Exercises in FRTN50 – Optimization for Learning

Mattias Fält, Pontus Giselsson,
Martin Morin, Hamed Sadeghi

Created: 2019
Latest update: November 4, 2019

Introduction

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

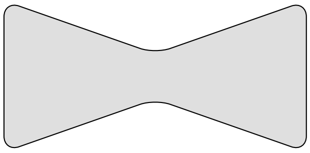
Exercises marked with (H) have hints available, listed in the end of each chapter. Not as fundamental or more difficult exercises are marked with $(\star)$.

Chapter 1

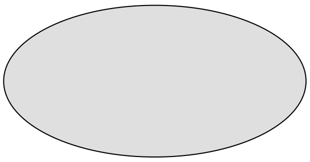
Convex Sets and Convex Functions

Exercise 1.1

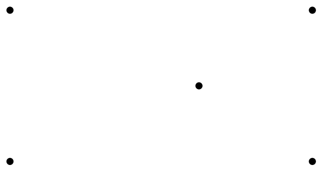
Given the following sets.



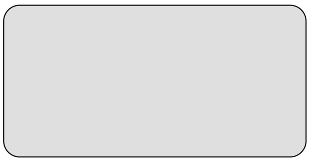
a.



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. Prove or disprove. You can 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. $S = \{x \in \mathbb{R}^n : l \leq x \leq u\}$

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. $S = \{X \in \mathbb{R}^{n \times n} : X$ positive semi-definite$\}$
10. $S = \{x \in \mathbb{R}^n : x = a\}$
11. $S = \{x \in \mathbb{R}^n : x = a$ or $x = b$ with $a \neq b\}$

Exercise 1.3

Suppose that C_1 and C_2 are convex sets.

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

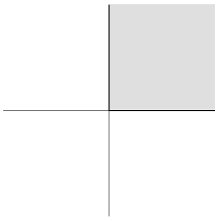
Exercise 1.4

Which of the following sets are affine?

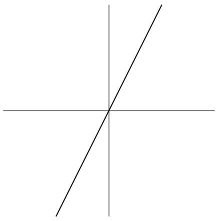
1. $V = \{x \in \mathbb{R}^n : x = a\}$
2. $V = \{x \in \mathbb{R}^n : x = \alpha a + (1 - \alpha)b, a \neq b, \alpha \in [0, 1]\}$
3. $V = \{x \in \mathbb{R}^n : x = \alpha a + (1 - \alpha)b, a \neq b, \alpha \in \mathbb{R}\}$

Exercise 1.5

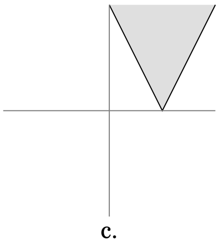
A set K is a cone if for all $x \in K$ also $\alpha x \in K$ for all $\alpha \geq 0$. Which of the following figures represent cones? Which of them are convex?



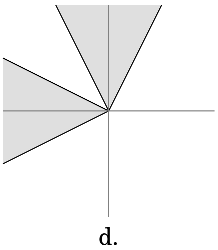
a.



b.



c.



d.

Exercise 1.6

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. $S = \{x \in \mathbb{R}^n : Ax = b \text{ with } b \neq 0\}$ with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$
3. $S = \{x \in \mathbb{R}^n : Ax \leq 0\}$ with $A \in \mathbb{R}^{m\times n}$
4. $S = \{x \in \mathbb{R}^n : Ax \leq b \text{ with } A \neq 0 \text{ and } b \neq 0\}$ with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$
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. $S = \{X \in \mathbb{R}^{n\times n} : X \text{ positive semidefinite}\}$

Exercise 1.7

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 $Q \in \mathbb{R}^{n\times n}$ positive semi-definite matrix
7. $f(x) = \text{dist}_C(x) = \inf_{y\in C}\{\|x - y\|\}$ where C is a convex set

Exercise 1.8

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. $f(x) = \sum_i(\max(0, 1 + x_i)) + \|x\|_2^2$
5. $f(x) = \sup_y(x^T y - g(y))$ (these will be called conjugate functions)

Exercise 1.9 (H)

Let $C = \{x : g(x) \leq 0\}$.

1. Suppose that $g : \mathbb{R}^n \to \mathbb{R}$ is a convex function for which $\bar{x} \in \mathbb{R}^n$ exists with $g(\bar{x}) < 0$. Show that C is a nonempty convex set.
2. Construct a *nonconvex* function $g : \mathbb{R} \to \mathbb{R}$ such that C is *convex*.
3. Construct a *nonconvex* function $g : \mathbb{R} \to \mathbb{R}$ such that C is *nonconvex*.

Exercise 1.10

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 1.11 (H)

Assume that $g_1 : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $g_2 : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are convex functions. Prove the following explicitly, without resorting to convexity preserving operations.

1. Show that $g(x) = g_1(x) + g_2(x)$ is convex
2. Show that $g(x) = \max(g_1(x), g_2(x))$ is convex

Exercise 1.12

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 1.13

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 1.14

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 1.15

Strong convexity and smoothness.

1. Show that $f(x) - \frac{\sigma}{2}\|x\|_2^2$ is convex (i.e., f is σ-strongly convex) if and only if
$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y) - \tfrac{\sigma}{2}\theta(1-\theta)\|x-y\|^2$$
for $\theta \in [0,1]$.

2. Show that $\frac{\beta}{2}\|x\|_2^2 - f(x)$ is convex (i.e., f is β-smooth) if and only if
$$f(\theta x + (1-\theta)y) \ge \theta f(x) + (1-\theta) f(y) - \tfrac{\beta}{2}\theta(1-\theta)\|x-y\|^2$$
for $\theta \in [0,1]$.

Exercise 1.16 (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 1.17

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 1.18

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 $f(x^\star) \leq f(x)$ 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 1.19

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

Exercise 1.20 (H) ($\star$)

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

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \tag{1.1}$$

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 1.21 ($\star$)

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, $x \in \mathbb{R}^n$ and define the convex set $C := \{x : Kx - b = 0\}$. 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 convex function and define the convex 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 1.22 ($\star$)

Suppose that f is convex and assume that $x^\star$ is locally optimal. That is, for all x such that $\|x - x^\star\| \leq \delta$, it satisfies $f(x^\star) \leq f(x)$. Show that $x^\star$ is a global minimum.

Exercise 1.23 ($\star$)

Jensen's inequality:

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

holds for convex functions f for all $n \geq 2$, where $\theta_i \geq 0$, and $\sum_{i=1}^n \theta_i = 1$. For $n = 2$, it reduces to the convexity definition. Prove the result for $n = 3$.

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

A function $g : \mathbb{R}^n \to \mathbb{R}$ 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 1.11

Prove that epi$g$ = epi$g_1 \cap$epig_2 in the second subproblem and conclude convexity from that.

Hint to exercise 1.16

See Exercise 1.15 for the smoothness and strong-convexity bounds.

Hint to exercise 1.20

The directional derivative at x in direction d satisfies

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

Chapter 2

Subdifferentials and Proximal Operators

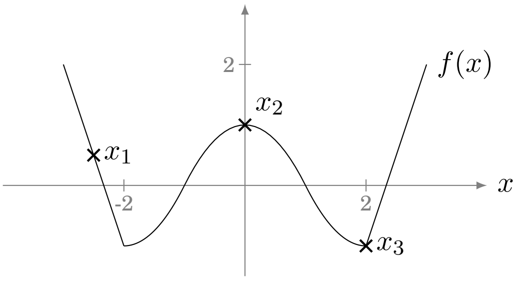
Exercise 2.1

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

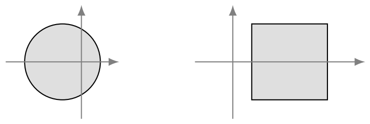
Consider the following even nonconvex function f.



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 $\partial g(y)$.



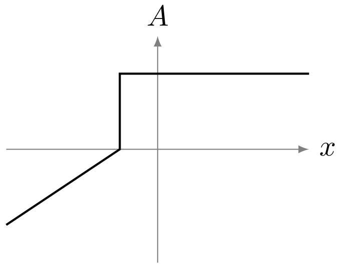
(a) (b)

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.

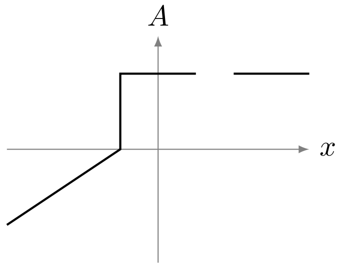
Exercise 2.4

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

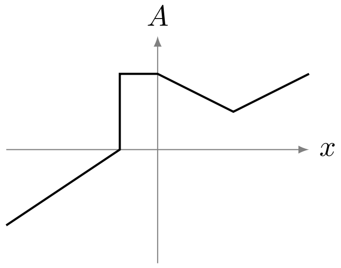
- Which are monotone?
- Which can be subdifferentials of convex functions?



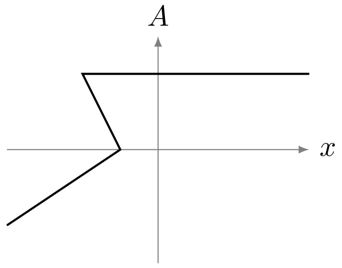
a.



b.



c.



d.

Exercise 2.5

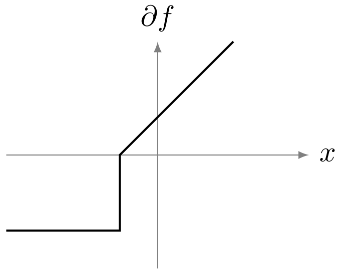
Let $\sigma : \mathbb{R} \to \mathbb{R}$ be differentiable and monotone. Prove or provide a counter-example to that the following functions are convex.

1. $f(x) = (\int \sigma(y)dy)(x)$ (means primitive function of σ).

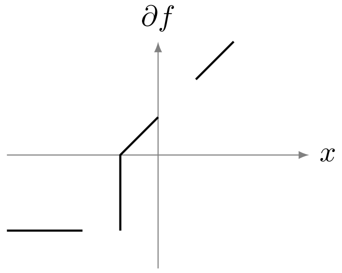
2. $f(x) = \|\sigma(x)\|_2^2$.

EXERCISE 2.6

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 2.7

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 2.8

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 $x \in \text{dom}\partial f$ and $y \in \mathbb{R}^n$ and $s \in \partial f(x)$.

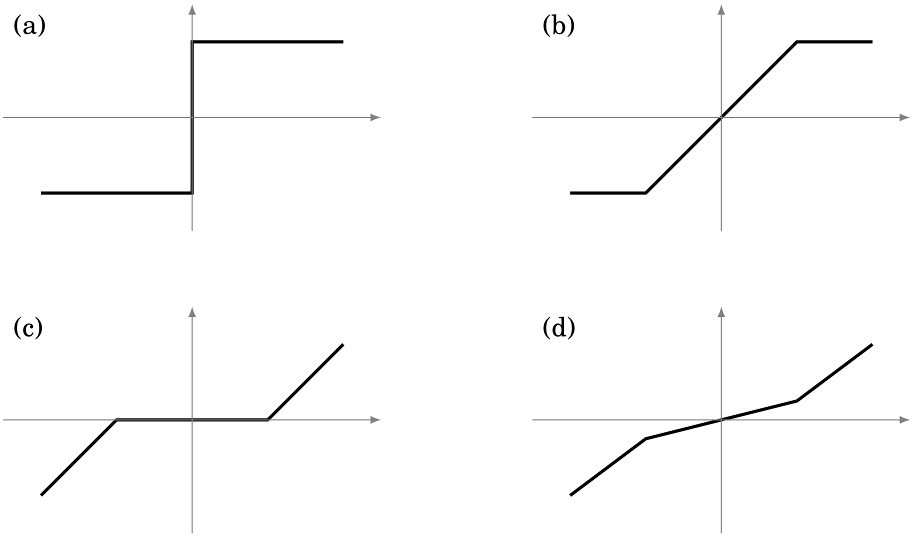
EXERCISE 2.9

1. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth, i.e., ∇f is β-Lipschitz continuous. Show that $g(x) = f(Ax + b)$ with $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^n$ is $\beta\|A\|^2$-smooth. Here, $\|A\| = \|A^T\|$ is the operator norm of A (and A^T) that satisfies $\|Ax\|_2 \leq \|A\|\|x\|_2$ for all x.

2. Suppose that $f : \mathbb{R}^n \to \mathbb{R}^n$ is σ-strongly convex. Show that $g(x) = f(Ax + b)$ is not necessarily strongly convex. (However, if A is positive definite, symmetric with all eigenvalues positive, g is strongly convex.)

Exercise 2.10

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 2.11 ($\star$)

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 2.12 ($\star$)

Assume that $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and that there exists y 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 2.13 ($\star$)

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

Exercise 2.14

Compute the proximal mapping for the following convex functions.

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

Exercise 2.15

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

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

Exercise 2.16 ($\star$)

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

Hints

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 semidefinite
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

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 3.3 (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 3.4 (H) ($\star$)

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

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 $\|s\|_2 \leq 1$.

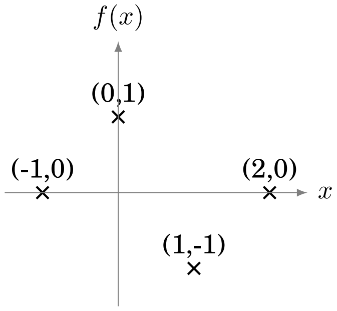
(c) Show that $f^*(s) = \infty$ for all s with $\|s\|_2 > 1$.

(d) Combine there results to state $f^*(s)$.

2. Use the conjugate to compute the subdifferential of f.

Exercise 3.5

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 3.6 ($\star$)

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 $\max(0, \max_i(s_i))$.

$$f(x) = \begin{cases} 0 & \text{if all } x_i \geq 0 \text{ and } \sum_i x_i = 1 \\ \infty & \text{else.} \end{cases}$$

Show that $f^*(s) = \max(s_i)$, i.e., the elementwise max.

Exercise 3.7

Show the following.

1. That

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

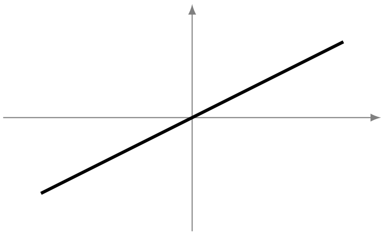
2. That the set of minimizers, $\operatorname{Argmin}_x f(x)$, for a convex function f satisfies

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

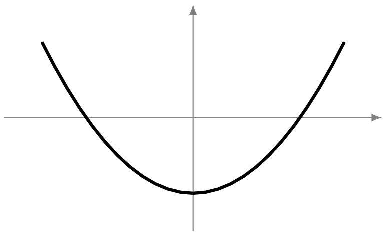
Exercise 3.8

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

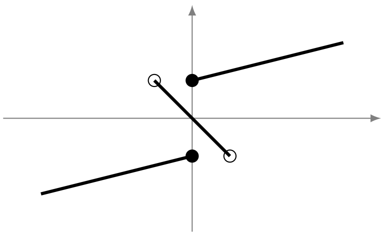
1. Draw the inverses, $A^{-1} : \mathbb{R} \to 2^{\mathbb{R}}$.
2. Which operators A are functions $f : \mathbb{R} \to \mathbb{R}$?
3. Which operator inverses A^{-1} are functions $f : \mathbb{R} \to \mathbb{R}$?



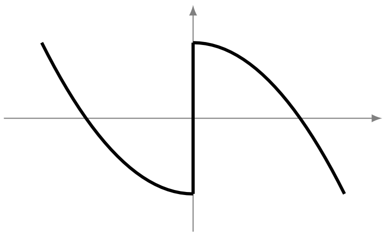
a.



b.



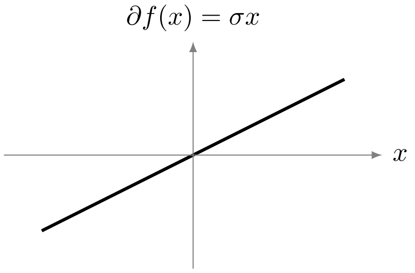
c.



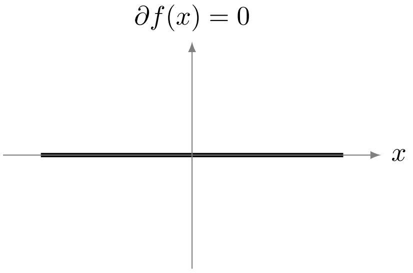
d.

Exercise 3.9

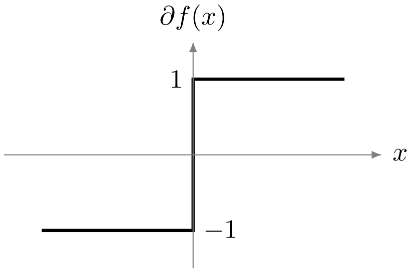
Consider the following four subdifferentials ∂f of convex functions. Decide ∂f^*, i.e., the subdifferential of the conjugate.



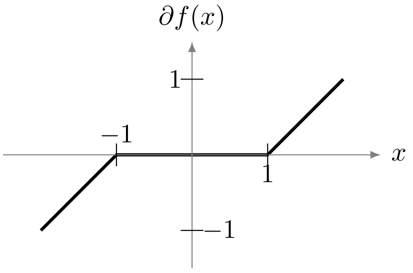
a.



b.



c.



d.

Exercise 3.10

Assume that f is convex. Show that $\text{prox}_{\gamma f}(z) = (I + \gamma\partial f)^{-1}(z)$, where the inverse means the operator inverse.

Exercise 3.11 (H)

Compute the proximal mapping for the following convex functions. 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 3.12 (H)

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

Exercise 3.13

Compute the prox_{f^*}, i.e., the prox of the conjugate, 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 3.14

Consider a primal problem the form

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

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 $\operatorname{relint}\operatorname{dom} f \cup \operatorname{relint}\operatorname{dom} g \neq \emptyset$

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

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

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

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

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 3.15 (H)

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 $L \in \mathbb{R}^{m\times n}$. Derive the dual problem

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

Exercise 3.16

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.

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 3.17 ($\star$)

Prove $f(x_0) = s_0^T x_0 - f^*(s_0) \Leftrightarrow s_0 \in \partial f(x_0)$ via the following steps.

1. Show that $f(s) + f(x) \geq s^T x$ for all x, s.

2. Suppose that $s_0 \in \partial f(x_0)$. Show that $f^*(s_0) \leq s_0^T x_0 - f(x_0)$, i.e., $f^*(s_0) = s_0^T x_0 - f(x_0)$.

3. Suppose that $f(x_0) = s_0^T x_0 - f^*(s_0)$. Show that $s_0 \in \partial f(x_0)$.

Exercise 3.18 ($\star$)

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) convex, then

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

   i.e., subdifferential of conjugate is inverse of subdifferential

$$\partial f^* = (\partial f)^{-1} \qquad \text{and} \qquad \partial f = (\partial f^*)^{-1}$$

Exercise 3.19 ($\star$)

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

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

Hints

Hint to exercise 3.3

Use the results from Exercise 3.1 and 3.2.

Hint to exercise 3.4

Cauchy-Schwarz inequality $s^T x \leq \|x\|_2 \|s\|_2$ holds for all x, s.

Hint to exercise 3.11

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

Hint to exercise 3.12

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

Hint to exercise 3.15

A very similar approach to Exercise 3.14 can be used.

Chapter 4

Learning

Exercise 4.1

Consider $f_1(x) = \frac{1}{2}\|Ax - b\|_2^2$, $f_2(x) = \|x\|_1$, $f_3(x) = \|x\|_2$, and $f_4 = \frac{1}{2}\|x\|_2^2$. For each f_i answer the following.

1. Is f_i convex?
2. Is f_i L-smooth? What is the smallest L?
3. Is f_i μ-strongly convex? What is the largest μ?

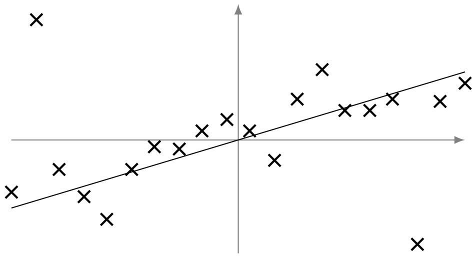
Exercise 4.2

In the same plot, sketch and compare the function $|\cdot|$ and $|\cdot|^2$. Use the results to pair the linear models plotted below with the following regression problems that generated them.

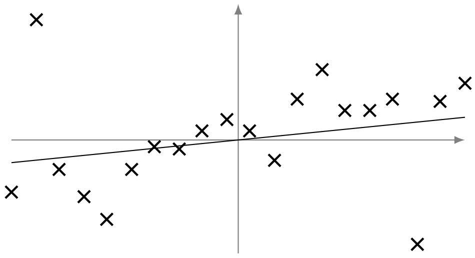
$$\text{Problem 1:} \quad \min_\theta \sum_{i=1}^n |x_i\theta - y_i| = \min_\theta \|X^T\theta - y\|_1$$

$$\text{Problem 2:} \quad \min_\theta \sum_{i=1}^n |x_i\theta - y_i|^2 = \min_\theta \|X^T\theta - y\|_2^2$$

The × markers are the data points (x_i, y_i) and the solid lines are the resulting functions $f(x) = \theta x$



a.



b.

Exercise 4.3

Consider a one dimensional regression problem with data $x_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$. We know that the dependency of y on x is periodic with length 1, i.e.

$y = f(x) = f(x+1)$ where f is the true relationship between x and y and what we wish to identify. y_i could for instance be an account balance and x_i could be time. It is reasonable to assume that the a spending/income patterns are similar from month to month. Spending depends of course on other things than just time but for simplicity we only consider the one dimensional case.

To identify the relationship between x_i and y_i we choose a parameterized model according to $y \approx m_w(x) = \phi(x)^T w$. w is the parameters and ϕ is a feature map to be determined. Periodic functions with period 1 can be written as a Fourier series

$$\tfrac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(2\pi n x) + b_n \sin(2\pi n x).$$

Use a truncated series, i.e. only use the first N terms of the sum, to design the feature map ϕ.

Exercise 4.4

Consider the Lasso problem

$$\min \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1$$

with $\lambda \geq \|A^T b\|_\infty$. Show $x = 0$ is a solution.

Exercise 4.5

Given some data points $x_i \in \mathbb{R}^n$ of some class $y_i \in \{-1, 1\}$ we model the probability of some data point x belonging to class $y = 1$ or $y = -1$ with the following logistic model.

$$P(y=1) = p_1(x) = \frac{1}{1 + e^{-(w^T x + b)}}$$
$$P(y=-1) = p_{-1}(x) = 1 - p_1(x) = \frac{1}{1 + e^{(w^T x + b)}}$$

where $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are the model parameters. With this model, the likelihood for measuring the data (x_i, y_i) for $i \in \{1, ..., N\}$ is

$$l_w(x, y) = \prod\nolimits_{i=1}^{N} p_{y_i}(x_i)$$

The model parameters w and b should be chosen such that this likelihood is maximized.

Show that the *maximum likelihood estimate* of (w, b) is given by the following logistic regression problem

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

Is this problem convex? L-smooth? μ-strongly convex?

Exercise 4.6

Consider the logistic regression problem in Exercise 4.5. Show that the problem is equivalent to

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

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

Exercise 4.7

Consider the logistic regression problem without regularization

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

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

Exercise 4.8

Consider the typical supervised learning problem

$$\min_{w} \sum\nolimits_{i=1}^{n} L(m_w(x_i), y_i)$$

where $x_i \in \mathbb{R}^d$ is the data, $y_i \in \mathbb{R}^l$ the response variable, $m_w : \mathbb{R}^d \to \mathbb{R}^l$ the parameterized model we wish to train, and $L : \mathbb{R}^d \times \mathbb{R}^l \to \mathbb{R}$ 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, $m_w(x) = w^T \phi(x)$ where $\phi : \mathbb{R}^d \to \mathbb{R}^f$.

2. $\sum_{i=1}^{n} L(m_w(x_i), y_i)$ is convex if a DNN model is used, $m_w(x) = \sigma_1(w_1^T \sigma_2(w_2^T ... \sigma_D(w_D^T x)...))$ where σ_i are some activation functions.

Exercise 4.9

Show that a kernel matrix K is positive definite, i.e. a matrix whose elements are given by

$$K_{ij} = k(x_i, x_j)$$

where $x_i, \forall i \in \{1, ..., n\}$ are elements in some input space $\mathcal{X}$ and k is a proper kernel. k is proper if it is given by an inner product in some inner product space $\mathcal{F}$, $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$ where $\phi : \mathcal{X} \to \mathcal{F}$. You may assume all things are defined over real numbers and not complex numbers.

(Note, the converse is also true. If $K \succeq 0$, then there exist some $\mathcal{F}$ and ϕ such that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$. For more information, see Mercer's theorem.)

Exercise 4.10

Consider the SVM (type) problem with bias term

$$\underset{w,b}{\text{minimize}}\, \underbrace{\mathbf{1}^T \max(0, 1 - (X^T w + b\mathbf{1}))}_{f(L(w,b))} + \underbrace{\tfrac{\lambda}{2}\|w\|_2^2}_{g(w,b)}$$

where $L = [X^T, \mathbf{1}]$.

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.

Exercise 4.11

Consider a classification problem with some data $x_i \in \mathbb{R}^p$ of some corresponding class y_i for $i \in \{1, ..., n\}$. There are two possible classes $y_i \in \{1, 2\}$.

We define a score function for each class $m_1(x) = w_1^T x$ and $m_2(x) = w_2^T x$. The idea is for the function to produce a high value if the tested data point is in the class, $m_\lambda(x_i) >> 0$ if $y_i = \lambda$, otherwise a low score, $m_\lambda(x_i) << 0$ if $y_i \neq \lambda$.

We further define the *confidence* in the score for x being of class 1 as the difference of the score for two classes, $c_1(x) = m_1(x) - m_2(x)$. Similarly we define the class 2 confidence as $c_2(x) = m_2(x) - m_1(x)$.

The parameters, w_1 and w_2, of the models can be found by minimizing the doubt (low confidence) on the known data, i.e.

$$\min_{w_1, w_2} \sum\nolimits_{i=1}^{n} \phi(c_{y_i}(x_i))$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a loss function that penalize doubt (low confidence), and $c_{y_i}(x_i)$ is the confidence of our model for x_i being the know correct class y_i, something that obviously want to be high. After solving this problem a prediction for a new unseen data point x is simply the class with the highest confidence, $\text{argmax}_{\lambda \in \{1,2\}}\, c_\lambda(x)$

- Draw/plot the following functions
  - $h(x) = \max(0, 1 - x)$ (Hinge-loss)
  - $l(x) = \log(1 + e^{-x})$ (Logistic loss)
- Show that using $\phi(x) = l(x)$ is equivalent to logistic regression, see Exercise 4.5.
- Show that using $\phi(x) = h(x)$ is equivalent to an unregularized SVM, $\min_w \sum_{i=1}^n \max(0, 1 - y_i x_i^T w)$.

- Other loss functions can of course be used. However, to also try to reward (negative loss) correct confidence/predictions can be problematic, for instance $\phi(x) = 1 - x$ can not be used, why?

Exercise 4.12

Consider the classification problem setup from Exercise 4.11 we now have K classes, $y_i \in \{1, ..., K\}$. This problem can be handled in many different ways. Arguably the simplest way is training several binary classifiers in a one-against-all or one-against-one fashion. Here we will instead look at two ways of changing the *confidence* measurement to consider multiple classes.

We define the average confidence as

$$\begin{aligned} c_\lambda^A(x) &= \tfrac{1}{K}\sum\nolimits_{k=1}^{K}(m_\lambda(x) - m_k(x)) = \tfrac{1}{K}\sum\nolimits_{k=1}^{K}(w_\lambda^T x - w_k^T x) \\ &= w_\lambda^T x - \tfrac{1}{K}\sum\nolimits_{k=1}^{K} w_k^T x \end{aligned}$$

and the worst case confidence as

$$\begin{aligned} c_\lambda^M(x) &= \min_{k \neq \lambda}(m_\lambda(x) - m_k(x)) = \min_{k \neq \lambda}(w_\lambda^T x - w_k^T x) \\ &= w_\lambda^T x - \max_{k \neq \lambda} w_k^T x. \end{aligned}$$

With these definitions we can define a prediction in the same way as before, i.e. we take the highest confidence prediction $\text{argmax}_{\lambda \in \{1,...,K\}}\, c_\lambda(x)$.

Using hinge-loss and adding a square 2-norm regularization, show that the minimization of the doubt is equivalent to the following problems.

1. Average confidence:

$$\min_w \underbrace{\sum\nolimits_{i=1}^{n} \max(0, 1 - A_i^T X_i^T w)}_{f(A^T X^T w)} + \underbrace{\tfrac{\gamma}{2}\|w\|_2^2}_{g(w)},$$

where $x_i, w_i \in \mathbb{R}^p$, $w = (w_1, \dots, w_K) \in \mathbb{R}^{pK}$, and

$$X_i = \begin{bmatrix} x_i & & \\ & \ddots & \\ & & x_i \end{bmatrix} \in \mathbb{R}^{pK \times K}, \quad A_i = e_{y_i} - \tfrac{1}{K}\mathbf{1} \in \mathbb{R}^K.$$

$\mathbf{1}$ is the vector of all ones in $\mathbb{R}^K$ while e_i is the unit vector in $\mathbb{R}^K$ of all zeros except the i:th element.

Further we have $f : \mathbb{R}^n \to \mathbb{R}$ satisfying $f(x) = \sum_{i=1}^n \max(0, 1 - x_i)$, $X = [X_1, ..., X_n] \in \mathbb{R}^{pK \times Kn}$, and $A = \text{blkdiag}(A_1, ..., A_n) \in \mathbb{R}^{Kn \times n}$.

2. Worst case confidence:

$$\min_w \underbrace{\sum\nolimits_{i=1}^{n} f_i(M_i^T X_i^T w)}_{f(M^T X^T w)} + \underbrace{\tfrac{\gamma}{2}\|w\|_2^2}_{g(w)},$$

where w_i, w, x_i, X_i, and X are the same for the average confidence. The functions, $f_i : \mathbb{R}^K \to \mathbb{R}$ satisfy $f_i(u_i) = \max(d_{y_i} - u_i)$ and

$$d_i = \mathbf{1} - e_i \in \mathbb{R}^K, \quad M_i = e_{y_i}\mathbf{1}^T - I \in \mathbb{R}^{K\times K}$$

where I is the identity matrix in $\mathbb{R}^{K\times K}$ while $\mathbf{1}$ and e_i are the same for the average confidence.

Further, $M = \text{blkdiag}(M_1, ..., M_n) \in \mathbb{R}^{Kn\times Kn}$ and the function $f : \mathbb{R}^{nK} \to \mathbb{R}$ satisfies $f(u) = \sum_{i=1}^n f_i(u_i)$ where $u = (u_1, \dots, u_n) \in \mathbb{R}^{Kn}$ and $u_i \in \mathbb{R}^K$.

Exercise 4.13

Consider the average confidence multiclass SVM problem from Exercise 4.12.

1. Find the dual problem

$$\min_{\mu} f^*(\mu) + g^*(-XA\mu).$$

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

Exercise 4.14

Consider the worst case confidence multiclass SVM problem from Exercise 4.12.

1. Find the dual problem

$$\min_{\mu} f^*(\mu) + g^*(-XM\mu).$$

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

Hints

Chapter 5

Algorithms

Exercise 5.1

Suppose that f is convex and differentiable. Consider the gradient descent algorithm

$$x^{k+1} = x^k - \lambda \nabla f(x^k)$$

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

Exercise 5.2

Suppose that f is convex and x is such that $x = \text{prox}_{\gamma f}(x)$ for $\gamma > 0$. Show that x minimizes f.

Exercise 5.3

Suppose that f and g are (closed) convex, f is differentiable, and x is such that $x = \text{prox}_{\gamma g}(x - \gamma \nabla f(x))$ for $\gamma > 0$. Show that x minimizes $f + g$.

Exercise 5.4

Consider the problem

$$\inf f(x)$$

and some iterative algorithm that generates a sequence $x^0, x^1, x^2, ...$ where the function values decrease, i.e.

$$f(x^{k+1}) \leq f(x^k).$$

1. Give an example of of a function f where this does not imply the convergence of the function values $f(x^k)$.
2. Assume the function is lower bounded $f(x) \geq B$. Prove that the sequence of function values converge to some value $f(x^k) \to b$.

3. Give an example of a function f that is bounded below and sequence x^k where $b \neq \inf f(x)$.

EXERCISE 5.5

Let f be an L-smooth function and consider the gradient descent algorithm, i.e. select $x_0 \in \mathbb{R}^n$ and for all $k \in \mathbb{N}$

$$x^{k+1} = x^k - \gamma \nabla f(x^k)$$

where $\frac{2}{L} > \gamma > 0$. Assume that f is lower bounded by B, $f(x) \geq B$ for all x.

1. Show that the sequence $(f(x^k))_{k \in \mathbb{N}}$ satisfies

$$f(x^{k+1}) \leq f(x^k) - \gamma(1 - \tfrac{L}{2}\gamma)\|\nabla f(x^k)\|_2^2.$$

2. Show that $\nabla f(x^k) \to 0$ as $k \to \infty$.
3. Assume that f is strongly convex, show that $x^k \to x^\star$ where $f(x^\star) = \min f(x)$.

Assuming that at least one minimum exists, it is possible to show that $x^k \to x^\star$ even in the smooth convex as well. However, it is not enough that $f(x^{k+1}) \leq f(x^k)$ and $\nabla f(x^k) \to 0$.

4. Give an example of a lower bounded convex function f and sequence x^k such that $f(x^{k+1}) \leq f(x^k)$ and $\nabla f(x^k) \to 0$ but where $f(x^k) \not\to \inf f(x)$.

EXERCISE 5.6

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

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

where $\gamma > 0$.

1. Show that $(f(x^k))_{k \in \mathbb{N}}$ is a decreasing sequence according to

$$f(x^{k+1}) \leq f(x^k) - \tfrac{1}{2\gamma}\|x^{k+1} - x^k\|_2^2.$$

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 $\text{dist}_{\partial f(x^k)}(0) \to 0$ where $\text{dist}_{\partial f(x)}(0) = \inf_{s \in \partial f(x)} \|s - 0\|$, 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 5.7

Consider the minimization problem

$$\min\, f(x) + g(x)$$

and the proximal gradient method

$$x^{k+1} = \mathbf{prox}_{\lambda g}(x^k - \lambda \nabla f(x^k)).$$

where $\lambda > 0$. Assume f is L-smooth and (closed) convexity of g. Prove that it is a descent method for sufficiently small λ and find the upper bound on λ.

Exercise 5.8

Which of the algorithms

- Gradient Descent
- Coordinate Gradient Descent
- Proximal Gradient
- Coordinate Proximal Gradient

are applicable to the minimization problem $\min_{x\in\mathbb{R}^n} h(x)$ 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 5.9

For each of the algorithms and functions in Excercise 5.8, which of the algorithms are applicable to some dual formulation of each of the problems?

Exercise 5.10

Estimate the per-iteration complexity for each of the primal algorithms from Exercise 5.8 on the problems

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

Exercise 5.11 ($\star$)

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 5.12

Consider the problem

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

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 minimize$_y$ $y^TV^TQVy + q^TVy$ has better properties. This is often called *preconditioning*. Find a diagonal matrix V so that the diagonal elements in V^TQV are 1.
5. What are (rougly) 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 5.13 ($\star$)

Let operator T be a contraction, i.e. $\|Tx - Ty\| \leq L\|x - y\|$ with $L < 1$ for all x, y. Given some x^0, show that the following fixed point iteration

$$x^{k+1} = Tx^k$$

converge to a fixed point, $x^\star = Tx^\star$, i.e. $\|x^k - x^\star\| \to 0$ as $k \to \infty$.

Exercise 5.14

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 5.15

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

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

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 5.16

Consider the problem and the proximal coordinate descent algorithm from Exercise 5.15 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 5.17 (*)

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 5.14 and make the same comparison.

Exercise 5.18 (H)

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

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

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 5.19 (*)

In this exercise we want to study the convergence of stochastic gradient descent. Consider the same problem as in Exercise 5.17, 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 5.17.

- Run the algorithm with a few different constant step sizes γ, for example $\lambda_{\text{max}}, \lambda_{\text{max}}/10, \lambda_{\text{max}}/100$, 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 5.18

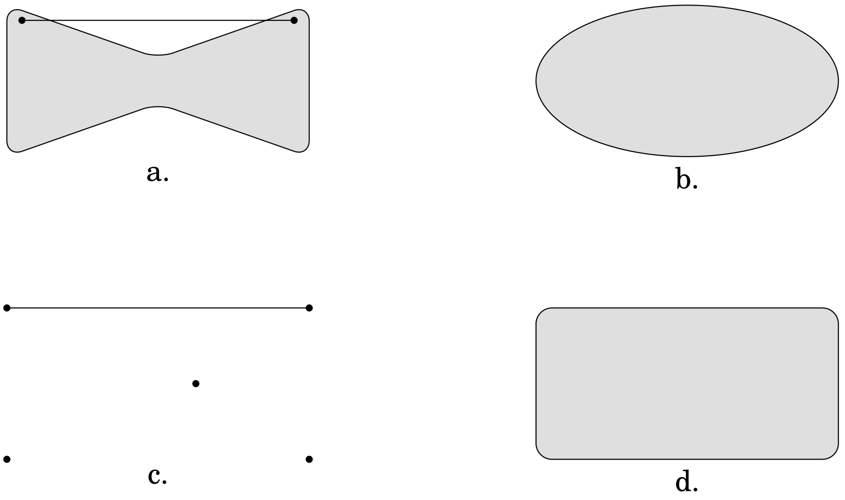
$$\mathbb{E}\|X - \mathbb{E}X\|^2 = \mathbb{E}\|X\|^2 - \|\mathbb{E}X\|^2$$

Solutions to Chapter 1

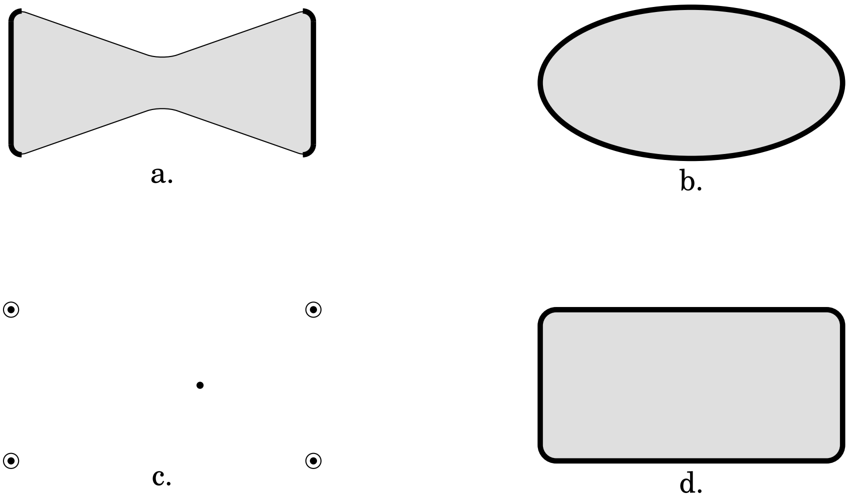
Solution 1.1

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

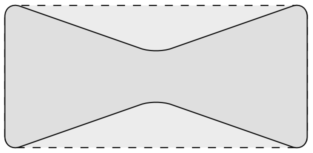
   Figures a. and c. represent nonconvex sets since the lines drawn below between two points in the respective sets are partially outside the sets.



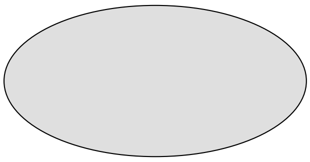
2. Figures b. and d. are convex so there exist supporting hyperplanes at the entire boundary.



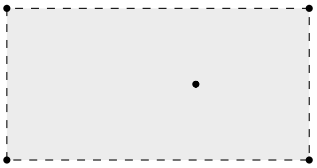
3. Figures b. and d. are convex so the convex hull is the set itself.



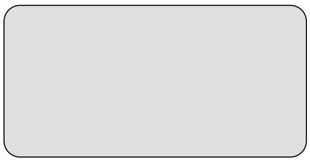
a.



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 $x^TXx \geq 0$ and $x^TYx \geq 0$ for all $x \in \mathbb{R}^n$, and for arbitrary $x \in \mathbb{R}^n$:
$$x^TZx = x^T(\theta X + (1-\theta)Y)x = \theta x^TXx + (1-\theta)x^TYx \geq 0.$$
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. 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 1.4

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 1.5

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

Solution 1.6

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 $a_j^T x = \alpha b_j > b_j$ and $\alpha x \notin S$.

   If $b_j < 0$ and $\alpha \in [0, 1)$, then $a_j^T x = \alpha b_j > b_j$ 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 1.7

1. Convex. We should prove that

$$\iota_C(\theta x + (1-\theta)y) \leq \theta\iota_C(x) + (1-\theta)\iota_C(y) \tag{5.1}$$

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 (5.1) holds. If $x \notin C$ or $y \notin C$, the RHS is ∞ which means that (5.1) 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 1.8

1. We know that $\|x\|$ is convex. Now, define

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

Since h is an increasing function for $p \geq 1$ and $\|x\|$ is convex, $h(\|x\|) = \|x\|^p$ 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.

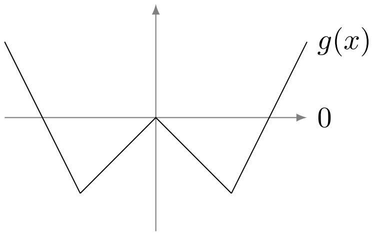
4. $max(0, 1 + x_i)$ max of convex functions, hence convex. Sum over these convex functions is convex. Second term is increasing function of (convex) norm, hence convex. Nonnegative sum is convex.

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.

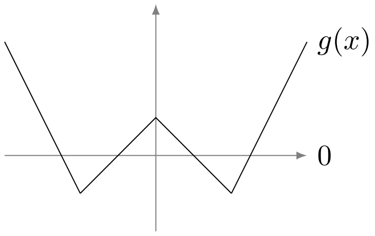
Solution 1.9

1. It is nonempty since obviously $\bar{x} \in C$. Now, let $x_1 \in C$ and $x_2 \in C$ be arbitrary. Then, $g(x_1) \leq 0$ and $g(x_2) \leq 0$. Now, by convexity of g, we have for all $\theta \in [0,1]$ that $x = \theta x_1 + (1-\theta)x_2$ satisfies $g(x) \leq \theta g(x_1) + (1-\theta)g(x_2) \leq 0$. Hence $x \in C$, and C is convex.

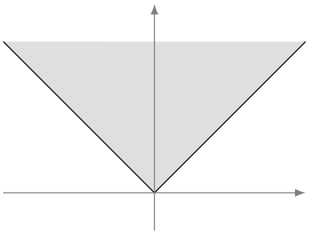
2. Let g be as follows:



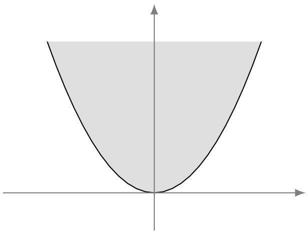
3. Let g be as follows:



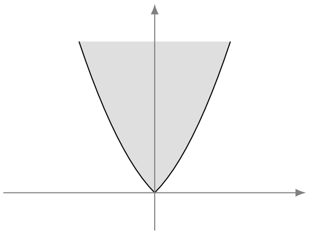
Solution 1.10



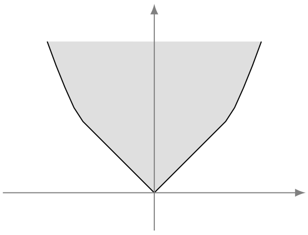
$f(x) = |x|$



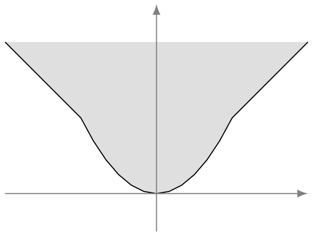
$f(x) = x^2$



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



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



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

Solution 1.11

1. We have

$$\begin{aligned} g(\theta x + (1-\theta)y) &= g_1(\theta x + (1-\theta)y) + g_2(\theta x + (1-\theta)y) \\ &\leq \theta g_1(x) + (1-\theta)g_1(y) + \theta g_2(x) + (1-\theta)g_2(y) \\ &= \theta[g_1(x) + g_2(x)] + (1-\theta)[g_1(y) + g_2(y)] \\ &= \theta g(x) + (1-\theta)g(y). \end{aligned}$$

   Hence g is convex.

2. We have

$$\begin{aligned} \text{epi}g &= \{(x,r) : g(x) \leq r\} \\ &= \{(x,r) : \max(g_1(x), g_2(x)) \leq r\} \\ &= \{(x,r) : g_1(x) \leq r \text{ and } g_2(x) \leq r\} \\ &= \{(x,r) : g_1(x) \leq r\} \cap \{(x,r) : g_2(x) \leq r\} \\ &= \text{epi}g_1 \cap \text{epi}g_2, \end{aligned}$$

   which is convex since g_1 and g_2 are convex. Hence g is a convex function.

Solution 1.12

Let $x, y \in \text{dom}f$. 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 $(\theta x + (1-\theta)y) \in \text{dom}f$ if $x, y \in \text{dom}f$ and $\text{dom}f$ is convex.

Solution 1.13

The epigraph of f is

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

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

Solution 1.14

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 1.15

1. That $f(x) - \frac{\sigma}{2}\|x\|_2^2$ is convex is equivalent to that for all x, y and $\theta \in [0,1]$ and $z = \theta x + (1-\theta)y$:

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

which is equivalent to that

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

Now,

$$\begin{aligned}
&\|\theta\, x + (1-\theta)y\|_2^2 - \theta\|x\|_2^2 - (1-\theta)\|y\|_2^2 \\
&\quad = (\theta^2 - \theta)\|x\|_2^2 + ((1-\theta)^2 - (1-\theta))\|y\|_2^2 + 2\theta(1-\theta)x^T y \\
&\quad = (\theta(1-\theta))(-\|x\|_2^2 - \|y\|_2^2 + 2x^T y) \\
&\quad = -(\theta(1-\theta))(\|x-y\|_2^2) && \text{(5.2)}
\end{aligned}$$

which proves the result.

2. That $\frac{\beta}{2}\|x\|_2^2 - f(x)$ is convex is equivalent to that for all x, y and $\theta \in [0,1]$ and $z = \theta x + (1-\theta)y$:

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

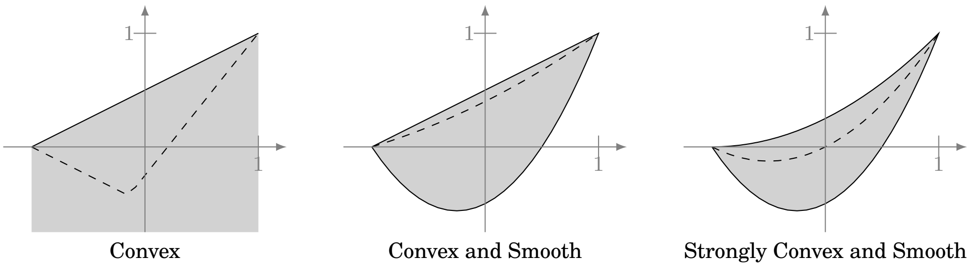
which is equivalent to that

$$f(z) \geq \theta f(x) + (1-\theta)f(y) + \tfrac{\beta}{2}(\|z\|_2^2 - \theta\|x\|_2^2 - (1-\theta)\|y\|_2^2).$$

Using (5.2) gives the result.

Solution 1.16

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.



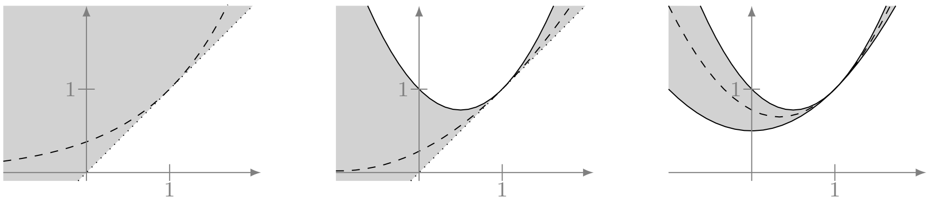
Convex

Convex and Smooth

Strongly Convex and Smooth

Solution 1.17

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

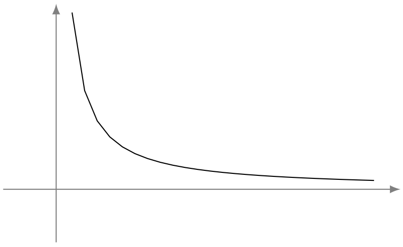
Solution 1.18

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.

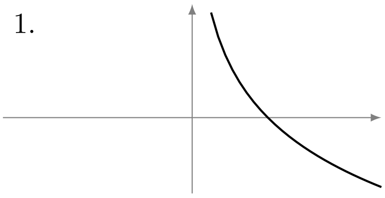
2. The function $f(x) = \frac{1}{x}$ with domain $x > 0$ is strictly convex with infimum 0. But no x exists such that $f(x) = 0$. See figure.



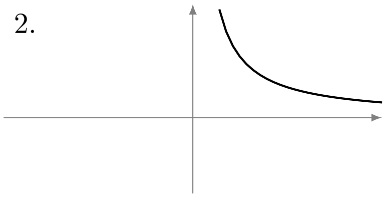
Solution 1.19

See figure below.

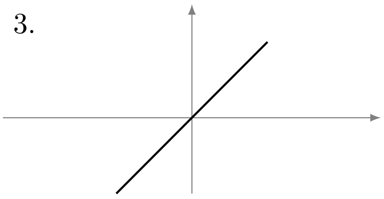
1.



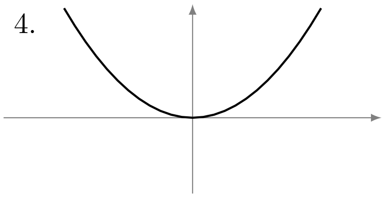
2.



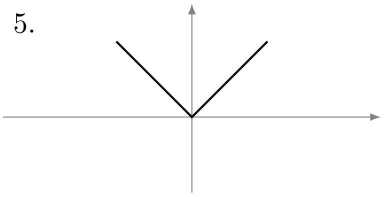
3.



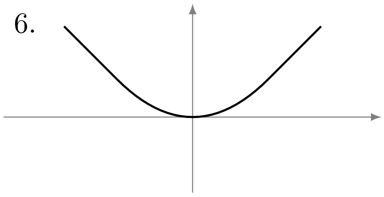
4.



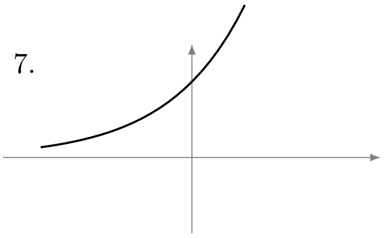
5.



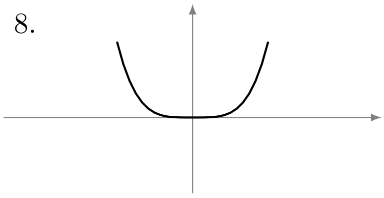
6.



7.



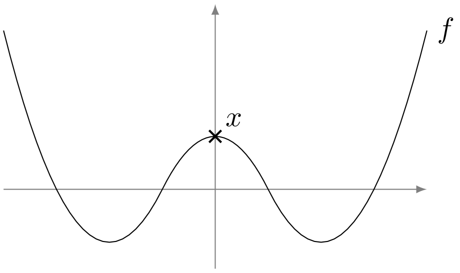
8.



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 1.20

1. Consider the following function f and point x:



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

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

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

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

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

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

Multiplying the first inequality by θ, the second by $1 - \theta$, and adding them gives (since $\theta \in [0, 1]$)

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

That is, f is convex.

Solution 1.21

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 1.22

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 1.23

We have

$$\begin{aligned}
f(\sum_{i=1}^{3} \theta_i x_i) &= f(\theta_1 x_1 + (1-\theta_1)(\sum_{i=2}^{3} \tfrac{\theta_i}{1-\theta_1} x_i)) \\
&\leq \theta_1 f(x_1) + (1-\theta_1) f(\sum_{i=2}^{3} \tfrac{\theta_i}{1-\theta_1} x_i) \\
&= \theta_1 f(x_1) + (1-\theta_1) f(\tfrac{\theta_2}{1-\theta_1} x_2 + (1 - \tfrac{\theta_2}{1-\theta_1}) \tfrac{\theta_3}{(1-\frac{\theta_2}{1-\theta_1})(1-\theta_1)} x_3)) \\
&= \theta_1 f(x_1) + (1-\theta_1) f(\tfrac{\theta_2}{1-\theta_1} x_2 + (1 - \tfrac{\theta_2}{1-\theta_1}) \tfrac{\theta_3}{1-\theta_1-\theta_2} x_3)) \\
&\leq \theta_1 f(x_1) + \theta_2 f(x_2) + (1-\theta_1-\theta_2) f(\tfrac{\theta_3}{1-\theta_1-\theta_2} x_3) \\
&= \theta_1 f(x_1) + \theta_2 f(x_2) + \theta_3 f(x_3)
\end{aligned}$$

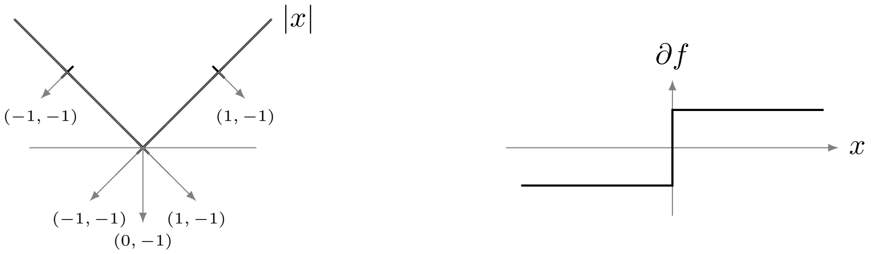
where the convexity definition has been used in the inequalities and that $\theta_3 = 1 - \theta_1 - \theta_2$ in the last equality.

Solutions to Chapter 2

Solution 2.1

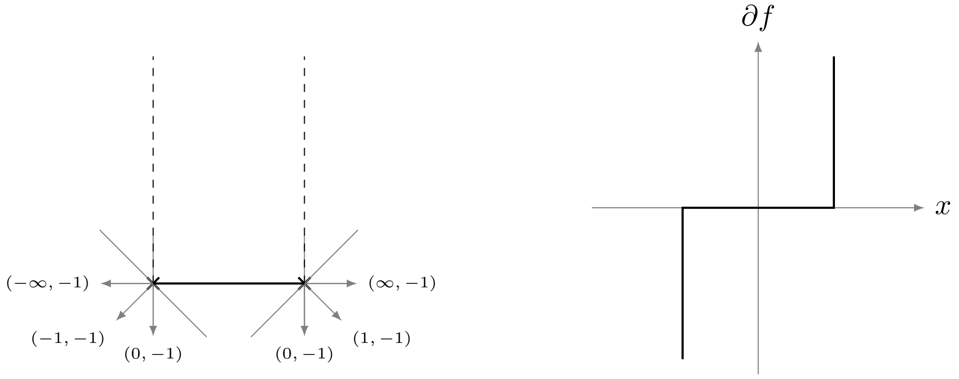
1. Function is convex and differentiable with $\nabla f(x) = x$. Hence $\partial f(x) = \{x\}$.
2. Function is convex and differentiable with $\nabla f(x) = Hx + h$. Hence $\partial f(x) = \{Hx + h\}$.
3. For $x < 0$, the function is $-x$ and differentiable with gradient -1. For $x > 0$, the function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

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



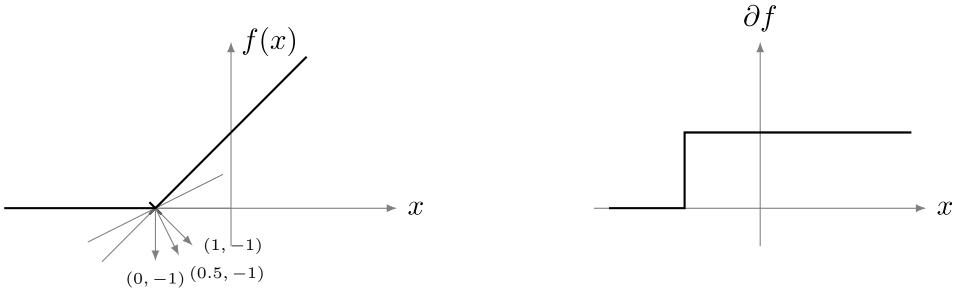
4. Whenever $x \in (-1, 1)$, the function is 0 with gradient 0, hence $\partial f(x) = 0$. When $x > 1$ or $x < -1$, x is outside the domain and $\partial f(x) = \emptyset$. When $x = 1$, all $s \geq 0$ are subgradients. When $x = -1$ all $s \geq 0$ are subgradient (see figure). Note that this subdifferential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

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



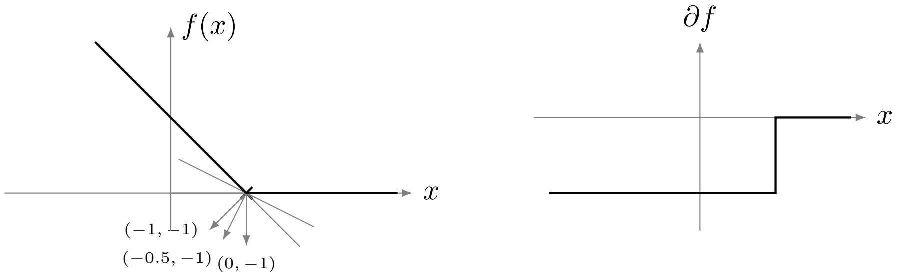
5. For $x < -1$, the function is 0 and the gradient is 0, hence $\partial f(x) = 0$. For $x > -1$, the function is $x+1$ and the gradient is 1, hence $\partial f(x) = 1$. For $x = -1$, all $s \in [0,1]$ are subgradients (see figure).

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



6. 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).

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



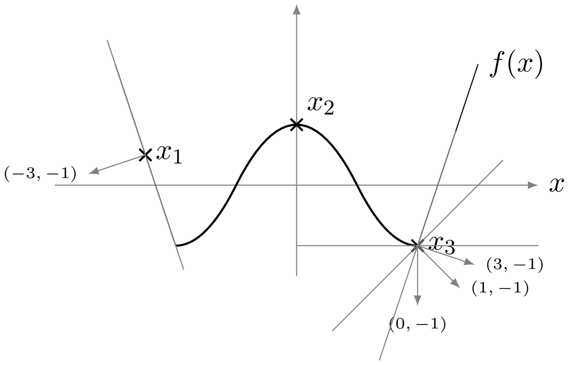
Solution 2.2

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.

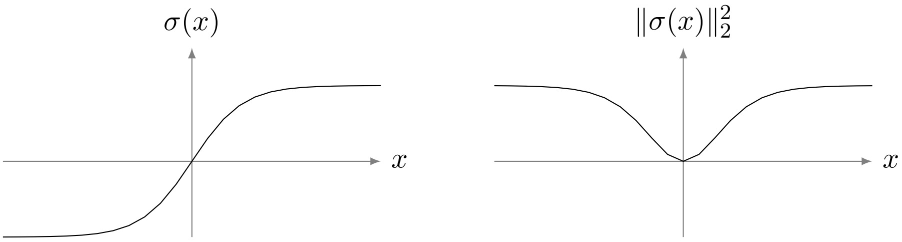
Solution 2.4

- 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 convex function. The subdifferential in b. is not maximally monotone, hence not the subdifferential of a convex function.

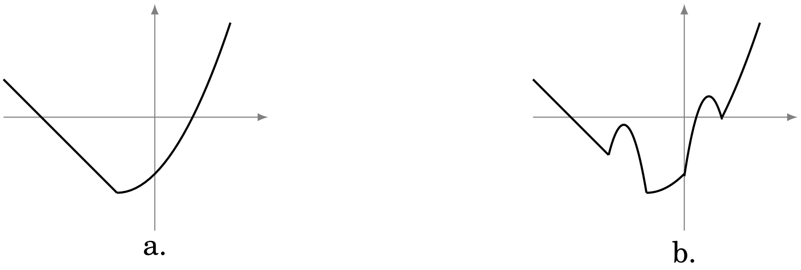
Solution 2.5

1. Convex. The gradient $\nabla f(x) = \sigma(x)$. Since σ is differentiable, so is f. Therefore, since $\sigma = \nabla f$ is monotone with full domain, f is convex.
2. Not convex in general for nonlinear σ. Consider, e.g., $\sigma(x) = \frac{2}{1+e^{-x}} - 1$ and corresponding $\|\sigma(x)\|_2^2$ below.



Solution 2.6

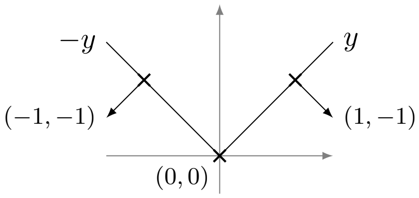
1. a. Since ∂f is maximally monotone, f is convex.
   b. Since ∂f is not maximally monotone, f is not convex.
2. The optimal point x^* satisfies $0 \in \partial f(x^*)$ (Fermat's rule). Hence, the minimizing x^* are the x where the graph crosses the x-axis for both a. and b..
3. No, since a constant offset of f is not visible in ∂f.
4. Below are example plots of f.



It is linear to the left of the minimum and quadratic to the right.

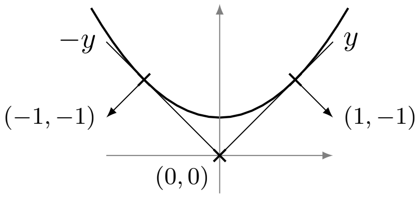
Solution 2.7

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 2.8

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 equilvalent to

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

Solution 2.9

1. We have:

$$\begin{aligned} \|\nabla g(x) - \nabla g(y)\|_2 &= \|A^T\nabla f(Ax+b) - A^T\nabla f(Ay+b)\|_2 \\ &= \|A^T(\nabla f(Ax+b) - \nabla f(Ay+b))\|_2 \\ &\leq \|A\|\|(\nabla f(Ax+b) - \nabla f(Ay+b))\|_2 \\ &\leq \|A\|\beta\|Ax+b-(Ay+b)\|_2 \\ &\leq \|A\|\beta\|A(x-y)\|_2 \\ &\leq \|A\|^2\beta\|x-y\|_2. \end{aligned}$$

Hence g is $\|A\|^2\beta$-smooth.

2. Set, e.g., $A = 0$. Then $g(x) = f(b)$, which is constant. A constant function is not lower bounded by a quadratic function with positive curvature. Hence g is not necessarily convex.

SOLUTION 2.10

(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 2.11

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 $s = (s_1, \dots, s_n)$ 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 $j \in \{1, \dots, n\}$ 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 2.12

All subgradients s must satisfy

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

Since there exists y such that $f(y) < \infty$ and $f(x) = \infty$, no subgradient s exists at $x \notin \text{dom} f$.

Solution 2.13

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 equilvalent 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 2.14

Fermat's rule says $x = \text{prox}_{\gamma f}(z)$ 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 2.15

We have

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

Solution 2.16

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.

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

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

   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,

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

3. We have

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

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

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

   Therefore

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

   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 $x = t_- \leq 0$ with $t_- \to -\infty$, which gives

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

since $s < -1$.

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

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

since $s > 1$.

For $s \in [-1, 1]$, we have by Cauchy-Schwarz that $sx \leq |x||s| \leq |x|$ for all x. Therefore $f^*(s) = \sup_x s^T x - |x| \leq \sup_x |x| - |x| = 0$. Further, $f^*(s) = \sup_x sx - |x| \geq \sup_x s0 - 0 = 0$. 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 $x = t_- \leq 0$ with $t_- \to -\infty$ and

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

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

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

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 $x = t_- \leq -1$ with $t_- \to -\infty$ and

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

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

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

   Hence, the conjugate is

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

Solution 3.2

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 3.3

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

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

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 3.4

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 $\max_{\|s\|_2 \leq 1}(s^T x) \geq \|x\|_2$.

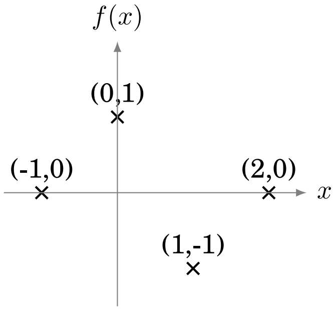
   Therefore

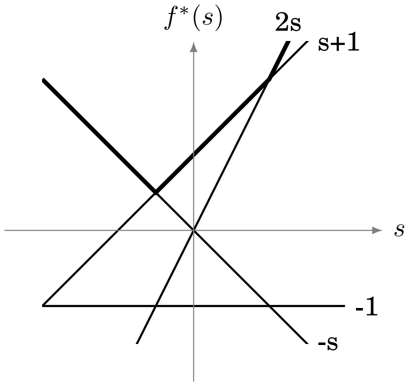
   $$\partial f(x) = \begin{cases} B(0, 1) & \text{if } x = 0 \\ x/\|x\|_2 & \text{else} \end{cases}$$

Solution 3.5

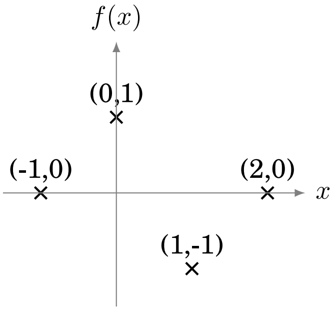
1. Since f is only defined in in four points, the conjugate is

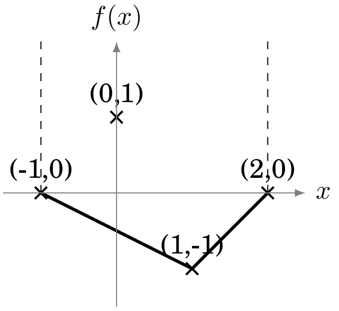
   $$f^*(s) = \sup_x(sx - f(x)) = \max(-s - 0, -1, s + 1, 2s)$$





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





Solution 3.6

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 3.7

1. We have

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

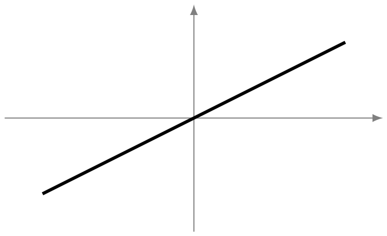
2. We have

$$\partial f^*(0) = \operatorname*{Argmax}_x(0^T x - f^{**}(x)) = \text{Argmax}(-f(x)) = \text{Argmin}\, f(x)$$

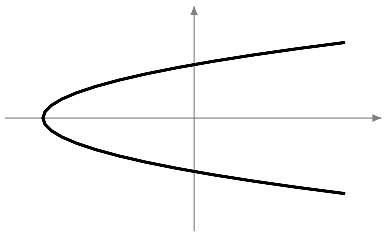
where we have used that $f^{**} = f$.

Solution 3.8

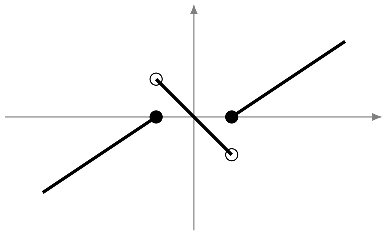
1. Since we are dealing with set valued mappings it is no problem if the inverses are set valued, i.e. we do not need to care about surjectivity and injectivity. The axis of the graphs are simply flipped.
2. Only a. and b. are functions. The other are set-valued.
3. Only the inverses of operators a. and c. are functions. The other are set-valued



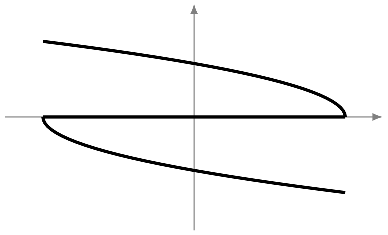
a.



b.



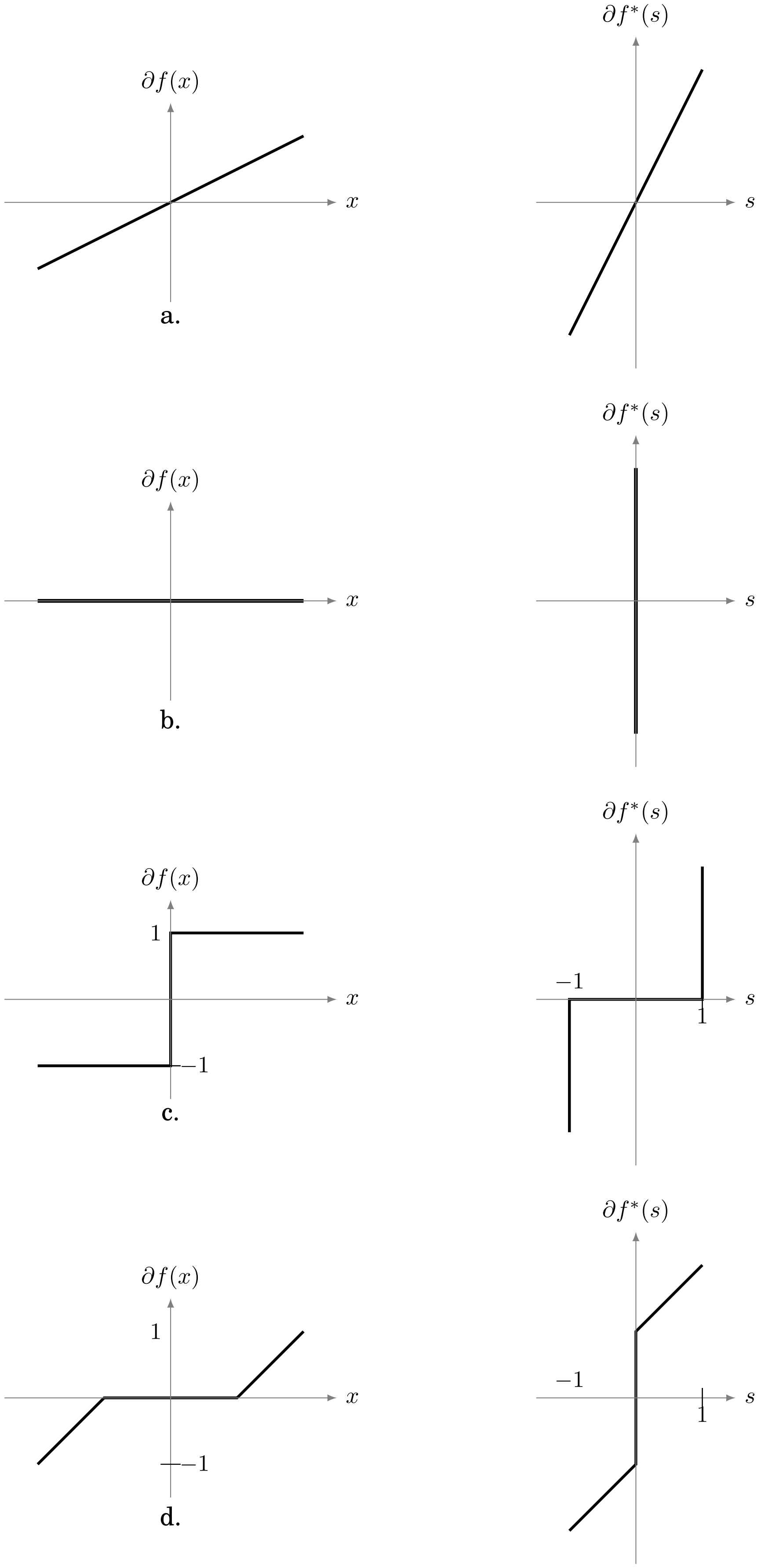
c.



d.

Solution 3.9

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


∂f(x)
x
a.
∂f*(s)
s
∂f(x)
x
b.
∂f*(s)
s
∂f(x)
1
x
−1
c.
∂f*(s)
−1
s
1
∂f(x)
1
x
−1
d.
∂f*(s)
−1
s
1

Solution 3.10

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

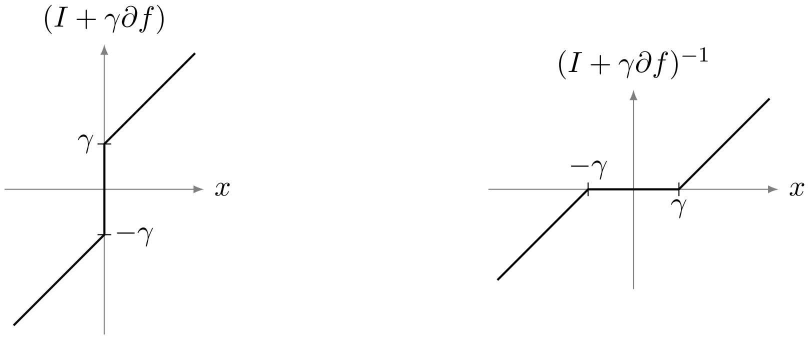
$$\begin{aligned} & 0 \in \partial f(z) + \gamma^{-1}(z - x) \\ \Leftrightarrow \qquad & x \in (I + \gamma\partial f)(z) \\ \Leftrightarrow \qquad & (I + \gamma\partial f)^{-1}x = z. \end{aligned}$$

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

Solution 3.11

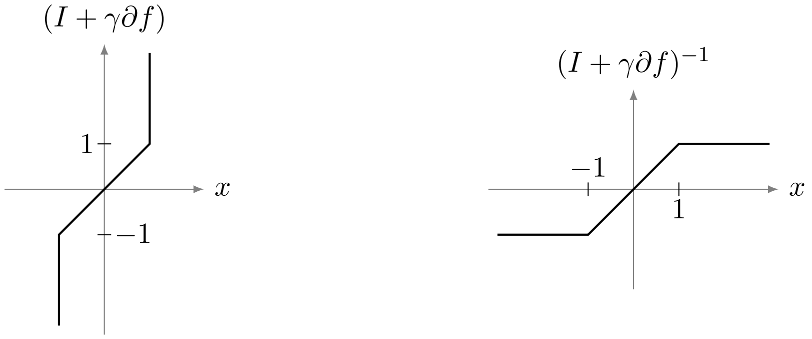
1. We will solve this graphically. Left plot shows $I + \gamma\partial f$ and the right shows $(I + \gamma\partial f)^{-1} = \text{prox}_{\gamma f}$.

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



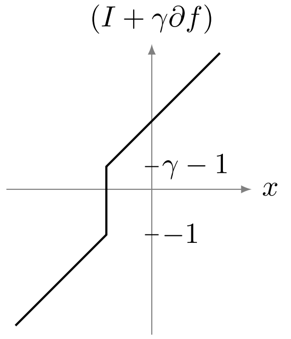
2. We will solve this graphically. Left plot shows $I + \gamma\partial f$ and the right shows $(I + \gamma\partial f)^{-1} = \text{prox}_{\gamma f}$. The prox does not depend on γ (since it is actually a projection).

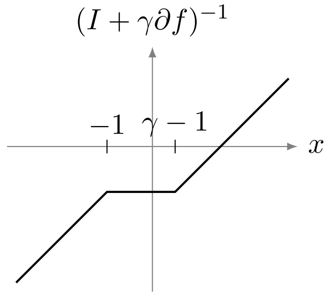
$$\text{prox}_{\gamma f}(x) = \begin{cases} -1 & \text{if } x \le -1 \\ x & \text{if } x \in [-1, 1] \\ 1 & \text{if } x \ge 1 \end{cases}$$



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

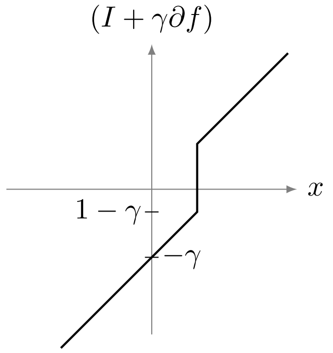
$$\text{prox}_{\gamma f}(x) = \begin{cases} x & \text{if } x \le -1 \\ -1 & \text{if } x \in [-1, \gamma - 1] \\ x - \gamma & \text{if } x \ge \gamma - 1 \end{cases}$$

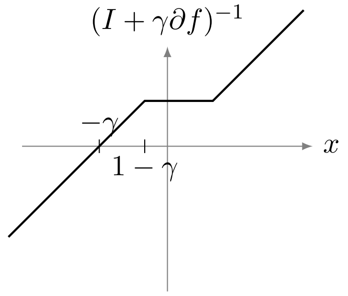




4. We will solve this graphically. Left plot shows $I + \gamma\partial f$ and the right shows $(I + \gamma\partial f)^{-1} = \text{prox}_{\gamma f}$.

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





Solution 3.12

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

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

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 $s = \mathrm{prox}_{(\gamma f)^*}(z)$ if and only if

$$\begin{aligned} s &= \operatorname*{Argmin}_y ((\gamma f)^*(y) + \tfrac{1}{2}\|y - z\|_2^2) \\ &= \operatorname*{Argmin}_y (\gamma f^*(\gamma^{-1} y) + \tfrac{1}{2}\|y - z\|_2^2) \\ &= \gamma \operatorname*{Argmin}_v (\gamma f^*(v) + \tfrac{1}{2}\|\gamma v - z\|_2^2) \\ &= \gamma \operatorname*{Argmin}_v (\gamma f^*(v) + \tfrac{\gamma^2}{2}\|v - \gamma^{-1} z\|_2^2) \\ &= \gamma \operatorname*{Argmin}_v (\gamma^{-1} f^*(v) + \tfrac{1}{2}\|v - (\gamma^{-1} z)\|_2^2) \\ &= \gamma \mathrm{prox}_{\gamma^{-1} f^*}(\gamma^{-1} z) \end{aligned}$$

4. Combine first and third subproblems.

Solution 3.13

Compute the prox_{f^*}, i.e., the prox of the conjugate, for the following f.

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

Moreau decomposition says $\mathrm{prox}_{f^*}(z) = z - \mathrm{prox}_f(z)$.

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

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

2. The prox of f satisfies

$$\mathrm{prox}_f(z) = \begin{cases} z & \text{if } z < -1 \\ -1 & \text{if } z \in [-1, \gamma - 1] \\ z - \gamma & \text{if } z > \gamma - 1 \end{cases}$$

which implies that

$$\mathrm{prox}_{f^*}(z) = z - \mathrm{prox}_f(z) = \begin{cases} 0 & \text{if } z < -1 \\ z + 1 & \text{if } z \in [-1, \gamma - 1] \\ \gamma & \text{if } z > \gamma - 1 \end{cases}$$

3. The prox of f 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

$$\mathbf{prox}_{f^*}(z) = z - \mathbf{prox}_f(z) = \begin{cases} -\gamma & \text{if } z < 1-\gamma \\ z-1 & \text{if } z \in [1-\gamma, 1] \\ 0 & \text{if } z > 1 \end{cases}$$

Solution 3.14

1. The functions are closed convex and constraint qualification holds so the primal problem is equivalent to

$$0 \in \partial f(x) + \partial g(x) \quad \Leftrightarrow \quad \begin{aligned} y &\in \partial f(x) \\ -y &\in \partial g(x) \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} x &\in \partial f^*(y) \\ x &\in \partial g^*(-y) \end{aligned}$$

since $\partial f = (\partial g^*)^{-1}$ for (closed) convex functions.

2. Eliminating the x gives:

$$\begin{aligned} x &\in \partial f^*(y) \\ x &\in \partial g^*(-y) \end{aligned} \quad \Leftrightarrow \quad 0 \in \partial f^*(y) - \partial g^*(-y)$$

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

$$\begin{aligned} x &\in \partial f^*(y^\star) \\ x &\in \partial g^*(-y^\star) \end{aligned}$$

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

$$\begin{aligned} x^\star &= \nabla f^*(y^\star) \\ x^\star &\in \partial g^*(-y^\star). \end{aligned}$$

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 3.15

Fermat's rule gives

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

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

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

Solution 3.16

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 $g^*(\nu) = \sum_{i=1}^n \max(0, 1 - \nu_i)$ (Exercise 3.1-6 and 3.2 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 3.3) and $g^*(\nu) = \frac{1}{2\lambda}\|\nu + b\|_2^2$ (Exercise 3.1-2). Therefore the dual problem is

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

   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 3.17

1. We have

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

2. We have

$$\begin{aligned} f^*(s_0) &= \sup_x (s_0^T x - f(x)) \le \sup_x (s_0^T x - (f(x_0) + s_0^T (x - x_0))) \\ &= s_0^T x_0 - f(x_0). \end{aligned}$$

   Equality holds in view of the first subproblem.

3. We have

$$\begin{aligned} & f^*(s_0) \le s_0^T x_0 - f(x_0) \\ \Leftrightarrow \quad & \sup_y \left\{ s_0^T y - f(y) \right\} \le s_0^T x_0 - f(x_0) \\ \Leftrightarrow \quad & s_0^T y - f(y) \le s_0^T x_0 - f(x_0) \text{ for all } y \\ \Leftrightarrow \quad & f(y) \ge f(x_0) + s_0^T y - x_0 \text{ for all } y \\ \Leftrightarrow \quad & s_0 \in \partial f(x_0) \end{aligned}$$

   And we have actually shown the full equivalence.

Solution 3.18

1. Since $f^{**} \le f$, we have

$$0 = f^*(s) + f(x) - s^T x \ge f^*(s) + f^{**}(x) - s^T x$$

   Fenchel Young's inequality says that other direction holds:

$$0 \le f^*(s) + f^{**}(x) - s^T x.$$

   This implies equality $0 = f^*(s) + (f^*)^*(x) - s^T x$ holds, which is equivalent to $x \in \partial f^*(s)$.

2. Apply previous result with f as f^*.

3. Use above results and that $f^{**} = f$ for convex (and closed) f.

Solution 3.19

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

$$g^*(s) = \sup_x (s^T x - h(Lx)) = -\inf_x (h(Lx) + l_s(x)),$$

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

$$\begin{aligned} h^*(\mu) &= \sup_y (\mu^T y - f(y + c)) = \sup_v (\mu^T (v - c) - f(v)) \\ &= \sup_v (\mu^T (v) - f(v)) - \mu^T c = f^*(\mu) - \mu^T c \\ l_s^*(\nu) &= \sup_x (\nu^T x + s^T x) = \sup_x ((\nu + s)^T x) = \iota_{\{0\}}(\nu + s) \end{aligned}$$

Now Fenchel strong duality (constraint qualification is satisfied since $\operatorname{dom} l_s = \mathbb{R}^n$) gives

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

Solutions to Chapter 4

Solution 4.1

- All are convex. f_2 and f_3 are norms which are convex. f_4 is convex since $\|\cdot\|$ is convex and positive and $(\cdot)^2$ is increasing for positive inputs. f_1 is convex since $\|Ax-b\|_2^2$ is a composition of affine and convex which is convex.

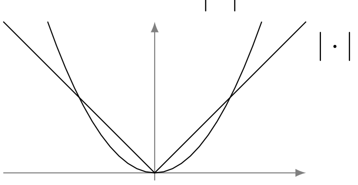
- - f_1 is $\|A^TA\| = \lambda_{\max}(A^TA)$ smooth. For smoothness is equivalent to the gradient being Lipschitz continuous.

    $$\begin{aligned}&\|A^T(Ax-b)-A^T(Ay-b)\|\\ &\quad = \|A^TAx - A^TAy\| = \|A^TA(x-y)\| \le \|A^TA\|\|x-y\|\end{aligned}$$

  - Both f_2 and f_3 are not smooth. $g = \frac{L}{2}\|\cdot\|_2^2 - f$ is not convex for any L, i.e. $g(\theta x + (1-\theta)y) \not\geq \theta g(x) + (1-\theta)g(y)$. Take for example $x = \frac{1}{L}$, $y = \frac{-1}{L}$ and $\theta = \frac{1}{2}$.
  - f_4 is smooth. $g = \frac{L}{2}\|\cdot\|_2^2 - f = \frac{L-1}{2}\|\cdot\|_2^2$ which is convex for $L \geq 1$.

- - f_1 is $\lambda_{\min}(A^TA)$-strongly convex. $f(x) - \frac{\mu}{2}\|x\|_2^2 = \frac{1}{2}(Ax-b)^T(Ax-b) - \frac{\mu}{2}x^Tx = \frac{1}{2}x^T(A^TA-\mu I)x - b^TAx + b^Tb$, which is convex as long as $A^TA - \mu I$ is positive semi-definite, i.e. $\lambda_{\min}(A^TA) - \mu \geq 0$.
  - Both f_2 and f_3 are not strongly convex. $g = f - \frac{\mu}{2}\|\cdot\|_2^2$ is not convex for any μ, i.e. $g(\theta x + (1-\theta)y) \not\geq \theta g(x) + (1-\theta)g(y)$. Take for example $x = \frac{3}{\mu}$, $y = \frac{-3}{\mu}$ and $\theta = \frac{1}{2}$.
  - f_4 is strongly convex. $g = f - \frac{\mu}{2}\|\cdot\|_2^2 = \frac{1-\mu}{2}\|\cdot\|_2^2$ which is convex for $\mu \leq 1$.

Solution 4.2



What differs between the two problems is how the error for each data point, $e_i = x_i\theta - y_i$ is penalized. Looking at the plot we see that the 1-norm put less penalty on large errors, $e_i > 1$ compared to the square 2-norm. For this reason should problem 1 be affected less by the large outliers, meaning it should be paired to figure a.

(From this figure we can also see why the square 2-norm does not promote sparsity in the error in same way the 1-norm does. The square flattens out near 0 which means that as the error becomes small there are diminishing returns for reducing the error further. The 1-norm does not have this problem and a reduction of the error will always yield the same cost reduction, regardless of the initial error. For this reason it is more likely that errors are driven to 0.)

Solution 4.3

The model becomes

$$\begin{aligned}
m_{a,b}(x) &= \tfrac{a_0}{2} + \sum_{n=1}^{N} a_n \cos(2\pi n x) + b_n \sin(2\pi n x) \\
&= \tfrac{a_0}{2} \\
&\quad + [\cos(2\pi x), \cos(2\pi 2x), ..., \cos(2\pi N x)] \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} \\
&\quad + [\sin(2\pi x), \sin(2\pi 2x), ..., \sin(2\pi N x)] \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix} \\
&= [\tfrac{1}{2}, \cos(2\pi x), ..., \cos(2\pi N x), \sin(2\pi x), ..., \sin(2\pi N x)] \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_N \\ b_1 \\ \vdots \\ b_N \end{bmatrix} \\
&= \phi(x)^T w
\end{aligned}$$

where $w = (a_0, a_1, ..., a_N, b_1, ..., b_N)$ and the feature map is

$$\phi(x) = (\tfrac{1}{2}, \cos(2\pi x), ..., \cos(2\pi N x), \sin(2\pi x), ..., \sin(2\pi N x))$$

Solution 4.4

- 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 Cauchy-Schwarz, 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 4.5

We first note that because of our class label choice we can write p_1 and p_{-1} in one expression as

$$p_y(x) = \frac{1}{1 + e^{-y(w^T x + b)}}$$

and the likelihood then becomes

$$l_w(x, y) = \prod_{i=1}^N p_{y_i}(x_i) = \prod_{i=1}^N \frac{1}{1 + e^{-y_i(w^T x_i + b)}}.$$

Maximizing $l_w(x, y)$ is the same as maximizing $\log(l_w(x, y))$ since the logarithm is monotonically increasing. Furthermore, maximizing $\log(l_w(x, y))$ is the same as minimizing $-\log(l_w(x, y))$, yielding

$$\begin{aligned} -\log(l_w(x, y)) &= -\log\left(\prod_{i=1}^N \frac{1}{1 + e^{-y_i(w^T x_i + b)}}\right) \\ &= \log\left(\prod_{i=1}^N 1 + e^{-y_i(w^T x_i + b)}\right) \\ &= \sum_{i=1}^N \log(1 + e^{-y_i(w^T x_i + b)}). \end{aligned}$$

This function is convex, $-y_i(w^T x_i + b)$ is affine, $1 + e^{(\cdot)}$ is convex, and $\log(\cdot)$ is increasing.
$\log(1 + e^x)$ is $\frac{1}{4}$-smooth,

$$\frac{d^2}{dx^2}\log(1 + e^x) = \frac{1}{1 + e^x}\Big(1 - \frac{1}{1 + e^x}\Big) \leq \tfrac{1}{4}$$

therefore the problem is $\frac{1}{4}\|A\|^2$-smooth, see Exercise 2.9.
We also see that $\frac{d^2}{dx^2}\log(1 + e^x) \to 0$ as $x \to \infty$ and $x \to -\infty$. Therefore it does not exist any positive lower bound and $\log(1 + e^x)$ and the logistic regression problem are not strongly convex.

Solution 4.6
Using the notation from Exercise 4.5 we have the following cost

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

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

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

Solution 4.7
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 4.8

1. True. See the model $m_w(x)$ as a function of w instead of x, $m_w(x) = f_x(w) = \phi(x)^T w$. $f_x(w)$ is clearly 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.

SOLUTION 4.9
Positive semi-definiteness is

$$a^T K a \geq 0, \quad \forall a.$$

Inserting K and using linearity gives

$$\begin{aligned}
a^T K a &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i K_{ij} a_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i k(x_i, x_j) a_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \langle \phi(x_i), \phi(x_j)\rangle_{\mathcal{F}} a_j \\
&= \sum_{i=1}^{n} a_i \langle \phi(x_i), \sum_{j=1}^{n} a_j \phi(x_j)\rangle_{\mathcal{F}} \\
&= \langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{j=1}^{n} a_j \phi(x_j)\rangle_{\mathcal{F}} \\
&= \langle \Phi, \Phi \rangle_{\mathcal{F}} \\
&= \|\Phi\|_{\mathcal{F}}^2 \geq 0
\end{aligned}$$

where $\Phi = \sum_{i=1}^{n} a_i \phi(x_i)$

SOLUTION 4.10

1. Since f is a sum of hinge-losses, the conjugate f^* is

$$f^*(\mu) = \sum_{i=1}^{N} \mu_i + \iota_{[-1,0]}(\mu_i) = \mathbf{1}^T\mu + \iota_{[-\mathbf{1},\mathbf{0}]}(\mu),$$

see Exercises 3.1-6 and 3.2. The conjugate g^* is

$$\begin{aligned}
g^*(\nu_w, \nu_b) &= \sup_{w,b}((\nu_w, \nu_b)^T(w, b) - \tfrac{\lambda}{2}\|w\|_2^2) \\
&= \sup_{w}(\nu_w^T w - \tfrac{\lambda}{2}\|w\|_2^2) + \sup_{b}(\nu_b b) \\
&= \tfrac{1}{2\lambda}\|\nu_w\|_2^2 + \iota_{\{0\}}(\nu_b),
\end{aligned}$$

(Exercise 3.1-1). We note that

$$g^*(-L^T\mu) = g^*(-\begin{bmatrix} X \\ \mathbf{1}^T \end{bmatrix}\mu) = \tfrac{1}{2\lambda}\| - X\mu\|_2^2 + \iota_{\{0\}}(-\mathbf{1}^T\mu)$$

The dual problem becomes

$$\min \mathbf{1}^T\mu + \iota_{[-\mathbf{1},\mathbf{0}]}(\mu) + \tfrac{1}{2\lambda}\| - X\mu\|_2^2 + \iota_{\{0\}}(-\mathbf{1}^T\mu).$$

or written differently

$$\begin{array}{ll}
\text{minimize} & \mathbf{1}^T\mu + \frac{1}{2\lambda}\mu^T X^T X\mu \\
\text{subject to} & -1 \leq \mu \leq 0 \\
 & \mathbf{1}^T\mu = 0
\end{array}$$

2. The optimal w is recovered from the following primal-dual optimality condition

$$\begin{aligned} L(w,b) &\in \partial f^*(\mu) \\ (w,b) &\in \partial g^*(-L^T\mu). \end{aligned}$$

The second condition yields

$$(w,b) \in (\tfrac{-1}{\lambda}X\mu, \{\mathbb{R} \text{ if } -\mathbf{1}^T\mu = 0\}),$$

i.e. $w = \frac{.1}{\lambda}X\mu$. However, since ∂g^* is not a singleton in the last element, we need something else to recover the optimal b.

The first condition then gives

$$[X^T, \mathbf{1}](w,b) = X^T w + b\mathbf{1} \in \partial f^*(\mu)$$

and we note that

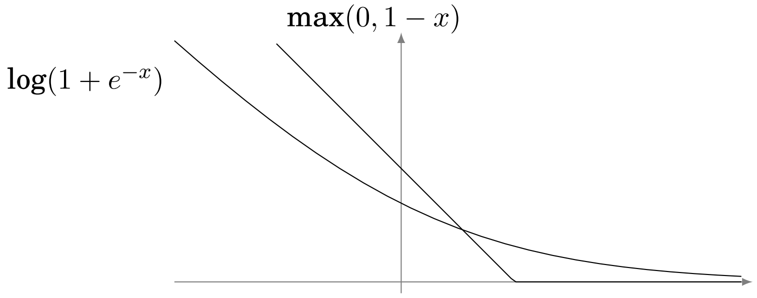
$$(\partial f^*(\mu))_i = \begin{cases} 1 & \text{if } -1 < \mu_i < 0 \\ [1,\infty] & \text{if } \mu_i = 0 \\ [-\infty, 1] & \text{if } \mu_i = -1 \end{cases}.$$

If $-1 < \mu_i < 0$ we see that we can determine b uniquely, i.e. take b such that

$$x_i^T w + b = 1 \iff b = 1 - x_i^T w$$

where i is such that $-1 < \mu_i < 0$.

Solution 4.11



- Start by simply inserting the model

$$\begin{aligned} \textstyle\sum_{i=1}^n \phi(c_{y_i}(x_i)) &= \textstyle\sum_{i=1}^n l(c_{y_i}(x_i)) \\ &= \sum_{\forall i: y_i = 1} l(c_1(x_i)) + \sum_{\forall i: y_i = 2} l(c_2(x_i)) \\ &= \sum_{\forall i: y_i = 1} l(m_1(x_i) - m_2(x_i)) + \sum_{\forall i: y_i = 2} l(m_2(x_i) - m_1(x_i)) \\ &= \sum_{\forall i: y_i = 1} l(w_1^T x_i - w_2^T x_i) + \sum_{\forall i: y_i = 2} l(w_2^T x_i - w_1^T x_i) \\ &= \sum_{\forall i: y_i = 1} l(x_i^T(w_1 - w_2)) + \sum_{\forall i: y_i = 2} l(x_i^T(w_2 - w_1)) \end{aligned}$$

From here we see that the loss only depends on the difference $w = w_1 - w_2$. We can also relabel our classes as $y = 1 \to \hat{y} = 1$ and $y = 2 \to \hat{y} = -1$, this gives

$$\begin{aligned} &= \sum_{\forall i:\hat{y}_i=1} l(x_i^T w) + \sum_{\forall i:\hat{y}_i=-1} l(-x_i^T w) \\ &= \sum\nolimits_{i=1}^n l(\hat{y}_i x_i^T w) \end{aligned} \tag{5.3}$$

Inserting the loss finally give

$$= \sum\nolimits_{i=1}^n \log(1 + e^{-\hat{y}_i x_i^T w})$$

- Starting from (5.3) but inserting $h(x)$ instead of $l(x)$ gives
$$= \sum\nolimits_{i=1}^n \max(0, 1 - \hat{y}_i x_i^T w).$$
- Starting from (5.3) but inserting $\phi(x) = 1 - x$ instead of $l(x)$ gives
$$\begin{aligned} &= \sum\nolimits_{i=1}^n 1 - \hat{y}_i x_i^T w \\ &= n - \left(\sum\nolimits_{i=1}^n \hat{y}_i x_i\right)^T w \end{aligned}$$

This is a linear function and therefore unbounded below unless $\sum_{i=1}^n \hat{y}_i x_i = 0$. Minimizing this loss would simply yield $-\infty$ and no w that attain it.

Solution 4.12

1. Using the notation from the assignment we see that

$$\begin{aligned} c_{y_i}^A(x_i) &= x_i^T w_{y_i} - \tfrac{1}{K} \sum\nolimits_{k=1}^K x_i^T w_k \\ &= [...0, 1, 0...] \begin{bmatrix} \vdots \\ x_i^T w_{y_i - 1} \\ x_i^T w_{y_i} \\ x_i^T w_{y_i+1} \\ \vdots \end{bmatrix} - \tfrac{1}{K}[1, ..., 1] \begin{bmatrix} x_i^T w_1 \\ \vdots \\ x_i^T w_K \end{bmatrix} \\ &= e_{y_i}^T \begin{bmatrix} \vdots \\ x_i^T w_{y_i - 1} \\ x_i^T w_{y_i} \\ x_i^T w_{y_i+1} \\ \vdots \end{bmatrix} - \tfrac{1}{K}\mathbf{1}^T \begin{bmatrix} x_i^T w_1 \\ \vdots \\ x_i^T w_K \end{bmatrix} \\ &= e_{y_i}^T \begin{bmatrix} x_i^T & & \\ & \ddots & \\ & & x_i^T \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_K \end{bmatrix} - \tfrac{1}{K}\mathbf{1}^T \begin{bmatrix} x_i^T & & \\ & \ddots & \\ & & x_i^T \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_K \end{bmatrix} \\ &= e_{y_i}^T X_i^T w - \tfrac{1}{K}\mathbf{1}^T X_i^T w = A_i^T X_i^T w \end{aligned}$$

Inserting this into the loss problem

$$\min_w \sum\nolimits_{i=1}^n \max(0, 1 - c_{y_i}^A(x_i)) + \tfrac{\gamma}{2}\|w\|_2^2$$

gives the desired problem formulation.

2. Inserting the confidence for one data point in the hinge-loss yields

$$\begin{aligned}
h(c_{y_i}^M(x_i)) &= \max(0, 1 - c_{y_i}^M(x_i)) \\
&= \max(0, 1 - (x_i^T w_{y_i} - \max_{k \neq y_i} x_i^T w_k)) \\
&= \max(0, 1 - x_i^T w_{y_i} + \max_{k \neq y_i} x_i^T w_k) \\
&= \max(0, \max_{k \neq y_i} 1 - x_i^T w_{y_i} + x_i^T w_k) \\
&= \max_k \begin{cases} 0 & \text{if } k = y_i \\ 1 - x_i^T w_{y_i} + x_i^T w_k & \text{otherwise} \end{cases} \\
&= \max \begin{bmatrix} \vdots \\ 1 \\ 0 \\ 1 \\ \vdots \end{bmatrix} - \begin{bmatrix} \vdots \\ x_i^T w_{y_i} \\ x_i^T w_{y_i} \\ x_i^T w_{y_i} \\ \vdots \end{bmatrix} + \begin{bmatrix} \vdots \\ x_i^T w_{y_i - 1} \\ x_i^T w_{y_i} \\ x_i^T w_{y_i + 1} \\ \vdots \end{bmatrix} \\
&= \max d_{y_i} - \begin{bmatrix} \vdots \\ 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} \begin{bmatrix} \ldots & 0 & 1 & 0 & \ldots \end{bmatrix} \begin{bmatrix} \vdots \\ x_i^T w_{y_i - 1} \\ x_i^T w_{y_i} \\ x_i^T w_{y_i + 1} \\ \vdots \end{bmatrix} + X_i^T w \\
&= \max d_{y_i} - \mathbf{1} e_{y_i}^T X_i^T w + X_i^T w \\
&= \max d_{y_i} - (\mathbf{1} e_{y_i}^T - I) X_i^T w \\
&= \max d_{y_i} - M_i^T X_i^T w.
\end{aligned}$$

Sum over all data points (i) and adding the regularization gives the desired problem.

Solution 4.13

1. Exercise 3.1-6 implies that

$$f_i^*(\mu_i) = \mu_i + \iota_{[-1,0]}(\mu_i)$$

Further

$$f^*(\mu) = \sum_{i=1}^N f_i^*(\mu_i) = \sum_{i=1}^n \mu_i + \iota_{[\mathbf{-1},\mathbf{0}]}(\mu_i) = \mathbf{1}^T \mu + \iota_{[\mathbf{-1},\mathbf{0}]}(\mu).$$

We also have $g^*(\nu) = \frac{1}{2\gamma}\|\nu\|_2^2$. Hence

$$g^*(-XA\mu) = \tfrac{1}{2\gamma}\mu^T A^T X^T X A \mu.$$

Hence, the dual problem is

$$\begin{array}{ll} \text{minimize} & \mu^T\mathbf{1} + \frac{1}{2\gamma}\mu^T A^T X^T X A\mu \\ \text{subject to} & \mu \in [-1, 0] \end{array}$$

2. Since g^* is differentiable, we can recover a primal solution from the primal-dual optimality condition

$$w = \partial g^*(-XA\mu) = \frac{-1}{\gamma} XA\mu.$$

SOLUTION 4.14

1. Exercise 3.6 implies that $\max_i(u_i)^* = \iota_{\Delta_K}(\mu_i)$ where

$$\Delta_K = \{x \in \mathbb{R}^K : x_i \geq 0 \text{ and } \sum\nolimits_{i=1}^K x_i = 1\}$$

Therefore

$$\begin{aligned} f_i^*(\mu_i) &= \sup_{u_i}(\mu_i^T u_i - \max(d_{y_i} - u_i)) \\ &= [v_i = d_{y_i} - u_i] \\ &= \sup_{v_i}(\mu_i^T(d_{y_i} - v_i) - \max(v_i)) \\ &= \sup_{v_i}((-\mu_i)^T v_i - \max(v_i)) + \mu_i^T d_{y_i} \\ &= \iota_{\Delta_K}(-\mu_i) + \mu_i^T d_{y_i}. \end{aligned}$$

Further

$$f^*(\mu) = \sum\nolimits_{i=1}^n f_i^*(\mu_i) = \sum\nolimits_{i=1}^n \iota_{\Delta_K}(-\mu_i) + \mu_i^T d_{y_i}.$$

where $\mu = (\mu_1, ..., \mu_n) \in \mathbb{R}^{Kn}$.
We also have $g^*(\nu) = \frac{1}{2\gamma}\|\nu\|_2^2$. Hence

$$g^*(-XM\mu) = \frac{1}{2\gamma}\mu^T M^T X^T XM\mu.$$

Hence, the dual problem is

$$\begin{array}{ll} \text{minimize} & \mu^T\mathbf{d} + \frac{1}{2\gamma}\mu^T M^T X^T XM\mu \\ \text{subject to} & -\mu_i \in \Delta_K \end{array}$$

where $\mathbf{d} = (d_{y_1}, d_{y_2}, ..., d_{y_n}) \in \mathbb{R}^{Kn}$

2. Since g^* is differentiable, we can recover a primal solution from the primal-dual optimality condition

$$w = \partial g^*(-XM\mu) = \frac{-1}{\gamma} XM\mu.$$

Solutions to Chapter 5

Solution 5.1

That $x^\star$ is a fixed point means tha t

$$\begin{aligned} & x^\star = x^\star - \lambda \nabla f(x^\star) \\ \iff & 0 = -\lambda \nabla f(x^\star) \\ \iff & 0 = \nabla f(x^\star) \end{aligned}$$

Convexity then gives that $x^\star$ minimizes f.

Solution 5.2

That $z = \mathrm{prox}_{\gamma f}(x)$ means

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

By Fermat's rule, the argmin z satisfies

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

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

Solution 5.3

$x = \mathrm{prox}_{\gamma g}(x - \gamma \nabla f(x))$ written out is

$$x = \underset{z}{\mathrm{argmin}}\, g(z) + \tfrac{1}{2\gamma}\|z - (x - \gamma \nabla f(x))\|^2$$

which due to convexity is equivalent to

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

Fermat's rule and convexity then gives that x is a minimum of $f + g$.

Solution 5.4

1. Any unbounded function, for instance $f(x) = x$. $f(x^k)$ can go to minus infinity.

2. First, since $f(x) \geq B$ there exists a largest lower bound on $f(x^k)$, i.e. $\inf f(x^k) = b \geq B$ (This comes from completeness of the real numbers).

   Second, $f(x^k) \to b$ means that for all ϵ there exist some $N \in \mathbb{N}$ such that $|f(x^k) - b| \leq \epsilon$ for all $k \geq N$. The inequality can equivalently be written as

   $$b - \epsilon \leq f(x^k) \leq b + \epsilon.$$

   The left inequality hold by definition of $b = \inf f(x^k)$. Furthermore, there exists an N such that $f(x^N) \leq b + \epsilon$, otherwise $b + \epsilon$ is a larger lower bound than b which is a contradiction. Since $f(x^k) \leq f(x^N)$ for $k \leq N$ the right inequality, $f(x^k) \leq b + \epsilon$, hold for all $k > N$ and $f(x^k) \to b$ have been established.

3. The most basic example would be any constant sequence $x^k = x$ where x is not the minimum. A slightly more interesting example would be $f(x, y) = x^2 + y^2$ and the sequence $x^k = \sin(k) + \frac{1}{k}$ and $y^k = \cos(k)$ 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 (useless) condition for a minimization algorithm.

Solution 5.5

1. Using L-smoothness gives

   $$\begin{aligned} f(x^{k+1}) &= f(x^k - \gamma \nabla f(x^k)) \\ &\leq f(x^k) + \nabla f(x^k)^T (x^k - \gamma \nabla f(x^k) - x^k) + \tfrac{L}{2} \|x^k - \gamma \nabla f(x^k) - x^k\|^2 \\ &= f(x^k) - \gamma \|\nabla f(x^k)\|^2 + \tfrac{L}{2} \gamma^2 \|\nabla f(x^k)\|^2 \\ &= f(x^k) - \gamma (1 - \tfrac{L}{2} \gamma) \|\nabla f(x^k)\|^2 \end{aligned}$$

2. Re-arranging the above inequality gives

   $$\gamma (1 - \tfrac{L}{2} \gamma) \|\nabla f(x^k)\|^2 \leq f(x^k) - f(x^{k+1}).$$

   Summing this inequality from $k = 0$ to $k = n$ gives

   $$\begin{aligned} \textstyle\sum_{k=0}^{n} \gamma (1 - \tfrac{L}{2} \gamma) \|\nabla f(x^k)\|^2 &\leq \textstyle\sum_{k=0}^{n} f(x^k) - f(x^{k+1}) \\ &= f(x^0) - f(x^{k+1}) \\ &\leq f(x^0) - B \end{aligned}$$

   This inequality hold for all n and the RHS is bounded which gives

   $$\gamma (1 - \tfrac{L}{2} \gamma) \textstyle\sum_{k=0}^{\infty} \|\nabla f(x^k)\|^2 < \infty$$

   Since $0 < \gamma < \frac{2}{L}$ is $\gamma (1 - \frac{L}{2} \gamma)$ positive and then must $\|\nabla f(x^k)\|^2$ be summable and

   $$\|\nabla f(x^k)\| \to 0$$

3. Strong convexity means that
$$\begin{aligned} f(x) &\geq f(x^\star) + \tfrac{\mu}{2}\|x - x^\star\|^2 \\ f(x^\star) &\geq f(x) + \nabla f(x)^T(x^\star - x) + \tfrac{\mu}{2}\|x - x^\star\|^2. \end{aligned}$$
Adding these two together yields
$$\nabla f(x)^T(x - x^\star) \geq \mu\|x - x^\star\|^2.$$
Using Cauchy-Schwarz on the LHS yields
$$\|\nabla f(x)\|\|x - y\| \geq \mu\|x - x^\star\|^2 \implies \|\nabla f(x)\| \geq \mu\|x - x^\star\|.$$
Therefore, if $\|\nabla f(x^k)\| \to 0$ then $\|x - x^\star\| \to 0$.

4. The function $f(x, y) = \frac{x^2}{y}$ for $y > 0$ is convex,
$$\nabla^2 f(x, y) = \tfrac{2}{y^3}\begin{bmatrix} y \\ -x \end{bmatrix}[y - x] > 0,$$
and has the minimum $f(0, y) = 0$. The sequence $(x^k, y^k) = (k + \frac{1}{k}, k^2)$ satisfy $\nabla f(x^k, y^k) \to 0$,
$$\nabla f(x^k, y^k) = \begin{bmatrix} \frac{2x^k}{y^k} \\ \frac{-(x^k)^2}{(y^k)^2} \end{bmatrix} = \begin{bmatrix} \frac{2(k+\frac{1}{k})}{k^2} \\ \frac{-(k+\frac{1}{k})^2}{k^4} \end{bmatrix} = \begin{bmatrix} 2(\frac{1}{k} + \frac{1}{k^3}) \\ (\frac{1}{k} + \frac{1}{k^3})^2 \end{bmatrix} \to \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$
and $f(x^{k+1}, y^{k+1}) \leq f(x^k, y^k)$,
$$\begin{aligned} f(x^{k+1}, y^{k+1}) &= \tfrac{x^2}{y} = \tfrac{(k+1+\frac{1}{k+1})^2}{(k+1)^2} = (1 + \tfrac{1}{(k+1)^2})^2 \\ &\leq (1 + \tfrac{1}{k^2})^2 = \tfrac{(k+\frac{1}{k})^2}{k^2} = f(x^k, y^k). \end{aligned}$$
However, from $f(x^k, y^k) = (1 + \frac{1}{k^2})^2$ we see that $f(x^k, y^k) \to 1$.

Solution 5.6

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 5.7

The proximal gradient update is

$$x^+ = \underset{z}{\text{argmin}}\, g(z) + \tfrac{1}{2\gamma}\|z - (x - \gamma\nabla f(x))\|^2$$

Due to convexity is this equivalent to

$$0 \in \partial g(x^+) + \nabla f(x) + \tfrac{1}{\gamma}(x^+ - x) \iff -(\nabla f(x) + \tfrac{1}{\gamma}(x^+ - x)) \in \partial g(x^+).$$

From the definition of subdifferential we have

$$\begin{aligned} & g(x) \geq g(x^+) - (\nabla f(x) + \tfrac{1}{\gamma}(x^+ - x))^T(x - x^+) \\ \iff & g(x^+) \leq g(x) - (\nabla f(x) + \tfrac{1}{\gamma}(x^+ - x))^T(x^+ - x) \\ & \qquad = g(x) - \nabla f(x)^T(x^+ - x) - \tfrac{1}{\gamma}\|x^+ - x\|^2 \end{aligned}$$

L-smoothness of f yields

$$f(x^+) \leq f(x) + \nabla f(x)^T(x^+ - x) + \tfrac{L}{2}\|x^+ - x\|^2.$$

Adding them together yields

$$f(x^+) + g(x^+) \leq f(x) + g(x) - (\tfrac{1}{\gamma} - \tfrac{L}{2})\|x^+ - x\|^2.$$

Here we see that we get descent for $\gamma < \frac{2}{L}$.

Solution 5.8

1. The function is smooth so gradient and coordinate-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, (coordinate) proximal gradient works but (coordiante) gradient-descent doesn't. Last part is separable so coordinate proximal gradient is efficient.

3. Both functions are smooth, second is separable and easy to prox on. (Coordinate) Gradient descent and (Coordinate) proximal gradient all work.

4. First function is smooth, second is easy to prox on but not separable nor smooth. Proximal gradient all 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. (Coordinate) proximal gradient works.

8. The second term is neither smooth nor simple to prox on, nether of the methods would be efficient.

9. From Excercise 4.5 we know that the first part is smooth, and the second part is trivially smooth, separable and easy to prox on. (Coordinate) Gradient descent and (coordinate) proximal gradient therefore works. However, the coordinate-wise algorithms will not be efficient, as shown in Excercise 5.10.

Solution 5.9

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, so there is no step length γ that can be selected that guarantees convergence. Thus neither of the methods work.

2. $\frac{1}{2}x^TQx + b^Tx$ is strongly convex so its dual is smooth. The dual of the last part is separable and easy to prox on but not smooth. (Coordinate) 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 the problem is $\min_x f(Ax) + g(x)$ and the dual can be written $\min_\mu f^*(\mu) + g^*(Ax)$. $f^*(\mu)$ is separable, smooth and proximable and $g^*(Ax)$ is smooth. 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 well.

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 and separable so the same is true for the dual. (Coordinate) 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 seperable and proximable so (conjugate) 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 5.10

- 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 5.11. The gradient of the first term $f(x) = \log(1 + e^{-w^Tx})$ is $\nabla f(x) = -w\frac{e^{-w^Tx}}{1+e^{-w^Tx}}$ 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 $(\text{prox}_{\gamma g}(x))_i = \max(0, x_i/(1+\gamma))$.
  - 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 5.11

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

$$(\textbf{prox}_{\gamma g}(z))_i = \textbf{prox}_{\gamma g_i}(z_i)$$

hence

$$x_i^{k+1} = (\textbf{prox}_{\gamma g}(x^k - \gamma \nabla f(x^k)))_i = \textbf{prox}_{\gamma g_i}((x^k - \gamma \nabla f(x^k)))_i) =$$
$$\textbf{prox}_{\gamma g_i}(x_i^k - \gamma(\nabla f(x^k))_i) = \max(0, (x_i^k - \gamma(\nabla f(x^k))_i)/(1+\gamma)).$$

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 5.12

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 5.13

We have

$$\|x^k - x^\star\| = \|Tx^{k-1} - x^\star\| = \|Tx^{k-1} - Tx^\star\| \le L\|x^{k-1} - x^\star\|.$$

Iterating this inequality back to $k = 0$ yields

$$\|x^k - x^\star\| \le L^k \|x^0 - x^\star\|.$$

Since $0 < L < 1$ the RHS goes to zero as $k \to \infty$, meaning $\|x^k - x^\star\| \to 0$ as $k \to \infty$.

Solution 5.14

- We have

  $$f(x) = \frac{1}{2}(Ax-b)^T(Ax-b) = \frac{1}{2}x^T A^T A x - b^T A x + \frac{1}{2}b^T b$$

  so

  $$f_{i,x}(\alpha) = \frac{1}{2}(x + \alpha e_i)^T A^T A (x + \alpha e_i) - b^T A (x + \alpha e_i) + \frac{1}{2} b^T b \quad (5.4)$$
  $$= \frac{1}{2}\alpha^2 e_i^T A^T A e_i - \alpha b^T A x e_i + \ldots \quad (5.5)$$

  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 $\|A^TA\|_2$ and

$$\begin{aligned}\|A^TA\|_2 = \sup_x \frac{\|A^TAx\|_2}{\|x\|_2} &\geq \frac{\|A^TAe_i\|_2}{\|e_i\|_2} = \|A^TAe_i\|_2 \\ &= \Big(\sum_j (A^TA)^2_{j,i}\Big)^{\frac{1}{2}} \geq \left((A^TA)^2_{i,i}\right)^{\frac{1}{2}} = (A^TA)_{i,i}.\end{aligned}$$

Solution 5.15

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 5.16

The exact same reasoning as Exercise 5.15 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 5.17

For implementations, see appendix. The function values are show in Figure 5.1.

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.

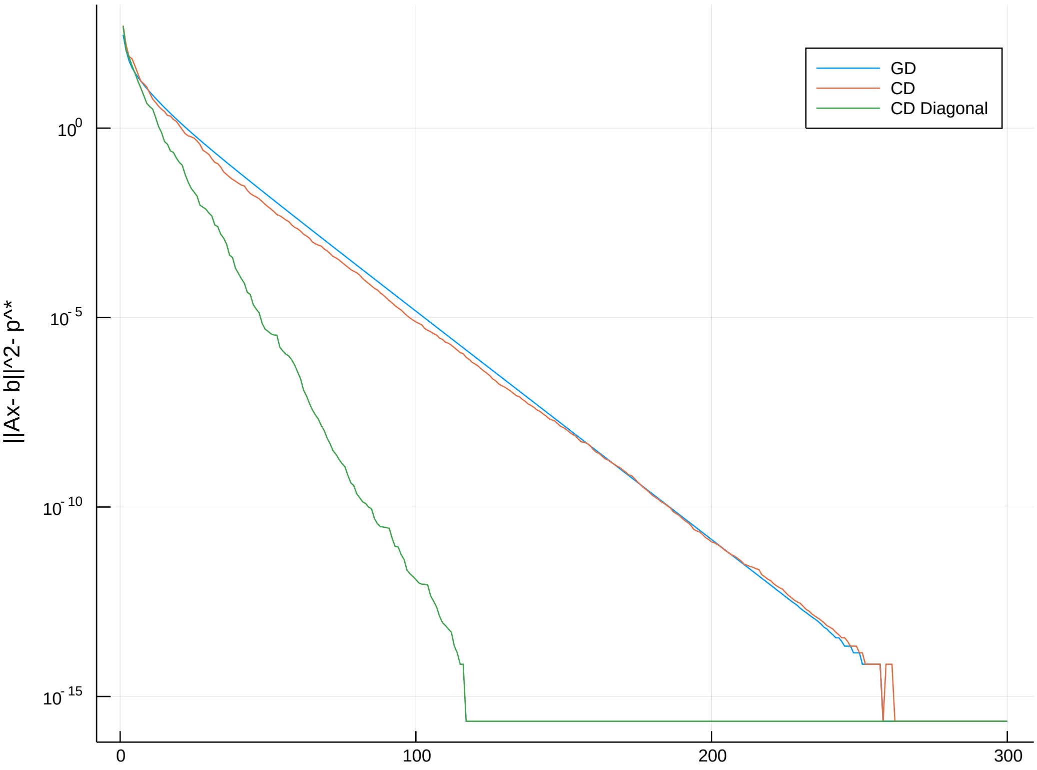


Figure 5.1: Function value for each iteration over full data with Gradient Descent, Coordinate Descent and Coordinate Descent with Diagonal scaling.

Solution 5.18

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

$$\begin{aligned}\textstyle\sum_{k=0}^{n}\gamma^k(1 - \tfrac{L}{2}\gamma^k)\mathbb{E}[\|\nabla F(x^k)\|^2] - (\gamma^k)^2\tfrac{L\sigma^2}{2} &\leq \mathbb{E}[F(x^0) - F(x^{k+1})]\\ &\leq \mathbb{E}F(x^0) - B\end{aligned}$$

since $F(x) \geq B$. Letting $n \to \infty$ gives

$$\textstyle\sum_{k=0}^{\infty}\gamma^k(1 - \tfrac{L}{2}\gamma^k)\mathbb{E}[\|\nabla F(x^k)\|^2] - (\gamma^k)^2\tfrac{L\sigma^2}{2} < \infty. \qquad (5.6)$$

Inserting γ^k gives

$$\textstyle\sum_{k=0}^{\infty}\tfrac{1}{2L}\mathbb{E}[\|\nabla F(x^k)\|^2 - \sigma^2] < \infty.$$

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 (5.6) yield

$$\textstyle\sum_{k=0}^{\infty}(\tfrac{1}{k} - \tfrac{L}{2}\tfrac{1}{k^2})\mathbb{E}[\|\nabla F(x^k)\|^2] - \tfrac{1}{k^2}\tfrac{L\sigma^2}{2} < \infty.$$

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 (5.6) 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 5.19

For implementations, see appendix. The function values are show in Figure 5.2.

- 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.

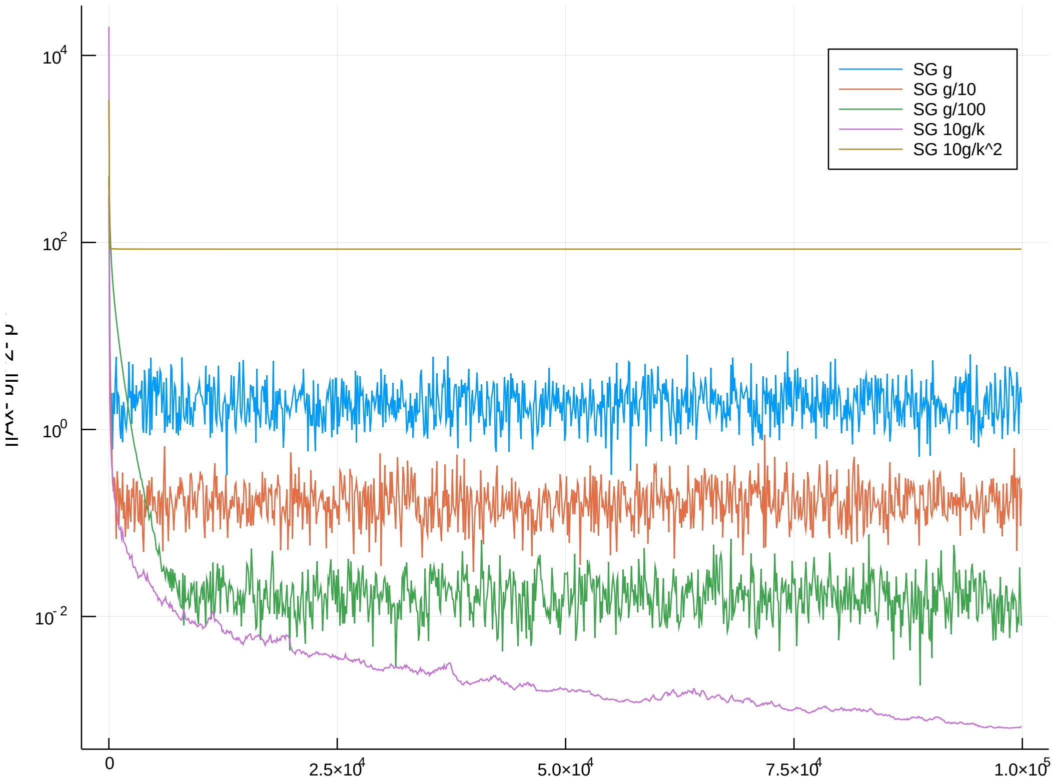


Figure 5.2: Stochastic gradient for different step lengths from Excercise 5.2, where $g = \lambda_{\max}$.

Julia Code

Implementation of Excercise 5.17

```
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 5.19

```
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
```