Cambridge University Press & Assessment
978-1-009-16085-8 — Large-Scale Convex Optimization
Ernest K. Ryu , Wotao Yin
Excerpt
More Information

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.

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 $\|\nabla f(x^k)\|^2 \le \mathcal{O}(1/k)$ and $f(x^k) - f(x^\star) \le \mathcal{O}(1/k)$. 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 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

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

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

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

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 $A^\intercal = A$, 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

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

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

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

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 $x \ne 0$.

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

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

for the maximum singular value of A and

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

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 $2/L$.)

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

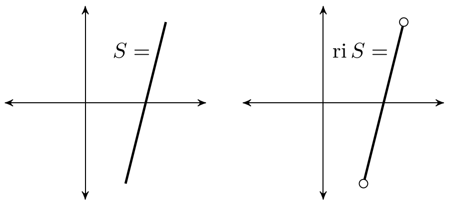
$$\operatorname{ri} C = \{x \in C \mid B(x, r) \cap \operatorname{aff} C \subseteq C \text{ for some } r > 0\}.$$

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.



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

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

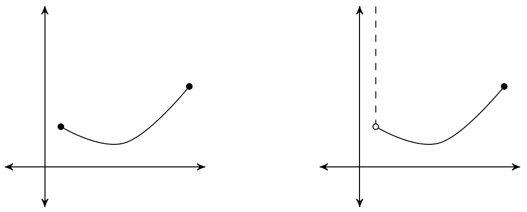
A function f is *strictly convex* if the inequality (1.1) is strict when $x \neq y$. We say f is (strictly) concave if $-f$ is (strictly) convex.

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 $\operatorname{dom} f$.

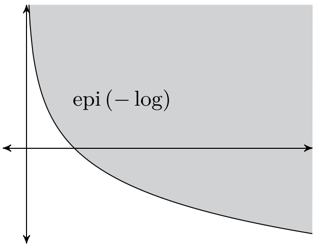


Closed convex function

Convex but not closed

The dashed line denotes the function value of ∞.

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 $\nabla f(x) = [\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x)]^\intercal$ 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

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

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*

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

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

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

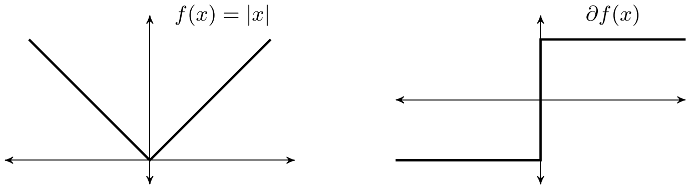
In other words, a subgradient provides an global affine lower bound of f. We call (1.2) the *subgradient inequality*. The *subdifferential* of a convex function f at x is

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

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.

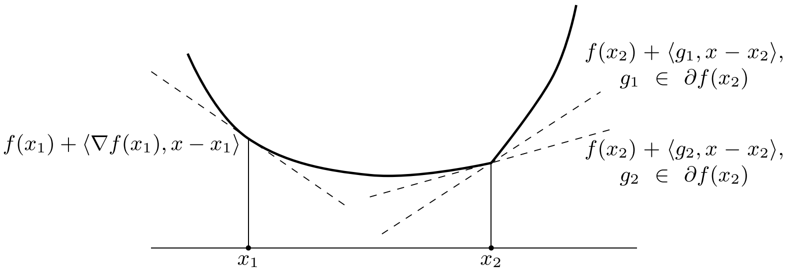
By definition, $x^\star \in \operatorname{argmin} f$ if and only if $0 \in \partial f(x^\star)$. This fact, called Fermat's rule, illustrates why subgradients are central in convex optimization.

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



Cambridge University Press & Assessment
978-1-009-16085-8 — Large-Scale Convex Optimization
Ernest K. Ryu , Wotao Yin
Excerpt
More Information

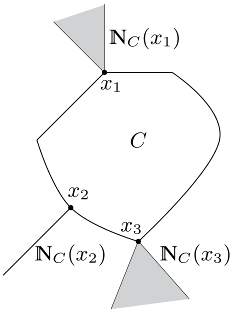
Example 1.5 At x_1 the convex function f is differentiable and $\partial f(x_1) = \{\nabla f(x_1)\}$. At x_2, f is not differentiable and has many subgradients.


$f(x_1) + \langle \nabla f(x_1), x - x_1 \rangle$
$f(x_2) + \langle g_1, x - x_2 \rangle,$
$g_1 \in \partial f(x_2)$
$f(x_2) + \langle g_2, x - x_2 \rangle,$
$g_2 \in \partial f(x_2)$
x_1
x_2

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

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

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 $\partial \delta_C$.

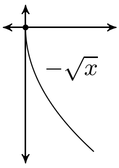
We say a convex f is subdifferentiable at x if $\partial f(x) \neq \emptyset$. When f is convex and proper, $\partial f(x) = \emptyset$ where $f(x) = \infty$. When f is convex and proper, $\partial f(x) \neq \emptyset$ for any $x \in \operatorname{ri} \operatorname{dom} f$. So a convex and proper function 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

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

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

Cambridge University Press & Assessment
978-1-009-16085-8 — Large-Scale Convex Optimization
Ernest K. Ryu , Wotao Yin
Excerpt
More Information



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.