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: November 24, 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

$$\|\mathbf{A}\|_{a,b}= \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_b: \|\mathbf{x}\|_a\leq 1\}. \tag{1.1}$$

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

$$\|\mathbf{A}\|_{a,b}= \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_b: \|\mathbf{x}\|_a= 1\} = \max_{\|\mathbf{x}\|_a\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a}. \tag{1.2}$$

Now we show that they are valid alternatives of (1.1) by proving two lemmas. The first alternative is exactly the following lemma.

Lemma 1.1. *The maximum points* $\mathbf{x}^*$ *of the RHS of* (1.1) *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

$$\|\mathbf{A}\mathbf{y}\|_b= \|\mathbf{A}(k\mathbf{x}^*)\|_b= |k|\|\mathbf{A}\mathbf{x}^*\|_b> \|\mathbf{A}\mathbf{x}^*\|_b \tag{1.3}$$

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* $\mathbf{x}\in\mathbb{R}^n$,

$$\|\mathbf{A}\|_{a,b}= \max_{\|\mathbf{x}\|_a\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a}. \tag{1.4}$$

Proof. An equivalent form of Lemma 1.1 is

$$\|\mathbf{A}\|_{a,b}= \max_{\mathbf{y}}\left\{\frac{\|\mathbf{A}\mathbf{y}\|_b}{\|\mathbf{y}\|_a} : \|\mathbf{y}\|_a= 1\right\} = \max_{\|\mathbf{y}\|_a=1} \frac{\|\mathbf{A}\mathbf{y}\|_b}{\|\mathbf{y}\|_a}. \tag{1.5}$$

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

$$\|\mathbf{A}\|_{a,b}= \max_{|k|\|\mathbf{x}\|_a=1} \frac{|k|\|\mathbf{A}\mathbf{x}\|_b}{|k|\|\mathbf{x}\|_a} = \max_{\|\mathbf{x}\|_a=1/|k|} \frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a} = \max_{\|\mathbf{x}\|_a\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a} \tag{1.6}$$

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 1.4.

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

$$\|\mathbf{A}\mathbf{x}\|_b\leq \|\mathbf{A}\|_{a,b}\|\mathbf{x}\|_a \tag{1.7}$$

holds.

Proof. According to Lemma 1.4, for any $\mathbf{x} \neq \mathbf{0}$, it follows that

$$\frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a} \leq \|\mathbf{A}\|_{a,b} \Longleftrightarrow \|\mathbf{A}\mathbf{x}\|_b \leq \|\mathbf{A}\|_{a,b}\|\mathbf{x}\|_a \tag{1.8}$$

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 ∂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:*

$$\mathrm{cl}(U) = \bigcap \{T : U \subseteq T,\ T \text{ is closed}\}. \tag{1.9}$$

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

$$\mathrm{cl}(U) = U \cup \mathrm{bd}(U). \tag{1.10}$$

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* $\mathrm{int}(U)$*:*

$$\mathrm{int}(U) = \{\mathbf{x} \in U : B(\mathbf{x}, r) \subseteq U \text{ for some } r > 0\}. \tag{1.11}$$

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).

$$\left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c \tag{1.12}$$

$$\left(\bigcap_{i\in I} A_i\right)^c = \bigcup_{i\in I} A_i^c \tag{1.13}$$

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.

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

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

Taking summation over i gives

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

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

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

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 $\mathbf{x}, \mathbf{y} \in \mathbf{C}^n$.

Extensions

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

Exercise 1.2

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

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

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

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

$$
\Downarrow
$$

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

$$
\Downarrow
$$

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

$$
\Downarrow
$$

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

□

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. □

Exercise 1.4

Prove the Cauchy-Schwarz inequality (Lemma 1.5)

$$
|\mathbf{x}^T \mathbf{y}| \le \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{1.14}
$$

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.

$$
\mathbf{x}^T \mathbf{y} = \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 \cdot \cos\theta \tag{1.15}
$$

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

$$
-\|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 \le \mathbf{x}^T \mathbf{y} = \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 \cdot \cos\theta \le \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 \tag{1.16}
$$

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 (1.16) in a compact way, then we get

$$|\mathbf{x}^T\mathbf{y}|\leq \|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{1.17}$$

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

$$\|\mathbf{A}+\mathbf{B}\|_{a,b}\leq \|\mathbf{A}\|_{a,b}+\|\mathbf{B}\|_{a,b} \tag{1.18}$$

holds.

Proof. By the definition of the induced norm, namely (1.1),

$$\begin{aligned}
\|\mathbf{A}+\mathbf{B}\|_{a,b} &= \max_{\mathbf{x}}\{\|(\mathbf{A}+\mathbf{B})\mathbf{x}\|_b\colon \|\mathbf{x}\|_a\leq 1\} && (1.19)\\
&= \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{x}\|_b\colon \|\mathbf{x}\|_a\leq 1\} && (1.20)\\
&\leq \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_b+\|\mathbf{B}\mathbf{x}\|_b\colon \|\mathbf{x}\|_a\leq 1\} && (1.21)\\
&\leq \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_b\colon \|\mathbf{x}\|_a\leq 1\} + \max_{\mathbf{x}}\{\|\mathbf{B}\mathbf{x}\|_b\colon \|\mathbf{x}\|_a\leq 1\} && (1.22)\\
&= \|\mathbf{A}\|_{a,b}+\|\mathbf{B}\|_{a,b} && (1.23)
\end{aligned}$$

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 δ 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 $\|\mathbf{x}\|> \|\mathbf{x}_0\|$,

$$\begin{aligned}
|f(\mathbf{x}) - f(\mathbf{x}_0)| &= \|\mathbf{x}\|-\|\mathbf{x}_0\| && (1.24)\\
&= \|\mathbf{x}-\mathbf{x}_0+\mathbf{x}_0\|-\|\mathbf{x}_0\| && (1.25)\\
&\leq \|\mathbf{x}-\mathbf{x}_0\|+\|\mathbf{x}_0\|-\|\mathbf{x}_0\| && (1.26)\\
&= \|\mathbf{x}-\mathbf{x}_0\|< \delta < \epsilon. && (1.27)
\end{aligned}$$

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

$$\begin{aligned}
|f(\mathbf{x}) - f(\mathbf{x}_0)| &= \|\mathbf{x}_0\|-\|\mathbf{x}\| && (1.28)\\
&= \|\mathbf{x}_0-\mathbf{x}+\mathbf{x}\|-\|\mathbf{x}\| && (1.29)\\
&\leq \|\mathbf{x}-\mathbf{x}_0\|+\|\mathbf{x}\|-\|\mathbf{x}\| && (1.30)\\
&= \|\mathbf{x}-\mathbf{x}_0\|< \delta < \epsilon. && (1.31)
\end{aligned}$$

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

$$\|\mathbf{A}\|_{a,b} = \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_b \colon \|\mathbf{x}\|_a = 1\}. \tag{1.32}$$

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

$$\|\mathbf{A}\|_{a,b} = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_b}{\|\mathbf{x}\|_a}. \tag{1.33}$$

Proof. This is exactly Lemma 1.2 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

$$\|\mathbf{A}\mathbf{B}\|_{a,c} \leq \|\mathbf{A}\|_{b,c}\|\mathbf{B}\|_{a,b}. \tag{1.34}$$

Proof. From Exercise 1.9, we have

$$\|\mathbf{A}\mathbf{B}\|_{a,c} \leq \frac{\|\mathbf{A}\mathbf{B}\mathbf{x}\|_c}{\|\mathbf{x}\|_a} \tag{1.35}$$

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,

$$\|\mathbf{A}\mathbf{B}\|_{a,c} \leq \frac{\|\mathbf{A}\mathbf{y}\|_c}{\|\mathbf{y}\|_b}\frac{\|\mathbf{B}\mathbf{x}\|_b}{\|\mathbf{x}\|_a} \leq \|\mathbf{A}\|_{b,c}\|\mathbf{B}\|_{a,b}. \tag{1.36}$$

This completes the proof. □

Exercise 1.11

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

$$\|\mathbf{A}\|_\infty = \max_{i=1,2,\ldots,m} \sum_{j=1}^{n} |A_{i,j}|. \tag{1.37}$$

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

$$\begin{aligned}
\|\mathbf{A}\|_\infty &= \max_{\mathbf{x}} \{\|\mathbf{A}\mathbf{x}\|_\infty \colon \|\mathbf{x}\|_\infty = 1\} && (1.38)\\
&= \max_{\mathbf{x}} \left\{ \max_{i=1,\ldots,m} \Big| \sum_{j=1}^{n} A_{ij}x_j \Big| \colon \max_{j=1,\ldots,n} |x_j| = 1 \right\} && (1.39)\\
&= \max_{\mathbf{x}} \left\{ \max_{i=1,\ldots,m} \sum_{j=1}^{n} | A_{ij}x_j | \colon \max_{j=1,\ldots,n} |x_j| = 1 \right\} && (1.40)\\
&= \max_{i=1,\ldots,m} \sum_{j=1}^{n} | A_{ij}\mathrm{sign}(A_{ij}) | = \max_{i=1,\ldots,m} \sum_{j=1}^{n} | A_{ij} | && (1.41)
\end{aligned}$$

where $\mathrm{sign}(A_{ij}) = 1$ if $A_{ij} \geq 0$ otherwise $\mathrm{sign}(A_{ij}) = -1$. Note that, besides the last line, (1.40) 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

$$\begin{aligned}
\|\mathbf{A}\|_2 &= \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 && \text{(Definition of } \|\mathbf{A}\|_2) && (1.42)\\
&= \max_{\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij}x_j \right)^2} && \text{(Definition of } \|\mathbf{A}\|_2) && (1.43)\\
&= \max_{\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^{m} \left(\sum_{j=1}^{n} |A_{ij}||x_j| \right)^2} && (\forall j,\ \mathrm{sgn}(x_j) \text{ does not change } \|\mathbf{x}\|_2) && (1.44)
\end{aligned}$$

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

$$\max_{\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^{m} \left(\sum_{j=1}^{n} |A_{ij}||x_j| \right)^2} \leq \max_{\|\mathbf{x}\|_\infty=1} \sqrt{\sum_{i=1}^{m} \left(\sum_{j=1}^{n} |A_{ij}||x_j| \right)^2} \quad (\{\mathbf{x} \mid \|x\|_2 = 1\} \subset \{\mathbf{x} \mid \|x\|_\infty = 1\}) \tag{1.45}$$

$$= \sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|A_{ij}|\right)^2} \qquad \text{(Maximum is attained at } |x_i|= 1\ \forall i) \tag{1.46}$$

$$\leq \sqrt{\sum_{i=1}^{m}\left(\max_{i=1,\ldots,m}\sum_{j=1}^{n}|A_{ij}|\right)^2} \qquad (u_i \leq \max_i |u_i|,\ \forall i) \tag{1.47}$$

$$= \sqrt{\sum_{i=1}^{m}(\|\mathbf{A}\|_\infty)^2} = \sqrt{m}\|\mathbf{A}\|_\infty \qquad \text{(Definition of } \|\mathbf{A}\|_\infty) \tag{1.48}$$

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

$$\max_{\|\mathbf{x}\|_2=1}\sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|A_{ij}||x_j|\right)^2} \geq \sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|A_{ij}|\cdot\frac{1}{\sqrt{n}}\right)^2} \qquad \left(\sum_{j=1}^{n}\left(\frac{1}{\sqrt{n}}\right)^2 = 1\right) \tag{1.49}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|A_{ij}|\right)^2} \qquad \left(\left(\frac{1}{\sqrt{n}}\right)^2 = \frac{1}{n}\right) \tag{1.50}$$

$$\geq \sqrt{\max_{i=1,\ldots,m}\frac{1}{n}\left(\sum_{j=1}^{n}|A_{ij}|\right)^2} \qquad \left(\sum_{i}|u_i|\geq \max_i |u_i|\ \forall i\right) \tag{1.51}$$

$$= \frac{1}{\sqrt{n}}\max_{i=1,\ldots,m}\sum_{j=1}^{n}|A_{ij}| = \frac{1}{\sqrt{n}}\|\mathbf{A}\|_\infty \qquad \text{(Definition of } \|\mathbf{A}\|_\infty) \tag{1.52}$$

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

$$\max_{\|\mathbf{x}\|_2=1}\sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|A_{ij}||x_j|\right)^2} = \sqrt{m}\cdot\max_{\|\mathbf{x}\|_2=1}\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|A_{ij}||x_j|}{m} \qquad \text{(AM-QM inequality)} \tag{1.53}$$

$$= \frac{1}{\sqrt{m}}\cdot\max_{\|\mathbf{x}\|_2=1}\sum_{j=1}^{n}|x_j|\left(\sum_{i=1}^{m}|A_{ij}|\right) \qquad \left(\forall m, n < \infty, \sum_{i=1}^{m}\sum_{j=1}^{n} = \sum_{j=1}^{n}\sum_{i=1}^{m}\right) \tag{1.54}$$

$$= \frac{1}{\sqrt{m}}\cdot\max_{\|\mathbf{x}\|_2=1}\sqrt{\sum_{j=1}^{n}|x_j|^2}\sqrt{\sum_{j=1}^{n}\left(\sum_{i=1}^{m}|A_{ij}|\right)^2} \qquad \text{(Cauchy-Schwarz inequality)} \tag{1.55}$$

$$= \frac{1}{\sqrt{m}}\sqrt{\sum_{j=1}^{n}\left(\sum_{i=1}^{m}|A_{ij}|\right)^2} \qquad (\|\mathbf{A}\|_2= 1) \tag{1.56}$$

$$\geq \frac{1}{\sqrt{m}} \sqrt{\max_{j=1,\dots,n} \left(\sum_{i=1}^{m} |A_{ij}| \right)^2} \qquad \left(\sum_i |u_i| \geq \max_i |u_i| \ \forall i \right) \tag{1.57}$$

$$= \frac{1}{\sqrt{m}} \max_{j=1,\dots,n} \sum_{i=1}^{m} |A_{ij}| = \frac{1}{\sqrt{m}} \|\mathbf{A}\|_1 \qquad (\text{Definition of } \|\mathbf{A}\|_1) \tag{1.58}$$

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.

$$\max_{\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^{m} \left(\sum_{j=1}^{n} |A_{ij}||x_j| \right)^2} \leq \max_{\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2 \sum_{j=1}^{n} x_j^2} \qquad (\text{Cauchy-Schwarz inequality}) \tag{1.59}$$

$$= \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2} \qquad (\|\mathbf{x}\|_2 = 1) \tag{1.60}$$

$$= \sqrt{\sum_{j=1}^{n} \sum_{i=1}^{m} |A_{ij}|^2} \qquad \left(\forall m, n < \infty, \sum_{i=1}^{m} \sum_{j=1}^{n} = \sum_{j=1}^{n} \sum_{i=1}^{m} \right) \tag{1.61}$$

$$\leq \sqrt{\sum_{j=1}^{n} \left(\sum_{i=1}^{m} |A_{ij}| \right)^2} \qquad \left(\forall a_i \geq 0, \sum_{i=1}^{m} a_i^2 \leq \left(\sum_{i=1}^{m} a_i \right)^2 \right) \tag{1.62}$$

$$\leq \sqrt{\sum_{j=1}^{n} \left(\max_{j=1,\dots,n} \sum_{i=1}^{m} |A_{ij}| \right)^2} \qquad (u_i \leq \max_i |u_i|, \ \forall i) \tag{1.63}$$

$$= \sqrt{n} \cdot \max_{j=1,\dots,n} \sum_{i=1}^{m} |A_{ij}| \qquad \left(\sum_{j=1}^{n} c = nc \right) \tag{1.64}$$

$$= \sqrt{n} \|\mathbf{A}\|_1 \qquad (\text{Definition of } \|\mathbf{A}\|_1) \tag{1.65}$$

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. □

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

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T \mathbf{A})} = \sigma_{\max}(\mathbf{A}) \tag{1.66}$$

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,

$$\|\mathbf{A}^T\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}\mathbf{A}^T)} = \sigma_{\max}(\mathbf{A}^T) \tag{1.67}$$

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,

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})} = \sqrt{\lambda_{\max}(\mathbf{A}\mathbf{A}^T)} = \|\mathbf{A}^T\|_2 \tag{1.68}$$

as desired.

Now we consider part (ii).

$$\begin{aligned}
\|\mathbf{A}\|_F^2 =& \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2 && \text{(Definition of Frobenius norm)} && (1.69)\\
=& \mathrm{Tr}(\mathbf{A}^T\mathbf{A}) && \text{(Definition of trace)} && (1.70)\\
=& \sum_{n=1}^{n} \lambda_i(\mathbf{A}^T\mathbf{A}) && && (1.71)
\end{aligned}$$

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

$$\begin{aligned}
p(t) =& \det(t\mathbf{I} - \mathbf{A}^T\mathbf{A}) && && (1.72)\\
=& t^n - \mathrm{Tr}(\mathbf{A}^T\mathbf{A})t^{n-1} + \cdots + (-1)^n \det(\mathbf{A}^T\mathbf{A}) && \text{(Definition of determinant)} && (1.73)
\end{aligned}$$

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

$$p(t) = (t-\lambda_1)(t-\lambda_2)\cdots(t-\lambda_n) \tag{1.74}$$

By comparing coefficients, we have

$$\mathrm{Tr}(\mathbf{A}^T\mathbf{A}) = \sum_{i=1}^{n} \lambda_i(\mathbf{A}^T\mathbf{A}) \tag{1.75}$$

which completes the proof. □

Exercise 1.14

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

$$\max_{\mathbf{x}} \{\mathbf{x}^T\mathbf{A}\mathbf{x} : \|\mathbf{x}\|^2 = 1\} = \lambda_{\max}(\mathbf{A}). \tag{1.76}$$

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 $\mathbf{x} = \mathbf{U}\mathbf{y}$.

$$\max_{\|\mathbf{x}\|_2^2=1} \mathbf{x}^T\mathbf{A}\mathbf{x} = \max_{\|\mathbf{U}\mathbf{y}\|_2^2=1} (\mathbf{U}\mathbf{y})^T\mathbf{A}\mathbf{U}\mathbf{y} \tag{1.77}$$

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

$$
\begin{aligned}
&= \max_{\|\mathbf{y}\|_2^2=1} \mathbf{y}^T\mathbf{U}^T\mathbf{A}\mathbf{U}\mathbf{y} && (\|\mathbf{U}\mathbf{y}\|_2^2 = \|\mathbf{y}\|_2^2) && (1.78)\\
&= \max_{\|\mathbf{y}\|_2^2=1} \mathbf{y}^T\mathbf{D}\mathbf{y} && (\mathbf{U}^T\mathbf{A}\mathbf{U} = \mathbf{D}) && (1.79)\\
&= \max_{\|\mathbf{y}\|_2^2=1} \sum_{i=1}^n d_i y_i^2 \le d_1 \max_{\|\mathbf{y}\|_2^2=1} \sum_{i=1}^n y_i^2 && (d_1 \ge d_2 \ge \cdots \ge d_n) && (1.80)\\
&= d_1 = \lambda_{\max}(\mathbf{A}) && && (1.81)
\end{aligned}
$$

□

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 $\mathrm{cl}(A \cap B) \subseteq \mathrm{cl}(A) \cap \mathrm{cl}(B)$. 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, $\mathrm{cl}(U) = U \cup \mathrm{bd}(U)$. since $A \cap B \subseteq A$, it follows that $\mathrm{cl}(A \cap B) \subseteq \mathrm{cl}(A)$. Likewise, $\mathrm{cl}(A \cap B) \subseteq \mathrm{cl}(B)$. Thus, $\mathrm{cl}(A \cap B) \subseteq \mathrm{cl}(A) \cap \mathrm{cl}(B)$ as desired.

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

Exercise 1.19

Let $A, B \subseteq \mathbb{R}^n$. Prove that $\mathrm{int}(A \cap B) = \mathrm{int}(A) \cap \mathrm{int}(B)$ and that $\mathrm{int}(A) \cup \mathrm{int}(B) \subseteq \mathrm{int}(A \cup B)$. 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. $\mathrm{int}(A \cap B) \subseteq \mathrm{int}(A) \cap \mathrm{int}(B)$ follows from

$$A \cap B \subseteq A \Rightarrow \mathrm{int}(A \cap B) \subseteq \mathrm{int}(A) \tag{1.82}$$
$$A \cap B \subseteq B \Rightarrow \mathrm{int}(A \cap B) \subseteq \mathrm{int}(B) \tag{1.83}$$
$$\Downarrow$$
$$\mathrm{int}(A \cap B) \subseteq \mathrm{int}(A) \cap \mathrm{int}(B). \tag{1.84}$$

[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

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

$$\mathrm{int}(A) \subseteq A, \quad \mathrm{int}(B) \subseteq B \tag{1.85}$$
$$\Downarrow$$
$$\mathrm{int}(A) \cap \mathrm{int}(B) \subseteq A \cap B. \tag{1.86}$$

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

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

$$\mathrm{int}(A) \subseteq A, \quad \mathrm{int}(B) \subseteq B \tag{1.87}$$
$$\Downarrow$$
$$\mathrm{int}(A) \cup \mathrm{int}(B) \subseteq A \cup B. \tag{1.88}$$

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

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

$$f_{\mathrm{RHS}}(t) = t^2 + \sqrt{13}t, \text{ with } 0 \leq t \leq 1. \tag{2.1}$$

Since $f'_{\mathrm{RHS}}(t) = 2t + \sqrt{13} \geq 0$, then $f_{\mathrm{RHS}}(t)$ 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 $f_{\mathrm{RHS}}(1) = 1 + \sqrt{13}$.

The left hand side is

$$f_{\text{LHS}}(t) = t^2 - \sqrt{13}t, \text{ with } 0 \le t \le 1 \tag{2.2}$$

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

$$\mathbf{a}^T\mathbf{x} \le \|\mathbf{a}\|\|\mathbf{x}\| \tag{2.3}$$

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

$$\mathbf{x}^T(\mathbf{A} + \mathbf{B})\mathbf{x} = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{B}\mathbf{x} \ge 0 \tag{2.4}$$

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

$$\mathbf{z}^T \begin{pmatrix} \mathbf{A} & \mathbf{0}_{n\times m} \\ \mathbf{0}_{m\times n} & \mathbf{B} \end{pmatrix} \mathbf{z} = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{y}^T\mathbf{B}\mathbf{y} \geq 0 \tag{2.5}$$

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 λ. 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

$$\mathbf{x}^T\mathbf{B}\mathbf{B}^T\mathbf{x} = (\mathbf{B}^T\mathbf{x})^T\mathbf{B}^T\mathbf{x} = \|\mathbf{B}^T\mathbf{x}\|_2^2 \geq 0. \tag{2.6}$$

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 (2.6) 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

$$A_{ij} = x_i x_j, \quad i, j = 1, 2, \ldots, n. \tag{2.7}$$

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

$$\mathbf{y}^T\mathbf{A}\mathbf{y} = \mathbf{y}^T\mathbf{x}\mathbf{x}^T\mathbf{y} = (\mathbf{x}^T\mathbf{y})^2 \geq 0 \tag{2.8}$$

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.

□

Exercise 2.8

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

$$\|\mathbf{x}\|_{\mathbf{Q}} = \sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x}} \tag{2.9}$$

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.

$$\|\mathbf{x}+\mathbf{y}\|_{\mathbf{Q}} \leq \|\mathbf{x}\|_{\mathbf{Q}} + \|\mathbf{y}\|_{\mathbf{Q}} \tag{2.10}$$
$$\Updownarrow \tag{2.11}$$
$$\sqrt{(\mathbf{x}+\mathbf{y})^T\mathbf{Q}(\mathbf{x}+\mathbf{y})} \leq \sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x}} + \sqrt{\mathbf{y}^T\mathbf{Q}\mathbf{y}} \tag{2.12}$$
$$\Updownarrow \tag{2.13}$$
$$(\mathbf{x}+\mathbf{y})^T\mathbf{Q}(\mathbf{x}+\mathbf{y}) \leq \mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{y}^T\mathbf{Q}\mathbf{y} + 2\sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x}\mathbf{y}^T\mathbf{Q}\mathbf{y}} \tag{2.14}$$
$$\Updownarrow \tag{2.15}$$
$$\mathbf{x}^T\mathbf{Q}\mathbf{y} + \mathbf{y}^T\mathbf{Q}\mathbf{x} \leq 2\sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x}\mathbf{y}^T\mathbf{Q}\mathbf{y}} \tag{2.16}$$

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 $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_n)$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{A}$. Let $\mathbf{U}\mathbf{x} = \tilde{\mathbf{x}}$ and $\mathbf{U}\mathbf{y} = \tilde{\mathbf{y}}$, then we have

$$\mathbf{x}^T\mathbf{U}^T\mathbf{D}\mathbf{U}\mathbf{y} + \mathbf{y}^T\mathbf{U}^T\mathbf{D}\mathbf{U}\mathbf{x} \le 2\sqrt{\mathbf{x}^T\mathbf{U}^T\mathbf{D}\mathbf{U}\mathbf{x}\mathbf{y}^T\mathbf{U}^T\mathbf{D}\mathbf{U}\mathbf{y}} \tag{2.17}$$

$$\Updownarrow \tag{2.18}$$

$$\sum_i^n d_i x_i y_i + \sum_i^n d_i x_i y_i \le 2\sqrt{(\sqrt{d_i}x_i)^2}\sqrt{(\sqrt{d_i}y_i)^2} \tag{2.19}$$

$$\Updownarrow \tag{2.20}$$

$$\sum_i^n (\sqrt{d_i}x_i)(\sqrt{d_i}y_i) \le \sqrt{(\sqrt{d_i}x_i)^2}\sqrt{(\sqrt{d_i}y_i)^2} \tag{2.21}$$

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 $i \neq j$

$$A_{ii}A_{jj} \ge A_{ij}^2 \tag{2.22}$$

(ii) Show that if for some $i \in \{1, 2, \ldots, n\}$ $A_{ii} = 0$, then the ith 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

$$(e_i x + e_j)^T\mathbf{A}(e_i x + e_j) \ge 0 \tag{2.23}$$

$$A_{ii}x^2 + 2A_{ij}x + A_{jj} \ge 0 \tag{2.24}$$

where $\mathbf{e}_i$ is a vector with all zeros except the ith 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.

$$4A_{ij}^2 - 4A_{ii}A_{jj} \le 0 \Rightarrow A_{ii}A_{jj} \ge A_{ij}^2. \tag{2.25}$$

(ii) With the result in the first part, if for some i, $A_{ii} = 0$, then for any $j \neq i$, we have $0 \times A_{jj} \ge A_{ij}^2$ which implies $A_{ij} = 0$. This shows that the ith 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

$$A_{ij} = \begin{cases} \alpha, & i = j, \\ 1, & i \neq j. \end{cases} \tag{2.26}$$

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

[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

$$\mathbf{x}^T \mathbf{A}^\alpha \mathbf{x} \geq 0 \Rightarrow 2\alpha - 2 \geq 0 \Rightarrow \alpha \geq 1. \tag{2.27}$$

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

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

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

Exercise 2.11

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

$$A_{ij} = \begin{cases} d_i - d_i^2, & i = j, \\ -d_i d_j, & i \neq j, \end{cases} \tag{2.28}$$

is positive semidefinite.

228

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

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \mathbf{x}^T (\mathrm{diag}(\mathbf{d}) - \mathbf{d}\mathbf{d}^T)\mathbf{x} = \mathbf{x}^T \mathrm{diag}(\mathbf{d})\mathbf{x} - \mathbf{x}^T \mathbf{d}\mathbf{d}^T \mathbf{x} = \sum_i^n (d_i - d_i^2) x_i^2 \geq 0 \tag{2.29}$$

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 2×2 matrix $\mathbf{A}$ is negative semidefinite if and only if $\mathrm{Tr}(\mathbf{A}) \leq 0$ and $\det(\mathbf{A}) \leq 0$.

230

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

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \tag{2.30}$$

Furthermore, the characteristic equation is given by

$$\det(\lambda \mathbf{I} - \mathbf{A}) = 0 \tag{2.31}$$

$$\begin{pmatrix} \lambda - a_{11} & -a_{12} \\ -a_{21} & \lambda - a_{22} \end{pmatrix} = 0 \tag{2.32}$$

$$\lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} = 0 \tag{2.33}$$

where λ denotes the two roots (λ_1 and λ_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 λ_1 and λ_2 are nonpositive. From the last equation above, we get

$$\begin{cases} \lambda_1 + \lambda_2 = a_{11} + a_{22} = \mathrm{Tr}(\mathbf{A}) \\ \lambda_1 \lambda_2 = a_{11}a_{22} - a_{21}a_{21} = \det(\mathbf{A}) \end{cases} \tag{2.34}$$

which implies

$$\begin{cases} \mathrm{Tr}(\mathbf{A}) \leq 0 \\ \det(\mathbf{A}) \geq 0 \end{cases} \iff \lambda_1, \lambda_2 \leq 0 \tag{2.35}$$

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

$$g(\mathbf{y}, \mathbf{z}) := \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix}^T \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \mathbf{y}^T\mathbf{A}\mathbf{y} + \mathbf{z}^T\mathbf{B}^T\mathbf{y} + \mathbf{y}^T\mathbf{B}\mathbf{z} + \mathbf{z}^T\mathbf{C}\mathbf{z} \geq 0, \quad \forall \mathbf{y}, \mathbf{z}. \tag{2.36}$$

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

$$0 \leq f(\mathbf{z}) := \min_{\mathbf{y}} g(\mathbf{y}, \mathbf{z}). \tag{2.37}$$

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

$$\nabla_{\mathbf{y}} g(\mathbf{y}, \mathbf{z}) = 2\mathbf{A}\mathbf{y} + 2\mathbf{B}\mathbf{z} = 0 \Longleftrightarrow \mathbf{y} = -\mathbf{A}^{-1}\mathbf{B}\mathbf{z}. \tag{2.38}$$

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

$$f(\mathbf{z}) = g(-\mathbf{A}^{-1}\mathbf{B}\mathbf{z}, \mathbf{z}) = \mathbf{z}^T(\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})\mathbf{z} \tag{2.39}$$

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.

$\square$

Exercise 2.16

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

$$\lim_{|x_1| \to \infty} f(x_1, \alpha x_1) = \lim_{|x_2| \to \infty} f(\alpha x_2, x_2) = \infty. \tag{2.40}$$

Solution: Consider the following function

$$f(x_1, x_2) = \frac{1 + x_1 x_2}{|x_1| + |x_2|} \tag{2.41}$$

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

$$\lim_{|x_1| \to \infty} f(x_1, \alpha x_1) = \lim_{|x_1| \to \infty} \frac{1 + x_1^2}{(1 + |\alpha|)|x_1|} = \infty. \tag{2.42}$$

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

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

(v) $f(x_1, x_2) = (x_1 - 2x_2)^4 + 64x_1 x_2$.

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

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

Solution:

(i) First,

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 16x_1(4x_1^2 - x_2) \\ -2(4x_1^2 - x_2) \end{pmatrix} \tag{2.43}$$

Hence, the stationary points are those satisfying

$$16x_1(4x_1^2 - x_2) = 0 \tag{2.44}$$

$$-2(4x_1^2 - x_2) = 0 \tag{2.45}$$

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

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 4x_1(x_1^2 - 1) \\ 2(x_2 + x_3) \\ 2(x_2 + 2x_3) \end{pmatrix}. \tag{2.46}$$

Therefore, the stationary points are those satisfying

$$4x_1(x_1^2 - 1) = 0 \tag{2.47}$$

$$2(x_2 + x_3) = 0 \tag{2.48}$$

$$2(x_2 + 2x_3) = 0. \tag{2.49}$$

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

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 4(3x_1^2 - 1) & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & 4 \end{pmatrix}. \tag{2.50}$$

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,

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 6x_1x_2 \\ 6x_2^2 - 12x_2 + 3x_1^2 \end{pmatrix} \tag{2.51}$$

Then the stationary points are those satisfying

$$6x_1x_2 = 0 \tag{2.52}$$

$$6x_2^2 - 12x_2 + 3x_1^2 = 0. \tag{2.53}$$

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

$$\nabla^2 f(\mathbf{x}) = 6 \begin{pmatrix} x_2 & x_1 \\ x_1 & 2x_2 - 2. \end{pmatrix} \tag{2.54}$$

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,

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 4x_1^3 + 4x_1x_2 - 8x_1 - 8 \\ 2x_1^2 + 2x_2 - 8 \end{pmatrix} \tag{2.55}$$

from which we know the stationary points are those that satisfy

$$x_1(x_1^2 + x_2) - 2x_1 - 2 = 0 \tag{2.56}$$

$$x_1^2 + x_2 - 4 = 0 \tag{2.57}$$

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

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 12x_1^2 + 4x_2 - 8 & 4x_1 \\ 4x_1 & 2 \end{pmatrix}. \tag{2.58}$$

Then we have

$$\nabla^2 f(1,3) = \begin{pmatrix} 16 & 4 \\ 4 & 2 \end{pmatrix} \succeq \mathbf{0} \tag{2.59}$$

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.

(v) First,

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 4(x_1 - 2x_2)^3 + 64x_2 \\ -8(x_1 - 2x_2)^3 + 64x_1 \end{pmatrix} = 0 \tag{2.60}$$

which has three solutions: $x_1 = x_2 = 0$, $x_1 = 1, x2 = -\frac{1}{2}$, and $x_1 = -1, x2 = \frac{1}{2}$. Then,

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 12(x_1 - 2x_2)^2 & -24(x_1 - 2x_2)^2 + 64 \\ -24(x_1 - 2x_2)^2 + 64 & 16(x_1 - 2x_2)^2 \end{pmatrix} \tag{2.61}$$

from which we get

$$\nabla^2 f(0,0) = \begin{pmatrix} 0 & 64 \\ 64 & 0 \end{pmatrix} \tag{2.62}$$

which is indefinite. Thus, $x_1 = x_2 = 0$ is a saddle point. It is easy to see

$$\nabla^2 f(1, -\frac{1}{2}) = \nabla^2 f(-1, \frac{1}{2}) = \begin{pmatrix} 48 & 16 \\ 16 & 64 \end{pmatrix} \tag{2.63}$$

is positive definite. When $\|\mathbf{x}\| \to \infty$, $f(\mathbf{x}) \to \infty$. Hence, f is coercive. Thus, f has a global minimum. Finally, $x_1 = 1, x2 = -\frac{1}{2}$ and $x_1 = -1, x2 = \frac{1}{2}$ are nonstrict global minimum points.

(vi) First,

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 2(2x_1 - x_2 + 1) \\ 6x_2 - 2x_1 - 3 \end{pmatrix} = 0 \tag{2.64}$$

which gives $x_1 = -3/10, x_2 = 4/10$. Then,

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 4 & -2 \\ -2 & 6 \end{pmatrix} \tag{2.65}$$

which is positive definite. Equivalently, $f(\mathbf{x}) = (x_1 - x_2)^2 + (x_1 - 1)^2 + 2(x_2 - 3/2)^2 - 11/2$, which is coercive. Hence, $x_1 = -3/10, x_2 = 4/10$ is a strict global minimum point.

(vii) First,

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 2x_1 + 4x_2 + 1 \\ 4x_1 + 2x_2 - 1 \end{pmatrix}. \tag{2.66}$$

Setting it to 0 gives $x_1 = -1/2, x_2 = 1/2$. The Hessian is given by

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix} \tag{2.67}$$

whose eigenvalues have a sum of 4 and a product of -12, which implies that $\nabla^2 f(\mathbf{x})$ has one positive eigenvalue and one negative eigenvalue. Hence, the Hessian is indefinite. Thus, $x_1 = -1/2, x_2 = 1/2$ is a saddle point.

□

Exercise 2.18

Let f be twice continuously differentiable function over $\mathbb{R}^n$. Suppose that $\nabla^2 f(\mathbf{x}) \succ \mathbf{0}$ for any $\mathbf{x} \in \mathbb{R}^n$. Prove that a stationary point of f is necessarily a strict global minimum point.

Proof. According to the linear approximation theorem, i.e. Theorem 1.24 in the textbook, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, there exists $\xi \in [\mathbf{x}, \mathbf{y}]$ such that

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) + \frac{1}{2} (\mathbf{y} - \mathbf{x})^T \nabla^2 f(\xi)(\mathbf{y} - \mathbf{x}). \tag{2.68}$$

Assume $\mathbf{x}^*$ is a strict global minimum point, we have

$$\nabla f(\mathbf{x}^*)^T (\mathbf{y} - \mathbf{x}^*) + \frac{1}{2} (\mathbf{y} - \mathbf{x}^*)^T \nabla^2 f(\xi)(\mathbf{y} - \mathbf{x}^*) = f(\mathbf{y}) - f(\mathbf{x}^*) > 0 \tag{2.69}$$

which implies

$$\nabla f(\mathbf{x}^*)^T (\mathbf{y} - \mathbf{x}^*) > -\frac{1}{2} (\mathbf{y} - \mathbf{x}^*)^T \nabla^2 f(\xi)(\mathbf{y} - \mathbf{x}^*) \tag{2.70}$$

where for any $\mathbf{y}$, the right hand side is always less than 0 since $\nabla^2 f(\xi) \succ \mathbf{0}$. This implies $\nabla f(\mathbf{x}^*) = 0$, otherwise the left hand side will not hold for arbitrary $\mathbf{y}$. Specifically, let $\mathbf{y} = t \nabla f(\mathbf{x}^*)/\|\nabla f(\mathbf{x}^*)\|^2$ where $t > 0$, then we substitute it into (2.70), which gives

$$\frac{t \nabla f(\mathbf{x}^*)^T \nabla^2 f(\xi) \nabla f(\mathbf{x}^*)}{2\|\nabla f(\mathbf{x}^*)\|^4} \geq 1 \tag{2.71}$$

which will not hold when t is small enough. This completes the proof. □

Exercise 2.19

Let $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + 2\mathbf{b}^T \mathbf{x} + c$, where $\mathbf{A}$ is symmetric, $\mathbf{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Suppose that $\mathbf{A} \succeq \mathbf{0}$. Show that f is bounded below over $\mathbb{R}^n$ if and only if $\mathbf{b} \in \text{Range}(\mathbf{A}) = \{\mathbf{A}\mathbf{y} : \mathbf{y} \in \mathbb{R}^n\}$.

Proof. One may use Lemma 2.41(b) in the textbook to show the claim. Lemma 2.41(b) says that given $\mathbf{A} \succeq \mathbf{0}$, $\mathbf{y}$ is a global minimum point if and only if $\mathbf{Ay} = -b$. However, a polynomial that is bounded below does not necessarily have a global minimum. For example, the function $f(x, y) = (1-xy)^2 + x^2$ is bounded below by 0, but 0 can not be attained although any small $\epsilon > 0$ can be attained[9].

Now we show the sufficiency. Given $\mathbf{A} \succeq \mathbf{0}$, if $\mathbf{Ax}^* = -\mathbf{b}$, equivalently, $\mathbf{b} \in \text{Range}(\mathbf{A})$, by Lemma 2.41(b), f has a global minimum at x^*, which means f is bounded below by $f(\mathbf{x}^*)$.

Alternatively, by the linear approximation theorem, i.e. Theorem 1.24 in the textbook, at the stationary point x^* satisfying $\mathbf{Ax}^* = -b$, there exists $\mathbf{z} \in [\mathbf{x}^*, \mathbf{x}]$ such that

$$f(\mathbf{x}) - f(\mathbf{x}^*) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \nabla^2 f(\mathbf{z})(\mathbf{x} - \mathbf{x}^*) = (\mathbf{x} - \mathbf{x}^*)^T \mathbf{A}(\mathbf{x} - \mathbf{x}^*) \geq 0 \tag{2.72}$$

where the inequality follows from $\mathbf{A} \succeq \mathbf{0}$.

For the necessity, we prove by contradiction. We need the result that $\text{Null}(\mathbf{A}^T)^\perp = \text{Range}(\mathbf{A})$[10]. Assume that $f(\mathbf{x}) = \mathbf{x}^T\mathbf{Ax} + 2\mathbf{b}^T\mathbf{x} + c$ is lower bounded by a constant d. If $\mathbf{b} \notin \text{Range}(\mathbf{A})$, that is to say $\mathbf{b} \notin \text{Null}(\mathbf{A}^T)^\perp$, then $\mathbf{b}^T\mathbf{x} \neq 0$ for any $\mathbf{A}^T\mathbf{x} = \mathbf{0} = \mathbf{Ax}$ since $\mathbf{A}$ is symmetric. In this case, $f(\mathbf{x}) = 2\mathbf{b}^T\mathbf{x} + c$. Let $t = \lambda \cdot \text{sign}(\mathbf{b}^T\mathbf{x})$ where $\lambda > 0$ and $\text{sign}(u) = 1$ if $u > 0$, $\text{sign}(u) = -1$ if $u < 0$, otherwise $\text{sign}(u) = 0$, then

$$f(t\mathbf{x}) = 2\lambda \text{sign}(\mathbf{b}^T\mathbf{x})\mathbf{b}^T\mathbf{x} \to -\infty, \tag{2.73}$$

as $\lambda \to \infty$. This contradicts the assumption. Therefore, $\mathbf{b} \in \text{Range}(\mathbf{A})$. This completes the proof. □

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

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

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

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

$$\begin{gathered}
\text{(RLS) has a unique solution} \iff \mathbf{A}^T\mathbf{A} + \lambda\mathbf{L}^T\mathbf{L} \succ \mathbf{0} \\
\Updownarrow \\
\mathbf{x}^T(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{L}^T\mathbf{L})\mathbf{x} > 0, \forall \mathbf{x} \neq \mathbf{0} \iff \|\mathbf{Ax}\|^2 + \lambda\|\mathbf{Lx}\|^2 > 0, \forall \mathbf{x} \neq \mathbf{0} \\
\Updownarrow \\
\text{There exists no nonzero } \mathbf{x} \text{ such that } \mathbf{Ax} = \mathbf{0} \text{ and } \mathbf{Lx} = \mathbf{0} \text{ hold simultaneously.} \\
\Updownarrow \\
\text{Null}(\mathbf{A}) \cap \text{Null}(\mathbf{L}) = \{\mathbf{0}\}.
\end{gathered}$$

This completes the proof. □

[9] https://math.stackexchange.com/questions/3820/does-a-polynomial-thats-bounded-below-have-a-global-minimum

[10] https://math.stackexchange.com/questions/318136/the-range-of-t-is-the-orthogonal-complement-of-kert

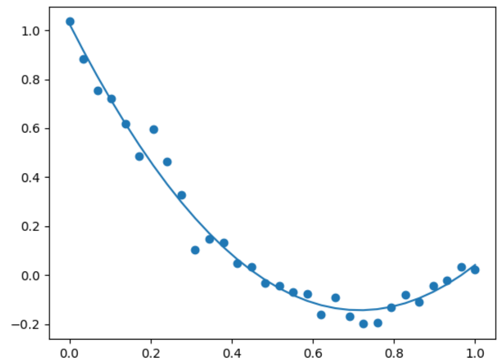

Figure 1: 30 points and their best quadratic least squares fit.

Exercise 3.2

Generate 30 points $(x_i, y_i), i = 1, 2, \ldots, 30$, by the Python code

```
import numpy as np
np.random.seed(2023)
x = np.linspace(0, 1, 30)
y = 2 * x**2 - 3*x + 1 + 0.05 * np.random.randn(len(x))
```

Find the quadratic function $y = ax^2 + bx + c$ that best fits the points in the least squares sense. Indicate what are the parameters a, b, c found by the least squares solution, and plot the points along with the derived quadratic function. The resulting plot should look like the one in Figure 1.

Solution: The data matrices $\mathbf{A}$ and $\mathbf{b}$ can be represented as follows.

$$\mathbf{A} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \cdots & & \\ x_{30}^2 & x_{30} & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{30} \end{bmatrix}. \tag{3.1}$$

Then we find the least squares solution via

$$\mathbf{s} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \tag{3.2}$$

which gives $a = 2.285, b = -3.270, c = 1.027$ by running its corresponding Python code.

```
import matplotlib.pyplot as plt
A = np.array([x**2, x, np.ones(len(x))]).transpose()
b = np.array(y).reshape(-1, 1)
sol = np.linalg.inv(A.transpose() @ A) @ A.transpose() @ b
plt.scatter(x, y)
plt.plot(x, sol[0]*x**2 + sol[1]*x + sol[2])
plt.show()
```

□

Exercise 3.3

Write a Python function *circle_fit* whose input is an $n \times m$ matrix $\mathbf{A}$; the columns of $\mathbf{A}$ are the m vectors in $\mathbb{R}^n$ to which a circle should be fitted. The call to the function will be of the form

```
x, r = circle_fit(A)
```

The output is the optimal solution of

$$\min_{\mathbf{x}\in\mathbb{R}^n, r\in\mathbb{R}_+} \sum_{i=1}^{m} (\|\mathbf{x} - a_i\|^2 - r^2)^2. \tag{3.3}$$

Use the code in order to find the best circle fit in the sense of (3.3) of the 5 points

$$\mathbf{a}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathbf{a}_2 = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}, \quad \mathbf{a}_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{a}_4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{a}_5 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{3.4}$$

Solution: According to Lemma 3.5 in the textbook, the solution is given by

$$\mathbf{x} = (\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})^{-1} \tilde{\mathbf{A}}^T \mathbf{b}, \tag{3.5}$$

where

$$\tilde{\mathbf{A}} = \begin{pmatrix} 2\mathbf{a}_1^T & -1 \\ 2\mathbf{a}_2^T & -1 \\ \vdots & \vdots \\ 2\mathbf{a}_5^T & -1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \|\mathbf{a}_1\|^2 \\ \|\mathbf{a}_2\|^2 \\ \vdots \\ \|\mathbf{a}_5\|^2 \end{pmatrix}. \tag{3.6}$$

With this, the code can be

```python
import numpy as np

def circle_fit(A):
    b = np.sum(A.T**2, axis=1, keepdims=True)
    A_tilde = np.concatenate([2 * A.T, -np.ones((A.shape[1], 1))], axis=1)
    sol = np.linalg.inv(A_tilde.T @ A_tilde) @ A_tilde.T @ b
    return sol[:2], np.sqrt(np.sum(sol[:2]**2) - sol[-1])

A = np.array([[0, 0], [0.5, 0], [1, 0], [1, 1], [0, 1]]).T
x, r = circle_fit(A)
```

which gives $x = [0.5, 0.54], r = 0.68$. □

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,

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

Exercise 4.1

Let $f \in C_L^{1,1}(\mathbb{R}^n)$ and let $\{\mathbf{x}^k\}_{k\geq 0}$ 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 $k \geq 0$, 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 $r > 0$ such that

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

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

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

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

Exercise 4.2

Consider the minimization problem

$$\min\{\mathbf{x}^T\mathbf{Q}\mathbf{x} : \mathbf{x} \in \mathbb{R}^2\} \tag{4.1}$$

where $\mathbf{Q}$ is a positive definite 2×2 matrix. Suppose we use the diagonal scaling matrix

$$\mathbf{D} = \begin{pmatrix} Q_{11}^{-1} & 0 \\ 0 & Q_{22}^{-1} \end{pmatrix}. \tag{4.2}$$

Show that the above scaling matrix improves the condition number of $\mathbf{Q}$ in the sense that

$$\chi(\mathbf{D}^{1/2}\mathbf{Q}\mathbf{D}^{1/2}) \leq \chi(\mathbf{Q}). \tag{4.3}$$

Proof. After simple algebra, we get

$$\mathbf{D}^{1/2}\mathbf{Q}\mathbf{D}^{1/2} = \begin{pmatrix} 1 & Q_{11}^{-1/2}Q_{12}Q_{22}^{-1/2} \\ Q_{11}^{-1/2}Q_{21}Q_{22}^{-1/2} & 1 \end{pmatrix}. \tag{4.4}$$

Denote the eigenvalues of $\mathbf{Q}$ and $\mathbf{D}^{1/2}Q\mathbf{D}^{1/2}$ by λ_1, λ_2 and λ_1', λ_2', respectively. Then we have

$$\lambda_1 + \lambda_2 = Q_{11} + Q_{22}, \lambda_1\lambda_2 = Q_{11}Q_{22} - Q_{12}Q_{21} \text{ and } \lambda_1' + \lambda_2' = 2, \lambda_1'\lambda_2' = 1 - \frac{Q_{12}Q_{21}}{Q_{11}Q_{22}}. \tag{4.5}$$

Moreover, we denote the condition numbers of these two matrices by χ and χ'. Then we have

$$\chi + \frac{1}{\chi} + 2 = \frac{(Q_{11} + Q_{22})^2}{Q_{11}Q_{22} - Q_{12}Q_{21}}, \quad \chi' + \frac{1}{\chi'} + 2 = \frac{2^2}{1 - \frac{Q_{12}Q_{21}}{Q_{11}Q_{22}}}. \tag{4.6}$$

Furthermore, we get

$$\frac{\chi' + \frac{1}{\chi'} + 2}{\chi + \frac{1}{\chi} + 2} = \frac{4Q_{11}Q_{22}}{(Q_{11}+Q_{22})^2} \leq 1 \tag{4.7}$$

which implies that

$$\chi' + \frac{1}{\chi'} \leq \chi + \frac{1}{\chi} \tag{4.8}$$

Since any condition number is greater than or equal to 1 and the function $x + \frac{1}{x}$ is monotonically increasing on $[1, \infty)$, then $\chi' \leq \chi$ as desired. $\square$

Exercise 4.3

Consider the quadratic minimization problem

$$\min\{\mathbf{x}^T\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathbb{R}^5\}, \tag{4.9}$$

where $\mathbf{A}$ is the 5×5 Hilbert matrix defined by

$$A_{i,j} = \frac{1}{i+j-1}, \quad i,j = 1,2,3,4,5. \tag{4.10}$$

The matrix can he constructed via the Scipy command *A=scipy.linalg.hilbert(5)*. Run the following methods and compare the number of iterations required by each of the methods when the initial vector is $\mathbf{x}_0 = (1,2,3,4,5)^T$ to obtain a solution $\mathbf{x}$ with $\|\nabla f(\mathbf{x})\| \leq 10^{-4}$:

- gradient method with backtracking stepsize rule and parameters $\alpha = 0.5, \beta = 0.5, s = 1$;
- gradient method with backtracking stepsize rule and parameters $\alpha = 0.1, \beta = 0.5, s = 1$;
- gradient method with exact line search;
- diagonally scaled gradient method with diagonal elements $D_{ii} = \frac{1}{A_{ii}}, i = 1,2,3,4,5$ and exact line search;
- diagonally scaled gradient method with diagonal elements $D_{ii} = \frac{1}{A_{ii}}, i = 1,2,3,4,5$ and backtracking line search with parameters $\alpha = 0.1, \beta = 0.5, s = 1$.

Solution: With the following code, the numbers of iterations are 5801, 3977, 1271, 235, and 263, respectively. From these results, in terms of the number of the iterations out of the outer loop, we can see that a smaller alpha may lead to faster convergence and diagonal scaling improves convergence significantly.

```python
import numpy as np
import scipy

def f(A, x):
    return x.T @ A @ x

def exact_line_search(A, grad, d):
    # refer to the closed form, i.e. eq. (4.3) in the textbook
    t = d.T @ grad / (2 * (d.T @ A @ d))
    return t

def backtracking(A, grad, d, alpha, beta, s, D=None):
```

```python
    i, t = 0, s
    while f(A, x) - f(A, x - t*d) < -alpha * t * grad.T @ d:
        t = s * beta**i
        i += 1
    return t

alpha = 0.1
beta = 0.5
s = 1.0
A = scipy.linalg.hilbert(5)
D = np.diag(1.0 / np.diag(A)) # None means without diagonal scaling
x = np.array([1,2,3,4,5]).reshape(-1,1).astype('float32')
for iter in range(10000):
    grad = 2.0 * A @ x
    grad_norm = scipy.linalg.norm(grad)
    if grad_norm <= 1e-4:
        print("exiting", iter, grad_norm)
        break
    if D is not None:
        d = D @ grad
    else:
        d = grad
    t = backtracking(A, grad, d, alpha, beta, s, D) # backtracking
    # t = exact_line_search(A, grad, d) # exact line search
    x -= t * d
```

□

Exercise 4.4

Consider the Fermat-Weber problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) \equiv \sum_{i=1}^{m} \omega_i \|\mathbf{x} - \mathbf{a}_i\| \right\}, \tag{4.11}$$

where $\omega_1, \ldots, \omega_m > 0$ and $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$ are m different points. Let

$$p \in \operatorname*{argmin}_{i=1,2,\ldots,m} f(\mathbf{a}_i). \tag{4.12}$$

Suppose that

$$\left\| \sum_{i\neq p} \omega_i \frac{\mathbf{a}_p - \mathbf{a}_i}{\|\mathbf{a}_p - \mathbf{a}_i\|} \right\| > \omega_p. \tag{4.13}$$

(i) Show that there exists a direction $\mathbf{d} \in \mathbb{R}^n$ such that $f'(\mathbf{a}_p; \mathbf{d}) < 0$.

(ii) Show that there exists $\mathbf{x}_0 \in \mathbb{R}^n$ satisfying $f(\mathbf{x}_0) < \min\{f(\mathbf{a}_1), f(\mathbf{a}_2), \ldots, f(\mathbf{a}_m)\}$. Explain how to compute such a vector.

Proof.

(i) First, we need the following result. For the function $f(\mathbf{x}) = \|\mathbf{x}\|$, its (sub)gradient is given by

$$\partial f(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}}{\|\mathbf{x}\|}, & \mathbf{x} \neq \mathbf{0}, \\ \mathbf{v}, & \mathbf{x} = \mathbf{0}, \end{cases} \tag{4.14}$$

where $\mathbf{v} = \{\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\| \leq 1\}$. With this in hand, we have

$$\partial f(\mathbf{a}_p) = \sum_{i \neq p} \omega_i \frac{\mathbf{a}_p - \mathbf{a}_i}{\|\mathbf{a}_p - \mathbf{a}_i\|} + \omega_p \mathbf{v} \tag{4.15}$$

Let $\mathbf{s}$ be the first term on the right hand side. Then given a direction $\mathbf{d} = -\mathbf{s}/\|\mathbf{s}\|$, its directional derivative is given by

$$\begin{aligned}
\mathbf{d}^T \partial f(\mathbf{x}) &= -\frac{\mathbf{s}^T}{\|\mathbf{s}\|}(\mathbf{s} + \omega_p \mathbf{v}) && (4.16)\\
&= -\|\mathbf{s}\| - \frac{\omega_p}{\|\mathbf{s}\|}\mathbf{s}^T \mathbf{v} && (4.17)\\
&< -\omega_p + \omega_p(-\frac{\mathbf{s}}{\|\mathbf{s}\|})^T \mathbf{v} && (4.18)\\
&\leq -\omega_p + \omega_p \|\mathbf{v}\| && (4.19)\\
&\leq -\omega_p + \omega_p = 0 && (4.20)
\end{aligned}$$

where the first inequality follows from $\|\mathbf{s}\| > \omega_p$, the second inequality follows from Cauchy-Schwarz inequality, and the third inequality follows from $\|\mathbf{v}\| \leq 1$. Thus, $\mathbf{d} = -\mathbf{s}/\|\mathbf{s}\|$ is a descent direction.

(ii) By Lemma 4.3, given $\alpha \in (0, 1)$ and $\mathbf{d} = -\mathbf{s}/\|\mathbf{s}\|$, there exists $\epsilon > 0$ such that

$$f(\mathbf{a}_p + t\mathbf{d}) \leq f(\mathbf{a}_p) + \alpha t \mathbf{d}^T \partial f(\mathbf{a}_p) < f(\mathbf{a}_p) \tag{4.21}$$

for all $t \in [0, \epsilon]$. The last inequality follows from the proved result that $\mathbf{d}$ is a descent direction. Thus, there exists $\mathbf{x}_0$ such that $f(\mathbf{x}_0) < \min\{f(\mathbf{a}_1), f(\mathbf{a}_2), \ldots, f(\mathbf{a}_m)\}$. To compute such a vector, we can solve the following minimization problem for the exact search method

$$\min_t f(\mathbf{a}_p + t\mathbf{d}) \equiv \sum_{i \neq p}^{m} \omega_i \|\mathbf{a}_p - \mathbf{a}_i + t\mathbf{d}\| = \sum_{i \neq p}^{m} \omega_i \|\mathbf{a}_p - \mathbf{a}_i + t\mathbf{d}\| + \omega_p t \tag{4.22}$$

Setting the derivative $f'_t(\mathbf{a}_p + t\mathbf{d})$ to 0 yields

$$t = -\frac{\omega_p + \sum_{i \neq p} \frac{\langle \mathbf{a}_p - \mathbf{a}_i, \mathbf{d} \rangle \omega_i}{\|\mathbf{a}_p - \mathbf{a}_i + t\mathbf{d}\|}}{\sum_{i \neq p} \frac{\mathbf{d}^T \mathbf{d} \omega_i}{\|\mathbf{a}_p - \mathbf{a}_i + t\mathbf{d}\|}}. \tag{4.23}$$

We can reformulate the optimality condition as $t = T(t)$, where T is the operator

$$T(t) = -\frac{\omega_p + \sum_{i \neq p} \frac{\langle \mathbf{a}_p - \mathbf{a}_i, \mathbf{d} \rangle \omega_i}{\|\mathbf{a}_p - \mathbf{a}_i + t\mathbf{d}\|}}{\sum_{i \neq p} \frac{\mathbf{d}^T \mathbf{d} \omega_i}{\|\mathbf{a}_p - \mathbf{a}_i + t\mathbf{d}\|}}. \tag{4.24}$$

Therefore, we can find an appropriate t such that $f(\mathbf{a}_p + t\mathbf{d}) < f(\mathbf{a}_p)$ by the iterations.

$$t_{k+1} = T(t_k). \tag{4.25}$$

We may need to show this fixed point operator is convergent. We leave it as future work. If we use backtracking, we need to make an initial guess $s > 0$, with $\alpha \in (0,1)$ and $\beta \in (0,1)$, and then increase i until the following

$$f(\mathbf{a}_p + s\beta^{i_k}\mathbf{d}) \leq f(\mathbf{a}_p) + \alpha s \beta^{i_k}\mathbf{d}^T \partial f(\mathbf{a}_p) \tag{4.26}$$

is satisfied. Finally, $t = s\beta^{i_k}$.

□

Exercise 4.5

In the "source localization problem" we are given m locations of sensors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m \in \mathbb{R}^m$, and approximate distances between the sensors and an unknown "source" located at $\mathbf{x} \in \mathbb{R}^m$:

$$d_i \approx \|\mathbf{x} - \mathbf{a}_i\|. \tag{4.27}$$

The problem is to find and estimate $\mathbf{x}$ given the locations $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m \in \mathbb{R}^m$ and the approximate distances $d_1, d_2, \ldots, d_m \in \mathbb{R}^m$. A natural formulation as an optimization problem is to consider the nonlinear least squares problem

$$\min \left\{ f(\mathbf{x}) \equiv \sum_{i=1}^{m} (\|\mathbf{x} - \mathbf{a}_i\| - d_i)^2 \right\}. \tag{4.28}$$

We will denote the set of sensors by $\mathcal{A} \equiv \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}$.

(i) Show that the optimality condition $\nabla f(\mathbf{x}) = \mathbf{0}(\mathbf{x} \notin \mathcal{A})$ is the same as

$$\mathbf{x} = \frac{1}{m}\left\{ \sum_{i=1}^{m} \mathbf{a}_i + \sum_{i=1}^{m} d_i \frac{\mathbf{x} - \mathbf{a}_i}{\|\mathbf{x} - \mathbf{a}_i\|} \right\}. \tag{4.29}$$

(ii) Show that the corresponding fixed point method

$$\mathbf{x}_{k+1} = \frac{1}{m}\left\{ \sum_{i=1}^{m} \mathbf{a}_i + \sum_{i=1}^{m} d_i \frac{\mathbf{x}_k - \mathbf{a}_i}{\|\mathbf{x}_k - \mathbf{a}_i\|} \right\}. \tag{4.30}$$

is a gradient method, assuming that $\mathbf{x}_k \notin \mathcal{A}$ for all $k \geq 0$. What is the stepsize?

Proof.

(i) After expanding the square, $f(\mathbf{x})$ can be written as

$$f(\mathbf{x}) = m\mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T \left(\sum_{i=1}^{m} \mathbf{a}_i \right) - 2\sum_{i=1}^{m} d_i \|\mathbf{x} - \mathbf{a}_i\| + \sum_{i=1}^{m} d_i. \tag{4.31}$$

The derivative of $f(\mathbf{x})$ with respect to $\mathbf{x}$ is given by

$$\nabla f(\mathbf{x}) = 2m\mathbf{x} - 2\sum_{i=1}^{m} \mathbf{a}_i - 2\sum_{i=1}^{m} d_i \frac{\mathbf{x} - \mathbf{a}_i}{\|\mathbf{x} - \mathbf{a}_i\|}. \tag{4.32}$$

Setting it to 0 yields

$$\mathbf{x} = \frac{1}{m}\left\{ \sum_{i=1}^{m} \mathbf{a}_i + \sum_{i=1}^{m} d_i \frac{\mathbf{x} - \mathbf{a}_i}{\|\mathbf{x} - \mathbf{a}_i\|} \right\} \tag{4.33}$$

as desired.

(ii) If such a stepsize t exists, (4.29) can be written as

$$\frac{1}{m}\left\{\sum_{i=1}^{m}\mathbf{a}_i+\sum_{i=1}^{m}d_i\frac{\mathbf{x}-\mathbf{a}_i}{\|\mathbf{x}-\mathbf{a}_i\|}\right\}=\mathbf{x}-t\nabla f(\mathbf{x}) \tag{4.34}$$

$$\frac{1}{m}\sum_{i=1}^{m}\frac{d_i}{\|\mathbf{x}-\mathbf{a}_i\|}\mathbf{x}+\frac{1}{m}\sum_{i=1}^{m}\left(1-\frac{d_i}{\|\mathbf{x}-\mathbf{a}_i\|}\right)\mathbf{a}_i=\mathbf{x}-t\left(2m\mathbf{x}-2\sum_{i=1}^{m}\mathbf{a}_i-2\sum_{i=1}^{m}d_i\frac{\mathbf{x}-\mathbf{a}_i}{\|\mathbf{x}-\mathbf{a}_i\|}\right) \tag{4.35}$$

$$\frac{1}{m}\sum_{i=1}^{m}\frac{d_i}{\|\mathbf{x}-\mathbf{a}_i\|}\mathbf{x}+\frac{1}{m}\sum_{i=1}^{m}\left(1-\frac{d_i}{\|\mathbf{x}-\mathbf{a}_i\|}\right)\mathbf{a}_i=\left(1-2(m-\sum_{i=1}^{m}\frac{d_i}{\|\mathbf{x}-\mathbf{a}_i\|})t\right)\mathbf{x}+2t\sum_{i=1}^{m}\left(1-\frac{d_i}{\|\mathbf{x}-\mathbf{a}_i\|}\right)\mathbf{a}_i \tag{4.36}$$

After comparing the terms containing no $\mathbf{x}$ on both sides, we arrive at

$$t=\frac{1}{2m}. \tag{4.37}$$

We can easily verify it by plugging t into $\mathbf{x}-t\nabla f(\mathbf{x})$. This completes the proof.

□

Exercise 4.6

Another formulation of the source localization problem consists of minimizing the following objective function:

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left\{f(\mathbf{x})\equiv\sum_{i=1}^{m}(\|\mathbf{x}-\mathbf{a}_i\|^2-d_i^2)^2\right\}. \tag{4.38}$$

This is of course a nonlinear least squares problem, and thus the Gauss-Newton method can be employed in order to solve it. We will assume that $n=2$.

(i) Show that as long as all the points $\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_m$ do not reside on the same line in the plane, the method is well-defined, meaning that the linear least squares problem solved at each iteration has a unique solution.

(ii) Write a Python function that implements the damped Gauss-Newton method employed on this problem with a backtracking line search strategy with parameters $s=1,\alpha=\beta=0.5,\epsilon=10^{-4}$. Run the function on the two-dimensional problem ($n=2$) with 5 anchors ($m=5$) and data generated by the Python commands

The columns of the 2×5 matrix $\mathbf{A}$ are the locations of the five sensors, $\mathbf{x}$ is the "true" location of the source, and $\mathbf{d}$ is the vector of noisy measurements between the source and the sensors. Compare your results (e.g., number of iterations) to the gradient method with backtracking and parameters $s=1,\alpha=\beta=0.5,\epsilon=10^{-4}$. Start both methods with the initial vector $(1000,-500)^T$.

Proof. 1. Let $g_i(\mathbf{x})=\|\mathbf{x}-\mathbf{a}_i\|^2$ for $\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_m$, then we have

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left\{f(\mathbf{x})\equiv\sum_{i=1}^{m}(g_i(x)-d_i^2)^2\right\}. \tag{4.39}$$

We adopt the notation in the textbook and then the above minimization problem is the following linear least squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^n}\|\mathbf{A}_k\mathbf{x}-\mathbf{b}_k\|^2, \tag{4.40}$$

where

$$\mathbf{A}_k = \begin{pmatrix} \nabla g_1(\mathbf{x}_k)^T \\ \nabla g_2(\mathbf{x}_k)^T \\ \vdots \\ \nabla g_m(\mathbf{x}_k)^T \end{pmatrix} = \begin{pmatrix} (\mathbf{x}_k - \mathbf{a}_1)^T \\ (\mathbf{x}_k - \mathbf{a}_2)^T \\ \vdots \\ (\mathbf{x}_k - \mathbf{a}_m)^T \end{pmatrix} = J(\mathbf{x}_k) \tag{4.41}$$

and

$$\mathbf{b}_k = J(\mathbf{x}_k)\mathbf{x}_k - F(\mathbf{x}_k) \tag{4.42}$$

where

$$F(\mathbf{x}_k) = \begin{pmatrix} g_1(\mathbf{x}_k) - d_1^2 \\ g_2(\mathbf{x}_k) - d_2^2 \\ \vdots \\ g_m(\mathbf{x}_k) - d_m^2 \end{pmatrix}. \tag{4.43}$$

The minimization in (4.40) produces a unique minimizer if and only if $\mathbf{A}_k$ is of full column rank. Since translation does not change the relative positions of points, the condition that $\mathbf{A}_k$ is of full column rank is equivalent to the condition that

$$\begin{pmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_m^T \end{pmatrix} \tag{4.44}$$

is of full column rank. Furthermore, we only need to guarantee the following

$$\begin{pmatrix} (\mathbf{a}_1 - \mathbf{a}_m)^T \\ (\mathbf{a}_2 - \mathbf{a}_m)^T \\ \vdots \\ \mathbf{a}_m^T \end{pmatrix} \tag{4.45}$$

is of full column rank. When $n = 2$, it is easy to see that its rank is 2 if and if only $\mathbf{a}_1 - \mathbf{a}_m$ and $\mathbf{a}_2 - \mathbf{a}_m$ are independent. In other words, $\mathbf{a}_1 - \mathbf{a}_m \neq \lambda(\mathbf{a}_1 - \mathbf{a}_m)$ with $\lambda \neq 0$, which is equivalent to that $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_m$ are not on the same line. Therefore, as long as all the points $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_m$ do not reside on the same line in the plane, the method is well-defined.

2. Note that since the initial vector deviates greatly from the target vector, the norm of the gradients is very large, which makes the optimization process unstable. To resolve this issue, we scale the norm of gradients to 1 if the norm is greater than 1.

```
from numpy.linalg import inv

def x_A_squares(x, A):
    return np.sum((x - A)**2, axis=0, keepdims=True)

def f(A, x, d):
    return np.sum(x_A_squares(x, A) - d**2)

def grad_f(A, x, d):
    return 4 * (x-A) @ (x_A_squares(x, A) - d**2).T

def backtracking(A, x, grad, d_k, alpha, beta, s):
    i, t = 0, s
```

```python
    while f(A, x, d) - f(A, x - t * grad, d) < -alpha * t * grad.T @ d_k:
        t = s * beta**i
        i += 1
    return t

def F(A, x, d):
    return x_A_squares(x, A) - d**2

def J_k(A, x):
    return 2 * (x - A)

def d_k(A, x, d):
    J = J_k(A, x)
    return inv((J @ J.T)) @ J @ F(A, x, d).T


import numpy as np
np.random.seed(2023)
A = np.random.randn(2, 5)
src = np.random.randn(2, 1)
d = np.sqrt(np.sum((A - src)**2, axis=0, keepdims=True)) \
    + 0.05 * np.random.randn(1, 5)
alpha = beta = 0.5 # 0.5
s = 1
th_norm = 1
x = np.array([1000, -500]).reshape(-1,1).astype('float64')
for iter in range(10000):
    grad = grad_f(A, x, d)
    grad_norm = np.sqrt(np.sum(grad**2))
    grad_Gauss_Newton = d_k(A, x, d)
    grad_Gauss_Newton_norm = np.sqrt(np.sum(grad_Gauss_Newton**2))
    if grad_norm <= 1e-4:
        print("exiting", iter, grad_norm)
        break
    elif grad_norm > th_norm:
        grad = grad / grad_norm * th_norm
    if grad_Gauss_Newton_norm > th_norm:
        grad_Gauss_Newton = grad_Gauss_Newton / grad_Gauss_Newton_norm * th_norm
    # t = backtracking(A, x, grad, grad, alpha, beta, s) # backtracking
    # x -= t * grad
    t = backtracking(A, x, grad, grad_Gauss_Newton, alpha, beta, s)
    x -= t * grad_Gauss_Newton
    if iter % 1000 == 0:
        print(f"{iter}, {t:3.3f}, {grad_norm:.3}, {grad_Gauss_Newton_norm:.3}")
```

$\square$

Exercise 4.7

Let $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c$, where $\mathbf{A}$ is a symmetric $n \times n$ matrix, $\mathbf{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Show that the smallest Lipschitz constant of ∇f is $2\|\mathbf{A}\|$.

Proof. First, we have $\nabla f(\mathbf{x}) = 2\mathbf{A}\mathbf{x} + 2\mathbf{b}$. Then

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| = \|2\mathbf{A}(\mathbf{x}-\mathbf{y})\| \leq 2\|\mathbf{A}\|\|\mathbf{x}-\mathbf{y}\|. \tag{4.46}$$

Thus, the smallest Lipschitz constant of ∇f is $2\|\mathbf{A}\|$. $\square$

Exercise 4.8

Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\mathbf{x}) = \sqrt{1 + \|\mathbf{x}\|^2}$. Show that $f \in C_1^{1,1}$.

Proof. [11] First, we get the gradient

$$\nabla f(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{1+\|\mathbf{x}\|^2}} \tag{4.47}$$

which is continuous everywhere. The Hessian is given by

$$\nabla^2 f(\mathbf{x}) = \frac{1}{\sqrt{1+\|\mathbf{x}\|^2}}\left(\mathbf{I} - \frac{\mathbf{x}\mathbf{x}^T}{1+\|\mathbf{x}\|^2}\right). \tag{4.48}$$

The Rayleigh quotient of $\nabla^2 f(\mathbf{x})$ can be computed as follows. For any $\mathbf{x} \neq \mathbf{0}$,

$$R(\mathbf{x}) = \frac{\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v}}{\|\mathbf{v}\|^2} = \frac{1}{\sqrt{1+\|\mathbf{x}\|^2}}\left(1 - \frac{\|\mathbf{x}\mathbf{v}^T\|^2}{(1+\|\mathbf{x}\|^2)\|\mathbf{v}\|^2}\right) \tag{4.49}$$

from which we can get

$$0 < R(\mathbf{x}) \leq \frac{1}{\sqrt{1+\|\mathbf{x}\|^2}} < 1. \tag{4.50}$$

According to Lemma 1.12, we have

$$\max_{\mathbf{x}\neq\mathbf{0}} R(\mathbf{x}) = \lambda_{\max}(\nabla^2 f(\mathbf{x})) = \|\nabla^2 f(\mathbf{x})\|_2 < 1. \tag{4.51}$$

For $\mathbf{x} = \mathbf{0}$, $\lambda_{\max}(\nabla^2 f(\mathbf{0})) = \lambda_{\max}(\nabla^2 f(\mathbf{I})) = 1$. Therefore, for any $\mathbf{x} \in \mathbb{R}^n$, we have $\|\nabla^2 f(\mathbf{x})\|_2 \leq 1$. By Theorem 4.20, we have that $\|\nabla^2 f(\mathbf{x})\|_2 \leq 1$ is equivalent to $f \in C_1^{1,1}$ as desired. $\square$

Exercise 4.9

Let $f \in C_L^{1,1}(\mathbb{R}^m)$, and let $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$. Show that the function defined by $g(\mathbf{x}) = f(\mathbf{A}\mathbf{x}+\mathbf{b})$ satisfies $g \in C_{\tilde{L}}^{1,1}(\mathbb{R}^n)$, where $\tilde{L} = \|\mathbf{A}\|^2 L$.

Proof. First, $\nabla g(\mathbf{x}) = \mathbf{A}^T \nabla f(\mathbf{A}\mathbf{x}+\mathbf{b})$. By the definition of the induced norm,

$$\|\mathbf{A}\|_{2,2} = \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{y}\|_2}{\|\mathbf{y}\|_\mathbf{2}} \geq \frac{\|\mathbf{A}\mathbf{y}\|_2}{\|\mathbf{y}\|_\mathbf{2}} \tag{4.52}$$

which gives

$$\|\mathbf{A}\mathbf{y}\|_2 \leq \|\mathbf{A}\|_{2,2}\|\mathbf{y}\|_\mathbf{2}. \tag{4.53}$$

Given this result, we have

$$\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{x})\|_2 = \|\mathbf{A}^T\nabla f(\mathbf{A}\mathbf{x}+\mathbf{b}) - \mathbf{A}^T\nabla f(\mathbf{A}\mathbf{y}+\mathbf{b})\|_2 \tag{4.54}$$

[11] https://math.stackexchange.com/questions/3546061/lipschitz-continuity-of-sqrt1-x-2-2

$$
\begin{aligned}
&\leq \|\mathbf{A}^T\|_{2,2}\|\nabla f(\mathbf{Ax}+\mathbf{b}) - \nabla f(\mathbf{Ay}+\mathbf{b})\|_2 && (4.55)\\
&\leq L\|\mathbf{A}^T\|_{2,2}\|\mathbf{Ax}+\mathbf{b}-\mathbf{Ay}-\mathbf{b}\|_2 && (4.56)\\
&= L\|\mathbf{A}^T\|_{2,2}\|\mathbf{A}(\mathbf{x}-\mathbf{y})\|_2 && (4.57)\\
&\leq L\|\mathbf{A}^T\|_{2,2}\|\mathbf{A}\|_{2,2}\|\mathbf{x}-\mathbf{y}\|_2 && (4.58)\\
&= L\|\mathbf{A}\|^2\|\mathbf{x}-\mathbf{y}\|_2 && (4.59)
\end{aligned}
$$

as desired. □

Exercise 4.10

Given an example of a function $f \in C_L^{1,1}(\mathbb{R})$ and a starting point $x_0 \in \mathbb{R}$ such that the problem $\min f(x)$ has an optimal solution and the gradient method with constant stepsize $t = 2/L$ diverges.

Solution: A good example is the following function

$$
f(x) = \begin{cases} x^2, & x \in [-1, 1] \\ |x|, & \text{otherwise.} \end{cases} \tag{4.60}
$$

It is easy to see $f \in C_2^{1,1}(\mathbb{R})$ which has an optimal solution at $x = 0$. Given a starting point $x_0 = 1/2$ and the stepsize $2/L = 1$, the gradient method will get stuck between the points $x = -1$ and $x = 1$. □

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:

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

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

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

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. □

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. □

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

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

where the second inequality follows from the result given in Exercise 7.36. □

Exercise 7.39

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

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

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:*

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

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

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

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

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

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

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

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

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:*

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

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

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

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.

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 $\underline{x \ne y \in C, \lambda \in (0, 1)}$,

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

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

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

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 $g(\mathbf{x}) \leq 0$. $\square$

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

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

If $\|\mathbf{x} - \mathbf{x}_0\| \leq r$, 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,

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

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

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

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[12].

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

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

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.

[12] 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].