Exercises in FRTN50 – Optimization for Learning

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

Created: 2019
Latest update: October 5, 2020

Introduction

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

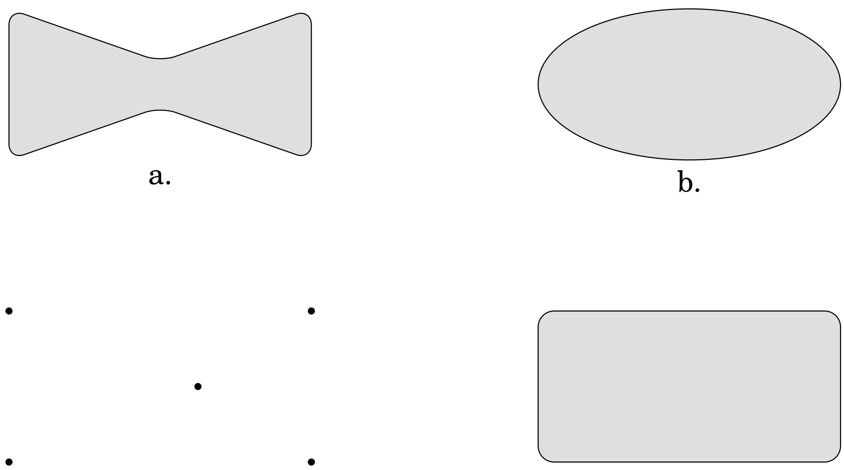
Exercises marked with (H) have hints available, listed in the end of each chapter. Not as fundamental or challenging exercises are marked with $(\star)$. Even more challenging exercises are marked with $(\star\star)$.

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 (H)

Which of the following sets are convex? If convex, prove it using the definition of convex sets, if not convex, disprove it. You may assume that the data defining the sets generate nonempty sets.

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. $S = \{X \in \mathbb{R}^{n \times n} : X \succeq 0\}$ (i.e. X is restricted to be positive semidefinite)
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 $a \neq b$

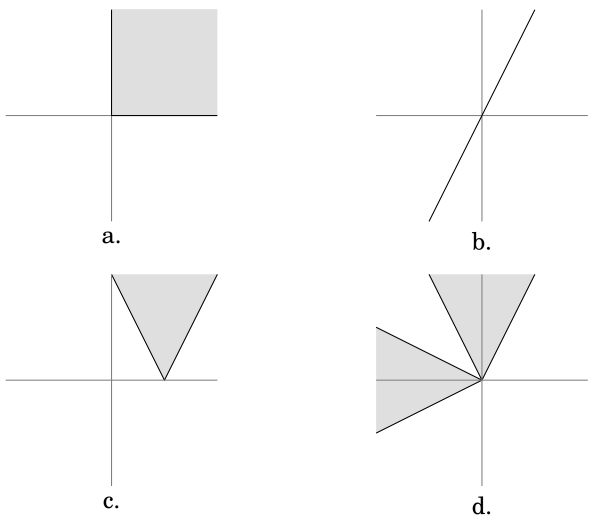
Exercise 1.3

Which of the following sets are affine?

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

Exercise 1.4

A set K is a cone if for all $x \in K$ also $\alpha x \in K$ for all $\alpha \geq 0$. 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. You can assume that the data defining the sets generate nonempty sets.

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 $A \neq 0$ and $b \neq 0$
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.

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 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 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

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

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

EXERCISE 1.9 (H)

Prove, without explicitly using the definition of convex sets, that each of the following sets are convex set. You may assume that the data defining the sets generate nonempty sets.

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 convex, i.e., let f satisfy

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

for all $\theta = [0, 1]$ and let X be (effective) domain of f, i.e. $X = \text{dom} f = \{x \in \mathbb{R}^n : f(x) < \infty\}$. Show that X is convex.

Exercise 1.12

Prove or disprove that the following functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are convex.

1. Indicator function of convex set C:

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

2. $f(x) = \|x\|$
3. $f(x) = -\|x\|$
4. $f(x, y) = xy$
5. $f(x) = a^T x + b$
6. $f(x) = \frac{1}{2} x^T Q x$ with $Q \in \mathbb{R}^{n \times n}$ such that $Q \succeq 0$
7. $f(x) = \text{dist}_C(x) = \inf_{y \in C} \|x - y\|$ where C is a convex set

Exercise 1.13

Draw the epigraph of the following functions.

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

Exercise 1.14

Let $f : \mathbb{R}^n \to \mathbb{R}$ be an affine function defined by $f(x) = a^T x + b$. Show that 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 epi$f = \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq r\}$. Show that f is convex if and only if epi$f$ is convex.

Exercise 1.16 (H)

For each $i \in \{1, \ldots, m\}$, 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 $i \in \{1, \ldots, m\}$
2. Show that $f(x) = \max_{i \in \{1,\ldots,m\}} f_i(x)$ 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. $f(x) = \|x\|^p$ with $p \geq 1$
2. $f(x) = \|Ax - b\|_2^2 + \|x\|_1$
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. $f(x) = \sup_y (x^T y - g(y))$ (these will be called conjugate functions)

EXERCISE 1.18 (H)

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

1. Suppose that g is a convex function for which $\bar{x} \in \mathbb{R}^n$ exists with $g(\bar{x}) < \alpha$. 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. $S = \{x \in \mathbb{R}^n : \|x\|_2 \le 1\}$
2. $S = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} : \|x\|_2 \le t\}$

EXERCISE 1.21 (H)

Suppose that $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a convex function and assume that $x^\star \in \mathbb{R}^n$ is locally optimal. That is, for all $x \in \mathbb{R}^n$ such that $\|x - x^\star\| \le \delta$, we get that $f(x^\star) \le f(x)$, for some $\delta > 0$. Show that $x^\star$ is a global minimum.

EXERCISE 1.22

Suppose that $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a strictly convex function.

1. Suppose that a point $x^\star$ exists such that $f(x^\star) \le f(x)$ for all $x \in \mathbb{R}^n$. Show that $x^\star$ is the unique minimizer of f.
2. Provide a strictly convex f whose minimum is not attained by any point $x^\star$.

For strongly convex functions (which are also strictly convex) the minimum always exists.

EXERCISE 1.23

Show for each of the following convex functions if it is smooth, strongly convex, strictly convex, or none of the above. Draw/plot the functions and decide from the drawings.

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

EXERCISE 1.24 (H)

Given some unknown function f where we know $f(1) = 1$, $f(-1) = 0$. For $x \in [-1, 1]$, draw the known bounds on f 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

Given some unknown differentiable function $f : \mathbb{R} \to \mathbb{R}$ where we know $f(1) = 1$, $f'(1) = 1$. 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) ($\star$)

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

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{1.1}$$

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

1. Provide a nonconvex differentiable function f and a point y for which (1.1) does not hold.
2. Prove the result.

Exercise 1.27 ($\star$)

The indicator function of a set C is defined as

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

Show the following.

1. Let $K \in \mathbb{R}^{m \times n}$ be a matrix, $b \in \mathbb{R}^m$ be a vector and define the convex set $C := \{x \in \mathbb{R}^n : Kx - b = 0\}$. Show that

   $$\iota_C(x) = \sup_{\mu} \mu^T (Kx - b)$$

   where $\mu \in \mathbb{R}^m$.

2. Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be a function and define the set $C := \{x : g(x) \leq 0\}$. Show that

$$\iota_C(x) = \sup_{\mu \geq 0} \mu^T g(x)$$

where $\mu \in \mathbb{R}^m$.

Exercise 1.28 ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex function. We get that

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

holds for for all integers $n \geq 2$, where $x_i \in \mathbb{R}^n$, $\theta_i \geq 0$, and $\sum_{i=1}^{n} \theta_i = 1$. This inequality is called *Jensen's inequality*. Prove Jensen's inequality.

Exercise 1.29 ($\star$)

Let $L(u, y)$ be convex in u for every fixed y.

1. Let $m(x; \theta) = \theta x$, where $x \in \mathbb{R}^m$ is fixed and $\theta \in \mathbb{R}^{n \times m}$. Is the function $L(m(x; \theta), y)$ convex in θ for all fixed x and y? Prove or provide counterexample.
2. Let $\theta = (\theta_1, \theta_2) \in \mathbb{R}^{n_1 \times m_1} \times \mathbb{R}^{n_2 \times m_2}$ and $m(x; \theta) = \theta_2 \sigma(\theta_1 x)$, where $\sigma : \mathbb{R}^{n_1} \to \mathbb{R}^{m_2}$ is differentiable and $x \in \mathbb{R}^{m_1}$ is fixed. Is $L(m(x; \theta), y)$ convex in θ for all fixed x and y and differentiable σ? Prove or provide counterexample.

Exercise 1.30 ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Show that $f(x) - \frac{\sigma}{2}\|x\|_2^2$ is convex (i.e., f is σ-strongly convex) if and only if

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

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

Exercise 1.31 ($\star\star$)

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 all $x, y \in \mathbb{R}^n$, 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) $\frac{\beta}{2}\|x\|_2^2 - f(x)$ and $f(x) + \frac{\beta}{2}\|x\|_2^2$ are convex

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

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

Show that these properties are equivalent.

Exercise 1.32 (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 all $x, y \in \mathbb{R}^n$, i.e., f is β-smooth

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

Hints

Hint to exercise 1.2

A matrix $Q \in \mathbb{R}^{n \times n}$ is positive semidefinite if and only if Q is symmetric ($Q = Q^T$) and $x^T Q x \geq 0$ for all $x \in \mathbb{R}^n$. This second condition is equivalent to that all eigenvalues are nonnegative.

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.

Hint to exercise 1.18

A function $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex if

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

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

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

See Exercises 1.30 and 1.31(IV) for the smoothness and strong-convexity bounds.

Hint to exercise 1.26

The directional derivative at $x \in \mathbb{R}^n$ in direction $d \in \mathbb{R}^n$ satisfies

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

(This is given by the chain-rule.)

Hint to exercise 1.32

Use Exercise 1.31 and the second order condition for convex functions.

Chapter 2

Subdifferentials and Proximal Operators

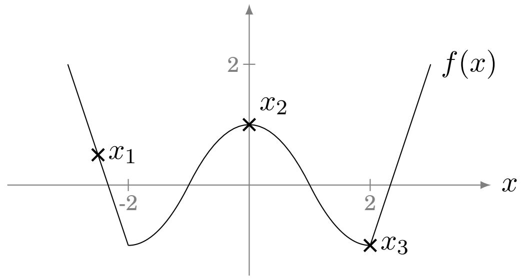
Exercise 2.1

Compute the subdifferentials for the following convex functions.

1. $f(x) = \frac{1}{2}\|x\|_2^2$
2. $f(x) = \frac{1}{2}x^T Hx + h^T x$ with H positive semidefinite
3. $f(x) = |x|$
4. $f(x) = \iota_{[-1,1]}(x)$
5. $f(x) = \max(0, 1 + x)$ (hinge loss)
6. $f(x) = \max(0, 1 - x)$

Exercise 2.2

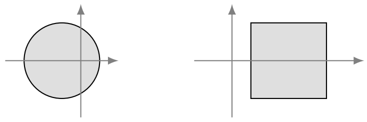
Consider the following even nonconvex function f.



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

Exercise 2.3

Figure (a) depicts $\partial f(x)$ and Figure (b) depicts $\partial g(y)$.



(a) (b)

1. Is x a minimum to f?
2. Is y a minimum to g?
3. Is f differentiable at x
4. Is g differentiable at y
5. Draw/explain examples of functions f and g that comply with the figure.

Exercise 2.4

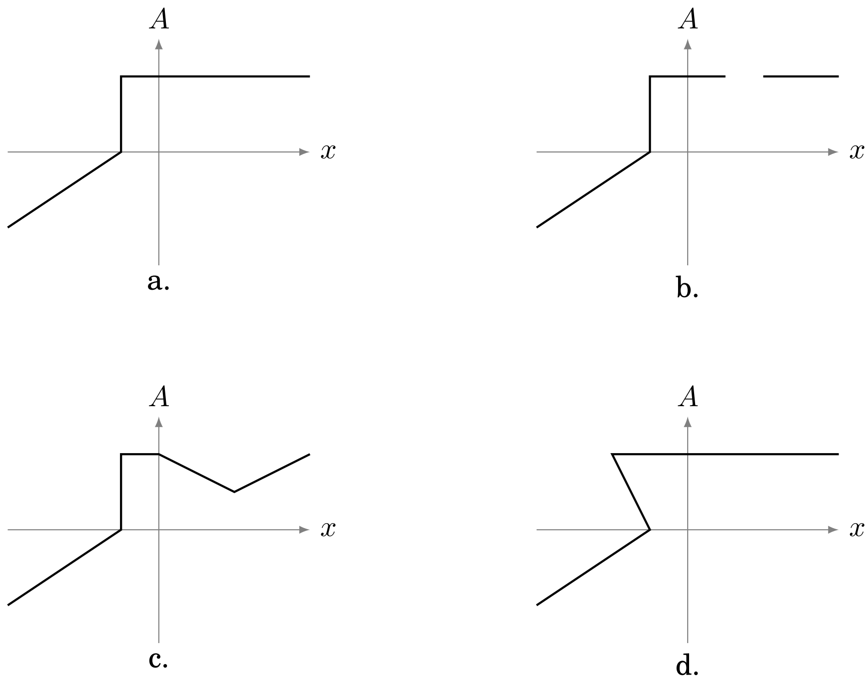
Suppose that $f : \mathbb{R} \to \mathbb{R}$ satisfies $f(-1) = 1$, $\partial f(-1) = \{-1\}$, $f(1) = 1$ and $\partial f(1) = \{1\}$.

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

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

- Which are monotone?
- Which can be subdifferentials of closed convex functions?



Exercise 2.6

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

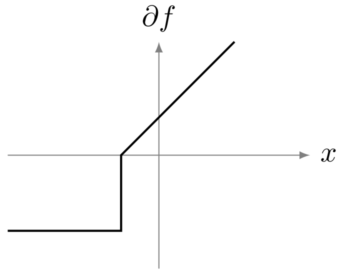
I) $f(y) \geq f(x) + \nabla f(x)^T(y - x)$ for all $x, y \in \mathbb{R}^n$, i.e. f is convex

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

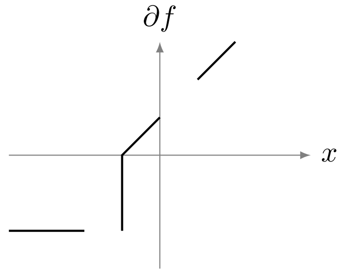
1. Show that I) implies II)
2. Show that II) implies I)

Exercise 2.9

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



a.



b.

1. Are the correspoding functions f convex?
2. Can you find the 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

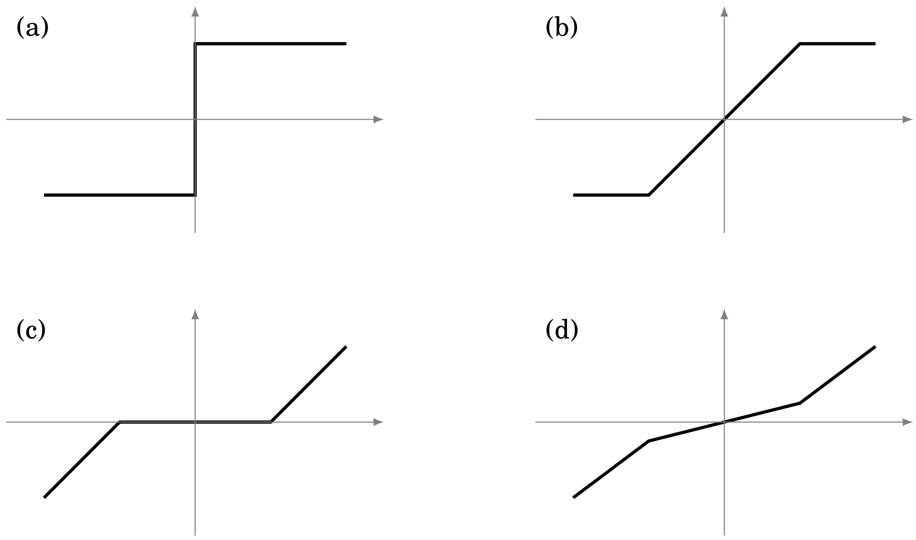
Assume that $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is σ-strongly convex. Show that

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

for all $x \in \mathrm{dom}\partial f = \{z \in \mathbb{R}^n : \partial f(z) \neq \emptyset\}$, $y \in \mathbb{R}^n$ and $s \in \partial f(x)$.

EXERCISE 2.11

The subdifferentials of four convex functions f are drawn below. State for each if f is differentiable, ∇f is Lipschitz continuous, f strongly convex. Also, estimate Lipschitz and strong convexity constants (given the axes are equal).



Exercise 2.12 ($\star$)

Suppose that $g(x) = \sum_{i=1}^{n} g_i(x_i)$, where $x = (x_1, \ldots, x_n)$. Show that $s \in \partial g(x)$ if and only if $s_i \in \partial g_i(x_i)$, where $s = (s_1, \ldots, s_n)$.

Exercise 2.13 ($\star$)

Assume that $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and that there exists $y \in \mathbb{R}^n$ with $f(y) < \infty$. Show that $\partial f(x)$ is empty for $x \notin \mathrm{dom} f$, i.e., for x such that $f(x) = \infty$.

Exercise 2.14 ($\star$)

Show that the subdifferential of the indicator function of a nonempty set C is the normal cone to C.

Exercise 2.15

Compute the proximal mapping for the following convex functions.

1. $f(x) = \frac{1}{2}\|x\|_2^2$
2. $f(x) = \frac{1}{2}x^T H x + h^T x$ with H positive semidefinite
3. $f(x) = |x|$
4. $f(x) = \iota_{[-1,1]}(x)$
5. $f(x) = \max(0, 1 + x)$
6. $f(x) = \max(0, 1 - x)$

Exercise 2.16

Suppose that $g(x) = \sum_{i=1}^{n} g_i(x_i)$, where $x = (x_1, \ldots, x_n)$. Show that

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

Hints

HINT TO EXERCISE 2.8

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

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

   by the chain-rule. This gives that

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

   for all $x, y \in \mathbb{R}^n$. 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 convex functions.

1. $f(x) = \frac{1}{2}\|x\|_2^2$
2. $f(x) = \frac{1}{2}x^T Hx + h^T x$ with $H \in \mathbb{R}^{n\times n}$ positive definite
3. $f(x) = \iota_{[-1,1]}(x)$
4. $f(x) = |x|$
5. $f(x) = \max(0, 1 + x)$
6. $f(x) = \max(0, 1 - x)$

Exercise 3.2

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

1. $f \le g$ implies that $f^* \ge g^*$
2. $f \le g$ implies that $f^{**} \le g^{**}$
3. $f = f^*$ 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

Assume that $g(x) = \sum_{i=1}^n g_i(x_i)$, i.e, g is separable. Show that $g^*(s) = \sum_{i=1}^n g_i^*(s_i)$, where g_i^* is the conjugate of g_i.

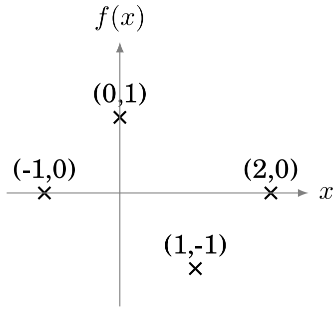
Exercise 3.6 (H)

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

1. $f(x) = \|x\|_1$
2. $f(x) = \iota_{[-\mathbf{1},\mathbf{1}]}(x)$, where $\mathbf{1} = (1, \ldots, 1)$

Exercise 3.7

Let f be the nonconvex function in the following figure. It satisfies $f(-1) = 0$, $f(0) = 1$, $f(1) = -1$, $f(2) = 0$, $f(x) = \infty$ for all $x \in \mathbb{R} \backslash \{-1, 0, 1, 2\}$.



1. Draw the conjugate f^* of f
2. Draw the bi-conjugate f^{**} of f

Exercise 3.8 (H) ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that $f(x) = \|x\|_2$.

1. Compute the conjugate f^* via the following steps
   (a) Show that $f^*(s) \geq 0$ for all s
   (b) Show that $f^*(s) \leq 0$ for all s with $\|s\|_2 \leq 1$
   (c) Show that $f^*(s) = \infty$ for all s with $\|s\|_2 > 1$
   (d) Combine there results to state $f^*(s)$
2. Use the conjugate to compute the subdifferential of f

Exercise 3.9 ($\star$)
Let Δ be the probability simplex

$$\Delta = \{x : x_i \geq 0 \text{ and } \sum_i x_i = 1\},$$

and let D be the similar set

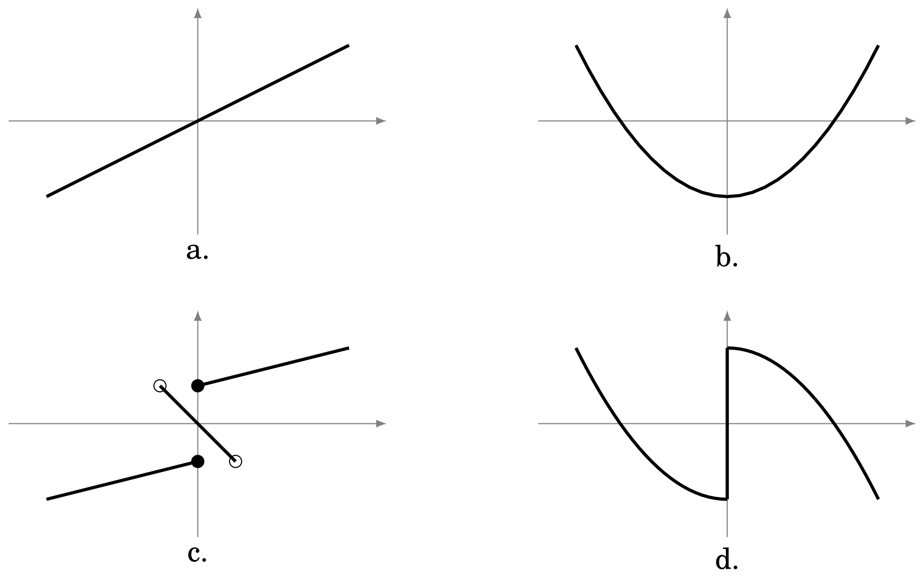
$$D = \{x : x_i \geq 0 \text{ and } \sum_i x_i \leq 1\}.$$

1. Let $f = \iota_\Delta$, where ι is the indicator function, show that $f^*(s) = \max_i(s_i)$, i.e., the element-wise max
2. Provide the conjugate of the $\max_i(s_i)$
3. Let $f = \iota_D$, where ι is the indicator function, show that $f^*(s) = \max(0, \max_i(s_i))$, where $\max_i(s_i)$ is the element-wise max
4. Provide the conjugate of $\max(0, \max_i(s_i))$

Exercise 3.10

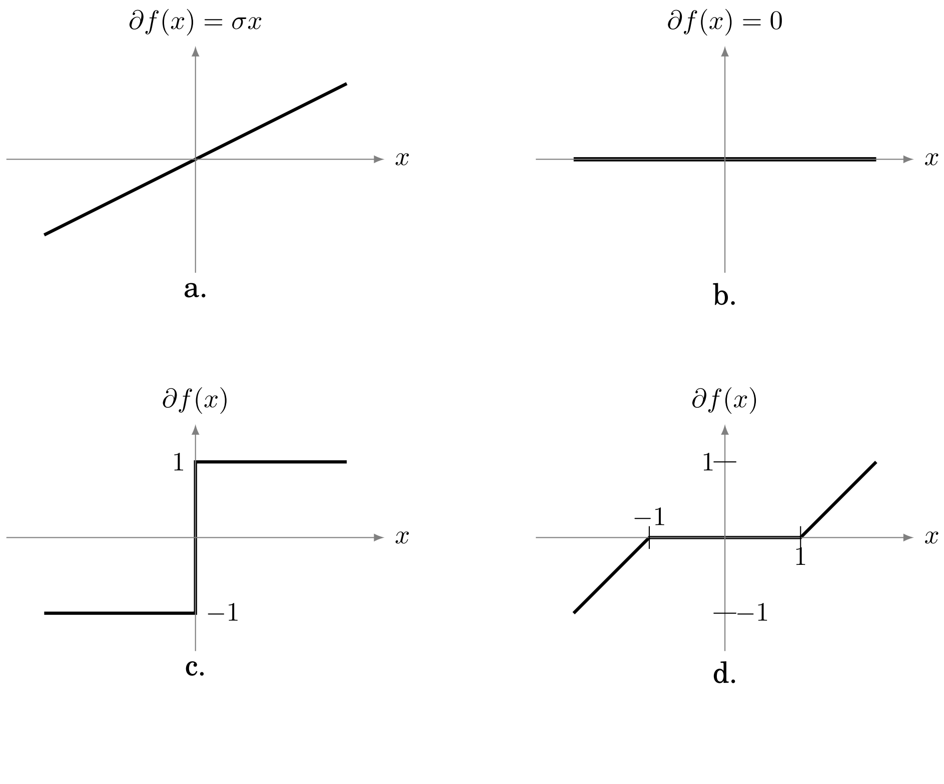
Consider the following set-valued operators $A : \mathbb{R} \to 2^{\mathbb{R}}$.

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 convex functions. Decide ∂f^*, i.e., the subdifferential of the conjugate.



Exercise 3.12

Assume that $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and $\gamma > 0$. Show that $\text{prox}_{\gamma f}(z) = (I + \gamma\partial f)^{-1}(z)$, where the inverse means the operator inverse.

Exercise 3.13 (H)

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

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

1. $\text{prox}_f(z) + \text{prox}_{f^*}(z) = z$
2. $(\gamma f)^*(s) = \gamma f^*(\gamma^{-1}s)$
3. $\text{prox}_{(\gamma f)^*}(z) = \gamma\text{prox}_{\gamma^{-1}f^*}(\gamma^{-1}z)$
4. $\text{prox}_{\gamma f}(z) + \gamma\text{prox}_{\gamma^{-1}f^*}(\gamma^{-1}z) = z$

Exercise 3.15

Compute the $\text{prox}_{(\gamma f)^*}$ for the following f.

1. $f(x) = \frac{1}{2}x^T Hx + h^T x$ with H positive definite
2. $f(x) = \max(0, 1 + x)$
3. $f(x) = \max(0, 1 - x)$

Exercise 3.16

Show the following.

1. That

$$\inf_x f(x) = -f^*(0)$$

2. That the set of minimizers, $\operatorname{Argmin}_x f(x)$, for a convex function f satisfies

$$\operatorname*{Argmin}_x f(x) = \partial f^*(0)$$

EXERCISE 3.17

Consider a primal problem of the form

$$\operatorname*{minimize}_x f(x) + g(x),$$

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

1. Show that this problem is equivalent to finding $x, y \in \mathbb{R}^n$ such that

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

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

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

that solves the dual problem

$$\operatorname{minimize} f^*(y) + g^*(-y)$$

3. Given a solution $y^\star$ to the dual condition (3.1) and a subgradient selector function, $s_{f^*}(y) : \mathbb{R}^n \to \mathbb{R}^n$ such that $s_{f^*}(y) \in \partial f^*(y)$. Can you recover a primal solution $x^\star$? What if f^* is differentiable?

EXERCISE 3.18 (H)

Consider primal problems of the form

$$\operatorname{minimize} f(Lx) + g(x),$$

where $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are (closed) convex functions, $L \in \mathbb{R}^{m \times n}$, and $\operatorname{relint}\operatorname{dom}(f \circ L) \cap \operatorname{relint}\operatorname{dom} g \neq \emptyset$. Derive the dual problem

$$\operatorname{minimize} f^*(y) + g^*(-L^T y).$$

EXERCISE 3.19

Consider primal problems of the form

$$\text{minimize } f(Lx) + g(x),$$

where $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, and $L \in \mathbb{R}^{m \times n}$. State the dual problem and show how to recover a primal solution from a dual solution for the following particular cases.

1. $f(y) = \frac{\lambda}{2}\|y\|_2^2$ where $\lambda > 0$ and $g(x) = \sum_{i=1}^{n} x_i + \iota_{[-1,0]}(x_i)$. Assume $m = n$ and L is invertible
2. $f(y) = \iota_{[-\mathbf{1},\mathbf{1}]}(y)$ and $g(x) = \frac{\lambda}{2}\|x\|_2^2 - b^T x$ where $\lambda > 0$

EXERCISE 3.20 ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Prove Fenchel-Young's equality,

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

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 $f^*(s) \leq s^T x - f(x)$. We may 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 $s \in \partial f(x)$

EXERCISE 3.21 ($\star$)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. 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 f closed and 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 $g(x) = f(Lx + c)$ where $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is closed convex, $L \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^m$ and $\operatorname{relint} \operatorname{dom} g \neq \emptyset$. Show that

$$g^*(s) = \inf_{\mu : s = L^T \mu} (f^*(\mu) - c^T \mu).$$

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

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

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 all x, s.

HINT TO EXERCISE 3.13

The subgradients for all functions have already been computed in previous exercises.

Hint to exercise 3.14

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

Hint to exercise 3.18

A very similar approach to Exercise 3.17 can be used.

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 descent algorithm

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

where $\gamma > 0$. Let $x^\star$ be a fixed point of this algorithm. Show that $x^\star$ minimizes f.

Exercise 4.2

Suppose that $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is closed convex and x is such that $x = \mathrm{prox}_{\gamma f}(x)$ for some $\gamma > 0$. Show that x minimizes f.

Exercise 4.3

Suppose that $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are closed convex, f is differentiable, and x is such that $x = \mathrm{prox}_{\gamma g}(x - \gamma \nabla f(x))$ for $\gamma > 0$. Show that x minimizes $f + g$.

Exercise 4.4

Which of the algorithms

- gradient method
- proximal gradient method

are applicable to the minimization problem $\text{minimize}_{x\in\mathbb{R}^n}\, h(x)$ where $h(x)$ is

1. $\frac{1}{2}\|Ax-b\|_2^2$, where $A \in \mathbb{R}^{m\times n}, m < n$
2. $\frac{1}{2}x^TQx + b^Tx + \|x\|_1$, where $Q \succ 0$
3. $\frac{1}{2}\|Ax-b\|_2^2 + \|x\|_2^2$, where $A \in \mathbb{R}^{m\times n}, m < n$
4. $\frac{1}{2}\|Ax-b\|_2^2 + \|x\|_2$, where $A \in \mathbb{R}^{m\times n}, m < n$
5. $\iota_{Ax=b}(x) + \iota_{[-1,1]}(x)$
6. $e^{\|x-y\|_2^4} + \iota_{[-1,1]}(x)$
7. $\frac{1}{2}x^TQx + \|Dx\|_1$, where $Q \succ 0, D$ diagonal
8. $\frac{1}{2}x^TQx + \iota_{[-1,1]}(Lx)$, where $Q \succ 0, L \in \mathbb{R}^{m\times n}$
9. $\log(1+e^{-w^Tx}) + \frac{1}{2}\sum_i \max(0,x_i)^2$

Exercise 4.5

For each of the algorithms and functions in Excercise 4.4, which of the algorithms are applicable to some dual formulation of each of the problems?

Exercise 4.6

Consider the problem

$$\underset{x}{\text{minimize}}\ \|x\|_1 + \tfrac{1}{2}x^TQx$$

where $Q \in \mathbb{R}^{n\times n}$ and $Q \succ 0$. The goal of this exercise is to state a dual problem and state the proximal gradient method for this dual problem. Let

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

The problem can be written as

$$\underset{x}{\text{minimize}}\ f(x) + g(x).$$

1. Compute f^*
2. Compute g^*
3. State a dual problem using general f^* and g^*
4. State a proximal gradient method step for this general dual problem. Specifically, assume that f is closed, convex and proximable, and g is closed and strongly convex (which in fact is true in our particular case). Construct a proximal gradient method 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

$$\underset{x}{\text{minimize}}\, f(Lx) + g(x),$$

where $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is closed convex and proximable, $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is 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

$$\underset{\mu}{\text{minimize}}\, f^*(\mu) + g^*(-L^T \mu).$$

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

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

Show that this equivalently can be written as

$$\begin{cases} x_k = \text{argmin}_x(g(x) + \mu_k^T Lx), \\ v_k = \mu_k + \gamma_k Lx_k, \\ \mu_{k+1} = v_k - \gamma_k \text{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} v_k). \end{cases} \tag{4.1}$$

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 these particular choices 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 4.6.

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 γ_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

$$x_{k+1} = \underset{y}{\text{argmin}}\, g(y)$$

for some function $g : \mathbb{R}^n \to \mathbb{R}$ such that $f \leq g$, 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.

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^n,$$

and use that $\gamma_k < 1/\beta$.

Chapter 5

Learning

EXERCISE 5.1

Consider the logistic regression problem

$$\min_{w,b} \sum_{i=1}^{N} \log(1 + e^{-y_i(x_i^T w + b)}),$$

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

$$\min_{w,b} \sum_{i=1}^{N} \left(\log(1 + e^{x_i^T w + b}) - y_i(x_i^T w + b) \right),$$

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

$$\min_{w,b} \sum_{i=1}^{N} \left(\log(1 + e^{x_i^T w + b}) - y_i(x_i^T w + b) \right),$$

with data points $x_i \in \mathbb{R}^n$ and class labels $y_i \in \{0, 1\}$, for each $i \in \{1, \ldots, N\}$. Assume that there exists $(\bar{w}, \bar{b}) \in \mathbb{R}^n \times \mathbb{R}$ such that $x_i^T \bar{w} + \bar{b} < 0$ if $y_i = 0$ and $x_i^T \bar{w} + \bar{b} > 0$ if $y_i = 1$, for each $i \in \{1, \ldots, n\}$. Show that the infimal value of the cost is 0, and that no (w, b) exists that attains the value.

EXERCISE 5.3

Consider the univariate Lasso problem

$$\min_{x \in \mathbb{R}} \tfrac{1}{2} \|ax - b\|_2^2 + \lambda |x|,$$

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 solution of this problem is simply $x = 0$). Prove that the optimal solution of this problem is

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

where

$$x_{\text{ls}} = \tfrac{a^T b}{\|a\|_2^2},$$

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

$$\min_{x \in \mathbb{R}^m} \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

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

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

Consider the following bivariate Lasso problem

$$\min_{x \in \mathbb{R}^2} \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1,$$

where $A \in \mathbb{R}^{n \times 2}$, $b \in \mathbb{R}^n$, $n \geq 2$, and $\lambda > 0$. Assume that the columns of A are normalized, $\|a_1\|_2 = \|a_2\|_2 = 1$, and A has full column rank (in particular, $|a_1^T a_2| < 1$). For each of the four possible sparsity patterns,

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

find the set $\Lambda_{i,j}$ s.t. if $x^\star \in X_{i,j}$ then $\lambda \in \Lambda_{i,j}$ where $x^\star$ is a solution to the problem. Verify that for a given problem the four ranges $\Lambda_{i,j}$ are disjoint and the number of zeros in the solution increase with λ.

Exercise 5.6

Consider the following SVM problem with an affine model

$$\min_{w \in \mathbb{R}^m, b \in \mathbb{R}} \sum_{i=1}^{n} \max\left(0, 1 - y_i(x_i^T w + b)\right) + \tfrac{\lambda}{2}\|w\|_2^2,$$

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

1. Consider the unregularized problem, $\lambda = 0$, where examples for 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 $x_i^T w + b < 0$ if $y_i = -1$ and $x_i^T w + b > 0$ if $y_i = 1$, for each $i \in \{1, \ldots, n\}$. Show the optimal cost is 0 and that the that the set of optimal (w, b) is unbounded.

2. Consider again the unregularized problem, $\lambda = 0$, but assume the data only consists of one class, e.g., there are no $i \in \{1, \ldots, n\}$ such that $y_i = -1$. Show that arbitrary w is optimal and the set of optimal (w, b) is unbounded.

3. Consider the regularized problem, $\lambda > 0$. Assume the data only consists of one class, e.g., there are no $i \in \{1, \ldots, n\}$ such that $y_i = -1$. Show that $w = 0$ is optimal and the set of optimal (w, b) is unbounded.

Exercise 5.7

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

$$\min_{w \in \mathbb{R}^m, b \in \mathbb{R}} \mathbf{1}^T \max\left(0, \mathbf{1} - (X^T w + b\phi)\right) + \tfrac{\lambda}{2}\|w\|_2^2,$$

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

Exercise 5.8

Consider the reformulated SVM problem in Exercise 5.7

$$\min_{w \in \mathbb{R}^m, b \in \mathbb{R}} \underbrace{\mathbf{1}^T \max(0, 1 - (X^T w + b\phi))}_{f(L(w,b))} + \underbrace{\tfrac{\lambda}{2}\|w\|_2^2}_{g(w,b)},$$

where $L = [X^T, \phi]$.

1. Derive the dual problem $\min_\mu f^*(\mu) + g^*(-L^T \mu)$.

2. Show how to recover a primal solution from a dual solution.

3. A support vector for this kind of *soft-margin* SVM is defined as any data point x of class y that lies on the wrong side of the margin, $1 \geq y(x^T w + b)$, for a model given by (w, b). It is easy to see that only the support vectors contribute to the cost of the objective (see objective in Exercise 5.6). Show that the non-zero elements of the optimal dual variable $\mu^\star$ correspond to support vectors of the optimal model $(w^\star, b^\star)$. Show that the optimal model can be recovered from the dual solution by only considering support vectors.

Exercise 5.9

Consider the typical supervised learning problem

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

where $x_i \in \mathbb{R}^d$ is data, $y_i \in \mathbb{R}^l$ the response variable, $m_w : \mathbb{R}^d \to \mathbb{R}^k$ the parameterized model we wish to train, and $L : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$ the loss comparing the model output $m_w(x_i)$ with the known correct output y_i.

Assume $L(\hat{y}, y)$ is convex in $\hat{y}$. Prove or disprove the following statements.

1. $\sum_{i=1}^{n} L(m_w(x_i), y_i)$ is convex if a feature mapped model is used. I.e. $m_w(x) = w^T \phi(x)$ where $\phi : \mathbb{R}^d \to \mathbb{R}^f$ for some $w \in \mathbb{R}^{f \times k}$.
2. $\sum_{i=1}^{n} L(m_w(x_i), y_i)$ is convex if a DNN model is used. I.e. $m_w(x) = \sigma_1(w_1^T \sigma_2(w_2^T ... \sigma_D(w_D^T x)...))$ where σ_i are some activation functions.

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 $\text{sgn}(x) = \text{sgn}(x)^{-1}$, $|x| = \text{sgn}(x)x$ and $\text{sgn}(x)\,\text{sgn}(y) = \text{sgn}(xy)$

Chapter 6

Algorithm Convergence

Exercise 6.1

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

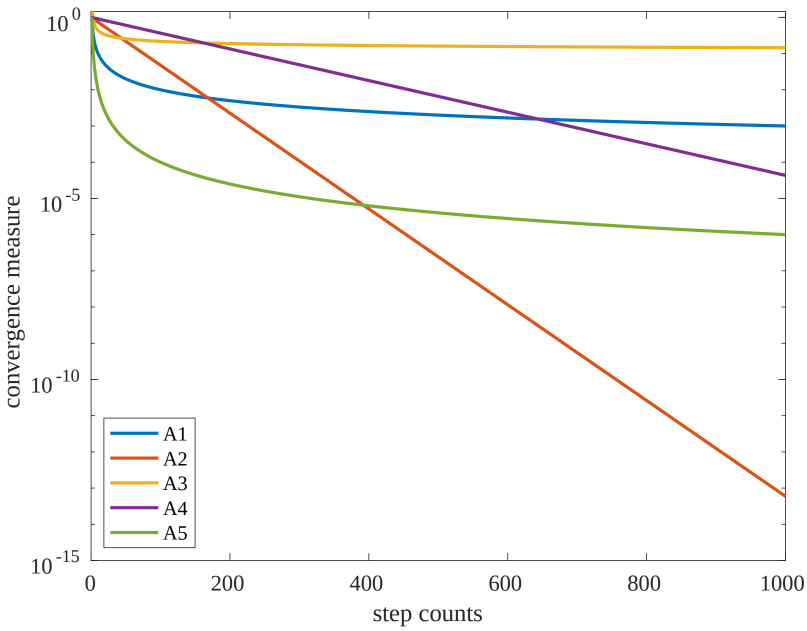
1. The first algorithm, performed 5000 floating point operations in each iteration and we ran it for 10^5 iterations
2. The second algorithm, performed 50 floating point operations in each iteration and we ran it for 2×10^6 iterations

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 $(V_k)_{k\in\natural}$ be a non-negative convergence measure. Show that

1. Q-linear rate

$$V_{k+1} \leq \rho V_k, \quad \forall k \in \mathbb{N},$$

for some $\rho \in [0,1)$, implies an R-linear rate

$$V_k \leq \rho^k C_L, \quad \forall k \in \mathbb{N},$$

for some $C_L \geq 0$ and find C_L

2. Q-quadratic rate

$$V_{k+1} \leq \rho V_k^2, \quad \forall k \in \mathbb{N},$$

for some $\rho \in [0,1)$, implies a local R-quadratic rate

$$V_k \leq \rho_Q^{2^k} C_Q, \quad \forall k \in \mathbb{N},$$

for some $\rho_Q \in [0,1)$ and $C_Q \geq 0$. We say that the rate is local since R-quadratic convergences depends on the initial value V_0. Find ρ_Q and C_Q and for which ρ and V_0 this R-quadratic rate imply that $V_k \to 0$ as $k \to \infty$.

EXERCISE 6.4

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

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

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

$$f(x_{k+1}) \leq f(x_k), \quad \forall k \in \mathbb{N}.$$

We call such a sequence $(x_k)_{k\in\mathbb{N}}$ a descent sequence (for f).

1. Give an example of a function f and descent sequence $(x_k)_{k\in\mathbb{N}}$ where this does not imply convergence of the sequence of function values $(f(x_k))_{k\in\mathbb{N}}$
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 $(f(x_k))_{k\in\mathbb{N}}$ converges
3. Give an example of a function f that is bounded from below and descent sequence $(x_k)_{k\in\mathbb{N}}$ such that $(f(x_k))_{k\in\mathbb{N}}$ does not converge to $\inf_{x\in\mathbb{R}^n} f(x)$

Exercise 6.5

Let $f(x) = e^x - 2x + x^2$ and $x \in \mathbb{R}$. Consider finding the minimizer of f using the standard Newton's method without line search

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

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 $|x_{k+1} - x^\star|/|x_k - x^\star|$ and $|x_{k+1} - x^\star|/|x_k - x^\star|^2$. Based on these ratios, estimate whether the sequence $\{|x_k - x^\star|\}_{k\in\mathbb{N}}$ is Q-linear or Q-quadratic convergent and find the corresponding rate parameter.

Exercise 6.6

A sequence $\{Q_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}$ is generated by some iterative algorithm and is found to satisfy the following inequality

$$0 \leq Q_k \leq \frac{V}{\psi_1(k)} + \frac{D}{\psi_2(k)}, \quad \forall k \in \mathbb{N},$$

where D and V are positive constants. The functions $\psi_1 : \mathbb{R} \to \mathbb{R}$ and $\psi_2 : \mathbb{R} \to \mathbb{R}$ are parameters of the algorithm that generated $\{Q_k\}_{k\in\mathbb{N}}$.

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

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

   (b) When

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

   and

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

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

   (c) When

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

   and

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

   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 $\{x_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^n$. A convergence analysis results in the following inequality

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

where V, D and b are positive constants, $x^\star$ is a minimizer of f and $\gamma_i > 0$, for each $i \in \mathbb{N}$, are step-size parameters.

1. Show $f(x_k)$ converges to $f(x^\star)$ if $\{\gamma_i^2\}_{i\in\mathbb{N}}$ is summable and $\{\gamma_i\}_{i\in\mathbb{N}}$ is not, i.e., $\sum_{i=0}^{\infty} \gamma_i^2 < \infty$ and $\sum_{i=0}^{\infty} \gamma_i = \infty$

2. Let $c > 0$ and estimate the convergence rates for the following step-sizes
   (a) $\gamma_i = c/(i+1)$
   (b) $\gamma_i = c/(i+1)^\alpha$ with $\alpha \in (0.5, 1)$
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 $\beta > 0$ and $x^\star \in \mathbb{R}^n$ a minimizer of f. Consider finding a minimizer of f, not necessarily $x^\star$, using the gradient descent method

$$x_{k+1} = x_k - \gamma \nabla f(x_k), \quad \forall k \in \mathbb{N},$$

where $x_0 \in \mathbb{R}^n$ is given 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

$$f(x_{k+1}) \leq f(x_k), \quad \forall k \in \mathbb{N},$$

i.e. $(x_k)_{k\in\mathbb{N}}$ is a descent sequence for f. Moreover, the following Lyapunov inequality

$$\|x_k - x^\star\|_2^2 \leq \|x_{k-1} - x^\star\|_2^2 - 2\gamma(f(x_k) - f(x^\star)), \quad \forall k \in \mathbb{N} \setminus \{0\},$$

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 minimizer $x^\star \in \mathbb{R}^n$, using a stochastic optimization algorithm, starting at some predetermined (deterministic) point $x_0 \in \mathbb{R}^n$. Analysis of the algorithm resulted in the following inequality

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

where G is a positive constant and the γ_k's are positive step-sizes of the algorithm satisfying $\sum_{k=0}^\infty \gamma_k = \infty$ and $\sum_{k=0}^\infty \gamma_k^2 < \infty$. In particular, $(x_k)_{k\in\mathbb{N}}$ 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

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

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.

$$\begin{cases} y_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k), \\ x_{k+1} = (1-\gamma_k)y_{k+1} + \gamma_k y_k, \end{cases} \qquad \forall k \in \mathbb{N},$$

for some initial points $x_0 = y_0 \in \mathbb{R}^n$, where

$$\gamma_k = \frac{1-\lambda_k}{\lambda_{k+1}}, \quad \forall k \in \mathbb{N} \quad \text{and} \quad \lambda_k = \begin{cases} 1 & \text{if } k = 0, \\ 0.5 + \sqrt{0.25 + \lambda_{k-1}^2} & \text{otherwise.} \end{cases}$$

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

$$V_{k+1} - V_k \leq \tfrac{2\lambda_k^2}{\beta}(f(x_k) - f(x^\star)) - \tfrac{2\lambda_{k+1}^2}{\beta}(f(x_{k+1}) - f(x^\star)), \quad \forall k \in \mathbb{N} \setminus \{0\}, \tag{6.1}$$

where $V_k = \|(\lambda_k - 1)(x_{k-1} - x_k) - x_k + x^\star\|^2$ for each $k \in \mathbb{N} \setminus \{0\}$.

1. Show that $f(x_k) \to f(x^\star)$ as $k \to \infty$ and find the rate of convergence
2. 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 $C = 2\beta V_1 + 4(f(x_1) - f(x^\star))$, 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 $\lambda_k \geq 1 + 0.5k$ for each $k \in \mathbb{N}$.

Chapter 7

Proximal Gradient Based Algorithms

Exercise 7.1

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a β-smooth function. Consider the gradient descent algorithm with a fixed step-size $\gamma > 0$ applied to this problem,

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

where $x_0 \in \mathbb{R}^n$ is given.

1. Find the Lyapunov inequality

$$V_{k+1} \leq V_k - \gamma(1 - \tfrac{\beta}{2}\gamma)\|\nabla f(x_k)\|_2^2$$

   where $V_k = f(x_k) - f(x^\star)$ and $x^\star$ is a minimizer of f, assuming it exists.

2. Show that $\|\nabla f(x_k)\|_2^2 \to 0$ if $0 < \gamma < \frac{2}{\beta}$. Find the convergence rate of $g_k = \min_{i \in \{0,\ldots,k\}} \|\nabla f(x_i)\|_2^2$.

Exercise 7.2

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a β-smooth convex function. Consider minimizing $f(x)$ using the gradient descent algorithm and with fixed step-size $\gamma \in (0, \frac{\beta}{2})$,

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

where $x_0 \in \mathbb{R}^n$ is given.

1. With $x^\star$ being a minimizer of f, show that the iterates satisfy

$$\begin{aligned}&\|x_{k+1} - x^\star\|_2^2 \\ &\quad \leq \|x_k - x^\star\|_2^2 - 2\gamma(f(x_{k+1}) - f(x^\star)) + \gamma^2(\beta\gamma - 1)\|\nabla f(x_k)\|_2^2\end{aligned}$$

   by, in order, expand the square $\|x_{k+1} - x^\star\|_2^2$, use the first order condition for convexity, and use the Lyapunov inequality from Exercise 7.1.

2. Show that $f(x_k) \to f(x^\star)$ as $k \to \infty$, by summing all inequalities for $i = 0, \ldots, k$ and using the results in Exercise 7.1 to bound $\sum_{i=0}^{k} \|\nabla f(x_k)\|_2^2$. What is the convergence rate?

Exercise 7.3

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth and μ-strongly convex. Consider minimizing $f(x)$ using the gradient descent algorithm and with fixed step-size $\gamma > 0$,

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

where $x_0 \in \mathbb{R}^n$ is given.

1. Show that the iterates satisfy the inequality

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

for $\gamma \in (0, 1/\beta]$ by using the same technique as in Exercise 7.2 but replacing the first order condition for convexity with the first order condition for strong convexity. Which step-size γ maximize the convergence rate?

2. A different approach is used to analyze the convergence in the lectures. There they get the following convergence inequality,

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

which hold for $\gamma \in (0, 2/\beta)$. What is the best step-size γ according to this inequality? Which approach gives the fastest possible rate?

Exercise 7.4

Consider the problem

$$\underset{x}{\text{minimize}}\, x^T Q x + q^T x, \quad \text{where } Q \succ 0$$

with the gradient descent algorithm $x^{k+1} = x^k - \gamma \nabla f(x^k)$ where $\gamma \in (0, 2/L)$ and $L = \|Q\|$.

1. Show that $\|x^{k+1} - x^*\| \leq \|(I - \gamma Q)\|\|x^k - x^*\|$ and that $\|I - \gamma Q\| < 1$, where x^* is the solution to the problem.

2. Let $\gamma = 1/L$ and find an expression of $\|(I - \gamma Q)\|$ in terms of the eigenvalues of Q.

Let the (geometric) convergence rate r be defined as the smallest r so that $\|x^k - x^*\| \leq r^k \|x^0 - x^*\|$ holds.

3. Let $Q = \begin{bmatrix} \epsilon & 0 \\ 0 & 1 \end{bmatrix}$ where $0 < \epsilon \ll 1$. What is the worst case convergence rate r we can expect given the result above? Let $q = 0$, can you find a point x^0 where this is the practical rate.

4. Let $Q = \begin{bmatrix} \epsilon & \epsilon/10 \\ \epsilon/10 & 1 \end{bmatrix}$. The eigenvalues of this matrix is approximately 1 and ϵ. Gradient descent will therefore be slow also on this problem. To improve the convergence rate, we want to find a variable change $x = Vy$, where V is invertible, so that the equivalent problem $\text{minimize}_y\, y^T V^T Q V y + q^T V y$ 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 (roughly) the eigenvalues of the new matrix V^TQV? 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

Consider the proximal point algorithm, i.e. select $x_0 \in \mathbb{R}^n$ and for all $k \in \mathbb{N}$

$$x^{k+1} = \mathbf{prox}_{\gamma f}(x^k)$$

where $\gamma > 0$.

1. Show that $(f(x^k))_{k\in\mathbb{N}}$ is a decreasing sequence according to
$$f(x^{k+1}) \leq f(x^k) - \tfrac{1}{2\gamma}\|x^{k+1} - x^k\|_2^2.$$
for all $\gamma > 0$.
2. Assume that f is lower bounded by B (i.e., B is such that $f(x) \geq B$ for all x). Show that $\|x^{k+1} - x^k\| \to 0$ as $k \to \infty$.
3. Assume (closed) convexity of f. Show that $\|x^{k+1} - x^k\| \to 0$ implies that $\mathrm{dist}_{\partial f(x^k)}(0) \to 0$ where $\mathrm{dist}_{\partial f(x)}(0) = \inf_{s\in\partial f(x)} \|s - 0\|$, i.e. the distance between the subdifferential and zero becomes arbitrary small.
4. Assume strong convexity of f, show that $x^k \to x^\star$ where $f(x^\star) = \min f(x)$.

Note about the last point. There exist weaker conditions than strong convexity for the sequence to converge but strong convexity is arguably the simplest.

Exercise 7.6

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and β-smooth and $g : \mathbb{R}^n \to \mathbb{R}\cup\infty$ be closed, convex and proper. We consider minimizing $f(x) + g(x)$ using proximal gradient algorithm given a $x_0 \in \mathbb{R}^n$, with fixed step-size $\gamma \in (0, 1/\beta]$

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

A procedure of proving convergence of the algorithm 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 an inequality on the form
$$V_{k+1} \leq V_k - Q_k$$
where Q_k is some non-negative convergence measure. Here we choose $V_k = \|x_k - x^\star\|_2^2$ as the Lyapunov function. We further define the residual mapping as $\mathcal{R}x_k = x_k - \mathbf{prox}_{\gamma g}(x_k - \gamma\nabla f(x_k))$. The proximal gradient update can then be written as
$$x_{k+1} = x_k - \mathcal{R}x_k \tag{7.1}$$

and we can use this to relate V_{k+1} to V_k by expanding the square,

$$V_{k+1} = V_k + \dots \tag{7.2}$$

2. The quantity $-2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2$ now needs to be bounded. We start by using (7.1) to re-write it as

$$-2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2 = -2(x_{k+1} - x^\star)^T(\mathcal{R}x_k) + \dots \tag{7.3}$$

3. We now turn to bounding $-2(x_{k+1}-x^\star)^T(\mathcal{R}x_k)$. From the proximal gradient update we have

$$\begin{aligned} x_k - \gamma\nabla f(x_k) - x_k + \mathcal{R}x_k &\in \gamma\partial g(x_{k+1}) \\ \implies \gamma^{-1}\mathcal{R}x_k - \nabla f(x_k) &\in \partial g(x_{k+1}). \end{aligned}$$

   The definition of a subgradient then gives

$$-2(x_{k+1} - x^\star)^T(\mathcal{R}x_k) \le \dots \tag{7.4}$$

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

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

   Adding these two together and rearranging yield

$$-2\gamma\nabla f(x_k)^T(x_{k+1} - x^\star) \le \dots \tag{7.5}$$

5. Inserting (7.5) into (7.4), (7.4) into (7.3), and finally (7.3) into (7.2) yield

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

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

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

   where

$$Q_k = \dots$$

   which is non-negative since $\gamma > 0$ and $\dots \ge \dots$ since $x^\star$ is a minimum.

7. Since $V_k \ge 0$ and $Q_k \ge 0$ we have that $Q_k \to 0$ which implies that

$$\dots \to \dots$$

   as $k \to \infty$.

Exercise 7.7

Consider the problem

$$\min_{x\in\mathbb{R}^n} f(x) + g(x) \tag{7.6}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \infty$ are closed functions that are μ_f-, and μ_g-strongly convex respectively. We further assume f is β-smooth and g is proximable. The problem can then be solved with the proximal gradient method,

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

where $\gamma > 0$ and $x_0 \in \mathbb{R}^n$ are given.

1. Show that the proximal gradient method satisfy

   $$\|x_{k+1} - x^\star\|_2^2 \leq \tfrac{\max(1-\mu_f\gamma,\beta\gamma-1)^2}{1+\mu_g\gamma}\|x_k - x^\star\|_2^2$$

   by inserting the definition of x_{k+1} in $\|x_{k+1} = x^\star\|_2^2$ and then use the following in order. A minimum $x^\star$ is a fixed point to the proximal gradient step. The proximal operator of a μ-strongly convex function is $\frac{1}{1+\mu\gamma}$-Lipschitz continuous. The gradient of a β-smooth and μ-strongly convex function h satisfies

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

   for all x, y. Then, in two different cases, use β-Lipschitz continuity of ∇f or the fact that the gradient of a μ-smooth function h satisfies

   $$\|\nabla h(x) - \nabla h(y)\|_2 \geq \mu\|x - y\|_2$$

   for all x, y.

2. For which step-sizes γ and combinations of $\mu_f \geq 0$ and $\mu_g \geq 0$ does the algoritm converge linearly?

3. It is sometimes possible to move the strong convexity between f and g. For instance, consider the following problem,

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

   where h is L-smooth closed convex and ϕ is closed convex and proximable. This can be written as a problem of the form (7.6) by choosing any $\delta \in [0, 1]$ and forming

   $$f(x) = h(x) + \delta\tfrac{\mu}{2}\|x\|_2^2, \quad g(x) = \phi(x) + (1 - \delta)\tfrac{\mu}{2}\|x\|_2^2.$$

   The objective $f + g$ will always be the same and will always be μ-strongly convex regardless of the choice of δ. However, the individual strong convexity of f and g will depend on δ and hence so will the convergence rate of the proximal gradient method [1].

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

[1] We assume the function g is proximable here. In general is it not easy to prox on the sum of two functions, even if both are proximable. However, since we add a square norm term which and the proximal operator already solves a square norm regularized problem, this is not too strong of an assumption.

Exercise 7.8 (H)

Consider the problem of minimizing the function $F(x)$ where

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x).$$

The stochastic gradient method is

$$\begin{aligned}&\text{Sample } i \text{ uniformly from } \{1, ..., N\}\\ &x^{k+1} = x^k - \gamma^k \nabla f_i(x^k).\end{aligned}$$

Note that $\mathbb{E}[\nabla f_i(x)|x] = \nabla F(x)$ given that x is known. Further assume that the variance is bounded, $\mathbb{E}[\|\nabla f_i(x) - \nabla F(x)\|^2|x] \leq \sigma^2$ for all x, and that F is lower bounded and L-smooth.

1. Show that stochastic gradient descent satisfies

$$\mathbb{E}[F(x^{k+1})|x^k] \leq F(x^k) - \gamma^k(1 - \tfrac{L}{2}\gamma^k)\|\nabla F(x^k)\|^2 + (\gamma^k)^2\tfrac{L\sigma^2}{2}.$$

2. Show that it is possible for $\mathbb{E}\|\nabla F(x^k)\| \to \sigma$ if $\gamma^k = \frac{1}{L}$.
3. Show that $\min_{k \leq T} \mathbb{E}\|\nabla F(x^k)\| \to 0$ as $T \to \infty$ if $\gamma^k = \frac{1}{k}$.
4. Show that it is possible for $\mathbb{E}\|\nabla F(x^k)\| \to c > 0$ if $\gamma^k = \frac{1}{k^2}$ for some constant c.

Exercise 7.9

For the two problems below, estimate the per-iteration computational complexity for gradient descent, proximal gradient descent and their coordinate-wise variants. For the proximal algorithms, apply the prox to the term for which it is cheapest to compute.

- $\frac{1}{2}x^TQx + b^Tx + \|x\|_2^2$, where $Q \succ 0$.
- $\log(1 + e^{-w^Tx}) + \sum_i \max(0, x_i)^2$

Exercise 7.10 ($\star$)

Show that it is possible to implement coordinate gradient (and coordinate proximal gradient) for the function $\log(1+e^{-w^Tx})+\sum_i \max(0, x_i)^2$ with a per-iteration cost that doesn't grow with the number of elements in x.

Exercise 7.11

One interpretation of coordinate descent is that you restrict the function to a line and take a gradient step of the function along this line. Let the direction we want to take a gradient step along be coordinate i, i.e. the direction e_i, where index i of e_i is 1 and the others are 0. Let $f_{i,x}(\alpha) := f(x + e_i\alpha)$, we can then formulate the problem as taking a gradient step of $f_{i,x}$ from $\alpha_0 = 0$, i.e

$$\bar{\alpha} = \alpha_0 - \gamma_i \nabla f_{i,x}(\alpha_0)$$

If $f_{i,x}$ is L_i-smooth, then we know that $f_{i,x}(\bar{\alpha}) \leq f_{i,x}(\alpha_0)$ as long as $\gamma_i \in (0, 2/L_i)$. With $\alpha_0 = 0$ we therefore get a non-increasing sequence

$$f(x^{k+1}) = f(x^k + e_i\bar{\alpha}) = f_{i,x^k}(\bar{\alpha}) \leq f_{i,x^k}(\alpha_0) = f(x^k)$$

when $x^{k+1} = x^k + e_i\bar{\alpha}$.

- Consider the function $f(x) = \frac{1}{2}\|Ax - b\|^2$. Find the smoothness constants L_i, i.e the bounds on γ_i.
- Show that $L_i \leq L$ for all i, where L is the smoothness constant for f, i.e, we are able to take longer steps with the coordinate gradient algorithm than with regular gradient descent.

Exercise 7.12

Consider the minimization problem

$$\min_{x \in \mathbb{R}^n} f(x) + \sum\nolimits_{i=1}^{n} g_i(x_i)$$

and the proximal coordinate descent algorithm

$$\begin{aligned}
&\text{Choose } i \text{ from } \{1, ..., n\} \\
&x_i^{k+1} = \mathbf{prox}_{\gamma g_i}(x_i^k - \gamma \nabla_i f(x^k)) \\
&x_j^{k+1} = x_j^k \quad \forall j \neq i
\end{aligned}$$

where $\nabla_i f(x)$ is the i:th coordinate of the gradient and $\gamma > 0$. Assume L-smoothness of f, (closed) convexity of g_i and that each $i \in \{1, ..., n\}$ is chosen an infinite number of times.

Show that this is an descent method for sufficiently small γ. Find the upper bound on γ.

Exercise 7.13

Consider the problem and the proximal coordinate descent algorithm from Exercise 7.12 but allow for coordinate-wise step-sizes,

$$\begin{aligned}
&x_i^{k+1} = \mathbf{prox}_{\gamma_i g_i}(x_i^k - \gamma_i \nabla_i f(x^k)) \\
&x_j^{k+1} = x_j^k \quad \forall j \neq i
\end{aligned}$$

where $\gamma_i > 0$. Find better upper bounds for each γ_i that still ensures descent under the following refined smoothness assumption on f.

For all x, y is

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \tfrac{1}{2}(y-x)^T M (y-x)$$

satisfied for some positive definite M. Since $\frac{1}{2}(y-x)^T M(y-x) \leq \frac{\lambda_{\max}(M)}{2}\|y-x\|^2$ this implies regular smoothness. Ordinary L-smoothness can also be written on this form with $M = LI$ where I is the identity matrix. However, allowing for arbitrary quadratic upper bound on f means it can be made tighter.

Exercise 7.14 (*)

In this exercise we want to study the convergence of gradient descent and coordinate gradient descent. Consider the simple problem

$$p^* = \min_x \frac{1}{2}\|Ax - b\|^2,$$

where we assume that $A \in \mathbb{R}^{m \times n}$. Let $m = 40, n = 20$ and generate a random matrix A and random vector b. The optimal point x^* can in this case can be found directly using the least squares solution in Julia: `xsol=A\b`.

- Implement gradient descent, and plot the cost $\frac{1}{2}\|Ax^k - b\|^2 - p^*$ as a function of the iteration k. Note that you can compute and save the matrix $A^T A$ and the vector $A^T b$ to reduce the number of computations needed at each iteration.
- Implement coordinate gradient descent and compare the cost $\frac{1}{2}\|Ax^k - b\|^2 - p^*$ to that of the full gradient descent. Take make the comparison fair, use the same initial point x^0 and same step-length, let the number of iterations be n times as many and plot the cost for every n iterations.
- Implement coordinate gradient descent with the step-lengths computed in Excercise 7.11 and make the same comparison.

Exercise 7.15 (*)

In this exercise we want to study the convergence of stochastic gradient descent. Consider the same problem as in Exercise 7.14, where we can write the cost as $\frac{1}{2}\|Ax^k - b\|^2 = \frac{1}{2}\sum_i (A_i x^k - b_i)^2$, where A_i is row i in A. Implement the stochastic gradient algorithm for this problem. Note that you may need significantly more iterations with this algorithm compared to Exercise 7.14.

- Run the algorithm with a few different constant step sizes γ, for example $\lambda_{\max}, \lambda_{\max}/10, \lambda_{\max}/100$, where λ_{max} is the largest eigenvalue of $A^T A$. What happens with the error $\frac{1}{2}\|Ax^k - b\|^2 - p^*$ after many iterations?
- Run the algorithm with a decreasing step size, for example γ/k or $10\gamma/k$. How does the behavior differ?
- What happens if we let gamma decrease faster, e.g. $10\gamma/k^2$?

Hints

Hint to exercise 7.8

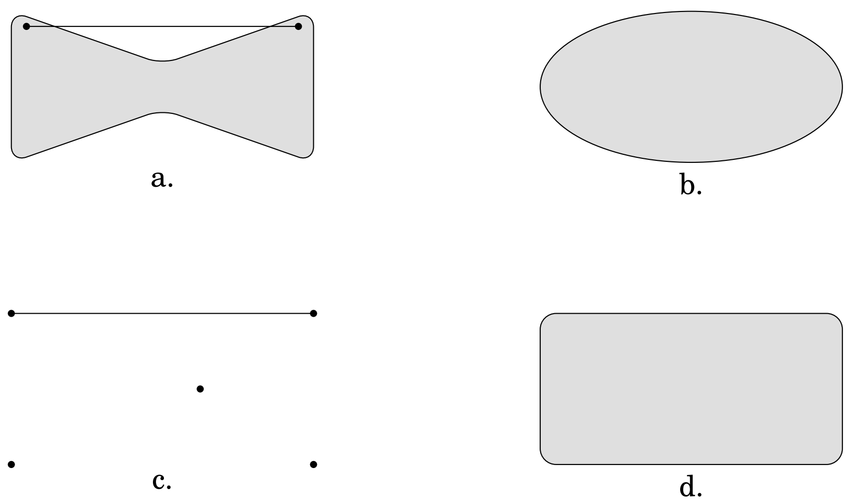
$$\mathbb{E}\|X - \mathbb{E}X\|^2 = \mathbb{E}\|X\|^2 - \|\mathbb{E}X\|^2$$

Solutions to Chapter 1

Solution 1.1

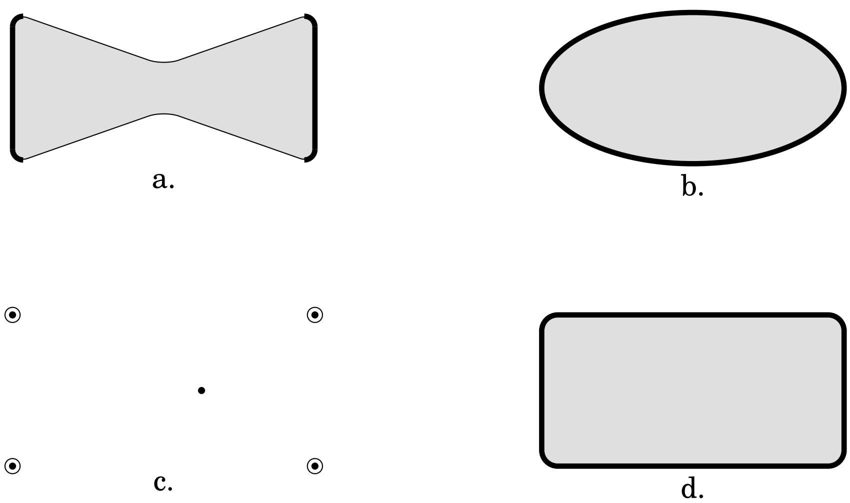
1. Figures b. and d. represent convex sets since the straight line connecting any two points with the sets are contained within the sets.

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



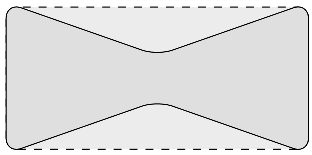
a. b. c. d.

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

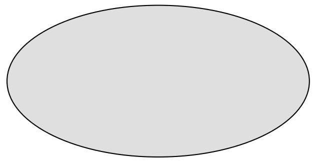


a. b. c. d.

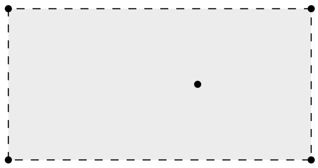
3. Figures b. and d. are convex so the convex hull is the set itself.



a.



b.



c.

d.

Solution 1.2

1. Take $x \in S$, $y \in S$, $\theta \in [0,1]$, and let $z = \theta x + (1-\theta)y$. Then $Ax = b$, $Ay = b$, and

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

Hence $z \in S$ and the set is convex. (This is an affine subspace/intersection of hyperplanes.)

2. Take $x \in S$, $y \in S$, $\theta \in [0,1]$, and let $z = \theta x + (1-\theta)y$. Then $Ax \leq b$, $Ay \leq b$, and

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

Hence $z \in S$ and the set is convex. (This is a polytope /intersection of halfspaces.)

3. Take $x \in S$, $y \in S$, $\theta \in [0,1]$, and let $z = \theta x + (1-\theta)y$. Then $x \geq 0$, $y \geq 0$, and

$$z = \theta x + (1-\theta)y \geq 0.$$

Hence $z \in S$ and the set is convex. (This is the non-negative orthant.)

4. Take $x \in S$, $y \in S$, $\theta \in [0,1]$, and let $z = \theta x + (1-\theta)y$. Then, since θ and $(1-\theta)$ are positive,

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

and

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

Hence $x \in S$ and the set is convex. (The constraints that defines the set are called box-constraints.)

5. Take $x \in S$, $y \in S$, $\theta \in [0,1]$, and let $z = \theta x + (1-\theta)y$. Then $\|x\|_2 \leq 1$, $\|y\|_2 \leq 1$, and

$$\|z\|_2 = \|\theta x + (1-\theta)y\|_2 \leq \theta\|x\|_2 + (1-\theta)\|y\|_2 \leq 1.$$

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

6. Consider $n = 1$, i.e., $x \in \mathbb{R}$. Let $x = -1$, $y = 1$, and $z = \frac{1}{2}(x+y) = 0$. Then $-\|x\|_2 = -1$ and $x \in S$. Similarly $-\|y\|_2 = -1$ and $y \in S$. However, $-\|z\|_2 = 0$ and $z \notin S$. Hence the set is not convex.

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

8. Take $(x, t_x) \in S$, $(y, t_y) \in S$, $\theta \in [0,1]$, and let $(z, t_z) = \theta(x, t_x) + (1-\theta)(y, t_y)$. Then $\|x\|_2 \leq t_x$, $\|y\|_2 \leq t_y$, and

$$\|z\|_2 = \|\theta x + (1-\theta)y\|_2 \leq \theta\|x\|_2 + (1-\theta)\|y\|_2 \leq \theta t_x + (1-\theta)t_y = t_z.$$

Hence $z \in S$ and the set is convex. (This set is called a second order cone and is shaped like an ice cream cone.)

9. Take $X \in S$, $Y \in S$, $\theta \in [0,1]$, and let $Z = \theta X + (1-\theta)Y$. Then $x^T X x \geq 0$ and $x^T Y x \geq 0$ for all $x \in \mathbb{R}^n$, and for arbitrary $x \in \mathbb{R}^n$:

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

In addition, Z is symmetric since X and Y are. Hence $z \in S$ and the set is convex.

10. Take $x \in S$, $y \in S$, $\theta \in [0,1]$, and let $z = \theta x + (1-\theta)y$. Then $x = a$, $y = a$, and

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

Hence $z \in S$ and the set is convex.

11. Consider $n = 1$, i.e., $x \in \mathbb{R}$. Let $x = a := -1$, $y = b := 1$, and $z = \frac{1}{2}(x+y) = 0$. Then $z \neq a$ and $z \neq b$, hence $z \notin S$ and the set is not convex.

Solution 1.3

1. Affine. Let $x \in V$ and $y \in V$. Then $x = y = a$ and $\alpha x + (1-\alpha)y = a \in V$ for all $\alpha \in \mathbb{R}$ and $x, y \in V$. Hence the set is affine.

2. Not affine. Affine means that $\beta x + (1-\beta)y \in V$ for all choices of $\beta \in \mathbb{R}$ and $x, y \in V$. But, for instance the choice of $x = a$, $y = b$ and $\beta = 2$ is quite obvious not in V.

   For a numerical example we can take $n = 1$, $a = -1$, $b = 1$. Then $V = [-1, 1]$ while $\beta x + (1-\beta)y =, 2 * (-1) + (-1) * 1 = -3 \notin V$.

3. Affine. Take $x, y \in V$. This means $\exists \beta_1, \beta_2 \in \mathbb{R}$ such that

$$
\begin{aligned}
x &= \beta_1 a + (1-\beta_1)b \\
y &= \beta_2 a + (1-\beta_2)b
\end{aligned}
$$

   Then for all $\alpha \in \mathbb{R}$,

$$
\begin{aligned}
&\alpha x + (1-\alpha)y \\
&\qquad = (\alpha\beta_1 + (1-\alpha)\beta_2)a + (\alpha(1-\beta_1) + (1-\alpha)(1-\beta_2))b \\
&\qquad = (\alpha\beta_1 + (1-\alpha)\beta_2)a + (1-(\alpha\beta_1 + (1-\alpha)\beta_2))b \\
&\qquad = \sigma a + (1-\sigma)b
\end{aligned}
$$

   where $\sigma = \alpha\beta_1 + (1-\alpha)\beta_2 \in \mathbb{R}$. Hence $\alpha x + (1-\alpha)y \in V$ and the set is affine.

SOLUTION 1.4

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

SOLUTION 1.5

All sets are in Exercise 1.2 shown to be convex. It is left to decide which sets that are cones.

1. Let $x \in S$, i.e., $Ax = 0$. Then $A(\alpha x) = \alpha Ax = 0$ for all $\alpha \geq 0$. Hence, $\alpha x \in S$ for all $\alpha \geq 0$ 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 all $\alpha \neq 1$ (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 all $\alpha \geq 0$. Hence $\alpha x \in S$ for all $\alpha \geq 0$ and S is a cone.
4. The inequality $Ax \leq b$ consists of m scalar inqualities $a_i^T x \leq b_i$ that all must hold. Let $x \in S$ and $j \in \{1, \ldots, m\}$ be such that $a_j^T x = b_j$ and $b_j \neq 0$ (such x always exists since $A \neq 0$ and since $b \neq 0$). Now, $a_j^T(\alpha x) = \alpha a_j^T x = \alpha b_j$ for all $\alpha \geq 0$.

   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 all $\alpha \geq 0$. Hence, $\alpha x \in S$ for all $\alpha \geq 0$ 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 all $\alpha \geq 0$. Hence $(\alpha x, \alpha t) \in S$ for all $\alpha \geq 0$ and S is a cone.

7. Let $X \in S$, i.e., X is symmetric and $x^T X x \geq 0$ holds for all $x \in \mathbb{R}^n$. Scaling X by α does not destroy symmetry. Also $x^T(\alpha X)x = \alpha x^T X x \geq 0$ for all $\alpha \geq 0$ and all $x \in \mathbb{R}^n$. Hence, $\alpha X \in S$ for all $\alpha \geq 0$ 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 all $\theta \in [0,1]$ 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 it is convex.
2. Union. Take $C_1 = \{0\}$ and $C_2 = \{1\}$. Then $C = \{0, 1\}$. This is not convex since, e.g., $0.5 \notin C$.

Solution 1.7

Note that for any $x, y \in \bigcap_{j \in J} C_j$ and any $\theta \in [0,1]$, we get that

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

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. Consider any $x, y \in h_{s,r}$ and any $\theta \in [0,1]$. 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. Consider any $x, y \in H_{s,r}$ and any $\theta \in [0,1]$. 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 therefore itself convex.

Solution 1.9

All of the sets are polytopes and therefore convex.

Solution 1.10

1. Consider any $y_1, y_2 \in f(C) = \{Ax + b : x \in C\}$ and any $\theta \in [0, 1]$. Then 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 get that $\theta x_1 + (1-\theta)x_2 \in C$ 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. Consider any $x_1, x_2 \in f^{-1}(D) = \{x : Ax + b \in D\}$ and any $\theta \in [0, 1]$. 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 $x, y \in \mathrm{dom} f$. Then, by definition of convexity, $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) < \infty$ for all $\theta \in [0, 1]$. That is $(\theta x + (1-\theta)y) \in \mathrm{dom} f$ if $x, y \in \mathrm{dom} f$ and $\mathrm{dom} f$ is convex.

Solution 1.12

1. Convex. We should prove that

$$\iota_C(\theta x + (1-\theta)y) \leq \theta\iota_C(x) + (1-\theta)\iota_C(y) \tag{7.7}$$

for all $x, y \in \mathbb{R}^n$ and all $\theta \in [0, 1]$. If $x, y \in C$, then the lefthand side and the righthand side are 0 by convexity of C, hence (7.7) holds. If $x \notin C$ or $y \notin C$, the RHS is ∞ which means that (7.7) is satisfied.

2. Convex. By the tringle inequality and positive homogenity of norms, we have for all $\theta \in [0,1]$:

$$\|\theta x + (1-\theta)y\| \leq \theta\|x\| + (1-\theta)\|y\|.$$

3. Not convex. By the tringle inequality and positive homogenity of norms, we have for all $\theta \in [0,1]$:

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

   Hence $f(x) = -\|x\|$ is only convex if we have equality for all x, y and $\theta \in [0,1]$. Now, let $y = -x \neq 0$ and $\theta = \frac{1}{2}$, which gives $0 \geq -\|x\|$. This holds with strict inequality for all $x \neq 0$. Hence f is not convex. (Another way to prove the second fact is that the convexity definition holds with equality everywhere if and only if f is affine.)

4. Not convex. The function is twice continuously differentiable. The gradient $\nabla f(x,y) = (y,x)$ and the Hessian

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

   This is not positive semidefinite (symmetric but eigenvalues -1,1). Hence f is not convex.

5. Convex. We have

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

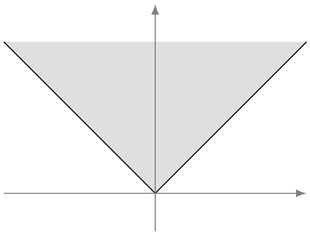
   and the convexity definition holds with equality.

6. Convex. The Hessian is $\nabla^2 f(x) = Q \succeq 0$, so f is convex.

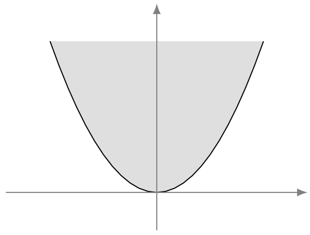
7. Convex. Let y_1 and y_2 be arbitrary. Let $x_1 \in C$ be the closest point in C from y_1 and let x_2 be the closed poitn in C from y_2. Further, let $\theta \in [0,1]$ and define $z = \theta x_1 + (1-\theta)x_2 \in C$ due to convexity of C. Then

$$\begin{aligned}
\theta \text{dist}_C(y_1) + (1-\theta)\text{dist}_C(y_2) &= \theta\|y_1 - x_1\| + (1-\theta)\|y_2 - x_2\| \\
&= \|\theta(y_1 - x_1)\| + \|(1-\theta)(y_2 - x_2)\| \\
&\geq \|\theta y_1 + (1-\theta)y_2 - (\theta x_1 + (1-\theta)x_2)\| \\
&= \|\theta y_1 + (1-\theta)y_2 - z\| \\
&\geq \text{dist}_C(\theta y_1 + (1-\theta)y_2).
\end{aligned}$$

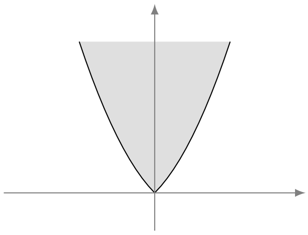
Solution 1.13



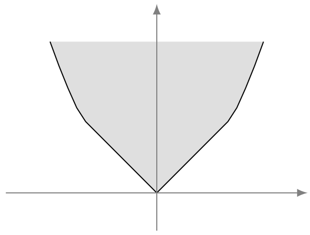
$f(x) = |x|$



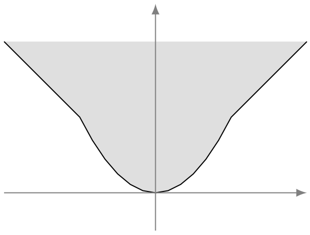
$f(x) = x^2$



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



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



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

Solution 1.14

The epigraph of f is

$$\begin{aligned}\mathbf{epi}f &= \{(x, r) : f(x) \leq r\} = \{(x, r) : a^T x + b \leq r\} \\ &= \{(x, r) : [a^T, -1] \begin{bmatrix} x \\ r \end{bmatrix} \leq -b\}\end{aligned}$$

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

Solution 1.15

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

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

We conclude that,

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

for all $(x_1, r_1), (x_2, r_2) \in \text{epi}f$ and all $\theta \in [0, 1]$. Thus, epi$f$ is convex.

Conversely, suppose that epi$f$ is convex. Consider the relationship

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2), \tag{7.8}$$

for $x_1, x_2 \in \mathbb{R}^n$ and $\theta \in [0, 1]$. If $x_1 \notin \text{dom}f$ or $x_1 \notin \text{dom}f$, relationship (7.8) clearly holds for all $\theta \in [0, 1]$. Thus, consider the case when $x_1, x_2 \in \text{dom}f$. Note that $(x_1, f(x_1)), (x_2, f(x_2)) \in \text{epi}f$. By convexity of epi$f$, we get that

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

for all $\theta \in [0, 1]$. In particular,

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

for all $x_1, x_2 \in \text{dom}f$ and all $\theta \in [0, 1]$. Thus, f is convex.

Solution 1.16

1. Consider any $x, y \in \mathbb{R}^n$ and any $\theta \in [0, 1]$. We get that

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

   Hence f is convex.

2. Recall that a function is convex if and only if the epigraph is convex. Note that

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

   We conclude that epi$f$ is convex since it is the intersection of convex sets.

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. First term: $\|z\|_2^2$ is convex and $\|Ax - b\|_2^2$ is convex since composition with affine mapping. $\|x\|_1$ convex since norm. Finally, sums of convex functions are convex.

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

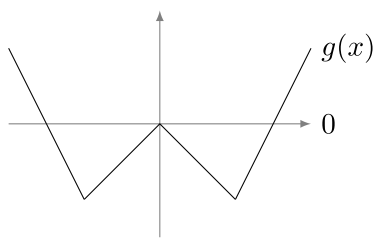
4. $max(0, 1+x_i)$ max of convex functions, hence convex. Sum over these convex functions is convex. Second term is increasing function of (convex) norm, hence convex. Nonnegative sum is convex.

5. Index all y using j from the uncountable index set J to get y_j. Further define $r_j = g(y_j)$. Then $a_j(x) = x^T y_j - r_j$ are affine functions of x and $f(x) = \sup_j(a_j(x) : j \in J)$. Since f is the supremum over a family of convex (affine) functions, it is convex.

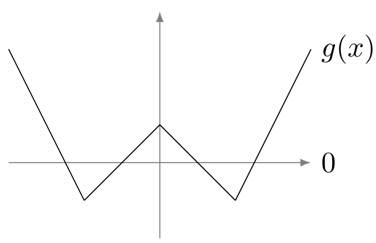
Solution 1.18

1. It is nonempty since obviously $\bar{x} \in C_\alpha$. Now, let $x_1 \in C_\alpha$ and $x_2 \in C_\alpha$ be arbitrary. Then, $g(x_1) \le \alpha$ and $g(x_2) \le \alpha$. Now, by convexity of g, we have for all $\theta \in [0, 1]$ that $x = \theta x_1 + (1-\theta)x_2$ satisfies $g(x) \le \theta g(x_1) + (1-\theta)g(x_2) \le \alpha$. Hence $x \in C_\alpha$, and 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 in nothing but a sublevel set of the convex function $\|x\|_2 - t$, and therefore a convex set.

Solution 1.21

Assume on the contrary that x^* is a local minimum, but not a global minimum, i.e., that there exists $\bar{x} \in \mathbb{R}^n$ such that $f(\bar{x}) < f(x^*)$ but that $f(x^*) \leq f(x)$ for all x such that $\|x - x^*\| \leq \delta$. Then, by convexity, for all $\theta \in (0, 1]$ we have

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

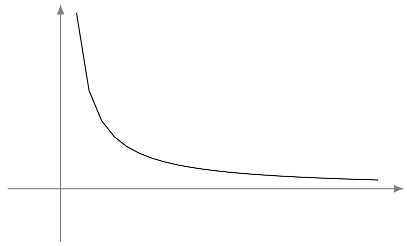
Now, let $x = (1-\theta)x^* + \theta\bar{x}$ and for small enough $\theta \in (0, 1]$ (for instance $\theta = \min(1, \frac{\delta}{\|x^* - \bar{x}\|})$) we have $\|x - x^*\| = \|(1-\theta)x^* + \theta\bar{x} - x^*\| = \theta\|x^* - \bar{x}\| \leq \delta$ but $f(x) < f(x^*)$, i.e., x^* is not a local minimum and we have reached a contradition. More specifically, we have shown that if x^* is not a global minimum, it is not a local minimum. Hence, if x^* is a local minimum, it must be a global minimum.

Solution 1.22

1. Assume on the contrary that two minimizers exist, i.e., that $x \neq x^*$ exists that satisfies $f(x) = f(x^*)$. Then, by strict convexity of f:

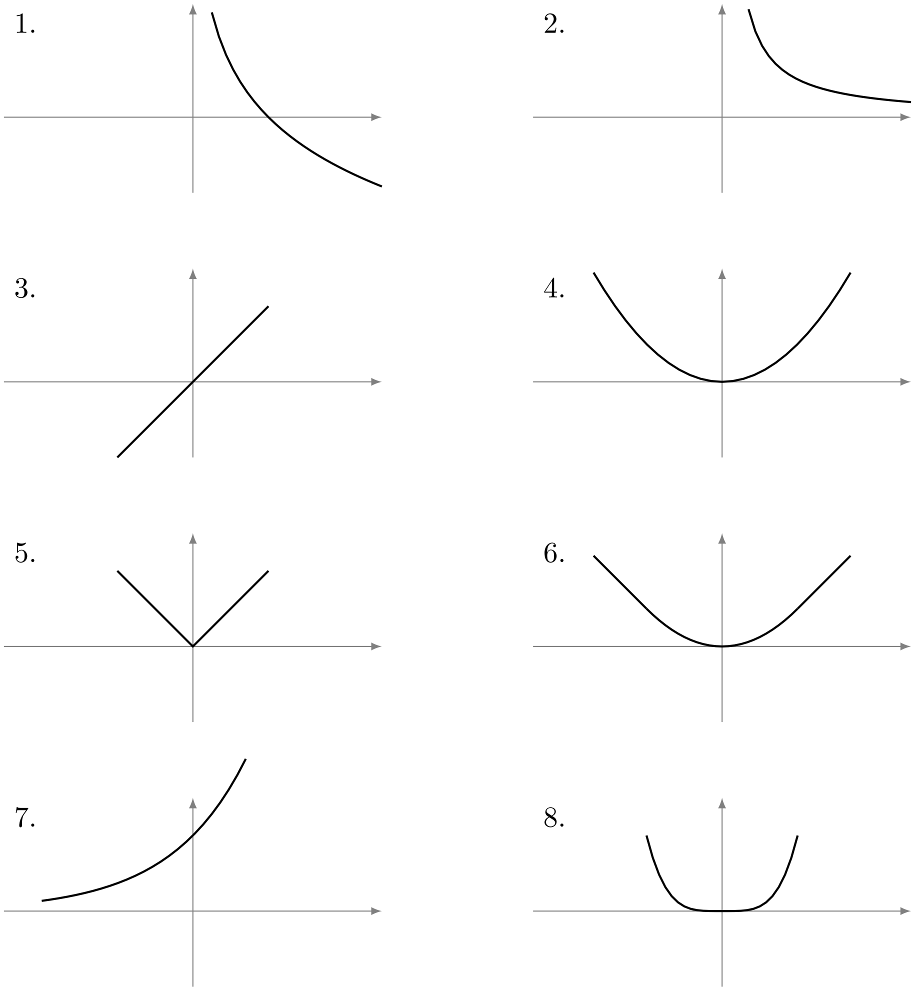
   $$f(\tfrac{1}{2}x + \tfrac{1}{2}x^*) < \tfrac{1}{2}(f(x) + f(x^*)) = f(x^*)$$

   which is a contradition. Hence, at most one minimizer can exist.
2. The function $f(x) = \frac{1}{x}$ with domain $x > 0$ is strictly convex with infimum 0. But no x exists such that $f(x) = 0$. See figure.



Solution 1.23

See figure below.

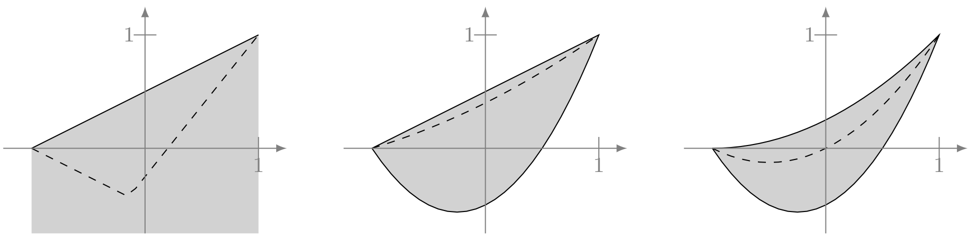


1. Not full domain, hence not smooth, strictly convex since no flat regions, not strongly convex since no quadratic lower bound.
2. Not full domain, hence not smooth, strictly convex since no flat regions, not strongly convex since no quadratic lower bound.
3. Smooth, not strictly convex since flat regions, not strongly convex.
4. Smooth, strictly convex, strongly convex.

5. Not smooth (no quadratic upper bound at 0), not strictly convex, not strongly convex.
6. Smooth since quadratic upper bounds everywhere, not strictly convex since flat regions, not strongly convex.
7. Not smooth since no uniform quadratic upper bound, stricly convex (since no flat regions), not strongly convex since no quadratic lower bound.
8. Not smooth since not uniform quadratic upper bound, strictly convex (since no flat regions), not strongly convex since no quadratic lower bound.

Solution 1.24

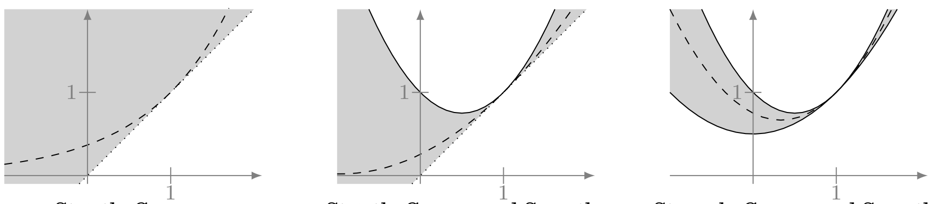
1. See the following figure. The graph a valid function must lie within the dark shaded areas. The dashed lines are examples of valid functions f. Note that smoothness implies differentiability. The example in the convex case can therefore not be used in the smooth case even though it lies within the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



Convex — Convex and Smooth — Strongly Convex and Smooth

Solution 1.25

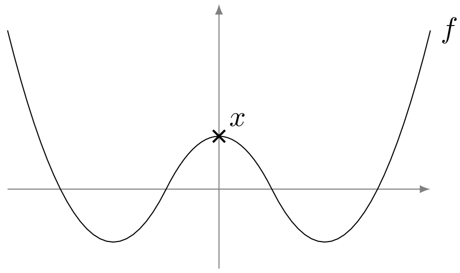
1. See the following figure. The graph a valid function must lie within the shaded areas. The dashed lines is are possible functions f.



Strictly Convex — Strictly Convex and Smooth — Strongly Convex and Smooth

Solution 1.26

1. Consider the following function f and point x:



2. Assume first that f is convex and $x, y \in \mathbb{R}^n$. By convexity of f

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

for all $\theta \in [0, 1]$. 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. That is, if f is convex, then (1.1) holds.

Now, assume instead that (1.1) holds. Choose any $x \neq y$, and $\theta \in [0, 1]$, 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 (since $\theta \in [0, 1]$)

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

That is, f is convex.

Solution 1.27

1. Let $f(x) := \sup_\mu \mu^T (Kx - b)$ and let $x \in C$, i.e., $Kx - b = 0$. Then $f(x) = \sup_\mu \mu^T 0 = 0$. That is, $f(x) = 0$ for all $x \in C$.

   If instead $x \notin C$, i.e., $Kx - b \neq 0$, then select $\mu = t(Kx - b)$ to get

   $$f(x) = \sup_\mu \mu^T (Kx - b) = \sup_t t\|Kx - b\|^2 \to \infty$$

   as $t \to \infty$. That is, $f(x) = \infty$ for all $x \notin C$.

2. Let $f(x) := \sup_{\mu \geq 0} \mu^T g(x)$ and let $x \in C$, i.e., $g(x) \leq 0$. Then for all $\mu \geq 0$, we have $\mu^T g(x) \leq 0$. In particular, for $\mu = 0$, we get $\mu^T g(x) = 0$. Hence $f(x) = \sup_\mu \mu^T g(x) = 0$. That is, $f(x) = 0$ for all $x \in C$.

If instead $x \notin C$, i.e., $g(x) > 0$, then select $\mu = tg(x)$ (which is nonnegative for all $t \geq 0$) to get

$$f(x) = \sup_{\mu} \mu^T g(x) = \sup_{t} t\|g(x)\|^2 \to \infty$$

as $t \to \infty$. That is, $f(x) = \infty$ for all $x \notin C$.

SOLUTION 1.28

We proceed by induction on n. The base case $n = 2$ is simply the definition of a convex function, and thus true. For the inductive step, assume that Jensen's inequality holds for $n = k$, were k is an integer greater than or equal to 2. We need to prove that Jensen's inequality holds for $n = k+1$. The case when $\theta_{k+1} = 1$ holds trivially, thus, we assume that $\theta_{k+1} < 1$. We get that

$$\begin{aligned}
f\left(\sum_{i=1}^{k+1} \theta_i x_i\right) &= f\left(\sum_{i=1}^{k} \theta_i x_i + \theta_{k+1} x_{k+1}\right) \\
&= f\left((1-\theta_{k+1})\left(\sum_{i=1}^{k} \frac{\theta_i}{1-\theta_{k+1}} x_i\right) + \theta_{k+1} x_{k+1}\right) \\
&\overset{\text{convexity of } f}{\leq} (1-\theta_{k+1}) f\left(\sum_{i=1}^{k} \frac{\theta_i}{1-\theta_{k+1}} x_i\right) + \theta_{k+1} f(x_{k+1}) \\
&\overset{\text{inductive assumption}}{\leq} (1-\theta_{k+1}) \sum_{i=1}^{k} \frac{\theta_i}{1-\theta_{k+1}} f(x_i) + \theta_{k+1} f(x_{k+1}) \\
&= \sum_{i=1}^{k+1} \theta_i f(x_i).
\end{aligned}$$

That is, the Jensen's inequality holds true for $n = k + 1$, establishing the inductive step. Thus, by mathematical induction, Jensen's inequality holds true for all integers $n \geq 2$.

SOLUTION 1.29

Since θ only affects the first argument of L, convexity w.r.t. the second is direct.

1. The function reads $L(\theta x, y)$ where x fixed. This is convex in the first argument w.r.t. θ since L is convex and it is a composition with an affine (linear) mapping θ.
2. Let, e.g., $\sigma(u) = u$, $L(u, y) = u$, $x = 1$ and $y \in \mathbb{R}$. Then $L(m(x;\theta), y) = m(x;\theta) = \theta_2\theta_1$, which is nonconvex. Hence, this formulation is nonconvex in general.

SOLUTION 1.30

$f(x) - \frac{\sigma}{2}\|x\|_2^2$ is convex, by definition, means that for all $x, y \in \mathbb{R}^n$ and all $\theta \in [0,1]$

$$f(z) - \tfrac{\sigma}{2}\|z\|_2^2 \leq \theta(f(x) - \tfrac{\sigma}{2}\|x\|_2^2) + (1-\theta)(f(y) - \tfrac{\sigma}{2}\|y\|_2^2),$$

where $z = \theta x + (1-\theta)y$. This is equivalent to

$$f(z) \leq \theta f(x) + (1-\theta)f(y) + \tfrac{\sigma}{2}(\|z\|_2^2 - \theta\|x\|_2^2 - (1-\theta)\|y\|_2^2). \tag{7.9}$$

Note that

$$\begin{aligned}
&\|z\|_2^2 - \theta\|x\|_2^2 - (1-\theta)\|y\|_2^2 \\
&= \|\theta\, x + (1-\theta)y\|_2^2 - \theta\|x\|_2^2 - (1-\theta)\|y\|_2^2 \\
&= (\theta^2 - \theta)\|x\|_2^2 + ((1-\theta)^2 - (1-\theta))\|y\|_2^2 + 2\theta(1-\theta)x^T y \\
&= (\theta(1-\theta))(-\|x\|_2^2 - \|y\|_2^2 + 2x^T y) \\
&= -(\theta(1-\theta))(\|x-y\|_2^2).
\end{aligned} \tag{7.10}$$

Inserting (7.10) into (7.9) gives the desired result.

Solution 1.31

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 $\beta > 0$.

<u>I) $\Rightarrow$ II)</u>: Assume that I) holds. Note that for $x, y \in \mathbb{R}^n$ and $t \in \mathbb{R}$,

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

by the chain-rule. This gives that

$$f(y) - f(x) = \int_0^1 \nabla f(x + t(y-x))^T (y-x) dt, \tag{7.11}$$

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

II) $\Rightarrow$ I): 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.

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

or

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

Adding this pair of inequalities yields

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^T (z - x) - \nabla f(y)^T (z - y) + \frac{\beta}{2}\|x - z\|_2^2 + \frac{\beta}{2}\|y - z\|_2^2 \\ &= f(x) - \nabla f(x)^T x + \nabla f(y)^T y + \frac{\beta}{2}\|x\|_2^2 + \frac{\beta}{2}\|y\|_2^2 + \beta\|z\|_2^2 + z^T(\nabla f(x) - \nabla f(y) - \beta x - \beta y) \\ &= f(x) - \nabla f(x)^T x + \nabla f(y)^T y + \frac{\beta}{2}\|x\|_2^2 + \frac{\beta}{2}\|y\|_2^2 \\ &\quad + \beta\|z + \frac{1}{2\beta}(\nabla f(x) - \nabla f(y) - \beta x - \beta y)\|_2^2 - \beta\|\frac{1}{2\beta}(\nabla f(x) - \nabla f(y) - \beta x - \beta y)\|_2^2. \end{aligned}$$

We are free to choose $z = -\frac{1}{2\beta}(\nabla f(x) - \nabla f(y) - \beta x - \beta y)$. This gives

$$\begin{aligned} f(y) &\leq f(x) - \nabla f(x)^T x + \nabla f(y)^T y + \frac{\beta}{2}\|x\|_2^2 + \frac{\beta}{2}\|y\|_2^2 - \frac{1}{4\beta}\|\nabla f(x) - \nabla f(y) - \beta x - \beta y\|_2^2 \\ &= f(x) - \frac{1}{4\beta}\|(\nabla f(x) - \nabla f(y)\|_2^2 \\ &\quad + \frac{\beta}{2}\|x\|_2^2 + \frac{\beta}{2}\|y\|_2^2 - \frac{\beta}{4}\|x + y\|_2^2 - \nabla f(x)^T x + \nabla f(y)^T y + \frac{1}{2}(\nabla f(x) - \nabla f(y))^T (x + y) \\ &= f(x) - \frac{1}{4\beta}\|\nabla f(x) - \nabla f(y)\|_2^2 + \frac{\beta}{4}\|x - y\|_2^2 + \frac{1}{2}(\nabla f(x) + \nabla f(y))^T (y - x). \end{aligned}$$

We may insert x for y and y for x in the in inequality above. This yields the pair of inequalities

$$\begin{cases} f(y) \leq f(x) - \frac{1}{4\beta}\|\nabla f(x) - \nabla f(y)\|_2^2 + \frac{\beta}{4}\|x - y\|_2^2 + \frac{1}{2}(\nabla f(x) + \nabla f(y))^T (y - x), \\ f(x) \leq f(y) - \frac{1}{4\beta}\|\nabla f(y) - \nabla f(x)\|_2^2 + \frac{\beta}{4}\|y - x\|_2^2 + \frac{1}{2}(\nabla f(y) + \nabla f(x))^T (x - y). \end{cases}$$

Adding the pair of inequalities gives

$$0 \leq -\frac{1}{2\beta}\|\nabla f(y) - \nabla f(x)\|_2^2 + \frac{\beta}{2}\|y - x\|_2^2,$$

i.e. I) holds.

II) $\Leftrightarrow$ III): 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

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

or

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

holds for all $x, y \in \mathbb{R}^n$. But this is II).

III) $\Leftrightarrow$ IV): Applying Exercise 1.30 to $\frac{\beta}{2}\|x\|_2^2 - f(x)$ and $f(x) + \frac{\beta}{2}\|x\|_2^2$ gives the result immediately.

SOLUTION 1.32

By Exercise 1.31 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

$$\beta I - \nabla^2 f(x) \succeq 0 \text{ and } \nabla^2 f(x) + \beta I \succeq 0, \text{ for all } x \in \mathbb{R}^n,$$

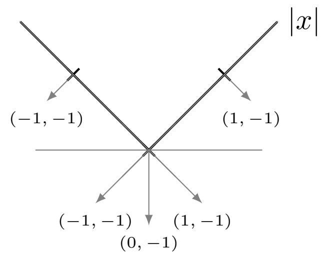
respectively. This is simply II). This establishes the desired equivalence.

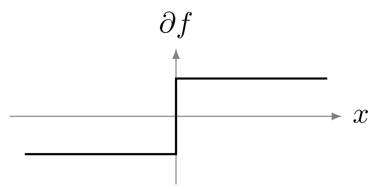
Solutions to Chapter 2

Solution 2.1

1. Function is convex and differentiable with $\nabla f(x) = x$. Hence $\partial f(x) = \{x\}$.
2. Function is convex and differentiable with $\nabla f(x) = Hx + h$. Hence $\partial f(x) = \{Hx + h\}$.
3. For $x < 0$, the function is $-x$ and differentiable with gradient -1. For $x > 0$, the function is x and differentiable with gradient 1. At $x = 0$, all elements in $[-1, 1]$ are subgradients (see figure).

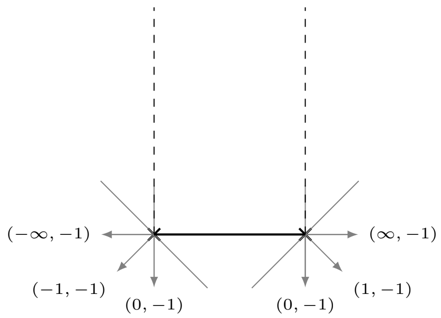
$$\partial f(x) = \begin{cases} -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

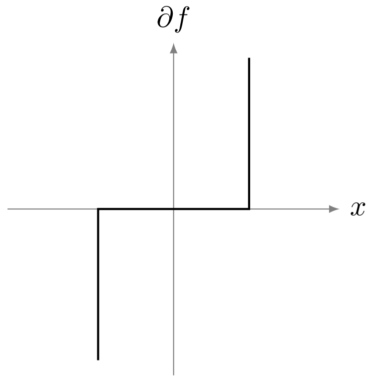




4. The function is convex. Whenever $x \in (-1, 1)$, the function is 0 with gradient 0, hence $\partial f(x) = 0$. When $x > 1$ or $x < -1$, x is outside the domain and $\partial f(x) = \emptyset$. When $x = 1$, all $s \geq 0$ are subgradients. When $x = -1$ all $s \leq 0$ are subgradient (see figure). Note that this subdifferential is the inverse of the subdifferential of $|x|$.

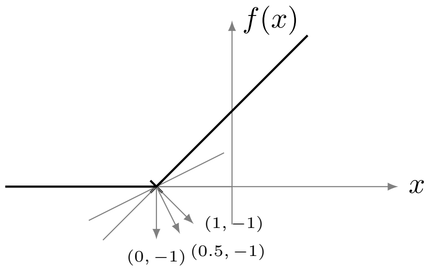
$$\partial f(x) = \begin{cases} [-\infty, 0] & \text{if } x = -1 \\ 0 & \text{if } x \in (-1, 1) \\ [0, \infty] & \text{if } x = 1 \\ \emptyset & \text{else} \end{cases}$$

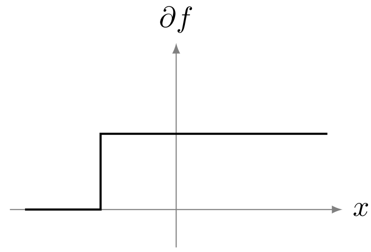




5. The function is convex. For $x < -1$, the function is 0 and the gradient is 0, hence $\partial f(x) = 0$. For $x > -1$, the function is $x + 1$ and the gradient is 1, hence $\partial f(x) = 1$. For $x = -1$, all $s \in [0, 1]$ are subgradients (see figure).

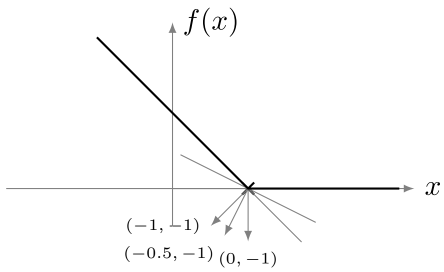
$$\partial f(x) = \begin{cases} 0 & \text{if } x < -1 \\ [0,1] & \text{if } x = -1 \\ 1 & \text{if } x > -1 \end{cases}$$

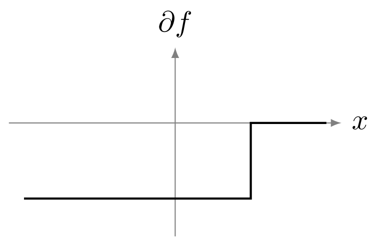




6. The function is convex. For $x > 1$, the function is 0 and the gradient is 0, hence $\partial f(x) = 0$. For $x < 1$, the function is $-x + 1$ and the gradient is -1, hence $\partial f(x) = -1$. For $x = 1$, all $s \in [-1, 0]$ are subgradients (see figure).

$$\partial f(x) = \begin{cases} -1 & \text{if } x < 1 \\ [-1,0] & \text{if } x = 1 \\ 0 & \text{if } x > 1 \end{cases}$$





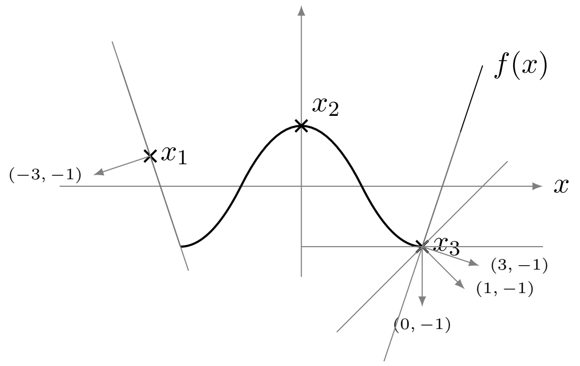
Solution 2.2

1. See figure below.

   x_1: There is one affine minorizor to f at x_1 with slope -3. Hence $\partial f(x_1) = \{-3\}$. 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 $\nabla f(x_2) = 0$.

   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.



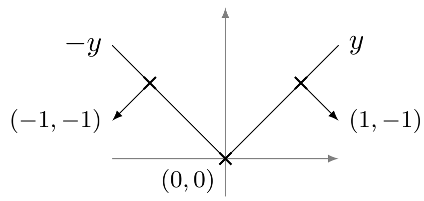
2. Fermat's rule $0 \in \partial f(x)$ holds for x_3 but not for x_1 and x_2. Therefore, x_3 is a global minimum to the nonconvex function f.

Solution 2.3

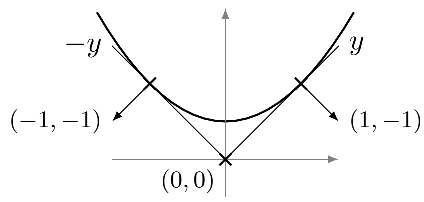
1. Yes, since $0 \in \partial f(x)$.
2. No, since $0 \notin \partial g(y)$.
3. No, since subdifferential not singleton (unique) at x.
4. No, since subdifferential not singleton (unique) at y.
5. See examples below.

Solution 2.4

1. The following function (which is the absolute value $|x|$) is a lower bound to 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 the subdifferential of a closed convex function. The subdifferential in b. is not maximally monotone, hence not the subdifferential of a closed convex function.

Solution 2.6

$A - \sigma I$ is monotone means that

$$((s_x - \sigma x) - (s_y - \sigma y))^T (x - y) \geq 0,$$

for all $x, y \in \mathrm{dom} A$, $s_x \in Ax$ and $s_y \in Ay$. The inequality is equivalent to

$$(s_x - s_y)^T (x - y) \geq \sigma \|x - y\|_2^2.$$

But this is the definition of A being σ-strongly monotone.

Solution 2.7

We know that we need to consider $n \geq 2$ since for $n = 1$, all monotone operators are subdifferentials of functions. Let $n = 2$ and set linear single-valued $A : \mathbb{R}^2 \to$

$\mathbb{R}^2$ as $A(x_1, x_2) = (x_2, -x_1)$, which can (with notation overloading) be represented by the matrix

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Then $A = -A^T$ (it 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.

It is not the gradient of a function since the matrix A would be the Hessian, but it is not symmetric.

Solution 2.8

1. 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. Using the hint we get that

$$\begin{aligned}&f(y) - f(x) - \nabla f(x)^T(y - x) \\ &= \int_0^1 t^{-1}(\nabla f(x + t(y - x)) - \nabla f(x))^T((x + t(y - x)) - x)dt \geq 0.\end{aligned}$$

   But this is I).

Solution 2.9

1. a. Since ∂f is maximally monotone, f is convex.
   b. Since ∂f is not maximally monotone, f is not convex.
2. The optimal point x^* satisfies $0 \in \partial f(x^*)$ (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, $g(x) := f(x) - \frac{\sigma}{2}\|x\|_2^2$ is convex. By the subdifferential sum rule, $\partial g(x) = \partial f(x) - \sigma x$. Now, by convexity of g, we have for all $s_g \in \partial g(x)$ that $s_g = s_f - \sigma x$ for some $s_f \in \partial f(x)$ and

$$\begin{aligned} f(y) - \tfrac{\sigma}{2}\|y\|_2^2 = g(y) \geq g(x) + s_g^T(y-x) &= f(x) - \tfrac{\sigma}{2}\|x\|_2^2 + s_g^T(y-x) \\ &= f(x) - \tfrac{\sigma}{2}\|x\|_2^2 + (s_f - \sigma x)^T(y-x). \end{aligned}$$

Now, since $\|y\|_2^2 - \|x\|_2^2 - 2x^T(y-x) = \|x-y\|_2^2$, this is equivalent to

$$f(y) = f(x) + s_f^T(y-x) + \tfrac{\sigma}{2}\|x-y\|_2^2.$$

Solution 2.11

(a) f not differentiable (∂f multivalued at 0), hence ∂f not Lipschitz. ∂f not strongly monotone (minimum slope 0), hence f not strongly convex.

(b) f differentiable (not multivalued anywhere). Max slope 1 implies ∂f 1-Lipschitz. ∂f not strongly monotone (minimum slope 0), hence f not strongly convex.

(c) f differentiable (not multivalued anywhere). Max slope 1 implies ∂f 1-Lipschitz. ∂f not strongly monotone (minimum slope 0), hence f not strongly convex.

(d) f differentiable (not multivalued anywhere). Max slope 1 implies ∂f 1-Lipschitz. ∂f strongly monotone with minimum slope 1/2. Since ∂f is also maximal, f is $\frac{1}{2}$-strongly convex.

Solution 2.12

Suppose that $s_i \in \partial g_i(x_i)$. Then

$$g_i(y_i) \geq g_i(x_i) + s_i(y_i - x_i).$$

Summing over i gives

$$g(y) \geq g(x) + \sum_{i=1}^{n} s_i(y_i - x_i) = g(x) + s^T(y - x)$$

and $s = (s_1, \ldots, s_n)$ is a subgradient of g.

Now suppose instead that $s \in \partial g(x)$. Then

$$\sum_{i=1}^{n} g_i(y_i) = g(y) \geq g(x) + s^T(x - y) = \sum_{i=1}^{n}(g_i(x_i) + s_i(y_i - x_i))$$

holds for all x, y. Let $j \in \{1, \ldots, n\}$ be arbitrary and set $x_i = y_i$ for all $i \neq j$, then this recues to

$$g_j(y_j) \geq g_j(x_j) + s_j(y_j - x_j),$$

i.e., $s_j \in \partial g_j$. Since j is arbitrary, the result follows.

Solution 2.13

For $x \notin \mathrm{dom} f$, subgradients $s \in \partial f(x)$ must satisfy

$$f(y) \geq f(x) + s^T(y - x) \quad \text{for all } y \in \mathbb{R}^n.$$

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

A vector s is in the subdifferential of the indicator function at x if

$$\iota_C(y) \geq \iota_C(x) + s^T(y - x)$$

for all y. Assume $x \in C$, then $\iota_C(y) \geq s^T(y - x)$ for all y, which is equivalent to that $s^T(y-x) \leq 0$ for all $y \in C$. Assume $x \notin C$ but $y \in C$. Then $0 \geq \infty + s^T(y-x)$ for all y. No such s exists and $\iota_C(x) = \emptyset$.

Solution 2.15

Fermat's rule says $x = \mathrm{prox}_{\gamma f}(z)$ if and only if $0 \in \partial f(x) + \gamma^{-1}(x - z)$.

1. We have $\partial f(x) = \{x\}$, which gives $0 = \gamma x + (x - z)$ or $x = (1+\gamma)^{-1} z$.
2. We have $\partial f(x) = \{Hx+h\}$, which gives $0 = \gamma(Hx+h)+(x-z)$ or $(I+\gamma H)x = z - \gamma h$ or $x = (I+\gamma H)^{-1}(z - \gamma h)$.
3. Let $x = \text{prox}_{\gamma f}(z)$, which means $0 \in \partial f(x) + \gamma^{-1}(x - z)$. The subdifferential satisfies

$$\partial f(x) = \begin{cases} -1 & \text{if } x < 0 \\ [-1,1] & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

   Let first $x < 0$ to have $\partial f(x) = \{-1\}$. Then

$$0 \in \partial f(x) + \gamma^{-1}(x - z)$$

   implies that $x = z + \gamma$. Now $x < 0$ if $z < -\gamma$.

   Let $x > 0$ to have $\partial f(x) = \{1\}$. Then

$$0 \in \partial f(x) + \gamma^{-1}(x - z)$$

   implies that $x = z - \gamma$. Now $x > 0$ if $z > \gamma$.

   Let $x = 0$ to have $\partial f(x) = [-1, 1]$ Then (since $x = 0$)

$$0 \in \partial f(0) + \gamma^{-1}(0 - z)$$

   implies that $z \in [-\gamma, \gamma]$.

   Hence, the prox becomes

$$\text{prox}_{\gamma f} = \begin{cases} z + \gamma & \text{if } z < -\gamma \\ 0 & \text{if } z \in [-\gamma, \gamma] \\ z - \gamma & \text{if } z > \gamma \end{cases}$$

4. The function is the indicator function of $C := [-1, 1]$, hence the prox reduces to the projection onto C. If $z \leq -1$, the projection is point is -1. If $z \in [-1, 1]$, the projection point is z since $z \in C$. If $z \geq 1$, the projection point is 1. Hence, the prox becomes

$$\text{prox}_{\gamma f} = \begin{cases} -1 & \text{if } z < -1 \\ z & \text{if } z \in [-1, 1] \\ 1 & \text{if } z > 1 \end{cases}$$

5. Let $x = \text{prox}_{\gamma f}(z)$, which means $0 \in \partial f(x) + \gamma^{-1}(x - z)$. The subdifferential satisfies

$$\partial f(x) = \begin{cases} 0 & \text{if } x < -1 \\ [0,1] & \text{if } x = -1 \\ 1 & \text{if } x > -1 \end{cases}$$

Let first $x < -1$ to have $\partial f(x) = \{0\}$. Then

$$0 \in \partial f(x) + \gamma^{-1}(x - z)$$

implies that $x = z$. Now $x < -1$ if $z < -1$.

Let $x > -1$ to have $\partial f(x) = \{1\}$. Then

$$0 \in \partial f(x) + \gamma^{-1}(x - z)$$

implies that $x = z - \gamma$. Now $x > -1$ if $z > \gamma - 1$.

Let $x = -1$ to have $\partial f(x) = [0, 1]$ Then (since $x = -1$)

$$0 \in \partial f(-1) + \gamma^{-1}(-1 - z)$$

implies that $z \in [-1, \gamma - 1]$.

Hence, the prox becomes

$$\text{prox}_{\gamma f} = \begin{cases} z & \text{if } z < -1 \\ -1 & \text{if } z \in [-1, \gamma - 1] \\ z - \gamma & \text{if } z > \gamma - 1 \end{cases}$$

6. Let $x = \text{prox}_{\gamma f}(z)$, which means $0 \in \partial f(x) + \gamma^{-1}(x - z)$. The subdifferential satisfies

$$\partial f(x) = \begin{cases} -1 & \text{if } x < 1 \\ [-1, 0] & \text{if } x = 1 \\ 0 & \text{if } x > 1 \end{cases}$$

Let first $x < 1$ to have $\partial f(x) = \{-1\}$. Then

$$0 \in \partial f(x) + \gamma^{-1}(x - z)$$

implies that $x = z + \gamma$. Now $x < 1$ if $z < 1 - \gamma$.

Let $x > 1$ to have $\partial f(x) = \{0\}$. Then

$$0 \in \partial f(x) + \gamma^{-1}(x - z)$$

implies that $x = z$. Now $x > 1$ if $z > 1$.

Let $x = 1$ to have $\partial f(x) = [-1, 0]$ Then (since $x = 1$)

$$0 \in \partial f(1) + \gamma^{-1}(1 - z)$$

implies that $z \in [1 - \gamma, 1]$.

Hence, the prox becomes

$$\text{prox}_{\gamma f} = \begin{cases} z + \gamma & \text{if } z < 1 - \gamma \\ 1 & \text{if } z \in [1 - \gamma, 1] \\ z & \text{if } z > 1 \end{cases}$$

SOLUTION 2.16

We have

$$
\begin{aligned}
\operatorname{prox}_{\gamma g}(z) &= \underset{x}{\operatorname{argmin}}(g(x) + \tfrac{1}{2\gamma}\|x - z\|_2^2) \\
&= \underset{x}{\operatorname{argmin}}(\sum_{i=1}^{n} g_i(x_i) + \tfrac{1}{2\gamma}\sum_{i=1}^{n}\|x_i - z_i\|_2^2) \\
&= \underset{x}{\operatorname{argmin}}(\sum_{i=1}^{n}(g_i(x_i) + \tfrac{1}{2\gamma}\|x_i - z_i\|_2^2)) \\
&= \begin{bmatrix} \operatorname{argmin}_{x_1}(g_1(x_1) + \tfrac{1}{2\gamma}\|x_1 - z_1\|_2^2) \\ \vdots \\ \operatorname{argmin}_{x_n}(g_n(x_n) + \tfrac{1}{2\gamma}\|x_n - z_n\|_2^2) \end{bmatrix} \\
&= \begin{bmatrix} \operatorname{prox}_{\gamma g_1}(z_1) \\ \vdots \\ \operatorname{prox}_{\gamma g_n}(z_n) \end{bmatrix}.
\end{aligned}
$$

Solutions to Chapter 3

Solution 3.1

Compute explicit expressions for the conjugates of the following convex functions.

1. We have

$$f^*(s) = \sup_x (s^T x - \tfrac{1}{2}\|x\|_2^2)$$

Now $\nabla f(x) = x$. Fermat's rule says x is a solution if and only if $0 = s - x$. Hence,

$$f^*(s) = s^T s - \tfrac{1}{2}\|x\|_2^2 = \tfrac{1}{2}\|s\|_2^2.$$

2. We have

$$f^*(s) = \sup_x (s^T x - \tfrac{1}{2}x^T H x - h^T x)$$

Now $\nabla f(x) = Hx + h$. Fermat's rule says x is a solution if and only if $0 = s - Hx - h$, i.e. $x = H^{-1}(s - h)$ (since H invertible). Hence,

$$\begin{aligned} f^*(s) &= s^T(H^{-1}(s-h)) - \tfrac{1}{2}(s-h)^T H^{-1} H H^{-1}(s-h) - h^T H^{-1}(s-h) \\ &= \tfrac{1}{2}(s-h)^T H^{-1}(s-h). \end{aligned}$$

3. We have

$$f^*(s) = \sup_{x \in [-1,1]} sx$$

For $s \le 0$, an optimal $x = -1$ and $f^*(s) = -s$.

For $x \ge 0$, an optimal $x = 1$ and $f^*(s) = s$.

Therefore

$$f^*(s) = \begin{cases} -s & \text{if } s \le 0 \\ s & \text{if } s \ge 0 \end{cases}$$

i.e., $f^*(s) = |s|$.

4. Since $\iota_{[-1,1]}$ is (closed) convex, $\iota^{**}_{[-1,1]} = \iota_{[-1,1]}$. In view of the above $f^* = |\cdot|^* = (\iota^*_{[-1,1]})^* = \iota^{**}_{[-1,1]} = \iota_{[-1,1]}$.

   It can also be proven explicitly. We have

$$f^*(s) = \sup_x (sx - |x|).$$

   For $s < -1$, let $x = t_- \le 0$ with $t_- \to -\infty$, which gives

$$f^*(s) = \sup_x (xs - |x|) \ge st_- - |t_-| = (s+1)t_- \to \infty$$

   since $s < -1$.

   For $s > 1$, let $x = t_+ \ge 0$ with $t_+ \to \infty$, which gives

$$f^*(s) = \sup_x (xs - |x|) \ge st_+ - |t_+| = (s-1)t_+ \to \infty$$

   since $s > 1$.

   For $s \in [-1,1]$, we have by Cauchy-Schwarz that $sx \le |x||s| \le |x|$ for all x. Therefore $f^*(s) = \sup_x s^T x - |x| \le \sup_x |x| - |x| = 0$. Further, $f^*(s) = \sup_x sx - |x| \ge s0 - |0| = 0$. Hence $f^*(s) = 0$ for all $s \in [-1,1]$.

   The conjugate becomes

$$f^*(s) = \begin{cases} 0 & \text{if } s \in [-1,1] \\ \infty & \text{else} \end{cases}$$

   i.e. $f^*(s) = \iota_{[-1,1]}(s)$

5. For all $s \in \partial f(x)$, the conjugate satisfies $f^*(s) = sx - f(x)$ (Fenchel-Young). The subdifferential is:

$$\partial f(x) = \begin{cases} 0 & \text{if } x < -1 \\ [0,1] & \text{if } x = -1 \\ 1 & \text{if } x > -1 \end{cases}$$

   Let $x < -1$, then $s = 0$ and $f^*(0) = 0 - f(x) = 0$ (since $x < -1$).

   Let $x > -1$, then $s = 1$ and $f^*(1) = x - f(x) = x - (x+1) = -1$ (since $x > -1$).

   Let $x = -1$, then $s \in [0,1]$ and $f^*(s) = -s - f(-1) = -s$ (since $x = -1$).

   The other s are not subgradients to f at any x. We verify that $f^*(s) = \infty$.

   For $s < 0$, let $x = t_- \le 0$ with $t_- \to -\infty$ and

$$f^*(s) \ge st_- - f(t_-) = st_- \to \infty.$$

   For $s > 1$, let $x = t_+ \ge 1$ with $t_+ \to \infty$ and

$$f^*(s) \ge st_+ - f(t_+) = (s-1)t_+ + 1) \to \infty.$$

   Hence, the conjugate is

$$f^*(s) = \begin{cases} -s & \text{if } s \in [0,1] \\ \infty & \text{else} \end{cases}$$

6. For all $s \in \partial f(x)$, the conjugate satisfies $f^*(s) = sx - f(x)$ (Fenchel-Young). The subdifferential is:

$$\partial f(x) = \begin{cases} -1 & \text{if } x < 1 \\ [-1, 0] & \text{if } x = 1 \\ 0 & \text{if } x > 1 \end{cases}$$

Let $x < 1$, then $s = -1$ and $f^*(-1) = -x - (1 - x) = -1$ (since $x < 1$).

Let $x > 1$, then $s = 0$ and $f^*(0) = 0 - f(x) = 0$ (since $x > 1$).

Let $x = 1$, then $s \in [-1, 0]$ and $f^*(s) = s - f(1) = s$ (since $x = 1$).

The other s are not subgradients to f at any x. We verify that $f^*(s) = \infty$.

For $s - 1$, let $x = t_- \le -1$ with $t_- \to -\infty$ and

$$f^*(s) \ge st_- - f(t_-) = (s + 1)t_- - 1 \to \infty.$$

For $s > 0$, let $x = t_+ \ge 0$ with $t_+ \to \infty$ and

$$f^*(s) \ge st_+ - f(t_+) = st_+ \to \infty.$$

Hence, the conjugate is

$$f^*(s) = \begin{cases} s & \text{if } s \in [-1, 0] \\ \infty & \text{else} \end{cases}$$

Solution 3.2

1. Assume that $f \le g$, i.e.

$$f(x) \le g(x),$$

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

$$s^T x - f(x) \ge s^T x - g(x),$$

for all $s, x \in \mathbb{R}^n$. In particular,

$$f^*(s) = \sup_x \left(s^T x - f(x)\right) \ge \sup_x \left(s^T x - g(x)\right) = g^*(s),$$

for all $s \in \mathbb{R}^n$. We conclude that $f^* \ge g^*$.

2. Assume that $f \le g$. For the previous subproblem we get that $f^* \ge g^*$, i.e.

$$f^*(s) \ge g^*(s),$$

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

$$x^T s - f^*(s) \le x^T s - g^*(s),$$

for all $s, x \in \mathbb{R}^n$. In particular,

$$f^{**}(x) = \sup_s \left(x^T s - f^*(s)\right) \le \sup_s \left(x^T s - g^*(s)\right) = g^{**}(x),$$

for all $x \in \mathbb{R}^n$. We conclude that $f^{**} \le g^{**}$.

3. Assume that $f = \frac{1}{2}\|\cdot\|_2^2$. From Exercise 3.1 we know that $f^* = \frac{1}{2}\|\cdot\|_2^2$. Therefore, $f = f^*$.

   Now assume that $f = f^*$. Note that

$$f(x) + f(s) = f(x) + f^*(s) \geq x^T s,$$

   for all $s, x \in \mathbb{R}^n$, by Fenchel-Young's inequality. If we pick $s = x$, we get that

$$f(x) \geq \frac{1}{2}\|x\|_2^2,$$

   for all $x \in \mathbb{R}^n$, i.e. $f \geq \frac{1}{2}\|\cdot\|_2^2$. However, we know from the first 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

By definition,

$$\left(\frac{|\cdot|^p}{p}\right)^*(s) = \sup_x \left(sx - \frac{|x|^p}{p}\right).$$

Using Fermat's rule, we get that x is a solution if and only if

$$0 \in \partial\left(\frac{|\cdot|^p}{p}\right)(x) - s \quad \text{or} \quad s \in \partial\left(\frac{|\cdot|^p}{p}\right)(x).$$

By the hint we have that

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

Let $x \neq 0$, then $s = x\,|x|^{p-2}$, and

$$\begin{aligned}
\left(\frac{|\cdot|^p}{p}\right)^*(s) &= sx - \frac{|x|^p}{p} \\
&= x\,|x|^p\,x - \frac{|x|^p}{p} \\
&= \left(1 - \frac{1}{p}\right)|x|^p \\
&= \frac{|x|^p}{q} \\
&= \frac{|x|^{(p-1)q}}{q} \\
&= \frac{\left|x\,|x|^{p-2}\right|^q}{q} \\
&= \frac{|s|^q}{q}.
\end{aligned}$$

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, for every $s \in \mathbb{R}^n$, it holds that

$$\begin{aligned}
(\alpha f + (1-\alpha) g)^*(s) &= \sup_x \left(s^T x - (\alpha f(x) + (1-\alpha) g(x))\right) \\
&= \sup_x \left(\alpha \left(s^T x - f(x)\right) + (1-\alpha)\left(s^T x - g(x)\right)\right) \\
&\leq \sup_x \left(\alpha \left(s^T x - f(x)\right)\right) + \sup_x \left((1-\alpha)\left(s^T x - g(x)\right)\right) \\
&= \alpha \sup_x \left(s^T x - f(x)\right) + (1-\alpha) \sup_x \left(s^T x - g(x)\right) \\
&= \alpha f^*(s) + (1-\alpha) g^*(s),
\end{aligned}$$

i.e. $(\alpha f + (1-\alpha) g)^* \leq \alpha f^* + (1-\alpha) g^*$ holds, as desired.

Solution 3.5

We have

$$\begin{aligned}
g^*(s) = \sup_x (x^T s - g(x)) &= \sup_x (\sum_{i=1}^n x_i s_i - g_i(x_i)) \\
&= \sum_{i=1}^n \sup_{x_i} (x_i s_i - g_i(x_i)) \\
&= \sum_{i=1}^n g_i^*(s_i).
\end{aligned}$$

Solution 3.6

1. The function $f(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$. Therefore

$$f^*(s) = \sum_{i=1}^n f_i^*(s_i) = \sum_{i=1}^n \iota_{[-1,1]}(s_i) = \iota_{\|\cdot\|_\infty \le 1}(s)$$

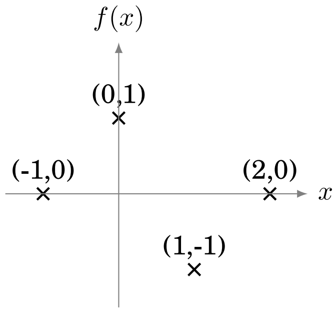
2. The function $f(x) = \iota_{[-\mathbf{1},\mathbf{1}]}(x) = \sum_{i=1}^n \iota_{[-1,1]}(x_i)$. Therefore

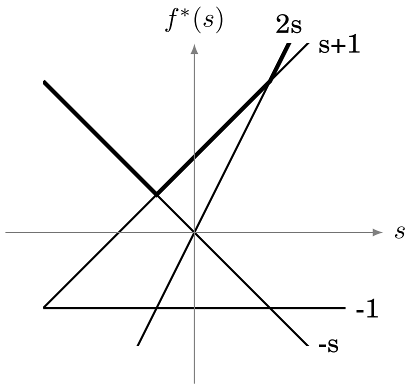
$$f^*(s) = \sum_{i=1}^n f_i^*(s_i) = \sum_{i=1}^n |s_i| = \|s\|_1.$$

Solution 3.7

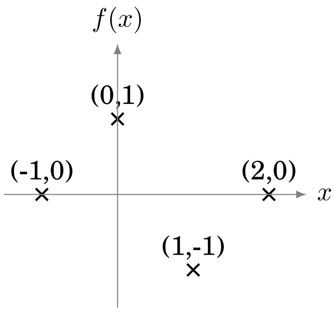
1. Since f is only defined in in four points, the conjugate is

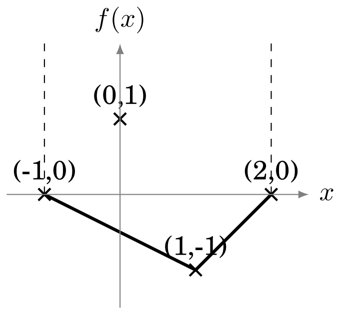
$$f^*(s) = \sup_x (sx - f(x)) = \max(-s - 0, -1, s + 1, 2s)$$





2. The biconjugate f^{**} is the convex envelope of f. See figure.





Solution 3.8

1. Conjugate $f^*(s) = \sup_x(\langle s, x\rangle - \|x\|_2)$

(a) The conjugate satisfies $f^*(s) \geq 0$ for all s since by selecting $x = 0$, we get $f^*(s) \geq \langle s, 0\rangle - \|0\|_2 = 0$.

(b) By Cauchy-Schwarz $\langle s, x\rangle \leq \|x\|_2\|s\|_2$, we have

$$f^*(s) = \sup_x(\langle s, x\rangle - \|x\|_2) \leq \sup_x(\|s\|_2\|x\|_2 - \|x\|_2) = \sup_x((\|s\|_2 - 1)\|x\|_2).$$

Hence, if $\|s\|_2 \leq 1$, $f^*(s) \leq 0$, which implies that $f^*(s) = 0$.

(c) Set $x = ts$ with $t \geq 0$ to get

$$f^*(s) = \sup_x(\langle s, x\rangle - \|x\|) \geq t\|s\|_2^2 - t\|s\|_2 = t\|s\|_2(\|s\|_2 - 1).$$

Whenever $\|s\|_2 > 1$, we let $t \to \infty$ to conclude that $f^*(s) = \infty$.

(d) To summarize

$$f^*(s) = \begin{cases} 0 & \text{if } \|s\|_2 \leq 1 \\ \infty & \text{else} \end{cases}$$

2. The subdifferential of f satisfies

$$\partial f(x) = \operatorname*{Argmax}_s(s^T x - f^*(s)) = \operatorname*{Argmax}_{\|s\|_2 \leq 1}(s^T x).$$

If $x = 0$, then the objective is 0 and all feasible points are optimal, i.e., $\partial f(0) = B(0,1) := \{s : \|s\|_2 \leq 1\}$.

If instead $x \neq 0$, then $\max_{\|s\|_2 \leq 1}(s^T x) \leq \max_{\|s\|_2 \leq 1} \|s\|_2\|x\|_2 = \|x\|_2$. Now, let $s = \frac{x}{\|x\|_2}$, to get $\max_{\|s\|_2 \leq 1}(s^T x) \geq \|x\|_2$.

Therefore

$$\partial f(x) = \begin{cases} B(0,1) & \text{if } x = 0 \\ x/\|x\|_2 & \text{else} \end{cases}$$

Solution 3.9

1. The claim says that

$$f^*(s) = \max_i(s_i) = \sup_{x \in \Delta} s^T x$$

Suppose that i is any index where $s_i = \max_i(s_i)$. First note that $x = e_i \in \Delta$ gives $s^T x = s_i$. Now let $x \neq e_i$ but $x \in \Delta$. Then

$$s^T x = \sum_{j=1}^n x_j s_j = x_i s_i + \sum_{j \neq i} s_j x_j \leq x_i s_i + s_i \sum_{j \neq i} x_j = s_i \sum_{j=1}^n x_j = s_i.$$

Hence, all points $x \in \Delta \backslash e_i$ satisfy $s^T x \leq s_i$. Therefore $\sup_{x \in \Delta} s^T x = \max_i(s_i)$.

2. Since element-wise max is closed and convex, the conjugate is ι_Δ.
3. The claim says that

$$f^*(s) = \max(0, \max_i(s_i)) = \sup_{x \in D} s^T x$$

   Suppose that i is any index where $s_i = \max_i(s_i)$ and that $s_i \geq 0$. First note that $x = e_i \in \Delta$ gives $s^T x = s_i$. Now let $x \neq e_i$ but $x \in D$. Then

$$s^T x = \sum_{j=1}^n x_j s_j = x_i s_i + \sum_{j \neq i} s_j x_j \leq x_i s_i + s_i \sum_{j \neq i} x_j = s_i \sum_{j=1}^n x_j \leq s_i,$$

   where the last step uses $s_i \geq 0$. Hence, whenever $s_i \geq 0$ all points $x \in D \backslash e_i$ satisfy $s^T x \leq s_i$. Therefore $\sup_{x \in D} s^T x = \max_i(s_i)$ for all s with at least one nonnegative element $s_i \geq 0$.

   Assume now all s_i are negative. Then for all $x \in D$:

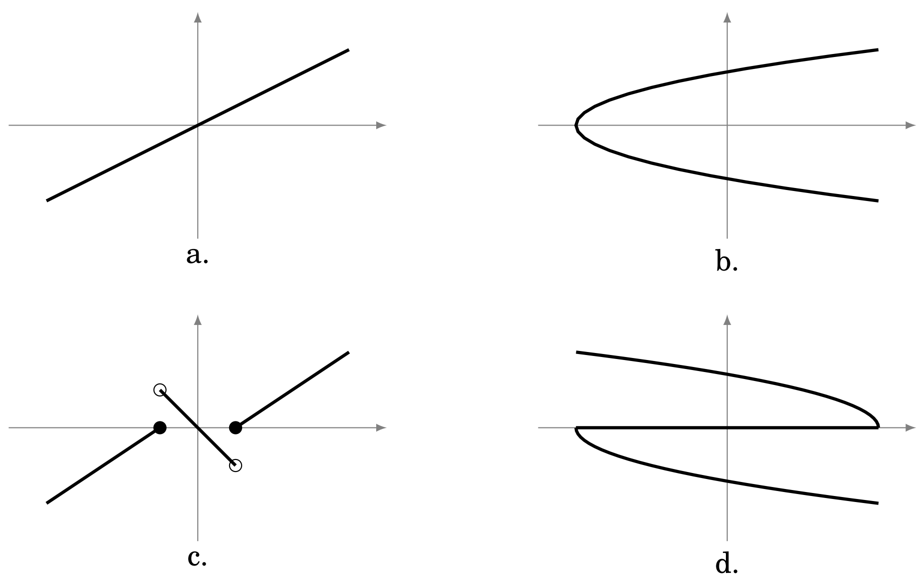
$$s^T x \leq 0 = s^T 0.$$

   Hence $x = 0$ is optimal and $f^*(s) = 0$ for s with all negative elements.

   Combined, this gives $f^*(s) = \max(0, \max_i(s_i))$.
4. Since $\max(0, \max_i(s_i))$ is closed and convex. The conjugate is ι_D.

Solution 3.10

1. 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

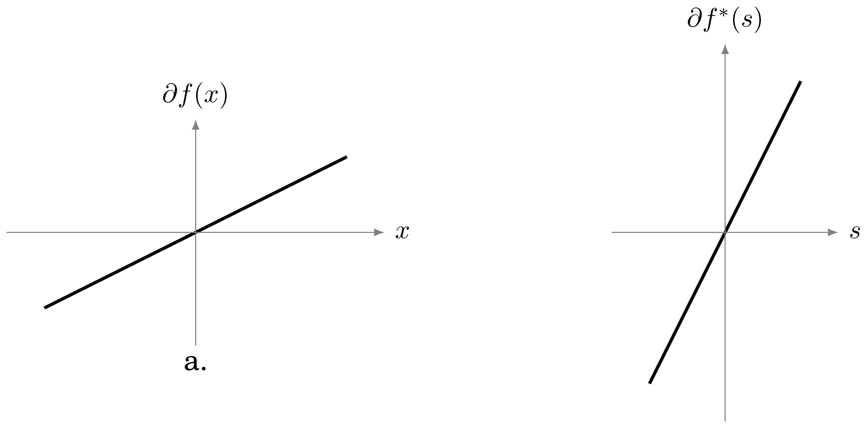


a. b.

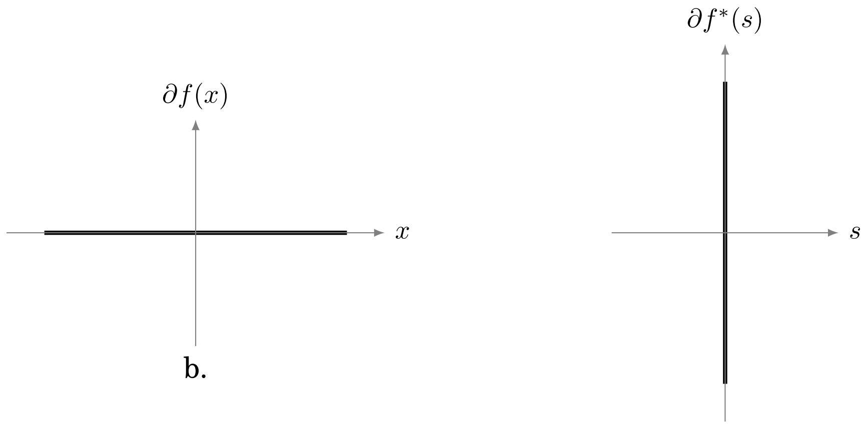
c. d.

SOLUTION 3.11

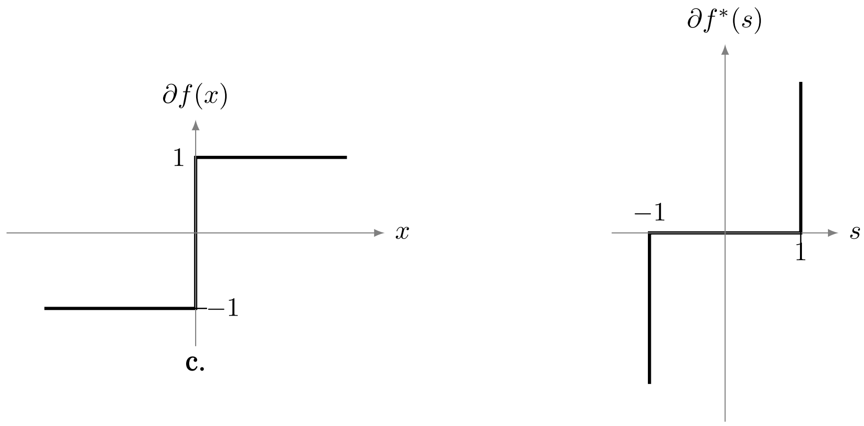
Since $\partial f^* = (\partial f)^{-1}$, we can flip the figures as follows.



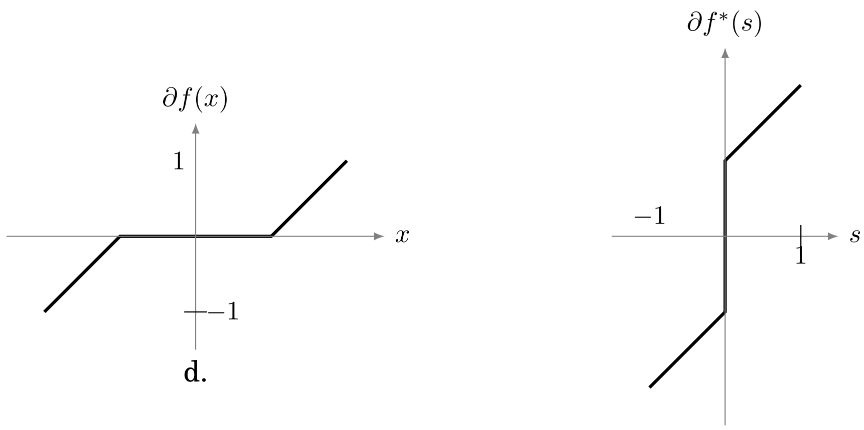
a.



b.



c.



d.

Solution 3.12

By Fermat's rule, $z = \text{prox}_{\gamma f}(x)$ if and only if

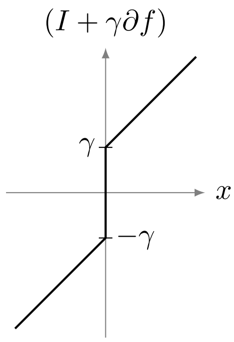
$$\begin{aligned} 0 &\in \partial f(z) + \gamma^{-1}(z - x) \\ \Leftrightarrow \qquad x &\in (I + \gamma\partial f)(z) \\ \Leftrightarrow \qquad (I + \gamma\partial f)^{-1}x &= z. \end{aligned}$$

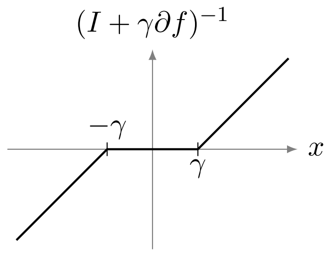
We have equality in the last step since we know that the prox is single-valued for convex functions.

Solution 3.13

1. We will solve this graphically. Left plot shows $I + \gamma\partial f$ and the right shows $(I + \gamma\partial f)^{-1} = \text{prox}_{\gamma f}$.

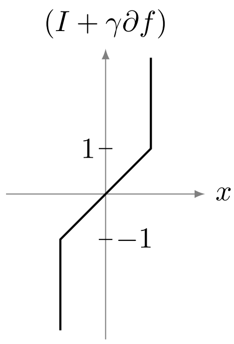
$$\text{prox}_{\gamma f}(x) = \begin{cases} x + \gamma & \text{if } x \leq -\gamma \\ 0 & \text{if } x \in [-\gamma, \gamma] \\ x - \gamma & \text{if } x \geq \gamma \end{cases}$$

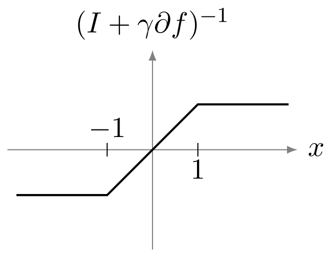




2. We will solve this graphically. Left plot shows $I + \gamma\partial f$ and the right shows $(I + \gamma\partial f)^{-1} = \text{prox}_{\gamma f}$. The prox does not depend on γ (since it is actually a projection).

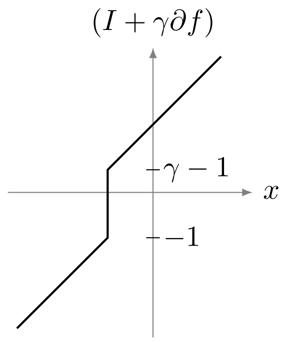
$$\text{prox}_{\gamma f}(x) = \begin{cases} -1 & \text{if } x \leq -1 \\ x & \text{if } x \in [-1, 1] \\ 1 & \text{if } x \geq 1 \end{cases}$$

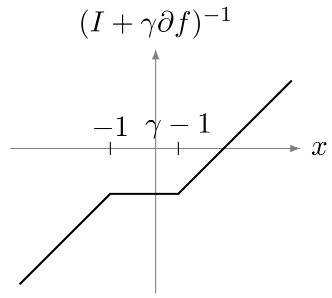




3. We will solve this graphically. Left plot shows $I + \gamma\partial f$ and the right shows $(I + \gamma\partial f)^{-1} = \mathrm{prox}_{\gamma f}$.

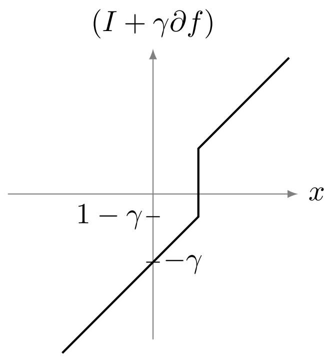
$$\mathrm{prox}_{\gamma f}(x) = \begin{cases} x & \text{if } x \le -1 \\ -1 & \text{if } x \in [-1, \gamma - 1] \\ x - \gamma & \text{if } x \ge \gamma - 1 \end{cases}$$

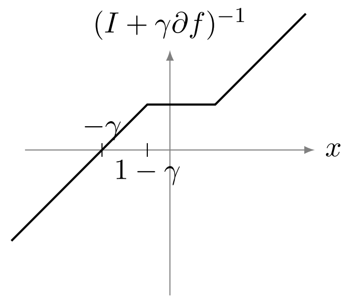




4. We will solve this graphically. Left plot shows $I + \gamma\partial f$ and the right shows $(I + \gamma\partial f)^{-1} = \mathrm{prox}_{\gamma f}$.

$$\mathrm{prox}_{\gamma f}(x) = \begin{cases} x + \gamma & \text{if } x \le 1 - \gamma \\ 1 & \text{if } x \in [1 - \gamma, 1] \\ x & \text{if } x \ge 1 \end{cases}$$





Solution 3.14

1. Let $u = z - x$. That $x = \mathrm{prox}_f(z)$ is equivalent to that

$$\begin{aligned} 0 \in \partial f(x) + x - z \quad &\Leftrightarrow \quad z - x \in \partial f(x) \\ &\Leftrightarrow \quad x \in \partial f^*(z - x) \\ &\Leftrightarrow \quad z - u \in \partial f^*(u) \\ &\Leftrightarrow \quad 0 \in \partial f^*(u) + u - z \\ &\Leftrightarrow \quad u = \mathrm{prox}_{f^*}(z). \end{aligned}$$

Since $u = z - x$ the result follows.

2. We have
$$(\gamma f^*)(s) = \sup_x (s^T x - \gamma f(x)) = \gamma \sup_x ((\gamma^{-1} s)^T x - f(x)) = \gamma f^*(\gamma^{-1} s).$$

3. We have $s = \mathbf{prox}_{(\gamma f)^*}(z)$ if and only if
$$\begin{aligned}
s &= \operatorname*{Argmin}_y ((\gamma f)^*(y) + \tfrac{1}{2}\|y - z\|_2^2) \\
&= \operatorname*{Argmin}_y (\gamma f^*(\gamma^{-1} y) + \tfrac{1}{2}\|y - z\|_2^2) \\
&= \gamma \operatorname*{Argmin}_v (\gamma f^*(v) + \tfrac{1}{2}\|\gamma v - z\|_2^2) \\
&= \gamma \operatorname*{Argmin}_v (\gamma f^*(v) + \tfrac{\gamma^2}{2}\|v - \gamma^{-1} z\|_2^2) \\
&= \gamma \operatorname*{Argmin}_v (\gamma^{-1} f^*(v) + \tfrac{1}{2}\|v - (\gamma^{-1} z)\|_2^2) \\
&= \gamma \mathbf{prox}_{\gamma^{-1} f^*}(\gamma^{-1} z)
\end{aligned}$$

4. Combine first and third subproblems.

Solution 3.15

The Moreau decomposition says $\mathbf{prox}_{(\gamma f)^*}(z) = z - \mathbf{prox}_{\gamma f}(z)$.

1. The prox of γf satisfies $\mathbf{prox}_{\gamma f}(z) = (I + \gamma H)^{-1}(z - \gamma h)$ which implies that
$$\mathbf{prox}_{(\gamma f)^*}(z) = z - (I + \gamma H)^{-1}(z - \gamma h)$$

2. The prox of γf satisfies
$$\mathbf{prox}_{\gamma f}(z) = \begin{cases} z & \text{if } z < -1 \\ -1 & \text{if } z \in [-1, \gamma - 1] \\ z - \gamma & \text{if } z > \gamma - 1 \end{cases}$$
which implies that
$$\mathbf{prox}_{(\gamma f)^*}(z) = z - \mathbf{prox}_{\gamma f}(z) = \begin{cases} 0 & \text{if } z < -1 \\ z + 1 & \text{if } z \in [-1, \gamma - 1] \\ \gamma & \text{if } z > \gamma - 1 \end{cases}$$

3. The prox of γf satisfies
$$\mathbf{prox}_{\gamma f} = \begin{cases} z + \gamma & \text{if } z < 1 - \gamma \\ 1 & \text{if } z \in [1 - \gamma, 1] \\ z & \text{if } z > 1 \end{cases}$$
which implies that
$$\mathbf{prox}_{(\gamma f)^*}(z) = z - \mathbf{prox}_{\gamma f}(z) = \begin{cases} -\gamma & \text{if } z < 1 - \gamma \\ z - 1 & \text{if } z \in [1 - \gamma, 1] \\ 0 & \text{if } z > 1 \end{cases}$$

Solution 3.16

1. We have
$$-f^*(0) = -\sup_x(0^T x - f(x)) = -\sup_x(-f(x)) = \inf_x f(x).$$

2. We have
$$\partial f^*(0) = \underset{x}{\text{Argmax}}(0^T x - f^{**}(x)) = \text{Argmax}(-f(x)) = \text{Argmin}\, f(x)$$
where we have used that $f^{**} = f$.

Solution 3.17

1. The functions are closed convex and constraint qualification holds so the primal problem is equivalent to
$$0 \in \partial f(x) + \partial g(x) \quad \Leftrightarrow \quad \begin{cases} y \in \partial f(x) \\ -y \in \partial g(x) \end{cases} \quad \Leftrightarrow \quad \begin{cases} x \in \partial f^*(y) \\ x \in \partial g^*(-y) \end{cases}$$
since $\partial f = (\partial g^*)^{-1}$ for (closed) convex functions.

2. Eliminating the x gives:
$$\begin{cases} x \in \partial f^*(y) \\ x \in \partial g^*(-y) \end{cases} \quad \Leftrightarrow \quad 0 \in \partial f^*(y) - \partial g^*(-y)$$

3. In general no. Inspired by $x \in \partial f^*(y)$ you could use the subgradient selector to generate a candidate solution $\hat{x} = s_{f^*}(y^\star)$. But
$$\begin{cases} x \in \partial f^*(y^\star) \\ x \in \partial g^*(-y^\star) \end{cases}$$
need not hold for all $x \in \partial f^*(y^\star)$ so
$$\hat{x} = s_{f^*}(y^\star) \in \partial f^*(y^\star) \nRightarrow \hat{x} \in \partial g^*(-y).$$
If f^* is differentiable $\partial f^*(y)$ is a singleton (unique) for all y. This means that for every $y^\star$, $x^\star$ is unique such that
$$\begin{cases} x^\star = \nabla f^*(y^\star) \\ x^\star \in \partial g^*(-y^\star). \end{cases}$$
In this case, the subgradient selector is the gradient and $\hat{x} = s_{f^*}(y^\star) = \nabla f^*(y^\star) = x^\star$ will recover the solution.

SOLUTION 3.18

Fermat's rule gives

$$\begin{aligned}
& 0 \in L^T \partial f(Lx) + \partial g(x) \\
\Leftrightarrow \quad & \begin{cases} \quad\; y \in \partial f(Lx) \\ -L^T y \in \partial g(x) \end{cases} \\
\Leftrightarrow \quad & \begin{cases} Lx \in \partial f^*(y) \\ \;\; x \in \partial g^*(-L^T y) \end{cases} \\
\Leftrightarrow \quad & 0 \in f^*(y) - L\partial g^*(-L^T y) \\
\Rightarrow \quad & 0 \in \partial(f^* + g^* \circ -L^T)(y)
\end{aligned}$$

which is Fermat's rule (optimality conditions) for the dual problem

$$\underset{y}{\text{minimize}}(f^*(y) + g^*(-L^T y)).$$

SOLUTION 3.19

The general dual problem is

$$\text{minimize}\, f^*(\mu) + g^*(-L^T\mu).$$

1. We have $f^*(\mu) = \frac{1}{2\lambda}\|\mu\|_2^2$ (Exercise 3.1-1 and 3.14-2) and $g^*(\nu) = \sum_{i=1}^n \max(0, 1-\nu_i)$ (Exercise 3.1-6 and 3.5 and using $g^{**} = g$). Therefore the dual problem is

$$\text{minimize}\, \tfrac{1}{2\lambda}\|\mu\|_2^2 + \sum_{i=1}^n \max(0, 1 + (L^T\mu)_i).$$

   Since the functions are convex and the primal and dual constraint qualifications hold, we can recover a primal solution from the primal-dual optimality condition $Lx = \partial f^*(\mu) = \frac{1}{\lambda}\mu \Rightarrow x = \frac{1}{\lambda}L^{-1}\mu$, which holds with equality since f^* is differentiable, i.e., $\partial f^*(\mu)$ is a singleton.

2. We have $f^*(\mu) = \|\mu\|_1$ (Exercise 3.6) and $g^*(\nu) = \frac{1}{2\lambda}\|\nu + b\|_2^2$ (Exercise 3.1-2). Therefore the dual problem is

$$\text{minimize}\, \|\mu\|_1 + \tfrac{1}{2\lambda}\| - L^T\mu + b\|_2^2.$$

   Since the functions are convex and the primal and dual constraint qualifications hold, we can recover a primal solution from the primal-dual optimality condition $x = \partial g^*(-L^T\mu) = \frac{1}{\lambda}(-L^T\mu + b)$, which holds with equality since g^* is differentiable, i.e., $\partial g^*(-L^T\mu)$ is a singleton.

SOLUTION 3.20

1. By definition, we have

$$f^*(s) = \sup_z (s^T z - f(z)) \geq s^T x - f(x),$$

as desired.

2. Suppose that $s \in \partial f(x)$. We have that

$$\begin{aligned}
& & s \in \partial f(x) \\
\Leftrightarrow & & f(y) \geq f(x) + s^T(y - x) \quad \text{for all } y \in \mathbb{R}^n \\
\Leftrightarrow & & s^T x - f(x) \geq s^T y - f(y) \quad \text{for all } y \in \mathbb{R}^n \\
\Leftrightarrow & & s^T x - f(x) \geq \sup_y (s^T y - f(y)) \\
\Leftrightarrow & & s^T x - f(x) \geq f^*(s) \\
\Leftrightarrow & & f^*(s) \leq s^T x - f(x),
\end{aligned}$$

as desired.

3. Suppose that $f^*(s) = s^T x - f(x)$. In partial, $f^*(s) \leq s^T x - f(x)$. However, the above equivalences gives that $s \in \partial f(x)$, as as desired.

Solution 3.21

1. Suppose that $s \in \partial f(x)$. 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 3.2-1). 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. Use above the results and that $f^{**} = f$ for closed convex f.

Solution 3.22

Introduce $h(y) = f(y + c)$. Then $g(x) = h(Lx)$ and

$$\begin{aligned}
g^*(s) &= \sup_x (s^T x - h(Lx)) \\
&= -\inf_x (h(Lx) + l_s(x)),
\end{aligned}$$

where $l_s(x) = -s^T x$. The conjugates satisfy

$$\begin{aligned}
h^*(\mu) &= \sup_y(\mu^T y - f(y+c)) \\
&= \sup_v(\mu^T(v-c) - f(v)) \\
&= \sup_v(\mu^T v - f(v)) - \mu^T c \\
&= f^*(\mu) - \mu^T c
\end{aligned}$$

$$\begin{aligned}
l_s^*(\nu) &= \sup_x(\nu^T x + s^T x) \\
&= \sup_x((\nu+s)^T x) \\
&= \iota_{\{0\}}(\nu+s).
\end{aligned}$$

Fenchel strong duality holds (constraint qualification is satisfied since $\operatorname{dom} l_s = \mathbb{R}^n$), and we get that

$$\begin{aligned}
g^*(s) &= -\inf_x(h(Lx) + l_s(x)) \\
&= -\sup_\mu(-h^*(\mu) - l_s^*(-L^T\mu)) \\
&= \inf_\mu(h^*(\mu) + l_s^*(-L^T\mu)) \\
&= \inf_{\mu:s=L^T\mu}(f^*(\mu) - \mu^T c).
\end{aligned}$$

Solution 3.23

Since f is closed convex, we have that $f(x) = f^{**}(x) = \sup_{s\in\mathbb{R}^n}(x^T s - f^*(s))$ for each $x \in \mathbb{R}^n$. Therefore,

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

is equal to

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

However, we may switch the supremums to get the equal problem

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

But this is equal to

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

since $g^*(s) = \sup_{x\in\mathbb{R}^n}(x^T s - g(x))$ 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

$$x^\star = x^\star - \gamma \nabla f(x^\star) \quad \Leftrightarrow \quad 0 = \nabla f(x^\star).$$

The first order condition for differentiable convex functions then gives that $x^\star$ minimizes f.

Solution 4.2

By definition,

$$z = \mathbf{prox}_{\gamma f}(x) = \underset{y}{\mathbf{argmin}}(f(y) + \tfrac{1}{2\gamma}\|y - x\|_2^2).$$

Fermat's rule and subdifferential calculus rules gives that z satisfies

$$0 \in \partial f(z) + \gamma^{-1}(z - x).$$

If $x = z = \mathbf{prox}_{\gamma f}(x)$, then this reduces to $0 \in \partial f(x)$, and x minimizes f.

Solution 4.3

By definition,

$$x = \mathbf{prox}_{\gamma g}(x - \gamma \nabla f(x)) = \underset{z}{\mathbf{argmin}} \left(g(z) + \tfrac{1}{2\gamma}\|z - (x - \gamma \nabla f(x))\|_2^2 \right).$$

Fermat's rule and subdifferential calculus rules gives that this is equivalent to

$$\begin{aligned} 0 &\in \partial g(x) + \tfrac{1}{\gamma}(x - (x - \gamma \nabla f(x))) \\ &= \partial g(x) + \nabla f(x) \\ &= \partial (g + f)(x). \end{aligned}$$

This gives that x is a minimum of $f + g$.

Solution 4.4

1. The function is smooth so gradient descent works. No need to use prox.
2. First two parts are smooth, third is not smooth but separable and easy to prox on, proximal gradient works but gradient-descent doesn't.
3. Both functions are smooth, second is separable and easy to prox on. Gradient descent and proximal gradient both work.
4. First function is smooth, second is easy to prox on but not smooth. Proximal gradient is the only alternative.
5. Neither of the functions are differentiable, so none of the methods work.
6. The first term is differentible, but not smooth (it grows too quick for large x), and the second is proximable but not differentiable. So none of the method works.
7. First term is smooth, second is proximable and seperable. Proximal gradient works.
8. The second term is neither smooth nor simple to prox on, nether of the methods would be efficient.
9. The first part is smooth, and the second part is smooth and easy to prox on. Both gradient descent and proximal gradient therefore works.

Solution 4.5

1. $\|Ax - b\|_2^2$ is not strongly convex unless 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, and the primal is not stronly convex. The dual will therefore not be smooth, thus neither of the methods work.
2. $\frac{1}{2}x^TQx + b^Tx$ is strongly convex since $Q \succ 0$ so its dual is smooth. The dual of the last part is easy to prox on but not smooth. Proximal gradient works.
3. First function is not strongly convex so dual of this part is not smooth and not proximable. However, if we let $f(x) = \frac{1}{2}\|x - b\|_2^2$ and $g(x) = \|x\|_2^2$, the problem is $\min_x f(Ax) + g(x)$ and the dual can be written $\min_\mu f^*(\mu) + g^*(-A^T\mu)$. $f^*(\mu)$ is convex, separable, smooth and proximable and $g^*(-A^T\mu)$ is smooth and strongly convex. Hence, any of the methods work.
4. First function is not strongly convex so dual is not smooth and it is not easy to prox on. Doing the same trick as for the previous problem doesn't work since $\|Ax\|_2$ is not smooth. Hence none of the methods works.
5. Neither is strongly convex so neither of the duals are smooth. None of the methods works.
6. Neither is strongly convex ($e^{\|x\|^4} \approx \|x^4\| + 1$ for small x) so neither of the duals are smooth. None of the methods works.
7. First term is strongly convex so dual is smooth, second is proximable so the same is true for the dual. Proximal gradient works.

8. With $f(x) = \iota_{[-1,1]}(x), g(x) = \frac{1}{2}x^TQx$, the primal problem can written as $\min_x f(Lx) + g(x)$ so the dual is $\min_\mu f^*(\mu) + g^*(-L\mu)$, where $g^*(\mu) = \frac{1}{2}x^TQ^{-1}x$, i.e $g^*(-L\mu) = \frac{1}{2}x^TL^TQ^{-1}Lx$ which is smooth. f^* is proximable so proximal gradient works.
9. Neither of the functions are strongly convex so the neither of the duals will be smooth. Hence, none of the algorithms work.

Solution 4.6

1. From Exercise 3.1 we know that $h^*(\mu_i) = \iota_{[-1,1]}(\mu_i)$ if $h(x_i) = |x_i|$. Since $f(x) = \|x\|_1 = \sum_{i=1}^n h(x_i)$ is separable, we know that $f^*(\mu) = \sum_{i=1}^n h^*(\mu_i) = \sum_{i=1}^n \iota_{[-1,1]}(\mu_i) = \iota_{-\mathbf{1}\leq\mu\leq\mathbf{1}}(\mu)$, where $\mathbf{1}$ is the vector of all ones.
2. From Exercise 3.1 we know that $g^*(\mu) = \frac{1}{2}\mu^TQ^{-1}\mu$.
3. One possible dual problem is given by
$$\underset{\mu}{\text{minimize}}\, f^*(\mu) + g^*(-\mu).$$
E.g., let $L = I$ in Exercise 3.18. Similarly, another dual problem is given by
$$\underset{\mu}{\text{minimize}}\, f^*(-\mu) + g^*(\mu).$$
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
$$\underset{\mu}{\text{minimize}}\, f^*(\mu) + g^*(-\mu),$$
we get, for some appropriate $\gamma_k > 0$, that
$$\mu_{k+1} = \mathbf{prox}_{\gamma_k f^*}\left(\mu_k - \gamma_k\nabla(g^* \circ -I)(\mu_k)\right),$$
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
$$\begin{aligned}\mathbf{prox}_{\gamma_k f^*}(z) &= \underset{\mu}{\text{argmin}}\left(\iota_{-\mathbf{1}\leq\mu\leq\mathbf{1}}(\mu) + \tfrac{1}{2\gamma_k}\|\mu - z\|_2^2\right)\\ &= \underset{-\mathbf{1}\leq\mu\leq\mathbf{1}}{\text{argmin}}\,\|\mu - z\|_2^2.\end{aligned}$$

Note that both the constraint set and the objective function of this argmin-problem are separable, yielding the following simple problem

$$(\mathrm{prox}_{\gamma_k f^*}(z))_i = \operatorname*{argmin}_{\mu_i \in [-1,1]} |\mu_i - z_i|^2 = \begin{cases} 1 & \text{if } z_i > 1 \\ -1 & \text{if } z_i < -1 \\ z_i & \text{otherwise} \end{cases} .$$

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

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

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

$$\begin{cases} x_k = \nabla g^*(-L^T\mu_k), \\ v_k = \mu_k + \gamma_k L x_k, \\ \mu_{k+1} = \mathrm{prox}_{\gamma_k f^*}(v_k). \end{cases}$$

Using Moreau decomposition, we have

$$\mathrm{prox}_{\gamma_k f^*}(z) = z - \gamma_k \mathrm{prox}_{\gamma_k^{-1} f^{**}}(\gamma_k^{-1} z) = z - \gamma_k \mathrm{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} z).$$

The last equality holds since $f = f^{**}$, by closed convexity of f. Using this, we can write the proximal gradient method step as

$$\begin{cases} x_k = \nabla g^*(-L^T\mu_k), \\ v_k = \mu_k + \gamma_k L x_k, \\ \mu_{k+1} = v_k - \gamma_k \mathrm{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} v_k). \end{cases}$$

Recall the subdifferential formula for g^*, i.e.

$$\partial g^*(\mu) = \operatorname*{Argmax}_x \left(\mu^T x - g^{**}(x)\right) = \operatorname*{Argmax}_x \left(\mu^T x - g(x)\right).$$

The last equality holds since $g = g^{**}$, by closed convexity of g. However, we know that g^* is smooth convex, and therefore, $\partial g^*(\mu) = \{\nabla g^*(\mu)\}$. In particular,

$$\nabla g^*(\mu) = \operatorname*{argmax}_x \left(\mu^T x - g(x)\right) = \operatorname*{argmin}_x \left(g(x) - \mu^T x\right).$$

This lets us write the proximal gradient method step as

$$\begin{cases} x_k = \textbf{argmin}_x \left(g(x) + \mu_k^T L x \right), \\ v_k = \mu_k + \gamma_k L x_k, \\ \mu_{k+1} = v_k - \gamma_k \textbf{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} v_k), \end{cases}$$

as desired.

Solution 4.8

Recall that Exercise 4.6 gives the dual proximal gradient method step

$$\begin{cases} v_k = \mu_k - \gamma_k Q^{-1} \mu_k \\ (\mu_{k+1})_i = \begin{cases} 1 & \text{if } (v_k)_i > 1, \\ -1 & \text{if } (v_k)_i < -1, \\ (v_k)_i & \text{otherwise}, \end{cases} \quad \forall i \in \{1, \ldots, n\}. \end{cases} \tag{7.12}$$

We must verify that

$$\begin{cases} x_k = \textbf{argmin}_x \left(g(x) + \mu_k^T x \right), \\ v_k = \mu_k + \gamma_k x_k, \\ \mu_{k+1} = v_k - \gamma_k \textbf{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} v_k), \end{cases} \tag{7.13}$$

gives the same step for

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

To verify correctness, note that

$$\underset{x}{\textbf{argmin}} \left(g(x) + \mu_k^T x \right) = \underset{x}{\textbf{argmin}} \left(\tfrac{1}{2} x^T Q x + x^T \mu_k \right) = -Q^{-1} \mu_k.$$

Thus, we can write (7.13) as

$$\begin{cases} v_k = \mu_k - \gamma_k Q^{-1} \mu_k, \\ \mu_{k+1} = v_k - \gamma_k \textbf{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} v_k). \end{cases} \tag{7.14}$$

Since $f(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$ is separable, so is $\textbf{prox}_{\gamma f}$. From Exercise 2.15 we then get

$$(\textbf{prox}_{\gamma f}(z))_i = \begin{cases} z_i + \gamma & \text{if } z_i < -\gamma, \\ 0 & \text{if } -\gamma \le z_i \le \gamma, \\ z_i - \gamma & \text{if } z_i > \gamma, \end{cases} \quad \forall i \in \{1, \ldots, n\}.$$

We can then calculate μ_{k+1} in (7.14) as

$$
\begin{aligned}
(\mu_{k+1})_i &= (v_k)_i - \gamma_k(\mathbf{prox}_{\gamma_k^{-1}f}(\gamma_k^{-1}v_k))_i \\
&= (v_k)_i - \gamma_k \begin{cases} \gamma_k^{-1}(v_k)_i + \gamma_k^{-1} & \text{if } \gamma_k^{-1}(v_k)_i < -\gamma_k^{-1}, \\ 0 & \text{if } -\gamma_k^{-1} \le \gamma_k^{-1}(v_k)_i \le \gamma_k^{-1}, \\ \gamma_k^{-1}(v_k)_i - \gamma_k^{-1} & \text{if } \gamma_k^{-1}(v_k)_i > \gamma_k^{-1}, \end{cases} \\
&= (v_k)_i - \begin{cases} (v_k)_i + 1 & \text{if } (v_k)_i < -1, \\ 0 & \text{if } -1 \le (v_k)_i \le 1, \\ (v_k)_i - 1 & \text{if } (v_k)_i > 1, \end{cases} \\
&= \begin{cases} -1 & \text{if } (v_k)_i < -1, \\ (v_k)_i & \text{if } -1 \le (v_k)_i \le 1, \\ 1 & \text{if } (v_k)_i > 1, \end{cases} \quad \forall i \in \{1, \ldots, n\}.
\end{aligned}
$$

This establishes the desired equality.

SOLUTION 4.9

Using the hint we get, with $x = x_k$, that

$$
\begin{aligned}
f(y) &\le f(x_k) + \nabla f(y)^T(x_k - y) + \tfrac{\beta}{2}\|x_k - y\|_2^2 \\
&< f(x_k) + \nabla f(y)^T(x_k - y) + \tfrac{1}{2\gamma_k}\|x_k - y\|_2^2, \quad \forall y \in \mathbb{R}^n.
\end{aligned}
$$

The function

$$
g(y) = f(x_k) + \nabla f(y)^T(x_k - y) + \tfrac{1}{2\gamma_k}\|x_k - y\|_2^2,
$$

is then a majorizer to f, i.e., $f \le g$. What remain to be shown is that

$$
x_{k+1} = \operatorname*{argmin}_y g(y).
$$

From Fermat's rule 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 cost yield

$$\begin{aligned}
&\sum_{i=1}^{N} \log(1 + e^{-y_i(w^T x_i + b)}) \\
&\quad = \sum_{\forall i: y_i = -1} \log(1 + e^{w^T x_i + b}) + \sum_{\forall i: y_i = 1} \log(1 + e^{-(w^T x_i + b)}) \\
&\quad = \sum_{\forall i: y_i = -1} \log(1 + e^{w^T x_i + b}) + \sum_{\forall i: y_i = 1} \log\left(\frac{1 + e^{w^T x_i + b}}{e^{w^T x_i + b}}\right) \\
&\quad = \sum_{\forall i: y_i = -1} \log(1 + e^{w^T x_i + b}) + \sum_{\forall i: y_i = 1} \log(1 + e^{w^T x_i + b}) - \sum_{\forall i: y_i = 1} \log(e^{w^T x_i + b}) \\
&\quad = \sum_{i=1}^{N} \log(1 + e^{w^T x_i + b}) - \sum_{\forall i: y_i = 1} \log(e^{w^T x_i + b}) \\
&\quad = \sum_{i=1}^{N} \log(1 + e^{w^T x_i + b}) - \sum_{\forall i: y_i = 1} w^T x_i + b.
\end{aligned}$$

From here we can go over to the new labels, $y_i = 1 \to \hat{y}_i = 1$ and $y_i = -1 \to \hat{y}_i = 0$.

$$\begin{aligned}
&= \sum_{i=1}^{N} \log(1 + e^{w^T x_i + b}) - \sum_{\forall i: \hat{y}_i = 1} w^T x_i + b \\
&= \sum_{i=1}^{N} \log(1 + e^{w^T x_i + b}) - \sum_{i=1}^{N} \hat{y}_i(w^T x_i + b) \\
&= \sum_{i=1}^{N} \Big(\log(1 + e^{w^T x_i + b}) - \hat{y}_i(w^T x_i + b)\Big).
\end{aligned}$$

Solution 5.2

We first note that all terms in the sum are positive for all finite (w, b). Let $u_i = x_i^T w + b$, and each term reduces to $\log(1 + e^{u_i}) - y_i(u_i)$. For $y_i = 0$, $\log(1 + e^{u_i}) > 0$ since $1 + e^{u_i} > 1$. For $y_i = 1$, $\log(1 + e^{u_i}) - u_i = \log(\frac{1+e^{u_i}}{e^{u_i}}) > 0$ since $\frac{1+e^{u_i}}{e^{u_i}} > 1$.

Now, take an i with $y_i = 0$. Let $(w, b) = t(\bar{w}, \bar{b})$, then

$$\begin{aligned}\log(1 + e^{x_i^T w + b}) - y_i(x_i^T w + b) &= \log(1 + e^{t(x_i^T \bar{w} + \bar{b})}) = \log(1 + e^t e^{x_i^T \bar{w} + \bar{b}}) \\ &\leq \log(1 + e^t) \to 0\end{aligned}$$

as $t \to -\infty$, where $e^{x_i^T \bar{w} + \bar{b}} \in (0, 1)$ (since $x_i^T \bar{w} + \bar{b} < 0$) has been used in the inequality.

Instead, take i with $y_i = 1$. Then

$$\begin{aligned}\log(1 + e^{x_i^T w + b}) - y_i(x_i^T w + b) &= \log(1 + e^{t(x_i^T \bar{w} + \bar{b})}) - t(x_i^T \bar{w} + \bar{b}) \\ &= \log(\tfrac{1 + e^t e^{x_i^T \bar{w} + \bar{b}}}{e^t e^{(x_i^T \bar{w} + \bar{b})}}) \\ &= \log(1 + e^{-t} e^{-(x_i^T \bar{w} + \bar{b})}) \\ &\leq \log(1 + e^{-t}) \to 0\end{aligned}$$

as $t \to \infty$, where $e^{-(x_i^T \bar{w} + \bar{b})} \in (0, 1)$ (since $x_i^T \bar{w} + \bar{b} > 0$) has been used in the inequality.

Hence, the infimum is 0 which is not attained by any (w, b) since the cost is positive for all (finite) (w, b).

Solution 5.3

First, consider the case when $\lambda = 0$. The objective function is

$$\tfrac{1}{2}\|ax - b\|_2^2.$$

By Fermat's rule, we get that the solution in this case is

$$0 = a^T(ax_{\text{ls}} - b) \quad \text{or} \quad x_{\text{ls}} = \tfrac{a^T b}{\|a\|_2^2}.$$

Now, consider the case when $\lambda > 0$. Using the subdifferential calculus rules (CQ holds since both functions have full domain) and Fermat's rule, the optimality condition is given by

$$0 \in \|a\|_2^2 x - a^T b + \lambda \begin{cases} \text{sgn}(x) & \text{if } x \neq 0, \\ [-1, 1] & \text{if } x = 0. \end{cases}$$

Thus, $x = 0$ is optimal 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 optimal if and only if

$$0 = \|a\|_2^2 x - a^T b + \lambda\, \text{sgn}(x) \quad \text{or} \quad x = \tfrac{a^T b}{\|a\|_2^2} - \tfrac{\lambda}{\|a\|_2^2}\, \text{sgn}(x).$$

However, since $|a^T b| > \lambda$ by assumption, $\text{sgn}(x) = \text{sgn}(a^T b) = \text{sgn}(x_{\text{ls}})$ must hold by necessity. Therefore, the solution in this case is given by

$$x = x_{\text{ls}} - \tfrac{\lambda}{\|a\|_2^2}\, \text{sgn}(x_{\text{ls}}).$$

This concludes the proof.

Solution 5.4

- Alternative 1:
  Optimality conditions are

$$0 \in A^T(Ax-b) + \lambda \begin{bmatrix} g(x_1) \\ \vdots \\ g(x_m) \end{bmatrix},$$

  where

$$g(x_i) = \begin{cases} -1 & \text{if } x_i < 0 \\ [-1,1] & \text{if } x_i = 0 \\ 1 & \text{if } x_i > 0 \end{cases}$$

  For $x = 0$, the optimality condition reads

$$0 \in -A^T b + \lambda[-1,1]^m$$

  which holds for all $\lambda \geq \max_i(|(A^T b)_i|) = \|A^T b\|_\infty$.

- Alternative 2:
  Let $f(x) = \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$. Using Hölder's inequality, we get the following lower bound.

$$\begin{aligned} f(x) &\geq \tfrac{1}{2}\|Ax-b\|_2^2 + \|A^T b\|_\infty \|x\|_1 \\ &\geq \tfrac{1}{2}\|Ax-b\|_2^2 + b^T Ax \\ &= \tfrac{1}{2}\|Ax\|_2^2 + \tfrac{1}{2}\|b\|_2^2 \\ &\geq \tfrac{1}{2}\|b\|_2^2. \end{aligned}$$

  Furthermore $f(0) = \frac{1}{2}\|b\|_2^2$ so the lower bound is attained at $x = 0$, therefore $x = 0$ is optimal.

Solution 5.5

Both functions of the problem have full domain so CQ holds. Fermat's rule then give that $x = (x_1, x_2)$ is a solution if

$$\begin{aligned} & 0 \in A^T Ax - A^T b + \lambda\partial(\|\cdot\|_1)(x) \\ \iff \quad & 0 \in \sum_{j=1}^{2} a_i^T a_j x_j - a_i^T b + \lambda\partial(|\cdot|)(x_i) \quad \forall i \in \{1,2\} \end{aligned}$$

where a_i is the ith column of A. The equivalence hold since $\|x\|_1 = |x_1| + |x_2|$. Inserting the subdifferential of $|\cdot|$ give

$$\begin{cases} 0 \in x_1 + a_1^T a_2 x_2 - a_1^T b + \lambda \begin{cases} \operatorname{sgn}(x_1) & \text{if } x_1 \neq 0 \\ [-1,1] & \text{if } x_1 = 0 \end{cases} \\ 0 \in a_2^T a_1 x_1 + x_2 - a_2^T b + \lambda \begin{cases} \operatorname{sgn}(x_2) & \text{if } x_2 \neq 0 \\ [-1,1] & \text{if } x_2 = 0 \end{cases} \end{cases} \tag{7.15}$$

where the fact that $\|a_i\|_2 = 1$ was used. With the optimality conditions in place, we can now look at the four cases.

- Assume $x \in X_{0,0}$, then (7.15) 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 $\max_i(|a_i^T b|) = \|Ab\|_\infty \leq \lambda$, i.e., $\Lambda_{0,0} = \{\lambda : \lambda \geq |a_1^T b|, \lambda \geq |a_2^T b|\}$.

- Assume $x \in X_{1,0}$, then (7.15) 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 $x \neq 0$ by assumption, i.e., if $a_1^T b = 0$ then $\Lambda_{1,0} = \emptyset$.

  From here on, we assume $a_1^T b \neq 0$. The first condition can be re-written to

$$0 = \operatorname{sgn}(x_1)|x_1| - a_1^T b + \lambda \operatorname{sgn}(x_1) \iff \frac{a_1^T b}{\operatorname{sgn}(x_1)} - |x_1| = \lambda$$

  Since $\lambda > 0$ this implies $\operatorname{sgn}(x_1) = \operatorname{sgn}(a_1^T b)$ and

$$0 < \lambda = |a_1^T b| - |x_1| < |a_1^T b|$$

  since by assumption is $x_1 \neq 0$. Multiplying both rows in the original condition with $\operatorname{sgn}(x_1) = \operatorname{sgn}(a_1^T b) = \frac{|a_1^T b|}{a_1^T b}$ gives

$$\begin{cases} |x_1| = |a_1^T b| - \lambda \\ 0 \in a_2^T a_1 |x_1| - |a_1^T b| \frac{a_2^T b}{a_1^T b} + \lambda[-1,1] \end{cases}$$
$$\implies 0 \in |a_1^T b|(a_2^T a_1 - \frac{a_2^T b}{a_1^T b}) - \lambda(a_2^T a_1 + [-1,1]).$$

  The last inclusion can be written as

$$\lambda(a_2^T a_1 - 1) \leq |a_1^T b|(a_2^T a_1 - \frac{a_2^T b}{a_1^T b}) \leq \lambda(a_2^T a_1 + 1)$$

  Since $\lambda < |a_1^T b|$ and $|a_2^T a_1| < 1$, this implies $\left|\frac{a_2^T b}{a_1^T b}\right| < 1$, if $\left|\frac{a_2^T b}{a_1^T b}\right| \geq 1$ then $\Lambda_{1,0} = \emptyset$.

We can re-formulate these conditions as

$$|a_1^T b| \frac{a_2^T a_1 - \frac{a_2^T b}{a_1^T b}}{a_2^T a_1 + \text{sgn}(a_2^T a_1 - \frac{a_2^T b}{a_1^T b})} \leq \lambda$$

Further simplification and including the $\lambda < |a_1^T b|$ condition finally give

$$\frac{|a_1^T b||\frac{a_2^T b}{a_1^T b} - a_1^T a_2|}{1 - a_1^T a_2 \,\text{sgn}(\frac{a_2^T b}{a_1^T b} - a_1^T a_2)} \leq \lambda < |a_1^T b|. \tag{7.16}$$

To summarize, the set of possible λ is then given by

$$\Lambda_{1,0} = \begin{cases} \emptyset & \text{if } a_1^T b = 0 \text{ or } |a_2^T b| \geq |a_1^T b| \\ \{\lambda \text{ s.t. (7.16) is satisfied.}\} & \text{otherwise} \end{cases}$$

- By symmetry is the set $\Lambda_{0,1}$ the same as $\Lambda_{1,0}$ but with the index swapped.
- Assume $x \in X_{1,1}$, then (7.15) is equivalent to

$$0 = A^T A x - A^T b + \lambda \begin{bmatrix} \text{sgn}(x_1) \\ \text{sgn}(x_2) \end{bmatrix}$$

where matrix $A^T A$ and it's inverse is given by the following

$$A^T A = \begin{bmatrix} 1 & a_1^T a_2 \\ a_1^T a_2 & 1 \end{bmatrix}, \quad (A^T A)^{-1} = \frac{1}{1-(a_1^T a_2)^2} \begin{bmatrix} 1 & -a_1^T a_2 \\ -a_1^T a_2 & 1 \end{bmatrix}.$$

The inverse exists by assumption since $|a_1^T a_2| < 1$. Multiplying the condition by the inverse gives

$$x = (A^T A)^{-1} A^T b - \lambda (A^T A)^{-1} \begin{bmatrix} \text{sgn}(x_1) \\ \text{sgn}(x_2) \end{bmatrix}.$$

We define

$$S = \begin{bmatrix} \text{sgn}(x_1) & 0 \\ 0 & \text{sgn}(x_2) \end{bmatrix}$$

and multiply with S from the left to get

$$0 < \begin{bmatrix} |x_1| \\ |x_2| \end{bmatrix} = S(A^T A)^{-1} A^T b - \lambda S (A^T A)^{-1} \begin{bmatrix} \text{sgn}(x_1) \\ \text{sgn}(x_2) \end{bmatrix}.$$

The last term is

$$\begin{aligned} & S(A^T A)^{-1} \begin{bmatrix} \text{sgn}(x_1) \\ \text{sgn}(x_2) \end{bmatrix} \\ &= \frac{1}{1-(a_1^T a_2)^2} \begin{bmatrix} \text{sgn}(x_1) & 0 \\ 0 & \text{sgn}(x_2) \end{bmatrix} \begin{bmatrix} 1 & -a_1^T a_2 \\ -a_1^T a_2 & 1 \end{bmatrix} \begin{bmatrix} \text{sgn}(x_1) \\ \text{sgn}(x_2) \end{bmatrix} \\ &= \frac{1 - a_1^T a_2 \,\text{sgn}(x_1 x_2)}{1-(a_1^T a_2)^2} \mathbf{1} \\ &> 0 \end{aligned}$$

since $|a_1^T a_2| < 1$. In order for the condition to have a solution we need $S(A^T A)^{-1} A^T b > 0$. In other words, $\text{sgn}(x_i) = \text{sgn}(\hat{x}_i)$ where $\hat{x} = (A^T A)^{-1} A^T b$ is the leasts squares solution. Inserting this back into the condition yields

$$0 < \begin{bmatrix} |\hat{x}_1| \\ |\hat{x}_2| \end{bmatrix} - \lambda \tfrac{1 - a_1^T a_2 \,\text{sgn}(\hat{x}_1 \hat{x}_2)}{1 - (a_1^T a_2)^2} \mathbf{1}.$$

To summarize

$$\Lambda_{1,1} = \{\lambda : \lambda < \tfrac{1 - (a_1^T a_2)^2}{1 - a_1^T a_2 \,\text{sgn}(\hat{x}_1 \hat{x}_2)} \min(|\hat{x}_1|, |\hat{x}_2|), \hat{x} = (A^T A)^{-1} A^T b\}.$$

To show that the $\Lambda_{i,j}$ sets are disjoint we note there are three different cases of interest regarding the data A and b. Either is $|a_1^T b| > |a_2^T b|$, $|a_2^T b| > |a_1^T b|$ or $|a_1^T b| = |a_2^T b|$. At least one of $\Lambda_{1,0}$ and $\Lambda_{0,1}$ is then empty since $|a_1^T b| > |a_2^T b|$ and $|a_2^T b| > |a_1^T b|$ can't hold at the same time. From symmetry it is enough to consider only one these cases, here we choose the case when $|a_1^T b| > |a_2^T b|$ and $\Lambda_{1,0}$ is non-empty. We will later cover the remaining case where $|a_1^T b| = |a_2^T b|$ and both $\Lambda_{1,0}$ and $\Lambda_{0,1}$ are empty.

Assume $\lambda_{0,0} \in \Lambda_{0,0}$, $\lambda_{1,0} \in \Lambda_{1,0}$ and $\lambda_{1,1} \in \Lambda_{1,1}$, if we show

$$\lambda_{1,1} < \lambda_{1,0} < \lambda_{0,0}$$

the three set are disjoint and sparsity increase with λ. Since $\lambda_{1,0} < |a_1^T b|$ and $|a_1^T b| \leq \lambda_{0,0}$ we have $\lambda_{1,0} < \lambda_{0,0}$. For $\lambda_{1,1}$ and $\lambda_{1,0}$ we have

$$\lambda_{1,1} < \frac{1 - (a_1^T a_2)^2}{1 - a_1^T a_2 \,\text{sgn}(\hat{x}_1 \hat{x}_2)} \min(|\hat{x}_1|, |\hat{x}_2|), \quad \frac{|a_1^T b| |\frac{a_2^T b}{a_1^T b} - a_1^T a_2|}{1 - a_1^T a_2 \,\text{sgn}(\frac{a_2^T b}{a_1^T b} - a_1^T a_2)} \leq \lambda_{1,0}$$

and we wish to show that these bounds are the same. We start by showing that $\min(|\hat{x}_1|, |\hat{x}_2|) = |\hat{x}_2|$, i.e., $|\hat{x}_2| \leq |\hat{x}_1|$. Using the definition of $\hat{x}$

$$\hat{x} = (A^T A)^{-1} A^T b = \tfrac{1}{1 - (a_1^T a_2)^2} \begin{bmatrix} a_1^T b - a_1^T a_2 a_2^T b \\ a_2^T b - a_1^T a_2 a_1^T b \end{bmatrix}$$

means that this is equivalent to

$$|a_2^T b - a_1^T a_2 a_1^T b| \leq |a_1^T b - a_1^T a_2 a_2^T b| \iff |\tfrac{a_2^T b}{a_1^T b} - a_1^T a_2| \leq |1 - a_1^T a_2 \tfrac{a_2^T b}{a_1^T b}|.$$

This can be equivalently written as

$$\begin{cases} \frac{a_2^T b}{a_1^T b} - a_1^T a_2 \leq 1 - a_1^T a_2 \frac{a_2^T b}{a_1^T b} \\ a_1^T a_2 - \frac{a_2^T b}{a_1^T b} \leq 1 - a_1^T a_2 \frac{a_2^T b}{a_1^T b} \end{cases}$$
$$\iff \begin{cases} 0 \leq 1 + a_1^T a_2 - a_1^T a_2 \frac{a_2^T b}{a_1^T b} - \frac{a_2^T b}{a_1^T b} = (1 + a_1^T a_2)(1 - \frac{a_2^T b}{a_1^T b}) \\ 0 \leq 1 - a_1^T a_2 - a_1^T a_2 \frac{a_2^T b}{a_1^T b} + \frac{a_2^T b}{a_1^T b} = (1 - a_1^T a_2)(1 + \frac{a_2^T b}{a_1^T b}) \end{cases}.$$

This hold since $|a_1^T a_2| < 1$ and $\frac{a_2^T b}{a_1^T b} < 1$ by assumption. The upper bound on $\lambda_{1,1}$ then reads

$$\lambda_{1,1} < \frac{1 - (a_1^T a_2)^2}{1 - a_1^T a_2 \,\text{sgn}(\hat{x}_1 \hat{x}_2)} |x_2| = \frac{|a_1^T b| |\frac{a_2^T b}{a_1^T b} - a_1^T a_2|}{1 - a_1^T a_2 \,\text{sgn}(\hat{x}_1 \hat{x}_2)}.$$

This is the same as the lower bound on $\lambda_{1,0}$ since

$$\begin{aligned}\operatorname{sgn}(\hat{x}_1\hat{x}_1) &= \operatorname{sgn}((a_1^Tb - a_1^Ta_2a_2^Tb)(a_2^Tb - a_1^Ta_2a_1^Tb)) \\ &= \operatorname{sgn}((1 - a_1^Ta_2\tfrac{a_2^Tb}{a_1^Tb})(\tfrac{a_2^Tb}{a_1^Tb} - a_1^Ta_2)) \\ &= \operatorname{sgn}(\tfrac{a_2^Tb}{a_1^Tb} - a_1^Ta_2)\end{aligned}$$

since $|a_1^Ta_2| < 1$ and $\frac{|a_2^Tb|}{|a_1^Tb|} < 1$. This concludes the proof for when $|a_1^Tb| > |a_2^Tb|$.

When $|a_1^Tb| = |a_2^Tb|$, both $\Lambda_{1,0} = \emptyset$ and $\Lambda_{0,1} = \emptyset$ and we want to show

$$\lambda_{1,1} < \tfrac{1-(a_1^Ta_2)^2}{1-a_1^Ta_2\operatorname{sgn}(\hat{x}_1\hat{x}_2)}\min(|\hat{x}_1|,|\hat{x}_2|) = |a_1^Tb| = |a_2^Tb| \le \lambda_{0,0}.$$

We know that

$$\begin{aligned}\tfrac{1}{1-a_1^Ta_2\operatorname{sgn}(\hat{x}_1\hat{x}_2)}|\hat{x}_1| &= \tfrac{1}{1-a_1^Ta_2\operatorname{sgn}(\hat{x}_1\hat{x}_2)}|a_2^Tb||\tfrac{a_1^Tb}{a_2^Tb} - a_1^Ta_2| \\ &= \tfrac{\operatorname{sgn}(\frac{a_1^Tb}{a_2^Tb}-a_1^Ta_2)}{1-a_1^Ta_2\operatorname{sgn}(\hat{x}_1\hat{x}_2)}|a_2^Tb|(\tfrac{a_1^Tb}{a_2^Tb} - a_1^Ta_2) \\ &= \tfrac{1}{\operatorname{sgn}(\frac{a_1^Tb}{a_2^Tb}-a_1^Ta_2)-a_1^Ta_2}|a_2^Tb|(\tfrac{a_1^Tb}{a_2^Tb} - a_1^Ta_2)\end{aligned}$$

where it was used that $\operatorname{sgn}(\hat{x}_1\hat{x}_2) = \operatorname{sgn}(\frac{a_1^Tb}{a_2^Tb} - a_1^Ta_2)$. We now note that $\frac{a_1^Tb}{a_2^Tb} = \operatorname{sgn}(\frac{a_1^Tb}{a_2^Tb})$ since $|a_1^Tb| = |a_2^Tb|$. Furthermore, we then also have

$$\operatorname{sgn}(\tfrac{a_1^Tb}{a_2^Tb} - a_1^Ta_2) = \operatorname{sgn}(\operatorname{sgn}(\tfrac{a_1^Tb}{a_2^Tb}) - a_1^Ta_2) = \operatorname{sgn}(\operatorname{sgn}(\tfrac{a_1^Tb}{a_2^Tb})) = \operatorname{sgn}(\tfrac{a_1^Tb}{a_2^Tb})$$

since $|a_1^Ta_2| < 1$. This yields

$$\begin{aligned}\tfrac{1}{1-a_1^Ta_2\operatorname{sgn}(\hat{x}_1\hat{x}_2)}|\hat{x}_1| &= \tfrac{1}{\operatorname{sgn}(\frac{a_1^Tb}{a_2^Tb})-a_1^Ta_2}|a_2^Tb|(\operatorname{sgn}(\tfrac{a_1^Tb}{a_2^Tb}) - a_1^Ta_2) \\ &= |a_2^Tb|.\end{aligned}$$

By symmetry, the analogue holds for $\frac{1}{1-a_1^Ta_2\operatorname{sgn}(\hat{x}_1\hat{x}_2)}|\hat{x}_2|$, which then yield the desired inequality

$$\lambda_{1,1} < |a_1^Tb| = |a_2^Tb| \le \lambda_{0,0}.$$

This concludes the proof for the $|a_1^Tb| = |a_2^Tb|$ case and for the statement in its entirety.

Note, in all cases can the distances $|\lambda_{1,0} - \lambda_{0,0}|$ and $|\lambda_{1,1} - \lambda_{1,0}|$ can be made arbitrary small. This is to be expected since otherwise there would be λ for which no solution exists. Since the Lasso problem is strongly convex for all $\lambda > 0$ we know that this is not the case.

SOLUTION 5.6

1. Let $t > 0$, then $t(w, b)$ also separate the data. Inserting this into the cost yields

$$\sum_{i=1}^{n} \max\left(0, 1 - ty_i(x_i^T w + b)\right) = \sum_{i=1}^{n} \max\left(0, 1 - t|x_i^T w + b|\right) \geq 0.$$

   Choosing any $t \geq (\max_i |x_i^T w + b|)^{-1}$ then give a cost of 0 and therefore $t(w, b)$ must be optimal. The set of optimal points is unbounded since $\|t(w, b)\| = t\|(w, b)\|$, $\|(w, b)\| > 0$ and $t \geq (\max_i |x_i^T w + b|)^{-1}$ can be made arbitrary large.

2. Choosing an arbitrary w and inserting into the cost gives

$$\sum_{i=1}^{n} \max\left(0, 1 - y_i(x_i^T w + b)\right) = \sum_{i=1}^{n} \max\left(0, 1 - x_i^T w - b\right) \geq 0.$$

   Choosing $b \geq 1 - \min_i x_i^T w$ give cost 0 and therefore (w, b) is optimal. The set of optimal points is unbounded since $\|(w, b)\|^2 = \|w\|^2 + |b|^2$, were $b \geq 1 - \min_i x_i^T w$, clearly can be made arbitrary large.

3. Letting $w = 0$ and inserting into the cost gives

$$\sum_{i=1}^{n} \max\left(0, 1 - y_i(x_i^T w + b)\right) + \tfrac{\lambda}{2}\|w\|_2^2 = \sum_{i=1}^{n} \max\left(0, 1 - b\right) \geq 0.$$

   Any $b \geq 1$ yields a cost of 0 and is therefore optimal. The set of optimal points is unbounded since $\|(w, b)\| = |b|$, were $b \geq 1$ can be made arbitrary large.

Solution 5.7

The regularization term is the same so we focus on the sum of hinge-losses.

$$\sum_{i=1}^{n} \max\left(0, 1 - y_i(x_i^T w + b)\right) = \mathbf{1}^T \begin{bmatrix} \max\left(0, 1 - y_1(x_1^T w + b)\right) \\ \vdots \\ \max\left(0, 1 - y_n(x_n^T w + b)\right) \end{bmatrix}$$

$$= \mathbf{1}^T \max\left(0, \begin{bmatrix} 1 - y_1(x_1^T w + b) \\ \vdots \\ 1 - y_n(x_n^T w + b) \end{bmatrix}\right) = \mathbf{1}^T \max\left(0, \mathbf{1} - \begin{bmatrix} y_1(x_1^T w + b) \\ \vdots \\ y_n(x_n^T w + b) \end{bmatrix}\right)$$

$$= \mathbf{1}^T \max\left(0, \mathbf{1} - \left(\begin{bmatrix} y_1 x_1^T w \\ \vdots \\ y_n x_n^T w \end{bmatrix} + \begin{bmatrix} y_1 b \\ \vdots \\ y_n b \end{bmatrix}\right)\right) = \mathbf{1}^T \max\left(0, \mathbf{1} - \left(\begin{bmatrix} y_1 x_1^T \\ \vdots \\ y_n x_n^T \end{bmatrix} w + b \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}\right)\right).$$

We can now identify $X = [y_1 x_1, \ldots, y_n x_n]$ and $\phi = (y_1, \ldots, y_n)$.

Solution 5.8

1. The function f is a separable sum of hinge-losses, $h(x) = \max(0, 1 - x)$, i.e.

$$f(x) = \sum_{i=1}^{n} \max(0, 1 - x_i) = \sum_{i=1}^{n} h(x_i).$$

The conjugate f^* is then

$$f^*(\mu) = \sum_{i=1}^{n} h^*(\mu_i) = \sum_{i=1}^{n} \mu_i + \iota_{[-1,0]}(\mu_i) = \mathbf{1}^T\mu + \iota_{[-\mathbf{1},0]}(\mu).$$

See Exercises 3.1 and 3.5. The conjugate of g is

$$\begin{aligned} g^*(\nu_w, \nu_b) &= \sup_{w,b}((\nu_w, \nu_b)^T(w, b) - \tfrac{\lambda}{2}\|w\|_2^2) \\ &= \sup_{w}(\nu_w^T w - \tfrac{\lambda}{2}\|w\|_2^2) + \sup_{b}(\nu_b b) \\ &= \tfrac{1}{2\lambda}\|\nu_w\|_2^2 + \iota_{\{0\}}(\nu_b). \end{aligned}$$

See Exercise 3.1. Note that

$$g^*\left(-L^T\mu\right) = g^*\left(-\begin{bmatrix} X \\ \phi^T \end{bmatrix}\mu\right) = \tfrac{1}{2\lambda}\|-X\mu\|_2^2 + \iota_{\{0\}}\left(-\phi^T\mu\right).$$

The dual problem becomes

$$\min \mathbf{1}^T\mu + \tfrac{1}{2\lambda}\|-X\mu\|_2^2 + \iota_{[-\mathbf{1},0]}(\mu) + \iota_{\{0\}}\left(-\phi^T\mu\right),$$

or written differently

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T\mu + \frac{1}{2\lambda}\mu^T X^T X\mu \\ \text{subject to} & -\mathbf{1} \le \mu \le 0 \\ & \phi^T\mu = 0 \end{array}$$

2. We first note that CQ hold for the dual problem as long as there are data points of both classes, i.e., ϕ has both positive and negative elements. Indeed, if all elements of ϕ have the same sign, the only μ in $\operatorname{dom} g^* \circ -L^T$ is $\mu = 0$, however, 0 is on the relative boundary of $\operatorname{dom} f^*$.

   A dual optimal point μ must then satisfy

$$\begin{aligned} & 0 \in \partial f^*(\mu) - L\partial g^*(-L^T\mu) \\ \iff & \begin{cases} L(w,b) \in \partial f^*(\mu) \\ (w,b) \in \partial g^*(-L^T\mu) \end{cases} \\ \iff & \begin{cases} \partial f(L(w,b)) \ni \mu \\ \partial g((w,b)) \ni -L^T\mu \end{cases} \\ \iff & 0 \in L^T\partial f(L(w,b)) + \partial g((w,b)). \end{aligned}$$

   Hence, (w, b) is a solution to the primal problem and we can recover (w, b) from

$$\begin{cases} L(w,b) \in \partial f^*(\mu) \\ (w,b) \in \partial g^*(-L^T\mu) \end{cases}$$

The second condition yields

$$w = \tfrac{-1}{\lambda} X\mu \quad \text{and} \quad b \in \mathbb{R},$$

since $\partial \iota_{\{0\}}(s) = \mathbb{R}$ if $s = 0$ and $\partial \iota_{\{0\}}(s) = \emptyset$ otherwise, and that $-\phi^T \mu = 0$ for any dual optimal μ. However, we need something else to recover the optimal b. The first condition gives

$$[X^T,\, \phi](w, b) = X^T w + b\phi \in \partial f^*(\mu),$$

and we note that

$$(\partial f^*(\mu))_i = \begin{cases} 1 & \text{if } -1 < \mu_i < 0, \\ [1, \infty] & \text{if } \mu_i = 0, \\ [-\infty, 1] & \text{if } \mu_i = -1, \\ \emptyset & \text{otherwise}. \end{cases}$$

If $-1 < \mu_i < 0$ for any i, we see that we can determine b uniquely, i.e., take b such that

$$y_i x_i^T w + b y_i = 1 \iff b = y_i^{-1} - x_i^T w = y_i - x_i^T w$$

where x_i and y_i denote data and class label defined in Exercise 5.6 and 5.7.

3. From the previous task we know that

$$w = \tfrac{1}{\lambda} X\mu = \tfrac{1}{\lambda} \sum_{i=1}^{n} x_i \mu_i = \tfrac{1}{\lambda} \sum_{i:\mu_i \neq 0} x_i \mu_i.$$

Clearly w can be recovered by only considering x_i s.t. $\mu_i \neq 0$. Since $b = y_i - x_i^T w$ for i s.t. $-1 < \mu_i < 0$ so can b.

It remains to show that $\mu_i < 0$ means that x_i is support vector. From the primal-dual optimality condition we know that

$$[X^T, \phi](w, b) = X^T w + b\phi \in \partial f^*(\mu).$$

The ith coordinate of the inclusion problem can be written as

$$y_i(x_i^T w + b) \in \begin{cases} 1 & \text{if } -1 < \mu_i < 0, \\ [1, \infty] & \text{if } \mu_i = 0, \\ [-\infty, 1] & \text{if } \mu_i = -1, \\ \emptyset & \text{otherwise}. \end{cases}$$

If $\mu_i < 0$ this means $y_i(x_i^T + b) \leq 1$, or equivalently $0 \leq 1 - y_i(x_i^T + b)$, and x_i is a support vector.

Solution 5.9

1. True. See the model $m_w(x)$ as a function of w instead of x, $m_w(x) = f_x(w) = w^T\phi(x)$. Clearly, $f_x(w)$ is linear in w since $\phi(x)$ does not depend on w and therefore is constant. Since y_i also does not depend on w is $L(m_w(x_i), y_i) = L(f_{x_i}(w), y_i)$ a composition between convex and linear, which is convex. The full cost is then a sum of convex functions which is convex.

2. False. Consider a two layer network with identity activation functions and $\mathbb{R} \to \mathbb{R}$ layers, $m_w(x) = w_1 w_2 x$. Take the square error loss and the data $x = 1$ and $y = 0$, then $L(m_w(x_i), y_i) = \|w_1 w_2\|_2^2 = (w_1 w_2)^2$ which is not convex. The points $(0, 1)$ and $(1, 0)$ both have function value 0 but the point $(0.5, 0.5)$ on the line between them have a positive function value.

Solutions to Chapter 6

Solution 6.1

To estimate the overall computational cost of an algorithm, we can roughly use (iterations count) $\times$ (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 any $k \in \mathbb{N}$. This implies an R-linear rate with $C_L = V_0$.

2. From the Q-quadratic rate definition we have that

$$
\begin{aligned}
&V_1 \leq \rho V_0^2 \\
&V_2 \leq \rho V_1^2 \leq \rho\rho^2 V_0^{2^2} \\
&V_3 \leq \rho V_2^2 \leq \rho\rho^2\rho^{2^2} V_0^{2^3} \\
&V_4 \leq \rho V_3^2 \leq \rho\rho^2\rho^{2^2}\rho^{2^3} V_0^{2^4} \\
&\quad\vdots \\
&V_k \leq \rho V_{k-1}^2 \leq \rho\rho^2\rho^{2^2}\rho^{2^3}\ldots\rho^{2^{k-1}} V_0^{2^k} = \rho^{2^k-1}V_0^{2^k} = \rho^{2^k}V_0^{2^k}/\rho,
\end{aligned}
$$

or

$$
V_k \leq (\rho V_0)^{2^k}\rho^{-1},
$$

holds inductively for any $k \in \mathbb{N}$. Here, we used that $1+2+2^2+\ldots+2^{k-1} = 2^k - 1$. This implies an R-quadratic rate with $\rho_Q = \rho V_0$ and $C_Q = \rho^{-1}$. Since $V_k \geq 0$ by assumption, it will converge to zero if the upper bound, $(\rho V_0)^{2^k}\rho^{-1}$, converge to zero. This upper bound only converge to zero if $\rho V_0 < 1$.

Solution 6.4

1. Let $n = 1$ and consider $f(x) = x$ and $x_k = -k$ for each $k \in \mathbb{N}$. Clearly, $(x_k)_{k\in\mathbb{N}}$ is a descent sequence and $f(x_k)$ goes to $-\infty$ as $k \to \infty$. I.e. the sequence of function values $(f(x_k))_{k\in\mathbb{N}}$ does not converge in $\mathbb{R}$.

2. Solution 1: Note that the sequence $(f(x_k))_{k\in\mathbb{N}}$ is monotone, by construction. Moreover, $(f(x_k))_{k\in\mathbb{N}}$ is bounded - from above by $f(x_0)$ and from below by B. Then, by the monotone convergence theorem, the sequence $(f(x_k))_{k\in\mathbb{N}}$ converges in $\mathbb{R}$.

   Solution 2: First, note that the non-empty set $\{f(x_k) : k \in \mathbb{N}\}$ in $\mathbb{R}$ is bounded from below by B or equivalently, $\{-f(x_k) : k \in \mathbb{N}\}$ 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 $\sup\{-f(x_k) : k \in \mathbb{N}\} = \tilde{b}$, or equivalently, $\inf\{f(x_k) : k \in \mathbb{N}\} = b$, 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 $\{f(x_k) : k \in \mathbb{N}\}$, i.e. $c \leq f(x_k)$ for each $k \in \mathbb{N}$, it holds that $c \leq b$.

   Third, we claim that $(f(x_k))_{k\in\mathbb{N}}$ converges to b, or written differently, $f(x_k) \to b$ as $k \to \infty$. This, by definition, means that for any $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that

$$
|f(x_k) - b| < \epsilon, \quad \text{for each} \quad k > N,
$$

or equivalently,

$$b - \epsilon < f(x_k) < b + \epsilon, \quad \text{for each} \quad k > N.$$

Indeed, let $\epsilon > 0$ be arbitrary. Since b is the greates lower bound of $\{f(x_k) : k \in \mathbb{N}\}$, we get that

$$b - \epsilon < b \leq f(x_k), \quad \forall k \in \mathbb{N}.$$

Moreover, there exists an $N \in \mathbb{N}$ 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 $\{f(x_k) : k \in \mathbb{N}\}$. But this would contradict the fact that b is the greates lower bound of $\{f(x_k) : k \in \mathbb{N}\}$, since $b < b + \epsilon$. Finally, note that

$$f(x_k) \leq f(x_N) < b + \epsilon, \quad \text{for each} \quad k > N,$$

by construction of the sequence $(x_k)_{k \in \mathbb{N}}$. I.e. we established that

$$b - \epsilon < f(x_k) < b + \epsilon, \quad \text{for each} \quad k > N,$$

as desired.

3. The most basic example would be to consider any function f that is bounded from below and let $x_k = x$ for each $k \in \mathbb{N}$, where $x \in \mathbb{R}^n$ is not the minimum. A slightly more interesting example would be $f(x, y) = x^2 + y^2$ and the sequence

$$(x_k, y_k) = \left(\sin(k)(1 + \tfrac{1}{k}), \cos(k)(1 + \tfrac{1}{k})\right), \quad \forall k \in \mathbb{N},$$

for which $f(x_k, y_k) = (1 + \frac{1}{k})^2$ 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

$$\nabla f(x) = e^x - 2 + 2x$$
$$\nabla^2 f(x) = e^x + 2,$$

which shows that f is strictly convex and thus has a unique minimizer. The Newton iteration is then explicitly written as

$$x_{k+1} = x_k - \frac{e^{x_k} - 2 + 2x_k}{e^{x_k} + 2}.$$

Solution 6.6

1. Note that

$$0 \le Q_k \le \frac{V}{\psi_1(k)} + \frac{D}{\psi_2(k)} \to 0 \quad \text{as} \quad k \to \infty.$$

   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 the following

   (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. We get that

$$0 \le f(x_k) - f(x^\star) \le \frac{V + D\sum_{i=0}^{k} \gamma_i^2}{b\sum_{i=0}^{k} \gamma_i} \le \frac{V + D\sum_{i=0}^{\infty} \gamma_i^2}{b\sum_{i=0}^{k} \gamma_i} \to 0 \quad \text{as} \quad k \to \infty.$$

   By the squeeze theorem, we have that $f(x_k) - f(x^\star) \to 0$ as $k \to \infty$, and therefore, $f(x_k) \to f(x^\star)$ as $k \to \infty$.

2. In each case, the function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$, where $\phi(i) = \gamma_i$, is non-negative and decreasing. Therefore, we obtain the following bound:

$$0 \leq \int_0^k \phi(t)dt \leq \sum_{i=0}^k \phi(i) = \sum_{i=0}^k \gamma_i, \quad \forall k \in \mathbb{N}.$$

Similarly, we also get the bound

$$\sum_{i=0}^k \gamma_i^2 = \sum_{i=0}^k \phi(k)^2 \leq \int_0^k \phi^2(t)dt + \phi^2(0) \leq \int_0^\infty \phi^2(t)dt + \phi^2(0), \quad \forall k \in \mathbb{N}.$$

Combining these bounds with the inequality given by the convergence analysis, we get the new inequality

$$f(x_k) - f(x^\star) \leq \frac{V + D\sum_{i=0}^k \gamma_i^2}{b\sum_{i=0}^k \gamma_i} \leq \frac{V + D\left(\int_0^\infty \phi^2(t)dt + \phi^2(0)\right)}{b\int_0^k \phi(t)dt}, \quad \forall k \in \mathbb{N}.$$

(a) Now let $\phi(i) = \gamma_i = c/(i+1)$. Note that

$$\begin{aligned}\int_0^\infty \phi^2(t)dt + \phi^2(0) &= c^2\left(\int_0^\infty \frac{1}{(1+t)^2}dt + 1\right)\\ &= c^2\left(\left[\frac{-1}{1+t}\right]_{t=0}^\infty + 1\right)\\ &= 2c^2,\end{aligned}$$

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

$$f(x_k) - f(x^\star) \leq \frac{V + 2Dc^2}{bc\log(k+1)}, \quad \forall k \in \mathbb{N},$$

which shows a $O(1/\log(k))$ sublinear rate of convergence.

(b) Next, let $\phi(i) = \gamma_i = c/(i+1)^\alpha$ with $\alpha \in (0.5, 1)$. 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

$$f(x_k) - f(x^\star) \le \frac{V + D\frac{2\alpha c^2}{2\alpha-1}}{\frac{bc}{1-\alpha}\left((k+1)^{1-\alpha} - 1\right)}, \quad \forall k \in \mathbb{N},$$

which shows a $O(1/k^{1-\alpha})$ sublinear rate of convergence.

3. $\alpha \in (0.5, 1)$ implies that $1 - \alpha \in (0, 0.5)$. Note that

$$\frac{1/k^{1-\alpha}}{1/\log(k)} = \frac{\log(k)}{k^{1-\alpha}} \to 0 \quad \text{as } k \to \infty.$$

Thus, step-size (b) gives the fastest convergence rate.

Solution 6.8

Let $V_k = \|x_k - x^\star\|_2^2$ for each $k \in \mathbb{N}$. The Lyapunov inequality can then be written as

$$V_k \le V_{k-1} - 2\gamma(f(x_k) - f(x^\star)), \quad \forall k \in \mathbb{N} \setminus \{0\}.$$

Recursively applying the inequality gives

$$V_k \le V_0 - 2\gamma \sum_{i=1}^{k} (f(x_i) - f(x^\star)), \quad \forall k \in \mathbb{N} \setminus \{0\},$$

and therefore

$$\sum_{i=1}^{k} (f(x_i) - f(x^\star)) \le \frac{V_0 - V_k}{2\gamma} \le \frac{V_0}{2\gamma}, \quad \forall k \in \mathbb{N} \setminus \{0\},$$

since $V_k \ge 0$ for each $k \in \mathbb{N}$. Then

$$k(f(x_k) - f(x^\star)) \le \sum_{i=1}^{k} (f(x_i) - f(x^\star)) \le \frac{V_0}{2\gamma}, \quad \forall k \in \mathbb{N} \setminus \{0\},$$

since $(x_k)_{k\in\mathbb{N}}$ is a descent sequence for f. Finally, note that

$$0 \le f(x_k) - f(x^\star) \le \frac{V_0}{2\gamma k} \to 0 \quad \text{as } k \to \infty.$$

By the squeeze theorem, we have that $f(x_k) - f(x^\star) \to 0$ as $k \to \infty$, and therefore, $f(x_k) \to f(x^\star)$ as $k \to \infty$. Moreover, we identify a $O(1/k)$ sublinear rate of convergence.

Solution 6.9

1. We start from the inequality

$$\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 \mid x_k\right] \le \|x_k - x^\star\|_2^2 - 2\gamma_k(f(x_k) - f(x^\star)) + \gamma_k^2 G^2, \quad \forall k \in \mathbb{N}.$$

By monotonicity and linearity of expectation, we get that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 \mid x_k\right]\right] &\le \mathbb{E}\left[\|x_k - x^\star\|_2^2 - 2\gamma_k(f(x_k) - f(x^\star)) + \gamma_k^2 G^2\right] \\ &= \mathbb{E}\left[\|x_k - x^\star\|_2^2\right] - \mathbb{E}\left[2\gamma_k(f(x_k) - f(x^\star))\right] + \mathbb{E}\left[\gamma_k^2 G^2\right] \\ &= \mathbb{E}\left[\|x_k - x^\star\|_2^2\right] - 2\gamma_k\,\mathbb{E}\left[f(x_k) - f(x^\star)\right] + \gamma_k^2 G^2,\end{aligned}$$

holds for each $k \in \mathbb{N}$. The law of total expectation yields

$$\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2\right] \le \mathbb{E}\left[\|x_k - x^\star\|_2^2\right] - 2\gamma_k\,\mathbb{E}\left[f(x_k) - f(x^\star)\right] + \gamma_k^2 G^2, \quad \forall k \in \mathbb{N}.$$

This is the Lyapunov inequality we pick.

2. Recursively applying the Lyapunov inequality above gives

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2\right] &\le \mathbb{E}\left[\|x_0 - x^\star\|_2^2\right] - 2\sum_{i=0}^{k}\gamma_i\,\mathbb{E}\left[f(x_i) - f(x^\star)\right] + G^2\sum_{i=0}^{k}\gamma_i^2 \\ &= \|x_0 - x^\star\|_2^2 - 2\sum_{i=0}^{k}\gamma_i\,\mathbb{E}\left[f(x_i) - f(x^\star)\right] + G^2\sum_{i=0}^{k}\gamma_i^2, \quad \forall k \in \mathbb{N},\end{aligned}$$

since $\|x_0 - x^\star\|_2^2$ is deterministic. Again, by monotonicity of expectation, we know that

$$0 \le \mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2\right], \quad \forall k \in \mathbb{N},$$

since

$$0 \le \|x_{k+1} - x^\star\|_2^2, \quad \forall k \in \mathbb{N}.$$

We conclude that

$$0 \le \|x_0 - x^\star\|_2^2 - 2\sum_{i=0}^{k}\gamma_i\,\mathbb{E}\left[f(x_i) - f(x^\star)\right] + G^2\sum_{i=0}^{k}\gamma_i^2, \quad \forall k \in \mathbb{N},$$

or by rearranging

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

as desired.

Solution 6.10

1. First, we prove the claim provided in the hint, i.e.
$$\lambda_k \geq 1 + 0.5k, \quad \forall k \in \mathbb{N}. \tag{7.17}$$
Clearly, (7.17) holds for $k = 0$. Note that $\lambda_k = 0.5 + \sqrt{0.25 + \lambda_{k-1}^2} \geq 0.5 + \lambda_{k-1}$ holds for each $k \in \mathbb{N} \setminus \{0\}$. Recursively applying this gives
$$\lambda_k \geq 0.5k + \lambda_0 = 1 + 0.5k, \quad \forall k \in \mathbb{N} \setminus \{0\}.$$
This establishes (7.17).

   Next, rearranging (6.1) and recursive application gives
$$\begin{aligned}
\tfrac{2\lambda_{k+1}^2}{\beta}(f(x_{k+1}) - f(x^\star)) &\leq V_k - V_{k+1} + \tfrac{2\lambda_k^2}{\beta}(f(x_k) - f(x^\star)) \\
&\leq V_1 - V_{k+1} + \tfrac{2\lambda_0^2}{\beta}(f(x_1) - f(x^\star)) \\
&= V_1 - V_{k+1} + \tfrac{2}{\beta}(f(x_1) - f(x^\star)) \\
&\leq V_1 + \tfrac{2}{\beta}(f(x_1) - f(x^\star)), \quad \forall k \in \mathbb{N} \setminus \{0\},
\end{aligned}$$
since $V_k \geq 0$ for each $k \in \mathbb{N} \setminus \{0\}$. Using (7.17), we get that
$$\begin{aligned}
f(x_{k+1}) - f(x^\star) &\leq \frac{V_1 + \frac{2}{\beta}(f(x_1) - f(x^\star))}{\frac{2\lambda_{k+1}^2}{\beta}} \\
&\leq \frac{\beta V_1 + 2(f(x_1) - f(x^\star))}{2(1 + 0.5(k+1))^2}, \quad \forall k \in \mathbb{N} \setminus \{0\},
\end{aligned}$$
or equivalently
$$f(x_k) - f(x^\star) \leq \frac{2\beta V_1 + 4(f(x_1) - f(x^\star))}{(k+2)^2}, \quad \forall k \in \mathbb{N} \setminus \{0, 1\}. \tag{7.18}$$
Note that
$$0 \leq f(x_k) - f(x^\star) \leq \frac{2\beta V_1 + 4(f(x_1) - f(x^\star))}{(k+2)^2} \to 0 \quad \text{as} \quad k \to \infty.$$
By the squeeze theorem, we have that $f(x_k) - f(x^\star) \to 0$ as $k \to \infty$, and therefore, $f(x_k) \to f(x^\star)$ as $k \to \infty$. Moreover, we identify a $O(1/k^2)$ sublinear rate of convergence.

2. From (7.18), if $k \geq 2$, we know that
$$f(x_k) - f(x^\star) \leq \frac{2\beta V_1 + 4(f(x_1) - f(x^\star))}{(k+2)^2}.$$
Therefore, if the integer $k \geq 2$ is so large such that
$$\frac{2\beta V_1 + 4(f(x_1) - f(x^\star))}{(k+2)^2} \leq \epsilon,$$
we obtain an ϵ-accurate objective value. This is equivalently to
$$k \geq \left\lceil \sqrt{\frac{2\beta V_1 + 4(f(x_1) - f(x^\star))}{\epsilon}} - 2 \right\rceil \quad \text{and} \quad k \geq 2,$$
or simply
$$k \geq \max\left(\left\lceil \sqrt{\frac{2\beta V_1 + 4(f(x_1) - f(x^\star))}{\epsilon}} - 2 \right\rceil, 2\right).$$

Solutions to Chapter 7

Solution 7.1

1. Since f is smooth use the descent lemma at x_k and x_{k+1},

$$f(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta}{2}\|x_{k+1} - x_k\|_2^2.$$

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

$$\begin{aligned} f(x_{k+1}) &\le f(x_k) + \nabla f(x_k)^T(-\gamma \nabla f(x_k)) + \tfrac{\beta}{2}\| - \gamma \nabla f(x_k)\|_2^2 \\ &= f(x_k) - \gamma(1 - \tfrac{\beta}{2}\gamma)\|\nabla f(x_k)\|_2^2. \end{aligned}$$

Now subtracting $f(x^\star)$ from both sides yields

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

2. We see that the Lyapunov inequality has a telescoping term in V_k. Sum over the inequalities for indices $i = 0, \dots, k$,

$$\begin{aligned} V^1 &\le V_0 - \gamma(1 - \tfrac{\beta}{2}\gamma)\|\nabla f(x_0)\|_2^2 \\ V^2 &\le V^1 - \gamma(1 - \tfrac{\beta}{2}\gamma)\|\nabla f(x_1)\|_2^2 \\ &\vdots \\ V_{k+1} &\le V_k - \gamma(1 - \tfrac{\beta}{2}\gamma)\|\nabla f(x_k)\|_2^2, \end{aligned}$$

gives

$$\gamma(1 - \tfrac{\beta}{2}\gamma)\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2 \le V_0 - V_{k+1}.$$

Since $x^\star$ is a minimizer of f, $V_k \ge 0$ for all k. Furthermore, if $0 < \gamma < \frac{2}{\beta}$ then $\gamma(1 - \frac{\beta}{2}\gamma) > 0$, this gives

$$\begin{aligned} V_0 &\ge V_0 - V_{k+1} \\ &\ge \gamma(1 - \tfrac{\beta}{2}\gamma)\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2 \\ &\ge \gamma(1 - \tfrac{\beta}{2}\gamma)(k+1)\min_{i \in \{0,\dots,k\}}\|\nabla f(x_i)\|_2^2 \\ &\ge 0 \end{aligned}$$

Dividing both sides with $\gamma(1-\frac{\beta}{2}\gamma)(k+1)$ yield

$$0 \leq \min_{i\in\{0,\dots,k\}} \|\nabla f(x_i)\|_2^2 = q_k \leq \tfrac{2V_2}{\gamma(1-\beta\gamma)(k+1)}.$$

The sequence $\{q_k\}_{k\in\mathbb{N}}$ then clearly converge to zero as the upper bound converge to zero. The rate is easily identified as $O(1/k)$ and is sublinear.

SOLUTION 7.2

1. We start by expanding the square

$$\begin{aligned}\|x_{k+1}-x^\star\|_2^2 &= \|x_k-\gamma\nabla f(x_k)-x^\star\|_2^2\\ &= \|x_k-x^\star\|_2^2 - 2\gamma(\nabla f(x_k))^T(x_k-x^\star) + \gamma^2\|\nabla f(x_k)\|_2^2.\end{aligned}$$

Using the first order condition for convexity we can bound the inner product as

$$\begin{aligned} & f(x^\star) \geq f(x_k) + (\nabla f(x_k))^T(x^\star - x_k)\\ \iff\ & -2\gamma(f(x_k)-f(x^\star)) \geq -2\gamma(\nabla f(x_k))^T(x_k-x^\star)\end{aligned}$$

which yields

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

From Exercise 7.1 we know

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

Adding $f(x^\star)$ to both sides, multiplying by 2γ and rearranging gives

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

Inserting this back into the original inequality gives

$$\begin{aligned}&\|x_{k+1}-x^\star\|_2^2\\ &\quad\leq \|x_k-x^\star\|_2^2 - 2\gamma(f(x_{k+1})-f(x^\star)) + \gamma^2(\beta\gamma-1)\|\nabla f(x_k)\|_2^2\end{aligned}$$

2. With $V_k = \|x_k - x^\star\|_2^2$, sum over the Lyapunov inequalities for $i=0,\dots,k$,

$$\begin{aligned} V_1 &\leq V_0 - 2\gamma(f(x_1)-f(x^\star)) + \gamma^2(\beta\gamma-1)\|\nabla f(x_0)\|_2^2\\ V_2 &\leq V_1 - 2\gamma(f(x_2)-f(x^\star)) + \gamma^2(\beta\gamma-1)\|\nabla f(x_1)\|_2^2\\ &\vdots\\ V_{k+1} &\leq V_k - 2\gamma(f(x_{k+1})-f(x^\star)) + \gamma^2(\beta\gamma-1)\|\nabla f(x_k)\|_2^2,\end{aligned}$$

which gives

$$V_{k+1} \leq V_0 - 2\gamma\sum_{i=0}^{k}(f(x^{i+1})-f(x^\star)) + \gamma^2(\beta\gamma-1)\sum_{i=1}^{k}\|\nabla f(x_i)\|_2^2$$

Rearranging and using $\gamma < \frac{2}{\beta}$ and $V_k \geq 0$ gives

$$\begin{aligned}2\gamma \sum_{i=0}^{k} (f(x^{i+1}) - f(x^\star)) &\leq V_0 - V_{k+1} + \gamma^2(\beta\gamma - 1) \sum_{i=1}^{k} \|\nabla f(x_i)\|_2^2 \\ &\leq V_0 + \gamma^2 \sum_{i=1}^{k} \|\nabla f(x_i)\|_2^2\end{aligned}$$

From Exercise 7.1 we know that $\gamma^2 \sum_{i=1}^{k} \|\nabla f(x_i)\|_2^2$ is bounded, way by $\overline{W}$, and that $f(x_k)$ is decreasing,

$$\begin{aligned}V_0 + \overline{W} &\geq V_0 + \gamma^2 \sum_{i=1}^{k} \|\nabla f(x_i)\|_2^2 \\ &\geq 2\gamma \sum_{i=0}^{k} (f(x^{i+1}) - f(x^\star)) \\ &\geq 2\gamma(k+1)(f(x^{k+1}) - f(x^\star)) \\ &\geq 0\end{aligned}$$

Dividing by $2\gamma(k+1)$ yield

$$0 \leq f(x^{k+1}) - f(x^\star) \leq \frac{V_0 + \overline{W}}{2\gamma(k+1)}.$$

This shows that $f(x_k) \to f(x^\star)$ with $O(1/k)$ sublinear rate.

Solution 7.3

1. We start by expanding the square

$$\begin{aligned}\|x_{k+1} - x^\star\|_2^2 &= \|x_k - \gamma\nabla f(x_k) - x^\star\|_2^2 \\ &= \|x_k - x^\star\|_2^2 - 2\gamma(\nabla f(x_k))^T(x_k - x^\star) + \gamma^2\|\nabla f(x_k)\|_2^2.\end{aligned}$$

Using the first order condition for strong convexity we can bound the inner product

$$\begin{aligned}& f(x^\star) \geq f(x_k) + (\nabla f(x_k))^T(x^\star - x_k) + \tfrac{\mu}{2}\|x^\star - x_k\|_2^2 \\ \iff & -2\gamma(f(x_k) - f(x^\star)) - \mu\gamma\|x_k - x^\star\|_2^2 \geq -2\gamma(\nabla f(x_k))^T(x_k - x^\star)\end{aligned}$$

which yields

$$\|x_{k+1} - x^\star\|_2^2 \leq (1 - \mu\gamma)\|x_k - x^\star\|_2^2 - 2\gamma(f(x_k) - f(x^\star)) + \gamma^2\|\nabla f(x_k)\|_2^2.$$

From Exercise 7.2 we know

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

Inserting this back into the original inequality gives

$$\begin{aligned}&\|x_{k+1} - x^\star\|_2^2 \\ &\quad \leq (1-\mu\gamma)\|x_k - x^\star\|_2^2 - 2\gamma(f(x_{k+1}) - f(x^\star)) + \gamma^2(\beta\gamma - 1)\|\nabla f(x_k)\|_2^2.\end{aligned}$$

Since $\gamma \leq \frac{1}{\beta}$ the last term is non-positive. The second to last term also non-positive since $\gamma > 0$ and $f(x_{k+1}) \geq f(x^\star)$ since $x^\star$ is optimal. This yield the desired inequality,

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

The rate is maximized when $1-\mu\gamma$ is minimized, i.e., when γ is maximized. The best step-size is therefore $\gamma = \frac{1}{\beta}$.

2. The best step-size minimize $h(\gamma) = \max(1-\mu\gamma, \beta\gamma - 1)$. Since h is the maximum of two affine functions it is closed and convex. Fermat's rule then give that the best step-size γ satisfies

$$0 \in \partial g(\gamma) = \begin{cases} -\mu & \text{if } 1-\mu\gamma > \beta\gamma - 1 \\ \beta & \text{if } 1-\mu\gamma < \beta\gamma - 1 \\ [-\mu, \beta] & \text{if } 1-\mu\gamma = \beta\gamma - 1 \end{cases}$$

This can clearly only hold when $1-\mu\gamma = \beta\gamma - 1$, i.e., when $\gamma = \frac{2}{\beta+\mu}$.

From the analysis earlier in this exercise we got that the best rate convergence was given by

$$\|x_{k+1} - x^\star\|_2^2 \leq (1-\tfrac{\mu}{\beta})\|x_k - x^\star\|_2^2.$$

From the analysis in the lectures we get that the best rate is given by

$$\|x_{k+1} - x^\star\|_2^2 \leq (1-\tfrac{2\mu}{\beta+\mu})^2\|x_k - x^\star\|_2^2.$$

Since 0 < $\mu \leq \beta$ we have

$$(1-\tfrac{2\mu}{\beta+\mu})^2 < 1-\tfrac{2\mu}{\beta+\mu} \leq 1-\tfrac{2\mu}{\beta+\beta} = 1-\tfrac{\mu}{\beta}$$

and hence does the convergence analysis in the lecture yields a faster rate.

Solution 7.4

1. We have

$$x^{k+1} = x^k - \gamma\nabla f(x^k) = x^k - \gamma Q x^k - \gamma q = (I-\gamma Q)x^k - \gamma q.$$

If x^* is a solution, then $0 = \nabla f(x^*)$ i.e.

$$x^* = x^* - \gamma\nabla f(x^*) = x^* - \gamma Q x^* - \gamma q = (I-\gamma Q)x^* - \gamma q$$

so

$$x^{k+1} - x^* = (I-\gamma Q)x^k - (I-\gamma Q)x^* = (I-\gamma Q)(x^k - x^*)$$

we therefore get $\|x^{k+1} - x^*\| = \|(I-\gamma Q)(x^k - x^*)\| \leq \|I-\gamma Q\|\|x^k - x^*\|$ Let $\lambda(M)$ be the set of eigenvalues for a matrix M. Then $0 < \lambda(Q) \leq \lambda_{\max}(Q)$, and since $L = \lambda_{\max}(Q)$ we get $\gamma \in (0, 2/L) = (0, 2/\lambda_{\max}(Q))$, which means that $\lambda(\gamma Q) \in (0,2)$, and lastly $\lambda(I-\gamma Q) \in (-1,1)$. Hence, $0 \leq \|I-\gamma Q\| < 1$

2. With $\gamma = 1/L$ we get $\lambda(\gamma Q) \in (0,1)$ and $\lambda(I - \gamma Q) \in (0,1)$. The eigenvalue of $I - \gamma Q$ with the largest absolute value therefore corresponds to the smallest eigenvalue of γQ, i.e. $\lambda_{\text{min}}(Q)/L$, where $L = \lambda_{\text{max}}(Q)$. The convergence rate therefore becomes $r = 1 - \lambda_{\text{min}}(Q)/\lambda_{\text{max}}(Q)$, where $\lambda_{\text{min}}(Q)/\lambda_{\text{max}}(Q)$ is known as the condition number.

3. The eigenvalues are 1 and ϵ. $L = 1$, so the eigenvalues of $I - \gamma Q$ are 0 and $1 - \epsilon$ with the rate set by $r = 1 - \epsilon$. When $q = 0$ then $x^* = 0$. If we let $x^0 = [1\ \ 0]^T$ then $x^k = [(1-\epsilon)^k\ \ 0]^T$ and the rate is achieved.

4. Let $V = \begin{bmatrix} 1/\sqrt{\epsilon} & 0 \\ 0 & 1 \end{bmatrix}$, we then get $V^T Q V = \begin{bmatrix} 1 & 0.01 \\ 0.01 & 1 \end{bmatrix}$ which has eigenvalues 0.99 and 1.01. The convergence will therefore be very fast. With $\gamma = 1/L = 1/1.01$ we get $r \approx 0.02$.

5. The prox if often computed on some function $g(x)$ that is separable. With a change of variables to $x = Vy$, we need to prox on the function $g(Vy)$ which is no longer separable, and computing the prox on this term generally becomes computationally expensive.

Solution 7.5

1. Let $x^+ = \text{prox}_{\gamma f}(x) = \text{argmin}_z(f(z) + \frac{1}{2\gamma}\|x - z\|_2^2)$. Therefore, for all z, it holds that

$$f(x^+) + \tfrac{1}{2\gamma}\|x^+ - x\|_2^2 \leq f(z) + \tfrac{1}{2\gamma}\|z - x\|_2^2.$$

Set in particular $z = x$ to get

$$f(x^+) + \tfrac{1}{2\gamma}\|x_+ - x\|_2^2 \leq f(x).$$

2. We have

$$\tfrac{1}{2\gamma}\|x^{k+1} - x^k\|_2^2 \leq f(x^k) - f(x^{k+1}).$$

Summing this inequality gives for all $n \in \mathbb{N}$:

$$\tfrac{1}{2\gamma}\sum_{k=0}^{n}\|x^{k+1} - x^k\|_2^2 \leq f(x^0) - f(x^{n+1}) \leq f(x^0) - B.$$

Letting $n \to \infty$ means that $\frac{1}{2\gamma}\sum_{k=0}^{\infty}\|x^{k+1} - x^k\|_2^2 < \infty$ and $\|x^{k+1} - x^k\| \to 0$ as $k \to \infty$.

3. From convexity of f and Fermat's rule we have that $x^+ = \text{prox}_{\gamma f}(x) = \text{argmin}_z(f(z) + \frac{1}{2\gamma}\|x - z\|_2^2)$ is equivalent to

$$0 \in \partial f(x^+) + \tfrac{1}{\gamma}(x^+ - x) \iff \tfrac{1}{\gamma}(x - x^+) \in \partial f(x^+).$$

This means that $\frac{1}{\gamma}(x^{k-1} - x^k) = s^k$ is a subgradient, $s^k \in \partial f(x^k)$. Then

$$0 \leq \text{dist}_{\partial f(x^k)}(0) \leq \|s^k - 0\| = \tfrac{1}{\gamma}\|x^{k-1} - x^k\| \to 0$$

4. Strong convexity means that

$$f(y) \geq f(x) + s^T(y-x) + \tfrac{\mu}{2}\|y-x\|^2$$

for all x, y and all $s \in \partial f(x)$. In particular we can take x^k with $\frac{1}{\gamma}(x^{k-1} - x^k) = s^k \in \partial f(x^k)$ and $x^\star$ with $0 \in \partial f(x^\star)$.

$$\begin{aligned} f(x^k) &\geq f(x^\star) + \tfrac{\mu}{2}\|x^k - x^\star\|^2 \\ f(x^\star) &\geq f(x^k) + (s^k)^T(x^\star - x^k) + \tfrac{\mu}{2}\|x^\star - x^k\|^2. \end{aligned}$$

Adding these two together and using Cauchy-Schwarz yields

$$\|\tfrac{1}{\gamma}(x^{k-1} - x^k)\| = \|s^k\| \geq \mu\|x^k - x^\star\|$$

which implies $x^k \to x^\star$.

Solution 7.6

The complete procedure is given below.

1. The goal is to get an inequality on the form

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

where Q_k is some non-negative convergence measure. Here we choose $V_k = \|x_k - x^\star\|_2^2$ as the Lyapunov function. We further define the residual mapping as $\mathcal{R}x_k = x_k - \mathbf{prox}_{\gamma g}(x_k - \gamma\nabla f(x_k))$. The proximal gradient update can then be written as

$$x_{k+1} = x_k - \mathcal{R}x_k \tag{7.19}$$

and we can use this to relate V_{k+1} to V_k by expanding the square,

$$\begin{aligned} V_{k+1} &= \|x_{k+1} - x^\star\|_2^2 \\ &= \|x_k - x^\star\|_2^2 - 2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2 \\ &= V_k - 2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2. \end{aligned} \tag{7.20}$$

2. The quantity $-2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2$ now needs to be bounded. We start by using (7.19) to re-write it as

$$\begin{aligned} &-2(x_k - x^\star)^T(\mathcal{R}x_k) + \|\mathcal{R}x_k\|_2^2 \\ &\quad= -2(x_k - x^\star)^T(\mathcal{R}x_k) + (\mathcal{R}x_k)^T(\mathcal{R}x_k) \\ &\quad= -2(x_k - x^\star)^T(\mathcal{R}x_k) + 2(\mathcal{R}x_k)^T(\mathcal{R}x_k) - (\mathcal{R}x_k)^T(\mathcal{R}x_k) \\ &\quad= -2(x_k - \mathcal{R}x_k - x^\star)^T(\mathcal{R}x_k) - (\mathcal{R}x_k)^T(\mathcal{R}x_k) \\ &\quad= -2(x_{k+1} - x^\star)^T(\mathcal{R}x_k) - \|\mathcal{R}x_k\|_2^2. \end{aligned} \tag{7.21}$$

3. We now turn to bounding $-2(x_{k+1} - x^\star)^T(\mathcal{R}x_k)$. From the proximal gradient update we have

$$\begin{aligned} &x_k - \gamma\nabla f(x_k) - x_k + \mathcal{R}x_k \in \gamma\partial g(x_{k+1}) \\ &\implies \gamma^{-1}\mathcal{R}x_k - \nabla f(x_k) \in \partial g(x_{k+1}). \end{aligned}$$

The definition of a subgradient then gives

$$\begin{aligned} & g(x^\star) \geq g(x_{k+1}) + (\gamma^{-1}\mathcal{R}x_k - \nabla f(x_k))^T(x^\star - x_{k+1}) \\ \implies & -2(x_{k+1} - x^\star)^T(\mathcal{R}x_k) \leq -2\gamma(g(x_{k+1}) - g(x^\star)) - 2\gamma\nabla f(x_k)^T(x_{k+1} - x^\star). \end{aligned} \tag{7.22}$$

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

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

Adding these two together and rearranging yield

$$\begin{aligned} & f(x_{k+1}) \leq f(x^\star) + \nabla f(x_k)^T(x_{k+1} - x^\star) + \tfrac{\beta}{2}\|\mathcal{R}x_k\|_2^2 \\ \implies & -2\gamma\nabla f(x_k)^T(x_{k+1} - x^\star) \leq -2\gamma(f(x_{k+1}) - f(x^\star)) + \gamma\beta\|\mathcal{R}x_k\|_2^2. \end{aligned} \tag{7.23}$$

5. Inserting (7.23) into (7.22), (7.22) into (7.21), and finally (7.21) into (7.20) yield

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

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

$$\begin{aligned} V_{k+1} &\leq V_k - (1 - \gamma\beta)\|\mathcal{R}x_k\|_2^2 - 2\gamma(g(x_{k+1}) + f(x_{k+1}) - g(x^\star) - f(x^\star)) \\ &\leq V_k - 2\gamma(g(x_{k+1}) + f(x_{k+1}) - g(x^\star) - f(x^\star)) \\ &= V_k - Q_k \end{aligned}$$

and where

$$Q_k = 2\gamma(g(x_{k+1}) + f(x_{k+1}) - g(x^\star) - f(x^\star))$$

which is non-negative since $\gamma > 0$ and $g(x_{k+1}) + f(x_{k+1}) \geq g(x^\star) + f(x^\star)$ since $x^\star$ is a minimum.

7. Since $V_k \geq 0$ and $Q_k \geq 0$ we have that $Q_k \to 0$ which implies that

$$f(x_k) + g(x_k) \to f(x^\star) + g(x^\star)$$

as $k \to \infty$.

Solution 7.7

1. Following the steps yield

$$
\begin{aligned}
\|x_{k+1}-x^\star\|_2^2 &= \|\mathbf{prox}_{\gamma g}(x_k-\gamma\nabla f(x_k))-x^\star\|_2^2\\
&= \|\mathbf{prox}_{\gamma g}(x_k-\gamma\nabla f(x_k))-\mathbf{prox}_{\gamma g}(x^\star-\gamma\nabla f(x^\star))\|_2^2\\
&\le \tfrac{1}{1+\mu_g\gamma}\|x_k-\gamma\nabla f(x_k)-x^\star+\gamma\nabla f(x^\star)\|_2^2\\
&= \tfrac{1}{1+\mu_g\gamma}\Big(\|x_k-x^\star\|_2^2\\
&\quad -2\gamma(\nabla f(x_k)-\nabla f(x^\star))^T(x_k-x^\star)+\gamma^2\|\nabla f(x_k)-\nabla f(x^\star)\|_2^2\Big)\\
&\le \tfrac{1}{1+\mu_g\gamma}\Big((1-\tfrac{2\beta\mu_f\gamma}{\beta+\mu_f})\|x_k-x^\star\|_2^2-\gamma(\tfrac{2}{\beta+\mu_f}-\gamma)\|\nabla f(x_k)-\nabla f(x^\star)\|_2^2\Big).
\end{aligned}
$$

If $0<\gamma\le\frac{2}{\beta+\mu_f}$, then the last term is negative and we can use

$$\|\nabla f(x_k)-\nabla f(x^\star)\|_2\ge\mu_f\|x_k-x^\star\|_2$$

to get

$$
\begin{aligned}
\|x_{k+1}-x^\star\|_2^2 &\le \tfrac{1}{1+\mu_g\gamma}\Big((1-\tfrac{2\beta\mu_f\gamma}{\beta+\mu_f})\|x_k-x^\star\|_2^2-\gamma(\tfrac{2}{\beta+\mu_f}-\gamma)\mu_f^2\|x_k-x^\star\|_2^2\Big)\\
&= \tfrac{1}{1+\mu_g\gamma}(1-\tfrac{2\beta\mu_f\gamma}{\beta+\mu_f}-\tfrac{2\mu_f^2\gamma}{\beta+\mu_f}+\gamma^2\mu_f^2)\|x_k-x^\star\|_2^2\\
&= \tfrac{1}{1+\mu_g\gamma}(1-\tfrac{2\beta\mu_f\gamma+2\mu_f^2\gamma}{\beta+\mu_f}+\gamma^2\mu_f^2)\|x_k-x^\star\|_2^2\\
&= \tfrac{1}{1+\mu_g\gamma}(1-2\mu_f\gamma+\gamma^2\mu_f^2)\|x_k-x^\star\|_2^2\\
&= \tfrac{(1-\mu_f\gamma)^2}{1+\mu_g\gamma}\|x_k-x^\star\|_2^2.
\end{aligned}
$$

If $\gamma\ge\frac{2}{\beta+\mu_f}$, the last term is positive and we can use Lipschitz continuity of ∇f to get

$$
\begin{aligned}
\|x_{k+1}-x^\star\|_2^2 &\le \tfrac{1}{1+\mu_g\gamma}\Big((1-\tfrac{2\beta\mu_f\gamma}{\beta+\mu_f})\|x_k-x^\star\|_2^2-\gamma(\tfrac{2}{\beta+\mu_f}-\gamma)\beta^2\|x_k-x^\star\|_2^2\Big)\\
&\le \tfrac{1}{1+\mu_g\gamma}(1-\tfrac{2\beta\mu_f\gamma}{\beta+\mu_f}-\tfrac{2\beta^2\gamma}{\beta+\mu_f}+\beta^2\gamma^2)\|x_k-x^\star\|_2^2\\
&\le \tfrac{1}{1+\mu_g\gamma}(1-\tfrac{2\beta\mu_f\gamma+2\beta^2\gamma}{\beta+\mu_f}+\beta^2\gamma^2)\|x_k-x^\star\|_2^2\\
&\le \tfrac{1}{1+\mu_g\gamma}(1-2\beta\gamma+\beta^2\gamma^2)\|x_k-x^\star\|_2^2\\
&\le \tfrac{(\beta\gamma-1)^2}{1+\mu_g\gamma}\|x_k-x^\star\|_2^2.
\end{aligned}
$$

To write these on one common form we use the fact that

$$\max(1-\mu_f\gamma,\beta\gamma-1)=\begin{cases}1-\mu_f\gamma & \text{if } \gamma\in[0,\frac{2}{\beta+\mu_f}]\\ \beta\gamma-1 & \text{if } \gamma\in[\frac{2}{\beta+\mu_f},\infty)\end{cases}$$

which gives the searched for inequality,

$$\|x_{k+1}-x^\star\|_2^2\le\tfrac{\max(1-\mu_f\gamma,\beta\gamma-1)^2}{1+\mu_g\gamma}\|x_k-x^\star\|_2^2.$$

2. The algorithm converge, $\|x_k-x^\star\|\to 0$, if $\frac{\max(1-\mu_f\gamma,\beta\gamma-1)^2}{1+\mu_g\gamma}<1$. First we assume $\mu_f>0$, this can be verified to hold if $0<\gamma<\frac{2}{\beta+\mu_f}$ so we only look

at the case when $\gamma \geq \frac{2}{\beta+\mu_f}$. This results in

$$\begin{aligned}\tfrac{(\beta\gamma-1)^2}{1+\mu_g\gamma} < 1 &\iff (\beta\gamma-1)^2 < 1+\mu_g\gamma \\ &\iff 1+\beta^2\gamma^2-2\beta\gamma < 1+\mu_g\gamma \\ &\iff \beta^2\gamma^2-\gamma\beta(2+\tfrac{\mu_g}{\beta}) < 0 \\ &\iff \beta\gamma < 2+\tfrac{\mu_g}{\beta} \\ &\iff \gamma < \tfrac{2}{\beta}+\tfrac{\mu_g}{\beta^2}.\end{aligned}$$

If $\mu_f > 0$, the algorithm converge if $0 < \gamma < \frac{2}{\beta} + \frac{\mu_g}{\beta^2}$.

Assume $\mu_f = 0$, this means $\frac{\max(1,\beta\gamma-1)^2}{1+\mu_g\gamma} < 1$ which can't hold if $\mu_g = 0$. Hence, the algorithm can't converge linearly if $\mu_f = \mu_g = 0$.

Assume $\mu_f = 0$ and $\mu_g > 0$, if $0 < \gamma \leq \frac{2}{\beta}$ then $\frac{\max(1-\mu_f\gamma,\beta\gamma-1)^2}{1+\mu_g\gamma} < 1$ is guaranteed to hold. If $\gamma > \frac{2}{\beta}$ the condition once again reads

$$\tfrac{(\beta\gamma-1)^2}{1+\mu_g\gamma} < 1 \iff \gamma < \tfrac{2}{\beta}+\tfrac{\mu_g}{\beta^2}.$$

To summarize, the algorithm converge if $0 < \gamma < \frac{2}{\beta} + \frac{\mu_g}{\beta^2}$ and at least one of $\mu_f > 0$ and $\mu_g > 0$ hold.

3. We first consider $\delta = 1$, this means $\beta = L + \mu$, $\mu_f = \mu$ and $\mu_g = 0$. The linear rate of $\|x_k - x^\star\|_2^2$ is then given by

$$\max(1-\mu\gamma, (L+\mu)\gamma - 1)^2.$$

In Exercise 7.3 we have already shown that this is minimized by $\gamma = \frac{2}{L+2\mu}$, resulting in the rate

$$\left(1 - \frac{2\mu}{L+2\mu}\right)^2 = \left(\frac{L}{L+2\mu}\right)^2.$$

Let's now consider $\delta = 0$ which give $\beta = L$, $\mu_f = 0$ and $\mu_g = \mu$. The rate is then given by

$$\frac{\max(1, L\gamma-1)^2}{1+\mu\gamma}.$$

We consider two cases, if $\gamma \leq \frac{2}{L}$, the rate is $\frac{1}{1+\mu\gamma}$ and hence should γ be chosen as large as possible, $\gamma = \frac{2}{L}$.

If $\frac{2}{L} \leq \gamma < \frac{2}{L} + \frac{\mu}{L^2}$ with the rate $\frac{(L\gamma-1)^2}{1+\mu\gamma}$. Taking the derivative of the rate gives

$$\begin{aligned}\frac{d}{d\gamma}\frac{(L\gamma-1)^2}{1+\mu\gamma} &= \frac{2L(L\gamma-1)}{1+\mu\gamma} - \tfrac{\mu(L\gamma-1)^2}{(1+\mu\gamma)^2} \\ &= \frac{L\gamma-1}{(1+\mu\gamma)^2}(2L(1+\mu\gamma) - \mu(L\gamma-1)) \\ &= \frac{L\gamma-1}{(1+\mu\gamma)^2}(2L + L\mu\gamma + \mu) \\ &\geq 0.\end{aligned}$$

Hence is the rate increasing in γ and the step-size should be chosen as small as possible $\gamma = \frac{2}{L}$.

The rate for when $\delta = 0$ is then

$$\frac{L}{L+2\mu} > \left(\frac{L}{L+2\mu}\right)^2$$

since $L > 0$ and $\mu > 0$. It is therefore advantageous two put the strong convexity in the gradient step.

Solution 7.8

1. L-smoothness gives

$$\begin{aligned} f(x^{k+1}) &= f(x^k - \gamma \nabla f_i(x^k)) \\ &\leq f(x^k) + \nabla F(x^k)^T(x^k - \gamma \nabla f_i(x^k) - x^k) + \tfrac{L}{2}\|x^k - \gamma \nabla f_i(x^k) - x^k\|^2 \\ &\leq f(x^k) - \gamma \nabla F(x^k)^T \nabla f_i(x^k) + \tfrac{L}{2}\gamma^2 \|\nabla f_i(x^k)\|^2. \end{aligned}$$

Taking expectation conditioned on x^k over both sides and using linearity yield

$$\begin{aligned} \mathbb{E}[f(x^{k+1})|x^k] \leq&\ f(x^k) - \gamma \nabla F(x^k)^T \mathbb{E}[\nabla f_i(x^k)|x^k] + \tfrac{L}{2}\gamma^2 \mathbb{E}[\|\nabla f_i(x^k)\|^2|x^k] \\ &= f(x^k) - \gamma \nabla F(x^k)^T \nabla F(x^k) + \tfrac{L}{2}\gamma^2 \mathbb{E}[\|\nabla f_i(x^k)\|^2|x^k] \\ &= f(x^k) - \gamma \|\nabla F(x^k)\|^2 + \tfrac{L}{2}\gamma^2 \mathbb{E}[\|\nabla f_i(x^k)\|^2|x^k]. \end{aligned}$$

Using the hint gives

$$\begin{aligned} \mathbb{E}[\|\nabla f_i(x^k)\|^2|x^k] &= \|\mathbb{E}[\nabla f_i(x^k)|x^k]\|^2 + \mathbb{E}[\|\nabla f_i(x^k) - \nabla F(x^k)\|^2|x^k] \\ &\leq \|\nabla F(x^k)\|^2 + \sigma^2. \end{aligned}$$

Inserting this into the first inequality gives the desired result.

2. Rearranging the first result

$$\gamma^k(1 - \tfrac{L}{2}\gamma^k)\|\nabla F(x^k)\|^2 - (\gamma^k)^2 \tfrac{L\sigma^2}{2} \leq \mathbb{E}[F(x^k) - F(x^{k+1})|x^k].$$

Taking total expectation yields

$$\gamma^k(1 - \tfrac{L}{2}\gamma^k)\mathbb{E}[\|\nabla F(x^k)\|^2] - (\gamma^k)^2 \tfrac{L\sigma^2}{2} \leq \mathbb{E}[F(x^k) - F(x^{k+1})].$$

Summing from $k = 0$ to $k = n$ yields

$$\begin{aligned} \sum\nolimits_{k=0}^{n} \gamma^k(1 - \tfrac{L}{2}\gamma^k)\mathbb{E}[\|\nabla F(x^k)\|^2] - (\gamma^k)^2 \tfrac{L\sigma^2}{2} &\leq \mathbb{E}[F(x^0) - F(x^{k+1})] \\ &\leq \mathbb{E}F(x^0) - B \end{aligned}$$

since $F(x) \geq B$. Letting $n \to \infty$ gives

$$\sum\nolimits_{k=0}^{\infty} \gamma^k(1 - \tfrac{L}{2}\gamma^k)\mathbb{E}[\|\nabla F(x^k)\|^2] - (\gamma^k)^2 \tfrac{L\sigma^2}{2} < \infty. \tag{7.24}$$

Inserting γ^k gives

$$\sum_{k=0}^{\infty} \tfrac{1}{2L}\mathbb{E}[\|\nabla F(x^k)\|^2 - \sigma^2] < \infty.$$

i.e. $\mathbb{E}\|\nabla F(x^k)\|^2 - \sigma^2$ must be summable and therefore must $\mathbb{E}\|\nabla F(x^k)\|^2 - \sigma^2 \to 0$. As a result we can not ensure that the gradient converge to 0 for a fixed step-size stochastic gradient descent. We only converge to a *noise ball* of size σ.

3. Inserting γ^k into (7.24) yield

$$\sum_{k=0}^{\infty} (\tfrac{1}{k} - \tfrac{L}{2}\tfrac{1}{k^2})\mathbb{E}[\|\nabla F(x^k)\|^2] - \tfrac{1}{k^2}\tfrac{L\sigma^2}{2} < \infty.$$

The $\frac{1}{k^2}\frac{L\sigma^2}{2}$ term will be summable there fore must the $(\frac{1}{k} - \frac{L}{2}\frac{1}{k^2})\mathbb{E}\|\nabla F(x^k)\|^2$ terms be summable to. For some finite C the following then hold

$$C > \sum_{k=K}^{T} (\tfrac{1}{k} - \tfrac{L}{2}\tfrac{1}{k^2})\mathbb{E}\|\nabla F(x^k)\|^2 \geq [\min_{k \leq T} \mathbb{E}\|\nabla F(x^k)\|^2] \sum_{k=K}^{T} (\tfrac{1}{k} - \tfrac{L}{2}\tfrac{1}{k^2})$$

for all $T \geq K$ where K is such that $\frac{1}{k} - \frac{L}{2}\frac{1}{k^2} > 0$ for all $k \geq K$. This give

$$0 \leq \min_{k \leq T} \mathbb{E}\|\nabla F(x^k)\|^2 \leq \frac{C}{\sum_{k=K}^{T} (\frac{1}{k} - \frac{L}{2}\frac{1}{k^2})} \to 0 \quad \text{as} \quad T \to \infty$$

since $\frac{1}{k}$ not is summable.

4. Inserting γ^k into (7.24) yield

$$\sum_{k=0}^{\infty} (\tfrac{1}{k^2} - \tfrac{L}{2}\tfrac{1}{k^4})\mathbb{E}[\|\nabla F(x^k)\|^2] - \tfrac{1}{k^4}\tfrac{L\sigma^2}{2} < \infty.$$

Once again is $\frac{1}{k^4}\frac{L\sigma^2}{2}$ summable, forcing $(\frac{1}{k^2} - \frac{L}{2}\frac{1}{k^4})\mathbb{E}[\|\nabla F(x^k)\|^2]$ to be summable to. But $(\frac{1}{k^2} - \frac{L}{2}\frac{1}{k^4}$ is here also summable, making it possible for $\mathbb{E}[\|\nabla F(x^k)\|^2 \to c^2 > 0$ without destroying summability. Clearly, having to fast decaying step-size could also hinder the convergence of the gradient.

Solution 7.9

- All methods were applicable. The gradients of the primal functions are $Qx + b$ and x. With $g(x) = \|x\|_2^2$ we have $(\text{prox}_{\gamma g}(z))_i = z_i/(1+\gamma)$.
  - One iteration of gradient descent therefore requires vector operations ($\mathcal{O}(n)$) as well as one matrix multiplications ($\mathcal{O}(n^2)$). Gradient descent therefore has complexity $\mathcal{O}(n^2)$ per iteration.
  - The gradient can be computed for each coordinate using only multiplication of one row in Q with x ($O(n)$). Per iteration complexity of coordinate gradient descent is therefore $\mathcal{O}(n)$.
  - The prox on g is separable, and the complexity for each coordinate is $O(1)$, so the complexity of the full prox is $O(n)$. The complexity of proximal gradient is therefore $O(n^2)$ and for coordinate proximal gradient it is $O(n)$.

- The following solution assumes a straight-forward implementation of the algorithms. It is possible to do some tricks to reduce the complexity of the coordiante-wise implementations, see Excercise 7.10. The gradient of the first term $f(x) = \log(1 + e^{-w^T x})$ is $\nabla f(x) = -w\frac{e^{-w^T x}}{1+e^{-w^T x}}$ and for the second term $g(x) = \frac{1}{2}\sum_i max(0, x_i)^2$ we get $(\nabla g(x))_i = \max(0, x_i)$, with the prox $(\mathbf{prox}_{\gamma g}(x))_i = \max(0, x_i/(1+\gamma))$.
  - Both gradient descent and proximal gradient will therefore have vector operations as the most costly operations, and the complexity is $O(n)$.
  - The coordinates of the gradient to f is given by $(\nabla f(x))_i = -w_i\frac{e^{-w^T x}}{1+e^{-w^T x}}$. The main cost here is the scalar product $w^T x$ which is $O(n)$. The coordinate-wise versions of the algorithms will therefore have a per iteration complexity of $O(n)$, which is the same as the full (proximal) gradient method. Iteration over all coordinates will therefore have a cost of $O(n^2)$ which means that the coordinate-wise methods are not competitive compared to the full algorithms.

Solution 7.10

At each iteration the algorithms update only one coordinate, i.e $x^{k+1} = x^k + \delta_{j_k}$ where δ_{j_k} is zero for all indices except j_k. Assume that $c^{k-1} := w^T x^{k-1}$ is already computed at iteration k. We can then calculate $c^k := w^T x^k = w^T(x^{k-1} + \delta_{j_{k-1}}) = w^T x^{k-1} + w_{j_{k-1}}\delta_{j_{k-1}} = c^{k-1} + w_{j_{k-1}}\delta_{j_{k-1}}$ using only scalar operations. The gradient of the term $f(x) = \log(1 + e^{-w^T x})$ at some index i can therefore be computed as

$$(\nabla f(x^k))_i = -w_i\frac{e^{-w^T x^k}}{1+e^{-w^T x^k}} = -w_i\frac{e^{-c^k}}{1+e^{-c^k}}$$

using only scalar operations. Since $g(x) = \sum_i g_i(x_i) = \max(0, x_i)^2$ is separable, so is the prox, i.e

$$(\mathbf{prox}_{\gamma g}(z))_i = \mathbf{prox}_{\gamma g_i}(z_i)$$

hence

$$x_i^{k+1} = (\mathbf{prox}_{\gamma g}(x^k - \gamma\nabla f(x^k)))_i = \mathbf{prox}_{\gamma g_i}((x^k - \gamma\nabla f(x^k)))_i) =$$
$$\mathbf{prox}_{\gamma g_i}(x_i^k - \gamma(\nabla f(x^k))_i) = \max(0, (x_i^k - \gamma(\nabla f(x^k))_i)/(1+\gamma)).$$

This means that if we start by computing $w^T x^0$, we are then able to do each of the following coordinate-wise updates using only scalar operations.

Solution 7.11

- We have

$$f(x) = \frac{1}{2}(Ax-b)^T(Ax-b) = \frac{1}{2}x^T A^T Ax - b^T Ax + \frac{1}{2}b^T b$$

so

$$f_{i,x}(\alpha) = \frac{1}{2}(x+\alpha e_i)^T A^T A(x+\alpha e_i) - b^T A(x+\alpha e_i) + \frac{1}{2}b^T b \tag{7.25}$$
$$= \frac{1}{2}\alpha^2 e_i^T A^T A e_i - \alpha b^T A x e_i + \dots \tag{7.26}$$

where the rest does not depend on α. We therefore get

$$\nabla f_{i,x}(\alpha) = \alpha e_i^T A^T A e_i - b^T A x e_i$$

and

$$\nabla^2 f_{i,x}(\alpha) = e_i^T A^T A e_i = (A^T A)_{i,i}$$

where $L_i = (A^T A)_{i,i}$ is the i:th diagonal element of $A^T A$.

- The Lipschitz constant of f is $\|A^T A\|_2$ and

$$\|A^T A\|_2 = \sup_x \frac{\|A^T A x\|_2}{\|x\|_2} \geq \frac{\|A^T A e_i\|_2}{\|e_i\|_2} = \|A^T A e_i\|_2$$
$$= \Big(\sum_j (A^T A)_{j,i}^2\Big)^{\frac{1}{2}} \geq \big((A^T A)_{i,i}^2\big)^{\frac{1}{2}} = (A^T A)_{i,i}.$$

Solution 7.12

The proximal update of the i:th coordinate is equivalent to

$$x_i^+ = \operatorname*{argmin}_z g_i(z) + \tfrac{1}{2\gamma}\|z - (x_i - \gamma\nabla_i f(x))\|^2$$

Due to convexity is this equivalent to

$$0 \in \partial g_i(x_i^+) + \nabla_i f(x) + \tfrac{1}{\gamma}(x_i^+ - x_i) \iff -(\nabla_i f(x) + \tfrac{1}{\gamma}(x_i^+ - x_i)) \in \partial g_i(x^+).$$

From the definition of a subgradient we get

$$g_i(x_i^+) \leq g_i(x_i) - \nabla_i f(x)^T(x_i^+ - x_i) + \tfrac{1}{\gamma}\|x_i^+ - x_i\|^2.$$

Since x^+ and x only differ in the i:th coordinate we have that $g_j(x_j^+) = g_j(x_j)$ for all $j \neq i$. This yields

$$G(x^+) \leq G(x) - \nabla_i f(x)^T(x_i^+ - x_i) + \tfrac{1}{\gamma}\|x_i^+ - x_i\|^2.$$

where $G(x) = \sum_{i=1}^n g_i(x_i)$. Furthermore, $\|x_i^+ - x_i\|^2 = \|x^+ - x\|^2$ and $\nabla_i f(x)^T(x_i^+ - x_i) = \nabla f(x)^T(x^+ - x)$ which yields

$$G(x^+) \leq G(x) - \nabla f(x)^T(x^+ - x) + \tfrac{1}{\gamma}\|x^+ - x\|^2.$$

Using L-smoothness of f yields

$$f(x^+) \leq f(x) + \nabla f(x)^T(x^+ - x) + \tfrac{L}{2}\|x^+ - x\|^2.$$

Adding these together yields

$$f(x^+) + G(x^+) \leq f(x) + G(x) - (\tfrac{1}{\gamma} - \tfrac{L}{2})\|x^+ - x\|^2.$$

which proves descent if $\gamma < \frac{2}{L}$.

Solution 7.13

The exact same reasoning as Exercise 7.12 yields

$$G(x^+) \leq G(x) - \nabla f(x)^T(x^+ - x) + \tfrac{1}{\gamma_i}\|x^+ - x\|^2.$$

The smoothness condition

$$\begin{aligned} f(x^+) &\leq f(x) + \nabla f(x)^T(x^+ - x) + \tfrac{1}{2}(x^+ - x)^T M(x^+ - x) \\ &= f(x) + \nabla f(x)^T(x^+ - x) + \tfrac{M_{ii}}{2}|x_i^+ - x_i|^2 \\ &= f(x) + \nabla f(x)^T(x^+ - x) + \tfrac{M_{ii}}{2}\|x^+ - x\|^2 \end{aligned}$$

where M_{ii} is the i:th diagonal element of M and the equalities hold since x^+ and x only differ in one coordinate. Adding the two inequalties together yields

$$f(x^+) + G(x^+) \leq f(x) + G(x) - (\tfrac{1}{\gamma_i} - \tfrac{M_{ii}}{2})\|x^+ - x\|^2.$$

and $\gamma_i < \frac{2}{M_{ii}}$ yields descent.

Solution 7.14

For implementations, see appendix. The function values are show in Figure 7.1.

We see that coordinate descent and gradient descent converge at approximately the same speed for the same amount of computations. However, by selecting a step length for each coordinate according to the individual smoothness constants as $\gamma_i = 1/(A^T A)_{i,i}$, we get considerably faster convergence.

Solution 7.15

For implementations, see appendix. The function values are show in Figure 7.2.

- We see that a larger step size will result in a quicker initial decrease of the function value. However, the error doesn't converge towards 0, and with a larger step-size the iterates will stay further away from the optimal point.
- The error keeps decreasing with this approach and we seem to get the benefit of both a large step size when we are far away, and a smaller step size when we are close to the solution. However, the convergence rate is still very slow compared to gradient descent.
- The error quickly converges (to something greater than 0) and the variance goes to 0. This is because the sequence $1/k^2$ is summable, i.e. $\sum_k \|x^{k+1} - x^k\| < c$ is bounded by some constant c, so the step lengths are not long enough to allow the iterates to go to the optimal point.

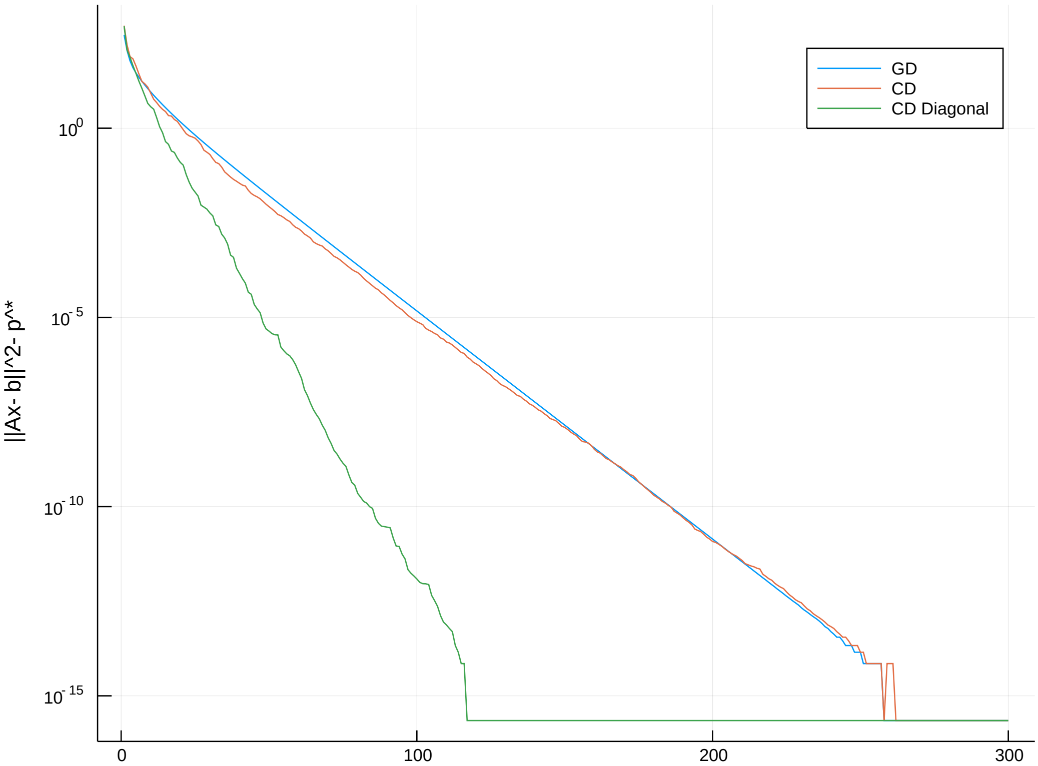


Figure 7.1: Function value for each iteration over full data with Gradient Descent, Coordinate Descent and Coordinate Descent with Diagonal scaling.

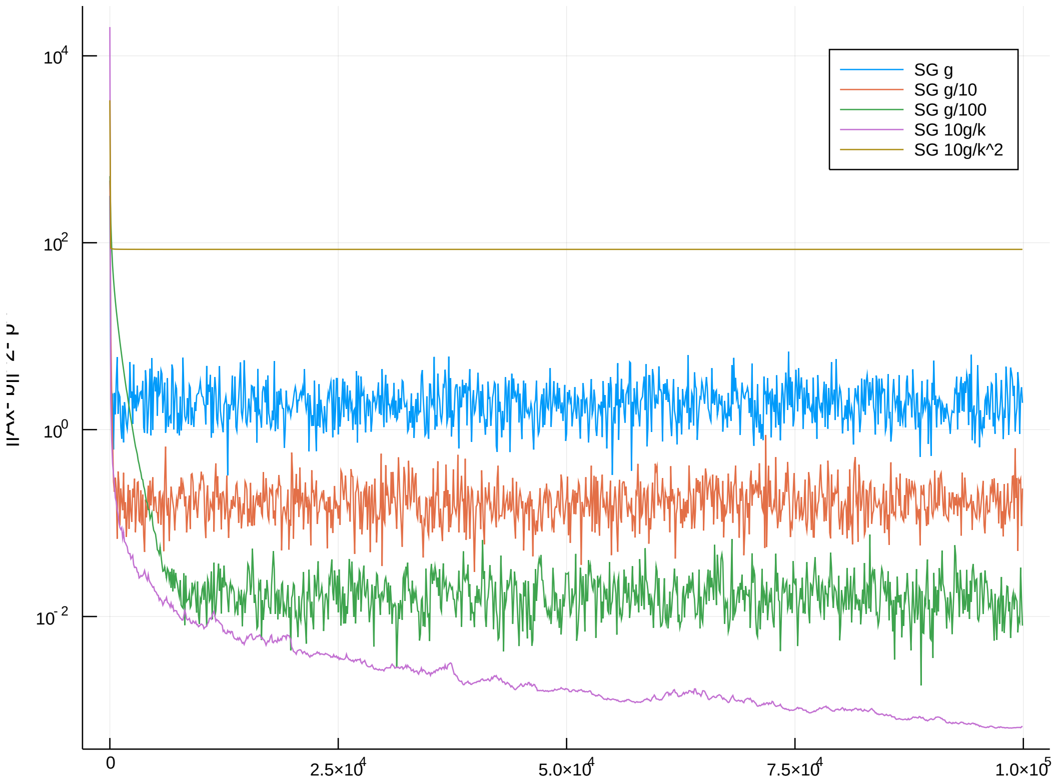


Figure 7.2: Stochastic gradient for different step lengths from Excercise 7.2, where $g = \lambda_{\max}$.

Julia Code

Implementation of Excercise 7.14

```
function grad_descent(A, b, x0, γ, kmax, xsol)
    x = copy(x0)
    res = zeros(kmax)
    err = zeros(kmax)
    AtA = A'A
    Atb = A'b
    for i = 1:kmax
        x = x .- γ.*(AtA*x .- Atb)
        res[i] = norm(A*x-b)^2
    end
    return x, res
end

coord_descent(A, b, x0, γ::Number, kmax, xsol) =
    coord_descent_efficient(A, b, x0, fill(γ, size(A,2)), kmax, xsol)

"""
    stochastic_gradient(A, b, x0, γs::AbstractArray, kmax, xsol)
    γs[i] should be γ a for index i
"""
function coord_descent(A, b, x0, γs::AbstractArray, kmax, xsol)
    n = size(A,2)

    x = copy(x0)
    res = zeros(kmax)
    err = zeros(kmax)
    # Store A*Aᵀ to avoid recomputing
    AAt = A'A
    Atb = A'b

    for i = 1:(kmax*n)
        # Random index
        j = rand(1:n)
        ∇j = view(AAt,:,j)'x - Atb[j]
        x[j] = x[j] - γs[j]*∇j
        if i%n == 0 # Every n iterations, compute error
            res[i÷n] = norm(A*x-b)^2
        end
    end
    return x, res
end
```

Implementation of Excercise 7.15

```julia
stochastic_gradient(A, b, x0, γ::Number, kmax, xsol) =
    stochastic_gradient(A, b, x0, fill(γ, kmax*size(A,1)), kmax, xsol)

"""
    stochastic_gradient(A, b, x0, γs::AbstractArray, kmax, xsol)
    γs[i] should be γ at batch i
"""
function stochastic_gradient(A, b, x0, γs::AbstractArray, kmax, xsol)
    n = size(A,1)

    x = copy(x0)
    # Only store every n iterations
    res = zeros(kmax)
    err = zeros(kmax)
    # Store Aᵀ since extracting rows is cheaper than columns
    At = copy(A')

    for i = 1:(kmax*n)
        j = rand(1:n)       # Random index
        Atj = view(At,:,j)  # For efficency, use views instead of direct index
        x .= x .- γs[(i-1)÷n+1].*Atj.*(Atj'*x - b[j])
        if i%n == 0         # Every n iterations, compute error
            res[i÷n] = norm(A*x-b)^2
        end
    end
    return x, res
end
```