Exercises in FRTN50 – Optimization for learning

Mattias Fält, Pontus Giselsson,
Martin Morin, Hamed Sadeghi, Manu Upadhyaya

Latest update: October 11, 2021

Introduction

The exercises are divided into problem areas that roughly match the lecture schedule.

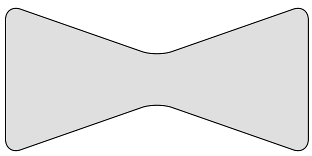
Exercises marked with (H) have hints available, listed in the end of each chapter. Challenging exercises are marked with ($\star$). Even more challenging exercises are marked with ($\star\star$).

Chapter 1

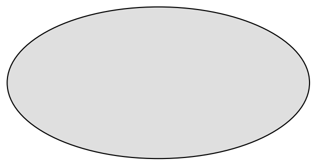
Convex sets and convex functions

Exercise 1.1

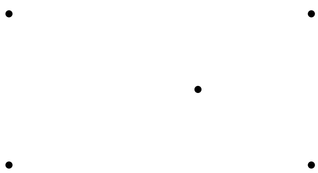
Given the following sets.



a.



b.



c.

d.

1. Which of the sets are convex. Motivate.
2. Mark all points the sets have supporting hyperplanes at.
3. Draw the convex hull of each set.

Exercise 1.2

Which of the following sets are convex? If convex, prove it using the definition of convex sets, if not convex, disprove it by finding a counter example.

1. $S = \{x \in \mathbb{R}^n : Ax = b\}$ with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$
2. $S = \{x \in \mathbb{R}^n : Ax \leq b\}$ with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$
3. $S = \{x \in \mathbb{R}^n : x \geq 0\}$
4. $S = \{x \in \mathbb{R}^n : l \leq x \leq u\}$ with $l, b \in \mathbb{R}^n$ such that $l \leq u$

5. $S = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$
6. $S = \{x \in \mathbb{R}^n : -\|x\|_2 \leq -1\}$
7. $S = \{x \in \mathbb{R}^n : -\|x\|_2 \leq 1\}$
8. $S = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : \|x\|_2 \leq t\}$
9. $S = \{X \in \mathbb{R}^{n \times n} : X \succeq 0\}$
10. $S = \{x \in \mathbb{R}^n : x = a\}$ with $a \in \mathbb{R}^n$
11. $S = \{x \in \mathbb{R}^n : x = a$ or $x = b\}$ with $a, b \in \mathbb{R}^n$ such that $a \neq b$

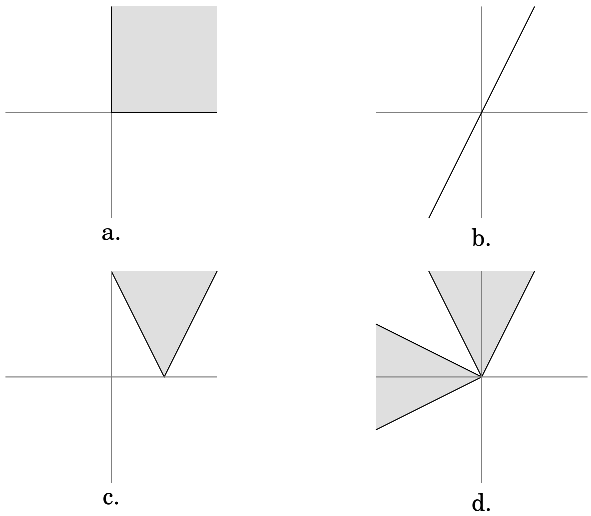
Exercise 1.3

Which of the following sets are affine?

1. $V = \{x \in \mathbb{R}^n : x = a\}$ for some $a \in \mathbb{R}^n$
2. $V = \{x \in \mathbb{R}^n : \exists \alpha \in [0, 1]$ such that $x = \alpha a + (1 - \alpha)b\}$ for some $a, b \in \mathbb{R}^n$ such that $a \neq b$
3. $V = \{x \in \mathbb{R}^n : \exists \alpha \in \mathbb{R}$ such that $x = \alpha a + (1 - \alpha)b\}$ for some $a, b \in \mathbb{R}^n$ such that $a \neq b$

Exercise 1.4

A set K is a cone if for each $x \in K$ also $\alpha x \in K$ for each $\alpha \geq 0$. Which of the following figures represent cones? Which of them are convex?



Exercise 1.5

Which of the following sets are convex cones? Prove or disprove. Assume that each set is nonempty.

1. $S = \{x \in \mathbb{R}^n : Ax = 0\}$ with $A \in \mathbb{R}^{m \times n}$
2. $S = \{x \in \mathbb{R}^n : Ax = b\}$ with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ such that $b \neq 0$
3. $S = \{x \in \mathbb{R}^n : Ax \leq 0\}$ with $A \in \mathbb{R}^{m \times n}$
4. $S = \{x \in \mathbb{R}^n : Ax \leq b\}$ with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ such that there exists at least one index $j \in \{1, \dots, m\}$ such that row j in the matrix A is nonzero and b_j is nonzero
5. $S = \{x \in \mathbb{R}^n : x \geq 0\}$
6. $S = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : \|x\|_2 \leq t\}$
7. $S = \{X \in \mathbb{R}^{n \times n} : X \succeq 0\}$

Exercise 1.6

Suppose that C_1 and C_2 are convex sets in $\mathbb{R}^n$.

1. Is the set $C = \{x \in \mathbb{R}^n : x \in C_1 \text{ and } x \in C_2\}$ the union or intersection of C_1 and C_2? Is it convex? Prove or provide a counter example
2. Is the set $C = \{x \in \mathbb{R}^n : x \in C_1 \text{ or } x \in C_2\}$ the union or intersection of C_1 and C_2? Is it convex? Prove or provide a counter example

Exercise 1.7

Let $\{C_j\}_{j \in J}$ be an indexed family of convex sets in $\mathbb{R}^n$, with index set J (J can be finite, countable or uncountable). Show that

$$\bigcap_{j \in J} C_j$$

is convex.

Exercise 1.8

Prove convexity for each of the following sets.

1. Affine hyperplanes. Recall that affine hyperplanes are written as $h_{s,r} = \{x \in \mathbb{R}^n : s^T x = r\}$ for some $s \in \mathbb{R}^n$ and $r \in \mathbb{R}$
2. Halfspaces. Recall that halfspaces are written as $H_{s,r} = \{x \in \mathbb{R}^n : s^T x \leq r\}$ for some $s \in \mathbb{R}^n$ and $r \in \mathbb{R}$
3. Polytopes. Recall that a polytope C can be represented as

   $$C = \{x \in \mathbb{R}^n : s_i^T x \leq r_i \text{ for } i \in \{1, \dots, m\} \text{ and } s_i^T x = r_i \text{ for } i \in \{m+1, \dots, p\}\},$$

   where $s_i \in \mathbb{R}^n$ and $r_i \in \mathbb{R}$ for each $i \in \{1, \dots, p\}$

Exercise 1.9 (H)

Prove, without explicitly using the definition of convex sets, that each of the following sets are convex set.

1. $S = \{x \in \mathbb{R}^n : Ax = b\}$ with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$
2. $S = \{x \in \mathbb{R}^n : Ax \leq b\}$ with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$
3. $S = \{x \in \mathbb{R}^n : x \geq 0\}$
4. $S = \{x \in \mathbb{R}^n : l \leq x \leq u\}$ with $l, b \in \mathbb{R}^n$ such that $l \leq u$
5. $S = \{x \in \mathbb{R}^n : x = a\}$ with $a \in \mathbb{R}^n$

Exercise 1.10 ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a function, and let $C \subseteq \mathbb{R}^n$ and $D \subseteq \mathbb{R}^m$ be two sets. The image of C under f is denote by $f(C)$ and is defined by

$$f(C) = \{f(x) : x \in C\}.$$

The inverse image of D under f is denote by $f^{-1}(D)$ and is defined by

$$f^{-1}(D) = \{x : f(x) \in D\}.$$

Now suppose that f is an affine function (or map), i.e. $f(x) = Ax+b$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, and let both sets C and D be convex. Show that

1. $f(C)$ is convex
2. $f^{-1}(D)$ is convex

Exercise 1.11 ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function, i.e., let f satisfy

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

for each $\theta = [0, 1]$ and for each $x, y \in \mathbb{R}^n$. The effective domain of f is defined as $\mathrm{dom} f = \{x \in \mathbb{R}^n : f(x) < \infty\}$. Show that $\mathrm{dom} f$ is convex.

Exercise 1.12

Show or disprove that the following functions are convex.

1. $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ equal to the indicator function of convex set $C \subseteq \mathbb{R}^n$, i.e.

$$f(x) = \iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{otherwise} \end{cases}$$

2. $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that $f(x) = \|x\|$ for each $x \in \mathbb{R}^n$

3. $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that $f(x) = -\|x\|$ for each $x \in \mathbb{R}^n$
4. $f : \mathbb{R}^2 \to \mathbb{R} \cup \{\infty\}$ such that $f(x, y) = xy$ for each $(x, y) \in \mathbb{R}^2$
5. $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that $f(x) = a^T x + b$ for each $x \in \mathbb{R}^n$, where $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$
6. $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that $f(x) = \frac{1}{2} x^T Q x$ for each $x \in \mathbb{R}^n$, where $Q \in \mathbb{S}^n_+$
7. $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that $f(x) = \mathrm{dist}_C(x) = \inf_{y \in C} \|x - y\|$ for each $x \in \mathbb{R}^n$, where $C \subseteq \mathbb{R}^n$ is a nonempty closed convex set

Exercise 1.13

Draw the epigraph of the following functions $f : \mathbb{R} \to \mathbb{R}$:

- $f(x) = |x|$
- $f(x) = x^2$
- $f(x) = |x| + x^2$
- $f(x) = \max(|x|, x^2)$
- $f(x) = \min(|x|, x^2)$

Exercise 1.14

Let $f : \mathbb{R}^n \to \mathbb{R}$ be an affine function defined by

$$f(x) = a^T x + b$$

for each $x \in \mathbb{R}^n$, where $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Show that $\mathrm{epi} f$ is a halfspace in $\mathbb{R}^{n+1}$.

Exercise 1.15

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a function. Recall that the epigraph of f is given by $\mathrm{epi} f = \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq r\}$. Show that f is convex if and only if $\mathrm{epi} f$ is convex.

Exercise 1.16 (H)

For each $i = 1, \ldots, m$, assume that the function $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a convex. Prove the following explicitly, without resorting to convexity preserving operations on functions.

1. Show that $f(x) = \sum_{i=1}^m \alpha_i f_i(x)$ is convex, where $\alpha_i \geq 0$ for each $i = 1, \ldots, m$
2. Show that $f(x) = \max_{i=1,\ldots,m} f_i(x)$ is convex

Exercise 1.17

Show that the following functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are convex. You may use convexity preserving operations.

1. $f(x) = \|x\|^p$ where $p \geq 1$
2. $f(x) = \|Ax - b\|_2^2 + \|x\|_1$ where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$
3. $f(x) = \max(\|x\|, \|x\|^2, \|x\|^3)$
4. $f(x) = \sum_{i=1}^n \max(0, 1 + x_i) + \|x\|_2^2$
5. $f(x) = \sup_{y \in \mathbb{R}^n} (x^T y - g(y))$ where $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is proper

Exercise 1.18

Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and define $C_\alpha = \{x \in \mathbb{R}^n : g(x) \leq \alpha\}$ for each $\alpha \in \mathbb{R}$.

1. Suppose that g is convex and suppose that there exists an $\bar{x} \in \mathbb{R}^n$ such that $g(\bar{x}) < \alpha$ for some $\alpha \in \mathbb{R}$. Show that C_α is a nonempty convex set.
2. For $n = 1$, construct a *nonconvex* function g such that C_0 is *convex*.
3. For $n = 1$, construct a *nonconvex* function g such that C_0 is *nonconvex*.

Exercise 1.19

Let $f : \mathbb{R}^{n_1} \to \mathbb{R} \cup \{\infty\}$ be a convex function and define a function $g : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R} \cup \{\infty\}$ such that $g(x, y) = f(x)$. Show that g is a convex function.

Exercise 1.20 (H)

Prove, without explicitly using the definition of convex sets, that each of the following sets are convex.

1. $S = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$
2. $S = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : \|x\|_2 \leq t\}$

Exercise 1.21 (H)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function. Suppose that $x^\star \in \mathbb{R}^n$ is a local optimum, i.e., there exists an $\delta > 0$ such that

$$f(x^\star) \leq f(x)$$

for each $x \in \mathbb{R}^n$ such that $\|x - x^\star\| \leq \delta$. Show that $x^\star$ is a global minimum, i.e.

$$f(x^\star) \leq f(x)$$

for each $x \in \mathbb{R}^n$.

Exercise 1.22

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a proper strictly convex function. Recall that f is called strictly convex if

$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta)f(y)$$

for each $x, y \in \mathrm{dom} f$ such that $x \neq y$ and for each $\theta \in (0,1)$. Completely analogous to Exercise 1.11, one can show that $\mathrm{dom} f$ must be convex.

1. Suppose that there exists a point $x^* \in \mathbb{R}^n$ such that
$$f(x^\star) \leq f(x) \tag{1.1}$$
for each $x \in \mathbb{R}^n$. Show that $x^\star$ is the unique minimizer of f.
2. Provide a strictly convex f whose infimum is not attained by any point $x^\star$.

Remark: For (proper, closed and) strongly convex functions, a minimizer always exists. Moreover, since strongly convex functions are strictly convex, the minimizer is unique.

Exercise 1.23

Decide which of the following convex functions $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ are

- smooth,
- strictly convex,
- strongly convex,

or none of the above. In this exercise, you only need to draw/plot the functions and decide from the drawings.

1. $f(x) = \begin{cases} -\log(x) & \text{if } x > 0 \\ \infty & \text{if } x \leq 0 \end{cases}$
2. $f(x) = \begin{cases} \dfrac{1}{x} & \text{if } x > 0 \\ \infty & \text{if } x \leq 0 \end{cases}$
3. $f(x) = x$
4. $f(x) = \frac{1}{2}x^2$
5. $f(x) = |x|$
6. $f(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq 1 \\ |x| - \frac{1}{2} & \text{else} \end{cases}$
7. $f(x) = e^x$
8. $f(x) = x^4$

Exercise 1.24 (H)

Suppose we are given some function $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ where we only know that $f(-1) = 0$ and $f(1) = 1$. For $x \in [-1, 1]$, draw the known bounds on $f(x)$ given the following assumptions:

- f is convex
- f is convex and 2-smooth
- f is 2-smooth and $\frac{1}{2}$-strongly convex

For each case, draw an example of a function that satisfies the assumptions.

Exercise 1.25 (H)

Suppose we are given some differentiable function $f : \mathbb{R} \to \mathbb{R}$ where we only know that $f(1) = 1$ and $f'(1) = 1$. Draw the known bounds on f given the following assumptions:

- f is strictly convex.
- f is strictly convex and 2-smooth.
- f is 2-smooth and 1-strongly convex.

For each case, draw an example of a function that satisfies the assumptions.

Exercise 1.26 (H)

Suppose that $a, b \geq 0$ and $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Show that

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \tag{1.2}$$

Inequality (1.2) is called Young's inequality.

Exercise 1.27 (H) ($\star$)

Consider the following statement: A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{1.3}$$

for each $x, y \in \mathbb{R}^n$.

1. Show that the statement is true.
2. Provide a nonconvex differentiable function f and a point x for which (1.3) does not hold.

Exercise 1.28 (H)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Suppose that the point $x \in \mathbb{R}^n$ satisfies $\nabla f(x) = 0$. Show that x is a global minimizer of f.

Exercise 1.29 ($\star$)

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function. Show that f is strictly convex if and only if

$$f(y) > f(x) + \nabla f(x)^T (y - x) \tag{1.4}$$

for each $x, y \in \mathbb{R}^n$ such that $x \neq y$.

Exercise 1.30 (H)

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function and let $\sigma > 0$. Show that f is σ-strongly convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\sigma}{2} \|x - y\|_2^2 \tag{1.5}$$

for each $x, y \in \mathbb{R}^n$.

Exercise 1.31 ($\star$)

The indicator function $\iota_C : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ of a set $C \subseteq \mathbb{R}^n$ is defined as

$$\iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{otherwise.} \end{cases}$$

Show the following:

1. Let $K \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $C = \{x \in \mathbb{R}^n : Kx - b = 0\}$. Show that

$$\iota_C(x) = \sup_{\mu \in \mathbb{R}^m} \mu^T (Kx - b).$$

2. Let $g : \mathbb{R}^n \to \mathbb{R}^m$ and $C = \{x \in \mathbb{R}^n : g(x) \leq 0\}$. Show that

$$\iota_C(x) = \sup_{\mu \in \mathbb{R}^m_+} \mu^T g(x).$$

Exercise 1.32 (H) ($\star$)

Solve the following problems:

1. Suppose that $h : \mathbb{R} \to \mathbb{R}$ is differentiable with nondecreasing derivative. Show that h is convex.

2. Let $p \geq 1$. Show that the function $h : \mathbb{R} \to \mathbb{R}$ given by

$$h(x) = \begin{cases} x^p & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

is a nondecreasing convex function.

Exercise 1.33 (H) ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function. Let $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathbb{R}^n$ and $\theta_1, \ldots, \theta_n \geq 0$ such that $\sum_{i=1}^n \theta_i = 1$. Show that

$$f\left(\sum_{i=1}^n \theta_i x_i\right) \leq \sum_{i=1}^n \theta_i f(x_i). \tag{1.6}$$

Inequality (1.6) is called *Jensen's inequality*.

Exercise 1.34 ($\star\star$)

Let $f : \mathbb{R}^n \to \mathbb{R}$. Show that f is affine if and only if f is convex and concave.

Exercise 1.35 ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and let $\sigma > 0$. Recall that f is called σ-strongly convex if $f - \frac{\sigma}{2} \|\|_2^2$ is convex. Show that f is σ-strongly convex if and only if

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) - \tfrac{\sigma}{2}\theta(1-\theta)\|x - y\|^2 \tag{1.7}$$

for each $x, y \in \mathbb{R}^n$ and for each $\theta \in [0, 1]$.

Exercise 1.36

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and let $\beta \geq 0$. Suppose that f is β-smooth, i.e., f is differentiable and ∇f is β-Lipschitz continuous. Let $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^n$. Let $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ such that

$$g(x) = f(Ax + b)$$

for each $x \in \mathbb{R}^m$. Show that g is $\beta \|A\|_2^2$-smooth.

Remark: Recall that $\|A\|_2$ is the spectral norm of the matrix A and that $\|A\|_2 = \left\|A^T\right\|_2$ holds.

Exercise 1.37 ($\star\star$)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and let $\beta \geq 0$. Consider the following properties

I) $\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2$, for each $x, y \in \mathbb{R}^n$, i.e., f is β-smooth

II) For each $x, y \in \mathbb{R}^n$

$$\begin{cases} f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2}\|x - y\|_2^2, \\ f(y) \geq f(x) + \nabla f(x)^T (y - x) - \frac{\beta}{2}\|x - y\|_2^2 \end{cases}$$

III) $\frac{\beta}{2}\|\cdot\|_2^2 - f$ and $f + \frac{\beta}{2}\|\cdot\|_2^2$ are convex

IV) For each $x, y \in \mathbb{R}^n$ and for each $\theta \in [0, 1]$

$$\begin{cases} f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) + \frac{\beta}{2}\theta(1-\theta)\|x - y\|_2^2, \\ f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta) f(y) - \frac{\beta}{2}\theta(1-\theta)\|x - y\|_2^2 \end{cases}$$

Show that these properties are equivalent.

Exercise 1.38 (H)($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function and let $\beta \geq 0$. Show that the following properties are equivalent:

I) $\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2$, for each $x, y \in \mathbb{R}^n$, i.e., f is β-smooth

II) $-\beta I \preceq \nabla^2 f(x) \preceq \beta I$, for each $x \in \mathbb{R}^n$

Hints

Hint to exercise 1.9

Use the results from Exercise 1.8.

Hint to exercise 1.16

For the second subproblem, use the fact that a function is convex if and only if its epigraph is convex, i.e. use Exercise 1.15.

Hint to exercise 1.20

Use the results from Exercise 1.18 and 1.19.

Hint to exercise 1.21

Use a proof by contradiction.

Hint to exercise 1.24
Recall that f is 2-smooth if and only if

$$\begin{cases} f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) + \theta(1-\theta)\|x-y\|_2^2, \\ f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta)f(y) - \theta(1-\theta)\|x-y\|_2^2 \end{cases}$$

for each $x, y \in \mathbb{R}$ and all $\theta \in [0,1]$ and that f is $\frac{1}{2}$-strongly convex if and only if

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \tfrac{1}{4}\theta(1-\theta)\|x-y\|^2$$

for each $x, y \in \mathbb{R}$ and all $\theta \in [0,1]$.

Hint to exercise 1.25
Recall that f is 2-smooth if and only if

$$\begin{cases} f(y) \leq f(x) + \nabla f(x)^T(y-x) + \|x-y\|_2^2, \\ f(y) \geq f(x) + \nabla f(x)^T(y-x) - \|x-y\|_2^2 \end{cases}$$

for each $x, y \in \mathbb{R}$ and that f is 1-strongly convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \tfrac{1}{2}\|x-y\|_2^2$$

for each $x, y \in \mathbb{R}$.

Hint to exercise 1.26
Consider the case $a = 0$ or $b = 0$ and the case $a > 0$ and $b > 0$ separately. Moreover, note that

$$x = \exp(\ln x)$$

for each $x > 0$.

Hint to exercise 1.27
The directional derivative of f at $x \in \mathbb{R}^n$ in direction $d \in \mathbb{R}^n$ satisfies

$$\lim_{\theta \to 0} \frac{f(x+\theta d) - f(x)}{\theta} = \nabla f(x)^T d.$$

Hint to exercise 1.28
Use Exercise 1.27.

Hint to exercise 1.30

Use Exercise 1.27.

Hint to exercise 1.32

1. The mean value theorem might be helpful.
2. Consider the cases $p = 1$ and $p > 1$ separately.

Hint to exercise 1.33
Use induction on n.

Hint to exercise 1.38
Use Exercise 1.37 and the second-order condition for convex functions.

Chapter 2

Subdifferentials and proximal operators

Exercise 2.1

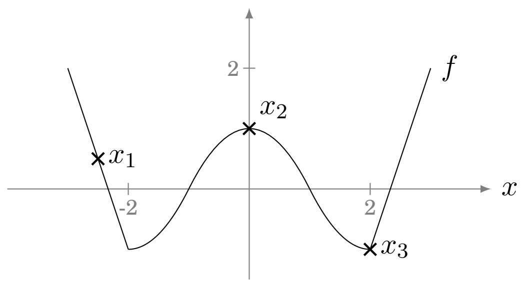
Compute the subdifferentials for the following proper closed convex functions:

1. $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x) = \frac{1}{2}\|x\|_2^2$ for each $x \in \mathbb{R}^n$
2. $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x) = \frac{1}{2}x^T Hx + h^T x$ for each $x \in \mathbb{R}^n$, where $H \in \mathbb{S}^n_+$
3. $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = |x|$ for each $x \in \mathbb{R}$
4. $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ such that $f(x) = \iota_{[-1,1]}(x)$ for each $x \in \mathbb{R}$
5. $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \max(0, 1 + x)$ for each $x \in \mathbb{R}$. This is known as the hinge loss
6. $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \max(0, 1 - x)$ for each $x \in \mathbb{R}$

You are allowed to rely on graphical arguments in this exercise.

Exercise 2.2

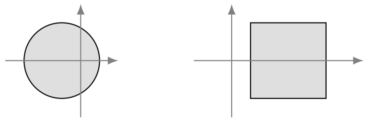
Consider the following even nonconvex function $f : \mathbb{R} \to \mathbb{R}$:



1. Compute (approximate) gradient and subdifferential at x_1, x_2, and x_3.
2. As which of the points x_1, x_2, and x_3 does Fermat's rule hold?

Exercise 2.3

Assume that f and g are two real-valued functions. Figure (a) depicts $\partial f(x)$ and Figure (b) depicts $\partial g(y)$.



(a) (b)

1. What are the domains for f and g? Note that we are not asking for the effective domains $\mathrm{dom} f$ and $\mathrm{dom} g$.
2. Is x a minimum to f?
3. Is y a minimum to g?
4. Is f differentiable at x?
5. Is g differentiable at y?
6. Draw/explain examples of functions f and g that comply with the figures

Exercise 2.4

Suppose that $f : \mathbb{R} \to \mathbb{R}$ satisfies

$$f(-1) = 1, \quad \partial f(-1) = \{-1\}$$

and

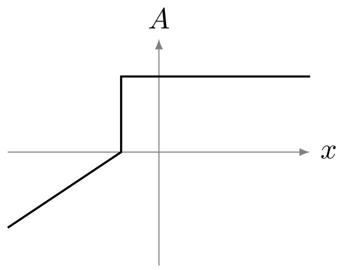
$$f(1) = 1, \quad \partial f(1) = \{1\}.$$

1. Draw a function that lower bounds f
2. Compute a lower bound to the optimal value of f
3. Draw a function f that complies with the requirements

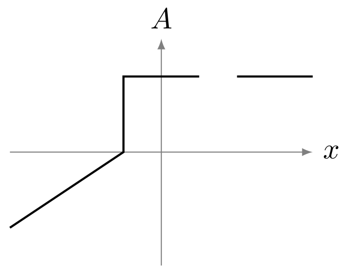
Exercise 2.5

Below a list of set-valued operators $A : \mathbb{R} \to 2^{\mathbb{R}}$ are given.

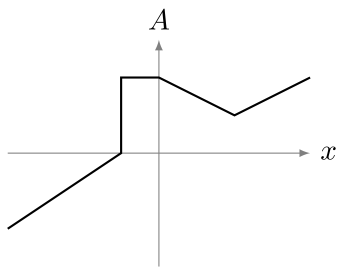
- Which of them are monotone?
- Which of them can be a subdifferential of a closed convex function?



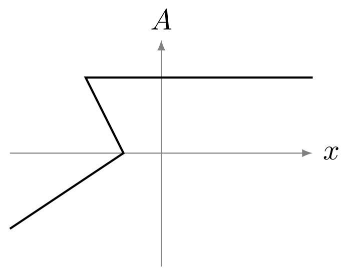
a.



b.



c.



d.

Exercise 2.6

Let $A : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ be an operator and let $\sigma > 0$. Show that A is σ-strongly monotone if and only if $A - \sigma I$ is monotone.

Remark: In particular, note that if $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the subdifferential ∂f is σ-strongly monotone if and only if $\partial f - \sigma I$ is monotone.

Exercise 2.7 ($\star$)

Provide a monotone operator $A : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ that is monotone but not the subdifferential of a function.

Exercise 2.8 (H)($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Then the following properties are equivalent:

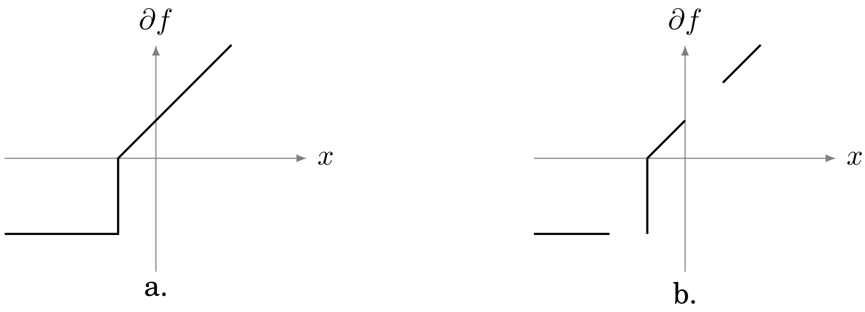
I) $f(y) \geq f(x) + \nabla f(x)^T(y - x)$ for each $x, y \in \mathbb{R}^n$, i.e. f is convex

II) $(\nabla f(y) - \nabla f(x))^T(y - x) \geq 0$ for each $x, y \in \mathbb{R}^n$, i.e. ∇f is monotone

1. Show that I) implies II)
2. Show that II) implies I)

Exercise 2.9

The subdifferential ∂f of two functions $f : \mathbb{R} \to \mathbb{R}$ are drawn below.



1. Are the corresponding functions f closed and convex?
2. Can you find an x^* that minimizes f. If so, where is it?
3. Can you compute the optimal value $f(x^*)$?
4. Draw examples of corresponding f

Exercise 2.10 ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be closed and let $\sigma > 0$. We denote the effective domain of the subdifferential ∂f as $\mathrm{dom}\partial f$ and define it as

$$\mathrm{dom}\partial f = \{x \in \mathbb{R}^n : \partial f(x) \neq \emptyset\}.$$

Assume that f is σ-strongly convex. Show that

$$f(y) \geq f(x) + s^T(y - x) + \frac{\sigma}{2} \|x - y\|_2^2$$

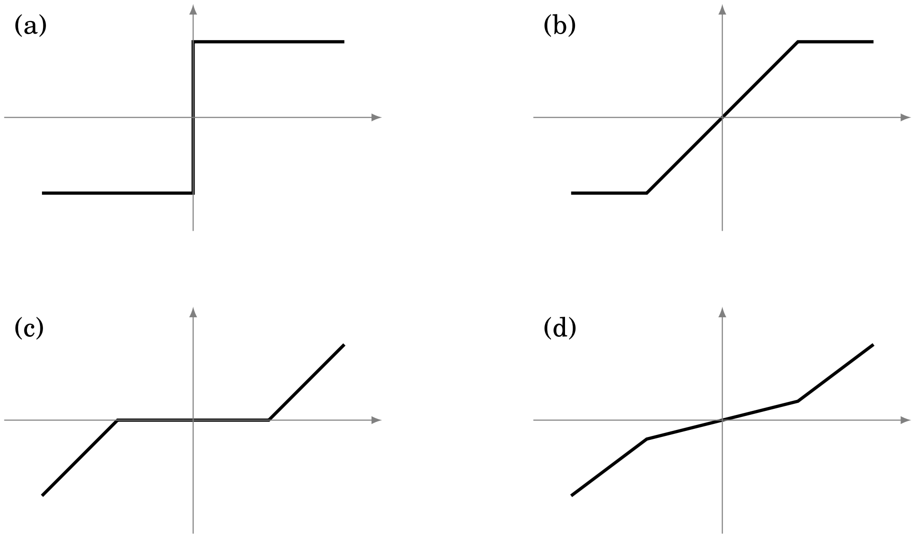
for each $y \in \mathbb{R}^n$, for each $x \in \mathrm{dom}\partial f$ and for each $s \in \partial f(x)$.

Exercise 2.11

The subdifferentials of four closed convex functions $f : \mathbb{R} \to \mathbb{R}$ are drawn below. State for each if

- f is differentiable,
- ∇f is Lipschitz continuous and
- f is strongly convex.

Also, if they exists, estimate the Lipschitz and the strong convexity parameters (given that the axes are equal).



Exercise 2.12 (⋆)
Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Suppose there exists $g_i : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ for each $i = 1, \ldots, n$ such that

$$g(x) = \sum_{i=1}^{n} g_i(x_i), \qquad \text{where} \qquad x = (x_1, \ldots, x_n)$$

for each $x \in \mathbb{R}^n$. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and let $s = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Show that $s \in \partial g(x)$ if and only if $s_i \in \partial g_i(x_i)$ for each $i = 1, \ldots, n$.

Exercise 2.13 (⋆)
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex and let $y \in \mathbb{R}^n$ be a point such that $f(y) < \infty$. Show that $\partial f(x)$ is empty for each $x \notin \mathrm{dom} f$.

Exercise 2.14 (⋆)

Show that the subdifferential of the indicator function of a nonempty set $C \subseteq \mathbb{R}^n$ is the normal cone to C.

Exercise 2.15
Compute the proximal mapping for the following proper closed convex functions:

1. $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x) = \frac{1}{2}\|x\|_2^2$ for each $x \in \mathbb{R}^n$
2. $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x) = \frac{1}{2}x^T Hx + h^T x$ for each $x \in \mathbb{R}^n$, where $H \in \mathbb{S}_+^n$
3. $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = |x|$ for each $x \in \mathbb{R}$

4. $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ such that $f(x) = \iota_{[-1,1]}(x)$ for each $x \in \mathbb{R}$
5. $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \max(0, 1 + x)$ for each $x \in \mathbb{R}$
6. $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \max(0, 1 - x)$ for each $x \in \mathbb{R}$

Exercise 2.16

Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a proper, closed and convex function. Suppose there exists $g_i : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ for each $i = 1, \ldots, n$ such that

$$g(x) = \sum_{i=1}^{n} g_i(x_i), \quad \text{where} \quad x = (x_1, \ldots, x_n)$$

for each $x \in \mathbb{R}^n$. Let $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ and let $\gamma > 0$. Show that

$$\operatorname{prox}_{\gamma g}(z) = \begin{bmatrix} \operatorname{prox}_{\gamma g_1}(z_1) \\ \vdots \\ \operatorname{prox}_{\gamma g_n}(z_n) \end{bmatrix}.$$

Hints

Hint to exercise 2.8

1. Add I) and I) with x and y swapped.
2. Let $x, y \in \mathbb{R}^n$ and $t \in \mathbb{R}$. Then

$$\frac{\partial}{\partial t} f(x + t(y - x)) = \nabla f(x + t(y - x))^T (y - x).$$

This gives that

$$f(y) - f(x) = \int_0^1 \nabla f(x + t(y - x))^T (y - x) dt. \tag{2.1}$$

Subtracting $\nabla f(x)^T (y - x)$ from the expression above yields

$$\begin{aligned} & f(y) - f(x) - \nabla f(x)^T (y - x) \\ &= \int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x) dt \\ &= \int_0^1 t^{-1} (\nabla f(x + t(y - x)) - \nabla f(x))^T ((x + t(y - x)) - x) dt. \end{aligned}$$

Chapter 3

Conjugate functions and duality

Exercise 3.1

Compute the conjugates for the following proper closed convex functions:

1. $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x) = \frac{1}{2}\|x\|_2^2$ for each $x \in \mathbb{R}^n$
2. $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x) = \frac{1}{2}x^T Hx + h^T x$ for each $x \in \mathbb{R}^n$, where $H \in \mathbb{S}_+^n$
3. $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ such that $f(x) = \iota_{[-1,1]}(x)$ for each $x \in \mathbb{R}$
4. $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = |x|$ for each $x \in \mathbb{R}$
5. $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \max(0, 1 + x)$ for each $x \in \mathbb{R}$
6. $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \max(0, 1 - x)$ for each $x \in \mathbb{R}$

Exercise 3.2

Let $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be two functions. Show that

1. $f^{**} \leq f$
2. $f \leq g$ implies that $f^* \geq g^*$
3. $f \leq g$ implies that $f^{**} \leq g^{**}$
4. $f = f^*$ if and only if $f = \frac{1}{2}\|\cdot\|_2^2$

Exercise 3.3 (H)

Let $p \in (1, \infty)$ and $q = p/(p-1)$. Show that

$$\left(\frac{|\cdot|^p}{p}\right)^* = \left(\frac{|\cdot|^q}{q}\right).$$

Exercise 3.4

Let $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $\alpha \in (0,1)$. Show that

$$(\alpha f + (1-\alpha)g)^* \leq \alpha f^* + (1-\alpha)g^*.$$

Exercise 3.5

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $f_i : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ for each $i = 1, \ldots, n$. Suppose that

$$f(x) = \sum_{i=1}^{n} f_i(x_i)$$

for each $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, i.e, f is separable. Show that

$$f^*(s) = \sum_{i=1}^{n} f_i^*(s_i)$$

for each $s = (s_1, \ldots, s_n) \in \mathbb{R}^n$, i.e, f^* is also separable.

Exercise 3.6 (H)

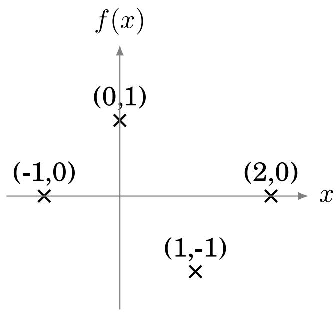
Compute the conjugates of the following functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$:

1. $f(x) = \|x\|_1$ for each $x \in \mathbb{R}^n$
2. $f(x) = \iota_{[-\mathbf{1},\mathbf{1}]}(x)$ for each $x \in \mathbb{R}^n$, where $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$

Exercise 3.7

Let $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be the nonconvex function in the figure below. It satisfies

$$f(x) = \begin{cases} 0 & \text{if } x = -1, \\ 1 & \text{if } x = 0, \\ -1 & \text{if } x = 1, \\ 0 & \text{if } x = 2, \\ \infty & \text{otherwise .} \end{cases}$$



1. Draw the conjugate f^* of f

2. Draw the biconjugate f^{**} of f

Exercise 3.8 (H) ($\star$)
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that

$$f(x) = \|x\|_2$$

for each $x \in \mathbb{R}^n$.

1. Compute the conjugate f^* via the following steps:
   (a) Show that $f^*(s) \geq 0$ for each $s \in \mathbb{R}^n$
   (b) Show that $f^*(s) \leq 0$ for each $s \in \mathbb{R}^n$ such that $\|s\|_2 \leq 1$
   (c) Show that $f^*(s) = \infty$ for each $s \in \mathbb{R}^n$ such that $\|s\|_2 > 1$
   (d) Combine the results and give $f^*(s)$
2. Use the conjugate to compute the subdifferential of f

Exercise 3.9 ($\star$)
Let Δ be the n-dimensional probability simplex, i.e.

$$\Delta = \left\{x \in \mathbb{R}^n : x \geq 0 \text{ and } \mathbf{1}^T x = 1\right\}.$$

Similarly, let D be the set

$$D = \left\{x \in \mathbb{R}^n : x \geq 0 \text{ and } \mathbf{1}^T x \leq 1\right\}.$$

1. Let $f = \iota_\Delta$. Show that

   $$f^*(s) = \max_{i=1,\ldots,n} s_i$$

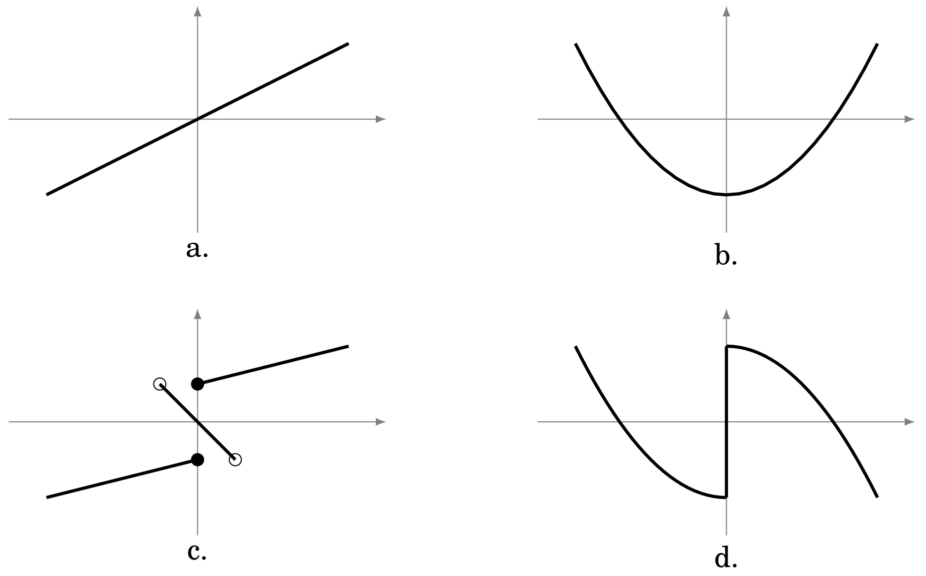
   for each $s = (s_1, \ldots, s_n) \in \mathbb{R}^n$
2. Find f^{**}
3. Let $g = \iota_D$. Show that

   $$g^*(s) = \max\left(0, \max_{i=1,\ldots,n} s_i\right)$$

   for each $s = (s_1, \ldots, s_n) \in \mathbb{R}^n$
4. Find g^{**}

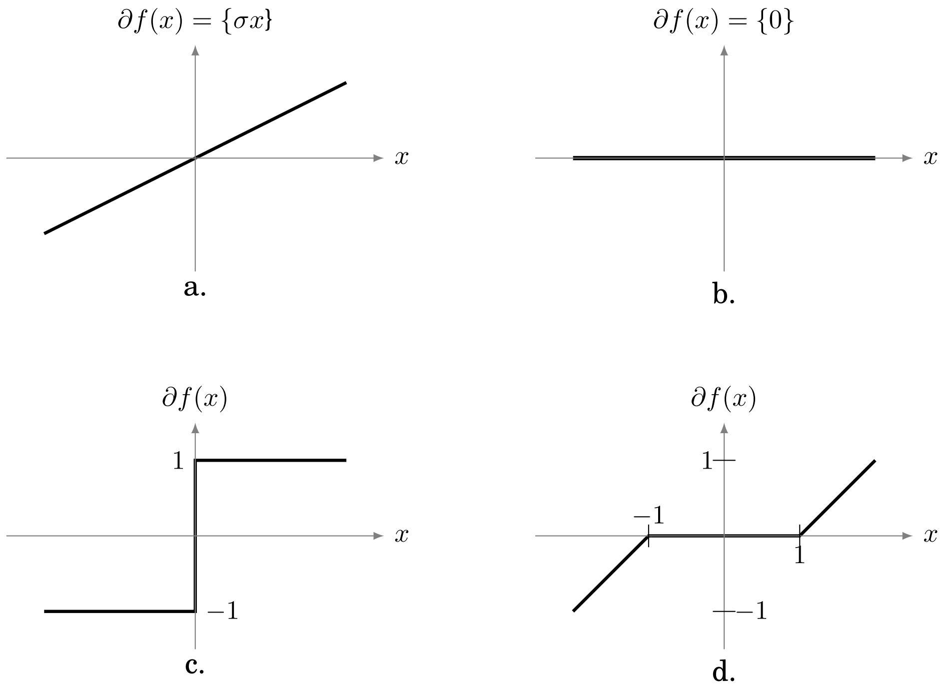
Exercise 3.10
Consider the following set-valued operators $A : \mathbb{R} \to 2^{\mathbb{R}}$:

1. Draw the inverses $A^{-1} : \mathbb{R} \to 2^{\mathbb{R}}$
2. Which operators A are functions $f : \mathbb{R} \to \mathbb{R}$?
3. Which operator inverses A^{-1} are functions $f : \mathbb{R} \to \mathbb{R}$?



Exercise 3.11

Consider the following four subdifferentials ∂f of proper closed convex functions in the figure below. Decide ∂f^*, i.e., the subdifferential of the conjugate.



Exercise 3.12

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper closed convex and $\gamma > 0$. Show that

$$\mathrm{prox}_{\gamma f}(z) = (I + \gamma \partial f)^{-1}(z)$$

for each $z \in \mathbb{R}^n$, where the inverse denotes the operator inverse.

Exercise 3.13 (H)

Compute the proximal mapping for the following convex functions on $\mathbb{R}$. Use graphical arguments and that $\mathrm{prox}_{\gamma f}(z) = (I + \gamma \partial f)^{-1}(z)$.

1. $f(x) = |x|$
2. $f(x) = \iota_{[-1,1]}(x)$
3. $f(x) = \max(0, 1 + x)$
4. $f(x) = \max(0, 1 - x)$

Exercise 3.14 (H)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper closed convex and $\gamma > 0$. Show that:

1. $\mathrm{prox}_f(z) + \mathrm{prox}_{f^*}(z) = z$ for each $z \in \mathbb{R}^n$
2. $(\gamma f)^*(s) = \gamma f^*(\gamma^{-1} s)$ for each $s \in \mathbb{R}^n$
3. $\mathrm{prox}_{(\gamma f)^*}(z) = \gamma \mathrm{prox}_{\gamma^{-1} f^*}(\gamma^{-1} z)$ for each $z \in \mathbb{R}^n$
4. $\mathrm{prox}_{\gamma f}(z) + \gamma \mathrm{prox}_{\gamma^{-1} f^*}(\gamma^{-1} z) = z$ for each $z \in \mathbb{R}^n$

Exercise 3.15

Let $\gamma > 0$. Compute the $\mathrm{prox}_{(\gamma f)^*}$ for the following f:

1. $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x) = \frac{1}{2} x^T H x + h^T x$ for each $x \in \mathbb{R}^n$, where $H \in \mathbb{S}^n_{++}$
2. $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \max(0, 1 + x)$ for each $x \in \mathbb{R}$
3. $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \max(0, 1 - x)$ for each $x \in \mathbb{R}$

Exercise 3.16

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$.

1. Show that

$$\inf_{x \in \mathbb{R}^n} f(x) = -f^*(0)$$

2. Suppose that f is proper closed convex. Show that

$$\operatorname*{Argmin}_{x\in\mathbb{R}^n} f(x) = \partial f^*(0)$$

Exercise 3.17
Consider a primal problem of the form

$$\operatorname*{minimize}_{x\in\mathbb{R}^n} f(x) + g(x),$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are proper closed convex functions and $\operatorname{relint}\operatorname{dom} f \cap \operatorname{relint}\operatorname{dom} g \neq \emptyset$.

1. Show that solving the primal problem is equivalent to finding $x, \mu \in \mathbb{R}^n$ such that

$$\begin{cases} x \in \partial f^*(\mu), \\ x \in \partial g^*(-\mu) \end{cases}$$

2. Show that this inclusion problem is equivalent to the following dual optimality condition

$$0 \in \partial f^*(\mu) - \partial g^*(-\mu), \tag{3.1}$$

that solves the dual problem

$$\operatorname*{minimize}_{\mu\in\mathbb{R}^n} f^*(\mu) + g^*(-\mu)$$

3. Suppose you are given a solution $\mu^\star$ to the dual condition (3.1) and a subgradient selector function $s_{f^*} : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$s_{f^*}(\mu) \in \partial f^*(\mu)$$

for each $\mu \in \mathbb{R}^n$. Can you recover a primal solution $x^\star$? What if f^* is differentiable?

Exercise 3.18
Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper closed convex functions. Let $L \in \mathbb{R}^{m\times n}$. Assume that $\operatorname{relint}\operatorname{dom}(f \circ L) \cap \operatorname{relint}\operatorname{dom} g \neq \emptyset$, i.e. constraint qualification holds. Consider the primal problem of the form

$$\operatorname*{minimize}_{x\in\mathbb{R}^n} f(Lx) + g(x). \tag{3.2}$$

Derive the Fenchel dual problem

$$\operatorname*{minimize}_{\mu\in\mathbb{R}^m} f^*(\mu) + g^*(-L^T\mu). \tag{3.3}$$

Exercise 3.19

Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $L \in \mathbb{R}^{m \times n}$. Consider the primal problem of the form

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\, f(Lx) + g(x).$$

State a Fenchel dual problem and show how to recover a primal solution from a dual solution for the following particular cases:

1. $$f(y) = \frac{\lambda}{2} \|y\|_2^2$$

   for each $y \in \mathbb{R}^m$, where $\lambda > 0$ and

   $$g(x) = \sum_{i=1}^{n} x_i + \iota_{[-1,0]}(x_i)$$

   for each $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Assume that L is square (i.e. $m = n$) and invertible.

2. $$f(y) = \iota_{[-\mathbf{1},\mathbf{1}]}(y)$$

   for each $y \in \mathbb{R}^m$ and

   $$g(x) = \frac{\lambda}{2} \|x\|_2^2 - b^T x$$

   for each $x \in \mathbb{R}^n$, where $\lambda > 0$ and $b \in \mathbb{R}^n$.

Exercise 3.20 ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a proper function. Let $x, s \in \mathbb{R}^n$. Fenchel-Young's equality states that

$$f^*(x) = s^T x - f(s) \quad \text{if and only if} \quad s \in \partial f(x). \tag{3.4}$$

Prove (3.4) via the following steps:

1. Prove Fenchel-Young's inequality, i.e. $f^*(s) \geq s^T x - f(x)$
2. Suppose that $s \in \partial f(x)$. Show that $f^*(s) \leq s^T x - f(x)$

   Remark: Combining the first and second subproblems, we conclude that $s \in \partial f(x)$ implies $f^*(s) = s^T x - f(x)$
3. Suppose that $f^*(s) = s^T x - f(x)$. Show that $s \in \partial f(x)$

Remark: Combining the second and third subproblems, we conclude that (3.4) holds.

Exercise 3.21 ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Let $x, s \in \mathbb{R}^n$. Show that:

1. $s \in \partial f(x)$ implies $x \in \partial f^*(s)$
2. $x \in \partial f^*(s)$ implies $s \in \partial f^{**}(x)$
3. Suppose that f in addition is closed convex. Then

$$s \in \partial f(x) \Leftrightarrow x \in \partial f^*(s)$$

   i.e. $(\partial f)^{-1} = \partial f^*$ (the inverse of the subdifferential is the subdifferential of the conjugate)

Exercise 3.22 ($\star$)

Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ be proper closed convex, $L \in \mathbb{R}^{m \times n}$ and $c \in \mathbb{R}^m$. Define $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that

$$g(x) = f(Lx + c)$$

for each $x \in \mathbb{R}^n$. Assume that $\operatorname{relint} \operatorname{dom} g \neq \emptyset$ and that there exists an $x_s^* \in \mathbb{R}^n$ such that

$$\begin{aligned} g^*(s) &= \sup_{x \in \mathbb{R}^n} \left(s^T x - g(x) \right) \\ &= s^T x_s^* - g(x_s^*) \end{aligned}$$

for each $s \in \mathbb{R}^n$. Show that

$$g^*(s) = \inf_{\substack{\mu \in \mathbb{R}^m \\ \text{s.t. } s = L^T \mu}} \left(f^*(\mu) - c^T \mu \right)$$

for each $s \in \mathbb{R}^n$.

Exercise 3.23 ($\star$)

In this exercise we study a type of duality in a nonconvex setting called Toland duality. Let $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be two functions, where f is closed convex and $\operatorname{dom} g \subseteq \operatorname{dom} f$. Show that

$$\sup_{x \in \mathbb{R}^n} (f(x) - g(x))$$

is equal to

$$\sup_{s \in \mathbb{R}^n} (g^*(s) - f^*(s)) .$$

Hints

Hint to exercise 3.3

Note that $\frac{|\cdot|^p}{p}$ is differentiable with gradient

$$\left(\nabla \frac{|\cdot|^p}{p}\right)(x) = \begin{cases} x\,|x|^{p-2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Hint to exercise 3.6

Use the results from Exercise 3.1 and 3.5.

Hint to exercise 3.8

Cauchy-Schwarz inequality $s^T x \leq \|x\|_2 \|s\|_2$ holds for each $x, s \in \mathbb{R}^n$.

Hint to exercise 3.13

The subdifferential for each function have already been computed in previous exercises.

Hint to exercise 3.14

For the first subproblem, let $x = \mathrm{prox}_f(z)$, introduce $u = z - x$ and show that $u = \mathrm{prox}_{f^*}(z)$. To prove this, use Fermat's rule on the definition of the prox.

Chapter 4

Proximal gradient method - basics

Exercise 4.1
Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable. Consider the gradient method with constant step-size:

- Pick some initial guess $x^0 \in \mathbb{R}^n$ and step-size $\gamma > 0$.
- For $k = 0, 1, 2, \ldots$, let
$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

Let $x^\star \in \mathbb{R}^n$ be a fixed-point of the gradient method. Show that $x^\star$ is a global minimizer of f.

Exercise 4.2
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper closed convex and $\gamma > 0$. Suppose that $x \in \mathbb{R}^n$ is such that

$$x = \operatorname{prox}_{\gamma f}(x).$$

Show that x is a global minimizer of f.

Exercise 4.3
Let $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper closed convex. Assume that f is differentiable. Let $\gamma > 0$. Suppose that $x \in \mathbb{R}^n$ is such that

$$x = \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x)).$$

Show that x is a global minimizer of $f + g$.

Exercise 4.4
Which of

- the gradient method, and
- the proximal gradient method

are applicable to the minimization problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, h(x)$$

where $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is the proper closed convex function:

1. $$h(x) = \frac{1}{2}\|Ax - b\|_2^2$$

   for each $x \in \mathbb{R}^n$ where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $m < n$

2. $$h(x) = \frac{1}{2}x^TQx + b^Tx + \|x\|_1$$

   for each $x \in \mathbb{R}^n$ where $Q \in \mathbb{S}^n_{++}$

3. $$h(x) = \frac{1}{2}\|Ax - b\|_2^2 + \|x\|_2^2$$

   for each $x \in \mathbb{R}^n$ where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $m < n$

4. $$h(x) = \frac{1}{2}\|Ax - b\|_2^2 + \|x\|_2$$

   for each $x \in \mathbb{R}^n$ where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $m < n$

5. $$h(x) = \iota_{\{z\in\mathbb{R}^n : Az=b\}}(x) + \iota_{[-\mathbf{1},\mathbf{1}]}(x)$$

   for each $x \in \mathbb{R}^n$ where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $m < n$ and $\{z \in \mathbb{R}^n : Az = b\} \neq \emptyset$

6. $$h(x) = e^{\|x-y\|_2^4} + \iota_{[-\mathbf{1},\mathbf{1}]}(x)$$

   for each $x \in \mathbb{R}^n$ where $y \in \mathbb{R}^n$

7. $$h(x) = \frac{1}{2}x^TQx + \|Dx\|_1$$

   for each $x \in \mathbb{R}^n$ where $Q \in \mathbb{S}^n_{++}$ and $D \in \mathbb{R}^{n\times n}$ is diagonal

8.

$$h(x) = \frac{1}{2}x^T Qx + \iota_{[-\mathbf{1},\mathbf{1}]}(Lx)$$

for each $x \in \mathbb{R}^n$ where $Q \in \mathbb{S}^n_{++}$ and $L \in \mathbb{R}^{m\times n}$

9.

$$h(x) = \log\left(1 + e^{-w^T x}\right) + \frac{1}{2}\sum_{i=1}^{n} \max(0, x_i)^2$$

for each $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ where $w \in \mathbb{R}^n$

Exercise 4.5

For the optimization methods and objective functions in Excercise 4.4, which are applicable to some dual formulation of the minimization problem?

Exercise 4.6

Consider the problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \|x\|_1 + \frac{1}{2}x^T Qx$$

where $Q \in \mathbb{S}^n_{++}$. The goal of this exercise is to state a Fenchel dual problem and find the proximal gradient update for this dual problem. Define the functions $f, g : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x) = \|x\|_1 \quad \text{and} \quad g(x) = \tfrac{1}{2}x^T Qx$$

for each $x \in \mathbb{R}^n$. The problem can be written as

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ f(x) + g(x).$$

1. Compute f^*
2. Compute g^*
3. State a Fenchel dual problem using general f^* and g^*
4. State a proximal gradient method step for this general dual problem. Specifically, assume that f is proper closed convex and proximable, and g is proper closed and strongly convex (which in fact is true in our particular case). Construct a proximal gradient method step that is computationally reasonable based on this information.
5. Specify the proximal gradient method step for the dual problem with our particular choice of f and g

Exercise 4.7 ($\star$)
Consider a primal problem of the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, f(Lx) + g(x)$$

where $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is proper closed convex and prox friendly, $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is proper closed and strongly convex, $L \in \mathbb{R}^{m\times n}$, and $\operatorname{relint}\operatorname{dom}(f \circ L) \cap \operatorname{relint}\operatorname{dom} g \neq \emptyset$. We know that a dual problem can be written as

$$\underset{\mu\in\mathbb{R}^m}{\text{minimize}}\, f^*(\mu) + g^*\left(-L^T\mu\right).$$

We also know that f^* is proper closed convex and prox friendly and that g^* is proper closed convex and smooth. If $\gamma_k > 0$, a proximal gradient method step can be written as

$$\mu_{k+1} = \operatorname{prox}_{\gamma_k f^*}\left(\mu_k - \gamma_k\nabla\left(g^* \circ -L^T\right)(\mu_k)\right).$$

Show that this equivalently can be written as

$$\begin{cases} x_k = \operatorname{argmin}_{x\in\mathbb{R}^n}\left(g(x) + \mu_k^T Lx\right), \\ v_k = \mu_k + \gamma_k Lx_k, \\ \mu_{k+1} = v_k - \gamma_k\operatorname{prox}_{\gamma_k^{-1}f}\left(\gamma_k^{-1}v_k\right). \end{cases} \tag{4.1}$$

I.e. we can perform the proximal gradient method step for the dual problem using only primal information (f and g).

Exercise 4.8
Consider the dual problem obtained in Exercise 4.6. For this particular choice of f and g, explicitly evaluate the dual proximal gradient method step and show that the resulting step is the same as the implicit step (4.1) obtained in Exercise 4.7.

Exercise 4.9 (H) ($\star$)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a β-smooth function for some $\beta > 0$. Consider the gradient method step

$$x_{k+1} = x_k - \gamma_k\nabla f(x_k),$$

for some $\gamma_k \in (0, 1/\beta)$. Show that the gradient method is a majorization-minimization algorithm. A majorization-minimization algorithm is an algorithm on the form

$$x_{k+1} = \underset{y\in\mathbb{R}^n}{\operatorname{argmin}}\, g(y)$$

for some function $g : \mathbb{R}^n \to \mathbb{R}$ such that $f \le g$, i.e. g is a majorizer of f. Thus, the goal is to find such a g.

Hints

Hint to exercise 4.9

Start from the decent lemma, i.e.

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \|y - x\|_2^2$$

for each $x, y \in \mathbb{R}^n$ and use that $\gamma_k < 1/\beta$.

Chapter 5

Learning

Exercise 5.1
Consider the logistic regression problem

$$\underset{(w,b)\in\mathbb{R}^n\times\mathbb{R}}{\text{minimize}} \sum_{i=1}^{N} \log\left(1+e^{-y_i\left(x_i^T w+b\right)}\right) \tag{5.1}$$

with data points $x_i \in \mathbb{R}^n$ and class labels $y_i \in \{-1, 1\}$, for each $i = 1, \ldots, N$. Show that (5.1) is equivalent to

$$\underset{(w,b)\in\mathbb{R}^n\times\mathbb{R}}{\text{minimize}} \sum_{i=1}^{N} \left(\log\left(1+e^{x_i^T w+b}\right) - y_i\left(x_i^T w+b\right)\right)$$

if the classes are labeled with $y_i \in \{0, 1\}$ instead of $y_i \in \{-1, 1\}$.

Exercise 5.2
Consider the logistic regression problem

$$\underset{(w,b)\in\mathbb{R}^n\times\mathbb{R}}{\text{minimize}} \sum_{i=1}^{N} \left(\log\left(1+e^{x_i^T w+b}\right) - y_i\left(x_i^T w+b\right)\right) \tag{5.2}$$

with data points $x_i \in \mathbb{R}^n$ and class labels $y_i \in \{0, 1\}$, for each $i = 1, \ldots, N$. Assume that there exists $(\bar{w}, \bar{b}) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$\begin{cases} x_i^T \bar{w} + \bar{b} < 0 & \text{if } y_i = 0 \\ x_i^T \bar{w} + \bar{b} > 0 & \text{if } y_i = 1 \end{cases}$$

for each $i = 1, \ldots, n$. Show that the optimal value of (5.2) is 0, and that no $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ exists that attains the optimal value 0.

Exercise 5.3

Consider the univariate Lasso problem

$$\underset{x\in\mathbb{R}}{\text{minimize}}\ \frac{1}{2}\left\|ax-b\right\|_2^2+\lambda\left|x\right| \tag{5.3}$$

where $a\in\mathbb{R}^n$, $b\in\mathbb{R}^n$ and $\lambda>0$ are given.

Assume that $a\neq 0$ and $b\neq 0$, since otherwise the optimal point of (5.3) is simply $x=0$. Prove that the optimal point of (5.3) is

$$x=\begin{cases}0 & \text{if } \lambda\geq\left|a^Tb\right|,\\ x_{\text{ls}}-\dfrac{\lambda}{\left\|a\right\|_2^2}\operatorname{sgn}\left(x_{\text{ls}}\right) & \text{if } \lambda<\left|a^Tb\right|\end{cases}$$

where

$$x_{\text{ls}}=\frac{a^Tb}{\left\|a\right\|_2^2}$$

corresponds to the solution of the problem for $\lambda=0$, i.e. the corresponding univariate least squares problem.

Exercise 5.4

Consider the Lasso problem

$$\underset{x\in\mathbb{R}^m}{\text{minimize}}\ \frac{1}{2}\left\|Ax-b\right\|_2^2+\lambda\left\|x\right\|_1 \tag{5.4}$$

where $A\in\mathbb{R}^{n\times m}$, $b\in\mathbb{R}^n$ and $\lambda\geq\left\|A^Tb\right\|_\infty$. Show $x=0$ is a solution.

Exercise 5.5 (H)($\star\star$)

Consider the following bivariate Lasso problem

$$\underset{x\in\mathbb{R}^2}{\text{minimize}}\ \frac{1}{2}\left\|Ax-b\right\|_2^2+\lambda\left\|x\right\|_1 \tag{5.5}$$

where $A\in\mathbb{R}^{n\times 2}$, $b\in\mathbb{R}^n$, $n\geq 2$ an integer and $\lambda>0$. Suppose that

$$A=\begin{bmatrix}a_1 & a_2\end{bmatrix}$$

has normalized columns, i.e. $\left\|a_1\right\|_2=\left\|a_2\right\|_2=1$, and that A has full (column) rank. This implies that $\left|a_1^Ta_2\right|<1$. Consider each of the four possible sparsity patterns of $x\in\mathbb{R}^2$ in (5.5), i.e.

$$\begin{aligned}X_{0,0}&=\left\{(0,0)\in\mathbb{R}^2\right\},\\ X_{1,1}&=\left\{(x_1,x_2)\in\mathbb{R}^2 : x_1\neq 0, x_2\neq 0\right\},\\ X_{1,0}&=\left\{(x,0)\in\mathbb{R}^2 : x\neq 0\right\},\\ X_{0,1}&=\left\{(0,x)\in\mathbb{R}^2 : x\neq 0\right\}.\end{aligned}$$

Find the set

$$\Lambda_{i,j} = \{\lambda > 0 : x \text{ optimal point for (5.5) using } \lambda \text{ and } x \in X_{i,j}\}$$

for each $i, j \in \{0, 1\}$. Verify that for a given problem the four ranges $\Lambda_{i,j}$ are disjoint and the number of zeros in the solution is nondecreasing with λ.

Exercise 5.6

Consider the SVM problem with an affine model

$$\underset{(w,b)\in\mathbb{R}^m\times\mathbb{R}}{\text{minimize}} \sum_{i=1}^{n} \max\left(0, 1 - y_i\left(x_i^T w + b\right)\right) + \frac{\lambda}{2}\|w\|_2^2 \tag{5.6}$$

with data points $x_i \in \mathbb{R}^m$ and class labels $y_i = \{-1, 1\}$ for each $i = 1, \ldots, n$, and a regularization parameter $\lambda \geq 0$.

1. Consider the unregularized problem, i.e. $\lambda = 0$, and assume that examples from both classes exists. Assume the data is fully separable, i.e. there exists a non-zero pair of parameters $(w, b) \in \mathbb{R}^m \times \mathbb{R}$ such that

   $$\begin{cases} x_i^T w + b < 0 & \text{if } y_i = -1 \\ x_i^T w + b > 0 & \text{if } y_i = 1 \end{cases}$$

   for each $i = 1, \ldots, n$. Show the optimal value of (5.6) is 0 and that the that the optimal set, i.e. the set of all optimal points, is unbounded.

2. Consider again the unregularized problem, i.e. $\lambda = 0$, but assume that the data only contains one class, e.g. there exists no $i = 1, \ldots, n$ such that $y_i = -1$. Show that an arbitrary $w \in \mathbb{R}^m$ is part of an optimal point of (5.6) and show that the optimal set is unbounded.

3. Consider the regularized problem, i.e. $\lambda > 0$. Assume the data only consists of one class, e.g. there exists no $i = 1, \ldots, n$ such that $y_i = -1$. Show that $w = 0 \in \mathbb{R}^m$ is part of an optimal point of (5.6) and show that the optimal set is unbounded.

Exercise 5.7

Find $X \in \mathbb{R}^{m\times n}$ and $\phi \in \mathbb{R}^n$ such that the SVM problem (5.6) in 5.6 can be reformulated as

$$\underset{(w,b)\in\mathbb{R}^m\times\mathbb{R}}{\text{minimize}} \ \mathbf{1}^T \max\left(0, \mathbf{1} - (X^T w + b\phi)\right) + \frac{\lambda}{2}\|w\|_2^2 \tag{5.7}$$

where the $\max$ function is applied element-wise and $\mathbf{1} \in \mathbb{R}^n$ is a vector of all ones.

Exercise 5.8($\star$)

Consider the reformulated SVM problem (5.7) in Exercise 5.7, i.e.

$$\underset{(w,b)\in\mathbb{R}^m\times\mathbb{R}}{\text{minimize}} \underbrace{\mathbf{1}^T \max\left(0, \mathbf{1} - \left(X^T w + b\phi\right)\right)}_{\underbrace{=\mathbf{1}^T \max\left(0, \mathbf{1} - L\begin{bmatrix} w \\ b \end{bmatrix}\right)}_{=f(L(w,b))}} + \underbrace{\frac{\lambda}{2}\|w\|_2^2}_{=g(w,b)}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is given by

$$f(u) = \mathbf{1}^T \max\left(0, \mathbf{1} - u\right)$$

for each $u \in \mathbb{R}^n$,

$$L = \begin{bmatrix} X^T & \phi \end{bmatrix}$$

and $g : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ is given by

$$g(w,b) = \frac{\lambda}{2}\|w\|_2^2$$

for each $(w,b) \in \mathbb{R}^m \times \mathbb{R}$. Assume that $\lambda > 0$ and that examples from both classes exists.

1. Find the Fenchel dual problem

$$\underset{\mu\in\mathbb{R}^n}{\text{minimize}}\, f^*(\mu) + g^*(-L^T\mu)$$

2. Show how to recover a primal solution from a dual solution and motivate when and why this is possible
3. A support vector for this kind of *soft-margin* SVM is defined as any data point $x \in \mathbb{R}^m$ of class $y \in \{1, -1\}$ that lies on the wrong side of the margin, i.e. $1 \geq y(x^T w + b)$, for a given model with parameters $(w,b) \in \mathbb{R}^m \times \mathbb{R}$. It is easy to see that only the support vectors contribute to the cost of the objective function (see objective function (5.6) in Exercise 5.6), if we ignore the regularization term.

   Suppose that $\mu^* \in \mathbb{R}^n$ is an optimal point for the dual problem. Show that the nonzero elements of $\mu^\star \in \mathbb{R}^n$ corresponds to support vectors of the corresponding model with optimal parameters $(w^\star, b^\star) \in \mathbb{R}^m \times \mathbb{R}$. Show that the optimal model parameters can be recovered from the dual solution by only considering support vectors

Exercise 5.9

Consider the typical supervised learning problem

$$\underset{w}{\text{minimize}} \sum_{i=1}^{n} L(m_w(x_i), y_i)$$

where n is the number of training examples, $x_i \in \mathbb{R}^d$ is a data point with corresponding response variable $y_i \in \mathbb{R}^l$, for each $i = 1, \ldots, n$, $m_w : \mathbb{R}^d \to \mathbb{R}^k$ is a model parameterized by w we wish to train, and $L : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$ is the loss function comparing the model output $m_w(x_i)$ with the known correct output y_i.

Assume that $L(\cdot, y)$ is convex for each $y \in \mathbb{R}^l$. Prove or disprove the following statements:

1. The objective function

$$w \mapsto \sum_{i=1}^{n} L(m_w(x_i), y_i)$$

is convex if a linear model with some feature map is used. I.e. if

$$m_w(x) = w^T \phi(x)$$

for each $x \in \mathbb{R}^d$ where $\phi : \mathbb{R}^d \to \mathbb{R}^f$ and $w \in \mathbb{R}^{f \times k}$

2. The objective function

$$w \mapsto \sum_{i=1}^{n} L(m_w(x_i), y_i)$$

is convex if a DNN model is used. I.e. if

$$m_w(x) = \sigma_1(w_1^T \sigma_2(w_2^T ... \sigma_D(w_D^T x)...))$$

for each $x \in \mathbb{R}^d$ where σ_i is an activation functions that act elements-wise, for each $i = 1, \ldots, D$, and

$$w = (w_1, \ldots, w_D)$$

such that

$$\begin{cases} w_1 \in \mathbb{R}^{f_1 \times k} \\ w_i \in \mathbb{R}^{f_i \times f_{i-1}} \text{ for } i = 2, \ldots, D-1 \\ w_D \in \mathbb{R}^{d \times f_{D-1}} \end{cases}$$

Hints

Hint to exercise 5.5

For $x^\star \in X_{1,0}$, first find the optimal $x_1^\star$. Use this together with the optimality condition for $x_2^\star = 0$ to find the bounds on λ. For $x^\star \in X_{1,1}$, first find the ordinary least squares solution and show the coordinates of the Lasso solution have the same signs. Use this, the optimality condition and $x^\star \neq 0$ to find the bound on λ. Useful identities are $\operatorname{sgn}(x) = \operatorname{sgn}(x)^{-1}$, $|x| = \operatorname{sgn}(x)x$ and $\operatorname{sgn}(x)\operatorname{sgn}(y) = \operatorname{sgn}(xy)$

Chapter 6

Algorithm convergence

Exercise 6.1

For a given optimization problem, we used two algorithms to solve it up to a desired precision.

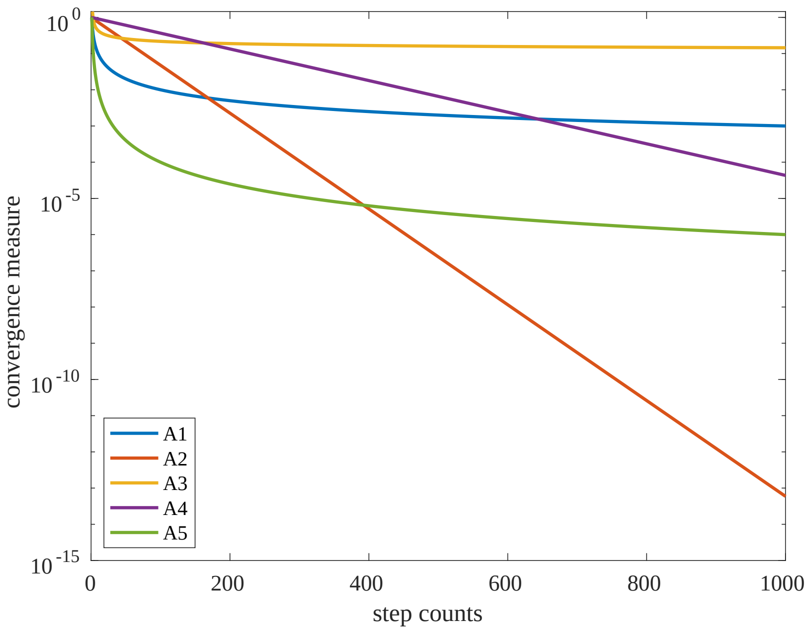
1. The first algorithm, performed 5000 floating point operations in each iteration and we ran it for 10^5 iterations
2. The second algorithm, performed 50 floating point operations in each iteration and we ran it for 2×10^6 iterations

Which algorithm had better performance?

Exercise 6.2

Match the following rates with the corresponding curve given in figure below. For each rate, specify if it is linear, sublinear or superlinear.

1. $O(\rho_1^k)$, with $0 < \rho_1 < 1$
2. $O(\rho_2^k)$, with $\rho_1 < \rho_2 < 1$
3. $O(1/\log(k))$
4. $O(1/k)$
5. $O(1/k^2)$



Exercise 6.3

Let $(V_k)_{k=0}^{\infty}$ be a nonnegative convergence measure.

1. Suppose that $(V_k)_{k=0}^{\infty}$ has a Q-linear rate, i.e. there exists a $\rho \in [0, 1)$ such that

   $$V_{k+1} \leq \rho V_k$$

   for each integer $k \geq 0$. Show that $(V_k)_{k=0}^{\infty}$ has a R-linear rate, i.e. there exists a $\rho_L \in [0, 1)$ and $C_L \geq 0$ such that

   $$V_k \leq \rho_L^k C_L$$

   for each integer $k \geq 0$

2. Suppose that $(V_k)_{k=0}^{\infty}$ has a Q-quadratic rate, i.e. there exists a $\rho \in [0, 1)$ such that

   $$V_{k+1} \leq \rho V_k^2 \tag{6.1}$$

   for each integer $k \geq 0$. Show that there exist $\rho_Q \geq 0$ and $C_Q \geq 0$ such that

   $$V_k \leq \rho_Q^{2^k} C_Q \tag{6.2}$$

   for each integer $k \geq 0$

3. Suppose that $(V_k)_{k=0}^{\infty}$ has a Q-quadratic rate as in (6.1). If $\rho_Q \in [0, 1)$ in (6.2), we say that $(V_k)_{k=0}^{\infty}$ has a R-quadratic rate and can conclude that

   $$V_k \to 0 \quad \text{as} \quad k \to \infty.$$

   However, $\rho_Q \in [0, 1)$ will only hold for certain initial values V_0 — which?

   Thus, R-quadratic rate is only achived localy, i.e. for certain initial values V_0

Exercise 6.4

Let $f : \mathbb{R}^n \to \mathbb{R}$ and consider the problem

$$\inf_{x \in \mathbb{R}^n} f(x).$$

Suppose that some iterative descent algorithm generates a sequence $(x_k)_{k=0}^{\infty}$ in $\mathbb{R}^n$, i.e.

$$f(x_{k+1}) \leq f(x_k)$$

for each integer $k \geq 0$. We call such a sequence $(x_k)_{k=0}^{\infty}$ a descent sequence for f.

1. Give an example of a function f and descent sequence $(x_k)_{k=0}^{\infty}$ for f such that the sequence of function values $(f(x_k))_{k=0}^{\infty}$ does not convergence
2. In addition, assume that the function f is bounded from below, i.e. there exists a $B \in \mathbb{R}$ such that $f(x) \geq B$ for all $x \in \mathbb{R}^n$. Prove that the sequence of function values $(f(x_k))_{k=0}^{\infty}$ converges
3. Give an example of a function f that is bounded from below and descent sequence $(x_k)_{k=0}^{\infty}$ of f such that $(f(x_k))_{k=0}^{\infty}$ does not converge to $\inf_{x \in \mathbb{R}^n} f(x)$

Exercise 6.5

Let $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = e^x - 2x + x^2$$

for each $x \in \mathbb{R}$. Consider finding a minimizer of f using the standard Newton's method without line search: Pick some initial point $x_0 \in \mathbb{R}$ and let

$$x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

for each integer $k \geq 0$. Below you find the 10 first iterations for when $x_0 = 5$.

k	x_k	$\lvert x_k - x^\star \rvert$
0	5.000000000000000	4.685076942154594
1	3.960109873126804	3.645186815281398
2	2.888130487596392	2.573207429750986
3	1.799138129515975	1.484215071670569
4	0.849076217909656	0.534153160064250
5	0.379763183818023	0.064840125972617
6	0.315791881094192	0.000868823248786
7	0.314923211324986	0.000000153479580
8	0.314923057845411	0.000000000000005
9	0.314923057845406	0.000000000000000

Calculate the ratios

$$\frac{\lvert x_{k+1} - x^\star \rvert}{\lvert x_k - x^\star \rvert}$$

and

$$\frac{\lvert x_{k+1} - x^\star \rvert}{\lvert x_k - x^\star \rvert^2}.$$

Based on these ratios, estimate whether the sequence $(|x_k - x^\star|)_{k=0}^{\infty}$ is Q-linear or Q-quadratic convergent and find the corresponding rate parameter.

Exercise 6.6

A sequence $(Q_k)_{k=0}^{\infty}$ in $\mathbb{R}$ is generated by some iterative algorithm. It is found to satisfy the following inequality

$$0 \leq Q_k \leq \frac{V}{\psi_1(k)} + \frac{D}{\psi_2(k)}$$

for each integer $k \geq 0$, where D and V are positive constants and $\psi_1, \psi_2 : \mathbb{R} \to \mathbb{R}_{++}$ are functions that depend on the algorithm that generated $(Q_k)_{k=0}^{\infty}$.

1. Show that $Q_k \to 0$ as $k \to \infty$ if

$$\begin{cases} \psi_1(k) \to \infty & \text{as} \quad k \to \infty, \\ \psi_2(k) \to \infty & \text{as} \quad k \to \infty \end{cases}$$

2. Let $c > 0$ and decide the rate of convergence for the following cases:

   (a) When

$$\psi_1(k) = \begin{cases} 1 & \text{if } k \leq 0, \\ 2c\sqrt{k} & \text{if } k > 0 \end{cases} \quad \text{and} \quad \psi_2(k) = \begin{cases} 1 & \text{if } k \leq 1, \\ \dfrac{\sqrt{k}}{c \log k} & \text{if } k > 1 \end{cases}$$

   (b) When

$$\psi_1(k) = \begin{cases} 1 & \text{if } k \leq 1, \\ \dfrac{2c(k^{1-\alpha} - 1)}{1 - \alpha} & \text{if } k > 1 \end{cases}$$

   and

$$\psi_2(k) = \begin{cases} 1 & \text{if } k \leq 1, \\ \dfrac{(1 - 2\alpha)(k^{1-\alpha} - 1)}{c(1 - \alpha)(k^{1-2\alpha} - 2\alpha)} & \text{if } k > 1 \end{cases}$$

   where $\alpha \in (0, 0.5)$

   (c) When

$$\psi_1(k) = \begin{cases} 1 & \text{if } k \leq 1, \\ \dfrac{2c(k^{1-\alpha} - 1)}{1 - \alpha} & \text{if } k > 1 \end{cases}$$

   and

$$\psi_2(k) = \begin{cases} 1 & \text{if } k \leq 1, \\ \dfrac{(1 - 2\alpha)(k^{1-\alpha} - 1)}{c(1 - \alpha)(k^{1-2\alpha} - 2\alpha)} & \text{if } k > 1 \end{cases}$$

   where $\alpha \in (0.5, 1)$

3. Which case above gives the fastest convergence rate?

Exercise 6.7

An iterative algorithm for minimizing a function $f : \mathbb{R}^n \to \mathbb{R}$ produces a sequence $(x_k)_{k=0}^{\infty}$ in $\mathbb{R}^n$. Suppose that $x^\star$ is a minimizer of f and $\gamma_i > 0$, for each integer $i \geq 0$, are the step-sizes used by the algorithm. A convergence analysis results in the following inequality:

$$f(x_k) - f(x^\star) \leq \frac{V + D\sum_{i=0}^{k} \gamma_i^2}{b\sum_{i=0}^{k} \gamma_i}$$

for each integer $k \geq 0$, where V, D and b are positive constants.

1. Show that $(f(x_k))_{k=0}^{\infty}$ converges to $f(x^\star)$ if $\left(\gamma_i^2\right)_{i=0}^{\infty}$ is summable and $(\gamma_i)_{i=0}^{\infty}$ is not, i.e. if

$$\sum_{i=0}^{\infty} \gamma_i^2 < \infty \quad \text{and} \quad \sum_{i=0}^{\infty} \gamma_i = \infty$$

2. Let $c > 0$ and estimate the convergence rates for the following step-sizes:
   (a) $\gamma_i = c/(i + 1)$ for each integer $i \geq 0$
   (b) $\gamma_i = c/(i + 1)^\alpha$ for each integer $i \geq 0$, where $\alpha \in (0.5, 1)$
3. Which step-size γ_i above gives the fastest convergence rate?

Exercise 6.8

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a β-smooth convex function for some $\beta > 0$. Let $x^\star \in \mathbb{R}^n$ be a minimizer of f. Consider finding a minimizer of f, not necessarily $x^\star$, using the gradient descent method:

$$x_{k+1} = x_k - \gamma\nabla f(x_k)$$

for each integer $k \geq 0$, where $x_0 \in \mathbb{R}^n$ is some given initial point and the step-size $\gamma \in (0, 1/\beta]$ is constant. In this case, the gradient descent method can be shown to be a descent algorithm, i.e.

$$f(x_{k+1}) \leq f(x_k)$$

for each integer $k \geq 0$. Put differently, $(x_k)_{k=0}^{\infty}$ is a descent sequence for f. Moreover, the Lyapunov inequality

$$\|x_k - x^\star\|_2^2 \leq \|x_{k-1} - x^\star\|_2^2 - 2\gamma(f(x_k) - f(x^\star)) \tag{6.3}$$

for each integer $k \geq 1$ can be shown to hold. Show that $f(x_k) \to f(x^\star)$ as $k \to \infty$ and find the convergence rate.

Exercise 6.9

Consider minimizing a function $f : \mathbb{R}^n \to \mathbb{R}$, with a minimizer $x^\star \in \mathbb{R}^n$, using a stochastic optimization algorithm and starting at some predetermined (deterministic) initial point $x_0 \in \mathbb{R}^n$. Analysis of the algorithm resulted in the inequality

$$\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 \mid x_k\right] \leq \|x_k - x^\star\|_2^2 - 2\gamma_k(f(x_k) - f(x^\star)) + \gamma_k^2 G^2$$

for each integer $k \geq 0$, where G is a positive constant and $\gamma_k > 0$ for each integer $k \geq 0$ are the deterministic step-sizes of the algorithm satisfying

$$\sum_{k=0}^{\infty} \gamma_k = \infty \quad \text{and} \quad \sum_{k=0}^{\infty} \gamma_k^2 < \infty.$$

In particular, $(x_k)_{k=0}^{\infty}$ is a stochastic process.

1. Apply an expectation to the above inequality to derive a Lyapunov inequality for the algorithm
2. Use the obtained Lyapunov inequality to show that

$$2\sum_{i=0}^{k} \gamma_i \,\mathbb{E}[f(x_i) - f(x^\star)] \leq \|x_0 - x^\star\|_2^2 + G^2 \sum_{i=0}^{k} \gamma_i^2$$

for each integer $k \geq 0$

Exercise 6.10 (H) ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a β-smooth convex function, for some $\beta > 0$. Consider finding a minimizer of f using Nesterov's accelerated gradient descent method, i.e.

$$\begin{cases} y_{k+1} = x_k - \dfrac{1}{\beta}\nabla f(x_k), \\ x_{k+1} = (1-\gamma_k)y_{k+1} + \gamma_k y_k \end{cases}$$

for each integer $k \geq 0$, for some initial points $x_0 = y_0 \in \mathbb{R}^n$, where

$$\gamma_k = \frac{1-\lambda_k}{\lambda_{k+1}}$$

and

$$\lambda_k = \begin{cases} 1 & \text{if } k = 0, \\ \dfrac{1+\sqrt{1+4\lambda_{k-1}^2}}{2} & \text{otherwise} \end{cases}$$

for each integer $k \geq 0$. Suppose that the function f has a minimum at $x^\star \in \mathbb{R}^n$. Nesterov's accelerated gradient descent method can be shown to satisfy

$$V_{k+1} - V_k \leq \frac{2\lambda_k^2}{\beta}(f(x_k) - f(x^\star)) - \frac{2\lambda_{k+1}^2}{\beta}(f(x_{k+1}) - f(x^\star)) \tag{6.4}$$

where

$$V_k = \|(\lambda_k - 1)(x_{k-1} - x_k) - x_k + x^\star\|^2$$

for each integer $k \geq 1$.

2. Show that $f(x_k) \to f(x^\star)$ as $k \to \infty$ and find the rate of convergence
3. Show that if the number of iterations k is as large or greater than

$$\max\left(\left\lceil\sqrt{\frac{C}{\epsilon}}-2\right\rceil, 2\right)$$

where

$$C = 2\beta V_1 + 4\lambda_1^2(f(x_1) - f(x^\star))$$

the methods achieves an ϵ-accurate objective value, i.e.

$$f(x_k) - f(x^\star) \leq \epsilon$$

Hints

Hint to exercise 6.10
For the first part, show that

$$\lambda_k \geq 1 + \frac{k}{2}$$

for each integer $k \geq 0$.

Chapter 7

Proximal gradient based algorithms

Exercise 7.1

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth for some $\beta > 0$. Consider the gradient method with constant step-size:

- Pick some initial guess $x_0 \in \mathbb{R}^n$ and step-size $\gamma > 0$.
- For $k = 0, 1, 2, \ldots$, let

$$x_{k+1} = x_k - \gamma \nabla f(x_k). \tag{7.1}$$

Suppose that $x^\star \in \mathbb{R}^n$ is a global minimizer of f.

1. Find the Lyapunov inequality

$$(f(x_{k+1}) - f(x^\star)) \leq (f(x_k) - f(x^\star)) - \gamma \left(1 - \frac{\beta\gamma}{2}\right) \|\nabla f(x_k)\|_2^2 \tag{7.2}$$

2. Show that

$$\|\nabla f(x_k)\|_2^2 \to 0 \quad \text{as} \quad k \to \infty$$

   if $0 < \gamma < \frac{2}{\beta}$

3. Find the convergence rate of

$$\min_{i=0,\ldots,k} \|\nabla f(x_i)\|_2^2$$

   if $0 < \gamma < \frac{2}{\beta}$

Exercise 7.2

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and β-smooth for some $\beta > 0$. Consider the gradient method with constant step-size:

- Pick some initial guess $x_0 \in \mathbb{R}^n$ and step-size $\gamma > 0$.

- For $k = 0, 1, 2, \ldots$, let

$$x_{k+1} = x_k - \gamma \nabla f(x_k). \tag{7.3}$$

Restrict the step-size to $0 < \gamma < \frac{2}{\beta}$. Suppose that $x^\star \in \mathbb{R}^n$ is a global minimizer of f.

1. Show that the iterates satisfy

$$\|x_{k+1} - x^\star\|_2^2 \le \|x_k - x^\star\|_2^2 - 2\gamma(f(x_{k+1}) - f(x^\star)) + \gamma^2(\beta\gamma - 1)\|\nabla f(x_k)\|_2^2. \tag{7.4}$$

   Do this by

   - expanding the square $\|x_{k+1} - x^\star\|_2^2$,
   - using the first order condition for convexity, and
   - using the Lyapunov inequality (7.2) from Exercise 7.1

2. Show that

$$f(x_k) \to f(x^\star) \quad \text{as} \quad k \to \infty$$

   and find the rate of convergence. Note that $\sum_{i=0}^{\infty} \|\nabla f(x_k)\|_2^2$ was shown to be bounded in Exercise 7.1.

Exercise 7.3

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is σ-strongly convex and β-smooth for some $\beta \ge \sigma > 0$. Consider the gradient method with constant step-size:

- Pick some initial guess $x_0 \in \mathbb{R}^n$ and step-size $\gamma > 0$.
- For $k = 0, 1, 2, \ldots$, let

$$x_{k+1} = x_k - \gamma \nabla f(x_k).$$

Suppose that $x^\star \in \mathbb{R}^n$ is the global minimizer of f.

1. Suppose that $\gamma \in (0, 1/\beta]$. Show that the iterates satisfy the inequality

$$\|x_{k+1} - x^\star\|_2^2 \le (1 - \sigma\gamma)\|x_k - x^\star\|_2^2.$$

   Use the same technique as in Exercise 7.2.1, but replace the first order condition for convexity with the first order condition for strong convexity.

   Which step-size γ gives the fastest convergence rate?

2. In the lectures a different approach is used to analyze the convergence. There it is shown that

$$\|x_{k+1} - x^\star\|_2 \le \max(1 - \sigma\gamma, \beta\gamma - 1)\|x_k - x^\star\|_2$$

   holds if $\gamma \in (0, 2/\beta)$. What is the best step-size γ according to this inequality?

3. Which approach gives the faster convergence rate?

Exercise 7.4

Consider the minimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \frac{1}{2} x^T Q x + q^T x$$

where $Q \in \mathbb{S}^n_{++}$ and $q \in \mathbb{R}^n$. We use the gradient method with constant step-size:

- Pick some initial guess $x_0 \in \mathbb{R}^n$ and step-size $\gamma > 0$.
- For $k = 0, 1, 2, \ldots$, let

$$x_{k+1} = x_k - \gamma \nabla f(x_k).$$

Suppose that $x^\star \in \mathbb{R}^n$ is the global minimizer of f. Moreover, let $\gamma \in (0, 2/\beta)$ where

$$\beta = \|Q\|_2 .$$

1. Show that

$$\|x_{k+1} - x^*\|_2 \le \|I - \gamma Q\|_2 \|x_k - x^*\|_2$$

   and that

$$\|I - \gamma Q\|_2 < 1$$

2. Let $\gamma = 1/\beta$ and find an expression of

$$\|I - \gamma Q\|_2$$

   in terms of the eigenvalues of Q

Let the linear convergence rate $\rho \in [0, 1)$ be defined as the smallest ρ so that

$$\|x_k - x^*\| \le \rho^k \|x_0 - x^*\|$$

holds for each integer $k \ge 0$.

3. Let $\gamma = 1/\beta$ and let

$$Q = \begin{bmatrix} \epsilon & 0 \\ 0 & 1 \end{bmatrix}$$

   where $0 < \epsilon < 1$. What is the worst case linear convergence rate ρ we can expect given the result above?

   Let $q = 0$. Can you find an initial point x_0 that achives this worst case convergence rate?

4. Let

$$Q = \begin{bmatrix} \epsilon & \frac{\epsilon}{10} \\ \frac{\epsilon}{10} & 1 \end{bmatrix}.$$

   where $0 < \epsilon < 1$ and assume that ϵ is much smallar than 1. The eigenvalues of this matrix are approximately 1 and ϵ. Gradient method will therefore be slow on this problem also.

To improve the convergence rate, we want to find a variable change $y = Ax$ for some invertible matrix $A \in \mathbb{R}^{n\times n}$ so that the equivalent problem

$$\underset{y\in\mathbb{R}^n}{\text{minimize}}\, \frac{1}{2} y^T A^T Q A y + q^T A y$$

has better properties. This is often called *preconditioning*. Find a diagonal matrix V so that the diagonal elements in $V^T Q V$ are 1.

5. What are the eigenvalues of the new matrix $V^T Q V$? What can we expect in terms of convergence rate of $\|y_k - y^*\|$?

6. When we have a problem where the proximal gradient method is needed instead of just gradient descent, why do we usually have to limit ourselves to diagonal scalings V?

Exercise 7.5

Let $f : \mathbb{R}^n \to \mathbb{R}$ be closed and convex. Consider the poximal point method:

- Pick some initial guess $x_0 \in \mathbb{R}^n$ and $\gamma > 0$.
- For $k = 0, 1, 2, \ldots$, let

$$x_{k+1} = \text{prox}_{\gamma f}(x_k).$$

1. Show that $(f(x_k))_{k=0}^{\infty}$ is a nonincreasing sequence by showing that

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2\gamma} \|x_{k+1} - x_k\|_2^2$$

2. Assume that f is lower bounded by $B \in \mathbb{R}$, i.e.

$$f(x) \geq B$$

for each $x \in \mathbb{R}^n$. Show that

$$\|x_{k+1} - x_k\|_2^2 \to 0 \quad \text{as} \quad k \to \infty.$$

3. Show that

$$\|x_{k+1} - x_k\|_2^2 \to 0 \quad \text{as} \quad k \to \infty.$$

implies that

$$\text{dist}_{\partial f(x_k)}(0) \to 0 \quad \text{as} \quad k \to \infty.$$

where

$$\text{dist}_{\partial f(x)}(y) = \inf_{s \in \partial f(x)} \|s - y\|_2$$

for each $x, y \in \mathbb{R}^n$. I.e. show that if the reisdual convergence to zero, then the distance between the subdifferential and zero convergence to zero.

4. In addition, assume that f is σ-strongly convex for some $\sigma > 0$. Let

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}^n} f(x).$$

Show that

$$x_k \to x^* \quad \text{as} \quad k \to \infty.$$

Remark: A note about the last point. There exist conditions weaker than strong convexity so that the sequence to converges to an optimal point, but strong convexity is arguably the simplest.

Exercise 7.6

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and β-smooth for some $\beta > 0$. Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper, closed and convex. Consider the proximal gradient method:

- Pick some initial guess $x_0 \in \mathbb{R}^n$ and $\gamma > 0$.
- For $k = 0, 1, 2, \ldots$, let

$$x_{k+1} = \operatorname{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k)).$$

Here we restict the step-size such that $\gamma \in (0, 1/\beta]$. Suppose that

$$x^* \in \operatorname*{Argmin}_{x \in \mathbb{R}^n} f(x) + g(x).$$

A procedure for proving convergence in function value of the method is given below. However, some of the steps are missing. Fill in the gaps marked by $\ldots$ to complete the procedure.

1. The goal is to get a Lyapunov inequality on the form

$$V_{k+1} \leq V_k - Q_k$$

for each integer $k \geq 0$, where $(Q_k)_{k=0}^{\infty}$ is some nonnegative convergence measure and

$$V_k = \|x_k - x^\star\|_2^2$$

for each integer $k \geq 0$. We further define the residual mapping $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\mathcal{R}x = x - \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x))$$

for each $x \in \mathbb{R}^n$. The proximal gradient update can then be written as

$$x_{k+1} = x_k - \mathcal{R}x_k. \tag{7.5}$$

We can use this to relate V_{k+1} to V_k by

$$V_{k+1} = V_k + \ldots \tag{7.6}$$

2. Next, we wish to upper bound the quantity $-2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2$. We start by using (7.5) to rewrite it as

$$-2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2 = -2(x_{k+1} - x^\star)^T(\mathcal{R}x_k) + \dots \tag{7.7}$$

3. We now turn to bounding $-2(x_{k+1} - x^\star)^T(\mathcal{R}x_k)$. Using Fermat's rule on the proximal gradient update gives that

$$0 \in \partial g(x_{k+1}) + \frac{1}{\gamma}\left(x_{k+1} - (x_k - \gamma\nabla f(x_k))\right)$$

which is equivalent to that

$$\gamma^{-1}\mathcal{R}x_k - \nabla f(x_k) \in \partial g(x_{k+1})$$

The definition of a subgradient then gives that

$$g(x^\star) \geq g(x_{k+1}) + (\gamma^{-1}\mathcal{R}x_k - \nabla f(x_k))^T(x^\star - x_{k+1})$$

which implies that

$$-2(x_{k+1} - x^\star)^T(\mathcal{R}x_k) \leq \dots \tag{7.8}$$

4. We continue to bound $-2\gamma\nabla f(x_k)^T(x_{k+1} - x^\star)$. Using the definition of β-smoothness of f and the first-order condition of convexity on f gives the two following inequalities:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{\beta}{2}\|x_{k+1} - x_k\|_2^2 \\ &= f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{\beta}{2}\|\mathcal{R}x_k\|_2^2 \\ f(x^\star) &\geq f(x_k) + \nabla f(x_k)^T(x^\star - x_k). \end{aligned}$$

Adding these two together and rearranging gives that

$$f(x_{k+1}) \leq f(x^\star) + \nabla f(x_k)^T(x_{k+1} - x^\star) + \frac{\beta}{2}\|\mathcal{R}x_k\|_2^2$$

which implies that

$$-2\gamma\nabla f(x_k)^T(x_{k+1} - x^\star) \leq \dots \tag{7.9}$$

5. Inserting (7.9) into (7.8), (7.8) into (7.7), and (7.7) into (7.6) gives that

$$V_{k+1} \leq V_k + \dots$$

6. Using the assumption $\gamma < \beta^{-1}$ gives that

$$V_{k+1} = V_k - Q_k$$

where

$$Q_k = \dots$$

which is nonnegative since $\gamma > 0$ and $\dots \geq \dots$ by assumption on $x^\star$.

7. Since $V_k \geq 0$ and $Q_k \geq 0$ we we know that

$$Q_k \to 0 \quad \text{as} \quad k \to \infty$$

which implies that

$$\ldots \to \ldots \quad \text{as} \quad k \to \infty.$$

Exercise 7.7

Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x) + g(x) \tag{7.10}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is σ_f-strongly convex and β-smooth, and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is proper, closed and σ_g-strongly convex, for some $0 \leq \sigma_f < \beta$ and $\sigma_g \geq 0$. The problem can then be solved using the proximal gradient method:

- Pick some initial guess $x_0 \in \mathbb{R}^n$ and $\gamma > 0$.
- For $k = 0, 1, 2, \ldots$, let

$$x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k)).$$

Let

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}^n} f(x) + g(x).$$

1. Show that the proximal gradient method satisfy

$$\|x_{k+1} - x^\star\|_2^2 \leq \frac{\max(1 - \sigma_f \gamma, \beta\gamma - 1)^2}{1 + \sigma_g \gamma} \|x_k - x^\star\|_2^2$$

by inserting the definition of x_{k+1} in $\|x_{k+1} - x^\star\|_2^2$ and then use the following:

- The minimum $x^\star$ is a fixed point to the proximal gradient step.
- The proximal operator of a σ-strongly convex function is $\frac{1}{1+\sigma\gamma}$-Lipschitz continuous.
- The gradient of f satisfies

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{\beta + \sigma_f} \|\nabla f(x) - \nabla f(y)\|_2^2 + \frac{\beta \sigma_f}{\beta + \sigma_f} \|x - y\|_2^2$$

for each $x, y \in \mathbb{R}^n$, since it is σ_f-strongly convex and β-smooth.

- Then, in two different cases, use that
  - ∇f is β-Lipschitz continuous, and that
  - the inequality

$$\|\nabla f(x) - \nabla f(y)\|_2 \geq \sigma_f \|x - y\|_2$$

holds for each $x, y \in \mathbb{R}^n$, since f is differentiable and σ_f-strongly convex

2. For which step-sizes γ and combinations of $\sigma_f \geq 0$ and $\sigma_g \geq 0$ does our analysis give that the proximal gradient method converge linearly?
3. Note that it is possible to "move" the strong convexity between f and g in some sense. In particular, consider the following problem

$$\min_{x \in \mathbb{R}^n} h(x) + \phi(x) + \frac{\sigma}{2} \|x\|_2^2$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and convex, $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is proper, closed and convex, and $\sigma, L > 0$. This can be written as a problem of the form (7.10) by choosing any $\delta \in [0, 1]$ and forming

$$f = h + \delta \frac{\sigma}{2} \|\cdot\|_2^2 \quad \text{and} \quad g = \phi + (1 - \delta) \frac{\sigma}{2} \|\cdot\|_2^2 .$$

The objective function $f + g$ will always be the same and will remain σ-strongly convex, regardless of the choice of δ. However, the individual strong convexity of f and g, and the smoothness of f, will depend on δ. Therefore, the same holds for the linear convergence rate of the proximal gradient method that we can prove.

Compare the convergence rates for the best choice of step-size γ when all strong convexity is put in the gradient step, i.e. $\delta = 1$, and when all is put in the proximal operator, i.e. $\delta = 0$.

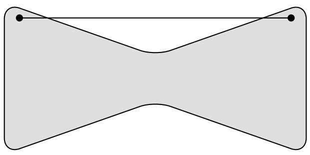
Hints

Solutions to chapter 1

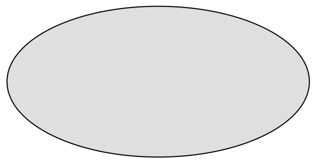
Solution 1.1

1. Figures b. and d. represent convex sets since the straight line connecting any two points with the sets are contained within the sets.

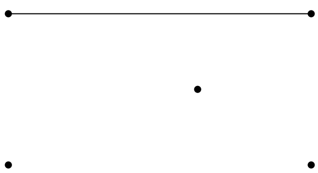
   Figures a. and c. represent nonconvex sets since the lines drawn below between two points in the respective sets are partially outside the sets.



a.



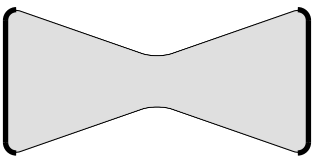
b.



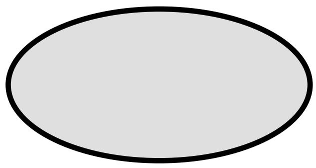
c.

d.

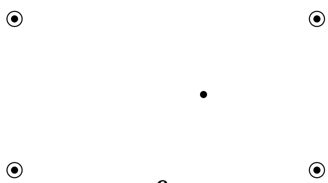
2. Figures b. and d. are convex so there exist supporting hyperplanes at the entire boundary.



a.



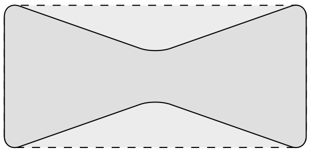
b.



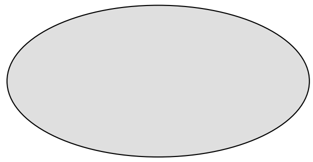
c.

d.

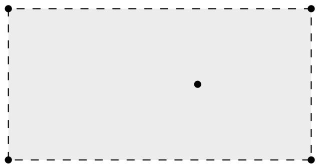
3. Figures b. and d. are convex so the convex hull is the set itself.



a.



b.



c.

d.

Solution 1.2

1. Let $x, y \in S$ and $\theta \in [0, 1]$. Then $Ax = b$ and $Ay = b$. Therefore,

$$A(\theta x + (1-\theta)y) = \theta Ax + (1-\theta)Ay = \theta b + (1-\theta)b = b$$

and we conclude that

$$\theta x + (1-\theta)y \in S.$$

Since, $x, y \in S$ and $\theta \in [0, 1]$ are arbitrary, the set S is convex. (This is an affine subspace/intersection of hyperplanes.)

2. Let $x, y \in S$ and $\theta \in [0, 1]$. Then $Ax \le b$ and $Ay \le b$. Since θ and $(1-\theta)$ are nonnegative, we have that

$$A(\theta x + (1-\theta)y) = \theta Ax + (1-\theta)Ay \le \theta b + (1-\theta)b = b$$

and we conclude that

$$\theta x + (1-\theta)y \in S.$$

Hence, the set S is convex. (This is a polytope /intersection of halfspaces.)

3. Let $x, y \in S$ and $\theta \in [0, 1]$. Then $x \ge 0$ and $y \ge 0$. Therefore, since θ and $(1-\theta)$ are nonnegative,

$$\theta x + (1-\theta)y \ge 0$$

and we conclude that

$$\theta x + (1-\theta)y \in S.$$

Hence, the set S is convex. (This is the non-negative orthant.)

4. Let $x, y \in S$ and $\theta \in [0,1]$. Then $l \leq x \leq u$ and $l \leq y \leq u$. Since θ and $(1-\theta)$ are nonnegative, we have that

$$\theta x + (1-\theta)y \leq \theta u + (1-\theta)u = u$$

and

$$\theta x + (1-\theta)y \geq \theta l + (1-\theta)l = l.$$

In particular,

$$l \leq \theta x + (1-\theta)y \leq u$$

and we conclude that

$$\theta x + (1-\theta)y \in S.$$

Hence, the set S is convex. (The constraints that defines the set are called box-constraints.)

5. Let $x, y \in S$ and $\theta \in [0,1]$. Then $\|x\|_2 \leq 1$ and $\|x\|_2 \leq 1$. Since θ and $(1-\theta)$ are nonnegative, we have that

$$\begin{aligned}
\|\theta x + (1-\theta)y\|_2 &\leq \|\theta x\|_2 + \|(1-\theta)y\|_2 \\
&= \theta \|x\|_2 + (1-\theta)\|y\|_2 \\
&\leq \theta + (1-\theta) \\
&= 1
\end{aligned}$$

and we conclude that

$$\theta x + (1-\theta)y \in S.$$

Hence, the set S is convex. (This is the unit 2-norm ball, i.e. all points with distance to the origin less than one.)

6. The set S is not convex. We prove this by finding a counter example to the definition of convexity. Let $x = e_1$ and $y = -e_1$. Then $-\|x\|_2 \leq -1$ and $-\|y\|_2 \leq -1$. In particular, $x, y \in S$. However, for the convex combination $(1/2)x + (1/2)y$ we have that

$$\left\| \frac{1}{2}x + \frac{1}{2}y \right\| = 0$$

and therefore

$$\frac{1}{2}x + \frac{1}{2}y \notin S.$$

This show that $(1/2)x + (1/2)y$ is a counter example to the definition of convexity, and therefore, we conclude that the set S is not convex, as desired.

7. The condition $-\|x\|_2 \leq 1$ holds for each $x \in \mathbb{R}^n$. Hence $S = \mathbb{R}^n$, which is convex.

8. Let $(x, t_x), (y, t_y) \in S$ and $\theta \in [0,1]$. Then $\|x\|_2 \leq t_x$ and $\|y\|_2 \leq t_y$. Since θ and $(1-\theta)$ are nonnegative, we have that

$$\begin{aligned}
\|\theta x + (1-\theta)y\|_2 &\leq \theta\|x\|_2 + (1-\theta)\|y\|_2 \\
&\leq \theta t_x + (1-\theta)t_y
\end{aligned}$$

and therefore

$$(\theta x + (1-\theta)y, \theta t_x + (1-\theta)t_y) \in S.$$

However,

$$\theta(x, t_x) + (1-\theta)(y, t_y) = (\theta x + (1-\theta)y, \theta t_x + (1-\theta)t_y)$$

and we conclude that

$$\theta(x, t_x) + (1-\theta)(y, t_y) \in S.$$

Hence, the set S is convex. (This set is called a second-order cone or Lorentz cone and is shaped like an ice cream cone.)

9. Let $X, Y \in S$ and $\theta \in [0,1]$. Note that $\theta X + (1-\theta)Y$ is symmetric since X and Y are. Also, $x^T X x \geq 0$ and $x^T Y x \geq 0$, for each $x \in \mathbb{R}^n$. Since θ and $(1-\theta)$ are nonnegative, we have that

$$x^T(\theta X + (1-\theta)Y)x = \theta x^T X x + (1-\theta)x^T Y x \geq 0$$

for each $x \in \mathbb{R}^n$, and therefore

$$\theta X + (1-\theta)Y \succeq 0$$

or

$$\theta X + (1-\theta)Y \in S.$$

Hence, the set S is convex.

10. Note that $S = \{a\}$, i.e. a singleton. Let $x, y \in S$ and $\theta \in [0,1]$. Then $x = a$, $y = a$. Note that

$$\theta x + (1-\theta)y = a$$

and therefore

$$\theta x + (1-\theta)y \in S.$$

Hence, the set S is convex. (In particular, all singletons are convex.)

11. Note that $S = \{a, b\}$. The set S is not convex. We prove this by finding a counter example to the definition of convexity. Let $x = a$ and $y = b$. Since $a \neq b$, there exists an index $i = 1, \dots, n$ such that $a_i \neq b_i$. Suppose without loss of generality that $a_i < b_i$. Create the convex combination

$$z = \frac{1}{2}x + \frac{1}{2}y.$$

Then $a_i < z_i < b_i$. Thus, $z \neq a$ and $z \neq b$. In particular,

$$z \notin S.$$

This show that z is a counter example to the definition of convexity, and therefore, we conclude that the set S is not convex, as desired.

Solution 1.3

1. Note that $V = \{a\}$, i.e. a singleton. The set V is affine. Let $x, y \in V$. Then $x = y = a$ and

$$\alpha x + (1-\alpha)y = a \in V$$

for each $\alpha \in \mathbb{R}$. Therefore, the set V is affine. (In particular, all singletons are affine.)

2. The set V is not affine. We prove this by finding a counter example to the definition of affine set. Note that $a, b \in V$. Since by assumption $a \neq b$, there exists an index $i = 1, \dots, n$ such that $a_i \neq b_i$. Suppose without loss of generality that $a_i < b_i$. But then

$$x_i \leq b_i$$

for each $x \in V$. Create the affine combination

$$z = (-1)a + (1-(-1))b = -a + 2b.$$

But it holds that $b_i < -a_i + 2b_i = z_i$. In particular, we must have that

$$z \notin V.$$

This show that z is a counter example to the definition of affine set, and therefore, we conclude that the set V is not an affine set, as desired.

3. The set V is affine. Let $x, y \in V$. But then there exists $\alpha_1, \alpha_2 \in \mathbb{R}$ such that

$$\begin{aligned} x &= \alpha_1 a + (1-\alpha_1)b \\ y &= \alpha_2 a + (1-\alpha_2)b. \end{aligned}$$

Note that

$$\begin{aligned} \alpha x + (1-\alpha)y &= (\alpha\alpha_1 + (1-\alpha)\alpha_2)a + (\alpha(1-\alpha_1) + (1-\alpha)(1-\alpha_2))b \\ &= (\alpha\alpha_1 + (1-\alpha)\alpha_2)a + (1-(\alpha\alpha_1 + (1-\alpha)\alpha_2))b \\ &= \beta_\alpha a + (1-\beta_\alpha)b \in V \end{aligned}$$

where $\beta_\alpha = \alpha\alpha_1 + (1-\alpha)\alpha_2$, for each $\alpha \in \mathbb{R}$. Thus, V is an affine set.

Solution 1.4

Figures (a), (b), and (d) are cones. Figures (a), (b) and (c) are convex.

Solution 1.5

All the sets in this exercises are in Exercise 1.2 and were shown to be convex. It remains to decide which of them are cones.

1. Let $x \in S$, i.e., $Ax = 0$. Then $A(\alpha x) = \alpha Ax = 0$ for each $\alpha \geq 0$. Hence, $\alpha x \in S$ for each $\alpha \geq 0$ and S is a cone.

2. Let $x \in S$, i.e., $Ax = b \neq 0$. Then $A(\alpha x) = \alpha Ax = \alpha b \neq b$ for each $\alpha \neq 1$ (unless $b = 0$), and therefore $\alpha x \notin S$. Hence S is not a cone.

3. Let $x \in S$, i.e., $Ax \leq 0$. Then $A(\alpha x) = \alpha Ax \leq 0$ for each $\alpha \geq 0$. Hence $\alpha x \in S$ for each $\alpha \geq 0$ and S is a cone.

4. The inequality $Ax \leq b$ consists of m scalar inequalities $a_i^T x \leq b_i$ that all must hold. Here, a_i is the ith row of the matrix A and b_i is the ith element of the vector b. Let $x \in S$ be such that $a_j^T x = b_j$ (such x always exists by assumption on j). Now, $a_j^T(\alpha x) = \alpha a_j^T x = \alpha b_j$ for each $\alpha \geq 0$.

   If $b_j > 0$ and $\alpha > 1$, then $a_j^T(\alpha x) = \alpha b_j > b_j$ and $\alpha x \notin S$.

   If $b_j < 0$ and $\alpha \in [0, 1)$, then $a_j^T(\alpha x) = \alpha b_j > b_j$ and $\alpha x \notin S$.

   Hence, S is not a cone.

5. Let $x \in S$, i.e., $x \geq 0$. Then $\alpha x \geq 0$ for each $\alpha \geq 0$. Hence, $\alpha x \in S$ for each $\alpha \geq 0$ and S is a cone.

6. Let $(x, t) \in S$, i.e., $\|x\|_2 \leq t$. Then $\|\alpha x\|_2 = \alpha \|x\|_2 \leq \alpha t$ for each $\alpha \geq 0$. Hence $(\alpha x, \alpha t) \in S$ for each $\alpha \geq 0$ and S is a cone.

7. Let $X \in S$, i.e., X is symmetric and $x^T X x \geq 0$ holds for each $x \in \mathbb{R}^n$. Scaling X by α does not destroy symmetry. Also, $x^T(\alpha X)x = \alpha x^T X x \geq 0$ for each $\alpha \geq 0$ and for each $x \in \mathbb{R}^n$. Hence, $\alpha X \in S$ for each $\alpha \geq 0$ and S is a cone.

Solution 1.6

1. Intersection. Take $x, y \in C$. Then $x, y \in C_1$ and $x, y \in C_2$. Therefore, by convexity of C_1 and C_2, we have for each $\theta \in [0, 1]$ that $\theta x + (1-\theta)y \in C_1$ and $\theta x + (1-\theta)y \in C_2$. Hence, $\theta x + (1 - \theta)y \in C$, which shows that C is convex.

2. Union. Take $C_1 = \{0\}$ and $C_2 = \{e_1\}$. Then $C = \{0, e_1\}$. This is not convex since, e.g., $0.5e_1 \notin C$.

Solution 1.7

Let $x, y \in \bigcap_{j \in J} C_j$ and let $\theta \in [0, 1]$. Then

$$\theta x + (1 - \theta)y \in C_j$$

by convexity of C_j, for each $j \in J$. Therefore,

$$\theta x + (1 - \theta)y \in \bigcap_{j \in J} C_j.$$

We conclude that the set $\bigcap_{j \in J} C_j$ is convex.

Solution 1.8

1. Let $x, y \in h_{s,r}$ and let $\theta \in [0,1]$. Note that

$$s^T(\theta x + (1-\theta)y) = \theta s^T x + (1-\theta)s^T y = \theta r + (1-\theta)r = r.$$

   Therefore, $\theta x + (1-\theta)y \in h_{s,r}$. We conclude that $h_{s,r}$ is convex.

2. Let $x, y \in H_{s,r}$ and let $\theta \in [0,1]$. Note that

$$s^T(\theta x + (1-\theta)y) = \theta s^T x + (1-\theta)s^T y \leq \theta r + (1-\theta)r = r$$

   Therefore, $\theta x + (1-\theta)y \in H_{s,r}$. We conclude that $H_{s,r}$ is convex.

3. Note that the set C can be written as an intersection of affine hyperplanes and halfspaces:

$$C = \left(\bigcap_{i \in \{1,\ldots,m\}} h_{s_i,r_i} \right) \bigcap \left(\bigcap_{i \in \{m+1,\ldots,p\}} H_{s_i,r_i} \right).$$

   In particular, we see that the set C is given by an intersection of convex sets, and is therefore itself convex.

Solution 1.9

All of the sets are polytopes and therefore convex.

Solution 1.10

1. Let $y_1, y_2 \in f(C) = \{Ax + b : x \in C\}$ and let $\theta \in [0,1]$. There exists $x_1, x_2 \in C$ such that

$$y_1 = Ax_1 + b \quad \text{and} \quad y_2 = Ax_2 + b.$$

   We have $\theta x_1 + (1-\theta)x_2 \in C$ since C is convex. Note that

$$\theta y_1 + (1-\theta)y_2 = A(\theta x_1 + (1-\theta)x_2) + b \in f(C).$$

   We conclude that $f(C)$ is convex.

2. Let $x_1, x_2 \in f^{-1}(D) = \{x : Ax + b \in D\}$ and let $\theta \in [0,1]$. We know that

$$Ax_1 + b \in D \quad \text{and} \quad Ax_2 + b \in D.$$

   By convexity of D we get that

$$\theta(Ax_1 + b) + (1-\theta)(Ax_2 + b) = A(\theta x_1 + (1-\theta)x_2) + b \in D.$$

   In particular, we note that $\theta x_1 + (1-\theta)x_2 \in f^{-1}(D)$. We conclude that $f^{-1}(D)$ is convex.

Solution 1.11

Let $x, y \in \mathrm{dom} f$ and $\theta \in [0,1]$. Then, by definition of convexity of f, we have that

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) < \infty.$$

This implies that

$$\theta x + (1-\theta)y \in \mathrm{dom} f.$$

We conclude that $\mathrm{dom} f$ is convex.

Solution 1.12

1. The function is convex. We need to prove that

$$\iota_C(\theta x + (1-\theta)y) \leq \theta \iota_C(x) + (1-\theta)\iota_C(y) \tag{7.11}$$

for each $x, y \in \mathbb{R}^n$ and for each $\theta \in [0,1]$. Moreover, recall that we are using arithmetics in the extended real numbers. In particular, we use the convention that

$$\begin{aligned} 0 \cdot \infty &= 0 \\ a \cdot \infty &= \infty \quad \text{for each} \quad a > 0 \\ a + \infty = \infty + a &= \infty \quad \text{for each} \quad a \in \mathbb{R} \\ \infty + \infty &= \infty \\ a &\leq \infty \quad \text{for each} \quad a \in \mathbb{R} \cup \{\infty\} \end{aligned}$$

- Suppose that $x, y \in C$. Then the lefthand side of (7.11) is 0 since $\theta x + (1-\theta)y \in C$ by convexity of C, and the righthand side of (7.11) is 0 since $\theta 0 + (1-\theta)0 = 0$. Thus, (7.11) holds in this case.
- Suppose that $x \notin C$ or $y \notin C$. If $\theta \in (0,1)$ then both θ and $1-\theta$ are positive, and the righthand side is ∞, which is always greater or equal to the lefthand side. Thus, (7.11) holds in this case. If $\theta \in \{0,1\}$ then at least of one of θ and $1-\theta$ is positive, and the righthand side is ∞, which is always greater or equal to the lefthand side. Thus, (7.11) holds in this case.

This covers all cases. Therefore, (7.11) always holds, and we conclude that the function f is a convex function.

2. The function is convex. Note that

$$\begin{aligned} f(\theta x + (1-\theta)y) &= \|\theta x + (1-\theta)y\| \\ &\leq \|\theta x\| + \|(1-\theta)y\| \\ &\leq \theta \|x\| + (1-\theta)\|y\| \\ &= \theta f(x) + (1-\theta) f(y) \end{aligned}$$

for each $x, y \in \mathbb{R}^n$ and for each $\theta \in [0,1]$. Therefore, f is a convex function.

3. The function is not convex. We will find $x, y \in \mathbb{R}^n$ and $\theta \in [0,1]$ such that

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) \tag{7.12}$$

fails. Indeed, pick $x = -y \neq 0$ and $\theta = 1/2$. Then

$$f(\theta x + (1-\theta)y) = -\|0\| = 0$$

and

$$\theta f(x) + (1-\theta) f(y) = -\frac{1}{2}\|x\| - \frac{1}{2}\|-x\| = -\|x\| < 0.$$

This example violates (7.12). Therefore, f is not a convex function.

4. The function is not convex. The function f is twice differentiable with Hessian

$$\nabla^2 f(x,y) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

for each $(x,y) \in \mathbb{R}^2$. Note that the Hessian is not positive semidefinite (it is symmetric but has eigenvalues 1 and -1). Therefore, by the second-order condition for convexity, we conclude that f is not a convex function.

5. The function convex. Note that

$$\begin{aligned} f(\theta x + (1-\theta)y) &= a^T(\theta x + (1-\theta)y) + b \\ &= \theta(a^T x + b) + (1-\theta)(a^T y + b) \\ &= \theta f(x) + (1-\theta) f(y) \end{aligned}$$

for each $x, y \in \mathbb{R}^n$ and for each $\theta \in [0,1]$. Therefore, the convexity definition holds with equality, and we conclude that f is a convex function.

6. The function is convex. Indeed, the function f is twice differentiable with Hessian $\nabla^2 f(x) = Q \succeq 0$ for each $x \in \mathbb{R}^n$. Therefore, by the second-order condition for convexity, we conclude that f is a convex function.

7. The function is convex. Note that

$$(x,y) \mapsto \iota_C(y)$$

is convex by Exercise 1.12.1 (and by Exercise 1.19) and that

$$(x,y) \mapsto \|x - y\|$$

is convex by Exercise 1.12.2 and the composition rule with a linear mapping. Therefore,

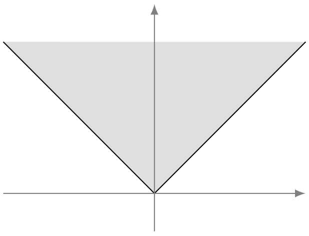
$$(x,y) \mapsto \|x-y\| + \iota_C(y) = h(x,y)$$

is a convex function since it is the sum of convex functions. Note that

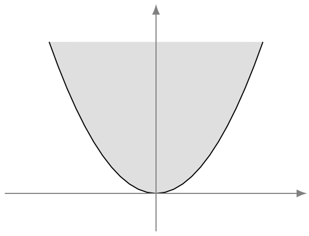
$$f(x) = \inf_{y \in C} h(x,y) = \inf_{y \in C} \|x - y\|$$

is convex by the convexity under partial minimization rule, establishing the desired result.

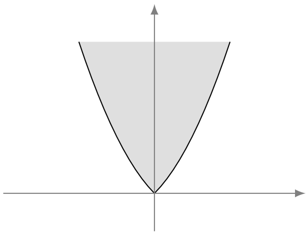
Solution 1.13



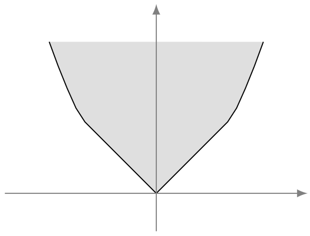
$f(x) = |x|$



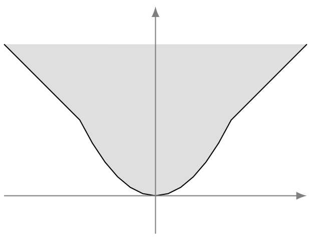
$f(x) = x^2$



$f(x) = |x| + x^2$



$f(x) = \max(|x|, x^2)$



$f(x) = \min(|x|, x^2)$

Solution 1.14

The epigraph of f is

$$\begin{aligned}\text{epi}f &= \{(x,r) \in \mathbb{R}^n \times \mathbb{R} : f(x) \le r\}\\ &= \{(x,r) \in \mathbb{R}^n \times \mathbb{R} : a^T x + b \le r\}\\ &= \{(x,r) \in \mathbb{R}^n \times \mathbb{R} : [a^T, -1]\begin{bmatrix} x \\ r \end{bmatrix} \le -b\}\end{aligned}$$

which is a halfspace in $\mathbb{R}^{n+1}$.

Solution 1.15

Suppose that f is convex. Let $(x_1, r_1), (x_2, r_2) \in \text{epi}f$ and let $\theta \in [0,1]$. By convexity of f, we get that

$$\begin{aligned}f(\theta x_1 + (1-\theta)x_2) &\le \theta f(x_1) + (1-\theta) f(x_2)\\ &\le \theta r_1 + (1-\theta) r_2\end{aligned}$$

since θ and $1-\theta$ are nonnegative. This implies that

$$\theta(x_1, r_1) + (1-\theta)(x_2, r_2) = (\theta x_1 + (1-\theta)x_2, \theta r_1 + (1-\theta)r_2) \in \text{epi}f.$$

Thus, $\text{epi}f$ is convex.

Conversely, suppose that $\text{epi}f$ is convex. The condition defining convexity is that

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2) \tag{7.13}$$

for each $x_1, x_2 \in \mathbb{R}^n$ and for each $\theta \in [0,1]$. If $x_1 \notin \text{dom}f$ or $x_1 \notin \text{dom}f$, condition (7.13) holds trivially for each $\theta \in [0,1]$. Thus, consider the case when $x_1, x_2 \in \text{dom}f$. But then $(x_1, f(x_1)), (x_2, f(x_2)) \in \text{epi}f$. Thus, by convexity of $\text{epi}f$, we get that

$$(\theta x_1 + (1-\theta)x_2, \theta f(x_1) + (1-\theta)f(x_2)) = \theta(x_1, f(x_1)) + (1-\theta)(x_2, f(x_2)) \in \text{epi}f$$

for each $\theta \in [0,1]$. This implies that

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2)$$

for each $\theta \in [0,1]$, i.e. condition (7.13) holds. This covers all cases and we conclude that f is convex.

Solution 1.16

1. Note that

$$\begin{aligned}
f(\theta x + (1-\theta)y) &= \sum_{i=1}^{m} \alpha_i f_i(\theta x + (1-\theta)y) \\
&\leq \sum_{i=1}^{m} \alpha_i \left[\theta f_i(x) + (1-\theta)f_i(y)\right] \\
&= \theta \sum_{i=1}^{m} \alpha_i f_i(x) + (1-\theta) \sum_{i=1}^{m} \alpha_i f_i(y) \\
&= \theta f(x) + (1-\theta)f(y).
\end{aligned}$$

for each $x, y \in \mathbb{R}^n$ and for each $\theta \in [0,1]$. We conclude that f is convex.

2. Recall that a function is convex if and only if the epigraph is convex (see Exercise 1.15). Thus, $\text{epi}f_i$ is convex for each $i = 1, \ldots, m$, by assumption. Note that

$$\begin{aligned}
\text{epi}f &= \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq r\} \\
&= \left\{(x, r) \in \mathbb{R}^n \times \mathbb{R} : \max_{i=1,\ldots,m} f_i(x) \leq r\right\} \\
&= \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : f_1(x) \leq r \text{ and } f_2(x) \leq r \ldots \text{ and } f_m(x) \leq r\} \\
&= \bigcap_{i=1,\ldots,m} \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : f_i(x) \leq r\} \\
&= \bigcap_{i=1,\ldots,m} \text{epi}f_i.
\end{aligned}$$

Therefore, $\text{epi}f$ is convex since it is the intersection of convex sets (see Exercise 1.7). We conclude that f is convex.

Solution 1.17

1. We know that $\|x\|$ is convex. Define

$$h(y) = \begin{cases} y^p & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Since h is a non-decreasing and convex, the composition $h(\|x\|) = \|x\|^p$ is convex.

2. The function

$$z \mapsto \|z\|_2^2$$

is convex by the previous subproblem and

$$x \mapsto \|Ax - b\|_2^2$$

is convex since it is a composition of a convex function with an affine mapping. The function

$$x \mapsto \|x\|_1$$

is convex since norms are convex. Therefore, the function

$$x \mapsto \|Ax - b\|_2^2 + \|x\|_1 = f(x)$$

is convex since it is a sum of convex functions.

3. All norms in the max expression are convex. The max operation preserves convexity.

4. The function

$$x \mapsto \max(0, 1 + x_i) \tag{7.14}$$

is convex since it is the maximum of two convex functions, and this holds for each $i = 1 \ldots, n$. The function

$$x \mapsto \sum_{i=1}^{n} \max(0, 1 + x_i) \tag{7.15}$$

is convex since it is the sum of convex functions. We have already established that

$$x \mapsto \|x\|_2^2$$

is a convex function. Therefore, the function

$$x \mapsto \sum_{i=1}^{n} \max(0, 1 + x_i) + \|x\|_2^2 = f(x)$$

is convex since it is a sum of convex functions.

5. Suppose that $y \in \mathbb{R}^n$ is fixed. The function

$$x \mapsto x^T y - g(y)$$

is an affine function and therefore also convex. Recall that the supremum of convex functions is convex. However, f is nothing but a supremum of convex functions, i.e.

$$f(x) = \sup_{y \in \mathbb{R}^n} (x^T y - g(y))$$

where $\mathbb{R}^n$ is the index set. We conclude that f is a convex function.

Solution 1.18

1. The set C_α is nonempty since $\bar{x} \in C_\alpha$. Let $x_1 \in C_\alpha$ and $x_2 \in C_\alpha$. Then, $g(x_1) \leq \alpha$ and $g(x_2) \leq \alpha$. By convexity of g, we have that

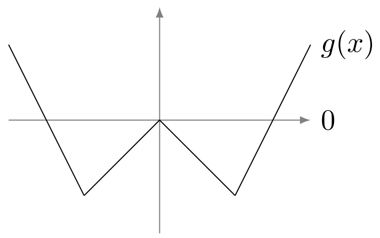
$$\begin{aligned} g(\theta x_1 + (1-\theta)x_2) &\leq \theta g(x_1) + (1-\theta) g(x_2) \\ &\leq \theta\alpha + (1-\theta)\alpha \\ &= \alpha \end{aligned}$$

and therefore

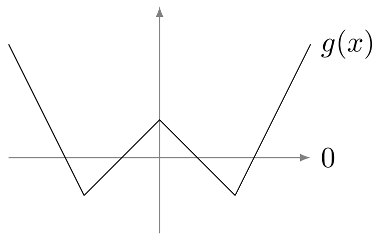
$$\theta x_1 + (1-\theta)x_2 \in C_\alpha$$

for each $\theta \in [0, 1]$. We conclude that C_α is convex.

2. Let g be as follows:



3. Let g be as follows:



Solution 1.19

Consider any $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and any $\theta \in [0, 1]$. Note that

$$\begin{aligned} g(\theta x_1 + (1-\theta)x_2, \theta y_1 + (1-\theta)y_2) &= f(\theta x_1 + (1-\theta)x_2) \\ &\leq \theta f(x_1) + (1-\theta)f(x_2) \\ &= \theta g(x_1, y_1) + (1-\theta)g(x_2, y_2) \end{aligned}$$

due to convexity of f. We conclude that g is convex.

Solution 1.20

1. The set is a sublevel set of a norm and norms are convex. We conclude that the set is convex.
2. The norm $\|x\|_2$ is convex in (x, y) and $-t$ is convex in (x, t). Therefore, their sum $\|x\|_2 - t$ is convex in (x, t). But the set is nothing but a sublevel set of the convex function $\|x\|_2 - t$, and therefore a convex set.

   Alternatively, the set is equal to the epigraph of the convex function $x \mapsto \|x\|_2$ and is therefore a convex set.

Solution 1.21
We proceed by a proof by contradiction. Assume on the contrary that x^* is not a global minimum (but still a local minimum with parameter δ). This means that there exists $\bar{x} \in \mathbb{R}^n \setminus \{x^*\}$ such that

$$f(\bar{x}) < f(x^*).$$

By convexity of f, we have

$$f((1-\theta)x^* + \theta\bar{x}) \leq (1-\theta)f(x^*) + \theta f(\bar{x}) < (1-\theta)f(x^*) + \theta f(x^*) = f(x^*)$$

or simply

$$f((1-\theta)x^* + \theta\bar{x}) < f(x^*) \tag{7.16}$$

for each $\theta \in (0, 1]$ (note that we must exclude the case $\theta = 0$ for the inequality above to hold). Now, let

$$x = (1-\theta)x^* + \theta\bar{x}$$

for some $\theta \in (0, 1]$ small enough (for instance, $\theta = \min\left(1, \frac{\delta}{\|x^* - \bar{x}\|}\right)$ will suffice here). Note that

$$\|x - x^*\| = \|(1-\theta)x^* + \theta\bar{x} - x^*\| = \theta\,\|x^* - \bar{x}\| \leq \delta$$

or simply

$$\|x - x^*\| \leq \delta.$$

However, note that (7.16) mush hold for this x, i.e.

$$f(x) < f(x^*).$$

But this is a contradiction to the fact that x^* is a local minimum of f (with parameter δ). Therefore, x^* must be a global minimum.

Solution 1.22

1. Since f is proper, we know that there exists a $y \in \mathbb{R}^n$ such that

   $$f(y) < \infty.$$

   By (1.1), we get that

   $$f(x^\star) \leq f(y) < \infty.$$

   This implies that $x^\star \in \mathrm{dom} f$. Next, we prove that $x^\star$ is the unique minimizer of f via a proof by contradiction. Assume on the contrary that there exists another minimizers $x \in \mathbb{R}^n$ of f, i.e., $x \neq x^*$ and $f(x) = f(x^*)$. This implies that $x \in \mathrm{dom} f$. Then, by strict convexity of f, we have that

   $$f\left(\frac{1}{2}x + \frac{1}{2}x^*\right) < \frac{1}{2}(f(x) + f(x^*)) = f(x^*)$$

   which is a contradiction. Hence, at most one minimizer can exist.

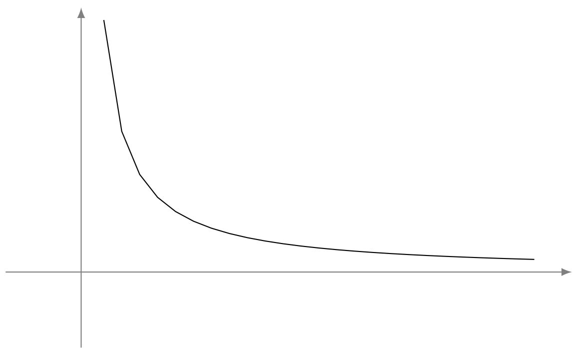
2. Consider the strictly convex function $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ such that

   $$f(x) = \begin{cases} \dfrac{1}{x} & \text{if } x > 0, \\ \infty & \text{otherwise.} \end{cases}$$

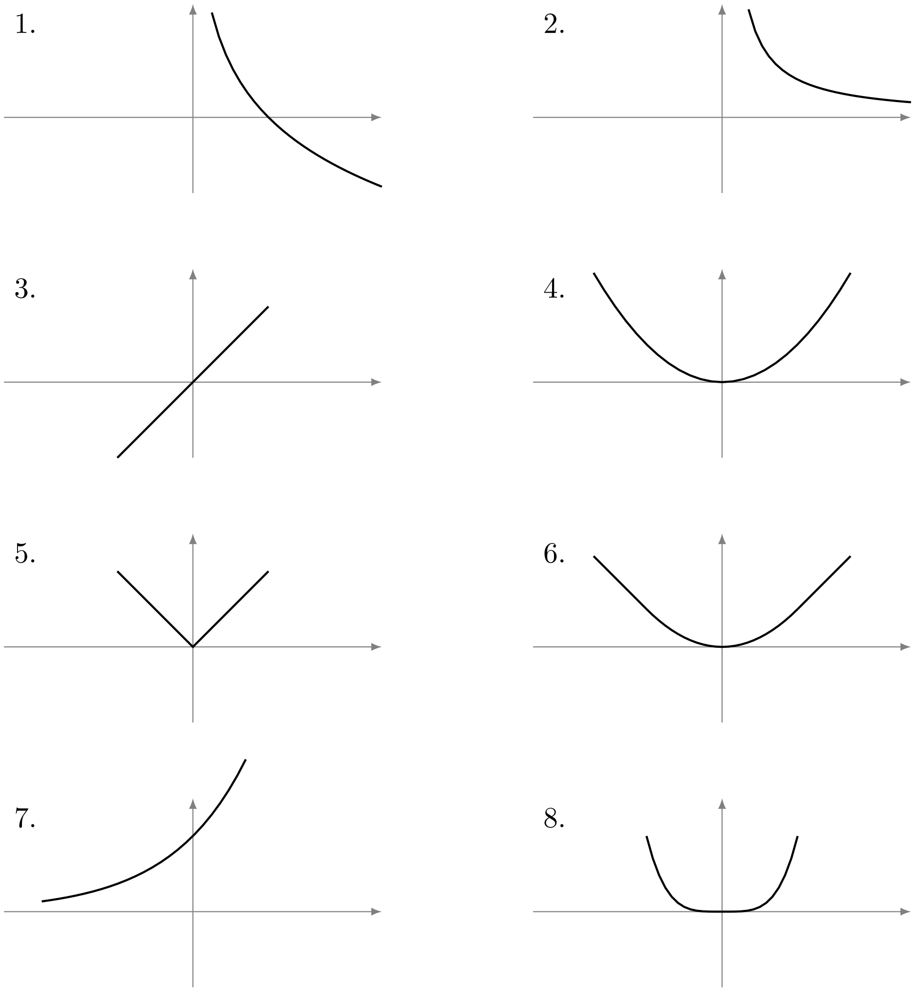
   Clearly,

   $$\inf_{x \in \mathbb{R}} f(x) = 0.$$

   However, there exists no $x \in \mathbb{R}$ such that $f(x) = 0$. See the figure below.



Solution 1.23
See figure below.



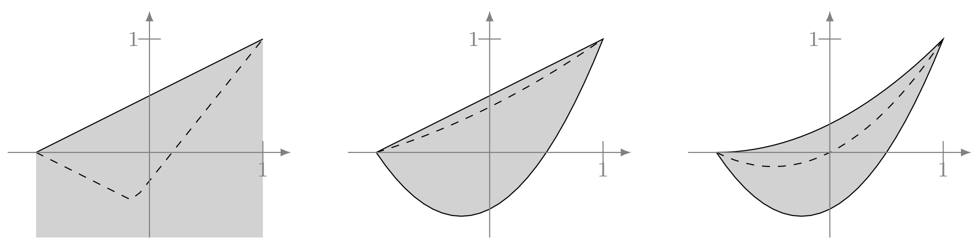
1.
   - Not smooth: The function does not have full effective domain. Hence, it can not be smooth.
   - Strictly convex: It is strictly convex since it has no flat regions.
   - Not strongly convex: It is not strongly convex since there is no quadratic lower bound.
2.
   - Not smooth: The function does not have full effective domain, hence it can not be smooth.
   - Strictly convex: It is strictly convex since it has no flat regions.
   - Not strongly convex: It is not strongly convex since there is no quadratic lower bound.
3.
   - Smooth: The function is smooth since it has quadratic upper bounds everywhere.
   - Not strictly convex: It is not strictly convex since it has flat regions.
   - Not strongly convex: It is not strongly convex since there is no quadratic lower bound.
4.
   - Smooth: The function is smooth since it has quadratic upper bounds everywhere.

- Strictly convex: It is strictly convex since it has no flat regions.
- Strongly convex: It is strongly convex since there is quadratic lower bounds everywhere.

5. - Not smooth: The function is not differentiable 0. Hence, it can not be smooth.
   - Not strictly convex: It is not strictly convex since it has flat regions.
   - Not strongly convex: It is not strongly convex since there is no quadratic lower bound.
6. - Smooth: The function is smooth since it has quadratic upper bounds everywhere.
   - Not strictly convex: It is not strictly convex since it has flat regions.
   - Not strongly convex: It is not strongly convex since there is no quadratic lower bound.
7. - Not smooth: The function is not smooth since it has no quadratic upper bounds.
   - Strictly convex: It is strictly convex since it has no flat regions.
   - Not strongly convex: It is not strongly convex since there is no quadratic lower bound.
8. - Not smooth: The function is not smooth since it has no quadratic upper bounds.
   - Strictly convex: It is strictly convex since it has no flat regions.
   - Not strongly convex: It is not strongly convex since there is no quadratic lower bound.

Solution 1.24

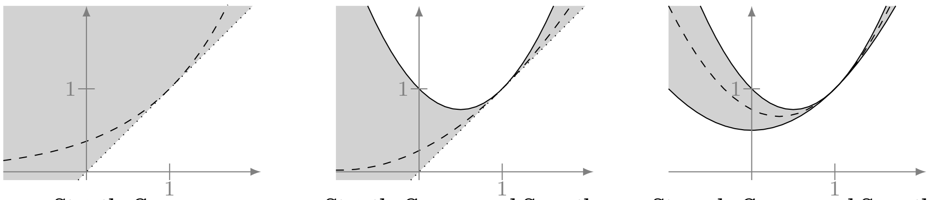
1. See the figure below. The graph a valid function must lie within the dark shaded areas. The dashed lines are examples of valid functions f. Note that smoothness always requires differentiability. The example in the convex case can therefore not be used in the smooth case even though it lies within the shaded region.



Convex — Convex and Smooth — Strongly Convex and Smooth

Solution 1.25

1. See the following figure. The graph a valid function must lie within the shaded areas. The dashed lines is are possible functions f.



Strictly Convex | Strictly Convex and Smooth | Strongly Convex and Smooth

Solution 1.26

If $a = 0$ or $b = 0$ the statement is obvious. Assume that $a, b > 0$. Let $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = \exp(x)$$

for each $x \in \mathbb{R}$. Note that

$$f''(x) = \exp(x) > 0$$

for each $x \in \mathbb{R}$. By the second-order condition for convex functions we conclude that f is convex. Note that

$$\begin{aligned} ab &= \exp\left(\frac{1}{p} p \ln a + \frac{1}{q} q \ln b\right) \\ &\overset{\text{convexity of } \exp}{\leq} \frac{1}{p}\exp(p \ln a) + \frac{1}{q}\exp(q \ln b) \\ &= \frac{a^p}{p} + \frac{b^q}{q} \end{aligned}$$

as desired.

Solution 1.27

1. Suppose that f is convex. Let $x, y \in \mathbb{R}^n$. By the convexity of f, we have that

$$f(x + \theta(y - x)) \leq (1 - \theta) f(x) + \theta f(y)$$

for each $\theta \in (0, 1]$. This can be written as

$$\theta f(y) \geq \theta f(x) + f(x + \theta(y - x)) - f(x)$$

for each $\theta \in (0, 1]$. If we divide both sides by θ and take the limit as $\theta \searrow 0$, we obtain

$$\begin{aligned} f(y) &\geq f(x) + \lim_{\theta \searrow 0} \frac{f(x + \theta(y - x)) - f(x)}{\theta} \\ &= f(x) + \nabla f(x)^T (y - x), \end{aligned}$$

where the equality follows from the hint. In particular, (1.3) holds.

Conversely, suppose that (1.3) holds. Let $x, y \in \mathbb{R}^n$, $\theta \in [0, 1]$, and let $z = \theta x + (1-\theta)y$. Then

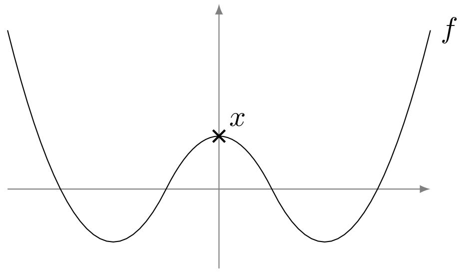
$$\begin{aligned} f(x) &\geq f(z) + \nabla f(z)^T(x - z) = f(z) + (1-\theta)\nabla f(z)^T(x - y), \\ f(y) &\geq f(z) + \nabla f(z)^T(y - z) = f(z) - \theta\nabla f(z)^T(x - y) \end{aligned}$$

Multiplying the first inequality by θ, the second by $1 - \theta$, and adding them gives

$$\theta f(x) + (1-\theta)f(y) \geq f(z) = f(\theta x + (1-\theta)y)$$

since $\theta \in [0, 1]$. We conclude that f is convex.

2. Consider the following function f and point x:



Solution 1.28

By Exercise 1.27 we know that

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) = f(x)$$

for each $y \in \mathbb{R}^n$. We see that x is a global minimizer of f.

Solution 1.29

Suppose that f is strictly convex. We know from Exercise 1.27 that we must have that

$$f(y) \geq f(x) + \nabla f(x)^T(y - x)$$

for each $x, y \in \mathbb{R}^n$. Suppose towards a contradiction that (1.4) does not hold, i.e. there exists $x, y \in \mathbb{R}^n$, $x \neq y$ such that

$$f(y) = f(x) + \nabla f(x)^T(y - x). \tag{7.17}$$

Define the function $\phi : \mathbb{R} \to \mathbb{R}$ such that

$$\phi(t) = f(x + t(y - x)) - f(x) - t\nabla f(x)^T(y - x)$$

for each $t \in \mathbb{R}$. Note that (7.17) can be written as

$$\phi(0) = \phi(1). \tag{7.18}$$

It not hard to show that ϕ is strictly convex and differentiable. Note that

$$\phi'(0) = 0.$$

By Exercise 1.28 we see that 0 is a minimizer of ϕ. But (7.18) gives that 1 is a minimizer of ϕ too. However, since ϕ is strictly convex, this gives a contradiction by Exercise 1.22 — strictly convex functions can only have an unique minimizer.

Conversely, suppose that (1.4) holds. Let $x, y \in \mathbb{R}^n$ such that $x \neq y$, $\theta \in (0,1)$, and let $z = \theta x + (1-\theta)y$. Then

$$\begin{aligned} f(x) &> f(z) + \nabla f(z)^T(x-z) = f(z) + (1-\theta)\nabla f(z)^T(x-y), \\ f(y) &> f(z) + \nabla f(z)^T(y-z) = f(z) - \theta\nabla f(z)^T(x-y) \end{aligned}$$

Multiplying the first inequality by θ, the second by $1-\theta$, and adding them gives

$$\theta f(x) + (1-\theta)f(y) > f(z) = f(\theta x + (1-\theta)y)$$

since $\theta \in (0,1)$. We conclude that f is strictly convex.

Solution 1.30

Suppose that f is σ-strongly i.e. $f - \frac{\sigma}{2}\left\|\cdot\right\|_2^2$ is convex. The derivative of $f - \frac{\sigma}{2}\left\|\cdot\right\|_2^2$ is

$$\nabla f(x) - \sigma x$$

for each $x \in \mathbb{R}$. Exercies 1.27 gives that $f - \frac{\sigma}{2}\left\|\cdot\right\|_2^2$ is convex if and only if

$$f(y) - \frac{\sigma}{2}\left\|y\right\|_2^2 \geq f(x) - \frac{\sigma}{2}\left\|x\right\|_2^2 + \left(\nabla f(x) - \sigma x\right)^T(y-x)$$

for each $x, y \in \mathbb{R}^n$. This is equivalent to that

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^T(y-x) + \frac{\sigma}{2}\left\|y\right\|_2^2 - \frac{\sigma}{2}\left\|x\right\|_2^2 - \sigma x^T(y-x) \\ &= f(x) + \nabla f(x)^T(y-x) + \frac{\sigma}{2}\left\|y\right\|_2^2 + \frac{\sigma}{2}\left\|x\right\|_2^2 - \sigma x^T y \\ &= f(x) + \nabla f(x)^T(y-x) + \frac{\sigma}{2}\left\|x-y\right\|_2^2 \end{aligned}$$

for each $x, y \in \mathbb{R}^n$. But this is (1.5). This completes the proof.

Solution 1.31

1. Consider the case $x \in C$. Then $Ax - b = 0$ and we get that

   $$\sup_{\mu \in \mathbb{R}^m} \mu^T \underbrace{(Kx-b)}_{=0} = 0 = \iota_C(x).$$

   Next, consider the case $x \notin C$. Then $Ax - b \neq 0$. Consider $\mu = t(Kx-b)$ where $t \in \mathbb{R}$. Then

   $$\mu^T(Kx-b) = t\underbrace{\left\|Kx-b\right\|^2}_{>0} \to \infty \quad \text{as} \quad t \to \infty.$$

In particular,

$$\sup_{\mu\in\mathbb{R}^m} \mu^T(Kx-b) = \infty = \iota_C(x).$$

2. Consider the case $x \in C$. Then $g(x) \leq 0$ and we have that

$$\mu^T g(x) \leq 0$$

for each $\mu \in \mathbb{R}^m_+$. Moreover,

$$\mu^T g(x) = 0$$

for $\mu = 0$. Therefore,

$$\sup_{\mu\in\mathbb{R}^m_+} \mu^T g(x) = 0 = \iota_C(x).$$

Next, consider the case $x \notin C$. Then there exists an index $i \in \{1,\ldots,m\}$ such that $(g(x))_i > 0$. Consider $\mu = te_i \in \mathbb{R}^n_+$ where $t \geq 0$. Then

$$\mu^T g(x) = t\underbrace{(g(x))_i}_{>0} \to \infty \quad \text{as} \quad t \to \infty.$$

In particular,

$$\sup_{\mu\in\mathbb{R}^m_+} \mu^T g(x) = \infty = \iota_C(x).$$

Solution 1.32

1. We want to show that

$$h(\theta x + (1-\theta)y) \leq \theta h(x) + (1-\theta)h(y) \tag{7.19}$$

for each $x, y \in \mathbb{R}$ and for each $\theta \in [0,1]$. If $x = y$ or $\theta = 0$ or $\theta = 1$, inequality (7.19) holds trivially. Thus, assume that $x \neq y$ and $\theta \in (0,1)$. We may without loss of generality assume that $x < y$. Then we have that

$$x < \theta x + (1-\theta)y < y.$$

By the mean value theorem, there exists $\xi_1, \xi_2 \in \mathbb{R}$ such that

$$\begin{aligned} \frac{h(\theta x + (1-\theta)y) - h(x)}{(1-\theta)(y-x)} &= \frac{h(\theta x + (1-\theta)y) - h(x)}{(\theta x + (1-\theta)y) - x} = h'(\xi_1), \\ \frac{h(y) - h(\theta x + (1-\theta))}{\theta(y-x)} &= \frac{h(y) - h(\theta x + (1-\theta))}{y - (\theta x + (1-\theta)y)} = h'(\xi_2) \end{aligned}$$

and

$$x < \xi_1 < \theta x + (1-\theta)y < \xi_2 < y.$$

Multiplying the first equality by $-\theta(1-\theta)(y-x)$, the second equality by $\theta(1-\theta)(y-x)$ and noting that $h'(\xi_1) \leq h'(\xi_2)$ gives that

$$-\theta(h(\theta x+(1-\theta)y)-h(x)) = -\theta(1-\theta)(y-x)h'(\xi_1) \geq -\theta(1-\theta)(y-x)h'(\xi_2),$$
$$(1-\theta)(h(y)-h(\theta x+(1-\theta))) = \theta(1-\theta)(y-x)h'(\xi_2)$$

Summing these and rearranging gives (7.19). We conclude that h is convex.

2. Suppose that $p = 1$. Then

$$h(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that h is not differentiable. However, it is trivial to check that h is nondecreasing and convex using the definitions. Next, suppose that $p > 1$. Then h is differentiable and

$$h'(x) = \begin{cases} px^{p-1} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Since $h' \geq 0$, we conclude that h is nondecreasing. If we can show that h' is nondecreasing, we know that h is convex by the previous subproblem. Thus, let $x < y$. If $x < y \leq 0$ or $x \leq 0 < y$ it is trivial to show that $h'(x) \leq h'(y)$. Therefore, assume that $0 < x < y$. But then

$$\begin{gathered} \ln x < \ln y \\ \Leftrightarrow \\ (p-1)\ln x < (p-1)\ln y \\ \Leftrightarrow \\ \exp((p-1)\ln x) < \exp((p-1)\ln y) \\ \Leftrightarrow \\ h'(x) = px^{p-1} = p\exp((p-1)\ln x) < p\exp((p-1)\ln y) = py^{p-1} = h'(y). \end{gathered}$$

This shows that h' is nondecreasing and thus, h is convex. This concludes the proof.

Solution 1.33

We proceed by induction on n. In the base case $n = 1$, inequality (1.6) holds trivially. For the inductive step, assume that inequality (1.6) holds for $n = k$, where $k \in \mathbb{N}$. We need to prove that inequality (1.6) holds for $n = k+1$. In the case $\theta_{k+1} = 1$, inequality

(1.6) holds trivially. Therefore, assume that $\theta_{k+1} < 1$. Note that

$$\begin{aligned}
f\left(\sum_{i=1}^{k+1} \theta_i x_i\right) &= f\left(\sum_{i=1}^{k} \theta_i x_i + \theta_{k+1} x_{k+1}\right) \\
&= f\left((1-\theta_{k+1})\left(\sum_{i=1}^{k} \frac{\theta_i}{1-\theta_{k+1}} x_i\right) + \theta_{k+1} x_{k+1}\right) \\
&\overset{\text{convexity of } f}{\leq} (1-\theta_{k+1}) f\left(\sum_{i=1}^{k} \frac{\theta_i}{1-\theta_{k+1}} x_i\right) + \theta_{k+1} f(x_{k+1}) \\
&\overset{\text{inductive assumption}}{\leq} (1-\theta_{k+1}) \sum_{i=1}^{k} \frac{\theta_i}{1-\theta_{k+1}} f(x_i) + \theta_{k+1} f(x_{k+1}) \\
&= \sum_{i=1}^{k+1} \theta_i f(x_i).
\end{aligned}$$

Thus, inequality (1.6) holds true for $n = k+1$, establishing the inductive step. By mathematical induction, inequality (1.6) holds true for each $n \in \mathbb{N}$.

Solution 1.34

First, we recall some definitions. The function f is called affine if the function

$$x \mapsto f(x) - f(0) \tag{7.20}$$

is linear. Moreover, the function f is called concave if $-f$ is convex. Thus, f is concave if and only if

$$f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta) f(y)$$

for each $x, y \in \mathbb{R}^n$ and for each $\theta \in [0,1]$.

First, suppose that f is affine. We then know that the function (7.20) is linear. Let $x, y \in \mathbb{R}^n$ and $\theta \in [0,1]$. Then

$$\begin{aligned}
f(\theta x + (1-\theta)y) - f(0) &= \theta(f(x) - f(0)) + (1-\theta)(f(y) - f(0)) \\
&= \theta f(x) + (1-\theta) f(y) - f(0)
\end{aligned}$$

which implies that

$$f(\theta x + (1-\theta)y) = \theta f(x) + (1-\theta) f(y).$$

In particular,

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

and

$$f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta) f(y)$$

holds, and we conclude that f is both convex and concave.

Conversely, suppose that f is concave and convex. Define the function $g : \mathbb{R}^n \to \mathbb{R}$ such that

$$g(x) = f(x) - f(0)$$

for each $x \in \mathbb{R}^n$. We need to show that g is linear. Note that g is concave and convex. This implies that

$$g(\theta x + (1-\theta)y) = \theta g(x) + (1-\theta)g(y)$$

for each $x, y \in \mathbb{R}^n$ and for each $\theta \in [0,1]$. Moreover, note that $g(0) = 0$. Let $x \in \mathbb{R}^n$. Then

$$\begin{aligned} 0 &= g(0) \\ &= g\left(\frac{1}{2}x + \frac{1}{2}(-x)\right) \\ &= \frac{1}{2}g(x) + \frac{1}{2}g(-x) \end{aligned}$$

which shows that g is an odd function, i.e.

$$g(-x) = -g(x).$$

We have the following two facts:

- *Claim:* The function g is homogeneous of degree 1, i.e.

  $$g(\alpha x) = \alpha g(x)$$

  for each $x \in \mathbb{R}^n$ and for each $\alpha \in \mathbb{R}$.

  Proof: Let $x \in \mathbb{R}^n$. The cases $\alpha = 0$ or $\alpha = 1$ hold trivially. Suppose that $\alpha \in (0,1)$. Then

  $$\begin{aligned} g(\alpha x) &= g(\alpha x + (1-\alpha)0) \\ &= \alpha g(x) + (1-\alpha)g(0) \\ &= \alpha g(x) + (1-\alpha)0 \\ &= \alpha g(x). \end{aligned}$$

  Suppose that $\alpha > 1$. Then

  $$\begin{aligned} g(x) &= g\left(\frac{1}{\alpha}(\alpha x) + \left(1 - \frac{1}{\alpha}\right)0\right) \\ &= \frac{1}{\alpha}g(\alpha x) + \left(1 - \frac{1}{\alpha}\right)g(0) \\ &= \frac{1}{\alpha}g(\alpha x) \end{aligned}$$

  which implies that

  $$g(\alpha x) = \alpha g(x).$$

  Suppose that $\alpha < 0$. Then

  $$\begin{aligned} g(\alpha x) &= g((-\alpha)(-x)) \\ &= (-\alpha)g(-x) \\ &= (-\alpha)(-g(x)) \\ &= \alpha g(x). \end{aligned}$$

  This covers all cases.

- *Claim:* The function g is addative with respect to addition, i.e.

$$g(x+y) = g(x) + g(y)$$

for each $x, y \in \mathbb{R}^n$.

Proof: Let $x, y \in \mathbb{R}^n$. Then

$$\begin{aligned} g(x+y) &= g\left(\frac{1}{2}(2x) + \frac{1}{2}(2y)\right) \\ &= \frac{1}{2}g(2x) + \frac{1}{2}g(2y) \\ &= \frac{1}{2}(2g(x)) + \frac{1}{2}(2g(y)) \\ &= g(x) + g(y). \end{aligned}$$

This proves the claim.

These two facts give that

$$g(\alpha x + \beta y) = \alpha g(x) + \beta g(y)$$

for each $x, y \in \mathbb{R}^n$ and each $\alpha, \beta \in \mathbb{R}$, i.e. g is linear. Thus, we conclude that f is affine.

Solution 1.35

Assume that f is σ-strongly convex, i.e. $f(x) - \frac{\sigma}{2}\|x\|_2^2$ is convex. By definition, this means that

$$f(z) - \tfrac{\sigma}{2}\|z\|_2^2 \leq \theta(f(x) - \tfrac{\sigma}{2}\|x\|_2^2) + (1-\theta)(f(y) - \tfrac{\sigma}{2}\|y\|_2^2) \tag{7.21}$$

where $z = \theta x + (1-\theta)y$, for each $x, y \in \mathbb{R}^n$ and for each $\theta \in [0,1]$. But (7.21) is equivalent to

$$f(z) \leq \theta f(x) + (1-\theta)f(y) + \tfrac{\sigma}{2}(\|z\|_2^2 - \theta\|x\|_2^2 - (1-\theta)\|y\|_2^2) \tag{7.22}$$

for each $x, y \in \mathbb{R}^n$ and for each $\theta \in [0,1]$. Note that

$$\begin{aligned} &\|z\|_2^2 - \theta\|x\|_2^2 - (1-\theta)\|y\|_2^2 \\ &= \|\theta\, x + (1-\theta)y\|_2^2 - \theta\|x\|_2^2 - (1-\theta)\|y\|_2^2 \\ &= (\theta^2 - \theta)\|x\|_2^2 + ((1-\theta)^2 - (1-\theta))\|y\|_2^2 + 2\theta(1-\theta)x^T y \\ &= (\theta(1-\theta))(-\|x\|_2^2 - \|y\|_2^2 + 2x^T y) \\ &= -(\theta(1-\theta))(\|x-y\|_2^2). \end{aligned} \tag{7.23}$$

Inserting (7.23) into (7.22) gives (1.7). This proves the equivalence.

Solution 1.36

Recall that the spectral norm $\|A\|_2$ of A is defined such that

$$\|A\|_2 = \max\{\|Ax\|_2 : x \in \mathbb{R}^m,\ \|x\|_2 \leq 1\}.$$

This definition implies that

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2$$

for each $x \in \mathbb{R}^m$. We have that

$$\nabla g(x) = A^T \nabla f(Ax - b)$$

for each $x \in \mathbb{R}^m$. Let $x, y \in \mathbb{R}^m$. Note that

$$\begin{aligned}
\|\nabla g(x) - \nabla g(x)\|_2 &= \left\|A^T \nabla f(Ax - b) - A^T \nabla f(Ay - b)\right\|_2 \\
&= \left\|A^T (\nabla f(Ax - b) - \nabla f(Ay - b))\right\|_2 \\
&= \left\|A^T\right\|_2 \|(\nabla f(Ax - b) - \nabla f(Ay - b))\|_2 \\
&= \beta \left\|A^T\right\|_2 \|(Ax - b) - (Ay - b)\|_2 \\
&= \beta \left\|A^T\right\|_2 \|A(x - y)\|_2 \\
&= \beta \left\|A^T\right\|_2 \|A\|_2 \|x - y\|_2 \\
&= \beta \|A\|_2^2 \|x - y\|_2 .
\end{aligned}$$

This shows that ∇g is $\beta \|A\|_2^2$-Lipschitz continuous. We conclude that g is $\beta \|A\|_2^2$-smooth, as desired.

Solution 1.37

We first prove the equivalence in the simple case when $\beta = 0$. Property I) is equivalent to f being affine. Moreover, property II)-IV) simply give that f is convex and concave. But this holds if and only if f is affine. Therefore, I)-IV) are equivalent.

Next, we consider the case when $\beta > 0$.

I) $\Rightarrow$ II): Assume that I) holds. Note that for $x, y \in \mathbb{R}^n$ and $t \in \mathbb{R}$,

$$\frac{\partial}{\partial t} f(x + t(y - x)) = \nabla f(x + t(y - x))^T (y - x).$$

This gives that

$$f(y) - f(x) = \int_0^1 \nabla f(x + t(y - x))^T (y - x) dt \tag{7.24}$$

for each $x, y \in \mathbb{R}^n$. Subtracting $\nabla f(x)^T (y - x)$ from the expression above and taking absolute value yields

$$\begin{aligned}
&\left|f(y) - f(x) - \nabla f(x)^T (y - x)\right| \\
&= \left|\int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x) dt\right| \\
&\leq \int_0^1 \left|(\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x)\right| dt \\
&\overset{\text{Cauchy-Schwartz}}{\leq} \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\|_2 \|y - x\|_2 dt \\
&\overset{\text{I)}}{\leq} \int_0^1 t\beta \|y - x\|_2^2 dt \\
&= \frac{\beta}{2} \|y - x\|_2^2.
\end{aligned}$$

I.e. II) holds.

<u>II) ⇒ I)</u>: Assume that II) holds. Consider any $x, y, z \in \mathbb{R}^n$. In II), insert z for y in the first inequality, and insert y for x and z for y in the second inequality. I.e.

$$\begin{cases} f(z) \le f(x) + \nabla f(x)^T(z-x) + \frac{\beta}{2}\|x-z\|_2^2, \\ f(z) \ge f(y) + \nabla f(y)^T(z-y) - \frac{\beta}{2}\|y-z\|_2^2, \end{cases}$$

or

$$\begin{cases} f(z) \le f(x) + \nabla f(x)^T(z-x) + \frac{\beta}{2}\|x-z\|_2^2, \\ f(y) \le f(z) - \nabla f(y)^T(z-y) + \frac{\beta}{2}\|y-z\|_2^2. \end{cases}$$

Adding this pair of inequalities yields

$$\begin{aligned} f(y) &\le f(x) + \nabla f(x)^T(z-x) - \nabla f(y)^T(z-y) + \frac{\beta}{2}\|x-z\|_2^2 + \frac{\beta}{2}\|y-z\|_2^2 \\ &= f(x) - \nabla f(x)^T x + \nabla f(y)^T y + \frac{\beta}{2}\|x\|_2^2 + \frac{\beta}{2}\|y\|_2^2 + \beta\|z\|_2^2 + z^T(\nabla f(x) - \nabla f(y) - \beta x - \beta y) \\ &= f(x) - \nabla f(x)^T x + \nabla f(y)^T y + \frac{\beta}{2}\|x\|_2^2 + \frac{\beta}{2}\|y\|_2^2 \\ &\quad + \beta\|z + \frac{1}{2\beta}(\nabla f(x) - \nabla f(y) - \beta x - \beta y)\|_2^2 - \beta\|\frac{1}{2\beta}(\nabla f(x) - \nabla f(y) - \beta x - \beta y)\|_2^2. \end{aligned}$$

We are free to choose $z = -\frac{1}{2\beta}(\nabla f(x) - \nabla f(y) - \beta x - \beta y)$. This gives

$$\begin{aligned} f(y) &\le f(x) - \nabla f(x)^T x + \nabla f(y)^T y + \frac{\beta}{2}\|x\|_2^2 + \frac{\beta}{2}\|y\|_2^2 - \frac{1}{4\beta}\|\nabla f(x) - \nabla f(y) - \beta x - \beta y\|_2^2 \\ &= f(x) - \frac{1}{4\beta}\|(\nabla f(x) - \nabla f(y)\|_2^2 \\ &\quad + \frac{\beta}{2}\|x\|_2^2 + \frac{\beta}{2}\|y\|_2^2 - \frac{\beta}{4}\|x+y\|_2^2 - \nabla f(x)^T x + \nabla f(y)^T y + \frac{1}{2}(\nabla f(x) - \nabla f(y))^T(x+y) \\ &= f(x) - \frac{1}{4\beta}\|\nabla f(x) - \nabla f(y)\|_2^2 + \frac{\beta}{4}\|x-y\|_2^2 + \frac{1}{2}(\nabla f(x) + \nabla f(y))^T(y-x). \end{aligned}$$

We may insert x for y and y for x in the in inequality above. This yields the pair of inequalities

$$\begin{cases} f(y) \le f(x) - \frac{1}{4\beta}\|\nabla f(x) - \nabla f(y)\|_2^2 + \frac{\beta}{4}\|x-y\|_2^2 + \frac{1}{2}(\nabla f(x) + \nabla f(y))^T(y-x), \\ f(x) \le f(y) - \frac{1}{4\beta}\|\nabla f(y) - \nabla f(x)\|_2^2 + \frac{\beta}{4}\|y-x\|_2^2 + \frac{1}{2}(\nabla f(y) + \nabla f(x))^T(x-y). \end{cases}$$

Adding the pair of inequalities gives

$$0 \le -\frac{1}{2\beta}\|\nabla f(y) - \nabla f(x)\|_2^2 + \frac{\beta}{2}\|y-x\|_2^2,$$

i.e. I) holds.

<u>II) ⇔ III)</u>: Note that the gradient of $\frac{\beta}{2}\|x\|_2^2 - f(x)$ and $f(x) + \frac{\beta}{2}\|x\|_2^2$ are $\beta x - \nabla f(x)$ and $\nabla f(x) + \beta x$, respectively. By the first-order condition for convexity, we get that $\frac{\beta}{2}\|x\|_2^2 - f(x)$ and $f(x) + \frac{\beta}{2}\|x\|_2^2$ are convex if and only if

$$\begin{cases} \frac{\beta}{2}\|y\|_2^2 - f(y) \ge \frac{\beta}{2}\|x\|_2^2 - f(x) + (\beta x - \nabla f(x))^T(y-x), \\ f(y) + \frac{\beta}{2}\|y\|_2^2 \ge f(x) + \frac{\beta}{2}\|x\|_2^2 + (\nabla f(x) + \beta x)^T(y-x), \end{cases}$$

or

$$\begin{cases} f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}\|x-y\|_2^2, \\ f(y) \geq f(x) + \nabla f(x)^T(y-x) - \frac{\beta}{2}\|x-y\|_2^2, \end{cases}$$

holds for each $x, y \in \mathbb{R}^n$. But this is II).

III) $\Leftrightarrow$ IV): Applying Exercise 1.35 (the statement in Exercise 1.35 generalizes to all $\sigma \in \mathbb{R}$ and the proof remains exactly the same) to $\frac{\beta}{2}\|x\|_2^2 - f(x)$ and $f(x) + \frac{\beta}{2}\|x\|_2^2$ gives the result immediately.

Solution 1.38

By Exercise 1.37 we get that I) is equivalent to that $\frac{\beta}{2}\|x\|_2^2 - f(x)$ and $f(x) + \frac{\beta}{2}\|x\|_2^2$ are convex function. However, by the second-order condition for convex functions, this is equivalent to

$$\beta I - \nabla^2 f(x) \succeq 0 \text{ and } \nabla^2 f(x) + \beta I \succeq 0, \text{ for each } x \in \mathbb{R}^n,$$

respectively. This is simply II). This establishes the desired equivalence.

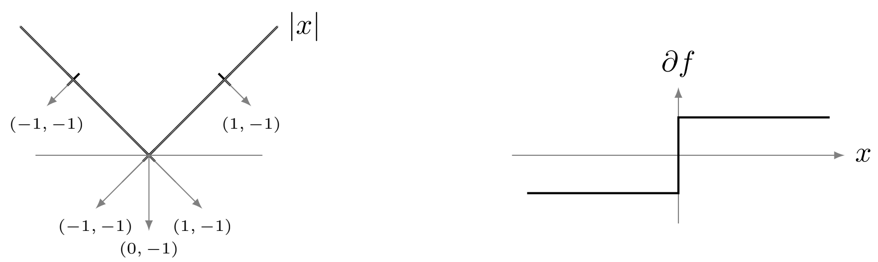
Solutions to chapter 2

Solution 2.1

1. The function is convex, finite-valued and differentiable with $\nabla f(x) = x$. Therefore, $\partial f(x) = \{\nabla f(x)\} = \{x\}$.
2. The function is convex, finite-valued and differentiable with $\nabla f(x) = Hx + h$. Therefore, $\partial f(x) = \{\nabla f(x)\} = \{Hx + h\}$.
3. The function is convex, finite-valued and differentiable except at $x = 0$.
    - For $x < 0$, the function is $f(x) = -x$ and differentiable with gradient $\nabla f(x) = -1$. Therefore, $\partial f(x) = \{\nabla f(x)\} = \{-1\}$ in this case.
    - For $x > 0$, the function is $f(x) = x$ and differentiable with gradient $\nabla f(x) = 1$. Therefore, $\partial f(x) = \{\nabla f(x)\} = \{1\}$ in this case.
    - At $x = 0$, all elements in $[-1, 1]$ are subgradients. See the figure below. Therefore, $\partial f(x) = [-1, 1]$ in this case.

   Thus,

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 0, \\ [-1, 1] & \text{if } x = 0, \\ \{1\} & \text{if } x > 0. \end{cases}$$



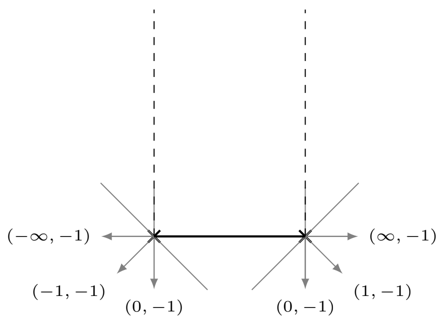
4. The function is convex.
    - For $x < -1$ or $x > 1$, we have that $x \notin \text{dom} f$. Therefore, $\partial f(x) = \emptyset$ in this case.
    - For $x \in (-1, 1) \subseteq \text{relintdom} f$, the function is $f(x) = 0$ and differentiable with gradient $\nabla f(x) = 0$. Therefore, $\partial f(x) = \{\nabla f(x)\} = \{0\}$ in this case.
    - For $x = 1$, each $s \geq 0$ is a subgradient. See the figure below. Therefore, $\partial f(x) = [0, \infty)$ in this case.

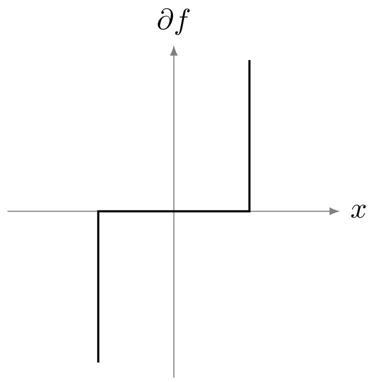
- For $x = -1$, each $s \leq 0$ is a subgradient. See the figure below. Therefore, $\partial f(x) = (-\infty, 0]$ in this case.

Thus,

$$\partial f(x) = \begin{cases} [-\infty, 0] & \text{if } x = -1, \\ \{0\} & \text{if } x \in (-1, 1), \\ [0, \infty] & \text{if } x = 1, \\ \emptyset & \text{otherwise.} \end{cases}$$

Remark: Note that this subdifferential is the inverse of the subdifferential of $|x|$.



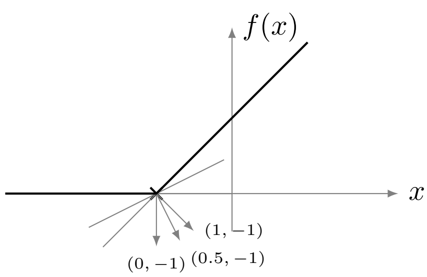


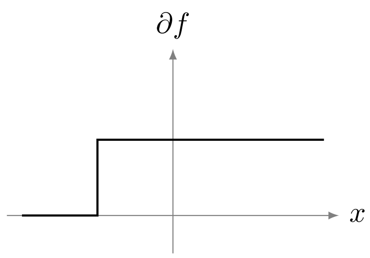
5. The function is convex and finite-valued.

   - For $x < -1$, the function is $f(x) = 0$ and differentiable with gradient $\nabla f(x) = 0$. Therefore, $\partial f(x) = \{\nabla f(x)\} = \{0\}$ in this case.
   - For $x > -1$, the function is $f(x) = 1 + x$ and differentiable with gradient $\nabla f(x) = 1$. Therefore, $\partial f(x) = \{\nabla f(x)\} = \{1\}$ in this case.
   - For $x = -1$, each $s \in [0, 1]$ is a subgradient. See the figure below. Therefore, $\partial f(x) = [0, 1]$ in this case.

   Thus,

$$\partial f(x) = \begin{cases} \{0\} & \text{if } x < -1, \\ [0, 1] & \text{if } x = -1, \\ \{1\} & \text{if } x > -1. \end{cases}$$





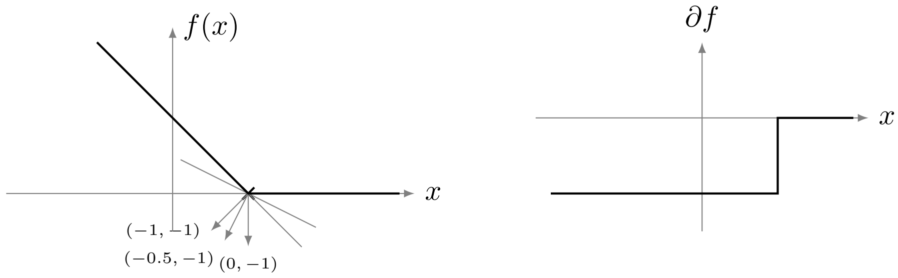
6. The function is convex and finite-valued.

   - For $x < 1$, the function is $f(x) = 1 - x$ and differentiable with gradient $\nabla f(x) = -1$. Therefore, $\partial f(x) = \{\nabla f(x)\} = \{-1\}$ in this case.

- For $x > 1$, the function is $f(x) = 0$ and differentiable with gradient $\nabla f(x) = 0$. Therefore, $\partial f(x) = \{\nabla f(x)\} = \{0\}$ in this case.
- For $x = 1$, each $s \in [-1, 0]$ is a subgradient. See the figure below. Therefore, $\partial f(x) = [-1, 0]$ in this case.

Thus,

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 1, \\ [-1, 0] & \text{if } x = 1, \\ \{0\} & \text{if } x > 1. \end{cases}$$



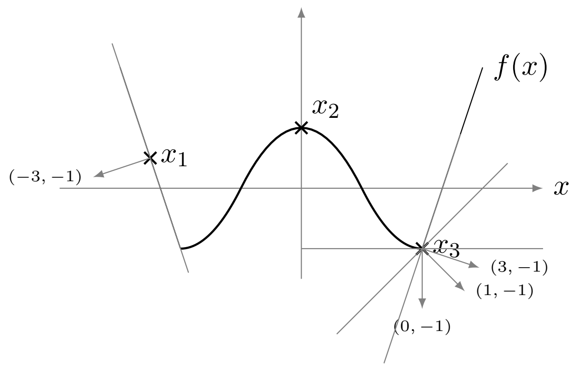
Solution 2.2

1. See figure below.

   x_1: There is one affine minorizor to f at x_1 with slope -3. Hence, $\partial f(x_1) = \{-3\}$. The function f is also differentiable at x_1 with gradient -3. Hence, $\nabla f(x_1) = -3$.

   x_2: There is no affine minorizor to f at x_2. Hence, $\partial f(x) = \emptyset$. However, f is differentiable at x_2 with gradient $\nabla f(x_2) = 0$.

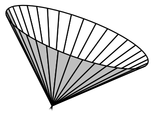
   x_3: There are several affine minorizors to f and x_3. Their slopes range from 0 to 3. Hence, $\partial f(x_3) = [0, 3]$. However, f is not differentiable at x_3.

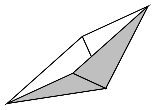


2. Fermat's rule $0 \in \partial f(x)$ holds for x_3 but not for x_1 and x_2. Therefore, x_3 is a global minimum to the nonconvex function f.

Solution 2.3

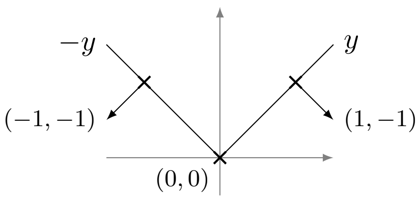
1. Since $\partial f(x)$ and $\partial g(y)$ are subsets of $\mathbb{R}^2$, a reasonable domain for both f and g is $\mathbb{R}^2$. I.e., we have that $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$. Also, we must have that $x, y \in \mathbb{R}^2$.
2. Yes, since $0 \in \partial f(x)$.
3. No, since $0 \notin \partial g(y)$.
4. No, since the subdifferential not a singleton (unique) at x.
5. No, since the subdifferential not a singleton (unique) at y.
6. See examples below.



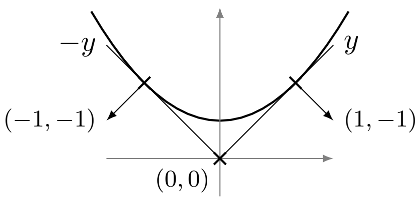


Solution 2.4

1. The following function (which is the absolute value $|x|$) is a lower bound to the function f:



2. Since the function above is a lower bound to f, its minimum 0 is a lower bound to the minimum of f.
3. An example of function f is given below. The function is $f(x) = \frac{1}{2}(x^2 + 1)$.



Solution 2.5

- From the definition of monotonicity, we know that the minimum slope is 0 and maximum is ∞. Therefore a. and b. are monotone while c. and d. are not.

- We rule out c. and d. since they are not monotone. Since operators $A : \mathbb{R} \to 2^{\mathbb{R}}$ for Figures a. and b. are monotone, there exist functions f such that $A = \partial f$. The subdifferential in a. is maximally monotone. Hence, a. is a subdifferential of a closed convex function. The subdifferential in b. is not maximally monotone Hence, b. is not a subdifferential of a closed convex function.

Solution 2.6

Suppose that $A - \sigma I$ is monotone, i.e

$$((s_x - \sigma x) - (s_y - \sigma y))^T(x - y) \geq 0$$

for each $x, y \in \mathrm{dom}A$, for each $s_x \in Ax$ and for each $s_y \in Ay$. However, this inequality is equivalent to that

$$(s_x - s_y)^T(x - y) \geq \sigma\|x - y\|_2^2$$

for each $x, y \in \mathrm{dom}A$, for each $s_x \in Ax$ and for each $s_y \in Ay$. But this is the definition of A being σ-strongly monotone. This proves the equivalence.

Solution 2.7

We know that we need to consider $n \geq 2$, since for $n = 1$, each monotone operator is a subdifferential of some function. Therefore, let $n = 2$ and let $A : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear single-valued operator such that

$$A(x_1, x_2) = (x_2, -x_1)$$

for each $(x_1, x_2) \in \mathbb{R}^2$. With some notation overloading, A can be represented by the matrix

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Then $A = -A^T$ (i.e. A is skew symmetric) and

$$\begin{aligned} (Ax - Ay)^T(x - y) &= (x - y)^T A^T (x - y) \\ &= -(x - y)^T(Ax - Ay) \\ &= -(Ax - Ay)^T(x - y). \end{aligned}$$

Hence $(Ax - Ay)^T(x - y) = 0$ and monotonicity holds with equality.

However, A is not the gradient of a function since the matrix A would be the Hessian, but it is not symmetric.

Solution 2.8

1. Assume that I) holds. Let $x, y \in \mathbb{R}^n$. Write I) and I) with x and y swapped,

$$\begin{cases} f(y) \geq f(x) + \nabla f(x)^T(y-x), \\ f(x) \geq f(y) + \nabla f(y)^T(x-y). \end{cases}$$

   Adding these gives

$$(\nabla f(y) - \nabla f(x))^T(y-x) \geq 0$$

   i.e. II).

2. Assume that II) holds. Let $x, y \in \mathbb{R}^n$. Using the hint we get that

$$\begin{aligned} &f(y) - f(x) - \nabla f(x)^T(y-x) \\ &= \int_0^1 t^{-1} \underbrace{(\nabla f(x+t(y-x)) - \nabla f(x))^T((x+t(y-x))-x)}_{\geq 0 \text{ by II)}} \, dt \\ &\geq 0. \end{aligned}$$

   But this is I).

Solution 2.9

1. a. Since ∂f is maximally monotone, f is closed and convex.
   b. Since ∂f is not maximally monotone, f is not closed and convex.
2. An optimal point x^* satisfies $0 \in \partial f(x^*)$ by Fermat's rule. Hence, the minimizing x^* are the x where the graph crosses the x-axis for both a. and b.
3. No, since a constant offset of f is not visible in ∂f.
4. Below are example plots of f.

a. b.

   It is linear to the left of the minimum and quadratic to the right.

Solution 2.10

Since f is σ-strongly convex and closed, the function $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that

$$g(x) = f(x) - \frac{\sigma}{2} \|x\|_2^2$$

for each $x \in \mathbb{R}^n$ is convex and closed. By the subdifferential sum rule, we have that

$$\begin{aligned}
\partial f(x) &= \partial \left(f - \frac{\sigma}{2}\left\|\cdot\right\|_2^2 + \frac{\sigma}{2}\left\|\cdot\right\|_2^2\right)(x) \\
&= \partial \left(g + \frac{\sigma}{2}\left\|\cdot\right\|_2^2\right)(x) \\
&= \partial g(x) + \partial \left(\frac{\sigma}{2}\left\|\cdot\right\|_2^2\right)(x) \\
&= \partial g(x) + \sigma x
\end{aligned}$$

for each $x \in \mathbb{R}^n$, which is equivalent to that

$$\partial g(x) = \partial f(x) - \sigma x \tag{7.25}$$

for each $x \in \mathbb{R}^n$. This implies that

$$\mathrm{dom}\partial f = \mathrm{dom}\partial g.$$

Let $x \in \mathrm{dom}\partial f$ and $s_f \in \partial f(x)$. Then (7.25) implies that there exists an $s_g \in \partial g(x)$ such that

$$s_g = s_f - \sigma x.$$

Let $y \in \mathbb{R}^n$. Note that

$$\begin{aligned}
f(y) - \frac{\sigma}{2}\|y\|_2^2 &= g(y) \\
&\overset{s_g \in \partial g(x)}{\geq} g(x) + s_g^T(y - x) \\
&= f(x) - \frac{\sigma}{2}\|x\|_2^2 + s_g^T(y - x) \\
&= f(x) - \frac{\sigma}{2}\|x\|_2^2 + (s_f - \sigma x)^T(y - x) \\
&= f(x) + s_f^T(y - x) - \frac{\sigma}{2}\|x\|_2^2 + \sigma x^T(y - x).
\end{aligned}$$

Now, since $\|y\|_2^2 - \|x\|_2^2 - 2x^T(y - x) = \|x - y\|_2^2$, this is implies that

$$f(y) \geq f(x) + s_f^T(y - x) + \tfrac{\sigma}{2}\|x - y\|_2^2$$

i.e. the desired result.

Solution 2.11

(a)
- The function f is not differentiable as ∂f is multivalued at 0.
- Since f is not differentiable, f does not have a Lipschitz continuous gradient.
- The subdifferential ∂f not strongly monotone since it has minimum slope 0. Hence, f is not strongly convex.

(b)
- The function f is differentiable as ∂f is a singleton everywhere.

- The subdifferential ∂f has maximum slope 1. Hence, ∇f is 1-Lipschitz.
- The subdifferential ∂f is not strongly monotone since it has minimum slope 0. Hence, f is not strongly convex.

(c)
- The function f is differentiable as ∂f is a singleton everywhere.
- The subdifferential ∂f has maximum slope 1. Hence, ∇f is 1-Lipschitz.
- The subdifferential ∂f is not strongly monotone since it has minimum slope 0. Hence, f is not strongly convex.

(d)
- The function f is differentiable as ∂f is a singleton everywhere.
- The subdifferential ∂f has maximum slope 1. Hence, ∇f is 1-Lipschitz.
- The subdifferential ∂f is 1/2-strongly monotone since it has minimum slope 1/2. Hence, f is 1/2-strongly convex.

Solution 2.12

Assume that $s \in \partial g(x)$. Then

$$\begin{aligned}\sum_{i=1}^{n} g_i(y_i) &= g(y) \\ &\geq g(x) + s^T(x-y) \\ &= \sum_{i=1}^{n}(g_i(x_i) + s_i(y_i - x_i)). \end{aligned} \tag{7.26}$$

for each $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Let $j \in \{1, \ldots, n\}$ and let $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ such that $y_i = x_i$ for each $i = 1, \ldots, n$ and $i \neq j$. Using this y in (7.26) gives that

$$g_j(y_j) \geq g_j(x_j) + s_j(y_j - x_j),$$

for each $y_j \in \mathbb{R}$. This implies that

$$s_j \in \partial g_j(x_j). \tag{7.27}$$

However, since $j \in \{1, \ldots, n\}$ is arbitrary, we get that (7.27) holds for each $j = 1, \ldots, n$.

Conversely, assume that $s_i \in \partial g_i(x_i)$ for each $i = 1, \ldots, n$. But then

$$g_i(y_i) \geq g_i(x_i) + s_i(y_i - x_i) \tag{7.28}$$

holds for each $y_i \in \mathbb{R}$ and for each $i = 1, \ldots, n$. Summing (7.28) over $i = 1, \ldots, n$ gives that

$$g(y) = \sum_{i=1}^{n} g_i(y_i) \geq \sum_{i=1}^{n}(g_i(x_i) + s_i(y_i - x_i)) = g(x) + s^T(y-x)$$

for each $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. In partial, $s \in \partial g(x)$ holds.

This proves the equivalence.

Solution 2.13
For $x \notin \mathrm{dom} f$, subgradients $s \in \partial f(x)$ must satisfy

$$f(y) \geq f(x) + s^T(y - x) \quad \text{for each } y \in \mathbb{R}^n.$$

Since there exists $y \in \mathbb{R}^n$ such that $f(y) < \infty$ and $f(x) = \infty$, we see that $\partial f(x)$ must by empty.

Solution 2.14
Recall that the normal cone to C at $x \in \mathbb{R}^n$ is given by

$$N_C(x) = \begin{cases} \left\{s \in \mathbb{R}^n : \forall y \in C,\, s^T(y - x) \leq 0\right\} & \text{if } x \in C \\ \emptyset & \text{if } x \notin C. \end{cases}$$

Let $x \in \mathbb{R}^n$. We have that $s \in \partial \iota_C(x)$ if and only if

$$\iota_C(y) \geq \iota_C(x) + s^T(y - x)$$

for each $y \in \mathbb{R}^n$, by definition.

- First, assume that $x \in C$ and $s \in \partial \iota_C(x)$. Then $\iota_C(y) \geq s^T(y - x)$ for each $y \in \mathbb{R}^n$, which is equivalent to that $s^T(y - x) \leq 0$ for each $y \in C$, since C is nonempty.
- Next, assume that $x \notin C$ and $s \in \partial \iota_C(x)$. Consider $y \in C$. Then $0 \geq \infty + s^T(y - x)$, which is impossible. Hence, $\iota_C(x) = \emptyset$ in this case.

We conclude that

$$\partial \iota_C(x) = N_C(x)$$

for each $x \in \mathbb{R}^n$.

Solution 2.15
Recall that Fermat's rule gives that $x = \mathrm{prox}_{\gamma f}(z)$ if and only if $0 \in \partial f(x) + \gamma^{-1}(x - z)$. We will use this multiple times throughout.

1. Let $z \in \mathbb{R}^n$, $\gamma > 0$ and $x = \mathrm{prox}_{\gamma f}(z)$. We have

   $$\partial f(x) = \{x\}.$$

   Therefore, we get that $0 = \gamma x + (x - z)$ or $x = (1 + \gamma)^{-1} z$, and conclude that

   $$\mathrm{prox}_{\gamma f}(z) = (1 + \gamma)^{-1} z.$$

2. Let $z \in \mathbb{R}^n$, $\gamma > 0$ and $x = \mathrm{prox}_{\gamma f}(z)$. We have

   $$\partial f(x) = \{Hx + h\}.$$

   Therefore, we get that $0 = \gamma(Hx + h) + (x - z)$ or $(I + \gamma H)x = z - \gamma h$ or $x = (I + \gamma H)^{-1}(z - \gamma h)$, and conclude that

   $$\mathrm{prox}_{\gamma f}(z) = (I + \gamma H)^{-1}(z - \gamma h).$$

3. Let $z \in \mathbb{R}$, $\gamma > 0$ and $x = \operatorname{prox}_{\gamma f}(z)$. We have

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 0, \\ [-1, 1] & \text{if } x = 0, \\ \{1\} & \text{if } x > 0. \end{cases}$$

   - For $x < 0$, we have $\partial f(x) = \{-1\}$. Therefore, we get that $0 = -\gamma + (x - z)$ or $x = \gamma + z$. Note that $z < -\gamma$ implies the condition $x < 0$.
   - For $x > 0$, we have $\partial f(x) = \{1\}$. Therefore, we get that $0 = \gamma + (x - z)$ or $x = z - \gamma$. Note that $z > \gamma$ implies the condition $x > 0$.
   - For $x = 0$, we have $\partial f(x) = [-1, 1]$. Therefore, we get that $0 \in [-\gamma, \gamma] - z$ or $z \in [-\gamma, \gamma]$.

   Thus,

$$\operatorname{prox}_{\gamma f}(z) = \begin{cases} z + \gamma & \text{if } z < -\gamma, \\ 0 & \text{if } z \in [-\gamma, \gamma], \\ z - \gamma & \text{if } z > \gamma. \end{cases}$$

4. Let $z \in \mathbb{R}$, $\gamma > 0$ and $x = \operatorname{prox}_{\gamma f}(z)$. Here, f is the indicator function of the set $[-1, 1]$. Recall that $\operatorname{prox}_{\gamma f}(z)$ then reduces to the projection onto $[-1, 1]$.

   - If $z \leq -1$, the projection is point is -1.
   - If $z \in [-1, 1]$, the projection point is z, since $z \in [-1, 1]$.
   - If $z \geq 1$, the projection point is 1.

   Thus,

$$\operatorname{prox}_{\gamma f} = \begin{cases} -1 & \text{if } z < -1, \\ z & \text{if } z \in [-1, 1], \\ 1 & \text{if } z > 1. \end{cases}$$

5. Let $z \in \mathbb{R}$, $\gamma > 0$ and $x = \operatorname{prox}_{\gamma f}(z)$. We have

$$\partial f(x) = \begin{cases} \{0\} & \text{if } x < -1, \\ [0, 1] & \text{if } x = -1, \\ \{1\} & \text{if } x > -1. \end{cases}$$

   - For $x < -1$, we have $\partial f(x) = \{0\}$. Therefore, we get that $0 = x - z$ or $x = z$. Note that $z < -1$ implies the condition $x < -1$.
   - For $x > -1$, we have $\partial f(x) = \{1\}$. Therefore, we get that $0 = \gamma + (x - z)$ or $x = z - \gamma$. Note that $z > \gamma - 1$ implies the condition $x > -1$.
   - For $x = -1$, we have $\partial f(x) = [0, 1]$. Therefore, we get that $0 \in [0, \gamma] + (-1 - z)$ or $z \in [-1, \gamma - 1]$.

   Thus,

$$\operatorname{prox}_{\gamma f} = \begin{cases} z & \text{if } z < -1, \\ -1 & \text{if } z \in [-1, \gamma - 1], \\ z - \gamma & \text{if } z > \gamma - 1. \end{cases}$$

6. Let $z \in \mathbb{R}$, $\gamma > 0$ and $x = \mathrm{prox}_{\gamma f}(z)$. We have

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 1, \\ [-1, 0] & \text{if } x = 1, \\ \{0\} & \text{if } x > 1. \end{cases}$$

- For $x < 1$, we have $\partial f(x) = \{-1\}$. Therefore, we get that $0 = -\gamma + (x - z)$ or $x = z + \gamma$. Note that $z < 1 - \gamma$ implies the condition $x < 1$.
- For $x > 1$, we have $\partial f(x) = \{0\}$. Therefore, we get that $0 = 0 + (x - z)$ or $x = z$. Note that $z > 1$ implies the condition $x > 1$.
- For $x = 1$, we have $\partial f(x) = [-1, 0]$. Therefore, we get that $0 \in [-\gamma, 0] + (1 - z)$ or $z \in [1 - \gamma, 1]$.

Thus,

$$\mathrm{prox}_{\gamma f} = \begin{cases} z + \gamma & \text{if } z < 1 - \gamma, \\ 1 & \text{if } z \in [1 - \gamma, 1], \\ z & \text{if } z > 1. \end{cases}$$

Solution 2.16

One can show that $g_i : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ must be proper, closed and convex, for each $i = 1, \ldots, n$. However, we may assume this without proof. We have that

$$\begin{aligned}
\mathrm{prox}_{\gamma g}(z) &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left(g(x) + \frac{1}{2\gamma} \|x - z\|_2^2 \right) \\
&= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left(\sum_{i=1}^n g_i(x_i) + \frac{1}{2\gamma} \sum_{i=1}^n (x_i - z_i)^2 \right) \\
&= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left(\sum_{i=1}^n g_i(x_i) + \frac{1}{2\gamma} (x_i - z_i)^2 \right) \\
&= \begin{bmatrix} \operatorname{argmin}_{x_1 \in \mathbb{R}} \left(g_1(x_1) + \frac{1}{2\gamma}(x_1 - z_1)^2 \right) \\ \vdots \\ \operatorname{argmin}_{x_n \in \mathbb{R}} \left(g_n(x_n) + \frac{1}{2\gamma}(x_n - z_n)_2^2 \right) \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{prox}_{\gamma g_1}(z_1) \\ \vdots \\ \mathrm{prox}_{\gamma g_n}(z_n) \end{bmatrix}.
\end{aligned}$$

Solutions to chapter 3

Solution 3.1

Recall that the conjugate function of a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is denoted as f^* and given by

$$f^*(s) = \sup_{x \in \mathbb{R}^n} \left(s^T x - f(x) \right)$$

for each $s \in \mathbb{R}^n$.

1. We have

$$\begin{aligned} f^*(s) &= \sup_{x \in \mathbb{R}^n} \left(s^T x - \tfrac{1}{2} \left\| x \right\|_2^2 \right) \\ &= - \inf_{x \in \mathbb{R}^n} \underbrace{\left(-s^T x + \tfrac{1}{2} \left\| x \right\|_2^2 \right)}_{=g(x)} \end{aligned}$$

for each $s \in \mathbb{R}^n$. Fermat's rule gives that $x \in \mathbb{R}^n$ is an optimal solution to the optimization problem above if and only if

$$0 \in \partial g(x). \tag{7.29}$$

Since g is finite-valued, convex and differentiable, we know that $\partial g(x) = \{\nabla g(x)\}$. Thus, (7.29) is equivalent to that

$$0 = -s + x \quad \text{or} \quad x = s.$$

Therefore,

$$\begin{aligned} f^*(s) &= s^T s - \tfrac{1}{2} \left\| s \right\|_2^2 \\ &= \tfrac{1}{2} \left\| s \right\|_2^2 \end{aligned}$$

for each $s \in \mathbb{R}^n$.

2. We have

$$\begin{aligned} f^*(s) &= \sup_{x \in \mathbb{R}^n} \left(s^T x - \tfrac{1}{2} x^T H x - h^T x \right) \\ &= - \inf_{x \in \mathbb{R}^n} \underbrace{\left(-s^T x + \tfrac{1}{2} x^T H x + h^T x \right)}_{=g(x)} \end{aligned}$$

for each $s \in \mathbb{R}^n$. Fermat's rule gives that $x \in \mathbb{R}^n$ is an optimal solution to the optimization problem above if and only if

$$0 \in \partial g(x). \tag{7.30}$$

Since g is finite-valued, convex and differentiable, we know that $\partial g(x) = \{\nabla g(x)\}$. Thus, (7.29) is equivalent to that

$$0 = -s + Hx + h \quad \text{or} \quad x = H^{-1}(s-h)$$

since H invertible. Therefore,

$$\begin{aligned} f^*(s) &= s^T(H^{-1}(s-h)) - \tfrac{1}{2}(s-h)^T H^{-1} H H^{-1}(s-h) - h^T H^{-1}(s-h) \\ &= \tfrac{1}{2}(s-h)^T H^{-1}(s-h). \end{aligned}$$

for each $s \in \mathbb{R}^n$.

3. We have

$$f^*(s) = \sup_{x\in[-1,1]} sx$$

for each $s \in \mathbb{R}$.

If $s \leq 0$, an optimal solution to the optimization problem above is $x = -1$ and get that $f^*(s) = -s$.

If $s \geq 0$, an optimal solution to the optimization problem above is $x = 1$ and get that $f^*(s) = s$.

Therefore,

$$f^*(s) = |s|$$

for each $s \in \mathbb{R}$.

4. *Alterative 1:* Since $\iota_{[-1,1]}$ is proper, closed and convex, we have that

$$\iota_{[-1,1]}^{**} = \iota_{[-1,1]}.$$

Recall from above that

$$\iota_{[-1,1]}^{*} = |\cdot|.$$

Therefore,

$$\begin{aligned} f^* &= |\cdot|^* \\ &= (\iota_{[-1,1]}^*)^* \\ &= \iota_{[-1,1]}^{**} \\ &= \iota_{[-1,1]}. \end{aligned}$$

Alterative 2: We have

$$f^*(s) = \sup_{x\in\mathbb{R}} (sx - |x|)$$

for each $s \in \mathbb{R}$.

Next, we consider three different cases.

- Suppose that $s < -1$. Let $x \leq 0$. Then

$$\begin{aligned} f^*(s) &\geq sx + x \\ &= \underbrace{(s+1)}_{<0} x \to \infty \quad \text{as} \quad x \to -\infty. \end{aligned}$$

  Thus, $f^*(s) = \infty$ in this case.

- Suppose that $s > 1$. Let $x \geq 0$. Then

$$\begin{aligned} f^*(s) &\geq sx - x \\ &= \underbrace{(s-1)}_{>0} x \to \infty \quad \text{as} \quad x \to \infty. \end{aligned}$$

  Thus, $f^*(s) = \infty$ in this case.

- Suppose that $s \in [-1, 1]$. Note that

$$\begin{aligned} f^*(s) &\geq s0 - |0| \\ &= 0. \end{aligned}$$

  By the Cauchy-schwarz inequality we have that

$$sx \leq |x||s| \leq |x|$$

  for each $x \in \mathbb{R}$, since $|s| \leq 1$. Therefore,

$$\begin{aligned} f^*(s) &\leq \sup_{x \in \mathbb{R}} (|x| - |x|) \\ &= 0. \end{aligned}$$

  Thus, $f^*(s) = 0$ for each $s \in [-1, 1]$.

We conclude that

$$f^* = \iota_{[-1,1]}.$$

5. Recall Fenchel-Young equality, i.e.

$$f^*(s) = sx - f(x)$$

if and only if

$$s \in \partial f(x).$$

Recall from Exercise 2.1-5 that

$$\partial f(x) = \begin{cases} \{0\} & \text{if } x < -1, \\ [0, 1] & \text{if } x = -1, \\ \{1\} & \text{if } x > -1. \end{cases}$$

Next, we consider different cases.

- Suppose that $x < -1$. Then $s = 0$. Therefore,

$$\begin{aligned} f^*(s) &= 0x - \underbrace{f(x)}_{=0 \text{ since } x<-1} \\ &= 0. \end{aligned}$$

- Suppose that $x > -1$. Then $s = 1$. Therefore,

$$\begin{aligned} f^*(s) &= x - \underbrace{f(x)}_{=1+x \text{ since } x>-1} \\ &= -1. \end{aligned}$$

- Suppose that $x = -1$. Then $s \in [0, 1]$. Therefore,

$$\begin{aligned} f^*(s) &= -s - \underbrace{f(x)}_{=0 \text{ since } x=-1} \\ &= -s. \end{aligned}$$

- Suppose that $s < 0$. Let $x \leq -1$. Then

$$f^*(s) \geq sx \to \infty \quad \text{as} \quad x \to -\infty.$$

  Therefore, $f^*(s) = \infty$ for each $s < 0$.
- Suppose that $s > 1$. Let $x \geq -1$. Then

$$\begin{aligned} f^*(s) &\geq sx - (1 + x) \\ &= \underbrace{(s-1)}_{>0} x - 1 \to \infty \quad \text{as} \quad x \to -\infty. \end{aligned}$$

  Therefore, $f^*(s) = \infty$ for each $s > 1$.

We conclude that

$$f^*(s) = \begin{cases} -s & \text{if } s \in [0, 1], \\ \infty & \text{otherwise.} \end{cases}$$

6. Recall Fenchel-Young equality, i.e.

$$f^*(s) = sx - f(x)$$

if and only if

$$s \in \partial f(x).$$

Recall from Exercise 2.1-6 that

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 1, \\ [-1, 0] & \text{if } x = 1, \\ \{0\} & \text{if } x > 1. \end{cases}$$

Next, we consider different cases.

- Suppose that $x < 1$. Then $s = -1$. Therefore,

$$\begin{aligned} f^*(s) &= -x - \underbrace{f(x)}_{=1-x \text{ since } x<1} \\ &= -1. \end{aligned}$$

- Suppose that $x > 1$. Then $s = 0$. Therefore,

$$\begin{aligned} f^*(s) &= 0x - \underbrace{f(x)}_{=0 \text{ since } x>1} \\ &= 0. \end{aligned}$$

- Suppose that $x = 1$. Then $s \in [-1, 0]$. Therefore,

$$\begin{aligned} f^*(s) &= s - \underbrace{f(x)}_{=0 \text{ since } x=1} \\ &= s. \end{aligned}$$

- Suppose that $s < -1$. Let $x \leq 1$. Then

$$\begin{aligned} f^*(s) &\geq sx - (1 - x) \\ &= \underbrace{(s+1)}_{<0} x - 1 \to \infty \quad \text{as} \quad x \to -\infty. \end{aligned}$$

  Therefore, $f^*(s) = \infty$ for each $s < -1$.

- Suppose that $s > 0$. Let $x \geq 1$. Then

$$f^*(s) \geq sx \to \infty \quad \text{as} \quad x \to \infty.$$

  Therefore, $f^*(s) = \infty$ for each $s > 0$.

We conclude that

$$f^*(s) = \begin{cases} s & \text{if } s \in [-1, 0], \\ \infty & \text{otherwise.} \end{cases}$$

Solution 3.2

1. Note that

$$\begin{aligned} f^*(s) &= \sup_{z \in \mathbb{R}^n} \left(s^T z - f(z) \right) \\ &\geq s^T x - f(x) \end{aligned}$$

   for each $s, x \in \mathbb{R}^n$. This implies that

$$f(x) \geq s^T x - f^*(s)$$

for each $s, x \in \mathbb{R}^n$. This implies that

$$\begin{aligned} f(x) &\geq \sup_{s\in\mathbb{R}^n} \big(s^T x - f^*(s)\big) \\ &= f^{**}(x) \end{aligned}$$

for each $x \in \mathbb{R}^n$ or simply

$$f^{**} \leq f$$

as desired.

2. Assume that $f \leq g$, i.e.

$$f(x) \leq g(x),$$

for each $x \in \mathbb{R}^n$. Then

$$s^T x - f(x) \geq s^T x - g(x),$$

for each $s, x \in \mathbb{R}^n$. In particular,

$$f^*(s) = \sup_{x\in\mathbb{R}^n} \big(s^T x - f(x)\big) \geq \sup_{x\in\mathbb{R}^n} \big(s^T x - g(x)\big) = g^*(s),$$

for each $s \in \mathbb{R}^n$. We conclude that $f^* \geq g^*$.

3. Assume that $f \leq g$. From the previous subproblem we get that $f^* \geq g^*$, i.e.

$$f^*(s) \geq g^*(s),$$

for each $s \in \mathbb{R}^n$. Then

$$x^T s - f^*(s) \leq x^T s - g^*(s),$$

for each $s, x \in \mathbb{R}^n$. In particular,

$$f^{**}(x) = \sup_{s\in\mathbb{R}^n} \big(x^T s - f^*(s)\big) \leq \sup_{s\in\mathbb{R}^n} \big(x^T s - g^*(s)\big) = g^{**}(x),$$

for each $x \in \mathbb{R}^n$. We conclude that $f^{**} \leq g^{**}$.

4. Suppose that $f = \frac{1}{2}\|\cdot\|_2^2$. From Exercise 3.1-1 we know that $f^* = \frac{1}{2}\|\cdot\|_2^2$. Therefore, $f = f^*$.

   Conversely, suppose that $f = f^*$. Note that

$$f(x) + f(s) = f(x) + f^*(s) \geq x^T s,$$

   for each $s, x \in \mathbb{R}^n$, by Fenchel-Young's inequality. If we pick $s = x$, we get that

$$f(x) \geq \frac{1}{2}\|x\|_2^2,$$

   for each $x \in \mathbb{R}^n$, i.e.

$$f \geq \frac{1}{2}\|\cdot\|_2^2.$$

   However, we know from the second subproblem above that this implies that

$$f = f^* \leq (\frac{1}{2}\|\cdot\|_2^2)^* = \frac{1}{2}\|\cdot\|_2^2.$$

   We conclude that $f = \frac{1}{2}\|\cdot\|_2^2$.

   This completes the proof.

Solution 3.3

The hint gives that

$$\left(\nabla\frac{|\cdot|^p}{p}\right)(x) = \begin{cases} x\,|x|^{p-2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

By definition,

$$\begin{aligned} \left(\frac{|\cdot|^p}{p}\right)^*(s) &= \sup_{x\in\mathbb{R}}\left(sx - \frac{|x|^p}{p}\right) \\ &= -\inf_{x\in\mathbb{R}} \underbrace{\left(-sx + \frac{|x|^p}{p}\right)}_{=g(x)} \end{aligned}$$

for each $s \in \mathbb{R}$. Fermat's rule gives that $x \in \mathbb{R}^n$ is an optimal solution to the optimization problem above if and only if

$$0 \in \partial g(x). \tag{7.31}$$

Since g is finite-valued, convex and differentiable, we know that $\partial g(x) = \{\nabla g(x)\}$. Thus, (7.31) is equivalent to that

$$0 = -s + \begin{cases} x\,|x|^{p-2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0 \end{cases} \quad \text{or} \quad s = \begin{cases} x\,|x|^{p-2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}.$$

Let $x \neq 0$, then $s = x\,|x|^{p-2}$, and

$$\begin{aligned} \left(\frac{|\cdot|^p}{p}\right)^*(s) &= sx - \frac{|x|^p}{p} \\ &= x\,|x|^{p-2}\,x - \frac{|x|^p}{p} \\ &= \left(1 - \frac{1}{p}\right)|x|^p \\ &= \frac{|x|^p}{q} \\ &= \frac{|x|^{(p-1)q}}{q} \\ &= \frac{\left|x\,|x|^{p-2}\right|^q}{q} \\ &= \frac{|s|^q}{q}. \end{aligned}$$

Let $x = 0$, then $s = 0$, and

$$\left(\frac{|\cdot|^p}{p}\right)^*(0) = 0 = \frac{|0|^q}{q}.$$

This covers all cases for $s \in \mathbb{R}$. We conclude that

$$\left(\frac{|\cdot|^p}{p}\right)^* = \left(\frac{|\cdot|^q}{q}\right),$$

as desired.

Solution 3.4
Note that

$$\begin{aligned}
(\alpha f + (1-\alpha)g)^*(s) &= \sup_{x\in\mathbb{R}^n} \left(s^T x - (\alpha f(x) + (1-\alpha)g(x))\right) \\
&= \sup_{x\in\mathbb{R}^n} \left(\alpha \left(s^T x - f(x)\right) + (1-\alpha)\left(s^T x - g(x)\right)\right) \\
&\leq \sup_{x\in\mathbb{R}^n} \left(\alpha \left(s^T x - f(x)\right)\right) + \sup_{x\in\mathbb{R}^n} \left((1-\alpha)\left(s^T x - g(x)\right)\right) \\
&= \alpha \sup_{x\in\mathbb{R}^n} \left(s^T x - f(x)\right) + (1-\alpha) \sup_{x\in\mathbb{R}^n} \left(s^T x - g(x)\right) \\
&= \alpha f^*(s) + (1-\alpha) g^*(s),
\end{aligned}$$

for every $s \in \mathbb{R}^n$, i.e.

$$(\alpha f + (1-\alpha)g)^* \leq \alpha f^* + (1-\alpha)g^*$$

as desired.

Solution 3.5
We have

$$\begin{aligned}
f^*(s) &= \sup_{x\in\mathbb{R}^n} \left(x^T s - f(x)\right) \\
&= \sup_{x\in\mathbb{R}^n} \left(\sum_{i=1}^n x_i s_i - f_i(x_i)\right) \\
&= \sum_{i=1}^n \sup_{x_i\in\mathbb{R}} \left(x_i s_i - f_i(x_i)\right) \\
&= \sum_{i=1}^n f_i^*(s_i).
\end{aligned}$$

for each $s = (s_1, \ldots, s_n) \in \mathbb{R}^n$.

Solution 3.6

1. The function f can be written as

$$\begin{aligned}
f(x) &= \|x\|_1 \\
&= \sum_{i=1}^n |x_i|
\end{aligned}$$

for each $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Therefore, by Exercise 3.5 and Exercise 3.1-4, we have that

$$\begin{aligned} f^*(s) &= \sum_{i=1}^{n} (|\cdot|)^*(s_i) \\ &= \sum_{i=1}^{n} \iota_{[-1,1]}(s_i) \\ &= \iota_{[-\mathbf{1},\mathbf{1}]}(s) \end{aligned}$$

for each $s = (s_1, \dots, s_n) \in \mathbb{R}^n$.

2. The function f can be written as

$$\begin{aligned} f(x) &= \iota_{[-\mathbf{1},\mathbf{1}]}(x) \\ &= \sum_{i=1}^{n} \iota_{[-1,1]}(x_i) \end{aligned}$$

for each $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Therefore, by Exercise 3.5 and Exercise 3.1-3, we have that

$$\begin{aligned} f^*(s) &= \sum_{i=1}^{n} \iota^*_{[-1,1]}(s_i) \\ &= \sum_{i=1}^{n} |s_i| \\ &= \|s\|_1 \end{aligned}$$

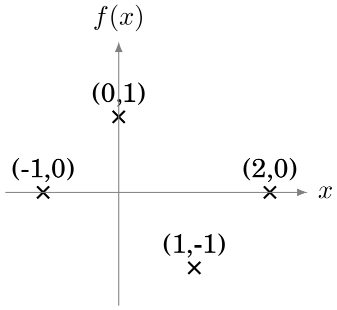
for each $s = (s_1, \dots, s_n) \in \mathbb{R}^n$.

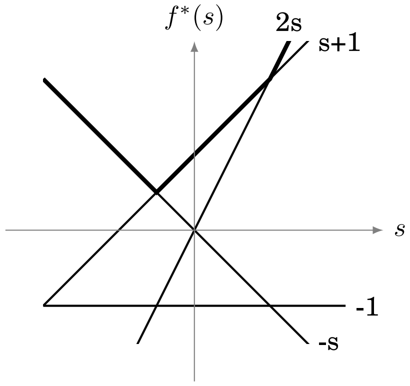
Solution 3.7

1. Since f is only defined in four points, the conjugate is

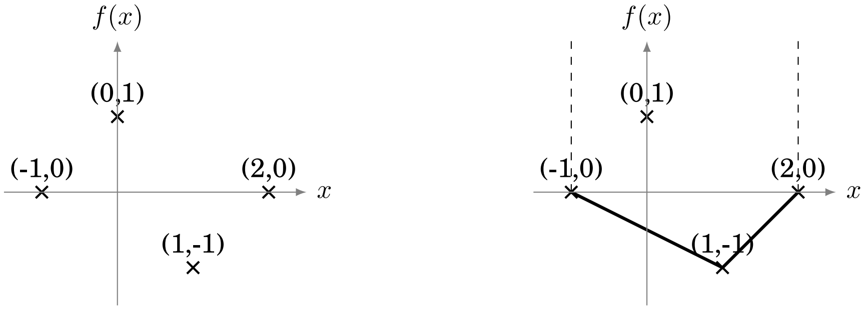
$$f^*(s) = \sup_{x \in \mathbb{R}} (sx - f(x)) = \max(-s - 0, -1, s + 1, 2s)$$

for each $s \in \mathbb{R}$.





2. The biconjugate f^{**} is the convex envelope of f. See the figure below.



Solution 3.8

1. We have that

$$f^*(s) = \sup_{x \in \mathbb{R}^n} \left(s^T x - \|x\|_2 \right)$$

for each $s \in \mathbb{R}^n$.

(a) Note that

$$\begin{aligned} f^*(s) &\geq s^T 0 - \|0\|_2 \\ &= 0 \end{aligned}$$

for each $s \in \mathbb{R}^n$.

(b) By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} f^*(s) &\leq \sup_{x \in \mathbb{R}^n} (\|s\|_2 \|x\|_2 - \|x\|_2) \\ &= \sup_{x \in \mathbb{R}^n} \left(\underbrace{(\|s\|_2 - 1)}_{\leq 0} \|x\|_2 \right) \\ &\leq 0 \end{aligned}$$

for each $s \in \mathbb{R}^n$ such that $\|s\|_2 \leq 1$. Combined with the previous sub-exercise, we see that $f^*(s) = 0$ for each $s \in \mathbb{R}^n$ such that $\|s\|_2 \leq 1$.

(c) Suppose that $s \in \mathbb{R}^n$ such that $\|s\|_2 > 1$. Let $x = ts$ for some $t \geq 0$. Then

$$\begin{aligned} f^*(s) &= \sup_{x \in \mathbb{R}} (s^T x - \|x\|_2) \\ &\geq t\|s\|_2^2 - t\|s\|_2 \\ &= t \underbrace{\|s\|_2}_{>1} \underbrace{(\|s\|_2 - 1)}_{>0} \to \infty \quad \text{as} \quad t \to \infty. \end{aligned}$$

Thus,

$$f^*(s) = \infty$$

for each $s \in \mathbb{R}^n$ such that $\|s\|_2 > 1$.

(d) To summarize, we have

$$f^*(s) = \begin{cases} 0 & \text{if } \|s\|_2 \leq 1, \\ \infty & \text{otherwise} \end{cases}$$

for each $s \in \mathbb{R}^n$.

2. Since f is closed and convex, the subdifferential of f satisfies

$$\begin{aligned} \partial f(x) &= \operatorname*{Argmax}_{s \in \mathbb{R}^n} (s^T x - f^*(s)) \\ &= \operatorname*{Argmax}_{s \in \mathbb{R}^n : \|s\|_2 \leq 1} \; s^T x. \end{aligned}$$

for each $x \in \mathbb{R}^n$.

- For $x = 0$, the objective is 0 and all feasible points are optimal, i.e.,

$$\begin{aligned} \partial f(x) &= \{s \in \mathbb{R}^n : \|s\|_2 \leq 1\} \\ &= B(0, 1). \end{aligned}$$

- For $x \neq 0$, note that

$$\begin{aligned} \max_{s \in \mathbb{R}^n : \|s\|_2 \leq 1} s^T x &\leq \max_{s \in \mathbb{R}^n : \|s\|_2 \leq 1} \|s\|_2 \, \|x\|_2 \\ &= \|x\|_2 \end{aligned}$$

with equality if and only if $s = x/\|x\|_2$, by the Cauchy-Schwarz inequality. Therefore,

$$\partial f(x) = \left\{ \frac{x}{\|x\|_2} \right\}.$$

We conclude that

$$\partial f(x) = \begin{cases} B(0, 1) & \text{if } x = 0, \\ \left\{ \frac{x}{\|x\|_2} \right\} & \text{otherwise} \end{cases}$$

for each $x \in \mathbb{R}^n$.

Solution 3.9

1. We have that

$$f^*(s) = \sup_{x \in \Delta} s^T x$$

for each $s \in \mathbb{R}^n$. The exercise claims that

$$f^*(s) = \max_{i=1,\ldots,n} s_i$$

for each $s = (s_1, \ldots, s_n) \in \mathbb{R}^n$. Let $s \in \mathbb{R}^n$. Let j be any index such that

$$s_j = \max_{i=1,\ldots,n} s_i.$$

First, note that

$$\begin{aligned} f^*(s) &\geq s^T \underbrace{e_j}_{\in \Delta} \\ &= s_j \end{aligned}$$

Next, let $x \in \Delta$ such that $x \neq e_j$. Then

$$\begin{aligned} s^T x &= \sum_{i=1}^{n} s_i x_i \\ &= s_j x_j + \sum_{\substack{i=0 \\ i \neq j}}^{n} s_i \underbrace{x_i}_{\geq 0} \\ &\leq s_j x_j + \sum_{\substack{i=0 \\ i \neq j}}^{n} s_j x_i \\ &= s_j x_j + s_j \sum_{\substack{i=0 \\ i \neq j}}^{n} x_i \\ &= s_j \underbrace{\sum_{i=1}^{n} x_i}_{=1} \\ &= s_j. \end{aligned}$$

Hence, all points $x \in \Delta \setminus \{e_j\}$ satisfy $s^T x \leq s_j$ and the point $e_j \in \Delta$ satisfy $s^T e_j = s_j$. Therefore,

$$\begin{aligned} f^*(s) &= \sup_{x \in \Delta} s^T x \\ &= \max_{i=1,\ldots,n} s_i, \end{aligned}$$

for each $s = (s_1, \ldots, s_n) \in \mathbb{R}^n$, as desired.

2. The function f is proper, closed and convex. Therefore,

$$f^{**} = f = \iota_\Delta.$$

3. We have that

$$g^*(s) = \sup_{x \in D} s^T x$$

for each $s \in \mathbb{R}^n$. The exercise claims that

$$g^*(s) = \max\left(0, \max_{i=1,\ldots,n} s_i\right)$$

for each $s = (s_1, \ldots, s_n) \in \mathbb{R}^n$. Let $s \in \mathbb{R}^n$.

- Suppose that $s < 0$. Then
$$s^T x \leq 0$$
for each $x \in D$ and with equality for $x = 0 \in D$. Therefore,
$$g^*(s) = 0$$
in this case.
- Suppose that $s < 0$ does not hold, i.e. the vector s has at least one nonnegative element. Let j be any index such that
$$s_j = \max_{i=1,\dots,n} s_i$$
and $s_j \geq 0$. First, note that
$$\begin{aligned} g^*(s) &\geq s^T \underbrace{e_j}_{\in D} \\ &= s_j \end{aligned}$$
Next, let $x \in D$ such that $x \neq e_j$. Then
$$\begin{aligned} s^T x &= \sum_{i=1}^{n} s_i x_i \\ &= s_j x_j + \sum_{\substack{i=0 \\ i \neq j}}^{n} s_i \underbrace{x_i}_{\geq 0} \\ &\leq s_j x_j + \sum_{\substack{i=0 \\ i \neq j}}^{n} s_j x_i \\ &= s_j x_j + s_j \sum_{\substack{i=0 \\ i \neq j}}^{n} x_i \\ &= \underbrace{s_j}_{\geq 0} \underbrace{\sum_{i=1}^{n} x_i}_{\leq 1} \\ &\leq s_j. \end{aligned}$$
Hence, all points $x \in D \setminus \{e_j\}$ satisfy $s^T x \leq s_j$ and the point $e_j \in D$ satisfy $s^T e_j = s_j$. Therefore,
$$g^*(s) = \max_{i=1,\dots,n} s_i$$
in this case.

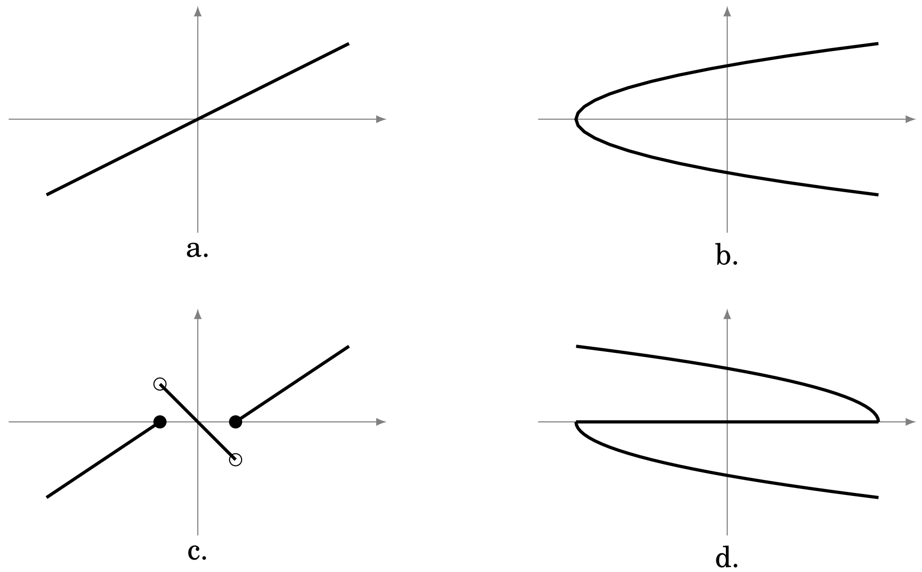
We conclude that
$$\begin{aligned} g^*(s) &= \sup_{x \in D} s^T x \\ &= \max\left(0, \max_{i=1,\dots,n} s_i\right), \end{aligned}$$
for each $s = (s_1, \dots, s_n) \in \mathbb{R}^n$, as desired.

4. The function g is proper, closed and convex. Therefore,

$$g^{**} = g = \iota_D.$$

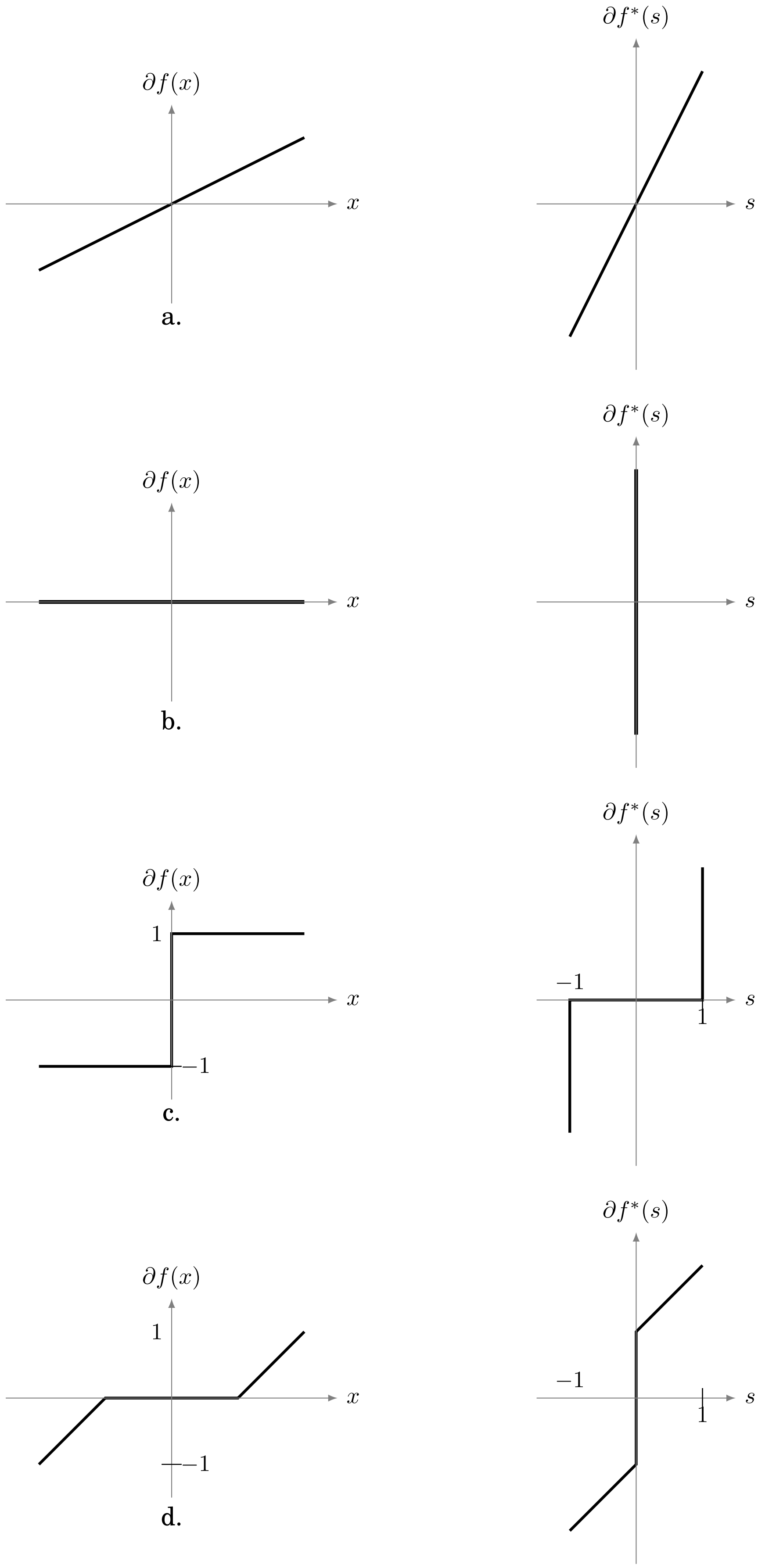
Solution 3.10

1. See the figure below. Since we are dealing with set valued mappings it is no problem if the inverses are set valued, i.e. we do not need to care about surjectivity and injectivity. The axis of the graphs are simply flipped.
2. Only a. and b. are functions. The other are set-valued.
3. Only the inverses of operators a. and c. are functions. The other are set-valued.



Solution 3.11

Since $\partial f^* = (\partial f)^{-1}$, we can flip the figures as follows:


∂f(x)
x
a.
∂f*(s)
s
∂f(x)
x
b.
∂f*(s)
s
∂f(x)
1
x
−1
c.
∂f*(s)
−1
s
1
∂f(x)
1
x
−1
d.
∂f*(s)
−1
s
1

Solution 3.12

Recall that

$$\operatorname{prox}_{\gamma f}(z) = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left(f(x) + \frac{1}{2\gamma} \|x - z\|_2^2 \right)$$

for each $z \in \mathbb{R}^n$. Let $z \in \mathbb{R}^n$. By Fermat's rule, $x = \operatorname{prox}_{\gamma f}(z)$ if and only if

$$\begin{gathered} 0 \in \partial f(x) + \gamma^{-1}(x - z) \\ \Leftrightarrow \\ z \in (I + \gamma \partial f)(x) \\ \Leftrightarrow \\ (I + \gamma \partial f)^{-1}(z) = x. \end{gathered}$$

We have equality in the last step since we know that the prox is single-valued for proper closed convex functions. Therefore,

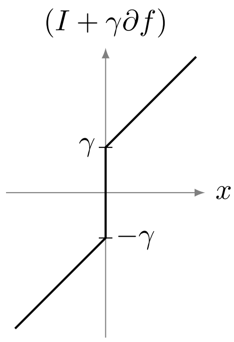
$$\operatorname{prox}_{\gamma f}(z) = (I + \gamma \partial f)^{-1}(z)$$

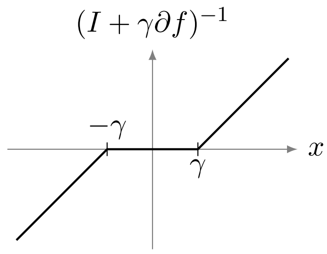
for each $z \in \mathbb{R}^n$, as desired.

Solution 3.13

1. We will solve this graphically. Left plot shows $I + \gamma \partial f$ and the right shows $(I + \gamma \partial f)^{-1} = \operatorname{prox}_{\gamma f}$. Therefore,

$$\operatorname{prox}_{\gamma f}(x) = \begin{cases} x + \gamma & \text{if } x \leq -\gamma, \\ 0 & \text{if } x \in [-\gamma, \gamma], \\ x - \gamma & \text{if } x \geq \gamma. \end{cases}$$



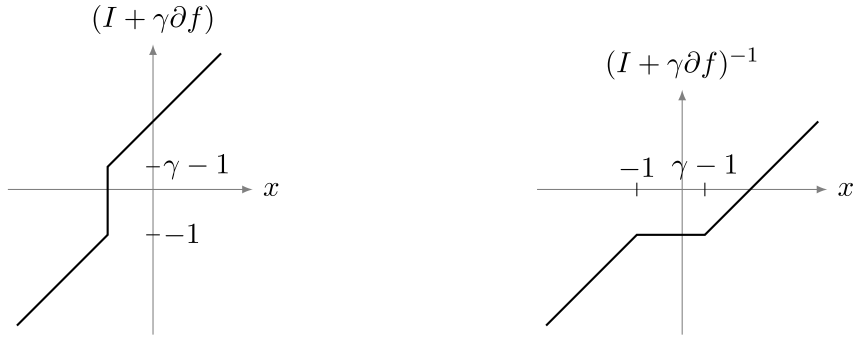


2. We will solve this graphically. Left plot shows $I + \gamma \partial f$ and the right shows $(I + \gamma \partial f)^{-1} = \operatorname{prox}_{\gamma f}$. The prox does not depend on γ (since it is actually a projection). Therefore,

$$\operatorname{prox}_{\gamma f}(x) = \begin{cases} -1 & \text{if } x \leq -1 \\ x & \text{if } x \in [-1, 1] \\ 1 & \text{if } x \geq 1 \end{cases}$$

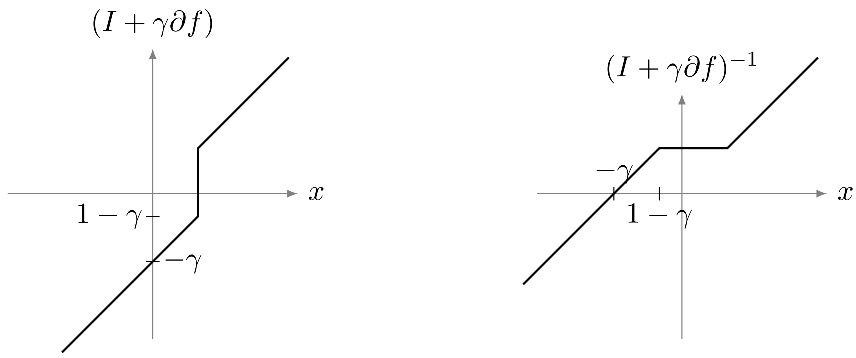
3. We will solve this graphically. Left plot shows $I + \gamma\partial f$ and the right shows $(I + \gamma\partial f)^{-1} = \mathrm{prox}_{\gamma f}$. Therefore,

$$
\mathrm{prox}_{\gamma f}(x) = \begin{cases} x & \text{if } x \leq -1, \\ -1 & \text{if } x \in [-1, \gamma - 1], \\ x - \gamma & \text{if } x \geq \gamma - 1. \end{cases}
$$



4. We will solve this graphically. Left plot shows $I + \gamma\partial f$ and the right shows $(I + \gamma\partial f)^{-1} = \mathrm{prox}_{\gamma f}$. Therefore,

$$
\mathrm{prox}_{\gamma f}(x) = \begin{cases} x + \gamma & \text{if } x \leq 1 - \gamma, \\ 1 & \text{if } x \in [1 - \gamma, 1], \\ x & \text{if } x \geq 1. \end{cases}
$$



Solution 3.14

1. Recall that

$$\operatorname{prox}_{\gamma f}(z) = \operatorname*{argmin}_{x\in\mathbb{R}^n} \left(f(x) + \frac{1}{2\gamma} \|x - z\|_2^2 \right)$$

for each $z \in \mathbb{R}^n$. Let $z \in \mathbb{R}^n$ and set $x = \operatorname{prox}_f(z)$. Let $u = z - x$. By Fermat's rule, $x = \operatorname{prox}_f(z)$ is equivalent to that

$$\begin{aligned}
0 \in \partial f(x) + x - z \quad &\Leftrightarrow \quad z - x \in \partial f(x) \quad \text{[subdifferential calculus rules]}\\
&\Leftrightarrow \quad x \in \partial f^*(z - x) \quad [f \text{ is proper closed convex}]\\
&\Leftrightarrow \quad z - u \in \partial f^*(u)\\
&\Leftrightarrow \quad 0 \in \partial f^*(u) + u - z\\
&\Leftrightarrow \quad u = \operatorname{prox}_{f^*}(z). \quad [\text{By Fermat's rule for } \operatorname{prox}_{f^*}(z)]
\end{aligned}$$

Therefore,

$$\begin{aligned}
z &= x + u\\
&= \operatorname{prox}_f(z) + \operatorname{prox}_{f^*}(z)
\end{aligned}$$

as desired.

2. We have

$$\begin{aligned}
(\gamma f^*)(s) &= \sup_{x\in\mathbb{R}^n} \left(s^T x - \gamma f(x)\right)\\
&= \gamma \sup_{x\in\mathbb{R}^n} \left((\gamma^{-1}s)^T x - f(x)\right)\\
&= \gamma f^*(\gamma^{-1}s)
\end{aligned}$$

for each $s \in \mathbb{R}^n$.

3. Let $z \in \mathbb{R}^n$ and set $u = \operatorname{prox}_{(\gamma f)^*}(z)$. Note that $u = \operatorname{prox}_{(\gamma f)^*}(z)$ is equivalent to that

$$\begin{aligned}
u &= \operatorname*{argmin}_{s\in\mathbb{R}^n} \left((\gamma f)^*(s) + \frac{1}{2}\|s - z\|_2^2 \right)\\
&= \operatorname*{argmin}_{s\in\mathbb{R}^n} \left(\gamma f^*(\gamma^{-1}s) + \frac{1}{2}\|s - z\|_2^2 \right)\\
&= \gamma \operatorname*{argmin}_{v\in\mathbb{R}^n} \left(\gamma f^*(v) + \frac{1}{2}\|\gamma v - z\|_2^2 \right)\\
&= \gamma \operatorname*{argmin}_{v\in\mathbb{R}^n} \left(\gamma f^*(v) + \frac{\gamma^2}{2}\|v - \gamma^{-1}z\|_2^2 \right)\\
&= \gamma \operatorname*{argmin}_{v\in\mathbb{R}^n} \left(\gamma^{-1} f^*(v) + \frac{1}{2}\|v - (\gamma^{-1}z)\|_2^2 \right)\\
&= \gamma \operatorname{prox}_{\gamma^{-1} f^*}(\gamma^{-1}z)
\end{aligned}$$

as desired.

4. Combine first and third subproblems.

Solution 3.15

Recall that the Moreau decomposition gives that

$$\mathrm{prox}_{(\gamma f)^*}(z) = z - \mathrm{prox}_{\gamma f}(z).$$

See Exercise 3.14.

1. Exercise 2.15-1 gives that the prox of γf is given by

$$\mathrm{prox}_{\gamma f}(z) = (I + \gamma H)^{-1}(z - \gamma h)$$

for each $z \in \mathbb{R}^n$, which by the Moreau decomposition implies that

$$\mathrm{prox}_{(\gamma f)^*}(z) = z - (I + \gamma H)^{-1}(z - \gamma h)$$

for each $z \in \mathbb{R}^n$.

2. Exercise 2.15-5 gives that the prox of γf is given by

$$\mathrm{prox}_{\gamma f}(z) = \begin{cases} z & \text{if } z < -1, \\ -1 & \text{if } z \in [-1, \gamma - 1], \\ z - \gamma & \text{if } z > \gamma - 1 \end{cases}$$

which by the Moreau decomposition implies that

$$\mathrm{prox}_{(\gamma f)^*}(z) = \begin{cases} 0 & \text{if } z < -1, \\ z + 1 & \text{if } z \in [-1, \gamma - 1], \\ \gamma & \text{if } z > \gamma - 1. \end{cases}$$

3. Exercise 2.15-6 gives that the prox of γf is given by

$$\mathrm{prox}_{\gamma f} = \begin{cases} z + \gamma & \text{if } z < 1 - \gamma, \\ 1 & \text{if } z \in [1 - \gamma, 1], \\ z & \text{if } z > 1 \end{cases}$$

which by the Moreau decomposition implies that

$$\mathrm{prox}_{(\gamma f)^*}(z) = \begin{cases} -\gamma & \text{if } z < 1 - \gamma, \\ z - 1 & \text{if } z \in [1 - \gamma, 1], \\ 0 & \text{if } z > 1. \end{cases}$$

Solution 3.16

1. Note that

$$\begin{aligned} -f^*(0) &= -\sup_{x \in \mathbb{R}^n} \left(0^T x - f(x)\right) \\ &= -\sup_{x \in \mathbb{R}^n} (-f(x)) \\ &= \inf_{x \in \mathbb{R}^n} f(x). \end{aligned}$$

2. Note that $f^{**} = f$, since f is proper closed convex. By the subdifferential formula for f^*, we have that

$$\begin{aligned}
\partial f^*(0) &= \operatorname*{Argmax}_{x\in\mathbb{R}^n}(0^T x - f^{**}(x)) \\
&= \operatorname*{Argmax}_{x\in\mathbb{R}^n}(-f(x)) \\
&= \operatorname*{Argmin}_{x\in\mathbb{R}^n} f(x).
\end{aligned}$$

Solution 3.17

1. The functions are proper closed convex and constraint qualification holds. Therefore, by Fermat's rule, $x \in \mathbb{R}^n$ is an optimal solution to the primal problem if and only if

$$\begin{gathered}
0 \in \partial f(x) + \partial g(x) \\
\Leftrightarrow \\
\begin{cases} \mu \in \partial f(x) \\ -\mu \in \partial g(x) \end{cases} \\
\Leftrightarrow \\
\begin{cases} x \in \partial f^*(\mu) \\ x \in \partial g^*(-\mu). \end{cases}
\end{gathered}$$

where $\mu \in \mathbb{R}^n$.

2. Eliminating x in the subproblem above gives that

$$\begin{gathered}
\begin{cases} x \in \partial f^*(\mu) \\ x \in \partial g^*(-\mu) \end{cases} \\
\Leftrightarrow \\
0 \in \partial f^*(\mu) - \partial g^*(-\mu).
\end{gathered}$$

3. In general no. Inspired by the condition

$$x \in \partial f^*(\mu)$$

you could use the subgradient selector s_{f^*} to generate a candidate solution

$$\hat{x} = s_{f^*}(\mu^\star) \in \partial f^*(\mu).$$

However, the full condition

$$\begin{cases} x \in \partial f^*(\mu^\star) \\ x \in \partial g^*(-\mu^\star) \end{cases}$$

need not necessarily hold for each $x \in \partial f^*(\mu^\star)$. I.e.

$$\hat{x} \in \partial f^*(\mu^\star) \not\Rightarrow \hat{x} \in \partial g^*(-\mu^\star).$$

If f^* is differentiable, we have that $\partial f^*(\mu) = \{\nabla f^*(\mu)\}$ for each $\mu \in \mathbb{R}^n$, since f^* is proper closed convex. This means that for every solution $\mu^\star$ to the dual condition (3.1), $x^\star$ is the unique point such that

$$\begin{cases} x^\star = \nabla f^*(\mu^\star) \\ x^\star \in \partial g^*(-\mu^\star). \end{cases}$$

In this case, the subgradient selector is the gradient and $\hat{x} = s_{f^*}(\mu^\star) = \nabla f^*(\mu^\star) = x^\star$ will recover the solution.

Solution 3.18

Fermat's rule gives that $x \in \mathbb{R}^n$ is an optimal solution to the primal problem (3.2) if and only if

$$0 \in \partial(f \circ L + g)(x). \tag{7.32}$$

Since $f \circ L$ and g are closed convex and $\operatorname{relint}\operatorname{dom}(f \circ L) \cap \operatorname{relint}\operatorname{dom} g \neq \emptyset$, the subdifferential sum rule gives that (7.32) is equivalent to

$$0 \in \partial(f \circ L)(x) + \partial g(x). \tag{7.33}$$

Moreover, since f is closed convex and $\operatorname{relint}\operatorname{dom}(f \circ L) \neq \emptyset$, the subdifferential composition rule gives that (7.33) is equivalent to

$$0 \in L^T \partial f(Lx) + \partial g(x).$$

This is equivalent to that there exits a point $\mu \in \mathbb{R}^m$ such that

$$\begin{cases} \mu \in \partial f(Lx) \\ -L^T\mu \in \partial g(x). \end{cases} \tag{7.34}$$

Since f and g are closed convex, we know that

$$(\partial f)^{-1} = \partial f^* \quad \text{and} \quad (\partial g)^{-1} = \partial g^*.$$

where f^* and g^* are the conjugate functions of f and g, respectively. Thus, (7.34) is equivalent to that

$$\begin{cases} Lx \in \partial f^*(\mu) \\ x \in \partial g^*(-L^T\mu). \end{cases}$$

This in turn is equivalent to that

$$0 \in \partial f^*(\mu) - L\partial g^*(-L^T\mu). \tag{7.35}$$

Note that it always holds that

$$-L\partial g^*(-L^T\mu) \subseteq \partial(g^* \circ -L^T)(\mu)$$

and

$$f^*(\mu) + \partial(g^* \circ -L^T)(\mu) \subseteq \partial(f^* + g^* \circ -L^T)(\mu).$$

This combined with (7.35), this implies that

$$0 \in \partial(f^* + g^* \circ -L^T)(\mu).$$

However, Fermat's rule gives that it is equivalent to that μ is an optimal solution to the optimization problem

$$\underset{\mu \in \mathbb{R}^m}{\text{minimize}}\, f^*(\mu) + g^*(-L^T\mu).$$

This is the Fenchel dual problem (3.3) we wanted to derive.

Solution 3.19

For the primal problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\, f(Lx) + g(x)$$

a Fenchel dual problem is

$$\underset{\mu \in \mathbb{R}^m}{\text{minimize}}\, f^*(\mu) + g^*(-L^T\mu). \tag{7.36}$$

1. Note that f and g are proper closed convex and constraint qualification holds, i.e.

   $$\text{relint dom}\,(f \circ L) \cap \text{relint dom}\, g \neq \emptyset$$

   for this particular case. Using Exercise 3.1-1 and 3.14-2, we have that

   $$f^*(\mu) = \frac{1}{2\lambda}\|\mu\|_2^2$$

   for each $\mu \in \mathbb{R}^m$. Exercise 3.1-6 gives that

   $$(\max(0, 1 - \cdot))^*(x_i) = x_i + \iota_{[-1,0]}(x_i)$$

   for each $x_i \in \mathbb{R}$. However, since $\max(0, 1 - \cdot)$ is proper closed convex, we know that

   $$\begin{aligned}(I + \iota_{[-1,0]})^*(\nu_i) &= (\max(0, 1 - \cdot))^{**}(\nu_i)\\ &= (\max(0, 1 - \cdot))(\nu_i)\\ &= \max(0, 1 - \nu_i)\end{aligned}$$

   for each $\nu_i \in \mathbb{R}$. Combining this with Exercise 3.5, we have that

   $$g^*(\nu) = \sum_{i=1}^{n} \max(0, 1 - \nu_i)$$

   for each $\nu = (\nu_1, \ldots, \nu_n) \in \mathbb{R}^n$.

Therefore, (7.36) becomes

$$\underset{\mu\in\mathbb{R}^m}{\text{minimize}}\ \frac{1}{2\lambda}\left\|\mu\right\|_2^2+\sum_{i=1}^{n}\max\left(0,1+\left(L^T\mu\right)_i\right) \tag{7.37}$$

in this case.

Note that f^* and g^* are proper closed convex, and that constraint qualification holds for (7.37), i.e.

$$\operatorname{relint}\operatorname{dom}\left(f^*\right)\cap\operatorname{relint}\operatorname{dom}g^*\circ -L^T\neq\emptyset.$$

Therefore, if $\mu\in\mathbb{R}^m$ is an optimal solution to (7.37), we can recover an optimal solution $x\in\mathbb{R}^n$ to the primal problem by considering any one of the primal dual necessary and sufficient optimality conditions. In particular, it must holds that

$$\begin{cases} Lx\in\partial f^*(\mu) \\ x\in\partial g^*\left(-L^T\mu\right). \end{cases} \tag{7.38}$$

Note that f^* is differentiable with gradient

$$\nabla f^*=\frac{1}{\lambda}I$$

Therefore, the first condition in (7.38) uniquely determines x, i.e.

$$x=\frac{1}{\lambda}L^{-1}\mu$$

since $\partial f^*(\mu)=\{\nabla f^*(\mu)\}$ and this x must then automatically fulfill the second condition in (7.38).

2. Note that f and g are proper closed convex and constraint qualification holds, i.e.

$$\operatorname{relint}\operatorname{dom}\left(f\circ L\right)\cap\operatorname{relint}\operatorname{dom}g\neq\emptyset$$

for this particular case. Using Exercise 3.6, we have that

$$f^*(\mu)=\left\|\mu\right\|_1$$

for each $\mu\in\mathbb{R}^m$. Using Exercise 3.1-2, we have that

$$g^*(\nu)=\frac{1}{2\lambda}\left\|\nu+b\right\|_2^2$$

for each $\nu\in\mathbb{R}^n$.

Therefore, (7.36) becomes

$$\underset{\mu\in\mathbb{R}^m}{\text{minimize}}\ \left\|\mu\right\|_1+\frac{1}{2\lambda}\left\|-L^T\nu+b\right\|_2^2 \tag{7.39}$$

in this case.

Note that f^* and g^* are proper closed convex, and that constraint qualification holds for (7.39), i.e.

$$\operatorname{relint}\operatorname{dom}\left(f^*\right)\cap\operatorname{relint}\operatorname{dom}g^*\circ -L^T\neq\emptyset.$$

Therefore, if $\mu \in \mathbb{R}^m$ is an optimal solution to (7.39), we can recover an optimal solution $x \in \mathbb{R}^n$ to the primal problem by considering any one of the primal dual necessary and sufficient optimality conditions. In particular, it must holds that

$$\begin{cases} Lx \in \partial f^*(\mu) \\ x \in \partial g^*\left(-L^T\mu\right). \end{cases} \tag{7.40}$$

Note that g^* is differentiable with gradient

$$\nabla g^*(\nu) = \frac{1}{\lambda}(\nu + b)$$

for each $\nu \in \mathbb{R}^n$. Therefore, the second condition in (7.40) uniquely determines x, i.e.

$$x = \frac{1}{\lambda}(-L^T\mu + b)$$

since $\partial g^*(-L^T\mu) = \{\nabla g^*(-L^T\mu)\}$ and this x must then automatically fulfill the first condition in (7.40).

Solution 3.20

1. By definition, we have

$$\begin{aligned} f^*(s) &= \sup_{z \in \mathbb{R}^n} \left(s^T z - f(z)\right) \\ &\geq s^T x - f(x) \end{aligned}$$

as desired.

2. Suppose that $s \in \partial f(x)$. This implies that $f(x) < \infty$. We have that

$$\begin{gathered} s \in \partial f(x) \\ \Leftrightarrow \\ f(y) \geq f(x) + s^T(y - x) \quad \text{for each } y \in \mathbb{R}^n \\ \Leftrightarrow \\ s^T x - f(x) \geq s^T y - f(y) \quad \text{for each } y \in \mathbb{R}^n \\ \Leftrightarrow \\ s^T x - f(x) \geq \sup_{y \in \mathbb{R}^n} \left(s^T y - f(y)\right) \\ \Leftrightarrow \\ s^T x - f(x) \geq f^*(s) \\ \Leftrightarrow \\ f^*(s) \leq s^T x - f(x) \end{gathered}$$

as desired.

3. Suppose that $f^*(s) = s^T x - f(x)$. This implies that $f^*(s) \leq s^T x - f(x)$. However, the above sequence of equivalences gives that $s \in \partial f(x)$, as as desired.

Solution 3.21

1. Suppose that $s \in \partial f(x)$. The function f is then proper. Fenchel-Young's equality (see Exercise 3.20) gives that

$$f^*(s) = s^T x - f(x).$$

We know that $f^{**} \leq f$ (see Exercise 3.2). We get that

$$0 = f^*(s) + f(x) - s^T x \geq f^*(s) + f^{**}(x) - s^T x \geq 0,$$

where the last inequality follows from Fenchel Young's inequality (see Exercise 3.20-1). Thus,

$$f^{**}(x) = s^T x - f^*(s),$$

which is equivalent to $x \in \partial f^*(s)$ by Fenchel-Young's equality.

2. Apply the previous result to f^*.
3. Combine the above the results and that $f^{**} = f$ for proper closed convex f.

Solution 3.22
Define $h : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ such that

$$h(y) = f(y + c)$$

for each $y \in \mathbb{R}^m$. Then $g = h \circ L$. Let $s \in \mathbb{R}^n$. We have that

$$\begin{aligned} g^*(s) &= \sup_{x \in \mathbb{R}^n} \big(s^T x - h(Lx)\big) \\ &= -\inf_{x \in \mathbb{R}^n} (h(Lx) + l_s(x)), \end{aligned} \tag{7.41}$$

where $l_s : \mathbb{R}^n \to \mathbb{R}$ is given by

$$l_s(x) = -s^T x$$

for each $x \in \mathbb{R}^n$. Note that

$$\begin{aligned} h^*(\mu) &= \sup_{y \in \mathbb{R}^m} \big(\mu^T y - f(y + c)\big) \\ &= \sup_{v \in \mathbb{R}^m} \big(\mu^T (v - c) - f(v)\big) \\ &= \sup_{v \in \mathbb{R}^m} \big(\mu^T v - f(v)\big) - \mu^T c \\ &= f^*(\mu) - \mu^T c \end{aligned}$$

for each $\mu \in \mathbb{R}^m$, and

$$\begin{aligned} l_s^*(\nu) &= \sup_{x \in \mathbb{R}^n} \big(\nu^T x + s^T x\big) \\ &= \sup_{x \in \mathbb{R}^n} \big((\nu + s)^T x\big) \\ &= \iota_{\{0\}}(\nu + s). \end{aligned}$$

for each $\mu \in \mathbb{R}^n$.

Consider the minimize problem in (7.41). We have that h and l_s are proper closed convex and that constraint qualification is satisfied since

$$\begin{aligned}\operatorname{relint}\operatorname{dom} h \circ L \cap \operatorname{relint}\operatorname{dom} l_s &= \operatorname{relint}\operatorname{dom} g \cap \operatorname{relint} \mathbb{R}^n \\ &= \operatorname{relint}\operatorname{dom} g \cap \mathbb{R}^n \\ &= \operatorname{relint}\operatorname{dom} g \\ &\neq \emptyset.\end{aligned}$$

Moreover, by assumption, we know that there is an $x_s \in \mathbb{R}^n$ that achieves the infimum in (7.41). Therefore, strong duality must hold, and we get

$$\begin{aligned}g^*(s) &= -\inf_{x\in\mathbb{R}^n} \left(h(Lx) + l_s(x)\right) \\ &= -\sup_{\mu\in\mathbb{R}^m} \left(-h^*(\mu) - l_s^*\left(-L^T\mu\right)\right) \\ &= \inf_{\mu\in\mathbb{R}^m} \left(h^*(\mu) + l_s^*\left(-L^T\mu\right)\right) \\ &= \inf_{\substack{\mu\in\mathbb{R}^m \\ \text{s.t. } s=L^T\mu}} \left(f^*(\mu) - c^T\mu\right).\end{aligned}$$

Solution 3.23

Since f is closed convex, we have that $f(x) = f^{**}(x) = \sup_{s\in\mathbb{R}^n} \left(x^T s - f^*(s)\right)$ for each $x \in \mathbb{R}^n$. Therefore,

$$\sup_{x\in\mathbb{R}^n} \left(f(x) - g(x)\right),$$

is equal to

$$\sup_{x\in\mathbb{R}^n} \sup_{s\in\mathbb{R}^n} \left(x^T s - f^*(s) - g(x)\right).$$

However, we may switch the supremums to get the equal problem

$$\sup_{s\in\mathbb{R}^n} \sup_{x\in\mathbb{R}^n} \left(x^T s - g(x) - f^*(s)\right).$$

But this is equal to

$$\sup_{s\in\mathbb{R}^n} \left(g^*(s) - f^*(s)\right),$$

since $g^*(s) = \sup_{x\in\mathbb{R}^n} \left(x^T s - g(x)\right)$ for each $s \in \mathbb{R}^n$. This completes the proof.

Solutions to chapter 4

Solution 4.1
That $x^\star$ is a fixed-point means that

$$x^\star = x^\star - \gamma \nabla f(x^\star).$$

This is equivalent to that

$$0 = \nabla f(x^\star).$$

Exercise 1.28 gives that $x^\star$ is a global minimizer of f.

Solution 4.2
Note that

$$\begin{aligned} z &:= \mathrm{prox}_{\gamma f}(x) \\ &= \underset{y \in \mathbb{R}^n}{\mathrm{argmin}} \left(f(y) + \frac{1}{2\gamma} \|y - x\|_2^2 \right). \end{aligned}$$

Fermat's rule and subdifferential calculus rules give that z satisfies

$$0 \in \partial f(z) + \gamma^{-1}(z - x).$$

The fixed-point assumption $z = x$ gives that

$$0 \in \partial f(x).$$

Fermat's rule gives that x is a global minimizer of f.

Solution 4.3
Note that

$$\begin{aligned} z :&= \mathrm{prox}_{\gamma g}(x - \gamma \nabla f(x)) \\ &= \underset{z \in \mathbb{R}^n}{\mathrm{argmin}} \left(g(z) + \frac{1}{2\gamma} \|z - (x - \gamma \nabla f(x))\|_2^2 \right). \end{aligned}$$

Fermat's rule and subdifferential calculus rules give that z satisfies

$$0 \in \partial g(z) + \frac{1}{\gamma} \left(z - (x - \gamma \nabla f(x)) \right).$$

The fixed-point assumption $z = x$ gives that

$$0 \in \partial g(x) + \nabla f(x). \tag{7.42}$$

The subdifferential sum rule gives that

$$\begin{aligned} \partial g(x) + \nabla f(x) &= \partial g(x) + \partial f(x) \\ &= \partial(g + f)(x) \end{aligned}$$

since $\partial f(x) = \{\nabla f(x)\}$. This combined with (7.42) gives that

$$0 \in \partial(g + f)(x).$$

Fermat's rule gives that x is a global minimizer of $f + g$.

Solution 4.4

1. The function is smooth so the gradient method works. No need to use the proximal gradient method.
2. The first two parts are smooth. The third part is not smooth but is separable and therefore prox friendly. Thus, the proximal gradient method works but the gradient method does not.
3. Both parts are smooth and the second part is separable and therefore prox friendly. Thus, the gradient method and the proximal gradient method both work.
4. First part is smooth. The second part is prox friendly but not smooth. Thus, the proximal gradient method works but not the gradient method.
5. Neither of the functions are differentiable, so none of the methods work.
6. The first part is differentible, but not smooth (it grows too quick for large x), and the second is prox friendly but not differentiable. Thus, none of the methods work.
7. First part is smooth. The second part is not smooth but is separable and therefore prox friendly. Thus, the proximal gradient method works but not the gradient method.
8. The second part is neither smooth nor prox friendly. Thus, none of the methods work.
9. Both parts are smooth and the second part is separable and therefore prox friendly. Thus, the gradient method and the proximal gradient method both work.

Solution 4.5

1. The part $\|Ax - b\|_2^2$ is strongly convex if and only if A^TA is invertible. Since $A \in \mathbb{R}^{m\times n}$ with $m < n$, A^TA has at most rank m and is therefore not invertible. Therefore, the primal objective is not strongly convex. The dual objective will therefore not be smooth. Thus, neither of the methods work.

2. The part $\frac{1}{2}x^TQx + b^Tx$ is strongly convex since $Q \in \mathbb{S}_{++}^{+}$, and therefore has a smooth conjugate. The conjugate of $\|x\|_1$ is prox friendly but not smooth. Thus, the proximal gradient method works but not the gradient method.

3. The first part is not strongly convex and will therefore have a nonsmooth conjugate. The conjugate of the first part is not prox friendly. However, if we let $f(y) = \frac{1}{2}\|y - b\|_2^2$ and $g(x) = \|x\|_2^2$, the problem can be written as

$$\min_{x\in\mathbb{R}^n} f(Ax) + g(x)$$

and a dual can be written

$$\min_{\mu\in\mathbb{R}^m} f^*(\mu) + g^*(-A^T\mu).$$

The function f^* is convex, smooth, separable and therefore prox friendly. The function $\mu \mapsto g^*(-A^T\mu)$ is smooth. Thus, the gradient method and the proximal gradient method both work.

4. The first part is not strongly convex and will therefore have a nonsmooth conjugate. The conjugate of the first part is not prox friendly. Doing the same trick as for the previous problem does not work since $\|x\|_2$ is not strongly convex and therefore it has a nonsmooth conjugate. Thus, neither of the methods work.

5. Neither part is strongly convex, therefore neither of the conjugates are smooth. Thus, neither of the methods work.

6. Neither part is strongly convex ($e^{\|x\|^4} \approx 1 + \|x^4\|$ for small x), therefore neither of the conjugates are smooth. Thus, neither of the methods work.

7. The first part is strongly convex and will therefore have a smooth conjugate. The second part is proximable, and therefore the same is true for the dual. However, the second part is not strongly convex and will therefore have a nonsmooth conjugate. Thus, the proximal gradient method works but not the gradient method.

8. With $f = \iota_{[-\mathbf{1},\mathbf{1}]}$ and $g(x) = \frac{1}{2}x^TQx$, the primal problem can written as

$$\min_{x\in\mathbb{R}^n} f(Lx) + g(x)$$

and has a dual

$$\min_{\mu\in\mathbb{R}^m} f^*(\mu) + g^*(-L^T\mu)$$

where $g^*(\mu) = \frac{1}{2}x^TQ^{-1}x$. Note that $\mu \mapsto g^*(-L\mu)$ is smooth. The function f^* is prox friendly but not smooth. Thus, the proximal gradient method works but not the gradient method.

9. Neither part is strongly convex, therefore neither of the conjugates are smooth. Thus, neither of the methods work.

Solution 4.6

1. Exercise 3.6-1 gives that

$$f^* = \iota_{[-\mathbf{1},\mathbf{1}]}$$

2. Exercise 3.1-1 gives that

$$g^*(\mu) = \tfrac{1}{2}\mu^T Q^{-1}\mu$$

for each $\mu \in \mathbb{R}^n$.

3. One possible dual problem is given by

$$\underset{\mu\in\mathbb{R}^n}{\text{minimize}}\, f^*(\mu) + g^*(-\mu)$$

(e.g., let $L = I$ in Exercise 3.18). Similarly, another dual problem is given by

$$\underset{\mu\in\mathbb{R}^n}{\text{minimize}}\, f^*(-\mu) + g^*(\mu).$$

In the remainder of the exercise, we will only consider the first dual problem.

4. Under the assumptions on f and g, we know that f^* is closed, convex and proximable, and g^* is closed convex and smooth. Therefore, for the dual problem

$$\underset{\mu\in\mathbb{R}^n}{\text{minimize}}\, f^*(\mu) + g^*(-\mu),$$

we get, for some appropriate $\gamma_k > 0$, that

$$\mu_{k+1} = \mathrm{prox}_{\gamma_k f^*}\left(\mu_k - \gamma_k \nabla(g^* \circ -I)(\mu_k)\right),$$

is a computationally reasonable step for the proximal gradient method.

5. Consider our particular choice of f^* and g^*. Differentiation yields

$$\nabla(g^* \circ -I)(\mu_k) = -\nabla g^*(-\mu_k) = Q^{-1}\mu_k.$$

By definition, the proximal operator of f^* is

$$\begin{aligned}
\mathrm{prox}_{\gamma_k f^*}(z) &= \underset{\mu\in\mathbb{R}^n}{\text{argmin}}\left(\iota_{[-\mathbf{1},\mathbf{1}]}(\mu) + \frac{1}{2\gamma_k}\|\mu - z\|_2^2\right)\\
&= \underset{\mu\in[-\mathbf{1},\mathbf{1}]}{\text{argmin}}\, \|\mu - z\|_2^2\\
&= \underset{\mu\in[-\mathbf{1},\mathbf{1}]}{\text{argmin}} \sum_{i=1}^{n}(\mu_i - z_i)^2
\end{aligned}$$

for each $z = (z_1,\ldots,z_n) \in \mathbb{R}^n$. Note that both the constraint set and the objective function of this argmin-problem are separable, yielding

$$\mathrm{prox}_{\gamma_k f^*}(z) = \begin{bmatrix} \text{argmin}_{\mu_1\in[-1,1]}(\mu_1 - z_1)^2 \\ \vdots \\ \text{argmin}_{\mu_1\in[-1,1]}(\mu_1 - z_1)^2 \end{bmatrix}$$

for each $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$. Note that

$$\underset{\mu_i \in [-1,1]}{\operatorname{argmin}} (\mu_i - z_i)^2 = \begin{cases} 1 & \text{if } z_i > 1 \\ -1 & \text{if } z_i < -1 \\ z_i & \text{otherwise} \end{cases}$$

for each $z_i \in \mathbb{R}$ and each $i = 1, \ldots, n$. Thus, the proximal gradient method step for the dual problem becomes

$$\begin{cases} v_k = \mu_k - \gamma_k Q^{-1} \mu_k \\ (\mu_{k+1})_i = \begin{cases} 1 & \text{if } (v_k)_i > 1, \\ -1 & \text{if } (v_k)_i < -1, \\ (v_k)_i & \text{otherwise,} \end{cases} \quad \forall i \in \{1, \ldots, n\}. \end{cases}$$

Solution 4.7

We start with

$$\mu_{k+1} = \operatorname{prox}_{\gamma_k f^*} \left(\mu_k - \gamma_k \nabla (g^* \circ -L^T)(\mu_k) \right).$$

Note that $\nabla(g^* \circ -L^T)(\mu_k) = -L \nabla g^*(-L^T \mu_k)$. Therefore, the proximal gradient method step can be rewritten as

$$\begin{cases} x_k = \nabla g^*(-L^T \mu_k), \\ v_k = \mu_k + \gamma_k L x_k, \\ \mu_{k+1} = \operatorname{prox}_{\gamma_k f^*} (v_k). \end{cases}$$

Using Moreau decomposition, we have

$$\operatorname{prox}_{\gamma_k f^*}(z) = z - \gamma_k \operatorname{prox}_{\gamma_k^{-1} f^{**}}(\gamma_k^{-1} z) = z - \gamma_k \operatorname{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} z).$$

for each $z \in \mathbb{R}^m$. The last equality holds since $f = f^{**}$, by closed convexity of the proper function f. Using this, we can write the proximal gradient method step as

$$\begin{cases} x_k = \nabla g^*(-L^T \mu_k), \\ v_k = \mu_k + \gamma_k L x_k, \\ \mu_{k+1} = v_k - \gamma_k \operatorname{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} v_k). \end{cases}$$

Recall the subdifferential formula for g^*, i.e.

$$\begin{aligned} \partial g^*(\mu) &= \underset{x \in \mathbb{R}^n}{\operatorname{Argmax}} \left(\mu^T x - g^{**}(x) \right) \\ &= \underset{x \in \mathbb{R}^n}{\operatorname{Argmax}} \left(\mu^T x - g(x) \right) \end{aligned}$$

for each $\mu \in \mathbb{R}^n$. The last equality holds since $g = g^{**}$, by closed convexity of the proper function g. However, we know that g^* is smooth and convex, and therefore, $\partial g^*(\mu) = \{\nabla g^*(\mu)\}$. Using this, we get that

$$\begin{aligned} \nabla g^*(\mu) &= \underset{x \in \mathbb{R}^n}{\operatorname{argmax}} \left(\mu^T x - g(x) \right) \\ &= \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \left(g(x) - \mu^T x \right) \end{aligned}$$

for each $\mu \in \mathbb{R}^n$. This lets us write the proximal gradient method step as

$$\begin{cases} x_k = \operatorname{argmin}_{x\in\mathbb{R}^n} \left(g(x) + \mu_k^T Lx\right), \\ v_k = \mu_k + \gamma_k Lx_k, \\ \mu_{k+1} = v_k - \gamma_k \operatorname{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} v_k), \end{cases}$$

as desired.

Solution 4.8

Recall that Exercise 4.6 gives the dual proximal gradient method step

$$\begin{cases} v_k = \mu_k - \gamma_k Q^{-1}\mu_k \\ (\mu_{k+1})_i = \begin{cases} 1 & \text{if } (v_k)_i > 1, \\ -1 & \text{if } (v_k)_i < -1, \\ (v_k)_i & \text{otherwise}, \end{cases} \quad \forall i \in \{1,\dots,n\}. \end{cases} \tag{7.43}$$

We must verify that

$$\begin{cases} x_k = \operatorname{argmin}_x \left(g(x) + \mu_k^T x\right), \\ v_k = \mu_k + \gamma_k x_k, \\ \mu_{k+1} = v_k - \gamma_k \operatorname{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} v_k), \end{cases} \tag{7.44}$$

gives the same step when $f, g : \mathbb{R}^n \to \mathbb{R}$ are given as

$$f(x) = \|x\|_1 = \sum_{i=1}^n |x_i| \quad \text{and} \quad g(x) = \frac{1}{2}x^T Q x$$

for each $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. To verify correctness, note that

$$\begin{aligned} \operatorname*{argmin}_{x\in\mathbb{R}^n} \left(g(x) + \mu_k^T x\right) &= \operatorname*{argmin}_{x\in\mathbb{R}^n} \left(\frac{1}{2}x^T Q x + x^T \mu_k\right) \\ &= -Q^{-1}\mu_k. \end{aligned}$$

Thus, we can write (7.44) as

$$\begin{cases} v_k = \mu_k - \gamma_k Q^{-1}\mu_k, \\ \mu_{k+1} = v_k - \gamma_k \operatorname{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} v_k). \end{cases} \tag{7.45}$$

Since f is separable, so is $\operatorname{prox}_{\gamma f}$, see Exercise 2.16. From Exercise 2.15-3 we get that

$$\left(\operatorname{prox}_{\gamma f}(z)\right)_i = \begin{cases} z_i + \gamma & \text{if } z_i < -\gamma, \\ 0 & \text{if } -\gamma \le z_i \le \gamma, \\ z_i - \gamma & \text{if } z_i > \gamma, \end{cases} \quad \forall i \in \{1,\dots,n\}.$$

We can then calculate μ_{k+1} in (7.45) as

$$
\begin{aligned}
(\mu_{k+1})_i &= (v_k)_i - \gamma_k \left(\operatorname{prox}_{\gamma_k^{-1} f}\left(\gamma_k^{-1} v_k\right)\right)_i \\
&= (v_k)_i - \gamma_k \begin{cases} \gamma_k^{-1}(v_k)_i + \gamma_k^{-1} & \text{if } \gamma_k^{-1}(v_k)_i < -\gamma_k^{-1}, \\ 0 & \text{if } -\gamma_k^{-1} \leq \gamma_k^{-1}(v_k)_i \leq \gamma_k^{-1}, \\ \gamma_k^{-1}(v_k)_i - \gamma_k^{-1} & \text{if } \gamma_k^{-1}(v_k)_i > \gamma_k^{-1}, \end{cases} \\
&= (v_k)_i - \begin{cases} (v_k)_i + 1 & \text{if } (v_k)_i < -1, \\ 0 & \text{if } -1 \leq (v_k)_i \leq 1, \\ (v_k)_i - 1 & \text{if } (v_k)_i > 1, \end{cases} \\
&= \begin{cases} -1 & \text{if } (v_k)_i < -1, \\ (v_k)_i & \text{if } -1 \leq (v_k)_i \leq 1, \\ 1 & \text{if } (v_k)_i > 1, \end{cases} \quad \forall i \in \{1, \ldots, n\}.
\end{aligned}
$$

This establishes the desired equality.

Solution 4.9

Using the hint with $x = x_k$ we get that

$$
\begin{aligned}
f(y) &\leq f(x_k) + \nabla f(x_k)^T (y - x_k) + \frac{\beta}{2} \|y - x_k\|_2^2 \\
&< f(x_k) + \nabla f(x_k)^T (y - x_k) + \frac{1}{2\gamma_k} \|y - x_k\|_2^2
\end{aligned}
$$

for each $y \in \mathbb{R}^n$. The function $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ given by

$$
g(y) = f(x_k) + \nabla f(x_k)^T (y - x_k) + \frac{1}{2\gamma_k} \|y - x_k\|_2^2
$$

for each $y \in \mathbb{R}^n$ is then a majorizer to f, i.e., $f \leq g$. What remain to be shown is that

$$
x_{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^n} g(y).
$$

By Fermat's rule and convex differentiability of g, we know this holds if and only if

$$
\nabla g(x_{k+1}) = 0.
$$

Straight forward calculations show that this is equivalent to

$$
x_{k+1} = x_k - \gamma_k \nabla f(x_k)
$$

as desired.

Solutions to chapter 5

Solution 5.1

Rearranging the objective function in (5.1) yields

$$
\begin{aligned}
&\sum_{i=1}^{N} \log\left(1+e^{-y_i\left(x_i^T w+b\right)}\right) \\
&= \sum_{\substack{i=1 \\ \text{s.t. } y_i=-1}}^{N} \log\left(1+e^{x_i^T w+b}\right) + \sum_{\substack{i=1 \\ \text{s.t. } y_i=1}}^{N} \log\left(1+e^{-\left(x_i^T w+b\right)}\right) \\
&= \sum_{\substack{i=1 \\ \text{s.t. } y_i=-1}}^{N} \log\left(1+e^{w^T x_i+b}\right) + \sum_{\substack{i=1 \\ \text{s.t. } y_i=1}}^{N} \log\left(\frac{1+e^{w^T x_i+b}}{e^{w^T x_i+b}}\right) \\
&= \sum_{\substack{i=1 \\ \text{s.t. } y_i=-1}}^{N} \log\left(1+e^{w^T x_i+b}\right) + \sum_{\substack{i=1 \\ \text{s.t. } y_i=1}}^{N} \log\left(1+e^{w^T x_i+b}\right) - \sum_{\substack{i=1 \\ \text{s.t. } y_i=1}}^{N} \log\left(e^{w^T x_i+b}\right) \\
&= \sum_{i=1}^{N} \log\left(1+e^{w^T x_i+b}\right) - \sum_{\substack{i=1 \\ \text{s.t. } y_i=1}}^{N} \log\left(e^{w^T x_i+b}\right) \\
&= \sum_{i=1}^{N} \log\left(1+e^{w^T x_i+b}\right) - \sum_{\substack{i=1 \\ \text{s.t. } y_i=1}}^{N} w^T x_i + b.
\end{aligned}
$$

From here, we can go over to the new labels, $y_i = 1 \to \hat{y}_i = 1$ and $y_i = -1 \to \hat{y}_i = 0$. We get that

$$
\begin{aligned}
&\sum_{i=1}^{N} \log\left(1+e^{w^T x_i+b}\right) - \sum_{\substack{i=1 \\ \text{s.t. } y_i=1}}^{N} w^T x_i + b \\
&= \sum_{i=1}^{N} \log\left(1+e^{w^T x_i+b}\right) - \sum_{i=1}^{N} \hat{y}_i \left(w^T x_i + b\right) \\
&= \sum_{i=1}^{N} \left(\log\left(1+e^{w^T x_i+b}\right) - \hat{y}_i \left(w^T x_i + b\right)\right)
\end{aligned}
$$

as desired.

Solution 5.2

Note that each term in the sum in (5.2) is positive for each $(w, b) \in \mathbb{R}^n \times \mathbb{R}$. Why is this true? Well, let $(w, b) \in \mathbb{R}^n \times \mathbb{R}$, $i = 1, \ldots, N$ and $u_i = x_i^T w + b$. Then

$$\log\left(1 + e^{x_i^T w + b}\right) - y_i\left(x_i^T w + b\right) = \log\left(1 + e^{u_i}\right) - y_i u_i \tag{7.46}$$

For $y_i = 0$, (7.46) becomes

$$\log\left(1 + e^{u_i}\right) > 0$$

since $1 + e^{u_i} > 1$. For $y_i = 1$, (7.46) becomes

$$\log\left(1 + e^{u_i}\right) - u_i = \log\left(\frac{1 + e^{u_i}}{e^{u_i}}\right) > 0$$

since $\frac{1+e^{u_i}}{e^{u_i}} > 1$.

Therefore, the objective function in (5.2) is positive for each $(w, b) \in \mathbb{R}^n \times \mathbb{R}$, since it is a sum of positive terms.

Let $(w, b) = t(\bar{w}, \bar{b})$ for some $t \in \mathbb{R}$. Suppose that $i = 1, \ldots, N$ is such that $y_i = 0$. Then

$$\begin{aligned}
\log\left(1 + e^{x_i^T w + b}\right) - y_i\left(x_i^T w + b\right) &= \log\left(1 + e^{t\left(x_i^T \bar{w} + \bar{b}\right)}\right) \\
&= \log\left(1 + \left(e^{x_i^T \bar{w} + \bar{b}}\right)^t\right) \to 0 \quad \text{as} \quad t \to \infty.
\end{aligned}$$

The limit above follows from $e^{x_i^T \bar{w} + \bar{b}} \in (0, 1)$, since $x_i^T \bar{w} + \bar{b} < 0$ by assumption on i. Suppose instead that $i = 1, \ldots, N$ is such that $y_i = 1$. Then

$$\begin{aligned}
\log\left(1 + e^{x_i^T w + b}\right) - y_i\left(x_i^T w + b\right) &= \log\left(1 + e^{t\left(x_i^T \bar{w} + \bar{b}\right)}\right) - t\left(x_i^T \bar{w} + \bar{b}\right) \\
&= \log\left(\frac{1 + e^{t\left(x_i^T \bar{w} + \bar{b}\right)}}{e^{t\left(x_i^T \bar{w} + \bar{b}\right)}}\right) \\
&= \log\left(1 + e^{-t\left(x_i^T \bar{w} + \bar{b}\right)}\right) \\
&= \log\left(1 + \left(e^{-\left(x_i^T \bar{w} + \bar{b}\right)}\right)^t\right) \to 0 \quad \text{as} \quad t \to \infty.
\end{aligned}$$

The limit above follows from $e^{-\left(x_i^T \bar{w} + \bar{b}\right)} \in (0, 1)$, since $x_i^T \bar{w} + \bar{b} > 0$ by assumption on i. In either case, the term goes to zero.

We conclude that the optimal value of (5.2) is 0, which is not attained, since the objective function is positive for each $(w, b) \in \mathbb{R}^n \times \mathbb{R}$.

Solution 5.3

First, consider the case when $\lambda = 0$. Then (5.3) becomes

$$\underset{x \in \mathbb{R}}{\text{minimize}}\ \frac{1}{2}\left\|ax - b\right\|_2^2.$$

By Fermat's rule, we get that the optimal point in this case is

$$0 = a^T\left(a x_{\text{ls}} - b\right) \quad \text{or} \quad x_{\text{ls}} = \frac{a^T b}{\|a\|_2^2}.$$

Now, consider the case when $\lambda > 0$. Using Fermat's rule and the subdifferential calculus rules (CQ holds since both functions have full effective domain), the optimality condition for (5.3) is given by

$$0 \in \|a\|_2^2 x - a^T b + \lambda \begin{cases} \operatorname{sgn}(x) & \text{if } x \neq 0, \\ [-1,1] & \text{if } x = 0. \end{cases}$$

Thus, $x = 0$ is an optimal point if and only if $a^T b \in [-\lambda, \lambda]$ or equivalently $\lambda \geq |a^T b|$. It remains to consider the case $\lambda < |a^T b|$. But then $x \neq 0$ by necessity, and x is an optimal point if and only if

$$0 = \|a\|_2^2 x - a^T b + \lambda \operatorname{sgn}(x) \quad \text{or} \quad x = \frac{a^T b}{\|a\|_2^2} - \frac{\lambda}{\|a\|_2^2} \operatorname{sgn}(x).$$

However, since $|a^T b| > \lambda$ by assumption, $\operatorname{sgn}(x) = \operatorname{sgn}(a^T b) = \operatorname{sgn}(x_{\text{ls}})$ must hold by necessity. Therefore, the solution in this case is given by

$$x = x_{\text{ls}} - \frac{\lambda}{\|a\|_2^2} \operatorname{sgn}(x_{\text{ls}}).$$

This concludes the proof.

Solution 5.4

- Alternative 1:
  Optimality conditions for (5.4) are given by

  $$0 \in A^T(Ax - b) + \lambda \begin{bmatrix} g(x_1) \\ \vdots \\ g(x_m) \end{bmatrix}$$

  where

  $$g(x_i) = \begin{cases} \{-1\} & \text{if } x_i < 0, \\ [-1,1] & \text{if } x_i = 0, \\ \{1\} & \text{if } x_i > 0. \end{cases}$$

  Thus, the optimality conditions above give that $x = 0$ is an optimal point to (5.4) if

  $$0 \in -A^T b + \lambda [-1,1]^m.$$

  This holds if and only if

  $$\begin{aligned} \lambda &\geq \max_{i=1,\ldots,m} |(A^T b)_i| \\ &= \|A^T b\|_\infty . \end{aligned}$$

- Alternative 2:
  Let $f : \mathbb{R}^m \to \mathbb{R}$ such that

  $$f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

  for each $x \in \mathbb{R}^m$. Using Hölder's inequality, we get the lower bound

  $$\begin{aligned} f(x) &\geq \frac{1}{2}\|Ax - b\|_2^2 + \left\|A^T b\right\|_\infty \|x\|_1 \\ &\geq \frac{1}{2}\|Ax - b\|_2^2 + \left\|b^T Ax\right\|_1 \\ &\geq \frac{1}{2}\|Ax - b\|_2^2 + b^T Ax \\ &= \frac{1}{2}\|Ax\|_2^2 + \frac{1}{2}\|b\|_2^2 \\ &\geq \frac{1}{2}\|b\|_2^2 \end{aligned}$$

  for each $x \in \mathbb{R}^m$. Furthermore, the lower bound is attained at $x = 0$, i.e. $f(0) = \frac{1}{2}\|b\|_2^2$. Therefore, $x = 0$ is an optimal point to (5.4).

Solution 5.5

CQ holds since both functions in (5.5) have full domain. Fermat's rule then gives that $x = (x_1, x_2) \in \mathbb{R}^2$ is an optimal point of (5.5) if and only if

$$\begin{gathered} 0 \in A^T Ax - A^T b + \lambda\partial(\|\cdot\|_1)(x) \\ \Leftrightarrow \\ 0 \in \sum_{j=1}^{2} a_i^T a_j x_j - a_i^T b + \lambda\partial(|\cdot|)(x_i), \quad \forall i \in \{1, 2\}. \end{gathered}$$

The equivalence hold since $\|x\|_1 = |x_1| + |x_2|$. Inserting the subdifferential of $|\cdot|$ gives

$$\begin{cases} 0 \in x_1 + a_1^T a_2 x_2 - a_1^T b + \lambda \begin{cases} \operatorname{sgn}(x_1) & \text{if } x_1 \neq 0 \\ [-1, 1] & \text{if } x_1 = 0 \end{cases} \\ 0 \in a_2^T a_1 x_1 + x_2 - a_2^T b + \lambda \begin{cases} \operatorname{sgn}(x_2) & \text{if } x_2 \neq 0 \\ [-1, 1] & \text{if } x_2 = 0 \end{cases} \end{cases} \tag{7.47}$$

where the assumption $\|a_1\|_2 = \|a_2\|_2 = 1$ was used. With the optimality conditions in place, we can now look at the four cases.

- Assume that $x \in X_{0,0}$. Then (7.47) is equivalent to

  $$\begin{cases} a_1^T b \in \lambda[-1, 1] \\ a_2^T b \in \lambda[-1, 1]. \end{cases}$$

  This in turn is equivalent to

  $$\begin{aligned} \lambda &\geq \max_{i=1,2} \left|a_i^T b\right| \\ &= \left\|A^T b\right\|_\infty . \end{aligned}$$

We conclude that

$$\Lambda_{0,0} = \left\{ \lambda > 0 : \lambda \geq \max_{i=1,2} \left| a_i^T b \right| \right\}.$$

- Assume that $x \in X_{1,0}$. Then (7.47) is equivalent to

$$\begin{cases} 0 = x_1 - a_1^T b + \lambda \operatorname{sgn}(x_1) \\ 0 \in a_2^T a_1 x_1 - a_2^T b + \lambda[-1,1]. \end{cases}$$

If $a_1^T b = 0$ the first condition can't be satisfied since $x_1 \neq 0$ by assumption. We conclude that

$$\Lambda_{1,0} = \emptyset$$

if $a_1^T b = 0$.

From here on, we assume $a_1^T b \neq 0$. The first condition can be re-written as

$$0 = \operatorname{sgn}(x_1)\, |x_1| - a_1^T b + \lambda \operatorname{sgn}(x_1)$$
$$\Leftrightarrow$$
$$\frac{a_1^T b}{\operatorname{sgn}(x_1)} - |x_1| = \lambda.$$

Since $\lambda > 0$, we get that $\operatorname{sgn}(x_1) = \operatorname{sgn}\left(a_1^T b\right)$ and

$$0 < \lambda = \left|a_1^T b\right| - |x_1| < \left|a_1^T b\right|$$

since $x_1 \neq 0$ by assumption. Multiplying both rows in the original condition with $\operatorname{sgn}(x_1) = \operatorname{sgn}\left(a_1^T b\right) = \frac{\left|a_1^T b\right|}{a_1^T b}$ gives

$$\begin{cases} |x_1| = \left|a_1^T b\right| - \lambda \\ 0 \in a_2^T a_1 \,|x_1| - \left|a_1^T b\right| \dfrac{a_2^T b}{a_1^T b} + \lambda[-1,1] \end{cases}$$
$$\Rightarrow 0 \in \left|a_1^T b\right| \left(a_2^T a_1 - \frac{a_2^T b}{a_1^T b} \right) - \lambda \left(a_2^T a_1 + [-1,1] \right).$$

The last inclusion can be written as

$$\lambda \left(a_2^T a_1 - 1 \right) \leq \left|a_1^T b\right| \left(a_2^T a_1 - \frac{a_2^T b}{a_1^T b} \right) \leq \lambda \left(a_2^T a_1 + 1 \right)$$

This implies that $\left|\frac{a_2^T b}{a_1^T b}\right| < 1$, since $0 < \lambda < \left|a_1^T b\right|$ and $a_1^T b \neq 0$. Thus, if $\left|\frac{a_2^T b}{a_1^T b}\right| \geq 1$, we must have $\Lambda_{1,0} = \emptyset$.

We can re-formulate these conditions as

$$\left|a_1^T b\right| \frac{a_2^T a_1 - \dfrac{a_2^T b}{a_1^T b}}{a_2^T a_1 + \operatorname{sgn}\left(a_2^T a_1 - \dfrac{a_2^T b}{a_1^T b} \right)} \leq \lambda$$

Further simplification and including the $\lambda < \left|a_1^T b\right|$ condition give

$$\frac{\left|a_1^T b\right| \left|\dfrac{a_2^T b}{a_1^T b} - a_1^T a_2\right|}{1 - a_1^T a_2 \operatorname{sgn}\left(\dfrac{a_2^T b}{a_1^T b} - a_1^T a_2\right)} \leq \lambda < \left|a_1^T b\right|. \tag{7.48}$$

To summarize this case, we have

$$\Lambda_{1,0} = \begin{cases} \emptyset & \text{if } \left|a_2^T b\right| \geq \left|a_1^T b\right| \\ \{\lambda > 0 : \text{(7.48) is satisfied}\} & \text{otherwise.} \end{cases}$$

Note that if $a_1^T b = 0$ then $\left|a_2^T b\right| \geq \left|a_1^T b\right|$, therefore, this cases is implicitly included above.

- By symmetry, the set $\Lambda_{0,1}$ is the same as $\Lambda_{1,0}$, but with the indices 1 and 2 swapped I.e. if

$$\frac{\left|a_2^T b\right| \left|\dfrac{a_1^T b}{a_2^T b} - a_1^T a_2\right|}{1 - a_1^T a_2 \operatorname{sgn}\left(\dfrac{a_1^T b}{a_2^T b} - a_1^T a_2\right)} \leq \lambda < \left|a_2^T b\right| \tag{7.49}$$

then

$$\Lambda_{0,1} = \begin{cases} \emptyset & \text{if } \left|a_1^T b\right| \geq \left|a_2^T b\right| \\ \{\lambda > 0 : \text{(7.49) is satisfied}\} & \text{otherwise.} \end{cases}$$

- Assume that $x \in X_{1,1}$. Then (7.47) is equivalent to the condition

$$0 = A^T A x - A^T b + \lambda \begin{bmatrix} \operatorname{sgn}(x_1) \\ \operatorname{sgn}(x_2) \end{bmatrix}$$

where matrix $A^T A$ and its inverse is given by

$$A^T A = \begin{bmatrix} 1 & a_1^T a_2 \\ a_1^T a_2 & 1 \end{bmatrix}, \quad (A^T A)^{-1} = \frac{1}{1 - \left(a_1^T a_2\right)^2} \begin{bmatrix} 1 & -a_1^T a_2 \\ -a_1^T a_2 & 1 \end{bmatrix}.$$

The inverse exists by assumption since $\left|a_1^T a_2\right| < 1$. Multiplying the condition from the left with $(A^T A)^{-1}$ gives

$$x = (A^T A)^{-1} A^T b - \lambda (A^T A)^{-1} \begin{bmatrix} \operatorname{sgn}(x_1) \\ \operatorname{sgn}(x_2) \end{bmatrix}.$$

Define the matrix

$$S = \begin{bmatrix} \operatorname{sgn}(x_1) & 0 \\ 0 & \operatorname{sgn}(x_2) \end{bmatrix}.$$

Multiply with S from the left gives

$$0 < \begin{bmatrix} |x_1| \\ |x_2| \end{bmatrix} = S(A^T A)^{-1} A^T b - \lambda S (A^T A)^{-1} \begin{bmatrix} \operatorname{sgn}(x_1) \\ \operatorname{sgn}(x_2) \end{bmatrix}.$$

The last term is

$$
\begin{aligned}
S(A^TA)^{-1}\begin{bmatrix}\operatorname{sgn}(x_1)\\ \operatorname{sgn}(x_2)\end{bmatrix} &= \frac{1}{1-\left(a_1^Ta_2\right)^2}\begin{bmatrix}\operatorname{sgn}(x_1) & 0\\ 0 & \operatorname{sgn}(x_2)\end{bmatrix}\begin{bmatrix}1 & -a_1^Ta_2\\ -a_1^Ta_2 & 1\end{bmatrix}\begin{bmatrix}\operatorname{sgn}(x_1)\\ \operatorname{sgn}(x_2)\end{bmatrix}\\
&= \frac{1-a_1^Ta_2\operatorname{sgn}(x_1x_2)}{1-\left(a_1^Ta_2\right)^2}\mathbf{1}\\
&> 0
\end{aligned}
$$

since $\left|a_1^Ta_2\right| < 1$. In order for the condition to have a solution we need

$$S(A^TA)^{-1}A^Tb > 0$$

In other words, $\operatorname{sgn}(x_i) = \operatorname{sgn}(\hat{x}_i)$ for $i = 1, 2$ where $\hat{x} = (A^TA)^{-1}A^Tb$ is the leasts squares solution. Inserting this back into the condition yields

$$0 < \begin{bmatrix}|\hat{x}_1|\\ |\hat{x}_2|\end{bmatrix} - \lambda\frac{1-a_1^Ta_2\operatorname{sgn}(\hat{x}_1\hat{x}_2)}{1-\left(a_1^Ta_2\right)^2}\mathbf{1}.$$

To summarize this case, we have

$$\Lambda_{1,1} = \left\{\lambda > 0 : \lambda < \frac{1-\left(a_1^Ta_2\right)^2}{1-a_1^Ta_2\operatorname{sgn}(\hat{x}_1\hat{x}_2)}\min\left(|\hat{x}_1|, |\hat{x}_2|\right),\ \hat{x} = (A^TA)^{-1}A^Tb\right\}.$$

In order to show the statement that the sets $\Lambda_{i,j}$ are disjoint and that the amount of sparsity is nondecreasing with λ, we need to consider different cases with respect to the data A and b. The case $A^Tb = 0$ gives that $\Lambda_{0,0} = \mathbb{R}_{++}$ and $\Lambda_{1,0} = \Lambda_{0,1} = \Lambda_{1,1} = \emptyset$ and statement holds. Thus, we consider $A^Tb \neq 0$ in the following. We can further divide into the cases

$$\left|a_1^Tb\right| > \left|a_2^Tb\right|, \quad \left|a_2^Tb\right| > \left|a_1^Tb\right| \quad \text{and} \quad \left|a_1^Tb\right| = \left|a_2^Tb\right|.$$

- One of $\Lambda_{1,0}$ and $\Lambda_{0,1}$ is empty since $\left|a_1^Tb\right| > \left|a_2^Tb\right|$ and $\left|a_2^Tb\right| > \left|a_1^Tb\right|$ can not hold at the same time. By symmetry, it is enough to consider only one of these cases. Here we consider the case $\left|a_1^Tb\right| > \left|a_2^Tb\right|$.

  Note that $\Lambda_{1,0}$ is nonempty and $\left|a_1^Tb\right| > 0$. Let $\lambda_{0,0} \in \Lambda_{0,0}$, $\lambda_{1,0} \in \Lambda_{1,0}$ and $\lambda_{1,1} \in \Lambda_{1,1}$. If we can show that

  $$\lambda_{1,1} < \lambda_{1,0} < \lambda_{0,0}$$

  we can conclude that the sets $\Lambda_{i,j}$ are disjoint and the amount of sparsity is nondecreasing with λ. Since $\lambda_{1,0} < |a_1^Tb|$ and $|a_1^Tb| \leq \lambda_{0,0}$ we have $\lambda_{1,0} < \lambda_{0,0}$. For $\lambda_{1,1}$ and $\lambda_{1,0}$ we have

  $$\lambda_{1,1} < \frac{1-\left(a_1^Ta_2\right)^2}{1-a_1^Ta_2\operatorname{sgn}(\hat{x}_1\hat{x}_2)}\min\left(|\hat{x}_1|, |\hat{x}_2|\right), \qquad \frac{\left|a_1^Tb\right|\left|\dfrac{a_2^Tb}{a_1^Tb} - a_1^Ta_2\right|}{1-a_1^Ta_2\operatorname{sgn}\left(\dfrac{a_2^Tb}{a_1^Tb} - a_1^Ta_2\right)} \leq \lambda_{1,0}$$

  and we will show that the upper bound of $\lambda_{1,1}$ and the lower bound of $\lambda_{1,0}$ are equal, proving that $\lambda_{1,1} < \lambda_{1,0}$. We start by showing that $\min\left(|\hat{x}_1|, |\hat{x}_2|\right) = |\hat{x}_2|$, i.e., $|\hat{x}_2| \leq |\hat{x}_1|$. Using the definition of $\hat{x}$, we have that

$$\begin{aligned}\hat{x} &= (A^T A)^{-1} A^T b \\ &= \frac{1}{1 - \left(a_1^T a_2\right)^2} \begin{bmatrix} a_1^T b - a_1^T a_2 a_2^T b \\ a_2^T b - a_1^T a_2 a_1^T b \end{bmatrix}.\end{aligned}$$

This implies that $|\hat{x}_2| \leq |\hat{x}_1|$ is equivalent to

$$\begin{gathered}\left|a_2^T b - a_1^T a_2 a_1^T b\right| \leq \left|a_1^T b - a_1^T a_2 a_2^T b\right| \\ \Leftrightarrow \\ \left|\frac{a_2^T b}{a_1^T b} - a_1^T a_2\right| \leq \left|1 - a_1^T a_2 \frac{a_2^T b}{a_1^T b}\right| = 1 - a_1^T a_2 \frac{a_2^T b}{a_1^T b}.\end{gathered}$$

The last equality holds since $\left|a_1^T a_2\right| < 1$ and $\left|\frac{a_2^T b}{a_1^T b}\right| < 1$ by assumption. The last inequality can equivalently be written as

$$\begin{gathered}\begin{cases} \dfrac{a_2^T b}{a_1^T b} - a_1^T a_2 \leq 1 - a_1^T a_2 \dfrac{a_2^T b}{a_1^T b} \\ a_1^T a_2 - \dfrac{a_2^T b}{a_1^T b} \leq 1 - a_1^T a_2 \dfrac{a_2^T b}{a_1^T b} \end{cases} \\ \Leftrightarrow \\ \begin{cases} 0 \leq 1 + a_1^T a_2 - a_1^T a_2 \dfrac{a_2^T b}{a_1^T b} - \dfrac{a_2^T b}{a_1^T b} = \left(1 + a_1^T a_2\right)\left(1 - \dfrac{a_2^T b}{a_1^T b}\right) \\ 0 \leq 1 - a_1^T a_2 - a_1^T a_2 \dfrac{a_2^T b}{a_1^T b} + \dfrac{a_2^T b}{a_1^T b} = \left(1 - a_1^T a_2\right)\left(1 + \dfrac{a_2^T b}{a_1^T b}\right) \end{cases}.\end{gathered}$$

But this holds since $\left|a_1^T a_2\right| < 1$ and $\left|\frac{a_2^T b}{a_1^T b}\right| < 1$ by assumption. The upper bound on $\lambda_{1,1}$ can now be written as

$$\frac{1 - \left(a_1^T a_2\right)^2}{1 - a_1^T a_2 \operatorname{sgn}\left(\hat{x}_1 \hat{x}_2\right)} |\hat{x}_2| = \frac{\left|a_1^T b\right| \left|\dfrac{a_2^T b}{a_1^T b} - a_1^T a_2\right|}{1 - a_1^T a_2 \operatorname{sgn}\left(\hat{x}_1 \hat{x}_2\right)}.$$

This is the same as the lower bound on $\lambda_{1,0}$ since

$$\begin{aligned}\operatorname{sgn}\left(\hat{x}_1 \hat{x}_1\right) &= \operatorname{sgn}\left(\left(a_1^T b - a_1^T a_2 a_2^T b\right)\left(a_2^T b - a_1^T a_2 a_1^T b\right)\right) \\ &= \operatorname{sgn}\left(\left(1 - a_1^T a_2 \frac{a_2^T b}{a_1^T b}\right)\left(\frac{a_2^T b}{a_1^T b} - a_1^T a_2\right)\right) \\ &= \operatorname{sgn}\left(\frac{a_2^T b}{a_1^T b} - a_1^T a_2\right)\end{aligned}$$

since $\left|a_1^T a_2\right| < 1$ and $\left|\frac{a_2^T b}{a_1^T b}\right| < 1$ by assumption. This concludes the proof for the case when $\left|a_1^T b\right| > \left|a_2^T b\right|$.

- Next, we consider the case $\left|a_1^T b\right| = \left|a_2^T b\right|$. Then $\left|a_1^T b\right| = \left|a_2^T b\right| > 0$ since $A^T b \neq 0$. Moreover, $\Lambda_{1,0} = \Lambda_{0,1} = \emptyset$. Let $\lambda_{0,0} \in \Lambda_{0,0}$ and $\lambda_{1,1} \in \Lambda_{1,1}$. We want to show that

$$\lambda_{1,1} \underbrace{<}_{\text{known}} \frac{1 - \left(a_1^T a_2\right)^2}{1 - a_1^T a_2 \operatorname{sgn}\left(\hat{x}_1 \hat{x}_2\right)} \min\left(\left|\hat{x}_1\right|, \left|\hat{x}_2\right|\right) \underbrace{=}_{\text{unknown}} \left|a_1^T b\right| = \left|a_2^T b\right| \underbrace{\leq}_{\text{known}} \lambda_{0,0}.$$

We know that

$$\begin{aligned}
\frac{1}{1-a_1^T a_2 \operatorname{sgn}(\hat{x}_1\hat{x}_2)}|\hat{x}_1| &= \frac{1}{1-a_1^T a_2 \operatorname{sgn}(\hat{x}_1\hat{x}_2)}|a_2^T b|\left|\frac{a_1^T b}{a_2^T b} - a_1^T a_2\right| \\
&= \frac{\operatorname{sgn}\left(\frac{a_1^T b}{a_2^T b} - a_1^T a_2\right)}{1-a_1^T a_2 \operatorname{sgn}(\hat{x}_1\hat{x}_2)}|a_2^T b|\left(\frac{a_1^T b}{a_2^T b} - a_1^T a_2\right) \\
&= \frac{1}{\operatorname{sgn}\left(\frac{a_1^T b}{a_2^T b} - a_1^T a_2\right) - a_1^T a_2}|a_2^T b|\left(\frac{a_1^T b}{a_2^T b} - a_1^T a_2\right)
\end{aligned}$$

where it was used that $\operatorname{sgn}(\hat{x}_1\hat{x}_2) = \operatorname{sgn}\left(\frac{a_1^T b}{a_2^T b} - a_1^T a_2\right)$. We now note that $\frac{a_1^T b}{a_2^T b} = \operatorname{sgn}\left(\frac{a_1^T b}{a_2^T b}\right)$ since $|a_1^T b| = |a_2^T b|$. Furthermore, we then also have

$$\begin{aligned}
\operatorname{sgn}\left(\frac{a_1^T b}{a_2^T b} - a_1^T a_2\right) &= \operatorname{sgn}\left(\operatorname{sgn}\left(\frac{a_1^T b}{a_2^T b}\right) - a_1^T a_2\right) \\
&= \operatorname{sgn}\left(\operatorname{sgn}\left(\frac{a_1^T b}{a_2^T b}\right)\right) \\
&= \operatorname{sgn}\left(\frac{a_1^T b}{a_2^T b}\right)
\end{aligned}$$

since $|a_1^T a_2| < 1$. This yields

$$\begin{aligned}
\frac{1}{1-a_1^T a_2 \operatorname{sgn}(\hat{x}_1\hat{x}_2)}|\hat{x_1}| &= \frac{1}{\operatorname{sgn}\left(\frac{a_1^T b}{a_2^T b}\right) - a_1^T a_2}|a_2^T b|\left(\operatorname{sgn}\left(\frac{a_1^T b}{a_2^T b}\right) - a_1^T a_2\right) \\
&= |a_2^T b|.
\end{aligned}$$

By symmetry, the analogue holds for $\frac{1}{1-a_1^T a_2 \operatorname{sgn}(\hat{x}_1\hat{x}_2)}|\hat{x}_2|$, i.e.

$$\frac{1}{1-a_1^T a_2 \operatorname{sgn}(\hat{x}_1\hat{x}_2)}|\hat{x_2}| = |a_1^T b|.$$

This gives us the desired inequality

$$\lambda_{1,1} < \lambda_{0,0}.$$

This concludes the proof for the $|a_1^T b| = |a_2^T b|$.

We have now covered all cases.

Note that, in all cases, the distances $|\lambda_{1,0} - \lambda_{0,0}|$ and $|\lambda_{1,1} - \lambda_{1,0}|$ can be made arbitrary small. This is expected since otherwise there would be λ for which no solution exists. Since problem (5.5) is strongly convex for all $\lambda > 0$, we know that this is not possible.

Solution 5.6

Note that (5.6) is always bounded below by zero.

1. Let $t > 0$. Then $t(w, b)$ also separates the data. Inserting this into the cost function of (5.6) gives that

$$\sum_{i=1}^{n} \max\left(0, 1 - ty_i\left(x_i^T w + b\right)\right) = \sum_{i=1}^{n} \max\left(0, 1 - t\left|x_i^T w + b\right|\right).$$

   Choosing any

$$t \geq \frac{1}{\min_{i=1,\ldots,n}\left|x_i^T w + b\right|}$$

   gives a cost of 0 and therefore $t(w, b)$ must be an optimal point. The set of optimal points is unbounded since $\|t(w, b)\|_2 = t\|(w, b)\|_2$, $\|(w, b)\|_2 > 0$ and $t \geq (\min_i |x_i^T w + b|)^{-1}$ can be made arbitrary large.

2. Choosing an arbitrary $w \in \mathbb{R}^m$ and inserting into the cost function of (5.6) gives

$$\sum_{i=1}^{n} \max\left(0, 1 - y_i\left(x_i^T w + b\right)\right) = \sum_{i=1}^{n} \max\left(0, 1 - x_i^T w - b\right).$$

   Choosing

$$b \geq 1 - \min_{i=1,\ldots,n} x_i^T w$$

   gives a cost of 0 and therefore (w, b) is an optimal point. The set of optimal points is unbounded since $\|(w, b)\|_2^2 = \|w\|_2^2 + |b|^2$, where $b \geq 1 - \min_{i=1,\ldots,n} x_i^T w$, can be made arbitrary large.

3. Letting $w = 0$ and inserting into the cost function of (5.6) gives

$$\sum_{i=1}^{n} \max\left(0, 1 - y_i(x_i^T w + b)\right) + \frac{\lambda}{2}\|w\|_2^2 = \sum_{i=1}^{n} \max\left(0, 1 - b\right) \geq 0.$$

   Any $b \geq 1$ yields a cost of 0 and (w, b) is therefore an optimal point. The set of optimal points is unbounded since $\|(w, b)\|_2 = |b|$, where $b \geq 1$, can be made arbitrary large.

Solution 5.7

Note that the regularization term is the same. Woking with the sum of hinge-losses we get that

$$\begin{aligned}
\sum_{i=1}^{n} \max\left(0, 1 - y_i\left(x_i^T w + b\right)\right) &= \mathbf{1}^T \begin{bmatrix} \max\left(0, 1 - y_1\left(x_1^T w + b\right)\right) \\ \vdots \\ \max\left(0, 1 - y_n\left(x_n^T w + b\right)\right) \end{bmatrix} \\
&= \mathbf{1}^T \max\left(0, \begin{bmatrix} 1 - y_1\left(x_1^T w + b\right) \\ \vdots \\ 1 - y_n\left(x_n^T w + b\right) \end{bmatrix}\right) \\
&= \mathbf{1}^T \max\left(0, \mathbf{1} - \begin{bmatrix} y_1\left(x_1^T w + b\right) \\ \vdots \\ y_n\left(x_n^T w + b\right) \end{bmatrix}\right) \\
&= \mathbf{1}^T \max\left(0, \mathbf{1} - \left(\begin{bmatrix} y_1 x_1^T w \\ \vdots \\ y_n x_n^T w \end{bmatrix} + \begin{bmatrix} y_1 b \\ \vdots \\ y_n b \end{bmatrix}\right)\right) \\
&= \mathbf{1}^T \max\left(0, \mathbf{1} - \left(\begin{bmatrix} y_1 x_1^T \\ \vdots \\ y_n x_n^T \end{bmatrix} w + b \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}\right)\right).
\end{aligned}$$

We can now identify

$$X = \begin{bmatrix} y_1 x_1 & \cdots & y_n x_n \end{bmatrix} \quad \text{and} \quad \phi = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Solution 5.8

1. The function f is a sum of hinge-losses and in particular separable, i.e.

   $$f(u) = \sum_{i=1}^{n} \max(0, 1 - u_i)$$

   for each $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$. Using Exercises 3.1 and 3.5, we get that

   $$\begin{aligned}
   f^*(\mu) &= \sum_{i=1}^{n} \left(\max(0, 1 - \cdot)\right)^* (\mu_i) \\
   &= \sum_{i=1}^{n} \mu_i + \iota_{[-1,0]}(\mu_i) \\
   &= \mathbf{1}^T \mu + \iota_{[-\mathbf{1},0]}(\mu)
   \end{aligned}$$

   for each $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$. Using Exercise 3.1, we get that the conjugate of

g is

$$
\begin{aligned}
g^*(\nu_w, \nu_b) &= \sup_{(w,b)\in\mathbb{R}^m\times\mathbb{R}} \left((\nu_w, \nu_b)^T (w, b) - \frac{\lambda}{2} \|w\|_2^2 \right) \\
&= \sup_{w\in\mathbb{R}^m} \left(\nu_w^T w - \frac{\lambda}{2} \|w\|_2^2 \right) + \sup_{b\in\mathbb{R}} (\nu_b b) \\
&= \frac{1}{2\lambda} \|\nu_w\|_2^2 + \iota_{\{0\}}(\nu_b).
\end{aligned}
$$

for each $(\nu_w, \nu_b) \in \mathbb{R}^m \times \mathbb{R}$. Note that

$$
\begin{aligned}
g^*\left(-L^T\mu\right) &= g^*\left(-\begin{bmatrix} X \\ \phi^T \end{bmatrix} \mu \right) \\
&= \frac{1}{2\lambda} \|-X\mu\|_2^2 + \iota_{\{0\}}\left(-\phi^T\mu\right) \\
&= \frac{1}{2\lambda} \|X\mu\|_2^2 + \iota_{\{0\}}\left(\phi^T\mu\right)
\end{aligned}
$$

for each $\mu \in \mathbb{R}^n$. Thus, the dual problem

$$
\underset{\mu\in\mathbb{R}^n}{\text{minimize}}\, f^*(\mu) + g^*(-L^T\mu)
$$

becomes

$$
\underset{\mu\in\mathbb{R}^n}{\text{minimize}}\, \mathbf{1}^T\mu + \frac{1}{2\lambda}\|X\mu\|_2^2 + \iota_{[-\mathbf{1},0]}(\mu) + \iota_{\{0\}}\left(\phi^T\mu\right)
$$

or written differently

$$
\begin{aligned}
\underset{\mu\in\mathbb{R}^n}{\text{minimize}} \quad & \mathbf{1}^T\mu + \frac{1}{2\lambda}\mu^T X^T X\mu \\
\text{subject to} \quad & -\mathbf{1} \le \mu \le 0, \\
& \phi^T\mu = 0.
\end{aligned}
$$

2. We claim that CQ holds for the dual problem, i.e.

$$
\operatorname{relint}\operatorname{dom} f^* \cap \operatorname{relint}\operatorname{dom} g^* \circ -L^T \neq \emptyset. \tag{7.50}
$$

Indeed, we have that

$$
\operatorname{relint}\operatorname{dom} f^* = (-\mathbf{1}, 0).
$$

Since we have examples from both classes, we know that

$$
\phi = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.
$$

has both 1 and -1 as elements. Thus, pick indices $i, j = 1, \ldots, n$, $i \neq j$ such that $\phi_i = 1$ and $\phi_j = -1$. Note that

$$
\{\mu \in \mathbb{R}^n : \mu_i = \mu_j,\ \mu_l = 0 \text{ for } l \in \{1, \ldots, n\} \setminus \{i, j\}\} \subseteq \operatorname{dom} g^* \circ -L^T
$$

and therefore

$$\{\mu \in \mathbb{R}^n : \mu_i = \mu_j,\ \mu_l = 0 \text{ for } l \in \{1, \ldots, n\} \setminus \{i, j\}\} \subseteq \operatorname{relint} \operatorname{dom} g^* \circ -L^T.$$

This show that the intersection in (7.50) is nonempty, as claimed.

Suppose that $\mu \in \mathbb{R}^n$ is an optimal point for the dual problem. By Fermat's rule, closed convexity of f and g, and since CQ holds for the dual problem, we know that

$$\begin{gathered}
0 \in \partial f^*(\mu) - L\partial g^*(-L^T\mu) \\
\Leftrightarrow \\
\begin{cases} L(w,b) \in \partial f^*(\mu) \\ \quad (w,b) \in \partial g^*(-L^T\mu) \end{cases} \\
\Leftrightarrow \\
\begin{cases} \qquad \mu \in \partial f(L(w,b)) \\ -L^T\mu \in \partial g(w,b) \end{cases} \\
\Leftrightarrow \\
0 \in L^T\partial f(L(w,b)) + \partial g(w,b) \\
\Rightarrow \\
0 \in \partial(f \circ L + g)(w,b).
\end{gathered}$$

(The last implication and be strengthened to an equivalence since CQ clearly holds for the primal problem, but the implication suffices to show what follows.) Hence, such a point $(w,b) \in \mathbb{R}^m \times \mathbb{R}$ must be an optimal point to the primal problem.

We can recover w from the second condition of

$$\begin{cases} L(w,b) \in \partial f^*(\mu) \\ \quad (w,b) \in \partial g^*(-L^T\mu). \end{cases} \tag{7.51}$$

Indeed, note that

$$\partial \iota_{\{0\}}(\nu) = \begin{cases} \mathbb{R} & \text{if } \nu = 0 \\ \emptyset & \text{if } \nu = \mathbb{R} \setminus \{0\} \end{cases}$$

Therefore,

$$\begin{aligned}
\partial g^*(\nu_w, \nu_b) &= \left\{ \begin{bmatrix} s_w \\ s_b \end{bmatrix} : s_w \in \partial\left(\frac{1}{2\lambda}\|\cdot\|_2^2\right)(\nu_w),\ s_b \in \partial\iota_{\{0\}}(\nu_b) \right\} \\
&= \begin{cases} \left\{ \begin{bmatrix} \frac{1}{\lambda}\nu_w \\ a \end{bmatrix} : a \in \mathbb{R} \right\} & \text{if } (\nu_w, \nu_b) \in \mathbb{R}^m \times \{0\} \\ \emptyset & \text{if } (\nu_w, \nu_b) \in \mathbb{R}^m \times (\mathbb{R} \setminus \{0\}) \end{cases}
\end{aligned}$$

Moreover, note that

$$-L^T\mu = \begin{bmatrix} -X\mu \\ -\phi^T\mu \end{bmatrix}$$

where we must have that $-\phi^T \mu = 0$ since μ is assumed to be an optimal point for the dual problem. Thus, we have that

$$\partial g^*(-L^T \mu) = \left\{ \begin{bmatrix} -\frac{1}{\lambda} X\mu \\ a \end{bmatrix} : a \in \mathbb{R} \right\}$$

and using the second condition of (7.51) we can uniquely determine w as

$$w = -\frac{1}{\lambda} X\mu.$$

However, this does not allow us to uniquely determine b.

Next, we determine b. Note that

$$(\partial f^*(\mu))_i = \begin{cases} \emptyset & \text{if } \mu_i < -1 \\ [-\infty, 1] & \text{if } \mu_i = -1, \\ \{1\} & \text{if } -1 < \mu_i < 0, \\ [1, \infty] & \text{if } \mu_i = 0, \\ \emptyset & \text{if } \mu_i > 0 \end{cases}$$

for each $\mu \in \mathbb{R}^n$ and each $i = 1, \dots, n$. The first condition of (7.51) gives the requirement

$$X^T w + b\phi \in \partial f^*(\mu), \tag{7.52}$$

Recall that

$$X = \begin{bmatrix} y_1 x_1 & \cdots & y_n x_n \end{bmatrix} \quad \text{and} \quad \phi = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Thus, under the condition that there exists an index $i = 1, \dots, n$ such that $-1 < \mu_i < 0$, we can uniquely determine b from

$$y_i x_i^T w + b y_i = 1 \iff b = y_i^{-1} - x_i^T w = y_i - x_i^T w.$$

3. Suppose that $\mu^* \in \mathbb{R}^n$ is an optimal point for the dual problem and that $(w^\star, b^\star) \in \mathbb{R}^m \times \mathbb{R}$ in an extracted optimal point for the primal problem. Clearly, it must hold that $-\mathbf{1} \leq \mu^* \leq 0$.

   First, we show that if $i = 1, \dots, n$ is an index such that $-1 \leq \mu_i^* < 0$, then x_i must be a support vector. Thus, let i be such an index. We repeat (7.52):

   $$X^T w^* + b^* \phi \in \partial f^*(\mu^*).$$

   The ith coordinate of this inclusion is

   $$y_i(x_i^T w^* + b^*) \in \begin{cases} \emptyset & \text{if } \mu_i^* < -1 \\ [-\infty, 1] & \text{if } \mu_i^* = -1, \\ \{1\} & \text{if } -1 < \mu_i^* < 0, \\ [1, \infty] & \text{if } \mu_i^* = 0, \\ \emptyset & \text{if } \mu_i^* > 0 \end{cases}$$

Since $-1 \leq \mu_i^* < 0$, we get that

$$y_i(x_i^T w^* + b^*) \leq 1$$

or equivalently

$$0 \leq 1 - y_i(x_i^T w^* + b^*)$$

and we conclude that x_i is a support vector.

Second, we show that we can recover $(w^\star, b^\star) \in \mathbb{R}^m \times \mathbb{R}$ only using support vectors. From the previous subproblem, we know that we can determine the optimal $w^* \in \mathbb{R}^m$ by

$$\begin{aligned} w^* &= -\frac{1}{\lambda} X\mu \\ &= -\frac{1}{\lambda} \sum_{i=1}^{n} y_i x_i \mu_i \\ &= -\frac{1}{\lambda} \sum_{\substack{i=1 \\ \text{s.t. } \mu_i \neq 0}}^{n} y_i x_i \mu_i \end{aligned}$$

i.e. we only utilize support vectors. The optimal parameter b^* can then be recovered as in the previous subproblem where a nonzero element of μ^* was utilized, i.e. a support vector.

Solution 5.9

1. True. Consider the model $m_w(x) = w^T \phi(x)$ as a function of w instead of x and note that it is linear in w since $\phi(x)$ does not depend on w. Since y_i also does not depend on w,

$$w \mapsto L(m_w(x_i), y_i)$$

is a convex function composed with a linear function and therefore itself convex. We see that the objective function is a sum of convex functions, and therefore itself convex.

2. False. Consider a two layer network, i.e. $D = 2$, with

$$d = l = k = f_1 = 1$$

and σ_1, σ_2 as identity functions. Then

$$m_w(x) = w_1 w_2 x$$

for each $x \in \mathbb{R}^n$. Take the L as the square error loss and consider a single ($n = 1$) data point $x_1 = 1$ with response variable $y_1 = 0$. Then we get the loss (and objective) function

$$\begin{aligned} L(m_w(x_1), y_1) &= \|w_1 w_2\|_2^2 \\ &= (w_1 w_2)^2. \end{aligned}$$

We claim that this is not convex as a function of $w = (w_1, w_2) \in \mathbb{R} \times \mathbb{R}$. The points $(0,1)$ and $(1,0)$ both have value 0 but the convex combination

$$\frac{1}{2}(0,1) + \frac{1}{2}(1,0) = (0.5, 0.5)$$

has a positive value. Therefore, the objective function is not convex in general.

Solutions to chapter 6

Solution 6.1

To estimate the overall computational cost of an algorithm, we can roughly use

$$\text{(iterations count)} \times \text{(per-iteration cost)}.$$

This quantity for the first algorithm is 5×10^8 and for the second one is 10^8. Hence, the second algorithm had a better performance.

Solution 6.2

1. $O(\rho_1^k) \leftrightarrow$ A2 (linear)
2. $O(\rho_2^k) \leftrightarrow$ A4 (linear)
3. $O(1/\log(k)) \leftrightarrow$ A3 (sublinear)
4. $O(1/k) \leftrightarrow$ A1 (sublinear)
5. $O(1/k^2) \leftrightarrow$ A5 (sublinear)

Solution 6.3

1. From the Q-linear rate definition, we have that

$$V_k \leq \rho V_{k-1} \leq \rho^2 V_{k-2} \leq \ldots \leq \rho^k V_0$$

or

$$V_k \leq \rho^k V_0$$

holds inductively for each integer $k \geq 0$. This implies an R-linear rate with $\rho_L = \rho$ and $C_L = V_0$.

2. From the Q-quadratic rate definition, we have that

$$
\begin{aligned}
&V_1 \leq \rho V_0^2 \\
&V_2 \leq \rho V_1^2 \leq \rho\rho^2 V_0^{2^2} \\
&V_3 \leq \rho V_2^2 \leq \rho\rho^2\rho^{2^2} V_0^{2^3} \\
&V_4 \leq \rho V_3^2 \leq \rho\rho^2\rho^{2^2}\rho^{2^3} V_0^{2^4} \\
&\quad\vdots \\
&V_k \leq \rho V_{k-1}^2 \leq \rho\rho^2\rho^{2^2}\rho^{2^3}\cdots\rho^{2^{k-1}} V_0^{2^k} = \rho^{2^k-1} V_0^{2^k} = \rho^{2^k} V_0^{2^k} \rho^{-1}
\end{aligned}
$$

or

$$V_k \leq (\rho V_0)^{2^k} \rho^{-1}$$

holds inductively for each integer $k \geq 0$. Here, we used that

$$1 + 2 + 2^2 + \ldots + 2^{k-1} = 2^k - 1.$$

We get that (6.2) holds with $\rho_Q = \rho V_0 \geq 0$ and $C_Q = \rho^{-1} \geq 0$ since $V_0 \geq 0$ and $\rho > 0$ by assumption.

3. If $\rho V_0 < 1$ or equivalently $V_0 \leq \rho^{-1}$, we get $\rho_Q = \rho V_0 \in [0, 1)$.

Solution 6.4

1. Let $n = 1$ and consider the function $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = x$$

for each $x \in \mathbb{R}$ and $x_k = -k$ for each integer $k \geq 0$. Clearly, $(x_k)_{k=0}^{\infty}$ is a descent sequence and

$$f(x_k) \to -\infty \quad \text{as} \quad k \to \infty.$$

I.e. the sequence of function values $(f(x_k))_{k=0}^{\infty}$ does not converge in $\mathbb{R}$.

2. Solution 1: Note that the sequence $(f(x_k))_{k=0}^{\infty}$ is monotone, by construction. Moreover, $(f(x_k))_{k=0}^{\infty}$ is bounded — from above by $f(x_0)$ and from below by B. Then, by the monotone convergence theorem, the sequence $(f(x_k))_{k=0}^{\infty}$ converges in $\mathbb{R}$.

   Solution 2: First, note that the nonempty set $\{f(x_k) : k \in \mathbb{N}_0\}$ in $\mathbb{R}$ is bounded from below by B or equivalently, $\{-f(x_k) : k \in \mathbb{N}_0\}$ is bounded from above by $-B$. By the least-upper-bound property of $\mathbb{R}$, there exists a real number, say $\tilde{b} \in \mathbb{R}$, such that

$$\sup \{-f(x_k) : k \in \mathbb{N}_0\} = \tilde{b}$$

   or equivalently

$$\inf \{f(x_k) : k \in \mathbb{N}_0\} = b$$

where $b = -\tilde{b}$. The least-upper-bound property of $\mathbb{R}$ can be taken as a completeness axoim of $\mathbb{R}$, or, proven as a theorem from some other completeness axoim, e.g., the convergence of every Cauchy sequence.

Second, recall that the definition of the infimum of a set is the greates lower bound of that set. In particular, for any $c \in \mathbb{R}$ that is a lower bound of $\{f(x_k) : k \in \mathbb{N}_0\}$, i.e. $c \leq f(x_k)$ for each integer $k \geq 0$, it holds that $c \leq b$.

Third, we claim that $(f(x_k))_{k=0}^{\infty}$ converges to b, or written differently,

$$f(x_k) \to b \quad \text{as} \quad k \to \infty.$$

This, by definition, means that for each $\epsilon > 0$, there exists an $N \in \mathbb{N}_0$ such that

$$|f(x_k) - b| < \epsilon$$

for each integer $k \geq N$, or equivalently,

$$b - \epsilon < f(x_k) < b + \epsilon$$

for each integer $k \geq N$. Indeed, let $\epsilon > 0$ be arbitrary. Since b is the greates lower bound of $\{f(x_k) : k \in \mathbb{N}_0\}$, we get that

$$b - \epsilon < b \leq f(x_k)$$

for each integer $k \geq 0$. Moreover, there exists an $N \in \mathbb{N}_0$ such that

$$f(x_N) < b + \epsilon.$$

Why does such an N exist? If there did not exists any such N, $b + \epsilon$ would be a lower bound of the set $\{f(x_k) : k \in \mathbb{N}_0\}$. But this would contradict the fact that b is the greates lower bound of $\{f(x_k) : k \in \mathbb{N}_0\}$, since $b < b + \epsilon$. Finally, note that

$$f(x_k) \leq f(x_N) < b + \epsilon$$

for each integer $k \geq N$, by construction of the sequence $(x_k)_{k=0}^{\infty}$. I.e. we established that

$$b - \epsilon < f(x_k) < b + \epsilon$$

for each integer $k \geq N$, as claimed.

3. The most basic example would be to consider any function f that is bounded from below and let $x_k = x$ for each $k \in \mathbb{N}_0$, where $x \in \mathbb{R}^n$ is not an optimal point. A slightly more interesting example would be $f : \mathbb{R}^2 \to \mathbb{R}$ such that

$$f(x, y) = x^2 + y^2$$

for each $(x, y) \in \mathbb{R}^2$ and the sequence

$$(x_k, y_k) = \left(\left(1 + \frac{1}{k}\right)\sin k, \left(1 + \frac{1}{k}\right)\cos k\right)$$

for each integer $k \geq 0$. We see that

$$f(x_k, y_k) = \left(1 + \frac{1}{k}\right)^2$$

is decreasing but does not converge to the optimum $f(0, 0) = 0$. There are plenty more examples. Function value decrease is a very weak (read: useless) condition for a minimization algorithm.

Solution 6.5

Below you see an expanded table with the asked for ratios. We see that the linear ratio is steadily decreasing while the quadratic ratio is more stable (up until machine precision is achieved). Clearly, the sequence appear to converge Q-quadratically. The parameter is given by the worst case ratio, i.e., $\rho \approx 0.24$.

k	x_k	$\lvert x_k - x^\star\rvert = d_k$	d_{k+1}/d_k	d_{k+1}/d_k^2
0	5.000000000000000	4.685076942154594	0.77804204	0.16606815
1	3.960109873126804	3.645186815281398	0.70591922	0.19365790
2	2.888130487596392	2.573207429750986	0.57679574	0.22415439
3	1.799138129515975	1.484215071670569	0.35988932	0.24247788
4	0.849076217909656	0.534153160064250	0.12138864	0.22725437
5	0.379763183818023	0.064840125972617	0.01339947	0.20665396
6	0.315791881094192	0.000868823248786	0.00017665	0.20332357
7	0.314923211324986	0.000000153479580	0.00000003	0.21226031
8	0.314923057845411	0.000000000000005	0.00000000	0.00000000
9	0.314923057845406	0.000000000000000	NA	NA

For the interested: The gradient and Hessian are

$$\nabla f(x) = e^x - 2 + 2x$$
$$\nabla^2 f(x) = e^x + 2$$

for each $x \in \mathbb{R}$, which shows that f is strongly convex and thus has a unique minimizer. The Newton iteration is then explicitly written as

$$x_{k+1} = x_k - \frac{e^{x_k} - 2 + 2x_k}{e^{x_k} + 2}$$

for each integer $k \geq 0$.

Solution 6.6

1. Note that
$$0 \leq Q_k \leq \frac{V}{\psi_1(k)} + \frac{D}{\psi_2(k)} \to 0 \quad \text{as} \quad k \to \infty.$$
Therefore, $Q_k \to 0$ as $k \to \infty$, by the squeeze theorem.

2. Since we have two terms (both converging to zero as $k \to \infty$) on the r.h.s. of the inequality, the slower term is the bottleneck and decides the rate of convergence, that is, the smaller between ψ_1 and ψ_2 determines the rate of convergence. When comparing we can ignore the constant terms. With that in mind, the rates are as follows:
   (a) $O(\log(k)/\sqrt{k})$ sublinear rate of convergence.
   (b) We should compare $O(1/k^{1-\alpha})$ and $O(\frac{1}{k^{1-\alpha}/k^{1-2\alpha}}) = O(1/k^{\alpha})$. Since $\alpha \in (0, 0.5)$, $O(1/k^{\alpha})$ is the rate of convergence.
   (c) We should compare $O(1/k^{1-\alpha})$ and $O(\frac{1}{k^{1-\alpha}/k^{1-2\alpha}}) = O(1/k^{\alpha})$. Since $\alpha \in (0.5, 1)$, $O(1/k^{1-\alpha})$ is the rate of convergence.

3. The cases in (b) and (c) are similar. We just need to compare them with case (a). Let $\alpha \in (0, 0.5)$ and note that
$$\frac{\log(k)/\sqrt{k}}{1/k^{\alpha}} = \frac{\log(k)}{k^{0.5-\alpha}} \to 0 \quad \text{as} \quad k \to \infty.$$
We conclude that case (a) gives the faster rate.

Solution 6.7

1. Note that
$$\begin{aligned} 0 \leq f(x_k) - f(x^\star) &\leq \frac{V + D\sum_{i=0}^{k} \gamma_i^2}{b\sum_{i=0}^{k} \gamma_i} \\ &\leq \frac{V + D\sum_{i=0}^{\infty} \gamma_i^2}{b\sum_{i=0}^{k} \gamma_i} \to 0 \quad \text{as} \quad k \to \infty. \end{aligned}$$
The squeeze theorem gives that
$$f(x_k) - f(x^\star) \to 0 \quad \text{as} \quad k \to \infty$$
or equivalently
$$f(x_k) \to f(x^\star) \quad \text{as} \quad k \to \infty$$
as desired.

2. In both cases, the function $\phi : \mathbb{R}_+ \to \mathbb{R}_{++}$ such that
$$\phi(i) = \gamma_i$$
for each $i \geq 0$ is decreasing. Therefore, we obtain the following bound:
$$\begin{aligned} 0 < \int_0^k \phi(t)dt &\leq \sum_{i=0}^{k} \phi(i) \\ &= \sum_{i=0}^{k} \gamma_i \end{aligned}$$
for each integer $k \geq 0$. Similarly, we also get the bound
$$\begin{aligned} \sum_{i=0}^{k} \gamma_i^2 &= \sum_{i=0}^{k} \phi^2(k) \\ &\leq \int_0^k \phi^2(t)dt + \phi^2(0) \\ &\leq \int_0^\infty \phi^2(t)dt + \phi^2(0) \end{aligned}$$

for each integer $k \geq 0$. Combining these bounds with the inequality given by the convergence analysis, we get the new inequality

$$\begin{aligned} f(x_k) - f(x^\star) &\leq \frac{V + D\displaystyle\sum_{i=0}^{k} \gamma_i^2}{b\displaystyle\sum_{i=0}^{k} \gamma_i} \\ &\leq \frac{V + D\left(\displaystyle\int_0^\infty \phi^2(t)dt + \phi^2(0)\right)}{b\displaystyle\int_0^k \phi(t)dt} \end{aligned}$$

for each integer $k \geq 0$.

(a) Let

$$\phi(t) = \frac{c}{t+1}$$

for each $t \geq 0$. Note that

$$\begin{aligned} \int_0^\infty \phi^2(t)dt + \phi^2(0) &= c^2\left(\int_0^\infty \frac{1}{(1+t)^2}dt + 1\right) \\ &= c^2\left(\left[\frac{-1}{1+t}\right]_{t=0}^\infty + 1\right) \\ &= 2c^2 \end{aligned}$$

and

$$\begin{aligned} \int_0^k \phi(t)dt &= c\int_0^k \frac{1}{t+1}dt \\ &= c\left[\log(t+1)\right]_{t=0}^k \\ &= c\log(k+1). \end{aligned}$$

We conclude that

$$f(x_k) - f(x^\star) \leq \frac{V + 2Dc^2}{bc\log(k+1)}$$

for each integer $k \geq 0$, which shows a $O(1/\log k)$ sublinear rate of convergence.

(b) Let

$$\phi(t) = \frac{c}{(t+1)^\alpha}$$

for each $t \geq 0$, where $\alpha \in (0.5, 1)$. Note that

$$\begin{aligned} \int_0^\infty \phi^2(t)dt + \phi^2(0) &= c^2\left(\int_0^\infty \frac{1}{(1+t)^{2\alpha}}dt + 1\right) \\ &= c^2\left(\left[\frac{1}{(1-2\alpha)(1+t)^{2\alpha-1}}\right]_{t=0}^\infty + 1\right) \\ &= \frac{2\alpha c^2}{2\alpha - 1} \end{aligned}$$

and

$$\begin{aligned}\int_0^k \phi(t)dt &= c\int_0^k \frac{1}{(t+1)^\alpha}dt \\ &= c\left[\frac{1}{(1-\alpha)(t+1)^{\alpha-1}}\right]_{t=0}^k \\ &= \frac{c}{1-\alpha}\left((k+1)^{1-\alpha}-1\right).\end{aligned}$$

We conclude that

$$f(x_k) - f(x^\star) \leq \frac{V + D\dfrac{2\alpha c^2}{2\alpha - 1}}{\dfrac{bc}{1-\alpha}\left((k+1)^{1-\alpha}-1\right)}$$

for each integer $k \geq 0$, which shows a $O(1/k^{1-\alpha})$ sublinear rate of convergence.

3. Note that $\alpha \in (0.5, 1)$ implies that $1-\alpha \in (0, 0.5)$. Therefore, we get that

$$\frac{1/k^{1-\alpha}}{1/\log k} = \frac{\log k}{k^{1-\alpha}} \to 0 \quad \text{as} \quad k \to \infty.$$

Thus, step-size (b) gives the fastest convergence rate.

Solution 6.8

The Lyapunov inequality (6.3) gives that

$$\|x_k - x^\star\|_2^2 \leq \|x_0 - x^\star\|_2^2 - 2\gamma\sum_{i=1}^k (f(x_i) - f(x^\star))$$

holds inductively for each integer $k \geq 1$. Therefore,

$$\begin{aligned}\sum_{i=1}^k (f(x_i) - f(x^\star)) &\leq \frac{\|x_0 - x^\star\|_2^2 - \|x_k - x^\star\|_2^2}{2\gamma} \\ &\leq \frac{\|x_0 - x^\star\|_2^2}{2\gamma}\end{aligned} \tag{7.53}$$

for each integer $k \geq 1$, since $\|x_k - x^\star\|_2^2 \geq 0$ for each integer $k \geq 1$. Furthermore,

$$k(f(x_k) - f(x^\star)) \leq \sum_{i=1}^k (f(x_i) - f(x^\star)) \tag{7.54}$$

for each integer $k \geq 1$, since $(x_k)_{k=0}^\infty$ is a descent sequence for f. Combining (7.53) and (7.54) gives

$$0 \leq f(x_k) - f(x^\star) \leq \frac{\|x_0 - x^\star\|_2^2}{2\gamma k} \to 0 \quad \text{as} \quad k \to \infty. \tag{7.55}$$

The squeeze theorem gives that

$$f(x_k) - f(x^\star) \to 0 \quad \text{as} \quad k \to \infty$$

or equivalently

$$f(x_k) \to f(x^\star) \quad \text{as} \quad k \to \infty.$$

Moreover, we identify a $O(1/k)$ sublinear rate of convergence from (7.55).

Solution 6.9

1. We start from the inequality

$$\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 \mid x_k\right] \le \|x_k - x^\star\|_2^2 - 2\gamma_k(f(x_k) - f(x^\star)) + \gamma_k^2 G^2$$

for each integer $k \ge 0$. By monotonicity and linearity of expectation, we get that

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 \mid x_k\right]\right] &\le \mathbb{E}\left[\|x_k - x^\star\|_2^2 - 2\gamma_k(f(x_k) - f(x^\star)) + \gamma_k^2 G^2\right] \\
&= \mathbb{E}\left[\|x_k - x^\star\|_2^2\right] - \mathbb{E}\left[2\gamma_k(f(x_k) - f(x^\star))\right] + \mathbb{E}\left[\gamma_k^2 G^2\right] \\
&= \mathbb{E}\left[\|x_k - x^\star\|_2^2\right] - 2\gamma_k \, \mathbb{E}\left[f(x_k) - f(x^\star)\right] + \gamma_k^2 G^2,
\end{aligned}$$

for each integer $k \ge 0$, since G and γ_k, for each integer $k \ge 0$, are deterministic. The law of total expectation gives that

$$\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2\right] \le \mathbb{E}\left[\|x_k - x^\star\|_2^2\right] - 2\gamma_k \, \mathbb{E}\left[f(x_k) - f(x^\star)\right] + \gamma_k^2 G^2$$

for each integer $k \ge 0$. This is the Lyapunov inequality we pick.

2. The Lyapunov inequality above gives that

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2\right] &\le \mathbb{E}\left[\|x_0 - x^\star\|_2^2\right] - 2\sum_{i=0}^{k} \gamma_i \, \mathbb{E}\left[f(x_i) - f(x^\star)\right] + G^2 \sum_{i=0}^{k} \gamma_i^2 \\
&= \|x_0 - x^\star\|_2^2 - 2\sum_{i=0}^{k} \gamma_i \, \mathbb{E}\left[f(x_i) - f(x^\star)\right] + G^2 \sum_{i=0}^{k} \gamma_i^2
\end{aligned}$$

holds inductively for each integer $k \ge 0$, since $\|x_0 - x^\star\|_2^2$ is deterministic. Again, by monotonicity of expectation, we know that

$$0 \le \mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2\right]$$

for each integer $k \ge 0$ since

$$0 \le \|x_{k+1} - x^\star\|_2^2$$

for each integer $k \ge 0$. We conclude that

$$0 \le \|x_0 - x^\star\|_2^2 - 2\sum_{i=0}^{k} \gamma_i \, \mathbb{E}\left[f(x_i) - f(x^\star)\right] + G^2 \sum_{i=0}^{k} \gamma_i^2$$

for each integer $k \ge 0$, or by rearranging

$$2\sum_{i=0}^{k} \gamma_i \, \mathbb{E}[f(x_i) - f(x^\star)] \le \|x_0 - x^\star\|_2^2 + G^2 \sum_{i=0}^{k} \gamma_i^2$$

for each integer $k \ge 0$, as desired.

Solution 6.10

1. First, we prove the claim provided in the hint, i.e.

$$\lambda_k \geq 1 + \frac{k}{2} \tag{7.56}$$

for each integer $k \geq 0$. Clearly, (7.56) holds for $k = 0$. Note that

$$\begin{aligned}\lambda_k &= \frac{1 + \sqrt{1 + 4\lambda_{k-1}^2}}{2} \\ &\geq \frac{1}{2} + \lambda_{k-1}\end{aligned}$$

holds for each integer $k \geq 1$. This gives that

$$\begin{aligned}\lambda_k &\geq k\frac{1}{2} + \lambda_0 \\ &= 1 + \frac{k}{2}\end{aligned}$$

holds inductively for each integer $k \geq 1$. This establishes (7.56).

Next, rearranging (6.4) and recursive application gives

$$\begin{aligned}\frac{2\lambda_{k+1}^2}{\beta}(f(x_{k+1}) - f(x^\star)) &\leq V_k - V_{k+1} + \frac{2\lambda_k^2}{\beta}(f(x_k) - f(x^\star)) \\ &\leq V_1 - V_{k+1} + \frac{2\lambda_1^2}{\beta}(f(x_1) - f(x^\star)) \\ &= V_1 - V_{k+1} + \frac{2\lambda_1^2}{\beta}(f(x_1) - f(x^\star)) \\ &\leq V_1 + \frac{2\lambda_1^2}{\beta}(f(x_1) - f(x^\star))\end{aligned}$$

for each integer $k \geq 1$, since $V_k \geq 0$ for each integer $k \geq 1$. Using (7.56), we get that

$$\begin{aligned}f(x_{k+1}) - f(x^\star) &\leq \frac{V_1 + \frac{2\lambda_1^2}{\beta}(f(x_1) - f(x^\star))}{\frac{2\lambda_{k+1}^2}{\beta}} \\ &\leq \frac{V_1 + \frac{2\lambda_1^2}{\beta}(f(x_1) - f(x^\star))}{\frac{2}{\beta}\left(1 + \frac{k+1}{2}\right)^2}\end{aligned}$$

for each integer $k \geq 1$, or equivalently

$$f(x_k) - f(x^\star) \leq \frac{2\beta V_1 + 4\lambda_1^2(f(x_1) - f(x^\star))}{(k+2)^2} \tag{7.57}$$

for each integer $k \geq 2$. Note that

$$0 \leq f(x_k) - f(x^\star) \leq \frac{2\beta V_1 + 4\lambda_1^2(f(x_1) - f(x^\star))}{(k+2)^2} \to 0 \quad \text{as} \quad k \to \infty.$$

The squeeze theorem gives that

$$f(x_k) - f(x^\star) \to 0 \quad \text{as} \quad k \to \infty$$

or equivalently

$$f(x_k) \to f(x^\star) \quad \text{as} \quad k \to \infty.$$

Moreover, we identify a $O(1/k^2)$ sublinear rate of convergence.

2. From (7.57), if $k \geq 2$, we know that

$$f(x_k) - f(x^\star) \leq \frac{2\beta V_1 + 4\lambda_1^2(f(x_1) - f(x^\star))}{(k+2)^2}.$$

Therefore, if the integer $k \geq 2$ is so large such that

$$\frac{2\beta V_1 + 4\lambda_1^2(f(x_1) - f(x^\star))}{(k+2)^2} \leq \epsilon$$

we obtain an ϵ-accurate objective value. This is equivalently to

$$k \geq \left\lceil \sqrt{\frac{2\beta V_1 + 4\lambda_1^2(f(x_1) - f(x^\star))}{\epsilon}} - 2 \right\rceil \quad \text{and} \quad k \geq 2$$

or simply

$$k \geq \max\left(\left\lceil \sqrt{\frac{2\beta V_1 + 4\lambda_1^2(f(x_1) - f(x^\star))}{\epsilon}} - 2 \right\rceil, 2 \right).$$

Solutions to chapter 7

Solution 7.1

1. We know that the inequality

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{\beta}{2} \|x_{k+1} - x_k\|_2^2$$

holds since f is β-smooth. Using the update rule (7.1) we get that

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T \left((x_k - \gamma \nabla f(x_k)) - x_k\right) + \frac{\beta}{2} \|(x_k - \gamma \nabla f(x_k)) - x_k\|_2^2 \\ &= f(x_k) - \gamma \left(1 - \frac{\beta\gamma}{2}\right) \|\nabla f(x_k)\|_2^2 \end{aligned}$$

Subtracting $f(x^\star)$ from both sides gives that

$$(f(x_{k+1}) - f(x^\star)) \leq (f(x_k) - f(x^\star)) - \gamma \left(1 - \frac{\beta\gamma}{2}\right) \|\nabla f(x_k)\|_2^2$$

as desired.

2. The Lyapunov inequality (7.2) can be written as

$$\gamma \left(1 - \frac{\beta\gamma}{2}\right) \|\nabla f(x_i)\|_2^2 \leq (f(x_i) - f(x^\star)) - (f(x_{i+1}) - f(x^\star))$$

for each integer $i \geq 0$. Summing over $i = 0, \ldots, k$ gives that

$$\begin{aligned} \gamma \left(1 - \frac{\beta\gamma}{2}\right) \sum_{i=0}^{k} \|\nabla f(x_i)\|_2^2 &\leq \sum_{i=0}^{k} \left((f(x_i) - f(x^\star)) - (f(x_{i+1}) - f(x^\star))\right) \\ &= (f(x_0) - f(x^\star)) - (f(x_{k+1}) - f(x^\star)) \\ &\leq f(x_0) - f(x^\star) \end{aligned}$$

for each integer $k \geq 0$, since $f(x_{k+1}) - f(x^\star) \geq 0$ by assumption. Suppose that $0 < \gamma < \frac{2}{\beta}$. Then we get that

$$\sum_{i=0}^{k} \|\nabla f(x_i)\|_2^2 \leq \frac{f(x_0) - f(x^\star)}{\gamma \left(1 - \frac{\beta\gamma}{2}\right)} \tag{7.58}$$

for each integer $k \geq 0$. In particular, we see that

$$\sum_{i=0}^{\infty} \|\nabla f(x_i)\|_2^2 \leq \frac{f(x_0) - f(x^\star)}{\gamma\left(1 - \frac{\beta\gamma}{2}\right)}$$

and conclude that

$$\|\nabla f(x_k)\|_2^2 \to 0 \quad \text{as} \quad k \to \infty$$

as desired.

3. Using inequality (7.58), we get that

$$\begin{aligned}(k+1) \min_{i=0,\ldots,k} \|\nabla f(x_i)\|_2^2 &\leq \sum_{i=0}^{k} \|\nabla f(x_i)\|_2^2 \\ &\leq \frac{f(x_0) - f(x^\star)}{\gamma\left(1 - \frac{\beta\gamma}{2}\right)}.\end{aligned}$$

for each integer $k \geq 0$. Dividing by $k+1$ gives that

$$0 \leq \min_{i=0,\ldots,k} \|\nabla f(x_i)\|_2^2 \leq \frac{f(x_0) - f(x^\star)}{\gamma\left(1 - \frac{\beta\gamma}{2}\right)(k+1)}.$$

for each integer $k \geq 0$. We identified a $O(1/k)$ sublinear rate of convergence.

Solution 7.2

1. Plugging in the update rule (7.3) into $\|x_{k+1} - x^\star\|_2^2$ and expanding gives that

$$\begin{aligned}\|x_{k+1} - x^\star\|_2^2 &= \|x_k - \gamma\nabla f(x_k) - x^\star\|_2^2 \\ &= \|x_k - x^\star\|_2^2 - 2\gamma\nabla f(x_k)^T(x_k - x^\star) + \gamma^2 \|\nabla f(x_k)\|_2^2.\end{aligned}$$

The first order condition for convexity gives that

$$f(x^\star) \geq f(x_k) + \nabla f(x_k)^T(x^\star - x_k)$$

which is equivalently to that

$$-\nabla f(x_k)^T(x_k - x^\star) \leq -\left(f(x_k) - f(x^\star)\right).$$

Therefore,

$$\|x_{k+1} - x^\star\|_2^2 \leq \|x_k - x^\star\|_2^2 - 2\gamma\left(f(x_k) - f(x^\star)\right) + \gamma^2 \|\nabla f(x_k)\|_2^2. \tag{7.59}$$

From Exercise 7.1, we have the Lyapunov inequality (7.2), i.e.

$$\left(f(x_{k+1}) - f(x^\star)\right) \leq \left(f(x_k) - f(x^\star)\right) - \gamma\left(1 - \frac{\beta\gamma}{2}\right) \|\nabla f(x_k)\|_2^2.$$

Adding $f(x^\star)$ to both sides, multiplying by 2γ and rearranging gives that

$$-2\gamma f(x_k) \leq -2\gamma f(x_{k+1}) - \gamma^2(2-\beta\gamma)\left\|\nabla f(x_k)\right\|_2^2. \tag{7.60}$$

Inserting this into (7.59) gives that

$$\|x_{k+1}-x^\star\|_2^2 \leq \|x_k - x^\star\|_2^2 - 2\gamma(f(x_{k+1})-f(x^\star)) + \gamma^2(\beta\gamma-1)\left\|\nabla f(x_k)\right\|_2^2$$

as desired.

2. The inequality (7.4) can be written as

$$2\gamma(f(x_{i+1})-f(x^\star)) \leq \|x_i - x^\star\|_2^2 - \|x_{i+1}-x^\star\|_2^2 + \gamma^2(\beta\gamma-1)\left\|\nabla f(x_i)\right\|_2^2$$

for each integer $i \geq 0$. Summing over $i = 0,\ldots,k$ gives that

$$\begin{aligned}
2\gamma\sum_{i=0}^{k}(f(x_{i+1})-f(x^\star)) &\leq \sum_{i=0}^{k}\left(\|x_i - x^\star\|_2^2 - \|x_{i+1}-x^\star\|_2^2 + \gamma^2(\beta\gamma-1)\left\|\nabla f(x_i)\right\|_2^2\right)\\
&= \|x_0 - x^\star\|_2^2 - \|x_{k+1}-x^\star\|_2^2 + \gamma^2(\beta\gamma-1)\sum_{i=0}^{k}\left\|\nabla f(x_i)\right\|_2^2\\
&\leq \|x_0 - x^\star\|_2^2 + \gamma^2(\beta\gamma-1)\sum_{i=0}^{k}\left\|\nabla f(x_i)\right\|_2^2
\end{aligned}$$

for each integer $k \geq 0$. Note that

$$\gamma < \frac{2}{\beta}$$

implies that

$$\beta\gamma - 1 < 1.$$

Therefore,

$$\begin{aligned}
2\gamma\sum_{i=0}^{k}(f(x_{i+1})-f(x^\star)) &\leq \|x_0 - x^\star\|_2^2 + \gamma^2\sum_{i=0}^{k}\left\|\nabla f(x_i)\right\|_2^2\\
&\leq \|x_0 - x^\star\|_2^2 + \gamma^2\sum_{i=0}^{\infty}\left\|\nabla f(x_i)\right\|_2^2 \qquad (7.61)
\end{aligned}$$

for each integer $k \geq 0$. Note that inequality (7.2) implies that $(x_i)_{i=0}^{\infty}$ is a decent sequence for f, i.e. $(f(x_i))_{i=1}^{\infty}$ is nonincreasing. This implies that

$$2\gamma(k+1)(f(x_{k+1})-f(x^\star)) \leq 2\gamma\sum_{i=0}^{k}(f(x_{i+1})-f(x^\star))$$

which combined with (7.61) gives that

$$0 \leq f(x_{k+1}) - f(x^\star) \leq \frac{\|x_0 - x^\star\|_2^2 + \gamma^2\sum_{i=0}^{\infty}\left\|\nabla f(x_i)\right\|_2^2}{2\gamma(k+1)} \to 0 \quad \text{as} \quad k \to \infty$$

since $\sum_{i=0}^{\infty}\left\|\nabla f(x_i)\right\|_2^2$ is bounded. We conclude that

$$f(x_k) \to f(x^\star) \quad \text{as} \quad k \to \infty$$

and identify a $O(1/k)$ sublinear rate of convergence.

Solution 7.3

1. Note that

$$\begin{aligned}\|x_{k+1} - x^\star\|_2^2 &= \|(x_k - \gamma\nabla f(x_k)) - x^\star\|_2^2 \\ &= \|x_k - x^\star\|_2^2 - 2\gamma\nabla f(x_k)^T(x_k - x^\star) + \gamma^2 \|\nabla f(x_k)\|_2^2 .\end{aligned}$$

The first order condition for strong convexity gives that

$$f(x^\star) \geq f(x_k) + \nabla f(x_k)^T(x^\star - x_k) + \frac{\sigma}{2}\|x^\star - x_k\|$$

which is equivalently to that

$$-\nabla f(x_k)^T(x_k - x^\star) \leq -(f(x_k) - f(x^\star)) - \frac{\sigma}{2}\|x^\star - x_k\| .$$

Therefore,

$$\|x_{k+1} - x^\star\|_2^2 = (1 - \sigma\gamma)\|x_k - x^\star\|_2^2 - 2\gamma(f(x_k) - f(x^\star)) + \gamma^2\|\nabla f(x_k)\|_2^2 . \quad \textbf{(7.62)}$$

Recall inequality (7.60) from Exercise 7.2, i.e.

$$-2\gamma f(x_k) \leq -2\gamma f(x_{k+1}) - \gamma^2(2 - \beta\gamma)\|\nabla f(x_k)\|_2^2 .$$

Using inequality (7.60) in (7.62) gives that

$$\begin{aligned}\|x_{k+1} - x^\star\|_2^2 &= (1 - \sigma\gamma)\|x_k - x^\star\|_2^2 - \underbrace{2\gamma(f(x_{k+1}) - f(x^\star))}_{\geq 0} + \underbrace{\gamma^2(\beta\gamma - 1)\|\nabla f(x_k)\|_2^2}_{\leq 0 \text{ since } \gamma \leq 1/\beta} \\ &\leq (1 - \sigma\gamma)\|x_k - x^\star\|_2^2\end{aligned}$$

as desired. The fastest convergence rate is obtained when $1 - \sigma\gamma$ is minimized which in turn happens when γ is maximized. Since γ is upper bounded by $1/\beta$, the fastest convergence rate is obtained when

$$\gamma = \frac{1}{\beta}$$

which gives the convergence rate

$$1 - \frac{\sigma}{\beta} = \frac{\beta - \sigma}{\beta}.$$

2. Let $g : \mathbb{R} \to \mathbb{R}$ such that

$$g(\gamma) = \max(1 - \sigma\gamma, \beta\gamma - 1)$$

for each $\gamma \in \mathbb{R}$. The step-size that gives the fastest convergence rate is the one that minimizes g. Since g is the maximum of two affine functions it is closed and convex. Fermat's rule then give that the best step-size γ satisfies

$$0 \in \partial g(\gamma) = \begin{cases} \{-\sigma\} & \text{if } 1 - \sigma\gamma > \beta\gamma - 1, \\ \{\beta\} & \text{if } 1 - \sigma\gamma < \beta\gamma - 1, \\ [-\sigma, \beta] & \text{if } 1 - \sigma\gamma = \beta\gamma - 1. \end{cases}$$

Clearly, this holds only when $1 - \sigma\gamma = \beta\gamma - 1$, i.e., when

$$\gamma = \frac{2}{\beta + \sigma}$$

which gives the convergence rate

$$g\left(\frac{2}{\beta + \sigma}\right) = \frac{\beta - \sigma}{\beta + \sigma}.$$

3. From the analysis earlier in the previous subproblem, we get that the fastest convergence rate is given by

$$\|x_{k+1} - x^\star\|_2^2 \leq \frac{\beta - \sigma}{\beta} \|x_k - x^\star\|_2^2.$$

From the analysis in the lectures we get that the fastest convergence rate is given by

$$\|x_{k+1} - x^\star\|_2^2 \leq \left(\frac{\beta - \sigma}{\beta + \sigma}\right)^2 \|x_k - x^\star\|_2^2.$$

Since $0 < \sigma \leq \beta$, we have that

$$\left(\frac{\beta - \sigma}{\beta + \sigma}\right)^2 \leq \frac{\beta - \sigma}{\beta + \sigma} \leq \frac{\beta - \sigma}{\beta}.$$

Thus, the convergence analysis in the lectures yields a faster convergence rate.

Solution 7.4

1. Let $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x) = \frac{1}{2}x^T Q x + q^T x$$

for each $x \in \mathbb{R}^n$. We have that

$$\nabla f(x) = Qx + q$$

and

$$\nabla^2 f(x) = Q \succeq \lambda_{\min}(Q) I$$

for each $x \in \mathbb{R}^n$, where $\lambda_{\min}(Q) > 0$ since $Q \in \mathbb{S}^n_{++}$. The second order condition for strong convexity gives that x^* is the unique global minimizer of f. Fermat's rule gives that x^* is the global minimizer of f if and only if

$$\nabla f(x^\star) = 0.$$

Note that

$$\begin{aligned} x_{k+1} &= x_k - \gamma \nabla f(x_k) \\ &= (I - \gamma Q) x_k - \gamma q \end{aligned}$$

and

$$\begin{aligned} x^* &= x^* - \gamma \nabla f(x^*) \\ &= (I - \gamma Q)x^* - \gamma q. \end{aligned}$$

Therefore,

$$\begin{aligned} \|x_{k+1} - x^*\|_2 &= \|(I - \gamma Q)(x_k - x^*)\|_2 \\ &\leq \|I - \gamma Q\|_2 \, \|x_k - x^*\|_2 \,. \end{aligned}$$

Note that

$$\lambda_i(I - \gamma Q) = 1 - \gamma \lambda_{n-i+1}(Q) \tag{7.63}$$

and therefore

$$-1 < \lambda_i(I - \gamma Q) < 1$$

for each $i = 1, \ldots, n$, since $\gamma \in (0, 2/\beta)$ where

$$\begin{aligned} \beta &= \|Q\|_2 \\ &= \lambda_{\max}(Q) \end{aligned}$$

since $Q \in \mathbb{S}^n_{++}$. We see that

$$\begin{aligned} \|I - \gamma Q\|_2 &= \sigma_{\max}(I - \gamma Q) \\ &= \sqrt{\lambda_{\max}\left((I - \gamma Q)^2\right)} \\ &= \max_{i=1,\ldots,n} |\lambda_i(I - \gamma Q)| \end{aligned}$$

and conclude that

$$0 \leq \|I - \gamma Q\|_2 < 1.$$

2. Suppose that $\gamma = 1/\beta$. This implies that

$$0 < \lambda_i(I - \gamma Q) < 1$$

by (7.63). Therefore $I - \gamma Q \in \mathbb{S}^n_{++}$ and

$$\begin{aligned} \|I - \gamma Q\|_2 &= \lambda_{\max}(I - \gamma Q) \\ &= 1 - \gamma \lambda_{\min}(Q) \\ &= 1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}. \end{aligned}$$

3. Note that

$$\lambda_{\min}(Q) = \epsilon \quad \text{and} \quad \lambda_{\max}(Q) = 1$$

and therefore

$$\begin{aligned} \rho &= \|I - \gamma Q\|_2 \\ &= 1 - \epsilon. \end{aligned}$$

If $q = 0$, then $x^* = 0$ is the unique global minimizer of f. If we pick

$$x_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

then

$$x_k = \begin{bmatrix} (1-\epsilon)^k \\ 0 \end{bmatrix}$$

and the linear convergence rate is achieved, i.e. the inequality

$$\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\|$$

becomes an equality. In particular, if ϵ is very small compared to 1, we get a slow convergence.

4. Let

$$V = \begin{bmatrix} \frac{1}{\sqrt{\epsilon}} & 0 \\ 0 & 1 \end{bmatrix}.$$

Then

$$V^T Q V = \begin{bmatrix} 1 & \frac{1}{10} \\ \frac{1}{10} & 1 \end{bmatrix}.$$

5. $V^T Q V$ has the eigenvalues 0.99 and 1.01. The convergence will therefore be very fast. Indeed, with

$$\gamma = \frac{1}{1.01}$$

we get the linear rate of convergence

$$\rho = 1 - \frac{0.99}{1.1} \approx 0.02.$$

6. Suppose that V is not diagonal. The proximal operator is often computed on some function g that is separable. With the change of variables to $x = Vy$, we need compute the proximal operator of the function $g \circ V$ which in general is no longer separable. Computing the proximal operator on this term generally becomes computationally expensive.

Solution 7.5

1. Recall that

$$\operatorname{prox}_{\gamma f}(x) = \underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \left(f(z) + \frac{1}{2\gamma} \|x - z\|_2^2 \right)$$

for each $x \in \mathbb{R}^n$. Thus, if

$$x_{k+1} = \operatorname{prox}_{\gamma f}(x_k)$$

then

$$f(x_{k+1}) + \frac{1}{2\gamma} \|x_{k+1} - x_k\|_2^2 \leq f(z) + \frac{1}{2\gamma} \|z - x_k\|_2^2$$

for each $z \in \mathbb{R}^n$. Setting $z = x_k$ gives that

$$f(x_{k+1}) + \frac{1}{2\gamma} \|x_{k+1} - x_k\|_2^2 \leq f(x_k)$$

or equivalently

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2\gamma} \|x_{k+1} - x_k\|_2^2 \tag{7.64}$$

as desired.

2. Inequality (7.64) can be written as

$$\frac{1}{2\gamma} \|x_{k+1} - x_k\|_2^2 \leq f(x_k) - f(x_{k+1})$$

Summing over $k = 0, \ldots, l$ gives that

$$\begin{aligned} \frac{1}{2\gamma} \sum_{k=0}^{l} \|x_{k+1} - x_k\|_2^2 &\leq \sum_{k=0}^{l} (f(x_k) - f(x_{k+1})) \\ &\leq f(x_0) - f(x_{l+1}) \\ &\leq f(x_0) - B \end{aligned}$$

and therefore

$$\sum_{k=0}^{l} \|x_{k+1} - x_k\|_2^2 \leq 2\gamma (f(x_0) - B)$$

for each integer $l \geq 0$. In particular, we see that

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|_2^2 \leq 2\gamma (f(x_0) - B) < \infty$$

and conclude that

$$\|x_{k+1} - x_k\|_2^2 \to 0 \quad \text{as} \quad k \to \infty$$

as desired.

3. Suppose that

$$\|x_{k+1} - x_k\|_2^2 \to 0 \quad \text{as} \quad k \to \infty.$$

Fermat's rule gives that

$$x_{k+1} = \operatorname{prox}_{\gamma f}(x_k)$$

is equivalent to

$$0 \in \partial f(x_{k+1}) + \frac{1}{\gamma} (x_{k+1} - x_k)$$

or equivalently

$$\frac{1}{\gamma}(x_{k-1} - x_k) \in \partial f(x_k).$$

This implies that

$$\begin{aligned} 0 \le \operatorname{dist}_{\partial f(x_k)}(0) &\le \left\| \frac{1}{\gamma}(x_{k-1} - x_k) - 0 \right\|_2 \\ &= \frac{1}{\gamma}\|x_{k-1} - x_k\|_2 \to 0 \quad \text{as} \quad k \to \infty. \end{aligned}$$

The squeeze theorem gives that

$$\operatorname{dist}_{\partial f(x_k)}(0) \to 0 \quad \text{as} \quad k \to \infty$$

as desired.

4. Note that f is lower bounded by $f(x^*)$. Therefore,

$$\|x_k - x_{k-1}\|_2 \to 0 \quad \text{as} \quad k \to \infty$$

by a previous subproblem. The σ-strong convexity of f implies that

$$f(y) \ge f(x) + s^T(y - x) + \frac{\sigma}{2}\|y - x\|_2^2$$

for each $x, y \in \mathbb{R}^n$ and each $s \in \partial f(x)$. In particular, we have that

$$\begin{aligned} \frac{1}{\gamma}(x_{k-1} - x_k) \in \partial f(x_k) &\Rightarrow f(x^*) \ge f(x_k) + \frac{1}{\gamma}(x_{k-1} - x_k)^T(x^* - x_k) + \frac{\sigma}{2}\|x^* - x_k\|_2^2 \\ 0 \in \partial f(x^\star) &\Rightarrow f(x_k) \ge f(x^\star) + \frac{\sigma}{2}\|x_k - x^\star\|_2^2. \end{aligned}$$

Adding these two inequalities together and unsing the Cauchy-Schwarz inequality gives that

$$\begin{aligned} \|x_k - x^\star\|_2^2 &\le \frac{1}{\gamma\sigma}(x_k - x_{k-1})^T(x^* - x_k) \\ &\le \frac{1}{\gamma\sigma}\|x_k - x_{k-1}\|_2\|x^* - x_k\|_2 \end{aligned}$$

and therefore

$$0 \le \|x_k - x^\star\|_2 \le \frac{1}{\gamma\sigma}\|x_k - x_{k-1}\|_2 \to 0 \quad \text{as} \quad k \to \infty.$$

The squeeze theorem gives that

$$\|x_k - x^\star\|_2 \to 0 \quad \text{as} \quad k \to \infty$$

or equivalently

$$x_k \to x^\star \quad \text{as} \quad k \to \infty$$

as desired.

Solution 7.6

The complete procedure is given below:

1. The goal is to get a Lyapunov inequality on the form

$$V_{k+1} \leq V_k - Q_k$$

for each integer $k \geq 0$, where $(Q_k)_{k=0}^{\infty}$ is some nonnegative convergence measure and

$$V_k = \|x_k - x^\star\|_2^2$$

for each integer $k \geq 0$. We further define the residual mapping $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\mathcal{R}x = x - \text{prox}_{\gamma g}(x - \gamma \nabla f(x))$$

for each $x \in \mathbb{R}^n$. The proximal gradient update can then be written as

$$x_{k+1} = x_k - \mathcal{R}x_k. \tag{7.65}$$

We can use this to relate V_{k+1} to V_k by

$$\begin{aligned} V_{k+1} &= \|x_{k+1} - x^\star\|_2^2 \\ &= \|(x_k - \mathcal{R}x_k) - x^\star\|_2^2 \\ &= \|x_k - x^\star\|_2^2 - 2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2 \\ &= V_k - 2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2 . \end{aligned} \tag{7.66}$$

2. Next, we wish to upper bound the quantity $-2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2$. We start by using (7.65) to rewrite it as

$$\begin{aligned} -2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2 &= -2(x_k - x^\star)^T(\mathcal{R}x_k) + (\mathcal{R}x_k)^T(\mathcal{R}x_k) \\ &= -2(x_k - x^\star)^T(\mathcal{R}x_k) + 2(\mathcal{R}x_k)^T(\mathcal{R}x_k) - (\mathcal{R}x_k)^T(\mathcal{R}x_k) \\ &= -2(x_k - \mathcal{R}x_k - x^\star)^T(\mathcal{R}x_k) - (\mathcal{R}x_k)^T(\mathcal{R}x_k) \\ &= -2(x_{k+1} - x^\star)^T(\mathcal{R}x_k) - \|\mathcal{R}x_k\|_2^2 . \end{aligned} \tag{7.67}$$

3. We now turn to bounding $-2(x_{k+1} - x^\star)^T(\mathcal{R}x_k)$. Using Fermat's rule on the proximal gradient update gives that

$$0 \in \partial g(x_{k+1}) + \frac{1}{\gamma}(x_{k+1} - (x_k - \gamma \nabla f(x_k)))$$

which is equivalent to that

$$\gamma^{-1}\mathcal{R}x_k - \nabla f(x_k) \in \partial g(x_{k+1})$$

The definition of a subgradient then gives that

$$g(x^\star) \geq g(x_{k+1}) + (\gamma^{-1}\mathcal{R}x_k - \nabla f(x_k))^T(x^\star - x_{k+1})$$

which implies that

$$-2(x_{k+1} - x^\star)^T(\mathcal{R}x_k) \leq -2\gamma(g(x_{k+1}) - g(x^\star)) - 2\gamma \nabla f(x_k)^T(x_{k+1} - x^\star). \tag{7.68}$$

4. We continue to bound $-2\gamma\nabla f(x_k)^T(x_{k+1}-x^\star)$. Using the definition of β-smoothness of f and the first-order condition of convexity on f gives the two following inequalities:

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T(x_{k+1}-x_k) + \frac{\beta}{2}\|x_{k+1}-x_k\|_2^2 \\
&= f(x_k) + \nabla f(x_k)^T(x_{k+1}-x_k) + \frac{\beta}{2}\|\mathcal{R}x_k\|_2^2 \\
f(x^\star) &\geq f(x_k) + \nabla f(x_k)^T(x^\star - x_k).
\end{aligned}$$

Adding these two together and rearranging gives that

$$f(x_{k+1}) \leq f(x^\star) + \nabla f(x_k)^T(x_{k+1}-x^\star) + \frac{\beta}{2}\|\mathcal{R}x_k\|_2^2$$

which implies that

$$-2\gamma\nabla f(x_k)^T(x_{k+1}-x^\star) \leq -2\gamma(f(x_{k+1}) - f(x^\star)) + \gamma\beta\|\mathcal{R}x_k\|_2^2. \tag{7.69}$$

5. Inserting (7.69) into (7.68), (7.68) into (7.67), and (7.67) into (7.66) gives that

$$\begin{aligned}
V_{k+1} &= V_k - 2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2 \\
&= V_k - \|\mathcal{R}x_k\|_2^2 - 2(x_{k+1}-x^\star)^T(\mathcal{R}x_k) \\
&\leq V_k - \|\mathcal{R}x_k\|_2^2 - 2\gamma(g(x_{k+1}) - g(x^\star)) - 2\gamma\nabla f(x_k)^T(x_{k+1}-x^\star) \\
&\leq V_k - \|\mathcal{R}x_k\|_2^2 - 2\gamma(g(x_{k+1}) - g(x^\star)) - 2\gamma(f(x_{k+1}) - f(x^\star)) + \gamma\beta\|\mathcal{R}x_k\|_2^2 \\
&\leq V_k - (1-\gamma\beta)\|\mathcal{R}x_k\|_2^2 - 2\gamma(g(x_{k+1}) + f(x_{k+1}) - g(x^\star) - f(x^\star)).
\end{aligned}$$

6. Using the assumption $\gamma < \beta^{-1}$ gives that

$$\begin{aligned}
V_{k+1} &\leq V_k - (1-\gamma\beta)\|\mathcal{R}x_k\|_2^2 - 2\gamma(g(x_{k+1}) + f(x_{k+1}) - g(x^\star) - f(x^\star)) \\
&\leq V_k - 2\gamma(g(x_{k+1}) + f(x_{k+1}) - g(x^\star) - f(x^\star)) \\
&= V_k - Q_k
\end{aligned}$$

where

$$Q_k = 2\gamma(g(x_{k+1}) + f(x_{k+1}) - g(x^\star) - f(x^\star))$$

which is nonnegative since $\gamma > 0$ and $g(x_{k+1}) + f(x_{k+1}) \geq g(x^\star) + f(x^\star)$ by assumption on $x^\star$.

7. Since $V_k \geq 0$ and $Q_k \geq 0$ we we know that

$$Q_k \to 0 \quad \text{as} \quad k \to \infty$$

which implies that

$$f(x_k) + g(x_k) \to f(x^\star) + g(x^\star) \quad \text{as} \quad k \to \infty.$$

Solution 7.7

1. Following the steps gives that

$$\begin{aligned}
\|x_{k+1} - x^\star\|_2^2 &= \left\|\mathrm{prox}_{\gamma g}(x_k - \gamma\nabla f(x_k)) - x^\star\right\|_2^2 \\
&= \left\|\mathrm{prox}_{\gamma g}(x_k - \gamma\nabla f(x_k)) - \mathrm{prox}_{\gamma g}(x^\star - \gamma\nabla f(x^\star))\right\|_2^2 \\
&\le \frac{1}{1+\sigma_g\gamma}\|x_k - \gamma\nabla f(x_k) - x^\star + \gamma\nabla f(x^\star)\|_2^2 \\
&= \frac{1}{1+\sigma_g\gamma}\left(\|x_k - x^\star\|_2^2 - 2\gamma(\nabla f(x_k) - \nabla f(x^\star))^T(x_k - x^\star) + \gamma^2\|\nabla f(x_k) - \nabla f(x^\star)\|_2^2\right) \\
&\le \frac{1}{1+\sigma_g\gamma}\left(\left(1 - \frac{2\beta\sigma_f\gamma}{\beta+\sigma_f}\right)\|x_k - x^\star\|_2^2 - \gamma\left(\frac{2}{\beta+\sigma_f} - \gamma\right)\|\nabla f(x_k) - \nabla f(x^\star)\|_2^2\right).
\end{aligned}$$

If $\gamma \ge \frac{2}{\beta+\sigma_f}$, then the last term is positive, and we can use β-Lipschitz continuity of ∇f to get that

$$\begin{aligned}
\|x_{k+1} - x^\star\|_2^2 &\le \frac{1}{1+\sigma_g\gamma}\left(\left(1 - \frac{2\beta\sigma_f\gamma}{\beta+\sigma_f}\right)\|x_k - x^\star\|_2^2 - \gamma\left(\frac{2}{\beta+\sigma_f} - \gamma\right)\beta^2\|x_k - x^\star\|_2^2\right) \\
&\le \frac{1}{1+\sigma_g\gamma}\left(1 - \frac{2\beta\sigma_f\gamma}{\beta+\sigma_f} - \frac{2\beta^2\gamma}{\beta+\sigma_f} + \beta^2\gamma^2\right)\|x_k - x^\star\|_2^2 \\
&\le \frac{1}{1+\sigma_g\gamma}\left(1 - \frac{2\beta\sigma_f\gamma + 2\beta^2\gamma}{\beta+\sigma_f} + \beta^2\gamma^2\right)\|x_k - x^\star\|_2^2 \\
&\le \frac{1}{1+\sigma_g\gamma}\left(1 - 2\beta\gamma + \beta^2\gamma^2\right)\|x_k - x^\star\|_2^2 \\
&\le \frac{(\beta\gamma - 1)^2}{1+\sigma_g\gamma}\|x_k - x^\star\|_2^2.
\end{aligned}$$

If $0 < \gamma \le \frac{2}{\beta+\sigma_f}$, then the last term is negative, and we can use

$$\|\nabla f(x_k) - \nabla f(x^\star)\|_2 \ge \sigma_f\|x_k - x^\star\|_2$$

to get that

$$\begin{aligned}
\|x_{k+1} - x^\star\|_2^2 &\le \frac{1}{1+\sigma_g\gamma}\left(\left(1 - \frac{2\beta\sigma_f\gamma}{\beta+\sigma_f}\right)\|x_k - x^\star\|_2^2 - \gamma\left(\frac{2}{\beta+\sigma_f} - \gamma\right)\sigma_f^2\|x_k - x^\star\|_2^2\right) \\
&= \frac{1}{1+\sigma_g\gamma}\left(1 - \frac{2\beta\sigma_f\gamma}{\beta+\sigma_f} - \frac{2\sigma_f^2\gamma}{\beta+\sigma_f} + \gamma^2\sigma_f^2\right)\|x_k - x^\star\|_2^2 \\
&= \frac{1}{1+\sigma_g\gamma}\left(1 - \frac{2\beta\sigma_f\gamma + 2\sigma_f^2\gamma}{\beta+\sigma_f} + \gamma^2\sigma_f^2\right)\|x_k - x^\star\|_2^2 \\
&= \frac{1}{1+\sigma_g\gamma}\left(1 - 2\sigma_f\gamma + \gamma^2\sigma_f^2\right)\|x_k - x^\star\|_2^2 \\
&= \frac{(1 - \sigma_f\gamma)^2}{1+\sigma_g\gamma}\|x_k - x^\star\|_2^2.
\end{aligned}$$

To write these on one common form we use the fact that

$$\max(1 - \sigma_f\gamma, \beta\gamma - 1) = \begin{cases} 1 - \sigma_f\gamma & \text{if } \gamma \in (0, \frac{2}{\beta+\sigma_f}] \\ \beta\gamma - 1 & \text{if } \gamma \in [\frac{2}{\beta+\sigma_f}, \infty) \end{cases} \tag{7.70}$$

which gives the desired inequality, i.e.

$$\|x_{k+1} - x^\star\|_2^2 \le \frac{\max(1 - \sigma_f\gamma, \beta\gamma - 1)^2}{1+\sigma_g\gamma}\|x_k - x^\star\|_2^2.$$

2. From the previous subproblem we see that

$$x_k \to x^* \quad \text{as} \quad k \to \infty$$

with linear convergence if

$$\frac{\max(1-\sigma_f\gamma, \beta\gamma - 1)^2}{1+\sigma_g\gamma} < 1. \tag{7.71}$$

- Assume that $\sigma_f > 0$.
  - If $0 < \gamma < \frac{2}{\beta+\sigma_f}$, (7.70) gives that

$$\begin{aligned}\frac{\max(1-\sigma_f\gamma, \beta\gamma - 1)^2}{1+\sigma_g\gamma} &= \frac{(1-\sigma_f\gamma)^2}{1+\sigma_g\gamma} \\ &\leq (1-\sigma_f\gamma)^2 \\ &< \max\left(1, \left(1 - 2\frac{\sigma_f}{\beta+\sigma_f}\right)^2\right) \\ &= 1.\end{aligned}$$

  - If $\gamma \geq \frac{2}{\beta+\sigma_f}$, (7.70) gives that

$$\begin{gathered}\frac{\max(1-\sigma_f\gamma, \beta\gamma - 1)^2}{1+\sigma_g\gamma} = \frac{(\beta\gamma-1)^2}{1+\sigma_g\gamma} < 1 \\ \Leftrightarrow \\ (\beta\gamma - 1)^2 < 1 + \sigma_g\gamma \\ \Leftrightarrow \\ 1 + \beta^2\gamma^2 - 2\beta\gamma < 1 + \sigma_g\gamma \\ \Leftrightarrow \\ \beta^2\gamma^2 - \gamma\beta(2 + \frac{\sigma_g}{\beta}) < 0 \\ \Leftrightarrow \\ \beta\gamma < 2 + \frac{\sigma_g}{\beta} \\ \Leftrightarrow \\ \gamma < \frac{2}{\beta} + \frac{\sigma_g}{\beta^2}.\end{gathered}$$

  The case $\sigma_f > 0$ can be summarize by that the proximal gradient method converges linearly if $0 < \gamma < \frac{2}{\beta} + \frac{\sigma_g}{\beta^2}$, since $\frac{2}{\beta+\sigma_f} < \frac{2}{\beta} + \frac{\sigma_g}{\beta^2}$.

- Assume that $\sigma_f = 0$. Then (7.71) becomes

$$\frac{\max(1, \beta\gamma - 1)^2}{1+\sigma_g\gamma} < 1$$

  which is impossible if $\sigma_g = 0$.

- Assume that $\sigma_g > 0$.

- If $0 < \gamma < \frac{2}{\beta+\sigma_f}$, then (7.70) gives that
$$\frac{\max(1-\sigma_f\gamma, \beta\gamma-1)^2}{1+\sigma_g\gamma} = \frac{(1-\sigma_f\gamma)^2}{1+\sigma_g\gamma}.$$
We have that
$$0 < \gamma < \frac{2}{\beta+\sigma_f} \le \frac{2}{\beta}$$
$$\Rightarrow$$
$$0 \le \sigma_f\gamma \le \frac{2\sigma_f}{\beta} < 2$$
$$\Rightarrow$$
$$-1 < 1-\sigma_f\gamma \le 1$$
$$\Rightarrow$$
$$(1-\sigma_f\gamma)^2 \le 1$$
Thus, (7.71) holds since $1+\sigma_g\gamma > 1$.

- If $\gamma \ge \frac{2}{\beta+\sigma_f}$, (7.70) gives that
$$\frac{\max(1-\sigma_f\gamma, \beta\gamma-1)^2}{1+\sigma_g\gamma} = \frac{(\beta\gamma-1)^2}{1+\sigma_g\gamma} < 1$$
$$\Leftrightarrow$$
$$\gamma < \frac{2}{\beta} + \frac{\sigma_g}{\beta^2}$$
where the equivalence is shown exactly as in the $\sigma_f > 0$ case above.

The case $\sigma_g > 0$ can be summarize by that the proximal gradient method converges linearly if $0 < \gamma < \frac{2}{\beta} + \frac{\sigma_g}{\beta^2}$, since $\frac{2}{\beta+\sigma_f} < \frac{2}{\beta} + \frac{\sigma_g}{\beta^2}$.

To summarize all cases, the proximal gradient method converges linearly if $0 < \gamma < \frac{2}{\beta} + \frac{\sigma_g}{\beta^2}$ and at least one of $\sigma_f > 0$ and $\sigma_g > 0$ holds.

3. • Let $\delta = 1$. Then $\beta = L + \sigma$, $\sigma_f = \sigma$ and $\sigma_g = 0$. The linear convergence rate is then given by
$$\max(1-\sigma\gamma, (L+\sigma)\gamma - 1)^2.$$
In Exercise 7.3 we have already shown that this is minimized by
$$\gamma = \frac{2}{L+2\sigma}$$
which is a valid step-size by the analysis above, and results in the linear convergence rate
$$\left(1 - \frac{2\sigma}{L+2\sigma}\right)^2 = \left(\frac{L}{L+2\sigma}\right)^2.$$

• Let $\delta = 0$. Then $\beta = L$, $\sigma_f = 0$ and $\sigma_g = \sigma$. The linear convergence rate is then given by
$$\frac{\max(1, L\gamma - 1)^2}{1+\sigma\gamma}.$$
Next, we split this up into two subcases with respect to valid step-sizes:

- Suppose that $0 < \gamma \leq \frac{2}{L}$. The rate is then

  $$\frac{1}{1+\sigma\gamma}.$$

  Hence, γ should be chosen as large as possible, i.e. $\gamma = \frac{2}{L}$, giving the linear convergence rate

  $$\frac{L}{L+2\sigma}.$$

- Suppose that $\frac{2}{L} \leq \gamma < \frac{2}{L} + \frac{\sigma}{L^2}$. The rate is then

  $$\frac{(L\gamma-1)^2}{1+\sigma\gamma}.$$

  Taking the derivative of the rate gives

  $$\begin{aligned}\frac{d}{d\gamma}\frac{(L\gamma-1)^2}{1+\sigma\gamma} &= \frac{2L(L\gamma-1)}{1+\sigma\gamma} - \frac{\sigma(L\gamma-1)^2}{(1+\sigma\gamma)^2} \\ &= \frac{L\gamma-1}{(1+\sigma\gamma)^2}(2L(1+\sigma\gamma) - \sigma(L\gamma-1)) \\ &= \frac{L\gamma-1}{(1+\sigma\gamma)^2}(2L + L\sigma\gamma + \sigma) \\ &\geq 0.\end{aligned}$$

  Hence, the rate increasing in γ and the step-size should be chosen as small as possible, i.e. $\gamma = \frac{2}{L}$, again giving the linear convergence rate

  $$\frac{L}{L+2\sigma}.$$

To summarize, the best linear convergence rate we can get for the case $\delta = 0$ is

$$\frac{L}{L+2\sigma}.$$

Note that

$$\frac{L}{L+2\sigma} > \left(\frac{L}{L+2\sigma}\right)^2$$

since $L > 0$ and $\sigma > 0$. It is therefore advantageous two put the strong convexity in the gradient step.