(x,r) \subseteq U^c$ with $r > 0$, which contradicts $x \in U'$ where $U'$ denotes the set of accumulation points of $U$. Specifically, since $x \in U'$, by Definition 1.4, there are infinitely many points of $B(x,r)$ belonging to $U$, which is impossible for $B(x,r) \subseteq U^c$.

Now we show the necessity. Given any point $x \in U^c$, it suffices to show that $x$ is an interior point of $U^c$. Obviously, $x \notin U$. Since $U$ is closed, $x$ is not an accumulation point of $U$. By Definition 1.5, this implies that there exists an open ball $B(x,r)$ such that $B(x,r) \bigcap U = \emptyset$. Thus, $B(x,r) \subseteq U^c$. This completes our proof. □

**Exercise 1.16**

1. Let $\{A_i\}_{i\in I}$ be a collection of open sets where $I$ is a given index set. Show that $\bigcup_{i\in I} A_i$ is an open Set. Show that if $I$ is finite, then $\bigcap_{i\in I} A_i$ is open.
2. Let $\{A_i\}_{i\in I}$ be a collection of closed sets where $I$ is a given index set. Show that $\bigcap_{i\in I} A_i$ is a closed Set. Show that if $I$ is finite, then $\bigcup_{i\in I} A_i$ is closed.

The following proof is taken from the proof of Theorem 11.1.5 in Chen et al. (2019).

*Proof.*

1. For any $\mathbf{x} \in \bigcup_{i\in I} A_i$, then there exists at least an $i \in I$ such that $\mathbf{x} \in A_i$. Since $A_i$ is an open set, then $\mathbf{x}$ is an interior point of $A_i$. Also, $\mathbf{x}$ is an interior point of $\bigcup_{i\in I} A_i$. Thus, $\bigcup_{i\in I} A_i$ is an open set.

   Since $I$ is finite, suppose there are $k$ sets in total. For any $\mathbf{x} \in \bigcap_{i\in I} A_i$, $x \in A_i$ for arbitrary $i = 1,\ldots,k$. Thus, for any $i \in I$, there exists an $r_i > 0$ such that $B(\mathbf{x}, r_i) \subset A_i$. Let $r = \min_{i\in I} r_i$, then $B(\mathbf{x}, r) \subset \bigcap_{i\in I} A_i$. Therefore, $\bigcap_{i\in I} A_i$ is open.

2. By De Morgan's Theorem (see Theorem 1.10), $(\bigcap_{i\in I} A_i)^c = \bigcup_{i\in I} A_i^c$. Since $A_i$ is closed, its complement $A_i^c$ is open. From the first part of this proof, $\bigcup_{i\in I} A_i^c$ is open. Thus, its complement $\bigcap_{i\in I} A_i$ is closed.

   If each $A_i$ is closed, then $A_i^c$ is open. If $I$ is finite, by the first part of this proof, $\bigcap_{i\in I} A_i^c$ is open. According to De Morgan's Theorem, its complement is $\bigcup_{i\in I} A_i$ which is closed. This completes the proof.

□

**Exercise 1.17**

Give an example of open sets $A_i$, $i \in I$ for which $\bigcap_{i \in I} A_i$ is not open.

The following solution is from Mathematics Stack Exchange[3].

**Solution:** Let $\mathbb{Z}_+$ denote the set of positive integers. When $A_i$ is defined as

$$A_i = (-\frac{1}{i}, \frac{1}{i}), \quad i \in \mathbb{Z}_+,$$

the intersection

$$\bigcap_{i \in \mathbb{Z}_+} A_i = [0]$$

is not open. However, it is a closed set. □

**Extensions**

Likewise, we can construct an example of closed sets $A_i$, $i \in \mathbb{Z}_+$ for which $\bigcup_{i \in \mathbb{Z}_+} A_i$ is not closed. For example, the union of the closed sets $A_i = [\frac{1}{i}, 2 - \frac{1}{i}], \forall i \in \mathbb{Z}_+$ is $(0, 2)$ which is an open set.

**Exercise 1.18**

Let $A, B \subseteq \mathbb{R}^n$. Prove that $\text{cl}(A \cap B) \subseteq \text{cl}(A) \cap \text{cl}(B)$. Give an example in which the inclusion is proper.

This proof is from Mathematics Stack Exchange[4].

*Proof.* By the definition of closure, i.e. Definition 1.7, $\text{cl}(U) = U \cup \text{bd}(U)$. since $A \cap B \subseteq A$, it follows that $\text{cl}(A \cap B) \subseteq \text{cl}(A)$. Likewise, $\text{cl}(A \cap B) \subseteq \text{cl}(B)$. Thus, $\text{cl}(A \cap B) \subseteq \text{cl}(A) \cap \text{cl}(B)$ as desired.

Given $A = (0, 1)$ and $B = (1, 2)$, then $A \cap B = \emptyset$ and $\text{cl}(A \cap B) = \emptyset$. On the other hand, $\text{cl}(A) = [0, 1]$ and $\text{cl}(B) = [1, 2]$. Thus, $\text{cl}(A) \cap \text{cl}(B) = \{1\}$. Obviously, $\emptyset \neq \{1\}$. Hence, the inclusion is proper in this case. □

**Exercise 1.19**

Let $A, B \subseteq \mathbb{R}^n$. Prove that $\text{int}(A \cap B) = \text{int}(A) \cap \text{int}(B)$ and that $\text{int}(A) \cup \text{int}(B) \subseteq \text{int}(A \cup B)$. Show an example in which the latter inclusion is proper.

*Proof.* The first part of the following proof is from a YouTube video[5].

1. $\text{int}(A \cap B) \subseteq \text{int}(A) \cap \text{int}(B)$ follows from

$$A \cap B \subseteq A \Rightarrow \text{int}(A \cap B) \subseteq \text{int}(A) \tag{82}$$
$$A \cap B \subseteq B \Rightarrow \text{int}(A \cap B) \subseteq \text{int}(B) \tag{83}$$
$$\Downarrow$$
$$\text{int}(A \cap B) \subseteq \text{int}(A) \cap \text{int}(B). \tag{84}$$

[3] https://math.stackexchange.com/questions/1460853/infinite-intersection-of-open-sets
[4] https://math.stackexchange.com/questions/1485869/closure-of-intersection-of-two-sets
[5] https://www.youtube.com/watch?v=uZZkMloQbd0

$\text{int}(A) \cap \text{int}(B) \subseteq \text{int}(A \cap B)$ follows from

$$\text{int}(A) \subseteq A, \quad \text{int}(B) \subseteq B \tag{85}$$
$$\Downarrow$$
$$\text{int}(A) \cap \text{int}(B) \subseteq A \cap B. \tag{86}$$

Since the finite intersection of open sets is an open set (see Exercise 1.16(i)), then $\text{int}(A) \cap \text{int}(B)$ is open. By definition, the interior of a set is the largest open subset of that set, so $\text{int}(A \cap B)$ contains $\text{int}(A) \cap \text{int}(B)$. In other words, $\text{int}(A) \cap \text{int}(B) \subseteq \text{int}(A \cap B)$. Therefore, $\text{int}(A \cap B) = \text{int}(A) \cap \text{int}(B)$.

2. $\text{int}(A) \cup \text{int}(B) \subseteq \text{int}(A \cup B)$ follows from

$$\text{int}(A) \subseteq A, \quad \text{int}(B) \subseteq B \tag{87}$$
$$\Downarrow$$
$$\text{int}(A) \cup \text{int}(B) \subseteq A \cup B. \tag{88}$$

In Exercise 1.16(i), we have shown that the union of open sets is open, so $\text{int}(A) \cup \text{int}(B)$ is an open set. By definition, the interior of $A \cup B$ is the largest open set of $A \cup B$. Thus, $\text{int}(A \cup B)$ contains $\text{int}(A) \cup \text{int}(B)$. Hence, $\text{int}(A) \cup \text{int}(B) \subseteq \text{int}(A \cup B)$.

For example, $A = (0, 1)$ and $B = [1, 2)$. It is easy to see that $\text{int}(A) \cup \text{int}(B) = (0, 1) \cup (1, 2)$, but $\int(A \cup B) = (1, 2)$. This inclusion is proper.

□

# 2 Chapter 2 Optimality Conditions for Unconstrained Optimization

**Exercise 2.1**

Find the global minimum and maximum points of the function $f(x, y) = x^2 + y^2 + 2x - 3y$ over the unit ball $S = B[0, 1] = \{(x, y) : x^2 + y^2 \leq 1\}$.

**Solution:** By applying Cauchy-Swcharz inequality on $2x - 3y$, we get

$$|2x - 3y| = \left| \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 2 \\ -3 \end{pmatrix} \right| \leq \sqrt{2^2 + (-3)^2}\sqrt{x^2 + y^2} = \sqrt{13}\sqrt{x^2 + y^2}$$
$$\Downarrow$$
$$-\sqrt{13}\sqrt{x^2 + y^2} \leq 2x - 3y \leq \sqrt{13}\sqrt{x^2 + y^2}$$

where the equalities hold when $-3x = 2y$. Thus,

$$x^2 + y^2 - \sqrt{13}\sqrt{x^2 + y^2} \leq x^2 + y^2 + 2x - 3y \leq x^2 + y^2 + \sqrt{13}\sqrt{x^2 + y^2}$$

Let $t = \sqrt{x^2 + y^2}$, then the right hand side can be written as

$$f_{\text{RHS}}(t) = t^2 + \sqrt{13}t, \text{ with } 0 \leq t \leq 1. \tag{89}$$

Since $f'_{\text{RHS}}(t) = 2t + \sqrt{13} \geq 0$, then $f_{\text{RHS}}(t)$ is increasing on $[0, 1]$. So, the maximum can be attained at $t = 1$. Thus, solving $x^2 + y^2 = 1$ and $-3x = 2y$ gives $x = 2/\sqrt{13}$ and $y = -3/\sqrt{13}$ and $f(2/\sqrt{13}, -3/\sqrt{13}) = 1 + \sqrt{13}$, which is equal to $f_{\text{RHS}}(1) = 1 + \sqrt{13}$.

The left hand side is

$$f_{\text{LHS}}(t) = t^2 - \sqrt{13}t, \text{ with } 0 \le t \le 1 \tag{90}$$

Its derivative with respect to $t$ is $f'_{\text{LHS}}(t) = 2t - \sqrt{13} < 0$ on $[0,1]$, which means $f_{\text{LHS}}(t)$ is strictly decreasing on $[0,1]$. The minimum can be achieved at $t = 1$, i.e. $x^2 + y^2 = 1$ and $f_{\text{LHS}}(1) = 1 - \sqrt{13}$. Given $-3x = 2y$, we obtain $x = -2/\sqrt{13}$ and $y = 3/\sqrt{13}$, which gives the desired $f(-2/\sqrt{13}, 3/\sqrt{13}) = 1 - \sqrt{13}$.

To sum up, the global minimum and maximum points are $(x, y) = (2/\sqrt{13}, -3/\sqrt{13})$ and $(x, y) = (-2/\sqrt{13}, 3/\sqrt{13})$, respectively. □

**Exercise 2.2**

Let $\mathbf{a} \in \mathbb{R}^n$ be a nonzero vector. Show that the maximum of $\mathbf{a}^T\mathbf{x}$ over $B[0,1] = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \le 1\}$ is attained at $\mathbf{x}^* = \frac{\mathbf{a}}{\|\mathbf{a}\|}$ and that the maximal value is $\|\mathbf{a}\|$.

*Proof.* According to Cauchy-Schwarz inequality, we have

$$\mathbf{a}^T\mathbf{x} \le \|\mathbf{a}\|\|\mathbf{x}\| \tag{91}$$

the equality holds if and only if $\mathbf{x} = \lambda\mathbf{a}$ where $0 \neq \lambda \in \mathbb{R}$. Since $\mathbf{x} \le 1$, the maximum of the right hand side can be achieved when $\|\mathbf{x}\| = 1$. Combining this with $\mathbf{x} = \lambda\mathbf{a}$, we get $\|\lambda\mathbf{a}\| = 1$ and $\lambda = \frac{1}{\|\mathbf{a}\|}$. Thus, $\mathbf{x}^* = \lambda\mathbf{a} = \frac{\mathbf{a}}{\|\mathbf{a}\|}$ and the maximum value is $\|\mathbf{a}\|\|\mathbf{x}\| = \|\mathbf{a}\|$. □

**Exercise 2.3**

Find the global minimum and maximum points of the function $f(x, y) = 2x - 3y$ over the set $S = \{(x, y) : 2x^2 + 5y^2 \le 1\}$.

**Solution:** We can make use of the result in Exercise 2.2. To do this, we need to perform a change of variables. Specifically, let $u = \sqrt{2}x$ and $v = \sqrt{5}y$. By doing this, the original problem is equivalently reformulated as finding the global minimum and maximum points of $\tilde{f}(u, v) = \sqrt{2}u - \frac{3\sqrt{5}}{5}v$ over the set $\tilde{S} = \{(u, v) : u^2 + v^2 \le 1\}$. In this case, $\mathbf{a} = (\sqrt{2}, -\frac{3\sqrt{5}}{5})^T$. It follows from that the maximum point is $\frac{\mathbf{a}}{\|\mathbf{a}\|} = (\frac{5\sqrt{2}}{19}, -\frac{3\sqrt{5}}{19})^T$. Changing back to the original variables gives $x = 5/19$ and $-3/19$. Similarly, the minimum point is $x = -5/19$ and $3/19$. □

**Exercise 2.4**

Show that if $\mathbf{A}, \mathbf{B}$ are $n \times n$ positive semidefinite matrices, then their sum $\mathbf{A} + \mathbf{B}$ is also positive semidefinite.

*Proof.* Since $\mathbf{A}, \mathbf{B}$ are semidefinite matrices, then $\mathbf{x}^T\mathbf{A}\mathbf{x} \ge 0$ and $\mathbf{x}^T\mathbf{B}\mathbf{x} \ge 0$ for every $\mathbf{x} \in \mathbb{R}^n$. It follows that

$$\mathbf{x}^T(\mathbf{A} + \mathbf{B})\mathbf{x} = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{B}\mathbf{x} \ge 0 \tag{92}$$

for every $\mathbf{x} \in \mathbb{R}^n$. Hence, $\mathbf{A} + \mathbf{B}$ is also positive semidefinite. This completes the proof. □

**Exercise 2.5**

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times m}$ be two symmetric matrices. Prove that the following two claims are equivalent:

(i) $\mathbf{A}$ and $\mathbf{B}$ are positive semidefinite.

(ii) $\begin{pmatrix} \mathbf{A} & \mathbf{0}_{n\times m} \\ \mathbf{0}_{m\times n} & \mathbf{B} \end{pmatrix}$ is positive semidefinite.

*Proof.* We first show (i)⇒(ii). Given $\mathbf{A}$ and $\mathbf{B}$ are positive semidefinite, we have $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ and $\mathbf{y}^T\mathbf{B}\mathbf{y} \geq 0$ for any $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Then for any $\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \in \mathbb{R}^{n+m}$, we have

$$\mathbf{z}^T \begin{pmatrix} \mathbf{A} & \mathbf{0}_{n\times m} \\ \mathbf{0}_{m\times n} & \mathbf{B} \end{pmatrix} \mathbf{z} = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{y}^T\mathbf{B}\mathbf{y} \geq 0 \tag{93}$$

as desired.

Now we consider (ii)⇒(i). Given $\begin{pmatrix} \mathbf{A} & \mathbf{0}_{n\times m} \\ \mathbf{0}_{m\times n} & \mathbf{B} \end{pmatrix}$ is positive semidefinite, for any $\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \in \mathbb{R}^{n+m}$, we have $\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{y}^T\mathbf{B}\mathbf{y} \geq 0$. Since $\mathbf{A}$ is a symmetric matrix, then its eigenvalues are real values. Without loss of generality, suppose $\mathbf{A}$ is not positive semidefinite, then it will have at least one negative eigenvalue $\lambda$. Then we get $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ and $\mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda\mathbf{x}^T\mathbf{x} = \lambda\|\mathbf{x}\|^2 < 0$ for any $\mathbf{x} \neq \mathbf{0}$. So, regardless of $\mathbf{y}$, as $\|\mathbf{x}\|^2 \to -\infty$, $\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{y}^T\mathbf{B}\mathbf{y} \to -\infty$, which contradicts that the block matrix is positive semidefinite. Thus, $\mathbf{A}$ must be positive semidefinite. Likewise, $\mathbf{B}$ must be positive semidefinite. This completes the proof. □

**Exercise 2.6**

Let $\mathbf{B} \in \mathbb{R}^{n\times k}$ and let $\mathbf{A} = \mathbf{B}\mathbf{B}^T$.

(i) Prove $\mathbf{A}$ is positive semidefinite.

(ii) Prove that $\mathbf{A}$ is positive definite if and only if $\mathbf{B}$ has a full row rank.

*Proof.*

(i) For any $\mathbf{x} \in \mathbb{R}^{n\times n}$, we have

$$\mathbf{x}^T\mathbf{B}\mathbf{B}^T\mathbf{x} = (\mathbf{B}^T\mathbf{x})^T\mathbf{B}^T\mathbf{x} = \|\mathbf{B}^T\mathbf{x}\|_2^2 \geq 0. \tag{94}$$

So, $\mathbf{A}$ is positive semidefinite.

(ii) If $\mathbf{B}$ has a full row rank, namely, $\mathbf{B}^T$ has a full column rank, then the columns of $\mathbf{B}^T$ are linearly independent. Then $\mathbf{B}^T\mathbf{x} = \mathbf{0}$ holds only if $\mathbf{x} = \mathbf{0}$. Hence, $\mathbf{A}$ is positive definite.

If $\mathbf{A}$ is positive definite, it follows from (94) that then $\|\mathbf{B}^T\mathbf{x}\|_2^2 > 0$ for any $\mathbf{x} \neq 0$. Therefore, the columns of $\mathbf{B}^T$ are linearly independent. Thus, $\mathbf{B}$ has a full row rank.

□

**Exercise 2.7**

(i) Let $\mathbf{A}$ be an $n \times n$ symmetric matrix. Show that $\mathbf{A}$ is positive semidefinite if and only if there exists a matrix $\mathbf{B} \in \mathbf{R}^{n\times n}$ such that $\mathbf{A} = \mathbf{B}\mathbf{B}^T$.

(ii) Let $\mathbf{x} \in \mathbb{R}^n$ and let $\mathbf{A}$ be defined as

$$A_{ij} = x_i x_j, \quad i, j = 1, 2, \ldots, n. \tag{95}$$

Show that $\mathbf{A}$ is positive semidefinite and that it is not a positive definite matrix when $n > 1$.

*Proof.* (i) The sufficiency has been shown in Exercixe 2.6(i). To show the necessity, by the spectral decomposition theorem, $\mathbf{A}$ can be represented as $\mathbf{U}\mathbf{D}\mathbf{U}^T$ with $\mathbf{U}$ is an orthogonal matrix and $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{A}$. Since $\mathbf{A}$ is positive semidefinite, we have that $d_1, d_2, \ldots, d_n \geq 0$. Let $B = \mathbf{U}\mathbf{D}^{1/2}\mathbf{U}^T$, then $\mathbf{B}\mathbf{B}^T = \mathbf{U}\mathbf{D}^{1/2}\mathbf{U}^T\mathbf{U}\mathbf{D}^{1/2}\mathbf{U}^T = \mathbf{U}\mathbf{D}\mathbf{U}^T$. This shows the necessity.

(ii) $\mathbf{A}$ can be represented as $\mathbf{x}\mathbf{x}^T$. For any $\mathbf{y} \in \mathbb{R}^n$, we have

$$\mathbf{y}^T\mathbf{A}\mathbf{y} = \mathbf{y}^T\mathbf{x}\mathbf{x}^T\mathbf{y} = (\mathbf{x}^T\mathbf{y})^2 \geq 0 \tag{96}$$

which shows $\mathbf{A}$ is positive semidefinite. When $n = 1$, $\mathbf{A}$ is a scalar, so it is positive definite when $x > 0$, otherwise it is not positive definite. Since there always exists a vector $\mathbf{y} \neq \mathbf{0}$ such that $\mathbf{x}^T\mathbf{y} = 0$, $\mathbf{y}^T\mathbf{A}\mathbf{y} > 0$ does not hold for arbitrary $\mathbf{y}$. By definition, $\mathbf{A}$ is not a positive definite matrix. This completes the proof.

$\square$

**Exercise 2.8**

Let $\mathbf{Q} \in \mathbb{R}^{n\times n}$ be a positive definite matrix. Show that the "Q-norm" defined by

$$\|\mathbf{x}\|_{\mathbf{Q}} = \sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x}} \tag{97}$$

is indeed a norm.

*Proof.* We need to check if the "Q-norm" satisfies the three properties of the definition of a norm. Since $\mathbf{Q}$ is positive definite, for any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^T\mathbf{Q}\mathbf{x} \geq 0$ and $\mathbf{x}^T\mathbf{Q}\mathbf{x} = 0$ if and only if $\mathbf{x} = \mathbf{0}$, so $\|\mathbf{x}\|_{\mathbf{Q}} \geq 0$. Thus, the nonnegativity is satisfied. For any $\mathbf{x} \in \mathbb{R}^n$, $\|\lambda\mathbf{x}\|_{\mathbf{Q}} = \sqrt{\lambda^2\mathbf{x}^T\mathbf{Q}\mathbf{x}} = |\lambda|\|\mathbf{x}\|_{\mathbf{Q}}$. Hence, the positive homogeneity is satisfied.

Before proving the triangle inequality for the $\mathbf{Q}$ norm, we need to assume $\mathbf{Q}$ is a symmetric matrix, otherwise it may have complex eigenvalues.

$$\|\mathbf{x} + \mathbf{y}\|_{\mathbf{Q}} \leq \|\mathbf{x}\|_{\mathbf{Q}} + \|\mathbf{y}\|_{\mathbf{Q}} \tag{98}$$

$$\Updownarrow \tag{99}$$

$$\sqrt{(\mathbf{x}+\mathbf{y})^T\mathbf{Q}(\mathbf{x}+\mathbf{y})} \leq \sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x}} + \sqrt{\mathbf{y}^T\mathbf{Q}\mathbf{y}} \tag{100}$$

$$\Updownarrow \tag{101}$$

$$(\mathbf{x}+\mathbf{y})^T\mathbf{Q}(\mathbf{x}+\mathbf{y}) \leq \mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{y}^T\mathbf{Q}\mathbf{y} + 2\sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x}\mathbf{y}^T\mathbf{Q}\mathbf{y}} \tag{102}$$

$$\Updownarrow \tag{103}$$

$$\mathbf{x}^T\mathbf{Q}\mathbf{y} + \mathbf{y}^T\mathbf{Q}\mathbf{x} \leq 2\sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x}\mathbf{y}^T\mathbf{Q}\mathbf{y}} \tag{104}$$

By the spectral decomposition theorem, $\mathbf{Q}$ can be written as $\mathbf{U}^T\mathbf{D}\mathbf{U}$ where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{A}$. Let $\mathbf{Ux} = \tilde{\mathbf{x}}$ and $\mathbf{Uy} = \tilde{\mathbf{y}}$, then we have

$$\mathbf{x}^T\mathbf{U}^T\mathbf{D}\mathbf{U}\mathbf{y} + \mathbf{y}^T\mathbf{U}^T\mathbf{D}\mathbf{U}\mathbf{x} \leq 2\sqrt{\mathbf{x}^T\mathbf{U}^T\mathbf{D}\mathbf{U}\mathbf{x}\mathbf{y}^T\mathbf{U}^T\mathbf{D}\mathbf{U}\mathbf{y}} \tag{105}$$

$$\Updownarrow \tag{106}$$

$$\sum_i^n d_i x_i y_i + \sum_i^n d_i x_i y_i \leq 2\sqrt{(\sqrt{d_i}x_i)^2}\sqrt{(\sqrt{d_i}y_i)^2} \tag{107}$$

$$\Updownarrow \tag{108}$$

$$\sum_i^n (\sqrt{d_i}x_i)(\sqrt{d_i}y_i) \leq \sqrt{(\sqrt{d_i}x_i)^2}\sqrt{(\sqrt{d_i}y_i)^2} \tag{109}$$

which is the Cauchy-Schwarz inequality. This completes the proof. □

**Exercise 2.9**

Let $\mathbf{A}$ be an $n \times n$ positive semidefinite matrix.

(i) Show that for any $i \neq j$

$$A_{ii}A_{jj} \geq A_{ij}^2 \tag{110}$$

(ii) Show that if for some $i \in \{1, 2, \ldots, n\}$ $A_{ii} = 0$, then the $i$th row of $\mathbf{A}$ consists of zeros.

*Proof.* [6]

(i) As stated in Section 2.2 of the textbook, $\mathbf{A}$ is symmetric. Given $\mathbf{A}$ is a positive semidefinite matrix, we always have

$$(e_i x + e_j)^T \mathbf{A}(e_i x + e_j) \geq 0 \tag{111}$$

$$A_{ii}x^2 + 2A_{ij}x + A_{jj} \geq 0 \tag{112}$$

where $\mathbf{e}_i$ is a vector with all zeros except the $i$th entry being 1, also $\mathbf{e}_j$ is defined in the same way, and $x \in \mathbb{R}$. Then the determinant is supposed to be nonpositive.

$$4A_{ij}^2 - 4A_{ii}A_{jj} \leq 0 \Rightarrow A_{ii}A_{jj} \geq A_{ij}^2. \tag{113}$$

(ii) With the result in the first part, if for some $i$, $A_{ii} = 0$, then for any $j \neq i$, we have $0 \times A_{jj} \geq A_{ij}^2$ which implies $A_{ij} = 0$. This shows that the $i$th row of $\mathbf{A}$ consists of zeros. This completes the proof.

□

**Exercise 2.10**

Let $\mathbf{A}^\alpha$ be the $n \times n$ matrix $(n > 1)$ defined by

$$A_{ij} = \begin{cases} \alpha, & i = j, \\ 1, & i \neq j. \end{cases} \tag{114}$$

Show that $\mathbf{A}^\alpha$ is positive semidefinite if and only if $\alpha \geq 1$.

[6] https://math.stackexchange.com/questions/3544963/product-of-diagonal-elements-of-positive-semidefinite-matrix

*Proof.* We first prove the necessity. Given $\mathbf{A}^\alpha$ is positive semidefinite and a vector $\mathbf{x}$ whose entries are all zeros except $x_i = 1$ and $x_j = -1$, we always have

$$\mathbf{x}^T\mathbf{A}^\alpha\mathbf{x} \geq 0 \Rightarrow 2\alpha - 2 \geq 0 \Rightarrow \alpha \geq 1. \tag{115}$$

Now we consider the sufficiency. $\mathbf{A}^\alpha$ can be represented as $(\alpha - 1)\mathbf{I} + \mathbf{1}\mathbf{1}^T$. Together with $\alpha \geq 1$, for any vector $\mathbf{x} \in \mathbb{R}^n$, we have

$$\mathbf{x}^T\mathbf{A}^\alpha\mathbf{x} = (\alpha-1)\mathbf{x}^T\mathbf{I}\mathbf{x} + \mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x} = (\alpha-1)\|\mathbf{x}\|^2 + \|\mathbf{1}^\mathbf{T}\mathbf{x}\|^2 \geq 0$$

227 which implies that $\mathbf{A}^\alpha$ is positive semidefinite. □

**Exercise 2.11**

Let $\mathbf{d} \in \Delta_n$ ($\Delta_n$ being the unit-simplex). Show that the $n \times n$ matrix $\mathbf{A}$ defined by

$$A_{ij} = \begin{cases} d_i - d_i^2, & i = j, \\ -d_i d_j, & i \neq j, \end{cases} \tag{116}$$

is positive semidefinite.

228

*Proof.* $\mathbf{A}$ can be represented as $\operatorname{diag}(\mathbf{d}) - \mathbf{d}\mathbf{d}^T$. For any vector $\mathbf{x} \in \mathbb{R}^n$, we have

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T(\operatorname{diag}(\mathbf{d}) - \mathbf{d}\mathbf{d}^T)\mathbf{x} = \mathbf{x}^T\operatorname{diag}(\mathbf{d})\mathbf{x} - \mathbf{x}^T\mathbf{d}\mathbf{d}^T\mathbf{x} = \sum_i^n (d_i - d_i^2)x_i^2 \geq 0 \tag{117}$$

229 where the last inequality follows from $0 \leq d_i \leq 1$ for any $i \in \{1, 2, \ldots, n\}$. □

**Exercise 2.12**

Prove that a $2 \times 2$ matrix $\mathbf{A}$ is negative semidefinite if and only if $\operatorname{Tr}(\mathbf{A}) \leq 0$ and $\det(\mathbf{A}) \leq 0$.

230

*Proof.* Without loss of generality, a $2 \times 2$ matrix $\mathbf{A}$ can be written as

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \tag{118}$$

Furthermore, the characteristic equation is given by

$$\det(\lambda\mathbf{I} - \mathbf{A}) = 0 \tag{119}$$

$$\begin{pmatrix} \lambda - a_{11} & -a_{12} \\ -a_{21} & \lambda - a_{22} \end{pmatrix} = 0 \tag{120}$$

$$\lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} = 0 \tag{121}$$

where $\lambda$ denotes the two roots ($\lambda_1$ and $\lambda_2$) of the characteristic equation, and also represents the set of the eigenvalues of $\mathbf{A}$. $\mathbf{A}$ is negative semidefinite if and only if both its two eigenvalues $\lambda_1$ and $\lambda_2$ are nonpositive. From the last equation above, we get

$$\begin{cases} \lambda_1 + \lambda_2 = a_{11} + a_{22} = \operatorname{Tr}(\mathbf{A}) \\ \lambda_1\lambda_2 = a_{11}a_{22} - a_{21}a_{21} = \det(\mathbf{A}) \end{cases} \tag{122}$$

which implies

$$\begin{cases} \operatorname{Tr}(\mathbf{A}) \leq 0 \\ \det(\mathbf{A}) \geq 0 \end{cases} \iff \lambda_1, \lambda_2 \leq 0 \tag{123}$$

231 which completes the proof. □

**Exercise 2.13**

For each of the following matrices determine whether they are positive/negative semidefinite/definite or indefinite:

(i) $\mathbf{A} = \begin{pmatrix} 2 & 2 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$

(ii) $\mathbf{B} = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 3 & 3 \\ 2 & 3 & 3 \end{pmatrix}$

(iii) $\mathbf{C} = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 2 & 1 \\ 3 & 1 & 2 \end{pmatrix}$

(iv) $\mathbf{D} = \begin{pmatrix} -5 & 1 & 1 \\ 1 & -7 & 1 \\ 1 & 1 & -5 \end{pmatrix}$

**Solution:**

(i) It is easy to know that $\mathbf{A}$ is diagonally dominant and its diagonal elements are positive. By Theorem 2.25 in the textbook, $\mathbf{A}$ is at least positive semidefinite. Since the principal minor $D_2(\mathbf{A}) = 0$, then $\mathbf{A}$ is not positive definite.

(ii) We observe that all the principal minors are nonnegative. Recall that the generalized Sylvester's criterion says that a hermitian matrix is positive-semidefinite if and only if all the principal minors are nonnegative[7]. Therefore, $\mathbf{B}$ is positive semidefinite.

(iii) It is easy to get $\mathrm{Tr}(\mathbf{C}) = 6$ and $\det(\mathbf{C}) = -2$, which implies that $\mathbf{C}$ has both positive and negative eigenvalues. This indicates $\mathbf{C}$ is indefinite.

(iv) Obviously, $-\mathbf{D}$ is a strictly diagonally dominant matrix whose diagonal elements are positive, so $-\mathbf{D}$ is positive definite. Hence, $\mathbf{D}$ is negative definite.

$\square$

**Exercise 2.14**

Let

$$\mathbf{D} = \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^T & c \end{pmatrix},$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{b} \in \mathbb{R}^n$, $c \in \mathbb{R}$. Suppose that $\mathbf{A} \succ \mathbf{0}$. Prove that $\mathbf{D} \succeq \mathbf{0}$ if and only if $c - \mathbf{b}^T \mathbf{A}^{-1} \mathbf{b} \geq 0$.

*Proof.* [8] Here we consider a more general case, i.e., $\mathbf{D} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix}$, where $\mathbf{B}$ and $\mathbf{C}$ are matrices instead of vectors or scalars, particularly, $\mathbf{C}$ is symmetric. Recall that $\mathbf{D}$ is positive semidefinite if

[7] https://en.wikipedia.org/wiki/Sylvester%27s_criterion

[8] https://inst.eecs.berkeley.edu/~ee127/sp21/livebook/thm_schur_compl.html

and only if $\mathbf{x}^T\mathbf{D}\mathbf{x} \geq 0$ for any vector $\mathbf{x}$. Let $\mathbf{x} = \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix}$, then

$$g(\mathbf{y}, \mathbf{z}) := \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix}^T \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \mathbf{y}^T\mathbf{A}\mathbf{y} + \mathbf{z}^T\mathbf{B}^T\mathbf{y} + \mathbf{y}^T\mathbf{B}\mathbf{z} + \mathbf{z}^T\mathbf{C}\mathbf{z} \geq 0, \quad \forall \mathbf{y}, \mathbf{z}. \tag{124}$$

This is equivalent to, for any $\mathbf{z}$,

$$0 \leq f(\mathbf{z}) := \min_{\mathbf{y}} g(\mathbf{y}, \mathbf{z}). \tag{125}$$

Since $\mathbf{A}$ is positive definite, $g(\mathbf{y}, \mathbf{z})$ is convex with respect to $\mathbf{y}$. Hence, minimizing $g(\mathbf{y}, \mathbf{z})$ w.r.t. $\mathbf{y}$ is an unconstrained convex problem. Setting the gradient $\nabla_{\mathbf{y}} g(\mathbf{y}, \mathbf{z})$ to 0, we get

$$\nabla_{\mathbf{y}} g(\mathbf{y}, \mathbf{z}) = 2\mathbf{A}\mathbf{y} + 2\mathbf{B}\mathbf{z} = 0 \Longleftrightarrow \mathbf{y} = -\mathbf{A}^{-1}\mathbf{B}\mathbf{z}. \tag{126}$$

Plugging this into $g(\mathbf{y}, \mathbf{z})$ yields

$$f(\mathbf{z}) = g(-\mathbf{A}^{-1}\mathbf{B}\mathbf{z}, \mathbf{z}) = \mathbf{z}^T(\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})\mathbf{z} \tag{127}$$

where $f(\mathbf{z}) \geq 0$ if and only if $\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}$ is positive semidefinite. $\square$

**Exercise 2.15**

For each of the following functions, determine whether it is coercive or not:

(i) $f(x_1, x_2) = x_1^4 + x_2^4$.

(ii) $f(x_1, x_2) = e^{x_1^2} + e^{x_2^2} - x_1^{200} + x_2^{200}$.

(iii) $f(x_1, x_2) = 2x_1^2 - 8x_1x_2 + x_2^2$.

(iv) $f(x_1, x_2) = 4x_1^2 + 2x_1x_2 + 2x_2^2$.

(v) $f(x_1, x_2, x_3) = x_1^3 + x_2^3 + x_3^3$.

(vi) $f(x_1, x_2) = x_1^2 - 2x_1x_2^2 + x_2^4$.

(vii) $f(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\|\mathbf{x}\|+1}$, where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is positive definite.

**Solution:**

(i)

$$\begin{aligned} f(x_1, x_2) &= x_1^4 + x_2^4 \\ &= (x_1^2 + x_2^2)^2 - 2x_1^2x_2^2 \\ &= (x_1^2 + x_2^2)^2 - \frac{(2x_1x_2)^2}{2} \geq (x_1^2 + x_2^2)^2 - \frac{(x_1^2 + x_2^2)^2}{2} \\ &= \frac{(x_1^2 + x_2^2)^2}{2} = \frac{\|\mathbf{x}\|^2}{2} \end{aligned}$$

which implies, as $\|\mathbf{x}\|^2 = x_1^2 + x_2^2 \to \infty$, $f(x_1, x_2) \to \infty$. Hence, $f(x_1, x_2)$ is coercive.

(ii) Since $e^x$ grows faster than $x^n$, $f(x_1, x_2)$ is coercive.

(iii) $f(x_1, x_2)$ can be written as $2(x_1 - 2x_2)^2 - 7x_2^2$. As $x_1^2 + x_2^2 \to \infty$ while $x_1 = 2x_2$, $f(x_1, x_2) \to -\infty$, which shows $f(x_1, x_2)$ is not coercive.

(iv) $f(x_1, x_2) = (x_1 + x_2)^2 + 3x_1^2 + x_2^2 \geq x_1^2 + x_2^2$. So, $f(x_1, x_2)$ is coercive since $x_1^2 + x_2^2 \to \infty$, $f(x_1, x_2) \to \infty$.

(v) $f(x_1, x_2, x_3)$ is not coercive since $f(x_1, x_2, x_3) \to -\infty$ as $x_1, x_2, x_3 \to -\infty$ while $x_1^2 + x_2^2 + x_3^2 \to \infty$.

(vi) $f(x_1, x_2)$ is not coercive since $f(x_1, x_2) = 0$ while for any $x_1, x_2$ satisfying $x_1 = x_2^2$.

(vii) $f(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A}\mathbf{x}}{\|\mathbf{x}\|+1} \leq \frac{\mathbf{x}^T \mathbf{A}\mathbf{x}}{\|\mathbf{x}\|}$ where the right hand side is the so-called Rayleigh quotient which is upper bounded by the maximum eigenvalue of $\mathbf{A}$ (see Lemma 1.11 in the textbook). Hence, $f(\mathbf{x})$ is not coercive.

□

**Exercise 2.16**

Find a function $f : \mathbb{R}^2 \to \mathbb{R}$ which is not coercive and satisfied that for any $\alpha \in \mathbb{R}$

$$\lim_{|x_1| \to \infty} f(x_1, \alpha x_1) = \lim_{|x_2| \to \infty} f(\alpha x_2, x_2) = \infty. \tag{128}$$

**Solution:** Consider the following function

$$f(x_1, x_2) = \frac{1 + x_1 x_2}{|x_1| + |x_2|} \tag{129}$$

which goes to $-\infty$ when $x_1^2 + x_2^2 \to \infty$ while $x_1 = -x_2$. Also, when $x_2 = \alpha x_1$, we have

$$\lim_{|x_1| \to \infty} f(x_1, \alpha x_1) = \lim_{|x_1| \to \infty} \frac{1 + x_1^2}{(1 + |\alpha|)|x_1|} = \infty. \tag{130}$$

The similar argument follows for the case where $\lim_{|x_2| \to \infty} f(\alpha x_2, x_2) = \infty$. □

**Exercise 2.17**

For each of the following functions, find all the stationary points and classify them according to whether they are saddle points, strict/nonstrict local/global minimum/maximum points:

(i) $f(x_1, x_2) = (4x_1^2 - x_2)^2$.

(ii) $f(x_1, x_2, x_3) = x_1^4 - 2x_1^2 + x_2^2 + 2x_2 x_3 + 2x_3^2$.

(iii) $f(x_1, x_2) = 2x_2^3 - 6x_2^2 + 3x_1^2 x_2$.

(iv) $f(x_1, x_2) = x_1^4 + 2x_1^2 x_2 + x_2^2 - 4x_1^2 - 8x_1 - 8x_2$.

**Solution:**

(i) First,

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 16x_1(4x_1^2 - x_2) \\ -2(4x_1^2 - x_2) \end{pmatrix} \tag{131}$$

Hence, the stationary points are those satisfying

$$16x_1(4x_1^2 - x_2) = 0 \tag{132}$$

$$-2(4x_1^2 - x_2) = 0 \tag{133}$$

The first equation means that either $x_1 = 0$ or $x_2 = 4x_1^2$. If $x_1 = 0$, then by the second equation, $x_2 = 0$. If $x_2 = 4x_1^2$, then the second equation is satisfied automatically. Hence, the stationary points are those satisfying $x_2 = 4x_1^2$. For the stationary points $(x_1, 4x_1^2)$, we have $f(x_1, 4x_1^2)) = 0$. Since $f(x_1, x_2)$ is lower bounded by 0, the points satisfying $x_2 = 4x_1^2$ are nonstrict global minimum points.

(ii) The gradient is given by

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 4x_1(x_1^2 - 1) \\ 2(x_2 + x_3) \\ 2(x_2 + 2x_3) \end{pmatrix}. \tag{134}$$

Therefore, the stationary points are those satisfying

$$4x_1(x_1^2 - 1) = 0 \tag{135}$$

$$2(x_2 + x_3) = 0 \tag{136}$$

$$2(x_2 + 2x_3) = 0. \tag{137}$$

The first equation gives $x_1 = 0$ or $x_1^2 = 1$. The second and the third equations give $x_2 = x_3 = 0$. So, the stationary points are $x_1 = 0, x_2 = 0, x_3 = 0$, $x_1 = 1, x_2 = 0, x_3 = 0$, and $x_1 = -1, x_2 = 0, x_3 = 0$. Furthermore, the Hessian is given by

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 4(3x_1^2 - 1) & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & 4 \end{pmatrix}. \tag{138}$$

Then, $\nabla^2 f(0,0,0)$ is indefinite, implying $x_1 = 0, x_2 = 0, x_3 = 0$ is a saddle point. Both $\nabla^2 f(1,0,0)$ and $\nabla^2 f(-1,0,0)$ are positive definite. Thus, both $x_1 = 1, x_2 = 0, x_3 = 0$ and $x_1 = -1, x_2 = 0, x_3 = 0$ are nonstrict minimum points. Moreover, $f(x_1, x_2, x_3)$ can be written as $x_1^2(x_1^2 - 2) + (x_2 + x_3)^2 + x_3^2$. As $\|\mathbf{x}\| \to \infty$, $f(x_1, x_2, x_3) \to \infty$. Hence, $f(x_1, x_2, x_3)$ is coercive and has a global minimum point. Since $f(1,0,0) = f(-1,0,0) = -1$, they are nonstrict global minimum points.

(iii) First,

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 6x_1x_2 \\ 6x_2^2 - 12x_2 + 3x_1^2 \end{pmatrix} \tag{139}$$

Then the stationary points are those satisfying

$$6x_1x_2 = 0 \tag{140}$$

$$6x_2^2 - 12x_2 + 3x_1^2 = 0. \tag{141}$$

From the first equation, $x_1 = 0$ or $x_2 = 0$. Combining with the second equation, if $x_1 = 0$, $x_2 = 0$ or $x_2 = 2$. If $x_2 = 0$, $x_1 = 0$. Therefore, the stationary points are $x_1 = 0$, $x_2 = 0$ and $x_1 = 0$, $x_2 = 2$. $f(x_1, x_2)$ can be written as $x_2(2(x_2 - 3/2)^2 - 9/2 + 3x_1^2)$, which implies that for any $x_1$, as $x_2 \to -\infty$, $f(x_1, x_2) \to -\infty$, and as $x_2 \to \infty$, $f(x_1, x_2) \to \infty$. Hence, $f(x_1, x_2)$ does not have global minimum and maximum points. Now consider the Hessian

$$\nabla^2 f(\mathbf{x}) = 6 \begin{pmatrix} x_2 & x_1 \\ x_1 & 2x_2 - 2. \end{pmatrix} \tag{142}$$

Then we have $\nabla^2 f(0,0) = 6\begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix} \preceq \mathbf{0}$ and $\nabla^2 f(0,2) = 6\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \succeq \mathbf{0}$. Thus, $x_1 = 0, x_2 = 0$ is a local maximum point and $x_1 = 0, x_2 = 2$ is a local minimum point.

(iv) First,

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 4x_1^3 + 4x_1x_2 - 8x_1 - 8 \\ 2x_1^2 + 2x_2 - 8 \end{pmatrix} \tag{143}$$

from which we know the stationary points are those that satisfy

$$x_1(x_1^2 + x_2) - 2x_1 - 2 = 0 \tag{144}$$

$$x_1^2 + x_2 - 4 = 0 \tag{145}$$

which gives $x_1 = 1$ and $x_2 = 3$. Now we consider the Hessian

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 12x_1^2 + 4x_2 - 8 & 4x_1 \\ 4x_1 & 2 \end{pmatrix}. \tag{146}$$

Then we have

$$\nabla^2 f(1,3) = \begin{pmatrix} 16 & 4 \\ 4 & 2 \end{pmatrix} \succeq \mathbf{0} \tag{147}$$

where the positive definiteness follows from Proposition 2.20 in the textbook. Due to the terms $x_1^4$ and $x_2^2$ in $f$, $f$ is coercive. Hence, $x_1 = 1, x_2 = 3$ is the global minimum point.

□

# 3 Chapter 3 Least Squares

**Exercise 3.1**

Let $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m, \mathbf{L} \in \mathbb{R}^{p\times n}$, and $\lambda \in \mathbb{R}_{++}$. Consider the regularized least squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda\|\mathbf{L}\mathbf{x}\|^2. \tag{RLS}$$

Show that (RLS) has a unique solution if and only if $\mathrm{Null}(\mathbf{A}) \cap \mathrm{Null}(\mathbf{L}) = \{\mathbf{0}\}$, where here for a matrix $\mathbf{B}$, $\mathrm{Null}(\mathbf{B})$ is the null space of $\mathbf{B}$ given by $\{\mathbf{x} : \mathbf{B}\mathbf{x} = \mathbf{0}\}$.

Note that it is supposed to be $\mathbf{b} \in \mathbb{R}^m$ instead of $\mathbf{b} \in \mathbb{R}^n$. In the textbook, this is a typo which is not yet mentioned at http://www.siam.org/books/mo19/mo19_err.pdf.

*Proof.* Since the Hessian of the objective function is $2(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{L}^T\mathbf{L}) \succeq \mathbf{0}$, it follows by Lemma 2.41 of the textbook that any stationary point is a global minimum point. Then, we have

$$\text{(RLS) has a unique solution} \iff \mathbf{A}^T\mathbf{A} + \lambda\mathbf{L}^T\mathbf{L} \succ \mathbf{0}$$

$$\Updownarrow$$

$$\mathbf{x}^T(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{L}^T\mathbf{L})\mathbf{x} > 0, \forall \mathbf{x} \neq \mathbf{0} \iff \|\mathbf{A}\mathbf{x}\|^2 + \lambda\|\mathbf{L}\mathbf{x}\|^2 > 0, \forall \mathbf{x} \neq \mathbf{0}$$

$$\Updownarrow$$

$$\text{There exists no nonzero } \mathbf{x} \text{ such that } \mathbf{A}\mathbf{x} = \mathbf{0} \text{ and } \mathbf{L}\mathbf{x} = \mathbf{0} \text{ hold simultaneously.}$$

$$\Updownarrow$$

$$\mathrm{Null}(\mathbf{A}) \cap \mathrm{Null}(\mathbf{L}) = \{\mathbf{0}\}.$$

This completes the proof. □

# 4 Chapter 4 The Gradient Method

Before working on the exercises of Chapter 4, we first introduce the notation of $f \in C_L^{k,p}(D)$. We write $f \in C_L^{k,p}(D)$ if

1. $f^{(k)}$ exists and is continuous on $D$.
2. $f^{(p)}$ is Lipschitz continuous with a constant $L$, namely,

$$\|f^{(p)}(y_1) - f^{(p)}(y_2)\| \le L\|y_1 - y_2\|, \quad \forall y_1, y_2 \in D.$$

**Exercise 4.1**

Let $f \in C_L^{1,1}(\mathbb{R}^n)$ and let $\{\mathbf{x}^k\}_{k\ge 0}$ be the sequence generated by the gradient method with a constant stepsize $t_k = \frac{1}{L}$. Assume that $\mathbf{x}_k \to \mathbf{x}^*$. Show that if $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ for all $k \ge 0$, then $\mathbf{x}^*$ is not a local maximum point.

*Proof.* Suppose $\mathbf{x}^*$ is a local maximum point, then there exists a ball $B(\mathbf{x}^*, r)$ with any $r > 0$ such that

$$f(\mathbf{x}^*) \ge f(\mathbf{x}_k), \quad \forall \mathbf{x}_k \in B(\mathbf{x}^*, r)$$

Since $t_k = \frac{1}{L}$, by the descent lemma (Lemma 4.22 in the textbook), we have

$$\begin{aligned} f(\mathbf{x}^*) &\le f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x}^* - \mathbf{x}_k) + \frac{L}{2}\|\mathbf{x}^* - \mathbf{x}_k\|^2 \\ &= f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(-\frac{1}{L}\nabla f(\mathbf{x}_k)) + \frac{L}{2}\|-\frac{1}{L}\nabla f(\mathbf{x}_k)\|^2 \\ &= f(\mathbf{x}_k) - \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2 \\ &< f(\mathbf{x}_k) \end{aligned}$$

where the last line follows from that $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ for all $k \ge 0$. This contradicts the supposition, which implies that $\mathbf{x}^*$ is not a local maximum point. This completes the proof. $\square$

# 5 Chapter 5 Newton's Method

# 6 Chapter 6 Convex Sets

# 7 Chapter 7 Convex Functions

**Exercise 7.36**

Prove that for any $x_1, x_2, \ldots, x_n \in \mathbb{R}_+$ the following inequality holds:

$$\frac{\sum_{i=1}^n x_i}{n} \le \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}}$$

*Proof.* According to Cauchy-Schwartz inequality which says that given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\|\mathbf{x}\|_2\|\mathbf{y}\|_2 \ge |\mathbf{x}^T\mathbf{y}|$, we have

$$\sqrt{\frac{\sum_{i=1}^n x_i^2}{n}} = \sqrt{\sum_{i=1}^n (\frac{|x_i|}{\sqrt{n}})^2} \cdot \overbrace{\sqrt{\sum_{i=1}^n (\frac{1}{\sqrt{n}})^2}}^{=1}$$

$$\geq \frac{\sum_{i=1}^n |x_i|}{n} \geq \frac{\sum_{i=1}^n x_i}{n},$$

where the equalities in the first and second inequalities hold if and only if $|x_1|=|x_2|=\cdots=|x_n|$ and $x_1=x_2=\cdots=x_n$, respectively. This completes the proof. $\square$

**Exercise 7.37**

Prove that for any $x_1, x_2, \ldots, x_n \in \mathbb{R}_{++}$ the following inequality holds:

$$\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i} \leq \sqrt{\frac{\sum_{i=1}^n x_i^3}{\sum_{i=1}^n x_i}}$$

*Proof.* Let $f(x)=x^2$ and then $f''(x)=2>0$ implying that $f$ is convex. Furthermore, given $\lambda_1, \lambda_2, \ldots, \lambda_n \in [0,1]$ satisfying $\sum_{i=1}^n \lambda_i = 1$, we have

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

By letting $\lambda_i = \frac{x_i}{\sum_{i=1}^n x_i}$, we have

$$\left(\sum_{i=1}^n \frac{x_i}{\sum_{i=1}^n x_i} x_i\right)^2 \leq \sum_{i=1}^n \frac{x_i}{\sum_{i=1}^n x_i} x_i^2 \iff \left(\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i}\right)^2 \leq \frac{\sum_{i=1}^n x_i^3}{\sum_{i=1}^n x_i} \iff \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i} \leq \sqrt{\frac{\sum_{i=1}^n x_i^3}{\sum_{i=1}^n x_i}}.$$

Note that the condition $\lambda_i \in [0,1]$ is satisfied automatically since $x_i > 0, \forall i = 1, 2, \ldots, n$. This completes our proof. $\square$

**Exercise 7.38**

Let $x_1, x_2, \ldots, x_n > 0$ satisfy $\sum_{i=1}^n x_i = 1$. Prove that

$$\sum_{i=1}^n \frac{x_i}{\sqrt{1-x_i}} \geq \sqrt{\frac{n}{n-1}}.$$

*Proof.* Define $f(x) = 1/\sqrt{1-x}$ and then $f''(x) = \frac{3}{4}(1-x)^{-5/2} > 0$. So $f(x)$ is convex. Since $\sum_{i=1}^n x_i = 1$, then we have

$$\begin{aligned}
\sum_{i=1}^n x_i f(x_i) \geq & f\left(\sum_{i=1}^n x_i \cdot x_i\right) = f\left(\sum_{i=1}^n x_i^2\right) \\
=& 1/\sqrt{1-\sum_{i=1}^n x_i^2} \\
\geq & 1/\sqrt{1-\frac{(\sum_{i=1}^n x_i)^2}{n}} \\
=& 1/\sqrt{1-\frac{1}{n}} = 1/\sqrt{\frac{n-1}{n}} \\
=& \sqrt{\frac{n}{n-1}}
\end{aligned}$$

where the second inequality follows from the result given in Exercise 7.36. $\square$

**Exercise 7.39**

Prove that for any $a, b, c > 0$ the following inequality holds:

$$\frac{9}{a+b+c} \leq 2\left(\frac{1}{a+b} + \frac{1}{b+c} + \frac{1}{c+a}\right)$$

To simplify the proof of Exercise 7.39, we introduce the following theorem which says that the **harmonic mean** (HM) is less than or equal to the **geometric mean** (GM).

**Theorem 7.1 (HM≤GM).** *For any $x_1, x_2, \cdots, x_n > 0$ the following inequality holds:*

$$\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \leq \sqrt[n]{\prod_{i=1}^{n} x_i}$$

*Proof.* According to AGM inequality, for any $a_1, a_2, \cdots, a_n \geq 0$, we have

$$\frac{1}{n}\sum_{i=1}^{n} a_i \geq \sqrt[n]{\prod_{i=1}^{n} a_i}.$$

Replacing $a_i$ with $\frac{1}{x_i}$ where $x_i > 0$ for $i \in \{1, 2, \ldots, n\}$, we get

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i} \geq \sqrt[n]{\prod_{i=1}^{n} \frac{1}{x_i}}.$$

Since both sides are positive, taking reciprocals and reversing the inequality yield

$$\frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}} \leq \frac{1}{\sqrt[n]{\prod_{i=1}^{n} \frac{1}{x_i}}}$$

$$\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \leq \sqrt[n]{\prod_{i=1}^{n} \frac{1}{x_i}},$$

as desired. □

Naturally, we get the following corollary in which AM is short for the arithmetic mean.

**Corollary 7.2 (HM≤GM≤AM).** *For any $x_1, x_2, \cdots, x_n > 0$ the following inequality holds:*

$$\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \leq \sqrt[n]{\prod_{i=1}^{n} x_i} \leq \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}$$

*Proof.* The first inequality and the second inequality are exactly Theorem 7.1 and AGM inequality, respectively. □

Now we prove Exercise 7.39 using Corollary 7.2.

*Proof.* Since HM≤AM, letting $x_1 = \frac{2}{a+b}$, $x_2 = \frac{2}{b+c}$ and $x_3 = \frac{2}{c+a}$ yields

$$\begin{aligned}
\frac{3}{\frac{1}{\frac{2}{a+b}} + \frac{1}{\frac{2}{b+c}} + \frac{1}{\frac{2}{c+a}}} &\le \frac{\frac{2}{a+b} + \frac{2}{b+c} + \frac{2}{c+a}}{3} \\
\frac{3}{a+b+c} &\le \frac{2}{3}\left(\frac{1}{a+b} + \frac{1}{b+c} + \frac{1}{c+a}\right) \\
\frac{9}{a+b+c} &\le 2\left(\frac{1}{a+b} + \frac{1}{b+c} + \frac{1}{c+a}\right),
\end{aligned}$$

315 as desired. □

**Exercise 7.40**

(i) Prove that the function $f(x) = \frac{1}{1+e^x}$ is strictly convex over $[0, \infty)$.

(ii) Prove that for any $a_1, a_2, \ldots, a_n \ge 1$ the equality

$$\sum_{i=1}^{n} \frac{1}{1+a_i} \ge \frac{n}{1 + \sqrt[n]{a_1 a_2 \cdots a_n}}$$

holds.

316

*Proof.* (i) The second derivative is given by

$$f''(x) = \frac{e^x(e^x - 1)}{(1+e^x)^3} > 0, \quad x > 0$$

Thus, $f(x)$ is strictly convex on $(0, +\infty)$. By Theorem 7.13 in the textbook, $f''(x) > 0$ is a sufficient, not necessary, condition for strict convexity. Even though $f''(x) = 0$ at the unique boundary point $x = 0$, this does not alter the strict convexity of $f(x)$. To see this, recall the definition of strict convexity, i.e. Definition 7.2, that is, for any $\underline{x \ne y \in C, \lambda \in (0,1)}$,

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y).$$

317 It is easy to see that for any $y > x = 0$, the above always holds for any $\lambda \in (0,1)$. Thus, $\frac{1}{1+e^x}$
318 is strictly convex over $[0, +\infty]$.

(ii) Let $a_i = e^{x_i}, i = 1, \ldots, n$. Then for any $a_i \ge 1$, $x_i \ge 0$. Since $f(x) = \frac{1}{1+e^x}$ is strictly convex, then

$$\begin{aligned}
\sum_{i=1}^{n} \frac{1}{n} \cdot \frac{1}{1+a_i} = \sum_{i=1}^{n} \frac{1}{n} \cdot \frac{1}{1+e^{x_i}} &\ge \frac{1}{1+e^{1/n * \sum_{i=1}^{n} x_i}} \\
&= \frac{1}{1+(e^{\sum_{i=1}^{n} x_i})^{1/n}} \\
&= \frac{1}{1+(\prod_{i=1}^{n} e^{x_i})^{1/n}} \\
&= \frac{1}{1+(\prod_{i=1}^{n} a_i)^{1/n}} = \frac{1}{1+\sqrt[n]{a_1 a_2 \cdots a_n}}
\end{aligned}$$

Multiplying both sides by $n$ gives the claim, namely,

$$\sum_{i=1}^{n} \frac{1}{1+a_i} \ge \frac{n}{1 + \sqrt[n]{a_1 a_2 \cdots a_n}}$$

Since $\frac{1}{1+e^x}$ is strictly convex, the equality holds if and only if $a_1 = a_2 = \cdots = a_n = 1$. This completes our proof.

$\square$

# 8 Chapter 8 Convex Optimization

**Exercise 8.1**

Consider the problem

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g(\mathbf{x}) \le 0 \\ & \mathbf{x} \in X \end{aligned} \tag{P}$$

where $f$ and $g$ are convex functions over $\mathbb{R}^n$ and $X \subseteq \mathbb{R}^n$ is a convex set. Suppose that $\mathbf{x}^*$ is an optimal solution of (P) that satisfies $g(\mathbf{x}^*) < 0$. Show that $\mathbf{x}^*$ is also an optimal solution of the problem

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in X \end{aligned}$$

*Proof.* We denote the feasible sets of (P) and the second problem by $C_p$ and $C$, respectively. Since $f(\mathbf{x}), g(\mathbf{x})$ and $X$ are convex, both $C_p$ and $C$ are convex sets with $C_p \subseteq C$. Since $g(\mathbf{x}^*) < 0$, $\mathbf{x}^* \in \text{int}(C_p)$. This indicates that the second problem has a local optimal solution on $C_p$, i.e. $\mathbf{x}^*$. By Theorem 8.1, we know that a local minimum is also a global minimum in terms of convex optimization. Hence, $\mathbf{x}^*$ is also an optimal solution of the problem without the constraint of $g(\mathbf{x}) \le 0$. $\square$

**Exercise 8.2**

Let $C = B[\mathbf{x}_0, r]$, where $\mathbf{x}_0 \in \mathbb{R}^n$ and $r > 0$ are given. Find a formula for the orthogonal projection operator $P_C$.

**Solution:** Given $\mathbf{x} \in \mathbb{R}^n$, we want to find its projection onto the closed ball $B[\mathbf{x}_0, r]$. Then the optimization problem associated with the computation of $P_C(\mathbf{x})$ is given by

$$\min_{\mathbf{y}} \{\|\mathbf{y} - \mathbf{x}\|^2 | \; \|\mathbf{y} - \mathbf{x}_0\|^2 \le r^2\}.$$

If $\|\mathbf{x} - \mathbf{x}_0\| \le r$, then obviously $\mathbf{y} = \mathbf{x}$ since it corresponds to the optimal value 0. When $\|\mathbf{x} - \mathbf{x}_0\| > r$, then the optimal solution must belong to the boundary of the ball due to Theorem 2.6 in the textbook. Specifically, Theorem 2.6 says that for a differentiable function $f(\mathbf{x})$, if $\mathbf{x}^*$ is a local optimum point, then $\nabla f(\mathbf{x}^*) = 0$. Accordingly,

$$2(\mathbf{y} - \mathbf{x}) = 0 \Longleftrightarrow \mathbf{y} = \mathbf{x},$$

which is impossible since $\mathbf{x} \notin C$. Thus, we conclude that in the case of $\|\mathbf{x} - \mathbf{x}_0\| > r$, the projection problem is equivalent to

$$\begin{aligned} & \min_{\mathbf{y}} \{\|\mathbf{y} - \mathbf{x}\|^2 | \; \|\mathbf{y} - \mathbf{x}_0\|^2 = r^2\} \\ \Longleftrightarrow & \min_{\mathbf{y}} \{\|\mathbf{y} - \mathbf{x}_0 + \mathbf{x}_0 - \mathbf{x}\|^2 | \; \|\mathbf{y} - \mathbf{x}_0\|^2 = r^2\} \\ \Longleftrightarrow & \min_{\mathbf{y}} \{\|\mathbf{y} - \mathbf{x}_0\|^2 + 2\langle \mathbf{y} - \mathbf{x}_0, \mathbf{x}_0 - \mathbf{x}\rangle + \|\mathbf{x}_0 - \mathbf{x}\|^2 | \; \|\mathbf{y} - \mathbf{x}_0\|^2 = r^2\} \\ \Longleftrightarrow & \min_{\mathbf{y}} \{r^2 + 2\langle \mathbf{y} - \mathbf{x}_0, \mathbf{x}_0 - \mathbf{x}\rangle + \|\mathbf{x}_0 - \mathbf{x}\|^2 | \; \|\mathbf{y} - \mathbf{x}_0\|^2 = r^2\}. \end{aligned}$$

After dropping those terms that are not depend on $\mathbf{y}$, we get the equivalent form as follows.

$$\underset{\mathbf{y}}{\operatorname{argmin}} \; \{\langle \mathbf{y}, \mathbf{x}_0 - \mathbf{x} \rangle \mid \|\mathbf{y} - \mathbf{x}_0\|^2 = r^2\}$$

By the Cauchy-Schwarz inequality, the objective function can be lower bounded by

$$\langle \mathbf{y}, \mathbf{x}_0 - \mathbf{x} \rangle \geq -\|\mathbf{y}\|\|\mathbf{x}_0 - \mathbf{x}\| = -r\|\mathbf{x}_0 - \mathbf{x}\|,$$

and this lower bound can be attained at $\mathbf{y} = r\frac{\mathbf{x}-\mathbf{x}_0}{\|\mathbf{x}-\mathbf{x}_0\|}$. Therefore, the orthogonal projection operator $P_C$ is

$$P_{B[\mathbf{x}_0, r]} = \begin{cases} \mathbf{x}, & \text{if } \|\mathbf{x}\| \leq r \\ r\frac{\mathbf{x}-\mathbf{x}_0}{\|\mathbf{x}-\mathbf{x}_0\|}, & \text{if } \|\mathbf{x}\| > r. \end{cases}$$

□

# 9 Chapter 9 Optimization over a Convex Set

**Exercise 9.1**

Let $f$ be a continuously differentiable convex function over a closed and convex set $C \subseteq \mathbb{R}^n$. Show that $x^* \in C$ is an optimal solution of the problem

$$\min \{f(\mathbf{x}) : \mathbf{x} \in C\} \tag{P}$$

if and only if

$$\langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle \leq 0 \text{ for all } \mathbf{x} \in C.$$

The necessity is easy to show, but proving the sufficiency is hard. On Math StackExchange, Parasseux Nguyen provides a beautiful proof for the sufficiency[9].

*Proof.* We first show the necessity. Since $x^* \in C$ is an optimal solution of (P), then we have

$$f(\mathbf{x}^*) - f(\mathbf{x}) \leq 0.$$

By the convexity of $f$, we have

$$f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle \leq f(\mathbf{x}^*) \iff \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle \leq f(\mathbf{x}^*) - f(\mathbf{x}) \leq 0.$$

Proving the sufficiency is not trivial. For all $\mathbf{x} \in C$, let $\mathbf{v} = \mathbf{x} - \mathbf{x}^*$ and then $\mathbf{x}^* + t\mathbf{v} = (1-t)\mathbf{x}^* + t\mathbf{x} \in C$. Define $g(t) = f(\mathbf{x}^* + t\mathbf{v})$ on $t \in [0,1]$. Since $f$ is continuously differentiable over $C$, then $g(t)$ is also continuously differentiable on $[0,1]$. Furthermore,

$$\begin{aligned} g'(t) &= \langle \nabla f(\mathbf{x}^* + t\mathbf{v}), \mathbf{v} \rangle \\ &= \frac{1}{t}\langle \nabla f(\mathbf{x}^* + t\mathbf{v}), t\mathbf{v} \rangle \\ &= \frac{1}{t}\langle \nabla f(\mathbf{x}^* + t\mathbf{v}), (\mathbf{x}^* + t\mathbf{v}) - \mathbf{x}^* \rangle \\ &= -\frac{1}{t}\langle \nabla f(\mathbf{x}^* + t\mathbf{v}), \mathbf{x}^* - (\mathbf{x}^* + t\mathbf{v}) \rangle \\ &\geq 0 \end{aligned}$$

where the inequality follows from the premise of $\langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle \leq 0$ for all $\mathbf{x} \in C$. □

*Note.* It is interesting to note that from the above proof, we can see that the convexity of $f$ is not required for the sufficiency and we only used the convexity of $C$.

[9] https://math.stackexchange.com/questions/4178673/if-nabla-fxt-x-x-leq-0-for-all-x-in-c-then-x-is-optimal-so?

# Bibliography

Chen, J., Yu, C., and Jin, L. (2019). *Mathematical Analysis, third edition.*

Horn, R. A. and Johnson, C. R. (2013). *Matrix Analysis, Second Edition.*

Jax (2016). Minkowski inequality for $0 < p < 1$. https://math.stackexchange.com/questions/73294/minkowski-inequality-for-p-le-1. [Online; accessed 25-May-2022].