1 Introduction and Preliminaries

Monotone operator theory is an elegant and powerful tool for analyzing first-order convex optimization methods and, as such, plays a central role in convex analysis and convex optimization theory. In this book, we use this tool to provide a unified analysis of many classical and modern convex optimization methods.

This book is organized into two parts. Part I presents analysis of convex optimization methods via monotone operators, the core content. The content of Part I has sequential dependence, so the chapters should be read in a linear order. Part II presents additional auxiliary topics. The chapters can be read independently of each other. A diagram in the preface illustrates the dependency of the chapters.

1.1 FIRST-ORDER METHODS IN THE MODERN ERA

Many convex optimization methods can be classified into first or second-order methods. First-order methods can be described and analyzed with gradients and subgradients, while second-order methods use second-order derivatives or their approximations.

In the early days of convex optimization, the 1970s through the 1990s, researchers focused primarily on second-order methods, as they were more effective in solving the relatively smaller optimization problems of the era. Within the past decade, however, the demand to solve ever-larger problems grew, and so did the popularity of first-order methods.

Second-order methods require relatively fewer iterations to solve the optimization problem to high accuracy, even up to machine precision. However, the computational cost per iteration quickly becomes expensive as the problem size grows. In contrast, first-order methods have a much lower computational cost per iteration. For some large-scale optimization problems, running even a single iteration of a second-order method is infeasible, while first-order methods can solve such problems to acceptable accuracy.

Another advantage of first-order methods is that they are extremely simple; we can usually describe the entire method with two or three lines of equations. This is a significant advantage in practice, as simpler methods are easy for practitioners to implement and try out quickly, and the simplicity tends to make efficient parallelization easier.

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

The two classes of methods are usually not in competition. When a high-accuracy solution is needed, second-order methods should be used. In large-scale problems, one should use first-order methods and tolerate inaccuracy. After all, most engineering applications require only a few digits of accuracy in their solution. If the problem size is small, one should use second-order methods since there is little reason to forgo the high accuracy.

The total cost of a method is

$$(\text{cost per iteration}) \times (\text{number of iterations}).$$

We can analyze the cost per iteration by examining the computational cost of the individual components of the method. We can analyze the number of iterations required for convergence by analyzing the *rate of convergence*.

In convex optimization, arguments advocating one method over another are often based on the cost per iteration. In fact, we just made this very argument in comparing first-order and second-order methods. However, it is important to keep in mind that these arguments are incomplete since the cost per iteration is only half of the equation, literally. A method with a low cost per iteration has the potential, not a guarantee, to be efficient.

Nevertheless, primarily focusing on the cost per iteration of a method is still a useful simplification, so we adopt it in this book. With the exception of §12 and §13, this book almost entirely focuses on establishing convergence without paying much attention to the rate of convergence. We do prove convergence rates, but the rates are discussed infrequently.

1.2 LIMITATIONS OF MONOTONE OPERATOR THEORY

One of the main goals of this book is to provide streamlined and simple convergence proofs, and we only discuss results that fit this approach. Such results are simple but often not the strongest. The strongest results in convex optimization usually involve arguments that go beyond monotone operator theory.

Proofs based on monotone operator theory use monotonicity, rather than convexity, as the key property. This line of analysis does not lead to results involving function values. For example, the gradient method $x^{k+1} = x^k - \alpha \nabla f(x^k)$ converges, under suitable assumptions, with rate $\|\nabla f(x^k)\|^2 \leq O(1/k)$ and $f(x^k) - f(x^\star) \leq O(1/k)$. We can prove the first result with properties of monotone operators, but the second result requires properties of convex functions. Also, topics such as line searching, Frank–Wolfe, and second-order methods are not explained very well with monotone operator theory. Monotone operators do play a central role, but convex optimization theory does go beyond monotone operators.

1.3 PRELIMINARIES

In this section, we quickly review preliminary topics. We simply state, without proof, many of the results based on convex analysis and refer interested readers to standard

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

references such as [Roc70d, Roc74, HL93, HL01, BV04, Nes04, BL06, NP06, Ber09, BV10, BC17a].

1.3.1 Sets

A set is empty when it contains no element. Let $\emptyset$ denote the empty set. When a set contains one element, we say it is a *singleton*.

A set S is *convex* if $x, y \in S$ implies $\theta x + (1-\theta)y \in S$ for all $\theta \in [0,1]$. The empty set, singletons, and $\mathbb{R}^n$ are also convex sets.

In this book, we overload the standard notation defined for points to sets. In particular, when $\alpha \in \mathbb{R}$, $x \in \mathbb{R}^n$, $A, B \subseteq \mathbb{R}^n$, and $M \in \mathbb{R}^{m\times n}$, we write

$$\begin{aligned} \alpha A &= \{\alpha a \mid a \in A\} \\ x + A &= \{x + a \mid a \in A\} \\ MA &= \{Ma \mid a \in A\} \\ A + B &= \{a + b \mid a \in A,\, b \in B\}. \end{aligned}$$

These operations preserve convexity; if A and B are convex, all of these sets are convex. The sum $A + B$ is called the *Minkowski sum*.

1.3.2 Linear Algebra

Write $\mathbb{R}^n$ for the n-dimensional Euclidean space. For any $x, y \in \mathbb{R}^n$, write

$$\langle x, y\rangle = x^\mathsf{T} y = \sum_{i=1}^{n} x_i y_i$$

for the standard inner product.

Given a matrix $A \in \mathbb{R}^{m\times n}$, write $\mathcal{R}(A)$ for the range of A and $\mathcal{N}(A)$ for the nullspace of A. If $A \in \mathbb{R}^{n\times n}$, we say A is a square matrix. If $A^\mathsf{T} = A$, which implies A is square, we say A is symmetric. If A is symmetric, the eigenvalues of A are real. Write $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ respectively for the largest and smallest eigenvalues of A, when A is symmetric.

If all eigenvalues of a symmetric matrix A are nonnegative, we say A is symmetric positive semidefinite and write $A \succeq 0$. If all eigenvalues of a symmetric matrix A are strictly positive, we say A is symmetric positive definite and write $A \succ 0$. We write $A \succeq B$ and $A \succ B$ if $A - B \succeq 0$ and $A - B \succ 0$, respectively.

Given $M \succeq 0$, write $M^{1/2}$ for the matrix square root, the unique symmetric positive semidefinite matrix that satisfies $(M^{1/2})^2 = M$. If $M \succ 0$, then $M^{1/2} \succ 0$, and we write $M^{-1/2} = (M^{1/2})^{-1}$.

Consider a symmetric matrix $X \in \mathbb{R}^{(m+n)\times(m+n)}$ partitioned as

$$X = \begin{bmatrix} A & B \\ B^\mathsf{T} & C \end{bmatrix},$$

where $A = A^\mathsf{T} \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{m\times n}$, and $C = C^\mathsf{T} \in \mathbb{R}^{n\times n}$. When A is invertible, we call the matrix

$$S = C - B^\mathsf{T} A^{-1} B$$

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

the *Schur complement* of A in X. Note that $S \in \mathbb{R}^{n\times n}$ is symmetric. Given $A \succ 0$, X is positive (semi)definite if and only if S is positive (semi)definite. Likewise, when C is invertible,

$$T = A - BC^{-1}B^\intercal$$

is the Schur complement of C in X. Given $C \succ 0$, X is positive (semi)definite if and only if T is positive (semi)definite. We use the Schur complement to assess whether a symmetric matrix is positive (semi)definite.

The 2-norm or the Euclidean norm is

$$\|x\| = \|x\|_2 = \sqrt{\langle x, x\rangle}.$$

In some cases, we will use the 1-norm and the ∞-norm respectively defined as

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \qquad \|x\|_\infty = \max_{i=1,\dots,n} |x_i|.$$

Given $A \succ 0$, define the A-norm as

$$\|x\|_A = \sqrt{x^\intercal A x}.$$

Given $A \succeq 0$, define the A-seminorm as

$$\|x\|_A = \sqrt{x^\intercal A x}.$$

Since this is a seminorm, the triangle inequality $\|x + y\|_A \le \|x\|_A + \|y\|_A$ and absolute homogeneity $\|\alpha x\|_A = |\alpha|\|x\|_A$ hold, but $\|x\|_A = 0$ is possible when $x \neq 0$.

Given a matrix $A \in \mathbb{R}^{m\times n}$, write

$$\sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^\intercal A)} = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$$

for the maximum singular value of A and

$$\sigma_{\min}(A) = \sqrt{\lambda_{\min}(A^\intercal A)} = \min_{x\neq 0} \frac{\|Ax\|}{\|x\|}$$

for the minimum singular value of A. While a real eigenvalue can be negative, all singular values are nonnegative.

We say $V \subseteq \mathbb{R}^n$ is a (linear) subspace if $0 \in V$, $x, y \in V$ implies $x + y \in V$, and $x \in V$ implies $\alpha x \in V$ for any $\alpha \in \mathbb{R}$. Under this definition, $\{0\}$ and $\mathbb{R}^n$ are also subspaces. For any $A \in \mathbb{R}^{m\times n}$, $\mathcal{R}(A)$ and $\mathcal{N}(A)$ are subspaces.

1.3.3 Analysis

For $L > 0$, we say that a mapping $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^m$ is L-Lipschitz (continuous) if

$$\|\mathbb{T}(x) - \mathbb{T}(y)\| \le L\|x - y\| \qquad \forall\, x, y \in \mathbb{R}^n.$$

We say $\mathbb{T}$ is Lipschitz (continuous) if $\mathbb{T}$ is L-Lipschitz for some unspecified $L \in (0, \infty)$. (One could say that a constant function is 0-Lipschitz, but we exclude this degenerate case from our definition, since we will later encounter quantities like $2/L$.)

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

If a mapping is Lipschitz, it is a continuous mapping. If $\mathbb{T}_1$ and $\mathbb{T}_2$ are respectively L_1- and L_2-Lipschitz, then $\mathbb{T}_1 \circ \mathbb{T}_2$ is L_1L_2-Lipschitz since

$$\|\mathbb{T}_1(\mathbb{T}_2(x)) - \mathbb{T}_1(\mathbb{T}_2(y))\| \le L_1\|\mathbb{T}_2(x) - \mathbb{T}_2(y)\| \le L_1L_2\|x - y\|.$$

If $\mathbb{T}_1$ and $\mathbb{T}_2$ are respectively L_1- and L_2-Lipschitz, then $\alpha_1\mathbb{T}_1 + \alpha_2\mathbb{T}_2$ is $(|\alpha_1|L_1 + |\alpha_2|L_2)$-Lipschitz.

A matrix $A \in \mathbb{R}^{m\times n}$ can be viewed as a mapping from x to Ax. Since

$$\|Ax\| \le \sigma_{\max}(A)\|x\|,$$

we can view A as a $\sigma_{\max}(A)$-Lipschitz mapping.

Write

$$B(x,r) = \{y \in \mathbb{R}^n \mid \|y - x\| \le r\}$$

for the closed ball of radius r centered at x. Define the interior of a set C as

$$\operatorname{int} C = \{x \in C \mid B(x,r) \subseteq C \text{ for some } r > 0\}.$$

Denote the closure of a set C as $\operatorname{cl} C$. Define the boundary of C as $\operatorname{cl} C \backslash \operatorname{int} C$.

An affine set A can be expressed as

$$A = x_0 + V,$$

where $x_0 \in \mathbb{R}^n$ and $V \subseteq \mathbb{R}^n$ is a subspace. The affine hull of C is defined as

$$\operatorname{aff} C = \{\theta_1x_1 + \cdots + \theta_kx_k \mid x_1, \ldots, x_k \in C,\ \theta_1 + \cdots + \theta_k = 1,\ k \ge 1\}.$$

The affine hull is the smallest affine set containing C; if $C \subseteq A$ and A is affine, then $\operatorname{aff}\operatorname{cl} C \subseteq A$.

Define the relative interior of a set C as

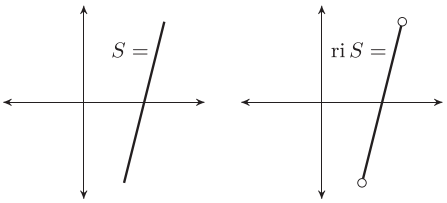
$$\operatorname{ri} C = \{x \in C \mid B(x,r) \cap \operatorname{aff} C \subseteq C \text{ for some } r > 0\}.$$

The relative interior of a nonempty convex set is nonempty. Under this definition, the relative interior of a singleton is the singleton itself. Define the relative boundary of C as $\operatorname{cl} C \backslash \operatorname{ri} C$. When we are dealing with low-dimensional sets placed in higher-dimensional spaces, the notion of relative interior is useful.

Example 1.1 Consider the line segment

$$S = \left\{(x,y) \in \mathbb{R}^2 \mid x \in [0.5, 1],\ y = 4x - 3\right\}.$$

The relative interior is the line segment with the end points excluded.



---

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

Define the distance of a point $x \in \mathbb{R}^n$ to a nonempty set $X \subseteq \mathbb{R}^n$ as

$$\operatorname{dist}(x, X) = \inf_{z \in X} \|z - x\|.$$

When X is nonempty and closed, the infimum is attained and $\operatorname{dist}(x, X) = 0$ if and only if $x \in X$. For notational convenience, write $\operatorname{dist}^2(x, X) = (\operatorname{dist}(x, X))^2$.

1.3.4 Functions

An *extended real-valued* function is a function that maps to the extended real line, $\mathbb{R} \cup \{\pm\infty\}$. Unless otherwise specified, functions in this book are extended real-valued. Write

$$\operatorname{dom} f = \{x \in \mathbb{R}^n \mid f(x) < \infty\}$$

for the (effective) domain of f. We use $\leq$, $<$, $\geq$, and $>$ for elements of the extended real line in the obvious way; for any finite α, we have $-\infty < \alpha < \infty$. We allow $\infty \leq \infty$ and $-\infty \leq -\infty$, but not $\infty < \infty$ or $-\infty < -\infty$.

A function f is *convex* if $\operatorname{dom} f$ is a convex set and

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y), \quad \forall\, x, y \in \operatorname{dom} f,\ \theta \in (0,1). \tag{1.1}$$

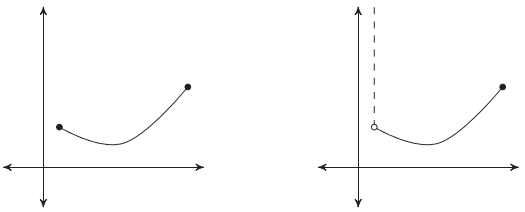
A function f is *strictly convex* if the inequality (1.1) is strict when $x \neq y$. We say f is (strictly) concave if $-f$ is (strictly) convex.

The *epigraph* of a function is defined as

$$\operatorname{epi} f = \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq \alpha\}.$$

A function f is convex if and only if $\operatorname{epi} f$ is convex. A function is *proper* if its value is never $-\infty$ and is finite somewhere. A proper function is *closed* if its epigraph is a closed set in $\mathbb{R}^{n+1}$. A proper function is closed if and only if it is lower semicontinuous. We say a function is CCP if it is closed, convex, and proper. As most convex functions of interest are closed and proper, we focus exclusively on CCP functions in this book. A function is CCP if and only if its epigraph is a nonempty closed convex set without a "vertical line," a line of the form $\{(x_0, t) \mid t \in \mathbb{R}\}$ for some $x_0 \in \mathbb{R}^n$.

Example 1.2 Whether a convex function f is closed is determined by f's behavior on the boundary of $\operatorname{dom} f$.

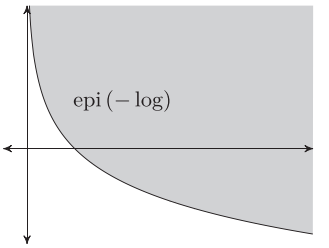


Closed convex function Convex but not closed

The dashed line denotes the function value of ∞.

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

Example 1.3 The epigraph of the CCP function $-\log$ is a nonempty closed convex set.



If f is a CCP function and $\alpha > 0$, then αf is CCP. If f and g are CCP functions and there is an x such that $f(x) + g(x) < \infty$, then $f + g$ is CCP. If f is a CCP function on $\mathbb{R}^n$, $A \in \mathbb{R}^{n\times m}$, and there is an $x \in \mathbb{R}^m$ such that $f(Ax) < \infty$, then $g(x) = f(Ax)$ is CCP.

We say $f\colon \mathbb{R}^n \to \mathbb{R}\cup\{\pm\infty\}$ is differentiable if $f\colon \mathbb{R}^n \to \mathbb{R}$ (so f is not extended real-valued), gradient $\nabla f(x) = [\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x)]^\intercal$ exists for all $x \in \mathbb{R}^n$, and

$$\lim_{h\to 0} \frac{f(x+h) - f(x) - \langle \nabla f(x), h\rangle}{\|h\|} = 0$$

for all $x \in \mathbb{R}^n$. A differentiable function f is convex if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle \qquad \forall\, x, y \in \mathbb{R}^n.$$

In other words, f is convex if its first-order Taylor expansion is a global lower bound of f. A twice continuously differentiable function f is convex if and only if $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^n$. (By the classic Schwarz's theorem, $\nabla^2 f(x) \in \mathbb{R}^{n\times n}$ is symmetric when f is twice continuously differentiable.) Intuitively speaking, $\nabla^2 f$ measures curvature, and f is convex if f is flat or has upward curvature everywhere. If f is a one-dimensional differentiable function, f is convex if and only if $f'(x)$ is monotonically nondecreasing. See the bibliographical notes for further discussion.

Write

$$\operatorname{argmin} f = \left\{ x \in \mathbb{R}^n \,\middle|\, f(x) = \inf_{z\in\mathbb{R}^n} f(z) \right\}$$

for the set of minimizers of f. When f is CCP, $\operatorname{argmin} f$ is a closed convex set, possibly empty. When f is strictly convex, $\operatorname{argmin} f$ has at most one point.

For $S \subseteq \mathbb{R}^n$, define the *indicator function*

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

If S is convex, closed, and nonempty, then δ_S is CCP.

1.3.5 Convex Optimization Problems

An *unconstrained optimization problem*

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

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

is convex if f is a convex function. We call f the *objective function*. The *constrained optimization problem*

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & x \in C \end{array}$$

is convex if f is a convex function and C is a convex set. We call $x \in C$ the *constraint*. When C is an affine set of the form $\{x \mid Ax = b\}$, we also write

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & Ax = b. \end{array}$$

In these problems, $x \in \mathbb{R}^n$ is the *optimization variable*. If a solution to an optimization problem exists, write superscript $\star$ to denote a solution. So if x is the optimization variable, $x^\star$ denotes a solution. If u is the optimization variable, $u^\star$ denotes a solution.

Indicator functions allow us to move the constraint into the objective function and treat a constrained problem as an unconstrained problem:

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

This use of indicator functions and extended value functions greatly simplifies the notation.

1.3.6 Subgradient

We say $g \in \mathbb{R}^n$ is a *subgradient* of a convex function f at x if

$$f(y) \geq f(x) + \langle g, y - x\rangle \qquad \forall\, y \in \mathbb{R}^n. \tag{1.2}$$

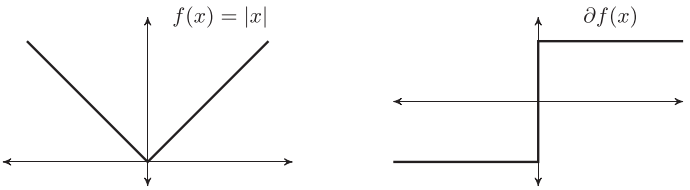
In other words, a subgradient provides an global affine lower bound of f. We call (1.2) the *subgradient inequality*. The *subdifferential* of a convex function f at x is

$$\partial f(x) = \{g \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle g, y - x\rangle,\ \forall\, y \in \mathbb{R}^n\}.$$

In other words, $\partial f(x)$ is the set of subgradients of f at x. It is straightforward to see that $\partial f(x)$ is a closed convex set, possibly empty. A convex function f is differentiable at x if and only if $\partial f(x)$ is a singleton.

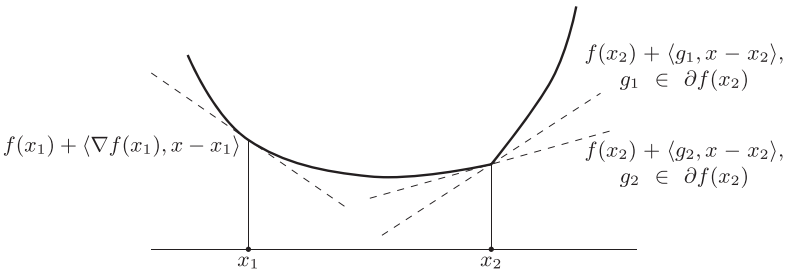
By definition, $x^\star \in \operatorname{argmin} f$ if and only if $0 \in \partial f(x^\star)$. This fact, called Fermat's rule, illustrates why subgradients are central in convex optimization.

Example 1.4 The absolute value function is differentiable everywhere except at 0.



https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

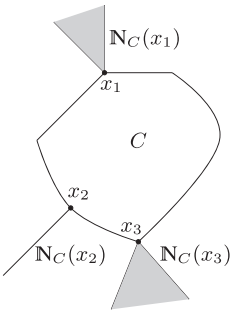
Example 1.5 At x_1 the convex function f is differentiable and $\partial f(x_1) = \{\nabla f(x_1)\}$. At x_2, f is not differentiable and has many subgradients.



Example 1.6 Let $C \subseteq \mathbb{R}^n$ be a closed convex set. Then $\partial \delta_C(x) = \mathbb{N}_C(x)$, where

$$\mathbb{N}_C(x) = \begin{cases} \emptyset & \text{if } x \notin C \\ \{y \mid \langle y, z - x\rangle \le 0 \ \forall z \in C\} & \text{if } x \in C \end{cases}$$

is the normal cone operator. For $x \in \operatorname{int} C$, $\mathbb{N}_C(x) = \{0\}$, and for $x \notin C$, $\mathbb{N}_C(x) = \emptyset$; $\mathbb{N}_C(x)$ is nontrivial only when x is on the boundary of C.



In this book, we will not pay too much attention to the meaning of $\mathbb{N}_C$. Rather, we use $\mathbb{N}_C$ as notational shorthand for $\partial \delta_C$.

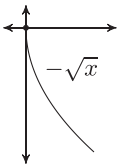
We say a convex f is subdifferentiable at x if $\partial f(x) \neq \emptyset$. When f is convex and proper, $\partial f(x) = \emptyset$ where $f(x) = \infty$. When f is convex and proper, $\partial f(x) \neq \emptyset$ for any $x \in \operatorname{ri} \operatorname{dom} f$. So a convex and proper function is not subdifferentiable outside its domain, is subdifferentiable within the relative interior of its domain, and may or may not be subdifferentiable on the relative boundary of its domain.

Example 1.7 The CCP function f defined as

$$f(x) = \begin{cases} -\sqrt{x} & \text{for } x \ge 0 \\ \infty & \text{for } x < 0 \end{cases}$$

is not subdifferentiable at $x = 0$. The slope is $-\infty$, but we do not allow infinite gradients.

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press



Several standard identities for gradients also hold for subdifferentials. Let f be CCP and $\alpha > 0$. Then

$$\partial(\alpha f)(x) = \alpha \partial f(x).$$

Let f be CCP and $\mathcal{R}(A) \cap \operatorname{ri} \operatorname{dom} f \neq \emptyset$. If $g(x) = f(Ax)$, then

$$\partial g(x) = A^\intercal \partial f(Ax). \tag{1.3}$$

Let f and g be CCP and $\operatorname{dom} f \cap \operatorname{int} \operatorname{dom} g \neq \emptyset$. Then

$$\partial(f + g)(x) = \partial f(x) + \partial g(x). \tag{1.4}$$

To clarify, $\partial f(x) + \partial g(x)$ is the Minkowski sum of the sets $\partial f(x)$ and $\partial g(x)$. Without the regularity conditions involving interiors, we can say

$$\partial g(x) \supseteq A^\intercal \partial f(Ax), \qquad \partial(f + g)(x) \supseteq \partial f(x) + \partial g(x).$$

Using the operator notation we define in §2, we can more concisely write

$$\partial \alpha f = \alpha \partial f, \qquad \partial g = A^\intercal \partial f A, \qquad \partial(f + g) = \partial f + \partial g,$$

provided the regularity conditions involving interiors hold.

1.3.7 Regularity Conditions

Say we have a mathematical statement "If P then Q". Then, if P "usually" holds, then Q "usually" holds. In this case, we say P is a *regularity condition*, since P is satisfied in the usual "regular" case. We just saw an example of this; if the regularity condition $\operatorname{dom} f \cap \operatorname{int} \operatorname{dom} g \neq \emptyset$ holds, then the identity $\partial(f + g) = \partial f + \partial g$ holds.

Statements in this book involving interiors and relative interiors can be considered regularity conditions. We keep track of these conditions, as they are necessary for a rigorous treatment of the subject. However, we do not focus on them.

1.3.8 Conjugate Function, Strong Convexity, and Smoothness

Define the conjugate function of f as

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x \rangle - f(x)\},$$

which is also known as the Fenchel conjugate or Legendre–Fenchel transform. When f is CCP, f^* is CCP and $f^{**} = f$; that is, the conjugate is CCP and the conjugate of the conjugate function is the original function. We call f^{**} the *biconjugate* of f. Note that we use the symbol $*$ for the notion of conjugate or dual, while we use the symbol $\star$ for the notion of optimality.

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

The conjugate function appears in optimization often because if f is CCP, then ∂f is an "inverse" of ∂f^* in the sense we define in §2.1. When f and f^* are both differentiable, then $(\nabla f)^{-1} = \nabla f^*$ as functions from $\mathbb{R}^n$ to $\mathbb{R}^n$.

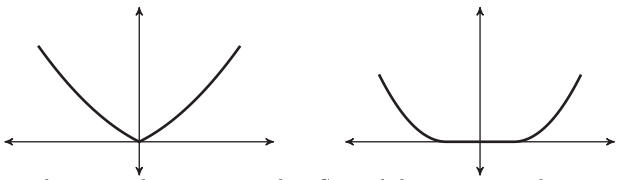
We say a CCP f is μ-strongly convex if any of the following equivalent conditions are satisfied:

- $f(x) - (\mu/2)\|x\|^2$ is convex.
- $\langle \partial f(x) - \partial f(y), x - y \rangle \geq \mu \|x - y\|^2$ for all x, y.
- $\nabla^2 f(x) \succeq \mu I$ for all x if f is twice continuously differentiable.

The second condition is written with set-valued notation; the left-hand side is a subset of $\mathbb{R}$, so the inequality means the subset lies in $[\mu\|x - y\|^2, \infty)$. In the third condition, $I \in \mathbb{R}^{n \times n}$ denotes the identity matrix.

Strongly convex CCP functions have unique minimizers. If f is μ-strongly convex and g is convex, then $f + g$ is μ-strongly convex. Informally speaking, a function is μ-strongly convex if it has upward curvature of at least μ, and we can think of nondifferentiable points to be points with infinite curvature. To clarify, strong convexity does not imply differentiability.

Example 1.8 Informally speaking, μ-strongly convex functions have upward curvature of at least μ and L-smooth convex functions have upward curvature of no more than L.



We say a CCP f is L-smooth if any of the following equivalent conditions are satisfied:

- $f(x) - (L/2)\|x\|^2$ is concave.
- f is differentiable and $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq (1/L)\|\nabla f(x) - \nabla f(y)\|^2$ for all x, y.
- f is differentiable and ∇f is L-Lipschitz.
- $\nabla^2 f(x) \preceq LI$ for all x if f is twice continuously differentiable.

(Remember, a function g is concave if $-g$ is convex.) The terminology "L-smoothness" is somewhat nonstandard; "smoothness" often means infinite differentiability in other fields of mathematics. Under our definition, L-smooth functions only need to be once-continuously differentiable.

Informally speaking, a convex function is L-strongly convex if it has upward curvature of at most L. Since non-differentiable points of convex functions can be thought of as points with infinite upward curvature, it is natural that smooth functions are differentiable.

If f is μ-strongly convex and L-smooth, then $\mu \leq L$. This follows from

$$\mu\|x - y\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \|\nabla f(x) - \nabla f(y)\| \|x - y\| \leq L\|x - y\|^2,$$

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

where we used the Cauchy–Schwartz inequality and the Lipschitz continuity of ∇f. Strong convexity and smoothness are dual properties; a CCP f is μ-strongly convex if and only if f^* is $(1/\mu)$-smooth. This follows from the fact that ∂f and ∂f^* are inverse operators, which we show in §2.1.

1.3.9 Convex Duality

In many introductory texts of convex optimization, one starts with a primal optimization problem and finds a corresponding dual problem. In this book, we take a slightly different viewpoint. We view the primal and dual problems as the two halves of a larger saddle point problem.

Let $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{\pm\infty\}$. We say $\mathbf{L}(x,u)$ is convex-concave if $\mathbf{L}$ is convex in x when u is fixed and concave in u when x is fixed. We say $(x^\star, u^\star)$ is a saddle point of $\mathbf{L}$ if

$$\mathbf{L}(x^\star, u) \le \mathbf{L}(x^\star, u^\star) \le \mathbf{L}(x, u^\star) \qquad \forall\, x \in \mathbb{R}^n,\, u \in \mathbb{R}^m.$$

We call

$$\operatorname*{minimize}_{x \in \mathbb{R}^n} \quad \sup_{u \in \mathbb{R}^m} \mathbf{L}(x,u)$$

the *primal problem* generated by $\mathbf{L}$ and write $p^\star = \inf_x \sup_u \mathbf{L}(x,u)$ for the primal optimal value. We call

$$\operatorname*{maximize}_{u \in \mathbb{R}^m} \quad \inf_{x \in \mathbb{R}^n} \mathbf{L}(x,u)$$

the *dual problem* generated by $\mathbf{L}$ and write $d^\star = \sup_u \inf_x \mathbf{L}(x,u)$ for the dual optimal value. In most engineering settings, one starts with an optimization problem, not a convex-concave saddle function. With this view of duality, the trick is to find a convex-concave saddle function that generates the primal problem of interest.

Example 1.9 Let f be a CCP function on $\mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. Consider the Lagrangian

$$\mathbf{L}(x,u) = f(x) + \langle u, Ax - b \rangle, \tag{1.5}$$

which generates the primal problem

$$\begin{array}{ll} \operatorname*{minimize}_{x \in \mathbb{R}^n} & f(x) \\ \text{subject to} & Ax = b \end{array} \tag{1.6}$$

and dual problem

$$\operatorname*{maximize}_{u \in \mathbb{R}^m} \quad -f^*(-A^\mathsf{T} u) - b^\mathsf{T} u. \tag{1.7}$$

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

The dual variable u is also called the Lagrange multipliers. If the constraint qualification

$$\{x \mid Ax = b\} \cap \operatorname{int} \operatorname{dom} f \neq \emptyset$$

holds, then $d^\star = p^\star$.

Example 1.10 Consider the Lagrangian

$$\mathbf{L}(x,u) = f(x) + \langle u, Ax\rangle - g^*(u), \tag{1.8}$$

which generates the primal problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x) + g(Ax) \tag{1.9}$$

and dual problem

$$\underset{u\in\mathbb{R}^m}{\text{maximize}} \quad -f^*(-A^\mathsf{T}u) - g^*(u). \tag{1.10}$$

If the constraint qualification

$$A\operatorname{dom} f \cap \operatorname{int} \operatorname{dom} g \neq \emptyset$$

holds, then $d^\star = p^\star$. This primal-dual problem pair is sometimes called the Fenchel–Rockafellar dual.

Weak duality, which states $d^\star \leq p^\star$, always holds. To prove this, note that for any x, u we have

$$\begin{aligned}
\inf_x \mathbf{L}(x,u) &\leq \mathbf{L}(x,u) \\
\sup_u \inf_x \mathbf{L}(x,u) &\leq \sup_u \mathbf{L}(x,u) \\
d^\star = \sup_u \inf_x \mathbf{L}(x,u) &\leq \inf_x \sup_u \mathbf{L}(x,u) = p^\star.
\end{aligned}$$

Strong duality, which states $d^\star = p^\star$, holds often but not always in convex optimization. Regularity conditions that ensure strong duality are sometimes called constraint qualifications. The constraint qualifications for strong duality are similar to the regularity conditions for subgradient identities. Again, interested readers can refer to standard references such as [Roc74, Ber09, Boţ10] for a careful discussion of this subject.

Total duality states that a primal solution exists, a dual solution exists, and strong duality holds. Total duality holds if and only if $\mathbf{L}$ has a saddle point. Solving the primal and dual optimization problems is equivalent to finding a saddle point of the saddle function generating the primal and dual problems, provided that total duality holds. We will see in §2 and §3 that total duality is the regularity condition that ensures primal-dual methods converge.

Let us prove the equivalence. Assume $\mathbf{L}$ has a saddle point $(x^\star, u^\star)$. Then

$$\begin{aligned}
\mathbf{L}(x^\star,u^\star) &= \inf_x \mathbf{L}(x,u^\star) \\
&\leq \sup_u \inf_x \mathbf{L}(x,u) = d^\star \\
&\leq \inf_x \sup_u \mathbf{L}(x,u) = p^\star \\
&\leq \sup_u \mathbf{L}(x^\star,u) = \mathbf{L}(x^\star,u^\star),
\end{aligned}$$

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

and equality holds throughout. Since $\inf_x \sup_u \mathbf{L}(x,u) = \sup_u \mathbf{L}(x^\star,u)$, $x^\star$ is a primal solution. Since $\inf_x \mathbf{L}(x,u^\star) = \sup_u \inf_x \mathbf{L}(x,u)$, $u^\star$ is a dual solution. Since $d^\star = \sup_u \inf_x \mathbf{L}(x,u) = \inf_x \sup_u \mathbf{L}(x,u) = p^\star$, strong duality holds.

On the other hand, assume total duality holds and $x^\star$ and $u^\star$ are primal and dual solutions. Then

$$\begin{aligned}\inf_x \mathbf{L}(x,u^\star) &= \sup_u \inf_x \mathbf{L}(x,u) = d^\star \\ &= \inf_x \sup_u \mathbf{L}(x,u) = p^\star \\ &= \sup_u \mathbf{L}(x^\star,u).\end{aligned}$$

Since

$$\mathbf{L}(x^\star,u^\star) \leq \sup_u \mathbf{L}(x^\star,u) = \inf_x \mathbf{L}(x,u^\star) \leq \mathbf{L}(x^\star,u^\star),$$

equality holds throughout and we conclude

$$\sup_u \mathbf{L}(x^\star,u) = \mathbf{L}(x^\star,u^\star) = \inf_x \mathbf{L}(x,u^\star),$$

that is, $(x^\star,u^\star)$ is a saddle point.

An *augmented Lagrangian* is a saddle function that has additional terms while sharing the same saddle points as its unaugmented counterpart.

Example 1.11 Consider the Lagrangian

$$\mathbf{L}(x,u) = f(x) + \langle u, Ax - b\rangle$$

with the associated primal problem

$$\begin{array}{ll}\underset{x\in\mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & Ax = b.\end{array}$$

We will often use the augmented Lagrangian

$$\mathbf{L}_\rho(x,u) = f(x) + \langle u, Ax - b\rangle + \frac{\rho}{2}\|Ax - b\|^2 \tag{1.11}$$

with $\rho > 0$. It is straightforward to show that (x,u) is a saddle point of $\mathbf{L}$ if and only if it is a saddle point of $\mathbf{L}_\rho$ for any $\rho > 0$.

Certain augmented Lagrangians arise naturally in monotone operator theory. In this book, we simply use these augmented Lagrangians without ascribing meaning to them.

1.3.10 Slater's Constraint Qualification

In the context of convex duality, regularity conditions that ensure strong duality are sometimes called *constraint qualifications*. The so-called Slater's constraint qualification is widely used, although not all constraint qualifications are due to Slater.

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

Consider the primal problem

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & f_0(x) \\ \text{subject to} & f_i(x) \le 0 \quad \text{for } i = 1,\dots,m \\ & Ax = b, \end{array}$$

where $f_0, f_1, \dots, f_m$ are CCP functions, $A \in \mathbb{R}^{p\times n}$, and $b \in \mathbb{R}^p$, generated by the Lagrangian

$$\mathbf{L}(x,\lambda,\nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \langle \nu, Ax - b\rangle - \delta_{\mathbb{R}^m_+}(\lambda),$$

where $\lambda \in \mathbb{R}^m$, $\nu \in \mathbb{R}^p$, and $\mathbb{R}^m_+ = \{(\lambda_1,\dots,\lambda_m) \mid \lambda_i \ge 0 \text{ for } i = 1,\dots,m\}$ is the nonnegative orthant.

Slater's constraint qualification states that if there exists an x such that

$$x \in \operatorname{ri} \bigcap_{i=0}^{m} \operatorname{dom} f_i, \quad f_i(x) < 0 \quad \text{for } i = 1,\dots,m, \quad Ax = b,$$

then strong duality holds (i.e., $d^\star = p^\star$), and if, furthermore, the optimal values are finite (i.e., $d^\star = p^\star > -\infty$), then a dual solution exists.

1.3.11 Proximal Operators

Let f be a CCP function on $\mathbb{R}^n$. Let $\alpha > 0$. We define the proximal operator with respect to αf as

$$\operatorname{Prox}_{\alpha f}(y) = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \alpha f(x) + \frac{1}{2}\|x - y\|^2 \right\}.$$

When $\alpha = 1$, we write Prox_f. If f is CCP, then $\operatorname{Prox}_{\alpha f}$ is well defined, that is, the argmin uniquely exists.

Let us prove the well-definedness of $\operatorname{Prox}_{\alpha f}$. Let $x_0 \in \operatorname{ri}\operatorname{dom} f$ and $g \in \partial f(x_0)$. (A CCP f has a nonempty domain, which is convex, the relative interior of a nonempty convex set is nonempty, and a CCP function is subdifferentiable on the relative interior of its domain.) Then, $f(x) \ge f(x_0) + \langle g, x - x_0\rangle$, and

$$\underbrace{\alpha f(x) + \frac{1}{2}\|x - y\|^2}_{=\tilde{f}(x)} \ge \underbrace{\alpha f(x_0) + \alpha\langle g, x - x_0\rangle + \frac{1}{2}\|x - y\|^2}_{=h(x)}.$$

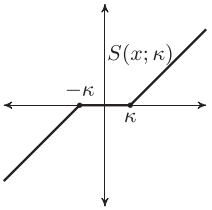
Since $\lim_{\|x\|\to\infty} h(x) = \infty$ and $\tilde{f} \ge h$, we have $\lim_{\|x\|\to\infty} \tilde{f}(x) = \infty$. Therefore, $\tilde{f}(x^k) \to \inf_x \tilde{f}(x)$ implies $x^0, x^1, \dots$ is bounded. For any convergent subsequence $x^{k_j} \to \bar{x}$, lower semicontinuity of $\tilde{f}$ implies $\tilde{f}(\bar{x}) \le \inf_x \tilde{f}(x)$. Thus $\tilde{f}(\bar{x}) = \inf_x \tilde{f}(x)$, that is, a solution exists. Finally, $\tilde{f}$ is strictly convex, so the minimizer is unique.

Example 1.12 The *soft-thresholding operator* $S(x;\kappa)$ for $x \in \mathbb{R}^n$ and $\kappa \ge 0$ is defined by

$$(S(x;\kappa))_i = \begin{cases} x_i - \kappa & \text{for } \kappa < x_i \\ 0 & \text{for } -\kappa \le x_i \le \kappa \\ x_i + \kappa & \text{for } x_i < -\kappa \end{cases}$$

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

for $i = 1, \ldots, n$. This is the proximal operator with respect to ℓ_1 norm, that is, $S(x;\kappa) = \mathrm{Prox}_{\kappa\|\cdot\|_1}(x)$.



Example 1.13 Let C be a nonempty closed convex set. Define the projection onto C as

$$\Pi_C(y) = \operatorname*{argmin}_{x \in C} \|x - y\|.$$

It is straightforward to check that $\mathrm{Prox}_{\alpha\delta_C} = \mathrm{Prox}_{\delta_C} = \Pi_C$ for any $\alpha > 0$. In this sense, proximal operators generalize projections.

In general, evaluating a proximal operator is an optimization problem itself. For many interesting convex functions, however, the proximal operator has a closed-form solution and, if so, is suitable to use as a subroutine. We loosely say a function is *proximable* if its proximal operator is computationally efficient to evaluate. Several references such as [CP11b], [PB14b, Section 6], [BSS16, Section 3], and website [CCCP] catalog a list of proximable functions.

The field of monotone operator and splitting methods revolve around the idea of decomposing a given optimization problem (which is presumably not simple as a whole) into smaller, simpler pieces and operating on them separately. These simple pieces are functions for which we can easily evaluate the gradient or the proximal operators.

1.3.12 Asymptotic Notation

Write $f(x_1, \ldots, x_r) = O(g(x_1, \ldots, x_r))$ if

$$\limsup_{x_1,\ldots,x_r \to \infty} \left| \frac{f(x_1, \ldots, x_r)}{g(x_1, \ldots, x_r)} \right| < \infty.$$

We call this the *O-notation* (and read it as "big O notation"). For example,

$$6n^2m + n^{3/2}m = O(n^2m).$$

Write $f(x_1, \ldots, x_r) = o(g(x_1, \ldots, x_r))$ if

$$\limsup_{x_1,\ldots,x_r \to \infty} \left| \frac{f(x_1, \ldots, x_r)}{g(x_1, \ldots, x_r)} \right| = 0.$$

We call this the *o-notation* (and read it as "little o notation"). For example,

$$\frac{1}{k \log k} = o(1/k).$$

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

Write $f(x_1,\dots,x_r) \sim g(x_1,\dots,x_r)$ if

$$\limsup_{x_1,\dots,x_r\to\infty} \frac{f(x_1,\dots,x_r)}{g(x_1,\dots,x_r)} = 1$$

and say f and g are *asymptotically equivalent*. For example,

$$2n^2m^3 + 3nm^3 \sim 2n^2m^3.$$

These are examples of *asymptotic notation*. Asymptotic notation is useful for identifying the limiting behavior of a function as the inputs tend toward a regime of interest. When discussing the convergence of methods, often the regime of interest is $k \to \infty$, where k is the iteration count, as we wish to know how the method eventually behaves. When discussing problem sizes, the regime of interest is $m,n \to \infty$, where m and n describe the problem size, because a method is judged by how well it can solve large (difficult) problems rather than small (easy) problems. That is not to say that non-asymptotic information is irrelevant. Sometimes we should ask at what iteration count or at what problem size the behavior described by the asymptotic notation becomes visible. Nevertheless, the asymptotic notation is a useful simplification.

BIBLIOGRAPHICAL NOTES

The 10-page lecture notes on subgradients by Boyd, Duchi, and Vandenberghe [BDV18] is a great resource to learn more about subgradients. Chapter 23 of Rockafellar's textbook [Roc70d] is another great resource providing a careful convex analytical treatment of subgradients.

The use of the conjugate function in convex analysis was pioneered by Fenchel in his unpublished lecture notes that were later distributed in mimeographed form [Fen53]. In particular, the result that $f = f^{**}$ when f is CCP is called the Fenchel–Moreau theorem and was first presented in [Fen49] and [Fen53, Theorem 37].

In careful treatments of calculus and analysis, the existence of partial derivatives, differentibility, and continuous differentiability are carefully distinguished. For convex functions, however, these notions coincide. By [Roc70d, Theorem 25.2], if f is a convex function and $x \in \mathbb{R}^n$ is a point such that $f(x) < \infty$, then f is differentiable at x if and only if

$$\frac{\partial f}{\partial x_i}(x) = \lim_{h\to 0} \frac{f(x + he_i)}{h}$$

exists and is finite for all $i = 1,\dots,n$ (where e_i is the ith unit vector and the limit is two-sided). By [Roc70d, Corollary 25.5.1], if $f\colon \mathbb{R} \to \mathbb{R}$ is convex and differentiable, then f is necessarily *continuously* differentiable, that is, when f is convex, existence of $\nabla f(x)$ for all $x \in \mathbb{R}^n$ implies $\nabla f(x)$ is continuous.

Showing that the equivalent definitions for strong convexity and smoothness are indeed equivalent is a relatively straightforward exercise in vector calculus, when the function is twice continuously differentiable. Proofs in the general case can be found in references

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

such as [Nes04]. The equivalence of the smoothness definitions is called the Baillon–Haddad theorem [BH77, Corollaire 10] [BC10].

There are multiple related but distinct viewpoints of convex duality. The view that primal-dual problem pairs are two halves of a larger saddle-point problem was developed in the mid 1960s by Dantzig, Eisenberg, and Cottle [DEC65], Stoer [Sto63, Sto64], and Mangasarian and Ponstein [MP65]. The presentation of this book closely follows Rockafellar's 1974 book [Roc74]. This 74-page book is still one of the best references on convex duality. Regularity conditions that ensure strong duality in optimization is an area with a large body of research. Slater's constraint qualification, the most widely used such condition, dates back to 1950 [Sla50]. Rockafellar's book [Roc74] provides a thorough discussion on this subject.

To expand on the discussion of §1.2, one can, in fact, establish an improved rate $\|\nabla f(x^k)\|^2 \leq O(1/k^2)$ for the gradient method using properties of convex functions [TB19, Theorem 3]; but this result cannot be established using only properties of monotone operators.

EXERCISES

1.1 Assume $\mathbb{T}_1 : \mathbb{R}^n \to \mathbb{R}^m$ is L_1-Lipschitz and $\mathbb{T}_2 : \mathbb{R}^n \to \mathbb{R}^m$ is L_2-Lipschitz. Show that $\alpha_1\mathbb{T}_1 + \alpha_2\mathbb{T}_2$ is $(|\alpha_1|L_1 + |\alpha_2|L_2)$-Lipschitz.

1.2 Let f be a convex function on $\mathbb{R}^n$. Show that $\partial f(x)$ is a closed convex set for all $x \in \mathbb{R}^n$.
Hint. Write $\partial f(x)$ as an intersection of closed half-spaces.
Remark. Remember that $\partial f(x)$ can be empty, but the empty set is a closed convex set.

1.3 Show that if f is a CCP function on $\mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, and $g(x) = f(Ax)$, then

$$\partial g(x) \supseteq A^\mathsf{T}\partial f(Ax)$$

for all $x \in \mathbb{R}^n$. Also show that if f and g are CCP functions on $\mathbb{R}^n$, then

$$\partial(f+g)(x) \supseteq \partial f(x) + \partial g(x)$$

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

1.4 Consider the function $f\colon \mathbb{R}^2 \to \mathbb{R} \cup \{\pm\infty\}$ defined as

$$f(x,y) = \begin{cases} x^2/y & \text{for } y > 0, \\ 0 & \text{for } x = y = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Clearly f is proper, and it is possible to show that f is convex. Show that
(a) f is closed, and
(b) $f|_{\operatorname{dom} f} : \operatorname{dom} f \to \mathbb{R}$ is not continuous at $(0,0)$, that is, show that f restricted to where it is finite is not continuous at $(0,0)$.
Remark. This example demonstrates that a CCP function need not be continuous on its domain. In convex optimization, lower semi-continuity, not continuity, is the regularity condition of interest. However, a proper convex function is continuous on the relative interior of its domain.

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

1.5 *Existence of a minimizer with Slater.* Let f be a CCP function on $\mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$. Assume $\mathcal{R}(A^\intercal) \cap \operatorname{ri}\operatorname{dom} f^* \neq \emptyset$. Consider the optimization problem

$$\begin{array}{ll}\underset{\mu\in\mathbb{R}^m,\,\nu\in\mathbb{R}^n}{\text{minimize}} & f^*(\nu) - \mu^\intercal y + \frac{1}{2}\|\mu\|^2 \\ \text{subject to} & A^\intercal\mu - \nu = 0\end{array}$$

generated by the Lagrangian

$$\mathbf{L}(\mu,\nu,x) = f^*(\nu) - \mu^\intercal y + \frac{1}{2}\|\mu\|^2 + \langle x, A^\intercal\mu - \nu\rangle.$$

Using Slater's constraint qualification, show

$$\operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{f(x) + (1/2)\|Ax - y\|^2\right\} \neq \emptyset.$$

1.6 *Saddle points of augmented Lagrangians.* Let f be a CCP function on $\mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. Show that the Lagrangian

$$\mathbf{L}(x,u) = f(x) + \langle u, Ax - b\rangle$$

and the augmented Lagrangian

$$\mathbf{L}_\alpha(x,u) = f(x) + \langle u, Ax - b\rangle + \frac{\alpha}{2}\|Ax - b\|^2,$$

where $\alpha > 0$, share the same set of saddle points.

1.7 Assume that a CCP function $f\colon \mathbb{R}^n \to \mathbb{R}\cup\{\pm\infty\}$ is proximable. Define $g\colon \mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}\cup\{\pm\infty\}$ as

$$g(x_1,x_2) = f(x_1 + x_2).$$

Show that

$$\operatorname{Prox}_g(x_1,x_2) = \frac{1}{2}\begin{bmatrix} x_1 - x_2 + \operatorname{Prox}_{2f}(x_1 + x_2) \\ x_2 - x_1 + \operatorname{Prox}_{2f}(x_1 + x_2)\end{bmatrix}.$$

Likewise, show that if

$$h(x_1,x_2) = f(x_1 - x_2),$$

then

$$\operatorname{Prox}_h(x_1,x_2) = \frac{1}{2}\begin{bmatrix} x_1 + x_2 + \operatorname{Prox}_{2f}(x_1 - x_2) \\ x_1 + x_2 + \operatorname{Prox}_{2f}(x_1 - x_2)\end{bmatrix}.$$

Hint. Note that $g = f\circ\begin{bmatrix} I & I\end{bmatrix}$ and show that $(y_1,y_2) = \operatorname{Prox}_g(x_1,x_2)$ if and only if there exists a $v \in \partial f(y_1 + y_2)$ such that

$$\begin{aligned} 0 &= v + (y_1 - x_1) \\ 0 &= v + (y_2 - x_2).\end{aligned}$$

1.8 Assume a CCP function $f\colon \mathbb{R}^n \to \mathbb{R}\cup\{\pm\infty\}$ is proximable. Assume $a = (a_1,\ldots,a_m) \in \mathbb{R}^m$ satisfies $a \neq 0$. Define $g\colon \mathbb{R}^{mn} \to \mathbb{R}\cup\{\pm\infty\}$ as

$$g(x_1,\ldots,x_m) = f(a_1x_1 + \cdots + a_mx_m).$$

Show that

$$v = \frac{1}{\|a\|^2}\left(a_1x_1 + \cdots + a_mx_m - \operatorname{Prox}_{\|a\|^2 f}(a_1x_1 + \cdots + a_mx_m)\right)$$

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press

$$\operatorname{Prox}_g(x_1,\dots,x_m) = \begin{bmatrix} x_1 - a_1 v \\ \vdots \\ x_m - a_m v \end{bmatrix}.$$

1.9 *Basic normal cone example.* Let $\mathbb{R}^n_+ = \{(x_1,\dots,x_n) \,|\, x_i \geq 0 \text{ for } i = 1,\dots,n\}$ be the nonnegative orthant.

(i) Characterize $\mathbb{N}_{\mathbb{R}^n_+}$, that is, describe the set $\mathbb{N}_{\mathbb{R}^n_+}(x)$ for all $x \in \mathbb{R}^n$.

(ii) Let $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be CCP and differentiable. Directly show, without using the subgradient identity $\partial(f+g) = \partial f + \partial g$, that x solves

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & x \geq 0 \end{array}$$

if and only if $-\nabla f(x) \in \mathbb{N}_{\mathbb{R}^n_+}(x)$.

1.10 *Linear programming duality.* Consider the convex–concave saddle function

$$\mathbf{L}(x,\nu,\mu) = \langle c,x\rangle + \langle Ax+b,\nu\rangle - \langle x,\mu\rangle - \delta_{\mathbb{R}^m_+}(\nu) - \delta_{\mathbb{R}^n_+}(\mu),$$

convex in $x \in \mathbb{R}^n$ and concave in $(\nu,\mu) \in \mathbb{R}^m \times \mathbb{R}^n$. Here, $\mathbb{R}^m_+$ and $\mathbb{R}^n_+$ denote the m and n-dimensional nonnegative orthants. Remember that δ_C denotes the indicator function with respect to the set C.

Show that the saddle function $\mathbf{L}$ generates the primal problem

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & c^\intercal x \\ \text{subject to} & Ax + b \leq 0 \\ & x \geq 0. \end{array}$$

Here, the inequalities denote element-wise nonnegativity. Show that $\mathbf{L}$ generates a dual problem that is equivalent to

$$\begin{array}{ll} \underset{\nu\in\mathbb{R}^m}{\text{maximize}} & b^\intercal \nu \\ \text{subject to} & c + A^\intercal \nu \geq 0 \\ & \nu \geq 0. \end{array}$$

https://doi.org/10.1017/9781009160865.002 Published online by Cambridge University Press