Convex Analysis and Nonsmooth Optimization

Dmitriy Drusvyatskiy

October 22, 2020

Contents

Chapter 1

Background

This chapter sets the notation and reviews the background material that will be used throughout the rest of the book. The reader can safely skim this chapter during the first pass and refer back to it when necessary. The discussion is purposefully kept brief. The comments section at the end of the chapter lists references where a more detailed treatment may be found.

Roadmap. Sections 1.1-1.3 review basic constructs of linear algebra, including inner products, norms, linear maps and their adjoints, as well as eigenvalue and singular value decompositions. Section 1.4 establishes notation for basic set operations, such as sums and images/preimages of sets. Section 1.5 focuses on topological preliminaries; the main results are the Bolzano-Weierstrass theorem and a variant of the extreme value theorem. The final Sections 1.6-1.8 formally define first and second-order derivatives of multivariate functions, establish estimates on the error in Taylor approximations, and deduce derivative-based conditions for local optimality. The material in Sections 1.6-1.8 is often covered superficially in undergraduate courses, and therefore we provide an entirely self-contained treatment.

1.1 Inner products and linear maps

Throughout, we fix an *Euclidean space* $\mathbf{E}$, meaning that $\mathbf{E}$ is a finite-dimensional real vector space endowed with an *inner product* $\langle \cdot, \cdot \rangle$. Recall that an inner-product on $\mathbf{E}$ is an assignment $\langle \cdot, \cdot \rangle \colon \mathbf{E} \times \mathbf{E} \to \mathbf{R}$ satisfying the following three properties for all $x, y, z \in \mathbf{E}$ and scalars $a, b \in \mathbf{R}$:

(Symmetry) $\langle x, y \rangle = \langle y, x \rangle$

(Bilinearity) $\langle ax + by, z\rangle = a\langle x, z\rangle + b\langle y, z\rangle$

(Positive definiteness) $\langle x, x\rangle \geq 0$ and equality $\langle x, x\rangle = 0$ holds if and only if $x = 0$.

The most familiar example is the Euclidean space of n-dimensional column vectors $\mathbf{R}^n$, which we always equip with the *dot-product*

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

One can equivalently write $\langle x, y\rangle = x^T y$. We will denote the coordinate vectors of $\mathbf{R}^n$ by e_i and for any vector $x \in \mathbf{R}^n$, the symbol x_i will denote the i'th coordinate of x. A basic result of linear algebra shows that all Euclidean spaces $\mathbf{E}$ can be identified with $\mathbf{R}^n$ for some integer n, once an orthonormal basis is chosen. Though such a basis-specific interpretation can be useful, it is often distracting, with the indices hiding the underlying geometry. Consequently, it is often best to think coordinate-free.

The space of real $m \times n$-matrices $\mathbf{R}^{m\times n}$ furnishes another example of an Euclidean space, which we always equip with the *trace product*

$$\langle X, Y\rangle := \operatorname{tr} X^T Y.$$

Some arithmetic shows the equality $\langle X, Y\rangle = \sum_{i,j} X_{ij} Y_{ij}$. Thus the trace product on $\mathbf{R}^{m\times n}$ coincides with the usual dot-product on the matrices stretched out into long vectors. An important Euclidean subspace of $\mathbf{R}^{n\times n}$ is the space of real symmetric $n\times n$-matrices $\mathbf{S}^n$, along with the trace product $\langle X, Y\rangle := \operatorname{tr} XY$.

For any linear mapping $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$, there exists a unique linear mapping $\mathcal{A}^*\colon \mathbf{Y} \to \mathbf{E}$, called the *adjoint*, satisfying

$$\langle \mathcal{A}x, y\rangle = \langle x, \mathcal{A}^* y\rangle \qquad \text{for all points} \qquad x \in \mathbf{E},\ y \in \mathbf{Y}.$$

In the most familiar case of $\mathbf{E} = \mathbf{R}^n$ and $\mathbf{Y} = \mathbf{R}^m$, any linear map $\mathcal{A}$ can be identified with a matrix $A \in \mathbf{R}^{m\times n}$, while the adjoint $\mathcal{A}^*$ may then be identified with the transpose A^T.

Exercise 1.1. Given a collection of real $m \times n$ matrices $A_1, A_2, \ldots, A_l$, define the linear mapping $\mathcal{A}\colon \mathbf{R}^{m\times n} \to \mathbf{R}^l$ by setting

$$\mathcal{A}(X) := (\langle A_1, X\rangle, \langle A_2, X\rangle, \ldots, \langle A_l, X\rangle).$$

Show that the adjoint is the mapping $\mathcal{A}^* y = y_1 A_1 + y_2 A_2 + \ldots + y_l A_l$.

Linear mappings $\mathcal{A}\colon \mathbf{E} \to \mathbf{E}$, between a Euclidean space $\mathbf{E}$ and itself, are called *linear operators*, and are said to be *self-adjoint* if equality $\mathcal{A} = \mathcal{A}^*$ holds. Self-adjoint operators on $\mathbf{R}^n$ are precisely those operators that are representable as symmetric matrices. A self-adjoint operator $\mathcal{A}$ is *positive semi-definite*, denoted $\mathcal{A} \succeq 0$, whenever

$$\langle \mathcal{A}x, x\rangle \geq 0 \quad \text{for all } x \in \mathbf{E}.$$

Similarly, a self-adjoint operator $\mathcal{A}$ is *positive definite*, denoted $\mathcal{A} \succ 0$, whenever

$$\langle \mathcal{A}x, x\rangle > 0 \quad \text{for all } 0 \neq x \in \mathbf{E}.$$

For any two linear operators $\mathcal{A}$ and $\mathcal{B}$, we will use the notation $\mathcal{A} - \mathcal{B} \succeq 0$ to mean $\mathcal{A} \succeq \mathcal{B}$. The notation $\mathcal{A} - \mathcal{B} \succ 0$ is defined similarly.

1.2 Norms

A *norm* on a vector space $\mathcal{V}$ is a function $\|\cdot\|\colon \mathcal{V} \to \mathbf{R}$ for which the following three properties hold for all point $x, y \in \mathcal{V}$ and scalars $a \in \mathbf{R}$:

(Absolute homogeneity) $\|ax\| = |a| \cdot \|x\|$

(Triangle inequality) $\|x + y\| \leq \|x\| + \|y\|$

(Positivity) Equality $\|x\| = 0$ holds if and only if $x = 0$.

The inner product in the Euclidean space $\mathbf{E}$ always induces a norm $\|x\| = \sqrt{\langle x, x\rangle}$. Unless specified otherwise, the symbol $\|x\|$ for $x \in \mathbf{E}$ will always denote this induced norm. For example, the dot product on $\mathbf{R}^n$ induces the usual 2-norm $\|x\|_2 := \sqrt{x_1^2 + \ldots + x_n^2}$, while the trace product on $\mathbf{R}^{m\times n}$ induces the *Frobenius norm* $\|X\|_F := \sqrt{\operatorname{tr}(X^T X)}$. The Cauchy–Schwarz inequality guarantees that the induced norm satisfies the estimate:

$$|\langle x, y\rangle| \leq \|x\| \cdot \|y\| \qquad \text{for all } x, y \in \mathbf{E}. \tag{1.1}$$

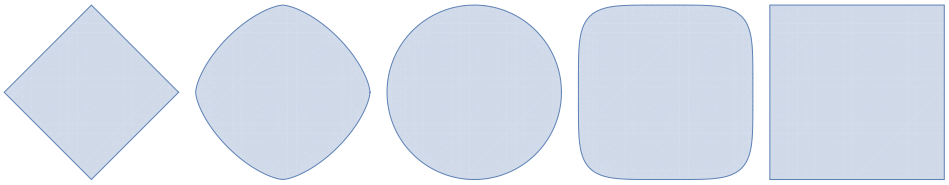
Other important examples of norms are the *l_p-norms* on $\mathbf{R}^n$:

$$\|x\|_p = \begin{cases} (|x_1|^p + \ldots + |x_n|^p)^{1/p} & \text{for } 1 \leq p < \infty \\ \max\{|x_1|, \ldots, |x_n|\} & \text{for } p = \infty \end{cases}.$$

The most notable of these are the l_1, l_2, and l_∞ norms; see Figure 1.1.

For an arbitrary norm $\|\cdot\|$ on $\mathbf{E}$, the dual norm $\|\cdot\|^*$ on $\mathbf{E}$ is defined by

$$\|v\|^* := \max\{\langle v, x\rangle : \|x\| \leq 1\}.$$



(a) $p = 1$ (b) $p = 1.5$ (c) $p = 2$ (d) $p = 5$ (e) $p = \infty$

Figure 1.1: Unit balls of ℓ_p-norms.

Thus $\|v\|^*$ is the maximal value that the linear function $x \mapsto \langle v, x\rangle$ takes over the closed unit ball of the norm $\|\cdot\|$. For example, the l_p and l_q norms on $\mathbf{R}^n$ are dual to each other whenever $p^{-1} + q^{-1} = 1$ and $p, q \in [1, \infty]$. In particular, the ℓ_2-norm on $\mathbf{R}^n$ is self-dual; the same goes for the Frobenius norm on $\mathbf{R}^{m\times n}$ (why?). More generally, it follows directly from (1.1) that the norm induced by the inner product in $\mathbf{E}$ is always self-dual. For an arbitrary norm $\|\cdot\|$ on $\mathbf{E}$, the generalized Cauchy-Schwarz inequality holds:

$$|\langle x, y\rangle| \leq \|x\| \cdot \|y\|^* \qquad \text{for all } x, y \in \mathbf{E}.$$

All norms on $\mathbf{E}$ are "equivalent" in the sense that any two are within a constant factor of each other. More precisely, for any two norms $\rho_1(\cdot)$ and $\rho_2(\cdot)$, there exist constants $\alpha, \beta > 0$ satisfying

$$\alpha\rho_1(x) \leq \rho_2(x) \leq \beta\rho_1(x) \qquad \text{for all } x \in \mathbf{E}.$$

Case in point, for any vector $x \in \mathbf{R}^n$, the relations hold:

$$\begin{aligned}\|x\|_2 \leq \|x\|_1 &\leq \sqrt{n}\|x\|_2\\ \|x\|_\infty \leq \|x\|_2 &\leq \sqrt{n}\|x\|_\infty\\ \|x\|_\infty \leq \|x\|_1 &\leq n\|x\|_\infty.\end{aligned}$$

For our purposes, the term "equivalent" is a misnomer: the proportionality constants α, β strongly depend on the (often enormous) dimension of the vector space $\mathbf{E}$. Hence measuring quantities in different norms can yield strikingly different conclusions.

1.3 Eigenvalue and singular value decompositions of matrices

The symbol $\mathbf{S}^n$ will denote the set of $n \times n$ real symmetric matrices

$$\mathbf{S}^n := \{X \in \mathbf{R}^{n\times n} : X^T = X\},$$

while $O(n)$ will denote the set of $n \times n$ real orthogonal matrices:

$$O(n) := \{X \in \mathbf{R}^{n\times n} : X^T X = XX^T = I\}.$$

A number $\lambda \in \mathbf{R}$ is an *eigenvalue* of a symmetric matrix $A \in \mathbf{S}^{n\times n}$ if there exists a vector $0 \neq v \in \mathbf{R}^n$ satisfying $Av = \lambda v$. Any such vector v is called an *eigenvector* corresponding to λ. Thus the eigenvalues of A are precisely the roots of the *characteristic polynomial*

$$\lambda \mapsto \det(A - \lambda I).$$

A central result of linear algebra shows that all n roots of this polynomial are real, when A is symmetric. We may therefore fix an ordering and denote the eigenvalues of A by

$$\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_n(A).$$

Any symmetric matrix $A \in \mathbf{S}^n$ admits an *eigenvalue decomposition*, meaning a factorization of the form

$$A = U\Lambda U^T, \tag{1.2}$$

where $U \in O(n)$ is orthogonal and $\Lambda \in \mathbf{S}^n$ is a diagonal matrix. The diagonal elements of Λ are precisely the eigenvalues of A and the columns of U are corresponding eigenvectors. A simple consequence of the decomposition (1.2) is the Rayleigh-Ritz theorem, which guarantees the relation:

$$\lambda_n(A) \leq \frac{\langle Au, u\rangle}{\langle u, u\rangle} \leq \lambda_1(A) \quad \text{for all } u \in \mathbf{R}^n \setminus \{0\}.$$

Thus the two conditions, $A \succeq 0$ and $\lambda_n(A) \geq 0$ are equivalent; similarly, $A \succ 0$ if and only $\lambda_n(A) > 0$. An important consequence of the eigenvalue decomposition (1.2) is that a matrix $A \in \mathbf{S}^n$ is positive semidefinite if and only if there exists a matrix $B \in \mathbf{S}^n$ satisfying $A = BB$ (why?). The matrix B is called the *square root* of A, and is denoted by $B = A^{1/2}$.

More generally, any rectangular matrix $A \in \mathbf{R}^{m\times n}$ admits a *singular value decomposition*, meaning a factorization of the form

$$A = U\Sigma V^T,$$

where $U \in O(m)$ and $V \in O(n)$ are orthogonal matrices and $\Sigma \in \mathbf{R}^{m\times n}$ is a diagonal matrix with nonnegative diagonal entries. The diagonal elements of Σ are uniquely defined and are called the *singular values* of A. Supposing

without loss of generality $m \leq n$, the singular values of A are precisely the square roots of the eigenvalues of AA^T, and we denote them by

$$\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_m(A) \geq 0.$$

In particular, the maximal singular-value $\sigma_1(A)$ coincides with the *operator norm* of A, defined as

$$\|A\|_{\text{op}} := \sup_{x:\|x\|\leq 1} \|Ax\|.$$

See Figure 1.2 for an illustration.

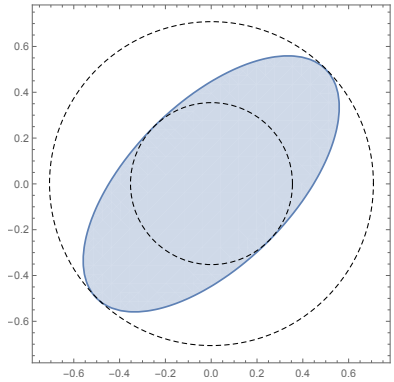


Figure 1.2: The shaded ellipse is the image of the unit disk by a nonsingular matrix $A \in \mathbf{R}^{2\times 2}$. The radii of the circumscribed and inscribed circles are $\sigma_1(A)$ and $\sigma_2(A)$, respectively.

Exercise 1.2. Given a positive definite matrix $A \in \mathbf{S}^n$, show that the assignment $\langle v, w\rangle_A := \langle Av, w\rangle$ is an inner product on $\mathbf{R}^n$, with the induced norm $\|v\|_A = \sqrt{\langle Av, v\rangle}$. Show that the dual norm with respect to the original inner product $\langle \cdot, \cdot \rangle$ is $\|v\|_A^* = \|v\|_{A^{-1}} = \sqrt{\langle A^{-1}v, v\rangle}$.
[**Hint:** Use the fact that any positive definite matrix A admits a square root.]

1.4 Set operations

In this section, we review notation for sums, generated cones, and images/preimages of sets. For any two sets $A, B \subset \mathbf{E}$ and $\lambda \in \mathbf{R}$, define the set operations:

$$\lambda A := \{\lambda a : a \in A\} \qquad \text{and} \qquad A + B := \{a + b : a \in A,\ b \in B\}.$$

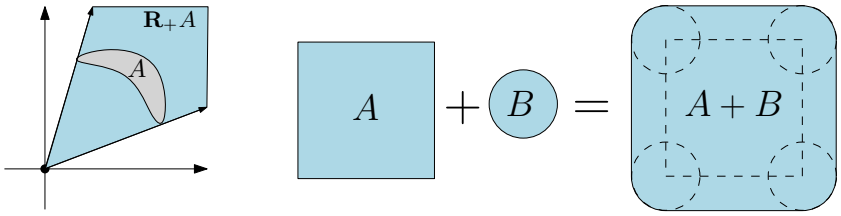
Thus the points in λA are simply the points in A scaled by λ. One can visualize the sum $A + B$ by writing it more suggestively as

$$A + B = \bigcup_{a \in A} (a + B).$$

Thus $A+B$ is formed from the union of the shifted sets $a+B$ over all points $a \in A$. In particular, forming the sum of a set $A \subset \mathbf{E}$ and a unit ball B in $\mathbf{E}$ has the affect of "fattening" A. The symbol $A - B$ is defined similarly. The cone generated by a set $A \subset \mathbf{E}$ will be denoted by

$$\mathbf{R}_+ A := \{\lambda x : x \in A, \lambda \geq 0\}.$$

See Figure 1.3 for an illustration of the generated cone and sum operation.



(a) Generated cone. (b) Disk plus square.

Figure 1.3: Sum and cone operations.

For any map $\mathcal{F}\colon \mathbf{E} \to \mathbf{Y}$ and sets $A \subset \mathbf{E}$ and $B \subset \mathbf{Y}$, define the two sets

$$\mathcal{F}A = \{\mathcal{F}(x) : x \in A\} \qquad \text{and} \qquad \mathcal{F}^{-1}B = \{x : \mathcal{F}x \in B\}.$$

The set $\mathcal{F}A$ is called the image of A under $\mathcal{F}$, while $\mathcal{F}^{-1}B$ is called the preimage of B under $\mathcal{F}$. Notice that the sum $A + B$ can also be written as the linear image of the product set $Q := A \times B$ under the map $\mathcal{F}(x, y) = x+y$.

1.5 Point-set topology and existence of minimizers

The symbol $B_r(x)$ will denote an open ball of radius r around a point x, namely $B_r(x) := \{y \in \mathbf{E} : \|y - x\| < r\}$. We will denote the open unit ball by $\mathbb{B}$. The *closure* of a set $Q \subset \mathbf{E}$, denoted $\operatorname{cl} Q$, consists of all points x such that the ball $B_\epsilon(x)$ intersects Q for all $\epsilon > 0$; the *interior* of Q, written as $\operatorname{int} Q$, is the set of all points x such that Q contains some open ball around x. We say that Q is an *open set* if it coincides with its interior and a *closed set* if it coincides with its closure. Any set Q in $\mathbf{E}$ that is closed

and bounded is called a *compact set*. We will often use the following result without explicitly quoting it.

Theorem 1.3 (Bolzano-Weierstrass)**.** *Any sequence in a compact set $Q \subset \mathbf{E}$ admits a subsequence converging to a point in Q.*

It will often be convenient to allow functions to take infinite values. Consequently, define the *extended real line* $\overline{\mathbf{R}} := \mathbf{R} \cup \{\pm\infty\}$. The *limit inferior* and *limit superior* of any sequence $\{r_i\} \subset \overline{\mathbf{R}}$ are defined by

$$\liminf_{i\to\infty} \; r_i = \lim_{i\to\infty} \left\{ \inf_{j\geq i} r_j \right\} \qquad \text{and} \qquad \limsup_{i\to\infty} \; r_i = \lim_{i\to\infty} \left\{ \sup_{j\geq i} r_j \right\}.$$

For any function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point $x \in \mathbf{E}$, we set

$$\liminf_{y\to x} f(y) = \lim_{r>0} \left\{ \inf_{y\in B_r(x)\setminus\{x\}} f(y) \right\}$$

The symbol $\limsup_{y\to x} f(y)$ is defined similarly, with sup replacing inf.

A basic question one can ask when minimizing a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is whether a minimizer even exists. For example, the infimal value of the function $f(x) = e^x$ is zero and yet this value is not attained at any point. A standard way to ensure that a function has minimizers, which we now discuss, is by assuming (1) compactness and (2) a mild continuity property.

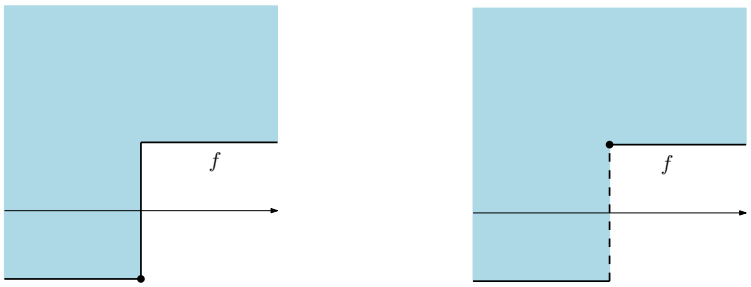
Definition 1.4 (Lower-semicontinuous)**.** A function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is *lower-semicontinuous at* $x \in \mathbf{E}$ if the inequality $\liminf_{y\to x} f(y) \geq f(x)$ holds. If f is lower-semicontinuous at every point $x \in \mathbf{E}$, then we call f *closed*.

Intuitively, lower-semicontinuity of f at x asserts that the function values cannot suddenly jump down as one moves slightly away from x. For example, the step function

$$f(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

is not lower-semicontinuous at $x = 0$ since $\lim_{i\to\infty} f(-i^{-1}) = -1 < f(0)$. If instead we redefine $f(0) = -1$, then the function becomes lower-semicontinuous; see Figure 1.4.

The following exercise shows that f is lower-semicontinuous at every point in $\mathbf{E}$ if and only if its epigraph—the set above the graph—is a closed set, thereby explaining why Definition 1.4 calls such functions closed. The geometry of the epigraph will play a central role in the later chapters.



(a) Closed. (b) Not closed.

Figure 1.4: Closed functions.

Exercise 1.5. ✍ Show that a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is closed if and only if the set, $\{(x, r) \in \mathbf{E} \times \mathbf{R} : f(x) \leq r\}$, is closed.

The following exercise shows that the infimal value of a closed function on a compact set is always attained.

Exercise 1.6 (Existence of minimizers on compact sets)**.** ✍ Consider a closed function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a nonempty compact set $Q \subset \mathbf{E}$. Then the infimum value $\inf_{x \in Q} f(x)$ is attained at some point in Q.

[**Hint:** Apply the Bolzano-Weierstrass Theorem to the sequence $x_i \in Q$ satisfying $f(x_i) \to \inf_Q f$ and invoke lower-semicontinuity.]

An important downside of the above exercise is it only guarantees existence of minimizers over compact sets. In light of the exponential example mentioned previously, if we wish to guarantee existence of minimizers over $\mathbf{E}$, then we must focus on a favorable class of functions.

Definition 1.7 (Coercive)**.** A function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is *coercive* if for any sequence x_i with $\|x_i\| \to \infty$, it must be that $f(x_i) \to +\infty$.

Equivalently, a function f is coercive precisely when the sublevel sets $\{x : f(x) \leq r\}$ are bounded for every $r \in \mathbf{R}$ (check this!). For example, the function $f(x) = e^{x^2}$ is coercive while the exponential $f(x) = e^x$ is not.

Exercise 1.8 (Existence of unconstrained minimizers)**.** ✍ Any coercive closed function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ has a minimizer.
[**Hint:** Choose $r \in \mathbf{R}$ such that the sublevel set $\mathcal{L} = \{x : f(x) \leq r\}$ is nonempty and apply Exercise 1.6.]

1.6 Differentiability

For the rest of the section, let $\mathbf{E}$ and $\mathbf{Y}$ be two Euclidean spaces, and U an open subset of $\mathbf{E}$. A mapping $F\colon Q \to \mathbf{Y}$, defined on a subset $Q \subset \mathbf{E}$, is *continuous* at a point $x \in Q$ if for any sequence x_i in Q converging to x, the values $F(x_i)$ converge to $F(x)$. We say that F is *continuous* if it is continuous at every $x \in Q$. We say that F is *L-Lipschitz continuous* if

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

If F if L-Lipschitz continuous with $L \in [0, 1)$, then we call F a *contraction*. If instead, F is 1-Lipschitz continuous, we say that F is *nonexpansive*.

A function $f\colon U \to \mathbf{R}$ is *differentiable* at a point x in U if there exists a vector, denoted by $\nabla f(x) \in \mathbf{E}$, satisfying

$$\lim_{h \to 0} \frac{f(x+h) - f(x) - \langle \nabla f(x), h \rangle}{\|h\|} = 0. \tag{1.3}$$

In words, the estimate (1.3) means that as h tends to zero, the error $f(x + h) - f(x) - \langle \nabla f(x), h \rangle$ tends to zero faster than any linear function. Rather than carrying fractions around, which can be cumbersome, it is convenient to introduce the following notation. The symbol $o(r)$ will always stand for a term satisfying $0 = \lim_{r \downarrow 0} o(r)/r$. Then the equation (1.3) simply amounts to the expression

$$f(x + h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|).$$

The term $o(\|h\|)$ is informally called a *first-order error* because it decays to zero faster than any linear function, as h tends to zero. The vector $\nabla f(x)$ is called the *gradient* of f at x. In the most familiar setting $\mathbf{E} = \mathbf{R}^n$, the gradient is simply the vector of partial derivatives

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}.$$

If the gradient mapping $x \mapsto \nabla f(x)$ is well-defined and continuous on U, we say that f is *C^1-smooth*. If the gradient satisfies the stronger Lipschitz property

$$\|\nabla f(y) - \nabla f(x)\| \leq \beta\|y - x\| \qquad \text{holds for all } x, y \in U,$$

then we say that f is *β-smooth*.

More generally, a mapping $F\colon U \to \mathbf{Y}$ is *differentiable* at $x \in U$ if there exists a linear mapping from $\mathbf{E}$ to $\mathbf{Y}$, denoted by $\nabla F(x)$, satisfying

$$F(x+h) = F(x) + \nabla F(x)h + o(\|h\|).$$

The linear mapping $\nabla F(x)$ is called the *Jacobian* of F at x. If the assignment $x \mapsto \nabla F(x)$ is continuous, we say that F is *C^1-smooth*. In the most familiar setting $\mathbf{E} = \mathbf{R}^n$ and $\mathbf{Y} = \mathbf{R}^m$, we can write F in terms of coordinate functions $F(x) = (F_1(x), \dots, F_m(x))$, and then the Jacobian is simply

$$\nabla F(x) = \begin{pmatrix} \nabla F_1(x)^T \\ \nabla F_2(x)^T \\ \vdots \\ \nabla F_m(x)^T \end{pmatrix} = \begin{pmatrix} \frac{\partial F_1(x)}{\partial x_1} & \frac{\partial F_1(x)}{\partial x_2} & \cdots & \frac{\partial F_1(x)}{\partial x_n} \\ \frac{\partial F_2(x)}{\partial x_1} & \frac{\partial F_2(x)}{\partial x_2} & \cdots & \frac{\partial F_2(x)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m(x)}{\partial x_1} & \frac{\partial F_m(x)}{\partial x_2} & \cdots & \frac{\partial F_m(x)}{\partial x_n} \end{pmatrix}.$$

Finally, we introduce second-order derivatives. A C^1-smooth function $f\colon U \to \mathbf{R}$ is *twice differentiable* at a point $x \in U$ if the gradient map $\nabla f\colon U \to \mathbf{E}$ is differentiable at x. Then the Jacobian of the gradient $\nabla(\nabla f)(x)$ is denoted by $\nabla^2 f(x)$ and is called the *Hessian* of f at x. Unraveling notation, the Hessian $\nabla^2 f(x)$ is characterized by the condition

$$\nabla f(x+h) = \nabla f(x) + \nabla^2 f(x)h + o(\|h\|).$$

If the map $x \mapsto \nabla^2 f(x)$ is continuous, we say that f is C^2-smooth. If f is indeed C^2-smooth, then a basic result of calculus shows that $\nabla^2 f(x)$ is a self-adjoint operator.

In the standard setting $\mathbf{E} = \mathbf{R}^n$, the Hessian is the matrix of second-order partial derivatives

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f_1(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}.$$

This matrix is symmetric, as long as it varies continuously with x in U.

Exercise 1.9. ✍ Define the function

$$f(x) = \tfrac{1}{2}\langle \mathcal{A}x, x\rangle + \langle v, x\rangle + c$$

where $\mathcal{A}\colon \mathbf{E} \to \mathbf{E}$ is a linear operator, v lies in $\mathbf{E}$, and c is a real number.

1. Show that if $\mathcal{A}$ is replaced by the self-adjoint operator $(\mathcal{A}+\mathcal{A}^*)/2$, the function values $f(x)$ remain unchanged.
2. Assuming $\mathcal{A}$ is self-adjoint, derive the equations:

$$\nabla f(x) = \mathcal{A}x + v \quad \text{and} \qquad \nabla^2 f(x) = \mathcal{A}.$$

3. Assuming $\mathcal{A}$ is self-adjoint, show that f is coercive if and only if $\mathcal{A}$ is positive definite.

Exercise 1.10. Define the function $f(x) = \frac{1}{2}\|F(x)\|^2$, where $F\colon \mathbf{E} \to \mathbf{Y}$ is a C^1-smooth mapping. Prove the identity $\nabla f(x) = \nabla F(x)^* F(x)$.

Exercise 1.11. ✍ Consider a function $f\colon \mathbf{E} \to \mathbf{R}$ and a linear mapping $\mathcal{A}\colon \mathbf{Y} \to \mathbf{E}$ and define the composition $h(x) = f(\mathcal{A}x)$.

1. Show that if f is differentiable at $\mathcal{A}x$, then

$$\nabla h(x) = \mathcal{A}^* \nabla f(\mathcal{A}x).$$

2. Show that if f is twice differentiable at $\mathcal{A}x$, then

$$\nabla^2 h(x) = \mathcal{A}^* \nabla^2 f(\mathcal{A}x)\mathcal{A}.$$

Exercise 1.12. ✍ Define the two sets

$$\begin{aligned} \mathbf{R}^n_{++} &:= \{x \in \mathbf{R}^n : x_i > 0 \text{ for all } i = 1, \dots, n\}, \\ \mathbf{S}^n_{++} &:= \{X \in \mathbf{S}^n : X \succ 0\}. \end{aligned}$$

Consider the two functions $f\colon \mathbf{R}^n_{++} \to \mathbf{R}$ and $F\colon \mathbf{S}^n_{++} \to \mathbf{R}$ given by

$$f(x) = -\sum_{i=1}^n \log x_i \qquad \text{and} \qquad F(X) = -\log\det(X),$$

respectively. Note, from basic properties of the determinant, the equality $F(X) = f(\lambda(X))$, where we set $\lambda(X) := (\lambda_1(X), \dots, \lambda_n(X))$.

1. Find the derivatives $\nabla f(x)$ and $\nabla^2 f(x)$ for $x \in \mathbf{R}^n_{++}$.
2. Using the property $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, prove $\nabla F(X) = -X^{-1}$ and $\nabla^2 F(X)[V] = X^{-1}VX^{-1}$ for any $X \succ 0$.

   [**Hint:** To compute $\nabla F(X)$, justify

$$F(X+tV) - F(X) + t\langle X^{-1}, V\rangle = -\log\det(I + tX^{-1/2}VX^{-1/2}) + t\cdot\operatorname{tr}(X^{-1/2}VX^{-1/2}).$$

By rewriting the expression in terms of eigenvalues of $X^{-1/2}VX^{-1/2}$, deduce that the right-hand-side is $o(t)$. To compute the Hessian, observe

$$(X+V)^{-1} = X^{-1/2}\left(I + X^{-1/2}VX^{-1/2}\right)^{-1} X^{-1/2},$$

and then use the expansion

$$(I+A)^{-1} = I - A + A^2 - A^3 + \ldots = I - A + O(\|A\|_{\text{op}}^2),$$

whenever $\|A\|_{\text{op}} < 1$.]

3. Show
$$\langle \nabla^2 F(X)[V], V\rangle = \|X^{-\frac{1}{2}}VX^{-\frac{1}{2}}\|_F^2$$
for any $X \succ 0$ and $V \in \mathbf{S}^n$. Deduce that the operator $\nabla^2 F(X)\colon \mathbf{S}^n \to \mathbf{S}^n$ is positive definite.

1.7 Accuracy in approximation

Recall that a set U in $\mathbf{E}$ is *convex* if for any two points $x, y \in U$ and real $\lambda \in [0,1]$, the point $\lambda x + (1-\lambda)y$ lies in U. In other words, a set U is convex if and only if the line segment joining any two point $x, y \in U$ lies entirely in U. Throughout the rest of the section, we let U be an open, convex subset of $\mathbf{E}$. Consider a function $f\colon U \to \mathbf{R}$ and a point $x \in U$. Multivariate calculus identifies the following two functions as the "best" linear and quadratic approximations of f near x, respectively:

$$\begin{aligned} l_x(y) &:= f(x) + \langle \nabla f(x), y - x\rangle, \\ Q_x(y) &:= f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y - x\rangle. \end{aligned}$$

The goal of this section is to quantify how closely $l_x(y)$ and $Q_x(y)$ approximate $f(y)$ under various smoothness assumptions on f. All results will follow quickly by restricting multivariate functions to line segments and then applying the fundamental theorem of calculus. To this end, the following observation plays a basic role.

Exercise 1.13. ✍ Consider a function $f\colon U \to \mathbf{R}$ and two points $x, y \in U$. Define the univariate function $\varphi\colon [0,1] \to \mathbf{R}$ given by $\varphi(t) = f(x + t(y-x))$ and let $x_t := x + t(y-x)$ for any t.

1. Show that if f is C^1-smooth, then equality
$$\varphi'(t) = \langle \nabla f(x_t), y - x\rangle \quad \text{holds for any } t \in (0,1).$$

2. Show that if f is C^2-smooth, then equality

$$\varphi''(t) = \langle \nabla^2 f(x_t)(y-x), y-x\rangle \quad \text{holds for any } t \in (0,1).$$

The following theorem precisely quantifies the gap between $f(y)$ and its linear and quadratic models, $l_x(y)$ and $Q_x(y)$.

Theorem 1.14 (Accuracy in approximation)**.** *Consider a C^1-smooth function $f\colon U \to \mathbf{R}$ and two points $x, y \in U$. Then we have*

$$f(y) = l_x(y) + \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x\rangle \, dt. \tag{1.4}$$

If f is C^2-smooth, then the equation holds:

$$f(y) = Q_x(y) + \int_0^1 \int_0^t \langle (\nabla^2 f(x + s(y-x)) - \nabla^2 f(x))(y-x), y-x\rangle \, ds\, dt.$$

Proof. Define the univariate function $\varphi(t) := f(x + t(y-x))$. The fundamental theorem of calculus yields the relation

$$\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)\, dt = \varphi'(0) + \int_0^1 \varphi'(t) - \varphi'(0)\, dt.$$

Using Exercise 1.13 directly yields (1.4). Suppose now that f is C^2-smooth. Applying the fundamental theorem of calculus twice yields

$$\begin{aligned}\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)\, dt &= \int_0^1 (\varphi'(0) + \int_0^t \varphi''(s)\, ds)\, dt \\ &= \varphi'(0) + \frac{1}{2}\varphi''(0) + \int_0^1 \int_0^t \varphi''(s) - \varphi''(0)\, ds\, dt.\end{aligned}$$

Appealing to Excercise 1.13, the result follows. □

Theorem 1.14 has a number of important consequences, two of which we derive now. The first consequence of Theorem 1.14 that we will often use is summarized in Corollary 1.15. The result shows that when the gradient mapping ∇f is β-Lipschitz continuous, one can replace the error term $o(\|y-x\|)$ in the definition of the gradient by a quadratic $\frac{\beta}{2}\|y-x\|^2$, with the estimation being accurate uniformly over all x and y.

Corollary 1.15 (Accuracy in approximation under Lipschitz conditions)**.** *Suppose that $f\colon U \to \mathbf{R}$ is a β-smooth function. Then for any points $x, y \in U$ the inequality*

$$\Big| f(y) - l_x(y) \Big| \le \frac{\beta}{2}\|y-x\|^2 \quad \textit{holds.} \tag{1.5}$$

Proof. Taking absolute values in (1.4) yields

$$\begin{aligned} |f(y) - l_x(y)| &\le \int_0^1 |\langle \nabla f(x + t(y-x)) - \nabla f(x), y - x\rangle|\, dt \\ &\le \int_0^1 \|\nabla f(x + t(y-x)) - \nabla f(x)\| \cdot \|y - x\|\, dt \qquad (1.6) \\ &\le \beta\|y-x\|^2 \cdot \left(\int_0^1 t\, dt\right) = \frac{\beta}{2}\|y-x\|^2, \qquad (1.7) \end{aligned}$$

where (1.6) follows from the Cauchy–Schwarz inequality and (1.7) uses Lipschitz continuity of ∇f. □

Exercise 1.16. Consider a function $f\colon U \to \overline{\mathbf{R}}$ that is C^2-smooth. Show that f is β-smooth if and only if the inequality $\|\nabla^2 f(x)\|_{\mathrm{op}} \le \beta$ holds.

The estimate (1.5) has a nice geometric interpretation. Observe that the inequality amounts to the two-sided bound

$$l_x(y) - \frac{\beta}{2}\|y-x\|^2 \le f(y) \le l_x(y) + \frac{\beta}{2}\|y-x\|^2.$$

Thus if f is β-smooth, then each point x yields two simple quadratics with amplitude β that upper-bound and lower-bound f, respectively, and agree with f at x. See Figure 1.5 for an illustration.

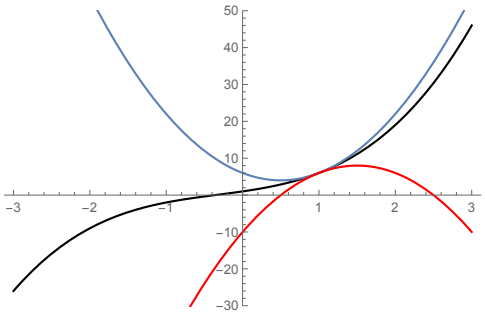


Figure 1.5: The black curve depicts the graph of a β-smooth function f; the blue and red curves depict graphs of the quadratics $l_x(\cdot) + \frac{\beta}{2}\|\cdot - x\|^2$ and $l_x(\cdot) - \frac{\beta}{2}\|\cdot - x\|^2$, respectively.

The second consequence of Theorem 1.14 that we will need is summarized in Corollary 1.17. The result shows that when f is C^2-smooth, the quadratic $Q_x(\cdot)$ is accurate up to a second-order error. Notice that this

is not immediate from the definition of the Hessian. Indeed, the Hessian $\nabla^2 f(x)$ a priori has no direct connection to the function values themselves, since it is defined as the Jacobian of the gradient map.

Corollary 1.17 (Second-order expansion)**.** *Suppose that $f\colon U \to \mathbf{R}$ is C^2-smooth. Then for any point $x \in U$, the estimate holds:*

$$\lim_{y \to x} \frac{f(y) - Q_x(y)}{\|y - x\|^2} = 0. \tag{1.8}$$

Proof. Fix two point $x, y \in U$ and define $x_s := x + s(y - x)$ for $s \in [0, 1]$. Using Theorem 1.14 and the Cauchy–Schwarz inequality, we compute

$$\begin{aligned} |f(y) - Q_x(y)| &\le \int_0^1 \int_0^t |\langle (\nabla^2 f(x_s) - \nabla^2 f(x))(y - x), y - x\rangle| \, ds \, dt \\ &\le \int_0^1 \int_0^t \|\nabla^2 f(x_s) - \nabla^2 f(x))(y - x)\| \cdot \|y - x\| \, ds \, dt \\ &\le \int_0^1 \int_0^t \|\nabla^2 f(x_s) - \nabla^2 f(x))\|_{\mathrm{op}} \cdot \|y - x\|^2 \, ds \, dt \\ &\le \|y - x\|^2 \cdot \max_{z \in [x,y]} \|\nabla^2 f(z) - \nabla^2 f(x)\|_{\mathrm{op}}. \end{aligned}$$

Since $\nabla^2 f$ is continuous, the function $z \mapsto \|\nabla f(z) - \nabla f(x)\|$ is uniformly continuous on any closed ball around x. Therefore, the term $\max_{z \in [x,y]} \|\nabla^2 f(z) - \nabla^2 f(x)\|_{\mathrm{op}}$ tends to zero as y tends to x. □

1.8 Optimality conditions for smooth optimization

We end the chapter with derivative-based necessary conditions and sufficient conditions for a point to be a local minimizer of a smooth function. A point x is called a *local minimizer* of a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ if there exists a neighborhood Q of x such that $f(x) \le f(y)$ for all $y \in Q$. Observe that naively checking if x is a local minimizer of f from the very definition requires evaluation of f at every point near x, an impossible task. We now derive a *verifiable necessary condition* for local optimality based on the gradient. Throughout the section, we let U be an open set in $\mathbf{E}$.

Theorem 1.18. *(First-order necessary conditions) Suppose that x is a local minimizer of a function $f\colon U \to \mathbf{R}$. If f is differentiable at x, then equality $\nabla f(x) = 0$ holds.*

Proof. Set $v := -\nabla f(x)$. Then for all small $t > 0$, the definition of differentiability implies

$$0 \le \frac{f(x+tv) - f(x)}{t} = -\|\nabla f(x)\|^2 + \frac{o(t)}{t}.$$

Letting t tend to zero yields $\nabla f(x) = 0$, as claimed. □

To obtain *verifiable sufficient conditions* for optimality, higher order derivatives are required.

Theorem 1.19. *(Second-order conditions)*
Consider a C^2-smooth function $f\colon U \to \mathbf{R}$ and fix a point $x \in U$. Then the following are true.

1. *(Necessary conditions) If $x \in U$ is a local minimizer of f, then*

$$\nabla f(x) = 0 \quad \text{and} \quad \nabla^2 f(x) \succeq 0.$$

2. *(Sufficient conditions) If the relations*

$$\nabla f(x) = 0 \quad \text{and} \quad \nabla^2 f(x) \succ 0$$

hold, then x is a local minimizer of f. More precisely, it holds:

$$\liminf_{y \to x} \frac{f(y) - f(x)}{\frac{1}{2}\|y - x\|^2} \ge \lambda_n(\nabla^2 f(x)).$$

Proof. Suppose first that x is a local minimizer of f. Then Theorem 1.18 guarantees $\nabla f(x) = 0$. Consider an arbitrary vector $v \in \mathbf{E}$. Then for all small $t > 0$, we deduce from a second-order expansion (1.8) the estimate

$$0 \le \frac{f(x+tv) - f(x)}{\frac{1}{2}t^2} = \langle \nabla^2 f(x)v, v\rangle + \frac{o(t^2)}{t^2}.$$

Letting t tend to zero yields $\langle \nabla^2 f(x)v, v\rangle \ge 0$ for all $v \in \mathbf{E}$, as claimed.

Suppose $\nabla f(x) = 0$ and $\nabla^2 f(x) \succ 0$. Let $\epsilon > 0$ be such that $B_\epsilon(x) \subset U$. Then for points y sufficiently close to x, the second-order expansion (1.8) yields the estimate

$$\begin{aligned}\frac{f(y) - f(x)}{\frac{1}{2}\|y - x\|^2} &= \left\langle \nabla^2 f(x) \left(\frac{y-x}{\|y-x\|} \right), \frac{y-x}{\|y-x\|} \right\rangle + \frac{o(\|y-x\|^2)}{\|y-x\|^2} \\ &\ge \lambda_n(\nabla^2 f(x)) + \frac{o(\|y-x\|^2)}{\|y-x\|^2}.\end{aligned}$$

Letting y tend to x, the result follows. □

The reader may be misled into believing that the role of the necessary conditions and the sufficient conditions for optimality (Theorem 1.19) is merely to determine whether a point x is a local minimizer of a smooth function f. Such a viewpoint is far too limited.

Necessary conditions serve as the basis for algorithm design. If necessary conditions for optimality fail at a point, then there must be some point nearby with a strictly smaller objective value. A method for discovering such a point is a first step for designing algorithms. Sufficient conditions play an entirely different role. In later chapters, we will later see that sufficient conditions for optimality at a point x guarantee that the function f is *strongly convex* on a neighborhood of x. Strong convexity, in turn, is essential for establishing rapid convergence of numerical methods.

1.9 Rates of convergence

A theoretically sound comparison of numerical methods relies on precise rates of progress in the iterates. For example, we will predominantly be interested in how fast the function gap $f(x_k) - \inf f$ or the distance to a minimizer $\|x_k - x^*\|$ tend to zero as a function of the counter k. In this section, we review three types of convergence rates that we will encounter.

Fix a sequence of real numbers $a_k > 0$ with $a_k \to 0$.

Sublinear rate. We will say that a_k converges *sublinearly* if there exist constants $c, q > 0$ satisfying

$$a_k \leq \frac{c}{k^q} \qquad \text{for all } k.$$

Larger q and smaller c indicates faster rates of convergence. In particular, given a target precision $\varepsilon > 0$, the inequality $a_k \leq \varepsilon$ holds for every $k \geq (\frac{c}{\varepsilon})^{1/q}$. The importance of the value of c should not be discounted; the convergence guarantee depends strongly on this value.

Linear rate. The sequence a_k is said to *converge linearly* if there exist constants $c > 0$ and $q \in (0, 1]$ satisfying

$$a_k \leq c \cdot (1 - q)^k \qquad \text{for all } k.$$

In this case, we call $1-q$ the *linear rate of convergence*. Fix a target accuracy $\varepsilon > 0$, and let us see how large k needs to be to ensure $a_k \leq \varepsilon$. To this end,

taking logs we get

$$c \cdot (1-q)^k \leq \varepsilon \quad \Longleftrightarrow \quad k \geq \frac{-1}{\ln(1-q)} \ln\left(\frac{c}{\varepsilon}\right).$$

Taking into account the inequality $\ln(1-q) \leq -q$, we deduce that the inequality $a_k \leq \varepsilon$ holds for every $k \geq \frac{1}{q}\ln(\frac{c}{\varepsilon})$. The dependence on q is strong, while the dependence on c is very weak, since the latter appears inside a log.

Quadratic rate. The sequence a_k is said to *converge quadratically* if there is a constant c satisfying

$$a_{k+1} \leq c \cdot a_k^2 \qquad \text{for all } k.$$

Observe then unrolling the recurrence yields

$$a_{k+1} \leq \frac{1}{c}(ca_0)^{2^{k+1}}.$$

The only role of the constant c is to ensure the starting moment of convergence. In particular, if $ca_0 < 1$, then the inequality $a_k \leq \varepsilon$ holds for all $k \geq \log_2 \ln(\frac{1}{c\varepsilon}) - \log_2(-\ln(ca_0))$. The dependence on c is negligible.

Comments

All results in this chapter can be found in standard textbooks in linear algebra and real analysis. For more details on the material in Sections 1.1-1.3, the reader may refer to the relevant sections Boyd-Vandenberghe [10], Halmos [16], and Strang [37]. The details of Section 1.5 can be found in Rudin [34]. The content of Sections 1.6-1.8 can be found in most advanced calculus textbooks, such as Apostol [1] and Folland [15].

Chapter 2

Convex geometry

This chapter introduces the basic geometric and topological properties of convex sets. The material presented here will, in turn, serve as the foundation for convex analysis developed in Chapter 3. The main goal for the reader should be to not only learn the formal theorems but to also develop intuition about convexity.

Roadmap. The chapter begins with Section 2.1 which recalls the definition of convex sets, introduces a few basic examples, and shows that convexity is preserved under various operations on sets, such as sums, intersections, and images/preimages by linear maps. Section 2.2 introduces the convex hull operation that associates to any set the smallest convex set that contains it. Section 2.3 discusses topological properties of convex sets. The key theorem proved in the section is that any nonempty convex set always has nonempty interior relative to the smallest affine space that contains it. Section 2.4 for the first time discusses the idea of hyperplane separation and duality. The main result is that any nonempty, closed, convex set admits a "dual description" as the intersection of all halfspaces containing it. An important construction motivated by such dual descriptions is the polar of a convex cone, discussed in Section 2.5. The final Section 2.6 introduces the cones of tangent and outward normal directions, which will play a central role in Chapter 3.

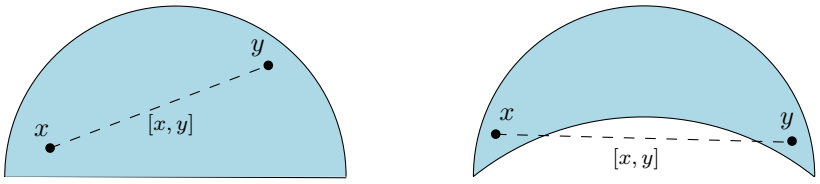
2.1 Operations preserving convexity

We begin with some convenient notation. For any two points x and y in $\mathbf{E}$, define the *closed line segment*

$$[x, y] := \{\lambda x + (1 - \lambda)y \ : \ 0 \leq \lambda \leq 1\}.$$

The open line segment (x, y) and the half-closed segments $[x, y)$ and $(x, y]$ are defined analogously. We have already encountered convex sets briefly in Section 1.7. Since convex set are the central objects of the current chapter, let us recall their defining property here.

Definition 2.1 (Convex sets)**.** A set $Q \subseteq \mathbf{E}$ is said to be *convex* if for any two points $x, y \in Q$, the entire line segment $[x, y]$ is contained in Q.



(a) Convex (b) Not convex.

Figure 2.1: Convexity.

Let us look at a few basic examples. First, it is immediate from the definition that linear subspaces are convex. More generally, a set $L \subset \mathbf{E}$ is called *affine* if it is a translate of a linear subspace. In other words L is affine if it has the form $L = v + S$ for some vector $v \in \mathbf{E}$ and a linear subspace $S \subset \mathbf{E}$. Since convexity is clearly preserved under translation, affine sets are convex. More interestingly, sets of the form $Q = \{x : \langle a, x\rangle \leq b\}$, for some $a \in \mathbf{E}$ and $b \in \mathbf{R}$, are convex. Such sets are called *half-spaces*. A quick computation also shows that unit balls of arbitrary norms are convex sets; see Figure 1.1 for an illustration. The reader should verify that the *nonnegative orthant*

$$\mathbf{R}^n_+ = \{x \in \mathbf{R}^n : x \geq 0\}$$

and the *cone of positive semi-definite matrices*

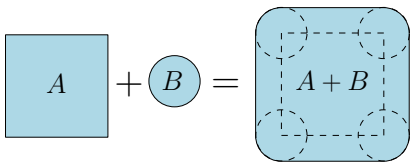
$$\mathbf{S}^n_+ = \{x \in \mathbf{S}^n : X \succeq 0\}$$

are convex. Here, the symbol "$\geq$" should be understood coordinatewise.

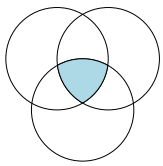
We thus have built a small (so far) library of convex sets. Verifying convexity from the definition is tedious and can often be avoided. The simplest way to argue that a set is convex is to recognize it as having been constructed from known convex sets (in our library) by a sequence of set operations that preserve convexity. In this section, we describe a few such convexity-preserving set operations. Refer to Section 1.4 for the sum, scaling, and image/preimage notation.

Exercise 2.2 (Preservation of convexity)**.** ✍ Prove the following statements.

1. *(Scaling)* For any convex set $A \subset \mathbf{E}$, the set $\mathbf{R}_+ A$ is convex.
2. *(Set addition)* For any two convex sets $Q_1, Q_2 \subset \mathbf{E}$, the sum $Q_1 + Q_2$ is convex. See Figure 2.2a for an example.
3. *(Intersection)* The intersection $\bigcap_{i \in I} Q_i$ of convex sets $Q_i \subset \mathbf{E}$, indexed by an arbitrary set I, is convex. See Figure 2.2b for an example.
4. *(Linear image/preimage)* For any convex sets $Q \subset \mathbf{E}$ and $L \subset \mathbf{Y}$ and a linear map $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$, the image $\mathcal{A}Q$ and the preimage $\mathcal{A}^{-1}L$ are convex sets.



(a) Disk plus square.



(b) Intersection of disks

Figure 2.2: Convexity preserving operations.

Let us look now at two notable examples of sets built from convexity preserving operations. A *polyhedron* is any set of the form

$$Q = \{x \in \mathbf{R}^n : Ax \geq c\},$$

for some $A \in \mathbf{R}^{m \times n}$ and $c \in \mathbf{R}^m$. Equivalently, we may write Q as an intersection of finitely many halfspaces or as the preimage $A^{-1}(c + \mathbf{R}^n_+)$. Appealing to Exercise (2.2), we deduce that polyhedra are convex. Linear programming refers to the problem of minimizing a linear function over a polyhedron.

More generally, a *spectrahedron* is any set of the form

$$Q = \{x \in \mathbf{R}^n : x_1 A_1 + x_2 A_2 + \ldots + x_n A_n \succeq C\},$$

for some matrices $A_i \in \mathbf{S}^m$ and $C \in \mathbf{S}^n$. Equivalently, we may write Q as the preimage $\mathcal{A}^{-1}(C + \mathbf{S}^n_+)$ for the linear map $\mathcal{A}(x) = \sum_{i=1}^n x_i A_i$. Appealing to Exercise (2.2), we deduce that spectrahedra are convex. Semidefinite programming refers to the problem of minimizing a linear function over a spectrahedron.

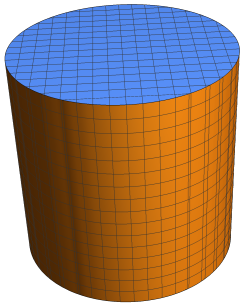
There are many more spectrahedra than polyhedra. For example, a quick computation shows that a cylinder can be written as the spectrahedron (do it!):

$$\left\{ (x,y,z) \in \mathbf{R}^3 : \begin{pmatrix} 1+x & y & 0 & 0 \\ y & 1-x & 0 & 0 \\ 0 & 0 & 1+z & 0 \\ 0 & 0 & 0 & 1-z \end{pmatrix} \succeq 0 \right\}.$$

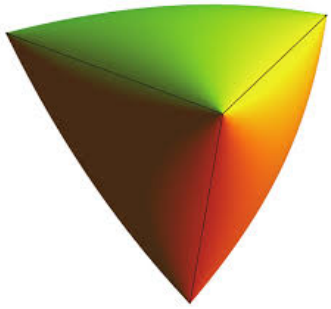
See Figure 2.3a for an illustration. A more interesting example, depicted in Figure 2.3b is the elliptope:

$$\left\{ (x,y,z) \in \mathbf{R}^3 : \begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix} \succeq 0 \right\}$$

The high dimensional version of this set appears in statistics as the set of correlation matrices and in combinatorial optimization when forming convex relaxations of NP-hard problems.



(a) Cylinder.



(b) The elliptope

Figure 2.3: Spectrahedra.

It will be important for the sets that we encounter to not only be convex but to also be closed. Not all set operations in Exercise 2.2 preserve closed sets. Whereas intersections and linear preimages of closed sets are closed, sums and linear images of closed sets need not be closed in general. The

following exercise presents two closed convex sets in $\mathbf{R}^3$ whose sum is not closed. Similarly, the image of a closed set under a linear map may also fail to be closed (why?). Though this pathology may seem like a technicality, it can have pronounced negative consequences; e.g. strong duality failing in convex optimization. Therefore, care must be taken when dealing with closure issues.

Exercise 2.3. Do the following exercises.

1. Show that if a closed set $Q \subset \mathbf{E}$ is bounded and does not contain the origin, then $\mathbf{R}_+Q$ is closed.
2. Show that if $Q_1, Q_2 \subset \mathbf{E}$ are closed sets and Q_1 is bounded, then the set $Q_1 + Q_2$ is closed.
3. Give an example of a closed set $Q \in \mathbf{R}^2$ such that $\mathbf{R}_+Q$ is not closed.
4. Define the two closed sets

   $$Q_1 = \{(x, y, r) \in \mathbf{R}^3 : \sqrt{x^2 + y^2} \leq r\} \qquad \text{and} \qquad Q_2 = \{(0, \lambda, \lambda) : \lambda \in \mathbf{R}\}.$$

   Show that the sum $Q_1 + Q_2$ is not a closed set.

We end the section with the following useful lemma that further highlights the interplay between convexity and set addition.

Lemma 2.4. *Consider a convex set $Q \subset \mathbf{E}$ and let $\lambda_1, \lambda_2 \geq 0$ be arbitrary. Then the equation holds:*

$$\lambda_1 Q + \lambda_2 Q = (\lambda_1 + \lambda_2)Q.$$

Proof. We may suppose $\lambda_1 + \lambda_2 \neq 0$, since otherwise the result is trivial. The inclusion $\supset$ clearly holds, independently of convexity. To see the converse, fix two points $x, y \in Q$. Convexity guarantees $\frac{\lambda_1}{\lambda_2+\lambda_2}x + \frac{\lambda_2}{\lambda_2+\lambda_2}y \in Q$. Multiplying through by $\lambda_1 + \lambda_2$ completes the proof. □

2.2 Convex hull

The notion of a linear combination of vectors plays a central role in linear algebra. Convex combinations of points play a similarly central role in convex geometry. To simplify notation, define the *unit simplex*

$$\Delta_n := \left\{ \lambda \in \mathbf{R}^n : \sum_{i=1}^{n} \lambda_i = 1, \lambda \geq 0 \right\}.$$

Definition 2.5 (Convex combination)**.** A point $x \in \mathbf{E}$ is a *convex combination* of points $x_1, \ldots, x_k \in \mathbf{E}$ if it can be written as $x = \sum_{i=1}^{k} \lambda_i x_i$ for some $\lambda \in \Delta_k$.

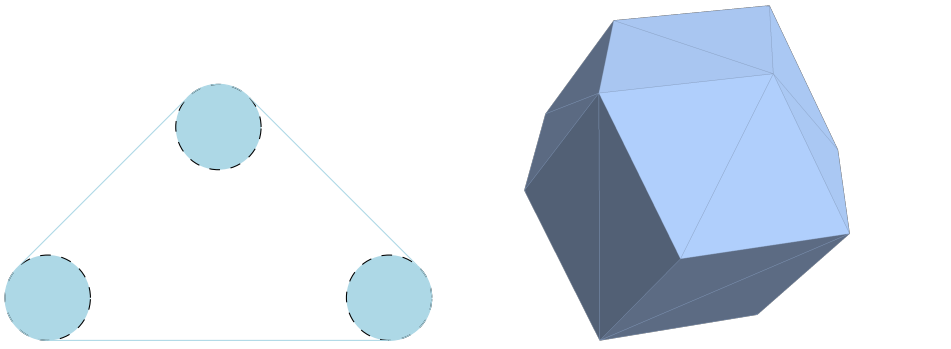
A useful way to think about a representation of x as a convex combination $x = \sum_{i=1}^{k} \lambda_i x_i$ is to regard x as a weighted average of $x_1, \ldots, x_k$ with $\lambda_1, \ldots, \lambda_k$ as the corresponding weights. Observe that convexity of a set $Q \subset \mathbf{E}$ guarantees that convex combinations of any two points of Q lie in Q; indeed, this property defines convexity. The following exercise shows that convexity of Q entails a seemingly stronger property: convex combinations of any finite number of points of Q lie in Q.

Exercise 2.6. ✍ Consider a convex set $Q \subset \mathbf{E}$ and let $k \in \mathbb{N}$ be arbitrary. Show that any convex combination of points $x_1, \ldots, x_k \in Q$ lies in Q.
[**Hint:** Rewrite $\sum_{i=1}^{k} \lambda_i x_i = (1-\lambda_k) \sum_{i=1}^{k-1} \frac{\lambda_i}{1-\lambda_k} x_i + \lambda_k x_k$ and reason inductively.]

For any nonconvex set Q, one can imagine forming the "minimal" convex set that contains Q. The resulting convex set is called the convex hull of Q.

Definition 2.7 (Convex hull)**.** The *convex hull* of a set $Q \subseteq \mathbf{E}$, denoted $\operatorname{conv}(Q)$, is the intersection of all convex sets containing Q.

Notice that by Exercise 2.2, the convex hull $\operatorname{conv}(Q)$ is a convex set. One can visualize the convex hull of a set $Q \subset \mathbf{R}^2$ by encircling Q with a rubber band and letting it contract. The outline of the rubber band marks the boundary of the convex hull. See Figure 2.4 for an illustration.



(a) Convex hull of three disks (b) $\operatorname{conv}\{(\pm 1, \pm 1, \pm 1), \pm 2e_1, \pm 2e_2, \pm 2e_3)\}$.

Figure 2.4: Convex hull.

The definition of the convex hull of Q is external: it involves sets that are larger than Q. The following exercise provides an equivalent internal description of $\operatorname{conv}(Q)$ as the set of all convex combinations of points in Q.

Exercise 2.8. ✍ For any set $Q \subset \mathbf{E}$, prove the equality:

$$\operatorname{conv}(Q) = \left\{ \sum_{i=1}^{k} \lambda_i x_i \ : \ k \in \mathbb{N},\ x_1, \dots, x_k \in Q,\ \lambda \in \Delta_k \right\}. \tag{2.1}$$

The description (2.1) does not rule out that one might have to take k arbitrarily large in order to obtain the entire convex hull $\operatorname{conv}(Q)$. On the contrary, the following theorem shows that it suffices to take $k \leq n+1$, where n is the dimension of $\mathbf{E}$.

Theorem 2.9 (Carathéodory)**.** *Consider a set $Q \subset \mathbf{E}$, where $\mathbf{E}$ is an n-dimensional Euclidean space. Then each point $x \in \operatorname{conv}(Q)$ can be written as a convex combination of at most $n+1$ points in Q.*

Proof. Since x belongs to $\operatorname{conv}(Q)$, we may write $x = \sum_{i=1}^{k} \lambda_i x_i$ for some integer k, points $x_1, \dots, x_k \in Q$, and weights $\lambda \in \Delta_k$. We may assume $k \geq n+2$, since otherwise there is nothing to prove. We claim that we may rewrite x as a convex combination of at most $k-1$ points.

We begin the argument by noticing that the vectors

$$x_2 - x_1, \dots, x_k - x_1$$

are linearly dependent, since there are at least $n+1$ of them. Therefore, there exist numbers μ_i for $i = 2, \dots, k$ not all zero and satisfying $0 = \sum_{i=2}^{k} \mu_i (x_i - x_1) = \sum_{i=2}^{k} \mu_i x_i - (\sum_{i=2}^{k} \mu_i) x_1$. Defining $\mu_1 := -\sum_{i=2}^{k} \mu_i$, we deduce $\sum_{i=1}^{k} \mu_i x_i = 0$ and $\sum_{i=1}^{k} \mu_i = 0$. Then for any $\alpha \in \mathbf{R}$, we compute

$$x = \sum_{i=1}^{k} \lambda_i x_i - \alpha \sum_{i=1}^{k} \mu_i x_i = \sum_{i=1}^{k} (\lambda_i - \alpha \mu_i) x_i$$

and

$$\sum_{i=1}^{k} (\lambda_i - \alpha \mu_i) = 1.$$

We will now choose α so that all the coefficients $\lambda_i - \alpha \mu_i$ are nonnegative and at least one of them is zero. To this end, observe that since the vector μ is not zero, it has at least one positive coordinate. Therefore, we may choose an index $i^* \in \operatorname{argmin}_i \{\lambda_i / \mu_i : \mu_i > 0\}$ and set $\alpha = \frac{\lambda_{i^*}}{\mu_{i^*}}$. Thus x is a convex combination of $k-1$ points, as the coefficient $\lambda_{i^*} - \alpha \mu_{i^*}$ is zero. Continuing this process, we will obtain a description of x as a convex combination of $k \leq n+1$ points. □

2.3 Affine hull and relative interior

Convex sets can easily have empty interior. For example, the unit simplex Δ_n has empty interior in its ambient space $\mathbf{R}^n$. The main result of this section shows that any nonempty convex set Q has nonempty interior "relative" to the smallest affine space that contains it. The main use of the relative interior in later sections will be to show that convex functions are very well-behaved within the relative interior of their domains.

Recall that a set is called affine if it has the form $L = v + S$ for some vector $v \in \mathbf{E}$ and a linear subspace $S \subset \mathbf{E}$. In particular, affine sets that contain the origin are linear subspaces (why?).

Definition 2.10 (Affine hull)**.** The *affine hull* of a set $Q \subset \mathbf{E}$, denoted by $\operatorname{aff} Q$, is the intersection of all affine sets that contain Q.

It is straightforward to check that $\operatorname{aff} Q$ is itself an affine set, and is by definition the smallest affine set that contains Q. See Figure 2.5 for an illustration. For example, the affine hull of the unit simplex Δ_n is the hyperplane $\{(x, y, z) : x+y+z = 1\}$. The reader should convince themselves that if Q contains the origin, then $\operatorname{aff} Q$ coincides with the linear span of Q.

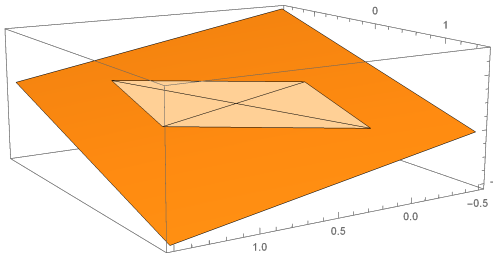


Figure 2.5: The convex set $\operatorname{conv}\{(e_1, e_2, e_3, ((1, -1, 1))\}$ and its affine hull.

Viewing $\operatorname{aff} Q$ as the ambient space of Q, it is appealing to focus on the interior of Q relative to this smaller ambient space.

Definition 2.11 (Relative interior and boundary)**.** The *relative interior* of a set $Q \subset \mathbf{E}$, denoted $\operatorname{ri} Q$, is the interior of Q relative to $\operatorname{aff}(Q)$. That is, we set

$$\operatorname{ri} Q := \{x \in Q : \exists \epsilon > 0 \text{ s.t. } B_\epsilon(x) \cap \operatorname{aff} Q \subseteq Q\}.$$

The *relative boundary* of Q is defined by $\operatorname{rb} Q := (\operatorname{cl} Q) \setminus (\operatorname{ri} Q)$.

The following is the main result of the section.

Theorem 2.12 (Relative interior is nonempty)**.** *For any nonempty convex set $Q \subset \mathbf{E}$, the relative interior* $\operatorname{ri} Q$ *is nonempty.*

Proof. Without loss of generality, we may translate Q to contain the origin. Then $\operatorname{aff} Q$ contains the origin and is therefore a linear subspace. Let d be the dimension of $\operatorname{aff} Q$ as a linear subspace. Observe that Q must contain some d linearly independent vectors $x_1, \ldots, x_d$, since otherwise $\operatorname{aff} Q$ would have a smaller dimension than d. Define the linear map $A\colon \mathbf{R}^d \to \operatorname{aff} Q$ by $A(\lambda_1, \ldots, \lambda_d) = \sum_{i=1}^d \lambda_i x_i$. Since the range of A contains $x_1, \ldots, x_d$, the map A is surjective. Hence A is a linear isomorphism. Consequently A maps the open set

$$\Omega := \left\{ \lambda \in \mathbf{R}^d : \sum_{i=1}^d \lambda_i < 1 \text{ and } \lambda_i > 0 \text{ for all } i \right\}$$

to an open subset $A(\Omega)$ of $\operatorname{aff} Q$. Note for any $\lambda \in \Omega$, we can write $A\lambda = \sum_{i=1}^d \lambda_i x_i + (1 - \sum_{i=1}^d \lambda_i) \cdot 0$. Hence, convexity of Q implies $A(\Omega) \subset Q$, thereby proving $\operatorname{ri} Q \neq \emptyset$. □

Exercise 2.13. ✍ Show that for any convex set $Q \subset \mathbf{E}$, the two sets, $\operatorname{cl} Q$ and $\operatorname{ri} Q$, are convex and have the same affine hull as Q itself.

One important consequence of (2.12) is that any closed convex set coincides with the closure of its relative interior. This result, proved in Corollary 2.15, will follow quickly from the following lemma.

Theorem 2.14 (Accessibility)**.** *Consider a convex set Q and two points $x \in \operatorname{ri} Q$ and $y \in \operatorname{cl} Q$. Then the line segment $[x, y)$ is contained in* $\operatorname{ri} Q$*.*

Proof. Without loss of generality, we may suppose that the affine hull of Q is all of $\mathbf{E}$. Fixing $\lambda \in (0, 1)$, we aim to show that the ball $(1-\lambda)x + \lambda y + \epsilon\mathbb{B}$ is contained in Q for some sufficiently small $\epsilon > 0$. Since y lies in $\operatorname{cl} Q$, the inclusion $y \in Q + \epsilon\mathbb{B}$ holds for all $\epsilon > 0$. We therefore deduce

$$\begin{aligned}(1-\lambda)x + \lambda y + \epsilon\mathbb{B} &\subset (1-\lambda)x + \lambda(Q + \epsilon\mathbb{B}) + \epsilon\mathbb{B} \\ &= (1-\lambda)\left(x + \tfrac{1+\lambda}{(1-\lambda)}\epsilon\mathbb{B}\right) + \lambda Q, \qquad (2.2)\end{aligned}$$

where (2.2) follows from Lemma 2.4. Since x lies in the interior of Q, the inclusion $x + \frac{1+\lambda}{1-\lambda}\epsilon\mathbb{B} \subset Q$ holds for all sufficiently small $\epsilon > 0$. Combining (2.2) with Lemma 2.4, we conclude

$$(1-\lambda)x + \lambda y + \epsilon\mathbb{B} \subset (1-\lambda)Q + \lambda Q = Q,$$

as we had to show. □

Corollary 2.15. *For any nonempty convex set Q in $\mathbf{E}$, equalities holds:*

$$\operatorname{cl}(\operatorname{ri} Q) = \operatorname{cl} Q \quad \textit{and} \quad \operatorname{ri}(\operatorname{cl} Q)) = \operatorname{ri} Q.$$

Proof. We begin by verifying the first equation. The inclusion $\operatorname{ri} Q \subseteq Q$ immediately implies $\operatorname{cl}(\operatorname{ri} Q) \subseteq \operatorname{cl} Q$. Conversely, fix a point $y \in \operatorname{cl} Q$. Since $\operatorname{ri} Q$ is nonempty by Theorem 2.12, we may also choose a point $x \in \operatorname{ri} Q$. Theorem 2.14 then guarantees $y \in \operatorname{cl}[x, y) \subseteq \operatorname{cl}(\operatorname{ri} Q)$. Since the point $y \in \operatorname{cl} Q$ is arbitrary, we have established the equality $\operatorname{cl}(\operatorname{ri} Q) = \operatorname{cl} Q$.

Next, we verify the second equation. Without loss of generality, we may suppose that Q contains the origin and therefore that $\operatorname{aff}(Q)$ is a linear subspace. The inclusion $\supset$ is clear. Fix any point $z \in \operatorname{ri}(\operatorname{cl} Q)$ and choose an arbitrary point $x \in \operatorname{ri} Q$. We may assume $x \neq z$, since otherwise the inclusion $z \in \operatorname{ri} Q$ would hold trivially. Observe from Exercise 2.13 the equality $\operatorname{aff} Q = \operatorname{aff}(\operatorname{cl} Q)$. Fix a constant $\mu > 0$ and define the point

$$y := z + \mu(z - x).$$

Since $\operatorname{aff} Q$ is a linear subspace, the inclusion $y \in \operatorname{aff} Q$ holds. Therefore the definition of the relative interior guarantees $y \in \operatorname{cl} Q$ for all sufficiently small $\mu > 0$. Rearranging the equation, we deduce

$$z = \frac{1}{1+\mu} y + \frac{\mu}{1+\mu} x \in (y, x).$$

Thus by Theorem 2.14, the inclusion $z \in \operatorname{ri} Q$ holds. □

2.4 Separation theorem

One of the most fruitful ways to study properties of sets is to instead focus on the functions that act on them. This is the principle of duality. This section introduces duality within the context of convex geometry. The main result is the principle of strict separation: any point y lying outside a closed, convex set Q can be separated from Q by a hyperplane. An important consequence is the dual description of convex sets. Tautologically a convex set Q is simply a collection of points. On the other hand, we will see that Q coincides with the intersection of all half-spaces containing it.

We begin with the following basic definitions. Along with any set $Q \subset \mathbf{E}$ we define the *distance function*

$$\operatorname{dist}_Q(y) := \inf_{x \in Q} \|x - y\|,$$

and the *projection*

$$\operatorname{proj}_Q(y) := \{x \in Q : \operatorname{dist}_Q(y) = \|x - y\|\}.$$

Thus $\operatorname{proj}_Q(y)$ consists of all the nearest points of Q to y; see Figure 2.6 for an illustration.

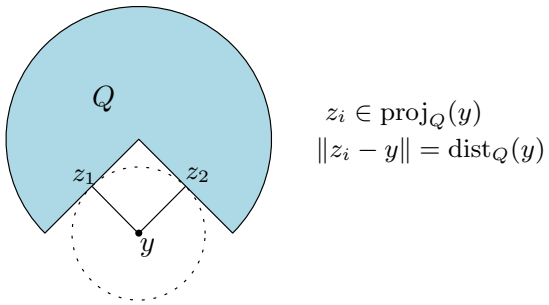


Figure 2.6: Nearest-point projection

Exercise 2.16. ✍ Show that for any nonempty set $Q \subseteq \mathbf{E}$, the function $\operatorname{dist}_Q \colon \mathbf{E} \to \mathbf{R}$ is 1-Lipschitz.

If Q is closed, then the nearest-point set $\operatorname{proj}_Q(y)$ is nonempty for any $y \in \mathbf{E}$. To see this, fix a point $y \in \mathbf{E}$ and define the function

$$\varphi(x) = \begin{cases} \|x - y\| & \text{if } x \in Q \\ +\infty & \text{if } x \notin Q \end{cases}.$$

The set of minimizers of φ coincides with $\operatorname{proj}_Q(y)$. Since φ is closed and coercive, Theorem 1.8 guarantees that φ has at least one minimizer, and therefore $\operatorname{proj}_Q(y)$ is nonempty.

The following theorem shows that when Q is closed and convex, the set $\operatorname{proj}_Q(y)$ is not only nonempty, but is also a singleton. Moreover, the nearest-point $z \in Q$ to y is characterized by the fact that the vector $y - z$ makes an obtuse angle with the vector $x - z$ for any $x \in Q$. See Figure 2.7 for an illustration.

Theorem 2.17 (Properties of the projection)**.** *For any nonempty, closed, convex set $Q \subset \mathbf{E}$, the set* $\operatorname{proj}_Q(y)$ *is a singleton. Moreover, the closest point $z \in Q$ to y is characterized by the property:*

$$\langle y - z, x - z \rangle \leq 0 \qquad \textit{for all } x \in Q. \tag{2.3}$$

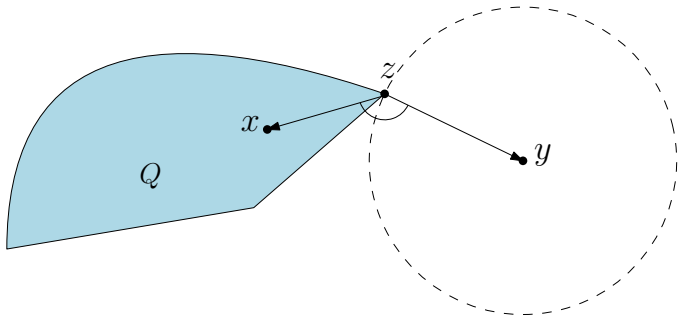


Figure 2.7: Nearest-point projection for convex sets

Proof. Fix a point $y \notin Q$. The claim that any point z satisfying (2.3) lies in $\mathrm{proj}_Q(y)$ is an easy exercise (verify it!). We therefore prove the converse. To this end, fix a point $z \in \mathrm{proj}_Q(y)$ and an arbitrary $x \in Q$. For each $t \in [0,1]$, define the point $x_t := z + t(x - z)$ and define the function $\varphi(t) := \frac{1}{2}\|y - x_t\|^2$. Convexity implies $x_t \in Q$ for all $t \in [0,1]$ and therefore

$$\varphi(t) \geq \tfrac{1}{2}\mathrm{dist}_Q^2(y) = \varphi(0).$$

Taking the derivative of φ, we therefore deduce

$$0 \leq \lim_{t \searrow 0} \frac{\varphi(t) - \varphi(0)}{t} = \varphi'(0) = -\langle y - z, x - z \rangle,$$

as claimed. Thus, a points z lies in $\mathrm{proj}_Q(y)$ if and only if (2.3) holds.

To see that $\mathrm{proj}_Q(y)$ is a singleton, consider any two points $z, z' \in \mathrm{proj}_Q(y)$. Then, the estimate (2.3) for z and z' (with $x = z'$ and $x = z$, respectively) becomes

$$\langle y - z, z' - z \rangle \leq 0 \qquad \text{and} \qquad \langle y - z', z - z' \rangle \leq 0.$$

Adding the two inequalities yields $0 \geq \langle z - z', z - z' \rangle = \|z - z'\|^2$, and therefore $z = z'$ as we had to show. □

Exercise 2.18. ✍ Show that for any nonempty, closed, convex set $Q \subset \mathbf{E}$, the map $x \mapsto \mathrm{proj}_Q(x)$ is 1-Lipschitz.

Theorem 2.17 allows to quickly prove the following fundamental property of convex sets. Given any closed convex set Q and a point $y \notin Q$, there exists a hyperplane that separates y from Q. See Figure 2.8 for an illustration.

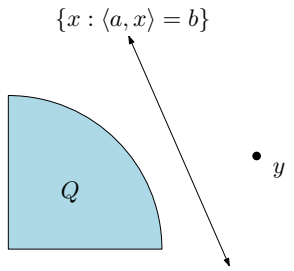


Figure 2.8: Basic separation

Theorem 2.19 (Strict separation)**.** *Consider a nonempty, closed, convex set $Q \subset \mathbf{E}$ and a point $y \notin Q$. Then there exists a nonzero vector $a \in \mathbf{E}$ and a number $b \in \mathbf{R}$ satisfying*

$$\langle a, x\rangle \leq b < \langle a, y\rangle \quad \text{for all } x \in Q.$$

Proof. Fix a point $y \notin Q$ and define the nonzero vector $a := y - \operatorname{proj}_Q(y)$. Then for any $x \in Q$, the condition (2.3) yields

$$\langle a, x\rangle \leq \langle a, \operatorname{proj}_Q(y)\rangle = \langle a, y\rangle - \|a\|^2 < \langle a, y\rangle,$$

as claimed. □

Exercise 2.20 (Supporting halfspaces)**.** Consider a convex set $Q \subset \mathbf{E}$. A halfspace $H \subset \mathbf{E}$ is said to *support Q at a point $x \in \operatorname{cl} Q$* if the inclusions, $x \in \operatorname{bd} H$ and $Q \subset H$, hold. Show that a convex set Q admits a supporting halfspace at a point $x \in \operatorname{cl} Q$ if and only if x lies on the boundary of Q. [**Hint:** Apply Theorem 2.19 with $y \notin Q$ tending to x.]

In particular, we can now establish the following "dual description" of convex sets, alluded to in the beginning of the section; see Figure 2.9.

Theorem 2.21. *Given a nonempty set $Q \subset \mathbf{E}$, define the set of halfspaces*

$$\mathcal{F}_Q := \{(a, b) \in \mathbf{E} \times \mathbf{R} : \langle a, x\rangle \leq b \quad \text{for all } x \in Q\}.$$

Then equality holds:

$$\operatorname{cl}\operatorname{conv}(Q) = \bigcap_{(a,b)\in\mathcal{F}_Q} \{x \in \mathbf{E} : \langle a, x\rangle \leq b\}. \tag{2.4}$$

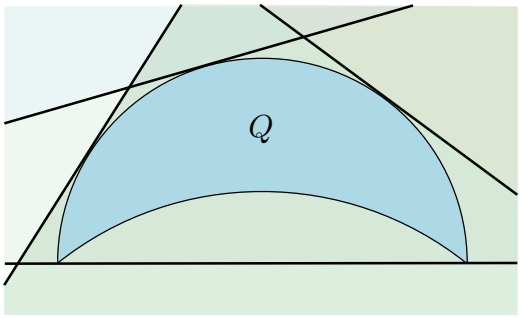


Figure 2.9: Closed convex envelope.

Proof. Let S be the set on the right side of (2.4). Since S is an intersection of half-spaces, it is closed and convex. Taking into account the inclusion $Q \subset S$, we deduce $\operatorname{cl}\operatorname{conv}(Q) \subset S$. To see the reverse inclusion, we argue by contradiction. Suppose that there exists a point $y \in S \setminus \operatorname{cl}\operatorname{conv}(Q)$. Then Theorem 2.19 yields $a \in \mathbf{E}$ and $b \in \mathbf{R}$ such that the halfspace $H = \{x : \langle a, x\rangle \leq b\}$ satisfies $\operatorname{cl}\operatorname{conv}(Q) \subset H$ and $y \notin H$. In particular, the inclusion $(a, b) \in \mathcal{F}_Q$ holds and therefore $S \subset H$. We thus arrive at a contradiction to the inclusions $y \in S \subset H$. The proof is complete. □

2.5 Cones and polarity

A particularly appealing class of sets consists of those that are invariant under scaling by nonnegative numbers.

Definition 2.22 (Cones)**.** A set $K \subseteq \mathbf{E}$ is called a *cone* if the inclusion $\lambda K \subset K$ holds for any $\lambda \geq 0$.

For example, the nonnegative orthant $\mathbf{R}^n_+$ and the set of positive semidefinite matrices $\mathbf{S}^n_+$ are closed convex cones.

Exercise 2.23. ✍ Show that a set $K \subset \mathbf{E}$ is a convex cone if and only if the point $\lambda x + \mu y$ lies in K for any two points $x, y \in K$ and numbers $\lambda, \mu \geq 0$.

Exercise 2.24. Prove for any convex cone $K \subset \mathbf{E}$ the equality

$$\operatorname{aff}(K) = K - K.$$

Duality ideas, explored in Theorem 2.21 simplify significantly for cones. Namely, consider a cone K and a halfspace

$$H = \{x : \langle a, x\rangle \leq b\}$$

that contains it. Since K contains the origin, b is nonnegative. Moreover, taking into account positive homogeneity of K, the halfspace

$$\overline{H} = \{x : \langle a, x\rangle \leq 0\}, \tag{2.5}$$

provides a tighter approximation $K \subset \overline{H} \subset H$. The set of all halfspaces of the form (2.5) that contain K comprise the polar cone.

Definition 2.25 (Polar cone)**.** The *polar cone* of a cone $K \subset \mathbf{E}$ is the set

$$K^\circ := \{v \in \mathbf{E} : \langle v, x\rangle \leq 0 \text{ for all } x \in K\}.$$

Thus K° consists of all vectors v that make an obtuse angle with every vector $x \in K$. Observe that K° is always closed and convex since it is defined as the intersection of infinitely many half-spaces. See Figure 2.10 for an illustration.

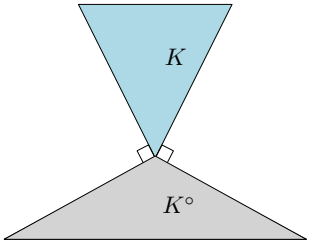


Figure 2.10: Polar cone

Exercise 2.26. Verify the following:

1. The polar cone of a linear subspace $L \subset \mathbf{E}$ is the orthogonal complement $L^\circ = L^\perp$.
2. $(\mathbf{R}^n_+)^\circ = \mathbf{R}^n_-$ and $(\mathbf{S}^n_+)^\circ = \mathbf{S}^n_-$

The following exercise shows a fundamental property of the polarity operation. Applying the polar operation twice to a cone K yields its closed convex hull. The proof is immediate from Theorem 2.21.

Exercise 2.27 (Double-polar theorem)**.** ✍ Prove for any nonempty cone $K \subset \mathbf{E}$ the equality:

$$(K^\circ)^\circ = \operatorname{cl}\operatorname{conv}(K).$$

Classically, the orthogonal complement to a sum of linear subspaces is the intersection of their orthogonal complements. In much the same way, the polarity operation satisfies "calculus rules".

Theorem 2.28 (Polarity calculus)**.** *For any linear mapping $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$ and a nonempty cone $K \subset \mathbf{Y}$, equality holds:*

$$(\mathcal{A}K)^\circ = (\mathcal{A}^*)^{-1}K^\circ. \tag{2.6}$$

In particular, for any two nonempty cones $K_1, K_2 \subset \mathbf{E}$, the sum rule holds:

$$(K_1 + K_2)^\circ = K_1^\circ \cap K_2^\circ \tag{2.7}$$

Proof. The definition of polarity yields the equivalences

$$\begin{aligned} y \in (\mathcal{A}K)^\circ &\iff \langle \mathcal{A}x, y\rangle \le 0 \text{ for all } x \in K \\ &\iff \langle x, \mathcal{A}^*y\rangle \le 0 \text{ for all } x \in K \\ &\iff \mathcal{A}^*y \in K^\circ \\ &\iff y \in (\mathcal{A}^*)^{-1}K^\circ. \end{aligned}$$

This establishes (2.6). The sum rule (2.7) follows from applying (2.6) to the expression $\mathcal{A}(K_1 \times K_2)$ with the mapping $\mathcal{A}(x, y) := x + y$. □

There is a convenient extension of polarity to general sets (not cones) that contain the origin. The idea is to "homogenize" the set and then apply the polarity operation for cones. Consider a set $Q \subset \mathbf{E}$ that contains the origin and let K be the cone generated by $Q \times \{1\} \subset \mathbf{E} \times \mathbf{R}$, that is

$$K = \{(\lambda x, \lambda) \in \mathbf{E} \times \mathbf{R} : x \in Q, \lambda \ge 0\}.$$

Since Q contains the origin, the polar cone K° is contained in $\mathbf{E} \times \mathbf{R}_-$. Consequently, it is natural to define the *polar set* as

$$Q^\circ := \{x \in \mathbf{E} : (x, -1) \in K^\circ\}.$$

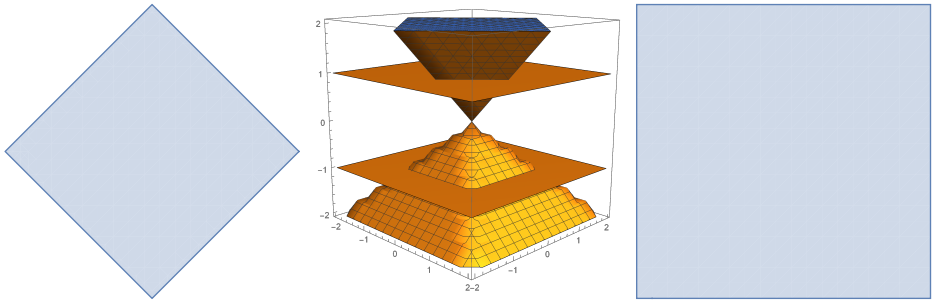
The reader should refer to Figure 2.11 for an illustration.

Unraveling the definitions, the following algebraic description of the polar appears.

Exercise 2.29. ✍ Show that for any set $Q \subset \mathbf{E}$ containing the origin, equality holds:

$$Q^\circ = \{v \in \mathbf{E} : \langle v, x\rangle \le 1 \text{ for all } x \in Q\}.$$

Thus, Q° can be identified with the intersection of the set $\mathcal{F}_Q$ from Theorem 2.21 with the slice $\mathbf{E} \times \{1\}$. The following exercise is a direct analogue of Exercise 2.27



(a) $Q = \{x : \|x\|_1 \leq 1\}$ (b) Homogenization (c) $Q^\circ = \{x : \|x\|_\infty \leq 1\}$

Figure 2.11: Figure 2.11a depicts Q, the unit ball of the ℓ_1-norm. Figure 2.11b depicts the homogenization of Q, namely $K = \{(x, y, r) : r \geq |x| + |y|\}$ and the polar cone $K^\circ = \{(x, y, r) : r \leq -\max\{|x|, |y|\}\}$, along with the parallel hyperplanes $\mathbf{R}^2 \times \{\pm 1\}$. Figure 2.11c depicts Q°, which can be identified with the intersection of K° with the hyperplane $\mathbf{R}^2 \times \{-1\}$.

Exercise 2.30 (Double polar)**.** For any set $Q \subset \mathbf{E}$ containing the origin, we have

$$(Q^\circ)^\circ = \operatorname{cl}\operatorname{conv}(Q).$$

Polarity of unit norm balls is in correspondence with duality of norms.

Exercise 2.31 (Polarity and dual norms)**.** Let $\rho(\cdot)$ be an arbitrary norm on $\mathbf{E}$ and define its unit ball

$$B_\rho = \{x \in \mathbf{E} : \rho(x) \leq 1\}.$$

Show that the polar of B_ρ is the unit ball of the dual norm:

$$B_\rho^\circ = B_{\rho^*}.$$

2.6 Tangents and normals

A principle technique in mathematical analysis is to reason about sets and functions using their first-order approximations. Both theory and algorithms rely on such approximations. This section revisits this idea by constructing first-order approximations of sets.

Consider a set $Q \subset \mathbf{E}$ and a point $\bar{x} \in Q$. Intuitively, we should think of a first-order approximation to Q at $\bar{x}$ as the set of all limits of rays $\mathbf{R}_+(x_i - \bar{x})$ over all possible sequences $x_i \in Q$ tending to $\bar{x}$. With this intuition in mind, we introduce the following definition.

Definition 2.32 (Tangent cone)**.** The *tangent cone* to a set $Q \subset \mathbf{E}$ at a point $\bar{x} \in Q$ is the set

$$T_Q(\bar{x}) := \left\{ \lim_{i\to\infty} \tau_i^{-1}(x_i - \bar{x}) : x_i \to \bar{x} \text{ in } Q,\ \tau_i \searrow 0 \right\}.$$

In the definition, the sequence $\tau_i > 0$ simply rescales the approach directions $x_i - \bar{x}$. See Figure 2.12 for an illustation. The reader should convince themselves that $T_Q(\bar{x})$ is a closed cone, which need not be convex in general. Whenever Q is convex, the definition simplifies drastically.

Exercise 2.33. ✍ Show for any convex set $Q \subset \mathbf{E}$ and a point $\bar{x} \in Q$ the equality:

$$T_Q(\bar{x}) = \text{cl } \mathbf{R}_+(Q - \bar{x}).$$

[**Hint:** The inclusion $\subset$ is clear and does not use convexity. The reverse inclusion follows from the definition of convexity and the fact that $T_Q(\bar{x})$ is closed.]

Thus for a convex set, the tangent cone $T_Q(\bar{x})$ is computed by shifting Q so that $\bar{x}$ becomes the origin and then taking the closure of all nonnegative scalings of the shifted set; see Figure 2.13.

Tangency concerns directions pointing "into the set". Alternatively, we can also think dually of outward normal vectors to a set Q at $\bar{x} \in Q$. Geometrically, it is intuitive to call a vector v an (outward) normal to Q at $\bar{x}$ if Q is fully contained in the half-space $\{x \in \mathbf{E} : \langle v, x - \bar{x}\rangle \le 0\}$ up to a first-order error.

Definition 2.34 (Normal cone)**.** The *normal cone* to a set $Q \subset \mathbf{E}$ at a point $\bar{x} \in Q$ is the set

$$N_Q(\bar{x}) := \{v \in \mathbf{E} : \langle v, x - \bar{x}\rangle \le o(\|x - \bar{x}\|) \quad \text{as } x \to \bar{x} \text{ in } Q\}.$$

Thus a vector v lies in $N_Q(\bar{x})$ if

$$\limsup_{x \underset{Q}{\to} \bar{x}} \frac{\langle v, x - \bar{x}\rangle}{\|x - \bar{x}\|} \le 0,$$

where the notation $x \underset{Q}{\to} \bar{x}$ means that x tends to $\bar{x}$ in Q. See Figure 2.12 for an illustration.

The following result shows that the normal cone $N_Q(\bar{x})$ is always the polar of the tangent cone $T_Q(\bar{x})$. In particular, $N_Q(\bar{x})$ is always a closed convex cone, even if Q is not convex. Consequently, taking into account the double polar formula (Exercise 2.27), equality $T_Q(\bar{x}) = (N_Q(\bar{x}))^\circ$ holds if and only if $T_Q(\bar{x})$ is convex.

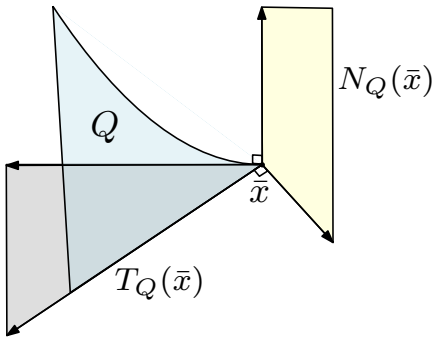


Figure 2.12: Illustration of the tangent and normal cones for nonconvex sets.

Lemma 2.35. ✍ *For any set $Q \subset \mathbf{E}$ and a point $\bar{x} \in Q$, the polarity relationship holds:*

$$N_Q(\bar{x}) = (T_Q(\bar{x}))^\circ.$$

Proof. We show the inclusion $\subset$ first. Fix vectors $v \in N_Q(\bar{x})$ and $w \in T_Q(\bar{x})$; we aim to show $\langle v, w\rangle \leq 0$. By definition of the tangent cone, there exist sequences $x_i \to \bar{x}$ in Q and $\tau_i \searrow 0$ satisfying $\tau_i^{-1}(x_i - \bar{x}) \to w$. We may suppose $w \neq 0$ (why?) and therefore $x_i \neq \bar{x}$ for all large i. The definition of the normal cones then yields

$$\langle v, w\rangle = \lim_{i\to\infty} \frac{\langle v, x_i - \bar{x}\rangle}{\tau_i} = \lim_{i\to\infty} \frac{\langle v, x_i - \bar{x}\rangle}{\|x_i - \bar{x}\|} \cdot \lim_{i\to\infty} \left\| \frac{x_i - \bar{x}}{\tau_i} \right\| \leq 0.$$

Since $w \in T_Q(\bar{x})$ was arbitrary, we deduce $v \in (T_Q(\bar{x}))^\circ$.

To see the reverse inclusion $\supset$, fix a vector $v \in (T_Q(\bar{x}))^\circ$. Thus the inequality $\langle v, w\rangle \leq 0$ holds for all $w \in T_Q(\bar{x})$. Consider now a sequence $x_i \to \bar{x}$ in Q, such that $x_i \neq \bar{x}$ for all large i. Defining $\tau_i = \|x_i - \bar{x}\|$, we deduce

$$\limsup_{i\to\infty} \frac{\langle v, x_i - \bar{x}\rangle}{\|x_i - \bar{x}\|} = \limsup_{i\to\infty} \ \langle v, \tau_i^{-1}(x_i - \bar{x})\rangle. \tag{2.8}$$

Passing to subsequence, we may assume that the real numbers $\langle v, \tau_i^{-1}(x_i - \bar{x})\rangle$ converge. Since the vectors $\tau_i^{-1}(x_i - \bar{x})$ all have unit norm, we may again pass to a subsequence to ensure that $\tau_i^{-1}(x_i - \bar{x})$ converge to some vector w. Since w clearly lies in $T_Q(\bar{x})$ while v lies in $(T_Q(\bar{x}))^\circ$, we deduce that the right-hand side of (2.8) is nonpositive. Therefore the inclusion $v \in N_Q(\bar{x})$ holds as claimed. □

When Q is convex, the definition of the normal cone simplifies.

Exercise 2.36. ✍ Show for any convex set $Q \subset \mathbf{E}$ and a point $\bar{x} \in Q$ the equality

$$N_Q(\bar{x}) = \{v \in \mathbf{E} : \langle v, x - \bar{x}\rangle \leq 0 \quad \text{for all } x \in Q\}.$$

[**Hint:** Appeal to Lemma 2.35.]

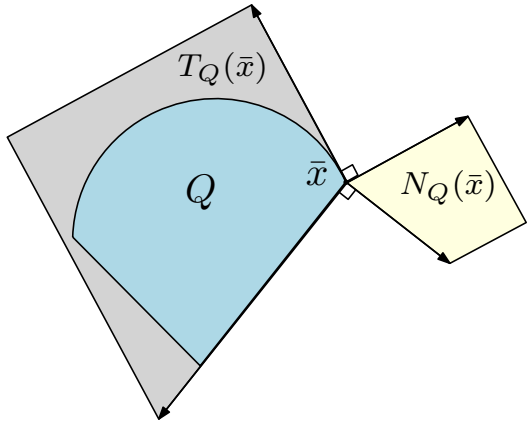


Figure 2.13: Illustration of the tangent and normal cones for a convex set.

Thus the $o(\|x-\bar{x}\|)$ error in the definition of the normal cone is irrelevant for convex set. That is, every vector $v \in N_Q(\bar{x})$ truly makes an obtuse angle with any direction $x - \bar{x}$ for $x \in Q$. Equivalently, every vector $v \in N_Q(\bar{x})$ corresponds to a halfspace $\{x : \langle v, x\rangle \leq \langle v, \bar{x}\rangle\}$ containing Q. See Figure 2.13.

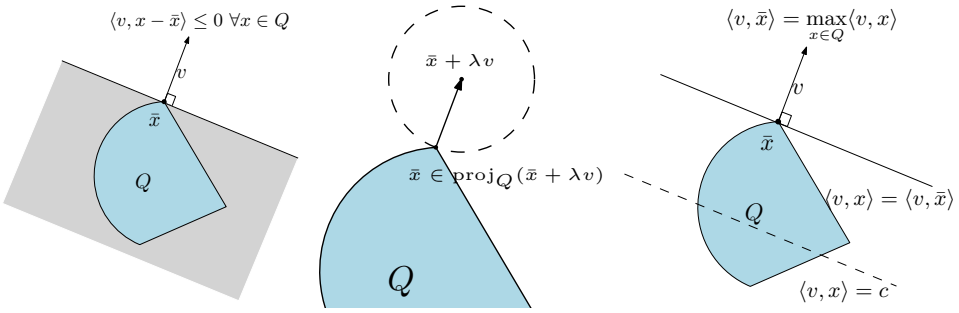
Some thought shows that normal cones and nearest-point projections are intimately related. The following exercise explores this connection and will be used extensively in the sequel; see the companion Figure 3.3.

Exercise 2.37. ✍ Prove that the following properties are equivalent for any nonempty, closed, convex set Q, a point $\bar{x} \in Q$, and a vector $v \in \mathbf{E}$.

1. $v \in N_Q(\bar{x})$,
2. $\bar{x} \in \operatorname{argmax}_{x \in Q}\langle v, x\rangle$.
3. $\operatorname{proj}_Q(\bar{x} + \lambda v) = \bar{x}$ for all $\lambda \geq 0$,
4. $\operatorname{proj}_Q(\bar{x} + \bar{\lambda} v) = \bar{x}$ for some $\bar{\lambda} > 0$.

Exercise 2.38. Show for any convex cone K and a point $x \in K$, the equality

$$N_K(x) = K^\circ \cap x^\perp.$$



(a) $v \in N_Q(\bar{x})$ (b) $\mathrm{proj}_Q(\bar{x} + \lambda v) = \bar{x}$ (c) $\bar{x} \in \mathrm{argmax}_{x \in Q}\langle v, x\rangle$

Figure 2.14: The figures depict the equivalences in Exercise 2.37.

Exercise 2.39. ✍ Show for any convex set Q and a point $x \in Q$, the equivalence

$$x \in \operatorname{int} Q \quad \Longleftrightarrow \quad N_Q(x) = \{0\}.$$

What is the relationship between normal vectors $v \in N_Q(x)$ and supporting halfspaces, defined in Exercise 2.20?

Comments

The results in Section 2.1-2.5 can be found in standard texts on convex geometry and analysis, such as Barvinok [2], Borwein-Lewis [8], Hiriart-Urruty and Lemaréchal [17], and Rockafellar [31]. The material in Section 2.6 blends the convex analytic viewpoint on tangency/normality with the more modern variation analytic perspective [33].

Chapter 3

Convex analysis

This chapter investigates analytic properties of convex functions. The material will form the backbone for the theory and algorithms presented in the rest of the book. The core principle of convex analysis is that analytic properties of a convex function are in direct correspondence with convex geometric properties of its epigraph—the set above the graph. This perspective is emphasized throughout the chapter, and should be kept in mind while reading.

Roadmap. The chapter begins by laying out basic notation and examples in Section 3.1, along with derivative-based characterizations of convexity for smooth funtions. Section 3.2 shows that convexity is preserved by a variety of functional operations. The main results in the section relate functional operations (sums, linear images/preimages) to geometric operations of epigraphs. Section 3.3 investigates the closed convex envelope of nonconvex functions. Section 3.4 takes the principle of duality in convex geometry much further by associating to a convex function its Fenchel conjugate. Section 3.5 introduces subderivatives and subgradients of a function, and shows that they are closely related to tangent and normal cones to epigraphs. Section 3.7 introduces strongly convex functions, the Moreau envelope, and the proximal map. A highlight theorem of the section is that a proper, closed, convex function is strongly convex if and only if its Fenchel conjugate is smooth, thereby formalizing the duality between the two notions. The final Section 3.6 illuminates the relationship between subgradients and Lipschitz continuity of a convex function. The main result of the section shows that a convex function is locally Lipschitz continuous at every point in the interior of its domain.

3.1 Basic definitions and examples

We will consider functions f mapping $\mathbf{E}$ to the extended-real-line $\overline{\mathbf{R}} = \mathbf{R} \cup \{\pm\infty\}$. To be completely precise, some care must be taken when working with $\pm\infty$. In particular, we set $0 \cdot \pm\infty = 0$ and will be careful to avoid expressions $(+\infty) + (-\infty)$ throughout. A function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is called *proper* if it never takes the value $-\infty$ and is finite at some point.

Given a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$, the *effective domain* and *epigraph* of f are

$$\begin{aligned} \operatorname{dom} f &:= \{x \in \mathbf{E} : f(x) < +\infty\}, \\ \operatorname{epi} f &:= \{(x, r) \in \mathbf{E} \times \mathbf{R} : f(x) \leq r\}, \end{aligned}$$

respectively. Thus $\operatorname{dom} f$ consists of all points x at which f is finite or evaluates to $-\infty$. The epigraph $\operatorname{epi} f$ is simply the set above the graph of the function on its domain. See Figure 3.1 for an illustration.

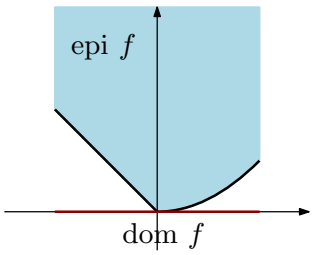


Figure 3.1: Epigraph and effective domain of the function that evaluates to $\max\{-x, \frac{1}{2}x^2\}$ on $[-1, 1]$ and to $+\infty$ elsewhere.

As we will see, much of convex analysis is guided by convex geometric properties of epigraphs.

Definition 3.1 (Convex functions)**.** A function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is *convex* if $\operatorname{epi} f$ is a convex set in $\mathbf{E} \times \mathbf{R}$.

Equivalently, a proper function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is convex if and only if the secant inequality

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

holds for all $x, y \in \mathbf{E}$ and $\lambda \in (0, 1)$. See Figure 3.2 for an illustration. In particular, for each $r \in \mathbf{R}$ the sublevel set $\{x : f(x) \leq r\}$ of a convex function is a convex set (verify this!).

We will often need to ensure that a function is proper. Though this seems like a technical annoyance, there are important settings where this

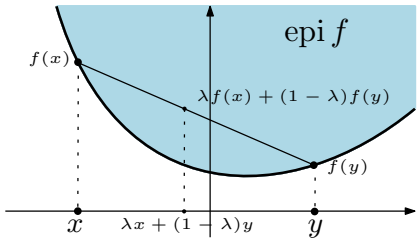


Figure 3.2: Secant inequality.

property is not for free. For example, the function $f(x) = \inf_y g(x, y)$ may be identically equal to $-\infty$ even if g is a proper function. The following exercise provides a convenient sufficient condition for a convex function to be proper.

Exercise 3.2. ✍ Let $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a convex function. Show that if there exists a point $x \in \operatorname{ri}(\operatorname{dom} f)$ with $f(x)$ finite, then f must be proper.

Exercise 3.3 (Jensen's Inequality)**.** ✍ Let $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a proper convex function. Show that the inequality $f(\sum_{i=1}^k \lambda_i x_i) \le \sum_{i=1}^k \lambda_i f(x_i)$ holds for any integer $k \in \mathbb{N}$, points $x_1, \dots, x_k \in \mathbf{E}$, and weights $\lambda \in \Delta_k$.

Exercise 3.4. ✍ Let $f_i\colon \mathbf{E} \to \overline{\mathbf{R}}$ be convex functions indexed by an arbitrary set I. Show that the the pointwise supremum function $f(x) := \sup_{i \in I} f_i(x)$ is convex.
[**Hint:** Express epi f in terms of epi f_i.]

There are a number of convex functions that naturally arise from convex sets. One example, which we have already seen, is the distance function. We now define three other useful functions associated to a set. Given a set $Q \subset \mathbf{E}$, define its *indicator function*

$$\delta_Q(x) = \begin{cases} 0, & x \in Q \\ +\infty, & x \notin Q \end{cases},$$

its *support function*

$$\delta_Q^\star(v) = \sup_{x \in Q} \langle v, x \rangle,$$

and its *gauge function*

$$\gamma_Q(x) = \inf\{\lambda \ge 0 \colon x \in \lambda Q\}.$$

The primary use of the indicator function δ_Q is to formally convert a constrained optimization problem into an unconstrained one. Namely, any optimization problem of the form $\min_{x\in Q} f(x)$ is equivalent to $\min_{x\in\mathbf{E}} f(x)+\delta_Q(x)$. The support function $\delta_Q^\star(v)$ evaluates the maximum that the linear function $\langle v,\cdot\rangle$ takes over Q. The reader might notice that we have encountered this function in Exercise 2.37. Indeed, the exercise showed the equivalence

$$v \in N_Q(x) \qquad \Longleftrightarrow \qquad \langle v, x\rangle = \delta_Q^\star(v),$$

provided Q is a closed convex set. The notation $\delta_Q^\star(v)$ may seem strange at first, since it is not clear what the support function $\delta_Q^\star(v)$ has to do with the indicator function $\delta_Q(x)$. The notation will make sense shortly, in light of Fenchel conjugacy (Section 3.4). Finally the gauge $\gamma_Q(\cdot)$ can be thought of as a generalization of a norm. In particular, the gauge of the closed unit ball of any norm is the norm itself. Norms are special among all gauges in that the set Q that generates them is centrally symmetric, meaning $Q = -Q$ (why?). The reader is referred to Figure 3.3 for an illustration of support functions and gauges.

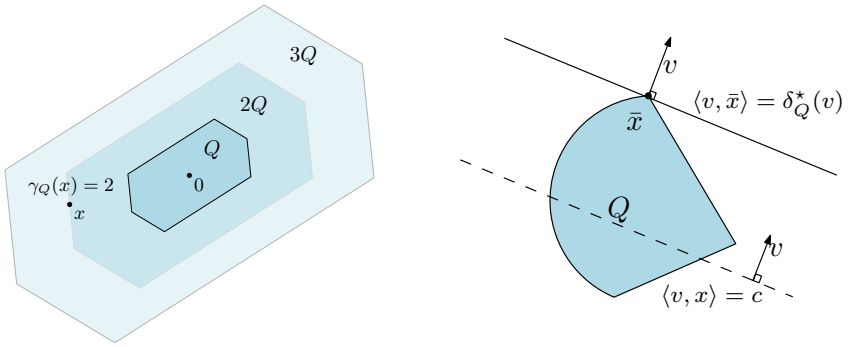


Figure 3.3: The figures depict the gauge γ_Q and the support functions $\delta_Q^\star$ of a set Q.

Exercise 3.5. ✍ Consider a set $Q \subset \mathbf{E}$,

1. Show that $\delta_Q^\star$ is a closed, convex function.
2. Show that if Q is convex, then δ_Q and dist_Q are convex
3. Show that if Q is convex, then γ_Q is convex.

Notice that support functions and gauges are positively homogeneous in the following sense.

Definition 3.6. A function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is *positively homogeneous* if its epigraph is a cone. Convex positively homogeneous functions are called *sublinear*.

The following exercise provides a useful characterization of proper sublinear functions that is analogous to Exercise 2.23.

Exercise 3.7. Let $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a proper function. Show that f is sublinear if and only if the inequality, $f(\lambda x+\mu y) \leq \lambda f(x)+\mu f(y)$, holds for all $x, y \in \mathbf{E}$ and $\lambda, \mu \geq 0$.

As noted previously, support functions and gauges of convex sets are sublinear. The forthcoming Exercise 3.21 shows essentially a converse: closed sublinear functions are support functions.

We end the section by showing that verifying convexity for a smooth function is often straightforward because it can be characterized entirely in terms of derivatives. In particular, the easiest way to verify that a smooth function is convex is to show that its Hessian is positive semi-definite everywhere. This observation opens the door to a number of interesting examples.

Theorem 3.8 (Differential characterizations of convexity)**.** *The following are equivalent for a C^1-smooth function $f\colon U \to \mathbf{R}$ defined on a convex open set $U \subset \mathbf{E}$.*

(a) **(convexity)** *f is convex.*

(b) **(gradient inequality)** *$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$ for all $x, y \in U$.*

(c) **(monotonicity)** *$\langle \nabla f(y) - \nabla f(x), y - x\rangle \geq 0$ for all $x, y \in U$.*

If f is C^2-smooth, then the following property can be added to the list:

(d) The relation $\nabla^2 f(x) \succeq 0$ holds for all $x \in U$.

Proof. Assume (a) holds, and fix two points x and y. For any $t \in (0, 1)$, convexity implies

$$f(x + t(y - x)) = f(ty + (1 - t)x) \leq tf(y) + (1 - t)f(x),$$

while the definition of the derivative yields

$$f(x + t(y - x)) = f(x) + t\langle \nabla f(x), y - x\rangle + o(t).$$

Combining the two expressions, canceling $f(x)$ from both sides, and dividing by t yields the relation

$$f(y) - f(x) \geq \langle \nabla f(x), y - x\rangle + o(t)/t.$$

Letting t tend to zero yields property (b).

Suppose now that (b) holds. Then for any $x, y \in U$, appealing to the gradient inequality, we deduce

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \qquad \text{and} \qquad f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle.$$

Adding the two inequalities yields (c).

Finally, suppose (c) holds. Define the function $\varphi(t) := f(x + t(y - x))$ and set $x_t := x + t(y - x)$. Then monotonicity shows that for any real numbers $t, s \in [0, 1]$ with $t > s$ the inequality holds:

$$\begin{aligned} \varphi'(t) - \varphi'(s) &= \langle \nabla f(x_t), y - x \rangle - \langle \nabla f(x_s), y - x \rangle \\ &= \frac{1}{t - s} \langle \nabla f(x_t) - \nabla f(x_s), x_t - x_s \rangle \geq 0. \end{aligned}$$

Thus the derivative φ' is nondecreasing, and hence for any $x, y \in U$, we have

$$f(y) = \varphi(1) = \varphi(0) + \int_0^1 \varphi'(r)\, dr \geq \varphi(0) + \varphi'(0) = f(x) + \langle \nabla f(x), y - x \rangle.$$

Some thought now shows that f admits the representation (check!)

$$f(y) = \sup_{x \in U} \{ f(x) + \langle \nabla f(x), y - x \rangle \}$$

for any $y \in U$. Since a pointwise supremum of an arbitrary collection of convex functions is convex (Excercise 3.4), we deduce that f is convex, establishing (a).

Suppose now that f is C^2-smooth. Then for any fixed $x \in U$ and $h \in \mathbf{E}$, and all small $t > 0$, property (b) and the second-order expansion (1.8) implies

$$f(x) + t\langle \nabla f(x), h \rangle \leq f(x + th) = f(x) + t\langle \nabla f(x), h \rangle + \frac{t^2}{2} \langle \nabla^2 f(x) h, h \rangle + o(t^2).$$

Canceling out like terms, dividing by t^2, and letting t tend to zero we deduce $\langle \nabla^2 f(x) h, h \rangle \geq 0$ for all $h \in \mathbf{E}$. Hence (d) holds. Conversely, suppose (d) holds. Then Theorem 1.14 immediately implies for all $x, y \in \mathbf{E}$ the inequality

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \int_0^1 \int_0^t \langle \nabla^2 f(x + s(y - x))(y - x), y - x \rangle \, ds \, dt \geq 0.$$

Hence (b) holds, and the proof is complete. □

Thus convexity for a C^1-smooth function f is equivalent to the property that the affine function $\ell(y) = f(x) + \langle \nabla f(x), y - x \rangle$ is a global underestimator of f for all x. Recall that without convexity, $\ell(\cdot)$ is an underestimator of f only up to a first-order error. Convexity is also equivalent to a monotonicity property of the gradient. In particular, convexity of a C^1-smooth univariate function f is equivalent to the derivative function $f'(\cdot)$ being nondecreasing. Finally, if f is C^2-smooth, then convexity is equivalent to the Hessian $\nabla^2 f(x)$ being positive semidefinite for all x. This property is usually the easiest to verify in examples.

Exercise 3.9. Show that negative logarithm and the negative log-determinant functions in Exercise 1.12 are convex.

Exercise 3.10. Consider a quadratic function $f(x) = \frac{1}{2}\langle \mathcal{A}x, x \rangle + \langle c, x \rangle + b$ for some self-adjoint linear operator $\mathcal{A}\colon \mathbf{E} \to \mathbf{E}$, a point $c \in \mathbf{E}$, and $b \in \mathbf{R}$. Show that f is convex if and only if $\mathcal{A}$ is positive semidefinite.

Exercise 3.11. Show that the following univariate functions are convex:

1. (Boltzmann-Shannon entropy)

$$f(x) = \begin{cases} x \log x & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ +\infty & \text{if } x < 0 \end{cases}$$

2. (Fermi-Dirac entropy)

$$f(x) = \begin{cases} x \log(x) + (1-x)\log(1-x) & \text{if } x \in (0,1) \\ 0 & \text{if } x \in \{-1, 1\} \\ +\infty & \text{otherwise} \end{cases}$$

3. (Hellinger)

$$f(x) = \begin{cases} -\sqrt{1-x^2} & \text{if } x \in [-1,1] \\ +\infty & \text{otherwise} \end{cases}$$

4. (Exponential) $f(x) = e^x$

5. (Log-exp) $f(x) = \log(1 + e^x)$

Recall from Section 1.7 that a function $f\colon \mathbf{E} \to \mathbf{R}$ is β-smooth if it is differentiable and its gradient map $x \mapsto \nabla f(x)$ is β-Lipschitz continuous. Exercise 1.15 showed that any β-smooth function f satisfies the estimate:

$$-\frac{\beta}{2}\|y-x\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y-x\rangle \leq \frac{\beta}{2}\|y-x\|^2, \tag{3.1}$$

for all $x, y \in \mathbf{E}$. This estimate will play a key role when designing algorithms. When f is convex, the left-hand-side of (3.1) can be replaced by zero. The following exercise shows that the resulting two-sided estimate *characterizes* smoothness and convexity.

Exercise 3.12. Consider a C^1-smooth function $f\colon \mathbf{R}^n \to \mathbf{R}$. Prove that each condition below holding for all points $x, y \in \mathbf{R}^n$ is equivalent to f being β-smooth and convex.

1. $0 \leq f(y) - f(x) - \langle \nabla f(x), y-x\rangle \leq \frac{\beta}{2}\|x-y\|^2$
2. $f(x) + \langle \nabla f(x), y-x\rangle + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y)$
3. $\frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x-y\rangle$
4. $0 \leq \langle \nabla f(x) - \nabla f(y), x-y\rangle \leq \beta\|x-y\|^2$

[**Hint:** Suppose first that f is convex and β-smooth. Then 1 is immediate. Suppose now 1 holds and define the function $\phi(y) = f(y) - \langle \nabla f(x), y-x\rangle$. Show using 1 that

$$\phi(x) = \min \phi \leq \phi\left(y - \frac{1}{\beta}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2\beta}\|\nabla\phi(y)\|^2.$$

Deduce the property 2. To deduce 3 from 2, add two copies of 2 with x and y reversed. Next applying Cauchy-Schwartz to 3 immediately implies that f is β-smooth and convex. Finally, show that 1 implies 4 by adding two copies of 1 with x and y reversed. Conversely, rewriting 4 deduce that the gradient of the function $\phi(x) = -f(x) + \frac{\beta}{2}\|x\|^2$ is monotone and therefore that ϕ is convex. Rewriting the gradient inequality for ϕ arrive at 1.]

3.2 Convex functions from epigraphical operations

We thus have built a small (so far) library of convex functions. Verifying convexity from the definition is tedious, as the reader has likely noticed,

and can often be avoided. The simplest way to argue that a function is convex is to recognize it as having been constructed from known convex functions (in our library) by a sequence of operations that preserve convexity. Strikingly, the typical operations one performs on functions, such as sums and compositions with linear maps, can be interpreted as set-operations on epigraphs. This viewpoint is very fruitful, since it couples convex analysis of functions to convex geometry of epigraphs. Table 3.1 records a few such functional operations and the corresponding operations on the level of epigraphs. Henceforth, for any linear map $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$, we define the linear map $\mathcal{A} \times I$ mapping $\mathbf{E} \times \mathbf{R}$ to $\mathbf{Y} \times \mathbf{R}$ by setting $(\mathcal{A} \times I)(x, r) := (\mathcal{A}x, r)$.

Function $h(x)$	**Epigraph** $\operatorname{epi} h$
$\lambda f(\frac{x}{\lambda})$	$\lambda \cdot \operatorname{epi} f$
$\sup_{i\in I} f_i(x)$	$\bigcap_{i\in I} \operatorname{epi} f_i$
$f(\mathcal{A}x)$	$[\mathcal{A} \times I]^{-1}\operatorname{epi} f$

Table 3.1: New functions from scaling ($\lambda > 0$), intersections, and preimages of epigraphs.

In particular, the intersection of epigraphs $\operatorname{epi} f_i$ for $i \in I$ is the epigraph of the pointwise supremum $h(x) = \sup_{i\in I} f_i(x)$, while the preimage of an epigraph $\operatorname{epi} f$ under a linear map $\mathcal{A} \times I$ is the epigraph of the precomposed function $h(x) = f(Ax)$.

Exercise 3.13. Establish the correspondences in Table 3.1. Deduce that if the functions f and f_i are convex, then so are the functions h in the table.

Going beyond the examples in Table 3.1, there is a convenient way to construct a convex function from sets in $\mathbf{E} \times \mathbf{R}$, which are not necessarily epigraphs. Fix a set $Q \subset \mathbf{E} \times \mathbf{R}$ and define the *lower envelope*

$$E_Q(x) := \inf\{r : (x, r) \in Q\}.$$

See Figure 3.4 for an illustration.

The following exercise follows directly from the definition of convexity.

Exercise 3.14. ✍ Let $Q \subset \mathbf{E} \times \mathbf{R}$ be a set. Establish the following statements.

1. If the infimum in the definition of $E_Q(x)$ is attained at every point x where it is finite, then equality holds:

$$\operatorname{epi} E_Q = Q + (\{0\} \times \mathbf{R}_+)\,.$$

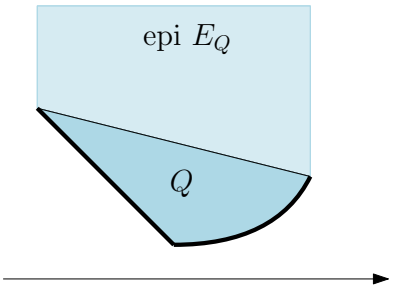


Figure 3.4: Lower envelope of Q.

What goes wrong if the infimum is not attained at some x with $E_Q(x)$ finite?

2. If Q is convex, then the envelope $E_Q \colon \mathbf{E} \to \overline{\mathbf{R}}$ is a convex function.

A primary example of the lower envelope construction arises from minimizing out a function in one of its variable. Namely, for any bivariate function $g \colon \mathbf{E} \times \mathbf{Y} \to \overline{\mathbf{R}}$, the function

$$h(x) := \inf_y \; g(x, y) \tag{3.2}$$

is called the *infimal projection* of g. A quick computation (do it!) shows that we may equivalently write

$$h(x) = \inf\{r : (x, r) \in \pi_{x,r}(\operatorname{epi} g)\}, \tag{3.3}$$

where $\pi_{x,r}$ is the canonical projection $\pi_{x,r}(x, y, r) = (x, r)$. Thus f is the lower envelope generated by the convex set $Q := \pi_{x,r}(\operatorname{epi} g)$. Using Exercise 3.14, we deduce that if g is convex, then so is h. Moreover, if the infimum in the definition (3.2) is attained at any x at which $h(x)$ is finite, then equality holds:

$$\operatorname{epi} h = \pi_{x,r}(\operatorname{epi} g).$$

An important example of the infimal projection, is the *infimal convolution* of two functions $f, g \colon \mathbf{E} \to \overline{\mathbf{R}}$ defined as

$$(f \,\square\, g)(x) := \inf_y \{f(y) + g(x - y)\}. \tag{3.4}$$

The name infimal convolution comes from the similarity of the construction with the integral convolution, where the integral and multiplication have

been replaced by infimum and sum, respectively. A quick computation (do it!) shows that we may equivalently write

$$(f \,\square\, g)(x) = \inf\{r : (x, r) \in \operatorname{epi} f + \operatorname{epi} g\}.$$

Hence, the infimal convolution is the lower envelope of $Q := \operatorname{epi} f + \operatorname{epi} g$. Using Exercise 3.14, we deduce that if f and g are convex, then so is the convolution $f \,\square\, g$. Moreover, if the infimum in the definition (3.4) is attained at any x at which $(f \,\square\, g)(x)$ is finite, then equality holds:

$$\operatorname{epi}(f \,\square\, g) = \operatorname{epi} f + \operatorname{epi} g. \tag{3.5}$$

Thus, under a mild regularity condition, addition of epigraphs corresponds to the infimal convolution operation.

As the final example of the epigraphical projection, fix a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a linear map $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$. Define the function

$$h(x) := \min_y \ \{f(y) : \mathcal{A}y = x\} \tag{3.6}$$

Notice that h is the infimal projection of $g(x, y) = f(y) + \delta_{\{0\}}(x - \mathcal{A}y)$. Unraveling notation, the reader should verify the equivalent representation

$$h(x) = \inf\{r : (x, r) \in (\mathcal{A} \times I)\operatorname{epi} f\}.$$

Hence h is the lower envelope of the linear image $Q := (\mathcal{A} \times I)\operatorname{epi} f$. Using Exercise 3.14, we deduce that if f is convex, then so is the value function h. Moreover, if the infimum in the definition (3.6) is attained at any x at which $h(x)$ is finite, then equality holds:

$$\operatorname{epi} h = (\mathcal{A} \times I)\operatorname{epi} f.$$

Table 3.2 summarizes the three examples (3.2), (3.4), and (3.6).

Function $h(x)$	**Lower envelope of** $Q \subset \mathbf{E} \times \mathbf{R}$
$\inf_y \ f(x, y)$	$\pi_{x,r}\operatorname{epi} f$
$\min_y\{f(y) : \mathcal{A}y = x\}$	$[\mathcal{A} \times I]\operatorname{epi} f$
$\inf_y \ f(y) + g(x - y)$	$\operatorname{epi} f + \operatorname{epi} g$

Table 3.2: New functions from sums, projections, and images of epigraphs.

A cautionary warning is now in order. Recall from Section 1.5 that we call a function f closed if $\operatorname{epi} f$ is a closed set. As discussed in Section 2.1,

linear images and sums of closed sets might not be closed. By the same token, the infimal convolution and epigraphical projection of closed functions might not be closed. Consequently, we will have to be careful with closure issues when dealing with these two functional operations.

Exercise 3.15. Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and define the set

$$Q := \mathbf{R}_{++}(\{1\} \times \operatorname{epi} f).$$

1. Show that Q is the epigraph of the *perspective function* $f^\pi\colon \mathbf{R}\times\mathbf{E} \to \overline{\mathbf{R}}$ defined by
$$f(\lambda, x) = \begin{cases} \lambda f(\frac{x}{\lambda}) & \text{if } \lambda > 0 \\ +\infty & \text{otherwise} \end{cases}.$$
Deduce that if f is convex, then so is f^π.

2. Define the quadratic-over-linear function
$$g(y) = \begin{cases} \frac{\|\mathcal{A}y-b\|^2}{\langle c,y\rangle - q} & \text{if } \langle c, y\rangle > q \\ +\infty & \text{otherwise} \end{cases}.$$
for some linear map $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$, points $b \in \mathbf{Y}$ and $c \in \mathbf{E}$, and $q \in \mathbf{R}$. Use the previous part of the exercise to show that g is convex.

[**Hint:** Write g as the perspective function of $\|\cdot\|^2$ composed with a linear map.]

3.3 The closed convex envelope

When encountering a nonconvex and possibly nonclosed function, it is convenient to introduce the following definition, which plays the role of the closed convex hull in convex geometry.

Definition 3.16 (Closed convex envelope)**.** The *closed convex envelope* of a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is the function whose epigraph is the closed convex hull of $\operatorname{epi} f$, and will be denoted by $\overline{\operatorname{co}}\, f$.

The reader should verify that the set $\operatorname{cl}(\operatorname{conv}(\operatorname{epi} f))$ is indeed an epigraph of some function. Though the definition of the closed convex envelope is geometrically pleasing, it is not convenient for computation. A better description arises from minorants, which facilitate a dual representation of the functions that is analogous to the dual representation of convex sets (Theorem 2.21)

Definition 3.17 (Minorants)**.** Given two functions f and g on $\mathbf{E}$, we say that g is a *minorant* of f if it satisfies $g(y) \leq f(y)$ for all $y \in \mathbf{E}$. If in addition, g has the form $g(x) = \langle a, x \rangle + b$ for some $a \in \mathbf{E}$ and $b \in \mathbf{R}$, then we call g an *affine minorant* of f. In the case $b = 0$, we call g a *linear minorant* of f.

The following theorem is a direct analogue of Theorem 2.21. Namely, Theorem 2.21 shows that we may write $\operatorname{epi}(\overline{\operatorname{co}} f)$ as an intersection of halfspaces in $\mathbf{E} \times \mathbf{R}$. The following theorem shows that the vertical halfspaces of the form $\{(x, r) : \langle a, x \rangle = b\}$ can be discarded from this dual description, and therefore that the envelope $\overline{\operatorname{co}} f$ can be written as the pointwise supremum of affine minorants of f. See Figure 3.5 for an illustration.

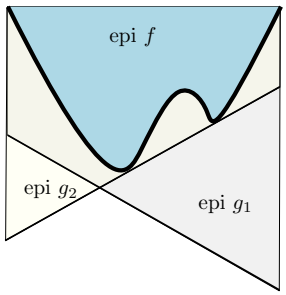


Figure 3.5: Affine envelope representation of f; the functions g_1 and g_2 are affine minorants of f.

Theorem 3.18 will play a crucial role when we discuss Fenchel conjugacy—a central topic in convex analysis.

Theorem 3.18 (Affine envelope representation)**.** *A proper function $f : \mathbf{E} \to \overline{\mathbf{R}}$ admits an affine minorant if and only if $(\overline{\operatorname{co}} f)$ is proper. Under these two equivalent conditions, equality holds:*

$$(\overline{\operatorname{co}} f)(x) = \sup\{g(x) : g \colon \mathbf{E} \to \overline{\mathbf{R}} \text{ is an affine minorant of } f\}. \tag{3.7}$$

Proof. Define the set $Q := \operatorname{cl}(\operatorname{conv}(\operatorname{epi} f))$. Clearly if f admits an affine minorant, then $\overline{\operatorname{co}}\ f$ never takes the value $-\infty$ and is therefore proper. Henceforth, we assume that $(\overline{\operatorname{co}} f)$ is proper. We will show that f admits at least one affine minorant and that (3.7) holds, thereby completing the proof. Applying Theorem 2.21, we deduce that Q can be written as an intersection of halfspaces in $\mathbf{E} \times \mathbf{R}$. Observe that one of the halfspaces in this representation must be nonvertical; otherwise, Q would be a union of vertical lines, thereby contradicting that $\overline{\operatorname{co}} f$ is proper. Let us write this

nonvertical halfspace as the epigraph of an affine minorant g_1 of f; we will use this function shortly.

Let $h(\cdot)$ be the function defined on the right-hand-side of (3.7). We will show that $(\overline{\mathrm{co}}\, f)$ and h have the same epigraphs. Since h is a pointwise supremum, we may write $\operatorname{epi} h$ as an intersection of halfspaces:

$$\operatorname{epi} h = \bigcap \{\operatorname{epi} g : g\colon \mathbf{E} \to \overline{\mathbf{R}} \text{ is an affine minorant of } f\}.$$

In particular, the inclusion $Q \subset \operatorname{epi} h$ clearly holds. Suppose now for the sake of contradiction that there exists a point $(\bar{x}, \bar{r}) \in \operatorname{epi} h$ that is not in Q. The separation theorem (Theorem 2.19) yields $(a, \mu) \in \mathbf{E} \times \mathbf{R}$ and $b \in \mathbf{R}$ such that the halfspace

$$H = \{(x, r) : \langle (a, \mu), (x, r) \rangle \leq b\}$$

contains Q and does not contain $(\bar{x}, \bar{r})$. By the nature of epigraphs, the inequality $\mu \leq 0$ holds (why?). If $\mu < 0$, then H is nonvertical, thereby contradicting the definition of h. Thus, we may assume $\bar{\mu} = 0$. The strategy now is to perturb H by using g_1 in order to make it nonvertical. To this end, define the function $g_2(x) := \langle a, x \rangle - b$ and observe $H = \{(x, r) : g_2(x) \leq 0\}$. In particular, every point $x \in \operatorname{dom} f$ satisfies $g_2(x) \leq 0$. We therefore deduce

$$\lambda g_2(x) + g_1(x) \leq f(x)$$

for all points $x \in \mathbf{E}$ and any $\lambda > 0$. Thus the function $g_3(x) := \lambda g_2(x) + g_1(x)$ is an affine minorant of f. Taking into account $g_2(\bar{x}) > 0$, we arrive at the contradiction

$$h(\bar{x}) \geq g_3(\bar{x}) = \lambda g_2(\bar{x}) + g_1(\bar{x}),$$

when $\lambda > 0$ is sufficiently large. □

The following exercise shows that for positively homogeneous functions, one may replace affine functions in the envelope description of Theorem 3.18 by linear functions.

Exercise 3.19. ✍ Consider a proper positively homogeneous function $h\colon \mathbf{E} \to \overline{\mathbf{R}}$. Show that h admits a linear minorant if and only if $\overline{\mathrm{co}}\, h$ is proper. Show that under these two equivalent conditions, equality holds:

$$(\overline{\mathrm{co}}\, h)(x) = \sup\{g(x) : g\colon \mathbf{E} \to \overline{\mathbf{R}} \text{ is a linear minorant of } h\}.$$

Exercise 3.20. Show that if $h\colon \mathbf{R} \to \overline{\mathbf{R}}$ is a proper positively homogeneous function, then equality holds

$$(\overline{\mathrm{co}}\, h) + \delta_{\mathbb{B}} = \overline{\mathrm{co}}\, (h + \delta_{\mathbb{B}}).$$

Deduce the expression $\inf\limits_{\|x\|\leq 1} h(x) = \inf\limits_{\|x\|\leq 1} (\overline{\mathrm{co}}\, h)(x)$.

The following exercise is a nice consequence of Exercise 3.19. It shows that any closed sublinear function is a support function of some set. This result will underly a key duality relationship between subdifferentials and subderivatives in Section 3.5.

Exercise 3.21. ✍ Let $h\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a proper, positively homogeneous function. Define the set

$$Q = \{x : \langle x, y\rangle \leq h(y) \ \ \forall y \in \mathbf{E}\}.$$

Show that $\overline{\mathrm{co}}\ h$ is proper if and only if Q is nonempty. Prove that under these two equivalent conditions, $\overline{\mathrm{co}}\ h$ is the support function of Q.
[**Hint:** Notice that Q parametrizes the set of linear minorants of h. Therefore, the support function of Q is the pointwise supremum of all linear minorants of h. Complete the proof by appealing to Exercise 3.19.]

3.4 The Fenchel conjugate

In convex geometry, one could associate with any convex cone its polar. Convex analysis takes this idea much further through a new operation on functions, called Fenchel conjugacy. Indeed, this construction subsumes all notions of duality we have encountered so far.

Definition 3.22. For a function $f : \mathbf{E} \to \overline{\mathbf{R}}$, define the *Fenchel conjugate* function $f^\star\colon \mathbf{E} \to \overline{\mathbf{R}}$ by

$$f^\star(y) = \sup_{x\in\mathbf{E}} \{\langle y, x\rangle - f(x)\}.$$

Though the definition might seem strange at first, this operation arises naturally from epigraphical geometry. From the very definition of the Fenchel conjugate, the epigraph $\operatorname{epi} f^\star$ consists of all pairs (y, r) satisfying $f(x) \geq \langle y, x\rangle - r$ for all points x. Thus $\operatorname{epi} f^\star$ encodes all affine minorants $x \mapsto$

$\langle y, x\rangle - r$ of f. An alternate insightful interpretation is through the support function to the epigraph. Observe

$$\begin{aligned} f^\star(y) &= \sup_{x\in\mathbf{E}}\{\langle (y,-1),(x,f(x))\rangle\} \\ &= \sup_{(x,r)\in \operatorname{epi} f}\{\langle (y,-1),(x,r)\rangle\} \\ &= \delta^\star_{\operatorname{epi} f}(y,-1). \end{aligned}$$

Thus the conjugate $f^\star(y)$ is exactly the support function of $\operatorname{epi} f$ evaluated at $(y,-1)$. Since the support function is sublinear, the appearance of -1 in the last coordinate simply serves as a normalization constant. See Figure 3.6 for an illustration of the two outlined viewpoints.

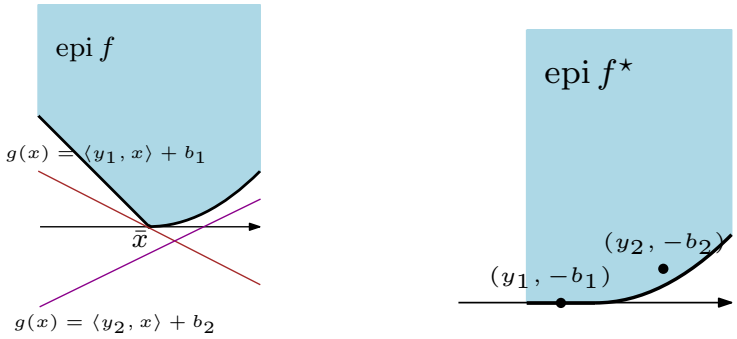


Figure 3.6: Epigraphs of the function $f = \max\{-x, \frac{1}{2}x^2\}$ and its conjugate $f^\star = \frac{1}{2}\max\{0,x\}^2 + \delta_{[-1,\infty)}(x)$.

Notice that $f^\star$ is a pointwise supremum of affine functions and is therefore closed and convex. Moreover, one can verify that $f^\star$ is proper as long as $\overline{\operatorname{co}} f$ is proper (check this!). Table 3.3 records the Fenchel conjugate of a few notable univariate functions.

We can now explain the symbol $\delta^\star_Q$ we have been using for the support function of a set Q. Observe that the Fenchel conjugate of the indicator function δ_Q is exactly the support function of Q, thereby explaining the notation $\delta^\star_Q$ for the latter.

Exercise 3.23. Show that the set $\mathcal{F}_Q$ of halfspaces in Theorem 2.21 coincides with $\operatorname{epi}\delta^\star_Q$.

We aim to show now that much like taking the double polar of a cone, taking the Fenchel conjugate twice results in the closed convex envelope of the function. As the first, step, the following exercise verifies this claim for affine functions.

$f(x)$	dom f	$f^\star(y)$	dom $f^\star$
$\lvert x\rvert$	$\mathbf{R}$	0	$[-1,1]$
$\frac{1}{p}\lvert x\rvert^p,\ \ p>1$	$\mathbf{R}$	$\frac{1}{q}\lvert y\rvert^q \quad (\frac{1}{p}+\frac{1}{q}=1)$	$\mathbf{R}$
$\sqrt{1+x^2}$	$\mathbf{R}$	$-\sqrt{1-y^2}$	[-1,1]
$-\log(x)$	$(0,\infty)$	$-1-\log(-y)$	$(-\infty,0)$
e^x	$\mathbf{R}$	$y\log(y)-y$	$[0,\infty)$
$x\log(x)$	$(0,\infty)$	e^y-1	$\mathbf{R}$
$\log(1+e^x)$	$\mathbf{R}$	$y\log(y)+(1-y)\log(1-y)$	$[0,1]$

Table 3.3: Examples of convex functions and their Fenchel conjugates.

Exercise 3.24. ✍ Show that for any affine function $f(x)=\langle a,x\rangle+b$, the conjugate takes the form $f^\star(y)=-b+\delta_{\{a\}}(y)$. Deduce the equality $f^{\star\star}=f$.

We will also use the following elementary observation.

Exercise 3.25. ✍ For any function $g\colon \mathbf{E}\times\mathbf{Y}\to\overline{\mathbf{R}}$, the estimate holds:

$$\sup_y \inf_x\ g(x,y)\le \inf_x \sup_y\ g(x,y).$$

We can now prove the biconjugacy theorem alluded to previously.

Theorem 3.26 (Biconjugacy)**.** *For any proper function $f\colon \mathbf{E}\to\overline{\mathbf{R}}$, equality $f^{\star\star}=\overline{\mathrm{co}}f$ holds.*

Proof. We begin by successively computing:

$$\begin{aligned}
(f^\star)^\star(x) &= \sup_y\{\langle x,y\rangle - f^\star(y)\}\\
&= \sup_y\{\langle x,y\rangle - \sup_z\{\langle z,y\rangle - f(z)\}\}\\
&= \sup_y \inf_z\{\langle y,x-z\rangle + f(z)\}\\
&\le \inf_z \sup_y\{\langle y,x-z\rangle + f(z)\} \qquad (3.8)\\
&= \inf_z \begin{cases} f(z) & x=z\\ +\infty & x\neq z\end{cases}\\
&= f(x),
\end{aligned}$$

where (3.8) follows from Exercise 3.25.

Thus we have established $f^{\star\star} \leq f$. Notice that $f^{\star\star}$ is by definition closed and convex. Hence the inequality $f^{\star\star} \leq \overline{\text{co}} f$ holds. To complete the proof, let $g(\cdot)$ be any affine minorant of f. By the definition of the conjugate, we see that conjugacy is order reversing and hence $g^\star \geq f^\star$. Exercise 3.24 then yields $g = (g^\star)^\star \leq (f^\star)^\star$. Taking the supremum over all affine minorants g of f and using Exercise 3.18 implies $\overline{\text{co}} f \leq f^{\star\star}$, as claimed. □

The biconjugacy theorem incorporates many duality ideas we have already seen in convex geometry. For example, let K be a nonempty cone. It is immediate from the definition of conjugacy that $\delta_K^\star = \delta_{K^\circ}$. Consequently, Theorem 3.26 shows

$$\delta_{\text{cl conv} K} = \overline{\text{co}}\ \delta_K = (\delta_K)^{\star\star} = \delta_{K^\circ}^\star = \delta_{K^{\circ\circ}}.$$

Hence we deduce $K^{\circ\circ} = \text{cl conv} K$. This is exactly the conclusion of Exercise 2.27.

Exercise 3.27. Show that for any closed, convex set $Q \subset \mathbf{E}$ containing the origin, it holds:

$$\gamma_Q(x) = \delta_{Q^\circ}^\star(x).$$

The Fenchel conjugacy interacts beautifully with the epigraphical set operations, as summarized in the Table 3.4.

Function $h(x)$	**Fenchel conjugate** $h^\star(y)$
$\lambda f(x)$	$\lambda f^\star(\frac{y}{\lambda})$
$f(x+b)$	$f^\star(y) - \langle b, y\rangle$
$\inf_z g(x,z)$	$g^\star(y,0)$
$f \,\square\, g$	$f^\star + g^\star$
$\inf_y\{f(y) : \mathcal{A}y = x\}$	$f^\star(\mathcal{A}y)$

Table 3.4: Duality in epigraphical operations; here, f and g are proper convex functions and $\lambda > 0$.

Exercise 3.28. Verify the correspondences in Table 3.4.

3.5 Subgradients and subderivatives

Section 2.6 introduced the tangent cone as a first-order approximation of a convex set, along with its polar, the normal cone. This section follows

a similar theme for an arbitrary function f, introducing the subderivatives $df(x)(\cdot)$ (analogue of directional derivative) and the subdifferential $\partial f(x)$ (set of generalized gradient). There is a tight connection between these two constructions and epigraphical geometry: the epigraph of the subderivative $df(x)$ is the tangent cone to the epigraph $T_{\operatorname{epi} f}(x, f(x))$, while the subdifferential $\partial f(x)$ can be identified with a slice of the normal cone $N_{\operatorname{epi} f}(x, f(x))$. The main result of the section establishes the following striking relationship: the subderivative coincides with the support function of the subdifferential, provided the latter is nonempty. The following two sections investigate the subdifferential and subderivative constructions, in turn.

3.5.1 Subdifferential

By their nature, nonsmooth functions cannot be approximated by affine functions up to a first-order error. Optimization problems, however, are inherently one-sided, that is one aims to either minimize or maximize a function. Consequently, it seems plausible that theory and algorithms can use one-sided affine approximations of functions. With this in mind, we introduce the following key definition, which simply replaces the equality in the definition of the gradient by an inequality.

Definition 3.29 (Subdifferential)**.** Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point x, with $f(x)$ finite. Then a vector $v \in \mathbf{E}$ is called a *subgradient* of f at x if the inequality holds:

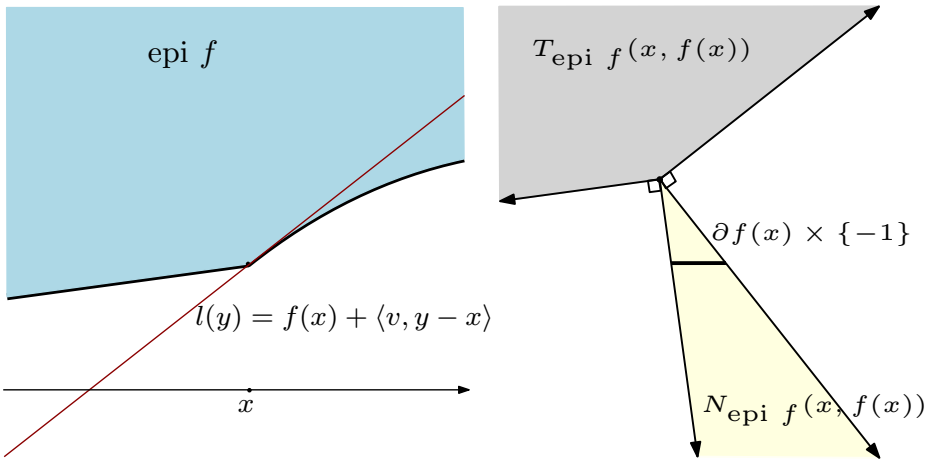
$$f(y) \geq f(x) + \langle v, y - x\rangle + o(\|y - x\|) \quad \text{as } y \to x. \tag{3.9}$$

The set of all such vectors v is called the *subdifferential of f at x*, and is denoted by $\partial f(x)$. For points x at which $f(x)$ is infinite, we set $\partial f(x) = \emptyset$.

Thus a vector v is a subgradient of f at x if the affine function $y \mapsto f(x) + \langle v, y - x\rangle$ minorizes f up to first-order near x; see Figure 3.7a. The set $\partial f(x)$ should therefore be thought of as a set of generalized gradients. In particular, the inclusion $0 \in \partial f(x)$ is a necessary condition for x to be a local minimizer.

Lemma 3.30 (Necessary condition for optimality)**.** *Let x be a local minimizer of a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$, and suppose that $f(x)$ is finite. Then the inclusion $0 \in \partial f(x)$ holds.*

Reassuringly, for differentiable functions, the subdifferential consists only of the gradient.



(a) Subgradient of nonconvex function (b) Tangents and normals

Figure 3.7: Subdifferential of a nonconvex function.

Exercise 3.31 (Subdifferential of differentiable functions)**.** ✍ Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ that is differentiable at x. Then equality $\partial f(x) = \{\nabla f(x)\}$ holds.
[**Hint:** The inclusion $\nabla f(x) \in \partial f(x)$ is clear from the definition of the gradient. Conversely, fix any subgradient $v \in \partial f(x)$ and a vector $h \in \mathbf{E}$.. Then for all small $t > 0$, the definition of the subgradient implies

$$f(x + th) \geq f(x) + t\langle v, h\rangle + o(t),$$

while differentiability yields

$$f(x + th) = f(x) + t\langle \nabla f(x), h\rangle + o(t).$$

Combine the two estimates to deduce $\nabla f(x) = v$.]

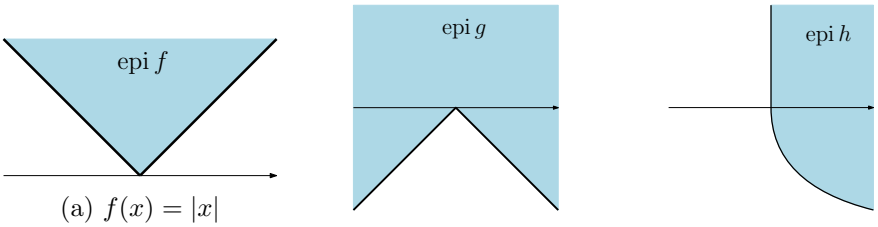
Let us look at a few simple examples to get a feeling for subdifferential computations. More interesting examples will appear at the end of the section, after we develop some finer tools.

Exercise 3.32. Define the univariate functions

$$f(x) := |x|, \qquad g(x) := -|x|, \qquad \text{and} \qquad h(x) := \begin{cases} -\sqrt{x} & \text{if } x \geq 0 \\ \infty & \text{if } x < 0 \end{cases}.$$

Show that $\partial f(x)$, $\partial g(x)$, and $\partial h(x)$ are given by the following expressions, respectively:

$$\begin{Bmatrix} -1 & \text{if } x < 0 \\ [-1,1] & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{Bmatrix}, \qquad \begin{Bmatrix} 1 & \text{if } x < 0 \\ \emptyset & \text{if } x = 0 \\ -1 & \text{if } x > 0 \end{Bmatrix}, \qquad \begin{Bmatrix} \emptyset & \text{if } x \leq 0 \\ \frac{-1}{2\sqrt{x}} & \text{if } x > 0 \end{Bmatrix}.$$



(a) $f(x) = |x|$ (b) $g(x) = -|x|$ (c) $h(x) = -\sqrt{x}$

Figure 3.8: Examples of subdifferentials.

Computing subdifferentials of more interesting functions, such as pointwise maxima $f(x) = \sup_{i=1,\dots,n} f_i(x)$, sums $f_1 + f_2$, and compositions $f(\mathcal{A}x)$, requires a calculus of subderivatives. Chapter 4 is devoted entirely to developing such a calculus for subdifferentials of convex functions. Consequently, subdifferential formulas for many more interesting functions will appear in Chapter 4. The following exercise develops two elementary calculus rules that we can already prove, and which will be useful in the sequel.

Exercise 3.33 (Separable sum)**.** Let $f_i \colon \mathbf{E}_i \to \overline{\mathbf{R}}$, for $i = 1, \dots, k$, be a proper function on a Euclidean spaces $\mathbf{E}_i$. Show that the subdifferential of the separable function $f(x_1, x_2, \dots, x_n) = \sum_{i=1}^{k} f_i(x_i)$ is given by

$$\partial f(x_1, x_2, \dots, x_n) = \partial f(x_1) \times \dots \times \partial f(x_k).$$

Deduce an expression for the subdifferential of the ℓ_1 norm $\|\cdot\|_1$.

Exercise 3.34 (Smooth plus nonsmooth)**.** ✍ Consider two proper functions $f, g \colon \mathbf{E} \to \overline{\mathbf{R}}$ and suppose that f is differentiable at a point x. Show the equality

$$\partial(f + g)(x) = \nabla f(x) + \partial g(x).$$

Given a set $Q \subset \mathbf{E}$, observe the equality $\partial \delta_Q(x) = N_Q(x)$ (check this!). Hence, the normal cone is an example of a subdifferential. The relationship between normal cones and subdifferentials runs much deeper, as the following exercise shows.

Theorem 3.35. ✍ *Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point x, with $f(x)$ finite. Then the equivalence holds:*

$$v \in \partial f(x) \qquad \Longleftrightarrow \qquad (v, -1) \in N_{\operatorname{epi} f}(x, f(x)).$$

Proof. The forward implication $\Longrightarrow$ is immediate from definitions; we leave the details for the reader to verify. Conversely, fix a pair $(v, -1) \in N_{\operatorname{epi} f}(x, f(x))$. The definition of the normal cone guarantees

$$r \geq f(x) + \langle v, y - x\rangle + o(\|(y, r)) - (x, f(x))\|), \tag{3.10}$$

as $(y, r) \to (x, f(x))$ in $\operatorname{epi} f$. Let us first argue that the quotients $\frac{f(y)-f(x)}{\|y-x\|}$ are bounded from below on a neighborhood of x. To see this, observe that (3.10) guarantees

$$r \geq f(x) + \langle v, y - x\rangle - \tfrac{1}{2}(\|y - x\| + |r - f(x)|), \tag{3.11}$$

for all (y, r) in $\operatorname{epi} f$ sufficiently close to $(x, f(x))$. In particular, if $r < f(x)$, then rearranging (3.11) and using the Cauchy-Schwarz inequality yields

$$r \geq f(x) + 2\langle v, y - x\rangle - \|y - x\| \geq f(x) - (1 + 2\|v\|)\|x - y\|. \tag{3.12}$$

We therefore deduce $\operatorname{liminf}_{y\to x} \frac{f(y)-f(x)}{\|y-x\|} > -\infty$, as claimed.

Seeking to verify the subgradient definition (3.9), consider any sequence $y_i \to x$, with $f(y_i)$ finite. Clearly, we may assume $f(y_i) \leq f(x) + \langle v, y_i - x\rangle$ for all indices i, since otherwise (3.9) holds trivially. The Cauchy-Schwarz inequality therefore guarantees $\operatorname*{limsup}_{i\to\infty} \frac{f(y_i)-f(x)}{\|y_i-x\|} < \infty$. Thus the quotients $\frac{|f(y_i)-f(x)|}{\|y_i-x\|}$ are uniformly bounded. Plugging in $r_i = f(y_i)$ into (3.10) yields

$$f(y_i) - f(x) - \langle v, y_i - x\rangle \geq o(\|(y_i, f(y_i)) - (x, f(x))\|). \tag{3.13}$$

Observe now that the vectors $\zeta_i := (y_i, f(y_i)) - (x, f(x))$ satisfy

$$\frac{o(\|\zeta_i\|)}{\|y_i-x\|} = \frac{o(\|\zeta_i\|)}{\|\zeta_i\|} \cdot \frac{\|\zeta_i\|}{\|y_i-x\|} = \frac{o(\|\zeta_i\|)}{\|\zeta_i\|} \cdot \sqrt{1 + \tfrac{(f(y_i)-f(x))^2}{\|y_i-x\|^2}} \to 0.$$

Therefore, the term $o(\|\zeta_i\|)$ on the right side of (3.13) is $o(\|y_i - x\|)$, which directly implies $v \in \partial f(x)$, as claimed. □

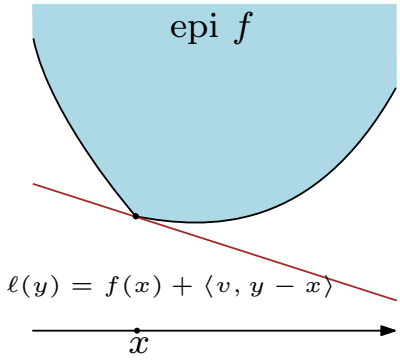
Thus, we may identify the subdifferential $\partial f(x)$ with the slice of the normal cone $N_{\operatorname{epi} f}(x, f(x)) \cap (\mathbf{E} \times \{-1\})$. See Figure 3.7b for an illustration. An important consequence of Theorem 3.35 is that the subdifferential $\partial f(x)$ is always a closed convex set. When the target function f is convex, the definition of the subdifferential simplifies, as one expects.

Exercise 3.36. ✍ Consider a convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point x, with $f(x)$ finite. Show that the inclusion $v \in \partial f(x)$ holds if and only if

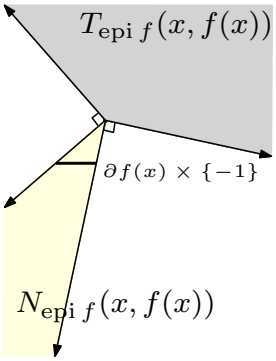
$$f(y) \geq f(x) + \langle v, y - x \rangle \qquad \text{holds for all } y \in \mathbf{E}.$$

[**Hint:** Use Exercises 2.36 and 3.35.]

Thus if f is convex, then a subgradient $v \in \partial f(x)$ has the property that the affine function $y \mapsto f(x) + \langle v, x - y \rangle$ globally minorizes f, with no error term. See Figure 3.9 for an illustration. This is the power of convexity and should be emphasized: subgradients of convex functions automatically furnish global affine minorants. In particular, the zero vector is a subgradient of a convex function f at x if and only if x is a global minimizer of f.



(a) Subgradient of a convex function



(b) Tangents and normals

Figure 3.9: Subdifferential of a convex function.

Corollary 3.37. *A point x is a global minimizer of a proper convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ if and only if the inclusion $0 \in \partial f(x)$ holds.*

Certainly if the subdifferential is empty at some point, then it provides no useful information. Looking back at Exercise 3.32, we see that a nonconvex function (e.g. $f(x) = -|\cdot|$) could easily have an empty subdifferential set at a point in the interior of its domain. In contrast, the following exercise shows that any convex function admits a subgradient at every point in the relative interior of its domain. This result is sharp in the sense that even a convex function may fail to admit a subgradient on the boundary of its domain. One example is the function $h(x) = -\sqrt{x}$ in Exercise 3.32.

Theorem 3.38 (Existence of subgradients)**.** *Consider a proper convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$. Then the subdifferential $\partial f(x)$ is nonempty at every point $x \in \operatorname{ri}(\operatorname{dom} f)$.*

Proof. We first prove the theorem when $\operatorname{dom} f$ has nonempty interior. For any point $x \in \operatorname{int}(\operatorname{dom} f)$, Exercise 2.39 yields a nonzero vector $(v, \alpha) \in N_{\operatorname{epi} f}(x, f(x))$. In the case $\alpha = 0$, the very definition of the normal cone guarantees $v \in N_{\operatorname{dom} f}(\bar{x})$, which is not possible since $\bar{x}$ lies in $\operatorname{int}(\operatorname{dom} f)$. Therefore we deduce $\alpha < 0$. Rescaling yields the inclusion $(-\alpha^{-1}v, -1) \in N_{\operatorname{epi} f}(x, f(x))$. Theorem 3.35 therefore guarantees $-\alpha^{-1}v \in \partial f(x)$, thereby completing the proof.

Suppose now that $\operatorname{dom} f$ has empty interior. Without loss of generality, we may assume that the domain of f contains the origin. Define the linear subspace $L := \operatorname{aff}(\operatorname{dom} f)$ and define the function $g\colon L \times L^{\perp} \to \overline{\mathbf{R}}$ by $g(x, y) = g(x)$. Clearly, g is convex and its domain $\operatorname{dom} g = \operatorname{dom} f \times L^{\perp}$ has nonempty interior. Therefore, by what we have already proved, for every point $x \in \operatorname{ri}(\operatorname{dom} f)$, the subdifferential set $\partial g(x, 0) = \partial f(x) \times \{0\}$ is nonempty. The proof is complete. □

The subdifferential and the Fenchel conjugate function are intimately linked by the following theorem; see the accompanying Figure 3.10.

Theorem 3.39 (Fenchel-Young Inequality)**.** *Consider a proper convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$. Then for any points $x, y \in \mathbf{E}$, the inequality*

$$f(x) + f^{\star}(y) \geq \langle x, y \rangle \qquad \text{holds,} \tag{3.14}$$

while equality holds if and only if $y \in \partial f(x)$.

Proof. Fix two points $x, y \in \mathbf{E}$. Observe

$$f^{\star}(y) = \sup_z \{\langle z, y \rangle - f(z)\} \geq \langle x, y \rangle - f(x),$$

establishing the claimed inequality (3.14). Next, Exercise 3.36 shows that the inclusion $y \in \partial f(x)$ holds if and only if $f(z) \geq f(x) + \langle y, z - x \rangle$ for all z, or equivalently if $\langle y, x \rangle - f(x) \geq \langle y, z \rangle - f(z)$ for all z. Taking supremum over z, this condition amounts to

$$\langle y, x \rangle - f(x) \geq \sup_z \{\langle y, z \rangle - f(z)\} = f^{\star}(y).$$

This is the reverse direction in the inequality (3.14). □

A crucial consequence of the Fenchel-Young inequality is that the conjugacy operation acts as an inverse on the level of subdifferentials.

Corollary 3.40. *Suppose $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is proper, closed, and convex. Then*

$$y \in \partial f(x) \quad \Longleftrightarrow \quad x \in \partial f^{\star}(y).$$

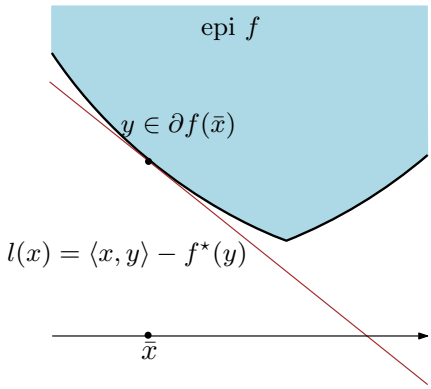


Figure 3.10: Tightness in the Fenchel-Young inequality

Proof. Theorem 3.39 shows that y lies in $\partial f(x)$ if and only if

$$\langle x, y\rangle = f(x) + f^\star(y).$$

On the other hand by Theorem 3.26, we have $f(x)+f^\star(y) = (f^\star)^\star(x)+f^\star(y)$. Applying Theorem 3.39 again with f replacing $f^\star$, we deduce $y \in \partial f(x)$ if and only if $x \in \partial f^\star(y)$. □

Using Corollary 3.40, we can now compute the subdifferential of a few more interesting functions.

Exercise 3.41. Let $Q \subset \mathbf{E}$ be a closed convex set. Show the equality

$$\partial \delta_Q^\star(v) = \operatorname*{argmax}_{x \in Q} \langle v, x\rangle \qquad \text{for all } x \in \mathbf{E}. \tag{3.15}$$

Use this expression to do the following exercises.

1. Show that the coordinate maximum function $\operatorname{mx} : \mathbf{R}^n \to \mathbf{R}$ defined by $\operatorname{mx}(x_1, \dots, x_n) = \max\{x_1, \dots, x_n\}$ satisfies

   $$\partial \operatorname{mx}(x) = \operatorname{conv}\{e_i : i \in I(x)\},$$

   where $I(x) := \{i : x_i = \operatorname{mx}(x)\}$ denotes the set of active indices.
2. Show that the ℓ_2-norm on $\mathbf{R}^n$ satisfies

   $$\partial \|x\|_2 = \begin{cases} \frac{x}{\|x\|_2} & \text{if } x \neq 0 \\ \operatorname{cl} \mathbb{B} & \text{if } x = 0 \end{cases}.$$

3. Show that the ℓ_∞ norm on $\mathbf{R}^n$ satisfies

$$\partial \|x\|_\infty = \operatorname{conv}\left(\bigcup_{i \in I(x)} \partial |x_i| \right),$$

where $I(x) := \{i : x_i = \|x\|_\infty\}$ denotes the set of active indices.

[**Hint:** Use Corollary 3.40 together with Exercise 2.37 to establish (3.15). Recognize that $\operatorname{mx}(\cdot)$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ functions are the support functions of the unit simplex Δ_n, the unit ℓ_2-ball, and the unit ℓ_1 ball, respectively. Use the expression (3.15) to complete the proof.]

We end with an illuminating exercise, which shows that the subdifferential of a convex function satisfies a nice closeness property.

Exercise 3.42. Consider a proper, closed, convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$. Show that for any sequences of point $x_i \in \mathbf{E}$ and $v_i \in \partial f(x_i)$, such that (x_i, v_i) converge to some pair $(\bar{x}, \bar{v})$, the inclusion $\bar{v} \in \partial f(\bar{x})$ must hold. Deduce that the graph of the subdifferential $\{(x, v) \in \mathbf{E} \times \mathbf{E} : v \in \partial f(x)\}$ is a closed set.

3.5.2 Subderivative

Subgradients of a function f describe affine minorants of f up to a first-order error. An interesting alternative construction would instead aim to measure the rate of change of f along a given direction. How should we measure the rate of change of a nonsmooth function f along a given direction $u \in \mathbf{E}$? The classical answer is the directional derivative

Definition 3.43 (Directional derivative)**.** Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and fix a point x, with $f(x)$ finite. The *directional derivative* of f at x in direction $u \in \mathbf{E}$ is defined by

$$f'(\bar{x}, u) := \lim_{\tau \searrow 0} \frac{f(\bar{x} + \tau u) - f(\bar{x})}{\tau} \tag{3.16}$$

provided the limit exists.

The definition of the directional derivative for general functions f is deeply flawed. First, the limit in the expression (3.16) might not exist. This problem turns out not to be too serious for most functions of interest; in particular, the forthcoming Lemma 3.48 shows that the limit does exist for

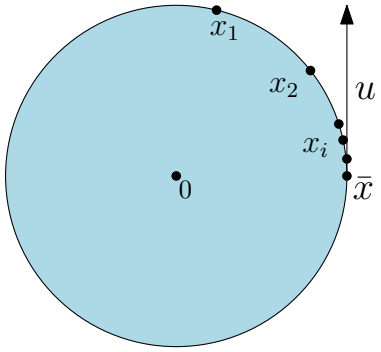


Figure 3.11: Example when the function $f'(\bar{x}, \cdot)$ is not closed.

convex functions. A much more serious deficit is that function $u \mapsto f'(\bar{x}, u)$ is not closed even if f is a closed, convex function; this is in contrast to the nice closedness properties of the subdifferential (Exercise 3.42).

A simple example will clarify what can go wrong. Let $f\colon \mathbf{R}^2 \to \overline{\mathbf{R}}$ be the indicator function of the closed unit disk and define the basepoint $\bar{x} = (1, 0)$. A quick computation yields the expression

$$f'(\bar{x}, u) = \begin{cases} 0 & \text{if } u_1 < 0 \\ +\infty & \text{otherwise} \end{cases}$$

Thus the function $u \mapsto f'(\bar{x}, u)$ is not closed. What goes wrong is that intuitively one expects $f'(\bar{x}, u) = 0$ for $u = (0, 1)$, and yet the value of the directional derivative is $f'(\bar{x}, u) = \infty$. What is at fault is the requirement that the approaching points $x_i = \bar{x} + \tau_i u$ in the difference quotients must lie on the ray $\{\bar{x}\} + \mathbf{R}_+\{u\}$. If we would have instead allowed a sequence x_i lying on the circle approaching $\bar{x}$, the problem would disappear. See Figure 3.11 for an illustration.

This example suggests that a better definition would compute difference quotients along a sequence of points that approaches $\bar{x}$ "directionally" along u, but not necessarily lying in the ray $\{\bar{x}\} + \mathbf{R}_+\{u\}$. That is, we should focus on points $x_i \to \bar{x}$ satisfying $\tau_i^{-1}(x_i - \bar{x}) \to u$ for some sequence $\tau_i \searrow 0$. Solving for x_i, we may write $x_i = \bar{x} + \tau_i w_i$ for some sequence $w_i \to u$. Thus a better measure of variation in direction u would involve the difference quotients $\frac{f(\bar{x}+\tau w)-f(\bar{x})}{\tau}$ as τ tends to zero and w tends to u. With this in mind, we introduce the following definition.

Definition 3.44 (Subderivative)**.** Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point $\bar{x}$, with $f(\bar{x})$ finite. Then the *subderivative* function $df(\bar{x})\colon \mathbf{E} \to \overline{\mathbf{R}}$ is

defined by

$$df(\bar{x})(u) = \liminf_{\substack{\tau \searrow 0 \\ w \to u}} \frac{f(\bar{x} + \tau w) - f(\bar{x})}{\tau}.$$

The following exercise shows that the subderivative of a function is closely tied to the tangent cone to its epigraph.

Exercise 3.45. ✍ Show that for any function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point x, with $f(x)$ finite, equality holds:

$$\operatorname{epi} df(x) = T_{\operatorname{epi} f}(x, f(x)).$$

For example, the grey region in Figure 3.7b coincides with the epigraph of the subderivative function $u \mapsto df(x)(u)$. An important consequence of Exercise 3.45 is that the function $df(x)(\cdot)$ is closed and positively homogeneous, since the tangent cone is always a closed cone.

The following exercise shows that when f is locally Lipschitz continuous around $\bar{x}$, the sequence w in the definition of the subderivative can be replaced simply by u—often, a convenient simplification. Note that local Lipschitz continuity excludes the example of the indicator function of the unit disk, discussed previously.

Exercise 3.46. Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ that is Lipschitz continuous on a neighborhood of a point $\bar{x}$. Then equality holds:

$$df(\bar{x})(u) = \liminf_{\tau \searrow 0} \frac{f(\bar{x} + \tau u) - f(\bar{x})}{\tau}.$$

For a smooth function f, the gradient and the directional derivative are tightly related by the expression $f'(x, u) = \langle \nabla f(x), u \rangle$. The following exercise shows that this relationship extends naturally to the nosmooth setting: the closed convex envelope of the subderivative is the support function of the subdifferential.

Theorem 3.47 (Subderivative as the support function)**.** ✍ *Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point $\bar{x}$, with $f(\bar{x})$ finite. Then the following two properties are equivalent:*

$$\partial f(\bar{x}) \neq \emptyset \qquad \Longleftrightarrow \qquad \overline{\operatorname{co}}\, df(\bar{x}) \text{ is proper.}$$

Under these two equivalent conditions, the envelope $\overline{\operatorname{co}}\, df(\bar{x})$ is the support function of the subdifferential $\partial f(\bar{x})$.

Proof. We will verify the relationship

$$\partial f(\bar{x}) = \{v \in \mathbf{E} : \langle u, v\rangle \leq df(\bar{x})(u) \ \ \forall u \in \mathbf{E}\}. \tag{3.17}$$

Then the statement of the theorem will follow immediately from Exercise 3.21. To this end, observe the following equivalences:

$$\begin{aligned} v \in \partial f(\bar{x}) \quad &\iff \quad (v,-1) \in N_{\operatorname{epi} f}(\bar{x}, f(\bar{x})) && (3.18)\\ &\iff \quad \langle (v,-1),(u,r)\rangle \leq 0 \qquad \forall (u,r) \in T_{\operatorname{epi} f}(\bar{x}, f(\bar{x})) && (3.19)\\ &\iff \quad \langle v, u\rangle \leq df(\bar{x})(u) \qquad \forall u \in \mathbf{E}, && (3.20) \end{aligned}$$

where (3.18) follows directly from Theorem 3.35, the equivalence (3.19) follows from Lemma 2.35, and (3.18) is a consequence of Exercise 3.45. We have thus proved that (3.17) holds, and the proof is complete. □

So far, the discussion of the subderivative $df(\bar{x})(\cdot)$ did not assume convexity of f. Our next goal is to show that when f is convex, the subderivative $df(\bar{x})(\cdot)$ coincides with the closure of the directional derivative function $\operatorname{cl} f'(\bar{x}, \cdot)$. We begin with the following lemma, which shows that the limit in the definition of the directional derivative (3.16) exists for convex functions.

Lemma 3.48. *Suppose $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is convex and fix a point x with $f(x)$ finite. Then for any point $u \in \mathbf{E}$, the quotients $\frac{f(x+\tau u)-f(x)}{\tau}$ are nondecreasing in τ, and therefore $f'(x,u)$ exists in $\overline{\mathbf{R}}$ for every $u \in \mathbf{E}$.*

Proof. Fix any reals $\hat{\tau}, \tau$ satisfying $0 < \hat{\tau} < \tau$. Using convexity, successively compute

$$\begin{aligned} \frac{f(x+\hat{\tau}u)-f(x)}{\hat{\tau}} &= \frac{f\left(\left(\frac{\tau-\hat{\tau}}{\tau}\right)x + \frac{\hat{\tau}}{\tau}(x+\tau u)\right) - f(x)}{\hat{\tau}}\\ &\leq \frac{\frac{\tau-\hat{\tau}}{\tau}f(x) + \frac{\hat{\tau}}{\tau}f(x+\tau u) - f(x)}{\hat{\tau}}\\ &= \frac{f(x+\tau u)-f(x)}{\tau} \end{aligned}$$

The result follows. □

Not surprisingly, the directional derivative $f'(x,\cdot)$ of a convex function f is itself a convex function.

Exercise 3.49. Let $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a proper convex function. Verify that for any point $x \in \operatorname{dom} f$, the directional derivative function $u \mapsto f'(x,u)$ is sublinear.

We are now ready to prove the following theorem

Theorem 3.50 (Subderivative and directional derivative)**.** *Consider a proper convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and fix a point $\bar{x} \in \operatorname{dom} f$. Then equality holds:*

$$df(\bar{x})(\cdot) = \operatorname{cl} f'(\bar{x}, \cdot). \tag{3.21}$$

Consequently, as long as $\partial f(\bar{x})$ is nonempty, we have

$$df(\bar{x}) = \operatorname{cl} f'(\bar{x}, \cdot) = \delta^\star_{\partial f(\bar{x})}.$$

Proof. Without loss of generality, we may suppose $\bar{x} = 0$ and $f(\bar{x}) = 0$ throughout the proof. Observe that since f is convex, the tangent cone $T_{\operatorname{epi} f}(\bar{x}, f(\bar{x}))$ is closed and convex. Therefore using Exercise 3.45, we deduce that $df(\bar{x})$ is a closed sublinear function. The inequality $df(\bar{x})(\cdot) \leq f'(x, \cdot)$ is immediate from definitions, and therefore we conclude $df(\bar{x})(\cdot) \leq \operatorname{cl} f'(x, \cdot)$. To see the reverse inequality, we will show the inclusion

$$\operatorname{epi} f - \{(\bar{x}, f(\bar{x}))\} \subset \operatorname{epi} f'(\bar{x}, \cdot). \tag{3.22}$$

To this end, fix an arbitrary point $(x, r) \in \operatorname{epi} f$, and compute

$$\frac{f(\bar{x} + (x - \bar{x})) - f(\bar{x})}{1} \leq r - f(\bar{x}).$$

Monotonicity of difference quotients (Lemma 3.48) guarantees $f'(\bar{x}, x - \bar{x}) \leq r - f(\bar{x})$, and therefore

$$(x, r) - (\bar{x}, f(\bar{x})) = (x - \bar{x}, r - f(\bar{x})) \in \operatorname{epi} f'(\bar{x}, \cdot)$$

Thus the inclusion (3.22) holds. Using Exercise 2.33 we deduce $T_{\operatorname{epi} f}(\bar{x}, f(\bar{x})) \subset \operatorname{cl} \operatorname{epi} f'(\bar{x}, \cdot)$. Exercise 3.45 therefore implies $\operatorname{cl} f'(\bar{x}, \cdot) \leq df(\bar{x})$, thereby completing the proof of (3.21). The final claim is now immediate from Exercise 3.47. □

3.6 Lipschitz continuity of convex functions

Theorem 3.38 showed that the subdifferential $\partial f(x)$ of a convex function f is nonempty at every point $x \in \operatorname{ri}(\operatorname{dom} f)$. This section proves a much stronger statement: a convex function f is Lipschitz continuous on any compact subset of $\operatorname{ri}(\operatorname{dom} f)$. To better appreciate the subtlety of this guarantee, it is worthwhile to see what can go wrong at points in the relative boundary of the domain. First, the example $h(x) = -\sqrt{x}$ discussed in Exercise 3.32, shows

that a closed convex function may fail to be locally Lipschitz continuous relative to its domain. A more interesting example is the function $f\colon \mathbf{R}^2 \to \overline{\mathbf{R}}$, depicted in Figure 3.12 and defined by

$$f(x,y) = \begin{cases} \frac{y^2}{x} & \text{if } x > 0 \\ 0 & \text{if } (x,y) = (0,0)\ . \\ +\infty & \text{otherwise} \end{cases} \tag{3.23}$$

It is straightforward to see that f is closed and convex, but is not even continuous at the origin relative to its domain.

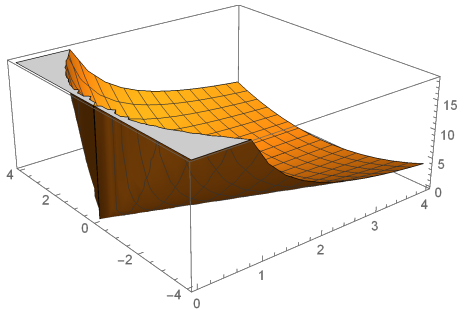


Figure 3.12: Plot of the function $f(x,y)$ in equation (3.23).

We begin with the following exercise that relates Lipschitz continuity to the size of subgradients.

Exercise 3.51. ✍ Let $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a proper convex function and let Q be any open subset of $\operatorname{dom} f$. Prove the identity:

$$\sup_{x,y\in Q} \frac{|f(x) - f(y)|}{\|x - y\|} = \sup_{x\in Q,\, v\in\partial f(x)} \|v\|. \tag{3.24}$$

[**Hint:** The inequality $\leq$ in (3.24) follows from the subgradient inequality and Cauchy–Schwarz. To see the reverse inequality, fix $x \in Q$ and $v \in \partial f(x)$, and estimate the error $f(x + tv) - f(x)$ using the subgradient inequality.]

Exercise 3.52. ✍ Consider a proper, closed, convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$. Show that f is L-Lipschitz continuous if and only if every vector in $\operatorname{dom} h^\star$ is bounded in norm by L.
[**Hint:** Using Corollary 3.40, verify $\sup_{x\in\mathbf{E},\, v\in\partial f(x)} \|v\| = \sup_{y\in\operatorname{dom} h^\star} \|y\|$. Complete the proof by appealing to Exercise 3.51.]

We are ready to prove the main result of this section.

Corollary 3.53. *Consider a proper convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and let Q be a compact subset of* $\operatorname{ri}(\operatorname{dom} f)$*. Then f is Lipschitz continuous on Q*

Proof. Without loss of generality, we may suppose that $\operatorname{dom} f$ has nonempty interior. Since Q is compact, we may choose a bounded open set U satisfying $Q \subset U$ and $\operatorname{cl} U \subset \operatorname{int}(\operatorname{dom} f)$ (verify this!). We will argue $\sup_{x\in U, v\in\partial f(x)} \|v\| < \infty$, and therefore f is Lipschitz continuous on Q by Exercise 3.51. To verify the claim, suppose for the sake of contradiction that there exist sequences $x_i \in U$ and $v_i \in \partial f(x_i)$ satisfying $\|v_i\| \to \infty$. Since U is bounded, we may pass to a subsequence and assume that x_i tends to some $\bar{x} \in \operatorname{cl} U \subset \operatorname{int}(\operatorname{dom} f)$ and $\frac{v_i}{\|v_i\|}$ tends to some unit vector $\bar{v}$. We will show that the inclusion $\bar{v} \in N_{\operatorname{dom} f}(\bar{x})$ holds, which is impossible since $\bar{x}$ lies in $\operatorname{int}(\operatorname{dom} f)$. To this end, observe that the subgradient inequality takes the form

$$f(y) \geq f(x_i) + \langle v_i, y - x_i\rangle \qquad \text{for all } y \in \operatorname{dom} f.$$

Dividing through by $\|v_i\|$ and passing to the limit yields

$$\liminf_{i\to\infty} \frac{f(x_i)}{\|v_i\|} \leq \langle \bar{v}, \bar{x} - y\rangle \qquad \text{for all } y \in \operatorname{dom} f. \tag{3.25}$$

Since f is proper, it admits an affine minorant and therefore f is clearly bounded from below on the compact set $\operatorname{cl} U$. Therefore the left side of (3.25) is nonnegative, and the inclusion $\bar{v} \in N_{\operatorname{dom} f}(\bar{x})$ holds, as claimed. □

We end the section with the following exercise, which shows that the converse of Exercise 3.31 holds for convex functions.

Exercise 3.54. ✍ A proper convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is differentiable at x if and only if $\partial f(x)$ is a singleton, in which case it must be that $\partial f(x) = \{\nabla f(x)\}$.
[**Hint:** The forward implication is Exercise 3.31. To see the converse, assume $\partial f(x) = \{v\}$ for some $v \in \mathbf{E}$. Next, verify the inclusion $N_{\operatorname{dom} f}(x) + \partial f(x) \subset \partial f(x)$ and deduce $x \in \operatorname{int}(\operatorname{dom} f)$. Consider any sequence $y_i \to x$ and choose any vectors $v_i \in \partial f(y_i)$, guaranteed to exist by Theorem 3.38. Use Exercise 3.51 and Corollary 3.53 to deduce that the sequence $\{v_i\}$ is bounded. Using the subgradient inequality and Exercise 3.42, argue $\operatorname{limsup}_{i\to\infty} \frac{f(y_i)-f(x)-\langle v, y_i - x\rangle}{\|y_i - x\|} \leq \lim_{i\to\infty} \|v - v_i\| = 0$.]

3.7 Strong convexity, Moreau envelope, and the proximal map

This section introduces class of strongly convex functions. Such functions play a fundamental role both in theory and algorithms, as we will see. In particular, typical algorithms for strongly convex functions converge faster than for functions that are merely convex. One of the main results of the section is the Baillon-Haddad theorem, which shows that a proper, closed, convex function is strongly convex if and only its Fenchel conjugate is smooth. In this sense, smoothness and strong convexity are dual property. On our way to proving this theorem, we will introduce the Moreau envelope, which serves as a smooth approximation of a convex function, and the proximal map that appears in the expression for its gradient. These three notions— strong convexity, Moreau envelope, and the proximal map—will appear often in later sections.

Definition 3.55 (Strong convexity)**.** A function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is called *μ-strongly convex* (with $\mu \geq 0$) if the perturbed function $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex.

Trivially, convex functions are strongly convex with parameter $\mu = 0$. More generally, if f is a convex function, then $f + \frac{\mu}{2}\|\cdot\|^2$ is μ-strongly convex. Indeed, this is the typical way in which strongly convex functions arise, as quadratic perturbations of convex functions. Checking strong convexity of a C^2-smooth function simply amounts to lower-bounding the minimal eigenvalue of the Hessian.

Exercise 3.56. Show that a C^2-smooth function $f\colon U \to \mathbf{R}$, defined on an open set $U \subset \mathbf{E}$, is α-strongly convex if and only if the relation $\nabla^2 f(x) \succeq -\alpha I$ holds for all $x \in U$.
[**Hint:** Apply Theorem 3.8.]

The main use of strong convexity stems from a strengthened sugradient inequality, established in the following theorem.

Theorem 3.57. *Let $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a μ-strongly convex function. Then for any $x \in \mathbf{E}$ and $v \in \partial f(x)$, the estimate holds:*

$$f(y) \geq f(x) + \langle v, y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \qquad \text{for all } y \in \mathbf{E}.$$

Proof. Fix a point $x \in \mathbf{R}$ and a vector $v \in \partial f(x)$. Define the convex function $g := f - \frac{\mu}{2}\|\cdot\|^2$ and note the equality $\partial g(x) = \partial f(x) - \mu x$ (Exercise 3.34).

The subgradient inequality for g therefore guarantees

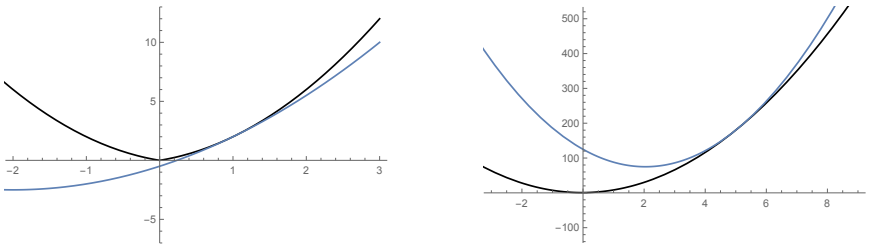
$$g(y) \geq g(x) + \langle v - \mu x, y - x \rangle.$$

Plugging in the definition of g and rearranging yields

$$\begin{aligned} f(y) &\geq f(x) + \langle v, y - x \rangle + \frac{\mu}{2} \left(\|y\|^2 + \|x\|^2 - 2\langle x, y \rangle \right) \\ &= f(x) + \langle v, y - x \rangle + \frac{\mu}{2} \|y - x\|^2, \end{aligned}$$

as we had to show. □

Thus, strong convexity of a function f guarantees that any point x and subgradient $v \in \partial f(x)$ yield a quadratic function $y \mapsto f(x) + \langle v, y - x \rangle + \frac{\mu}{2}\|y - x\|^2$ that minorizes f. Notice that in this way, strong convexity plays an opposite role to smoothness. Recall from Exercise 1.15 that if f is β-smooth, then the quadratic $y \mapsto f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2$ is an upper-estimator of f for any x. See Figure 3.13 for an illustration.



(a) Strong convexity (b) Smoothness

Figure 3.13: Depiction of strong convexity (left) and smoothness (right). Left: $f(x) = |x| + x^2$ (black), quadratic lower model formed at $x = 1$ (blue); Right: $f(x) = \log(1 + e^x) + 7x^2$ (black), quadratic upper model formed at $x = 5$ (blue)

The following is a direct consequence of Theorem 3.57.

Exercise 3.58. ✍ Any proper, closed, μ-strongly convex function $f : \mathbf{E} \to \overline{\mathbf{R}}$ has a unique minimizer x, which moreover satisfies

$$f(y) - f(x) \geq \frac{\mu}{2}\|y - x\|^2 \qquad \text{for all } y \in \mathbf{E}. \tag{3.26}$$

[**Hint:** Argue using Theorem 3.57 that f is coercive, and apply Exercise 1.6 to deduce existence of a minimizer x. Deduce the claimed inequality (3.26) from Theorem 3.57. Conclude uniqueness from (3.26).]

A minimizer of any function f, by definition, satisfies $f(y)-f(x) \geq 0$ for all y. The estimate (3.26) guarantees that if f is μ-strongly convex, then we can squeeze out an extra quadratic term $\frac{\mu}{2}\|y-x\|^2$ on the right-hand-side. We will use Exercise 3.58 heavily in what follows.

As we alluded to previously, smoothness and strong convexity play opposite roles: one has to do with upper quadratic approximations and the other with lower quadratic approximations. We will now see that this observation can be made precise through the language of duality. We will prove that a closed convex function is β-smooth if and only if its Fenchel conjugate is $\frac{1}{\beta}$-strongly convex. The argument will be based on the following two constructions, which are critically important in their own right.

Definition 3.59. For any function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and real $\alpha > 0$, define the *Moreau envelope* and the *proximal map*, respectively:

$$f_\alpha(x) := \min_y \; f(y) + \frac{1}{2\alpha}\|x-y\|^2$$
$$\operatorname{prox}_{\alpha f}(x) := \operatorname*{argmin}_y f(y) + \frac{1}{2\alpha}\|x-y\|^2.$$

The reader should recognize the Moreau envelope simply as the infimal convolution

$$f_\alpha = f \,\square\, \left(\frac{1}{2\alpha}\|\cdot\|^2\right).$$

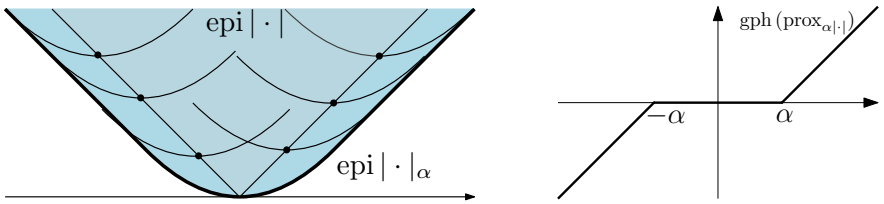
In particular, when f is proper, closed, and convex the perturbed function $f + \frac{1}{2\alpha}\|x-\cdot\|^2$ is proper, closed, and strongly convex. Consequently, Exercise 3.58 guarantees that $\operatorname{prox}_{\alpha f}(x)$ is a singleton for every x. Taking into account the expression (3.5), we recognize the epigraph of f_α as the sum:

$$\operatorname{epi} f_\alpha = \operatorname{epi} f + \operatorname{epi}\left(\frac{1}{2\alpha}\|\cdot\|^2\right), \tag{3.27}$$

An example will help to gain some intuition. Letting f be the absolute value function$f = |\cdot|$, a quick computation shows (Figure 3.14)

$$f_\alpha(x) = \begin{Bmatrix} \frac{1}{2\alpha}|x|^2 & \text{if } |x| \leq \alpha \\ |x| - \frac{1}{2}\alpha & \text{otherwise} \end{Bmatrix}, \qquad \operatorname{prox}_{\alpha f}(x) = \begin{Bmatrix} x-\alpha & \text{if } x \geq \alpha \\ 0 & \text{if } |x| \leq \alpha \\ x+\alpha & \text{if } x \leq -\alpha \end{Bmatrix}.$$

The envelope f_α appears often in the statistics literature, under the name of the Huber function. Notice that the Huber function coincides with a simple



(a) $\operatorname{epi}|\cdot|_\alpha = \operatorname{epi}|\cdot| + \operatorname{epi}\frac{1}{2}|\cdot|^2$ (b) $\operatorname{gph}(\operatorname{prox}_{\alpha|\cdot|})$

Figure 3.14: Moreau envelope and the proximal map of $|\cdot|$.

quadratic near the origin and then grows linearly. The associated proximal map $\operatorname{prox}_{\alpha f}$ is called soft-thresholding in statistics and signal processing.

In the simplest case that f is the indicator function of a closed convex set Q, the Moreau envelope reduces to the squared distance $f_\alpha(x) = \frac{1}{2\alpha}\operatorname{dist}_Q^2(x)$, while the proximal map becomes the nearest point projection $\operatorname{prox}_{\alpha f}(x) = \operatorname{proj}_Q(x)$. Consequently, a great deal of intuition can be gained by treating the proximal map as a generalization of the nearest-point projection for functions. The following theorem, which plays an analogous role to Exercise 2.18 for projections, shows that the proximal map of a proper, closed, convex function is 1-Lipschitz continuous.

Theorem 3.60. *Consider a proper, closed, convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$. Then the set $\operatorname{prox}_f(x)$ is a singleton for every point $x \in \mathbf{E}$. Moreover, for any points $x, y \in \mathbf{E}$ the estimate holds:*

$$\|\operatorname{prox}_f(x) - \operatorname{prox}_f(y)\|^2 \leq \langle \operatorname{prox}_f(x) - \operatorname{prox}_f(y), x - y\rangle.$$

In particular, the proximal map $x \mapsto \operatorname{prox}_f(x)$ is 1-Lipschitz continuous.

Proof. Notice that for every point x, the function $z \mapsto f(z) + \frac{1}{2}\|z - x\|^2$ is proper, closed, and 1-strongly convex, and hence has a unique minimizer (Exercise 3.58). Thus the proximal set prox_f is a singleton everywhere on $\mathbf{E}$, as claimed. Next, consider any two points $x, y \in \mathbf{E}$ and define $x^+ = \operatorname{prox}_f(x)$ and $y^+ = \operatorname{prox}_f(y)$. By definition x^+ is the minimizer of the function $f + \frac{1}{2}\|\cdot - x\|^2$ and y^+ is the minimizer of $f + \frac{1}{2}\|\cdot - y\|^2$. Using strong

convexity and Exercise 3.58, we therefore deduce

$$\begin{aligned} f(x^+) + \frac{1}{2}\|x^+ - x\|^2 &\leq \left(f(y^+) + \frac{1}{2}\|y^+ - x\|^2\right) - \frac{1}{2}\|y^+ - x^+\|^2 \\ &= f(y^+) + \frac{1}{2}\|y^+ - y\|^2 - \frac{1}{2}\|y^+ - x^+\|^2 \\ &\quad + \frac{1}{2}\|y^+ - x\|^2 - \frac{1}{2}\|y^+ - y\|^2 \\ &\leq \left(f(x^+) + \frac{1}{2}\|x^+ - y\|^2\right) - \|y^+ - x^+\|^2 \\ &\quad + \frac{1}{2}\|y^+ - x\|^2 - \frac{1}{2}\|y^+ - y\|^2. \end{aligned}$$

Rearranging yields the claimed estimate:

$$\begin{aligned} \|y^+ - x^+\|^2 &\leq \frac{1}{2}\left(\|x^+ - y\|^2 - \|y^+ - y\|^2 + \|y^+ - x\|^2 - \|x^+ - x\|^2\right) \\ &= \langle x^+ - y^+, x - y\rangle. \end{aligned}$$

The Cauchy-Schwarz inequality then implies $\|y^+ - x^+\| \leq \|x - y\|$ as claimed. □

The proximal map of a convex function and that of its conjugate are closely related, as the following theorem shows.

Theorem 3.61 (Moreau decomposition)**.** *For any proper, closed, convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$, equality holds*

$$\operatorname{prox}_f(x) + \operatorname{prox}_{f^\star}(x) = x \qquad \forall x \in \mathbf{E}.$$

Proof. Using the definition of the proximal map, we successively deduce

$$\begin{aligned} z = \operatorname{prox}_f(x) &\iff 0 \in \partial\left(f + \frac{1}{2}\|\cdot - x\|^2\right)(z) \\ &\iff x - z \in \partial f(z) \\ &\iff z \in \partial f^\star(x - z) \qquad (3.28) \\ &\iff 0 \in \partial f^\star(x - z) - z \\ &\iff 0 \in \partial\left(f^\star + \frac{1}{2}\|\cdot - x\|^2\right)(x - z) \\ &\iff x - z = \operatorname{prox}_{f^\star}(x). \end{aligned}$$

where (3.28) follows from Corollary 3.40. This completes the proof. □

Theorem 3.61 is geometrically appealing when specialized to indicator functions of cones. Letting f be the indicator function of a closed, convex cone $K \subset \mathbf{E}$, the theorem guarantees:

$$\operatorname{proj}_K(x) + \operatorname{proj}_{K^\circ}(x) = x \qquad \forall x \in \mathbf{E}.$$

See Figure 3.15 for an illustration.

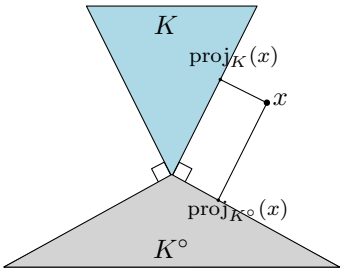


Figure 3.15: Polar decomposition of a point $x = \operatorname{proj}_K(x) + \operatorname{proj}_{K^\circ}(x)$

Thus every point x can be decomposed as a sum of a point in K and a point in K°. When K is a linear subspace, we simply recover the orthogonal decomposition. Continuing with the linear algebraic analogy, the following exercise shows that the polar decomposition is unique.

Exercise 3.62. Let $K \subset \mathbf{E}$ be a nonempty, closed, convex cone. Show that for any three points $x, y, z \in \mathbf{E}$, the following conditions are equivalent.

1. $z = x + y$ with $x \in K$, $y \in K^\circ$, $\langle x, y \rangle = 0$,
2. $x = \operatorname{proj}_K(z)$ and $y = \operatorname{proj}_{K^\circ}(z)$.

Exercise 3.63. Compute the proximal operator $\operatorname{prox}_{\alpha f}$ of the following functions: $f(x) = \|x\|_1$, $f(x) = \|x\|_2$, $f(x) = \|x\|_\infty$.

Looking at Figure 3.14b, it appears that the Moreau envelope of the absolute value function is smooth. This is not a coincidence. An intuitive reason stems from the expression (3.27) for the epigraph of the Moreau envelope. The addition of epi $\frac{1}{2\alpha}\|\cdot\|^2$ to an epigraph of a convex function tends to "smooth out" its nonsmooth features. The following theorem provides a formal justification.

Theorem 3.64. *For any proper, closed, convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$, the envelope f_α is continuously differentiable on $\mathbf{E}$ with gradient*

$$\nabla f_\alpha(x) = \alpha^{-1}(x - \operatorname{prox}_{\alpha f}(x)). \tag{3.29}$$

Consequently ∇f_α is Lipschitz continuous with parameter α^{-1}.

Proof. Suppose first $\alpha = 1$ and fix a point $x \in \mathbf{E}$. We aim to show that the subdifferential of f_α is a singleton at every point. To this end, we successively deduce the equivalences

$$\begin{aligned} z \in \partial f_\alpha(x) &\iff x \in \partial(f \,\square\, \tfrac{1}{2}\|\cdot\|^2)^\star(z) && (3.30)\\ &\iff x \in \partial\left(f^\star + \left(\tfrac{1}{2}\|\cdot\|^2\right)^\star\right)(z) && (3.31)\\ &\iff x \in \partial f^\star(z) + z \\ &\iff 0 \in \partial(f^\star + \tfrac{1}{2}\|\cdot - x\|^2)(z) \\ &\iff z = \operatorname{prox}_{f^\star}(x) \\ &\iff z = x - \operatorname{prox}_f(x), && (3.32) \end{aligned}$$

where (3.30) follows from Corollary 3.40 and the definition of the Moreau envelope, the second equivalence (3.31) uses the conjugate formula for the infimal convolution in Exercise 3.28, and the last equivalence (3.32) follows from Theorem 3.61. Thus we deduce that the subdifferential $\partial f_\alpha(x)$ is a singleton. Hence by Exercise 3.54, the envelope f_α is differentiable with $\nabla f_\alpha(x) = x - \operatorname{prox}_f(x)$. It follows immediately from Theorem 3.60 that ∇f_α is 1-Lipschitz continuous. Returning to the general setting $\alpha \neq 1$, observe that $\alpha \cdot f_\alpha$ is the Moreau-Yosida envelope of αf with parameter 1. Applying what we have already proved for the case $\alpha = 1$ completes the argument. □

Exercise 3.65 (Mean-value theorem)**.** Consider a proper convex function $f\colon \mathbf{R} \to \overline{\mathbf{R}}$ and fix two points $x, y \in \operatorname{ri}(\operatorname{dom} f)$. Show that there exists a point $z \in [x, y]$ and a subgradient $v \in \partial f(z)$ satisfying

$$f(y) - f(x) = \langle v, y - x\rangle.$$

[**Hint:** Explain why without loss of generality you can assume that $\operatorname{dom} f$ has nonempty interior. Then apply the mean value theorem to the Moreau envelope $f_\alpha(x)$ and note the inclusion $\nabla f_\alpha(x) \in \partial f(\operatorname{prox}_{\alpha f}(x))$. Complete the proof by letting α tend to zero. Why is it important that x and y lie in the interior of $\operatorname{dom} f$?]

We end the section with the following theorem, which formalizes the duality between smoothness and strong convexity.

Theorem 3.66 (Baillon-Haddad)**.** *A proper, closed, convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is μ-strongly convex if and only if the conjugate $f^\star$ is μ^{-1}-smooth.*

Proof. Suppose that f is μ-strongly convex and define the convex function $g(x) := f(x) - \frac{\mu}{2}\|x\|^2$. We may then write

$$f^\star = \left(g + \tfrac{\mu}{2}\|\cdot\|^2\right)^\star = g^\star \,\square\, \tfrac{1}{2\mu}\|\cdot\|^2.$$

The right-hand-side is simply the Moreau envelope of $g^\star$ with parameter μ, and is therefore μ^{-1}-smooth by Theorem 3.64.

To see the converse, assume that $f^\star$ is μ^{-1}-smooth. To simplify notation, define $h := f^\star$ and $\beta := 1/\mu$. We will show that $h^\star$ is $\frac{1}{\beta}$-strongly convex, thereby completing the proof. To this end, define the function $g(x) := \frac{\beta}{2}\|\cdot\|^2 - h$. Taking derivatives, we compute

$$\langle \nabla g(y) - \nabla g(x), y - x\rangle = \beta\|y-x\|^2 - \langle \nabla h(y) - \nabla h(x), y - x\rangle \geq 0,$$

for all $x, y \in \mathbf{E}$, where the last inequality follows from Exercise 3.12 . Thus g is convex by Theorem 3.8. We now express h as

$$\begin{aligned} h(y) = \tfrac{\beta}{2}\|y\|^2 - g(y) &= \tfrac{\beta}{2}\|y\|^2 - g^{\star\star}(y) \\ &= \tfrac{\beta}{2}\|y\|^2 - \sup_x \{\langle y, x\rangle - g^\star(x)\} \\ &= \inf_x \left\{\tfrac{\beta}{2}\|y\|^2 - \langle y, x\rangle + g^\star(x)\right\}. \end{aligned}$$

Using this expression for h, we compute the conjugate

$$\begin{aligned} h^\star(z) &= \sup_y \{\langle z, y\rangle - h(y)\} \\ &= \sup_y \left\{\langle z, y\rangle - \inf_x \left\{\tfrac{\beta}{2}\|y\|^2 - \langle y, x\rangle + g^\star(x)\right\}\right\} \\ &= \sup_x \sup_y \left\{\langle z, y\rangle - \tfrac{\beta}{2}\|y\|^2 + \langle y, x\rangle - g^\star(x)\right\} \\ &= \sup_x \left\{\sup_y \left\{\langle z + x, y\rangle - \tfrac{\beta}{2}\|y\|^2\right\} - g^\star(x)\right\} \\ &= \sup_x \ \tfrac{1}{2\beta}\|z + x\|^2 - g^\star(x). \end{aligned}$$

Subtracting $\frac{1}{2\beta}\|z\|^2$ from both sides yields

$$h^\star(z) - \tfrac{1}{2\beta}\|z\|^2 = \sup_x \ \tfrac{1}{\beta}\langle z, x\rangle + \tfrac{1}{2\beta}\|x\|^2 - g^\star(x).$$

The right-hand-side is a pointwise supremum of affine functions in z and is therefore convex. This completes the proof. $\square$

Exercise 3.67. Show that any β-smooth function $h\colon \mathbf{E} \to \overline{\mathbf{R}}$ can be written as the Moreau envelope $f_{1/\beta}$ of some closed convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$.

Another notable consequence of the Baillon-Haddad theorem is that strongly convex functions are subgradient dominated.

Theorem 3.68. *Any proper, closed, α-strongly convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ satisfies the subgradient dominance condition:*

$$f(x) - \min f \leq \frac{1}{\alpha}\|v\|^2 \qquad \text{for all } x \in \mathbf{E},\ v \in \partial f(x).$$

Proof. Let x^* be a minimizer of f. Fix any $x \in \mathbf{E}$ and $v \in \partial f(x)$. We compute

$$\begin{aligned} f(x) - f^* \leq \langle v, x - x^* \rangle &\leq \|v\| \cdot \|x - x^*\| \\ &= \|v\| \cdot \|\nabla f^\star(v) - \nabla f^\star(0)\| \leq \frac{1}{\alpha}\|v\|^2, \end{aligned} \tag{3.33}$$

where (3.33) follows from Corollary 3.40 and Theorem 3.66. □

Theorem 3.68 plays an important role when analyzing algorithms. We will encounter a number of algorithms for minimizing a convex function f that generate sequences x_i and $v_i \in \partial f(x_i)$ satisfying $v_i \to 0$. When f is strongly convex, Theorem 3.68 allows to translate estimates on the norms $\|v_i\|$ into estimates on the function gap $f(x_i) - \min f$.

3.8 Monotone operators and the resolvant

Optimization problems in applications often arise in the structured form

$$(P) \qquad \min_x \ h(\mathcal{A}x) + g(x)$$

for some proper, closed, convex functions $g\colon \mathbf{E} \to \overline{\mathbf{R}}$ and $h\colon \mathbf{Y} \to \overline{\mathbf{R}}$ and $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$ a linear map. Such optimization problems will be the main focus of Chapter 4. An important result of the chapter (Corollary 4.11) will show that under mild conditions, a point x is optimal for (P) if and only if there exists a "dual variable" y such that the pair (x, y) jointly satisfies the inclusion

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} 0 & \mathcal{A}^* \\ -\mathcal{A} & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \partial g(x) \times \partial h^\star(y). \tag{3.34}$$

Thus, letting $T(x, y)$ denote the set on the right-side of (3.34), the problem (P) is equivalent to finding a pair (x, y) satisfying $0 \in T(x, y)$. Notice that

$T(x, y)$ is a sum of a skew symmetric linear operator and the subdifferential of a proper, closed, convex function. In this section, we will see that even though $T(x, y)$ is not the subdifferential of any convex function, it does satisfy a nice "monotonicity property". It it this monotonicity property that enables application of a number of efficient algorithms to the system (3.34). In this section, we will use convex analytic techniques to study general monotone operators. The obtained results will serve as the foundation for the primal-dual algorithms developed in Chapter 7.

3.8.1 Notation and basic properties

We begin with some basic notation. A *set-valued map* T, denoted $T\colon \mathbf{E} \rightrightarrows \mathbf{Y}$, is an assignment that maps $\mathbf{E}$ to the powerset of $\mathbf{Y}$. Thus T assigns to every point $x \in \mathbf{E}$ a subset $T(x) \subset \mathbf{Y}$. We have already encountered an important set-valued map: the subdifferential ∂f of a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is a set-valued map from $\mathbf{E}$ to $\mathbf{E}$. The *domain*, *range*, and *graph* of a map $T\colon \mathbf{E} \rightrightarrows \mathbf{Y}$ are defined by

$$\begin{aligned} \operatorname{dom} T &= \{x \in \mathbf{E} : T(x) \neq \emptyset\}, \\ \operatorname{range} T &= \bigcup_{x \in \mathbf{E}} T(x), \\ \operatorname{gph} T &= \{(x, y) \in \mathbf{E} \times \mathbf{Y} : y \in T(x)\}, \end{aligned}$$

respectively. Thus the domain of T consists of all points x where $T(x)$ is nonempty. The range of T consists of all points y that satisfy $y \in T(x)$ for some point x. Finally the graph of T simply "stacks" the images of T; see Figure 3.16 for an illustration. In particular, a set-valued map is fully described by its graph. The inverse map $T^{-1}\colon \mathbf{Y} \rightrightarrows \mathbf{E}$ is defined as

$$T^{-1}(y) = \{x \in \mathbf{E} : y \in T(x)\}.$$

Notice that the inverse map is always well-defined, as a set-valued map. The graph of T^{-1} is simply the image of the graph of T under the reflection $(x, y) \mapsto (y, x)$.

A set-valued map $T\colon \mathbf{E} \rightrightarrows \mathbf{Y}$ is called *surjective* if equality, $\operatorname{range} T = \mathbf{Y}$, holds. We say that T is *single-valued* if $T(x)$ is a singleton set for every $x \in \mathbf{E}$. Single-valued maps T correspond to point-to-point maps in the conventional sense. The sum $T_1 + T_2$ of two set-valued maps $T_1\colon \mathbf{E} \rightrightarrows \mathbf{Y}$ and $T_2\colon \mathbf{E} \rightrightarrows \mathbf{Y}$ is the set-valued map defined through set-addition

$$(T_1 + T_2)(x) = T_1(x) + T_2(x).$$

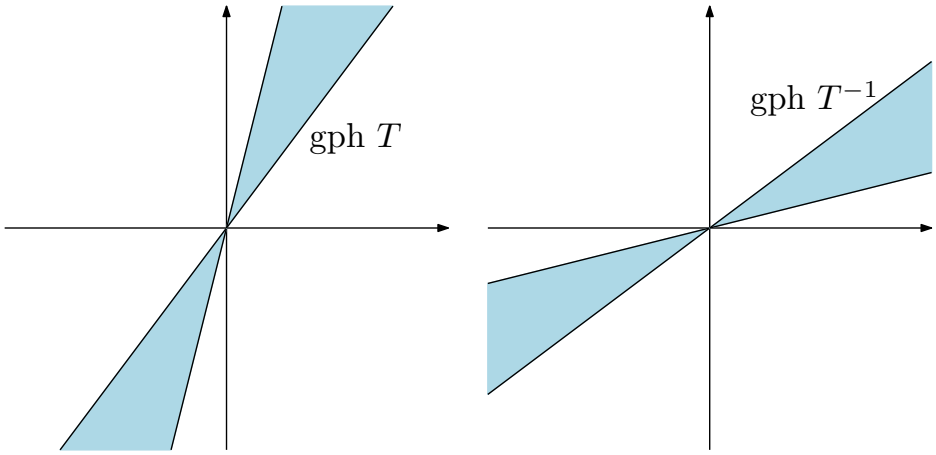


Figure 3.16: Graph of the set-valued map $T(x) = \operatorname{conv}\left\{\frac{4}{3}x, 4x\right\}$ and that of its inverse $T^{-1}(y) = \frac{1}{4}\operatorname{conv}\{y, 3y\}$.

In this section, we will be primarily interested in set-valued maps that satisfy the following monotonicity property.

Definition 3.69 (Monotone operator)**.** A set-valued map $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ is a *monotone operator* if it satisfies

$$\langle y_1 - y_2, x_1 - x_2\rangle \geq 0 \qquad \text{for all } (x_1, y_1),\ (x_2, y_2) \in \operatorname{gph} T.$$

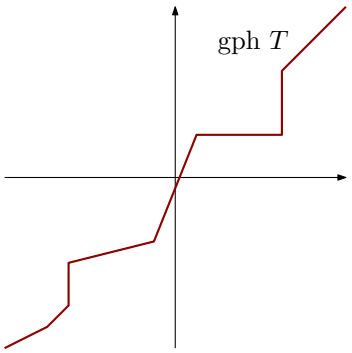


Figure 3.17: Graph of a monotone map $T\colon \mathbf{R} \rightrightarrows \mathbf{R}$.

To gain some intuition, let us look at a univariate set-valued map $T\colon \mathbf{R} \rightrightarrows \mathbf{R}$. Then by definition T is monotone if the inequality $(y_1 - y_2)(x_1 - x_2) \geq 0$ holds for all pairs (x_1, y_1), $(x_2, y_2) \in \operatorname{gph} T$. That is, whenever $x_1 \geq x_2$, the

inequality $y_1 \geq y_2$ must hold for all $y_1 \in T(x_1)$ and $y_2 \in T(x_2)$; see Figure 3.17 for an illustration. In particular, observe that the graph depicted in Figure 3.17 is very "thin" within its ambient space. Exercise 3.80 will formalize this observation for all monotone operators.

The main two examples of monotone operators are skew symmetric linear operators and subdifferentials of convex functions. Recall that a linear operator $\mathcal{A}\colon \mathbf{E} \to \mathbf{E}$ is skew-symmetric if it satisfies $\mathcal{A}^* = -\mathcal{A}$.

Exercise 3.70. Establish the following.

1. Show that a linear map $\mathcal{A}\colon \mathbf{E} \to \mathbf{E}$ (not necessarily self-adjoint) is monotone if and only if $\langle \mathcal{A}x, x\rangle \geq 0$ for all $x \in \mathbf{E}$.
2. Show that any skew symmetric linear operator $\mathcal{A}\colon \mathbf{E} \to \mathbf{E}$ is monotone.
3. Show that the subdifferential $\partial f\colon \mathbf{E} \rightrightarrows \mathbf{E}$ of any convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is monotone.
4. Show that a skew symmetric linear operator $\mathcal{A}\colon \mathbf{E} \to \mathbf{E}$ is the subdifferential of a convex function if and only if $\mathcal{A}$ is identically zero.

Monotonicity is preserved under a variety of operations. In particular, the sum of a skew symmetric linear operator and the subdifferential of a convex function is monotone.

Exercise 3.71. Let T, T_1, and T_2 be monotone operators on $\mathbf{E}$ and fix $\lambda > 0$. Show that T^{-1}, $T_1 + T_2$, and λT are monotone operators.

General monotone operators can be quite pathological. For example, given a monotone operator T on $\mathbf{E}$ and an arbitrary set $S \subset \mathbf{E}$, one can redefine $T(x) = \emptyset$ for all $x \in S$ and maintain monotonicity. The full power of monotonicity becomes available only once the the map in question is maximally monotone in the following sense.

Definition 3.72 (Maximal monotone operators)**.** A monotone mapping $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ is *maximally monotone* if $\operatorname{gph} T$ is not properly contained in the graph of any other monotone operator.

Thus a monotone operator $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ is maximally monotone if and only if no enlargement of its graph is possible without destroying monotonicity. More concretely, in order to show that a monotone map T is maximal monotone, one must argue that any pair $(\hat{x}, \hat{y}) \in \mathbf{E} \times \mathbf{E}$ satisfying

$$\langle \hat{y} - y, \hat{x} - x\rangle \geq 0 \quad \forall (x, y) \in \operatorname{gph} T$$

must satisfy the inclusion $(\hat{x}, \hat{y}) \in \operatorname{gph} T$.

Coming back to the two main examples, we will now show that skew symmetric linear operators and subdifferentials of proper, closed, convex functions are indeed maximally monotone. The following lemma, in particular, guarantees that skew symmetric linear operators are indeed maximally monotone.

Lemma 3.73. *Any continuous monotone map $T\colon \mathbf{E} \to \mathbf{E}$ is maximally monotone.*

Proof. Consider a pair $(\hat{x}, \hat{y}) \in \mathbf{E} \times \mathbf{E}$ satisfying $\langle \hat{y} - T(x), \hat{x} - x \rangle \geq 0$ for all $x \in \mathbf{E}$. Then setting $x = \hat{x} - tu$ for some $t > 0$ and $u \in \mathbf{E}$, we deduce $\langle \hat{y} - T(\hat{x} - tu), u \rangle \geq 0$. Letting t tend to zero and appealing to continuity of T yields the estimate $\langle \hat{y} - T(\hat{x}), u \rangle \geq 0$ for all $u \in \mathbf{E}$. Setting $u = T(\hat{x}) - \hat{y}$, we conclude $\|\hat{y} - T(\hat{x})\|^2 \leq 0$ and therefore $\hat{y} = T(\hat{x})$, as we had to show. □

Checking maximal monotonicity from the definition is often difficult. The following lemma provides a convenient sufficient condition for a monotone operator T to be maximal monotone, namely surjectivity of the operator $I + T$. Maximal monotonicity of the subdifferential of a proper, closed, convex function will follow quickly from this lemma. We will later prove that surjectivity of $I + T$ is not only sufficient for maximal monotonicity but also necessary—a much deeper result.

Lemma 3.74 (Surjectivity is sufficient)**.** *Let $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ be a monotone operator such that $I + T$ is surjective. Then T is maximal monotone.*

Proof. Fix a pair $(\hat{x}, \hat{y}) \in \mathbf{E} \times \mathbf{E}$ satisfying

$$\langle \hat{y} - y, \hat{x} - x \rangle \geq 0 \qquad \forall (x, y) \in \operatorname{gph} T. \tag{3.35}$$

Since $I + T$ is surjective, there exists $x \in \mathbf{E}$ satisfying $\hat{x} + \hat{y} \in (I + T)(x)$. Therefore we may set $y = \hat{x} + \hat{y} - x$ in (3.35), thereby deducing

$$0 \leq \langle \hat{y} - (\hat{x} + \hat{y} - x), \hat{x} - x \rangle = -\|\hat{x} - x\|^2.$$

We conclude $x = \hat{x}$ and therefore $\hat{y} \in T(\hat{x})$, as we had to show. □

Lemma 3.75 (Convex subdifferential)**.** *Let $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a proper, closed, convex function. Then the subdifferential $\partial f\colon \mathbf{E} \rightrightarrows \mathbf{E}$ is maximal monotone.*

Proof. In light of Lemma 3.74, it suffices to argue that the map $I + \partial f$ is surjective. To this end, observe that $I + \partial f$ is precisely the subdifferential of

the function $g = f + \frac{1}{2}\|\cdot\|^2$. The conjugate $g^\star$ is the Moreau envelope of $f^\star$ and is therefore C^1-smooth by Theorem 3.64. Therefore given any $v \in \mathbf{E}$, the point $x := \nabla g^\star(v)$ satisfies $v \in \partial g(x)$ by Corollary 3.40. Thus ∂g is indeed surjective as claimed. □

Coming back to our two running examples, we have already observe that the sum of a skew symmetric linear operator and the subdifferential of a proper, closed, convex function is monotone. More importantly, the following exercises shows that the sum is *maximal* monotone.

Exercise 3.76. Let $\mathcal{A}\colon \mathbf{E} \to \mathbf{E}$ be a skew symmetric linear operator and let $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ be a maximal monotone operator. Show that the sum $\mathcal{A} + T$ is maximal monotone.
[**Hint:** Recall that if $\mathcal{A}$ is skew-symmetric, then equality $\langle \mathcal{A}x, x\rangle = 0$ holds for all $x \in \mathbf{E}$.]

3.8.2 The resolvant and the Minty parametrization

We next discuss an important generalization of the proximal map to monotone operators, which will form the core of primal-dual algorithms explored in Chapter **??**. Recall that the proximal operator of a proper, closed, convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is defined by

$$\operatorname{prox}_f(x) = \operatorname*{argmin}_z \left\{ f(z) + \frac{1}{2}\|z - x\|^2 \right\}.$$

Let us rewrite prox_f purely in terms of the subdifferential of f. Appealing to first-order optimality conditions, we may write:

$$z = \operatorname{prox}_f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(z) + (z - x) \quad \Longleftrightarrow \quad x \in (I + \partial f)(z).$$

Thus we may identify the proximal map as the set-valued inverse

$$\operatorname{prox}_f(x) = (I + \partial f)^{-1}(x).$$

This expression for the proximal map suggests a generalization to any set-valued operator.

Definition 3.77 (Resolvant)**.** Let $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ be a set-valued operator. Then the *resolvant operator* $\mathcal{R}_T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ is defined by

$$\mathcal{R}_T(x) = (I + T)^{-1}.$$

Similar to the proximal map, we expect the resolvant of a maximal monotone operator to be single-valued everywhere and non-expensive. This is indeed the case.

Theorem 3.78 (Properties of the resolvant)**.** *Let $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ be a maximal monotone operator. Then the resolvant $\mathcal{R}_T\colon \mathbf{E} \to \mathbf{E}$ is a globally defined single-valued map satisfying*

$$\|\mathcal{R}_T(x) - \mathcal{R}_T(y)\|^2 \leq \langle \mathcal{R}_T(x) - \mathcal{R}_T(y), x - y\rangle \qquad \forall x, y \in \mathbf{E}. \tag{3.36}$$

In particular, the resolvant $\mathcal{R}_T$ is non-expansive.

The most involved part of the proof of Theorem 3.78 is to show that the domain of $\mathcal{R}_T$ is all of $\mathbf{E}$, or equivalently that the operator $I + T$ is surjective. Notice this is precisely the converse of Lemma 3.74. We state this result as Theorem 3.79, but postpone its proof until Section 3.8.3, since it is somewhat long.

Theorem 3.79 (Surjectivity)**.** *Let $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ be a monotone operator. Then T is maximal monotone if and only if $I + T$ is surjective.*

With Theorem 3.79 at hand, the proof of Theorem 3.78 is straightforward.

Proof of Theorem 3.78. Theorem 3.79 guarantees that the resolvant $\mathcal{R}_T$ has as its domain all of $\mathbf{E}$. For any pairs $(x_1, y_1), (x_2, y_2) \in \operatorname{gph} \mathcal{R}_T$, the definition of the resolvant guarantees the inclusions $x_i - y_i \in T(y_i)$ for $i = 1, 2$. Using monotonicity of T, we deduce

$$0 \leq \langle (x_1 - y_1) - (x_2 - y_2), y_1 - y_2\rangle = \langle x_1 - x_2, y_1 - y_2\rangle - \|y_1 - y_2\|^2.$$

Rearranging, yields the guarantee

$$\|y_1 - y_2\|^2 \leq \langle x_1 - x_2, y_1 - y_2\rangle. \tag{3.37}$$

In particular, in the case $x_1 = x_2$, the right side becomes zero and therefore $y_1 = y_2$. We conclude that $\mathcal{R}_T(x)$ is a singleton for every x. The estimate (3.37) is then exactly (3.36). Combining (3.36) with the Cauchy-Schwarz inequality directly implies that $\mathcal{R}_T$ is non-expensive. □

The following exercise explores further properties of the resolvant. In particular, it shows that if $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ is a maximal monotone operator, then the map $z \mapsto (\mathcal{R}_T(z), \mathcal{R}_{T^{-1}}(z))$ is a bijection from $\mathbf{E}$ to $\operatorname{gph} T$, whose

inverse is simply the restriction of the linear map $(x, y) \mapsto x + y$ to $\operatorname{gph} T$. Moreover, the bijection is Lipschitz continuous in both directions. In this sense, the graph of any monotone operator is "thin" inside its ambient space $\mathbf{E} \times \mathbf{E}$.

Exercise 3.80 (Minty parametrization)**.** Consider a map $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$.

1. Show the inverse resolvant identity

$$\mathcal{R}_{T^{-1}} = I - \mathcal{R}_T.$$

Recognize this identity as the generalization of Theorem 3.61.

[**Hint:** Unrolling notation, the claimed equation amounts to the identity $I - (I + T)^{-1} = (I + T^{-1})^{-1}$. This identity can be proved directly from the definition of the inverse map. Show first that the inclusion $z \in (I - (I + T)^{-1})(x)$ is equivalent to $x - z \in T^{-1}(z)$; then add z to be both sides and conclude the result.]

2. Suppose now that T is maximal monotone and define the map

$$H(z) \mapsto (\mathcal{R}_T(z), \mathcal{R}_{T^{-1}}(z)).$$

Show that H is a Lipschitz continuous bijection from $\mathbf{E}$ to $\operatorname{gph} T$, whose inverse is the restriction of the linear map $(x, y) \mapsto x + y$ to $\operatorname{gph} T$. See Figure 3.18 for an illustration. Deduce that H furnishes a Lipschitz homeomorphism between $\operatorname{gph} T$ and $\mathbf{E}$.

[**Hint:** The fact that the image of H is contained in $\operatorname{gph} T$ follows directly from the definition of the resolvant and part 1. Lipschitz continuity of H follows from Theorem 3.78. The fact that the inverse of H is the restriction of the linear map $(x, y) \mapsto x + y$ to $\operatorname{gph} T$ again follows from the definition of the resolvant and part 1.

3.8.3 Proof of the surjectivity theorem.

In this section, we prove the surjectivity Theorem 3.79. The argument will be based on analyzing the subdifferential of a certain convex function associated to a monotone operator.

Definition 3.81 (Fitzpatrick function)**.** With any monotone operator $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$, we associate the *Fitzpatrick function* $F_T\colon \mathbf{E} \times \mathbf{E} \to \overline{\mathbf{R}}$ defined by

$$F_T(x, y) = \langle x, y \rangle - \inf_{(\bar{x}, \bar{y}) \in \operatorname{gph} T} \langle \bar{y} - y, \bar{x} - x \rangle.$$

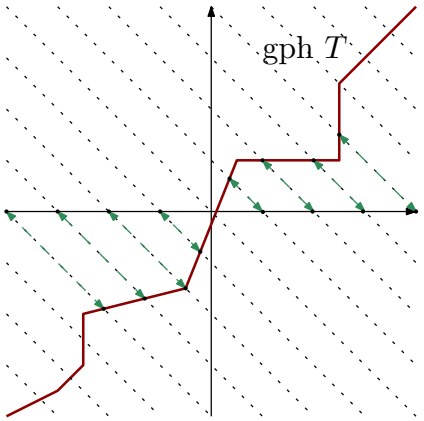


Figure 3.18: Illustration of the Minty parametrization of $\operatorname{gph} T$ in Exercise 3.80. The dashed black lines are the level sets of the linear map $(x, y) \mapsto x + y$, which restricts to a bijection between $\operatorname{gph} T$ and $\mathbf{R}$. The green arrows indicate the resulting bijection.

The main use of the Fitzpatrick function F_T is that it enables a generalization of the Fenchel-Young inequality for maximal monotone operators. Indeed, one should think of F_T as playing a similar role to the sum $f(x) + f^\star(y)$ in the case when $T = \partial f$ is the subdifferential of a proper, closed, and convex function f.

Lemma 3.82 (Fitzpatrick inequality)**.** *Let $T \colon \mathbf{E} \rightrightarrows \mathbf{E}$ be a maximal monotone operator. Then the Fitzpatrick function F_T is proper, closed, and convex. Moreover, for any pair $(x, y) \in \mathbf{E} \times \mathbf{E}$ the inequality*

$$F_T(x, y) \geq \langle x, y \rangle \qquad \text{holds,} \tag{3.38}$$

while equality holds if and only if $y \in T(x)$.

Proof. Observe that we may express the Fitzpatrick function as

$$\begin{aligned} F_T(x, y) &= \sup_{(\bar{x}, \bar{y}) \in \operatorname{gph} T} \{\langle x, y \rangle - \langle \bar{y} - y, \bar{x} - x \rangle\} \\ &= \sup_{(\bar{x}, \bar{y}) \in \operatorname{gph} T} \{\langle \bar{y}, x \rangle + \langle \bar{x}, y \rangle - \langle \bar{y}, \bar{x} \rangle\}. \end{aligned}$$

Thus F_T is a pointwise supremum of affine functions, and is therefore closed and convex. Next, observe that monotonicity of T guarantees

$$\inf_{(\bar{x}, \bar{y}) \in \operatorname{gph} T} \langle \bar{y} - y, \bar{x} - x \rangle = 0 \qquad \forall (x, y) \in \operatorname{gph} T,$$

while maximal monotonicity of T implies

$$\inf_{(\bar{x},\bar{y})\in \operatorname{gph} T} \langle \bar{y}-y, \bar{x}-x\rangle < 0 \qquad \forall (x,y)\in (\mathbf{E}\times\mathbf{E})\setminus \operatorname{gph} T.$$

Thus the inequality (3.38) always holds, while equality holds in (3.38) if and only if (x,y) lies in $\operatorname{gph} T$. Finally, the fact that F_T is proper is now immediate from (3.38) (why?). □

The next lemma shows that the inclusion $y\in T(x)$ can be verified using the subdifferential of the Fitzpatrick function.

Lemma 3.83 (Subdifferential of the Fitzpatrick function)**.** *Let $T\colon \mathbf{E}\rightrightarrows\mathbf{E}$ be a maximal monotone operator and fix a pair $(x_1,y_1)\in\mathbf{E}\times\mathbf{E}$ and a subgradient $(y_2,x_2)\in\partial F_T(x_1,y_1)$. Then the inequality*

$$\langle y_1-y_2, x_1-x_2\rangle \leq 0 \qquad \textit{holds.} \tag{3.39}$$

Moreover, if equality holds, then the inclusion $y_2\in T(x_2)$ is valid.

Proof. We compute

$$\begin{aligned}\langle y_1-y_2, x_1-x_2\rangle &= \langle y_1,x_1\rangle - \langle y_1,x_2\rangle - \langle y_2,x_1\rangle + \langle y_2,x_2,\rangle \\ &\leq F_T(x_1,y_1) - \langle y_1,x_2\rangle - \langle y_2,x_1\rangle + \langle y_2,x_2\rangle,\end{aligned} \tag{3.40}$$

where the inequality follows from Lemma 3.82. Aiming to upper bound $F_T(x_1,y_1)$, fix a pair $(\bar{x},\bar{y})\in\operatorname{gph} T$. Then using the subgradient inequality, we deduce

$$\begin{aligned}F_T(x_1,y_1) &\leq F_T(\bar{x},\bar{y}) + \langle (y_2,x_2),(x_1,y_1)-(\bar{x},\bar{y})\rangle \\ &= \langle \bar{x},\bar{y}\rangle + \langle y_2, x_1-\bar{x}\rangle + \langle x_2, y_1-\bar{y}\rangle,\end{aligned}$$

where the last equality follows from Lemma 3.82. Combining this estiamate with (3.40), we conclude

$$\langle y_1-y_2,x_1-x_2\rangle \leq \langle \bar{x},\bar{y}\rangle - \langle y_2,\bar{x}\rangle - \langle x_2,\bar{y}\rangle + \langle y_2,x_2\rangle = \langle \bar{y}-y_2,\bar{x}-x_2,\rangle.$$

Hence taking the infumum of the right-side over $(\bar{x},\bar{y})\in\operatorname{gph} T$ we conclude

$$\langle y_1-y_2,x_1-x_2\rangle \leq \inf_{(\bar{x},\bar{y})\in\operatorname{gph} T}\langle \bar{y}-y_2,\bar{x}-x_2,\rangle = \langle x_2,y_2\rangle - F_T(x_2,y_2).$$

An application of Lemma 3.82 completes the proof. □

We are now ready to prove Theorem 3.79.

Proof of Theorem 3.79. The backward implication is Lemma 3.74. Conversely, suppose that T is maximal monotone. We will show that $v = 0$ lies in the range of T. Surjectivity of T will then follow simply by replacing T with $T - \{v\}$ for an arbitrary vector $v \in \mathbf{E}$.

Suppose for the moment that we can find a pair $(x, y) \in \mathbf{E} \times \mathbf{E}$ satisfying the inclusion $-(x, y) \in \partial F_T(x, y)$. Then Lemma 3.83 immediately implies

$$0 \geq \langle y - (-x), x - (-y) \rangle = \|x + y\|^2 \geq 0$$

Thus equality holds throughout, and we conclude $-x \in T(-y)$ and $x = -y$. Therefore zero indeed lies in the range of $T + I$. It remains to construct the pair (x, y). To this end, define the strongly convex function

$$g(x, y) = F_T(x, y) + \frac{1}{2}(\|x\|^2 + \|y\|^2).$$

Let (x, y) be the minimizer of g. Then the optimality condition $(0, 0) \in \partial g(x, y)$ becomes $-(x, y) \in \partial F_T(x, y)$, and therefore (x, y) is exactly the pair we seek. □

Comments

Almost all of the material in this section can be found in standard textbooks on convex analysis (Bauschke-Combettes [4], Borwein-Lewis [8], Rockafellar [31], Hiriart-Urruty and Lemaréchal [17]) and variational analysis (Rockafellar-Wets [33], Mordukhovich [25]), with some significant variation in the order in which the material is presented and in the proofs. The driving theme of translating geometric properties of epigraphs to analytic properties of functions originates in [31]. The proof of the biconjugacy Theorem 3.26 is taken from [8], while the proof of the Baillon-Haddad Theorem 3.66 is taken from Bauschke-Combettes [3]. Section 3.8 is an introduction to monotone operator theory. For more details on the subject and historical references we refer the reader to Bauschke-Combettes [4]. The proof of the surjectivity theorem in Section 3.8.3 is due to Simons-Zălinescu [36]. The original proofs of the surjectivity theorem and maximal monotonicity of the subdifferential of a proper, closed, convex function is due to Rockafellar [32]. The Fitzpatrick function was introduced by Fitzpatrick in [14], and has now become a standard tool in monotone operator theory.

Chapter 4

Subdifferential calculus and primal/dual problems

Thus far, the only procedure we have for computing the subdifferential is directly from the definition. In this way, we have computed the subdifferentials of all the ℓ_p-norms and of the pointwise maximum $\mathrm{mx}\,(\cdot)$ in Exercise 3.41. Computing subdifferentials for a broader class of functions requires developing a calculus of subgradients. For example, we need calculus rules for computing subdifferentials of sums and of compositions of convex functions with linear maps. The goal of this chapter is to develop such a calculus. Along the way, we will see that subdifferential calculus is intimately tied to two seemingly unrelated topics: stability of an (1) optimization problem to perturbations and (2) existence of solutions to an auxiliary (dual) problem. Such primal/dual pairs of optimization problems are important in their own right, independently of calculus. Indeed, most of the chapter is devoted to the primal/dual formalism, with calculus one of its many consequences.

Calculus. Plainly put, the reason why subdifferential calculus is subtle is that functions may take infinite values. Before delving into technical details, it is best first to see a (pathological) example where a subdifferential rule may actually fail. Let A and B be two closed convex sets in $\mathbf{E}$. The sum rule of subdifferentials, which we expect to hold, would say

$$\partial(\delta_A + \delta_B)(x) = \partial\delta_A(x) + \partial\delta_B(x),$$

or equivalently

$$N_{A\cap B}(x) = N_A(x) + N_B(x). \tag{4.1}$$

The very definition of the normal cone shows that the inclusion $\supset$ always holds for any $x \in A \cap B$ (check this!). The reverse inclusion $\subset$ is much less clear at first sight. Indeed, verifying this inclusion amounts to showing that any vector $v \in N_{A\cap B}(x)$ can be decomposed into a sum of normals to individual sets A and B—in essence, an existence statement.

For the sake of building intuition, let us contrast two explicit instances of A and B, depicted in Figure 4.1.

Unstable Intersection: Define the shifted unit disks (Figure 4.1a)

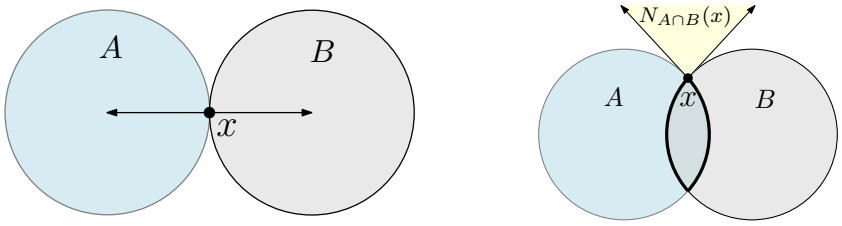
$$A := \operatorname{cl} B_1(0,-1) \qquad \text{and} \qquad B := \operatorname{cl} B_1(0,1) \tag{4.2}$$

The intersection $A \cap B$ consists only of the origin, and therefore the left-side of (4.1) evaluated at the origin is all of $\mathbf{R}^2$. The right-side in contrast equals $\mathbf{R} \times \{0\}$; thus, equality in (4.1) fails. The intuitive explanation for what has gone wrong is that the intersection $A \cap B$ is highly unstable under perturbations: translating A by a small vector v may yield an empty intersection $(A + v) \cap B = \emptyset$. In particular, the interiors of A and B do not meet.

Stable Intersection: Define the shifted disks (Figure 4.1b)

$$A := \operatorname{cl} B_1(0,-3/4) \qquad \text{and} \qquad B := \operatorname{cl} B_1(0,3/4) \tag{4.3}$$

and set $x = (0, \frac{\sqrt{7}}{4})$. A quick computation shows that the sum rule (4.1) holds. In contrast with the first example (4.2), the intersection $A \cap B$ is stable because the interiors of A and B do meet.



(a) Sum rule fails (b) Sum rule holds

Figure 4.1: Normal cone to an intersection.

The two instances (4.2) and (4.3) illustrate the subtlety of subdifferential calculus. In particular, it is clear that the sum rule will require some mild assumption on the way that the domains of the functions intersect.

Primal-dual pairs of optimization problems. We will see that validity of subdifferential sum and chain rules is intimately tied to success of a certain lower-bounding procedure. To illustrate, consider the problem of linear programming

$$(P) \qquad \min_{x \in \mathbf{R}^n} \ \langle b, x \rangle \qquad \text{subject to} \qquad Ax \geq c,$$

for some matrix $A \in \mathbf{R}^{m \times n}$ and vectors $b \in \mathbf{R}^n$ and $c \in \mathbf{R}^m$. How can we judge the quality of a putative feasible solution x? If we knew a good lower bound ℓ on the optimal value, we could compute the gap $f(x) - \ell$. One simple lower bound can be generated as follows. For any feasible point $x \in \mathbf{R}^n$ and any $y \geq 0$, the quantity $\langle y, c - Ax \rangle$ is negative. Therefore the estimate holds:

$$\langle b, x \rangle \geq \langle b, x \rangle + \langle y, c - Ax \rangle = \langle b - A^T y, x \rangle + \langle c, y \rangle.$$

Thus if we further impose the restriction $b = A^T y$, the estimate becomes independent of x and we obtain a lower bound on the optimal value of (P). Choosing the best such lower bound amounts to a new optimization problem

$$(D) \qquad \max_{y \in \mathbf{R}^m} \ \langle c, y \rangle \qquad \text{subject to} \qquad A^T y = b, \ y \geq 0.$$

This problem is called the dual. A central result of linear programming is the strong duality theorem: as long as the optimal value of (P) or (D) is finite, the two are equal and are attained. In particular, the best lower bound achieved by the outlined procedure matches the true optimal value of (P). Strong duality is an existence statement, much like the subdifferential sum rule, since it ensure existence of an optimal dual solution.

The Fenchel-Rockafellar and Lagrangian dualities extend the formalism outlined about to nonlinear convex optimization problems. Strong duality holding in these more general settings is equivalent to validity of a sum and chain rules for subdifferentials.

Roadmap. We begin developing subdifferential calculus by computing the subdifferential of the infimal projection $p(y) = \inf_x F(x, y)$ in Section 4.1. The main observation is that $\partial p(0)$ consists of solutions of an auxiliary optimization problem. Building on this, Section 4.2 introduces Fenchel-Rockafellar, Lagrangian, and minimax primal/dual pairs, and deduces the sum and chain rules for subdifferential calculus. Section 3.8 introduces basic properties of monotone operators, with the key surjectivity theorem proved

using strong duality. The final Section 4.3 is devoted to computing the subdifferential of spectral matrix functions; these are functions of symmetric matrices that depend on the matrix only through its eigenvalues.

4.1 The subdifferential of the value function

When solving an optimization problem, one is often interested not only in the optimal value, but also in how this value changes under small perturbation to the problem data. More formally, consider an arbitrary convex function $F\colon \mathbf{E} \times \mathbf{Y} \to \overline{\mathbf{R}}$ and the parametric optimization problem:

$$p(y) := \inf_x \; F(x, y)$$

The reader should think of y as a perturbation parameter and the problem corresponding to $p(0)$ as the original "primal" problem. The assignment $y \mapsto p(y)$ is called the *value function*. The reader should recognize p as the infimal projection of F.

A worthwhile goal is to study how $p(y)$ varies as y is perturbed around the origin. Thus with the machinery we have developed, we aim to compute the subdifferential $\partial p(0)$. Subdifferential calculus and duality theory—the two themes of the chapter—will follow quickly under an appropriate choice of the function F. Remarkably, we will see that $\partial p(0)$ coincides with the set of maximizers of an auxiliary convex optimization problem. To make this idea precise, define the new parametric family of problems

$$q(x) := \sup_y \; -F^\star(x, y).$$

Let us call the problem corresponding to $q(0)$ the *parametric dual*. The terminology is easy to explain. Indeed, looking at the Table 3.4 and invoking Exercise 3.28 yields the equality $p^\star(y) = F^\star(0, y)$, and therefore

$$p^{\star\star}(0) = \sup_y \; \{\langle 0, y\rangle - p^\star(y)\} = \sup_y -F^\star(0, y) = q(0). \tag{4.4}$$

Thus $q(0)$ is the biconjugate of p evaluated at zero. Hence we expect that under mild conditions, equality $p(0) = q(0)$ should hold. This is indeed the case, and moreover, the subdifferential $\partial p(0)$ consists precisely of the optimal solutions y to the parametric dual problem.

Theorem 4.1 (Parametric optimization)**.** *Suppose that $F\colon \mathbf{E} \times \mathbf{Y} \to \overline{\mathbf{R}}$ is proper, closed, and convex. Then the following are true.*

1. **(Weak duality)** *The inequality* $p(0) \geq q(0)$ *always holds.*
2. **(Subdifferential)** *If* $p(0)$ *is finite, then the inclusion holds:*

$$\partial p(0) \subset \operatorname*{argmax}_y \; -F^\star(0, y). \tag{4.5}$$

   If in addition, the inclusion $0 \in \operatorname{ri}(\operatorname{dom} p)$ *holds, then* (4.5) *holds with equality.*
3. **(Strong duality)** *If the subdifferential* $\partial p(0)$ *is nonempty, then equality* $p(0) = q(0)$ *holds and the supremum* $q(0)$ *is attained.*

Proof. Part (1) is immediate from the inequality $p(0) \geq p^{\star\star}(0)$. Next suppose $p(0)$ is finite. Observe that a vector ϕ satisfies $\phi \in \partial p(0)$ if and only if for all y it holds:

$$p(0) \leq p(y) - \langle \phi, y \rangle = \inf_x \left\{ F(x, y) - \left\langle \begin{pmatrix} 0 \\ \phi \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix} \right\rangle \right\}.$$

Taking the infimum over y, we deduce $\phi \in \partial p(0)$ if and only if $p(0) \leq -F^\star(0, \phi)$. Thus in light of (1) any subgradient $\phi \in \partial p(0)$ is dual optimal, that is (4.5) holds. Suppose now $p(0)$ is finite and the inclusion $0 \in \operatorname{ri} \operatorname{dom} p$ holds. Observe that then p is proper (Exercise 3.2). Moreover, since p is lower-semicontinous on a neighborhood of 0, the equality $p^{\star\star}(0) = (\overline{\operatorname{co}}\, p)(0) = p(0)$ holds. Consider a vector $\phi \in \operatorname{argmax}_y -F^\star(0, y)$. Since $p^\star$ coincides with $F^\star(0, \cdot)$, we conclude $0 \in \partial p^\star(\phi)$. Corollary 3.40 therefore guarantees $\phi \in \partial p^{\star\star}(0) \subset \partial p(0)$. Thus equality holds in (4.5).

Finally, suppose that there exists a subgradient $\phi \in \partial p(0)$. Since the function $y \mapsto p(0) + \langle \phi, y \rangle$ is an affine minorant of p, using Theorem 3.18 we deduce

$$p(0) + \langle \phi, 0 \rangle \leq (\overline{\operatorname{co}}\, p)(0) = p^{\star\star}(0) \leq p(0).$$

Thus, equality $p(0) = q(0)$ holds, while the supremum $q(0)$ is clearly attained by (2). The proof is complete. □

4.2 Duality and subdifferential calculus

The idea of duality has appeared throughout the previous sections, culminating in the definition of the Fenchel conjugate. In this section, we will use these ideas to investigate the so-called *primal-dual pairs* of convex optimization problems. Roughly speaking, we will see that for a number of well-structured convex minimization problems, there are natural ways to

obtain lower bounds on their optimal value. The task of finding the largest such lower bound is itself another optimization problem, called the dual. A central question is therefore to determine conditions ensuring that the best lower-bound matches the primal optimal value. As a bonus, we would also like to know that optimality in the primal can indeed be certified by a dual feasible solution, or in other words that the dual optimal value is attained. The answer for both questions will come directly from Theorem 4.1. The sum and chain rules for subdifferentials will follow immediately from dual attainment.

4.2.1 Fenchel-Rockafellar duality

Consider a general class of structured optimization problems

$$(P) \qquad \inf_{x \in \mathbf{E}} \; h(\mathcal{A}x) + g(x),$$

where $h \colon \mathbf{Y} \to \overline{\mathbf{R}}$ and $g \colon \mathbf{E} \to \overline{\mathbf{R}}$ are some proper functions and $\mathcal{A} \colon \mathbf{E} \to \mathbf{Y}$ is a linear map. Let us call this problem (P) the *primal*. Define now a new convex optimization problem, called the *dual*:

$$(D) \qquad \sup_{y \in \mathbf{Y}} \; - h^\star(y) - g^\star(-\mathcal{A}^* y).$$

The dual problem (D) arises naturally from a lower-bounding procedure. Let us try to find simple lower bounds for $\mathrm{val}(P)$, the optimal value of (P). From the Fenchel-Young inequality, any $\bar{y} \in \operatorname{dom} h^\star$ yields the lower bound:

$$\begin{aligned}
\mathrm{val}(P) &= \inf_x \, h^{\star\star}(\mathcal{A}x) + g(x), \\
&\geq \inf_x \; \langle \bar{y}, \mathcal{A}x \rangle - h^\star(\bar{y}) + g(x) \\
&= -h^\star(\bar{y}) - \sup_{x \in \mathbf{E}} \; \{ \langle -\mathcal{A}^* \bar{y}, x \rangle - g(x) \} \\
&= -h^\star(\bar{y}) - g^\star(-\mathcal{A}^* \bar{y}).
\end{aligned}$$

The right-hand-side is exactly the evaluation of the dual objective function at $\bar{y}$. Thus $\mathrm{val}(D)$ is the supremum over all lower-bounds on $\mathrm{val}(P)$ that can be obtained in this way. In particular, we have deduced the *weak-duality* inequality

$$\mathrm{val}(P) \geq \mathrm{val}(D).$$

Table 4.1 lists a few notable examples of Fenchel-Rockafellar dual problems.

Primal (P)	**Dual** (D)
$\min_x \ \frac{1}{2}\|Ax-b\|_2^2 + \|x\|_1$	$\max_y \ \left\{-\frac{1}{2}\|y\|^2 - \langle b, y\rangle : \|A^T y\|_\infty \le 1\right\}$
$\min_{x:\,\|x\|_q \le 1} \|Ax-b\|_p$	$\max_{y:\|y\|_{\bar{p}}\le 1} \ -\|A^T y\|_{\bar{q}} - \langle b, y\rangle$
$\min_x \ \{\langle c, x\rangle : Ax = b, x \in K\}$	$\max_y \ \{\langle b, y\rangle : A^T y - c \in K^\circ\}$
$\min_x \ \left\{\frac{1}{2}\langle Qx, x\rangle + \langle c, x\rangle : Ax \ge b\right\}$	$\max_{y\ge 0} -\frac{1}{2}\langle Q^{-1}(c-A^T y), c - A^T y\rangle + \langle b, y\rangle$

Table 4.1: Fenchel-Rockafellar dual pairs. The parameters are: K is a convex cone, $Q \succ 0$, and $p, \bar{p}, q, \bar{q} \in [1, \infty]$ satisfy $p^{-1} + \bar{p}^{-1} = q^{-1} + \bar{q}^{-1} = 1$.

Exercise 4.2. Verify that the problems in the second column of Table 4.1 are the Fenchel-Rockafellar duals of the problems in the first column.
[**Hint:** You may find Exercise 3.27 helpful.]

The goal now is to show that when h and g are proper, closed, and convex and a mild compatibility condition holds, we can be sure that *strong duality* holds: $\mathrm{val}(P) = \mathrm{val}(D)$ and the dual optimal value is attained. The argument proceeds by interpreting Fenchel-Rockafellar duality within the parametric framework of Theorem 4.1. The perturbation function will take the following simple form

$$p(y) = \inf_x h(\mathcal{A}x + y) + g(x).$$

In order to guarantee the inclusion $0 \in \mathrm{ri}\,(\mathrm{dom}\, p)$, we will require the following simple lemma establishing a calculus rule for the relative interior.

Lemma 4.3. *(Calculus of relative interiors) For any two nonempty convex sets $Q \in \mathbf{E}$, $P \in \mathbf{Y}$ and a linear map $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$, inclusions hold:*

$$\mathcal{A}(\mathrm{ri}\, Q) + \mathrm{ri}\, P \subset \mathrm{ri}\,(\mathcal{A}(Q) + P), \tag{4.6}$$

$$\mathrm{ri}\,(Q) \cap \mathcal{A}^{-1}(\mathrm{ri}\, P) \subset \mathrm{ri}\,(Q \cap \mathcal{A}^{-1}P). \tag{4.7}$$

Proof. We first prove (4.6). Without loss of generality, we may assume $P = \{0\}$, since otherwise we may replace $\mathcal{A}$ with the new linear map $(x, y) \mapsto \mathcal{A}x + y$ and replace Q with $Q \times P$ (check this!). Translating Q if necessary, we may assume that Q contains the origin. Let $\overline{\mathcal{A}}\colon \mathrm{aff}\, Q \to \mathcal{A}(\mathrm{aff}\, Q)$ be the linear map obtained by restricting the domain and range of $\mathcal{A}$. By definition, $\overline{\mathcal{A}}$ is a surjective linear map and therefore maps open sets to open sets. In

particular, the set $\mathcal{A}(\operatorname{ri} Q)$ is contained in $\mathcal{A}(Q)$ and is open in $\mathcal{A}(\operatorname{aff} Q)$. Therefore the inclusion $\mathcal{A}(\operatorname{ri} Q) \subset \operatorname{ri} \mathcal{A}(Q)$ holds, as claimed.

Next, we prove (4.7). Without loss of generality, we may assume $Q = \mathbf{E}$, since otherwise we may redefine the linear map $\mathcal{A}$ to be $x \mapsto (x, \mathcal{A}x)$ and replace P with $Q \times P$ (check this!). Define the map $\overline{\mathcal{A}}\colon \mathcal{A}^{-1}(\operatorname{aff} P) \to \operatorname{aff} P$ to be the restriction of $\mathcal{A}$. By continuity of $\bar{\mathcal{A}}$, the preimage $\mathcal{A}^{-1}(\operatorname{ri} P)$ is open in $\mathcal{A}^{-1}(\operatorname{aff} P)$ and is contained in $\mathcal{A}^{-1}P$. The inclusion $\mathcal{A}^{-1}(\operatorname{ri} P) \subset \operatorname{ri}(\mathcal{A}^{-1}P)$ follows immediately, as claimed. □

We are now ready to prove the main result of the section.

Theorem 4.4 (Fenchel-Rockafellar duality)**.** *Consider the problems:*

$$\begin{aligned} (P) \qquad & \min_x \; h(\mathcal{A}x) + g(x) \\ (D) \qquad & \max_y \; -g^\star(-\mathcal{A}^* y) - h^\star(y). \end{aligned}$$

where $g\colon \mathbf{E} \to \overline{\mathbf{R}}$ and $h\colon \mathbf{Y} \to \overline{\mathbf{R}}$ are proper, closed convex functions, and $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$ is a linear map. Suppose that the regularity condition holds:

$$0 \in \operatorname{ri}(\operatorname{dom} h) - \mathcal{A}(\operatorname{ri} \operatorname{dom} g) \tag{4.8}$$

Then the primal and dual optimal values are equal and the dual optimal value is attained, if finite.

Proof. We first show that the two problems (P) and (D) fit the perturbation framework of Theorem 4.1 with

$$F(x, y) = h(\mathcal{A}x + y) + g(x).$$

To this end, observe

$$F^\star(x, y) = \sup_{z,w} \{\langle (z, w), (x, y)\rangle - h(\mathcal{A}z + w) - g(z)\}.$$

Making the substitution $v := \mathcal{A}z + w$ yields

$$\begin{aligned} F^\star(x, y) &= \sup_{z,v} \{\langle z, x\rangle + \langle v - \mathcal{A}z, y\rangle - h(v) - g(z)\} \\ &= \sup_z \{\langle z, x - \mathcal{A}^* y\rangle - g(z)\} + \sup_v \{\langle v, y\rangle - h(v)\} \\ &= g^\star(x - \mathcal{A}^* y) + h^\star(y). \end{aligned}$$

Thus the Fenchel dual problem (D) is exactly $q(0) = \sup_y -F^\star(0, y)$. A quick computation shows

$$\operatorname{dom} p = \operatorname{dom} h - \mathcal{A}(\operatorname{dom} g). \tag{4.9}$$

Lemma 4.3, allows us to take relative interiors of both sides in (4.9). Therefore, assumption (4.8) implies $0 \in \operatorname{ri}(\operatorname{dom} p)$. If $p(0) = -\infty$, there is nothing to prove. Hence we may suppose that $p(0)$ is finite. Theorem 3.2 then implies that p is proper, while Theorem 3.38 in turn guarantees that the subdifferential $\partial p(0)$ is nonempty. Finally, an application of Theorem 4.1 completes the proof. □

In many applications, the primal problem (P) looks slightly different:

$$(P) \qquad \min_x \ \langle c, x\rangle + h(b - \mathcal{A}x) + g(x),$$

for some vectors $b \in \mathbf{Y}$ and $c \in \mathbf{E}$. Some thought shows that this formulation can be put into the standard form of Theorem 4.4 simply by replacing g with $g + \langle c, \cdot\rangle$ and replacing h with $h(b - \cdot)$. Then the dual reads:

$$(D) \qquad \max_y \ \langle b, y\rangle - g^\star(\mathcal{A}^* y - c) - h^\star(y),$$

and the regularity condition (4.8) becomes

$$b \in \mathcal{A}(\operatorname{ri} \operatorname{dom} g) + \operatorname{ri}(\operatorname{dom} h).$$

Among the most important consequences of Theorem 4.4 are a sum and a chain rule for subdifferentials.

Theorem 4.5 (Subdifferential calculus)**.** *Let $g\colon \mathbf{E} \to \overline{\mathbf{R}}$ and $h\colon \mathbf{Y} \to \overline{\mathbf{R}}$ be proper, closed convex functions and $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$ a linear map. Then for any point x, the inclusion holds:*

$$\partial(g + h \circ \mathcal{A})(x) \supset \partial g(x) + \mathcal{A}^* \partial h(\mathcal{A}x). \tag{4.10}$$

Moreover, equality holds under the regularity condition

$$0 \in \operatorname{ri}(\operatorname{dom} h) - \mathcal{A}(\operatorname{ri} \operatorname{dom} g). \tag{4.11}$$

Proof. The inclusion 4.10 follows immediately by adding the subgradient inequalities for g at x and for h at $\mathcal{A}x$. Assume now (4.11) holds. Fix a

vector $v \in \partial(g + h \circ \mathcal{A})(x)$. Without loss of generality we may assume $v = 0$, since otherwise, we may replace g by $g - \langle v, \cdot \rangle$. Thus x minimizes

$$(P) \qquad \min_z \; h(\mathcal{A}x) + g(x).$$

Consider now the dual problem

$$(D) \qquad \max_y \; -g^\star(-\mathcal{A}^* y) - h^\star(y).$$

Theorem 4.4 guarantees that the primal and dual optimal values are equal and the dual optimal value is attained. Letting y be any dual optimal solution, we deduce

$$\begin{aligned} 0 &= (g(x) + h(\mathcal{A}x)) + (g^\star(-\mathcal{A}^* y) + h^\star(y)) \\ &= (g(x) + g^\star(-\mathcal{A}^* y)) + (h(\mathcal{A}x) + h^\star(y)) \\ &\geq \langle x, -\mathcal{A}^* y \rangle + \langle \mathcal{A}x, y \rangle = 0, \end{aligned} \tag{4.12}$$

where the (4.12) follows from the Fenchel-Young inequality (Theorem 3.39). Hence equality holds throughout and we learn

$$g(x) + g^\star(-\mathcal{A}^* y) = \langle x, -\mathcal{A}^* y \rangle \qquad \text{and} \qquad h(\mathcal{A}x) + h^\star(y) = \langle \mathcal{A}x, y \rangle.$$

Using the characterization of equality in Theorem 3.39 yields the decomposition

$$0 = -\mathcal{A}^* y + \mathcal{A}^* y \in \partial g(x) + \mathcal{A}^* \partial h(\mathcal{A}x),$$

as claimed. □

The calculus of normal cones is now an immediate consequence

Corollary 4.6 (Normal cone calculus)**.** *Let $A, B \subset \mathbf{E}$ be nonempty, closed, convex sets. Then for any point $x \in A \cap B$, the inclusion holds:*

$$N_{A \cap B}(x) \supset N_A(x) + N_B(x).$$

Moreover, equality holds under the regularity condition $(\operatorname{ri} A) \cap (\operatorname{ri} B) \neq \emptyset$.

We derived subdifferential sum and chain rules as a consequence of strong duality. The following exercise, shows an essential converse.

Exercise 4.7. Consider the primal and dual problems in Theorem 4.4, where g and h are proper, closed, convex functions. Suppose that x is a minimizer of the primal (P) and it satisfies the calculus rule:

$$0 \in \mathcal{A}^* \partial h(\mathcal{A}x) + \partial g(x).$$

Show that strong duality holds: the optimal values of (P) and (D) are equal, and the dual admits a maximizer.
[**Hint:** By the assumptions, we may write $0 = \mathcal{A}^*y + w$ for some $y \in \partial h(\mathcal{A}x)$ and $w \in \partial g(x)$. Using the inversion formula (Corollary 3.40), arrive at the inclusion $0 \in \partial h^\star(y) - \mathcal{A}\partial g^\star(-\mathcal{A}^*y)$. Deduce that y is optimal for the dual (D). To show that the optimal values of (P) and (D) are equal, simplify the quantity $(g(x) + h(\mathcal{A}x)) + (g^\star(-\mathcal{A}^*y) + h^\star(y))$ using the Fenchel-Young inequality.

The following two exercises provide two useful calculus rules that follow by applying Theorem 4.6 to epigraphs.

Exercise 4.8 (Pointwise maximum)**.** Consider arbitrary convex functions $f_i \colon \mathbf{E} \to \mathbf{R}$ for $i = 1, \ldots, k$ and define the function $g(x) = \max\{f_1(x), \ldots, f_k(x)\}$. Prove the expression:

$$\partial g(x) = \operatorname{conv} \bigcup_{i \in I} \partial f_i(x),$$

where $I = \{i : f_i(x) = g(x)\}$ is the active index set.
[**Hint:**. Compute the normal cone to the intersection $\bigcap_{i=1}^k \operatorname{epi} f_i$ using the Corollary 4.6.]

Exercise 4.9 (Normal cone to sublevel sets)**.** Consider a proper, closed, convex function $f \colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point $\bar{x} \in \operatorname{int}(\operatorname{dom} f)$ such that $f(\bar{x})$ is not the minimum of f. Define the sublevel set $Q := \{x : f(x) \leq f(\bar{x})\}$. Prove the formula

$$N_Q(\bar{x}) = \mathbf{R}_+ \partial f(\bar{x}).$$

[**Hint:** Focus on the epigraphs.]

In the case of linear programming, strong duality holds as long as the optimal value is finite. That is, the regularity condition (4.11) is not needed. The following exercise proves this fact.

Exercise 4.10. A convex function $f \colon \mathbf{R}^n \to \overline{\mathbf{R}}$ is called polyhedral if its epigraph is a polyhedron. Any proper polyhedral function can be written as

$$f(x) = \max_{i=1,\ldots,k} \{\langle u_i, x \rangle + q_i\} + \delta_P(x),$$

for some $u_i \in \mathbf{R}^n$ and $q_i \in \mathbf{R}$ and some polyhedron $P \subset \mathbf{R}^n$. (You don't need to prove this).

1. Show that if a polyhedral function is finite at some point, then it must be proper.
2. Show that a polyhedral function f has a subgradient at every point x at which it is finite. You may use the following two intuitive facts without proof: (i) the image of a polyhedron under a linear map is a polyhedron and (ii) the infimum of a linear function over a polyhedron is always attained, if finite.
3. Consider the linear programming problem and its dual:

$$\begin{aligned} (P) \qquad & \min_{x \in \mathbf{R}^n} \ \langle b, x \rangle \qquad \text{subject to} \qquad Ax \geq c, \\ (D) \qquad & \max_{y \in \mathbf{R}^m} \ \langle c, y \rangle \qquad \text{subject to} \qquad A^T y = b, \ y \geq 0. \end{aligned}$$

   for some matrix $A \in \mathbf{R}^{m \times n}$ and vectors $b \in \mathbf{R}^n$ and $c \in \mathbf{R}^m$. Show that if the optimal value of (P) is finite, then the optimal values of (P) and (D) are equal and the dual (D) admits a maximizer.

[**Hint:** For the second part, use Exercise 4.8. For the third part, define the value function

$$p(y) = \inf_{x \in \mathbf{R}^n} \ \langle b, x \rangle + \delta_{\{c\} + \mathbf{R}^m_+}(Ax + y).$$

Recognize (D) as the parametric dual and argue that $p(\cdot)$ is a proper polyhedral function. Use part two of the exercise and Theorem 4.1 to complete the proof.]

Notice that the optimality conditions for the primal (P) and the dual (D) problems, under mild regularity conditions, read as

$$\begin{Bmatrix} 0 \in \mathcal{A}^* \partial h(\mathcal{A}x) + \partial g(x) \\ 0 \in -\mathcal{A} \partial g^\star(-\mathcal{A}^\star y) + \partial h^\star(y) \end{Bmatrix}.$$

This system of inclusions is not very convenient for two reasons. First, the variable x and y appear unrelated, even though they are closely related as we will see. Second, the fact that the subdifferentials ∂h and ∂g are evaluated at points in the image of $\mathcal{A}$ and $\mathcal{A}^*$, respectively, is inconvenient for computation. We end the section by deriving "primal-dual optimality conditions" that simultaneously capture optimality for the primal and dual problems in a convenient form. Such optimality conditions will play an important role in Chapter 7, when designing primal-dual algorithms. The proof follows similar reasoning as that of Theorem 4.5.

Corollary 4.11 (Primal-dual optimality conditions)**.** *Consider the problems:*

$$\begin{aligned} (P) \qquad & \min_x \; h(\mathcal{A}x) + g(x) \\ (D) \qquad & \max_y \; -g^\star(-\mathcal{A}^* y) - h^\star(y). \end{aligned}$$

where $g\colon \mathbf{E} \to \overline{\mathbf{R}}$ *and* $h\colon \mathbf{Y} \to \overline{\mathbf{R}}$ *are proper, closed, convex functions, and* $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$ *is a linear map. Suppose that the optimal values of* (P) *and* (D) *are equal, as is implied for example by either of the two regularity conditions:*

$$\begin{aligned} 0 &\in \operatorname{ri}(\operatorname{dom} h) - \mathcal{A}(\operatorname{ri} \operatorname{dom} g) \\ 0 &\in \operatorname{ri}(\operatorname{dom} g^\star) + \mathcal{A}^*(\operatorname{ri} \operatorname{dom} h^\star). \end{aligned}$$

Then x *is the minimizer of* (P) *and* y *is the maximizer of* (D) *if and only if the inclusion holds:*

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} 0 & \mathcal{A}^* \\ -\mathcal{A} & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \partial g(x) \times \partial h^\star(y). \tag{4.13}$$

Proof. Since the primal and dual optimal values are equal, we deduce that x is a minimizer of (P) and y is a maximizer of (D) if and only if equality holds:

$$0 = (h(\mathcal{A}x) + g(x)) + (g^\star(-\mathcal{A}^* y) + h^\star(y)). \tag{4.14}$$

The Fenchel-Young inequality (Theorem 3.39) guarantees

$$h(\mathcal{A}x) + h^\star(y) \geq \langle \mathcal{A}x, y \rangle \qquad \text{and} \qquad g^\star(-\mathcal{A}^* y) + g(x) \geq \langle -\mathcal{A}^* y, x \rangle. \tag{4.15}$$

Adding the two inequalities in (4.15), we see that the right side of (4.14) is always lower-bounded by zero. We therefore deduce that (4.14) holds if and only if the inequalities (4.15) hold as equalities. Theorem 3.39 guarantees that this happens precisely when the inclusions, $\mathcal{A}x \in \partial h^\star(y)$ and $-\mathcal{A}^* y \in \partial g(x)$, hold. These two inclusions are exactly the system (4.13). □

4.2.2 Lagrangian Duality

We next look at the principle of duality for a problem class that does not directly fit in the framework of Fenchel-Rockafellar. The standard problem of *nonlinear programming* takes the shape

$$\begin{aligned} \min_x \;\; & f(x) \\ \text{s.t.} \;\; & g_i(x) \leq 0 \qquad \forall i = 1, \ldots, k \\ & g_i(x) = 0 \qquad \forall i = k+1, \ldots, m \end{aligned} \tag{P}$$

for some proper functions $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and $g_i\colon \mathbf{E} \to \mathbf{R}$. Notice that we may write this problem in a form resembling the Fenchel-Rockafellar framework,

$$\min_x \ f(x) + h(G(x)),$$

by setting $G(x) = (g_1(x), \dots, g_m(x))$ and $h = \delta_{\mathbf{R}^k_- \times \{0\}_{m-k}}$. Since G is nonlinear, we can not appeal to Fenchel-Rockafeller duality. Instead, we will use a direct lower-bounding procedure to derive a dual formulation, which will be called the Lagrange dual.

To simplify notation, define

$$G(x) := (g_1(x), \dots, g_m(x)) \qquad \text{and} \qquad K := \mathbf{R}^k_- \times \{0\}_{m-k}.$$

Not that the polar of K is given by $K^\circ := \mathbf{R}^k_+ \times \mathbf{R}^{m-k}$. Next, define the *Lagrangian function*

$$L(x, y) := f(x) + \langle y, G(x) \rangle.$$

A quick computation shows (verify this!)

$$\sup_{y \in K^\circ} L(x, y) = \begin{cases} f(x) & \text{if } G(x) \in K \\ +\infty & \text{otherwise} \end{cases}.$$

Consequently, taking the infimum over $x \in \mathbf{E}$, we deduce

$$\inf_x \sup_{y \in K^\circ} L(x, y) = \operatorname{val}(P).$$

Hence any $\bar{y} \in K^\circ$ certifies a lower bound on the primal optimal value:

$$\operatorname{val}(P) \geq \inf_x L(x, \bar{y}).$$

Finding the best lower bound that is achievable in this way is the Lagrange dual optimization problem:

$$\sup_{y \in K^\circ} \Phi(y) \qquad \text{where} \qquad \Phi(y) := \inf_x L(x, y). \tag{D}$$

The following exercise asks the reader to compute the Lagrangian dual of a few explicit problems.

Exercise 4.12. Compute the Lagrangian duals of the following problems

1. $\min_x \ \{\langle c, x \rangle : \mathcal{A}x = b, x \in Q\}$, where Q is a closed convex cone.

2. $\min_x\{\frac{1}{2}\langle Qx, x\rangle : Ax \leq b\}$, where $Q \succ 0$.
3. $\min_{x\in\mathbf{R}^n}\{\sum_{i=1}^n f(x_i) : Ax \leq b\}$, for a univariate function $f\colon \mathbf{R} \to \overline{\mathbf{R}}$.

We will see that under reasonable convexity and compatibility assumptions, we can be sure that *strong duality* holds: $\operatorname{val}(P) = \operatorname{val}(D)$ and the dual optimal value is attained, if finite. First, we need the following observation.

Exercise 4.13. ✍ Show that if $\{g_i\}_{i=1}^k$ are convex and $\{g_i\}_{i=k+1}^m$ are affine functions, then the function $(x, y) \mapsto \delta_K(G(x) + y)$ is convex.

To establish strong duality, we simply apply Theorem 4.1 to the Lagrange primal-dual pair.

Theorem 4.14. *Consider the constrained optimization problem (P), where $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is closed, proper, and convex, $g_i\colon \mathbf{E} \to \mathbf{R}$ are convex for $i = 1, \ldots, k$, and $g_i\colon \mathbf{E} \to \mathbf{R}$ are affine functions for $i = k+1, \ldots, m$. Then the primal (P) and the dual problems (D) fit the perturbation framework of Theorem 4.1 with*

$$F(x, y) = f(x) + \delta_K(G(x) + y). \tag{4.16}$$

In particular, the regularity condition

$$\exists x \in \operatorname{ri}(\operatorname{dom} f) \text{ with } G(x) \in \mathbf{R}^k_{--} \times \{0\}_{m-k} \tag{4.17}$$

guarantees that the primal and dual optimal values are equal, and the dual optimal value is attained, if finite.

Proof. Let us verify the assumptions of Theorem 4.1. To this end, let $F(\cdot, \cdot)$ be the function defined in (4.16). Note that F is clearly proper and closed, and is convex by Exercise 4.13. We successively compute

$$\begin{aligned}
F^\star(0, y) &= \sup_{x,\, v} \ \langle (0, y), (x, v)\rangle - F(x, v) \\
&= \sup_{x,\, v} \ \langle v, y\rangle - f(x) - \delta_K(G(x) + v) \\
&= \sup_{x,\, z} \ \langle y, z - G(x)\rangle - f(x) - \delta_K(z) \\
&= \sup_{z\in K} \ \langle y, z\rangle - \inf_x\{\langle y, G(x)\rangle + f(x)\} \\
&= \delta_{K^\circ}(y) - \Phi(y).
\end{aligned}$$

Thus the Lagrange dual (D) is precisely the problem $q(0) = \max_y -F^\star(0, y)$. Next, observe the equality

$$\operatorname{dom} F = (\operatorname{dom} f \times \mathbf{R}^m) \cap \{(x, y) : G(x) + y \in K\}.$$

Let x be a point satisfying Assumption (4.17). Lemma 4.3 and the continuity assumption on G guarantee the inclusion $(x, 0) \in \operatorname{ri}(\operatorname{dom} F)$ (verify this!). Since the domain of $p(\cdot)$ is the image of $\operatorname{dom} F$ under the projection $(x, y) \mapsto y$, Lemma 4.3 guarantees the inclusion $0 \in \operatorname{ri}(\operatorname{dom} p)$. If $p(0) = -\infty$, there is nothing to prove. Hence we may suppose that $p(0)$ is finite. Exercise 3.2 then implies that p is proper, while Theorem 3.38 in turn guarantees that the subdifferential $\partial p(0)$ is nonempty. An application of Theorem 4.4 completes the proof. □

Exercise 4.15. Consider the problem

$$\begin{array}{lll} (P) & \min & \langle Q_0 x, x\rangle \\ & \text{s.t.} & \langle Q_j x, x\rangle = b_j \quad \text{for } j = 1, \ldots, m \end{array}$$

where Q_j (for $j = 0, 1, \ldots, m$) are $n \times n$ symmetric matrices and $b \in \mathbf{R}^m$ is a vector.

1. Prove that the Lagrangian dual of this problem is the *Semi-definite Program*

$$\begin{array}{lll} (D) & \max\limits_{y} & y^T b \\ & \text{s.t.} & Q_0 - \sum\limits_{j=1}^{m} y_j Q_j \succeq 0 \end{array} \tag{4.18}$$

2. Use Lagrangian (or Fenchel) duality to derive the dual problem of (D):

$$\begin{array}{lll} (\widehat{P}) & \min\limits_{X} & \langle Q_0, X\rangle \\ & \text{s.t.} & \langle Q_j, X\rangle = b_j \quad \text{for } j = 1, \ldots, m \\ & & X \succeq 0. \end{array}$$

3. When there exists $y \in \mathbf{R}^m$ such that the matrix $Q_0 - \sum_{j=1}^m y_j Q_j$ is *positive definite*, the optimal values of $(\widehat{P})$ and (D) are equal. What is then the relationship between the optimal value of (P) and that of $(\widehat{P})$? The problem $(\widehat{P})$ is called a *convex relaxation* of (P). Why?

4.2.3 Minimax duality

We motivated both Fenchel-Rockafellar and Lagrangian duality from the viewpoint of a lower-bounding procedure. In this section, we discuss an alternative viewpoint that is applicable in many other contexts. Namely,

suppose that we are given a function $\Phi\colon \mathcal{X} \times \mathcal{Y} \to \mathbf{R}$ of two variables $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are some sets. One may then define the *primal* and the *dual problems*, respectively:

$$\begin{array}{lll} (P) & \inf\limits_x \ \varphi(x) & \text{where} \quad \varphi(x) = \sup\limits_y \ \Phi(x,y) \\ (D) & \sup\limits_y \ \psi(y) & \text{where} \quad \psi(y) = \inf\limits_x \ \Phi(x,y) \end{array}. \tag{4.19}$$

Exercise 3.25 guarantees the weak duality inequality:

$$\operatorname{val}(P) = \inf_x \sup_y \Phi(x,y) \geq \sup_y \inf_x \Phi(x,y) = \operatorname{val}(D). \tag{4.20}$$

The question of when equality $\operatorname{val}(P) = \operatorname{val}(D)$ holds amounts to determining conditions on Φ, which ensure that we may freely interchange $\inf_x$ and $\sup_y$. The following exercise shows that both Fenchel-Rockafellar and Lagrangian duality can be understood from this perspective.

Exercise 4.16. Show that the following are true.

1. Let $h\colon \mathbf{Y} \to \overline{\mathbf{R}}$ and $g\colon \mathbf{E} \to \overline{\mathbf{R}}$ be proper, closed convex functions, and let $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$ be a linear map. Set $\mathcal{X} = \operatorname{dom} g$ and $\mathcal{Y} = \operatorname{dom} h^\star$ and define the function

   $$\Phi(x,y) = g(x) + \langle \mathcal{A}x, y\rangle - h^\star(y).$$

   Show that the problems (P) and (D) in (4.19) are exactly the Fenchel-Rockafellar primal-dual pair of Section 4.2.1.

2. Fix some proper functions $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and $g_i\colon \mathbf{E} \to \mathbf{R}$ for $i = 1, \ldots, m$. Set $\mathcal{X} = \operatorname{dom} f$ and $\mathcal{Y} = \mathbf{R}^k_+ \times \mathbf{R}^{m-k}$ and define the Lagrangian function

   $$\Phi(x,y) = f(x) + \sum_{i=1}^m y_i g_i(x).$$

   Show that the problems (P) and (D) in (4.19) are exactly the Lagrangian primal-dual pair of Section 4.2.2.

In this section, we aim to find conditions that ensure primal and dual attainment along with the equality $\operatorname{val}(P) = \operatorname{val}(D)$. To this end, we will need the following definition.

Definition 4.17. A pair $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is a saddle point of Φ if it satisfies

$$\Phi(x^*, y) \leq \Phi(x^*, y^*) \leq \Phi(x, y^*) \qquad \forall x \in \mathcal{X},\ y \in \mathcal{Y}.$$

Equivalently (x^*, y^*) is a saddle point of Φ if simultaneously x^* minimizes $\Phi(\cdot, y^*)$ and y^* maximizes $\Phi(x^*, \cdot)$. Using the functions φ and ψ, this amounts to the equation

$$\varphi(x^*) = \Phi(x^*, y^*) = \psi(y^*).$$

Existence of saddle points completely characterizes primal-dual attainment and zero duality gap.

Lemma 4.18 (Saddle-points and duality)**.** *The following are equivalent.*

1. $x^* \in \operatorname{argmin} \varphi$, $y^* \in \operatorname{argmax} \psi$, *and* $\varphi(x^*) = \psi(y^*)$,
2. (x^*, y^*) *is a saddle-point of* Φ.

Proof. Suppose (1) holds. We then deduce

$$\varphi(x^*) = \psi(y^*) = \inf_x \Phi(x, y^*) \leq \Phi(x^*, y^*),$$
$$\psi(y^*) = \varphi(x^*) = \sup_y \Phi(x^*, y) \geq \Phi(x^*, y^*).$$

Thus (2) holds as claimed. Conversely, suppose (2) holds. Then we compute

$$\inf_x \sup_y \Phi(x, y) \leq \sup_y \Phi(x^*, y) = \Phi(x^*, y^*) = \inf_x \Phi(x, y^*) \leq \sup_y \inf_x \Phi(x, y).$$

Taking into account the weak duality inequality (4.20), we deduce that equality holds throughout. The condition (1) follows immediately. □

We are now ready to establish conditions on Φ that ensure equality $\operatorname{val}(P) = \operatorname{val}(D)$. These conditions will be stated in terms "marginal functions" $p_y \colon \mathbf{E} \to \overline{\mathbf{R}}$ and $q_x \colon \mathbf{Y} \to \overline{\mathbf{R}}$, defined by

$$p_y(x) = \left\{ \begin{array}{ll} \Phi(x, y) & \text{if } x \in \mathcal{X} \\ +\infty & \text{otherwise} \end{array} \right\} \quad \text{and} \quad q_x(y) = \left\{ \begin{array}{ll} -\Phi(x, y) & \text{if } y \in \mathcal{Y} \\ +\infty & \text{otherwise} \end{array} \right\}.$$

The proof of the following theorem, not surprisingly, is based yet again on recognizing (P) and (D) as a parametric primal-dual pair for the perturbation function: Define the function $p \colon \mathbf{Y} \to \overline{\mathbf{R}}$ by

$$p(z) := \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \{\Phi(x, y) + \langle z, y \rangle\}. \tag{4.21}$$

The reader should convince themselves that when specialized to the Fenchel-Rockafellar and Lagrangian frameworks, this function reduces precisely to the perturbation function we used to establish strong duality in Theorems 4.4 and Theorem 4.14.

Theorem 4.19 (Convex minimax)**.** *Consider a function* $\Phi\colon \mathcal{X} \times \mathcal{Y} \to \mathbf{R}$ *defined on some sets* $\mathcal{X} \subset \mathbf{E}$ *and* $\mathcal{Y} \subset \mathbf{Y}$*. Assume the following.*

1. **(convex-concave)** *The functions* p_y *and* q_x *are proper, closed, and convex for all* $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$*.*
2. **(coercivity)** *For some point* $\bar{y} \in \mathcal{Y}$*, the function* $p_{\bar{y}}$ *is coercive.*

Then equality holds:

$$\inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} \Phi(x,y) = \sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} \Phi(x,y),$$

whenever the left-side is finite.

Proof. Define the value function $p\colon \mathbf{Y} \to \overline{\mathbf{R}}$ by (4.21). In particular, equality $\operatorname{val}(P) = p(0)$ holds. Our immediate goal is to show the equality $\operatorname{val}(D) = p^{\star\star}(0)$. To this end, we may express p as the infimal projection

$$p(z) = \inf_{x\in\mathbf{E}} F(x,z), \qquad \text{where} \qquad F(x,z) := \sup_{y\in\mathcal{Y}} \{p_y(x) + \langle z, y\rangle\}.$$

Observe that F is closed and convex, since it is the pointwise supremum of closed convex functions. In particular, we conclude that p is a convex function. Recall from (4.4) the expression for the double conjugate $p^{\star\star}(0) = \sup_y -F^\star(0,y)$. Using the definition of the Fenchel conjugate, we successively compute

$$\begin{aligned} F^\star(0,y) &= \sup_{(x,z)} \left\{ \langle (x,z),(0,y)\rangle - \sup_{u\in\mathcal{Y}} \{\langle z,u\rangle + p_u(x)\} \right\} \\ &= \sup_{x\in\mathcal{X}} \sup_{z\in\mathbf{Y}} \left\{ \langle z,y\rangle - \sup_{u\in\mathbf{Y}} \{\langle z,u\rangle - q_x(u)\} \right\} \\ &= \sup_{x\in\mathcal{X}} \sup_{z\in\mathbf{Y}} \{\langle z,y\rangle - q_x^\star(z)\} \\ &= \sup_{x\in\mathcal{X}} q_x^{\star\star}(y). \end{aligned}$$

Taking into account that q_x is proper, closed, and convex for every $x \in \mathcal{X}$, we conclude

$$p^{\star\star}(0) = \sup_{y\in\mathbf{Y}} -F^\star(0,y) = \sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} \Phi(x,y) = \operatorname{val}(D),$$

as we set out to show. Suppose now that $p(0)$ is finite, and fix a point $\bar{y} \in \mathcal{Y}$ such that $p_{\bar{y}}$ is coercive. We will show that p is lower-semicontinuous at 0,

from which equality $p^{\star\star}(0) = p(0)$ follows immediately. To this end, consider a sequence $z_i \to 0$. Passing to a subsequence, we may assume that the values $p(z_i)$ converge in $[-\infty, +\infty)$. Choose now points $x_i \in \mathcal{X}$ satisfying

$$F(x_i, z_i) \leq p(z_i) + i^{-1}. \tag{4.22}$$

We claim that the sequence $\{x_i\}$ is bounded. Indeed, observe

$$p(z_i) + i^{-1} \geq F(x_i, z_i) \geq p_{\bar{y}}(x_i) + \langle z_i, y\rangle \geq p_{\bar{y}}(x_i) - \|z_i\| \cdot \|\bar{y}\|.$$

We therefore deduce $\limsup_{i\to\infty} p_{\bar{y}}(x_i) < \infty$ Since $p_{\bar{y}}$ is coercive, we conclude that the sequence $\{x_i\}$ is bounded. Therefore passing to a subsequence, we may be sure that x_i converges to some point $\bar{x} \in \mathbf{E}$. Taking into account that F is closed, we conclude

$$p(0) = \inf_{x\in\mathbf{E}} F(x, 0) \leq F(\bar{x}, 0) \leq \liminf_{i\to\infty} F(x_i, z_i) \leq \liminf_{i\to\infty} p(z_i),$$

where the last inequality follows from (4.22). Thus p is lower-semicontinuous at $z = 0$, as we had to show. The proof is complete. □

An immediate and useful consequence is the Kakutani minimax theorem.

Exercise 4.20 (Convex compact minimax)**.** Consider a continuous function $\Phi\colon \mathbf{E}\times\mathbf{Y} \to \mathbf{R}$, and let $\mathcal{X} \subset \mathbf{E}$ and $\mathcal{Y} \subset \mathbf{Y}$ be two nonempty compact convex sets. Suppose that $\Phi(\cdot, y)$ is convex for all $y \in \mathcal{Y}$ and $\Phi(x, \cdot)$ is concave for all $x \in \mathcal{X}$. Show that Φ has a saddle point on $\mathcal{X} \times \mathcal{Y}$.

Exercise 4.21 (Direction of steepest descent)**.** Consider a convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point $x \in \operatorname{int}(\operatorname{dom} f)$. Verify the equality

$$\min_{v:\|v\|\leq 1} f'(x, v) = -\min_{v\in\partial f(x)} \|v\|.$$

Define the vector $\bar{v} \in \operatorname{argmin}_{v\in\partial f(x)} \|v\|$ and assume $\bar{v} \neq 0$. Show that there exists $\epsilon, \alpha > 0$ such that

$$f(x) - f(x - t\bar{v}) \geq \alpha t\|\bar{v}\| \qquad \text{for all } t \in (0, \epsilon).$$

[**Hint:** Use Theorem 3.50 and Exercise 4.20.]

4.3 Spectral functions

The previous sections focused on two closely related ideas of duality and subdifferential calculus for convex functions. The goal of this section is to derive an expression for the subdifferential of a special class of functions (called *spectral*) on the space of symmetric matrices. These are the functions that depend on the matrix only through its eigenvalues. Remarkably, we will be able to explicitly compute Fenchel conjugates, subdifferentials, and proximal maps of such functions. The expression for the subdifferential, in particular, can be regarded as a special type of chain rule.

We begin with some notation. The symbol $\mathbf{S}^n$, as before, will denote the Euclidean space of symmetric matrices, while $O(n)$ will denote the set of $n \times n$ orthogonal matrices. The symbol $\Pi(n)$ will denote set of all $n \times n$ permutation matrices. In particular, for any vector $x \in \mathbf{R}^n$ and permutation $\pi \in \Pi(n)$, the product πx is the same vector as x but with coordinates permuted according to π. We next record the following two key definitions, whose close relationship will be become clear shortly.

Definition 4.22. A function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ is called *symmetric* if it satisfies

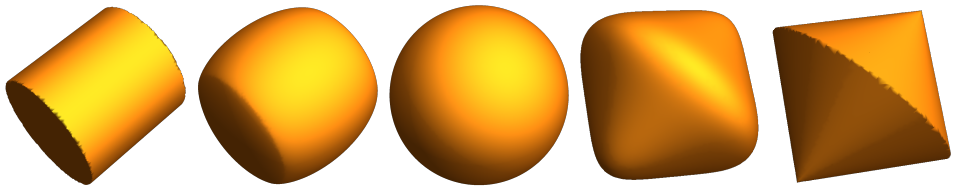
$$f(\pi x) = f(x), \qquad \text{for all } x \in \mathbf{R}^n, \pi \in \Pi(n).$$

Definition 4.23. A function $F\colon \mathbf{S}^n \to \overline{\mathbf{R}}$ is called *spectral* if it satisfies

$$F(UXU^T) = F(X), \qquad \text{for all } X \in \mathbf{S}^n, U \in O(n).$$

Thus a function f on $\mathbf{R}^n$ is symmetric if its value does not depend on the ordering of the coordinates of its argument. For example, all ℓ_p-norms are symmetric. A function F on $\mathbf{S}^n$ is spectral if it is invariant under conjugation of its argument by orthogonal matrices. For example, any function F on $\mathbf{S}^n$ that factors as $F = f(\lambda(X))$ is spectral, since the eigenvalues are invariant under conjugation by orthogonal matrices. Thus, the negative log-determinant $F(X) = -\ln\det(X)$ on $\mathbf{S}^n_{++}$ (see Exercise 1.12) and all Schatten ℓ_p-norms $\|X\|_p := \|\lambda(X)\|_p$ are spectral functions. The latter function class in particular includes the nuclear norm $F(X) = \sum_{i=1}^n |\lambda_i(X)|$, the operator norm $F(X) = \max_{i=1,\ldots,n} |\lambda_i(X)|$, and the Frobenius norm $F(X) = \sqrt{\sum_{i=1}^n \lambda_i^2(X)}$. See Figure 4.2 for an illustration of the unit balls of these norms.

All of the examples of spectral functions we have seen so far can be factored as $F(X) = f(\lambda(X))$ for some simple symmetric function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$. Indeed, all spectral functions factor in this way.



(a) $p = 1$ (b) $p = 1.5$ (c) $p = 2$ (d) $p = 5$ (e) $p = \infty$

Figure 4.2: Unit balls of Schatten ℓ_p-norms $\|X\|_p = \|\lambda(X)\|_p$ over $\mathbf{S}^2$.

Exercise 4.24. A function $F\colon \mathbf{S}^n \to \overline{\mathbf{R}}$ is spectral if and only if one can write $F = f \circ \lambda$ for some symmetric function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$.
[**Hint:** Explicitly, set $f(x) = F(\operatorname{Diag}(x))$.]

In typical circumstances, the spectral function F may appear highly complicated, whereas f is simple (e.g. polyhedral). For example, it is not clear at all that the Schatten ℓ_p-norms are indeed norms on $\mathbf{S}^n$, or that they are even convex functions. In this section, we will see that numerous analytic properties of F can be described purely in terms of the analogous properties of f. Before delving into the details, we will need the following two results. For a vector x in $\mathbf{R}^n$, let $x^\uparrow$ denote the vector with the same components permuted in nonincreasing order.

Exercise 4.25. ✍ For any two vectors $x, y \in \mathbf{R}^n$, the inequality holds:

$$\langle x, y\rangle \le \langle x^\uparrow, y^\uparrow\rangle.$$

Moreover, equality holds if and only if there exists a permutation π satisfying $\pi(x) = x^\uparrow$ and $\pi(y) = y^\uparrow$.

Exercise 4.26. Consider a convex symmetric function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$. Then whenever the inclusion $v \in \partial f(x^\uparrow)$ holds, so does the inclusion $v^\uparrow \in \partial f(x^\uparrow)$.
[**Hint:** Use the Fenchel-Young inequality together with Exercise 4.25.]

The key tool we will use to build a parallel between analytic properties of f and those of $F = f \circ \lambda$ is Theorem 4.27. The result shows that if two symmetric matrices X and Y are well-aligned, then so are their vectors of eigenvalues $\lambda(X)$ and $\lambda(Y)$. The proof will use the following basic linear algebraic fact: two symmetric matrices $X, Y \in \mathbf{S}^n$ commute if and only if they can be simultaneously diagonalized, meaning that there exists $U \in O(n)$ such that both UXU^T and UYU^T are diagonal matrices.

Theorem 4.27 (Trace Inequality)**.** *All matrices $X, Y \in \mathbf{S}^n$ satisfy the inequality:*

$$\langle \lambda(X), \lambda(Y) \rangle \geq \langle X, Y \rangle, \tag{4.23}$$

with equality if and only if X and Y admit a simultaneous ordered spectral decomposition, meaning that there exists $U \in O(n)$ satisfying

$$U^T X U = \operatorname{Diag}(\lambda(X)), \qquad U^T Y U = \operatorname{Diag}(\lambda(Y)). \tag{4.24}$$

The reader should verify that the estimate in (4.23) can be equivalently written as

$$\|\lambda(X) - \lambda(Y)\|_2 \leq \|X - Y\|_F.$$

Therefore, Theorem 4.27 shows that the eigenvalue map is 1-Lipschitz continous and characterizes those pairs of matrices for which the Lipschitz bound is tight. We postpone the proof of Theorem 4.27 to Section 4.3.3, so as to not distract from the main theme of the section.

4.3.1 Fenchel conjugate and the Moreau envelope

Armed with Theorem 4.27, we can now prove a precise relationship between the Fenchel conjugate and the Moreau envelope of a spectral function $f \circ \lambda$ and those of f.

Theorem 4.28. *Consider a symmetric function $f \colon \mathbf{R}^n \to \overline{\mathbf{R}}$. Then the inequalities*

$$(f \circ \lambda)^\star(Y) = (f^\star \circ \lambda)(Y) \qquad \textit{and} \qquad (f \circ \lambda)_\alpha(X) = f_\alpha(\lambda(X)),$$

hold for all $X, Y \in \mathbf{S}^n$.

Proof. We begin with the computation of the Fenchel conjugate. To this end, we successively compute

$$\begin{aligned}(f \circ \lambda)^\star(Y) &= \sup_{X \in \mathbf{S}^n} \{\langle X, Y \rangle - f(\lambda(X))\} \\ &\leq \sup_{X \in \mathbf{S}^n} \{\langle \lambda(X), \lambda(Y) \rangle - f(\lambda(X))\} \\ &\leq \sup_{z \in \mathbf{R}^n} \{\langle z, \lambda(Y) \rangle - f(z)\} = f^\star(\lambda(Y)),\end{aligned} \tag{4.25}$$

where (4.25) follows from Theorem 4.27. To see the reverse inequality, fix an eigenvalue decomposition $Y = U\mathrm{Diag}\,(\lambda(Y))U^T$ with $U \in O(n)$. Observe

$$\begin{aligned} f^\star(\lambda(Y)) &= \sup_{z\in\mathbf{R}^n}\{\langle z, \lambda(Y)\rangle - f(z)\} \\ &\le \sup_{X\in\mathbf{S}^n}\{\langle \lambda(X), \lambda(Y)\rangle - f(\lambda(X))\} \qquad (4.26) \\ &= \sup_{X\in\mathbf{S}^n}\{\langle U\mathrm{Diag}\,(\lambda(X))U^T, Y\rangle - f(\lambda(X))\} \\ &\le \sup_{Z\in\mathbf{S}^n}\{\langle Z, Y\rangle - f(\lambda(Z))\} = (f\circ\lambda)^\star(Y), \end{aligned}$$

where (4.26) uses symmetry of f and Exercise 4.25. We conclude $(f \circ \lambda)^\star(X) = (f^\star \circ \lambda)(X)$, as claimed.

Next, we verify the analogous expression for the Moreau envelope. Fix a matrix X and an eigenvalue decomposition $X = U\mathrm{Diag}\,(\lambda(X))U^T$ with $U \in O(n)$. We then deduce

$$\begin{aligned} (f\circ\lambda)_\alpha(X) &= \inf_{Z\in\mathbf{S}^n}\{f(\lambda(Z)) + \frac{1}{2\alpha}\|Z - X\|_F^2\} \\ &\le \inf_{z\in\mathbf{R}^n}\{f(z) + \frac{1}{2\alpha}\|U\mathrm{Diag}\,(z)U^T - U\mathrm{Diag}\,(\lambda(X))U^T\|_F^2\} \\ &\le \inf_{z\in\mathbf{R}^n}\{f(z) + \frac{1}{2\alpha}\|z - \lambda(X)\|_2^2\} = f_\alpha(\lambda(X)). \end{aligned}$$

Conversely, observe

$$\begin{aligned} f_\alpha(\lambda(X)) &= \inf_{z\in\mathbf{R}^n}\{f(z) + \frac{1}{2\alpha}\|z - \lambda(X)\|_2^2\} \\ &\le \inf_{Z\in\mathbf{S}^n}\{f(\lambda(Z)) + \frac{1}{2\alpha}\|\lambda(Z) - \lambda(X)\|_2^2\} \\ &\le \inf_{Z\in\mathbf{S}^n}\{f(\lambda(Z)) + \frac{1}{2\alpha}\|Z - X\|_F^2\} = (f\circ\lambda)_\alpha(X) \qquad (4.27) \end{aligned}$$

where (4.27) follows from Theorem 4.27. This completes the proof. □

Theorem 4.28 has a number of important corollaries. The first is that a closed symmetric function f is convex if and $f \circ \lambda$ is convex.

Corollary 4.29 (Convexity of spectral functions)**.** *Let $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ be a proper, closed, symmetric function. Then f is convex if and only if $f \circ \lambda$ is convex.*

Proof. If $F := f \circ \lambda$ is convex. then writing $f(x) = F(\mathrm{Diag}\,(x))$, we deduce that f is convex. Conversely, suppose that f is convex. Is it straightforward to check that the Fenchel conjugate of a symmetric function is itself symmetric. Therefore, applying Theorem 4.28 twice yields the equality

$$(f \circ \lambda)^{\star\star} = f^{\star\star} \circ \lambda = f \circ \lambda.$$

Since $f \circ \lambda$ coincides with its double conjugate, it must be convex. □

4.3.2 Proximal map and the subdifferential

With Theorem 4.28 at hand, we can now obtain an explicit expression for the proximal map of $f \circ \lambda$ in terms of the proximal map of f.

Theorem 4.30 (Proximal map)**.** *Consider a symmetric function $f\colon \mathbf{R}^n \to \mathbf{R}$ and an arbitrary matrix $X \in \mathbf{S}^n$. Then the expression holds:*

$$\mathrm{prox}_{f\circ\lambda}(X) = \left\{U\mathrm{Diag}\,(w)U^T : w \in \mathrm{prox}_f(\lambda(X)),\ U \in O_X\right\},$$

where

$$O_X := \{U \in O(n) : X = U\mathrm{Diag}\,(\lambda(X))U^T\}.$$

Proof. Fix a matrix $U \in O_X$ and a vector $w \in \mathrm{prox}_f(\lambda(X))$. Using Theorem 4.28 we compute

$$\begin{aligned}(f \circ \lambda)_\alpha(X) &= f_\alpha(\lambda(X)) \\ &= f(w) + \frac{1}{2\alpha}\|\lambda(X) - w\|_2^2 \\ &= f(U\mathrm{Diag}\,(w)U^T) + \frac{1}{2\alpha}\|X - U\mathrm{Diag}\,(w)U^T\|_F^2.\end{aligned}$$

Thus $U\mathrm{Diag}\,(w)U^T$ lies in $\mathrm{prox}_{f\circ\lambda}(X)$. To see the reverse inclusion, fix a matrix $W \in \mathrm{prox}_{f\circ\lambda}(X)$. Using Theorem 4.28 we compute

$$\begin{aligned}f_\alpha(\lambda(X)) &= (f \circ \lambda)_\alpha(X) \\ &= f(\lambda(W)) + \frac{1}{2\alpha}\|X - W\|_F^2 \\ &\geq f(\lambda(W)) + \frac{1}{2\alpha}\|\lambda(X) - \lambda(W)\|_F^2 \qquad (4.28) \\ &\geq f_\alpha(\lambda(X)) \qquad (4.29)\end{aligned}$$

where (4.28) follows from Theorem 4.27. Thus equality holds throughout. In particular, we deduce

$$\|X - W\|_F^2 = \|\lambda(X) - \lambda(W)\|_F^2 \qquad \text{and} \qquad \lambda(W) \in \mathrm{prox}_f(\lambda(X)).$$

Theorem 4.27 therefore guarantees that there exists a matrix $U \in O_X$ satisfying $W = U^T \mathrm{Diag}\,(\lambda(W))U$. The proof is complete. □

Thus the recipe for computing the proximal map of a spectral function $F = f \circ \lambda$ at a matrix X is simple: form an eigenvalue decomposition $X = U\mathrm{Diag}\,(\lambda(X))U^T$, compute any vector $w \in \mathrm{prox}_f(\lambda(X))$, and form the matrix $U\mathrm{Diag}\,(w)U^T$.

Exercise 4.31. Consider a proper, closed, convex, symmetric function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$. Show that the set $\mathrm{prox}_{f\circ\lambda}(X)$ consists of all matrices $W \in \mathbf{S}^n$ that admit a simultaneous ordered spectral decomposition with X and satisfy $\lambda(W) \in \mathrm{prox}_f(\lambda(X))$.

Next, we establish an analogous expression for the subdifferential of convex spectral functions.

Theorem 4.32 (Subdifferential of convex spectral functions)**.** *Consider a proper, closed, convex, symmetric function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ and an arbitrary matrix $X \in \mathbf{S}^n$. Then the expression holds:*

$$\partial(f \circ \lambda)(X) = \left\{U\mathrm{Diag}\,(y)U^T : y \in \partial f(\lambda(X)),\ U \in O_X\right\},$$

where

$$O_X := \{U \in O(n) : X = U\mathrm{Diag}\,(\lambda(X))U^T\}.$$

Proof. Fix a matrix $Y \in \partial(f \circ \lambda)(X)$. Then Theorem 3.39 guarantees the equality

$$(f \circ \lambda)^\star(Y) + (f \circ \lambda)(X) = \langle Y, X\rangle.$$

We therefore deduce

$$\begin{aligned}
\langle \lambda(Y), \lambda(X)\rangle &\leq f^\star(\lambda(Y)) + f(\lambda(X) && (4.30)\\
&= (f \circ \lambda)^\star(Y) + (f \circ \lambda)(X) && (4.31)\\
&= \langle Y, X\rangle \\
&\leq \langle \lambda(Y), \lambda(X)\rangle. && (4.32)
\end{aligned}$$

where (4.31) follows from Theorem 4.28 and (4.32) follows from Theorem 4.27. Hence equality holds throughout. Theorem 4.27 therefore guarantees Therefore combining equality in (4.32) with Theorem 4.27, we deduce that the matrices X and Y admit a simultaneously ordered decomposition. Equality in (4.30), in turn, guarantees the inclusion $\lambda(Y) \in \partial f(\lambda(X))$.

Conversely, consider a matrix $Y = U\mathrm{Diag}\,(y)U^T$ for some $U \in O_X$ and $y \in \partial f(\lambda(X))$. Then by similar reasoning as before, we have

$$\begin{aligned}\langle y, \lambda(X)\rangle &= f^\star(y) + f(\lambda(X)) \\ &= (f\circ\lambda)^\star(Y) + (f\circ\lambda)(X) \\ &\geq \langle Y, X\rangle \qquad (4.33)\\ &= \langle \lambda(Y), \lambda(X)\rangle \\ &\geq \langle y, \lambda(X)\rangle, \qquad (4.34)\end{aligned}$$

where (4.34) follows from Exercise 4.25. Thus equality holds throughout. From equality in (4.33) we conclude $Y \in \partial(f\circ\lambda)(X)$, as claimed. □

Thus the recipe for computing the subdifferential of a convex spectral function $F = f\circ\lambda$ at a matrix X is simple: form an eigenvalue decomposition $X = U\mathrm{Diag}\,(\lambda(X))U^T$, compute any subgradient $y \in \partial f(\lambda(X))$, and form the matrix $U\mathrm{Diag}\,(y)U^T$.

Exercise 4.33. Consider a proper, closed, convex, symmetric function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$. Show that the subdifferential $\partial(f\circ\lambda)(X)$ consists of all matrices $Y \in \mathbf{S}^n$ that admit a simultaneous ordered spectral decomposition with X and satisfy $\lambda(Y) \in \partial f(\lambda(X))$.

Exercise 4.34. Define the function $F\colon \mathbf{S}^n \to \mathbf{R}$ by $F(X) = \|X\|_{\mathrm{op}}$. Prove the expression

$$\partial(f\circ\lambda)(I) = \{Y \succeq 0 : \mathrm{tr}(Y) = 1\}.$$

We end the section with the promised proof of the trace inequality.

4.3.3 Proof of the trace inequality

Proof of Theorem 4.27. Fix two matrices $X, Y \in \mathbf{S}^n$ and define the set

$$\mathcal{L} = \{UXU^T : U \in O(n)\}.$$

Since $\mathcal{L}$ is compact (why?), the linear function $Z \mapsto \langle Z, Y\rangle$ attains its maximum over $\mathcal{L}$. Let Z be any maximizer of this linear function over $\mathcal{L}$. Fix now an arbitrary skew-symmetric matrix W meaning $W^T = -W$. Notice that for any $t \in \mathbf{R}$, the matrix exponential $E = e^{tW}$ is an orthogonal matrix, due to the computation

$$E^T = (e^{tW})^T = e^{tW^T} = e^{-tW} = E^{-1}.$$

Define the function $\varphi \in \mathbf{R} \to \mathbf{R}$ by

$$\varphi(t) = \langle e^{tW} Z e^{-tW}, Y \rangle.$$

Using the Taylor expansion of the exponential, it is straightforward to verify

$$\varphi'(0) = \langle YZ - ZY, W \rangle.$$

Since Z is a maximizer of φ, equality $\varphi'(0) = 0$ holds. In particular, setting $W := YZ - ZY$, we deduce $YZ = ZY$. Since Z and Y commute, they must be simultaneously diagonalizable. Thus there exist $U \in O(n)$ and a permutation π satisfying

$$Y = U\mathrm{Diag}\,(\pi \cdot \lambda(Y))U^T, \qquad Z = U\mathrm{Diag}\,(\lambda(Z))U^T. \tag{4.35}$$

Moreover, since Z and X both lie in $\mathcal{L}$, equality $\lambda(Z) = \lambda(X)$ holds. Thus

$$\langle X, Y \rangle \leq \langle Z, Y \rangle = \langle \pi \cdot \lambda(Y), \lambda(X) \rangle \leq \langle \lambda(Y), \lambda(X) \rangle, \tag{4.36}$$

where the last inequality follows from Exercise 4.25. This establishes the inequality (4.23).

Suppose now that equality holds in (4.23). Then we can set $Z = X$ in the first place. Thus Y and X are simultaneously diagonalizable and the estimate (4.36) guarantees the equality

$$\langle \pi \cdot \lambda(Y), \lambda(X) \rangle = \langle \lambda(Y), \lambda(X) \rangle.$$

Applying Exercise 4.25, we deduce that there exists a permutation $\hat{\pi}$ satisfying $\lambda(X) = \hat{\pi}\lambda(X)$ and $\pi\lambda(Y) = \hat{\pi}\lambda(Y)$. Consequently, using (4.35) we deduce

$$Y = U\mathrm{Diag}\,(\hat{\pi}\lambda(Y))U^T, \qquad X = U\mathrm{Diag}\,(\hat{\pi}\lambda(X))U^T.$$

Permuting the columns of U according to $\hat{\pi}^{-1}$ yields the factorization (4.24), thereby completing the proof. □

4.3.4 Orthogonally invariant functions of rectangular matrices

Thus far, this section has focused on orthogonally invariant functions on the space of symmetrice matrices $\mathbf{S}^n$. It is possible to develop a completely parallel theory for orthogonally invariant functions on the Euclidean space

of rectangular matrices $\mathbf{R}^{m\times n}$. We now outline such results, leaving the details as exercises.

We begin with the following two definitions that parallel those in the symmetric setting. Throughout, we let $\Pi_{\pm}(m)$ denote the set of $m \times m$. Thus any matrix $\pi \in \Pi_{\pm}(m)$ has exactly one entry ± 1 in each row and each column.

Definition 4.35. A function $f\colon \mathbf{R}^m \to \overline{\mathbf{R}}$ is called *absolutely symmetric* if it satisfies

$$f(\pi x) = f(x), \qquad \text{for all } x \in \mathbf{R}^n,\ \pi \in \Pi_{\pm}(m).$$

Definition 4.36. A function $F\colon \mathbf{S}^n \to \overline{\mathbf{R}}$ is called *orthogonally invariant* if it satisfies

$$F(UXV^T) = F(X), \qquad \text{for all } X \in \mathbf{R}^{m\times n},\ U \in O(m),\ V \in O(n).$$

Without loss of generality, suppose $m \leq n$. It is straightforward to see that a function F is orthogonally invariant if and only if it factors as $F = f \circ \sigma$, where f is some absolutely symmetric function on $\mathbf{R}^m$ and σ is the singular value map. The trace inequality then takes the following form. The proof follows along similar lines as its symmetric counterpart (Theorem 4.27), and is therefore omitted.

Theorem 4.37 (Trace inequality for rectangular matrices)**.**
All matrices $X, Y \in \mathbf{R}^{m\times n}$ satisfy the inequality:

$$\langle \sigma(X), \sigma(Y)\rangle \geq \langle X, Y\rangle,$$

with equality if and only if X and Y admit a simultaneous ordered singular value decomposition, meaning that there exist $U \in O(m)$ and $V \in O(n)$ satisfying

$$U^T X V = \operatorname{Diag}(\sigma(X)), \qquad U^T Y V = \operatorname{Diag}(\sigma(Y)).$$

The following result is the direct extension of Theorem 4.28, 4.30, and 4.32 to the rectangular setting, with an identical proof.

Theorem 4.38. *Consider an absolutely symmetric, proper, closed function $f\colon \mathbf{R}^m \to \overline{\mathbf{R}}$. Then the expressions hold:*

$$(f \circ \sigma)^\star = f^\star \circ \sigma \qquad \text{and} \qquad (f \circ \sigma)_\alpha = f_\alpha \circ \sigma,$$

and

$$\operatorname{prox}_{f\circ\sigma}(X) = \left\{U\operatorname{Diag}(w)V^T : w \in \operatorname{prox}_f(\sigma(X)),\ (U,V) \in O_X\right\},$$

where we define

$$O_X := \{(U,V) \in O(m) \times O(n) : X = U\operatorname{Diag}(\sigma(X))V^T\}.$$

Moreover, $f \circ \sigma$ is convex if and only if f is convex, in which case the subdifferential admits the form

$$\partial(f\circ\sigma)(X) = \left\{U\operatorname{Diag}(y)V^T : y \in \partial f(\sigma(X)),\ (U,V) \in O_X\right\}.$$

References.

The material in Section 4.1 follows the discussion in [33, Section 11.H]. The proof the subdifferential sum and chain rules (Theorem 4.5) using strong duality appears in [9]. The discussion of the Lagrangian duality is similar to that in [33, Section 11.H]. The material in Section 4.3 can be found in the manuscripts [12, 20–22].

Chapter 5

First-order algorithms for black-box convex optimization

This chapter introduces a number of foundational iterative methods for minimizing a convex function f on $\mathbf{E}$. The algorithms we consider will only use first-order information: gradients when f is smooth and subgradients when f is nonsmooth. The efficiency of any such first-order algorithm is measured by the number of (sub)gradient evaluations required to produce a point x_ϵ satisfying $f(x_\epsilon) - \min f \leq \epsilon$. In particular, we will only be interested in efficiency guarantees that are independent of the dimension of the ambinet space $\mathbf{E}$, which the reader should assume is huge. We focus on three algorithms: gradient descent, accelerated gradient descent, and the subgradient method. We will see that the accelerated gradient method has the best possible efficiency guarantees among any first-order method for minimizing smooth convex functions. Similarly, the subgradient method is best in class for minimizing Lipschitz convex functions. The final section of the chapter provides a more modern view of the three methods based on two-sided and one-sided model approximation. This viewpoint will lead to important extensions of these methods to a wider class of problems in Chapter **??**.

Roadmap. We begin with Section 5.1, which introduces gradient descent and the accelerated gradient method for minimizing β-smooth convex functions. Section 5.2 introduces the subgradient method for minimizing L-Lipschitz convex functions. The final Section 5.3 outlines an illuminating viewpoint of (accelerated) gradient and subgradient methods based on model

approximation.

5.1 Algorithms for smooth convex minimization

Throughout this section, we consider the optimization problem

$$\min_{x\in\mathbf{E}} f(x), \tag{5.1}$$

where $f\colon \mathbf{E} \to \mathbf{R}$ is a β-smooth and α-strongly convex function. Here, $\alpha, \beta \geq 0$ are nonnegative real numbers. In particular, the setting $\alpha = 0$ simply corresponds to convexity. Thus, as a consequence of Corollary 1.15 and Theorem 3.57, the function f satisfies the two sided bound

$$\frac{\alpha}{2}\|y-x\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y-x\rangle \leq \frac{\beta}{2}\|y-x\|^2, \tag{5.2}$$

for all $x, y \in \mathbf{E}$. Geometrically, this estimate means that f is sandwiched between the two quadratics

$$\begin{aligned} Q_x(y) &:= f(x) + \langle \nabla f(x), y-x\rangle + \frac{\beta}{2}\|y-x\|^2, \\ q_x(y) &:= f(x) + \langle \nabla f(x), y-x\rangle + \frac{\alpha}{2}\|y-x\|^2. \end{aligned}$$

See Figure 5.1 for a geometric illustration. The ratio $\kappa := \beta/\alpha$ is called the *condition number* of f. Intuitively, κ measures the "scaling" of the problem. In particular, equality $\kappa = 1$ holds if and only if f is a spherical quadratic. We will see that the condition number κ strongly influences convergence guarantees of numerical methods.

Throughout, we assume that f has at least one minimizer, which is automatic if $\alpha > 0$ (why?). The symbols f^* and x^* will denote the minimal value of f and an arbitrary minimizer of f, respectively.

5.1.1 Gradient descent

Gradient descent is the most basic numerical method for the problem (5.1). In each iteration, the method simply takes a step in the direction of the negative gradient:

$$x_{t+1} = x_t - \eta \nabla f(x_t),$$

where the parameter $\eta > 0$ is to be determined. The classical motivation for moving along the negative gradient direction is that this is the direction

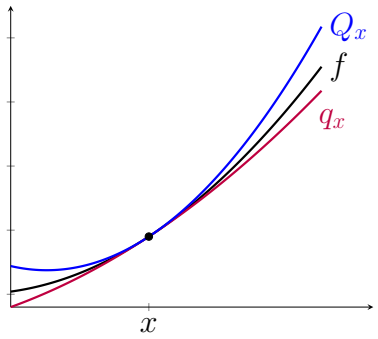


Figure 5.1: Illustration of a β-smooth and α-strongly convex function f, where $Q_x(y) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2$ is an upper estimator based at x and $q_x(y) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2$ is a lower estimator based at x.

of maximal instantaneous decrease

$$\operatorname*{argmin}_{v:\, \|v\| \leq 1} f'(x, v) = -\frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

In order to determine an appropriate parameter $\eta > 0$, let us estimate the functional decrease achieved by a single gradient step.

Lemma 5.1 (Descent)**.** *The gradient step $x^+ = x - \eta \nabla f(x_t)$ satisfies*

$$f(x^+) \leq f(x) - \eta \left(1 - \frac{\eta\beta}{2}\right) \|\nabla f(x)\|^2.$$

Proof. Using the right-hand-side of (5.2), we compute

$$\begin{aligned} f(x - \eta \nabla f(x)) &\leq f(x) - \langle \nabla f(x), \eta \nabla f(x) \rangle + \frac{\beta}{2} \|\eta \nabla f(x)\|^2 \\ &= f(x) - \eta \left(1 - \frac{\eta\beta}{2}\right) \|\nabla f(x)\|^2, \end{aligned}$$

as claimed. □

Thus the decrease in the function value achieved by a single gradient step is proportional to $\eta \left(1 - \frac{\eta\beta}{2}\right) \|\nabla f(x)\|^2$. Consequently, it is appealing to choose η to be the maximizer of the concave quadratic $\eta \mapsto \eta \left(1 - \frac{\eta\beta}{2}\right)$.

A quick computation yields the choice $\eta = \frac{1}{\beta}$. Algorithm 1 records the resulting procedure. Using this value of η in Lemma 5.1 shows that the iterates generated by Algorithm 1 satisfy the guarantee

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2 \qquad \text{for all } t \geq 0. \tag{5.3}$$

In words, this estimate means that the functional improvement in each step $f(x_t) - f(x_{t+1})$ is at least on the same order as the measure of optimality $\|\nabla f(x_t)\|^2$ at the current iterate. Thus, if x_t is highly suboptimal, we expect that $f(x_{t+1})$ will be siginificantly smaller than $f(x_t)$.

Algorithm 1: Gradient descent

Input: Starting point $x_0 \in \mathbf{E}$, parameter $\beta > 0$, iteration $T \in \mathbb{N}$.

Step $t = 0, 1, \ldots, T-1$:

Set $x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$

The following theorem establishes a sublinear rate of convergence of gradient descent for minimizing smooth convex functions. Notice that the guarantee of the theorem does not depend on the strong convexity constant α, and is valid even when $\alpha = 0$.

Theorem 5.2 (Gradient descent under convexity)**.** *Let $f\colon \mathbf{E} \to \mathbf{R}$ be a convex and β-smooth function. Then the iterates generated by Algorithm 1 satisfy*

$$f(x_t) - f^* \leq \frac{\beta\|x_0 - x^*\|^2}{2t}.$$

Proof. We successively compute

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|^2 && (5.4)\\
&= f(x_t) + \langle \nabla f(x_t), x^* - x_t\rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|^2 + \langle \nabla f(x_t), x_{t+1} - x^*\rangle && (5.5)\\
&\leq f^* + \frac{\beta}{2}\|x_{t+1} - x_t\|^2 + \langle \nabla f(x_t), x_{t+1} - x^*\rangle && (5.6)\\
&= f^* + \frac{\beta}{2}\left(\|x_{t+1} - x_t\|^2 - 2\langle x_{t+1} - x_t, x_{t+1} - x^*\rangle\right) && (5.7)\\
&= f^* + \frac{\beta}{2}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right), && (5.8)
\end{aligned}$$

where (5.4) and (5.7) follow from (5.2), while equalities (5.5), (5.7), and (5.8) follow from algebraic manipulations.

Subtracting f^* from both sides and summing for $i = 0, \dots, t-1$, the terms on the right side telescope, yielding

$$\sum_{i=0}^{t-1}(f(x_{i+1}) - f^*) \leq \frac{\beta}{2}\sum_{i=0}^{t-1}\left(\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2\right) \leq \frac{\beta}{2}\|x_0 - x^*\|^2.$$

Taking into account that that the function values $\{f(x_i)\}_{i\geq 0}$ are nonincreasing (Lemma 5.1), we deduce

$$f(x_t) - f^* \leq \frac{1}{t}\sum_{i=0}^{t-1}(f(x_{i+1}) - f^*) \leq \frac{\beta\|x_0 - x^*\|^2}{2t},$$

as claimed. □

The key part of the proof of Theorem 5.2 is the estimate (5.8). Though this inequality might seem like a lucky coincidence at first, Section **??** will provide a more appealing geometric explanation.

Theorem 5.2 shows that gradient descent, applied to a convex β-smooth function, drives the function gap $f(x_t) - f^*$ to zero at a sublinear rate $1/t$. When f is in addition α-strongly convex with $\alpha > 0$, gradient descent convergences at a linear rate.

Theorem 5.3 (Gradient descent under strong convexity)**.** *Let $f\colon \mathbf{E} \to \mathbf{R}$ be an α-strongly convex and β-smooth function. Then the iterates generated by Algorithm 1 satisfy*

$$f(x_{t+1}) - f^* \leq \left(1 - \frac{1}{2\kappa}\right)(f(x_t) - f^*), \tag{5.9}$$

$$\|x_{t+1} - x^*\|^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)\|x_t - x^*\|^2. \tag{5.10}$$

Proof. To see (5.9), we combine Lemma 5.1 with Theorem 3.68 to deduce

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\beta}\|\nabla f(x_t)\|^2 \leq -\frac{1}{2\kappa}(f(x_t) - f^*).$$

Adding and subtracting f^* from the left-side yields

$$(f(x_{t+1}) - f^*) - (f(x_t) - f^*) \leq -\frac{1}{2\kappa}(f(x_t) - f^*).$$

Rearranging completes the proof of (5.9).

Next, we prove (5.10). To this end, we successively compute

$$\begin{aligned}\|x_{t+1} - x^*\|^2 &= \|(x_t - x^*) - \beta^{-1}\nabla f(x_t)\|^2 \\ &= \|x_t - x^*\|^2 + \frac{2}{\beta}\langle \nabla f(x_t), x^* - x_t\rangle + \frac{1}{\beta^2}\|\nabla f(x_t)\|^2 \\ &\leq \|x_t - x^*\|^2 + \frac{2}{\beta}\Big(f^* - f(x_t) - \frac{\alpha}{2}\|x_t - x^*\|^2\Big) + \frac{1}{\beta^2}\|\nabla f(x_t)\|^2 \qquad (5.11) \\ &= \left(1 - \frac{\alpha}{\beta}\right)\|x_t - x^*\|^2 + \frac{2}{\beta}\Big(f^* - f(x_t) + \frac{1}{2\beta}\|\nabla f(x_t)\|^2\Big), \qquad (5.12)\end{aligned}$$

where (5.11) follows from strong convexity. Lemma 5.1 guarantees that the second term in (5.12) is nonpositive, and therefore the quantity $\|x_{t+1} - x^*\|^2$ tends to zero at the linear rate $1 - \kappa^{-1}$. We can establish a slightly faster rate by a more careful argument. Namely, strong convexity and Lemma 5.1 guarantee

$$f^* + \frac{\alpha}{2}\|x_{t+1} - x^*\|^2 \leq f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2,$$

and therefore

$$f^* - f(x_t) + \frac{1}{2\beta}\|\nabla f(x_t)\|^2 \leq -\frac{\alpha}{2}\|x_{t+1} - x^*\|^2.$$

Combining this estimate with (5.12) and rearranging yields (5.10). □

Thus gradient descent drives both the quantities $\|x_t - x^*\|^2$ and $f(x_t) - f^*$ to zero at a linear rate $1 - \kappa^{-1}$. In particular, in light of Section 1.9, we can be sure that the inequality $f(x_t) - f^* \leq \varepsilon$ holds after $t = \mathcal{O}(\kappa \cdot \ln\left(\frac{f(x_0) - f^*}{\varepsilon}\right))$ iterations. Combining Theorems 5.2 and 5.3, we see that gradient descent satisfies the guarantee:

$$f(x_t) - f^* \leq \min\left\{\frac{1}{2t}, \left(1 - \frac{1}{2\kappa}\right)^t\right\} \cdot \beta\|x_0 - x^*\|^2 \qquad \text{for all } t \geq 0.$$

Thus $f(x_t) - f^*$ is simultaneously bounded by two sequences, one converging sub-linearly and the other converging linearly to zero. Typically, the sublinear rate is observed in the early iterations of the algorithm, while the linear rate is observed towards the end (if at all). See Figure 5.2 for an illustration.

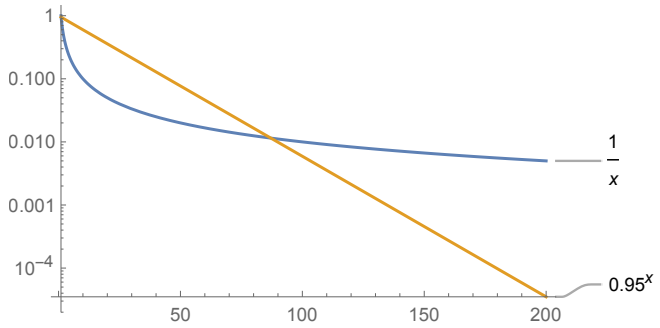


Figure 5.2: Sublinear vs. linear rates

5.1.2 Accelerated gradient descent

The reader may now wonder whether there is a faster algorithm for smooth convex minimization, in terms of the number of gradient evaluations, than gradient descent. In particular, can the sublinear rate $1/t$ be improved? The answer is yes! The accelerated gradient method, outlined in Algorithm 2, achieves the much faster sublinear rate $1/t^2$.

Algorithm 2: Accelerated gradient method

Input: Starting point $x_0 \in \mathbf{E}$.

Set $t = 0$ and $a_0 = a_{-1} = 1$, $x_{-1} = x_0$;

for $t = 0, \dots, T$ **do**

Set

$$u_t = x_t + a_t(a_{t-1}^{-1} - 1)(x_t - x_{t-1})$$

$$x_{t+1} = u_t - \frac{1}{\beta}\nabla f(u_t)$$

Set the extrapolation coefficient

$$a_{t+1} = \frac{\sqrt{a_t^4 + 4a_t^2} - a_t^2}{2}.$$

end

The idea of the method is to maintain two sequences of points x_t and u_t. The gradient steps are computed along the points u_t, while the iterates x_t are used to encode the history, and in particular the momentum, of the algorithm. The choice of the extrapolation sequence a_t is subtle and is entirely motivated by the convergence analysis of the method.

Theorem 5.4. *Let $f\colon \mathbf{E} \to \mathbf{R}$ be a convex and β-smooth function. Then the iterates $\{x_t\}$ generated by Algorithm 2 satisfy*

$$f(x_t) - f^* \leq \frac{2\beta\|x^* - y_0\|^2}{(t+1)^2}.$$

We postpone the proof of Theorem 5.4 until Section **??**, where a more general result will be established. Meanwhile, let us take the validity of Theorem 5.4 for granted. Recall that gradient descent converges at the linear rate $1 - \kappa^{-1}$ for smooth strongly convex functions. In contrast, to make Algorithm 2 linearly convergent, one must modify the method. There are two such approaches in the literature: (1) modify the definition of a_t in (6.20) using the strong convexity constant or (2) periodically restart the algorithm. We omit the first approach, since it is fairly technical; instead, we focus on the second approach, which is elementary and has many other uses.

The idea of restarts is useful in many contexts, often allowing one to boost sublinear rates to linear rates of convergence under strong convexity assumptions. Imagine that we run the basic accelerated gradient method on f for a number of iterations (an epoch) and then restart. Let x_t^i be the t'th iterate generated in epoch i. Theorem 5.4 along with strong convexity yields the guarantee

$$f(x_t^i) - f^* \leq \frac{2\beta\|x^* - x_0^i\|^2}{(t+1)^2} \leq \frac{4\kappa}{(t+1)^2} \cdot (f(x_0^i) - f^*). \tag{5.13}$$

Suppose that in each epoch, we run an accelerated gradient method for T iterations, and initialize each epoch with the final iterate of the previous epoch. The estimate (5.13) then guarantees that the function gap shrinks by a factor of $\frac{4\kappa}{(T+1)^2}$ after each epoch. The idea is now to optimally choose the number of epochs I and the length of each epoch T in order to reach desired accuracy ϵ with minimal computational effort.

Theorem 5.5. *Let $f\colon \mathbf{E} \to \mathbf{R}$ be an α-strongly convex and β-smooth function. Fix a target accuracy $\epsilon > 0$ and set the algorithmic parameters*

$$T = \left\lceil 2e\sqrt{\kappa} \right\rceil \qquad \text{and} \qquad I = \left\lceil \ln\left(\tfrac{f(x_0)-f^*}{\varepsilon}\right) \right\rceil.$$

Then the iterate $\{x_T^I\}$ generated by Algorithm 3 satisfies $f(x) - f^* \leq \epsilon$, while the run of the algorithm uses at most

$$\left\lceil 2e\sqrt{\kappa} \right\rceil \cdot \left\lceil \ln\left(\frac{f(x_0) - f^*}{\varepsilon}\right) \right\rceil$$

gradient evaluations.

Algorithm 3: Accelerated gradient method with restarts

Input: Initial$x_0 \in \mathbf{E}$, epoch length $T \in \mathbb{N}$, and epoch number $I \in \mathbb{N}$
Set $x_T^{-1} = x_0$.
for $i = 0, \ldots, I$ **do**
  Set $x_0^i = x_T^{i-1}$.
  Let x_T^i be the T'th iterate generated by Algorithm 2, initialized with x_0^i.
end

Proof. The estimate (5.13) immediately guarantees:

$$f(x_T^I) - f^* \leq \left(\frac{4\kappa}{(T+1)^2}\right)^I (f(x_0) - f^*) \leq e^{-I}(f(x_0) - f^*) \leq \epsilon, \quad (5.14)$$

thereby completing the proof. □

The efficiency guarantee $\mathcal{O}(\sqrt{\kappa} \cdot \ln\left(\frac{f(x_0)-f^*}{\epsilon}\right)$ of Algorithm (3) is much better than the efficiency guarantee for gradient descent $\mathcal{O}(\kappa \cdot \ln\left(\frac{f(x_0)-f^*}{\epsilon}\right)$. Indeed, we will see in Section 5.4 that the sublinear and linear efficiency estimates of the accelerated gradient method are the best possible among all first-order algorithms for smooth convex minimization.

5.2 Algorithms for nonsmooth convex minimization

In this section, we dispense with the smoothness assumption and focus on the optimization problem

$$\min_{x \in Q} f(x), \quad (5.15)$$

where $f\colon \mathbf{E} \to \mathbf{R}$ is a convex function that is L-Lipschitz continuous on a neighborhood of a closed convex set $Q \subset \mathbf{E}$. Thus in comparison with the previous section, we have weakened the smoothness assumption to Lipschitz continuity, while also allowing a convex constraint set. The analysis of the projected subgradient method—the content of this section—is essentially the same whether Q is present or not. Extensions of gradient descent to the constrained setting, in contrast, require a different argument and will appear in Section 6.1 .

5.2.1 Subgradient method

The projected subgradient method, summarized in Algorithm 4, is the basic algorithm for the problem (5.15). In each iteration t, the algorithm chooses a subgradient $v_t \in \partial f(x_t)$ and declares

$$x_{t+1} = \text{proj}_Q(x_t - \eta_t v_t),$$

for some user specified sequence $\{\eta_t\}$. Thus the method travels in the opposite direction to a subgradient v_t and then performs a projection operation to restore feasibility. Notice that in particular, the projection $\text{proj}_Q(\cdot)$ needs to be efficiently computable in order for the algorithm to be implementable.

Algorithm 4: Projected subgradient method

Input: Initial $x_0 \in \mathbf{E}$, iteration $T \in \mathbb{N}$, sequence $\{\eta_t\} \subset (0, \infty)$.
Step $t = 0, 1, \ldots, T-1$:
Choose $v_t \in \partial f(x_t)$
Set $x_{t+1} = \text{proj}_Q(x_t - \eta_t v_t)$

There is an important difference between the smooth and nonsmooth settings. A crucial property of gradient descent is that the function values $f(x_t)$ decrease along the iterate sequence. Indeed, our reasoning for setting $\eta = \frac{1}{\beta}$ was entirely motivated by the desire to force the function value to decrease as much as possible in a single iteration (Lemma 5.1). In contrast, the projected subgradient algorithm for nonsmooth optimization is not a descent method. Indeed, if $v \in \partial f(x)$ is an arbitrary nonzero sugradient, then following the direction $-v$ might not lead to function decrease. Moreover, even if we could find a vector $v \in \partial f(x_t)$ of minimal norm, which by Exercise 4.21 defines a descent direction, it is not possible to uniformly control the amount by which the function decreases.

Instead, we monitor a different quantity, the distance to an optimal solution $\|x - x^*\|$. To see why this is a reasonable strategy, observe that if x is not a minimizer of f, then any subgradient $v \in \partial f(x)$ satisfies

$$\langle v, x^* - x \rangle \leq f(x^*) - f(x) < 0. \tag{5.16}$$

Thus any subgradient $v \in \partial f(x)$ makes an obtuse angle with the vector $x^* - x$. It follows immediately that for all sufficiently small $\eta > 0$, the updated point $x - \eta v$ moves closer to x^*. Moreover, (5.16) suggests that if the ratio $\frac{f(x)-f^*}{\|v\| \cdot \|x-x^*\|}$ is large, then we should be able to make good progress towards x^* by following the direction $-v$. More formally, the convergence

analysis of the projected subgradient method (Theorem 5.6) will rely on the elementary estimate

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - 2\eta_t(f(x_t) - f^*) + \eta_t^2\|v_t\|^2.$$

Intuitively, since the third term on the right scales as η_t^2, it can be made small relative to the second term, which scales as η_t. Thus the function gap $f(x_t) - f^*$ indeed controls the decrease in the distance to the solution. Since, the distance $\|x_t - x^*\|$ is lower bounded by zero, the function values $f(x_t) - f^*$ must tend to zero. With this intuition in mind, we are now ready to establish the convergence guarantees for the projected subgradient method. It is worthwhile to note that in the nonsmooth setting, rather than measuring the suboptimality of any individual iterate x_t, it is most natural to measure the suboptimality of the running average of the iterates.

Theorem 5.6 (Subgradient method under convexity)**.** *Let $f\colon \mathbf{E} \to \mathbf{R}$ be a convex function that is L-Lipschitz continuous on a neighborhood of a closed convex set $Q \subset \mathbf{E}$. Then the iterates generated by Algorithm 4 satisfy*

$$f\left(\frac{1}{\sum_{i=0}^t \eta_i}\sum_{i=0}^t \eta_i x_i\right) - f^* \leq \frac{\|x_0 - x^*\|^2 + L^2\sum_{i=0}^t \eta_i^2}{2\sum_{i=0}^t \eta_i}. \tag{5.17}$$

In particular, when using the constant parameter $\eta_t = \frac{R}{L\sqrt{T+1}}$ for a fixed $R \geq \|x_0 - x^*\|$, the efficiency estimate becomes

$$f\left(\frac{1}{T+1}\sum_{t=0}^T x_t\right) - f^* \leq \frac{RL}{\sqrt{T+1}}. \tag{5.18}$$

Proof. We successively compute

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|\mathrm{proj}_Q(x_t - \eta_t v_t) - x^*\|^2 \\
&= \|\mathrm{proj}_Q(x_t - \eta_t v_t) - \mathrm{proj}_Q(x^*)\|^2 \\
&\leq \|(x_t - x^*) - \eta_t v_t\|^2 && (5.19)\\
&= \|x_t - x^*\|^2 - 2\eta_t\langle v_t, x_t - x^*\rangle + \eta_t^2\|v_t\|^2, && (5.20)\\
&\leq \|x_t - x^*\|^2 - 2\eta_t(f(x_t) - f^*) + \eta_t^2 L^2, && (5.21)
\end{aligned}$$

where (5.19) uses that proj_Q is 1-Lipschitz continuous (Theorem 3.60) and (5.21) uses convexity and Lipschitz continuity of f. Iterating the recursion yields

$$\|x_{T+1} - x^*\|^2 \leq \|x_0 - x^*\|^2 - 2\sum_{t=0}^T \eta_t(f(x_t) - f^*) + L^2\sum_{t=0}^T \eta_t^2.$$

Lower-bounding the left side by zero and rearranging, we conclude

$$\sum_{t=0}^{T} \eta_t (f(x_t) - f^*) \leq \frac{\|x_0 - x^*\|^2 + L^2 \sum_{t=0}^{T} \eta_t^2}{2}. \tag{5.22}$$

Finally using convexity, observe

$$f\left(\frac{1}{\sum_{t=0}^{T} \eta_t} \sum_{t=0}^{T} \eta_t x_t\right) - f^* \leq \frac{\sum_{t=0}^{T} \eta_t (f(x_t) - f^*)}{\sum_{i=0}^{t} \eta_t}.$$

Combining this estimate with (5.22) completes the proof of (5.17). Setting $\eta_t = \eta$ for all $t = 0, \ldots, T-1$ in (5.17) yields the guarantee

$$f\left(\frac{1}{T+1} \sum_{t=0}^{T} x_t\right) - f^* \leq \frac{\|x_0 - x^*\|^2}{2(T+1)\eta} + \frac{L^2 \eta}{2}.$$

Optimizing the right side of (5.17) in η yields the choice $\eta = \frac{R}{L\sqrt{T+1}}$ and the guarantee (5.18). □

Thus, (5.18) shows that setting η_t to be a constant, the uniform average of the iterates $\{x_t\}_{t=0}^{T}$ is suboptimal by an additive error $\frac{RL}{\sqrt{T+1}}$. When f is α-strongly convex with $\alpha > 0$, a faster rate is achievable by making the judicious choice $\eta_t = \frac{2}{\alpha(t+1)}$ and averaging the iterates non-uniformly, namely by assigning higher weight to the latter iterates.

Theorem 5.7 (Subgradient method under strong convexity)**.** *Let $f\colon \mathbf{E} \to \mathbf{R}$ be an α-strongly convex function that is L-Lipschitz continuous on a neighborhood of a closed convex set $Q \subset \mathbf{E}$. Then the iterates generated by Algorithm 4 with $\eta_t = \frac{2}{\alpha(t+1)}$ satisfy*

$$f\left(\tfrac{2}{t(t+1)} \sum_{i=1}^{t} i x_i\right) - f^* \leq \frac{2L^2}{\alpha(t+1)}.$$

Proof. Starting from (5.20) and using Lipschitz continuity and strong convexity of f, we compute

$$\begin{aligned} \|x_{t+1} - x^*\|^2 &\leq \|x_t - x^*\|^2 + 2\eta_t \langle v_t, x^* - x_t \rangle + \eta_t^2 \|v_t\|^2 \\ &\leq \|x_t - x^*\|^2 + 2\eta_t \big(f^* - f(x_t) - \tfrac{\alpha}{2} \|x^* - x_t\|^2\big) + \eta_t^2 L^2. \end{aligned}$$

Rearranging and diving through by $2\eta_t$ yields the expression

$$f(x_t) - f^* \leq \left(\frac{1-\alpha\eta_t}{2\eta_t}\right)\|x_t - x^*\|_2^2 - \frac{1}{2\eta_t}\|x_{t+1} - x^*\|_2^2 + \frac{\eta_t}{2}L^2.$$

Plugging in $\eta_t := \frac{2}{\alpha(t+1)}$ and multiplying through by t, we obtain

$$t\big(f(x_t) - f(x^*)\big) \leq \frac{\alpha t(t-1)}{4}\|x_t - x^*\|^2 - \frac{\alpha t(t+1)}{4}\|x_{t+1} - x^*\|^2 + \frac{t}{\alpha(t+1)}L^2.$$

Summing the estimate for $i = 1\ldots, t$, the first two terms on the right-side telescope, yielding

$$\sum_{i=1}^{t} i\big(f(x_i) - f(x^*)\big) \leq \sum_{i=1}^{t} \tfrac{i}{\alpha(i+1)}L^2 \leq \frac{tL^2}{\alpha}.$$

Dividing through by $\sum_{i=1}^{t} i = \frac{t(t+1)}{2}$ and taking into account that f is convex, we conclude

$$f\left(\tfrac{2}{t(t+1)}\sum_{i=1}^{t} i x_i\right) - f^* \leq \left(\frac{1}{\sum_{i=1}^{t} i}\right)\cdot\sum_{i=1}^{t} i\big(f(x_i) - f(x^*)\big) \leq \frac{2L^2}{\alpha(t+1)},$$

as claimed. □

5.3 Model-based view of first-order methods

Sections 5.1 and 5.2 presented the "classical" motivation and analysis of gradient descent and the projected subgradient method. The current section revisits both algorithms from a more modern perspective, rooted in model approximation. Sections 6.1 and 6.2 will leverage this viewpoint to develop new algorithms for a wider class of problems.

Setting the stage, fix a point $x \in \mathbf{E}$ and a vector $v \in \mathbf{E}$, and define the update

$$x^+ = \mathrm{proj}_Q(x - \eta v), \tag{5.23}$$

where Q is a closed convex set. Clearly, gradient decent and the projected subgradient method are examples of this update rule, under the settings $v = \nabla f(x)$ and $v \in \partial f(x)$, respectively. The key observation now is that the update (5.23) can be equivalently written in the variational form

$$x^+ = \operatorname*{argmin}_{y\in Q}\ f_x(y) + \frac{1}{2\eta}\|y - x\|^2. \tag{5.24}$$

where $f_x(\cdot)$ is the affine function

$$f_x(y) := f(x) + \langle v, y - x \rangle. \tag{5.25}$$

To see this, successively compute

$$\begin{aligned} x_+ = \operatorname*{argmin}_{y \in Q} \ \|y - (x - \eta v)\|^2 &= \operatorname*{argmin}_{y \in Q} \ 2\eta \langle v, y - x \rangle + \|y - x\|^2 \\ &= \operatorname*{argmin}_{y \in Q} \ f_x(y) + \frac{1}{2\eta} \|y - x\|^2. \end{aligned}$$

Thus x^+ is obtained by minimizing the sum of the affine model f_x of f and a simple quadratic. The assumptions of α-strong convexity and β-smoothness amount to the lower and upper estimates on approximation quality, respectively:

$$f(y) \leq f_x(y) + \frac{\beta}{2} \|y - x\|^2 \qquad \forall x, y \in \mathbf{E}, \tag{5.26}$$

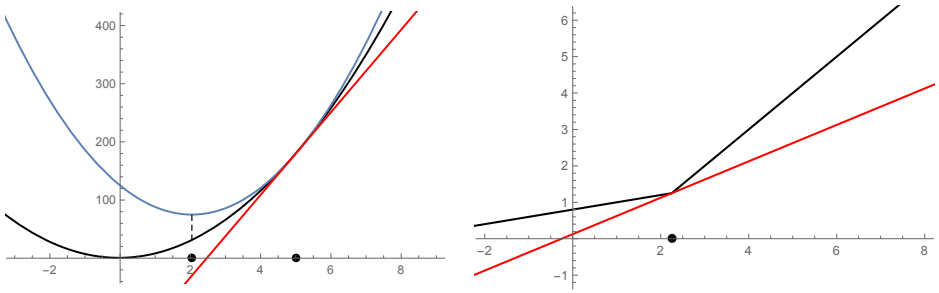
$$f(y) \geq f_x(y) + \frac{\alpha}{2} \|y - x\|^2 \qquad \forall x, y \in \mathbf{E}. \tag{5.27}$$

Observe that in the smooth setting, the function $f_x + \frac{\beta}{2} \| \cdot - x\|^2$ used in the update rule (5.24) for gradient descent globally upper bounds f. An immediate consequence, already well known to us, is that gradient descent generates iterates with decreasing function values. See Figure 5.3 for an illustration. In contrast, the function $f_x + \frac{1}{2\eta} \| \cdot - x\|^2$ defining the subgradient step in the nonsmooth setting might not be an upper model of f for any $\eta > 0$. Herein lies the distinction between smooth and nonsmooth optimization.

In Sections 6.1 and 6.2, we will see that the functional form of the models (5.25) is completely irrelevant for convergence guarantees of gradient descent and the projected subgradient method. Any algorithm that proceeds according to (5.24), with arbitrary convex models $f_x(\cdot)$ satisfying the two-sided accuracy (5.26)&(5.27), enjoys convergence guarantees that parallel those of gradient descent. Similarly, any algorithm that proceeds according to (5.24), with arbitrary convex models $f_x(\cdot)$ satisfying the one-sided accuracy (5.27), enjoys convergence guarantees that parallel those of the subgradient method. The ability to use more interesting models f_x will open the door to a number of new algorithm.

5.4 Lower complexity bounds

This chapter has discussed at great length convergence guarantees of (accelerated) gradient descent and of the subgradient method. Tables 5.1 and 5.2 summarize the efficiency estimates of these algorithms.



(a) Two-sided approximation

(b) One-sided approximation

Figure 5.3: Left: the black curve depicts the graph of a β-smooth function f, the red curve depicts the graph of the function $f_x(y) = f(x) + \langle \nabla f(x), y - x \rangle$, the blue curve depicts the graph of the function $y \mapsto f_x(y) + \frac{\beta}{2}\|y - x\|^2$. Right: the black curve depicts the graph of a nonsmooth function f, the red curve depicts the graph of the function $f_x(y) = f(x) + \langle v, y - x \rangle$ with $v \in \partial f(x)$.

	convex, β-smooth	α-strongly convex, β-smooth
Gradient descent	$\frac{\beta\|x_0 - x^*\|^2}{\epsilon}$	$\kappa \cdot \log(\frac{f(x_0) - f^*}{\epsilon})$
Accel. grad. descent	$\sqrt{\frac{\beta\|x_0 - x^*\|^2}{\epsilon}}$	$\sqrt{\kappa} \cdot \log(\frac{f(x_0) - f^*}{\epsilon})$

Table 5.1: Number of gradient evaluations to find x satisfying $f(x) - f^* \leq \epsilon$

This section switches gears and instead focuses on *lower complexity bounds*, which express limitations on the convergence guarantees that any algorithm can have. In particular, we will see that the convergence guarantees of accelerated gradient descent are the best possible among any algorithm for minimizing β-smooth convex functions. Similarly, the convergence guarantees of the subgradient method are the best possible among any algorithm for minimizing L-Lipschitz convex functions.

In order to make such results precise, we specify how an algorithm gathers information about the objective function. Consider an optimization problem

$$\min_{x \in \mathbf{E}} \; f(x), \tag{5.28}$$

where $f \colon \mathbf{E} \to \mathbf{R}$ is a finite-valued convex function. We will assume that an algorithm accesses information about f by querying a first-order oracle, which on input $x \in \mathbf{E}$ returns some subgradient $v \in \partial f(x)$. In particular, if f is smooth, then the oracle on input x simply returns the gradient $\nabla f(x)$. When f is nonsmooth, the oracle returns an arbitrary subgradi-

	convex, L-Lipschitz	α-strongly convex, L-Lipschitz
Subgrad. method	$\frac{L^2R^2}{\epsilon^2}$	$\frac{L^2}{\alpha\epsilon}$

Table 5.2: Number of subgradient evaluations to find x satisfying $f(x)-f^* \leq \epsilon$, where an upper bound $R \geq \|x_0 - x^*\|$ is assumed to be known.

ent $v \in \partial f(x)$. We will prove lower-complexity bounds for a large class of algorithms, summarized in the following definition.

Definition 5.8 (Linearly-expanding first-order method)**.** An algorithm for (5.28) is called a *linearly-expanding first-order method* if it generates an iterate sequence $\{x_k\}$ satisfying

$$x_t \in x_0 + \operatorname{span}\{v_0, \ldots, v_{t-1}\} \qquad \text{for } t \geq 1,$$

where $v_i \in \partial f(x_i)$ is generated by a call to the first-order oracle of f with input x_i.

5.4.1 Lower-complexity bound for nonsmooth convex optimization

We begin by proving a lower-complexity bound for minimizing an L-Lipschitz continuous convex function on a ball.

Theorem 5.9 (Lower-complexity bound for nonsmooth convex optimization)**.** *Fix a dimension $n \in \mathbb{N}$, an iteration counter $t \leq n$, and a real $L > 0$. Then there exists a convex function $f\colon \mathbf{R}^n \to \mathbf{R}$ that is L-Lipschitz continuous on a ball $B_R(0)$, for some $R > 0$, and such that any linear expanding first-order method initialized at the origin satisfies*

$$\min_{k=1,\ldots,t-1} f(x_k) - \min_{x \in B_R(0)} f(x) \geq \frac{RL}{2(1+\sqrt{t})}.$$

There also exists an α-strongly convex function $f\colon \mathbf{R}^n \to \mathbf{R}$ that is Lipschitz continuous on $B_R(0)$, for some $R > 0$, and such that any linear expanding first-order method initialized at the origin satisfies

$$\min_{k=1,\ldots,t-1} f(x_k) - \min_{x \in B_R(0)} f(x) \geq \frac{L^2}{8\alpha t}.$$

Proof. Fix two constants $\gamma, \alpha > 0$, which will be specified shortly. We will consider the outcome of applying any linearly expanding first-order method

to the α-strongly convex function:

$$f(x) = \gamma \max_{i=1,\dots,t} x_i + \frac{\alpha}{2}\|x\|^2.$$

Observe the expression for the subdifferential (Exercise 3.41)

$$\partial f(x) = \gamma \cdot \operatorname{conv}\{e_i : i \in I(x)\} + \alpha x,$$

where $I(x)$ consists of all indices $i = 1, \dots, t$ satisfying $x_i = \max_{j=1,\dots,t} x_j$. Taking into account that the function $x \mapsto \max_{i=1,\dots,t} x_i$ is 1-Lipschtiz continuous, while the quadratic $\frac{1}{2}\|x\|^2$ is R-Lipschitz on the ball $B_R(0)$, we deduce that f is Lipschitz continuous on $B_R(0)$ with constant $\gamma + \alpha R$.

Next we describe a first-order oracle for f. When asked for a subgradient at x, the oracle returns $\gamma e_i + \alpha x$, where i is the smallest coordinate in the set of active indices $I(x)$. In particular, if the algorithm is initialized at $x_0 = 0$, then the oracle returns e_1. Therefore the next iterate x_1 lies on the line generated by e_1. A simple induction shows that the iterate x_k lies in the linear span of $e_1, \dots, e_k$ for all $k = 1, \dots, t$. In particular, the t'th coordinate of x_{t-1} is zero and therefore $f(x_{t-1}) \geq 0$. Finally, let's compute the minimal value of f. We claim that the minimizer of f is the point x^* defined by setting $x_i^* = \frac{-\gamma}{\alpha t}$ for $1 \leq i \leq t$ and $x_i^* = 0$ for $t+1 \leq i \leq n$. To see this, simply observe $0 = \gamma \sum_{i=1}^t \frac{1}{t} e_i + \alpha x^* \in \partial f(x^*)$. Therefore, the minimal value of f is

$$f^* = f(x^*) = -\frac{\gamma^2}{\alpha t} + \frac{\alpha}{2}\frac{\gamma^2}{\alpha^2 t} = -\frac{\gamma^2}{2\alpha t}.$$

Thus we conclude

$$f(x_i) - f^* \geq \frac{\gamma^2}{2\alpha t}, \qquad \text{for all } i = 1, \dots, t-1.$$

Setting $\gamma = \frac{L}{2}$ and $R = \frac{L}{2\alpha}$ proves the lower bound for α-strongly convex functions. Taking $\alpha = \frac{L}{R}\frac{1}{1+\sqrt{t}}$ and $\gamma = L\frac{\sqrt{t}}{1+\sqrt{t}}$ completes the the proof of the lower-bound for non-strongly convex functions. In both cases, a short computation (do it!) verifies that x^* indeed lies in $B_R(0)$. □

In light of Theorem 5.9, we see that the projected subgradient method has the best possible efficiency estimate when minimizing Lipschitz (strongly) convex functions on a Euclidean ball. It is worthwhile to note that Theorem 5.9 stipulates that the dimension n of the ambient space is larger than

the number of iterations t. This is not an artifact of the proof. In the setting $n \ll t$, there exist much faster first-order methods than the subgradient algorithm, indeed ones that converge at a linear rate that depends on n. The reader may consult for example the recent survey [11] or the classical text [26].

5.4.2 Lower-complexity bound for smooth convex optimization

We now prove a lower complexity bound for minimizing β-smooth convex functions.

Theorem 5.10 (Lower-complexity bound for smooth convex optimization)**.** *Fix a dimension $n \in \mathbb{N}$, an iteration counter $1 \leq t \leq (n-1)/2$, and a real $\beta > 0$. Then there exists a convex β-smooth function $f\colon \mathbf{R}^n \to \mathbf{R}$ so that the iterates generated by any linearly-expanding first-order method started at x_0 satisfy*

$$f(x_t) - \min f \geq \frac{3\beta\|x_0 - x^*\|^2}{32(t+1)^2}, \tag{5.29}$$

where x^* is any minimizer of f.

Without loss of generality, assume $x_0 = 0$. The argument proceeds by constructing a uniformly worst function for all linearly-expanding first-order methods. The construction will guarantee that in the k'th iteration of such a method, the iterate x_t will lie in the subspace $\mathbf{R}^t \times \{0\}^{n-t}$. This will cause the function value at the iterates to be far from the optimal value.

Here is the precise construction. Fix a constant $\beta > 0$ and define the following family of quadratic functions

$$f_t(z_1, z_2, \ldots, z_n) = \tfrac{\beta}{4}\Big(\tfrac{1}{2}(z_1^2 + \sum_{i=1}^{t-1}(z_i - z_{i+1})^2 + z_t^2) - z_1\Big)$$

indexed by $t = 1, \ldots, n$. It is easy to check that f_t is convex and β-smooth. Indeed, a quick computation shows

$$\langle \nabla^2 f_t(x)v, v\rangle = \tfrac{\beta}{4}\Big((v_1^2 + \sum_{i=1}^{t-1}(v_i - v_{i+1})^2 + v_t^2)\Big)$$

and therefore

$$0 \leq \langle \nabla^2 f_t(x)v, v\rangle \leq \tfrac{\beta}{4}\Big((v_1^2 + \sum_{i=1}^{t-1} 2(v_i^2 + v_{i+1}^2) + v_t^2)\Big) \leq \beta\|v\|^2.$$

The following exercise establishes a few basic properties of the function f_t that we will need.

Exercise 5.11. Establish the following properties of f_t.

1. Define the point $\bar{x}_t \in \mathbf{R}^n$ whose i'th coordinate is given by

$$\begin{cases} 1 - \frac{i}{t+1}, & \text{if } i = 1, \ldots, t \\ 0 & \text{if } i = t+1, \ldots, n \end{cases}$$

   Using first-order conditions for optimality, show that $\bar{x}_t$ is a minimizer of f_t with optimal value $f_t^* = \frac{\beta}{8}\left(-1 + \frac{1}{t+1}\right)$.

2. Taking into account the standard inequalities,

$$\sum_{i=1}^{t} i = \frac{t(t+1)}{2} \qquad \text{and} \qquad \sum_{i=1}^{t} i^2 \leq \frac{(t+1)^3}{3},$$

   show the estimate $\|\bar{x}_t\|^2 \leq \frac{1}{3}(t+1)$.

3. Fix indices $1 < i < j < n$ and a point $x \in \mathbf{R}^i \times \{0\}^{n-i}$. Show that equality $f_i(x) = f_j(x)$ holds and that the gradient $\nabla f_t(x)$ lies in $\mathbf{R}^{i+1} \times \{0\}^{n-(i+1)}$.

Proof of Theorem 5.10. Proving Theorem 5.10 is now straightforward. Fix t and apply the linearly-expanding first order method to $f := f_{2t+1}$ starting at $x_0 = 0$. Let x^* be the minimizer of f and f^* the minimum of f. By Exercise 5.11 (part 3), the iterate x_t lies in $\mathbf{R}^t \times \{0\}^{n-t}$. Therefore by the same exercise, we have $f(x_t) = f_t(x_t) \geq \min f_t$. Taking into account parts 1 and 2 of Exercise 5.11, we deduce

$$\frac{f(x_t) - f^*}{\|x_0 - x^*\|^2} \geq \frac{\frac{\beta}{8}\left(-1 + \frac{1}{t+1}\right) - \frac{\beta}{8}\left(-1 + \frac{1}{2t+2}\right)}{\frac{1}{3}(2t+2)} = \frac{3\beta}{32(t+1)^2}.$$

This proves the result. □

The lower-complexity bounds in Theorem 5.10 do not depend on the strong convexity constant. When the target function class consists of β-smooth and α-strongly convex functions, the analogous complexity bound becomes

$$f(x_t) - f^* \geq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{2t} \|x_0 - x^*\|^2. \tag{5.30}$$

where x^* is any minimizer of f. The proof is similar to that of Theorem 5.10, where one modifies the definition of f_t by adding a multiple of the quadratic $\|\cdot\|^2$.

In conclusion, we see that the accelerated gradient method achieves the best possible efficiency estimate for minimizing β-smooth convex functions. Similarly, the restarted accelerated gradient method achieves the best possible efficiency estimate for minimizing α-strongly convex and β-smooth functions.

References: The earliest convergence guarantees for gradient descent and the subgradient method can be found for example in Nemirovsky-Yudin [26] and Shor [35]. The accelerated gradient method and its restarted variant were developed by Nesterov in [28]. The proof of Theorem 5.7 appears in [18], though an analogous guarantee that is suboptimal by a log factor appears in [26]. The complexity viewpoint for first-order methods, which we follows here, originates in the monograph of Nemirovsky-Yudin [26]. The treatment of lower-complexity bounds in Section 5.4 follows the simplified treatment in Nesterov [27]. Contemporary texts focusing on first-order methods include Beck [6], Bertsekas [7], Bubeck [11], and Nesterov [27].

5.5 Additional exercises

Exercise 5.12 (Ridge Regression)**.** In this exercise, you will consider the ridge regression problem:

$$\min_{x\in\mathbf{R}^n} \ \frac{1}{2}\|Ax-y\|_2^2+\frac{\lambda}{2}\|x\|_2^2,$$

which aims to recover a point x_{opt} from (noisy) linear measurements y. For $n = 100$, $m = 80$, and $\lambda = 1$, generate data as follows:

- the underlying signal is drawn $x_{opt} \sim N(0, I)$,
- the measurement matrix $A \in \mathbf{R}^{m\times n}$ is drawn with independent standard Gaussian rows $A_i \sim N(0, I)$,
- the observed data $y \in \mathbf{R}^m$ is

$$y = Ax_{opt} + \epsilon, \qquad \epsilon_i \sim N(0, 0.25) \text{ i.i.d.}$$

Note that the objective function is $\beta-$smooth with $\beta = \|A^*A\|_{op} + \lambda$, and $\alpha-$strongly convex with $\alpha = \lambda$. In your experiments, you may set $\beta =$

$4(m+n)+\lambda$.[1] Write code that generates the problem data as above and implement the following algorithms:

1. Gradient descent,
2. Accelerated gradient descent.

For each algorithm, plot the the function value over the first 50 iterations on a semilog plot.

Exercise 5.13 (Huber regression)**.** Consider the huber-ized version of ridge regression:

$$\min_{x\in\mathbf{R}^n} \sum_{i=1}^{m} h_\eta(y_i - a_i^T x) + \frac{\lambda}{2}\|x\|_2^2$$

where a_i^T are the rows of A, and h_η is the huber function with parameter η:

$$h_\eta(w) = \begin{cases} \frac{1}{2}w^2, & \text{if } |w| \le \eta \\ \eta(|w| - \eta/2), & \text{otherwise.} \end{cases}$$

The huber function penalizes large deviations less than the quadratic cost function, and is therefore often used when outliers are present in the data. Note that the objective function in this case is $\beta-$smooth with $\beta > m\eta + \lambda$ and $\alpha-$strongly convex with $\alpha = \lambda$.

Generate x_{opt} and A as in Exercise 5.12, and generate $y = Ax_{opt} + \epsilon$ where ϵ contains 5 outliers drawn from $N(0, 25)$ and the rest of its entries are again independent $N(0, 0.25)$. Implement the same two algorithms as in Exercise 5.12 on this problem, and plot the function values for the first 100 iterations in a semilog plot.

Exercise 5.14 (Logistic regression)**.** Consider the regularized logistic regression problem:

$$\min_{\theta\in\mathbf{R}^n} \sum_{i=1}^{m} \ln\left(1 + \exp(-y_i\theta^T x_i)\right) + \frac{\lambda}{2}\|\theta\|_2^2.$$

Here $y_i \in \{\pm 1\}$ are binary labels for the data points $x_i \in \mathbf{R}^n$, and we are trying to find the best vector θ of parameters for the model

$$P(y = 1|x, \theta) = \frac{1}{1 + e^{-\theta^T x}}.$$

[1] By Gaussian concentration, with high probability we have $\|A\|_{op} \lesssim \sqrt{m} + \sqrt{n}$.

(We are doing this by minimizing the negative log likelihood function plus a regularization term.)

Implement the same two algorithms as before, with $n = 50$ and $m = 100$ and $\lambda = 1$. Generate data as follows:

1. $\theta_{opt} = (1, \ldots, 1)^T$;
2. $X \in \mathbf{R}^{m \times n}$ has i.i.d. standard normal entries, with rows x_i^T;
3. $z = X\theta_{opt}$ and the true vector of probabilities is $p = 1/(1 + \exp(-z))$, where the operations are applied entrywise to the vector z;
4. $y \in \{\pm 1\}^m$ has independent Bernoulli entries with $P(y_i = 1) = p_i$.
5. θ_0 has i.i.d. standard normal entries.

Implement the same two algorithms as in Exercise 5.12 on this problem, and plot the function values for the first 100 iterations in a semilog plot. You may take the smoothness parameter of the objective function to be $\beta = 100$ and the strong convexity parameter to be $\alpha = 1$.

Exercise 5.15 (Polyak stepsize)**.** Consider a differentiable convex function $f : \mathbf{E} \to \mathbf{R}$ and let x^* be any of its minimizers. Consider the gradient descent iterates

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

for some sequence $\gamma_k \geq 0$.

1. By writing the term $\|x_{k+1} - x^*\|^2 = \|(x_{k+1} - x_k) + (x_k - x^*)\|^2$ and expanding the square, deduce the estimate

$$\frac{1}{2}\|x_{k+1} - x^*\|^2 \leq \frac{1}{2}\|x_k - x^*\|^2 - \gamma_k(f(x_k) - f(x^*)) + \frac{\gamma_k^2}{2}\|\nabla f(x_k)\|^2. \tag{5.31}$$

2. Supposing that you know the minimal value f^* of f, show that the sequence $\gamma_k = \frac{f(x_k) - f^*}{\|\nabla f(x_k)\|^2}$ minimizes the right-hand-side of (5.31) in γ, thereby yielding the guarantee

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \left(\frac{f(x_k) - f^*}{\|\nabla f(x_k)\|}\right)^2.$$

3. Let x_k be the sequence generated by the gradient method with $\gamma_k = \frac{f(x_k)-f^*}{\|\nabla f(x_k)\|^2}$. Supposing that f is β-smooth, conclude the estimate

$$f\left(\frac{1}{k}\sum_{i=0}^{k-1} x_i\right) - f^* \leq \frac{2\beta\|x_0 - x^*\|^2}{k}.$$

If f is in addition α-strongly convex, derive the guarantee

$$\|x_{k+1} - x^*\|^2 \leq \left(1 - \frac{\alpha}{4\beta}\right)\|x_k - x^*\|^2.$$

Chapter 6

Algorithms for additive composite problems

In the previous chapter, we saw that there is a sharp distinction in the complexity of smooth and nonsmooth optimization. Namely, the accelerated gradient method finds an ϵ-approximate minimizer of a β-smooth convex function using on the order of $\sqrt{\frac{\beta\|x_0-x^*\|^2}{\epsilon}}$ gradient evaluations, while the subgradient method finds an ϵ-approximate minimizer of an L-Lipschitz convex function using on the order of $\left(\frac{LR}{\epsilon}\right)^2$ gradient evaluations, where $R \geq \|x_0 - x^*\|$ is assumed to be known. Both efficiency estimates are the best possible for the two problem classes. That being said, the lower-complexity bounds we derived make the fundamental assumption that the only access to the objective functions is through a first-order oracle. Typical nonsmooth problems that arise in practice, however, are highly structured and one can hope to use this structure to develop faster algorithms. In this section, we focus on one well-structured class of problems, for which this is indeed possible.

Throughout the section, we consider the optimization problem in *additive composite* form:

$$\min_x \ \varphi(x) = f(x) + r(x), \tag{6.1}$$

where $f\colon \mathbf{E} \to \mathbf{R}$ is a "complicated" function and $r\colon \mathbf{E} \to \overline{\mathbf{R}}$ is "simple". The meaning of the words "complicated" and "simple" will become clear shortly. The two functions f and r will play different roles algorithmically. The reader should think of f as a difficult function, which needs to be replaced by a simpler model within numerical schemes. In contrast, r is already simple, such as an ℓ_p-norm. Let us look at the two most important

examples of this problem class.

Example 6.1 (Smooth plus simple)**.** Consider the setting where

1. f is convex and β-smooth, and
2. r is closed and convex, but possibly nonsmooth.

Let us assume moreover that r is simple in the sense that its proximal map is efficiently computable. Then a natural algorithm, which we will motivate shortly, is the *proximal gradient method*:

$$x_{t+1} = \operatorname{prox}_{r/\beta}\left(x_t - \tfrac{1}{\beta}\nabla f(x)\right).$$

Thus, the method accesses f by computing its gradient, while accessing r through its proximal map. We will show that even though the sum $\varphi = f + r$ is nonsmooth, the proximal gradient method converges at a similar rate as gradient descent for smooth minimization. Moreover, the proximal gradient method can be accelerated by using inertia.

Example 6.2 (Lipschitz plus simple)**.** Suppose next that

1. f is convex and L-Lipschitz, and
2. r is closed and convex, but has a huge Lipschitz constant (maybe infinite).

Let us assume again that the proximal map of r is efficiently computable. Then the *proximal subgradient method*, which we will motivate shortly, in each iteration t chooses a subgradient $v_t \in \partial f(x_t)$ and declares

$$x_{t+1} = \operatorname{prox}_{\eta_t r}\left(x_t - \eta_t v_t\right),$$

where $\eta_t > 0$ is a user-specified control sequence. We will show that even though the sum $\varphi = f + r$ might not be globally Lipschitz, the proximal subgradient method converges at a similar rate as the basic subgradient method for minimizing Lipschitz convex functions.

Moreover, even when r is Lipschitz, the proximal subgradient method may have advantages oven the vanilla subgradient method. For example, when $r = \|\cdot\|_1$ is the ℓ_1-norm on $\mathbf{R}^n$, the application of the proximal map $\operatorname{prox}_{\eta_t\|\cdot\|_1}$ tends to generate iterates x_t that are sparse and are therefore easier to store and manipulate. The iterates generated by the basic subgradient method, in contrast, are not sparse.

Both the proximal gradient and the proximal subgradient methods are examples of a more general class of algorithms. Namely, in this section we analyze procedures that simply iterate the steps

$$x_{t+1} = \operatorname*{argmin}_x \; f_{x_t}(x) + r(x) + \frac{1}{2\eta_t}\|x - x_t\|^2,$$

where $\eta_t > 0$ is a user specified control sequence and $f_{x_t}(\cdot)$ is some "model" of the function f. We will place relevant assumptions on f_{x_t} shortly. Not surprisingly, the convergence guarantees of Algorithm 5 will strongly depend on how accurately the models $f_x(\cdot)$ approximate f.

Algorithm 5: Model Based Algorithm

Input: Starting point $x_0 \in \mathbf{E}$, parameters $\eta_t > 0$, iteration $T \in \mathbb{N}$.

Step $t = 0, 1, \ldots, T-1$:

Set $x_{t+1} = \operatorname*{argmin}_x \; f_{x_t}(x) + r(x) + \frac{1}{2\eta_t}\|x - x_t\|^2$.

6.1 Proximal methods based on two-sided models

In this section, we analyze Algorithm 5, when we have two-sided control on the error $f(y) - f_x(y)$. Formally, we impose the following assumption throughout the section.

Assumption 6.3 (Two-sided model)**.** Let $f_x \colon \mathbf{E} \to \mathbf{R}$ be a family of functions indexed by points $x \in \operatorname{dom} r$. Assume that there exist real $\alpha_1, \alpha_2 \geq 0$ and $\beta > 0$ satisfying the following properties.

(A1) **(Two-sided accuracy)** The estimate holds:

$$\frac{\alpha_1}{2}\|y - x\|^2 \leq f(y) - f_x(y) \leq \frac{\beta}{2}\|y - x\|^2 \qquad \forall x, y \in \mathbf{E}.$$

(A2) **(Convexity)** The functions $f_x(\cdot) + r(\cdot)$ are α_2-strongly convex for all $x \in \mathbf{E}$.

In a nutshell, under assumption (6.3), the convergence guarantees of Algorithm 5 directly parallel those of gradient descent. Before delving into the convergence analysis, let us look at three important instantiations of Algorithm 5 with two-sided models.

Example 6.4 (Proximal gradient method)**.** The first and most important example occurs when $f\colon \mathbf{E} \to \mathbf{R}$ is β-smooth and $r\colon \mathbf{E} \to \overline{\mathbf{R}}$ is closed and and convex. In this case, we may simply use the affine models

$$f_x(y) = f(x) + \langle \nabla f(x), y - x \rangle.$$

Algorithm 5, equipped with these models is called the *proximal gradient method*. In particular, setting $\eta_t = \frac{1}{\beta}$ the proximal gradient method iterates:

$$x_{t+1} = \operatorname*{argmin}_x \; f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + r(x) + \frac{\beta}{2}\|x - x_t\|^2. \tag{6.2}$$

Equivalently, by completely the square, we may write

$$\begin{aligned} x_{t+1} &= \operatorname*{argmin}_x \; r(x) + \frac{\beta}{2}\left\|x - (x_t - \tfrac{1}{\beta}\nabla f(x_t))\right\|^2 \\ &= \operatorname{prox}_{r/\beta}\left(x_t - \tfrac{1}{\beta}\nabla f(x_t)\right). \end{aligned}$$

When f is α-strongly convex, we may set $\alpha_1 = \alpha$. Similarly if r is α-strongly convex, we may set $\alpha_2 = \alpha$. Notice that in order to apply the proximal gradient method, the proximal map $\operatorname{prox}_{r/\beta}(x)$ must be computable. In particular, when r is identically zero, the method reduces to gradient descent. More generally, if r is the indicator function of a closed convex set Q, the method reduces to the projected gradient algorithm.

Example 6.5 (Partial-linearization)**.** As the second example, suppose that f can be written as a pointwise maximum

$$f(x) = \max\{f_1(x), f_2(x), \ldots, f_k(x)\},$$

for some β-smooth functions $f_i\colon \mathbf{E} \to \mathbf{R}$. Then we may choose the polyhderal models

$$f_x(y) = \max_{i=1,\ldots,k} \; \{f_i(x) + \langle \nabla f_i(x), y - x \rangle\}.$$

Each iteration of Algorithm 5 with $\eta_t = \frac{1}{\beta}$ then amounts to solving

$$x_{t+1} = \operatorname*{argmin}_x \; \max\{f_i(x_t) + \langle \nabla f_i(x_t), x - x_t \rangle\} + r(x) + \frac{1}{2\beta}\|x - x_t\|^2. \tag{6.3}$$

If r is a polyhedral function, then one can rewrite the subproblem (6.3) as minimizing a simple quadratic function over a polyhedron (why?), for which specialized algorithms are available. If each function f_i is α-strongly convex, then we may set $\alpha_1 = \alpha$. Similarly if r is α-strongly convex, we may set $\alpha_2 = \alpha$.

Example 6.6 (Proximal point method)**.** The most accurate model of a function is itself $f_x = f$ for all $x \in \mathbf{E}$. With this choice of the models, we may set $\beta > 0$ arbitrarily. Algorithm 5 is then called the *proximal point method* and it simply iterates:

$$x_{t+1} = \operatorname*{argmin}_x \ \varphi(x) + \frac{\beta}{2}\|x - x_t\|^2. \tag{6.4}$$

Equivalently, the reader should recognize the right side as the evaluation of the proximal map

$$x_{t+1} = \operatorname{prox}_{\varphi/\beta}(x_t),$$

Invoking Theorem 3.64, yet another equivalent description of the method emerges as gradient descent on the Moreau envelope

$$x_{t+1} = x_t - \frac{1}{\beta}\nabla\varphi_{1/\beta}(x_t).$$

The proximal point method is a conceptual algorithm since each subproblem (6.4) typically does not admit a solution in closed form. Instead, the implementation of the method requires invoking an auxiliary optimization algorithm, albeit on the strongly convex function $\varphi + \frac{\beta}{2}\|\cdot - x_t\|^2$. Instead, one should think of the proximal point method as a conceptual algorithm, guiding the design and analysis of other algorithms that try to amulate it. For example, one can think of Algorithm 5, with any models f_x satisfying Assumption 6.3, as an "approximate proximal point method". This viewpoint will be useful for us shortly.

Our main tool for analyzing Algorithm 5 will be the observation that each iterate x_{t+1} is by definition the minimizer of the strongly convex function $h(x) := f_{x_t}(x) + r(x) + \frac{1}{2\eta_t}\|x - x_t\|^2$. Exercise 3.58 therefore guarantees the estimate:

$$h(x) \geq h(x_{t+1}) + \frac{\alpha}{2}\|x_{t+1} - x\|^2 \qquad \forall x \in \mathbf{E}, \tag{6.5}$$

where α is the strong convexity constant of h. The convergence analysis of Algorithm 5 will be based entirely on this observation.

6.1.1 Sublinear rate

We are now ready to establish a sublinear rate of convergence for Algorithm 5, which directly parallels Theorem 5.2 for gradient descent. Looking ahead, it will also be useful to introduce the *gap function*

$$\Delta_f(x, y) = f(y) - f_x(y).$$

Observe that Δ_f is always nonnegative. We will see that not only does the residual $f(x_t) - f^*$ tend to zero at a controlled rate, but so does the average gap $\frac{1}{t}\sum_{i=0}^{t-1} \Delta_f(x_t, x^*)$. The consequences of convergence in the average gap will be discussed in Section 7.1.

Theorem 6.7. *The iterates generated by Algorithm 5 with $\eta_t = \frac{1}{\beta}$ satisfy*

$$\varphi(x_t) - \varphi(x) + \frac{1}{t}\sum_{i=0}^{t-1} \Delta_f(x_i, x) \leq \frac{\beta\|x_0 - x\|^2}{2t} \qquad \forall x \in \mathbf{E}.$$

Proof. Since x_{t+1} is the minimizer of the β-strongly convex function $f_{x_t} + r + \frac{\beta}{2}\|\cdot - x_t\|^2$, we deduce

$$\begin{aligned}
\varphi(x_{t+1}) &\leq f_{x_t}(x_{t+1}) + r(x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - x_t\|^2 && (6.6)\\
&\leq f_{x_t}(x) + r(x) + \frac{\beta}{2}\|x - x_t\|^2 - \frac{\beta}{2}\|x - x_{t+1}\|^2 && (6.7)\\
&= \varphi(x) - \Delta_f(x_t, x) + \frac{\beta}{2}\left(\|x_t - x\|^2 - \|x_{t+1} - x\|^2\right), && (6.8)
\end{aligned}$$

where (6.6) and (6.8) follow from (A1). Subtracting $\varphi(x) - \Delta_f(x_t, x)$ from both sides, summing for $i = 0, \ldots, t-1$, and dividing by t completes the proof. □

6.1.2 Linear rate

We next turn to proving a linear rate of convergence for Algorithm 5 that depends on the constants $\alpha_1, \alpha_2 \geq 0$. To this end, the view of Algorithm 5 as an approximate proximal point method will be particularly fruitful. Namely, fix a constant $\eta > 0$ and define the map $\mathcal{G}_\eta \colon \mathbf{E} \to \mathbf{E}$ by

$$\mathcal{G}_\eta(x) = \eta^{-1}(x - x^+),$$

where x^+ denotes the update

$$x^+ = \operatorname*{argmin}_y \; f_x(y) + r(y) + \frac{1}{2\eta}\|y - x\|^2.$$

The map $\mathcal{G}_\eta$ is called the *proximal gradient* because in the setting $f_x = f$, it coincides with the gradient of the Moreau envelope (Theorem 3.64). Continuing with the analogy, it seems plausible that $\mathcal{G}_\eta$ should act as an "approximate gradient" for f itself. This is the content of the following theorem.

Theorem 6.8. *Fix a point $x \in \mathbf{E}$ and set*

$$x^+ = \operatorname*{argmin}_y \; f_x(y) + r(y) + \frac{\beta}{2}\|y - x\|^2.$$

Then the estimate

$$\begin{aligned} \varphi(y) \geq \varphi(x^+) + \langle \mathcal{G}_{1/\beta}(x), y - x\rangle + \frac{1}{2\beta}\|\mathcal{G}_{1/\beta}(x)\|^2 \\ + \frac{\alpha_1}{2}\|y - x\|^2 + \frac{\alpha_2}{2}\|y - x^+\|^2, \end{aligned} \tag{6.9}$$

holds for all $x, y \in \mathbf{E}$.

In particular, in the setting $\alpha_1 = \alpha_2 = 0$, we may rewrite the estimate (6.9) as

$$\begin{aligned} \varphi(y) &\geq \varphi(x^+) + \langle \mathcal{G}_{1/\beta}(x), y - x\rangle + \frac{1}{2\beta}\|\mathcal{G}_{1/\beta}(x)\|^2 \\ &= \varphi(x^+) + \langle \mathcal{G}_{1/\beta}(x), y - x^+\rangle - \frac{1}{2\beta}\|\mathcal{G}_{1/\beta}(x)\|^2 \end{aligned}$$

for all $y \in \mathbf{E}$. That is, φ is lower bounded by the affine function $\varphi(x^+) + \langle \mathcal{G}_{1/\beta}(x), \cdot - x^+\rangle$ up to a constant offset $\frac{1}{2\beta}\|\mathcal{G}_{1/\beta}(x)\|^2$. See Figure 6.1 for an illustration.

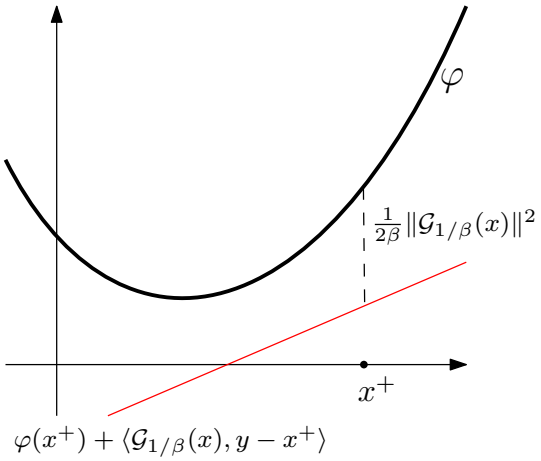


Figure 6.1: Depiction of Theorem 6.8

Proof. Since x^+ is the minimizer of the $(\beta+\alpha_2)$-strongly convex function $f_x + r + \frac{\beta}{2}\|y-x\|^2$, we deduce

$$\begin{aligned} f_x(y) + r(y) + \frac{\beta}{2}\|y-x\|^2 \geq f_x(x^+) + r(x^+) + \frac{\beta}{2}\|x^+-x\|^2 \\ + \frac{\alpha_2+\beta}{2}\|y-x^+\|^2. \end{aligned} \tag{6.10}$$

Property $(A1)$ in turn yields the two estimates

$$\begin{aligned} f_x(y) + r(y) &\leq \varphi(y) - \frac{\alpha_1}{2}\|y-x\|^2, \\ f_x(x^+) + r(x^+) &\geq \varphi(x^+) - \frac{\beta}{2}\|x^+-x\|^2, \end{aligned} \tag{6.11}$$

while algebraic manipulations yield the expression

$$\begin{aligned} \|x^+-x\|^2 - \|y-x\|^2 + \|y-x^+\|^2 &= 2\langle y-x^+, x-x^+\rangle \\ &= 2\langle y-x, x-x^+\rangle + 2\|x-x^+\|^2. \end{aligned} \tag{6.12}$$

Combining (6.10),(6.11), and (6.12) completes the proof. □

There are a few useful consequences of Theorem 6.8 worth highlighting, which are summarized in the following corollary.

Corollary 6.9. *Fix a point $x \in \mathbf{E}$, define $\eta = \frac{1}{\beta}$, and set*

$$x^+ = \operatorname*{argmin}_y \; f_x(y) + r(y) + \frac{1}{2\eta}\|y-x\|^2.$$

Then the estimates hold:

$$\varphi(x) \geq \varphi(x^+) + \frac{\eta}{2}\|\mathcal{G}_\eta(x)\|^2, \tag{6.13}$$

$$\langle \mathcal{G}_\eta(x), x-x^*\rangle \geq \frac{\eta}{2}\|\mathcal{G}_\eta(x)\|^2 + \frac{\alpha_1}{2}\|x-x^*\|^2 + \frac{\alpha_2}{2}\|x^+-x^*\|^2, \tag{6.14}$$

$$\|\mathcal{G}_\eta(x)\| \geq \frac{\alpha_1}{2}\|x-x^*\|. \tag{6.15}$$

Proof. Setting $y = x$ in (6.9) immediately yields (6.13). Next, set $y = x^*$ in (6.9). The inequality (6.14) then follows immediately after using the lower bound $\varphi(x^+) \geq \varphi(x^*)$. Dividing the inequality (6.14) by $\|x-x^*\|$ and using the upper bound $\langle \mathcal{G}_\eta(x), x-x^*\rangle \leq \|\mathcal{G}_\eta(x)\| \cdot \|x-x^*\|$ yields (6.15). □

Having proved Corollary 6.9, we can now quickly establish the linear rate of convergence for Algorithm 5 that is analogous to Theorem 5.3.

Theorem 6.10. *Suppose that Assumption 6.3 holds. Then the iterates generated by Algorithm 5 with $\eta_t = \frac{1}{\beta}$ satisfy*

$$\varphi(x_{t+1}) - \varphi^* \leq \left(1 - \frac{\alpha_1}{4\beta}\right)(\varphi(x_t) - \varphi^*), \tag{6.16}$$

$$\|x_{t+1} - x^*\|^2 \leq \left(\frac{\beta - \alpha_1}{\beta + \alpha_2}\right)\|x_t - x^*\|^2. \tag{6.17}$$

Proof. To simplify notation, set $\eta = \frac{1}{\beta}$. We estimate

$$\begin{aligned}\|x_{t+1} - x^*\|^2 &= \|x_t - x^* - \eta\mathcal{G}_\eta(x_t)\|^2 \\ &= \|x_t - x^*\|^2 - 2\eta\langle\mathcal{G}_\eta(x_t), x_t - x^*\rangle + \eta^2\|\mathcal{G}_\eta(x_t)\|^2 \\ &\leq \|x_t - x^*\|^2 - \alpha_1\eta\|x_t - x^*\|^2 - \alpha_2\eta\|x_{t+1} - x^*\|^2,\end{aligned}$$

where the last inequality follows from (6.14). Rearranging yields (6.17).

Next, setting $y = x^*$ in (6.9) and using (6.15), we deduce

$$\begin{aligned}\varphi(x_{t+1}) - \varphi(x^*) &\leq \langle\mathcal{G}_\eta(x_t), x_t - x^*\rangle - \frac{\eta}{2}\|\mathcal{G}_\eta(x_t)\|^2 \\ &\leq \|\mathcal{G}_\eta(x_t)\|^2\left(\frac{\|x_t - x^*\|}{\|\mathcal{G}_\eta(x_t)\|} - \frac{\eta}{2}\right) \\ &\leq \|\mathcal{G}_\eta(x_t)\|^2\left(\frac{2}{\alpha_1} - \frac{\eta}{2}\right).\end{aligned}$$

Combining this estimate with (6.13), we compute

$$\varphi(x_{t+1}) - \varphi(x_t) \leq -\frac{\eta}{2}\|\mathcal{G}_\eta(x_t)\|^2 \leq \frac{1}{1 - \frac{4\beta}{\alpha_1}} \cdot (\varphi(x_{t+1}) - \varphi(x^*)).$$

Adding and subtracting φ^* from the left-side and rearranging yields (6.16), thereby completing the proof. □

6.1.3 Accelerated algorithm

In this section, we analyze an accelerated variant of Algorithm 5 that is in exactly the same spirit as the accelerated gradient method for smooth minimzation. Algorithm 6 summarizes the resulting procedure.

Exercise 6.11. ✍ Suppose $a_0 = 1$ and a_t is given by (6.20) for each index $t \geq 1$.

Algorithm 6: Accelerated proximal method

Input: Starting point $x_0 \in \mathbf{E}$.

Set $t = 0$ and $a_0 = a_{-1} = 1$;

for $t = 0, \dots, T$ **do**

Set

$$u_t = x_t + a_t(a_{t-1}^{-1} - 1)(x_t - x_{t-1}) \tag{6.18}$$

$$x_{t+1} = \operatorname*{argmin}_x \; f_{u_t}(x) + r(x) + \frac{\beta}{2}\|x - u_t\|^2 \tag{6.19}$$

Set

$$a_{t+1} = \frac{\sqrt{a_t^4 + 4a_t^2} - a_t^2}{2} \tag{6.20}$$

$t \leftarrow t + 1$.

end

1. Show that the relation holds:

$$\frac{1 - a_{t+1}}{a_{t+1}^2} = \frac{1}{a_t^2} \qquad \forall t \geq 0. \tag{6.21}$$

2. Using induction, establish $\sum_{i=0}^{t} \frac{1}{a_i} = \frac{1}{a_t^2}$ and $a_t \leq \frac{2}{t+2}$, for each $t \geq 0$.

[**Hint:** In order to prove $\sum_{i=0}^{t} \frac{1}{a_i} = \frac{1}{a_t^2}$, rewrite (6.21) as $\frac{1}{a_{t+1}^2} - \frac{1}{a_t^2} = \frac{1}{a_{t+1}}$ and sum up the equation.]

The following theorem establishes the convergence rate of Algorithm 6. In particular, Theorem 5.4 whose proof was omitted earlier, is a direct consequence. The main idea of the argument is based on the following algebraic trick. Recall that the proof of Theorem 6.7 for the unaccelerated algorithm was based on comparing the value of the function $f_{x_t} + r$ at x_{t+1} with the value at x^* based on strong convexity of the function $f_{x_t} + r + \frac{\beta}{2}\|\cdot - x_t\|^2$. We now use the the interpolated comparison point $a_t x^* + (1 - a_t)x_t$ instead of x^*. The flexibility in choosing $a_t \in (0, 1)$ is a key ingredient that facilitates an accelerated rate of convergence.

Theorem 6.12. *The iterates generated by Algorithm 6 satisfy*

$$\varphi(x_{t+1}) - \varphi(x) + a_t^2 \cdot \sum_{i=0}^{t} \frac{\Delta_f(x_i, x)}{a_i} \leq \frac{a_t^2 \beta \|x_0 - x\|^2}{2} \qquad \forall x \in \mathbf{E},$$

and therefore the estimate holds:

$$\varphi(x_{t+1}) - \varphi^* \leq \frac{2\beta\|x_0 - x^*\|^2}{(t+2)^2},$$

Proof. Let x_t be the iterates generated by Algorithm 6 and define the function $m_t(x) := f_{u_t}(x) + r(x)$ for each index t. Since x_{t+1} is the minimizer of the β-strongly convex function $m_t + \frac{\beta}{2}\|\cdot - u_t\|^2$, we successively estimate

$$\begin{aligned}
\varphi(x_{t+1}) &\leq m_t(x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - u_t\|^2 && (6.22)\\
&\leq m_t(a_t x + (1-a_t)x_t) \\
&\quad + \frac{\beta}{2}\left(\|a_t x + (1-a_t)x_t - u_t\|^2 - \|a_t x + (1-a_t)x_t - x_{t+1}\|^2\right) && (6.23)\\
&\leq a_t m_t(x) + (1-a_t)m_t(x_t) \\
&\quad + \frac{\beta a_t^2}{2}\left(\|x - [x_t - a_t^{-1}(x_t - _t)]\|^2 - \|x - [x_t - a_t^{-1}(x_t - x_{t+1})]\|^2\right) && (6.24)\\
&\leq a_t\varphi(x) + (1-a_t)\varphi(x_t) - a_t\Delta_f(u_t, x) \\
&\quad + \frac{\beta a_t^2}{2}\left(\|x - [x_t - a_t^{-1}(x_t - u_t)]\|^2 - \|x - [x_t - a_t^{-1}(x_t - x_{t+1})]\|^2\right), && (6.25)
\end{aligned}$$

where (6.22) and (6.25) follow from Assumption $(A1)$ and (6.24) uses convexity of m_t. Subtracting $\varphi(x)$ from both sides and dividing by a_t^2 then yields

$$\begin{aligned}
\frac{1}{a_t^2}(\varphi(x_{t+1}) - \varphi(x)) \leq \frac{1-a_t}{a_t^2}(\varphi(x_t) - \varphi(x)) - \frac{\Delta_f(u_t, x)}{a_t} \\
+ \frac{\beta}{2}\Big(\|x - [x_t - a_t^{-1}(x_t - u_t)]\|^2 \\
- \|x - [x_t - a_t^{-1}(x_t - x_{t+1})]\|^2\Big). && (6.26)
\end{aligned}$$

The update rule (6.18) is precisely designed to force telescoping in the last two lines of (6.26). To see this, define an auxiliary sequence $z_t = x_t - a_t^{-1}(x_t - u_t)$. Observe that z_{t+1} then satisfies

$$\begin{aligned}
z_{t+1} = x_{t+1} - a_{t+1}^{-1}(x_{t+1} - u_{t+1}) &= x_{t+1} + (a_t^{-1} - 1)(x_{t+1} - x_t) \\
&= x_t - a_t^{-1}(x_t - x_{t+1}).
\end{aligned}$$

Thus the inequality (6.26) becomes

$$\begin{aligned}\frac{1}{a_t^2}(\varphi(x_{t+1}) - \varphi(x)) + \frac{\beta}{2}\|x - z_{t+1}\|^2 &\leq \frac{1-a_t}{a_t^2}(\varphi(x_t) - \varphi(x)) + \frac{\beta}{2}\|x - z_t\|^2 \\ &\quad - \frac{\Delta_f(u_t, x)}{a_t} \\ &= \frac{1}{a_{t-1}^2}(\varphi(x_t) - \varphi(x)) + \frac{\beta}{2}\|x - z_t\|^2 \\ &\quad - \frac{\Delta_f(u_t, x)}{a_t}\end{aligned}$$

where the last equality follows directly from the definition of a_t in (6.20). Iterating the recurrence yields

$$\frac{1}{a_t^2}(\varphi(x_{t+1}) - \varphi(x)) \leq \frac{1-a_0}{a_0}(\varphi(x_0) - \varphi(x)) + \frac{\beta}{2}\|x - z_0\|^2 - \sum_{i=0}^{t} \frac{\Delta_f(u_i, x)}{a_i}.$$

Taking into account $a_0 - 1 = 0$, $z_0 = x_0 - a_0^{-1}(x_0 - u_0) = u_0$, and $a_t \leq \frac{2}{t+2}$ (Exercise 6.11) completes the proof. □

Recall that one approach to making the accelerated gradient method linearly convergent under a strong convexity assumption was to periodically restart the algorithm. The resulting procedure was summarized in Algorithm 3 with its convergence guarantees developed in Theorem 5.5. Exactly the same strategy boosts the sublinear rate of Algorithm 6 to an accelerated linear rate, whenever φ is strongly convex. We leave the details for the reader.

6.2 Proximal methods based on lower models

In this section, we consider the additive composite problem class (6.1) but weaken the Assumption 6.3. Namely, we only assume the one-sided bound $f_x \leq f$.

Assumption 6.13 (One-sided model)**.** Assume that there exist real $\mu \geq 0$ and $L \geq 0$ such that the following properties hold:

(B1) **(One-sided accuracy)** The estimate holds:

$$f_x(x) = f(x) \quad \text{and} \quad f_x(y) \leq f(y) \qquad \forall x, y \in \operatorname{dom} r.$$

(B2) **(Model convexity)** The functions $f_x(\cdot) + r(\cdot)$ are α-strongly convex for all $x \in \operatorname{dom} r$.

(B3) **(Lipschitz property)** The function $f_x(\cdot)$ is L-Lipschitz continuous on $\operatorname{dom} r$ for every $x \in \operatorname{dom} r$.

Throughout the section, we suppose that Assumption 6.13 holds. Before establishing convergence guarantees of Algorithm 5, let us look at two examples.

Example 6.14 (Proximal subgradient method)**.** As the first example, suppose that r is closed and convex, and f is convex and L-Lipschitz continuous on a neighborhood of $\operatorname{dom} r$. The *proximal subgradient method* then simply uses the affine models

$$f_x(y) = f(x) + \langle v_x, y - x \rangle,$$

for some vectors $v \in \partial f(x)$. Each step of the algorithm then takes the form

$$x_{t+1} = \operatorname{prox}_{\eta_t r}\left(x_t - \eta_t v_t\right),$$

for some vectors $v_t \in \partial f(x_t)$. If r is strongly convex, we may set α to be its strong convexity constant.

Example 6.15 (Clipped subgradient method)**.** Suppose again that we are in the setting of Example 6.14, but we also have available a lower-bound ℓ^* on the optimal value f^*. Then we can use the tighter models

$$f_x(y) = \max\{f(x) + \langle v, y - x \rangle, \ell\},$$

for some $v \in \partial f(x)$. The update of Algorithm 5 then takes the form

$$x_{t+1} = \operatorname*{argmin}_x \ \max\{f(x_t) + \langle v_t, x - x_t \rangle, \ell\} + r(x) + \frac{1}{2\eta_t}\|x - x_t\|^2.$$

In particular, in the case $r = 0$ a quick computation (do it!) yields the explicit update

$$x_{t+1} = x_t - \min\left\{\tfrac{f(x_t) - \ell}{\|v_t\|^2}, \eta_t\right\} \cdot v_t.$$

Not surprisingly, we will see that Algorithm 5 exhibits similar convergence guarantees as the basic subgradient method for minimizing Lipschitz convex functions. Before passing to the convergence analysis, we record the following simple lemma.

Lemma 6.16. *The function f is L-Lipschitz continuous on* $\operatorname{dom} r$.

Proof. Assumption $(B1)$ guarantees $f(x)-f(y) \leq f_x(x)-f_x(y) \leq L\|x-y\|$, for all $x, y \in \operatorname{dom} r$. Switching the roles of x and y completes the proof. □

Recall that the convergence guarantees for the subgradient method fundamentally relied on the relation

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - 2\eta_t(f(x_t) - f^*) + \eta_t^2 L^2,$$

where f is the Lipschitz convex function to be minimized. The following lemma establishes an analogous guarantee for Algorithm 5, from which convergence guarantees will follow quickly.

Lemma 6.17. *For every index $t \geq 0$ and $x \in \operatorname{dom} r$, the inequality holds:*

$$(1 + \eta_t\alpha)\|x_{t+1} - x\|^2 \leq \|x_t - x\|^2 - 2\eta_t(\varphi(x_{t+1}) - \varphi(x)) + 2\eta_t^2 L^2. \quad (6.27)$$

Proof. Taking into account that x_{t+1} is the minimizer of the $(\alpha + \eta_t^{-1})$-strongly convex function $f_{x_t} + r + \frac{1}{2\eta}\|\cdot - x_t\|^2$, we deduce

$$\begin{aligned}
&\frac{\alpha + \eta_t^{-1}}{2}\|x - x_{t+1}\|^2 + \frac{1}{2\eta_t}\|x_{t+1} - x_t\|^2 - \frac{1}{2\eta}\|x - x_t\|^2 \\
&\leq r(x) + f_{x_t}(x) - r(x_{t+1}) - f_{x_t}(x_{t+1}) \\
&\leq r(x) + f_{x_t}(x) - r(x_{t+1}) - f_{x_t}(x_t) + L\|x_{t+1} - x_t\| \qquad (6.28) \\
&\leq r(x) + f(x) - r(x_{t+1}) - f(x_t) + L\|x_{t+1} - x_t\| \qquad (6.29) \\
&\leq r(x) + f(x) - r(x_{t+1}) - f(x_{t+1}) + 2L\|x_{t+1} - x_t\| \qquad (6.30)
\end{aligned}$$

where (6.28) follows from Assumption (B3), inequality (6.29) follows from (B1), and (6.30) follows from Lemma 6.16.

Define $\delta_t := \|x_{t+1} - x_t\|$. Rearranging (6.30) and multiplying through by $2\eta_t$, we immediately deduce

$$\begin{aligned}
(1 + \eta_t\alpha)\|x - x_{t+1}\|^2 &\leq \|x - x_t\|^2 - 2\eta_t(\varphi(x_{t+1}) - \varphi(x)) + 4L\delta_t\eta_t - \delta_t^2 \\
&\leq \|x - x_t\|^2 - 2\eta_t(\varphi(x_{t+1}) - \varphi(x)) + \max_{\delta \in \mathbf{R}}\{4L\delta\eta_t - \delta^2\} \\
&= \|x - x_t\|^2 - 2\eta_t(\varphi(x_{t+1}) - \varphi(x)) + 4L^2\eta_t^2.
\end{aligned}$$

The proof is complete. □

Convergence guarantees of Algorithm 5 now follow quickly, both in the convex and strongly convex settings.

Theorem 6.18 (Convergence rate)**.** *The iterates generated by Algorithm 5 satisfy*

$$\varphi\left(\tfrac{1}{\sum_{t=0}^{T}\eta_t}\sum_{t=0}^{T}\eta_t x_{t+1}\right)-\varphi(x)\leq\frac{\frac{1}{2}\|x_0-x\|^2+\mathsf{L}^2\sum_{t=0}^{T}\eta_t^2}{\sum_{t=0}^{T}\eta_t},\tag{6.31}$$

for any point $x\in\operatorname{dom} r$. In particular, if Algorithm 5 uses the constant parameter $\eta_t=\frac{R}{L\sqrt{2(T+1)}}$, for some real $R>\|x_0-x^*\|$, the estimate holds:

$$\varphi\left(\tfrac{1}{T+1}\sum_{t=1}^{T+1}x_t\right)-\varphi(x^*)\leq\frac{\sqrt{2}LR}{\sqrt{T+1}}.\tag{6.32}$$

Proof. Lemma 6.17 guarantees

$$2\eta_t(\varphi(x_{t+1})-\varphi(x^*))\leq\|x_t-x\|^2-\|x_{t+1}-x^*\|^2+2L^2\eta_t^2.$$

The estimate (6.31) then follows by summing across $t=0,\dots,T$, dividing through by $\sum_{t=0}^{T}\eta_t$, and using convexity of φ. The estimate (6.32) is immediate from (6.31). □

Theorem 6.19 (Convergence rate under strong convexity)**.** *The iterates generated by Algorithm 5 with $\eta_t=\frac{2}{\alpha(t+1)}$ satisfy*

$$\varphi\left(\tfrac{2}{(T+2)(T+3)-2}\sum_{t=1}^{T+1}(t+1)x_t\right)-\varphi(x)\leq\frac{\alpha\|x_0-x\|^2}{(T+2)^2}+\frac{8\mathsf{L}^2}{\alpha(T+2)}.$$

for any point $x\in\operatorname{dom} r$.

Proof. For each index t, define $\Delta_t:=\frac{1}{2}\|x-x_t\|^2$. Lemma 6.17 yields

$$\varphi(x_{t+1})-\varphi(x)\leq\frac{1}{\eta_t}\Delta_t-\frac{1+\eta_t\alpha}{\eta_t}\Delta_{t+1}+L^2\eta_t.$$

Plugging in $\eta_t:=\frac{2}{\alpha(t+1)}$, multiplying through by $t+2$, and summing, we get

$$\begin{aligned}\sum_{t=0}^{T}(t+2)(\varphi(x_{t+1})-\varphi(x))&\leq\sum_{t=0}^{T}\left(\tfrac{\alpha(t+1)(t+2)}{2}\Delta_t-\tfrac{\alpha(t+2)(t+3)}{2}\Delta_{t+1}\right)+\sum_{t=0}^{T}\tfrac{2L^2(t+2)}{\alpha(t+1)}\\&\leq\alpha\Delta_0+\frac{4L^2(T+1)}{\alpha}.\end{aligned}$$

Dividing through by the sum $\sum_{t=0}^{T}(t+2) = \frac{(T+2)(T+3)}{2} - 1$ and using convexity of φ, we conclude

$$\begin{aligned}\varphi\left(\tfrac{2}{(T+2)(T+3)-2}\sum_{t=0}^{T}(t+2)x_{t+1}\right) - \varphi(x) &\leq \tfrac{\alpha\|x_0-x\|^2}{(T+2)(T+3)-2} + \tfrac{8L^2(T+1)}{\alpha((T+2)(T+3)-2)} \\ &\leq \frac{\alpha\|x_0-x\|^2}{(T+2)^2} + \frac{8L^2}{\alpha(T+2)}, \end{aligned} \tag{6.33}$$

where (6.33) uses the estimate $(T+2)(T+3)-2 \geq (T+2)^2$. The proof is complete. □

References. Most of the material in Section 6.1 is a direct generalization of the analogous results in Beck and Teboulle [5] and Nesterov [30] for the (accelerated) proximal gradient method. Namely, the proofs of Theorems 6.7 and 6.12 appear in [5]. The proof of the estimate (6.17) appears in [30]. The short proof of (6.16) follows the same ideas as put forth by Luo-Tseng [24] when r is an indicator function of a closed convex set. Section 7.1 follows the discussion in Tseng [38], though similar techniques appear also Lan-Lu-Manteiro [19], Lu [23], and Nesterov [29]. The material in Section 6.2 follows Davis-Drusvyatskiy [13].

Chapter 7

Smoothing and primal-dual algorithms

In this section, we develop specialized algorithms for problems of the form

$$\min_x \ \varphi(x) = h(\mathcal{A}x) + g(x),$$

where g and h are "simple" convex functions.

7.1 Proximal (accelerated) gradient method solves the dual

A remarkable feature of the proximal gradient method and its accelerated variant is that both algorithms not only solve the target problem (6.1) but also its dual! This section explains this phenomenon. Throughout, we make the following assumptions.

Assumption 7.1. We assume the following are true for some $\beta, R \geq 0$.

1. **(Smooth+simple)** Assume that f is convex and β-smooth, while r is proper, closed, and convex and satisfies $\sup_{x,z \in \operatorname{dom} r} \|x - z\| \leq R$. We set

$$f_x(y) = f(x) + \langle \nabla f(x), y - x \rangle \qquad \forall x, y \in \mathbf{E}.$$

2. **(Dual representation)** f admits a "dual description" as

$$f(x) = \max_{y \in \mathbf{Y}} \ g(x, y)$$

for some function $g\colon \mathbf{E} \times \mathbf{Y} \to \overline{\mathbf{R}}$ that is convex in $x \in \mathbf{E}$ and concave in $y \in \mathbf{Y}$. We suppose moreover that for every $x \in \mathbf{E}$, there exists

$$y \in \operatorname{argmax} g(x, \cdot) \quad \text{satisfying} \quad \partial_x g(x, y) \neq \emptyset. \tag{7.1}$$

With this setup, define the *dual objective function*

$$\psi(y) = \inf_{x \in \mathbf{E}} \{g(x, y) + r(x)\}.$$

The smooth plus simple assumption is by now well familiar to the reader. In particular, Algorithm 5 becomes the proximal gradient method, while Algorithm 6 is the accelerated proximal gradient algorithm. To better internalize the existence of the dual representation, let us look at the most important example of the Fenchel-Rockafellar duality.

Suppose that we are interested in the optimization problem in the form:

$$(P) \qquad \min_x \ \varphi(x) = h(\mathcal{A}x) + g(x),$$

for some γ-smooth function $h\colon \mathbf{Y} \to \mathbf{R}$, a linear map $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$, and a proper, closed convex function $r\colon \mathbf{E} \to \overline{\mathbf{R}}$. That is, we aim to solve the additive composite problem (6.1) under the identifications $f = h \circ \mathcal{A}$ and $r = g$. A quick computation shows that we may set the smoothness parameter of f as $\beta = \|\mathcal{A}\|_{\text{op}}^2 \gamma$. We may then express f in the conjugate form

$$f(x) = \sup_{y \in \mathbf{Y}} \ \langle \mathcal{A}x, y \rangle - h^\star(y). \tag{7.2}$$

Observe that the function $g(x, y) := \langle \mathcal{A}x, y \rangle - h^\star(y)$ is indeed convex in x and concave in y. Moreover, (7.1) holds automatically. Indeed, since h is γ-smooth, the conjugate $h^\star$ is strongly convex (Theorem 3.66) and therefore for every $x \in \mathbf{E}$ the supremum in (7.2) is attained at a unique point y. Moreover, differentiating $g(x, \cdot)$ in y and using Corollary 3.40 yields the equivalences:

$$y \in \operatorname{argmin} g(x, \cdot) \quad \iff \quad \mathcal{A}x \in \partial h^\star(y) \quad \iff \quad y = \nabla h(\mathcal{A}x).$$

The reader should verify that the dual objective ψ is then simply the objective function in the Fenchel-Rockafellar dual problem

$$(D) \qquad \max_y \ \psi(y) = -h^\star(y) - g^\star(-\mathcal{A}y).$$

The flexibility of the dual representation in Assumption 7.1 is convenient, since it is not directly tied to conjugacy. For the rest of the section, we suppose that Assumption 7.1 holds. We will need the following intuitive lemma expressing the gradient $\nabla f(x)$ in terms of the gradient of g.

Lemma 7.2. *Fix two points $x \in \mathbf{E}$ and $y \in \mathbf{Y}$ satisfying (7.1). Then $g(\cdot, y)$ is differentiable at x and equality $\{\nabla f(x)\} = \nabla_x g(x, y)$ holds.*

Proof. Fix any subgradient $v \in \partial_x g(x, y)$. Then for any point $z \in \mathbf{E}$, we compute

$$f(z) = \sup g(z, \cdot) \geq g(z, y) \geq g(x, y) + \langle v, z - x\rangle = f(x) + \langle v, z - x\rangle.$$

Thus v is a subgradient of f at x. Since f is differentiable at x, we conclude that $\partial_x f(x, y)$ is a singleton, and therefore $g(\cdot, y)$ is differentiable at x by Exercise 3.54. □

We are now ready to show that the proximal gradient method and its accelerated variant automatically produce iterates that approximately solve the dual problem $\sup_y \psi(y)$. In a nutshell, if x_t are the iterates generates by the two methods, then the running average of the corresponding dual points y_t approximately solves the dual. In particular, within the Fenchel-Rockafellar framework, the dual iterates are the simply the running average of the gradient $\nabla h(\mathcal{A}x_t)$. The main idea of the argument is to relate the gap function $\Delta_f(x, z)$ to the dual objective ψ. The following lemma provides the first indication that the two are closely related.

Lemma 7.3. *Fix two point $x \in \mathbf{E}$ and $y \in \mathbf{Y}$ satisfying (7.1). Then the estimate holds:*

$$\Delta_f(x, z) \geq f(z) - g(z, y) \qquad \text{for all } z \in \mathbf{E}.$$

Proof. Taking into account convexity of $g(\cdot, y)$, we compute

$$\begin{aligned} \Delta_f(x, z) &= f(z) - f(x) - \langle \nabla f(x), z - x\rangle \\ &= f(z) - g(x, y) - \langle \nabla_x g(x, y), z - x\rangle \\ &\geq f(z) - g(z, y), \end{aligned} \tag{7.3}$$

as claimed. □

The following two main results of the section now follow quickly by combining the convergence guarantees of the (accelerated) proximal gradient method (Theorems 6.7 and 6.12) with Lemma 7.3.

Theorem 7.4 (Dual convergence of prox-gradient)**.** *Let $\{x_t\}$ be the iterates generated by Algorithm 5 with $\eta_t = \frac{1}{\beta}$. Define the dual sequence $y_t \in \operatorname{argmax}_y g(x_t, y)$ satisfying $\partial_x g(x_t, y_t) \neq \emptyset$ and define the running average $\bar{y}_t = \frac{1}{t}\sum_{i=0}^{t} y_t$. Then the estimate holds:*

$$\varphi(x_t) - \psi(\bar{y}_{t-1}) \leq \frac{\beta R^2}{2t}.$$

Proof. Using Theorem 6.7 and the Lemma 7.3 we deduce

$$\begin{aligned}
\frac{\beta\|x_0 - x\|^2}{2t} &\geq \varphi(x_t) - \varphi(x) + \frac{1}{t}\sum_{i=0}^{t-1} \Delta_f(x_i, x) \\
&\geq \varphi(x_t) - \varphi(x) + \frac{1}{t}\sum_{i=0}^{t-1} (f(x) - g(x, y_i)) \\
&\geq \varphi(x_t) - r(x) - g\left(x, \bar{y}_{t-1}\right),
\end{aligned}$$

where the last inequality follows from concavity of $g(x, \cdot)$. Taking the supremum over $x \in \operatorname{dom} r$ completes the proof. □

Theorem 7.5 (Dual convergence of accelerated prox-gradient)**.** *Let $\{x_t\}$ and $\{u_t\}$ be the iterates generated by Algorithm 6. Define the dual sequence $y_t \in \operatorname{argmax}_y g(u_t, y)$ satisfying $\partial_x g(u_t, y_t) \neq \emptyset$ and define the running average $\bar{y}_t = \sum_{i=0}^{t} \frac{a_i^2}{a_i} y_t$. Then the estimate holds:*

$$\varphi(x_{t+1}) - \psi(\bar{y}_t) \leq \frac{2\beta R^2}{(t+2)^2}.$$

Proof. Using Theorem 6.12 and Lemma 7.3 we deduce

$$\begin{aligned}
\frac{\beta\|x_0 - x\|^2}{2} &\geq \frac{1}{a_t^2}(\varphi(x_{t+1}) - \varphi(x)) + \sum_{i=0}^{t} \frac{\Delta_f(u_i, x)}{a_i} \\
&\geq \frac{1}{a_t^2}(\varphi(x_{t+1}) - \varphi(x)) + \sum_{i=0}^{t} \frac{f(x) - g(x, y_i))}{a_i} \\
&= \frac{1}{a_t^2}(\varphi(x_{t+1}) - \varphi(x)) + \left(\sum_{i=0}^{t} \frac{1}{a_i}\right) f(x) - \sum_{i=0}^{t} \frac{g(x, y_i)}{a_i} \\
&\geq \frac{1}{a_t^2}(\varphi(x_{t+1}) - \varphi(x)) + \left(\sum_{i=0}^{t} \frac{1}{a_i}\right) f(x) - \left(\sum_{i=0}^{t} \frac{1}{a_i}\right) \cdot g(x, \bar{y}_t) \qquad (7.4) \\
&\geq \frac{1}{a_t^2}(\varphi(x_{t+1}) - r(x) - g(x, \bar{y}_t)) \qquad (7.5)
\end{aligned}$$

where (7.4) follows from concavity of $g(x, \cdot)$ and (7.5) follows from Lemma 6.11. Taking the supremum over $x \in \operatorname{dom} r$ completes the proof. □

7.2 Smoothing technique

In this section, we consider the optimization problem

$$\min_{x \in \mathbf{E}} \ \varphi(x) = h(\mathcal{A}x) + g(x),$$

under the following assumptions:

1. $h \colon \mathbf{Y} \to \mathbf{R}$ is a L-Lipschitz continuous convex function,
2. $g \colon \mathbf{E} \to \mathbf{R}$ is a proper, closed, and convex function with a computable proximal map,
3. $\mathcal{A} \colon \mathbf{E} \to \mathbf{Y}$ is a linear map.

Notice that the function $f(x) = h(\mathcal{A}x)$ is Lipschitz continuous with constant $L\|\mathcal{A}\|_{\mathrm{op}}$. In particular, the only method we have available for this problem is the proximal subgradient method, which will find a point x satisfying $\varphi(x) - \varphi^* \leq \epsilon$ after $\mathcal{O}\left(\frac{L^2\|\mathcal{A}\|_{op}^2 R^2}{\epsilon^2}\right)$ iterations, where $R \geq \|x_0 - x^*\|$ is a known upper bound. Each iteration t of the method consists of a single proximal iteration $\operatorname{prox}_{\eta_t g}$, an evaluation of a subgradient $v \in \partial h(\mathcal{A}x)$, and formation of the vector $\mathcal{A}^* v \in \partial f(x_t)$. In typical applications, the cost of

computing the proximal map $\mathrm{prox}_{\eta t g}$ and evaluating a subgradient v of h is negligible compared to the two matrix-vector multiplications $\mathcal{A}x$ and $\mathcal{A}^* v$. We will now see that if h is sufficiently simple (e.g. with a computable proximal map), then there are algorithms for the target problem that require much fewer matrix-vector multiplications.

An appealing strategy, which we now describe, is to simply replace h by a smooth approximation $\widetilde{h}$ and then apply an accelerated proximal gradient method. There are two important parameters that will determine the overall efficiency: the error in approximation $\sup_{x \in \mathrm{dom}\, r} h(x) - \widetilde{h}$ and the Lipschitz constant of the gradient $\nabla \widetilde{h}$. One appealing choice for the smoothing function is the Moreau envelope h_ν.

Lemma 7.6. *The gradient ∇h_η is $\frac{1}{\eta}$-Lipschitz continuous and the estimate holds:*

$$0 \leq h(x) - h_\eta(x) \leq \frac{\eta L^2}{2}. \tag{7.6}$$

Proof. The fact that ∇h_ν is $\frac{1}{\eta}$-Lipschitz continuous was already proved in Theorem 3.64. Next, observe that since h is L-Lipschitz, the norm of every vector $y \in \mathrm{dom}\, h^\star$ is bounded by L (Exercise 3.52). We therefore deduce

$$\begin{aligned} h_\eta(x) &= \left(h \,\square\, \tfrac{1}{2\eta}\|\cdot\|^2\right)^{\star\star}(x) \\ &= \left(h^\star + \tfrac{\eta}{2}\|\cdot\|^2\right)^\star(x) \qquad (7.7) \\ &= \sup_y \; \left\{\langle y, x\rangle - h^\star(y) - \tfrac{\eta}{2}\|y\|^2\right\} \\ &\geq \sup_y \; \{\langle y, x\rangle - h^\star(y)\} - \frac{\eta L^2}{2} = h(x) - \frac{\eta L^2}{2}, \end{aligned}$$

where (7.7) follows from Table 3.4 (row four). The proof is complete. □

In light of Lemma 7.6, we see that the accuracy in approximation $h - h_\eta$ scales as $\sim \eta$, while the Lipschitz constant of the gradient of ∇h_η scales as $\sim \eta^{-1}$. Using Theorem 6.12, we deduce that the accelerated proximal gradient method on the smoothed problem

$$\min_{x \in \mathbf{E}} \; \widetilde{\varphi}(x) := h_\eta(\mathcal{A}x) + g(x),$$

will generate a sequence x_t satisfying

$$\widetilde{\varphi}(x_{t+1}) - \widetilde{\varphi}(x^*) \leq \frac{2\|\mathcal{A}\|_{\mathrm{op}}^2 \|x_0 - x^*\|^2}{\eta(t+2)^2},$$

where x^* is the minimizer of φ. Therefore, we deduce

$$\begin{aligned}\varphi(x_{t+1}) - \varphi(x^*) &\leq \widetilde{\varphi}(x_{t+1}) - \widetilde{\varphi}(x^*) + \frac{\eta L^2}{2} \\ &\leq \frac{2\|\mathcal{A}\|_{\mathrm{op}}^2 \|x_0 - x^*\|^2}{\eta(t+2)^2} + \frac{\eta L^2}{2}.\end{aligned}$$

The first term on the right side is the optimization error, while the second term is the approximation error. Given a target accuracy $\epsilon > 0$, we may set each of the terms to be $\frac{\epsilon}{2}$ yielding the parameter settings

$$\eta = \frac{\epsilon}{L^2} \qquad \text{and} \qquad t = 1 + \left\lceil \frac{2L\|\mathcal{A}\|_{\mathrm{op}}\|x_0 - x^*\|}{\epsilon} \right\rceil.$$

With this parameter choices, the iterate x_{t+1} generated by the accelerated proximal subgradient method satisfies $\varphi(x_{t+1}) - \varphi(x^*) \leq \epsilon$.

In summary, if we assume that the cost of evaluating the proximal maps of h and g are negligible compared to the cost of applying the linear map $\mathcal{A}$, then the efficiency estimate $\mathcal{O}\left(\frac{L\|\mathcal{A}\|_{\mathrm{op}}\|x_0 - x^*\|}{\epsilon}\right)$ of the outlined smoothing-based method is much better than the guarantee $\mathcal{O}\left(\frac{L^2\|\mathcal{A}\|_{op}^2\|x_0 - x^*\|^2}{\epsilon^2}\right)$ for the proximal subgradient method.

The reader should note that the only properties of the Moreau envelope h_η that were used so far are summarized by Lemma, namely that the accuracy in approximation $h - h_\eta$ and the Lipschitz constant of the gradient of ∇h_η are inversely proportional. Any smooth approximation of h satisfying this inverse relationship can be used instead of the Moreau envelope, yielding efficiency guarantees that still scale as $1/\epsilon$.

7.3 Proximal point method

Consider the problem dual pair

$$\begin{aligned}(P) \qquad & \min_x \; h(\mathcal{A}x) + g(x) \\ (D) \qquad & \max_y \; -g^\star(-\mathcal{A}^* y) - h^\star(y).\end{aligned} \tag{7.8}$$

where $g\colon \mathbf{E} \to \overline{\mathbf{R}}$ and $h\colon \mathbf{Y} \to \overline{\mathbf{R}}$ are proper, closed, convex functions, and $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$ is a linear map. We moreover assume that the individual proximal operators $\mathrm{prox}_{tf}(x)$ and $\mathrm{prox}_{tg}(x)$ are easily computable. Corollary 4.11 showed that under mild conditions, x is the minimizer of (P) and y is the maximizer of (D) if and only if the inclusion holds:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} 0 & \mathcal{A}^* \\ -\mathcal{A} & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \partial g(x) \times \partial h^\star(y). \tag{7.9}$$

In this section we design algorithm that explicitly attempt to solve the system (7.9). The main observation we will use is that the right side of (7.9) is a maximal monotone operator, as was shown in Exercise 3.76. To this end, let us abstract away from (7.9) and instead consider the task of solving the system

$$0 \in T(x) \tag{7.10}$$

where $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ is a maximal-monotone operator. A conceptual algorithm for this problem is the *proximal point method*

$$x_{t+1} = \mathcal{R}_T(x_t) \qquad \forall t \geq 0.$$

Notice that the proximal point method is not directly implementable because the resolvant $\mathcal{R}_T$ is in general difficult to compute. Nonetheless, the proximal point method will motivate a number of interesting implementable algorithm, which can be thought of as approximations of the basic proximal point method.

Convergence guarantees for the proximal point method follow quickly from the basic properties of the resolvant established in Theorem 3.78 and the following elementary exercise.

Theorem 7.7 (Proximal point method)**.** *Consider a maximal monotone operator $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ and suppose that there exists at least one point x satisfying $0 \in T(x)$. Then the iterates generated by the proximal point method:*

$$x_{t+1} = \mathcal{R}_T(t)$$

converge to some point x^* satisfying $0 \in T(x^*)$. Moreover, the residuals satisfy

$$\min_{i=0,\dots,t} \|\mathcal{R}_T(x_i) - x_i\| \leq \sqrt{\frac{\|x_0 - x^*\|^2}{t+1}}.$$

Proof. As the first step, observe the equivalences

$$x = \mathcal{R}_T(x) \qquad \iff \qquad (I+T)(x) = x \qquad \iff \qquad 0 \in T(x).$$

Thus x is a fixed point of the resolvant $\mathcal{R}_T$ if and only if it satisfies the desired inclusion $0 \in T(x)$. To simplify notation, set $S = \mathcal{R}_T$. Algebraic manipulations show that the estimate (3.36) can be equivalently written as

$$\|S(x) - S(y)\|^2 + \|(I-S)x - (I-S)y\|^2 \leq \|x-y\|^2 \qquad \forall x, y \in \mathbf{E}.$$

Setting $y = x_t$ and $x = x^*$, we deduce

$$\begin{aligned}\|x_{t+1} - x^*\|^2 = \|S(x_t) - S(x^*)\|^2 &\leq \|x_t - x^*\|^2 - \|(I-S)x_t - (I-S)x^*\|^2 \\ &= \|x_t - x^*\|^2 - \|S(x_t) - x_t\|^2.\end{aligned}$$

Iterating the recursion and lower bounding the left-side by zero yields

$$\sum_{i=0}^{t} \|S(x_i) - x_i\|^2 \leq \|x_0 - x^*\|^2. \tag{7.11}$$

Dividing by $t+1$, we conclude

$$\min_{i=0,\ldots,t} \|S(x_i) - x_i\|^2 \leq \frac{1}{t+1} \sum_{i=0}^{t} \|S(x_i) - x_i\|^2 \leq \frac{\|x_0 - x^*\|^2}{t+1}, \tag{7.12}$$

where the first inequality uses that the minimum of finitely many positive numbers is smaller than their average. Thus (7.7) holds. Next, observe from (7.11) that the the sequence $\{\|S(x_i) - x_i\|^2\}_{i\geq 0}$ is summable and therefore tends to zero. Using continuity of S, we deduce that every limit point of $\{x_t\}_{t\geq 0}$ is a fixed point of S. Convergence of the entire sequence $\{x_t\}_{t\geq 0}$ follows from Exercise 7.8, whose verification we leave for the reader. □

Exercise 7.8 (Fejér monotone)**.** Consider a sequence of points $\{x_t\}_{t\geq 0}$ and a set $\mathcal{F} \subset \mathbf{E}$ such that if the inequality

$$\|x_{t+1} - x\| \leq \|x_t - x\|,$$

holds for all $x \in \mathcal{F}$ and all indices $t \geq 0$. Show that if a limit point of the sequence $\{x_t\}_{\geq 0}$ lies in $\mathcal{F}$, then the entire sequence must converge.

Looking back at the proof of Theorem 7.7, it is evident that only two properties of the resolvant were used. Indeed, an identical result is valid if $T\colon \mathbf{E} \rightrightarrows \mathbf{E}$ is an arbitrary set-valued map and $\mathcal{R}_T$ is replaced by any map $S\colon \mathbf{E} \to \mathbf{E}$ satisfying the two properties:

(*i*) the set of fixed points of S coincides with $T^{-1}(0)$,

(*ii*) S is *firmly nonexpansiveness*, meaning

$$\|S(x) - S(y)\|^2 \leq \langle S(x) - S(y), x - y\rangle \qquad \forall x, y \in \mathbf{E}.$$

As observed previously, the proximal point method is a conceptual algorithm because the resolvant $\mathcal{R}_T$, in general, can not be evaluated in closed form, even for the well-structured system (7.9). We now present two algorithms that can be thought of as an implementable modification or approximation of the proximal point method, specifically tailored to the primal-dual system (7.9).

7.3.1 Proximal point method for saddle point problems

Let $\Phi: : \mathbf{E} \times \mathbf{Y} \to \mathbf{R}$ be a function and let $\mathcal{X} \subset \mathbf{E}$ and $\mathcal{Y} \subset \mathbf{Y}$ be closed convex sets. Suppose that $\Phi(\cdot, y)$ is convex and $\Phi(x, \cdot)$ is concave for all $x \in \mathcal{X}$ and $y \in \mathbf{Y}$. Consider the saddle point problem

$$\min_x \max_y \ \Phi(x, y).$$

Define the primal and dual objective values

$$\varphi(x) = \sup_y \ \Phi(x, y) \qquad \text{and} \qquad \psi(y) = \inf_x \ \Phi(x, y).$$

Then a pair (x^*, y^*) is a saddle point of the problem if and only if

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} \partial_x \Phi(x^*, y^*) \\ -\partial_y \Phi(x^*, y^*) \end{bmatrix}.$$

The proximal point method for this inclusion simply iterates

$$\eta \begin{bmatrix} x_t \\ y_t \end{bmatrix} \in \begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} + \begin{bmatrix} \partial_x \Phi(x_{t+1}, y_{t+1}) \\ -\partial_y \Phi(x_{t+1}, y_{t+1}) \end{bmatrix}.$$

Equivalently, (x_{t+1}, y_{t+1}) is the saddle-point of the regularized problem

$$\min_x \max_y \ \Phi(x, y) + \frac{1}{2\eta}\|x - x_t\|^2 - \frac{1}{2\eta}\|y - y_t\|^2.$$

Theorem 7.9. *Suppose that there exists $R > 0$ such that*

$$\|z_1 - z_2\| \le R \qquad \forall z_1, z_2 \in \mathcal{X} \times \mathcal{Y}.$$

Let x_t and y_t be the the iterates produced by the proximal point method and define the averages $\bar{x}_t = \frac{1}{t}\sum_{i=0}^t x_t$ and $\bar{y}_t = \frac{1}{t}\sum_{i=0}^t y_t$. Then the estimate holds:

$$\varphi(\bar{x}_t) - \psi(\bar{y}_t) \le \frac{R^2}{2(t+1)}.$$

Proof. Since x_{t+1} is the minimizer of the η^{-1}-strongly convex function $\Phi(\cdot, y_{t+1}) + \delta_{\mathcal{X}} + \frac{1}{2\eta}\|\cdot - x_t\|^2$, we deduce

$$\Phi(x_{t+1}, y_{t+1}) + \frac{1}{2\eta}\|x_{t+1} - x_t\|^2 \le \Phi(x, y_{t+1}) + \frac{1}{2\eta}\|x - x_t\|^2 - \frac{1}{2\eta}\|x_{t+1} - x\|^2,$$

for all $x \in \mathcal{X}$. Similarly, since y_{t+1} is the maximizer of the η^{-1}-strongly concave function $\Phi(x_{t+1}, y) + \delta_{\mathcal{Y}} - \frac{1}{2\eta}\|\cdot - y_t\|^2$, we deduce

$$\Phi(x_{t+1}, y_{t+1}) - \frac{1}{2\eta}\|y_{t+1} - y_t\|^2 \geq \Phi(x_{t+1}, y) - \frac{1}{2\eta}\|y - y_t\|^2 + \frac{1}{2\eta}\|y_{t+1} - y\|^2,$$

for all $y \in \mathcal{Y}$. Adding the two estimates yields

$$\Phi(x_{t+1}, y) - \Phi(x, y_{t+1}) \leq \frac{1}{2\eta}\left(\|z - z_t\|^2 - \|z - z_{t+1}\|^2 - \|z_{t+1} - z_t\|^2\right).$$

Upper bounding $-\|z_{t+1} - z_t\|^2$ by zero and summing across the iterations, the right side telescopes and we deduce

$$\sum_{i=0}^{t+1} \Phi(x_i, y) - \sum_{i=0}^{t+1} \Phi(x, y_{i+1}) \leq \frac{1}{2\eta}\left(\|z - z_0\|^2 - \|z - z_{t+1}\|^2\right).$$

Dividing through by $t + 1$ and using that Φ is convex-concave, we conclude

$$\Phi(\bar{x}_{t+1}, y) - \Phi(x, \bar{y}_{t+1}) \leq \frac{\|z - z_0\|^2}{2\eta(t+2)}.$$

Finally taking the supremum over x and y completes the proof. □

7.4 Preconditioned proximal point method

To motivate the first approach, consider solving an inclusion

$$0 \in T(x), \tag{7.13}$$

for some maximal monotone operator $T\colon \mathbf{R} \rightrightarrows \mathbf{E}$, whose resolvant $\mathcal{R}_T$ may be difficult to compute. Let $\mathcal{D}\colon \mathbf{E} \to \mathbf{E}$ be a positive definite linear operator. Then clearly, the inclusion (7.13) is equivalent to

$$0 \in (\mathcal{D}^{-1} \circ T)(x).$$

Moreover the operator $(\mathcal{D}^{-1} \circ T)$ is maximal monotone in the modified inner product $\langle x, y\rangle_{\mathcal{D}} = \langle \mathcal{D}x, y\rangle$. The possible advantage of such a reformulation is that in concrete circumstances, it may be possible to choose the "preconditioner" $\mathcal{D}$ so that the resolvant of the operator $\mathcal{D}^{-1} \circ T$ is easily computable.

This is indeed the case for the operator T corresponding to the primal-dual system (7.9). To see this, choose real numbers $t, s > 0$ such that the linear operator

$$\mathcal{D} := \begin{bmatrix} \frac{1}{\tau} I & -\mathcal{A}^* \\ -\mathcal{A} & \frac{1}{s} I \end{bmatrix} \tag{7.14}$$

is positive definite.

Exercise 7.10. Show that the linear operator $\mathcal{D}$ in (7.14) is positive definite as long as $\tau s\|\mathcal{A}\|_{op}^2 < 1$

Computing the resolvant of $\mathcal{D}^{-1}\circ T$ at (x, y) amounts to finding $(x^+, y^+) \in (I + \mathcal{D}^{-1} \circ T)^{-1}(x, y)$, or equivalently

$$\begin{bmatrix} \frac{1}{\tau}I & -A^* \\ -A & \frac{1}{s}I \end{bmatrix} \begin{bmatrix} x - x^+ \\ y - y^+ \end{bmatrix} \in \begin{bmatrix} 0 & \mathcal{A}^* \\ -\mathcal{A} & 0 \end{bmatrix} \begin{bmatrix} x^+ \\ y^+ \end{bmatrix} + \partial g(x^+) \times \partial h^\star(y^+).$$

Rearranging the inclusion yields the system

$$\begin{aligned} x - \tau\mathcal{A}^* y &\in x^+ + \tau\partial g(x^+) \\ y + s\mathcal{A}(2x^+ - x) &\in y^+ + s\partial h^\star(y^+) \end{aligned}$$

Thus the proximal-point method on $\mathcal{D}^{-1} \circ T$ simply becomes

$$\begin{aligned} x_{t+1} &= \mathrm{prox}_{\tau g}(x_t - \tau\mathcal{A}^* y_t) \\ y_{t+1} &= \mathrm{prox}_{sh^\star}(y_t + s\mathcal{A}(2x_{t+1} - x_t)) \end{aligned}$$

This algorithm is called the preconditioned proximal point method and is recorded in Algorithm 7.

Algorithm 7: Preconditioned Proximal Point Algorithm (PPPA)

Input: Initial $x_0 \in \mathrm{dom}\, g$, $y_0 \in \mathrm{dom}\, h$, parameters $s, \tau > 0$, iteration $T \in \mathbb{N}$.

Step $t = 0, 1, \ldots, T$: compute

$$\begin{aligned} x_{t+1} &= \mathrm{prox}_{\tau g}(x_t - \tau A^* y_t) \\ y_{t+1} &= \mathrm{prox}_{sh^*}(y_t + sA(2x_{t+1} - x_t)) \end{aligned}$$

The generic guarantees for the proximal point method (Theorem 7.7) clearly apply to Algorithm 7. Namely, the iterates generated by the algorithm converge to a solution of the system (7.9), provided one exists. We will now prove a rate of convergence on the suboptimality gap. To this end, define the function

$$\Phi(x, y) = g(x) + \langle Ax, y\rangle - h^\star(y).$$

Then we express the primal and dual objective values as

$$\begin{aligned} \varphi(x) &= \sup_y \ \Phi(x, y) = h(\mathcal{A}x) + g(x) \\ \psi(y) &= \inf_x \ \Phi(x, y) = -g^\star(-\mathcal{A}^* y) - h^\star(y). \end{aligned}$$

Theorem 7.11 (Convergence of PPPA)**.** *Consider the primal-dual pair* (7.8)*, where* $g\colon \mathbf{E} \to \overline{\mathbf{R}}$ *and* $h\colon \mathbf{Y} \to \overline{\mathbf{R}}$ *are proper, closed, convex functions, and* $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$ *is a linear map. Suppose moreover that there exists* $R > 0$ *such that*

$$\|z_1 - z_2\|_{\mathcal{D}} \leq R \qquad \forall z_1, z_2 \in (\operatorname{dom} g) \times (\operatorname{dom} h^\star).$$

Let x_t *and* y_t *be the the iterates produced by Algorithm 7 and define the averages* $\bar{x}_t = \frac{1}{t}\sum_{i=0}^{t} x_t$ *and* $\bar{y}_t = \frac{1}{t}\sum_{i=0}^{t} y_t$*. Then the estimate holds:*

$$\varphi(\bar{x}_t) - \psi(\bar{y}_t) \leq \frac{R^2}{2(t+1)}.$$

Proof. A quick computation shows that the update of Algorithm 7 can be equivalently written as

$$\begin{aligned} x_{t+1} &= \operatorname*{argmin}_x \; g(x) + \langle \mathcal{A}x, y_t\rangle + \frac{1}{2\tau}\|x - x_t\|^2 \\ y_{t+1} &= \operatorname*{argmin}_y \; h^\star(y) - \langle \mathcal{A}(2x_{t+1} - x_t), y\rangle + \frac{1}{2s}\|y - y_t\|^2. \end{aligned}$$

Since x_{t+1} is the minimizer of the $\frac{1}{\tau}$-strongly convex function $g + \langle \mathcal{A}\cdot, y_t\rangle + \frac{1}{2\tau}\|\cdot - x_t\|^2$, we deduce

$$\begin{aligned} g(x_{t+1}) + \langle \mathcal{A}x_{t+1}, y_t\rangle + \frac{1}{2\tau}\|x_{t+1} - x_t\|^2 \leq g(x) + \langle \mathcal{A}x, y_t\rangle + \frac{1}{2\tau}\|x - x_t\|^2 \\ - \frac{1}{2\tau}\|x_{t+1} - x\|^2 \end{aligned} \tag{7.15}$$

for all $x \in \mathbf{E}$. Exactly the same reasoning for y_{t+1} shows the estimate

$$\begin{aligned} h^\star(y_{t+1}) - \langle \mathcal{A}(2x_{t+1} - x_t), y_{t+1}\rangle + \frac{1}{2s}\|y_{t+1} - y_t\|^2 \\ \leq h^\star(y) - \langle \mathcal{A}(2x_{t+1} - x_t), y\rangle + \frac{1}{2s}\|y - y_t\|^2 - \frac{1}{2s}\|y - y_t\|^2, \end{aligned} \tag{7.16}$$

for all $y \in \mathbf{Y}$. Summing (7.15) and (7.16) yields (miraculously!) the guarantee

$$\Phi(x_{t+1}, y) - \Phi(x, y_{t+1}) \leq \frac{1}{2}\|z_t - z\|_{\mathcal{D}}^2 - \frac{1}{2}\|z_{t+1} - z\|_{\mathcal{D}}^2 - \frac{1}{2}\|z_{t+1} - z_t\|_{\mathcal{D}}^2,$$

where we define $z = (x, y)$, $z_t = (x_t, y_t)$, and $z_{t+1} = (x_{t+1}, y_{t+1})$. Upper-bounding the last term by zero and summing the inequalities for the iterates $0, \dots, t+1$, we conclude

$$\sum_{i=0}^{t+1} \Phi(x_i, y) - \Phi(x, y_i) \leq \frac{1}{2}\|z_0 - z\|_{\mathcal{D}}^2.$$

Dividing through by $t+2$ and noting that $\Phi(\cdot, y)$ is convex and $\Phi(x, \cdot)$ is concave, we arrive at

$$\Phi(\bar{x}_{t+1}, y) - \Phi(x, \bar{y}_{t+1}) \leq \frac{\|z_0 - z\|_{\mathcal{D}}^2}{2(t+2)}.$$

Taking the supremum over y and x completes the proof □

7.5 Extragradient method

Consider the optimization problem

$$\min_x h(\mathcal{A}x) + g(x),$$

where $h\colon \mathbf{E} \to \overline{\mathbf{R}}$ and $g\colon \mathbf{E} \to \overline{\mathbf{R}}$ are closed convex functions. We can equivalently write it as a saddle point problem

$$\min_x \max_y g(x) + \langle \mathcal{A}x, y\rangle - h^*(y).$$

In this section, we show how to solve such saddle point problems by iterating jointly in the (x, y) variables.

Consider a function $\Phi\colon \mathbf{E} \times \mathbf{Y} \to \overline{\mathbf{R}}$ and define the primal and dual objectives

$$\varphi(x) = \sup_y \ \Phi(x, y) \qquad \text{and} \qquad \psi(y) = \inf_x \ \Phi(x, y)$$

Assumption 7.12. We assume

$$\Phi(x, y) = p_1(x) + g(x, y) + p_2(y),$$

where $g\colon \mathbf{E} \times \mathbf{Y} \to \mathbf{R}$, $p_1\colon \mathbf{E} \to \overline{\mathbf{R}}$, $p_2\colon \mathbf{Y} \to \overline{\mathbf{R}}$ satisfy:

1. g is a β-smooth function,
2. $g(\cdot, y)$ is convex for all $y \in \mathbf{Y}$ and $g(x, \cdot)$ is concave for all $x \in \mathbf{E}$,

3. p_1 and $-p_2$ are proper, closed, convex functions, and the diameter of $(\operatorname{dom} p_1) \times (\operatorname{dom} -p_2)$ is upper bounded by R.

Define

$$F(z) = [\nabla_x g(x,y), -\nabla_y g(x,y)] \qquad P(x,y) = p_1(x) - p_2(y).$$

Everywhere, we will use notation $z = (x, y)$.

Algorithm 8: Extra-gradient method

Input: Starting point $x_0 \in \mathbf{E}$, parameter $\beta > 0$, iteration $T \in \mathbb{N}$.
Step $t = 0, 1, \ldots, T-1$: compute

$$\begin{aligned} \hat{z}_t &= \operatorname{prox}_{P/\beta}\left(z_t - \frac{1}{\beta}F(z_t)\right), \\ z_{t+1} &= \operatorname{prox}_{P/\beta}\left(z_t - \frac{1}{\beta}F(\hat{z}_t)\right). \end{aligned} \tag{7.17}$$

The convergence analysis will be based on monitoring the following quantity along the iterate sequence:

$$\sup_z \{P(z_t) - P(z) + \langle F(z_t), z_t - z\rangle\}. \tag{7.18}$$

The following lemma provides a simple path to translate an upper bound on (7.18) into guarantees on the difference between the primal and dual objective values.

Lemma 7.13. *For any points $z = (x, y)$ and $\hat{z} = (\hat{x}, \hat{y})$, the estimate holds:*

$$P(\hat{z}) - P(z) + \langle F(\hat{z}), \hat{z} - z\rangle \geq \Phi(\hat{x}, y) - \Phi(x, \hat{y}).$$

Proof. Taking into account convexity of $g(\cdot, \hat{y})$ and concavity of $g(\hat{x}, \cdot)$ we deduce

$$\begin{aligned} \langle F(\hat{z}), \hat{z} - z\rangle &= \langle \nabla_x g(\hat{x}, \hat{y}), \hat{x} - x\rangle - \langle \nabla_y g(\hat{x}, \hat{y}), \hat{y} - y\rangle \\ &\geq g(\hat{x}, \hat{y}) - g(x, \hat{y}) + g(\hat{x}, y) - g(\hat{x}, \hat{y}) \\ &= g(\hat{x}, y) - g(x, \hat{y}), \end{aligned}$$

as claimed. Adding $P(\hat{x}) - P(z)$ to both sides and regrouping terms completes the proof.

□

We are not ready to analyze the algorithm. Notice that we may rewrite the update equations (7.17) as

$$\hat{z}_t = \operatorname*{argmin}_z \{\langle F(z_t), z\rangle + P(z) + \frac{\beta}{2}\|z - z_t\|^2\}$$
$$z_{t+1} = \operatorname*{argmin}_z \{\langle F(\hat{z}_t), z\rangle + P(z) + \frac{\beta}{2}\|z - z_t\|^2\}$$

Theorem 7.14. *Let $z_t = (x_t, y_t)$ be the iterates generated by Algorithm 8 and define the running averages $\bar{x}_t = \sum_{i=0}^t \hat{x}_t$ and $\bar{y}_t = \sum_{i=0}^t \hat{y}_t$. Then the following estimate holds:*

$$\varphi(\bar{x}_t) - \varphi(\bar{y}_t) \leq \frac{\beta R^2}{2(t+1)}.$$

Proof. Since $\hat{z}_t$ is the minimizer of the β-strongly convex function $\langle F(z_t), \cdot\rangle + P + \frac{\beta}{2}\|\cdot - z_t\|^2$, we deduce

$$\begin{aligned}\langle F(z_t), \hat{z}_t\rangle + \frac{\beta}{2}\|\hat{z}_t - z_t\|^2 + P(\hat{z}_t) \leq \langle F(z_t), u\rangle + \frac{\beta}{2}\|u - z_t\|^2 + P(u) \\ - \frac{\beta}{2}\|u - \hat{z}_t\|^2\end{aligned} \tag{7.19}$$

for all $u \in \mathbf{E} \times \mathbf{Y}$. Similarly, since x_{t+1} is the minimizer of the β-strongly convex function $\langle F(\hat{z}_t), \cdot\rangle + P + \frac{\beta}{2}\|\cdot - z_t\|^2$, we deduce

$$\begin{aligned}\langle F(\hat{z}_t), z_{t+1}\rangle + \frac{\beta}{2}\|z_{t+1} - z_t\|^2 + P(z_{t+1}) \leq \langle F(\hat{z}_t), z\rangle + \frac{\beta}{2}\|z - z_t\|^2 + P(z) \\ - \frac{\beta}{2}\|z - z_{t+1}\|^2\end{aligned} \tag{7.20}$$

for all $z \in \mathbf{E} \times \mathbf{Y}$. Next set $u = z_{t+1}$ in (7.19). Rearranging (7.19) and (7.20) yields the following two estimates, respectively:

$$\begin{aligned}\langle F(z_t), \hat{z}_t - z_{t+1}\rangle &\leq \frac{\beta}{2}\left(\|z_{t+1} - z_t\|^2 - \|z_{t+1} - \hat{z}_t\|^2 - \|\hat{z}_t - z_t\|^2\right) \\ &\quad + P(z_{t+1}) - P(\hat{z}_t) \\ \langle F(\hat{z}_t), z_{t+1} - z\rangle &\leq \frac{\beta}{2}\left(\|z - z_t\|^2 - \|z - z_{t+1}\|^2 - \|z_{t+1} - z_t\|^2\right) \\ &\quad + P(z) - P(z_{t+1}).\end{aligned} \tag{7.21}$$

Seeking to bound the quantities $\langle F(\hat{z}_t), \hat{z}_t - z\rangle$, let us decompose it as

$$\langle F(\hat{z}_t), \hat{z}_t - z\rangle = \langle F(\hat{z}_t) - F(z_t), \hat{z}_t - z_{t+1}\rangle + \langle F(z_t), \hat{z}_t - z_{t+1}\rangle + \langle F(\hat{z}_t), z_{t+1} - z\rangle$$

Bounding the last two terms using (7.21) yields the estimate

$$\begin{aligned}\langle F(\hat{z}_t), \hat{z}_t - z\rangle \leq \frac{\beta}{2}\left(\|z - z_t\|^2 - \|z - z_{t+1}\|^2\right) + P(z) - P(\hat{z}_t) \\ + \langle F(\hat{z}_t) - F(z_t), \hat{z}_t - z_{t+1}\rangle - \tfrac{\beta}{2}\left(\|z_{t+1} - \hat{z}_t\|^2 + \|\hat{z}_t - z_t\|^2\right).\end{aligned} \tag{7.22}$$

Using the Cauchy–Schwarz inequality and Lipschitz continuity of F, we deduce that the term on the last line of (7.22) is upper-bounded by

$$\beta\|\hat{z}_t - z_{t+1}\| \cdot \|\hat{z}_t - z_t\| - \frac{\beta}{2}\left(\|z_{t+1} - \hat{z}_t\|^2 + \|\hat{z}_t - z_t\|^2\right) = 0.$$

Summing (7.22) for $i = 0, \ldots, t-1$ and dividing by t yields

$$\frac{1}{t}\sum_{t=0}^{t-1}(P(\hat{z}_i) - P(z) + \langle F(\hat{z}_i), \hat{z}_i - z\rangle) \leq \frac{\beta\|z - z_0\|^2}{2t}. \tag{7.23}$$

Invoking Lemma 7.13 and using convexity of $\Phi(\cdot, y)$ and concavity of $\Phi(x, \cdot)$, we conclude

$$\begin{aligned}\frac{1}{t}\sum_{t=0}^{t-1}(P(\hat{z}_i) - P(z) + \langle F(\hat{z}_i), \hat{z}_i - z\rangle) &\geq \frac{1}{t}\sum_{t=0}^{t-1}(\Phi(\hat{x}_i, y) - \Phi(x, \hat{y}_i)) \\ &\geq \Phi(\bar{x}_{t-1}, y) - g(x, \bar{y}_{t-1}).\end{aligned}$$

Combining with (7.23) we deduce

$$\Phi(\bar{x}_{t-1}, y) - \Phi(x, \bar{y}_{t-1}) \leq \frac{\beta\|z - z_0\|^2}{2t}.$$

Taking the supremum over $(x, y) \in \mathbf{E} \times \mathbf{Y}$ completes the proof. □

Chapter 8

Introduction to Variational Analysis

In this chapter, we depart from convexity and analyze computational problems that may be neither smooth nor convex.

8.1 An introduction to variational techniques.

The typical technique for answering existence questions in various mathematical disciplines is based on fixed point theorems, such as those of Banach, Brower, and Kakutani. Variational analysis in large part replaces fixed point arguments with a *variational technique* that invokes existence of minimizers for a judiciously chosen potential function. This section presents a few variational arguments of this type to familiarize the reader with the technique.

As the first example, the following lemma shows that the subdifferential map of any closed superlinearly growing function is surjective. It is worthwhile for the reader to pause here and try their hand at a direct proof to better appreciate the elegance of the variational technique.

Lemma 8.1. *Consider a proper, closed function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ that is superlinearly growing, meaning*

$$\lim_{\|x\|\to\infty} \frac{f(x)}{\|x\|} = +\infty. \tag{8.1}$$

Then the subdifferential $\partial f\colon \mathbf{E} \rightrightarrows \mathbf{E}$ is surjective.

Proof. For any vector $v \in \mathbf{E}$, define the potential function $f_v(x) = f(x) - \langle v, x\rangle$. The assumption (8.1) immediately implies that f_v is coercive. Con-

sequently, f_v admits a minimizer x. First-order conditions for optimality therefore yield the inclusion $0 \in \partial f_v(x) = \partial f(x) - v$, or equivalently $v \in \partial f(x)$. Since $v \in \mathbf{E}$ is arbitrary, we conclude that ∂f is surjective. □

To motivate the next example, consider the exponential function $f(x) = e^x$. Clearly f does not admit any minimizers or even critical points. Nonetheless, there does exist an asymptotically critical sequence x_i along which the gradients $\nabla f(x_i)$ tend to zero. The following exercise uses the variational technique to establish existence of asymptotically critical sequences for any proper, closed, function that is bounded from below.

Exercise 8.2 (Asymptotic criticality)**.** Suppose $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is proper, closed, and bounded from below. Then there exist sequences $x_i \in \mathbf{E}$ and $v_i \in \partial f(x_i)$ satisfying $v_i \to 0$ and $f(x_i) \to \inf f$.
[**Hint:** Fix a sequence $r_i \to 0$ and define the functions $f_i(y) = f(y) + \frac{1}{2r_i}\|y\|^2$. Argue that f_i is coercive and closed and therefore admits a minimizer x_i. See Figure 8.1.]

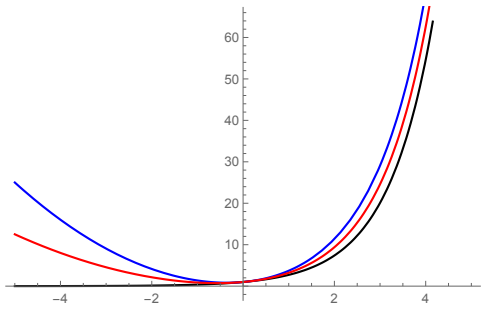


Figure 8.1: Quadratic perturbations of the exponential function.

Recall that the subdifferential of any convex function f is nonempty at any point in the relative interior of its domain (Theorem 3.38). The subdifferential would not be a very useful tool for nonconvex functions if it were often empty. As the final example, the following lemma shows that the domain of the subdifferential is dense in the domain of the function.

Lemma 8.3 (Density of the subdifferential)**.** *Consider a proper closed function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$. Then the set* $\operatorname{dom} \partial f$ *is dense in* $\operatorname{dom} f$.

Proof. Fix a point $x \in \operatorname{dom} f$. We will show that there exists a sequence $x_i \to x$ such that the subdifferential $\partial f(x_i)$ is nonempty. Since f is closed, the exists a closed ball Q around x such that the estimate $f(y) \geq f(x) - 1$

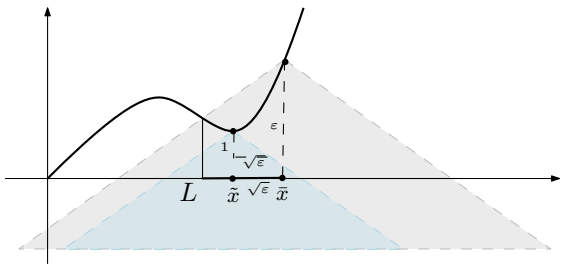


Figure 8.2: Illustration of the proof of Ekeland's principle.

holds for all $y \in Q$. Fix a sequence $r_i \searrow 0$ and define the potential function $f_i(y) = f(y) + \delta_Q(y) + \frac{1}{2r_i}\|y - x\|^2$. Then f_i is clearly closed and coercive and therefore admits a minimizer x_i. Observe $f(x_i) + \frac{1}{2r_i}\|x_i - x\|^2 \leq f(x)$ and therefore $\|x_i - x\|^2 \leq 2r_i(f(x) - f(x_i)) \leq 2r_i$. We deduce $x_i \to x$. In particular, x_i lies in the interior of Q for all large indices i. Necessary conditions for optimality thus become $0 \in \partial f_i(x_i) = \partial f(x_i) + \frac{1}{r_i}(x_i - x)$, or equivalently $\frac{1}{r_i}(x - x_i) \in \partial f(x_i)$. The proof is complete. □

8.2 Variational principles.

We next prove a fundamental existence theorem, which has an appealing geometric interpretation and will be extensively used in the later sections. Setting the stage, consider a proper closed function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point $\bar{x}$ that is ε-minimal in the sense that $f(\bar{x}) - \inf f \leq \varepsilon$. The following theorem shows that there exists another point that is $\sqrt{\varepsilon}$-close to $\bar{x}$ and is a *true minimizer* of the slightly perturbed function $f + \sqrt{\varepsilon}\|\cdot - \tilde{x}\|$. Conceptually, this existence result is useful because it allows to invoke first-order conditions for optimality for a nearby function at a nearby point.

Theorem 8.4 has an intriguing geometric interpretation, which motivates its proof. Namely, if $\bar{x}$ were a true minimizer of f, then clearly f would admit a supporting affine minorant at $\bar{x}$ with zero slope. When $\bar{x}$ is only an ε-approximate minimizer, the theorem shows that f admits a conic supporting minorant $f(\tilde{x}) - \sqrt{\varepsilon}\|\cdot - \tilde{x}\|$ at some nearby point $\tilde{x} \in B_{\sqrt{\varepsilon}}(\bar{x})$. See Figure 8.2 for an illustration. The width of the conic region scales as $1/\sqrt{\varepsilon}$, and it therefore closely approximates a supporting halfspace as ε tends to zero.

Theorem 8.4 (Nonsmooth variational principle)**.** *Let $f\colon \mathbf{R} \to \overline{\mathbf{R}}$ be a proper closed function. Fix a point $\bar{x}$ satisfying $f(\bar{x}) - \inf f \leq \varepsilon$ for some finite $\varepsilon > 0$. Then for any $\delta > 0$, there exists a point $\tilde{x}$ satisfying*

1. $\|\bar{x} - \tilde{x}\| \leq \frac{\varepsilon}{\delta}$,
2. $f(\tilde{x}) \leq f(\bar{x})$,
3. $\{\tilde{x}\} = \operatorname{argmin}_x\{f(x) + \delta\|x - \tilde{x}\|\}$.

Proof. The goal is to show that f admits a conic minorant $f(\tilde{x}) - \delta\|\cdot - \tilde{x}\|$ at some point $\tilde{x}$ near $\bar{x}$. To this end, define the function $C(x) = f(\bar{x}) - \delta\|x - \bar{x}\|$. If the estimate $f(x) > C(x)$ holds for all $x \neq \bar{x}$, then we may simply set $\tilde{x} = \bar{x}$. Suppose therefore that this is not the case and define the set

$$L := \{x : f(x) \leq C(x)\} \quad \text{or equivalently} \quad L = \{x : f(x) + \delta\|x - \bar{x}\| \leq f(\bar{x})\}.$$

Clearly, L is nonempty and compact since it is a sublevel set of the proper, closed, coercive function $f + \delta\|\cdot - \bar{x}\|$. Consequently, the function $f + \delta_L$ admits some minimizer $\tilde{x}$; see Figure 8.2 for an illustration. Since $\tilde{x}$ lies in L, the inequality holds:

$$f(\tilde{x}) + \delta\|\tilde{x} - \bar{x}\| \leq f(\bar{x}). \tag{8.2}$$

The two estimates $\|\bar{x} - \tilde{x}\| \leq \frac{\varepsilon}{\delta}$ and $f(\tilde{x}) \leq f(\bar{x})$ follow immediately. Next, observe that for all $x \in L$, we have $f(\tilde{x}) \leq f(x)$. Moreover, any point $x \notin L$ by definition satisfies $f(x) + \delta\|x - \bar{x}\| > f(\bar{x})$. Lower-bounding the right side using (8.2) and rearranging yields

$$f(\tilde{x}) < f(x) + \delta\|x - \bar{x}\| - \delta\|\tilde{x} - \bar{x}\| \leq f(x) + \delta\|x - \tilde{x}\|,$$

where we used the reverse triangle inequality. Thus the equality $\{\tilde{x}\} = \operatorname{argmin}_x\{f(x) + \delta\|x - \tilde{x}\|\}$ holds as claimed. □

Remarkably, Theorem 8.4 is true in any complete metric space, where one simply replaces the norm in the statement by the metric. This generalization is called the Ekeland's variational principle. One reasonable critique of Theorem 8.4 is that it involves a nonsmooth perturbation of f that is proportional to the norm. The following elementary theorem replaces this perturbation with the squared Euclidean norm. The geometric consequence is that supporting conics in Theorem 8.4 are replaced by supporting quadratics with small slope.

Theorem 8.5 (Smooth variational principle)**.** *Let $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a proper closed function. Fix a point $\bar{x}$ satisfying $f(\bar{x}) - \inf f \leq \varepsilon$ for some finite $\varepsilon > 0$. Then for any $\delta > 0$, there exists a point $\tilde{x}$ satisfying*

1. $\|\bar{x} - \tilde{x}\| \leq \frac{\varepsilon}{\delta}$,

2. $f(\tilde{x}) \leq f(\bar{x})$,

3. $\tilde{x} \in \operatorname{argmin}_x\{f(x) + \frac{\delta^2}{\varepsilon}\|x - \bar{x}\|^2\}$.

Proof. Since the function $f + \frac{\delta^2}{\varepsilon}\|\cdot - \bar{x}\|^2$ is proper, closed, and coercive, it admits a minimizer $\tilde{x}$. Consequently, we deduce $f(\tilde{x}) + \frac{\delta^2}{\varepsilon}\|\tilde{x} - \bar{x}\|^2 \leq f(\bar{x})$. The claimed properties follow trivially. □

A subtle generalization of Theorem 8.5 to complete metric spaces is called the Borwein-Preiss variational principle. Theorems 8.4 and 8.5 can typically be used interchangeably in finite dimensions. In the later sections, we will mostly use Theorem 8.4 as it yields slightly better bounds.

8.3 Descent principle and stability of sublevel sets.

Level and sublevel sets of functions often appear as constraints in optimization problems. Consequently, the geometry and stability of such sets with respect to perurbation play an important role. It is important to keep in mind that level sets even of C^∞-smooth functions can be unwieldy in general. Indeed, the celebrated Whitney's extension theorem guarantees that *any closed set* in $\mathbf{E}$, however pathological, can be realized as a level set of some C^∞-smooth function. Consequently, without further assumptions, there is no difference between closed sets (e.g. fractals) and those that are cut out by smooth equations! A closely related pathology is that the level-set mapping $\alpha \mapsto [f = \alpha]$ may be highly irregular. To illustrate, consider a monotone univariate function $f\colon \mathbf{R} \to \mathbf{R}$. Then the mapping $\alpha \mapsto [f = \alpha]$ is simply the inverse f^{-1} and, as such, its derivative is given by $1/f'(x)$. Thus the local Lipschitz constant of the level-set map is determined by the reciprocal of the derivative. In this section, we prove a far reaching generalization of this phenomenon for sublevel sets of any proper closed function.

Throughout, will measure the deviation between any two sets $X, Y \subset \mathbf{E}$ using the *Hausdorff distance*:

$$\operatorname{dist}(X, Y) := \max\left\{\sup_{x\in X} \operatorname{dist}(x, Y), \sup_{y\in Y} \operatorname{dist}(y, X)\right\}.$$

8.3.1 Level sets of smooth functions.

Before addressing stability of sublevel sets of nonsmooth functions, it is instructive to consider first the smooth settings, where classical (nonvariational) arguments suffice. To this end, a clear prerequisite for analyzing

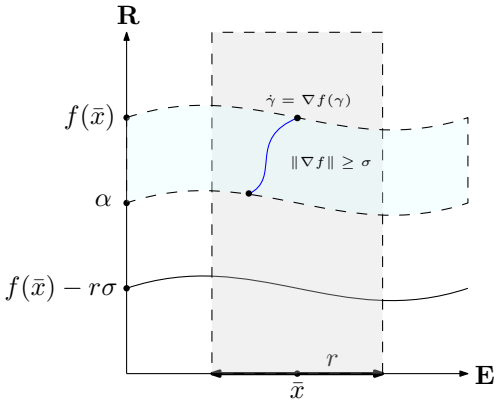


Figure 8.3: Illustration of the descent principle (Theorem 8.6).

stability of the map $\alpha \mapsto [f = \alpha]$ is to estimate the range of parameters α for which the level set $[f = \alpha]$ is nonempty. The following theorem provides such an estimate, based on an assumed lower bound on the gradient norm $\|\nabla f(x)\| \geq \sigma > 0$ near a point $\bar x$; see Figure 8.3 for an illustration. The theorem moreover shows that for a range of right-hand sides α, the distance from $\bar x$ to the sublevel set $[f = \alpha]$—a quantity that is difficult to estimate in general—is linearly bounded by the easily computable residual value $|f(x) - \alpha|$. Estimates of this type are often called error bounds, and will play a prominent role in later sections.

Theorem 8.6 (Smooth descent principle)**.** *Let $f\colon \mathbf{E} \to \mathbf{R}$ be a C^1-smooth function that has a locally Lipschitz continuous gradient. Suppose that for some point $\bar x \in \mathbf{E}$, there are constants $\alpha < f(\bar x)$ and $\sigma, r > 0$ satisfying inequality*

$$\|\nabla f(x)\| \geq \sigma \qquad \text{for all } x \in B_r(\bar x) \cap [\alpha < f \leq f(\bar x)].$$

If in addition $f(\bar x) - \alpha < r\sigma$, then the estimate holds:

$$\operatorname{dist}(\bar x, [f = \alpha]) \leq \sigma^{-1}(f(\bar x) - \alpha). \tag{8.3}$$

Proof. The existence theorem for ordinary differential equations yields a differentiable curve $\gamma\colon [0, \eta) \to \mathbf{E}$ satisfying $\dot\gamma(t) = -\nabla f(\gamma(t))$ for all $t \in [0, \eta)$, where $[0, \eta)$ is a maximal domain of definition. Define the set

$$U := B_r(\bar x) \cap [\alpha < f \leq f(\bar x)]$$

and define the exit time $t^* := \inf\{t \geq 0 : \gamma(t) \notin U\}$. Observe that for any $t \in [0, t^*)$ the fundamental theorem of calculus gives

$$\begin{aligned}\alpha - f(\bar{x}) \leq f(\gamma(t)) - f(\gamma(0)) &= \int_0^t \frac{d}{ds}(f \circ \gamma)(s)\, ds \\ &= \int_0^t \langle \nabla f(\gamma(s), \dot{\gamma}(s)\rangle\, ds \\ &= \int_0^t -\|\nabla f(\gamma(s))\|^2\, ds \qquad (8.4) \\ &\leq -t\sigma^2. \qquad (8.5)\end{aligned}$$

We claim that t^* is finite. Indeed, in the case $\eta = \infty$, the estimate (8.5) immediately guarantees that the exit time time t^* must be finite. In the complementary case $\eta < \infty$, the ODE extension theorem implies $\|\gamma(t)\| \to \infty$ and therefore γ eventually leaves U, as well.

Next, set $x^* := \gamma(t^*)$. By continuity, one of the the two conditions, $\|x^* - \bar{x}\| = r$ or $f(x^*) = \alpha$, must hold. We therefore deduce the lower bound on the length

$$\int_0^{t^*} \|\dot{\gamma}(s)\|\, ds \geq \|x^* - \bar{x}\| \geq \min\{r, \operatorname{dist}(\bar{x}, [f = \alpha])\}.$$

Continuing (8.4) with $t = t^*$, we therefore conclude

$$f(x^*) - f(\bar{x}) = -\int_0^{t^*} \|\nabla f(\gamma(s))\| \cdot \|\dot{\gamma}(s)\|\, dt \leq -\sigma \min\{r, \operatorname{dist}(\bar{x}, [f = \alpha])\}. \qquad (8.6)$$

Rearranging yields

$$\min\{r, \operatorname{dist}(\bar{x}, [f = \alpha])\} \leq \sigma^{-1}(f(\bar{x}) - f(x^*)) \leq \sigma^{-1}(f(\bar{x}) - \alpha).$$

Taking into account the assumed inequality $r > \sigma^{-1}(f(\bar{x}) - \alpha)$ completes the proof. □

An easy consequence of Theorem 8.6 is a gradient-based characterization of local Lipschitz continuity of the level set map with respect to the Hausdorff distance.

Corollary 8.7 (Lipschitz continuity of level sets.)**.** *Consider a C^1-smooth function $f\colon \mathbf{E} \to \mathbf{R}$ with a locally Lipschitz continuous gradient. Fix $\sigma > 0$ and two real numbers $\alpha < \beta$. Then the following two conditions are equivalent.*

1. **(Noncriticality)** *For all $x \in [\alpha < f < \beta]$ the estimate holds:*

$$\|\nabla f(x)\| \geq \sigma.$$

2. **(Lipschitz continuity)** *For all $s, t \in (\alpha, \beta)$ the estimate holds:*

$$\operatorname{dist}([f = s], [f = t]) \leq \sigma^{-1}|s - t|.$$

Proof. The implication (1) $\Rightarrow$ (2) follows directly by applying Theorem 8.6 to f and $-f$. To see the implication (2) $\Rightarrow$ (1), fix a point $x \in [\alpha < f < \beta]$. Fix a sequence $\gamma_i \to f(x)$ but distinct from $f(x)$ and choose an arbitrary sequence $x_i \in \operatorname{proj}_{[f=\gamma_i]}(x)$. Assumption (2) guarantees

$$\|x - x_i\| = \operatorname{dist}(x, [f = \gamma_i]) \leq \sigma^{-1}|f(x) - \gamma_i| \to 0.$$

Thus x_i tends to x and we therefore conclude

$$\|\nabla f(x)\| \geq \lim_{i\to\infty} \frac{\|f(x_i) - f(x)\|}{\|x_i - x\|} = \lim_{i\to\infty} \frac{|\gamma_i - f(x)|}{\operatorname{dist}(x, [f = \gamma_i])} \geq \sigma,$$

as we had to show. □

8.3.2 Sublevel sets of nonsmooth functions.

Our goal for the rest of the section is to generalize Theorem 8.6 to any proper closed function f. To this end, clearly we must find a replacement for the norm of the gradient. One appealing candidate is the minimal subgradient norm $\operatorname{dist}(0, \partial f(x))$. A technical problem with this idea is that the subdifferential $\partial f(x)$ may be empty even if f is a Lipschitz continuous function. As a result, it will be convenient to use a slightly different construction, which coincides with $\operatorname{dist}(0, \partial f(x))$ whenever $\partial f(x)$ is nonempty. Henceforth, for any real number r, we define the positive part $r^+ = \max\{r, 0\}$.

Definition 8.8 (Slope)**.** Consider a function $f\colon \mathbf{R} \to \overline{\mathbf{R}}$ and a point x with $f(x)$ finite. The *slope* of f at x, denoted by $|\nabla f|(x)$, is the quantity

$$|\nabla f|(x) := \limsup_{y\to x} \frac{(f(x) - f(y))^+}{\|x - y\|}.$$

Thus when x happens to be a local minimizer of f, the slope $|\nabla f|(x)$ is zero. If this is not the case, then the slope $|\nabla f|(x)$ measures the maximal instanteneous rate of decrease of f at x. The following exercise records a few basic properties of the slope.

Exercise 8.9 (Basic properties)**.** Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point x with $f(x)$ finite. Verify the following.

1. The equality $|\nabla f|(x) = -\inf_{\|u\|\le 1} df(x)(u)$ holds.
2. If f is differentiable at x, then equality $|\nabla f|(x) = \|\nabla f(x)\|$ holds.
3. The inequality, $|\nabla f|(x) \le \operatorname{dist}(0, \partial f(x))$, always holds. Moreover, equality holds as long as $\partial f(x)$ is nonempty.

[**Hint:** The first claim follows from the definition of the slope $|\nabla f|(x)$ and the subderivative $df(x)(u)$. To see the second claim, verify $|\nabla f|(x) \le \operatorname{dist}(0, \partial f(x))$ directly from definitions. Suppose now that $\partial f(x)$ is nonempty. Theorem 3.47 then guarantees that the closed convex envelope $\overline{\operatorname{co}}\, df(x)$ is the support function of $\partial f(x)$. Using Exercise 3.20 conclude $\inf_{\|u\|\le 1} df(x)(u) = \inf_{\|u\|\le 1} \overline{\operatorname{co}}\, df(x)(u)$. Use this expression to complete the proof.]

We are now ready to establish a generalization of Theorem 8.6 to nonsmooth functions with the slope replacing the norm of the gradient. The argument is based on the nonsmooth variational principle. This is sharp contrast to the proof of Theorem 8.6, which relied on the existence theorem for ODEs and the chain rule for differentiation.

Theorem 8.10 (Decrease principle)**.** *Let $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a proper closed function. Suppose that for some point $\bar{x} \in \operatorname{dom} f$, there are constants $\alpha < f(\bar{x})$ and $\sigma, r > 0$ so that the implication*

$$|\nabla f|(x) \ge \sigma \qquad \text{for all } x \in B_r(\bar{x}) \cap [\alpha < f \le f(\bar{x})]$$

holds. If in addition $f(\bar{x}) - \alpha < r\sigma$, then the estimate holds:

$$\operatorname{dist}(\bar{x}, [f \le \alpha]) \le \sigma^{-1}(f(\bar{x}) - \alpha).$$

Proof. Define the function $g(u) = (f(u) - \alpha)^+$ and observe

$$g(\bar{x}) - \inf g \le f(\bar{x}) - \alpha.$$

We aim to apply the nonsmooth variational principle (Theorem 8.4) to the function g. To this end, set $\varepsilon := f(\bar{x}) - \alpha$. Taking into account the assumed inequality $\sigma^{-1}\varepsilon < r$, we may choose $\delta \in (0, \sigma)$ satisfying $\delta^{-1}\varepsilon \le r$. Applying Theorem 8.4 yields a point $\tilde{x}$ satisfying $g(\tilde{x}) \le g(\bar{x})$, $\|\tilde{x} - \bar{x}\| \le \delta^{-1}\varepsilon$, and

$$\{\tilde{x}\} = \operatorname*{argmin}_u \{g(u) + \delta\|u - \tilde{x}\|\}. \tag{8.7}$$

We claim that $\tilde{x}$ lies in the sublevel set $[f \leq \alpha]$. To see this, observe that the equality (8.7) along with the very definition of the slope implies $|\nabla g|(\tilde{x}) \leq \sigma$. Consequently, if the inequality $f(\tilde{x}) > \alpha$ were to hold, then f and g would coincide on a neighborhood of $\tilde{x}$ (why?) thereby yielding the contradiction $|\nabla f|(\tilde{x}) = |\nabla g|(\tilde{x}) \leq \delta < \sigma$. We conclude $\operatorname{dist}(\bar{x}, [f \leq \alpha]) \leq \|\bar{x} - \tilde{x}\| \leq \delta^{-1}(f(\bar{x}) - \alpha)$. Letting δ tend to σ completes the proof. □

A slope-based characterization of local Lipschitz continuity of the sublevel set map follows quickly.

Corollary 8.11 (Lipschitz continuity of level sets.)**.** *Consider a proper closed function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$. Fix $\sigma > 0$ and two real numbers $\alpha < \beta$. Then the following two conditions are equivalent.*

1. **(Noncriticality)** *For all $x \in [\alpha < f < \beta]$ the estimate holds:*

$$|\nabla f|(x) \geq \sigma.$$

2. **(Lipschitz continuity)** *For all $s, t \in (\alpha, \beta)$ the estimate holds:*

$$\operatorname{dist}([f \leq s], [f \leq t]) \leq \sigma^{-1}|s - t|.$$

Proof. To see the implication (1) ⇒ (2), fix $s, t \in (\alpha, \beta)$. Without loss of generality suppose $s < t$. Then for any $x \in [f \leq s]$, we trivially have $\operatorname{dist}(x, [f \leq t]) = 0$. On the other hand, for any $x \in [f \leq t]$, Theorem 8.10 directly implies $\operatorname{dist}(x, [f \leq t]) \leq \sigma^{-1}(t - s)$. Thus (2) holds.

To see the reverse implication (2) ⇒ (1), fix a point $x \in [\alpha < f < \beta]$. Fix a strictly increasing sequence γ_i converging to $f(x)$ and choose an arbitrary sequence $x_i \in \operatorname{proj}_{[f \leq \gamma_i]}(x)$. Assumption (2) guarantees

$$\|x - x_i\| = \operatorname{dist}(x, [f \leq \gamma_i]) \leq \sigma^{-1}|f(x) - \gamma_i| \to 0.$$

Thus x_i tends to x and we therefore conclude

$$|\nabla f|(x) \geq \limsup_{i \to \infty} \frac{(f(x) - f(x_i))}{\|x - x_i\|} \geq \limsup_{i \to \infty} \frac{f(x) - \gamma_i}{\operatorname{dist}(x, [f \leq \gamma_i])} \geq \sigma,$$

as we had to show. □

8.4 Limiting subdifferential and limiting slope.

Both the slope $|\nabla f|$ and the Fréchet subdifferential ∂f have an important deficiency: they fail to have good semi-continuity properties with respect to their arguments even if f is Lipschitz continuous. As a simple illustration, consider the univariate function $f(x) = -|x|$. It is straightforward to verify that the graph of the subdifferential $\operatorname{gph} \partial f$ is not closed and the slope function $x \mapsto |\nabla f|(x)$ is not lower-semicontinuous at zero. Though this seems like a technical issue at first, it is fundamental. For example, desirable calculus rules fail for the Fréchet subdifferential precisely for this reason. To illustrate, define the univariate function $g(x) = |x|$ and set $f(x) = -|x|$ as before. Since the sum $f + g$ is identically zero, the inclusion $0 \in \partial(f + g)(0)$ clearly holds. On the other hand, the sum of the subdifferentials $\partial f(0) + \partial g(0)$ is empty since $\partial f(0)$ is empty. Therefore, the desired inclusion $\partial(f + g)(0) \subset \partial f(0) + \partial g(0)$ fails in this setting.

An appealing remedy is to define a new subdifferential that is slightly larger than ∂f but has good closure properties. Perhaps the first attempt would be to declare a vector v a "limiting subgradient" of a function f at x if there exist sequences $(x_i, v_i) \in \operatorname{gph} \partial f$ converging to (x, v). This construction, however, is not well aligned with epigraphical geometry when f is discontinuous because the points $(x_i, f(x_i))$ may not converge to $(x, f(x))$. Consequently, when defining the new subdifferential we should instead focus on sequences x_i satisfying $(x_i, f(x_i)) \to (x, f(x))$. To this end, we say that a sequence x_i *converges f-attentively to* x, denoted $x_i \xrightarrow{f} x$, if it satisfies $(x_i, f(x_i)) \to (x, f(x))$. With this in mind, we define the following two constructions, which will play a central role in the next sections.

Definition 8.12 (Limiting subdifferential and limiting slope)**.** Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point x with $f(x)$ finite.

1. A vector v is a *limiting subgradient* of f at x, written $v \in \partial_L f(x)$, if there exist sequences $x_i \in \mathbf{E}$ and $v_i \in \partial f(x_i)$ satisfying $(x_i, f(x_i), v_i) \to (x, f(x), v)$.
2. The *limiting slope* of f at x as

$$\overline{|\nabla f|}(x) = \operatorname*{liminf}_{y \xrightarrow{f} x} |\nabla f|(y).$$

As a simple example, the reader should verify the expressions for the

function $f(x) = -|x|$:

$$\partial_L f(x) = \begin{cases} \{-1\} & x < 0 \\ \{-1, 1\} & x = 0 \\ \{1\} & x > 0 \end{cases} \quad \text{and} \quad \overline{|\nabla f|}(x) = \begin{cases} -1 & x < 0 \\ -1 & x = 0 \\ 1 & x > 0 \end{cases}.$$

Naturally for any set $Q \subset \mathbf{E}$ and $x \in Q$, we define the *limiting normal cone* $N_Q^L(x) := \partial_L \delta_Q(x)$. Thus for any closed set Q, the inclusion $v \in N_Q^L(x)$ means that there are sequences $(x_i, v_i) \in \operatorname{gph} N_Q$ satisfying $(x_i, v_i) \to (x, v)$; in other words, equality $\operatorname{gph} N_Q^L = \operatorname{cl}(\operatorname{gph} N_Q)$ holds. See Figure 8.4 for an illustration.

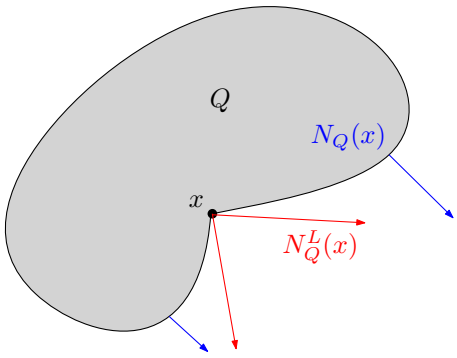


Figure 8.4: Limiting normal cone.

The following exercises establish some basic properties of the limiting constructions.

Exercise 8.13 (Basic properties)**.** Consider a proper closed function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point x with $f(x)$ finite.

1. Show that $\partial_L f(x)$ is a closed set, though is possibly not convex.
2. Show that if some sequences $x_i \in \mathbf{E}$ and $v_i \in \partial_L f(x_i)$ satisfy $(x_i, f(x_i), v_i) \to (x, f(x), v)$, then the inclusion $v \in \partial_L f(x)$ holds.
3. Show that if f is locally Lipschitz continuous around x, then $\partial_L f(x)$ is nonempty.

Exercise 8.14 (Limiting normals to epigraph)**.** Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point x with $f(x)$ finite. Show the equality

$$\partial_L f(x) = \{v \in \mathbf{E} : (v, -1) \in N_{\operatorname{epi} f}^L(x, f(x))\}.$$

Recall from Exercise 8.9 that the slope and the Fréchet subdifferential of a closed function are related by the expression:

$$|\nabla f|(x) = \operatorname{dist}(0, \partial f(x)),$$

as long as $\partial f(x)$ is nonempty. In the next section, we will see somewhat remarkably that the analogous equality holds for the limiting constructions, unconditionally.

8.5 Subdifferential calculus

Calculus of subdifferentials is of utmost importance both for theory and algorithms. Chapter 4 was in large part devoted to developing the calculus of subdifferentials of convex functions. In this section, we use the variational techniques to establish calculus rules in the nonconvex settings.

We will make use of two simple exercises. Recall that a Fréchet subgradient $v \in \partial f(x)$ is characterized by the inequality

$$f(y) \geq f(x) + \langle v, y - x \rangle + o(\|y - x\|) \qquad \text{as } y \to x.$$

The function on the right-hand side may in principle be nonsmooth in y, due to the presence of the little-o term. The following two exercises show that we may without loss of generality assume that the little-o term on the right is indeed C^1-smooth.

Exercise 8.15. Consider a function $\varphi\colon [0, +\infty) \to \mathbf{R}_+$ satisfying $\varphi(t) = o(t)$ as $t \searrow 0$. Then there exists a function $\psi\colon \mathbf{R} \to \mathbf{R}_+$ satisfying

1. $\psi(t) \geq \varphi(|t|)$ for all $t \in \mathbf{R}$,
2. ψ is C^1-smooth with $\psi(0) = \psi'(0) = 0$.

Exercise 8.16 (Supporting smooth minorants)**.** Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point x with $f(x)$ finite. Then the inclusion $v \in \partial f(x)$ holds if and only if there exists a C^1-smooth function $\omega\colon U \to \mathbf{R}$ defined on a neighborhood U of x satisfying $\omega(x) = f(x)$, $\nabla\omega(x) = v$, and $\omega(y) < f(y)$ for all $y \in U \setminus \{x\}$.

We now have all the ingredients to establish a chain rule for subdifferentials. Recall that even in the convex case, subdifferential calculus required some nondegeneracy assumptions. Rather then imposing nondegeneracy conditions directly, we prove a "fuzzy chain rule" which is always valid.

Namely, the following theorem shows that when composing a closed function $h(\cdot)$ with a differentiable map $c(\cdot)$, the inclusion holds:

$$\partial(h \circ c)(\bar{x}) \subset \nabla c(x)^* \partial h(y) + \varepsilon \mathbb{B},$$

where x and y are ε perturbations of $\bar{x}$ and $c(\bar{x})$, respectively.

Theorem 8.17 (Fuzzy chain rule)**.** *Consider the function $f(x) = h(c(x))$, where $h\colon \mathbf{Y} \to \overline{\mathbf{R}}$ is a proper, closed function and $c\colon \mathbf{E} \to \mathbf{Y}$ is differentiable around a point $\bar{x} \in \mathbf{E}$. Then the inclusion holds:*

$$\partial f(\bar{x}) \supset \nabla c(\bar{x})^* \partial h(c(\bar{x})).$$

Conversely, for every subgradient $v \in \partial f(\bar{x})$ and $\epsilon > 0$, there exist points $x \in \mathbf{E}$ and $y \in \mathbf{Y}$ satisfying

$$\|x - \bar{x}\| \leq \epsilon, \qquad \|y - c(\bar{x})\| \leq \epsilon, \qquad |h(y) - h(c(\bar{x}))| \leq \epsilon,$$

and such that the inclusion holds:

$$v \in \nabla c(x)^* \partial h(y) + \epsilon \mathbb{B}.$$

Proof. Fix a vector $v \in \partial f(\bar{x})$. Theorem 8.16 provides for us a C^1-smooth function $h\colon U \to \mathbf{R}$ defined on a neighborhood U of $\bar{x}$ satisfying $h(\bar{x}) = f(\bar{x})$, $\nabla h(\bar{x}) = v$, and $\omega(y) < f(y)$ for all $y \in U \setminus \{\bar{x}\}$. In particular, $\bar{x}$ is a strict local minimizer of $f - \omega$ on U. Shrinking U, we may assume U is a closed ball and c is differentiable on U. Let V be a compact neighborhood of $c(\bar{x})$. Fix a sequence $r_i \searrow 0$ and consider the decoupled minimization problem

$$\min_{x \in U, y \in V} \; F_i(x, y) := h(y) - \omega(x) + \frac{1}{2r_i}\|y - c(x)\|^2. \tag{8.8}$$

Since F_i is closed and coercive, the problem (8.8) admits a minimizer (x_i, y_i). Let us show that the sequence (x_i, y_i) tends to $(\bar{x}, c(\bar{x}))$ as i tends to infinity. To this end, passing to a subsequence, we may assume (x_i, y_i) tends to some pair (x^*, y^*). Observe $F_i(x_i, y_i) \leq F_i(\bar{x}, \bar{y}) = 0$ and therefore

$$\frac{1}{2r_i}\|y_i - c(x_i)\|^2 \leq \omega(x_i) - h(y_i).$$

Since the right-hand is bounded, we deduce $y_i - c(x_i) \to 0$ and therefore $y^* = c(x^*)$. Lower-semicontinuity of h guarantees

$$\begin{aligned} f(x^*) - \omega(x^*) &= h(y^*) - \omega(x^*) \\ &\leq \liminf_{i \to \infty} \; h(y_i) - \omega(x_i) + \frac{1}{2r_i}\|y_i - c(x_i)\|^2 \\ &= \liminf_{i \to \infty} F_i(x_i, y_i) \\ &\leq \liminf_{i \to \infty} F_i(\bar{x}, c(\bar{x})) = 0. \end{aligned} \tag{8.9}$$

Thus x^* is a minimizer of $f - \omega$ on C. Since x^* is the unique minimizer of $f - \omega$ on C, we deduce $x^* = \bar{x}$. Consequently equality holds throughout (8.9). It follow immediately that $h(y_i)$ tends to $h(y^*)$. Finally, since for all large indices i, the points x_i lie in the interior of U, first-order necessary optimality conditions for (8.8) become

$$\begin{aligned} 0 &\in -\nabla\omega(x_i) - \frac{1}{r_i}\nabla c(x_i)^*(y_i - c(x_i)) \\ 0 &\in \partial h(y_i) + \frac{1}{r_i}(y_i - c(x_i)). \end{aligned}$$

Substituting the second inclusion into the first and adding and subtracting v yields

$$v \in \nabla c(x_i)^* \partial h(y_i) + (v - \omega(x_i)).$$

This completes the proof. □

An appealing question is what happens when we allow ϵ to tend to zero in Theorem 8.17. Let us look at this question more closely. Namely for any vector $v \in \partial(h \circ c)(\bar{x})$, the theorem guarantees

$$\|v - \nabla c(x_\epsilon)^* w_\epsilon\| \leq \epsilon,$$

for some points $x_\epsilon \to \bar{x}$, $y_\epsilon \to c(\bar{x})$ satisfying $h(y_\epsilon) \to h(c(\bar{x}))$, and $w_\epsilon \in \partial h(y_\epsilon)$. The main problems that may arise when taking the limit as $\epsilon \to 0$ is that the subgradients w_ϵ might be unbounded. In order to rule out this degeneracy, we must ensure that the directions along which subgradient blow up are disjoint from the null space of $\nabla c(\bar{x})^*$. To make this assumption precise, we introduce the following set, which captures such directions of blow up.

Definition 8.18 (Horizon subdifferential)**.** Consider a function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point $\bar{x}$ with $f(\bar{x})$ finite. We say that v is a *horizon subgradient* of f at x, written $v \in \partial^\infty f(x)$ if there exist sequences $x_i \in \mathbf{E}$, $v_i \in \partial f(x_i)$, and $\tau_i \searrow 0$ satisfying $(x_i, f(x_i), \tau_i v_i) \to (x, f(x), v)$.

The following exercise establishes a few basic properties of the horizon subdifferential. In particular, the horizon subdifferential has an intuitive geometric interpretation in terms of horizontal normals to the epigraph.

Exercise 8.19 (Horizon subdifferential: basic properties)**.** Consider a proper, closed function $f\colon \mathbf{E} \to \mathbf{R}$ and a point x with $f(x)$ finite. Show the following.

1. The set $\partial^\infty f(x)$ is a closed cone.

2. The equality, $\partial^\infty f(x) = \{v \in \mathbf{E} : (v, 0) \in N^L_{\operatorname{epi} f}(x, f(x))\}$, holds.
3. If f is locally Lipschitz around x, then equality $\partial^\infty f(x) = \{0\}$ holds.
4. If f is convex, then the equality, $\partial^\infty f(x) = N_{\operatorname{dom} f}(x)$, holds.
5. If f is an indicator function of a closed set $Q \subset \mathbf{E}$, then we have $\partial^\infty f(x) = N_Q(x)$.

We are now ready to pass to the limit $\epsilon \searrow 0$ in the fuzzy chain rule.

Corollary 8.20 (Chain rule in limiting form)**.** *Consider the function $f(x) = h(c(x))$, where $h\colon \mathbf{Y} \to \overline{\mathbf{R}}$ is a proper, closed function and $c\colon \mathbf{E} \to \mathbf{Y}$ is C^1-smooth around a point $\bar{x} \in \mathbf{E}$. Then the inclusion holds:*

$$\partial f(\bar{x}) \supset \nabla c(\bar{x})^* \partial h(c(\bar{x})).$$

Moreover, the transversality condition

$$\partial^\infty h(\bar{x}) \cap \operatorname{Ker}\, (\nabla c(\bar{x})^*) = \{0\},$$

guarantees the reverse inclusion

$$\partial_L f(\bar{x}) \subset \nabla c(\bar{x})^* \partial_L h(c(\bar{x})).$$

Proof. We know that for every subgradient $v \in \partial f(\bar{x})$ and $\epsilon > 0$, there exist points $x \in \mathbf{E}$ and $y \in \mathbf{Y}$ satisfying

$$\|x - \bar{x}\| \leq \epsilon, \qquad \|y - c(\bar{x})\| \leq \epsilon, \qquad |h(y) - h(c(\bar{x}))| \leq \epsilon,$$

and such that the inclusion holds:

$$v \in \nabla c(x)^* w + \epsilon \mathbb{B} \qquad \text{for some } w \in \partial h(y).$$

Now let $\epsilon \searrow 0$. We claim that w is bounded. Indeed if this were not the case, we would deduce

$$\frac{v}{\|w\|} \in \nabla c(x)^* \left(\frac{w}{\|w\|} \right) + \frac{\epsilon}{\|w\|} \mathbb{B}.$$

Thus any limit point $\bar{w}$ of $w/\|w\|$ satisfies $0 = \nabla c(\bar{x})^* \bar{w}$ while at the same time $\bar{w} \in \partial^\infty h(\bar{x})$, which is a contradiction. □

As usual, the sum rule is an immediate consequence of the chain rule.

Corollary 8.21 (Sum rule)**.** *Consider two proper, closed functions $f_1, f_2 \colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point $x \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$. Then the inclusion holds:*

$$\partial(f_1 + f_2)(x) \supset \partial f_1(x) + \partial f_2(x).$$

Moreover, the transversality condition

$$\partial^\infty f_1(x) \cap -\partial^\infty f_2(x) = \{0\} \tag{8.10}$$

ensures the reverse inclusion

$$\partial_L(f_1 + f_2)(x) \subset \partial_L f_1(x) + \partial_L f_2(x).$$

Proof. Apply Corollary 8.20 with $c(x) = (x, x)$ and $h(y, z) = f_1(y) + f_2(z)$. □

Note in particular that the transversality condition (8.10) holds trivially if either f_1 or f_2 is Lipschitz continuous around x, since the corresponding horizon subdiffferential consists only of the origin (Exercise 8.19).

We end the section with two intriguing consequences of calculus rules. The first is a mean-value theorem for nonsmooth and nonconvex functions; the second shows that the limiting slope coincides with the minimal norm of limiting subgradients.

Theorem 8.22 (Mean-value I)**.** *Consider a proper closed function $f \colon \mathbf{E} \to \overline{\mathbf{R}}$ and fix two points $x_0, x_1 \in \operatorname{dom} f$. Then for every $\varepsilon > 0$, there exists a point x and a subgradient $v \in \partial f(x)$ satisfying $x \in [x_0, x_1] + \varepsilon\mathbb{B}$, $|f(y) - \min_{[x_0,x_1]} f| \leq \varepsilon$, and*

$$f(x_1) - f(x_0) \leq \langle v, x_1 - x_0 \rangle + \epsilon.$$

Proof. Define the map $c \colon \mathbf{R} \to \mathbf{E}$ by $c(t) = (1-t)x_0 + tx_1$ and set

$$\varphi(t) = f(c(t)) - (1-t)f(x_0) - tf(x_1) + \delta_{[0,1]}(t).$$

Observe the equality $\varphi(0) = \varphi(1) = 0$. Since φ is proper, closed, and coercive it admits some minimizer $\hat{t} \in [0, 1]$. Applying the fuzzy chain and sum rules, we deduce that for every $\epsilon > 0$, there exists $t_1 \in \mathbf{R}$, $t_2 \in [0, 1]$, and $x \in \mathbf{E}$ satisfying the proximity conditions

$$\max\left\{|t_1 - \hat{t}|, |t_2 - \hat{t}|, |x - c(t_1)|, |f(x) - f(c(\hat{t}))|\right\} \leq \varepsilon$$

and the the inclusion

$$0 \in \langle \partial f(y), x_1 - x_0 \rangle - (f(x_1) - f(x_0)) + N_{[0,1]}(t_2) + [-\varepsilon, \varepsilon].$$

The conclusion of the theorem follows immediately. □

Exercise 8.23 (Mean-value II)**.** Consider a function f that is continuous on a neighborhood of the line segment $[x_1, x_2] \subset \operatorname{dom} f$. Then for every $\epsilon > 0$, there exist a point $x \in [x_0, x_1] + \epsilon\mathbf{B}$ and a vector v that either lies in $\partial f(x)$ or in $-\partial(-f)(x)$ and satisfies

$$|f(x_1) - f(x_0) - \langle v, x_1 - x_0\rangle| \leq \epsilon.$$

[**Hint**: Proceed as in the proof of Theorem 8.22 but argue that φ always admits a maximizer or a minimizer that lies in the open interval $(0, 1)$.]

Theorem 8.24 (Limiting slope and limiting subdifferential)**.** *Consider a proper closed function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point x with $f(x)$ finite. Then the equality holds:*

$$\overline{|\nabla f|}(x) = \operatorname{dist}(0, \partial_L f(x)). \tag{8.11}$$

Proof. The inequality $\overline{|\nabla f|}(x) \leq \operatorname{dist}(0, \partial_L f(x))$ follows directly from Exercise 8.9 (verify this!). To see the converse, set $u := \overline{|\nabla f|}(x)$. If u is infinite, then the estimate $\overline{|\nabla f|}(x) \leq \operatorname{dist}(0, \partial_L f(x))$ directly implies that $\partial_L f(x)$ is empty. Therefore, in this case, the claimed equality (8.11) holds trivially. Consequently, for the rest of the proof we suppose that u is finite.

Fix an arbitrary $\varepsilon > 0$ and let y be a point satisfying

$$\|y - x\| < \varepsilon, \qquad |f(y) - f(x)| < \varepsilon, \quad \text{and} \quad |\nabla f|(x) < u + \varepsilon.$$

Define the function $g(z) := f(z) + (u + \varepsilon)\|z - y\|$. The lower bound on the slope ensures that y is a local minimizer of g and therefore

$$0 \in \partial_L g(y) \subset \partial_L f(y) + (u + \varepsilon)\mathbb{B}.$$

Rearranging yields $\operatorname{dist}(0, \partial_L f(y)) \leq u + \varepsilon$. Letting ε tend to zero, we conclude $\operatorname{dist}(0, \partial_L f(x)) \leq \overline{|\nabla f|}(x)$, as we had to show. □

Bibliography

[1] Tom M. Apostol. *Calculus. Vol. II: Multi-variable calculus and linear algebra, with applications to differential equations and probability.* Second edition. Blaisdell Publishing Co. Ginn and Co., Waltham, Mass.-Toronto, Ont.-London, 1969.

[2] Alexander Barvinok. *A course in convexity*, volume 54. American Mathematical Soc., 2002.

[3] Heinz H Bauschke and Patrick L Combettes. The baillon-haddad theorem revisited. *arXiv preprint arXiv:0906.0807*, 2009.

[4] H.H. Bauschke and P.L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces.* CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham, second edition, 2017. With a foreword by Hédy Attouch.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[6] Amir Beck. *First-order methods in optimization*, volume 25 of *MOS-SIAM Series on Optimization.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2017.

[7] Dimitri P Bertsekas and Athena Scientific. *Convex optimization algorithms.* Athena Scientific Belmont, 2015.

[8] J.M. Borwein and A.S. Lewis. *Convex analysis and nonlinear optimization.* CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 3. Springer-Verlag, New York, 2000. Theory and examples.

[9] J.M. Borwein and A.S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2006.

[10] Stephen Boyd and Lieven Vandenberghe. *Introduction to applied linear algebra: vectors, matrices, and least squares*. Cambridge university press, 2018.

[11] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[12] A. Daniilidis, A.S. Lewis, J. Malick, and H. Sendov. Prox-regularity of spectral functions and spectral sets. *J. Convex Anal.*, 15(3):547–560, 2008.

[13] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[14] Simon Fitzpatrick et al. Representing monotone operators by convex functions. In *Workshop/Miniconference on Functional Analysis and Optimization*, pages 59–65. Centre for Mathematics and its Applications, Mathematical Sciences Institute . . . , 1988.

[15] Gerald B. Folland. *Advanced Calculus*. First edition. Pearson, 2001.

[16] Paul R Halmos. *Finite-dimensional vector spaces*. Courier Dover Publications, 2017.

[17] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[18] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

[19] Guanghui Lan, Zhaosong Lu, and Renato DC Monteiro. Primal-dual first-order methods with $o(1/epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29, 2011.

[20] A.S. Lewis. Convex analysis on the Hermitian matrices. *SIAM J. Optim.*, 6(1):164–177, 1996.

[21] A.S. Lewis. Nonsmooth analysis of eigenvalues. *Math. Program.*, 84(1, Ser. A):1–24, 1999.

[22] A.S. Lewis and H.S. Sendov. Nonsmooth analysis of singular values. I. Theory. *Set-Valued Anal.*, 13(3):213–241, 2005.

[23] Zhaosong Lu. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19(4):1807–1827, 2009.

[24] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[25] B.S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*. Grundlehren der mathematischen Wissenschaften, Vol 330, Springer, Berlin, 2006.

[26] A.S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[27] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[28] Yu. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[29] Yu Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16(1):235–249, 2005.

[30] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[31] R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[32] RT Rockafellar. On the maximality of sums of nonlinear monotone operators. *Transactions of the American Mathematical Society*, 149(1):75–88, 1970.

[33] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.

[34] Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill Book Co., New York-Auckland-Düsseldorf, third edition, 1976. International Series in Pure and Applied Mathematics.

[35] Naum Zuselevich Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.

[36] S Simons and C Zălinescu. A new proof for rockafellar's characterization of maximal monotone operators. *Proceedings of the American Mathematical Society*, 132(10):2969–2972, 2004.

[37] Gilbert Strang, Gilbert Strang, Gilbert Strang, and Gilbert Strang. *Introduction to linear algebra*, volume 3. Wellesley-Cambridge Press Wellesley, MA, 1993.

[38] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 1, 2008.