Convex Optimization for Statistics and Machine Learning, Volume I: Analysis

Ryan J. Tibshirani

Contents

Part 1

Introduction

Why Read This Book?

How To Read This Book

Part 2

Fundamentals

Convex Sets and Functions

3.1. Convex sets

Convex sets are the main building blocks for the study of convex functions and their properties. Essentially everything that can be said about convex functions can be traced back to a statement about convex sets. A set $C \subseteq \mathbb{R}^d$ is called *convex* if it satisfies

$$x, y \in C \implies tx + (1-t)y \in C, \quad \text{for all } t \in [0,1]. \tag{3.1}$$

This says that the line segment $\{tx + (1-t)y : t \in [0,1]\}$ joining x and y lies entirely in C. See Figure 3.1. A *convex combination* of points $x_1, \ldots, x_n \in \mathbb{R}^d$ is one of the form

$$\sum_{i=1}^n t_i x_i, \quad \text{where } t_i \geq 0, \text{ for } i = 1, \ldots, n \text{ and } \sum_{i=1}^n t_i = 1.$$

The *convex hull* of C is the set of all convex combinations of points in C,

$$\operatorname{conv}(C) = \left\{ \sum_{i=1}^n t_i x_i : n \geq 1,\, x_i \in C,\, t_i \geq 0, \text{ for } i = 1, \ldots, n, \text{ and } \sum_{i=1}^n t_i = 1 \right\}.$$

The convex hull $\operatorname{conv}(C)$ is itself a convex set, for any set C; in fact, it is the smallest convex set containing C, meaning $\operatorname{conv}(C) \subseteq D$ for any convex set $D \supseteq C$.

Example 3.1. The following are examples of convex sets.

a. The empty set $\emptyset$, and all of Euclidean space $\mathbb{R}^d$.
b. A line $\{x + ty : t \in \mathbb{R}\}$, ray $\{x + ty : t \geq 0\}$, and line segment $\{x + ty : t \in [0,1]\}$.
c. A linear subspace $\{x : Ax = 0\}$, and affine subspace $\{x : Ax = b\}$.
d. A hyperplane $\{x : a^\mathsf{T} x = b\}$, and halfspace $\{x : a^\mathsf{T} x \leq b\}$.
e. A *norm ball* $\{x : \|x\| \leq t\}$, where $\|\cdot\|$ is a norm on $\mathbb{R}^d$, and $t \geq 0$.
f. A *polyhedron* $\{x : a_i^\mathsf{T} x \leq b_i,\, i = 1, \ldots, m\}$. This is the intersection of a finite number of halfspaces. We can express this succinctly as $\{x : Ax \leq b\}$, where here and throughout we read the inequality componentwise.

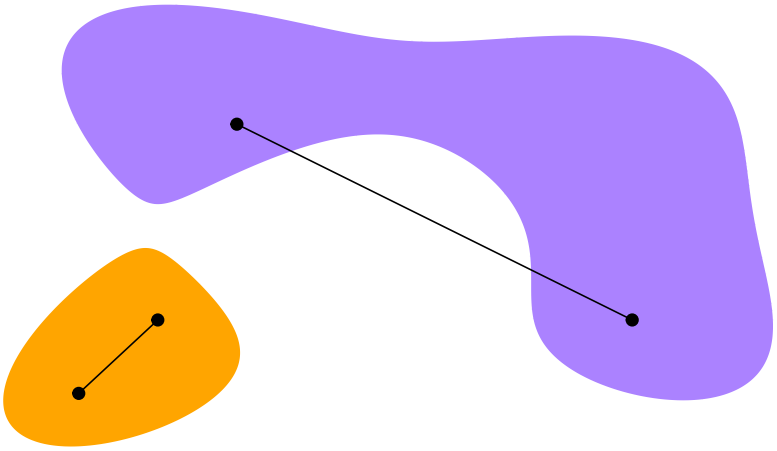

Figure 3.1. Lower left: convex set, such that the line segment joining any two elements will lie entirely in the set. Upper right: nonconvex set, such that this does not hold.

It will be important to think about convexity for sets of matrices (and convexity for functions of matrices), since some interesting optimization problems are formulated over matrices. By viewing $\mathbb{R}^{k\times d}$—the space of real $k \times d$ matrices—as a vector space of dimension kd, everything we cover for convex sets in $\mathbb{R}^d$ (and convex functions on $\mathbb{R}^d$) can be translated over to $\mathbb{R}^{k\times d}$.

The same can be said about $\mathbb{S}^d$—the space of symmetric real $d \times d$ matrices—which we can view as a vector space of dimension $d(d+1)/2$. A subset of $\mathbb{S}^d$ of particular interest is

$$\mathbb{S}^d_+ = \{X \in \mathbb{S}^d : X \succeq 0\}. \tag{3.2}$$

Here we write $X \succeq 0$ to denote that X is *positive semidefinite*: it satisfies $a^\mathsf{T} X a \geq 0$, for all $a \in \mathbb{R}^d$. We call $\mathbb{S}^d_+$ the *positive semidefinite cone* (of dimension d). This is a convex set, indeed a *convex cone* which means it satisfies $X, Y \in \mathbb{S}^d \implies sX + tY \in \mathbb{S}^d$ for all $s, t \geq 0$. To see this, note that for any $s, t \geq 0$ and $a \in \mathbb{R}^d$, we have

$$a^\mathsf{T}(sX + tY)a = sa^\mathsf{T}Xa + ta^\mathsf{T}Ya \geq 0,$$

provided that X, Y are positive semidefinite to begin with.

Given a set C and a point $x \in C$, another interesting convex cone is the *normal cone* to C at x,

$$\mathcal{N}_C(x) = \{g : g^\mathsf{T}x \geq g^\mathsf{T}y, \text{ for all } y \in C\}. \tag{3.3}$$

For any set C (convex or not), the associated normal cone $\mathcal{N}_C(x)$ at any $x \in C$ is always a convex cone (which can be verified from the definition).

We finish this short section with two key theorems on properties of convex sets.

Theorem 3.2 (Separating hyperplane theorem)**.** *If C, D are nonempty disjoint convex sets, then there exists $a \neq 0$ and b such that $C \subseteq \{x : a^\mathsf{T}x \leq b\}$ and $D \subseteq \{x : a^\mathsf{T}x \geq b\}$. The set $\{x : a^\mathsf{T}x = b\}$ is called a* separating hyperplane *between C, D.*

Theorem 3.3 (Supporting hyperplane theorem)**.** *If C is a convex set and $x_0 \in \mathrm{bd}(C)$ (the boundary of C), then there exists $a \neq 0$ and b such that $a^\mathsf{T}x_0 = b$ and $C \subseteq \{x : a^\mathsf{T}x \leq b\}$. The set $\{x : a^\mathsf{T}x = b\}$ is called a* supporting hyperplane *to C at x_0.*

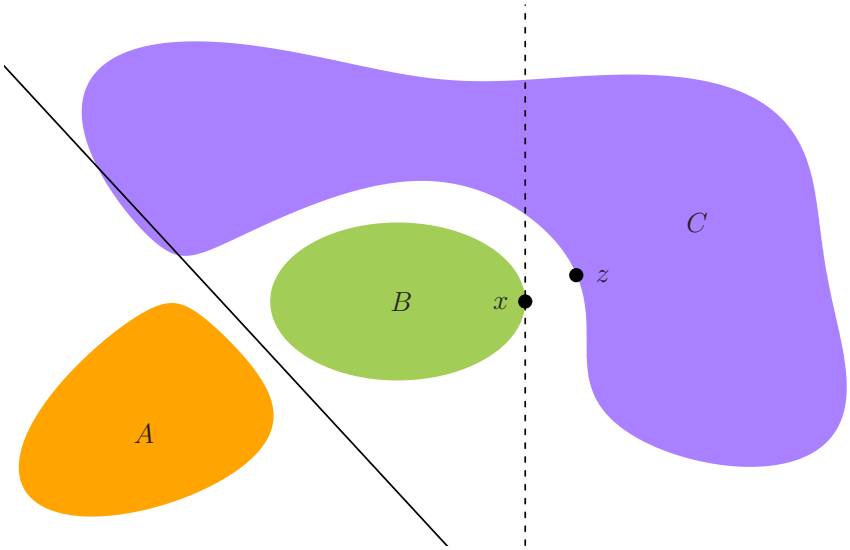

Figure 3.2. Lower left, A and B: two disjoint convex sets, which must hence admit a separating hyperplane, illustrated as the solid line running between them. The set B, again by convexity, has a supporting hyperplane at every boundary point, illustrated by the dashed line supporting it at x. Upper right, B and C: two disjoint sets which have no separating hyperplane, which is possible as C is nonconvex. The set C has no supporting hyperplane at z, possible again by nonconvexity.

The separating hyperplane theorem can be proved using basic arguments, and the supporting hyperplane theorem can be proved from the separating hyperplane theorem. These results are highly intuitive and the role of convexity can be made clear pictorially, see Figure 3.2. Furthermore, they have important consequences in optimization. For example, the separating hyperplane theorem can be used to prove what are called *theorems of alternatives* (such as *Farkas' lemma*); see Exercise 3.8. Also, the supporting hyperplane theorem can be used to prove that subgradients always exist for a convex function (on the relative interior of its effective domain); see Exercise 3.9.

3.2. Convex functions

We now move from sets to functions, which may be a more familiar object of study to some readers. For reasons that will be apparent, it is convenient to allow functions to take both real and infinite values. Throughout, by default (without further specification), a function is defined on all of $\mathbb{R}^d$ and takes values in the *extended real numbers* $[-\infty, \infty]$. For such $f : \mathbb{R}^d \to [-\infty, \infty]$, we denote

$$\operatorname{dom}(f) = \{x : f(x) < \infty\},$$

and call this set the *effective domain* of f. Note that, as we are allowing functions to take infinite values, there is no loss of generality in considering functions defined on all of $\mathbb{R}^d$: given $S \subseteq \mathbb{R}^d$ and $f : S \to [-\infty, \infty]$, we can always be extend f to all of $\mathbb{R}^d$ by setting it equal to ∞ outside of S.

We say that $f : \mathbb{R}^d \to (-\infty, \infty]$ is *convex* if $\operatorname{dom}(f)$ is a convex set and

$$f\big(tx + (1-t)y\big) \le tf(x) + (1-t)f(y), \quad \text{for all } x, y \in \operatorname{dom}(f) \text{ and } t \in [0,1]. \tag{3.4}$$

This says that the line segment joining $(x, f(x))$ and $(y, f(y))$ lies above the graph of f, as shown in Figure 3.3. We call f *strictly convex* provided that strict inequality holds in the above statement,

$$f\big(tx + (1-t)y\big) < tf(x) + (1-t)f(y), \quad \text{for all } x \ne y \in \operatorname{dom}(f) \text{ and } t \in (0,1). \tag{3.5}$$

In a sense, this says that f is "more convex" than a linear function, as a linear function would (by definition) have equal left and right hand sides in (3.5). A stronger notion than strict convexity is

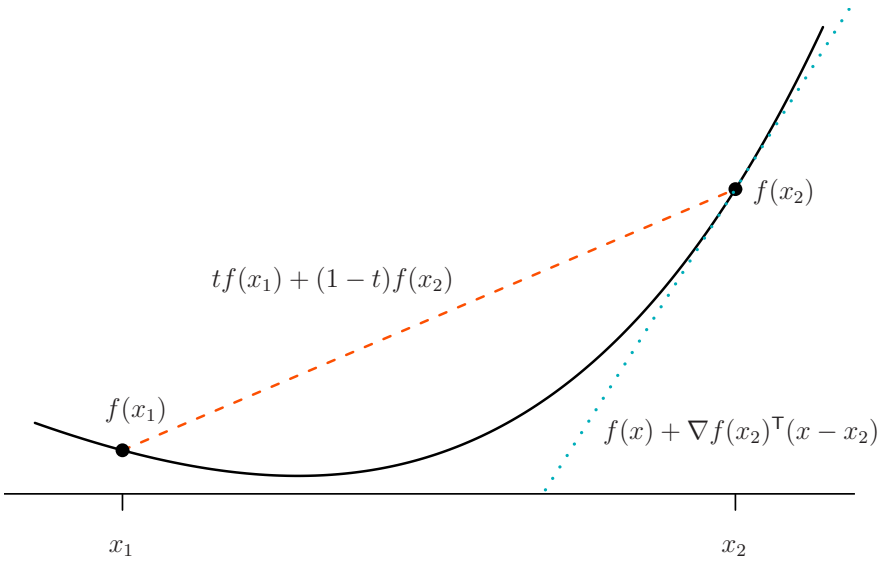

Figure 3.3. Convex function f, such that the line segment between any two points on its graph lies above the function, as illustrated by the dashed line joining $(x_1, f(x_1))$ and $(x_2, f(x_2))$. Since f is differentiable, convexity is equivalent to f lying everywhere above its tangent line at any point, as illustrated by the dotted line running tangent to f at x_2.

strong convexity, which means, for a parameter $m > 0$, the function f_m defined by

$$f_m(x) = f(x) - \frac{m}{2}\|x\|_2^2$$

is convex. Like strict convexity requires more curvature than a linear function, strong convexity requires that f be "more convex" than a quadratic function.

A companion notion to convexity is *concavity*. A function $f : \mathbb{R}^d \to [-\infty, \infty)$ is called *concave* provided that $-f$ is convex, or equivalently: $\mathrm{dom}(-f)$ is a convex set and

$$f\big(tx + (1-t)y\big) \geq tf(x) + (1-t)f(y), \quad \text{for all } x, y \in \mathrm{dom}(-f) \text{ and } t \in [0,1]. \tag{3.6}$$

Similarly, we say that f is *strictly concave* or *strongly concave* provided that $-f$ is strictly convex or strongly convex, respectively.

In general, when it is not clear from the context, we will say that f convex "on C" to indicate that we are interpreting its effective domain to be $\mathrm{dom}(f) = C$ (think: we set f to be ∞ outside of C), and similarly for concave functions.

Example 3.4. The following are examples of convex or concave functions. In all cases, the expressions should be interpreted as functions of x.

a. The power function x^a is convex on $\mathbb{R}_+ = \{x : x \geq 0\}$ (the nonnegative real numbers) for $a \geq 1$, and concave on $\mathbb{R}_{++} = \{x : x > 0\}$ (the positive real numbers) for $0 < a < 1$. It is also convex on $\mathbb{R}_{++}$ for $a < 0$.
b. The exponential function e^x is convex on $\mathbb{R}$. The logarithm function $\log(x)$ is concave on $\mathbb{R}_{++}$. The *negative entropy* function $x \log(x)$ is convex on $\mathbb{R}_{++}$.
c. A linear function $a^\mathsf{T}x + b$ is both convex and concave on $\mathbb{R}^d$.
d. A quadratic function $\frac{1}{2}x^\mathsf{T}Ax + b^\mathsf{T}x + c$ is convex on $\mathbb{R}^d$ for $A \succeq 0$ (positive semidefinite) and concave on $\mathbb{R}^d$ for $A \preceq 0$ (negative semidefinite).

e. A norm $\|x\|$ is always convex. For example, on $\mathbb{R}^d$ we have the ℓ_p *norms*, defined as:

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p}, \quad \text{for } p \geq 1, \tag{3.7}$$

$$\|x\|_\infty = \max_{i=1,\dots,d} |x_i|, \tag{3.8}$$

and on $\mathbb{R}^{k\times d}$ we have the *trace norm* and *operator norm*, defined as:

$$\|X\|_{\mathrm{tr}} = \sum_{i=1}^{k} \sigma_i(X), \tag{3.9}$$

$$\|X\|_{\mathrm{op}} = \sigma_1(X), \tag{3.10}$$

respectively (these are also called the *nuclear norm* and *spectral norm*, respectively). Here $\sigma_1(X) \geq \cdots \geq \sigma_k(X) \geq 0$ denote the singular values of X.

f. The *characteristic function* of a set C,

$$I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}, \tag{3.11}$$

is convex provided that C is a convex set.

g. The *support function* of a set C,

$$h_C(x) = \sup_{z\in C} z^{\mathsf{T}} x, \tag{3.12}$$

is always convex (regardless of the set C).

An important property of convex functions is *Jensen's inequality*, which says that for a convex function $f : \mathbb{R}^d \to (-\infty, \infty]$ and a random variable X that is supported on $\mathrm{dom}(f)$, we have

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)], \tag{3.13}$$

provided the expectations exist. Oppositely, if f is concave then $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$, again provided the expectations exist. These inequalities can be seen as generalizations of the defining properties of convexity and concavity, (3.4) and (3.6), respectively, from discrete to arbitrary distributions.

Remark 3.5. In this book, as per our definitions above, we do *not allow convex functions to take the value* $-\infty$ and do *not allow concave functions to take the value* ∞. This is different from the approach taken by some other authors and it excludes what are known as *improper* convex functions from consideration; see the chapter notes for further discussion. Restricting convex and concave functions in this way will be sufficient for our purposes, and simplifies matters that would otherwise require a more nuanced treatment.

To see one advantage of restricting convex functions to take values in $(-\infty, \infty]$, observe that for $f : \mathbb{R}^d \to (-\infty, \infty]$ with $\mathrm{dom}(f)$ convex, we can equivalently write (3.4) as

$$f\big(tx + (1-t)y\big) \leq tf(x) + (1-t)f(y), \quad \text{for all } x, y \in \mathbb{R}^d \text{ and } t \in [0,1].$$

(This is possible because we never encounter the undefined expressions $\infty - \infty$ or $-\infty + \infty$ in the above statement, as f cannot take the value $-\infty$.) This offers an alternative view for convexity, based on well-defined arithmetic within the extended real numbers, that can be more fluid when working with functions that can take infinite values.

3.3. Equivalent characterizations

There are a number of interesting alternative characterizations of convexity for functions (conditions other than the definition that are either equivalent to convexity or imply convexity), outlined next.

A. Epigraph characterization. A function $f : \mathbb{R}^d \to (-\infty, \infty]$ is convex if and only if

$$\text{epi}(f) = \{(x, t) \in \text{dom}(f) \times \mathbb{R} : f(x) \leq t\}, \tag{3.14}$$

called its *epigraph*, is a convex set.

B. Lines characterization. A function $f : \mathbb{R}^d \to (-\infty, \infty]$ convex if and only if the restriction of f to every line is convex, that is, the function $g(t) = f(x + tv)$ is convex on $\{t : x + tv \in \text{dom}(f)\}$ for all $x, v \in \mathbb{R}^d$.

C. First-order characterization. A differentiable function $f : \mathbb{R}^d \to (-\infty, \infty]$ is convex if and only if $\text{dom}(f)$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^\mathsf{T}(y - x), \quad \text{for all } x, y \in \text{dom}(f), \tag{3.15}$$

where $\nabla f(x)$ denotes the gradient of f at x. This says that first-order Taylor approximation to f around any point x must globally under-approximate f; this is illustrated in Figure 3.3. A similar result holds for strict convexity: a differentiable function $f : \mathbb{R}^d \to (-\infty, \infty]$ is strictly convex if and only if $\text{dom}(f)$ is convex and

$$f(y) > f(x) + \nabla f(x)^\mathsf{T}(y - x), \quad \text{for all } x \neq y \in \text{dom}(f).$$

D. Second-order characterization. A twice differentiable function $f : \mathbb{R}^d \to (-\infty, \infty]$ is convex if and only if $\text{dom}(f)$ is convex and

$$\nabla^2 f(x) \succeq 0, \quad \text{for all } x \in \text{dom}(f), \tag{3.16}$$

where $\nabla^2 f(x)$ denotes the Hessian of f at x. This condition says that the function f, at any point x, must be "curved upwards" in any direction; more precisely, for $v \in \mathbb{R}^d$, letting $g(t) = f(x + tv)$, we see that $g''(0) = v^\mathsf{T} \nabla^2 f(x) v \geq 0$ (by positive semidefiniteness). For twice differentiable $f : \mathbb{R}^d \to (-\infty, \infty]$, if $\text{dom}(f)$ is convex and

$$\nabla^2 f(x) \succ 0, \quad \text{for all } x \in \text{dom}(f),$$

meaning $\nabla^2 f(x)$ is positive definite for all $x \in \text{dom}(f)$, then f is strictly convex. However, we note that this is *not* a necessary condition; for example, $f(x) = x^4$ is strictly convex but $f''(0) = 0$.

For concavity, each of the above equivalent characterizations has an analog (we just apply the characterization to $-f$). For alternative characterizations of strong convexity, see Theorem 3.9.

Example 3.6. The convexity or concavity of the following two functions can be confirmed using the second-order characterization (3.16).

a. The *log-sum-exp function*, defined by

$$f(x) = \log\left(\sum_{i=1}^{d} e^{x_i}\right),$$

is convex on $\mathbb{R}^d$.

b. The *log-det function*, defined by $f(X) = \log(\det(X))$, is concave on $\mathbb{S}^d_{++} = \{X \in \mathbb{S}^d : X \succ 0\}$ (the space of positive definite $d \times d$ matrices).

3.4. Operations preserving convexity

To check the convexity of a given set or function, we can of course apply the definition directly, or for functions, we can appeal to the alternative characterizations from Chapter 3.3. For complicated sets or functions, this can become tedious. It is often easier to instead (a) memorize a few key base examples of convex sets and functions (for example, those given in Chapters 3.1 and 3.2), and (b) remember that various transformations preserve convexity (as we will detail below). Then, checking convexity for a given set or function becomes a task of trying to relate it to one of the key base examples by one (or any sequence of) convexity-preserving transformations.

Here are some useful operations that preserve convexity for sets.

A. Intersection. If C_s is convex for each $s \in S$, then the intersection $\cap_{s \in S} C_s$ is convex.

B. Scaling and translation. If $C \subseteq \mathbb{R}^d$ is convex, $a \in \mathbb{R}$, and $b \in \mathbb{R}^d$, then

$$aC + b = \{ax + b : x \in C\}$$

is convex.

C. Linear images and preimages. If $C \subseteq \mathbb{R}^d$ is convex and $f(x) = Ax + b$ for $A \in \mathbb{R}^{k \times d}$ and $b \in \mathbb{R}^k$, then the image of C under f,

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

is convex. Also, if $D \subseteq \mathbb{R}^k$ is convex, then the preimage (or inverse image) of D under f,

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

is convex.

D. Perspective images and preimages. The $(d+1)$-variate *perspective function* is defined for $x \in \mathbb{R}^d$ and $z > 0$ as

$$P(x, z) = x/z.$$

Note $\mathrm{dom}(P) = \mathbb{R}^d \times \mathbb{R}_{++}$ (where recall $\mathbb{R}_{++}$ denotes the positive reals). If $C \subseteq \mathrm{dom}(P)$ is convex then the perspective image $P(C)$ is convex; and if $D \subseteq \mathbb{R}^d$ is convex then the perspective preimage $P^{-1}(D)$ is convex.

E. Linear-fractional images and preimages. A *linear-fractional function* is the composition of a perspective function with a linear function, that is, a function f of the form

$$f(x) = \frac{Ax + b}{c^\mathsf{T} x + e},$$

with $\mathrm{dom}(f) = \{x : c^\mathsf{T} x + e > 0\}$. Supposing that $A \in \mathbb{R}^{k \times d}$, if $C \subseteq \mathbb{R}^d$ is convex and $C \subseteq \mathrm{dom}(f)$ then the linear-fractional image $f(C)$ is convex; and if $D \subseteq \mathbb{R}^k$ is convex then the linear-fractional preimage $f^{-1}(D)$ is convex.

Here are now some useful operations that preserve convexity for functions (note the analogy to the operations for sets, in many cases).

F. Nonnegative linear combination. If $f_1, \ldots, f_n$ are convex functions and $a_1, \ldots, a_n \geq 0$ then the nonnegative linear combination $F = a_1 f_1 + \cdots + a_n f_n$, the function defined as

$$F(x) = a_1 f_1(x) + \cdots + a_n f_n(x)$$

is convex.

G. Partial supremum. Let f be a function acting on a block variable (x, z). If $f(\cdot, z)$ is convex for each $z \in Z$ (meaning, $x \mapsto f(x, z)$ is convex), then the partial supremum $F = \sup_{z \in Z} f(\cdot, z)$, the function defined as

$$F(x) = \sup_{z \in Z} f(x, z),$$

is convex. An important special case, corresponding to a finite set Z: if f_i, $i = 1, \ldots, k$ are convex, then their pointwise maximum F, defined as $F(x) = \max_{i=1,\ldots,k} f_i(x)$, is convex.

H. Partial infimum. Let f be a function acting on a block variable (x, z). If f is convex (to be clear and to emphasize the difference to the above, here we mean $(x, z) \mapsto f(x, z)$ is convex) and Z is a convex set, then the partial infimum $F = \inf_{z \in Z} f(\cdot, z)$, the function defined as

$$F(x) = \inf_{z \in Z} f(x, z),$$

is convex, provided that F is nowhere equal to $-\infty$.

I. Linear composition. Let $f : \mathbb{R}^k \to (-\infty, \infty]$ be convex, and let $A \in \mathbb{R}^{k \times d}$, $b \in \mathbb{R}^d$. Then the function F defined as $F(x) = f(Ax + b)$ is convex.

J. Perspective transformation. The *perspective transform* of f is the function defined as

$$F(x, t) = tf(x/t),$$

with $\mathrm{dom}(F) = \mathrm{dom}(f) \times \mathbb{R}_{++}$. If f is convex, then so is its perspective transform.

K. General composition. Let $f : \mathbb{R}^k \to (-\infty, \infty]$, $g : \mathbb{R}^d \to \mathbb{R}^k$, and denote their composition by $F = f \circ g$, that is, $F(x) = f(g(x))$. Then F is convex if f is convex and, for each $i = 1, \ldots, k$, either of the following holds (where g_i denotes the i^{th} component function of g):

- f is nondecreasing in its i^{th} argument and g_i is convex; or
- f is nonincreasing in its i^{th} argument and g_i is concave.

To develop intuition for this rule, it helps to think of univariate twice differentiable f, g (though the rule of course applies more generally). In this case, we can use the chain rule to compute

$$F''(x) = f''(g(x))(g'(x))^2 + f'(g(x))g''(x).$$

In order to have $F'' \geq 0$, we see that it suffices to have $f'' \geq 0$ and $g'' \geq 0$ (that is, f and g convex) as well as $f' \geq 0$ (that is, f nondecreasing). It would also work to have $f'' \geq 0$ and $g'' \leq 0$ (that is, f convex and g concave) as well as $f' \leq 0$ (that is, f nonincreasing).

Example 3.7. The following claims about convexity can be checked by identifying the right base examples and convexity-preserving transformations.

a. For $A_1, \ldots, A_d, B \in \mathbb{S}^n$, the set $\{x : x_1 A_1 + x_2 A_2 + \cdots + x_d A_d \preceq B\}$ is convex.
b. For $C \subseteq \mathbb{R}^d$ and a norm $\|\cdot\|$, the function giving the supremum distance to C,

$$f(x) = \sup_{z \in C} \|x - z\|,$$

is convex.
c. For $C \subseteq \mathbb{R}^d$ and a norm $\|\cdot\|$, the function giving the infimum distance to C,

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

is convex, provided that C is convex.

Many common loss functions in statistical estimation and machine learning are convex, and this can be readily checked using the results from this section or the last; see Exercise 3.6.

3.5. Smoothness and growth*

We say that a function F is *Lipschitz continuous* with parameter $L > 0$ if

$$\|F(x) - F(y)\|_2 \leq L\|x - y\|_2, \quad \text{for all } x, y \in \mathrm{dom}(f).$$

We say that a function f is *Lipschitz smooth* with parameter $L > 0$ provided f is differentiable and its gradient $F = \nabla f$ satisfies the above condition.

Lipschitz continuity can be seen as a type of growth condition on the function in question. To develop intuition for this, consider the case of a real-valued function F: Lipschitz continuity means that, along any line segment, F cannot grow faster than a linear function with slope L. Thus for real-valued f, when ∇f is Lipschitz continuous (f is Lipschitz smooth), one might imagine that f cannot grow faster than a quadratic function. The next result makes this precise, and provides several other conditions related to Lipschitz smoothness.

Theorem 3.8. *For differentiable $f : \mathbb{R}^d \to (-\infty, \infty]$ and $L > 0$, consider the statements:*

(i) $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$, *for all* $x, y \in \mathrm{dom}(f)$;
(ii) the function $-f_L$ *is convex, where* $f_L(x) = f(x) - \frac{L}{2}\|x\|_2^2$;
(iii) $f(y) \leq f(x) + \nabla f(x)^\mathsf{T}(y - x) + \frac{L}{2}\|y - x\|_2^2$, *for all* $x, y \in \mathrm{dom}(f)$;
(iv) $(\nabla f(x) - \nabla f(y))^\mathsf{T}(x - y) \leq L\|x - y\|_2^2$, *for all* $x, y \in \mathrm{dom}(f)$.

Then the following relations hold:

$$\textit{(i)} \implies \textit{(ii)} \iff \textit{(iii)} \iff \textit{(iv)}.$$

If f is twice continuously differentiable, then (ii)–(iv) are also equivalent to the statement:

(v) $\nabla^2 f(x) \preceq LI$, *for all* $x \in \mathrm{dom}(f)$.

Lastly, for convex f, statements (i)–(iv) are all equivalent, and for twice continuously differentiable convex f, statements (i)–(v) are all equivalent.

Interestingly, the concept of strong convexity admits a similar string of equivalences.

Theorem 3.9. *For differentiable $f : \mathbb{R}^d \to (-\infty, \infty]$ and $m > 0$, consider the statements:*

(i) $\|\nabla f(x) - \nabla f(y)\|_2 \geq m\|x - y\|_2$, *for all* $x, y \in \mathrm{dom}(f)$;
(ii) the function f_m *is convex, where* $f_m(x) = f(x) - \frac{m}{2}\|x\|_2^2$;
(iii) $f(y) \geq f(x) + \nabla f(x)^\mathsf{T}(y - x) + \frac{m}{2}\|y - x\|_2^2$, *for all* $x, y \in \mathrm{dom}(f)$;
(iv) $(\nabla f(x) - \nabla f(y))^\mathsf{T}(x - y) \geq m\|x - y\|_2^2$, *for all* $x, y \in \mathrm{dom}(f)$.

Then the following relations hold:

$$\textit{(i)} \impliedby \textit{(ii)} \iff \textit{(iii)} \iff \textit{(iv)}.$$

If f is twice continuously differentiable, then (ii)–(iv) are also equivalent to the statement:

(v) $\nabla^2 f(x) \succeq mI$, *for all* $x \in \mathrm{dom}(f)$.

Comparing Theorems 3.8 and 3.9, we might view Lipschitz smoothness and strong convexity as dual concepts, informally speaking, since they give rise to conditions that are symmetric in nature. (Perhaps not surprisingly, the proofs of these conditions also proceed symmetrically, as outlined in

Exercise 3.11.) In fact, as we will see later in Chapter 8.3, there is a deeper (formal) dual relation at play: Lipschitz smoothness of a function is intimately tied to strong convexity of its conjugate.

Much of this chapter studies convex functions through the lens of their gradients (or Hessians). Of course, a convex function need not be differentiable (or twice differentiable). Take, for example, $f(x) = |x|$: convex, but not differentiable at $x = 0$. Meanwhile, for a univariate convex function, the only possibility for a point of nondifferentiability seems intuitively to be a "kink", like the behavior of the absolute value function at the origin. This leads us to wonder, more generally: to what extent can a convex function lack smoothness? Can a convex function be arbitrarily nonsmooth? Or does the property of convexity impose restrictions on how "much" nondifferentiability is "allowed"? The next theorem answers this last question affirmatively.

Theorem 3.10. *Let f be a convex function. Then the following hold:*

(i) f is continuous at every point in $\mathrm{int}(\mathrm{dom}(f))$*;*

(ii) in fact, f is locally Lipschitz continuous on $\mathrm{int}(\mathrm{dom}(f))$*, meaning that for each compact set $C \subseteq \mathrm{int}(\mathrm{dom}(f))$, there is a constant $L > 0$ (this can depend on C) such that*

$$|f(x) - f(y)| \leq L\|x - y\|_2, \quad \text{for all } x, y \in C;$$

(iii) hence, f is differentiable at almost every point in $\mathrm{int}(\mathrm{dom}(f))$*;*

(iv) moreover, f is twice differentiable at almost every point in $\mathrm{int}(\mathrm{dom}(f))$*.*

We remark that property (i) actually holds at every point in $\mathrm{relint}(\mathrm{dom}(f))$, and we only write $\mathrm{int}(\mathrm{dom}(f))$ in the theorem statement theorem to preserve symmetry in the presentation of (i)–(iv). We also remark that property (iii) follows from (ii): that locally Lipschitz functions are differentiable almost everywhere is a result known as *Rademacher's theorem*. Property (iv), the almost everywhere twice differentiability of convex functions, is known as *Aleksandrov's theorem*. To be explicit, when we say that a certain property holds for almost every point in a set S, we mean that it holds on a set $E \subseteq S$ such that $S \setminus E$ has Lebesgue measure zero.

While the almost everywhere differentiability and twice differentiability of convex functions is certainly interesting, we must be mindful not to overinterpret the implications of Theorem 3.10 as they pertain to convex optimization. Generally speaking, sets of Lebesgue measure zero cannot be ignored in mathematical optimization (unlike, say, functional analysis), particularly because many interesting optimization problems lack smoothness at their minima; when this happens, as we will see in the subsequent chapters, we must account for it, both analytically and algorithmically, and subgradients will be one of our primary tools for doing so.

3.6. Cones and polyhedra*

We cover cones and polyhedra in a bit more detail, two classes of sets that have special structure and play a central role in convex optimization.

3.6.1. Cones. A set $C \subseteq \mathbb{R}^d$ is called a *cone* if it satisfies

$$x \in C \implies tx \in C, \quad \text{for all } t \geq 0. \tag{3.17}$$

A special type of cone of particular interest is a *convex cone*, which is simply a set that is both a cone and convex. Combining (3.1) and (3.17), we see that a convex cone C is defined by

$$x, y \in C \implies sx + ty \in C, \quad \text{for all } s, t \geq 0. \tag{3.18}$$

Recall that the positive semidefinite cone (3.2) and normal cone (3.3) are both convex cones. Another noteworthy example is the *norm cone*, defined for a norm $\|\cdot\|$ by

$$\{(x,t) \in \mathbb{R}^d \times \mathbb{R} : \|x\| \le t\}. \tag{3.19}$$

The *conic hull* of C is the set of all nonnegative linear combinations of points in C,

$$\text{cone}(C) = \left\{ \sum_{i=1}^n t_i x_i : n \ge 1,\ x_i \in C,\ t_i \ge 0, \text{ for } i = 1, \dots, n \right\}.$$

The conic hull $\text{cone}(C)$ is itself a convex cone, for any set C; in fact, it is the smallest convex cone containing C, meaning that $\text{cone}(C) \subseteq D$ for any convex cone $D \supseteq C$.

Next we present a classical result on conic and convex hulls.

Theorem 3.11 (Carathéodory's theorem)**.** *Let $C \subseteq \mathbb{R}^d$.*

(i) Every nonzero element in the conic hull $\text{cone}(C)$ can be represented as a positive linear combination of linearly independent (and thus at most d) elements of C.

(ii) Every element in the convex hull $\text{conv}(C)$ can be represented as a convex combination of affinely independent (and thus at most $d+1$) elements of C.

This has various important implications; see part f of Exercise 3.3 for a basic one.

3.6.2. Polyhedra. A *polyhedron* $P \subseteq \mathbb{R}^d$ is the intersection of a finite number of halfspaces, which we can express generically as

$$P = \{x : Ax \le b\}$$

for a matrix A and vector b, where recall we interpret this inequality componentwise. It should be clear that $\{x : Ax \ge b\}$, $\{x : \ell \le Ax \le u\}$, and $\{x : Ax \le b,\ Gx = h\}$ are all also polyhedra, as we can rewrite each of them in the form of the above display, for appropriately (re)defined A, b.

Less obvious is the fact that an optimization problem over a polyhedron can always be equated with a higher-dimensional optimization problem over the intersection of an affine subspace and the nonnegative orthant, $\{x : Ax = b,\ x \ge 0\}$. We will revisit this when we discuss linear programs in standard form, in Chapter 5.1.

We call a bounded polyhedron a *polytope*. A fundamental fact about polytopes is as follows.

Theorem 3.12. *Every polytope can be represented as the convex hull of a finite set of points.*

This result is both highly intuitive and highly nontrivial; later, in Chapter 12.2, we will return to it from the perspective of duality. Now we simply introduce some nomenclature to help remember Theorem 3.12: we call $P = \{x : Ax \le b\}$ a *halfspace representation* or *H-representation* of P, and $Q = \text{conv}\{x_1, \dots, x_n\}$ a *vertex representation* or *V-representation* of Q. Then in this language, the theorem says that every polytope has an equivalent H-representation and V-representation.

An interesting aspect of polyhedra is their facial structure. A *face* F of a polyhedron $P \subseteq \mathbb{R}^d$ is a set F such that

$$F = \emptyset,\ F = P, \text{ or } F = P \cap H \text{ for a supporting hyperplane } H \text{ of } P.$$

The faces $F = \emptyset$ and $F = P$ are said to be *improper*; each other face of F of P is called *proper*. A face F is said to have dimension k (or, called a k-face) if its affine span $\text{aff}(F)$ is k-dimensional. A maximal proper face (proper face of maximum degree) of P is called a *facet* of P. If $F = \{x\}$ is a 0-face of P, then we call x an *vertex* of P.

To be fair, these concepts are not specific to polyhedra and the same definitions apply to convex sets in general. A small difference in nomenclature: if $F = \{x\}$ is a 0-face of a convex set C, then we call x an *exposed point* of C (for polyhedra, vertices and exposed points are equivalent, but not in general). Even in the general convex setting, these concepts lead to interesting developments; for example, for any compact (closed and bounded) convex set C,

$$C = \operatorname{cl}\big(\operatorname{conv}\{x : x \text{ is an exposed point of } C\}\big),$$

a result called *Straszewicz' theorem*. (For polytopes, the above is true without the surrounding $\operatorname{cl}(\cdot)$ because the set of exposed points—that is, the vertex set—is finite.) For more, see Exercise 3.13.

What makes polyhedra so special, however, is the *structure* exhibited by their faces. The faces of a polyhedron P obey a beautiful recursive structure: each face of a face of P is also a face of P, each face of P can be expressed as an intersection of facets of P, and so on. We will not go into this in any detail, but we will leverage the relationship between faces of P and faces of its dual polytope P^* when we prove Theorem 3.12 in Chapter 12.2.

Chapter Notes

A lot can be said about convex sets and functions, far more than what is said in this chapter. For readers seeking a more in-depth treatment, there are many excellent references on the subject, for example, [**Roc70**] (Chapters 1–11, 17–22), [**BV04**] (Chapters 2, 3), [**Ber09**] (Chapters 1, 2), to name just a few. [**BV04**] gives a particularly comprehensive treatment of operations that preserve convexity. For more details on the smoothness properties of convex functions, we refer readers to [**Roc70**] (Chapters 10, 25), or [**EG15**] (Chapters 6.3, 6.4). To learn more about polyhedra and the study of their facial structure, a classic reference is [**Grü03**].

As briefly discussed in Remark 3.5, our definition of a convex function does not allow it to take the value $-\infty$ (and likewise, we do not allow a concave function to take the value ∞). This is in contradiction with the standard approach in convex analysis, for example [**Roc70, Ber09**] (but it is consistent with the approach of others, for example [**BV04**]). The standard approach in convex analysis allows a convex function to take values in $[-\infty, \infty]$, and defines a *proper* convex function f as one such that $f \neq \infty$ and $f > -\infty$ (f is not identically ∞, and never $-\infty$). While proper convex functions are the real topic of interest, *improper* convex functions do arise in various situations, and there are some nontrivial improper convex functions that are not just identically $\pm\infty$, for example

$$f(x) = \begin{cases} -\infty & \|x\| < 1 \\ 0 & \|x\| = 1 \\ \infty & \|x\| > 1 \end{cases},$$

for a norm $\|\cdot\|$. Thus, from a general mathematical perspective, it is important to be able to deal with them. Altogether, for our purposes, we find it simpler to rule them out completely, though we admit this does come with its own set of drawbacks, for example we must occasionally add explicit qualifiers to statements about certain functions lying above $-\infty$, as we did in Property 3.4.H, the convexity-preserving rule for a partial infimum.

Exercises

3.1 Prove that if X is a random variable supported on a convex set $C \subseteq \mathbb{R}^d$ (and $\mathbb{E}(X)$ exists) then $\mathbb{E}(X) \in C$.

3.2 Show that if f is a convex function, then $\{x : f(x) \leq t\}$, called the *sublevel set* of f at level t, is a convex set. Show that the converse is not true: give an example of a nonconvex function whose sublevel sets are convex (for all levels t).

3.3 We explore some of the differences between preserving convexity and closedness of sets.

a. Prove Property 3.4.A, that an intersection of (any number of) convex sets is convex.
b. Prove that the intersection of (any number of) closed sets is closed.
c. Prove Property 3.4.C, that linear images and preimages of convex sets are convex.
d. Show that the preimage of a closed set under a linear map is closed; but the image of a closed set under a linear map need not be closed (give a counterexample).
e. Give an example of a closed set whose convex hull is not closed.
f. Prove that the convex hull of a compact (closed and bounded) set is compact. Hint: use part (ii) of Carathéodory's theorem (Theorem 3.11), and sequential compactness.

3.4 We explore some properties of the various "flavors" of convexity.

a. Give an example of a strictly convex function that is not strongly convex.
b. Give an example of a strongly convex function that is not differentiable.
c. Give an example of a strictly convex function f such that $f(x) \to -\infty$ as $\|x\|_2 \to \infty$.
d. Prove that for any differentiable strongly convex function f, we must have $f(x) \to \infty$ as $\|x\|_2 \to \infty$. Hint: use part (iii) of Theorem 3.9.

   Note: differentiability is not actually required here; strong convexity alone is enough as we can use part (iii) of Theorem 6.6; see Exercise 6.17.

3.5 We explore Property 3.4.K from a few perspectives.

a. Prove Property 3.4.K by verifying directly that F satisfies the definition of convexity.
b. Prove that the second bullet point in Property 3.4.K can be collapsed into the first: if the conditions of the property hold, and there exists i such that f nonincreasing in its i^{th} argument and g_i concave, then we can simply reparametrize by flipping the signs of the appropriate arguments/component functions so that we can write $F = \tilde{f} \circ \tilde{g}$ for $\tilde{f}$ nondecreasing in each argument, and $\tilde{g}$ with all component functions convex.
c. Prove that Property 3.4.K can be extended to the case where g takes infinite values, in the following way. If $f : \mathbb{R}^k \to (-\infty, \infty]$ is convex and nondecreasing in each argument, and $g : \mathbb{R}^d \to (-\infty, \infty]^k$ is convex in each component, then $F = f \circ g$ is convex, where we set $F(x) = \infty$ for any x with at least one component equal to ∞.

3.6 In this exercise, we practice checking convexity, focusing on functions that commonly appear in statistical estimation and machine learning. In some instances, it might be easiest to use convexity-preserving operations from Chapter 3.4, in others, it might be easier to prove the given claim straight from the definition.

a. *Linear regression loss.* Prove that $f(\beta) = \|y - X\beta\|_2^2$ is convex, for any $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$.
b. *Logistic regression loss.* Prove that $f(\beta) = -y^\mathsf{T} X\beta + \sum_{i=1}^n \log(1 + e^{x_i^\mathsf{T}\beta})$ is convex, for any $X \in \mathbb{R}^{n \times d}$ and $y \in \{0,1\}^n$, where $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ denote the rows of X.
c. *Hinge classification loss.* Prove that $f(\beta) = \sum_{i=1}^n (1 - y_i x_i^\mathsf{T}\beta)_+$ is convex, where $a_+ = \max\{a, 0\}$, for any $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ and $y \in \{-1, 1\}^n$,

d. *Kullback-Leibler divergence between discrete distributions.* Prove that

$$f(p,q) = \sum_{i=1}^{d} p_i \log(p_i/q_i)$$

is a convex function of $(p,q) \in \{x \in \mathbb{R}^{2d} : 1^\mathsf{T} x = 1,\ x > 0\}$.

e. *Gaussian log likelihood for precision matrix.* Prove that

$$f(\Sigma^{-1}) = -\frac{n}{2}\log(\det(\Sigma)) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^\mathsf{T}\Sigma^{-1}(x_i - \mu)$$

is a concave function of $\Sigma^{-1} \in \mathbb{S}^d_{++}$ (the inverse covariance matrix, called the *precision matrix*), for any $x_i \in \mathbb{R}^d$, $i = 1, \ldots, d$ and $\mu \in \mathbb{R}^d$. Hint: use the relationship between $\det(A)$ and $\det(A^{-1})$ for an invertible matrix A, and also linearity of the trace operator $\mathrm{tr}(\cdot)$.

f. Let f be strictly convex with $\mathrm{dom}(f) = \mathbb{R}^d$. By Property 3.4.I, the linear composition rule, we know that $g(\beta) = f(X\beta)$ is convex for any $X \in \mathbb{R}^{n\times d}$. Prove that g is strictly convex if and only if $\mathrm{rank}(X) = d$.

3.7 Now we practice checking nonconvexity. In some cases, a counterexample to (3.4) should be simple to produce; in others, inspecting the second-order condition (3.16) might be easier.

a. ℓ_0 *"norm".* Prove that $f(x) = \|x\|_0$, where

$$\|x\|_0 = \sum_{i=1}^{d} 1\{x_i \neq 0\} \tag{3.20}$$

is the ℓ_0 "norm", is nonconvex over $\mathbb{R}^d$.

Note: we use quotation marks since $\|\cdot\|_0$ is not a norm (it lacks positive homogeneity) and calling it a *pseudonorm* would be more appropriate; however, we adopt the common terminology henceforth and simply refer to it as a norm (without quotations).

b. *Matrix rank.* Prove that $f(x) = \mathrm{rank}(X)$ is nonconvex over $\mathbb{R}^{n\times d}$.

c. *Gaussian negative log likelihood for mean and variance.* Prove that

$$f(\mu, \sigma^2) = \log\sigma + \frac{(y-\mu)^2}{2\sigma^2}$$

is nonconvex over $\mathbb{R} \times \mathbb{R}_{++}$.

d. *Least squares for two-layer neural network.* Prove that

$$f(W,u,v) = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k} v_j \phi(w_j^\mathsf{T} x_i + u_j)\right)^2$$

is nonconvex over $\mathbb{R}^{k\times p} \times \mathbb{R}^k \times \mathbb{R}^k$, where $\phi : \mathbb{R} \to \mathbb{R}$ can be taken (for simplicity) to be twice differentiable and monotone.

3.8 The following is a "strict" variant of the separating hyperplane theorem: if C, D are nonempty disjoint closed convex sets, and (say) D is bounded, then there exists $a \neq 0$ and b such that

$$a^\mathsf{T} x > b \text{ for all } x \in C \text{ and } a^\mathsf{T} x < b \text{ for all } x \in D,$$

that is, the hyperplane $\{x : a^\mathsf{T} x = b\}$ strictly separates C, D. Use this to prove *Farkas' lemma*: given $A \in \mathbb{R}^{k\times d}$, $b \in \mathbb{R}^k$, exactly one of the following statements is true:

- there exists $x \in \mathbb{R}^d$ such that $Ax = b$, $x \geq 0$;
- there exists $y \in \mathbb{R}^k$ such that $A^\mathsf{T} y \geq 0$, $y^\mathsf{T} b < 0$.

Hint: take $C = \{Ax : x \geq 0\}$.

3.9 The following is a "strict" variant of the supporting hyperplane theorem: if C is convex and $x_0 \in \text{relbd}(C)$, then there exists $a \neq 0$ and b such that $a^\mathsf{T} x_0 = b$ and

$$a^\mathsf{T} x \leq b \text{ for all } x \in C, \text{ with } a^\mathsf{T} x < b \text{ for } x \in \text{relint}(C),$$

where $\text{relbd}(C)$ and $\text{relint}(C)$ denote the relative boundary and relative interior, respectively, of C. We will use this to prove the existence of (what will be later defined as) subgradients of a convex function, on the relative interior of its effective domain.

a. Let f be a convex function and $x \in \text{relint}(\text{dom}(f))$. Prove that $(x, f(x)) \in \text{relbd}(\text{epi}(f))$, where $\text{epi}(f)$ is the epigraph of f, as defined in (3.14).

b. Apply the strict version of the supporting hyperplane theorem given above to prove that there exists some nonzero $a = (s, v) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$s^\mathsf{T} x + v f(x) \geq s^\mathsf{T} y + vt, \quad \text{for all } (y, t) \in \text{epi}(f),$$

with strict inequality when $f(y) < t$.

c. Prove that we must have $v < 0$, thus by rescaling (s, v) and rearranging the last display,

$$f(y) \geq f(x) + s^\mathsf{T}(y - x), \quad \text{for all } y \in \text{dom}(f),$$

which says that s is a subgradient of f at x, denoted $s \in \partial f(x)$, as we will learn later in Chapter 6.1.

d. Prove that restricting $x \in \text{relint}(\text{dom}(f))$ is necessary in general for the existence of a subgradient, by giving an example where f is convex, $x \in \text{dom}(f) \setminus \text{relint}(\text{dom}(f))$, but f has no subgradients at x. Hint: for this example, we want the supporting hyperplane to $\text{epi}(f)$ at $(x, f(x))$ to be vertically-oriented, so if $a = (s, v)$ denotes the normal vector to this hyperplane, then $v = 0$.

3.10 Prove, using the first-order characterization for convexity (3.15), that a differentiable convex function f has a gradient ∇f that acts as a *monotone operator*, which means that it satisfies

$$\big(\nabla f(x) - \nabla f(y)\big)^\mathsf{T}(x - y) \geq 0, \quad \text{for all } x, y \in \text{dom}(f). \tag{3.21}$$

Prove that the converse is also true, for differentiable f: if f satisfies the above property then it must be convex. Hint: set $g(t) = f(x + tv)$ for $v = y - x$ and $t \in [0, 1]$, and consider what the monotone gradient condition implies about $g'(t)$.

3.11 In this exercise, we will work through the proofs of Theorems 3.8 and 3.9.

a. Beginning with Theorem 3.8, show that (ii) $\iff$ (iii) $\iff$ (iv) by invoking equivalent characterizations of convexity (including (3.21)).

b. Show that (i) $\implies$ (iv) by the Cauchy-Schwarz inequality, which proves the first display in Theorem 3.8.

c. When f is twice continuously differentiable, show that (v) $\implies$ (iii) using a first-order Taylor expansion of f with remainder; show that (iv) $\implies$ (v) by expressing $\nabla^2 f(x)h$ as the directional derivative of $\nabla f(x)$ in the direction h, applying (iv), and taking h to be the top eigenvector of $\nabla^2 f(x)$. This establishes the second part of Theorem 3.8.

d. Moving now to Theorem 3.9, show that (ii) $\iff$ (iii) $\iff$ (iv) by invoking equivalent characterizations of convexity (including (3.21)).

e. Show that (iv) $\implies$ (i) by the Cauchy-Schwarz inequality, which proves the first display in Theorem 3.9.

f. When f is twice continuously differentiable, show that (v) $\implies$ (iii) using a first-order Taylor expansion of f with remainder; show that (iv) $\implies$ (v) by expressing $\nabla^2 f(x)h$ as the directional derivative of $\nabla f(x)$ in the direction h, applying (iv), and taking h to be the bottom eigenvector of $\nabla^2 f(x)$. This establishes the second part of Theorem 3.9.

g. For convex f, prove the final part of Theorem 3.8 by showing (iv) $\implies$ (i), using the following steps. First, define $F_x(z) = f(z) - \nabla f(x)^\mathsf{T} z$. Argue that F_x satisfies (iv), and thus also satisfies (iii), that is,

$$F_x(z) \leq F_x(y) + \nabla F_x(y)^\mathsf{T}(z-y) + \frac{L}{2}\|z-y\|_2^2.$$

Minimize each side of the above display over z. (Hint: use the fact that a differentiable convex function is minimized by setting its gradient to zero applied to each of the left- and right-hand sides separately.) Show that, after rearrangement, this yields

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

Exchange the roles of x, y, and add the resulting statement to the above display to get

$$\big(\nabla f(x) - \nabla f(y)\big)^\mathsf{T}(x-y) \geq \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|_2^2.$$

Use Cauchy-Schwarz to conclude that f is Lipschitz smooth, as desired.

A remark on the proof technique: there is an interesting parallel between the argument here and that used to prove that the conjugate a strongly convex function is Lipschitz smooth; see Exercise 8.10. The last display above is equivalent to the conclusion that the conjugate f^* of f is strongly convex with parameter $1/L$. To check this, set $u = \nabla f(x)$, $v = \nabla f(y)$, and note that $\nabla f^*(u) = x$ and $\nabla f^*(v) = y$. In comparison, the argument in Exercise 8.10 starts with the premise of strong convexity, and then proceeds in the other direction, by essentially reversing the steps in the above argument.

3.12 We explore Theorems 3.8 and 3.9 a bit further, to better understand their statements of results and the necessity of certain assumptions.

a. First, show that the conditions $L > 0$, $m > 0$ in the theorems are unnecessary, in the sense that the stated results still hold for $L = 0$ and $m = 0$. Hint: what is a Lipschitz smooth function with $L = 0$? What is a strongly convex function with $m = 0$?

b. Next, show that each of the four statements, (ii)–(v), in Theorem 3.9 imply that f is convex (meaning, convexity is implicit in all of the equivalences/implications).

c. Now, show that convexity is *not* implied by any one of the five statements, (i)–(v), in Theorem 3.8.

d. Lastly, show that convexity of f is indeed necessary for equating (i) with the rest of the statements in Theorem 3.8: give an example of a nonconvex function that satisfies one of (ii)–(v), but not (i).

3.13 A point $x \in C$ is said to be an *extreme point* of a convex set C if

$$x = ty + (1-t)z,\ t \in (0,1),\ y, z \in C \implies x = y = z.$$

In other words, x cannot lie in the relative interior of any line segment joining distinct points in C. In this exercise, we will explore the properties of extreme points and their differences to exposed points, which recall from Chapter 3.6.2, are points $x \in C$ such that

$$\{x\} = C \cap H \text{ for a supporting hyperplane } H \text{ of } C.$$

We will write $\mathrm{ext}(C)$ and $\exp(C)$ for the sets of extreme and exposed points, respectively, of C. Throughout this exercise we take C to be convex.

a. Prove that if F is a face of C and C is closed, then $\mathrm{ext}(F) = F \cap \mathrm{ext}(C)$.

b. Prove that if C is compact, then $C = \mathrm{conv}(\mathrm{ext}(C))$. Hint: note that $C \supseteq \mathrm{conv}(\mathrm{ext}(C))$ by the definition of extreme points; for the other direction (the opposite containment),

use induction on the dimension d of $\text{aff}(C)$ (for $d = 1$, the only compact sets are closed bounded intervals) and use part a.

c. Prove that $\exp(C) \subseteq \text{ext}(C)$.

d. Prove that this is not an equality in general, by giving an example such that $\exp(C) \subsetneq \text{ext}(C)$ (an example with extreme points that are not exposed).

A note on Straszewicz' theorem, from Chapter 3.6.2: we can interpret this, in light of parts b, c, d, as saying that for any compact convex set C, the set of exposed points $\exp(C)$ is dense in the set of extreme points $\text{ext}(C)$. For polytopes, these two sets are finite, and hence must be equal, which gives us another way of understanding Theorem 3.12.

3.14 Let f be a convex function and $P \subseteq \text{dom}(f)$ a polytope (bounded polyhedron). Prove that the maximum of f over P is attained by one of the vertices of P. Hint: use the V-representation for P that we know exists from Theorem 3.12 (that is, express P as the convex hull of a finite set of vertices).

3.15 Let f be a convex function and $C \subseteq \text{dom}(f)$ be a compact convex set. As a generalization of the last exercise, show that the maximum of f over C is attained by one of the extreme points of C. Hint: use Exercise 3.13 part b.

Optimization Basics

4.1. Optimization problems

Equipped with a working knowledge of convex sets and functions, the basic principles of optimization are now within reach. In this book, we denote an *optimization problem* by

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & g_i(x) \leq 0,\ i = 1, \ldots, m \\ & h_j(x) = 0,\ j = 1, \ldots, k. \end{array} \tag{4.1}$$

Here f, g_i, $i = 1, \ldots, m$, and h_j, $j = 1, \ldots, k$ are functions, from $\mathbb{R}^d$ to $[-\infty, \infty]$. In problem (4.1), the minimization is implicitly restricted to the intersection of relevant effective domains:

$$D = \operatorname{dom}(f) \cap \bigcap_{i=1}^{m} \operatorname{dom}(g_i) \cap \bigcap_{j=1}^{k} \operatorname{dom}(h_j).$$

The function f in (4.1) is called the *objective* or *criterion*. A *feasible point* is a point in D such that all constraints (inequality and equality constraints) are met in problem (4.1). The infimal criterion value among all feasible points is denoted

$$f^\star = \inf \big\{ f(x) : x \in D,\ g_i(x) \leq 0,\ i = 1, \ldots, m,\ h_j(x) = 0,\ j = 1, \ldots, k \big\},$$

and called the *optimal value* in (4.1). A feasible point that achieves the optimal value is denoted $x^\star$ (note that $f^\star = f(x^\star)$), and is called a *solution* or *minimizer*.

It is worth being clear at the outset that a solution need not exist in an optimization problem in general. This can happen for two distinct reasons. First, a solution never exists in an optimization problem that is *infeasible*, which means that it has no feasible points. In this case, we set $f^\star = \infty$ by convention. A second, more interesting case: even in a feasible optimization problem ($f^\star < \infty$), a solution fails to exist when the optimal value $f^\star$ is not attained. For example, informally speaking, this happens when the criterion is minimized "somewhere off at infinity" (as in $f(x) = e^{-x}$). The existence of solutions (in feasible problems) is itself an interesting topic, and is covered further in Chapter 4.4.

A *convex optimization problem* is one of the form (4.1) such that f and g_i, $i = 1, \ldots, m$ are all convex functions, and h_j, $j = 1, \ldots, k$ are all affine functions. In other words, the problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \ i = 1, \ldots, m \\ & Ax = b, \end{array} \tag{4.2}$$

is convex whenever f and g_i, $i = 1, \ldots, m$ are convex (and A and b are arbitrary).

In general, we will say that two optimization problems are *equivalent* if solutions of one can be computed from solutions of the other, and vice versa. Clearly, problem (4.1) is equivalent to:

$$\begin{array}{ll} \underset{x}{\text{maximize}} & -f(x) \\ \text{subject to} & g_i(x) \leq 0, \ i = 1, \ldots, m \\ & h_j(x) = 0, \ j = 1, \ldots, k, \end{array}$$

As f is convex if and only if $-f$ is concave, the convex minimization problem (4.2) is hence also equivalent to a concave maximization problem. In this book, we will use "optimization" to refer to both minimization and maximization (and we carry all definitions and notations introduced above over to maximization problems); likewise, we will use "convex optimization" to refer to both convex minimization and concave maximization.

Remark 4.1. In statistics, we rarely denote the parameter in an optimization problem by x; we typically use β or θ (or something else, as x is usually reserved for an input feature vector). We also rarely denote the solution to an optimization problem by $\beta^\star$ or $\theta^\star$; we commonly use $\hat{\beta}$ or $\hat{\theta}$, as these are typically seen as estimates of population-level parameters of interest. In this book, when discussing abstract properties of mathematical optimization, we will adhere to the standard notation in optimization, as demonstrated above; but when discussing problems in statistics or machine learning, we will switch fluidly to the notation more common in these fields. This should not cause any confusion, as the meaning (for example, what is a parameter and what is a feature) should be clear from the context.

Example 4.2. The following are examples of two central optimization problems in statistics and machine learning. Both problems are convex.

a. Given responses $y_i \in \mathbb{R}$ and associated feature vectors $x_i \in \mathbb{R}^d$, for $i = 1, \ldots, n$, the *least absolute selection and shrinkage operator (lasso)* is a sparse estimator of the coefficients β in a linear model (to y_i predict from $x_i^\mathsf{T}\beta$), defined by the optimization problem:

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^\mathsf{T}\beta)^2 + \lambda \sum_{j=1}^{d} |\beta_j|. \tag{4.3}$$

Here $\lambda \geq 0$ is a tuning parameter governing the tradeoff between goodness-of-fit (first term) and sparsity (second term). The above also has a more compact form, denoting by $y \in \mathbb{R}^n$ the response vector and $X \in \mathbb{R}^{n \times d}$ the feature matrix (whose i^{th} row is x_i^T):

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \|\beta\|_1. \tag{4.4}$$

For insight into the sparsity-inducing property of ℓ_1 penalties, from the perspective of proximal mappings, see Chapter 7.3.

Note: by default, we will always omit an intercept term in the lasso regression model as it can be accounted for by centering y and each column of X, before solving (4.4); see Example 4.3.a for a generalization.

b. Given class labels $y_i \in \{-1, 1\}$ and feature vectors $x_i \in \mathbb{R}^d$, for $i = 1, \ldots, n$, the *support vector machine (SVM)* is a large-margin linear classifier (to predict y_i from the sign of $\beta_0 + x_i^T \beta$), defined by the optimization problem:

$$\begin{array}{ll} \underset{\beta_0, \beta, \xi}{\text{minimize}} & \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} & y_i(\beta_0 + x_i^T\beta) \geq 1 - \xi_i, \ i = 1, \ldots, n \\ & \xi \geq 0. \end{array} \tag{4.5}$$

Here $C \geq 0$ is a tuning parameter governing the tradeoff between the size of the margin (the first term above is actually the *inverse* margin; see Exercise 4.1) and violations to the margin condition (the second term is the sum of violation costs, and each violation is an instance of a prediction $\beta_0 + x_i^T\beta$ lying on the wrong side of the margin).

It is helpful to introduce some additional notation for an optimization problem (4.1) in which the criterion is identically zero $f(x) = 0$ (or, equal to any finite constant). We call this a *feasbility problem*. A feasibility problem effectively seeks whether the constraints can be satisfied, and if so, seeks any point $x^\star$ that is feasible. We write it as

$$\begin{array}{ll} \text{find} & x \\ \text{subject to} & g_i(x) \leq 0, \ i = 1, \ldots, m \\ & h_j(x) = 0, \ j = 1, \ldots, k. \end{array} \tag{4.6}$$

A *convex feasibility problem* is one in which the constraint functions satisfy the usual requirements: g_i, $i = 1, \ldots, m$ are convex and h_j, $j = 1, \ldots, k$ are affine.

4.2. Properties of convex problems

Next we describe several important properies of convex optimization problems, beginning with the most important one.

A. Local solutions are global solutions. A point $\bar{x}$ is called a *local solution* of the problem (4.1) if it is feasible and there is some $\delta > 0$ such that

$$f(\bar{x}) \leq f(x), \quad \text{for all feasible } x \text{ such that } \|x - \bar{x}\|_2 \leq \delta.$$

For a convex optimization problem, the following holds: any local solution $\bar{x}$ must also be a global solution, in that $f(\bar{x}) \leq f(x)$ for all feasible points x. (This is also simply called a solution, and we add the modifier "global" here to emphasize the difference to the local condition.) This result is so important that it may as well be called the *fundamental theorem of convex optimization*.

The proof of this fundamental result is elementary. It can be broken into two steps. The first step is to check that for the convex problem (4.2), the set of feasible points is a convex set, which follows from convexity of the set D, convexity of the functions f and g_i, $i = 1, \ldots, m$, and the linear structure of the equality constraints $Ax = b$ (Exercise 4.2 part a).

The second step proceeds by contradiction. If $\bar{x}$ is not a global solution, then there is a feasible point y such that $f(y) < f(\bar{x})$. Since $\bar{x}$ is locally optimal, we must have $\|y - \bar{x}\|_2 > \delta$. However,

we can choose some $t \in (0,1)$ such that $x = ty + (1-t)\bar{x}$ satisfies $\|x - \bar{x}\|_2 \leq \delta$ (for example, take $t = \delta/\|y - \bar{x}\|_2$). By convexity of the feasible set, we know that x is feasible. By convexity of f,

$$f(x) \leq tf(y) + (1-t)f(\bar{x}) < f(\bar{x}),$$

where the last inequality is strict since $f(y) < f(\bar{x})$ and $t > 0$. The above display is a contradiction of local optimality, which proves that such a y cannot exist, and $\bar{x}$ must be globally optimal.

B. Solution sets are convex. A related property of a convex optimization problem is that its solution set, which we can denote by

$$S^\star = \{x^\star : x^\star \text{ is a solution in problem (4.2)}\},$$

is itself a convex set. The proof is similar to the proof that the feasible set of a convex problem is convex (Exercise 4.2 part b). Convexity of $S^\star$ has the following interesting implication: a convex problem can have 0, 1, or infinitely many solutions—no other number is possible!

An important refinement is possible when the criterion f in a convex problem is strictly convex. In this case, the solution set $S^\star$—if nonempty—must be a singleton (Exercise 4.2 part d).

C. First-order optimality condition. Denote by

$$C = \{x \in D : g_i(x) \leq 0,\ i = 1, \ldots, m,\ Ax = b\}$$

the feasible set of the convex optimization problem (4.2). For a differentiable criterion f, a point $x \in C$ is a solution if and only if

$$\nabla f(x)^\mathsf{T}(y - x) \geq 0, \quad \text{for all } y \in C. \tag{4.7}$$

This is called the *first-order optimality condition* for problem (4.2). It can be interpreted as follows: any move from x in the direction of another feasible point cannot decrease the criterion f, according to the first-order Taylor expansion of f at x. In the unconstrained case (which means $m = r = 0$, so there are no constraints), the feasible set is $C = \mathrm{dom}(f)$, and since a differentiable function has an open effective domain (by assumption, see Appendix B.1), the first-order optimality condition (4.7) reduces to the more familiar zero-gradient condition

$$\nabla f(x) = 0. \tag{4.8}$$

To see this observe that for sufficiently small $\delta > 0$, we have $y = x + \delta v \in \mathrm{dom}(f)$ for any v, which from (4.7) will lead us to the conclusion that $\nabla f(x)^\mathsf{T} v = 0$ for any v, implying (4.8).

We note that the first-order optimality condition (4.7) is actually a special case of what we will call the subgradient optimality condition, to be encountered later in Chapter 6.3.

Example 4.3. The following are three examples of the first-order optimality condition for convex optimization.

a. For a differentiable convex optimization problem with only equality constraints,

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad Ax = b,$$

the first-order optimality condition (4.7) reduces to

$$\nabla f(x) + A^\mathsf{T} v = 0, \quad \text{for some } v. \tag{4.9}$$

The argument justifying (4.9) is similar to that used in the unconstrained case to justify (4.8) (see Exercise 4.3). The condition (4.9) is known as a *Lagrange multiplier condition*, and it will be generalized by the Karush-Kuhn-Tucker conditions, which we will study later in Chapter 11.

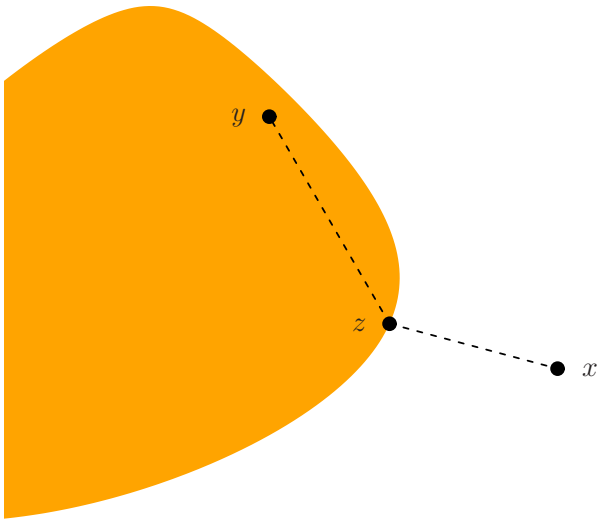

Figure 4.1. Illustration of the variational inequality that determines the optimality of the projection z of a point x onto a convex set. The angle between the line segments zx and zy must be at least a right angle (at least 90°), for any other point y in the set.

b. The Euclidean projection operator P_C onto a convex set $C \subseteq \mathbb{R}^d$ can be formulated in terms of differentiable convex optimization. In particular, for any $x \in \mathbb{R}^d$, its projection $P_C(x)$ onto C is the unique solution of the optimization problem

$$\underset{z}{\text{minimize}} \quad \|x - z\|_2^2 \quad \text{subject to} \quad z \in C.$$

The condition (4.7) (now interpreted with respect to z) gives, after rearrangement,

$$(x - z)^\mathsf{T}(z - y) \geq 0, \quad \text{for any } y \in C. \tag{4.10}$$

This is often referred to as the *variational inequality* for a projection. It says that the vector pointing from z to x must have a positive inner product with the vector pointing from y to z, for any $y \in C$. See Figure 4.1.

c. For a convex quadratic function $f(x) = \frac{1}{2}x^\mathsf{T}Ax + b^\mathsf{T}x + c$ (where $A \succeq 0$) the first-order optimality condition (4.8) says that x minimizes f if and only if

$$Ax + b = 0 \tag{4.11}$$

We can now reason in cases.

(i) If A is invertible, then the only solution to (4.11) is $x^\star = -A^{-1}b$.

(ii) If A is singular and $b \in \text{col}(A)$, then there are infinitely many solutions to (4.11), of the form $x^\star = x_0 + v$ where x_0 is one particular solution ($Ax_0 + b = 0$) and v is any element in $\text{null}(A)$, the null space of A.

(iii) If A is singular and $b \notin \text{col}(A)$, then (4.11) has no solution, which means that f does not have a minimizer (it does not attain its infimum). An example in which this occurs is the 2-dimensional convex quadratic $f(x) = x_1^2 - x_2$.

4.3. Problem transformations

We walk through various general transformations of optimization problems that can make a problem easier to solve or easier to understand. In each case, we do not assume convexity outright, but we do highlight the situations in which more can be said about convex optimization.

A. Characteristic formulation. Denote the feasible set of the optimization problem (4.1) by

$$C = \{x \in D : g_i(x) \leq 0,\ i = 1, \ldots, m,\ h_j(x) = 0,\ j = 1, \ldots, k\}$$

Then (4.1) is equivalent to the unconstrained optimization problem

$$\underset{x}{\text{minimize}} \quad f(x) + I_C(x), \tag{4.12}$$

where I_C is the characteristic function of C, defined in (3.11). For convex optimization, recall that the criterion f and feasible set C must both be convex, which implies $f + I_C$ is also convex. In others words, problem (4.12) is convex provided that the original problem (4.1) is.

We refer to (4.12) as the *characteristic formulation* of problem (4.1). Though the characteristic formulation is not (typically) more practically convenient than the original problem, it can still lead to useful insights. For example, in the convex case (with f and C convex), applying the subgradient optimality condition to (4.12) yields a natural generalization of the first-order optimality condition in (4.7). This will be covered in Chapter 6.3.

B. Partial optimization. For any function g, it holds that (Exercise 4.5 part a):

$$\inf_{x_1, x_2} g(x_1, x_2) = \inf_{x_1} G(x_1),$$

where $G(x_1) = \inf_{x_2} g(x_1, x_2)$. If we work from the characteristic formulation (4.12), for a general optimization problem, and we treat x as a block variable $x = (x_1, x_2)$, then we can apply the above result to $g = f + I_C$: this tells us that we can perform partial optimization—namely, optimization over x_2—to produce an equivalent optimization problem, in x_1 alone.

This statement can be made more concrete in the case that, for each fixed x_1, the infimum of $g(x_1, x_2)$ over x_2 is attained, and we denote a minimizer by $x_2^\star(x_1)$. Then $G(x_1) = g(x_1, x_2^\star(x_1))$, and

$$\underset{x_1, x_2}{\text{minimize}} \quad f(x_1, x_2) \quad \text{subject to} \quad (x_1, x_2) \in C,$$

is equivalent to

$$\underset{x_1}{\text{minimize}} \quad F(x_1) \quad \text{subject to} \quad x_1 \in \tilde{C},$$

with $F(x_1) = \inf_{x_2 : (x_1, x_2) \in C} f(x_1, x_2) = f(x_1, x_2^\star(x_1))$, and $\tilde{C} = \{x_1 : (x_1, x_2^\star(x_1)) \in C\}$. It is worth noting that the setting of convex optimization, where f and C are convex in the former problem (second-to-last display), we know by Property 3.4.H that the latter problem (last display) is also convex, provided that F is nowhere equal to $-\infty$.

Example 4.4. The following are two examples of partial optimization in convex problems.

a. For responses $y \in \mathbb{R}^n$, and features $X_1 \in \mathbb{R}^{n \times d_1}$ and $X_2 \in \mathbb{R}^{n \times d_2}$, consider

$$\underset{\beta_1, \beta_2}{\text{minimize}} \quad \frac{1}{2}\|y - X_1\beta_1 - X_2\beta_2\|_2^2 + \lambda\|\beta_1\|_1.$$

This is like the (usual form) lasso problem (4.4), but where the β_2 block of coefficients is not penalized. Since the above problem is simply a quadratic in β_2, we can minimize over it, for fixed β_1, and find that a minimizer $\hat{\beta}_2(\beta_1)$ (which is unique if and only if X_2 is full column rank) satisfies

$$X_2\hat{\beta}_2(\beta_1) = P_{X_2}(y - X_1\beta_1)$$

where $P_{X_2} = X_2(X_2^T X_2)^+ X_2$ is the projection onto $\mathrm{col}(X_2)$. By plugging this into the above original problem, we get an equivalent problem

$$\underset{\beta_1}{\text{minimize}} \quad \frac{1}{2}\|P_{X_2}^\perp y - P_{X_2}^\perp X_1 \beta_1\|_2^2 + \lambda\|\beta_1\|_1,$$

where $P_{X_2}^\perp = I - P_{X_2}$ is the projection onto the orthocomplement $\mathrm{col}(X_2)$. Note that this is now a (usual form) lasso problem with response $P_{X_2}^\perp y$ and features $P_{X_2}^\perp X_1$.

b. Consider the SVM problem (4.5). Note that, after rearrangement, the linear constraints are equivalent to

$$\xi_i \geq \big[1 - y_i(\beta_0 + x_i^T\beta_0)\big]_+, \quad i = 1, \ldots, n,$$

where $a_+ = \max\{a, 0\}$ denotes the positive part of a. Looking at the criterion in (4.5), we can see that a minimizer over the ξ block of variables, as a function of the others, achieves all equalities in the above set of inequalities (and this is unique if $C > 0$):

$$\hat{\xi}_i(\beta_0, \beta) = \big[1 - y_i(\beta_0 + x_i^T\beta_0)\big]_+, \quad i = 1, \ldots, n.$$

This holds as increasing ξ_i from $\hat{\xi}_i(\beta_0, \beta)$ by any amount $\delta \geq 0$ increases the criterion by $C\delta \geq 0$. Plugging this in results in what is called the *hinge form* of the SVM problem:

$$\underset{\beta_0, \beta}{\text{minimize}} \quad C\sum_{i=1}^n \big[1 - y_i(\beta_0 + x_i^T\beta_0)\big]_+ + \frac{1}{2}\|\beta\|_2^2. \tag{4.13}$$

This has a familiar "loss + penalty" form, much like ridge regression, but with the hinge loss replacing what would be the squared loss in the ridge regression problem.

C. Monotone criterion transformation. If ϕ is increasing then problem (4.1), with feasible set abbreviated by C, is equivalent to

$$\underset{x}{\text{minimize}} \quad \phi(f(x)) \quad \text{subject to} \quad x \in C,$$

and hence also equivalent to, for any decreasing ψ,

$$\underset{x}{\text{maximize}} \quad \psi(f(x)) \quad \text{subject to} \quad x \in C.$$

It is worth noting that monotone transformations can influence convexity or concavity in nontrivial ways. For example, $f(x) = e^{-x^2}$ is neither convex nor concave, but for $\psi(u) = -\log(u)$ (decreasing) we get $\psi(f(x)) = x^2$, which is convex. Therefore it can be useful to apply a monotone transformation to the criterion in an optimization problem, because it can turn a nonconvex problem into a convex problem. An important example of this is maximum likelihood estimation in log-concave families, as discussed in Chapter 4.5.

It is also worth noting that monotone transformations can just as well be applied in order to recast constraints; for example, if ϕ is increasing and ψ decreasing, then

$$g(x) \leq 0 \iff \phi(g(x)) \leq \phi(0) \iff \psi(g(x)) \geq \psi(0).$$

D. One-to-one variable transformation. If $\phi : \mathbb{R}^k \to \mathbb{R}^d$ is one-to-one (meaning, $u \neq v \implies \phi(u) \neq \phi(v)$), and $\mathrm{ran}(\phi)$ contains the feasible set C in problem (4.1), then (4.1) is equivalent to

$$\underset{y}{\text{minimize}} \quad f(\phi(y)) \quad \text{subject to} \quad \phi(y) \in C.$$

Such a variable transformation can be useful for various reasons: for example, it can be dimension-reducing ($k < d$), or it can transform a nonconvex problem into a convex one.

Example 4.5. An optimization problem with linear equality constraints,

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & g_i(x) \leq 0,\ i = 1, \ldots, m \\ & Ax = b, \end{array}$$

can always be recast, via variable transformation, into one that has only inequality constraints. Assuming $Ax = b$ has a solution x_0 (otherwise the above problem is infeasible), we can write any and all solutions as $x = x_0 + Vy$, where the columns of $V \in \mathbb{R}^{d \times k}$ form a basis for $\mathrm{null}(A)$ and $y \in \mathbb{R}^k$. Therefore the above problem is equivalent to

$$\begin{array}{ll} \underset{y}{\text{minimize}} & f(x_0 + Vy) \\ \text{subject to} & g_i(x_0 + Vy) \leq 0,\ i = 1, \ldots, m. \end{array}$$

Though the above two problems are always equivalent, it is not always a good idea, practically speaking, to seek to solve the second instead of the first—and this is not a simple matter of comparing k to d. Eliminating equality constraints can alter problem structure in nontrivial ways. For example, if A sparse, then it might not be easy (or even possible) to find a sparse basis V for $\mathrm{null}(A)$, and such sparsity could be lost in moving from the first problem to the second. This could make a big difference in practice, depending on the specifics (the size of the problem, the algorithm one has in mind for computing a solution, etc.).

E. Slack variables. Proceeding in somewhat of an opposite direction to the last example (eliminating equality constraints), we can replace all inequality constraints in (4.1) with equality constraints, using the following equivalent formulation:

$$\begin{array}{ll} \underset{x,s}{\text{minimize}} & f(x) \\ \text{subject to} & g_i(x) + s_i = 0,\ i = 1, \ldots, m \\ & h_j(x) = 0,\ j = 1, \ldots, k, \\ & s \geq 0. \end{array}$$

The variables s_i, $i = 1, \ldots, m$ here are called *slack variables*. Comparing the above problem to (4.1), we have increased the problem dimension (from d to $d + m$); however, in some settings, reducing the inequality constraints to simple nonnegativity constraints brings other advantages.

In general, introducing slack variables can turn a convex problem into a nonconvex one; only for an affine function g_i is the equality constraint $g_i(x) + s_i = 0$ affine in x, s.

F. Relaxation. If $\tilde{C} \supseteq C$, then

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad x \in \tilde{C}$$

is called a *relaxation* of

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad x \in C.$$

The following properties are easily verified:

(i) if $f^\star(C), f^\star(\tilde{C})$, denote the optimal values in the original and relaxed problems, respectively, then it always holds that $f^\star(\tilde{C}) \leq f^\star(C)$;

(ii) if $x^\star$ denotes a solution in the relaxed problem and $x^\star \in C$, then $x^\star$ is also a solution in the original problem and $f^\star(\tilde{C}) = f^\star(C)$.

If $f^\star(\tilde{C}) = f^\star(C)$, then the relaxation is said to be *tight*. If the relaxation is a convex optimization problem, then it is called a *convex relaxation*.

Example 4.6. Consider the problem of finding the best rank k approximation of $X \in \mathbb{R}^{n\times d}$, for a given positive integer $k \leq \mathrm{rank}(X)$,

$$\underset{\Theta}{\text{minimize}} \quad \|X - \Theta\|_F^2 \quad \text{subject to} \quad \mathrm{rank}(\Theta) = k, \tag{4.14}$$

where the Frobenius norm of a matrix $A \in \mathbb{R}^{n\times d}$ is denoted $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^d A_{ij}^2}$. The above problem is nonconvex because of its constraint ($\mathrm{rank}(\cdot)$ is not an affine function over matrices; it is not even a convex function). Nevertheless, it has a well-known solution, by the *Eckart-Young-Mirsky theorem*, in terms of the singular value decomposition (SVD) of X:

$$\hat{\Theta} = \sum_{i=1}^k \sigma_i u_i v_i^\mathsf{T}. \tag{4.15}$$

Here σ_i denotes the i^{th} largest singular value of X, and $u_i \in \mathbb{R}^n, v_i \in \mathbb{R}^d$ are the corresponding left and right singular vectors of X. The solution in (4.15) is unique if and only if $\sigma_k \neq \sigma_{k+1}$. In statistics—where X would typically represent a data matrix (n samples observed over d features)—we call $\hat{\Theta}$ the reconstruction of X from its top k *principal components*.

As it turns out, the nonconvex problem (4.14) admits, after reformulation, a tight convex relaxation. To see this, we start by reformulating the problem as (Exercise 4.7 part a):

$$\underset{P}{\text{minimize}} \quad \|X - XP\|_F^2 \quad \text{subject to} \quad P \in \mathcal{P}_k, \tag{4.16}$$

where $\mathcal{P}_k$ denotes the set of rank k projection matrices, from $\mathbb{R}^d$ to $\mathbb{R}^d$. (A solution here is given by $\hat{P} = \sum_{i=1}^k v_i v_i^\mathsf{T}$, which is again unique iff $\sigma_k \neq \sigma_{k+1}$.) Now define an inner product on matrices $A, B \in \mathbb{R}^{n\times d}$ by $\langle A, B\rangle = \mathrm{tr}(A^\mathsf{T} B)$, and observe that $\|A\|_F^2 = \langle A, A\rangle$. After some basic manipulations (Exercise 4.7 part b), we arrive at the equivalent problem

$$\underset{P}{\text{maximize}} \quad \langle S, P\rangle \quad \text{subject to} \quad P \in \mathcal{P}_k, \tag{4.17}$$

where $S = X^\mathsf{T} X$. A convex relaxation of (4.17) is thus given by

$$\underset{P}{\text{maximize}} \quad \langle S, P\rangle \quad \text{subject to} \quad P \in \mathcal{F}_k, \tag{4.18}$$

where $\mathcal{F}_k = \mathrm{conv}(\mathcal{P}_k)$. Furthermore, the relaxation (4.18) is tight, which follows by arguing that the maximum of the linear criterion in (4.18) is attained by one of the extreme points of $\mathcal{F}_k$, which are exactly the rank k projection matrices in $\mathcal{P}_k$ (Exercise 4.7 part c).

It is interesting to note that the set $\mathcal{F}_k = \mathrm{conv}(\mathcal{P}_k)$, the convex hull of rank k projection matrices, can be written explicitly as (Exercise 4.7 part d):

$$\mathcal{F}_k = \{P \in \mathbb{S}^d : 0 \preceq P \preceq I, \mathrm{tr}(P) = k\}. \tag{4.19}$$

This is often referred to as the *Fantope* of order k, named after mathematician Ky Fan.

4.4. Existence of minima*

It is not hard to see that some notion of continuity is required in order to guarantee the existence of minima (guarantee a function attains its infimum) in general. For example, the function

$$f(x) = \begin{cases} x^2 & x \neq 0 \\ 1 & x = 0 \end{cases} \tag{4.20}$$

has an infimum of 0, but no minimizer. In order to prevent such behavior, we need some condition that restricts what can happen as we move along a sequence that drives a function to its infimum.

A weak notion of continuity that is useful for reasoning about the existence of minima is called *lower semicontinuity*. A function $f : \mathbb{R}^d \to [-\infty, \infty]$ is lower semicontinuous at a point x if, for any sequence such that $\lim_{k\to\infty} x_k = x$, it holds that

$$f(x) \leq \liminf_{k\to\infty} f(x_k).$$

(We emphasize that, in our definition, the point x does not need to be in $\mathrm{dom}(f)$, and the sequence $\{x_k\}_{k=1}^{\infty}$ does need to be contained in $\mathrm{dom}(f)$.) We call a function f lower semicontinuous if it is lower semicontinuous at each $x \in \mathbb{R}^d$. The next lemma collects a few relevant equivalences.

Lemma 4.7. *For any $f : \mathbb{R}^d \to [-\infty, \infty]$, the following statements are equivalent:*

(i) f is lower semicontinuous;
(ii) its sublevel sets $\{x : f(x) \leq t\}$, $t \in \mathbb{R}$ are all closed;
(iii) its epigraph $\mathrm{epi}(f) = \{(x, t) : f(x) \leq t\}$ is closed.

We call a function *closed* if its epigraph is closed, and we use this terminology interchangeably with lower semicontinuity (since, according to the above, they are equivalent properties). Recall by Theorem 3.10 part (i) that a convex function f is necessarily continuous, thus lower semicontinuous, at each $x \in \mathrm{int}(\mathrm{dom}(f))$; in this light, we can see that a for convex function, closedness is really just a matter of its behavior on the boundary (if nonempty) of its effective domain.

Lower semicontinuity does rule out examples such as (4.20), but cannot by itself guarantee the existence of minima. For example, consider $f(x) = e^{-x}$. Thus, it seems that in order to guarantee a function f attains its infimum, we should pair lower semicontinuity with a condition that rules out the possibility of f being minimized "somewhere off at infinity".

First, we recount (a generalization of) *Weierstrass' theorem*, which applies broadly to a function with compact sublevel sets. Here, and in the rest of the results in this section, we implicitly assume that $f \neq \infty$, that is, $\mathrm{dom}(f) \neq \emptyset$, to rule out a trivial case in which f does not attain its infimum.

Theorem 4.8 (Weierstrass' theorem)**.** *A closed function $f : \mathbb{R}^d \to (-\infty, \infty]$ has a nonempty and compact set of minima if there exists some $t \in \mathbb{R}$ such that $\{x : f(x) \leq t\}$ is nonempty and bounded. Two simple sufficient conditions for the latter to hold (existence of a nonempty bounded sublevel set) are:*

(i) $\mathrm{dom}(f)$ is bounded;
(ii) f is coercive*, which means that $\lim_{k\to\infty} f(x_k) \to \infty$ whenever $\lim_{k\to\infty} \|x_k\|_2 \to \infty$.*

The role of closedness of f in Theorem 4.8 is to ensure closedness of its sublevel sets; combined with the assumption that $\{x : f(x) \leq t\}$ is nonempty and bounded, for some $t \in \mathbb{R}$, we see that such a sublevel is indeed compact. Hence the sublevel sets of f at all levels $s \leq t$ are also compact. The result follows by recognizing that the set of minima of f is the intersection of its nonempty sublevel sets; see Exercise 4.11.

In the constrained case, for minimization of f over a set $C \subseteq \mathrm{dom}(f)$, the following is a special case of the Weierstrass theorem, obtained by applying Theorem 4.8 to $f + I_C$.

Corollary 4.9. *For a function $f : \mathbb{R}^d \to (-\infty, \infty]$, lower semicontinuous at every point in a nonempty closed set $C \subseteq \mathrm{dom}(f)$, the optimization problem*

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad x \in C$$

has a nonempty and compact solution set if there exists $t \in \mathbb{R}$ such that $\{x \in C : f(x) \leq t\}$ is nonempty and bounded. Two simple sufficient conditions for the latter to hold are:

(i) C is bounded;

(ii) f is coercive *on C, which means that $\lim_{k\to\infty} f(x_k) \to \infty$ whenever $\{x_k\}_{k=1}^{\infty} \subseteq C$ and $\lim_{k\to\infty} \|x_k\|_2 \to \infty$.*

For a convex function, more can be said. Towards this end, it is helpful to introduce the notion of a *direction of recession* of a convex function f: this is a direction $v \in \mathbb{R}^d$ such that

$$\lambda \mapsto f(x + \lambda v) \text{ is a nonincreasing function of } \lambda \in \mathbb{R}, \quad \text{for all } x \in \mathrm{dom}(f). \tag{4.21}$$

Viewed geometrically, this is a statement about the existence of rays (running parallel to v) within the sublevel sets of f. The next lemma provides the details. It helps to analogously define a *direction of recession* of a convex set $C \subseteq \mathbb{R}^d$: this is $v \in \mathbb{R}^d$ such that $\{x + \lambda v : \lambda \geq 0\} \subseteq C$, for all $x \in C$.

Lemma 4.10. *For a closed convex function f, the following are equivalent:*

(i) v is a direction of recession of f;

(ii) v is a direction of recession of every $V_t = \{x : f(x) \leq t\}$, $t \in \mathbb{R}$;

(iii) there exists $t \in \mathbb{R}$ and $x \in V_t$ such that $\{x + \lambda v : \lambda \geq 0\} \subseteq V_t$;

(iv) there exists $x \in \mathrm{dom}(f)$ such that $\lambda \mapsto f(x + \lambda v)$ is nonincreasing.

Furthermore, as for the other direction, the following are equivalent:

(v) f has no directions of recession;

(vi) no sublevel set of f contains a ray;

(vii) every sublevel set of f is bounded.

The key fact underlying the above lemma is that, for a closed convex set C:

$$\{x + \lambda v : \lambda \geq 0\} \subseteq C, \text{ for one } x \in C \iff \{x + \lambda v : \lambda \geq 0\} \subseteq C, \text{ for all } x \in C. \tag{4.22}$$

This is investigated in Exercise 4.12, along with the proof of Lemma 4.10.

Looking at statement (vii) in Lemma 4.10, we see that for closed convex f, having no directions of recession implies (indeed, is equivalent to) boundedness of its sublevel sets. Furthermore, as all sublevel sets of f are closed (since f is), this gives us the compactness condition needed in order to apply Weierstrass' theorem.

Theorem 4.11. *A closed convex function f has a nonempty and compact set of minima if it has no directions of recession.*

Alternatively, if the only directions of recession of f are directions in which it is constant then f has a nonempty but noncompact set of minima, which has the form $S^\star = S + L$ for a compact set S and linear subspace L.

The second part of the above theorem goes beyond what is expected from Weierstrass' theorem: even if f has a direction of recession, as long as this is a direction in which it is constant, then for a

closed convex function this constant value must be equal to its infimum, which is therefore attained (Exercise 4.13).

For constrained convex minimization, the following corollary is obtained by applying the previous theorem to $f + I_C$. We introduce one more concept: if v and $-v$ are both directions of recession of a closed convex set C, then we say v is direction in which C is *linear*.

Corollary 4.12. *For a closed convex function f and nonempty closed convex set $C \subseteq \mathrm{dom}(f)$, the convex optimization problem*

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad x \in C$$

has a nonempty and compact solution set if f and C have no common directions of recession.

Alternatively, if the only directions of recession that f and C share are directions in which f is constant and C is linear, then the above problem has a nonempty but noncompact solution set, which has the form $S^\star = S + L$ for a compact set S and linear subspace L.

4.5. Maximum likelihood*

Maximum likelihood estimation is a central problem in statistics: given independent and identically distributed (i.i.d.) samples $Z_1, \ldots, Z_n$ from a distribution P_θ, we estimate the parameter θ of P_θ in such a way that the probability of the observed data is as large as possible, that is, by solving

$$\underset{\theta}{\text{maximize}} \quad \prod_{i=1}^n L(\theta; Z_i). \tag{4.23}$$

Here $\prod_{i=1}^n L(\theta; Z_i)$ is called the *likelihood function* (interpreted as a function of θ), and defined via $L(\theta; z) = p_\theta(z)$, where p_θ is the probability density function associated with P_θ (or probability mass function, in the discrete case). For completeness, we note that maximum likelihood estimation as a concept is by no means actually limited to the i.i.d. case, and we restrict our attention to it in this section only for simplicity.

An equivalent problem to (4.23) (using a monotone criterion transformation, Property 4.3.C) is

$$\underset{\theta}{\text{minimize}} \quad -\sum_{i=1}^n \ell(\theta; Z_i), \tag{4.24}$$

where $\ell(\theta; z) = \log(L(\theta; z))$. In somewhat of an abuse of nomenclature, both (4.23) and (4.24) are referred to as maximum likelihood estimation, and a solution $\hat{\theta}$, if one exists, is called a maximum likelihood estimator (MLE).

In general, a function f is said to be *log-convex* if $\log(f)$ is convex, and similarly *log-concave* if $\log(f)$ is concave. Clearly, if the map $\theta \mapsto L(\theta; z)$ is log-concave (for any fixed z), then maximum likelihood estimation, in its equivalent form (4.24), is a convex problem.

Of special interest among log-concave likelihoods are those associated with exponential family distributions, which have a particularly nice structure. A probability density (or mass) function is said to be of *exponential family* form if it can be written as

$$p_\eta(z) = \exp\big(T(z)^\mathsf{T}\eta - \psi(\eta)\big)h(z), \tag{4.25}$$

Here $\eta \in \mathbb{R}^d$ is called the *natural parameter* of the exponential family (we use η, rather than θ, to adhere to the standard notation in statistics). The function ψ is called the *log-partition function*. A remarkable fact: in any exponential family distribution, the log-partition function ψ is automatically convex, by virtue of the fact that p_η must be a bona fide density and therefore must integrate to one

(sum to one, in the discrete case); see Exercise 4.14. This means that the map $\eta \mapsto p_\eta(z)$ is always log-concave (for fixed z), and the resulting maximum likelihood problem

$$\underset{\eta}{\text{minimize}} \quad -\frac{1}{n}\Big(\sum_{i=1}^n T(Z_i)\Big)^\mathsf{T}\eta + \psi(\eta)$$

is always convex. Some examples of well-known exponential family distributions are the Gaussian, Bernoulli, Poisson, gamma, and beta distributions (Exercise 4.15).

Exponential families provide the foundation for the study of *generalized linear models* (GLMs). In this setting, we observe independent draws of a response variable y_i, $i = 1, \ldots, n$, with each one sampled from an exponential family distribution. The form of this exponential family distribution is common across samples (the functions T, ψ, h are common), however, the natural parameter is now sample-specific: to each y_i we assign a separate natural parameter η_i, and model it as $\eta_i = x_i^\mathsf{T}\beta$ for an observed feature vector $x_i \in \mathbb{R}^d$ and parameter $\beta \in \mathbb{R}^d$. Maximum likelihood becomes:

$$\underset{\beta}{\text{minimize}} \quad \sum_{i=1}^n \Big(-T(y_i)x_i^\mathsf{T}\beta + \psi(x_i^\mathsf{T}\beta)\Big). \tag{4.26}$$

Below we specify three cases for the underlying exponential family distribution, namely, Gaussian: $\psi(u) = \frac{u^2}{2}$, Bernoulli: $\psi(u) = \log(1+e^u)$, and Poisson: $\psi(u) = e^u$, and examine the corresponding GLM optimization problem. In each case, the function T (the natural sufficient statistic) equals the identity. As we see, this recovers the linear, logistic, and Poisson regression problems.

$$\text{Gaussian}: \quad \underset{\beta}{\text{minimize}} \quad \sum_{i=1}^n \Big(-y_i x_i^\mathsf{T}\beta + \frac{(x_i^\mathsf{T}\beta)^2}{2}\Big) \qquad \text{(linear regression)} \tag{4.27}$$

$$\text{Bernoulli}: \quad \underset{\beta}{\text{minimize}} \quad \sum_{i=1}^n \Big(-y_i x_i^\mathsf{T}\beta + \log(1+\exp(x_i^\mathsf{T}\beta))\Big) \qquad \text{(logistic regression)} \tag{4.28}$$

$$\text{Poisson}: \quad \underset{\beta}{\text{minimize}} \quad \sum_{i=1}^n \Big(-y_i x_i^\mathsf{T}\beta + \exp(x_i^\mathsf{T}\beta)\Big) \qquad \text{(Poisson regression)} \tag{4.29}$$

(These are each written without an explicit intercept term in the model; to keep the same notation but accomodate an intercept term, we can append an entry of 1 to each feature vector $x_i \in \mathbb{R}^d$, or equivalently, a column of all 1s to the corresponding feature matrix $X \in \mathbb{R}^{n\times d}$.)

The first problem (4.27) can of course be equivalently posed as minimization of $\sum_{i=1}^n (y_i - x_i^\mathsf{T}\beta)^2$, or in a more compact form,

$$\underset{\beta}{\text{minimize}} \quad \|y - X\beta\|_2^2,$$

where $y \in \mathbb{R}^n$ is the response vector and $X \in \mathbb{R}^{n\times d}$ is the feature matrix. This is the more familiar least squares formulation of linear regression, and we can clearly recognize the lasso problem (4.4) as the ℓ_1 regularized version of the above. In the same vein, regularization can be applied to any GLM maximum likelihood problem. This recipe—a GLM loss plus a convex penalty (usually, a norm or seminorm)—is a core problem form that we will frequently use to motivate our study in the chapters that follow.

We finish this section by reviewing a few classical results on the existence of solutions in linear, logistic, and Poisson regression. These results are, in effect, consequences of Weierstrass' theorem for convex functions, Theorem 4.11. Later, in Chapter 12.1, we will derive a generalization via duality arguments (this will also provide simple proofs for the logistic and Poisson results).

Theorem 4.13. *In problems (4.27), (4.28), (4.29), let $y \in \mathcal{Y}^n$ denote the response vector and $X \in \mathbb{R}^{n \times d}$ the feature matrix.*

(i) In linear regression (4.27) where $\mathcal{Y} = \mathbb{R}$, a solution always exists. Further, the solution set $\hat{S}$ is an affine subspace, namely $\hat{S} = (X^T X)^+ X^T y + \mathrm{null}(X)$.

(ii) In logistic regression (4.28), where $\mathcal{Y} = \{0, 1\}$, a solution exists provided:

$$\text{there does not exist } \beta \neq 0 \text{ such that } (2y_i - 1)x_i^T \beta \geq 0,\ i = 1, \ldots, n. \tag{4.30}$$

(iii) In Poisson regression (4.29), where $\mathcal{Y} = \mathbb{N}$ (recall that $\mathbb{N} = \{0, 1, 2, \ldots\}$ denotes the set of natural numbers), a solution exists provided:

$$\text{there exists } \delta \in \mathrm{null}(X^T) \text{ such that } y_i + \delta_i > 0,\ i = 1, \ldots, n. \tag{4.31}$$

Remark 4.14. The result for linear regression, in part (i), was essentially already given in Example 4.2.c (on general convex quadratic functions). The results for logistic and Poisson regression, in parts (ii) and (iii), are well-known results from the statistics literature (see the bibliographic references given shortly, at the end of the chapter). Interestingly, the conditions (4.30), (4.31) are known from the literature to be not only sufficient (as Theorem 4.13 states) but also *necessary* for the existence of an MLE.

Chapter Notes

Just as with the last chapter, a lot can be said about the basic principles of (convex) optimization; for readers seeking more, there are many excellent references, such as [**Roc70**] (Chapters 27, 28), [**BV04**] (Chapter 4), and [**Ber09**] (Chapter 3). Some remarks on nomenclature: some authors, for example [**Roc70, Ber09**], use the term *feasible solution* to mean what we call a feasible point and the term *optimal solution* to mean what we call a solution; Theorem 4.8, which we call Weierstrass' theorem, is actually a generalization of Weierstrass' classical extreme value theorem, from [**Ber09**], Proposition 3.2.1. Several related exercises presented at the end of the current chapter are adapted from proofs of results in [**Ber09**]. To learn more about directions of recession, and how this relates to the existence of solutions, see [**Roc70**] (Chapters 7, 27) and [**Ber09**] (Chapters 1.4, 3.2).

The Eckart-Young-Mirsky theorem is named after [**EY36**], who proved the result for the case of the Frobenius norm (they showed (4.15) solves problem (4.14)), and after [**Mir60**], who generalized this to *unitarily invariant* matrix norms (he showed (4.15) solves a generalized version of (4.14), in which the criterion uses any norm $\|\cdot\|$ that satisfies $\|UAV\| = \|A\|$, for any A and orthogonal U, V). However, the Frobenius norm result was actually proved much earlier by [**Sch07**].

The lasso was first proposed by [**Tib96**], and the idea also appeared independently in [**CDS98**]. The support vector machine first appeared in [**BGV92**], although Vladimir Vapnik had apparently developed the idea much earlier; the standard soft-margin formulation, which is used in this book, is due to [**CV95**]. The lasso and the SVM (especially its kernel version) are of course cornerstone methods in modern statistics and machine learning, and there are many excellent books that cover them in great detail, including [**HTF09, SS02, HTJ15**]. We will also dive into greater detail on the lasso and the SVM in Chapters 13 and 14, respectively.

Lastly, maximum likelihood plays a central role in classical statistical inference, and there are numerous books that cover the topic nicely, for example [**CH74, Sil75, LC98, Was04**], and also [**MN89**] (on GLMs in particular). On the existence of MLEs in GLMs: the condition (4.30) in the logistic case can be found in [**AA84**], and (4.31) in the Poisson case can be found in [**Hab73**].

Exercises

4.1 Prove that the margin (that is, the minimum ℓ_2 distance) between the two hyperplanes

$$H_+ = \{x : \beta_0 + x^\mathsf{T}\beta = 1\} \quad \text{and} \quad H_- = \{x : \beta_0 + x^\mathsf{T}\beta = -1\}$$

is equal to $2/\|\beta\|_2^2$. Hint: pick two points $x_+ \in H_+$ and $x_- \in H_-$, and argue that the margin is the inner product of $x_+ - x_-$ with the outward normal $\beta/\|\beta\|_2$ pointing from H_- to H_+.

4.2 We explore the convexity of various important sets that can be defined relative to a convex optimization problem (4.2), as referenced in Properties 4.2.A and 4.2.B.

a. Prove that the feasible set of (4.2) is convex.

b. Prove that the *ϵ-suboptimal set* of (4.2), defined by

$$S_\epsilon = \{x : x \text{ is feasible for problem (4.2)}, \text{ and } f(x) - f^\star \leq \epsilon\},$$

is a convex set, for any $\epsilon \geq 0$. Hint: recognize that S_ϵ is the intersection of the feasible set with a sublevel set of f, which are both convex sets (by part a, and Exercise 3.2).

c. Argue that $S^\star$ is convex as a special case of part b.

d. If f is strictly convex, then prove that $S^\star$ must either be nonempty or a singleton.

e. If $f = F(Mx + d)$, where F is a strictly convex function, M an arbitrary matrix, and d an arbitrary vector, then prove that

$$MS^\star + d = \{Mx^\star + d : x^\star \in S^\star\}$$

must either be nonempty or a singleton. In other words, if a solution $x^\star$ in (4.2) exists, then $Mx^\star + d$ must be unique.

4.3 Show that the more general first-order condition (4.7) is equivalent to the Lagrange multiplier condition (4.9) when the feasible set is an affine subspace, $C = \{x : Ax = b\}$. Hint: proceed in two steps. For the first step, argue that $\nabla f(x)^\mathsf{T} v = 0$ for all $v \in \text{null}(A)$. For the second step, use the fact that the null space and row space of a matrix are orthocomplements.

4.4 Prove that the optimization problem (4.1) is equivalent to

$$\underset{x}{\text{minimize}} \quad \sup_{u \geq 0,\, v} \underbrace{\left\{ f(x) + \sum_{i=1}^m u_i g_i(x) + \sum_{j=1}^k v_j h_j(x) \right\}}_{L(x,u,v)}. \tag{4.32}$$

This is sometimes referred to as the *saddle point* or *min-max* form of problem (4.1); and the innermost function, labeled $L(x, u, v)$, is what we will later define as the Lagrangian of (4.1).

4.5 We explore various rules for combining or exchanging infimums and supremums. Throughout, $f : \mathbb{R}^d \to [-\infty, \infty]$ is a general function acting on a block variable $x = (x_1, x_2)$.

a. Prove that

$$\inf_{x_1, x_2} f(x_1, x_2) = \inf_{x_1} \inf_{x_2} f(x_1, x_2).$$

b. By just exchanging the roles of x_1, x_2, argue that

$$\inf_{x_1} \inf_{x_2} f(x_1, x_2) = \inf_{x_2} \inf_{x_1} f(x_1, x_2).$$

c. Argue that the results in parts a, b hold when we replace infimums with supremums.

d. Now prove that

$$\inf_{x_1} \sup_{x_2} f(x_1, x_2) \geq \sup_{x_2} \inf_{x_1} f(x_1, x_2).$$

e. Give an example to show that the inequality in part d can be loose, in general.

A remark on the interpretation: in the language of duality, covered in Chapter 10.1, part d says that weak duality holds in general, for any optimization problem; and part e reminds us that strong duality need not hold in general, without further assumptions.

4.6 In this exercise, we will look at various reformulations of optimization problems involving ℓ_1 penalties.

a. First prove that the function $g : \mathbb{R} \times \mathbb{R}_+ \to (-\infty, \infty]$, defined as

$$g(x,y) = \begin{cases} x^2/y + y & y > 0 \\ 0 & x = 0,\ y = 0 \\ \infty & x \neq 0,\ y = 0 \end{cases},$$

is convex. Show that $\inf_{y \geq 0} g(x,y) = 2|x|$.

b. For any function $f : \mathbb{R}^n \to [-\infty, \infty]$, prove that

$$\underset{\theta}{\text{minimize}} \quad f(\theta) + \lambda \|\theta\|_1$$

and

$$\underset{\theta,\sigma}{\text{minimize}} \quad f(\theta) + \frac{\lambda}{2} \sum_{i=1}^n g(\theta_i, \sigma_i) \quad \text{subject to} \quad \sigma \geq 0$$

are equivalent problems. Hint: apply partial optimization in the last problem, and use the result of part a.

c. Prove that the first problem in part b is also equivalent to

$$\underset{u,v}{\text{minimize}} \quad f(u \odot v) + \frac{\lambda}{2}(\|u\|_2^2 + \|v\|_2^2),$$

where $u \odot v = (u_1 v_1, \ldots, u_n v_n)$ is the Hadamard (elementwise) product between u, b.

d. Argue that if f is convex, then $h : \mathbb{R}^n \times \mathbb{R}^n \to (-\infty, \infty]$, defined by $h(u,v) = f(u \odot v)$, need not be convex. Thus the reformulation in part c does not retain convexity, though that in part b does.

4.7 We work through the details of the equivalences described in Example 4.6.

a. By decomposing $\Theta = \Theta P_X + \Theta P_X^\perp$, where P_X denotes the projection onto $\mathrm{col}(X)$ and $P_X^\perp = I - P_X$, argue using orthogonality that

$$\|X - \Theta\|_F^2 = \|X - P_X\Theta\|_F^2 + \|P_X^\perp \Theta\|_F^2,$$

for any $\Theta \in \mathbb{R}^{n \times d}$. Argue, therefore, that the solution in problem (4.14) must occur at Θ with $P_X^\perp \Theta = 0$, that is, $\Theta = P_X \Theta$. Prove that any Θ satisfying this condition, that also satisfies $\mathrm{rank}(\Theta) = k$, can be written as $\Theta = XP$ for a rank k projection matrix P, completing the equivalence between (4.14) and (4.16).

b. Prove that $\|X - XP\|_F^2 = \|X\|_F^2 - \langle S, P \rangle$, using the fact that $P^\mathsf{T} = P$ (symmetry) and $P^2 = P$ (idempotency) for a projection matrix P, where $S = X^\mathsf{T} X$, and thus argue that (4.16) and (4.17) are equivalent.

c. As an application of Exercise 3.15, show that the maximum in (4.18) is attained by an element of $\mathcal{P}_k$, certifying the equivalence between (4.17) and (4.18).

d. Prove that the set $\mathcal{P}_k$ of rank k projection matrices (from $\mathbb{R}^d$ to $\mathbb{R}^d$) can be written as

$$\mathcal{P}_k = \{P \in \mathbb{S}^d : \lambda_i(P) \in \{0,1\},\ i = 1, \ldots, d,\ \mathrm{tr}(P) = k\},$$

where $\lambda_i(P)$ is the i^{th} largest eigenvalue of P. Prove from this representation that its convex hull $\mathcal{F}_k = \mathrm{conv}(\mathcal{P}_k)$, the Fantope of order k, takes the form claimed in (4.19).

4.8 In this exercise, we will prove the best rank k approximation problem (4.14) admits (4.15) as a solution. We will also show that the same result is true when the Frobenius norm $\|\cdot\|_F$ in (4.14) is replaced by the operator norm $\|\cdot\|_{\text{op}}$.

a. First show that for any matrices A, B and any i, j,

$$\sigma_{i+j+1}(A+B) \leq \sigma_{i+1}(A) + \sigma_{j+1}(B). \tag{4.33}$$

For convenience, here and in what follows, we let $\sigma_\ell(M) = 0$ when $\ell > \text{rank}(M)$, for a matrix M. The above result is called *Weyl's singular value pertubation inequality*. Hint: use the min-max (Courant-Fischer) representation for the singular values of a matrix, as reviewed in Appendix C.1.

b. Using (4.33), show that for any rank k matrix $\Theta \in \mathbb{R}^{n\times d}$ and any i,

$$\sigma_i(X - \Theta) \geq \sigma_{k+i}(X).$$

Hint: take $A = X - \Theta$ and $B = \Theta$; use the fact that Θ has rank k.

c. Use the last part to argue that

$$\|X - \Theta\|_F^2 \geq \sum_{i=k+1}^{d} \sigma_i(X).$$

Show that the right-hand side above is $\|X - \hat{\Theta}\|_F^2$, for $\hat{\Theta}$ as in (4.15), which establishes that $\hat{\Theta}$ is indeed a solution.

d. Based on similar arguments, prove that the same result holds when the Frobenius norm in (4.14) is replaced by the operator norm.

4.9 Prove Lemma 4.7. Hint: show that (i) $\Longrightarrow$ (iii) $\Longrightarrow$ (ii) $\Longrightarrow$ (i).

4.10 In preparation for the next exercise, prove that a sequence $C_1 \supseteq C_2 \supseteq C_3 \supseteq \cdots$ of nonempty nested compact sets must have a nonempty intersection, $\cap_{k=1}^{\infty} C_k \neq \emptyset$. This is called *Cantor's intersection theorem*. Hint: proceed by contradiction, and in doing so, define $U_k = C_1 \setminus C_k$, $k = 1, 2, 3, \ldots$, then use the topological definition of compactness (Heine-Borel theorem).

4.11 We work through the proof of Weierstrass' theorem, Theorem 4.8.

a. Let t be such that the sublevel set $\{x : f(x) \leq t\}$ is bounded (as assumed to exist in the theorem) Argue that $\{x : f(x) \leq s\}$ is also bounded for all $s \leq t$.

b. Show that the set of minima $S^\star$ of f is nonempty and compact, by part a and Cantor's intersection theorem (Exercise 4.10). Hint: to get started, observe that for any sequence $\{t_k\}_{k=1}^{\infty}$ with $t_k \to f^\star = \inf_x f(x)$ as $k \to \infty$, we have $S^\star = \cap_{k=1}^{\infty}\{x : f(x) \leq t_k\}$.

c. Show that conditions (i) and (ii) in the theorem each imply the existence of t such that $\{x : f(x) \leq t\}$ is bounded.

4.12 We work through the proof of Lemma 4.10. We introduce the concept of the *recession cone* of a convex set $C \subseteq \mathbb{R}^d$, denoted R_C: this is the set of vectors that are directions of recession of C. It is straightforward to check that R_C is indeed a (closed and convex) cone.

a. Show that (i) $\Longleftrightarrow$ (ii) and (iii) $\Longleftrightarrow$ (iv).

b. Prove that each sublevel set V_t, $t \in \mathbb{R}$ of f has the same recession cone, and denoting it by R_f, this cone satisfies

$$R_f = \big\{d : (d, 0) \in R_{\text{epi}(f)}\big\},$$

where $R_{\text{epi}(f)}$ denotes the recession cone of $\text{epi}(f)$, the epigraph of f.

c. Prove property (4.22), for a closed convex set C. Hint: the forward implication here is challenging. For this direction, let $x \in C$ have the given property and assume (without

a loss of generality) that $\|v\|_2 = 1$. Fix any other $y \in C$, let $\{\lambda_k\}_{k=1}^{\infty}$ be a nonnegative and diverging sequence ($\lambda_k \to \infty$ as $k \to \infty$), and define

$$x_k = x + \lambda_k v \quad \text{and} \quad v_k = \frac{x_k - y}{\|x_k - y\|_2}.$$

Prove that $v_k \to v$ as $k \to \infty$, and use this to establish that $\{y + \lambda v : \lambda \geq 0\} \subseteq C$.

d. Putting the results in parts b and c together, argue that (ii) $\iff$ (iii), which, together with part a, establishes the equivalence of (i)–(iv).

e. Now show that (v) $\iff$ (vi).

f. Show that (vi) $\iff$ (vii). Together with the last part, this establishes the equivalence of (v)–(vii). Hint: the forward direction (vi) $\implies$ (vii) is challenging. For this, consider proving the contrapositive: any unbounded closed convex set must contain a ray, which can be shown using a construction similar to that from part c.

4.13 We work through the proof of the second part of Theorem 4.11. (The first part follows from combining Lemma 4.10 and Theorem 4.8, as explained above the statement of Theorem 4.11.) We first introduce the concept of the *lineality space* of a convex set $C \subseteq \mathbb{R}^d$, denoted L_C: this is the set of vectors $v \in \mathbb{R}^d$ such that both v and $-v$ are directions of recession of C. We can see that $L_C = R_C \cap (-R_C)$, where R_C is the recession cone of C, as defined in Exercise 4.12. Further, it is not hard to check that the set L_C is a linear subspace.

The following are some helpful facts about the sublevel sets of a closed convex function f and their lineality spaces. First, all sublevel sets share the same lineality space, denoted L_f (analogous to the fact about recession cones, from Exercise 4.12 part b). Second, any $v \in L_f$ is actually a direction in which f is *constant*, which means $f(x + \lambda v) = f(x)$ for all $\lambda \in \mathbb{R}$ and $x \in \mathrm{dom}(f)$. (To see this, note that f must be nonincreasing along both the ray pointing from any x in the direction v, and along the ray pointing from any x in the direction $-v$, by part (iv) of Lemma 4.10.) With these facts in hand, we proceed with the exercise itself.

a. First prove that for a nonempty convex set C, it holds that $C = L_C + (C \cap L_C^{\perp})$, where $L_C^{\perp}$ denotes the orthocomplement of the linear subspace L_C. Prove furthermore that if the only directions of recession of C are directions in which it is linear, $R_C = L_C$, then $C \cap L_C^{\perp}$ is bounded.

b. Let $C_1 \supseteq C_2 \supseteq C_3 \supseteq \cdots$ be a sequence of nonempty nested closed convex sets such that the following two properties hold: (i) $R_{C_k} = L_{C_k}$ for each k, and (ii) $L_{C_k} = L$ for all k. In words, for each set in the sequence, its only directions of recession are directions of linearity; and all sets share the same lineality space. Prove that such a sequence must have a nonempty intersection, and furthermore,

$$\bigcap_{k=1}^{\infty} C_k = L + S,$$

where S is nonempty and compact. Hint: use the result in part a to write each set C_k as a sum of L and a compact set S_k. Then note $(L + A) \cap (L + B) \supseteq L + (A \cap B)$ for any sets A, B, and judiciously apply Cantor's intersection theorem (Exercise 4.10) to prove that $\cap_{k=1}^{\infty} C_k$ is nonempty.

c. Apply part b to the sublevel sets of f to prove the second part of Theorem 4.11. Hint: use the helpful facts given at the start of this exercise.

4.14 In this exercise, we will prove that for an exponential family distribution, whose probability density or mass function is of the form (4.25), the log-partition function ψ must be convex.

a. Prove that ψ must satisfy

$$\psi(\eta) = \begin{cases} \log\left(\int \exp\left(T(z)^\mathsf{T}\eta\right)h(z)\,\mathsf{d}z\right) & \text{(continuous case)} \\ \log\left(\sum_z \exp\left(T(z)^\mathsf{T}\eta\right)h(z)\right) & \text{(discrete case)} \end{cases}.$$

b. Show that $\mathrm{dom}(\psi)$ is a convex set. Hint: use part a.

c. Show that ψ is twice differentiable. Hint: use part a, then use the Leibniz integral rule to interchange differentiation and integration.

d. Show that $\nabla^2\psi$ is positive semidefinite everywhere, and conclude that ψ is convex.

4.15 In each of the following, express the probability density or mass function of the distribution in question in exponential family form (4.25), and check directly that the log-partition function ψ is convex. For clarity: below when we say a parameter is "fixed", it is to be excluded from the natural parameter vector, when writing the distribution exponential family form.

a. $N(\mu, 1)$, the normal distribution with mean μ and fixed variance 1.

b. $N(\mu, \sigma^2)$, the normal distribution with mean μ and variance σ^2.

c. Bernoulli(p), the Bernoulli distribution with success probability p.

d. Binomial(N, p), the binomial distribution with a fixed number of trials N and success probability p.

e. Poisson(μ), the Poisson distribution with mean μ.

f. Gamma(α, β), the gamma distribution with shape parameter α and rate parameter β.

g. Beta(α, β), the beta distribution with shape parameters α, β.

4.16 In maximum likelihood estimation, let (θ, η) denote a block decomposition of a parameter that determines the distribution of i.i.d. samples $Z_1, \ldots, Z_n$. Abbreviating $L_i(\theta, \eta) = L(\theta, \eta; Z_i)$ for $i = 1, \ldots, n$, we denote the full likelihood by

$$L(\theta, \eta) = \prod_{i=1}^n L_i(\theta, \eta).$$

The *profile likelihood* for θ is defined by

$$\tilde{L}(\theta) = \sup_\eta\, L(\theta, \eta).$$

Denoting by $\ell(\theta, \eta) = \log(L(\theta, \eta))$ the full log likelihood, the profile log likelihood is similarly

$$\tilde{\ell}(\theta) = \sup_\eta\, \ell(\theta, \eta).$$

The MLE $\hat{\theta}$ for θ, the first block in the parameter $(\hat{\theta}, \hat{\eta})$ that maximizes the full likelihood L (equivalently, maximizes the log likelihood ℓ), can also be obtained by maximizing the profile likelihood $\tilde{L}$ (equivalently, $\tilde{\ell}$). Notice that this is simply the principle of partial optimization, Property 4.3.B.

In this exercise, we demonstrate profile likelihood in the context of survival analysis. The Cox model for the hazard function of the survival time of a subject with features $x_i \in \mathbb{R}^d$ is

$$\lambda(t; x_i) = \lambda_0(t)\exp(x_i^\mathsf{T}\theta). \tag{4.34}$$

Here λ_0 is some unspecified base hazard function, and $\theta \in \mathbb{R}^d$ is a parameter to be estimated. Given n subjects with survival times t_i, $i = 1, \ldots, n$, the Cox partial likelihood is

$$L(\theta) = \prod_{i=1}^n \frac{\exp(x_i^\mathsf{T}\theta)}{\sum_{j:t_j \geq t_i} \exp(x_j^\mathsf{T}\theta)}. \tag{4.35}$$

a. Let $\Lambda_0(t) = \int_0^t \lambda_0(t)$ denote the base cumulative hazard function. Show that under the Cox model (4.34), the density of the survival time t_i for subject i is

$$f(t; \theta, \Lambda_0) = \lambda_0(t) \exp(x_i^\mathsf{T}\theta) e^{-\Lambda_0(t)\exp(x_i^\mathsf{T}\theta)}.$$

In other words, the likelihood function for subject i is

$$L_i(\theta, \Lambda_0) = \lambda_0(t_i) \exp(x_i^\mathsf{T}\theta) e^{-\Lambda_0(t_i)\exp(x_i^\mathsf{T}\theta)}.$$

and the full likelihood is

$$L(\theta, \Lambda_0) = \prod_{i=1}^n \lambda_0(t_i) \exp(x_i^\mathsf{T}\theta) e^{-\Lambda_0(t_i)\exp(x_i^\mathsf{T}\theta)}.$$

Hint: recall that for a cumulative distribution function F, having density f, the hazard function is related by $\lambda = f/S$, where $S = 1 - F$ denotes the survival function, and the cumulative hazard satisfies $\Lambda = -\log(S)$.

b. Based on the full likelihood in part a, argue that in maximum likelihood we can restrict our attention to functions Λ_0 that are piecewise constant with jumps at t_i, $i = 1, \ldots, n$. Hence if we write η_i for the jump at t_i, and collect these into a vector $\eta = (\eta_1, \ldots, \eta_n)$, then we can reparametrize the likelihood in a finite-dimensional form,

$$L(\theta, \eta) = \prod_{i=1}^n \eta_i \exp(x_i^\mathsf{T}\theta) e^{-\left(\sum_{j:t_j \le t_i} \eta_j\right)\exp(x_i^\mathsf{T}\theta)}.$$

c. Show that, for any fixed $\theta \in \mathbb{R}^d$, the likelihood $L(\theta, \eta)$ from the last display is maximized over η when

$$\frac{1}{\hat{\eta}_i(\theta)} = \sum_{j:t_j \ge t_i} \exp(x_j^\mathsf{T}\theta), \quad i = 1, \ldots, n.$$

Plugging this in, show that the profile likelihood for θ is thus precisely the Cox partial likelihood (4.35).

d. Repeat the profile likelihood calculation in the case of right censoring: here, instead of observing t_i for each subject i, we observe $y_i = \min\{t_i, c_i\}$ and $\delta_i = 1\{t_i \le c_i\}$, where c_i is a censoring time independent of t_i (and t_i follows the Cox model (4.34), as before).

An note on the difference between this and the previous case: you will have to restrict the cumulative baseline hazard function Λ_0 to be piecewise constant with jumps at the failure times (the subset of y_i, $i = 1, \ldots, n$ such that $\delta_i = 1$), *by assumption*. Under censoring, this no longer falls out of maximizing the full likelihood.

Canonical Problem Forms

5.1. Linear programs

It will be useful to develop a categorization of convex optimization problems, as this will help us reason about problems from a variety of perspectives, in the remainder of this book. The first class we study is that of *linear programs (LPs)*. An LP is an optimization problem of the form

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^{\mathsf{T}} x \\ \text{subject to} & Ax \leq b \\ & Gx = h, \end{array} \tag{5.1}$$

for $c \in \mathbb{R}^d$ and matrix-vector pairs A, b and G, h of compatible dimensions. Its name refers to the fact that the criterion and all constraint functions are linear (to be precise, affine) functions. Note that the constraints above can be written as $x \in P$, for a polyhedron P. Importantly, observe that an LP is always a convex optimization problem.

Example 5.1. The following are classic examples of linear programs.

a. The *diet problem* is an LP to find the cheapest combination of foods that satisfies some nutritional requirements:

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \displaystyle\sum_{j=1}^{n} c_j x_j \\ \text{subject to} & \displaystyle\sum_{j=1}^{n} a_{ij} x_j \geq b_i,\ i = 1, \ldots, m \\ & x \geq 0, \end{array}$$

where c_j is the cost per unit of food j, b_i is the minimum required intake of nutrient i, and a_{ij} is the content of nutrient i per unit of food j. At the solution, $x_j^\star$ is the number of units of food j in the optimal diet.

b. The *transportation problem* is an LP to find the cheapest way to ship items from given sources to destinations:

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} \\ \text{subject to} & \displaystyle\sum_{j=1}^{n} x_{ij} \le s_i,\ i = 1,\dots,m \\ & \displaystyle\sum_{i=1}^{m} x_{ij} \ge d_j,\ j = 1,\dots,n \\ & x \ge 0, \end{array}$$

where c_{ij} is the per unit shipping cost from i to j, s_i is the supply at source i, and d_j is the demand at destination j. At the solution, $x^\star_{ij}$ is the number of units shipped from i to j in the optimal shipping scheme.

In general, not only for LPs but for all problem classes (QPs, SDPs, and so on), we will call a problem an LP provided that it is equivalent to one. For example, we still call (5.1) an LP when the minimization is replaced by maximization.

Next we list some basic properties of linear programs.

A. Standard form. A linear program can always be written in the form (Exercise 5.1 part a)

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^T x \\ \text{subject to} & Ax = b \\ & x \ge 0, \end{array} \tag{5.2}$$

which is known as *standard form*.

B. Recasting ℓ_1 and ℓ_∞ penalties. In an optimization problem, ℓ_1 and ℓ_∞ norm penalties can always be recast using linear penalties and constraints: the problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) + \lambda\|x\|_1 + \gamma\|x\|_\infty \\ \text{subject to} & x \in C \end{array}$$

is equivalent to (Exercise 5.2 part a)

$$\begin{array}{ll} \underset{x,y,z}{\text{minimize}} & f(x) + \lambda 1^T y + \gamma z \\ \text{subject to} & -y \le x \le y \\ & -z1 \le x \le z1 \\ & x \in C,\ y, z \ge 0. \end{array}$$

When f is linear and C is a polyhedron, the above problem is an LP. Furthermore, the analogous equivalence holds for ℓ_1 and ℓ_∞ constraints (see Exercise 5.2 part b).

Example 5.2. The first two examples below are problems that are related to the lasso, and are LPs by Property 5.1.B. The last two are related to the SVM, and are LPs by inspection.

a. Given a response vector $y \in \mathbb{R}^n$ and feature matrix $X \in \mathbb{R}^{n \times d}$, the *basis pursuit* problem seeks a sparse subset of the columns of X that can serve as a linear basis for y:

$$\begin{array}{ll} \underset{\beta}{\text{minimize}} & \|\beta\|_1 \\ \text{subject to} & X\beta = y. \end{array} \tag{5.3}$$

This can be seen as a special case of the lasso problem (4.4) as $\lambda \to 0$.

b. Under the same setup as in the last example, the *Dantzig selector* seeks a sparse subset of the columns of X that can serve an approximate linear basis for y:

$$\begin{array}{ll} \underset{\beta}{\text{minimize}} & \|\beta\|_1 \\ \text{subject to} & \|X^T(y - X\beta)\|_\infty \leq \lambda, \end{array} \tag{5.4}$$

where $\lambda \geq 0$ is a tuning parameter. The constraint in (5.4) can be seen as a relaxation of the familiar zero-gradient condition $X^T(y - X\beta) = 0$ for the least squares problem, in which we minimize $\|y - X\beta\|_2^2$ over β.

c. Given class labels $y_i \in \{-1, 1\}$ and feature vectors $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, the following problem seeks a linear classifier (to predict y_i from the sign of $\beta_0 + x_i^T\beta$) with minimal sum of violation costs to the margin condition:

$$\begin{array}{ll} \underset{\beta_0, \beta, \xi}{\text{minimize}} & \sum_{i=1}^n \xi_i \\ \text{subject to} & y_i(\beta_0 + x_i^T\beta) \geq 1 - \xi_i, \ i = 1, \ldots, n \\ & \xi \geq 0. \end{array} \tag{5.5}$$

This can be seen as a special case of the SVM problem (4.5) as $C \to \infty$.

d. Under the same setup as in the last example, consider the feasibility problem where we seek a linear classifier that linearly separates the two classes:

$$\begin{array}{ll} \text{find} & (\beta_0, \beta) \\ \text{subject to} & y_i(\beta_0 + x_i^T\beta) \geq 1, \ i = 1, \ldots, n. \end{array} \tag{5.6}$$

This is called a *linear feasibility problem* (an LP that is also a feasibility problem).

5.2. Quadratic programs

The second class we consider is that of *quadratic programs (QPs)*. A QP is of the form

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^Tx + \frac{1}{2}x^TQx \\ \text{subject to} & Ax \leq b \\ & Gx = h, \end{array} \tag{5.7}$$

for $c \in \mathbb{R}^d$, $Q \in \mathbb{S}_+^d$, and matrix-vector pairs A, b and G, h of compatible dimensions. The name here refers to the fact that the criterion is a quadratic function; the constraint functions are still linear, as in (5.1). We emphasize that in our definition, the matrix Q that determines the quadratic in (5.7) is *assumed to be positive semidefinite*, and thus a QP in this book is always convex. To distinguish, we will refer to a problem of the form (5.7) with $Q \not\succeq 0$ as a *nonconvex QP*.

Next we list some basic properties of quadratic programs.

A. LPs are QPs. Observe that every LP (5.1) is a QP (5.7) (simply with $Q = 0$).

B. Standard form. As with an LP, a QP can always be written in the form (Exercise 5.1 part b)

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^\mathsf{T}x + \frac{1}{2}x^\mathsf{T}Qx \\ \text{subject to} & Ax = b \\ & x \geq 0, \end{array} \tag{5.8}$$

which is again known as standard form.

C. Closed-form solution for equality constraints only. The QP

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^\mathsf{T}x + \frac{1}{2}x^\mathsf{T}Qx \\ \text{subject to} & Ax = b \end{array}$$

has a closed-form solution which can be found by solving the Lagrange multiplier condition (4.9) (a special case of the first-order optimality condition), together with the linear constraints:

$$\begin{bmatrix} Q & A^\mathsf{T} \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ -b \end{bmatrix}.$$

The matrix on the left-hand side above is often called the *KKT matrix*, as the above condition can be derived using the Karush-Kuhn-Tucker (KKT) conditions, which we cover later in Chapter 11.

Example 5.3. The first example below is a classic quadratic program from economics. The second is a QP by Property 5.1.B, and the third is a QP by inspection.

a. The *portfolio selection problem* is a QP to construct a financial portfolio by trading off performance and risk:

$$\begin{array}{ll} \underset{x}{\text{maximize}} & \mu^\mathsf{T}x - \frac{\gamma}{2}x^\mathsf{T}\Sigma x \\ \text{subject to} & 1^\mathsf{T}x = 1 \\ & x \geq 0, \end{array}$$

where μ is the vector of expected returns on the assets, Σ is the covariance matrix of returns on the assets, and $\gamma \geq 0$ is tuning parameter called the risk tolerance factor in this context. The solution $w^\star$ is the vector of optimal portfolio holdings.

b. The lasso problem (4.4) is a QP.

c. The SVM problem (4.5) is a QP.

5.3. Semidefinite programs

Now we move on to consider the class of *semidefinite programs* (SDPs). An SDP is of the form

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^\mathsf{T}x \\ \text{subject to} & x_1A_1 + \cdots + x_dA_d \preceq B \\ & Gx = h, \end{array} \tag{5.9}$$

for $c \in \mathbb{R}^d$, symmetric matrices $A_1, \ldots, A_d, B$ of equal dimensions, and a matrix-vector pair G, h of compatible dimensions. Though less obvious than the case of LPs and QPs, an SDP (5.9) is always a convex problem; this follows from the fact that the constraints in (5.9), which are known as *linear matrix inequalities*, always form a convex set (Exercise 5.3). Notably, we do not require the matrices $A_1, \ldots, A_d, B$ to be positive semidefinite (they are only assumed to be symmetric). One might then

ask: where does the name semidefinite program come from? It is a reflection of the fact that in an SDP, the usual ordering $\leq$ on vectors (recall that this is interpreted componentwise) is replaced by the ordering $\preceq$ induced by the positive semidefinite cone (recall that we write $X \preceq Y$ to mean $X - Y \succeq 0$ for matrices X, Y).

Below we list some basic properties of semidefinite programs.

A. LPs are SDPs. Every LP (5.1) is an SDP (5.9). To see this, first note that for vectors x, y it holds that $x \leq y \iff \operatorname{diag}(x) \preceq \operatorname{diag}(y)$, where $\operatorname{diag}(x)$ denotes the diagonal matrix with diagonal elements equal to the elements of x and likewise for $\operatorname{diag}(y)$. Then, note that $Ax \leq b$ is equivalent to the linear matrix inequality

$$x_1 \operatorname{diag}(A_1) + x_2 \operatorname{diag}(A_2) + \cdots x_d \operatorname{diag}(A_d) \preceq \operatorname{diag}(b),$$

where $A_1, \ldots, A_d$ denote the columns of A, proving that (5.1) is of the form (5.9).

B. QPs are SDPs. Every QP (5.7) is an SDP (5.9), though this is less obvious than the result for LPs. To see this, note that for any matrix $Q \succeq 0$, vector x, and $t \geq 0$ we have

$$x^T Q x \leq t \iff \begin{bmatrix} tI & Q^{1/2}x \\ x^T Q^{1/2} & t \end{bmatrix} \succeq 0, \tag{5.10}$$

using properties of Schur complements. (Here $Q^{1/2}$ denotes the symmetric square root of Q.) As the right-hand side above can be shown to be equivalent to a linear matrix inequality, a QP can therefore be rewritten in SDP form (Exercise 5.6).

C. Standard form. As with LPs and QPs, an SDP has an equivalent standard form. However, unlike LPs and QPs, this is somewhat of a big jump from its original form (5.9): as we will see, it brings us from a vector-valued variable in the optimization problem to matrix-valued one. To recall some notation first: for symmetric matrices X, Y, we write their inner product as $\langle X, Y \rangle = \operatorname{tr}(XY)$. Now we are ready to present the standard form of an SDP:

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \langle C, X \rangle \\ \text{subject to} \quad & \langle A_i, X \rangle = b_i, \; i = 1, \ldots, m \\ & X \succeq 0. \end{aligned} \tag{5.11}$$

for symmetric matrices $C, A_1, \ldots, A_m$ and scalars $b_1, \ldots, b_m$. Establishing the equivalence between (5.9) and (5.11) is much more nontrivial than the corresponding standard form result for LPs and QPs, and each direction requires its own proof (whereas for LPs and QPs, it was obvious by direct inspection that the standard form programs were themselves LPs and QPs according to the original definitions); see Exercise 5.4.

Just as there is a clear link between an LP and SDP in their original forms (5.1) and (5.9), the latter being a matrix-based generalization of the former, there is a clear link between their standard forms (5.2) and (5.11), the latter again being a matrix-based generalization of the former. This link is formally pursued in Exercise 5.5.

D. Recasting operator norm penalties. In an optimization problem, we can always recast an operator norm penalty using semidefinite constraints: the problem

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & f(X) + \lambda \|X\|_{\text{op}} \\ \text{subject to} \quad & x \in C \end{aligned}$$

(where recall $\|X\|_{\text{op}}$ is the largest singular value of X, as in (3.10)) is equivalent to

$$\begin{aligned} \underset{X,t}{\text{minimize}} \quad & f(X) + \lambda t \\ \text{subject to} \quad & \begin{bmatrix} tI & X \\ X^T & tI \end{bmatrix} \succeq 0 \\ & X \in C. \end{aligned}$$

This can be shown using elementary arguments (properties of Schur complements), and when f is linear and C is the intersection of a linear subspace and the positive semidefinite cone, as in (5.11), the above problem is an SDP (Exercise 5.7 part a). A similar equivalence holds for operator norm constraints (Exercise 5.7 part b).

E. Recasting trace norm penalties. A trace norm penalty also has an equivalent semidefinite form: the problem

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & f(X) + \lambda \|X\|_{\text{tr}} \\ \text{subject to} \quad & x \in C \end{aligned}$$

(where recall $\|X\|_{\text{tr}}$ is the sum of singular values of X, as in (3.9)) is equivalent to

$$\begin{aligned} \underset{X,U,V}{\text{minimize}} \quad & f(X) + \lambda(\text{tr}(U) + \text{tr}(V)) \\ \text{subject to} \quad & \begin{bmatrix} U & \frac{1}{2}X^T \\ \frac{1}{2}X & V \end{bmatrix} \succeq 0 \\ & X \in C. \end{aligned}$$

If f is linear and C is the intersection of a linear subspace and the positive semidefinite cone, as in (5.11), then the above problem is again an SDP. Trace norm constraints yield a similar equivalence. Compared to the operator norm equivalences in Property 5.3.D, these trace norm equivalences are more intricate to prove; we return to them later through the lens of SDP duality in Exercise 12.1.

Example 5.4. The first two examples in what follows are important SDPs in statistics and machine learning. The last is a famous example of an SDP from theoretical computer science.

a. Finding the best rank k approximation of a matrix, which is the optimization problem underlying principal components analysis (PCA), is equivalent (recall Example 4.6) to maximizing a linear function (4.18) over the Fantope (4.19). This is an SDP. Further, we can use an ℓ_1 penalty in order to sparsify the approximation,

$$\underset{P}{\text{maximize}} \quad \langle S, P \rangle - \lambda \|\text{vec}(P)\|_1 \quad \text{subject to} \quad P \in \mathcal{F}_k, \tag{5.12}$$

where $\text{vec}(\cdot)$ is the vectorization operator, which is still an SDP by Property 5.1.B (and linearity of the vectorization operator).

b. Suppose that we observe only some of the entries of a matrix $X \in \mathbb{R}^{n \times d}$, namely, those indexed by a set $\Omega \subseteq \{1, \dots, n\} \times \{1, \dots, d\}$. The following problem seeks a low rank approximation to the observed entries in X:

$$\underset{\Theta}{\text{minimize}} \quad \frac{1}{2}\|P_\Omega(X - \Theta)\|_F^2 + \lambda \|\Theta\|_{\text{tr}}. \tag{5.13}$$

Here P_Ω is the linear map that acts as the identity on entries in the set Ω, and returns zero otherwise; that is, $P_\Omega(Z)$ is matrix of the same dimensions as Z, with entries

$$[P_\Omega(Z)]_{ij} = \begin{cases} Z_{ij} & (i,j) \in \Omega \\ 0 & (i,j) \notin \Omega \end{cases};$$

and $\lambda \geq 0$ is a tuning parameter. The solution in (5.13) is a known as a kind of *matrix completion* estimator; in particular, one defined via trace norm penalization. Combining Properties 5.3.B and 5.3.E, we can see that problem (5.13) is an SDP.

c. Let G be an undirected weighted graph, with nodes labeled $1, \ldots, n$, and where $w_{ij} \geq 0$ denotes the weight on the edge between nodes i, j. For a subset $S \subseteq \{1, \ldots, n\}$, we say that S and $S^c = \{1, \ldots, n\} \setminus S$ form a *cut* in the graph G, whose size is defined as the sum of edge weights among edges that "cross the cut":

$$\text{cut}_G(S) = \sum_{i \in S,\, j \in S^c} w_{ij}.$$

The *max cut* problem (as its name suggests) seeks the cut in G that with maximal size among all possible cuts, which can formulated as the following optimization problem:

$$\begin{array}{ll} \underset{u}{\text{maximize}} & \frac{1}{2} \sum_{i<j} w_{ij}(1 - u_i u_j) \\ \text{subject to} & |u_i| = 1, \; i - 1, \ldots, n. \end{array} \tag{5.14}$$

In (5.14), any feasible point u has entries that are either 1 or -1, indicating membership in the sets S or S^c that determine the cut, and the criterion is precisely $\text{cut}_G(S)$, since each summand contributes $2w_{ij}$ if u_i and u_j disagree in sign, and 0 otherwise. This is clearly not a convex problem (the constraint set $\{-1, 1\}^n$ is nonconvex). Consider:

$$\begin{array}{ll} \underset{U}{\text{maximize}} & \frac{1}{2} \sum_{i<j} w_{ij}(1 - u_i^T u_j) \\ \text{subject to} & \|u_i\|_2 = 1, \; i - 1, \ldots, n. \end{array} \tag{5.15}$$

In (5.14), each u_i is a scalar, but in (5.15), each u_i is a vector, in (say) $\mathbb{R}^d$. Furthermore, the latter is a relaxation of the former, because in (5.15), if we restrict each u_i to be aligned with e_i, the i^{th} coordinate basis vector in $\mathbb{R}^d$, then we recover (5.14).

Importantly, it turns out that (5.15) is equivalent to an SDP. In general, a matrix $X \in \mathbb{S}^n$ is positive semidefinite if and only if we can factorize it as $X = U^T U$ for some $U \in \mathbb{R}^{n \times d}$ and $d \leq n$ (one direction of the equivalence here is obvious, the other can be verified using an eigendecomposition). Hence, interpreting u_i as the i^{th} column of U, for $i = 1, \ldots, n$, we can reparametrize (5.15) as

$$\begin{array}{ll} \underset{X}{\text{maximize}} & \frac{1}{2} \sum_{i<j} w_{ij}(1 - X_{ij}) \\ \text{subject to} & X_{ii} = 1, \; i - 1, \ldots, n \\ & X \succeq 0, \end{array} \tag{5.16}$$

which is an instance of a standard form SDP (5.11). The Goemans-Williamson max cut approximation algorithm solves (5.16) and then applies a randomized rounding scheme to form a set S'. It can be shown that $\text{cut}_G(S') \geq 0.878 \cdot \text{cut}_G(S^\star)$, where $S^\star$ is formed from the solution in (5.14) (that is, $\text{cut}_G(S^\star)$ is the max cut in G).

5.4. Cone programs*

The most general form we consider is the *cone program*, which can be written as

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^T x \\ \text{subject to} & Ax + b \in K \\ & Gx = h, \end{array} \tag{5.17}$$

for $c \in \mathbb{R}^d$, matrix-vector pairs A, b and G, h of compatible dimensions, and a convex cone K. A cone program, as defined in (5.17), is always a convex optimization problem.

We list some basic properties of cone programs.

A. SDPs are cone programs. Every SDP (5.9) is a cone program (5.17). To check this claim, we need to show how to translate the linear matrix inequality $x_1A_1 + \cdots + x_dA_d \preceq B$ into the form $Ax + b \in K$ for some convex cone K. This can be achieved by taking A to be the linear map such that $Ax = -\operatorname{vec}(x_1A_1 + \cdots + x_dA_d)$, $b = \operatorname{vec}(B)$, and $K = \operatorname{vec}(\mathbb{S}^n_+)$, the vectorization of the positive semidefinite cone in dimension n, where n denotes the number of rows (or columns) of the symmetric matrices $A_1, \ldots, A_d, B$.

Combined with the containments established thus far, this establishes the following hierarchy of problem classes in convex optimization (written informally, but intuitively):

$$\text{LPs} \subseteq \text{QPs} \subseteq \text{SDPs} \subseteq \text{cone programs}.$$

Exercise 5.10 gives a refinement of this hierarchy.

B. Standard form. Like LPs, QPs, and SDPs, a cone program has an equivalent standard form:

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^T x \\ \text{subject to} & Ax = b \\ & x \in K. \end{array} \tag{5.18}$$

The proof that any cone program (5.17) can be written in the form (5.18) follows from similar steps to the analogous result for SDPs, established in Exercise 5.4 parts a and b.

Example 5.5. Demonstrating the generality of cone programs, the following two examples show how arguably the most common GLM optimization problems outside of linear regression (4.27) (itself a QP) are cone programs. See Exercise 5.11 for details.

a. Logistic regression (4.28) is a cone program.
b. Poisson regression (4.29) is a cone program.

Chapter Notes

LPs, QPs, SDPs, and cone programs are of great interest in convex optimization, and there are many excellent books that focus on just one of these problem classes alone. These classes not only form a clean hierarchy, but in a sense they also serve as interesting historical landmarks, reflecting a progression in the focus of research in mathematical optimization over the years: systematic study of LPs began in the 1940s, QPs in the 1950s, GPs (see Exercise 5.11) in the 1960s, and SDPs and cone programs only more recently, in the 1990s. We will not attempt to give even a brief account of the history, nor a list of classic references on LPs, SDPs, and so on, as comprehensive historical

review and extensive bibliographies can be readily found elsewhere; for example, the bibliography section in Chapter 4 of [**BV04**] provides pointers to many nice books and review articles.

Basis pursuit was proposed by [**CDS98**], and the Dantzig selector by [**CT07**] (the name of the latter was chosen as a tribute to George Dantzig's contributions to linear programming). It is not as easy to trace back the origins of the linear classification problems (5.5) and (5.6). The latter dates back to at least [**Ros58**], who proposed the *perceptron* algorithm for solving (5.6). The formulation of portfolio selection as a QP is due to [**Mar52**], which is considered the birth of modern portfolio theory (and sparked increased interest in quadratic programming).

Matrix completion via trace norm regularization was first studied in [**CR09, CT10**], though it appears that [**MHT10**] were first to promote the "noisy" version of the problem that we consider in (5.13) to serious consideration. The ℓ_1-penalized Fantope projection problem in (5.12) was proposed by [**VCLR13**]. The Goemans-Williamson max cut approximation algorithm is due to [**GW95**] (and sparked increased interest in semidefinite programming).

Exercises

5.1 In this exercise, we show that every LP and every QP can be written in standard form.

a. Starting with an LP (5.1), argue that we can transform the inequality constraints into equality constraints by introducing slack variables. Argue further that we can decompose $x = x^+ - x^-$, where $x^+, x^- \geq 0$ (often referred to as a decomposition into *positive and negative parts*), thus showing (5.1) is equivalent to a problem with only linear equality constraints and nonnegativity constraints, that is, a problem in standard form (5.2).

b. Apply the same argument as in the last part to a QP (5.7).

5.2 We will prove that ℓ_1 and ℓ_∞ penalties and constraints in optimization problems can be recast as linear ones.

a. Beginning with the equivalence claimed in Property 5.1.B, consider the second problem stated there. Prove that at optimality, the criterion equals $f(x) + \lambda\|x\|_1 + \gamma\|x\|_\infty$. Use this to argue that the two problems in Property 5.1.B are equivalent.

b. Now for the analogous constrained equivalence, prove that

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & \|x\|_1 \leq s \\ & \|x\|_\infty \leq t \\ & x \in C \end{array}$$

and

$$\begin{array}{ll} \underset{x,y,z}{\text{minimize}} & f(x) \\ \text{subject to} & 1^T y \leq s \\ & z \leq t \\ & -y \leq x \leq y \\ & -z1 \leq x \leq z1 \\ & x \in C,\ y, z \geq 0. \end{array}$$

are equivalent problems.

5.3 We study the convexity of sets defined by linear matrix inequalities.

a. Prove that we can subsume equality constraints into linear matrix inequalities; that is, a set of the form

$$\{x : x_1 A_1 + \cdots + x_d A_d \preceq B,\ Gx = h\}$$

can always be rewritten into one of the form

$$\{x : x_1 A_1 + \cdots + x_d A_d \preceq B\}$$

with $A_1, \ldots, A_d, B$ redefined appropriately. This allows us to check that the constraint set in (5.9) is convex by just checking the convexity of the linear matrix inequality set in the last display.

b. Prove that the set in the last display is convex for any symmetric matrices $A_1, \ldots, A_d, B$, by simply checking the definition of convexity.

c. Reprove the result in the part last by using the fact that the positive semidefinite cone is convex, and applying an appropriate a convexity-preserving transformation.

5.4 We will show that (5.9) and (5.11) are equivalent problem forms.

a. We begin by showing that (5.9) can be rewritten in the form (5.11). Using Exercise 5.3 part a, argue that we can ignore the equality constraints $Gx = h$ in (5.9) without a loss of generality; and by decomposing x into positive and negative parts as in Exercise 5.1 part a (and relabeling variables appropriately), argue that we can assume $x \geq 0$ in (5.9) without a loss of generality. Thus to be clear, we have transformed (5.9) into

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^T x \\ \text{subject to} & x_1 A_1 + \cdots + x_d A_d \preceq B \\ & x \geq 0, \end{array}$$

without a loss of generality.

b. By defining $Y = B - \sum_{i=1}^n x_i A_i$ and

$$Z = \begin{bmatrix} \operatorname{diag}(x) & 0 \\ 0 & Y \end{bmatrix},$$

argue that the problem from the last part can be rewritten in terms of a linear criterion in Z, subject to linear equality constraints and $Z \succeq 0$, that is, rewritten in the standard form (5.11).

c. Now we show that (5.11) can be recast in the form (5.9). To do so, we will make use of the vectorization operator $\operatorname{vec}(\cdot)$: this takes a matrix and returns a vector by appending the columns of the matrix one after another. Argue that the criterion in problem (5.11) can be written as $\langle C, X \rangle = \operatorname{vec}(C)^T \operatorname{vec}(X)$. Argue further that the equality constraints $\langle A_i, X \rangle = b_i$, $i = 1, \ldots, m$ can be written as

$$\begin{bmatrix} \operatorname{vec}(A_1)^T \operatorname{vec}(X) & 0 & \cdots & 0 \\ 0 & \operatorname{vec}(A_2)^T \operatorname{vec}(X) & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \operatorname{vec}(A_m)^T \operatorname{vec}(X) \end{bmatrix} = b.$$

d. Show that the last display is equivalent to a linear matrix inequality in $\operatorname{vec}(X)$; and the positive semidefinite constraint $X \succeq 0$ is also equivalent to a linear matrix inequality in $\operatorname{vec}(X)$. Putting this together proves that (5.11) is of SDP form (5.9).

5.5 Show that an LP (5.2) in standard form is a special case of an SDP (5.11) in standard form. Hint: use $x \geq 0 \iff \operatorname{diag}(x) \succeq 0$, and note that we can impose the condition that a matrix X is diagonal via linear equality constraints on X.

5.6 We prove that a QP (5.7) is a special case of an SDP (5.9).

a. Prove the equivalence in (5.10) using properties of Schur complements, as reviewed in Appendix C.

b. Show that the right-hand side in (5.10) can be expressed as a linear matrix inequality in (x, t), and use this to show that (5.7) can be transformed into SDP form.

5.7 We will prove that operator norm penalties and constraints in optimization problems exhibit equivalent forms involving linear matrix inequalities and positive semidefinite constraints.

a. Beginning with the equivalence claimed in Property 5.3.D, consider the second problem stated there. Prove using properties of Schur complements that

$$\begin{bmatrix} tI & X \\ X^T & tI \end{bmatrix} \succeq 0 \iff \|X\|_{\text{op}} \leq t.$$

Hence prove that at optimality, the criterion equals $f(x) + \lambda \|X\|_{\text{op}}$, and argue that the two problems in Property 5.3.D are equivalent.

b. Now for the constrained analogs, by similar arguments, prove that

$$\begin{array}{ll}\underset{X}{\text{minimize}} & f(X) \\ \text{subject to} & \|X\|_{\text{op}} \leq s \\ & x \in C\end{array}$$

and

$$\begin{array}{ll}\underset{X}{\text{minimize}} & f(X) \\ \text{subject to} & \begin{bmatrix} sI & X \\ X^T & sI \end{bmatrix} \succeq 0 \\ & X \in C\end{array}$$

are equivalent problems.

5.8 A *second-order cone program* (SOCP) is of the form

$$\begin{array}{ll}\underset{x}{\text{minimize}} & c^T x \\ \text{subject to} & \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \ i = 1, \ldots, m \\ & Gx = h,\end{array} \tag{5.19}$$

for $c \in \mathbb{R}^d$, matrix-vector pairs A_i, b_i, $i = 1, \ldots, m$ and G, h of compatible dimensions, as well as $c_i \in \mathbb{R}^d$, $d_i \in \mathbb{R}$, $i = 1, \ldots, m$. Prove that an SOCP (5.19) is actually a cone program (5.17). Hint: observe that $\|A_i x + b_i\|_2 \leq c_i^T x + d_i$ is equivalent to $(A_i x + b_i, c_i^T x + d_i)$ lying in what is called a *second-order cone*, which is simply a norm cone (3.19) with $\|\cdot\| = \|\cdot\|_2$.

5.9 A generalization of a QP is a *quadratically-constrained quadratic program* (QCQP), which is of the form

$$\begin{array}{ll}\underset{x}{\text{minimize}} & \frac{1}{2} x^T Q x + c^T x \\ \text{subject to} & \frac{1}{2} x^T Q_i x + c_i^T x + b_i \leq 0, \ i = 1, \ldots, m \\ & Gx = h,\end{array} \tag{5.20}$$

for $Q \in \mathbb{S}_+^d$, $c \in \mathbb{R}^d$, a matrix-vector pair of compatible dimensions G, h, and $Q_i \in \mathbb{S}_+^d$, $c \in \mathbb{R}^d$, $b_i \in \mathbb{R}$, $i = 1, \ldots, m$. Note that when $Q_i = 0$, $i = 1, \ldots, m$, this reduces to a QP. Show that a QCQP (5.20) is a second-order cone program (5.19). Hint: show first that, without a loss of generality, we may consider a problem of the form (5.20) with $Q = 0$. Then observe that (in general):

$$\frac{1}{2} x^T Q x \leq t \iff \left\| \left(\frac{1}{\sqrt{2}} Q^{1/2} x, \ \frac{1}{2}(1 - t) \right) \right\|_2 \leq \frac{1}{2}(1 + t),$$

where $Q^{1/2}$ denotes the symmetric square root of Q.

5.10 Show, using the property in (5.10) for $Q = I$, that an SOCP (5.19) is an SDP (5.9).

(Note that this establishes, combined with the last exercise and all of the containments in this chapter, the following hierarchy for classes of convex problems:

$$\text{LPs} \subseteq \text{QPs} \subseteq \text{QCQPs} \subseteq \text{SOCPs} \subseteq \text{SDPs} \subseteq \text{cone programs},$$

written informally, but intuitively.)

5.11 A *geometric program* (GP) is of the form

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \displaystyle\sum_{k=1}^{K_0} \exp(a_{0k}^\mathsf{T} x + b_{0k}) \\ \text{subject to} & \displaystyle\sum_{k=1}^{K_i} \exp(a_{ik}^\mathsf{T} x + b_{ik}) \leq 1, \; i = 1, \ldots, m \\ & Gx = h, \end{array} \tag{5.21}$$

where $a_{ik} \in \mathbb{R}^d$, $b_{ik} \in \mathbb{R}$, for $k = 1, \ldots, K_i$ and $i = 0, \ldots, m$, as well as a matrix-vector G, h of compatible dimensions. Note that when $K_i = 1$ for each $i = 0, \ldots, m$, this reduces to an LP by taking a log transform of the criterion and both sides of all inequality constraints. But a GP is much more general than an LP. In this exercise we explore connections to cone programs, logistic regression, and Poisson regression.

a. Prove that (5.21) can be rewritten as

$$\begin{array}{ll} \underset{x,y}{\text{minimize}} & c_0^\mathsf{T} y_0 \\ \text{subject to} & c_i^\mathsf{T} y_i \leq 1, i = 1, \ldots, m \\ & \exp(a_{ik}^\mathsf{T} x) \leq y_{ik}, \; k = 1, \ldots, K_i, \; i = 0, \ldots, m \\ & Gx = h, \end{array}$$

for suitably defined $c_i \in \mathbb{R}_+^{K_i}$, $i = 0, \ldots, m$.

b. Using the representation from the last part, prove that a GP is a cone program (5.17). Hint: consider the convex cone in $\mathbb{R}^3$ given by

$$\{(u, v, w) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R} : v \log(v/u) \leq w\}.$$

To see that this is a convex set, note that the map $(u, v) \to v \log(v/u)$ (over $\mathbb{R}_{++}^2$) is the perspective transform of the negative logarithm.

c. Prove that logistic regression (4.28) is a GP (5.21), and thus a cone program by part b, which establishes the claim in Example 5.4.a. Hint: observe that $\log(1 + e^u) \leq v \iff e^{-v} + e^{u-v} \leq 1$.

d. Lastly, using similar arguments to part b, show that Poisson regression (4.29) is a cone program, establishing the claim in Example 5.4.b.

Part 3

Subdifferential Theory

Subgradients

6.1. Definition and properties

As we saw already in Chapters 3 and 4, derivatives play a key role in understanding the properties of convex functions and convex optimization more broadly. But of course, not all convex functions are differentiable. In this chapter, we will develop a generalized notion of lower differentiability.

For a function f on $\mathbb{R}^d$, we say that $s \in \mathbb{R}^d$ is a *subgradient* of f at $x \in \mathrm{dom}(f)$ provided that

$$f(y) \geq f(x) + s^\mathsf{T}(y - x), \quad \text{for all } y \in \mathrm{dom}(f). \tag{6.1}$$

This is analogous to the first-order characterization for convexity (3.15), where s playes the role of $\nabla f(x)$: here, s defines a linear map that passes through f at x, and lies below f everywhere. See Figure 6.1 for an illustration.

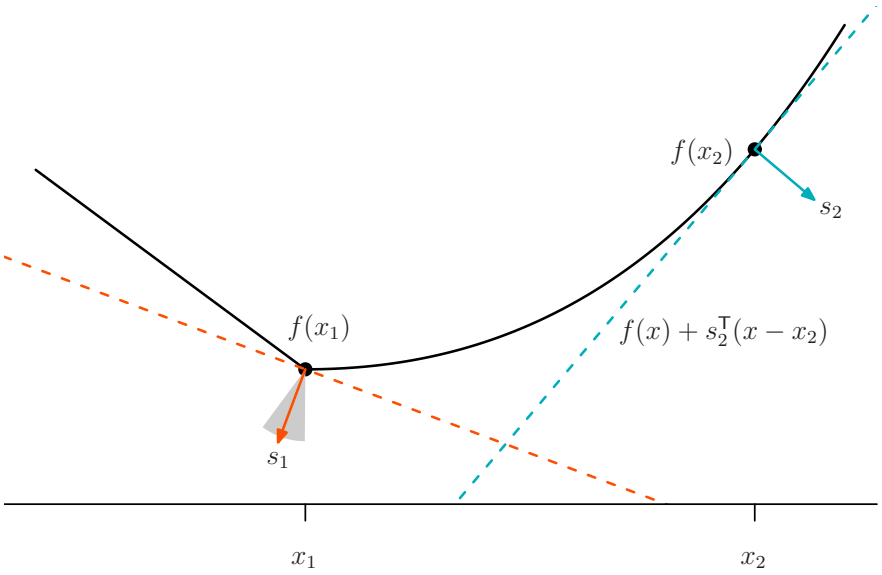

Figure 6.1. Two example subgradients s_1 and s_2 of a function f at two points x_1 and x_2, respectively. Each s_i defines a linear map that passes through x_i and lies below f everywhere (that is, it defines a supporting hyperplane to $\mathrm{epi}(f)$ at x_i, whose normal is $(s_i, -1)$). At x_2, f is differentiable, and $s_2 = \nabla f(x_2)$ is the only subgradient; at x_1 there are many possible subgradients (as illustrated by the gray wedge).

The set of all subgradients of f at x is called its *subdifferential* at x, which we denote by $\partial f(x)$. One can check that $\partial f(x)$ is always a closed convex set (regardless of whether f itself is convex). If $\partial f(x)$ is nonempty, then f is said to be *subdifferentiable* at x. Sometimes we will use a subscript on the subdifferential operator to emphasize the variable under consideration. For example, when f is a function of a block variable (x, y), we use $\partial_x f(x, y)$ for the subdifferential of the function $f(\cdot, y)$ at x, with its second argument fixed at y.

The notions defined above are apparently general, in that we have not assumed convexity of the function f in question. However, as we will see in our discussion of some of the basic properties of subgradients and subdifferentials below, these concepts are in fact intimately tied to convexity.

A. Existence of subgradients. By rearranging the inequality in (6.1), one finds that

$$s \in \partial f(x) \iff \text{epi}(f) \text{ has a supporting hyperplane at } (x, f(x)) \text{ with normal vector } (s, -1).$$

We see that the existence of a subgradient of f at a point is linked to the existence of a supporting hyperplane at a certain point of $\text{epi}(f)$. Given the existence of supporting hyperplanes for convex sets, it is reasonable to expect that subgradients also exist in some generality for convex functions (whose epigraphs are convex). The next result gives the details.

Theorem 6.1. *A convex function f is subdifferentiable at every point in* $\text{relint}(\text{dom}(f))$*, the relative interior of its effective domain. In particular, this means that a convex function that is finite on all of $\mathbb{R}^d$ is subdifferentiable at every point.*

Conversely, if $f : \mathbb{R}^d \to (-\infty, \infty]$ is closed, the set $\text{dom}(f)$ *is convex, and f is subdifferentiable on* $\text{relint}(\text{dom}(f))$*, then f must be convex.*

As alluded to earlier, the first part of Theorem 6.1, on the existence of subgradients for a convex function, is based on applying the supporting hyperplane theorem to the epigraph representation of subgradients; the proof is outlined in Exercise 3.9. The second part, on subdifferentiability implying convexity, is based on a converse supporting hyperplane theorem; see Exercises 6.1 and 6.2.

B. Uniqueness of subgradients. The subgradient of f at a given point $x \in \text{dom}(f)$ need not be unique, that is, the subdifferential $\partial f(x)$ (when nonempty) need not be a singleton. So, when it is unique at x, what can we say about f at x, vis-a-vis differentiability? The answer is nothing, in general (Exercise 6.3). However, for convex f, the subgradient of f is unique at x if and only it is differentiable at x. This result is stated formally next; Exercise 6.6 walks through its proof.

Theorem 6.2. *Let f be a convex function. If f is differentiable at a point $x \in \text{dom}(f)$, then it has a unique subgradient $s = \nabla f(x)$ at x. Conversely, if f has a unique subgradient s at x, then it is differentiable at x and $\nabla f(x) = s$.*

An interesting practical consequence of the last result is that we can infer the differentiability of a convex function at a point (based on it having only one subgradient), when it may be otherwise hard to see this from first principles. This is the case in several of the next examples.

Example 6.3. The following are examples of subgradients of common norms. For the first two examples, the claims can be checked directly using the definition. The others follow from Property 6.2.C and the dual representation of the norm in question.

a. For $f(x) = |x|$, we have

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

b. For $f(x) = \|x\|_1$, subgradients $s \in \partial f(x)$ are points of the form:

$$s_i \in \begin{cases} \{+1\} & x_i > 0 \\ \{-1\} & x_i < 0 \\ [-1, 1] & x_i = 0 \end{cases}, \quad i = 1, \ldots, d.$$

We see that the ℓ_1 norm is differentiable at each point x that does not lie on any of the coordinate axes ($x_i \neq 0$ for all $i = 1, \ldots, d$).

c. For $f(x) = \|x\|_p$, and $1 < p < \infty$, let $1 < q < \infty$ be such that $1/p + 1/q = 1$. Then for $x \neq 0$, we have (Exercise 6.7 part a):

$$s_i = \text{sign}(x_i)|x_i|^{p/q} \cdot \|x\|_p^{-p/q}, \quad i = 1, \ldots, d$$

as the unique subgradient in $\partial f(x)$, and for $x = 0$, we have $\partial f(0) = \{s : \|s\|_q \leq 1\}$. We see that the ℓ_p norm, $1 < p < \infty$, is differentiable at any $x \neq 0$. Note that for $p = 2$ in particular, we have:

$$\partial f(x) = \begin{cases} \{x/\|x\|_2\} & x \neq 0 \\ \{s : \|s\|_2 \leq 1\} & x = 0 \end{cases}.$$

d. For $f(x) = \|x\|_\infty$, subgradients $s \in \partial f(x)$ are points of the form (Exercise 6.7 part b):

$$s_i \in \begin{cases} [0, 1] & x_i = \|x\|_\infty \\ [-1, 0] & -x_i = \|x\|_\infty \\ \{0\} & |x_i| < \|x\|_\infty \end{cases}, \quad i = 1, \ldots, d,$$

where $\|s\|_1 \leq 1$. We see that the ℓ_∞ norm is differentiable at any x that has a unique maximum absolute coordinate.

e. For $f(X) = \|X\|_{\text{op}}$, the operator norm, we have (Exercise 6.8):

$$\partial f(X) = \text{conv}\{uv^T : \|u\|_2 \leq 1, \, \|v\|_2 \leq 1, \, u^T X v = \|X\|_{\text{op}}\}.$$

In other words, subgradients are given by convex combinations of outer products of top left and right singular vectors of X. If the top singular value of X has multiplicity one, then this outer product is unique (there is only one top left and right singular vector, up to sign flips), and the operator norm is differentiable at X with derivative uv^T.

f. For $f(X) = \|X\|_{\text{tr}}$, the trace norm, letting $X = U\Sigma V^T$ denote the SVD of X, we have (Exercise 6.9):

$$\partial f(X) = \{UV^T + W : \|W\|_{\text{op}} \leq 1, \, U^T W = 0, \, WV = 0\}.$$

In other words, subgradients are given by adding to UV^T a matrix W of at most unit operator norm that is orthogonal to the columns of U and the rows of V. If X has full column rank or full row rank ($X^T X$ or XX^T is invertible), then only $W = 0$ satisfies these constraints, and the trace norm is differentiable at X with derivative UV^T.

6.2. Subgradient calculus

We describe rules that will be helpful in the calculation of subgradients.

A. Scaling. For any function f, $x \in \mathrm{dom}(f)$, and $a > 0$, it holds that $(\partial af)(x) = a\partial f(x)$. This is also trivially valid for $a = 0$, provided that $\partial f(x) \neq \emptyset$.

B. Sum. For any functions $f_1, \ldots, f_n$, their sum $F = f_1 + \cdots + f_n$ satisfies, for any $x \in \mathrm{dom}(F)$,

$$\partial F(x) \supseteq \partial f_1(x) + \cdots + \partial f_n(x),$$

where on the right we interpret $U + V = \{u + v : u \in U,\, v \in V\}$, the set sum (Minkowski sum) of U, V. Moreover, if $f_1, \ldots, f_n$ are convex, and $\cap_{i=1}^n \mathrm{relint}(\mathrm{dom}(f_i)) \neq \emptyset$, then for any $x \in \mathrm{dom}(F)$,

$$\partial F(x) = \partial f_1(x) + \cdots + \partial f_n(x).$$

C. Partial supremum. Let f be a function acting on a block variable (x, z). If $f(\cdot, z)$ is convex for each $z \in Z$, then the partial supremum $F = \sup_{z \in Z} f(\cdot, z)$ satisfies, for any $x \in \mathrm{dom}(F)$,

$$\partial F(x) \supseteq \mathrm{cl}\left(\mathrm{conv}\left(\bigcup_{z \in \bar{Z}(x)} \partial_x f(x, z)\right)\right),$$

where $\bar{Z}(x) = \{z \in Z : f(x, z) = F(x)\}$, the set of points at which the supremum at x is achieved. In other words, for any z such that $f(x, z)$ achieves the supremum, the subgradients of the function $f(\cdot, z)$ at x are subgradients of F at x, along with limits of convex combinations of them. Moreover, if $\mathrm{relint}(\mathrm{dom}(F)) \neq \emptyset$, Z is compact, f is continuous on $\mathrm{relint}(\mathrm{dom}(F)) \times Z$, and $f(\cdot, z)$ is *closed* and convex for each $z \in Z$, then

$$\partial F(x) = \mathrm{conv}\left(\bigcup_{z \in \bar{Z}(x)} \partial_x f(x, z)\right), \tag{6.2}$$

a result known as the *Danskin-Bertsekas theorem*. An important special case corresponds to a finite set Z: if f_i, $i = 1, \ldots, k$ are closed and convex, then their pointwise maximum $F = \max_{i=1,\ldots,k} f_i$ satisfies $\partial F(x) = \mathrm{conv}(\cup_{i: f_i(x) = F(x)} \partial f_i(x))$, for any $x \in \mathrm{dom}(F)$.

D. Partial infimum. Let f be a function acting on a block variable (x, z). If f is convex and Z is a convex set, then the partial infimum $F = \inf_{z \in Z} f(\cdot, z)$ satisfies, for any $x \in \mathrm{dom}(F)$ such that $F(x) > -\infty$, where we now denote $\bar{Z}(x) = \{z \in Z : f(x, z) = F(z)\}$,

$$\partial F(x) \supseteq \mathrm{cl}\left(\mathrm{conv}\left(\bigcup_{z \in \bar{Z}(x)} \{s : (s, 0) \in \partial f(x, z)\}\right)\right),$$

E. Linear composition. Let $f : \mathbb{R}^k \to (-\infty, \infty]$ be convex, and let $A \in \mathbb{R}^{k \times d}$, $b \in \mathbb{R}^d$. Then the function F defined as $F(x) = f(Ax + b)$ satisfies, for any $x \in \mathrm{dom}(F)$,

$$\partial F(x) \supseteq A^{\mathsf{T}} \partial f(Ax + b).$$

Moreover, if $(\mathrm{col}(A) + b) \cap \mathrm{relint}(\mathrm{dom}(f)) \neq \emptyset$, then for any $x \in \mathrm{dom}(F)$,

$$\partial F(x) = A^{\mathsf{T}} \partial f(Ax + b).$$

F. General composition. Let $f : \mathbb{R}^k \to (-\infty, \infty]$, $g : \mathbb{R}^d \to \mathbb{R}^k$, and denote their composition by $F = f \circ g$, that is, $F(x) = f(g(x))$. If f is convex and nondecreasing in each argument, and each component function g_i, $i = 1, \ldots, k$ is convex, then

$$\partial F(x) \supseteq \{r_1 s_1 + \cdots + r_k s_k : r = (r_1, \ldots, r_k) \in \partial f(g(x)),\ s_i \in \partial g_i(x),\ i = 1, \ldots, k\}.$$

Note that the conditions here are no less general than those in Property 3.4.K, on the convexity of a composition; see Exercise 3.5 part b.

Example 6.4. The following examples can be checked using subgradient calculus rules.

a. For $f(x) = h_C(X)$, the support function (3.12) of a compact set C, we have

$$\partial h_C(x) = \left\{ y : y^T x = \sup_{z \in C} z^T x \right\},$$

A norm $\|\cdot\|$ is in fact a support function, corresponding to the set $C = \{x : \|x\|_* \leq 1\}$. Here $\|\cdot\|_*$ is itself another norm that we call the dual norm of $\|\cdot\|$, and thus C can be called the dual norm unit ball. The connection between $\|\cdot\|$ and $\|\cdot\|_*$ is detailed later, in Chapter 10.3, when we cover duality; for now we simply observe that we can write

$$\|x\| = \sup_{\|z\|_* \leq 1} z^T x, \tag{6.3}$$

and therefore by the Danskin-Bertsekas theorem (6.2),

$$\partial \|x\| = \{y : \|y\|_* \leq 1,\ y^T x = \|x\|\}. \tag{6.4}$$

This can be used as a starting point to establish the results on subgradients of specific norms given in Example 6.3 (Exercises 6.7–6.9).

b. For $f(x) = \|Ax\|$, where $A \in \mathbb{R}^{k \times d}$ and $\|\cdot\|$ is a norm, we have

$$\partial f(x) = \{A^T y : \|y\|_* \leq 1,\ y^T A x = \|Ax\|\}.$$

Here $\|\cdot\|_*$ the dual norm of $\|\cdot\|$, as in the last example.

c. For $f(x) = \inf_{z \in C} \|x - z\|_2$, where C is closed and convex, suppose $x \notin C$. Letting P_C denote the (Euclidean) projection operator onto C, which satisfies $\|x - P_C(x)\|_2 = f(x)$, we have

$$\partial f(x) = \left\{ \frac{x - P_C(x)}{\|x - P_C(x)\|_2} \right\}. \tag{6.5}$$

6.3. Subgradient optimality condition

Subgradients have an important connection to the minimization of a function. The following can be checked directly from the definition of a subgradient in (6.1): $f(x) \leq f(y)$ for all y if and only if

$$0 \in \partial f(x). \tag{6.6}$$

This is called the *subgradient optimality condition* for f. Note that it generalizes the more familiar zero-gradient condition (4.8) for differentiable convex f, since for differentiable convex f, the only subgradient at x is the gradient $\nabla f(x)$ (Theorem 6.2). We will soon find the subgradient optimality condition (6.6) very useful, when we discuss proximal operators in the next chapter.

This can be extended to an optimality condition for the constrained problem,

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad x \in C.$$

for a convex function f and convex set C, where $\mathrm{relint}(\mathrm{dom}(f)) \cap \mathrm{relint}(C) \neq \emptyset$. We rewrite this problem in characteristic form

$$\underset{x}{\text{minimize}} \quad f(x) + I_C(x),$$

and observe that $x \in C$ is a solution if and only if zero is a subgradient of the criterion at x, which (using convexity of f and I_C, and the subgradient rule for a sum in Property 6.2.B), can be written as $0 \in \partial f(x) + \partial I_C(x)$, that is,

$$-s \in I_C(x), \quad \text{for some } s \in \partial f(x).$$

It remains to compute the subdifferential of I_C, the characteristic function of the convex set C. A straightforward calculation reveals

$$\partial I_C(x) = \mathcal{N}_C(x) = \{s : s^T x \geq s^T y, \text{ for all } y \in C\}.$$

the normal cone to C at x, and the second-to-last display, the (necessary and sufficient) subgradient optimality condition for a constrained convex problem, becomes

$$s^T(y - x) \geq 0, \quad \text{for some } s \in \partial f(x) \text{ and all } y \in C, \tag{6.7}$$

which nicely generalizes the first-order optimality condition (4.7) covered in Chapter 4.2.

6.4. Subgradient monotonicity*

The subdifferential of a function f is a *monotone operator*, which means that it satisfies

$$(s_x - s_y)^T(x - y) \geq 0, \quad \text{for all } x, y \in \mathrm{dom}(f), \text{ and } s_x \in \partial f(x),\ s_y \in \partial f(y). \tag{6.8}$$

For given x, y, the inequality can be seen by simply adding together the two conditions defining the subgradients, $f(y) \geq f(x) + s_x^T(y - x)$ and $f(x) \geq f(y) + s_y^T(x - y)$, and rearranging. Note that this generalizes the monotone gradient condition (3.21) satisfied by a differentiable convex function.

It is natural to ask whether there is a converse to the subgradient monotonicity condition. For gradients, recall, there is a converse result: if f is differentiable, and $(\nabla f(x) - \nabla(y))^T(x - y) \geq 0$ for all x, y, then f is convex (Exercise 3.10). However, for subgradients we will need to be careful at the outset, in even just posing the question precisely. For example, if $\partial f(x) = \emptyset$ for all x, as would be the case for (say) differentiable and strictly concave f, then (6.8) is vacuously true, but clearly f is not convex. On the other hand, if we assumed that subgradients exist on $\mathrm{relint}(\mathrm{dom}(f))$, then f would already be convex (provided f is closed; recall Theorem 6.1).

A way to pose the converse question precisely is as follows. Given a set-valued operator T, with $T(x) \subseteq \mathbb{R}^d$ for each $x \in \mathbb{R}^d$ (and possibly $T(x) = \emptyset$ for some x), if

$$(s_x - s_y)^T(x - y) \geq 0, \quad \text{for all } x, y, \text{ and } s_x \in T(x),\ s_y \in T(y), \tag{6.9}$$

then does there exist a convex function f for which $T \subseteq \partial f$? This is a kind of *embedding* problem: we are asking whether the graph of T embeds into the graph of the subdifferential ∂f of some convex function f; or equivalently, given a set of pairs $\{(x_i, s_i) : i \in I\}$ (the graph of T, where I is possibly infinite), we are seeking a convex function f that satisfies the relations

$$f(y) \geq f(x_i) + s_i^T(y - x_i), \quad i \in I.$$

Interestingly, monotonicity of T, as in (6.9), is not enough, but a suitable generalization of this condition is. To motivate this, let (x_i, s_i), $i = 1, \ldots, n$ denote any n points in the graph of ∂f; that is, satisfying $s_i \in \partial f(x_i)$, $i = 1, \ldots, n$. Writing $x_{n+1} = x_1$ for convenience, adding together the n subgradient inequalities

$$f(x_{i+1}) \geq f(x_i) \geq s_i^T(x_{i+1} - x_i), \quad i = 1, \ldots, n,$$

and rearranging, gives

$$s_1^T(x_2 - x_1) + s_2^T(x_3 - x_2) + \cdots + s_{n-1}^T(x_n - x_{n-1}) + s_n^T(x_1 - x_n) \leq 0. \tag{6.10}$$

An operator T that satisfies (6.10) for all $n \geq 1$ and all (x_i, s_i), $i = 1, \ldots, n$ in its graph is said to be *cyclically monotone*. The argument just given shows that ∂f is itself cyclically monotone.

The next result, which we call *Rockafellar's embedding theorem*, gives a complete characterization of subgradients and monotonicity.

Theorem 6.5. *Let T be a set-valued operator, where $T(x) \subseteq \mathbb{R}^d$ for $x \in \mathbb{R}^d$. There exists a convex function f with $T \subseteq \partial f$ (which means that $T(x) \subseteq \partial f(x)$ for all $x \in \mathbb{R}^d$) if and only if T is cyclically monotone.*

Moreover, there exists a closed convex function with $T = \partial f$ if and only if T is maximal cyclically monotone (where maximal means that there is no other cyclically monotone map T' with $T \subseteq T'$). Finally, when it exists, the solution f (to the equation $T = \partial f$) is unique up to an arbitrary additive constant.

Note that the "only if" direction for the containment result $T \subseteq \partial f$ in Theorem 6.5 was already established in the motivating discussion before the theorem statement. For other parts of the proof of Theorem 6.5, see Exercise 6.14.

6.5. Subgradients and growth*

An important use of the monotonicity characterization (3.21) for the convexity of a differentiable function was that it played a key role in the proofs of Theorems 3.8 and 3.9. These were theorems about equivalent growth conditions, for smooth functions. With subgradients in place of gradients, the next result essentially generalizes Theorem 3.9.

Theorem 6.6. *For closed convex $f : \mathbb{R}^d \to (-\infty, \infty]$ and $m > 0$, consider the statements:*

(i) $\|s_x - s_y\|_2 \geq m\|x - y\|_2$, for all $x, y \in \mathrm{dom}(f)$ and $s_x \in \partial f(x)$, $s_y \in \partial f(y)$;
(ii) the function f_m is convex, where $f_m(x) = f(x) - \frac{m}{2}\|x\|_2^2$;
(iii) $f(y) \geq f(x) + s_x^T(y - x) + \frac{m}{2}\|y - x\|_2^2$, for all $x, y \in \mathrm{dom}(f)$ and $s_x \in \partial f(x)$;
(iv) $(s_x - s_y)^T(x - y) \geq m\|x - y\|_2^2$, for all $x, y \in \mathrm{dom}(f)$ and $s_x \in \partial f(x)$, $s_y \in \partial f(y)$.

Then the following relations hold:

$$\textit{(i)} \impliedby \textit{(ii)} \iff \textit{(iii)} \iff \textit{(iv)}.$$

The proof of Theorem 6.6 is outlined in Exercise 6.16, divided in two main parts. The first part shows that (ii) $\iff$ (iii) $\implies$ (iv) $\implies$ (i), using arguments entirely analogous to those used for Theorem 3.9 (Exercise 3.11). The second part proves (iv) $\implies$ (ii), using Rockafellar's embedding theorem, and a projection argument that reduces consideration to set-valued operators on $\mathbb{R}$ (where monotone and cyclically monotone operators coincide). An alternative proof for (iv) $\implies$ (ii) can be given using generalized subgradients; see the chapter notes for more details.

It is worth noting that an important consequence of Theorem 6.6 is that it establishes strongly convex functions are coercive. See Exercise 6.17.

Next we give one more simple but important growth result, on the boundedness of subgradients over compact sets. We remark that part (ii) of the next theorem in fact establishes that a convex function is locally Lipschitz, as claimed in part (iii) of Theorem 3.10.

Theorem 6.7. *Let f be a convex function, and let $C \subseteq \mathrm{int}(\mathrm{dom}(f))$ be compact. Then:*

(i) $\cup_{x \in C} \partial f(x)$ is nonempty and bounded;

(ii) f is Lipschitz continuous on C, meaning (recall)

$$|f(x) - f(y)| \leq L\|x - y\|_2, \quad \text{for all } x, y \in C,$$

with Lipschitz constant $L = \sup_{s \in \cup_{x \in C} \partial f(x)} \|s\|_2$.

Further, as a converse to (i), the condition $x \in \mathrm{int}(\mathrm{dom}(f))$ is also necessary for $\partial f(x)$ to be nonempty and bounded.

6.6. Subgradients and geometry*

Subgradients possess a deep connection to convex geometry. Recall based on the discussion before and after Theorem 6.1 that their existence is tied to the fact that a subgradient of a function defines a normal vector to its epigraph. A related interpretation is as follows: if $s \in \partial f(x)$, then we observe straight from the definition (6.1) that

$$f(y) \leq f(x) \implies s^\mathsf{T} x \geq s^\mathsf{T} y.$$

In other words, s defines the normal to a supporting hyperplane of a sublevel set $\{y : f(y) \leq f(x)\}$ at x. As the normal cone to $\{y : f(y) \leq f(x)\}$ at x is, by definition, the collection of all such normal vectors, we can also write this conclusion as

$$\partial f(x) \subseteq \mathcal{N}_{\{y : f(y) \leq f(x)\}}(x).$$

Interestingly, as the next result shows, the subgradients at x not only lie in the normal cone to the sublevel set at x, they also generate it.

Theorem 6.8. *Let f be a convex function, and $x \in \mathrm{relint}(\mathrm{dom}(f))$ be a point at which f does not attain its infimum. Then*

$$\mathcal{N}_{\{y : f(y) \leq f(x)\}}(x) = \mathrm{cone}\big(\partial f(x)\big).$$

Just like the subgradient optimality condition (6.6) generalizes the zero-gradient condition (4.8) from classical smooth analysis, we can view Theorem 6.8 as a generalization of the classical result that the gradient vector ∇f lies orthogonal to tangent planes of level sets a smooth function f. See Figure 6.2 for an illustration.

Lastly, we discuss a categorization of points of subdifferentiability of f, and we show how the geometric perspective provided in Theorem 6.8 leads to an interesting conclusion about the number of points lying in the "least smooth" category, along any slice through the graph of a convex function f. The categorization is as follows: when $k = \dim(\mathrm{span}(\partial f(x)))$, the dimension of the linear span $\partial f(x)$, we say that x is *category* k point in the subdifferential $\partial f(x)$. Thus, for convex f, at a point x where f does not attain its infimum:

- x is category 0 $\iff$ f not subdifferentiable at x;
- x is category 1 $\iff$ f is differentiable at x;
- x is category 2 $\iff$ f has two linearly independent subgradients at x, say s_1 and s_2, where $\mathrm{span}\{s_1, s_2\} = \partial f(x)$;
- and so on, up through category d, where $\mathrm{dom}(f) \subseteq \mathbb{R}^d$.

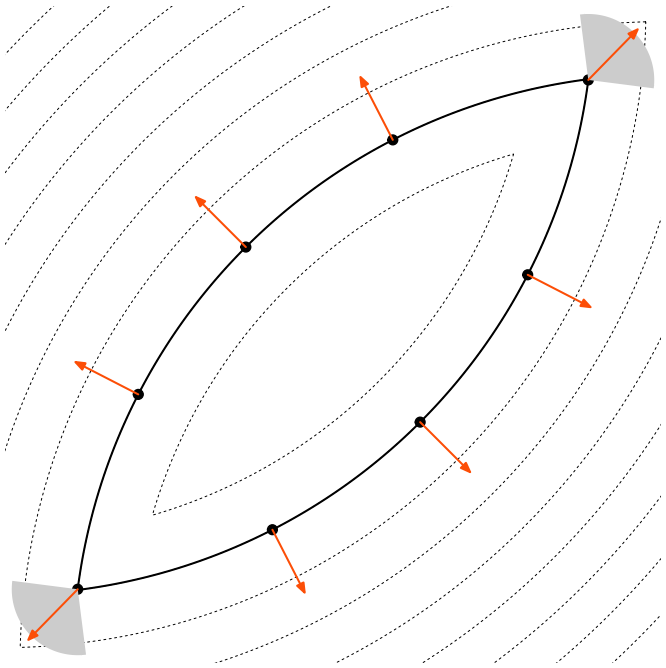

Figure 6.2. Contours of a convex function f, with a particular level set of interest highlighted in solid black. At various points x along the level set, subgradients of f at x are drawn, represented as normal vectors emanating from the level set at x. At two points along the level set, in the bottom left and top right, f is not differentiable, and the subdifferentials are represented by gray wedges.

Points in category 1 can be thought of as the "most smooth", and points in category d as the "least smooth" (among categories 1 through d, excluding 0): for a point x in category d the normal cone to the sublevel set $\{y : f(y) \leq f(x)\}$ at x is d-dimensional, which means the sublevel set looks "pointy" at x, like a vertex on a polyhedron, or the bottom left and top right points along the level set drawn in Figure 6.2.

We already know from Theorem 3.10 part (iii) that categories 2 through d are, altogether, "rare" compared to category 1: the union of category 2 through d points only make up a set of Lebesgue measure zero in $\text{int}(\text{dom}(f))$. But in fact, we can say something much finer along any slice through the graph of a convex function f: on any such slice, it turns out that f can only have a *countable* number of points in category d, the "least smooth" category.

Theorem 6.9. *Let f be convex with $\text{dom}(f) \subseteq \mathbb{R}^d$, and let $t \in \mathbb{R}$ be arbitrary. Then f has a countable number of category d subdifferentiable points along the level set $\{x : f(x) = t\}$.*

The proof of Theorem 6.9 is elementary, and it rests critically on the normal cone formulation of subgradients in Theorem 6.8. Essentially, it uses the normal cone relationship (and convexity of the sublevel set) to argue that the interiors of conic hulls of category d subdifferentials along the level set form a collection of disjoint open sets in $\mathbb{R}^d$, and hence by classical arguments in analysis, there can be at most a countable number of them. See Exercise 6.18.

Chapter Notes

Jean Jacques Moreau and R. Tyrrell Rockafellar are widely considered to be the "founding fathers" of subdifferentiability, each having worked separately to develop the subject in the early 1960s. To

learn more, beyond what is covered in this chapter, a classic reference (and masterful treatment) is [**Roc70**] (Chapters 23–25); other nice references are [**HUL01**] (Chapter D) and [**Ber09**] (Chapter 5.4). The study of monotone operators is closely linked to the study of subgradients (and to that of proximal operators, which will be covered in the next chapter). Two nice very books that develop this connection, from different perspectives (analytic versus algorithmic) are [**BC11**, **RY22**].

The Danskin-Bertsekas theorem is named after the extension given in Bertsekas' Ph.D. thesis [**Ber71**] of an earlier result by Danskin [**Dan67**]. Even more can be said about the subdifferential of a partial supremum of convex functions; see, for example, [**HLZ08**] for a recent overview and connections to other subgradient calculus rules. Rockafellar's embedding theorem (which is not, as far as we know, a commonly used name for the result stated in this chapter) is due to [**Roc66**].

Finally, we note that generalizations of the definition of a subgradient have been developed for nonconvex functions, notably by Francis H. Clarke and Rockafellar, in the 1980s. These definitions reduce to the usual notion of a subgradient for convex functions, and also reduce the usual notion of a gradient for differentiable (and possibly nonconvex) functions, but apply much more broadly. Two authoritative references, each of which covers much more than just generalized subgradients, are [**Cla90**, **RW09**]. Generalized subgradients are not only critical tools in nonsmooth and variational analysis, they can also be useful in *convex* analysis. For example, the generalized subgradient used in variational analysis, as defined in [**RW09**], provides an alternative, short proof that (iv) $\implies$ (ii) in Theorem 6.6; see Exercise 12.59 in [**RW09**].

Exercises

6.1 This exercise proves the following converse to the supporting hyperplane theorem: if C is a closed set, $\text{relint}(C) \neq \emptyset$, and every $x_0 \in \text{relbd}(C)$ has a supporting hyperplane (there exists $a \neq 0$ and b such that $a^T x_0 = b$ and $a^T x \leq b$ for all $x \in C$), then C must be convex.

We will assume, without a loss of generality, that C is full-dimensional ($\text{aff}(C) = \mathbb{R}^d$), so its relative interior and relative boundary are its interior and boundary, respectively. (Note: having proved this full-dimensional result, to accommodate the case when $\text{aff}(C)$ is a proper subspace of $\mathbb{R}^d$, we can reparametrize to the affine subspace, and apply the result just proved to conclude that the set D—the reparametrization of C to the affine subspace—is convex in this coordinate system, and then simply view C as an affine image of D in order to conclude that C itself is convex.)

Let H be the intersection of all supporting halfspaces to C,

$$H = \bigcap_{x_0 \in \text{bd}(C)} \underbrace{\{x : a_{x_0}^T x \leq b_{x_0}\}}_{H_{x_0}},$$

where for each $x_0 \in \text{bd}(C)$, we write a_{x_0} and b_{x_0} for the normal vector and offset, respectively, that define the supporting hyperplane to C at x_0. Note that H is convex, and that $C \subseteq H$.

a. Fix $y \notin C$, and let $x \in \text{int}(C)$. Let

$$t = \sup\{s \geq 0 : x + s(y - x) \in C\}.$$

Prove that $s \in (0, 1)$. (Hint: use that $x \in \text{int}(C)$, and the fact that $C^c = \mathbb{R}^d \setminus C$ is open.) Prove also that $ty \in \text{bd}(C)$. (Hint: use again the fact that C is closed.)

b. By inspecting the supporting hyperplane at $x_0 = ty$, argue that since $x \in H_{x_0}$, we must have $y \notin H_{x_0}$, and thus $y \notin H$.

Since y was arbitrary, note that we have shown $C^c \subseteq H^c$, that is, $H \subseteq C$, which together with the observation $C \subseteq H$, proves that $C = H$ and hence that C is convex.

6.2 We will now prove the second part of Theorem 6.1, using the converse supporting hyperplane theorem from Exercise 6.1. Let f satisfy the conditions of the theorem: $f : \mathbb{R}^d \to (-\infty, \infty]$ is closed and has a subgradient and every point in $\text{relint}(\text{dom}(f))$, where $\text{dom}(f)$ is convex.

a. Abbreviate $S = \text{relint}(\text{dom}(f))$. Prove that any point (x, t) on the relative boundary of $\text{epi}(f)$ can be written in one of two "types":

$$\begin{aligned} &\text{type I}: \quad x \in S,\ t = f(x), \\ &\text{type II}: \quad x \in \text{dom}(f) \setminus S,\ t \geq f(x). \end{aligned}$$

b. Prove that every relative boundary point of type I has a supporting hyperplane. Hint: use the existence of subgradients on S.

c. Prove that every relative boundary point of type II has a supporting hyperplane. Hint: note that $x \in \text{bd}(\text{dom}(f))$, and use the fact that $\text{dom}(f)$ is itself convex and thus must admit a supporting hyperplane, by the supporting hyperplane theorem.

d. Complete the proof by applying the converse supporting hyperplane theorem to $\text{epi}(f)$.

6.3 Give two distinct examples of a function f that is nondifferentiable at a point x and yet still has a unique subgradient at x, where in one example f is discontinuous at x, and in another continuous. Hint: by Theorem 6.2, the functions f here will have to be nonconvex.

6.4 Prove for a convex function f, by using the existence of subgradients on $\mathrm{relint}(\mathrm{dom}(f))$, that f is equal to the pointwise supremum of its affine minorants:

$$f(x) = \sup\{g(x) : g \text{ is affine, and } g \leq f\}, \quad \text{for all } x \in \mathrm{relint}(\mathrm{dom}(f)), \tag{6.11}$$

where we write $g \leq f$ to mean that $g(x) \leq f(x)$ for all x.

6.5 For a function f on $\mathbb{R}^d$, we define its (one-sided) *directional derivative* with respect to $v \in \mathbb{R}^d$ at point $x \in \mathrm{dom}(f)$ by the (one-sided) limit

$$f'(x; v) = \lim_{t \to 0^+} \frac{f(x + tv) - f(x)}{t}, \tag{6.12}$$

if it exists (with $\pm\infty$ limits being allowed). Note the difference between the usual (two-sided) directional derivative from multivariate calculus, as reviewed in Appendix B.2. In this exercise, we will explore the relationship between directional derivatives and subgradients. We take f to be convex in what follows.

a. Prove that (6.12) always exists (again, with $\pm\infty$ limits being allowed). Hint: prove that $(f(x + tv) - f(x))/t$ is nonincreasing as $t \to 0^+$, using convexity of f.

b. Prove that $v \mapsto f'(x; v)$ is convex and positively homogeneous, where the latter means $f'(x; \alpha v) = \alpha f'(x; v)$ for all $\alpha \geq 0$.

c. For $x \in \mathrm{relint}(\mathrm{dom}(f))$, prove that

$$s \in \partial f(x) \iff f'(x; v) \geq s^T v, \text{ for all } v.$$

d. For $x \in \mathrm{relint}(\mathrm{dom}(f))$, prove that

$$f'(x; v) = \sup_{s \in \partial f(x)} s^T v.$$

Hint: use the last part to establish $\geq$ in the above display. For the other direction, fix x, denote $F_x(v) = f(x; v)$, and consider Exercise 6.4 applied to F_x. Show that, by positive homogeneity of F_x, we in fact have

$$F_x(v) = \sup\{G(v) : G \text{ is linear, and } G \leq F_x\},$$

and lastly, show that G linear with $G \leq F_x$ implies $G(v) = s^T v$ for $s \in \partial f(x)$, to upper bound the supremum on the right-hand side above, completing the proof.

6.6 We will work through the proof of Theorem 6.2.

a. If f is differentiable at x, then show $s = \nabla f(x)$ is its only subgradient at x by using the relation in Exercise 6.5 part c between subgradients and directional derivatives. Hint: if f is differentiable at x, then note that $f'(x; v) = \nabla f(x)^T v$.

b. If s is the unique subgradient at x, then denote $F_x(v) = f(x + v) - f(x) - s^T v$, and argue that F_x has 0 as its unique subgradient at the origin. Show that this implies

$$\lim_{v \to 0} \frac{F_x(v)}{\|v\|_2} = 0,$$

which means (by definition) that f is differentiable at x with $\nabla f(x) = s$. Hint: to prove the above display, first note that for any v with unit norm, by Exercise 6.5 part d (which we know applies, because s cannot be unique if $x \notin \mathrm{int}(\mathrm{dom}(f))$, by Theorem 6.7),

$$0 = F_x'(0; v) = \lim_{t \to 0^+} \frac{F_x(tv)}{t}.$$

Then use the above pointwise convergence, along with convexity of $v \mapsto F_x(tv)/t$, to prove that in fact $F_x(tv)/t \to 0$ as $t \to 0^+$ *uniformly* over the unit ball $\{v : \|v\|_2 \leq 1\}$, which leads to the desired conclusion.

6.7 We establish results on the subgradients of the ℓ_p norm, $1 < p < \infty$, and the ℓ_∞ norm.

a. Prove the statement in Example 6.1.c by using the dual representation

$$\|x\|_p = \sup_{\|z\|_q \le 1} z^\mathsf{T} x,$$

where $1 < q < \infty$ is such that $1/p + 1/q = 1$, and applying (6.4) (which follows from an application of the Danskin-Bertsekas theorem). Hint: in Hölder's inequality,

$$a^\mathsf{T} b \le \|a\|_p \|b\|_q,$$

equality holds for nonzero $a, b \in \mathbb{R}^d$ if and only if $a_i^p / \|a\|_p^p = b_i^q / \|b\|_q^q$, $i = 1, \ldots, d$.

b. Prove the statement in Example 6.1.d similarly, via the dual representation,

$$\|x\|_\infty = \sup_{\|z\|_1 \le 1} z^\mathsf{T} x,$$

then applying (6.4). An alternative is to observe that the Danskin-Bertsekas theorem (in the case where Z is finite) can be applied directly to $\|x\|_\infty = \max_{i=1,\ldots,d} |x_i|$.

6.8 We establish the result in Example 6.1.e, on the subgradients of the operator norm. We use the dual representation

$$\|X\|_{\mathrm{op}} = \sup_{\|Z\|_{\mathrm{tr}} \le 1} \langle Z, X \rangle,$$

covered later in Chapter 10.3 (where recall we write $\langle Z, X \rangle = \mathrm{tr}(Z^\mathsf{T} X)$), along with (6.4).

a. Prove that if $Z = \sum_{i=1}^k u_i v_i^\mathsf{T}$, where $\|u_i\|_2 \le 1$, $\|v_i\|_2 \le 1$, $u_i^\mathsf{T} X v_i = \|X\|_{\mathrm{op}}$, $i = 1, \ldots, k$, then $\|Z\|_{\mathrm{tr}} \le 1$ and $\langle Z, X \rangle = \|X\|_{\mathrm{op}}$.

b. Prove the opposite direction: if $\|Z\|_{\mathrm{tr}} \le 1$ and $\langle Z, X \rangle = \|X\|_{\mathrm{op}}$ then Z must be of the above form ($Z = \sum_{i=1}^k u_i v_i^\mathsf{T}$, where $\|u_i\|_2 \le 1$, $\|v_i\|_2 \le 1$, $u_i^\mathsf{T} X v_i = \|X\|_{\mathrm{op}}$, $i = 1, \ldots, k$). Hint: use an SVD of Z, and consider what the conditions on Z say about the factors.

6.9 We establish the result in Example 6.1.f, on the subgradients of the trace norm. We use the dual representation

$$\|X\|_{\mathrm{tr}} = \sup_{\|Z\|_{\mathrm{op}} \le 1} \langle Z, X \rangle,$$

covered later in Chapter 10.3 (where recall we write $\langle Z, X \rangle = \mathrm{tr}(Z^\mathsf{T} X)$), along with (6.4). As in the example, we write $X = U \Sigma V^\mathsf{T}$ for an SVD of X.

a. Prove that if $Z = UV^\mathsf{T} + W$, for $\|W\|_{\mathrm{op}} \le 1$, $U^\mathsf{T} W = 0$, and $WV = 0$, then $\|Z\|_{\mathrm{op}} \le 1$ and $\langle Z, X \rangle = \|X\|_{\mathrm{tr}}$. Hint: use an SVD of Z, and consider what the conditions on W say about the factors.

b. Prove the opposite direction: if $\|Z\|_{\mathrm{op}} \le 1$ and $\langle Z, X \rangle = \|X\|_{\mathrm{tr}}$ then Z must be of the above form ($Z = UV^\mathsf{T} + W$, where $\|W\|_{\mathrm{op}} \le 1$, $U^\mathsf{T} W = 0$, and $WV = 0$). Hint: define $W = Z - UV^\mathsf{T}$, and consider what the conditions on Z say about W.

6.10 Prove the result in Property 6.2.F, on subgradients of a composition of convex functions.

6.11 Let f be a function of a block variable (x, z), and let

$$F(x) = \mathbb{E}_z[f(x, z)] = \int f(x, z)\, dP(z),$$

for a distribution P over z. Fix any x, and suppose that for each z, we have $s(z) \in \partial_x f(x, z)$. Prove that $s = \mathbb{E}_z[s(z)] \in \partial F(x)$.

6.12 For a convex function f, suppose $s \in \mathbb{R}^d$ is such that at a point $x \in \mathrm{dom}(f)$,

$$f(y) \ge f(x) + s^\mathsf{T}(y - x), \quad \text{for all } y \in \mathrm{dom}(f) \text{ such that } \|x - y\| \le \delta,$$

for some $\delta > 0$. Prove that $s \in \partial f(x)$. Note: this, combined with subgradient optimality (6.6), gives another way of seeing the fundamental theorem of convex optimization, Property 4.2.A: a zero subgradient locally implies a zero subgradient globally.

6.13 Prove or disprove, for the subgradient optimality condition for a constrained convex problem: if (6.7) holds, then there can still exist another $v \in \partial f(x)$, $v \neq s$, such that

$$v^\mathsf{T}(y - x) < 0, \quad \text{for some } y \in C.$$

In other words, one subgradient s is telling us not to move away from x at all (restricting the search directions to those keeping us in C), yet another subgradient v is telling us to move in the direction of another feasible point y.

6.14 In this exercise, we will work through the proof of Theorem 6.5.

a. To show the "if" direction in the containment $T \subseteq \partial f$, fix any x_1 such that $s_1 \in T(x_1)$, and define

$$f(x) = \sup \left\{ \sum_{i=1}^{n-1} s_i^\mathsf{T}(x_{i+1} - x_i) + s_n^\mathsf{T}(x - x_n) : n \geq 1,\ s_i \in \partial f(x_i),\ i = 2, \ldots, n \right\}.$$

Prove that f is convex and $T \subseteq \partial f$.

b. To show the "if" direction in the equality $T = \partial f$, observe that if T is maximal, and $T \subseteq \partial f$, then we must have $T = \partial f$. Prove that, additionally, T is the subdifferential of a *closed* convex function. Hint: define $g(x) = \liminf_{y \to x} f(x)$, and show that g is closed with $\partial f \subseteq \partial g$. Then then invoke maximality.

A note on the remaining part, the "if" direction in the equality $T = \partial f$ and uniqueness of the solution f to $T = \partial f$, up to an additive constant: the arguments required for this part are more subtle. What we can say immediately is that, for any closed convex f, we have $\partial f \subseteq \partial g$ for some convex closed g, where ∂g is maximal cyclically monotone; this follows from the fact that each cyclically monotone operator must be embedded in some maximal one (using Zorn's lemma), and that this maximal one can always be written as ∂g for closed convex g (by part b of this exercise). However, to prove the next step:

$$f, g \text{ closed convex with } \partial f \subseteq \partial g \implies f = g + c \text{ for some constant } c, \tag{6.13}$$

requires a more involved argument. We refer the reader to the proofs of Theorems 3 and 4 in [**Roc66**], which are based on approximate subgradients and duality; see also Theorems 12.17 and 12.25 in [**RW09**], which take a variational approach.

6.15 We examine some properties of cyclical monotonicity, which, recall, for a set-valued operator T on $\mathbb{R}^d$, means that it satisfies (6.10) for all $n \geq 1$ and all (x_i, s_i), $i = 1, \ldots, n$ in its graph.

a. Show that an equivalent condition for T to be cyclically monotone is

$$\sum_{i=1}^n s_i^\mathsf{T} x_i \geq \sum_{i=1}^n s_{\sigma(i)}^\mathsf{T} x_i,$$

for all $n \geq 1$, permutations σ on $1, \ldots, n$, and (x_i, s_i), $i = 1, \ldots, n$ in the graph of T.

b. Show that when $d = 1$ (so that T is set-valued on $\mathbb{R}$), monotonicity (6.9) is equivalent to cyclical monotonicity. Hint: a cyclically monotone operator is monotone. For the other direction, use the condition from part a, and consider the rearrangement inequality.

6.16 In this exercise, we will work through the proof of Theorem 6.6.

a. Prove that (ii) $\iff$ (iii) by using the subgradient characterization of convexity for f_m, from Theorem 6.1. Prove that (iii) $\implies$ (iv) using subgradient monotonicity (applied to

f_m), and (iv) $\implies$ (i) using the Cauchy-Schwarz inequality. Note that these arguments are analogous to those given in Exercise 3.11 for Theorem 3.9.

b. Now we will work towards (iv) $\implies$ (ii), in steps laid out over the next couple of parts. First, fix any $x \in \mathrm{dom}(f)$, $v \in \mathbb{R}^d$, and let $g(t) = f_m(x + tv)$. Define

$$T(t) = \{s^\mathsf{T} v : s \in \partial f(x + vt) - m(x + vt)\}.$$

Use (iv) to argue that T is monotone, in fact, cyclically monotone by Exercise 6.15 part b (as T set-valued in $\mathbb{R}$).

c. Apply Rockafellar's embedding theorem and the closedness of f to conclude g is convex, and as x, v were arbitrary, f_m itself must be convex. Hint: by the embedding theorem, we know that $T \subseteq \partial h$ for some closed convex function h. Rewrite this containment as

$$\partial g(x) \subseteq \partial h(x) + m(x + vt), \quad \text{for all } x.$$

Now use closedness of g and (6.13) to prove that g is convex.

6.17 Prove that a strongly convex function is coercive. Hint: use part (iii) of Theorem 6.6.

6.18 We will prove Theorem 6.9. Let $\mathcal{X}$ denote the set of category d subdifferentiable points along the level set $\{x : f(x) = t\}$, and denote $U_x = \mathrm{int}(\mathrm{cone}(\partial f(x)))$, for $x \in \mathcal{X}$.

a. Prove that $U_x \neq \emptyset$, $x \in \mathcal{X}$.

b. Prove that $U_x \cap U_y = \emptyset$, for each $x \neq y$. Hint: use convexity of $\{x : f(x) \leq t\}$, and the normal cone representation of the subdifferentials, from Theorem 6.8.

c. Prove that any collection of open disjoint subsets in $\mathbb{R}^d$ must be countable, completing the proof of Thoerem 6.9. Hint: use the fact that the rationals are dense in $\mathbb{R}^d$.

Proximal Mappings

7.1. Definition and properties

In this chapter we cover a useful concept that generalizes the notion of (Euclidean) projection onto a given set: the *proximal mapping* (also called proximal map, or proximal operator) associated with a function $f : \mathbb{R}^d \to [-\infty, \infty]$. This is a set-valued map, denoted by prox_f, that maps each $x \in \mathbb{R}^d$ to a subset of $\mathbb{R}^d$ (possibly the empty set), defined by

$$\text{prox}_f(x) = \underset{z}{\text{argmin}} \left\{ \frac{1}{2}\|x - z\|_2^2 + f(z) \right\}. \tag{7.1}$$

Here and throughout, we denote by $\text{argmin}_{z \in S} F(z)$ the set of minimizers of a function F over a set S, and denote by $\text{argmin}_z F(z)$ the set of minimizers of F over its entire effective domain. We will abide by the common convention that argmin returns the minimizer itself when the latter is unique (not the set containing the minimizer).

As mentioned, we can think of prox_f as a generalization of the projecton map onto a set. Indeed for $f = I_C$, the characteristic function of a set C,

$$\text{prox}_f(x) = P_C(x) = \underset{z \in C}{\text{argmin}} \ \|x - z\|_2. \tag{7.2}$$

Keeping this connection in mind will generally be helpful for developing intuition about proximal maps because—as we will see in what follows—the proximal map associated with a closed convex function shares several useful properties of projection onto a closed convex set.

A slight variation on (7.1) will be of central interest in this chapter,

$$\text{prox}_{\lambda f}(x) = \underset{z}{\text{argmin}} \left\{ \frac{1}{2\lambda}\|x - z\|_2^2 + f(z) \right\}, \tag{7.3}$$

which is simply the proximal map associated with the scaled function λf for $\lambda > 0$. An intimately related object is

$$f_\lambda(x) = \inf_z \left\{ \frac{1}{2\lambda}\|x - z\|_2^2 + f(z) \right\}, \tag{7.4}$$

called the *Moreau envelope* of λf, which will be studied a bit later in Chapter 7.6.

We now discuss some basic properties and interpretations of proximal mappings. We generally assume henceforth that $f \neq \infty$ ($\text{dom}(f) \neq \emptyset$) to rule out trivialities when studying proximal maps.

A. Existence and uniqueness. First we discuss a case where the proximal mapping acts as a well-defined function, from $\mathbb{R}^d$ to $\mathbb{R}^d$. If f is closed and convex, then for any $\lambda > 0$ the function

$$z \mapsto \frac{1}{2\lambda}\|x - z\|_2^2 + f(z)$$

is closed and strongly convex. For the proximal mapping (7.3) to be a function (single-valued rather than set-valued), we need a minimizer of the function in the above display to exist and be unique. Existence comes from Weirstrass' theorem (Theorem 4.8), along with the fact that strongly convex functions are coercive (Exercise 6.17). Uniqueness comes from the fact that strictly (thus strongly) convex functions have at most one minimizer (Exercise 4.2 part d). We summarize this next.

Theorem 7.1. *For a closed and convex function $f : \mathbb{R}^d \to (-\infty, \infty]$ with nonempty domain, and for any $\lambda > 0$, the proximal map (7.3) is a well-defined function: it maps each point in $x \in \mathbb{R}^d$ to a point $\mathrm{prox}_{\lambda f}(x) \in \mathbb{R}^d$.*

B. Subgradient characterization. For closed convex f, the optimization problem

$$\underset{z}{\text{minimize}} \quad \frac{1}{2\lambda}\|x - z\|_2^2 + f(z)$$

has a subgradient optimality condition: $0 \in (z - x)/\lambda + \partial f(z)$, that is, $x - z \in \lambda \partial f(z)$. (Convexity is used here to write the subdifferential of a sum as a sum of subdifferentials, by Property 6.2.B.) Therefore by subgradient optimality, we get $z = \mathrm{prox}_{\lambda f}(x)$ if and only if

$$x - z \in \lambda \partial f(z), \tag{7.5}$$

which we call the subgradient characterization of the proximal operator associated with λf.

For $f = I_C$, the characteristic function of a convex set C, if we set (say) $\lambda = 1$, and recall that $\partial I_C(z) = \mathcal{N}_C(z)$, the normal cone to C at z, then we can see that (7.5) reduces to the variational inequality (4.10) for the projection $z = P_C(x)$ of x onto C.

C. Gradient step interpretation. The subgradient characterization (7.5) leads to a nice interpretation for $\mathrm{prox}_{\lambda f}(x)$ when f convex and differentiable. In this case, note that (7.5) becomes a fixed-point equation for $z = \mathrm{prox}_{\lambda f}(x)$,

$$z = x - \lambda \nabla f(z).$$

For small λ, it is reasonable to assume that $z = \mathrm{prox}_{\lambda f}(x)$ will be close to x (as the term $\|x - z\|_2^2$ in the criterion in (7.3) will be multiplied by a large weight), so replacing $\nabla f(z)$ with $\nabla f(x)$ in the above fixed-point equation gives the approximation

$$\mathrm{prox}_{\lambda f}(x) \approx x - \lambda \nabla f(x). \tag{7.6}$$

That is, for small λ, we can interpret $\mathrm{prox}_{\lambda f}(x)$ as performing a gradient descent step starting at x, with step size λ. We will return to this interpretation in Chapter 7.6, when we will be able to make a more precise statement involving the Moreau envelope.

D. Resolvent of subdifferential. A transformation of the subgradient characterization (7.5) yields another important interpretation for the proximal operator. Rearranging the right-hand side in (7.5), and using I for the identity map, we get

$$\begin{aligned} z = \mathrm{prox}_{\lambda f}(x) &\iff x \in (I + \lambda \partial f)(z) \\ &\iff z \in (I + \lambda \partial f)^{-1}(x). \end{aligned}$$

Here, we should interpret $I + \lambda\partial f$ as a set-valued map: in general, it evaluates to $(I + \lambda\partial f)(z) = \{z + \lambda s : s \in \partial f(z)\}$, and its preimage to $(I + \lambda\partial f)^{-1}(x) = \{z : x \in (I + \lambda\partial f)(z)\}$. Inspecting the above display carefully, we can see that it reveals something quite interesting: the preimage in the last line must actually be *single-valued*, since there is exactly one point $z = \text{prox}_{\lambda f}(x)$ that satisfies $z \in (I + \lambda\partial f)^{-1}(x)$. This allows us to write

$$\text{prox}_{\lambda f} = (I + \lambda\partial f)^{-1}, \tag{7.7}$$

where we interpret both the left- and right-hand sides in (7.7) as single-valued mappings. This is a representation for the proximal map in terms of what is called the *resolvent* of the subdifferential.

Example 7.2. The following are examples of proximal maps of general interest in statistics and machine learning. In each case, the stated result can be derived using the subgradient characterization (7.5) for the proximal operator.

a. For $f(x) = \frac{1}{2}x^T Ax + b^T x + c$ with $A \succeq 0$, we have

$$\text{prox}_{\lambda f}(x) = (I + \lambda A)^{-1}(x - \lambda b).$$

b. For $f(x) = \|x\|_1$, we have (Exercise 7.1 part a):

$$[\text{prox}_{\lambda f}(x)]_i = [S_\lambda(x)]_i = [x_i - \text{sign}(x_i)\lambda]_+, \quad i = 1, \ldots, d.$$

where recall $a_+ = \max\{a, 0\}$ gives the positive part of a, and we interpret $\text{sign}(0) = 1$. The operator S_λ is called (coordinatewise) *soft-thresholding* at the level λ, equivalently

$$[S_\lambda(x)]_i = \begin{cases} x_i - \lambda & x_i > \lambda \\ 0 & |x_i| \leq \lambda \\ x_i + \lambda & x_i < -\lambda \end{cases}, \quad i = 1, \ldots, d. \tag{7.8}$$

c. For $f(x) = \|x\|_2$, we have (Exercise 7.1 part b):

$$\text{prox}_{\lambda f}(x) = \left(1 - \frac{\lambda}{\|x\|_2}\right)_+ x.$$

d. For $f(X) = \|X\|_{\text{tr}}$, the trace norm, and $X = U\Sigma V^T$ denoting the SVD of X, we have (Exercise 7.2 part a):

$$\text{prox}_{\lambda f}(X) = U\,\text{diag}(S_\lambda(\sigma))V^T.$$

where σ is the vector of singular values (the diagonal of Σ), S_λ is the soft-thresholding operator defined above, and we use $A = \text{diag}(a)$ to construct a diagonal matrix A from a vector of diagonal elements a. The proximal operator here is typically denoted M_λ, and called *matrix soft-thresholding*.

e. For $f(X) = \|X\|_F$, the Frobenius norm, we have (Exercise 7.2 part b):

$$\text{prox}_{\lambda f}(X) = \left(I - \frac{\lambda}{\|X\|_F}\right)_+ X,$$

where A_+ returns the elementwise positive part of a matrix A.

f. For $f(x) = \|x\|_0$, the ℓ_0 norm (recall this returns the number of nonzero values, as in (3.20); it is nonconvex and *not* actually a norm) we have (Exercise 7.1 part c):

$$[\text{prox}_{\lambda f}(x)]_i = [H_\lambda(x)]_i = x_i 1\{|x_i| > \lambda\}, \quad i = 1, \ldots, d.$$

The operator H_λ is known as (coordinatewise) *hard-thresholding* at the level λ.

g. For $f(x) = \text{rank}(X)$ (recall this is nonconvex), and $X = U\Sigma V^\mathsf{T}$ denoting the SVD of X, we have (Exercise 7.3 part c):

$$\text{prox}_{\lambda f}(X) = U \,\text{diag}(H_\lambda(\sigma))V^\mathsf{T},$$

where H_λ is the hard-thresholding operator as defined above.

What can we do when the proximal operator is not available in closed-form? In some settings, fast computaton may still be possible using specialized techniques.

Example 7.3. The examples below highlight interesting proximal mappings that cannot be computed in closed-form, but can still be efficiently evaluated using specialized algorithms.

a. Let $y_i \in \mathbb{R}$, $i = 1, \ldots, n$ be independent draws from an exponential family distribution with distinct natural parameters $\eta_i \in \mathbb{R}$, $i = 1, \ldots, n$, but a common sufficient statistic $T : \mathbb{R} \to \mathbb{R}$ and log-partition function $\psi : \mathbb{R} \to \mathbb{R}$. The negative log likelihood is

$$f(\eta) = \sum_{i=1}^n \Big(\psi(\eta_i) - T(y_i)\eta_i\Big),$$

where we have dropped terms not depending on η. The associated proximal operator

$$\text{prox}_{\lambda f}(\eta) = \underset{u}{\text{argmin}} \left\{ \frac{1}{2}\|\eta - u\|_2^2 + \lambda \sum_{i=1}^n \Big(\psi(u_i) - T(y_i)u_i\Big) \right\}$$

is not generally available in closed-form, but is efficiently computable for differentiable convex ψ (recall that ψ is always convex in any exponential family distribution). First note that the proximal map decomposes across coordinates: for each $i = 1, \ldots, n$,

$$[\text{prox}_{\lambda f}(\eta)]_i = \underset{u_i}{\text{argmin}} \left\{ \frac{1}{2}(\eta_i - u_i)^2 + \lambda\Big(\psi(u_i) - T(y_i)u_i\Big) \right\}.$$

The first-order optimality condition for the above problem is

$$u_i + \lambda\psi'(u_i) = \eta_i + \lambda T(y_i),$$

a nonlinear equation in univariate u_i. The left-hand side is monotone nondecreasing in u_i, so the solution can be iteratively approximated with (say) binary search.

b. The *total variation (TV) penalty* is defined as

$$f(x) = \sum_{i=1}^{d-1} |x_{i+1} - x_i|. \tag{7.9}$$

This is a convex function (it is a seminorm). Its proximal operator

$$\text{prox}_{\lambda f}(z) = \underset{z}{\text{argmin}} \left\{ \frac{1}{2}\|x - z\|_2^2 + \lambda \sum_{i=1}^{d-1} |z_{i+1} - z_i| \right\}$$

is efficiently computable (in linear-time, by which we mean the number of operations grows linearly in the dimension d) with dynamic programming or taut string methods.

c. The *sorted ℓ_1 penalty (SLOPE)* is defined, for constants $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$, as

$$f(x) = \sum_{i=1}^d \lambda_i |x|_{(i)}, \tag{7.10}$$

where $|x|_{(i)}$ denotes the i^{th} largest element of $|x_1|, \ldots, |x_d|$. This is a convex function (it is indeed a norm, reducing to the ℓ_1 norm for $\lambda_1 = \cdots = \lambda_d$). Its proximal map

$$\operatorname{prox}_f(z) = \operatorname*{argmin}_z \left\{ \frac{1}{2}\|x - z\|_2^2 + \sum_{i=1}^{d} \lambda_i |z|_{(i)} \right\}$$

can be reduced to isotonic projection, as in (7.16), and is thus efficiently computable (in linear-time) with the pool adjacent violators algorithm (PAVA).

7.2. Proximal calculus

We describe rules that are helpful in the calculation of proximal operators; throughout f, f_1, f_2 are assumed to be closed and convex functions with effective domains in $\mathbb{R}^d$. The last two paragraphs are actually "non-rules" in the general sense reflected in their titles, as they are not generally applicable to *all* sums of functions, and *all* linear compositions, respectively. This is especially noteworthy in contrast to the subgradient rules for sums and linear compositions (recall Properties 6.2.B and 6.2.E, respectively), which do apply generally.

A. Scaling and translation. If $F(x) = f(ax + b)$, where $a \neq 0$ and $b \in \mathbb{R}^d$, then

$$\operatorname{prox}_F(x) = \frac{1}{a}\big(\operatorname{prox}_{a^2 f}(ax + b) - b\big).$$

B. Separable sum. If $F(x) = f_1(x_1) + f_2(x_2)$ for a block variable $x = (x_1, x_2)$, then

$$\operatorname{prox}_F(x) = \operatorname{prox}_{f_1}(x_1) + \operatorname{prox}_{f_2}(x_2).$$

C. Linear sum. If $F(x) = f(x) + a^{\mathsf{T}}x$, then $\operatorname{prox}_F(x) = \operatorname{prox}_f(x - a)$.

D. Quadratic sum. If $F(x) = f(x) + \frac{m}{2}\|x - a\|_2^2$, where $m \geq 0$, then

$$\operatorname{prox}_F(x) = \operatorname{prox}_{tf}\big(tx + (1 - t)a\big),$$

where $t = 1/(1 + m)$.

E. General sum. If $F(x) = f(x) + g(x)$ (a general sum and *not* separable), then prox_F is not in general easily computable from prox_f and prox_g. However, in some special cases the proximal map of F can be expressed via composition of the maps of f, g:

$$\operatorname{prox}_F = \operatorname{prox}_f \circ \operatorname{prox}_g \,. \tag{7.11}$$

This is the case for the linear sum rule above, which can be recast in terms of compositions:

$$\operatorname{prox}_{f + \langle a, \cdot \rangle} = \operatorname{prox}_f \circ \operatorname{prox}_{\langle a, \cdot \rangle},$$

where we use $\langle a, \cdot \rangle$ for the map $x \mapsto a^{\mathsf{T}}x$. As another example, the quadratic sum rule above (with $a = 0$) implies that for any positively homogeneous f (Exercise 7.6):

$$\operatorname{prox}_{f + \frac{m}{2}\|\cdot\|_2^2} = \operatorname{prox}_{\frac{m}{2}\|\cdot\|_2^2} \circ \operatorname{prox}_f \,. \tag{7.12}$$

Finally, denoting by $\|\cdot\|_{\text{TV}}$ the TV seminorm defined in (7.9), for any permutation invariant f and $\lambda > 0$, it holds that (Exercise 7.8):

$$\operatorname{prox}_{f + \lambda\|\cdot\|_{\text{TV}}} = \operatorname{prox}_f \circ \operatorname{prox}_{\lambda\|\cdot\|_{\text{TV}}} \,. \tag{7.13}$$

See the chapter notes for further discussion of the proximal decomposition phenomenon (7.11).

F. Linear composition. If $F(x) = f(Ax)$ for a matrix $A \in \mathbb{R}^{k\times d}$, then prox_F is not in general easily computable from A and prox_f. However, if A is orthogonal then

$$\text{prox}_F(x) = A^T \text{prox}_f(Ax).$$

(See Exercise 7.9 part b for a slight generalization.)

7.3. Proximal optimality condition

For closed and convex f, it turns out that $f(x) \leq f(y)$ for all y if and only if

$$x = \text{prox}_f(x), \tag{7.14}$$

that is, x minimizes f if and only if x is a fixed-point of the proximal operator of f, which is called the *proximal optimality condition*. This is easy to verify from the subgradient characterization (7.5): we can see that $x = \text{prox}_f(x)$ if and only if $0 \in \partial f(x)$, which holds if and only x minimizes f, by the subgradient optimality condition (6.6).

The proximal fixed-point equation (7.14) is rarely of direct use for deriving minima of a given function, but it does has a number of useful consequences, both algorithmic and conceptual. To see this, let us further develop the fixed-point perspective by considering a function $F = f + g$, for closed and convex f, g, where f is differentiable. In this case, x minimizes F if and only if, for any $\lambda > 0$,

$$x = \text{prox}_{\lambda g}\big(x - \lambda \nabla f(x)\big). \tag{7.15}$$

The proof that (7.15) is a necessary and sufficient condition for optimality is based on the resolvent characterization (7.7) of the proximal operator; see Exercise 7.10.

The fixed-point equation (7.15) for minimizing the composite function $F = f + g$, and generalizations thereof (see Exercise 7.17), underlie various proximal algorithms for iterative optimization. Another nice consequence is that it allows us to rigorously certify the solution structure induced by certain convex penalties, as discussed next.

Example 7.4. The examples below show how we can use the analytic form of proximal maps (when available) to confirm the structure of solutions in more general optimization problems. In each example, there is nothing special about squared loss, and the exact same conclusion holds for an arbitrary loss f.

a. How do we know that the ℓ_1 penalty in the lasso problem (4.4) induces sparsity in its solution(s)? From the proximal optimality condition (7.15), with $f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$ and $g(\beta) = \lambda\|\beta\|_1$ (and denoting by t the proximal parameter), we can see that β is a lasso solution if and only if, for any $t > 0$,

$$\beta = S_{\lambda t}\big(\beta + tX^T(y - X\beta)\big),$$

where $S_{\lambda t}$ denotes soft-thresholding operator at the level λt, as in (7.8). This certifies that the lasso problem admits a sparse solution (as the output of $S_{\lambda t}$ is sparse), with generally a greater degree of sparsity for a larger λ.

b. How do we know that the trace norm penalty in the matrix completion problem (5.13) induces a low-rank structure in its solution(s)? Again, from proximal optimality (7.15) with $f(\Theta) = \frac{1}{2}\|P_\Omega(X - \Theta)\|_F^2$ and $g(\Theta) = \lambda\|\Theta\|_{\text{tr}}$, we learn that Θ is a solution if and only if, for any $t > 0$,

$$\Theta = M_{\lambda t}\big(\Theta + tP_\Omega(X - \Theta)\big),$$

with $M_{\lambda t}$ denoting matrix soft-thresholding at the level λt, as in Example 7.1.d. This certifies that the matrix completion problem admits a low-rank solution (as the output of $M_{\lambda t}$ is low-rank), with generally a smaller rank for a larger λ.

The idea demonstrated in the last example extends beyond the case of a closed-form proximal mapping: if we have a specialized algorithm for computing the proximal mapping of g and its steps reveal certain structure (for example, the dynamic programming algorithm for the proximal map of the TV penalty (7.9) shows that the output admits a piecewise constant structure), then proximal optimality (7.15) confirms that this structure persists more generally, when minimizing $f + g$.

7.4. Euclidean projection

Now we turn to discussing (Euclidean) projection, which is the special case (7.2) of a proximal map when $f = I_C$, the characteristic function of a set C. From Theorem 7.1, we know that projection P_C is a well-defined function for any closed and convex set C.

Projections play a key role in various parts of convex analysis, and the design of optimization algorithms. We have already covered various aspects of projections in earlier chapters of this book, to do with optimality conditions (4.10) and subgradients (6.5). In this section, we cover numerous examples of projections and discuss two of their central properties, which serve as motivation for analogous properties held by proximal operators.

Example 7.5. The following are examples of projections onto some fundamental classes of convex sets.

a. For $C = \{x : Ax = b\}$, an affine subspace, we have $P_C(x) = x + A^\mathsf{T}(AA^\mathsf{T})^+(b - Ax)$.

b. As a special case, for the column space, row space, and null space of given a matrix A, we have, respectively,

$$\begin{aligned} P_{\mathrm{col}(A)} &= A^\mathsf{T}(A^\mathsf{T}A)^+A^\mathsf{T} = AA^\mathsf{T}(AA^\mathsf{T})^+, \\ P_{\mathrm{row}(A)} &= A(AA^\mathsf{T})^+A = A^\mathsf{T}A(A^\mathsf{T}A)^+, \\ P_{\mathrm{null}(A)} &= I - A(AA^\mathsf{T})^+A = I - A^\mathsf{T}A(A^\mathsf{T}A)^+. \end{aligned}$$

c. For $C = \{x : a^\mathsf{T}x = b\}$ with $a \neq 0$, a hyperplane, we have $P_C(x) = x + (b - a^\mathsf{T}x)a/\|a\|_2$; whereas for $C = \{x : a^\mathsf{T}x \leq b\}$, a halfspace, we have

$$P_C(x) = \begin{cases} x + (b - a^\mathsf{T}x)a/\|a\|_2 & a^\mathsf{T}x > b \\ x & a^\mathsf{T}x \leq b \end{cases}.$$

d. For $C = [a_1, b_1] \times \cdots \times [a_d, b_d]$, a hyperrectangle, we have

$$[P_C(x)]_i = \begin{cases} a_i & x_i < a_i \\ x_i & a \leq x_i \leq b \\ b_i & x_i > b_i \end{cases}, \quad i = 1, \ldots, d.$$

e. As a special case, for the nonnegative orthant, we have $P_{\mathbb{R}^d_+}(x) = (x_i)_+$, $i = 1, \ldots, d$.

f. For $C = \{x : \|x\|_2 \leq 1\}$, the unit ℓ_2 ball, we have

$$P_C(x) = \begin{cases} x/\|x\|_2 & \|x\|_2 > 1 \\ x & \|x\|_2 \leq 1 \end{cases}.$$

g. For $C = \mathbb{S}^d_+$, the positive semidefinite cone, the projection operator onto C (from the ambient space $\mathbb{S}^d$ of symmetric $d \times d$ matrices) is $P_C(X) = U\Sigma_+ U^\mathsf{T}$, where $X = U\Sigma U^\mathsf{T}$ is the eigendecomposition of X, and Σ_+ is the elementwise positive part of Σ.

Like proximal operators, even when not available in closed-form, some projections may still be efficiently computable using specialized algorithms.

Example 7.6. The following are examples of some interesting projection operators that are not available in closed-form, but can be efficiently computed with specialized algorithms.

a. For $C = \{x : 1^\mathsf{T} x = 1,\ x \geq 0\}$, a polyhedron which is called the *probability simplex*, the projection $P_C(x)$ can be computed efficiently with a specialized algorithm whose cost is dominated by sorting the entries of x (and is thus nearly linear-time); see Exercise 11.2.

b. For $C = \{x : \|x\|_1 \leq 1\}$, the unit ℓ_1 ball, the projection operator P_C can be reduced to projection onto the probability simplex (Exercise 7.11), and thus the former projection map is again efficiently computable (in nearly linear-time).

c. For $C = \{x : x_1 \leq \cdots \leq x_d\}$, the isotonic cone, the projection map

$$P_C(x) = \operatorname*{argmin}_{z:\, z_1 \leq \cdots \leq z_d} \|x - z\|_2^2, \tag{7.16}$$

can be computed efficiently (in linear-time) with an algorithm called the pool adjacent violators algorithm (PAVA).

Below we describe two key properties of projection maps. Both properties admit direct analogs for proximal maps, the first covered shortly in Chapter 7.5, and the second later in Chapter 8.4.1.

A. Nonexpansiveness. The projection map P_C onto any convex set C is *nonexpansive*, meaning

$$\|P_C(x) - P_C(y)\|_2 \leq \|x - y\|_2, \quad \text{for all } x, y. \tag{7.17}$$

Equivalently, this says that the map P_C is Lipschitz continuous with Lipschitz constant $L = 1$. For convex sets, this property is quite intuitive; see Figure 7.1 for an illustration. For nonconvex sets, it is actually no longer true in general; the same figure gives an illustration.

Beyond nonexpansiveness of the projection map itself, the *residual projection map* $I - P_C$ for convex C, defined as $(I - P_C)(x) = x - P_C(x)$, is also nonexpansive:

$$\|(I - P_C)(x) - (I - P_C)(y)\|_2 \leq \|x - y\|_2, \quad \text{for all } x, y. \tag{7.18}$$

Indeed, both (7.17) and (7.18) are consequences of a more general property for convex projections that is called *firm nonexpansiveness*:

$$\big(P_C(x) - P_C(y)\big)^\mathsf{T}(x - y) \geq \|P_C(x) - P_C(y)\|_2^2, \quad \text{for all } x, y. \tag{7.19}$$

Firm nonexpansiveness follows from variational inequality (4.10); details are given in Exercise 7.12.

That firm nonexpansiveness (7.19) implies nonexpansiveness (7.17) and residual nonexpansiveness (7.18) is also covered in Exercise 7.12. Of particular note: Exercise 7.12 part d shows that firm nonexpansiveness implies the property:

$$\|P_C(x) - P_C(y)\|_2^2 + \|(I - P_C)(x) - (I - P_C)(y)\|_2^2 \leq \|x - y\|_2^2, \quad \text{for all } x, y, \tag{7.20}$$

from which (7.17) and (7.18) clearly follow.

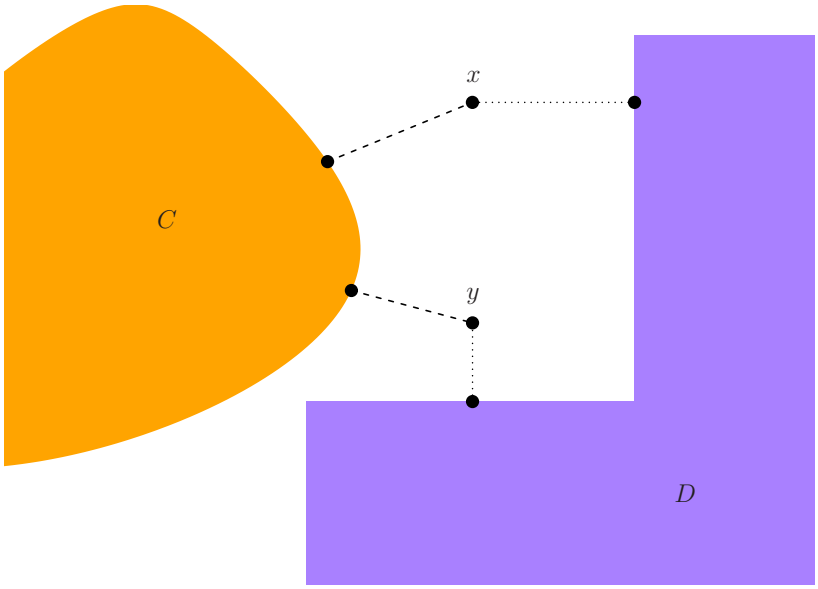

Figure 7.1. Illustration of properties of Euclidean projection operators onto convex and nonconvex sets, C and D. The projections of x and y onto C (visualized via dashed lines) can only grow closer in Euclidean norm, whereas their projections onto D (dotted lines) can spread apart.

B. Orthogonal decomposition. For a linear subspace L, it holds that

$$P_L(x) + P_{L^\perp}(x) = x, \quad \text{for all } x, \tag{7.21}$$

where $L^\perp = \{x : x^\mathsf{T} y = 0, \text{ for all } y \in L\}$ denotes the orthogonal complement of L, which, as it turns out, is itself a linear subspace. An interesting application of the orthogonal decomposition property (7.21) occurs when $L = \mathrm{row}(A)$ and $L^\perp = \mathrm{null}(A)$, the row and null space of a matrix A, and this explains the relationship between the projection maps onto $\mathrm{row}(A)$ and $\mathrm{null}(A)$ in Example 7.4.b.

The orthogonal decomposition fact (7.21) for linear subspaces is straightforward to verify. From the variational inequality (4.10) for projection onto L, denoting $z = P_L(x)$, we first note that this is actually equivalent to a variational *equality*

$$(x - z)^\mathsf{T}(z - y) = 0, \quad \text{for any } y \in L. \tag{7.22}$$

as $v = z - y$ with $z, y \in L$ implies $v \in L$, and hence $-v \in L$. The corresponding variational equality for $z' = P_{L^\perp}(x)$ is therefore

$$(x - z')^\mathsf{T}(z' - y') = 0, \quad \text{for any } y' \in L^\perp.$$

This is satisfied for $z' = x - z$, as we get $z^\mathsf{T}(x - z - y') = z^\mathsf{T}(x - z) + z^\mathsf{T} y' = 0 + 0$ for any $y' \in L^\perp$, the first term being zero due to (7.22) with $y = 0$, and the second term being zero due to the fact that $z \in L$ and $y' \in L$. This confirms $P_{L^\perp}(x) = x - P_L(x)$ and proves (7.21).

7.5. Proximal nonexpansiveness*

We examine firm nonexpansiveness of the proximal mapping. As usual, we take f to be closed and convex, and $\lambda > 0$. Just as we saw for projection maps onto convex sets, it turns out that prox_f is *firmly nonexpansive*,

$$\big(\mathrm{prox}_{\lambda f}(x) - \mathrm{prox}_{\lambda f}(y)\big)^\mathsf{T}(x - y) \geq \|\mathrm{prox}_{\lambda f}(x) - \mathrm{prox}_{\lambda f}(y)\|_2^2, \quad \text{for all } x, y, \tag{7.23}$$

This is a powerful property, but verifying it is actually straightforward (easier shorter than proving from first principles the corresponding result for projections; recall Exercise 7.12), once we invoke

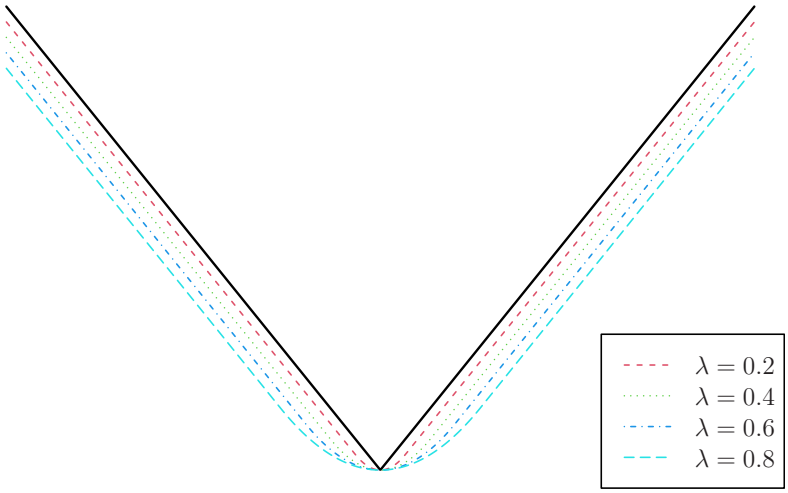

Figure 7.2. Convex function (solid line) along with its Moreau envelope plotted at four values of $\lambda > 0$ (dashed or dotted lines of various patterns). The Moreau envelope is always a differentiable convex minorant to the original function.

the appropriate tools: the proximal subgradient characterization, and subgradient monotonicity. To see this, without loss of generality, we take $\lambda = 1$, and abbreviate $z_x = \mathrm{prox}_f(x)$ and $z_y = \mathrm{prox}_f(y)$. Observe that

$$(z_x - z_y)^\mathsf{T}(x - y) = (z_x - z_y)^\mathsf{T}\big((x - z_x) - (y - z_y)\big) + \|z_x - z_y\|_2^2.$$

By the proximal subgradient characterization (7.5), we have $x - z_x \in \partial f(z_x)$ and $y - z_y \in \partial f(z_y)$, and by subgradient monotonicity (6.8), the first term on the right-hand side above is nonnegative. This proves the firm nonexpansiveness of prox_f, as desired.

Again, as in the case of projections (proved in part d of Exercise 7.12), firm nonexpansiveness (7.23) implies the more evocative property

$$\|\mathrm{prox}_{\lambda f}(x) - \mathrm{prox}_{\lambda f}(y)\|_2^2 + \|(I - \mathrm{prox}_{\lambda f})(x) - (I - \mathrm{prox}_{\lambda f})(y)\|_2^2 \leq \|x - y\|_2^2, \quad \text{for all } x, y, \tag{7.24}$$

from which we can see that proximal *nonexpansiveness* and *residual nonexpansiveness* follow,

$$\|\mathrm{prox}_{\lambda f}(x) - \mathrm{prox}_{\lambda f}(y)\|_2 \leq \|x - y\|_2, \quad \text{for all } x, y, \tag{7.25}$$

$$\|(I - \mathrm{prox}_{\lambda f})(x) - (I - \mathrm{prox}_{\lambda f})(y)\|_2 \leq \|x - y\|_2, \quad \text{for all } x, y, \tag{7.26}$$

respectively.

7.6. Moreau-Yosida regularization*

In this last section, we return to discussing the Moreau envelope f_λ of a closed convex function f, as defined in (7.4). Repeating its definition here for convenience, for $\lambda > 0$:

$$f_\lambda(x) = \inf_z \left\{ \frac{1}{2\lambda}\|x - z\|_2^2 + f(z) \right\}.$$

The Moreau envelope f_λ can be seen as a smoothed or regularized version of f, and accordingly it is also known as *Moreau-Yosida regularization*. Figure 7.2 gives an illustration. We note that more can be said about the precise connection between f_λ and regularization, and this will be revisited in the next chapter, from the perspective of conjugate functions.

Several important properties of f_λ are apparent. First, the Moreau envelope f_λ is itself convex (by Property 3.4.H). Second, f_λ has full domain, $\mathrm{dom}(f_\lambda) = \mathbb{R}^d$, even when the original function f does not. Third, f is minorized by its Moreau envelope,

$$f_\lambda(x) \leq f(x), \quad \text{for all } x.$$

Fourth, it is not hard to see that f_λ and f have the same set of minimizers (Exercise 7.14 part a), which means that minimization of f and of f_λ are equivalent problems. Fifth, a property tied to the fourth one (Exercise 7.14 part b), the point $z = \mathrm{prox}_{\lambda f}(x)$ achieves the infimum in (7.4), that is,

$$f_\lambda(x) = \frac{1}{2\lambda}\|x - \mathrm{prox}_{\lambda f}(x)\|_2^2 + f\big(\mathrm{prox}_{\lambda f}(x)\big). \tag{7.27}$$

Sixth, and perhaps most importantly, the Moreau envelope f_λ is differentiable, and satisfies

$$\nabla f_\lambda(x) = \frac{1}{\lambda}\big(x - \mathrm{prox}_{\lambda f}(x)\big). \tag{7.28}$$

This fact is always true regardless of the smoothness of f; it is not as immediately apparent as the above facts, and its proof is outlined in Exercise 7.14 parts c and d.

This latter property, on smoothness of the Moreau envelope, is useful for many reasons. First, it has important algorithmic implications, because (as already mentioned) minimization of f and f_λ are equivalent problems, and the latter function is always smooth and admits a clean formula for its gradient, as seen in (7.28). Second, it has gives a more rigorous perspective on the approximation presented in (7.6): rearranged, it says that for all x and all $\lambda > 0$, we have the equality

$$\mathrm{prox}_{\lambda f}(x) = x - \lambda \nabla f_\lambda(x).$$

For small enough values of λ, the Moreau envelope f_λ and the original function f should be similar, as should their gradients. This supports the idea of substituting $\nabla f(x)$ for $\nabla f_\lambda(x)$ on the right-hand side above, which yields (7.6).

The third implication of the gradient relation (7.28) requires a bit more background to explain. It turns out that the Moreau envelope is a uniquely identifying transform, in the following sense: if two closed convex functions have matching Moreau envelopes, then they must be the same function. Furthermore, the relationship between the proximal operator and the gradient of the Moreau envelope (7.28) then transfers this identifiability property to the former (up to an additive constant). These results are stated next, to conclude the chapter.

Theorem 7.7. *For closed convex $f, g : \mathbb{R}^d \to (-\infty, \infty]$ with nonempty domains, the following properties hold:*

(i) if $f_\lambda = g_\lambda$ for any $\lambda > 0$, then $f = g$;

(ii) if $\mathrm{prox}_{\lambda f} = \mathrm{prox}_{\lambda g}$ for any $\lambda > 0$, then $f = g + c$, for some constant $c \in \mathbb{R}$.

We call this the *identification theorem* for Moreau envelopes and proximal mappings; its proof is outlined in Exercise 7.16.

Chapter Notes

The notion of a proximal mapping is due to Jean Jacques Moreau, who published seminal work on the topic in the early and mid 1960s. Moreau studied the ways in which proximal maps generalize projections, and developed (what is now called) the Moreau envelope. The Moreau envelope is also sometimes called Moreau-Yosida regularization, to pay homage to earlier related work in functional analysis by Kosaku Yosida (cf. the Yosida approximation of operators). For readers seeking to learn

more about proximal operators, including extensive and important historical references, we refer to [**RW09**] (Chapters 1.G, 2.D, 3.D). We also refer to the recent monograph [**PB13**], which offers numerous interesting interpretations and examples, and useful references. The connection between the proximal map and the resolvent of the subdifferential operator is due to [**Roc76**]. For an elegant treatment of this and related topics from the perspective of monotone operators, see [**BC11**].

The proximal map of the total variation (TV) seminorm defines a minimization problem known as *total variation denoising*. This was proposed by [**ROF92**], and led to a large following in research across applied mathematics, signal processing, and statistics. An important contribution in statistics that develops this idea further, under the name *fused lasso*, is [**TSR+05**]. The taut string algorithm for TV denoising—that is, for computing the proximal map of the TV seminorm—was developed by [**DK01**], building on earlier work by [**MvdG97**]. A dynamic programming approach for solving the TV denoising problem was given in [**Joh13**]. The taut string and dynamic programming algorithms are both linear-time. To be clear, this is all in reference to the univariate TV denosing problem; for generalizations defined over multivariate lattices or arbitrary undirected graphs, the computation can be more challenging. Important algorithmic contributions to more general TV denoising settings include [**CD09**, **CP11**, **BS18**], among many others.

The sorted ℓ_1 penalty (SLOPE) was proposed by [**BvdBS+15**]. These same authors presented a fast algorithm for its proximal map by reducing the problem to isotonic projection, which can be solved in linear-time using the pool adjacent violators algorithm (PAVA) of [**BBBB72**]. See also [**dLHM10**] for a nice review of various fast algorithms for isotonic regression and related problems. Projection onto the probability simplex (and projection onto the ℓ_1 ball, which is reducible to the former) can be done in in nearly linear-time with a deterministic algorithm, or expected linear-time with a randomized algorithm, see [**DSSSC08**, **Con16**].

Insights into the proximal decomposition phenomenon (7.11), including fairly general sufficient conditions, were given in [**Yu13**]. This reproduces (and generalizes) previously known results about proximal decomposition in special cases, such as those found in [**FHHT07**] on the elastic net and fused lasso penalties. Exercise 7.7 below is based on [**Yu13**].

Exercises

7.1 We establish the proximal results for the ℓ_1, ℓ_2, and ℓ_0 norms in Example 7.2.

a. Using the subgradient characterization (7.5) and the subgradients of the ℓ_1 norm from Example 6.1.b, prove the result in Example 7.1.b.

b. Using the subgradient characterization (7.5) and the subgradients of the ℓ_2 norm from Example 6.1.c, prove the result in Example 7.1.c.

c. Now prove the result in Example 7.1.f on the ℓ_0 norm, by breaking the argument down into cases (for $|x_i| > \lambda$ and $|x_i| \leq \lambda$), and simply arguing directly that the minimizer is as claimed in each case.

7.2 We establish the proximal results for the trace and Frobenius norms in Example 7.2. To be entirely clear, for a function f defined on $\mathbb{R}^{k \times d}$, its proximal map is

$$\text{prox}_{\lambda f}(X) = \underset{z}{\text{argmin}} \left\{ \frac{1}{2}\|X - Z\|_F^2 + \lambda f(Z) \right\}, \tag{7.29}$$

which is consistent with interpreting the space $k \times d$ of matrices as a kd-dimensional vector space defined by its entries, and applying the usual definition (7.3) of the proximal map. The subgradient optimality condition for (7.29) says that $Z = \text{prox}_{\lambda f}(X)$ if and only if

$$X - Z \in \lambda \partial f(Z), \tag{7.30}$$

which is just the appropriate translation of the subgradient characterization (7.5) to the matrix domain.

a. Using the subgradient characterization (7.30) along with the subgradients of the trace norm from Example 6.1.f, prove the result in Example 7.1.d.

b. Using the subgradient characterization (7.30), where the subgradients of the Frobenius norm can derived analogously to those for the ℓ_2 norm as in Example 6.1.c, prove the result in Example 7.1.e.

7.3 Now we derive proximal mappings for a class of functions on $\mathbb{R}^{k \times d}$, of the form

$$f(X) = g(\sigma(X)), \tag{7.31}$$

for a function $g : \mathbb{R}^d \to \mathbb{R}$, where we write $\sigma(X) = (\sigma_1(X), \ldots, \sigma_k(X)) \in \mathbb{R}^k$ for the vector of singular values of X. An important example here is the *Schatten p-norm*, defined for $p \geq 1$ as

$$\|X\|_p = \|\sigma(X)\|_p, \tag{7.32}$$

with $\|\cdot\|_p$ on the right-hand side denoting the usual ℓ_p norm on vectors. Note that in terms of Shatten norms (that is, the left-hand side in all equalities below represents a Schatten norm):

- $\|\cdot\|_1 = \|\cdot\|_{\text{tr}}$, the trace norm;
- $\|\cdot\|_2 = \|\cdot\|_F$, the Frobenius norm; and
- $\|\cdot\|_\infty = \|\cdot\|_{\text{op}}$, the operator norm.

An important fact about functions f of the form (7.31): they are *unitarily invariant*, meaning $f(UXV) = f(X)$ for any orthogonal matrices U, V (because left- or right-muliplication by orthogonal matrices does not change the singular values).

a. For $X = U\Sigma V^\mathsf{T}$ denoting the SVD of X, and $\sigma = \sigma(X)$ is the diagonal of Σ, prove

$$\text{prox}_{\lambda f}(X) = U \,\text{diag}(\text{prox}_g(\sigma))V^\mathsf{T},$$

where we use $A = \text{diag}(a)$ to construct a diagonal matrix from a vector a.

b. Using the Schatten p-norm connections listed above, check that the result from part a reproduces the results for the trace and Frobenius norms, in Examples 7.1.d and 7.1.e.

c. Now use the result from part a to derive the proximal map of the matrix rank function, from Example 7.1.g.

7.4 For $f(x) = \inf_{z \in C} \|x - z\|_2 = \|x - P_C(x)\|_2$, where C is a closed convex set, prove that

$$\text{prox}_{\lambda f}(x) = \begin{cases} x - \lambda \dfrac{x - P_C(x)}{\|x - P_C(x)\|_2} & \|x - P_C(x)\|_2 > \lambda \\ P_C(x) & \|x - P_C(x)\|_2 \leq \lambda \end{cases}.$$

Hint: recall that f is differentiable at each $x \notin C$, with gradient given in (6.5).

7.5 Prove the proximal quadratic sum rule in Property 7.2.D.

7.6 Prove the proximal decomposition fact (7.12) for positively homogeneous f. Hint: show that positive homogeneity implies that $\text{prox}_{tf}(tx) = t\,\text{prox}_f(x)$, for any $t > 0$ and any x; use this together with the quadratic sum rule from Property 7.2.D to prove the desired result.

7.7 In this exercise, we will establish a sufficient condition for the proximal decomposition given in (7.11), following the development of Theorem 1 and Proposition 2 in [**Yu13**]. Let f, g be closed and convex.

a. First show that, for any x, the point $\text{prox}_f(\text{prox}_g(x))$ satisfies

$$0 \in \text{prox}_f(\text{prox}_g(x)) - x + \partial g(\text{prox}_g(x)) + \partial f\big(\text{prox}_f(\text{prox}_g(x))\big).$$

Hint: use the proximal subgradient characterization (7.5) for $\text{prox}_f(\text{prox}_g(x))$, as well as for $\text{prox}_g(x)$, and add these two expressions together.

b. Assume that

$$\partial g(x) \subseteq \partial g(\text{prox}_f(x)), \quad \text{for all } x. \tag{7.33}$$

Show that, for any x, the point $\text{prox}_f(\text{prox}_g(x))$ must then satisfy

$$0 \in \text{prox}_f(\text{prox}_g(x)) - x + \partial g\big(\text{prox}_f(\text{prox}_g(x))\big) + \partial f\big(\text{prox}_f(\text{prox}_g(x))\big).$$

Hint: use part a. Conclude using the subgradient characterization for $\text{prox}_{f+g}(x)$ and single-valuedness of the proximal map that $\text{prox}_{f+g}(x) = \text{prox}_f(\text{prox}_g(x))$, and as x was arbitrary, $\text{prox}_{f+g} = \text{prox}_f \circ \text{prox}_g$.

c. Let $g = \sum_{i=1}^k g_i$, with each g_i closed and convex, and $\cap_{i=1}^k \text{relint}(\text{dom}(g_i)) \neq \emptyset$. Prove that (7.33) is implied by

$$\partial g_i(x) \subseteq \partial g_i(\text{prox}_f(x)), \quad \text{for all } x, \text{ and } i = 1, \ldots, k, \tag{7.34}$$

and thus the latter is a sufficient condition for $\text{prox}_{f+g} = \text{prox}_f \circ \text{prox}_g$.

7.8 Now we follow the development of Corollary 4 in [**Yu13**]. Let g be a polyhedral function,

$$g(x) = \max_{i=1,\ldots,k} a_i^T x.$$

a. Show that the sufficient condition in (7.34), from Exercise 7.7 part c, is equivalent to

$$\text{prox}_g(K_i) \subseteq K_i, \quad i = 1, \ldots, k,$$

where each $K_i = \{x : a_i^T x = g(x)\}$ is a polyhedral cone. Hint: use the subgradient rule for a pointwise maximum from Property 6.2.C.

b. Let f be permutation invariant, which means that $f(x) = f(x_\pi)$, for all x and permutations π (where we denote $x_\pi = (x_{\pi_1}, \ldots, x_{\pi_d})$ for $\pi = (\pi_1, \ldots, \pi_d)$). Show that prox_f is then order-preserving,

$$x_i \geq x_j \implies [\text{prox}_f(x)]_i \geq [\text{prox}_f(x)]_j, \quad \text{for all } x \in \mathbb{R}^d \text{ and } i, j.$$

c. Let $\|x\|_{\mathrm{TV}} = \sum_{(i,j)\in E} |x_i - x_j|$ be a generalized TV seminorm, defined with respect to an edge set E (notice that this generalizes (7.9), which corresponds to $E = \{(i, i+1) : i = 1, \ldots, d-1\}$. Prove that (7.13) holds for any permutation invariant f and any edge set E. Hint: check that the condition in part a holds, for the appropriate instantiation of cones K_i, $i = 1, \ldots, k$; and for this, use the order-preserving property from part b.

7.9 We examine linear composition rules for proximal mappings.

a. Prove the rule in Property 7.2.F for orthogonal A (that is, $A^\mathsf{T}A = AA^\mathsf{T} = I$).

b. Prove that more generally, if A has orthonormal rows (and it need not be square), and $b \in \mathbb{R}^d$ is arbitrary, then $F(x) = f(Ax + b)$ has proximal mapping

$$\mathrm{prox}_F(x) = x - A^\mathsf{T}\big(Ax + b - \mathrm{prox}_f(Ax + b)\big).$$

7.10 In this exercise, we prove that the fixed-point equation (7.15) is necessary and sufficient for minimizing $F = f + g$, where f, g are closed and convex and f is differentiable. Let $\lambda > 0$ be arbitrary.

a. Show that x minimizes F if and only if

$$x - \lambda\nabla f(x) \in x + \lambda\partial g(x).$$

b. Show that the above display is equivalent to (7.15) using the resolvent characterization $\mathrm{prox}_{\lambda g} = (I + \lambda\partial g)^{-1}$, and the fact that this inverse is single-valued.

7.11 Consider the projection problems for the probability simplex and the unit ℓ_1 ball,

$$\underset{z}{\text{minimize}} \quad \|x - z\|_2^2 \quad \text{subject to} \quad 1^\mathsf{T}x = 1,\ x \geq 0, \tag{7.35}$$

$$\underset{z}{\text{minimize}} \quad \|x - z\|_2^2 \quad \text{subject to} \quad \|x\|_1 \leq 1, \tag{7.36}$$

respectively.

a. Starting with (7.36), show that if z is a solution, then $\mathrm{sign}(z_i) = \mathrm{sign}(x_i)$, $i = 1, \ldots, d$.

b. Use part a to argue that we can reduce any ℓ_1 projection to a simplex projection; that is, in order to solve (7.36), we can first solve a problem of the form (7.35) and then we can postprocess the solution to obtain a solution to the original problem.

7.12 This exercise investigates firm nonexpansiveness (7.19) of the projection operator P_C onto a convex set C. For any x, y, let us abbreviate $p_x = P_C(x)$ and $p_y = P_C(y)$. From the variational inequality (4.10), note that

$$(x - p_x)^\mathsf{T}(p_x - y) \geq 0,$$
$$(y - p_y)^\mathsf{T}(p_y - x) \geq 0.$$

a. Use the above pair of inequalities to prove that

$$\big(x - p_x - (y - p_y)\big)^\mathsf{T}(p_x - p_y) \geq 0, \quad \text{for all } x, y.$$

Hint: expand $p_x - y = p_x - p_y + p_y - y$ in the first inequality, $p_y - x = p_y - p_x + p_x - x$ in the second inequality, then subtract the two and simplify.

b. Show that the condition from part a is equivalent to firm nonexpansiveness (7.19), and show that it also equivalent to

$$(r_x - r_y)^\mathsf{T}(x - y) \geq \|r_x - r_y\|_2^2, \quad \text{for all } x, y,$$

where we abbreviate $r_x = x - P_C(x)$ and $r_y = x - P_C(y)$.

c. Show using the Cauchy-Schwarz inequality that (7.19) implies (7.17). Show similarly using part b and the Cauchy-Schwarz inequality that (7.19) implies (7.18).

d. Show finally that firm nonexpansiveness implies

$$\|p_x - p_y\|_2^2 + \|r_x - r_y\|_2^2 \leq \|x - y\|_2^2, \quad \text{for all } x, y,$$

which gives another way of seeing that (7.19) leads to both (7.17) and (7.18).

7.13 In this exercise, we study a *contractive* property of proximal operators (and projections): for closed convex f and $\lambda > 0$, show that if $f(x^\star) = f^\star = \inf_x f(x)$, then

$$\|\operatorname{prox}_{\lambda f}(x) - x^\star\|_2 < \|x - x^\star\|_2, \quad \text{for all } x \text{ with } f(x) > f^\star.$$

Note the strict inequality. In the special case of $f = I_C$, the characteristic function of a closed convex set C, this reduces to

$$\|P_C(x) - y\|_2 < \|x - y\|_2, \quad \text{for all } x \notin C \text{ and } y \in C.$$

Hint: recall the implication (7.24) from firm nonexpansiveness, and the proximal fixed-point characterization of minimizers in (7.14).

7.14 We examine key facts about the Moreau envelope f_λ of a closed convex function f, for $\lambda > 0$.

a. Prove that f and f_λ share the same set of minimizers. Hint: consider the subgradient optimality condition, applied to the definition of f_λ.

b. Show that the subgradient calculation from part a implies (7.27).

c. Prove that if we already knew that f_λ was differentiable, then its gradient must be as in (7.28), by using the subgradient rule for a partial infimum in Property 6.2.D.

d. Now prove that f_λ is differentiable with gradient as given in (7.28). Hint: abbreviating $f_x = f_\lambda(x)$ and $z_x = \operatorname{prox}_{\lambda f}(x)$, simply verify the definition of differentiability directly:

$$\lim_{y \to x} \frac{f_y - f_x - (y - x)^\mathsf{T}(x - z_x)/\lambda}{\|y - x\|_2^2} = 0.$$

To check this, use (7.27) and direct algebra to argue that $f_y - f_x \geq (y - x)^\mathsf{T}(x - z_x)/\lambda$; by swapping the roles of x and y, argue that also $f_y - f_x \leq (y - x)^\mathsf{T}(y - z_y)/\lambda$. These two inequalities can be used in combination to verify the differentiability condition.

7.15 Let $f(x) = |x|$. Prove that its Moreau envelope f_λ is the *Huber function* (which is commonly used as a loss in robust statistics, being smooth and having linear growth away from zero):

$$f_\lambda(x) = \begin{cases} x^2/(2\lambda) & |x| \leq \lambda \\ |x| - \lambda/2 & |x| > \lambda \end{cases}.$$

This is in fact the Moreau envelope that is visualized in Figure 7.2.

7.16 In this exercise, we prove the identification theorem, Theorem 7.7, for Moreau envelopes and proximal maps. Following the development of Theorem 3.34 in [**RW09**], it helps to introduce the general concept of an *infimal convolution* of functions f, g:

$$(f \,\#\, g)(x) = \inf_z \Big\{ f(z) + g(x - z) \Big\}.$$

In this notation, observe that $f_\lambda = f \,\#\, \|\cdot\|_2^2/(2\lambda)$, that is, the Moreau envelope is the infimal convolution of f and the map $x \mapsto \|x\|_2^2/(2\lambda)$. Below, let f, g be closed and convex.

a. Assume that $f \,\#\, h = g \,\#\, h$ for a closed, convex, and coercive function h. Prove that for any value of $\lambda > 0$,

$$\inf_x \Big\{ f_\lambda(x) - v^\mathsf{T} x \Big\} = \inf_x \Big\{ g_\lambda(x) - v^\mathsf{T} x \Big\}, \quad \text{for all } v.$$

Hint: first observe that $f_\lambda \,\#\, h = g_\lambda \,\#\, h$ by the commutative property of infimal convolutions. Then use basic algebra to decompose $\inf_x \{ f_\lambda \,\#\, h(x) - v^\mathsf{T} x \}$ into two terms, one

depending only on f_λ, and the other only on h. Compare this to the expression we get when f is replaced by g.

b. Now under the same assumption $f \,\#\, h = g \,\#\, h$, show that $f = g$. Hint: if $f \neq g$ then for some x we will have (without a loss of generality) $f(x) > g(x)$, and thus $f_\lambda(x) > g_\lambda(x)$ for small enough $\lambda > 0$. Show that for $v = \nabla f_\lambda(x)$, we get

$$\inf_y \Big\{ f_\lambda(y) - v^T y \Big\} = f_\lambda(x) - v^T x > g_\lambda(x) - v^T x \geq \inf_y \Big\{ g_\lambda(y) - v^T y \Big\},$$

which would violate the result from part a.

c. Fix $\lambda > 0$. Apply the result from part b to $h(x) = \|x\|_2^2/(2\lambda)$, in order to prove part (i) in Theorem 7.7. Use the gradient relation (7.28) to prove part (ii) in the theorem.

7.17 We explore a generalization of the proximal map known as the *scaled proximal map*. We fix a positive definite matrix $H \succ 0$, and then define the scaled proximal mapping with respect to H, acting on a given function f, by

$$\text{prox}_f^H(x) = \underset{z}{\text{argmin}} \left\{ \frac{1}{2}\|x - z\|_H^2 + f(z) \right\}, \tag{7.37}$$

where $\|\cdot\|_H$ denotes the scaled Euclidean norm, defined by $\|x\|_H^2 = x^T H x$. The parts of this exercise that follow walk through several nice properties of the scaled proximal mapping (7.37), generalizing those of the (unscaled) proximal mapping (7.3). Note that the former is a special case of the latter with $H = (1/\lambda)I$, with I denoting the identity matrix. Below f is assumed to be closed and convex.

a. *Existence and uniqueness.* Prove that the minimizer in (7.37) exists and is unique.

b. *Subgradient characterization.* Prove that $z = \text{prox}_f^H(x)$ if and only if $H(x - z) \in \partial f(z)$.

c. *Steepest descent interpretation.* Argue (using heuristic arguments) that

$$\text{prox}_f^H(x) \approx x - H^{-1}\nabla f(x),$$

which we can interpret as performing a steepest descent step starting at x, which respect to the scaled norm $\|\cdot\|_H$.

d. *Resolvent characterization.* Prove that $\text{prox}_f^H = (H + \partial f)^{-1} H$, where the inverse here is interpreted as single-valued.

e. *Fixed-point equation.* Prove that x minimizes f if and only if $x = \text{prox}_f^H(x)$; furthermore, x minimizes $F = f + g$ for closed convex f, g, with f differentiable, if and only if

$$x = \text{prox}_g^H\big(x - H^{-1}\nabla f(x)\big).$$

f. *Firm nonexpansiveness.* Prove that prox_f^H is firmly nonexpansive in the $\|\cdot\|_H$ norm,

$$\big(\text{prox}_f^H(x) - \text{prox}_f^H(y)\big)^T (x - y) \geq \|\text{prox}_f^H(x) - \text{prox}_f^H(y)\|_H^2, \quad \text{for all } x, y.$$

Show that this in turn implies that it is both nonexpansive and residual nonexpansive, again in the $\|\cdot\|_H$ norm.

Convex Conjugates

8.1. Definition and properties

This chapter introduces a concept that is deeply connected to subgradients and proximal mappings (as covered in the last two chapters), and to duality theory (as will be covered next). In fact, just as with subgradients, proximal operators, and duality, the topic of the current chapter is simultaneously simple and elementary, as well as highly nontrivial and powerful.

Given a function $f : \mathbb{R}^d \to [-\infty, \infty]$, its *convex conjugate* f^* (also simply called its conjugate) is another function on $\mathbb{R}^d$ defined as

$$f^*(u) = \sup_x \left\{ u^\mathsf{T} x - f(x) \right\}. \tag{8.1}$$

The mapping from $f \mapsto f^*$ is also called the *Legendre-Fenchel transform*. At the outset, we remark that f^* is always convex, by the partial supremum rule in Property 3.4.G (the map $u \mapsto u^\mathsf{T} x - f(x)$ is affine and thus convex for each x). Moreover, the function f^* is always closed, since its epigraph can be expressed as an intersection of closed halfspaces, which is a closed set (Exercise 8.1). To be clear, these properties—the closedness and convexity of f^*—hold regardless of whether f itself is closed or convex.

A useful interpretation of f^* is as follows: at each point u, the value $f^*(u)$ is the maximum gap between a linear function with "slope" u and f. Figure 8.1 gives an illustration. This interpretation suggests that there may be some interesting geometry at play that underlies the convex conjugate, which we revisit shortly when we discuss double conjugation.

Next we describe several important properties of convex conjugates. We generally assume that $f \neq \infty$ ($\mathrm{dom}(f) \neq \emptyset$) henceforth to avoid trivialities when studying convex conjugates.

A. Fenchel's inequality. For any u, observe that by the definition of the convex conjugate (8.1) it holds that $f^*(u) \geq u^\mathsf{T} x - f(x)$, for any x. Rearranging yields what is called *Fenchel's inequality*,

$$f(x) + f^*(u) \geq x^\mathsf{T} u, \quad \text{for all } x, u. \tag{8.2}$$

Equality holds in (8.2) if and only if x achieves the supremum in (8.1). As we show next, this can be further characterized using subgradients of f.

B. Subgradient equivalences. The supremum in (8.1) is attained at x if and only if x solves

$$\underset{x}{\text{minimize}} \quad f(x) - u^\mathsf{T} x,$$

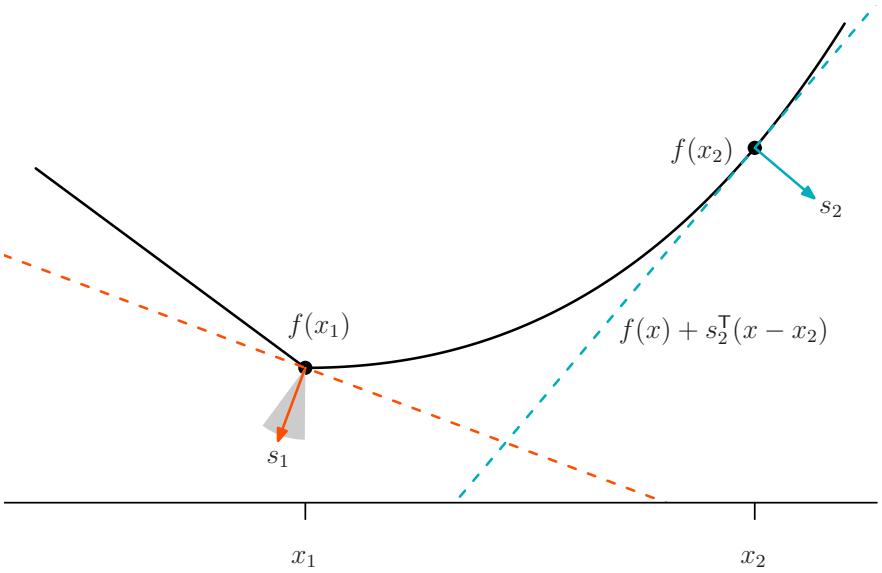

Figure 8.1. The conjugate f^* at u is the maximum gap between a linear function with slope u and $f(x)$, which is illustrated by the dotted line. The double conjugate f^{**} is the pointwise supremum of all affine minorants to f, that is, the greatest closed convex minorant to f, which is illustrated by the dashed line.

which for convex f is equivalent to $0 \in \partial f(x) - u$, that is, $u \in \partial f(x)$, using subgradient optimality. (Note that we use convexity of f to split the subdifferential of a sum into a sum of subdifferentials, by Property 6.2.B.) Meanwhile, by the rule for subgradients of a partial supremum (Property 6.2.C) we know that $\partial f^*(u)$ contains all points of the form

$$\nabla_u \big(u^T x - f(x)\big) = x,$$

such that x that achieves the supremum in (8.1). In fact, if f is *closed* and convex, then using the fact that $f^{**} = f$, to be established below, we can conclude that this describes *all* of the elements of $\partial f^*(u)$ (Exercise 8.2). The next result records these equivalences.

Theorem 8.1. *For closed and convex f with nonempty domain, the following statements are all equivalent:*

(i) x *achieves the supremum in* (8.1)*;*
(ii) $f(x) + f^*(u) = x^T u$*;*
(iii) $u \in \partial f(x)$*;*
(iv) $x \in \partial f^*(x)$*.*

C. Double conjugation. The conjugate of f^* is known as the *double conjugate* of f and denoted f^{**}. Simply applying the definition (8.1) with f^* in place of f, we get

$$f^{**}(x) = \sup_u \Big\{x^T u - f^*(u)\Big\}, \tag{8.3}$$

and by the same rationale given previously, we note that f^{**} is always closed and convex. Applying Fenchel's inequality (8.2), that is, $f^*(u) \geq x^T u - f(x)$, to the term inside the supremum gives

$$f^{**} \leq f,$$

or, in other words, the double conjugate f^{**} minorizes the original function f. In fact, the double conjugate function f^{**} is not just any (convex) minorant of f, it is the pointwise supremum of all

affine minorants of f (Exercise 8.3):

$$f^{**}(x) = \sup\{g(x) : g \text{ is affine, and } g \leq f\}, \quad \text{for all } x \in \mathrm{dom}(f). \tag{8.4}$$

See Figure 8.1 again for an illustration. Lastly, an important special reduction occurs for closed and convex f: in this case, we get

$$f^{**} = f. \tag{8.5}$$

Exercises 8.4–8.6 walk through the proof of this and related facts.

Example 8.2. Below are examples of convex conjugates for a few well-known functions of interest. The calculations for most are straightforward, and for others we defer the details to exercises.

a. For $f(x) = \frac{1}{2}x^T Q x$, where $Q \succ 0$, its conjugate is $f^*(u) = \frac{1}{2}u^T Q^{-1} u$.
b. For $f(x) = \sum_{i=1}^d x_i \log(x_i)$, its conjugate is $f^*(u) = \sum_{i=1}^d \exp(u_i - 1)$.
c. For $f = I_C$, the characteristic function of an arbitrary set C, its conjugate is
$$f^* = h_C,$$
where recall $h_C(u) = \sup_{x \in C} u^T x$ denotes the support function corresponding to C.
d. For $f = h_C$, the support function corresponding to a closed, convex, and nonempty set C, its conjugate is (Exercise 8.7):
$$f^* = I_C,$$
the characteristic function of C.
e. For $f(x) = \|x\|$, where $\|\cdot\|$ is an arbitrary norm, its conjugate is (Exercise 8.8):
$$f^* = I_{\{u \,:\, \|u\|_* \leq 1\}},$$
the characteristic function of the unit ball in the dual norm $\|\cdot\|_*$. (In Chapter 10.3, we will develop the connection between $\|\cdot\|$ and $\|\cdot\|_*$ in more detail.)

8.2. Conjugate calculus

We describe rules that will be helpful in calculating convex conjugates.

A. Scaling. It helps to first introduce some notation. For a function f and $a > 0$, we write af to denote the function defined by $(af)(x) = af(x)$, whereas we write fa for the function defined by $(fa)(x) = af(x/a)$. We refer to the operation $f \mapsto af$ as *left scaling*, and the operation $f \mapsto fa$ as *right scaling*.

Now we are ready to describe the relationship between scaling and conjugacy: for any function f and $a > 0$, it holds that $(fa)^* = af^*$. Moreover, for closed and convex f, we have $(af)^* = f^* a$. In short, for closed and convex functions, left and right scaling are dual operations under conjugacy.

B. Translation. The simplest translation rule is as follows: for any function f and $a \in \mathbb{R}$, writing $f + a$ for the function defined by $(f + a)(x) = f(x) + a$, it holds that $(f + a)^* = f^* - a$. That is, addition and subtraction by a real constant are dual to each other.

Another translation rule: if $F(x) = f(x - a)$, for any f and $a \in \mathbb{R}^d$, then $F^*(u) = f^*(u) + a^T u$, and if $F(x) = f(x) + a^T x$, then $F^*(u) = f^*(u - a)$. That is, translation of the domain by a vector a and addition by a linear function with "slope" a are dual to each other.

C. Linear composition. Similar to scaling, we first introduce some notation. For $A \in \mathbb{R}^{d\times k}$, and a function f on $\mathbb{R}^d$, we write fA to denote the function (on $\mathbb{R}^k$) defined by $(fA)(x) = f(Ax)$. Also, for a function f on $\mathbb{R}^k$, we write Af to denote the function (on $\mathbb{R}^d$) defined by

$$(Af)(y) = \inf_{Ax=y} f(x).$$

We refer to the operation $f \mapsto Af$ as *left composition*, and $f \mapsto fA$ as *right composition*.

Now we can describe the relationship between composition and conjugacy: for any function f on $\mathbb{R}^k$ and $A \in \mathbb{R}^{d\times k}$, it holds that $(Af)^* = f^*A^\mathsf{T}$. Moreover, for a closed and convex function f on $\mathbb{R}^d$, we have $(fA)^* = A^\mathsf{T} f^*$. In other words, for closed and convex functions, left and right composition are dual to each other.

D. Separable sum. If $F(x) = f_1(x_1) + f_2(x_2)$ for a block variable $x = (x_1, x_2)$, then $F^*(u) = f_1^*(u_1) + f_2^*(u_2)$ for $u = (u_1, u_2)$.

E. General sum. In order to explain what happens for a general sum of functions, which is one of the most interesting calculus rule for convex conjugates, we need to recall the notion of an *infimal convolution* of functions f, g (first introduced in Exercise 7.16): this is a function denoted $f \,\#\, g$, and defined by

$$(f \,\#\, g)(x) = \inf_z \Big\{ f(z) + g(x - z) \Big\}.$$

For any f, g, a straightforward calculation shows that $(f \,\#\, g)^* = f^* + g^*$. Meanwhile, for closed and convex f, g, we have $(f + g)^* = f^* \,\#\, g^*$. That is, for closed and convex functions, infimal convolution and addition are dual to each other. This provides an interesting perspective on the Moreau envelope (a special case of an infimal convolution), which we return to in Chapter 8.4.2.

8.3. Conjugates and smoothness*

discuss relationships between smoothness of f, f*

define Legendre function (essential smoothness, essential strict convexity). gradient map is a homeomorphism (with inverse being graident of conjugate)

Chapter 26 of Rockafellar

8.4. Proximal connections*

8.4.1. Moreau decomposition.

8.4.2. Moreau envelope, revisited.

Chapter Notes

Chapters 11 and 12 or Rockafellar masterful geometric treatment of conjugacy. Exercise 8.4 gives a glimpse of what the geometric view can provide

The connection between the moreau envelope and regularization is due to Hedy Attouch in 1977

infimal convolution and conjugates and addition

relationship between conjugates and moreau envelopes, which shows it as a form of smoothing

Exercises

8.1 Prove that the epigraph of f^* in (8.1) can be expressed as

$$\text{epi}(f^*) = \{(u, s) : u^\mathsf{T} x \leq f(x) + s, \text{ for all } x\},$$

which as an intersection of closed halfspaces, and is hence closed.

8.2 This exercise examines subgradients of the conjugate function.

a. Show that $x \in \partial f^*(u)$ for any x that achieves the supremum in (8.1). Hint: recall the rule in Property 6.2.C.

b. When f is closed and convex, show further that $x \in \partial f^*(u)$ if and only if x achieves the supremum in (8.1). Hint: first show $x \in \partial f^*(u)$ if and only if u achieves the supremum in (8.3), by subgradient optimality. Then use the fact that u achieves the supremum in (8.3) if and only if $f^{**}(x) + f^*(u) = x^\mathsf{T} u$. Lastly, reinterpret the last condition using the fact that $f^{**} = f$ for closed convex f.

8.3 We prove the fact in (8.4), for $x \in \text{dom}(f)$, by proving separately that the left-hand side is at most and at least the right-hand side.

a. Prove that $f^{**}(x) \geq \sup\{g(x) : g \text{ is affine, and } g \leq f\}$. Hint: if $g(x) = u^\mathsf{T} x + b$ satisfies $g \leq f$, then show that $f^*(u) \leq -b$, and thus $f^{**}(x) \geq u^\mathsf{T} x + b = g(x)$.

b. Prove that $f^{**}(x) \leq \sup\{g(x) : g \text{ is affine, and } g \leq f\}$. Hint: observe that $f^{**}(x)$ is the supremum of $g(x)$ over all affine minorants g that take the form $g(y) = u^\mathsf{T} y - f^*(u)$.

8.4 This exercise proves a refined version of the fact from Exercise 6.4: for closed and convex f,

$$f(x) = \sup\{g(x) : g \text{ is affine, and } g \leq f\}, \quad \text{for all } x \in \text{dom}(f). \tag{8.6}$$

Note that we really only need to prove that the equality holds for $x \in \text{dom}(f) \setminus \text{relint}(\text{dom}(f))$, as the result was already established on $\text{relint}(\text{dom}(f))$ (using the existence of subgradients) in Exercise 6.4. Nonetheless, in this exercise we will adopt a different approach to establishing (8.6) that seamlessly applies to all $x \in \text{dom}(f)$.

a. First we establish a geometric analog of (8.6). For a closed convex set C, prove that:

$$C \text{ is the intersection of all closed halfspaces } H \supseteq C. \tag{8.7}$$

Hint: this is basically the same as the fact proved in Exercise 6.1 part a. You can either translate this result appropriately in order to prove (8.7), or you can prove (8.7) directly using the strict version of the separating hyperplane theorem, from Exercise 3.8: show that the intersection of all closed halfspaces containing C must exclude any point $a \notin C$ by applying the strict separating hyperplane theorem to C and $D = \{a\}$.

b. Apply the fact from part a to the closed and convex set $C = \text{epi}(f)$ (since f is assumed to be closed and convex) to yield:

$$\text{epi}(f) \text{ is the intersection of all closed "upper" or "vertical" halfspaces } H \supseteq C,$$

where if H is a halfspace defined by the normal vector $(a, b) \in \mathbb{R}^d \times \mathbb{R}$, then we call H an "upper" halfspace when $b > 0$, and a "vertical" halfspace when $b = 0$. Hint: a "lower" halfspace, with $b < 0$, can never contain $\text{epi}(f)$.

c. Show that we can exclude vertical halfspaces from the last display:

$$\text{epi}(f) \text{ is the intersection of all closed "upper" halfspaces } H \supseteq C.$$

Hint: it is sufficient to show that for any closed vertical halfspace $V \supseteq \text{epi}(f)$, and any point $(x_0, t_0) \notin V$, there exists a closed upper halfspace $H_0 \supseteq \text{epi}(f)$ such that $(x_0, t_0) \notin$

H_0 as well. This can be constructed by as follows. Denote $V = \{(x,t) : a_1^T x \leq c_1\}$, and let $H = \{(x,t) : a_2^T x + b_2 t \leq c_2\}$ be any upper halfspace (such that $b_2 > 0$) that contains $\text{epi}(f)$. For $\lambda > 0$, define $H_0^\lambda = \{(x,t) : (\lambda a_1 + a_2)^T x + b_2 t \leq \lambda c_1 + c_2\}$. Then show that for any $\lambda > 0$, the upper halfspace H_0^λ contains $\text{epi}(f)$, while for sufficiently large $\lambda > 0$, it excludes (x_0, t_0).

d. Show that the result from part d is equivalent to (8.6).

8.5 Show that

$$f^{**}(x) = \sup\{g(x) : g \text{ is closed and convex, and } g \leq f\}, \quad \text{for all } x \in \text{dom}(f). \tag{8.8}$$

This provides the view of the double conjugate f^{**} as the greatest closed convex minorant of f. Hint: start with the representation in (8.4), then show separately that the right-hand side in (8.4) is at most and at least the right-hand side in (8.8). For the latter direction, consider applying the fact from Exercise 8.4 to each closed convex function g in the supremum.

8.6 Prove (8.5) for closed convex f. Hint: use (8.4) and (8.6).

8.7 Prove the statement in Example 8.1.d. Hint: argue that, since C is a closed and convex set, I_C is a closed and convex function, then use (8.5) and the fact from Example 8.1.c.

8.8 Prove the statement in Example 8.1.e. Hint: argue that, since $\|\cdot\|_*$ is a norm, it is a closed and convex function, then use the relation in (6.3) and the fact from Example 8.1.d.

8.9 Prove that for $p, q \geq 1$ with $1/p + 1/q = 1$, and $f(x) = \|x\|_p^p/p$, we have $f^*(u) = \|u\|_q^q/q$.

8.10

8.11 linf proximal mapping from Moreau decomposition

8.12 Operator norm proximal map from linf and Exercise 7.3

Part 4

Duality and Optimality

Duality in Linear Programs

Chapter 10

Duality in General Problems

Chapter 11

Karush-Kuhn-Tucker Conditions

Exercises

11.1 Compare these two? And how about basis pursuit?

11.2

Dual Correspondences

12.1. Conjugates and dual problems

gives general relationship between primal and dual in terms of conjugates

discuss 0 being in int(dom(f)) being equivalent to f* having no directions of recession. give nice proposition/theorem about existence of solutions of

$$\underset{\theta}{\text{minimize}} \quad g(X\theta) + h_C(x),$$

where C is a convex set. prove in exercises. specialize to generalized linear model case, interpret sufficient conditions, and prove theorems in Chapter 4.5 about existence of logistic and Poisson regression solutions.

MAKE Sure to talk about interpretation of logistic interpretation

12.2. Dual cones and polar sets*

dual in terms of gauge of polar set

Exercises

12.1

Part 5

Case Studies

Lasso

Chapter 14

Support Vector Machines

Part 6

Advanced Topics

Basic Topology

interior, closure, boundary, linear span affine span affine and line dimension relative counterparts set operators applied directly to sets lack parantheses inf, sup, min, max minkowski sum Euclidean projection

Multivariate Calculus

B.1. Derivative

B.2. Directional derivative

directional derivative derivative existence and continuity of directional derivatives imply differentiability

in fact, directional derivatives being linear iff differentiable

for convex functions, existence of partial derivatives is enough Theorem 25.2 of Rockafellar

Linear Algebra

column vectors are the default inner product, transpose column space, row space, null space row and null are orthocomplements introduce span operator, null operator pseudoinverse. eigendecompositions symmetric square root positive definite, positive semidefinite loewner ordering projectors orthocomplement

matrices can be treated as a vector space vectorization operator

eigenvalues of block matrices schur complement:

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0 \iff A - BC^{-1}B^T \succeq 0$$

for A, C symmetric and $C \succ 0$

===== Basic linear algebra facts, here $\lambda(X) = (\lambda_1(X), \ldots, \lambda_n(X))$:

$$X \in \mathbb{S}^n \implies \lambda(X) \in \mathbb{R}^n$$
$$X \in \mathbb{S}^n_+ \iff \lambda(X) \in \mathbb{R}^n_+$$
$$X \in \mathbb{S}^n_{++} \iff \lambda(X) \in \mathbb{R}^n_{++}$$

We can define an inner product over $\mathbb{S}^n$: given $X, Y \in \mathbb{S}^n$,

$$\langle X, Y \rangle = \operatorname{tr}(XY)$$

We can define a partial ordering over $\mathbb{S}^n$: given $X, Y \in \mathbb{S}^n$,

$$X \succeq Y \iff X - Y \in \mathbb{S}^n_+$$

Note: for $x, y \in \mathbb{R}^n$, $\operatorname{diag}(x) \succeq \operatorname{diag}(y) \iff x \geq y$ (recall, the latter is interpreted elementwise) =====

C.1. Singular value decomposition

singular value decomposition. rotation, stretch, rotation relationship between singular values and eigenvalues NOT true unless positive semidefinite! look at the operator norm of a nonpositive definite matrix ... can't relate it to the eigenvalues

min-max (courant-fischer-weyl) theorem https://qchu.wordpress.com/2017/03/13/singular-value-decomposition/

Bibliography

[AA84] Adelin Albert and John A. Anderson, *On the existence of maximum likelihood estimates in logistic regression models*, Biometrika **71** (1984), no. 1, 1–10.

[BBBB72] Richard E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk, *Statistical inference under order restrictions: The theory and application of isotonic regression*, Wiley, 1972.

[BC11] Heinz H. Bauschke and Patrick L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, Springer, 2011.

[Ber71] Dimitri P. Bertsekas, *Control of uncertain systems with set-membership description of the uncertainty*, Ph.D. thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, 1971.

[Ber09] Dimitri P. Bertsekas, *Convex optimization theory*, Athena Scientific, 2009.

[BGV92] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik, *A training algorithm for optimal margin classifiers*, ACM Workshop on Computational Learning Theory, 1992.

[BS18] Alvaro Barbero and Suvrit Sra, *Modular proximal optimization for multidimensional total-variation regularization*, Journal of Machine Learning Research **19** (2018), no. 56, 1–82.

[BV04] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[BvdBS+15] Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès, *Adaptive variable selection via convex optimization*, Annals of Applied Statistics **9** (2015), no. 3, 1103–1140.

[CD09] Antonin Chambolle and Jerome Darbon, *On total variation minimization and surface evolution using parametric maximum flows*, International Journal of Computer Vision **84** (2009), no. 3, 288–307.

[CDS98] Scott Chen, David L. Donoho, and Michael Saunders, *Atomic decomposition for basis pursuit*, SIAM Journal on Scientific Computing **20** (1998), no. 1, 33–61.

[CH74] David R. Cox and David V. Hinkley, *Theoretical statistics*, Chapman & Hall/CRC Press, 1974.

[Cla90] Francis H. Clarke, *Optimization and nonsmooth analysis*, Society for Industrial and Applied Mathematics, 1990.

[Con16] Laurent Condat, *Fast projection onto the simplex and the l1 ball*, Mathematical Programming **158** (2016), no. 1, 575–585.

[CP11] Antonin Chambolle and Thomas Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of Mathematical Imaging and Vision **40** (2011), no. 1, 120–145.

[CR09] Emmanuel J. Candès and Benjamin Recht, *Exact matrix completion via convex optimization*, Foundations of Computational Mathematics **9** (2009), no. 6, 717–772.

[CT07] Emmanuel J. Candès and Terence Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Annals of Statistics **35** (2007), no. 6, 2313–2351.

[CT10] ———, *The power of convex relaxation: Near-optimal matrix completion*, IEEE Transactions on Information Theory **56** (2010), no. 5, 2053–2080.

[CV95] Corinna Cortes and Vladimir N. Vapnik, *Support-vector networks*, Machine Learning **20** (1995), no. 3, 273–297.

[Dan67] John M. Danskin, *The theory of max-min and its applications to weapons allocation problems*, Springer, 1967.

[DK01] P. Laurie Davies and Arne Kovac, *Local extremes, runs, strings and multiresolution*, Annals of Statistics **29** (2001), no. 1, 1–65.

[dLHM10] Jan de Leeuw, Kurt Hornik, and Patrick Mair, *Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods*, Journal of Statistical Software **32** (2010), no. 5, 1–24.

[DSSSC08] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra, *Efficient projections onto the l1-ball for learning in high dimensions*, International Conference on Machine Learning, 2008.

[EG15] Lawrence C. Evans and Ronald F. Gariepy, *Measure theory and fine properties of functions*, CRC Press, 2015, Revised edition.

[EY36] Carl Eckart and Gale Young, *The approximation of one matrix by another of lower rank*, Psychometrika **1** (1936), no. 3, 211–218.

[FHHT07] Jerome Friedman, Trevor Hastie, Holger Hoefling, and Robert Tibshirani, *Pathwise coordinate optimization*, Annals of Applied Statistics **1** (2007), no. 2, 302–332.

[Grü03] Branko Grünbaum, *Convex polytopes*, Springer, 2003, Second edition.

[GW95] Michel X. Goemans and David P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, Journal of the Association for Computing Machinery **42** (1995), no. 6, 1115–1145.

[Hab73] Shelby J. Haberman, *Log-linear models for frequency data: Sufficient statistics and likelihood equations*, Annals of Statistics **1** (1973), no. 4, 617–632.

[HLZ08] Abderrahim Hantoute, Marco Antonio López, and Constantin Zălinescu, *Subdifferential calculus rules in convex analysis: A unifying approach via pointwise supremum functions*, SIAM Journal on Optimization **19** (2008), no. 2, 863–882.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning: Data mining, inference and prediction*, Springer, 2009, Second edition.

[HTJ15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright J., *Statistical learning with sparsity: The lasso and generalizations*, Chapman & Hall, 2015.

[HUL01] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal, *Fundamentals of convex analysis*, Springer, 2001.

[Joh13] Nicholas Johnson, *A dynamic programming algorithm for the fused lasso and l_0-segmentation*, Journal of Computational and Graphical Statistics **22** (2013), no. 2, 246–260.

[LC98] Erich L. Lehmann and George Casella, *Theory of point estimation*, Springer, 1998, Second edition.

[Mar52] Harry Markowitz, *Portfolio selection*, Journal of Finance **7** (1952), no. 1, 77–91.

[MHT10] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani, *Spectral regularization algorithms for learning large incomplete matrices*, Journal of Machine Learning Research **11** (2010), 2287–2322.

[Mir60] Leon Mirsky, *Symmetric gauge functions and unitarily invariant norms*, The Quarterly Journal of Mathematics **11** (1960), no. 1, 50–59.

[MN89] Peter McCullaugh and John A. Nelder, *Generalized linear models*, Chapman & Hall/CRC Press, 1989, Second edition.

[MvdG97] Enno Mammen and Sara van de Geer, *Locally apadtive regression splines*, Annals of Statistics **25** (1997), no. 1, 387–413.

[PB13] Neil Parikh and Stephen Boyd, *Proximal algorithms*, Foundations and Trends in Machine Learning **1** (2013), no. 3, 123–231.

[Roc66] R. Tyrrell Rockafellar, *Characterization of the subdifferentials of convex functions*, Pacific Journal of Mathematics **17** (1966), no. 3, 497–510.

[Roc70] ———, *Convex analysis*, Princeton University Press, 1970.

[Roc76] ———, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization **14** (1976), no. 5, 877–898.

[ROF92] Leonid I. Rudin, Stanley Osher, and Emad Faterni, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena **60** (1992), no. 1, 259–268.

[Ros58] Frank Rosenblatt, *The perceptron: A probabilistic model for information storage and organization in the brain*, Psychological Review **65** (1958), no. 6, 386–408.

[RW09] R. Tyrrell Rockafellar and Roger J-B Wets, *Variational analysis*, Springer, 2009, Third printing.

[RY22] Ernest K. Ryu and Wotao Yin, *Large-scale convex optimization via monotone operators*, Cambridge University Press, 2022.

[Sch07] Erhard Schmidt, *Zur theorie der linearen und nichtlinearen Integralgleichungen*, Mathematische Annalen **63** (1907), 433–476.

[Sil75] Samuel D. Silvey, *Statistical inference*, Chapman & Hall/CRC Press, 1975.

[SS02] Bernhard Scholkopf and Alexander Smola, *Learning with kernels*, The MIT Press, 2002.

[Tib96] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B **58** (1996), no. 1, 267–288.

[TSR+05] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight, *Sparsity and smoothness via the fused lasso*, Journal of the Royal Statistical Society: Series B **67** (2005), no. 1, 91–108.

[VCLR13] Vincent Q. Vu, Juhee Cho, Jing Lei, and Karl Rohe, *Fantope projection and selection: A near-optimal convex relaxation of sparse PCA*, Advances in Neural Information Processing Systems, 2013.

[Was04] Larry Wasserman, *All of statistics: A concise course in statistical inference*, Springer, 2004.

[Yu13] Yaoling Yu, *On decomposing the proximal map*, Advances in Neural Information Processing Systems, 2013.

Index