Chapter 2
Convex Optimization Theory

Check for updates

Many machine learning tasks can be formulated as an optimization problem given in the form of

$$\min_{x \in X} f(x), \tag{2.0.1}$$

where f, x, and X denote the objective function, decision variables, and feasible set, respectively. Unfortunately, solving an optimization problem is challenging. In general, we cannot guarantee whether one can find an optimal solution, and if so, how much computational effort one needs. However, it turns out that we can provide such guarantees for a special but broad class optimization problems, namely convex optimization, where X is a convex set and f is a convex function. In fact, many machine learning models we formulated so far, such as least square linear regression, logistic regression, and support vector machine, are convex optimization problems.

Our goal in this chapter is to provide a brief introduction to the basic convex optimization theory, including convex sets, convex functions, strong duality, and KKT conditions. We will also briefly discuss some consequences of these theoretic results in machine learning, e.g., the representer theorem, Kernel trick, and dual support vector machine. We include proofs for some important results but the readers can choose to skip them for the first pass through the text.

2.1 Convex Sets

2.1.1 Definition and Examples

We begin with the definition of the notion of a convex set.

Definition 2.1. A set $X \subseteq \mathbb{R}^n$ is said to be convex if it contains all of its segments, that is

$$\lambda x + (1-\lambda) y \in X, \quad \forall (x, y, \lambda) \in X \times X \times [0,1].$$

© Springer Nature Switzerland AG 2020

G. Lan, *First-Order and Stochastic Optimization Methods for Machine Learning*,
Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-39568-1_2

Note that the point $\lambda x + (1-\lambda)y$ is called a convex combination of x and y. Figure 2.1 show the examples of a convex set (left) and a nonconvex set (right).

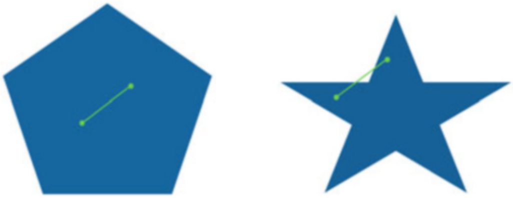

Fig. 2.1 Convex vs. nonconvex sets

It is easy to check that the following sets are convex.

(a) n-dimensional Euclidean space, $\mathbb{R}^n$. Given $x, y \in \mathbb{R}^n$, we must have $\lambda x + (1-\lambda)y \in \mathbb{R}^n$.

(b) Nonnegative orthant, $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x_i \geq 0, i = 1, \dots, n\}$. Let $x, y \in \mathbb{R}^n_+$ be given. Then for any $\lambda \in [0,1]$,

$$(\lambda x + (1-\lambda)y)_i = \lambda x_i + (1-\lambda)y_i \geq 0.$$

(c) Balls defined by an arbitrary norm, $\{x \in \mathbb{R}^n | \|x\| \leq 1\}$ (e.g., the l_2 norm $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ or l_1 norm $\|x\|_1 = \sum_{i=1}^n |x_i|$ balls). To show this set is convex, it suffices to apply the Triangular inequality and the positive homogeneity associated with a norm. Suppose that $\|x\| \leq 1, \|y\| \leq 1$ and $\lambda \in [0,1]$. Then

$$\|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\| = \lambda \|x\| + (1-\lambda)\|y\| \leq 1.$$

(d) Affine subspace, $\{x \in \mathbb{R}^n | Ax = b\}$. Suppose $x, y \in \mathbb{R}^n$, $Ax = b$, and $Ay = b$. Then

$$A(\lambda x + (1-\lambda)y) = \lambda Ax + (1-\lambda)Ay = b.$$

(e) Polyhedron, $\{x \in \mathbb{R}^n | Ax \leq b\}$. For any $x, y \in \mathbb{R}^n$ such that $Ax \leq b$ and $Ay \leq b$, we have

$$A(\lambda x + (1-\lambda)y) = \lambda Ax + (1-\lambda)Ay \leq b$$

for any $\lambda \in [0,1]$.

(f) The set of all positive semidefinite matrices S^n_+. S^n_+ consists of all matrices $A \in \mathbb{R}^{n \times n}$ such that $A = A^T$ and $x^T A x \geq 0$ for all $x \in \mathbb{R}^n$. Now consider $A, B \in \mathsf{S}^n_+$ and $\lambda \in [0,1]$. Then we must have

$$[\lambda A + (1-\lambda)B]^T = \lambda A^T + (1-\lambda)B^T = \lambda A + (1-\lambda)B.$$

Moreover, for any $x \in \mathbb{R}^n$,

$$x^T(\lambda A + (1-\lambda)B)x = \lambda x^T A x + (1-\lambda)x^T B x \geq 0.$$

(g) Intersections of convex sets. Let $X_i, i = 1,\ldots,k$, be convex sets. Assume that $x,y \in \cap_{i=1}^k X_i$, i.e., $x,y \in X_i$ for all $i = 1,\ldots,k$. Then for any $\lambda \in [0,1]$, we have $\lambda x + (1-\lambda)y \in X_i$ by the convexity of X_i, $i = 1,\ldots,k$, whence $\lambda x + (1-\lambda)y \in \cap_{i=1}^k X_i$.

(h) Weighted sums of convex sets. Let $X_1,\ldots,X_k \subseteq \mathbb{R}^n$ be nonempty convex subsets and $\lambda_1,\ldots,\lambda_k$ be reals. Then the set

$$\begin{aligned}&\lambda_1 X_1 + \ldots + \lambda_k X_k \\ \equiv\ &\{x = \lambda_1 x_1 + \ldots + \lambda_k x_k : x_i \in X_i, 1 \le i \le k\}\end{aligned}$$

is convex. The proof also follows directly from the definition of convex sets.

2.1.2 Projection onto Convex Sets

In this subsection we define the notion of projection over a convex set, which is important to the theory and computation of convex optimization.

Definition 2.2. Let $X \subset \mathbb{R}^n$ be a closed convex set. For any $y \in \mathbb{R}^n$, we define the closest point to y in X as:

$$\mathrm{Proj}_X(y) = \operatorname*{argmin}_{x \in X} \|y - x\|_2^2. \tag{2.1.2}$$

$\mathrm{Proj}_X(y)$ is called the projection of y onto X.

In the above definition, we require the set X to be closed in order to guarantee the existence of projection. On the other hand, if X is not closed, then the projection over X is not well defined. As an example, the projection of the point $\{2\}$ onto the interval $(0,1)$ does not exist. The existence of the projection over a closed convex set is formally stated as follows.

Proposition 2.1. *Let $X \subset \mathbb{R}^n$ be a closed convex set, and $y \in \mathbb{R}^n$ be given. Then* $\mathrm{Proj}_X(y)$ *must exist.*

Proof. Let $\{x_i\} \subseteq X$ be a sequence such that

$$\|y - x_i\|_2 \to \inf_{x \in X} \|y - x\|_2, \ i \to \infty.$$

The sequence $\{x_i\}$ clearly is bounded. Passing to a subsequence, we may assume that $x_i \to \bar{x}$ as $i \to \infty$. Since X is closed, we have $\bar{x} \in X$, and

$$\|y - \bar{x}\|_2 = \lim_{i \to \infty} \|y - x_i\|_2 = \inf_{x \in X} \|y - x\|_2.$$

■

The following result further shows that the projection onto a closed convex set X is unique.

Proposition 2.2. *Let X be a closed convex set, and $y \in \mathbb{R}^n$ be given. Then* $\mathrm{Proj}_X(y)$ *is unique.*

Proof. Let a and b be the two closet points in X to the given point y, so that $\|y-a\|_2 = \|y-b\|_2 = d$. Since X is convex, the point $z = (a+b)/2 \in X$. Therefore $\|y-z\|_2 \geq d$. We now have

$$\underbrace{\|(y-a)+(y-b)\|_2^2}_{=\|2(y-z)\|_2^2 \geq 4d^2} + \underbrace{\|(y-a)-(y-b)\|_2^2}_{=\|a-b\|^2} = \underbrace{2\|y-a\|_2^2 + 2\|y-b\|_2^2}_{4d^2},$$

whence $\|a-b\|_2 = 0$. Thus, the closest to y point in X is unique. ■

In many cases when the set X is relatively simple, we can compute $\mathrm{Proj}_X(y)$ explicitly. In fact, in Sect. 1.4, we computed the distance from a given point $y \in \mathbb{R}^n$ to a given hyperplane $H := \{x \in \mathbb{R}^n | w^T x + b = 0\}$ by using projection. Following the same reasoning (see Fig. 2.2a), we can write down

$$\mathrm{Proj}_H(y) = y - \tfrac{(w^T y + b) w}{\|w\|_2^2}.$$

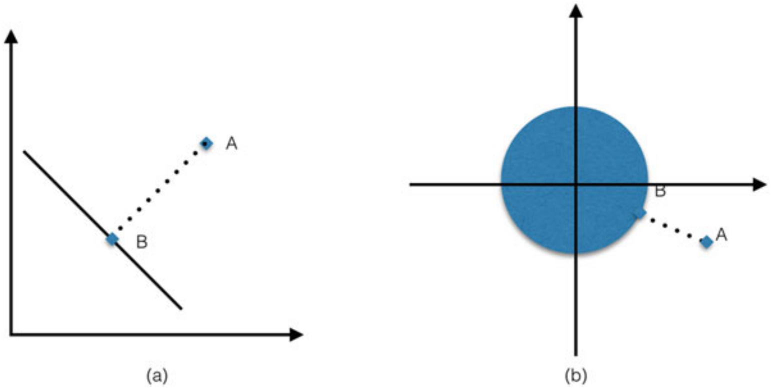

Fig. 2.2 Projection over convex sets

As another example, let us consider the projection of $y \in \mathbb{R}^n$ onto the standard Euclidean ball defined as $\mathsf{B} := \{x \in \mathbb{R}^n | \|x\|_2 \leq 1\}$ (see Fig. 2.2b). We can easily see that $\mathrm{Proj}_{\mathsf{B}}(y) = \frac{y}{\|y\|_2}$.

Projection over a convex set will be used extensively as a subroutine for solving more complicated optimization problems later in this book.

2.1.3 Separation Theorem

One fundamental result in convex analysis is the Separation Theorem. In this section, we will prove this theorem based on the projection onto a closed convex set and discuss some of its consequences.

We first discuss the separation of a point from a closed convex set.

Theorem 2.1. *Let $X \subseteq \mathbb{R}^n$ be a nonempty closed convex set, and a point $y \notin X$ be given. Then there exists $w \in \mathbb{R}^n$, $w \neq 0$ such that*

$$\langle w, y \rangle < \langle w, x \rangle, \ \ \forall x \in X.$$

Proof. Our proof is based on projecting y onto the set X. In particular, let $\text{Proj}_X(y)$ be defined in (2.1.2), we show that the vector $w = y - \text{Proj}_X(y)$ separates y and X. Note that $w \neq 0$ since $y \notin X$. Also let $x \in X$ be given and denote $z = tx + (1 - t)\text{Proj}_X(y)$ for any $t \in [0,1]$. Then we must have $z \in X$ and hence

$$\begin{aligned}\|y - \text{Proj}_X(y)\|_2^2 \le \|y - z\|^2 &= \|y - [tx + (1-t)\text{Proj}_X(y)]\|_2^2 \\ &= \|y - \text{Proj}_X(y) - t(x - \text{Proj}_X(y))\|_2^2 \\ &= \|w - t(x - \text{Proj}_X(y))\|_2^2.\end{aligned}$$

Define $\phi(t) := \|y - \text{Proj}_X(y) - t(x - \text{Proj}_X(y))\|_2^2$. It then follows from the above inequality that $\phi(0) \le \phi(t)$ for any $t \in [0,1]$. We have

$$0 \le \phi'(0) = -2w^T(x - \text{Proj}_X(y)),$$

which implies that

$$\forall x \in X : w^T x \le w^T \text{Proj}_X(y) = w^T(y - w) = w^T y - \|w\|_2^2.$$

■

We can generalize the above theorem to separate a closed convex set from another compact convex set.

Corollary 2.1. *Let X_1, X_2 be two nonempty closed convex sets and $X_1 \cap X_2 = \emptyset$. If X_2 is bounded, there exists $w \in \mathbb{R}^n$ such that*

$$\sup_{x \in X_1} w^T x < \inf_{x \in X_2} w^T x. \tag{2.1.3}$$

Proof. The set $X_1 - X_2$ is convex (the weighted sum of convex sets) and closed (the difference of a closed with a compact set). Moreover, $X_1 \cap X_2 = \emptyset$ implies $0 \notin X_1 - X_2$. So by Theorem 2.1, there exists w such that

$$\sup_{y \in X_1 - X_2} w^T y < w^T 0 = 0.$$

Or equivalently,

$$\begin{aligned} 0 &> \sup_{x_1 \in X_1, x_2 \in X_2} w^T(x_1 - x_2) \\ &= \sup_{x_1 \in X_1} w^T x_1 + \sup_{x_2 \in X_2} w^T(-x_2) \\ &= \sup_{x_1 \in X_1} w^T x_1 - \inf_{x_2 \in X_2} w^T x_2. \end{aligned}$$

Since X_2 is bounded, the last infimum becomes a min. Moreover, it is finite and can be moved to the left-hand side. ∎

When X_2 is unbounded, Corollary 2.1 may fail. One possible fix is to replace the strict inequality in (2.1.3) by an inequality. However, this might cause some problems. For example, consider the line segment connecting $(-1,0)$ and $(0,0)$, and the other one connecting $(-1,0)$ and $(2,0)$. Observe that the inner product between $(0,1)$ and any point from these line segments equals 0. The vector $w = (0,1)$ appears to "separate" these two line segments, while apparently they are not separable.

To address this issue, we say that a linear form $w^T x$ properly separates nonempty sets S and T if and only if

$$\begin{aligned} \sup_{x \in S} w^T x &\le \inf_{y \in T} w^T y \\ \inf_{x \in S} w^T x &< \sup_{y \in T} w^T y. \end{aligned} \tag{2.1.4}$$

In this case, the hyperplanes associated with w that separate S and T are exactly the hyperplanes

$$\{x : w^T x - b = 0\} \text{ with } \sup_{x \in S} w^T x \le b \le \inf_{y \in T} w^T y.$$

The proper separation property holds under quite general assumptions on the intersection $X_1 \cap X_2$. To state this more general result, we need to introduce the notion of *relative interior* $\mathrm{ri}(X)$, defined as the interior of X when we view it as subset of the affine subspace it generates. Without specific mention, we assume that the set X is full dimensional so that $\mathrm{int}(X) = \mathrm{ri}(X)$.

Theorem 2.2. *If the two nonempty convex sets X_1 and X_2 satisfy $\mathrm{ri}(X_1) \cap \mathrm{ri}(X_2) = \emptyset$, they can be properly separated.*

The above separation theorem can be derived from Theorem 2.1, but requiring us to establish a few technical results. We will first prove the result about the separability of a set in $\mathbb{R}^n$.

Lemma 2.1. *Every nonempty subset $S \subseteq \mathbb{R}^n$ is separable: one can find a sequence $\{x_i\}$ of points from S which is dense in S such that every point $x \in S$ is the limit of an appropriate subsequence of the sequence.*

Proof. Let $r_1, r_2, \ldots$ be the countable set of all rational vectors in $\mathbb{R}^n$. For every positive integer t, let $X_t \subset S$ be the countable set given by the following construction: we examine, one after another, at the points $r_1, r_2, \ldots$ and for every point r_s check whether there is a point $z \in S$ which is at most at the distance $1/t$ away from r_s. If points z with this property exist, we take one of them and add it to X_t and then pass to r_{s+1}, otherwise directly pass to r_{s+1}.

It is clear that every point $x \in S$ is at the distance at most $2/t$ from certain point of X_t. Indeed, since the rational vectors are dense in $\mathbb{R}^n$, there exists s such that r_s is at the distance $\leq \frac{1}{t}$ from x. Therefore, when processing r_s, we definitely add to X_t a point z which is at the distance $\leq 1/t$ from r_s and thus is at the distance $\leq 2/t$ from x. By construction, the countable union $\cup_{t=1}^{\infty} X_t$ of countable sets $X_t \subset S$ is a countable set in S, and by the fact that every point $x \in S$ is at most $2/t$ from X_t, this set is dense in S. ■

With the help of Lemma 2.1, we can refine the basic separation result stated in Theorem 2.1 by removing the "closedness" assumption, and using the notion of proper separation.

Proposition 2.3. *Let $X \subseteq \mathbb{R}^n$ be a nonempty convex set and $y \in \mathbb{R}^n, y \notin X$ be given. Then there exists $w \in \mathbb{R}^n$, $w \neq 0$ such that*

$$\sup_{x \in X} w^T x \leq w^T y,$$
$$\inf_{x \in X} w^T x < w^T y.$$

Proof. First note that we can perform the following simplification.

- Shifting X and $\{y\}$ by $-y$ (which clearly does not affect the possibility of separating the sets), we can assume that $\{0\} \not\subset X$.
- Replacing, if necessary, $\mathbb{R}^n$ with $\text{Lin}(X)$, we may further assume that $\mathbb{R}^n = \text{Lin}(X)$, i.e., the linear subspace generated by X.

In view of Lemma 2.1, let $\{x_i \in X\}$ be a sequence which is dense in X. Since X is convex and does not contain 0, we have

$$0 \notin \text{Conv}(\{x_1, \ldots, x_i\}) \; \forall i.$$

Noting that $\text{Conv}(\{x_1, \ldots, x_i\})$ are closed convex sets, we conclude from Theorem 2.1 that

$$\exists w_i : 0 = w_i^T 0 > \max_{1 \leq j \leq i} w_i^T x_j. \tag{2.1.5}$$

By scaling, we may assume that $\|w_i\|_2 = 1$. The sequence $\{w_i\}$ of unit vectors possesses a converging subsequence $\{w_{i_s}\}_{s=1}^{\infty}$ and the limit w of this subsequence is also a unit vector. By (2.1.5), for every fixed j and all large enough s we have $w_{i_s}^T x_j < 0$, whence

$$w^T x_j \leq 0 \; \forall j. \tag{2.1.6}$$

Since $\{x_j\}$ is dense in X, (2.1.6) implies that $w^T x \leq 0$ for all $x \in X$, and hence that

$$\sup_{x \in X} w^T x \leq 0 = w^T 0. \tag{2.1.7}$$

Now, it remains to verify that

$$\inf_{x \in X} w^T x < w^T 0 = 0.$$

Assuming the opposite, (2.1.7) would imply that $w^T x = 0$ for all $x \in X$, which is impossible, since $\mathrm{Lin}(X) = \mathbb{R}^n$ and w is nonzero. ■

We can now further show that two nonempty convex sets (not necessarily bounded or closed) can be properly separated.

Proposition 2.4. *If the two nonempty convex sets X_1 and X_2 satisfy $X_1 \cap X_2 = \emptyset$, they can be properly separated.*

Proof. Let $\widehat{X} = X_1 - X_2$. The set $\widehat{X}$ clearly is convex and does not contain 0 (since $X_1 \cap X_2 = \emptyset$). By Proposition 2.3, $\widehat{X}$ and $\{0\}$ can be separated: there exists f such that

$$\begin{aligned} \sup_{x \in X_1} w^T s - \inf_{y \in X_2} w^T y = \sup_{x \in X_1, y \in X_2} [w^T x - w^T y] \leq 0 = \inf_{z \in \{0\}} w^T z, \\ \inf_{x \in X_1} w^T x - \sup_{y \in X_2} w^T y = \inf_{x \in X_1, y \in X_2} [w^T x - w^T y] < 0 = \sup_{z \in \{0\}} w^T z, \end{aligned}$$

whence

$$\begin{aligned} \sup_{x \in X_1} w^T x \leq \inf_{y \in X_2} w^T y, \\ \inf_{x \in X_1} w^T x < \sup_{y \in X_2} w^T y. \end{aligned}$$

■

We are now ready to prove Theorem 2.2, which is even stronger than Proposition 2.4 in the sense that we only need $\mathrm{ri}(X_1) \cap \mathrm{ri}(X_2) = \emptyset$. In other words, these two sets can possibly intersect on their boundaries.

Proof of Theorem 2.2. The sets $X_1' = \mathrm{ri}(X_1)$ and $X_2' = \mathrm{ri}(X_2)$ are convex and nonempty, and these sets do not intersect. By Proposition 2.4, X_1' and X_2' can be separated: for properly chosen w, one has

$$\begin{aligned} \sup_{x \in X_1'} w^T x \leq \inf_{y \in X_2'} w^T y, \\ \inf_{x \in X_1'} w^T x < \sup_{y \in X_2'} w^T y. \end{aligned}$$

Since X_1' is dense in X_1 and X_2' is dense in X_2, inf's and sup's in the above relations remain the same when replacing X_1' with X_1 and X_2' with X_2. Thus, w separates X_1 and X_2. ■

In fact, we can show the reverse statement of Theorem 2.2 also holds.

Theorem 2.3. *If the two nonempty convex sets X_1 and X_2 can be properly separated, then* $\mathrm{ri}(X_1) \cap \mathrm{ri}(X_2) = \emptyset$.

Proof. We will first need to prove the following claim.

Claim. Let X be a convex set, $f(x) = w^T x$ be a linear form, and $a \in \mathrm{ri}(X)$. Then

$$w^T a = \max_{x \in X} w^T x \Leftrightarrow f(x) = \mathrm{const}\ \forall x \in X.$$

Indeed, shifting X, we may assume $a = 0$. Let, on the contrary to what should be proved, $w^T x$ be nonconstant on X, so that there exists $y \in X$ with $w^T y \neq w^T a = 0$. The case of $w^T y > 0$ is impossible, since $w^T a = 0$ is the maximum of $w^T x$ on X. Thus, $w^T y < 0$. Since $0 \in \mathrm{ri}(X)$, all points $z = -\varepsilon y$ belong to X, provided that $\varepsilon > 0$ is small enough. At every point of this type, $w^T z > 0$, which contradicts the fact that $\max_{x\in X} w^T x = w^T a = 0$.

Now let us use the above claim to prove our main result. Let $a \in \mathrm{ri}X_1 \cap \mathrm{ri}X_2$. Assume, on contrary to what should be proved, that $w^T x$ separates X_1 and X_2, so that

$$\sup_{x\in X_1} w^T x \leq \inf_{y\in X_2} w^T y.$$

Since $a \in X_2$, we get $w^T a \geq \sup_{x\in X_1} w^T x$, that is, $w^T a = \max_{x\in X_1} w^T x$. By the above claim, $w^T x = w^T a$ for all $x \in X_1$. Moreover, since $a \in X_1$, we get $w^T a \leq \inf_{y\in X_2} w^T y$, that is, $w^T a = \min_{y\in T} w^T y$. By the above claim, $w^T y = w^T a$ for all $y \in X_2$. Thus,

$$z \in X_1 \cup X_2 \Rightarrow w^T z \equiv w^T a,$$

so that w does not properly separate X_1 and X_2, which is a contradiction. ■

As a consequence of Theorem 2.2, we have the following supporting hyperplane theorem.

Corollary 2.2. *Let $X \subseteq \mathbb{R}^n$ be a convex set, and y be a point from its relative boundary. Then there exists $w \in \mathbb{R}^n$ and $w \neq 0$ such that*

$$\langle w,y\rangle \geq \sup_{x\in X}\langle w,x\rangle, \textit{ and } \langle w,y\rangle > \inf_{x\in X}\langle w,x\rangle.$$

The hyperplane $\{x|\langle w,x\rangle = \langle w,y\rangle\}$ is called a supporting hyperplane of X at y.

Proof. Since y is a point from the relative boundary of X, it is outside the relative interior of X and therefore $\{x\}$ and $\mathrm{ri}X$ can be separated by the Separation Theorem. The separating hyperplane is exactly the desired supporting hyperplane to X at y. ■

2.2 Convex Functions

2.2.1 Definition and Examples

Let $X \subseteq \mathbb{R}^n$ be a given convex set. A function $f : X \to \mathbb{R}$ is said to be convex if it always lies below its chords (Fig. 2.3), that is

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall (x,y,\lambda) \in X \times X \times [0,1]. \tag{2.2.8}$$

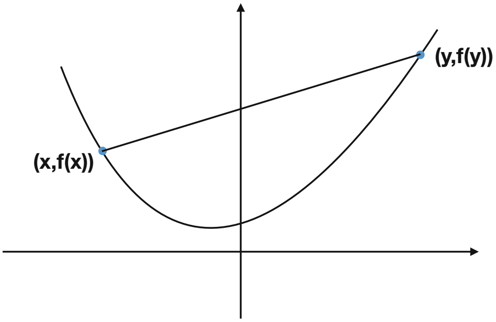

Fig. 2.3 The graph of a convex function

We say a function is strictly convex if (2.2.8) holds with strict inequality for any $x \neq y$ and $\lambda \in (0,1)$. We say that f is concave if $-f$ is convex, and similarly that f is strictly concave if $-f$ is strictly concave.

Some examples of convex functions are given as follows.

(a) Exponential, $f(x) = \exp(ax)$ for any $a \in \mathbb{R}$.
(b) Negative logarithm, $f(x) = -\log x$ with $x > 0$.
(c) Affine functions, $f(x) = w^T x + b$.
(d) Quadratic functions, $f(x) = \frac{1}{2}x^T Ax + b^T x$ with $A \in \mathsf{S}^n_+$ or $A \succeq 0$.
(e) Norms, $f(x) = \|x\|$.
(f) Nonnegative weighted sums of convex functions. Let $f_1, f_2, \ldots, f_k$ be convex functions and $w_1, w_2, \ldots, w_k$ be nonnegative real numbers. Then $f(x) = \sum_{i=1}^k w_i f_i(x)$ is a convex function.

2.2.2 Differentiable Convex Functions

Suppose that a function $f : X \to \mathbb{R}$ is differentiable over its domain. Then f is convex if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

for any $x, y \in X$, where ∇f denotes the gradients of f. The function $f(x) + \langle \nabla f(x), y - x \rangle$ is the first-order Taylor approximation of f at the point x. The above first-order condition for convexity says that f is convex if and only if the tangent line underestimates f everywhere in its domain. Similar to the definition of convexity, f will be strictly convex if this condition holds with strict inequality, concave if the inequality is reversed, and strictly concave if the reverse inequality is strict.

Suppose that a function $f : X \to \mathbb{R}$ is twice differentiable. Then f is convex if and only if its Hessian is positive semidefinite, i.e.,

$$\nabla^2 f(x) \succeq 0.$$

In one dimension, this is equivalent to the condition that the second-order derivative $f''(x)$ is nonnegative. Again analogous to both the definition and the first-order conditions for convexity, f is strictly convex if its Hessian is positive definite, concave if the Hessian is negative semidefinite, and strictly concave if the Hessian is negative definite. The function f is said to be strongly convex modulus μ with respect to the norm $\|\cdot\|$ if

$$f(y) \geq f(x) + \langle g, y-x \rangle + \tfrac{\mu}{2}\|y-x\|^2$$

for some $\mu > 0$. Clearly, strong convexity implies strict convexity.

2.2.3 Non-differentiable Convex Functions

Note that convex functions are not always differentiable everywhere over its domain. For example, the absolute value function $f(x) = |x|$ is not differentiable when $x = 0$. In this subsection, we will introduce an important notion about convex functions, i.e., subgradients, to generalize the gradients for differentiable convex functions.

Definition 2.3. $g \in \mathbb{R}^n$ is a subgradient of f at $x \in X$ if for any $y \in X$

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

The set of subgradients of f at x is called the subdifferential, denoted by $\partial f(x)$.

In order to show the existence of the subgradients for a convex function, we need to use the epigraph of a function $f : X \to \mathbb{R}$ given by

$$\text{epi}(f) = \{(x,t) \in X \times \mathbb{R} : f(x) \leq t\}.$$

It can be easily shown that f is convex if and only if epi(f) is a convex set.

The next result establishes the existence of subgradients for convex functions.

Proposition 2.5. *Let $X \subseteq \mathbb{R}^n$ be convex and $f : X \to \mathbb{R}$. If $\forall x \in X$, $\partial f(x) \neq \emptyset$, then f is convex. Moreover, if f is convex, then for any $x \in \text{ri}(X)$, $\partial f(x) \neq \emptyset$.*

Proof. The first claim is obvious. Let $g \in \partial f(\lambda x + (1-\lambda)y)$ for some $\lambda \in [0,1]$. Then by definition we have

$$\begin{aligned} f(y) &\geq f(\lambda x + (1-\lambda)y) + \lambda \langle g, y-x \rangle, \\ f(x) &\geq f(\lambda x + (1-\lambda)y) + (1-\lambda)\langle g, x-y \rangle. \end{aligned}$$

Multiplying the first inequality by $1-\lambda$ and the second one by λ, and then summing them up, we show the convexity of f.

We now show that f has subgradients in the interior of X. We will construct such a subgradient by using a supporting hyperplane to the epigraph of f. Let $x \in X$. Then $(x, f(x)) \in \text{epi}(f)$. By the convexity of epi(f) and the separating hyperplane theorem, there exists $(w,v) \in \mathbb{R}^n \times \mathbb{R}$ ($(w,v) \neq 0$) such that

$$\langle w,x\rangle + vf(x) \geq \langle w,y\rangle + vt, \ \ \forall (y,t) \in \mathrm{epi}(f). \tag{2.2.9}$$

Clearly, by tending t to infinity, we can see that $v \leq 0$. Now let us assume that x is in the interior of X. Then for $\varepsilon > 0$ small enough, $y = x + \varepsilon w \in X$, which implies that $v \neq 0$, since otherwise, we have $0 \geq \varepsilon \|w\|_2^2$ and hence $w = 0$, contradicting with the fact that $(w,v) \neq 0$. Letting $t = f(y)$ in (2.2.9), we obtain

$$f(y) \geq f(x) + \tfrac{1}{v}\langle w, y-x\rangle,$$

which implies that w/v is a subgradient of f at x. ■

Let f be a convex and differentiable function. Then by definition,

$$\begin{aligned} f(y) &\geq \tfrac{1}{\lambda}\left[f((1-\lambda)x+\lambda y) - (1-\lambda)f(x)\right] \\ &= f(x) + \tfrac{1}{\lambda}\left[f((1-\lambda)x+\lambda y) - f(x)\right]. \end{aligned}$$

Tending λ to 0, we show that $\nabla f(x) \in \partial f(x)$.

Below we provide some basic subgradient calculus for convex functions. Observe that many of them mimic the calculus for gradient computation.

(a) Scaling: $\partial(af) = a\partial f$ provided $a > 0$. The condition $a > 0$ makes function f remain convex.
(b) Addition: $\partial(f_1 + f_2) = \partial(f_1) + \partial(f_2)$.
(c) Affine composition: if $g(x) = f(Ax+b)$, then $\partial g(x) = A^T \partial f(Ax+b)$.
(d) Finite pointwise maximum: if $f(x) = \max_{i=1,\dots,m} f_i(x)$, then

$$\partial f(x) = \mathrm{conv}\left\{\cup_{i:f_i(x)=f(x)} \partial f_i(x)\right\},$$

which is the convex hull of union of subdifferentials of all active $i : f_i(x) = f(x)$ functions at x.
(e) General pointwise maximum: if $f(x) = \max_{s\in S} f_s(x)$, then under some regularity conditions (on S and f_s),

$$\partial f(x) = \mathrm{cl}\left\{\mathrm{conv}\left(\cup_{s:f_s(x)=f(x)} \partial f_s(x)\right)\right\}.$$

(f) Norms: important special case, $f(x) = \|x\|_p$. Let q be such that $1/p + 1/q = 1$, then

$$\partial f(x) = \{y : \|y\|_q \leq 1 \,\mathrm{and}\, y^T x = \max\{z^T x : \|z\|_q \leq 1\}.$$

Other notions of convex analysis will prove to be useful. In particular the notion of closed convex functions is convenient to exclude pathological cases: these are convex functions with closed epigraphs (see Sect. 2.4 for more details).

2.2.4 Lipschitz Continuity of Convex Functions

Our goal in this section is to show that convex functions are Lipschitz continuous inside the interior of its domain.

We will first show that a convex function is locally bounded.

Lemma 2.2. *Let f be convex and $x_0 \in \operatorname{int}\operatorname{dom} f$. Then f is locally bounded, i.e., $\exists \varepsilon > 0$ and $M(x_0, \varepsilon) > 0$ such that*

$$f(x) \le M(x_0, \varepsilon) \; \forall \, x \in B_\varepsilon(x_0) := \{x \in \mathbb{R}^n : \|x - x_0\|_2 \le \varepsilon\}.$$

Proof. Since $x_0 \in \operatorname{int}\operatorname{dom} f$, $\exists \varepsilon > 0$ such that the vectors $x_0 \pm \varepsilon e_i \in \operatorname{int}\operatorname{dom} f$ for $i = 1, \ldots, n$, where e_i denotes the unit vector along coordinate i. Also let $H_\varepsilon(x_0) := \{x \in \mathbb{R}^n : \|x - x_0\|_\infty \le \varepsilon\}$ denote the hypercube formed by the vectors $x_0 \pm \varepsilon e_i$. It can be easily seen that $B_\varepsilon(x_0) \subseteq H_\varepsilon(x_0)$ and hence that

$$\max_{x \in B_\varepsilon(x_0)} f(x) \le \max_{x \in H_\varepsilon(x_0)} f(x) \le \max_{i=1,\ldots,n} f(x_0 \pm \varepsilon e_i) =: M(x_0, \varepsilon).$$

■

Next we show that f is locally Lipschitz continuous.

Lemma 2.3. *Let f be convex and $x_0 \in \operatorname{int}\operatorname{dom} f$. Then f is locally Lipschitz, i.e., $\exists \varepsilon > 0$ and $\bar{M}(x_0, \varepsilon) > 0$ such that*

$$|f(y) - f(x_0)| \le \bar{M}(x_0, \varepsilon)\|x - y\|, \; \forall y \in B_\varepsilon(x_0) := \{x \in \mathbb{R}^n : \|x - x_0\|_2 \le \varepsilon\}. \tag{2.2.10}$$

Proof. We assume that $y \ne x_0$ (otherwise, the result is obvious). Let $\alpha = \|y - x_0\|_2/\varepsilon$. We extend the line segment connecting x_0 and y so that it intersects the ball $B_\varepsilon(x_0)$, and then obtain two intersection points z and u (see Fig. 2.4). It can be easily seen that

$$y = (1 - \alpha)x_0 + \alpha z, \tag{2.2.11}$$

$$x_0 = [y + \alpha u]/(1 + \alpha). \tag{2.2.12}$$

It then follows from the convexity of f and (2.2.11) that

$$\begin{aligned} f(y) - f(x_0) &\le \alpha[f(z) - f(x_0)] = \tfrac{f(z) - f(x_0)}{\varepsilon}\|y - x_0\|_2 \\ &\le \tfrac{M(x_0,\varepsilon) - f(x_0)}{\varepsilon}\|y - x_0\|_2, \end{aligned}$$

where the last inequality follows from Lemma 2.2. Similarly, by the convexity f, (2.2.11), and Lemma 2.2, we have

$$f(x_0) - f(y) \le \|y - x_0\|_2 \tfrac{M(x_0,\varepsilon) - f(x_0)}{\varepsilon}.$$

Combining the previous two inequalities, we show (2.2.10) holds with $\bar{M}(x_0, \varepsilon) = [M(x_0, \varepsilon) - f(x_0)]/\varepsilon$. ■

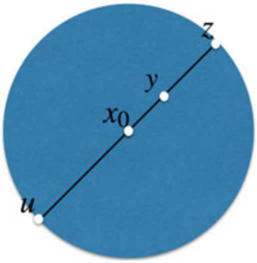

Fig. 2.4 Local Lipschitz continuity of a convex function

The following simple result shows the relation between the Lipschitz continuity of f and the boundedness of subgradients.

Lemma 2.4. *The following statements hold for a convex function f.*

(a) If $x_0 \in \operatorname{int}\operatorname{dom} f$ and f is locally Lipschitz (i.e., (2.2.10) holds), then $\|g(x_0)\| \le \bar{M}(x_0,\varepsilon)$ for any $g(x_0) \in \partial f(x_0)$.
(b) If $\exists g(x_0) \in \partial f(x_0)$ and $\|g(x_0)\|_2 \le \bar{M}(x_0,\varepsilon)$, then $f(x_0) - f(y) \le \bar{M}(x_0,\varepsilon)\|x_0 - y\|_2$.

Proof. We first show part (a). Let $y = x_0 + \varepsilon g(x_0)/\|g(x_0)\|_2$. By the convexity of f and (2.2.10), we have

$$\varepsilon\|g(x_0)\|_2 = \langle g(x_0), y - x_0\rangle \le f(y) - f(x_0) \le \bar{M}(x_0,\varepsilon)\|y - x_0\| = \varepsilon\bar{M}(x_0,\varepsilon),$$

which implies part (a). Part (b) simply follows the convexity of f, i.e.,

$$f(x_0) - f(y) \le \langle g(x_0), x_0 - y\rangle \le \bar{M}(x_0,\varepsilon)\|x_0 - y\|_2.$$

■

Below we state the global Lipschitz continuity of a convex function in its interior of domain.

Theorem 2.4. *Let f be a convex function and let K be a closed and bounded set contained in the relative interior of the domain* $\operatorname{dom} f$ *of f. Then f is Lipschitz continuous on K, i.e., there exists constant M such that*

$$|f(x) - f(y)| \le M_K\|x - y\|_2 \quad \forall x, y \in K. \tag{2.2.13}$$

Proof. The result directly follows from the local Lipschitz continuity of a convex function (see Lemmas 2.3 and 2.4) and the boundedness of K. ■

Remark 2.1. All three assumptions on K—i.e., (a) closedness, (b) boundedness, and (c) $K \subset \operatorname{ridom} f$—are essential, as it is seen from the following three examples:

- $f(x) = 1/x$, dom$f = (0,+\infty)$, $K = (0,1]$. We have (b), (c) but not (a); f is neither bounded, nor Lipschitz continuous on K.
- $f(x) = x^2$, dom$f = \mathbb{R}$, $K = \mathbb{R}$. We have (a), (c) and not (b); f is neither bounded nor Lipschitz continuous on K.
- $f(x) = -\sqrt{x}$, dom$f = [0,+\infty)$, $K = [0,1]$. We have (a), (b) and not (c); f is not Lipschitz continuous on K although is bounded. Indeed, we have $\lim_{t\to+0} \frac{f(0)-f(t)}{t} = \lim_{t\to+0} t^{-1/2} = +\infty$, while for a Lipschitz continuous f the ratios $t^{-1}(f(0)-f(t))$ should be bounded.

2.2.5 Optimality Conditions for Convex Optimization

The following results state the basic optimality conditions for convex optimization.

Proposition 2.6. *Let f be convex. If x is a local minimum of f, then x is a global minimum of f. Furthermore this happens if and only if $0 \in \partial f(x)$.*

Proof. It can be easily seen that $0 \in \partial f(x)$ if and only if x is a global minimum of f. Now assume that x is a local minimum of f. Then for $\lambda > 0$ small enough one has for any y,

$$f(x) \le f((1-\lambda)x + \lambda y) \le (1-\lambda)f(x) + \lambda f(y),$$

which implies that $f(x) \le f(y)$ and thus that x is a global minimum of f. ■

The above result can be easily generalized to the constrained case. Given a convex set $X \subseteq \mathbb{R}^n$ and a convex function $f : X \to \mathbb{R}$, we intend to

$$\min_{x\in X} f(x).$$

We first define the indicator function of the convex set X, i.e.,

$$I_X(x) := \begin{cases} 0, & x \in X, \\ \infty, & Otherwise. \end{cases}$$

By definition of subgradients, we can see that the subdifferential of I_X is given by the normal cone of X, i.e.,

$$\partial I_X(x) = \{w \in \mathbb{R}^n | \langle w, y-x\rangle \le 0, \forall y \in X\}. \tag{2.2.14}$$

Proposition 2.7. *Let $f : X \to \mathbb{R}$ be a convex function and X be a convex set. Then x^* is an optimal solution of $\min_{x\in X} f(x)$ if and only if there exists $g^* \in \partial f(x^*)$ such that*

$$\langle g^*, y-x^*\rangle \ge 0, \forall y \in X.$$

Proof. Clearly the problem is equivalent to $\min_{x\in\mathbb{R}^n} f(x)+I_X(x)$, where the I_X denotes the indicator function of X. The results then immediately follow from (2.2.14) and Proposition 2.6. ■

In particular, if $X=\mathbb{R}^n$, then we must have $0\in\partial f(x)$, which reduces to the case in Proposition 2.6.

2.2.6 Representer Theorem and Kernel

In this subsection, we introduce a very important application of the optimality condition for convex optimization in machine learning.

Recall that many supervised machine learning models can be written in the following form:

$$f^* := \min_{x\in\mathbb{R}^n} \left\{ f(x) := \textstyle\sum_{i=1}^N L(x^T u_i, v_i) + \lambda r(x) \right\}, \tag{2.2.15}$$

for some $\lambda \ge 0$. For the sake of simplicity we assume that $r(x)=\|x\|_2^2/2$. It turns out that under these assumptions, we can always write the solutions to problem (2.2.15) as a linear combination of the input variables u_i's as shown in the following statement.

Theorem 2.5. *The optimal solution of (2.2.15) with $r(x)=\|x\|_2^2/2$ can be written as*

$$x^* = \textstyle\sum_{i=1}^N \alpha_i u^{(i)}$$

for some real-valued weights α_i.

Proof. Let $L'(z,v)$ denote a subgradient of L w.r.t. z. Then by the chain rule of subgradient computation, the subgradients of f can be written in the form of

$$f'(x) = \textstyle\sum_{i=1}^N L'(x^T u^{(i)}) u^{(i)} + \lambda x.$$

Noting that $0\in\partial f(x^*)$ and letting $w_i = L'(x^T u^{(i)})$, there must exist w_i's such that

$$x = -\tfrac{1}{\lambda}\textstyle\sum_{i=1}^N w_i u^{(i)}.$$

The result then follows by setting $\alpha_i = -1/(\lambda w_i)$. ■

This result has some important consequence in machine learning. For any inner product of $x^T u$ in machine learning models, we can replace it with

$$x^T u = u^T x = \textstyle\sum_{i=1}^N \alpha_i (u^{(i)})^T u^{(i)},$$

and then view these α_i, $i=1,\ldots,N$, as unknown variables (or parameters).

More generally, we may consider a nonlinear transformation of our original input variables u. Recall in our regression example in Chap. 1, we have an input variable u, i.e., the rating of a friend (say Judy), and we can consider regression using the features u, u^2, and u^3 to obtain a cubic function. We can use $\phi(u)$ to define such a nonlinear mapping from the original input to a new feature space.

Rather than learning the parameters associated with the original input variables u, we may instead learn using these expanded features $\phi(u)$. To do so, we simply need to go over our previous models, and replace u everywhere in it with $\phi(u)$.

Since the model can be written entirely in terms of the inner products $\langle u, z\rangle$, we can replace all those inner products with $\langle \phi(u), \phi(z)\rangle$. Given a feature mapping ϕ, let us define the so-called kernel

$$K(u,z) = \phi(u)^T \phi(z).$$

Then, we replace everywhere we previously had $\langle u, z\rangle$ with $K(u,z)$. In particular, we can write the new objective function as

$$\begin{aligned}\Phi(\alpha) = f(x) &= \textstyle\sum_{i=1}^N L(x^T u^{(i)}, v^{(i)}) + \frac{\lambda}{2}\|x\|_2^2 \\ &= \textstyle\sum_{i=1}^N L\left(\phi(u^{(i)})^T \sum_{j=1}^N \alpha_j \phi(u^{(j)}), v_i\right) + \frac{\lambda}{2}\|\sum_{j=1}^N \alpha_j \phi(u^{(j)})\|_2^2 \\ &= \textstyle\sum_{i=1}^N L\left(\phi(u^{(i)})^T \sum_{j=1}^N \alpha_j \phi(u^{(j)}), v_i\right) \\ &\quad + \textstyle\frac{\lambda}{2}\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \phi(u^{(i)})^T \phi(u^{(j)}) \\ &= \textstyle\sum_{i=1}^N L\left(\sum_{j=1}^N \alpha_j K(u^{(i)}, u^{(j)}), v_i\right) \\ &\quad + \textstyle\frac{\lambda}{2}\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j K(u^{(i)}, u^{(j)}).\end{aligned}$$

In this way, we can write the objective function in terms of the Kernel matrix

$$K = \{K(u^{(i)}, v^{(j)})\}_{i,j=1}^N.$$

Even more interestingly, in many cases, we do not need to compute the nonlinear mapping $\phi(u)$ explicitly for every u, since the Kernel might be easier to compute than ϕ. One commonly used Kernel is the Gaussian or Radial Basis Function (RBF) kernel given by

$$K(u,z) = \exp\left(-\tfrac{1}{2\tau^2}\|u - z\|_2^2\right)$$

applicable to data in any dimension and the other one is the min-kernel given by $K(x,z) = \min(x,z)$ applicable to data in $\mathbb{R}$.

2.3 Lagrange Duality

In this section, we consider differentiable convex optimization problems of the form

$$\begin{array}{ll} f^* \equiv \min_{x\in X} & f(x) \\ \text{s.t.} & g_i(x) \le 0, i = 1, \ldots, m, \\ & h_j(x) = 0, j = 1, \ldots, p, \end{array} \tag{2.3.16}$$

where $X \subseteq \mathbb{R}^n$ is a closed convex set, $f : X \to \mathbb{R}$ and $g_i : X \to \mathbb{R}$ are differentiable convex functions, and $h_j : \mathbb{R}^n \to \mathbb{R}$ are affine functions. Our goal is to introduce Lagrange duality and a few optimality conditions for these convex optimization problems with function constraints.

2.3.1 Lagrange Function and Duality

We define the Lagrangian function L for (2.3.16) as

$$\mathrm{L}(x,\lambda,y) = f(x) + \textstyle\sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p y_j h_j(x),$$

for some $\lambda_i \in \mathbb{R}_+$, $i = 1,\ldots,m$, and $y_j \in \mathbb{R}, j = 1,\ldots,p$. These λ_i's and y_j's are called dual variables or Lagrange multipliers.

Intuitively, the Lagrangian function L can be viewed as a relaxed version of the objective function in the original problem (2.3.16) by allowing violation of the constraints ($g_i(x) \le 0$ and $h_j(x) \le 0$).

Let us consider the minimization of $\mathrm{L}(x,\lambda,y)$ w.r.t. x. Suppose that $\lambda \ge 0$ and $y \in \mathbb{R}^p$ are given. Let us define

$$\phi(\lambda,y) := \min_{x\in X} \mathrm{L}(x,\lambda,y).$$

Clearly, for any feasible point x to (2.3.16) (i.e., $x \in X$, $g_i(x) \le 0$, and $h_j(x) = 0$), we have

$$\begin{aligned}\mathrm{L}(x,\lambda,y) &= f(x) + \textstyle\sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p y_j h_j(x)\\ &= f(x) + \textstyle\sum_{i=1}^m \lambda_i g_i(x) \le f(x).\end{aligned}$$

In particular, letting $x = x^*$ be the optimal solution of (2.3.16), we must have

$$\phi(\lambda,y) \le f^*.$$

In other words, $\phi(\lambda,y)$ gives us a lower bound on the optimal value f^*. In order to obtain the strongest lower bound, we intend to maximize $\phi(\lambda,y)$ w.r.t. $\lambda \ge 0$ and $y \in \mathbb{R}^p$, and thus define the Lagrange dual as

$$\phi^* \equiv \max_{\lambda\ge 0,y} \left\{ \phi(\lambda,y) := \min_{x\in X} \mathrm{L}(x,\lambda,y) \right\}. \tag{2.3.17}$$

By construction, we must have

$$\phi^* \le f^*.$$

This relation is the so-called *weak duality*. What is more interesting is that under certain conditions, we have $\phi^* = f^*$ as stated in Theorem 2.6. The proof of this result, however, is more involved. Hence, we provide this proof separately in Sect. 2.3.2.

Theorem 2.6. *Suppose that (2.3.16) is below bounded and that there exists $\bar{x} \in \mathrm{int} X$ s.t. $g(\bar{x}) < 0$ and $h(\bar{x}) = 0$. Then the Lagrange dual is solvable and we must have*

$$\phi^* = f^*.$$

The above theorem says that as long as the primal problem (2.3.16) has a strictly feasible solution (called *Slater condition*), the optimal value for the Lagrange dual must be equal to the optimal value of the primal. This result is called *strong duality*. In practice, nearly all convex problems satisfy this type of constraint qualification, and hence the primal and dual problems have the same optimal value.

2.3.2 Proof of Strong Duality

In this subsection, we provide a proof for the strong duality for convex optimization. The proof follows from the separation theorem and the consequent convex theorem on alternatives. For the sake of simplicity, we focus on the case when there exist only nonlinear inequality constraints. The readers can easily adapt the proof to the case when affine constraints do exist, or even further refine the results if there only exist affine constraints.

Before proving Theorem 2.6, we will first establish the Convex Theorem on Alternative (CTA). Consider a system of constraints on x

$$\begin{aligned} f(x) &< c, \\ g_j(x) &\le 0, \; j = 1, \ldots, m, \\ x &\in X, \end{aligned} \tag{I}$$

along with system of constraints on λ:

$$\begin{aligned} \inf_{x \in X} [f(x) + \textstyle\sum_{j=1}^m \lambda_j g_j(x)] &\ge c, \\ \lambda_j &\ge 0, \; j = 1, \ldots, m. \end{aligned} \tag{II}$$

We first discuss the trivial part of the CTA.

Proposition 2.8. *If (II) is solvable, then (I) is insolvable.*

What is more interesting is that the reverse statement is also true under the slater condition.

Proposition 2.9. *If (I) is insolvable and the subsystem*

$$\begin{aligned} g_j(x) &< 0, \; j = 1, \ldots, m, \\ x &\in X \end{aligned}$$

is solvable, then (II) is solvable.

Proof. Assume that (I) has no solutions. Consider two sets S and T in $\mathbb{R}^{m+1}$:

$$S := \left\{ u \in \mathbb{R}^{m+1} : u_0 < c, u_1 \le 0, \ldots, u_m \le 0 \right\},$$
$$T := \left\{ u \in \mathbb{R}^{m+1} : \exists x \in X : \begin{array}{c} f(x) \le u_0 \\ g_1(x) \le u_1 \\ \ldots\ldots\ldots \\ g_m(x) \le u_m \end{array} \right\}.$$

First observe that S and T are nonempty convex sets. Moreover, S and T do not intersect (otherwise (I) would have a solution). By Theorem 2.2, S and T can be separated: $\exists (a_0, \ldots, a_m) \neq 0$ such that

$$\inf_{u \in T} a^T u \ge \sup_{u \in S} a^T u$$

or equivalently,

$$\inf_{x \in X} \inf_{\substack{u_0 \ge f(x) \\ u_1 \ge g_1(x) \\ \vdots \\ u_m \ge g_m(x)}} [a_0 u_0 + a_1 u_1 + \ldots + a_m u_m] \ge \sup_{\substack{u_0 < c \\ u_1 \le 0 \\ \vdots \\ u_m \le 0}} [a_0 u_0 + a_1 u_1 + \ldots + a_m u_m].$$

In order to bound the RHS, we must have $a \ge 0$, whence

$$\inf_{x \in X} [a_0 f(x) + a_1 g_1(x) + \ldots + a_m g_m(x)] \ge a_0 c. \tag{2.3.18}$$

Finally, we observe that $a_0 > 0$. Indeed, otherwise $0 \neq (a_1, \ldots, a_m) \ge 0$ and

$$\inf_{x \in X} [a_1 g_1(x) + \ldots + a_m g_m(x)] \ge 0,$$

while $\exists \bar{x} \in X : g_j(\bar{x}) < 0$ for all j. Now, dividing both sides of (2.3.18) by a_0, we have

$$\inf_{x \in X} \left[f(x) + \textstyle\sum_{j=1}^m \left(\frac{a_j}{a_0} \right) g_j(x) \right] \ge c.$$

By setting $\lambda_j = a_j / a_0$ we obtain the result. ■

We are now ready to prove the strong duality.

Proof of Theorem 2.6. The system

$$f(x) < f^*, \; g_j(x) \le 0, \; j = 1, \ldots, m, \; x \in X$$

has no solutions, while the system

$$g_j(x) < 0,\ j = 1, \dots, m,\ x \in X$$

has a solution. By CTA,

$$\exists \lambda^* \geq 0 : f(x) + \sum_j \lambda_j^* g_j(x) \geq f^*\ \forall x \in X,$$

whence

$$\phi(\lambda^*) \geq f^*.$$

Combined with Weak Duality, the above inequality says that

$$f^* = \phi(\lambda^*) = \phi^*.$$

■

2.3.3 Saddle Points

Now let us examine some interesting consequences of strong duality. In particular, we can derive a few optimality conditions for convex optimization in order to check whether an $x^* \in X$ is optimal to (2.3.16) or not.

The first one is given in the form of a pair of *saddle points.*

Theorem 2.7. *Let $x^* \in X$ be given.*

(a) If x^* can be extended, by a $\lambda^* \geq 0$ and $y^* \in \mathbb{R}^p$, to a saddle point of the Lagrange function on $X \times \{\lambda \geq 0\}$:

$$\mathrm{L}(x, \lambda^*, y^*) \geq \mathrm{L}(x^*, \lambda^*, y^*) \geq \mathrm{L}(x^*, \lambda, y)\ \forall (x \in X, \lambda \geq 0, y \in \mathbb{R}^p),$$

then x^* is optimal for (2.3.16).

(b) If x^* is optimal for (2.3.16) which is convex and satisfies the Slater condition, then x^* can be extended, by a $\lambda^* \geq 0$ and $y^* \in \mathbb{R}^p$, to a saddle point of the Lagrange function on $X \times \{\lambda \geq 0\} \times \mathbb{R}^p$.

Proof. We first prove part (a). Clearly,

$$\sup_{\lambda \geq 0, y} \mathrm{L}(x^*, \lambda, y) = \begin{cases} +\infty, & x^* \text{ is infeasible} \\ f(x^*), & \text{otherwise} \end{cases}$$

Thus, $\lambda^* \geq 0,\ \mathrm{L}(x^*, \lambda^*, y^*) \geq \mathrm{L}(x^*, \lambda, y)\ \forall \lambda \geq 0 \forall y$ is equivalent to

$$g_j(x^*) \leq 0 \forall j,\ \lambda_i^* g_i(x^*) = 0 \forall i,\ h_j(x^*) = 0 \forall j.$$

Consequently, $\mathrm{L}(x^*, \lambda^*, y^*) = f(x^*)$, hence

$$\mathrm{L}(x, \lambda^*, y^*) \geq \mathrm{L}(x^*, \lambda^*, y^*)\ \forall x \in X$$

reduces to

$$\mathrm{L}(x,\lambda^*,y^*) \geq f(x^*)\ \forall x.$$

Since for $\lambda \geq 0$ and y, one has $f(x) \geq \mathrm{L}(x,\lambda,y)$ for all feasible x, the above inequality then implies that

$$x \text{ is feasible } \Rightarrow f(x) \geq f(x^*).$$

We now show part (b). By Lagrange Duality, $\exists \lambda^* \geq 0$, y^*:

$$f(x^*) = \phi(\lambda^*,y^*) \equiv \inf_{x\in X}\left[f(x) + \Sigma_i \lambda_i^* g_i(x) + \Sigma_j y_j^* h_j(x)\right]. \qquad (2.3.19)$$

Since x^* is feasible, we have

$$\inf_{x\in X}\left[f(x) + \Sigma_i \lambda_i^* g_i(x) + \Sigma_j y_j^* h_j(x)\right] \leq f(x^*) + \Sigma_i \lambda_i^* g_i(x^*) \leq f(x^*).$$

By (2.3.19), the last " $\leq$ " here should be " $=$ ". This identity, in view of the fact that $\lambda^* \geq 0$, is possible if and only if $\lambda_j^* g_j(x^*) = 0 \forall j. Therefore, we have$

$$f(x^*) = \mathrm{L}(x^*,\lambda^*,y^*) \geq \mathrm{L}(x^*,\lambda,y)\ \forall \lambda \geq 0, \forall y \in \mathbb{R}^p,$$

where the last inequality follows from the definition of L (or weak duality). Now (2.3.19) reads $\mathrm{L}(x,\lambda^*,y^*) \geq f(x^*) = \mathrm{L}(x^*,\lambda^*,y^*)$. ■

2.3.4 Karush–Kuhn–Tucker Conditions

We are now ready to derive the Karush–Kuhn–Tucker (KKT) optimality conditions for convex programming.

Theorem 2.8. *Let (2.3.16) be a convex program, let x^* be its feasible solution, and let the functions f, $g_1,\ldots,g_m$ be differentiable at x^*. Then*

(a) Exist Lagrange multipliers $\lambda^* \geq 0$ and y^* such that the following KKT conditions

$$\nabla f(x^*) + \Sigma_{i=1}^m \lambda_i^* \nabla g_i(x^*) + \Sigma_{i=1}^p y_j^* \nabla h_j(x^*) \in N_X^*(x^*) \ \text{[stationarity]}$$
$$\lambda_i^* g_i(x^*) = 0,\ 1 \leq i \leq m \ \text{[complementary slackness]}$$
$$g_i(x^*) \leq 0, 1 \leq i \leq m; h_j(x^*) = 0,\ 1 \leq j \leq p \ \text{[primal feasibility]}$$

are sufficient for x^* to be optimal.

(b) If (2.3.16) satisfies restricted Slater condition: *$\exists \bar{x} \in \mathrm{rint}X : g_i(\bar{x}) \leq 0, h_j(\bar{x}) = 0$ for all constraints and $g_j(\bar{x}) < 0$ for all nonlinear constraints, then the KKT conditions are necessary and sufficient for x^* to be optimal.*

Proof. We first prove part (a). Indeed, complementary slackness plus $\lambda^* \geq 0$ ensure that

$$\mathrm{L}(x^*,\lambda^*,y^*) \geq \mathrm{L}(x^*,\lambda,y) \quad \forall \lambda \geq 0, \forall y \in \mathbb{R}^p.$$

Further, $\mathrm{L}(x,\lambda^*,y^*)$ is convex in $x \in X$ and differentiable at $x^* \in X$, so that the stationarity condition implies that

$$\mathrm{L}(x,\lambda^*,y^*) \geq \mathrm{L}(x^*,\lambda^*,y^*) \quad \forall x \in X.$$

Thus, x^* can be extended to a saddle point of the Lagrange function and therefore is optimal for (2.3.16).

We now show that part (b) holds. By Saddle Point Optimality condition, from optimality of x^* it follows that $\exists \lambda^* \geq 0$ and $y^* \in \mathbb{R}^p$ such that (x^*,λ^*,y^*) is a saddle point of $\mathrm{L}(x,\lambda,y)$ on $X \times \{\lambda \geq 0\} \times \mathbb{R}^p$. This is equivalent to $h_j(x^*) = 0$,

$$\lambda_i^* g_i(x^*) = 0 \,\forall i,$$

and

$$\min_{x \in X} \mathrm{L}(x,\lambda^*,y^*) = \mathrm{L}(x^*,\lambda^*,y^*).$$

Since the function $\mathrm{L}(x,\lambda^*,y^*)$ is convex in $x \in X$ and differentiable at $x^* \in X$, the last identity implies the stationarity condition. ■

Let us look at one example.

Example 2.1. Assuming $a_i > 0$, $p \geq 1$, show that the solution of the problem

$$\min_x \left\{ \sum_i \frac{a_i}{x_i} : x > 0, \sum_i x_i^p \leq 1 \right\}$$

is given by

$$x_i^* = \frac{a_i^{1/(p+1)}}{\left(\sum_j a_j^{p/(p+1)}\right)^{1/p}}.$$

Proof. Assuming $x^* > 0$ is a solution such that $\sum_i (x_i^*)^p = 1$, the KKT stationarity condition reads

$$\nabla_x \left\{ \sum_i \frac{a_i}{x_i} + \lambda \left(\sum_i x_i^p - 1\right) \right\} = 0 \Leftrightarrow \frac{a_i}{x_i^2} = p\lambda x_i^{p-1}$$

whence $x_i = [a_i/(p\lambda)]^{1/(p+1)}$. Since $\sum_i x_i^p$ should be 1, we get

$$x_i^* = \frac{a_i^{1/(p+1)}}{\left(\sum_j a_j^{p/(p+1)}\right)^{1/p}}.$$

This x^* is optimal because the problem is convex and the KKT conditions are satisfied at this point. ■

By examining the KKT conditions, we can obtain explicit solutions for many simple convex optimization problems, which can be used as subproblems in iterative algorithms for solving more complicated convex or even nonconvex optimization problems.

2.3.5 Dual Support Vector Machine

In this subsection, we discuss one interesting application of the optimality conditions for convex programming in support vector machines.

Recall that the support vector machine can be formulated as

$$\begin{array}{ll} \min_{w,b} & \frac{1}{2}\|w\|^2 \\ \text{s.t.} & v^{(i)}(w^T u^{(i)} + b) \geq 1, i = 1, \ldots, m. \end{array}$$

We can write the constraints equivalently as

$$g_i(w,b) = -v^{(i)}(w^T u^{(i)} + b) + 1 \leq 0.$$

Thus Lagrangian function L for our problem is given by

$$\mathrm{L}(w,b,\lambda) = \tfrac{1}{2}\|w\|^2 - \textstyle\sum_{i=1}^N \lambda_i [v^{(i)}(w^T u^{(i)} + b) - 1].$$

For fixed λ_i's, the problem is unconstrained. Let us minimize $\mathrm{L}(w,b,\lambda)$ w.r.t. w and b. Setting the derivatives of L w.r.t. w and b to zero, i.e.,

$$\nabla_w \mathrm{L}(w,b,\lambda) = w - \textstyle\sum_{i=1}^N \lambda_i v^{(i)} u^{(i)} = 0,$$

we have

$$w = \textstyle\sum_{i=1}^m \lambda_i v^{(i)} u^{(i)}. \tag{2.3.20}$$

Moreover, we have

$$\nabla_b \mathrm{L}(w,b,\lambda) = \textstyle\sum_{i=1}^m \lambda_i v^{(i)} = 0. \tag{2.3.21}$$

Plugging the above definition of w into $\mathrm{L}(w,b,\lambda)$, we obtain

$$\mathrm{L}(w,b,\lambda) = \textstyle\sum_{i=1}^N \lambda_i - \tfrac{1}{2}\sum_{i,j=1}^N v^{(i)} v^{(j)} \lambda_i \lambda_j (u^{(i)})^T u^{(j)} - b\sum_{i=1}^m \lambda_i v^{(i)}.$$

Since by (2.3.21), the last term must be zero, we have

$$\mathrm{L}(w,b,\lambda) = \textstyle\sum_{i=1}^N \lambda_i - \tfrac{1}{2}\sum_{i,j=1}^N v^{(i)} v^{(j)} \lambda_i \lambda_j (u^{(i)})^T u^{(j)}.$$

Therefore, we can write the dual SVM problem as

$$\begin{array}{ll} \max_\lambda & \sum_{i=1}^N \lambda_i - \frac{1}{2}\sum_{i,j=1}^N v^{(i)} v^{(j)} \lambda_i \lambda_j (u^{(i)})^T u^{(j)} \\ \text{s.t.} & \lambda_i \geq 0, i = 1, \ldots, m \\ & \sum_{i=1}^N \lambda_i v^{(i)} = 0. \end{array}$$

Once we find the optimal λ^*, we can use (2.3.20) to compute optimal w^*. Moreover, with optimal w^*, we can easily solve the primal problem to find the intercept term b as

$$b^* = -\frac{\max_{i:v^{(i)}=-1} w^{*T} v^{(i)} + \min_{i:v^{(i)}=1} w^{*T} v^{(i)}}{2}.$$

It is also interesting to observe that the dual problem only depends on the inner product and we can generalize it easily by using the Kernel trick (Sect. 2.2.6).

2.4 Legendre–Fenchel Conjugate Duality

2.4.1 Closure of Convex Functions

We can extend the domain of a convex function $f : X \to \mathbb{R}$ to the whole space $\mathbb{R}^n$ by setting $f(x) = +\infty$ for any $x \notin X$. In view of the definition of a convex function in (2.2.8), and our discussion about the epigraphs in Sect. 2.2, a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex if and only if its epigraph

$$\operatorname{epi}(f) = \{(x,t) \in \mathbb{R}^{n+1} : f(x) \le t\}$$

is a nonempty convex set.

As we know, closed convex sets possess many nice topological properties. For example, a closed convex set is comprised of the limits of all converging sequences of elements. Moreover, by the Separation Theorem, a closed and nonempty convex set X is the intersection of all closed half-spaces containing X. Among these half-spaces, the most interesting ones are the supporting hyperplanes touching X on the relative boundary.

In functional Language, the "closedness" of epigraph corresponds to a special type of continuity, i.e., the lower semicontinuity. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a given function (not necessarily convex). We say that f is lower semicontinuous (l.s.c.) at a point $\bar{x}$, if for every sequence of points $\{x_i\}$ converging to $\bar{x}$ one has

$$f(\bar{x}) \le \liminf_{i \to \infty} f(x_i).$$

Of course, $\liminf$ of a sequence with all terms equal to $+\infty$ is $+\infty$. f is called lower semicontinuous, if it is lower semicontinuous at every point.

A trivial example of a lower semicontinuous function is a continuous one. Note, however, that a semicontinuous function is not necessarily continuous. What it is obliged is to make only "jumps down." For example, the function

$$f(x) = \begin{cases} 0, & x \neq 0 \\ a, & x = 0 \end{cases}$$

is lower semicontinuous if $a \le 0$ ("jump down at $x = 0$ or no jump at all"), and is not lower semicontinuous if $a > 0$ ("jump up").

The following statement links lower semicontinuity with the geometry of the epigraph.

Proposition 2.10. *A function f defined on $\mathbb{R}^n$ and taking values from $\mathbb{R} \cup \{+\infty\}$ is lower semicontinuous if and only if its epigraph is closed (e.g., due to its emptiness).*

Proof. We first prove the "only if" part (from lower semicontinuity to closed epigraph). Let (x,t) be the limit of the sequence $\{x_i, t_i\} \subset \operatorname{epi} f$. Then we have $f(x_i) \leq t_i$. Thus the following relation holds: $t = \lim_{i\to\infty} t_i \geq \lim_{i\to\infty} f(x_i) \geq f(x)$.

We now show the "if" part (from closed epigraph to lower semicontinuity). Suppose for contradiction that $f(x) > \gamma > \lim_{i\to\infty} f(x_i)$ for some constant γ, where x_i converges to x. Then there exists a subsequence $\{x_{i_k}\}$ such that $f(x_{i_k}) \leq \gamma$ for all i_k. Since the epigraph is closed, then x must belong to this set, which implies that $f(x) \leq \gamma$, which is a contradiction. ∎

As an immediate consequence of Proposition 2.10, the upper bound

$$f(x) = \sup_{\alpha \in \mathscr{A}} f_\alpha(x)$$

of arbitrary family of lower semicontinuous functions is lower semicontinuous. Indeed, the epigraph of the upper bound is the intersection of the epigraphs of the functions forming the bound, and the intersection of closed sets always is closed.

Now let us look at proper lower semicontinuous (l.s.c.) convex functions. According to our general convention, a convex function f is proper if $f(x) < +\infty$ for at least one x and $f(x) > -\infty$ for every x. This implies that proper convex functions have convex nonempty epigraphs. Also as we just have seen, "lower semicontinuous" means "with closed epigraph." Hence proper l.s.c. convex functions always have closed convex nonempty epigraphs.

Similar to the fact that a closed convex set is intersection of closed half-spaces, we can provide an outer description of a proper l.s.c. convex function. More specifically, we can show that a proper l.s.c. convex function f is the upper bound of all its affine minorants given in the form of $t \geq d^T x - a$. Moreover, at every point $\bar{x} \in \operatorname{ri}\operatorname{dom} f$ from the relative interior of the domain f, f is even not the upper bound, but simply the maximum of its minorants: there exists an affine function $f_{\bar{x}}(x)$ which underestimates $f(x)$ everywhere in $\mathbb{R}^n$ and is equal to f at $x = \bar{x}$. This is exactly the first-order approximation $f(\bar{x}) + \langle g(\bar{x}), x - \bar{x}\rangle$ given by the definition of subgradients.

Now, what if the convex function is not lower semicontinuous (see Fig. 2.5)? A similar question also arises about convex sets—what to do with a convex set which is not closed? To deal with these convex sets, we can pass from the set to its closure and thus get a "normal" object which is very "close" to the original one. Specifically, while the "main part" of the original set—its relative interior—remains unchanged, we add a relatively small "correction," i.e., the relative boundary, to the set. The same approach works for convex functions: if a proper convex function f is not l.s.c. (i.e., its epigraph, being convex and nonempty, is not closed), we can "correct" the function—replace it with a new function with the epigraph being the closure of $\operatorname{epi}(f)$. To justify this approach, we should make sure that the closure of the epigraph of a convex function is also an epigraph of such a function.

Thus, we conclude that the closure of the epigraph of a convex function f is the epigraph of certain function, referred to as *the closure* $\operatorname{cl} f$ *of* f. Of course, this

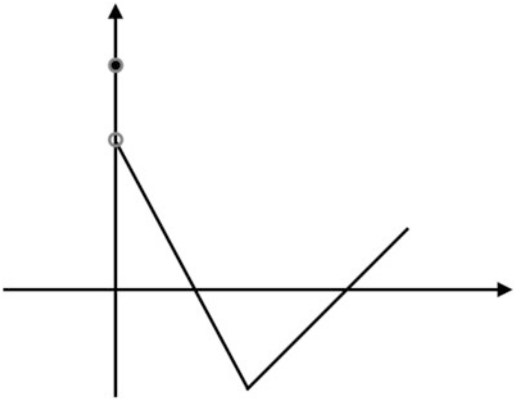

Fig. 2.5 Example for an upper semicontinuous function. The domain of this function is $[0,+\infty)$, and it "jumps up" at 0. However, the function is still convex

latter function is convex (its epigraph is convex—it is the closure of a convex set), and since its epigraph is closed, cl$f$ is proper. The following statement gives direct description of cl$f$ in terms of f:

(i) For every x one has $\mathrm{cl}f(x) = \lim\limits_{r\to+0} \inf\limits_{x':\|x'-x\|_2\le r} f(x')$. In particular,

$$f(x) \geq \mathrm{cl}f(x)$$

for all x, and

$$f(x) = \mathrm{cl}f(x)$$

whenever $x \in \mathrm{ridom}f$, or equivalently whenever $x \notin \mathrm{cldom}f$. Thus, the "correction" $f \mapsto \mathrm{cl}f$ may vary f only at the points from the relative boundary of dom$f$,

$$\mathrm{dom}f \subset \mathrm{domcl}f \subset \mathrm{cldom}f,$$

hence

$$\mathrm{ri}\,\mathrm{dom}f = \mathrm{ri}\,\mathrm{domcl}f.$$

(ii) The family of affine minorants of cl$f$ is exactly the family of affine minorants of f, so that

$$\mathrm{cl}f(x) = \sup\{\phi(x) : \phi \text{ is an affine minorant of } f\},$$

due to the fact that cl$f$ is l.s.c. and is therefore the upper bound of its affine minorants, and the sup in the right-hand side can be replaced with max whenever $x \in \mathrm{ri}\,\mathrm{domcl}f = \mathrm{ridom}f$.

2.4.2 Conjugate Functions

Let f be a convex function. We know that f "basically" is the upper bound of all its affine minorants. This is exactly the case when f is proper, otherwise the corresponding equality takes place everywhere except, perhaps, some points from the relative boundary of dom$f$. Now, when an affine function $d^T x - a$ is an affine minorant of f? It is the case if and only if

$$f(x) \geq d^T x - a$$

for all x or, which is the same, if and only if

$$a \geq d^T x - f(x)$$

for all x. We see that if the slope d of an affine function $d^T x - a$ is fixed, then in order for the function to be a minorant of f we should have

$$a \geq \sup_{x \in \mathbb{R}^n} [d^T x - f(x)].$$

The supremum in the right-hand side of the latter relation is certain function of d and this function, denoted by f^*, is called the *Legendre–Fenchel conjugate* of f:

$$f^*(d) = \sup_{x \in \mathbb{R}^n} [d^T x - f(x)].$$

Geometrically, the Legendre–Fenchel transformation answers the following question: given a slope d of an affine function, i.e., given the hyperplane $t = d^T x$ in $\mathbb{R}^{n+1}$, what is the minimal "shift down" of the hyperplane which places it below the graph of f?

From the definition of the conjugate it follows that this is a proper l.s.c. convex function. Indeed, we lose nothing when replacing $\sup_{x \in \mathbb{R}^n} [d^T x - f(x)]$ by $\sup_{x \in \mathrm{dom} f} [d^T x - f(x)]$, so that the conjugate function is the upper bound of a family of affine functions. This bound is finite at least at one point, namely, at every d coming from affine minorant of f, and we know that such a minorant exists. Therefore, f^* must be a proper l.s.c. convex function, as claimed.

The most elementary (and the most fundamental) fact about the conjugate function is its symmetry.

Proposition 2.11. *Let f be a convex function. Then $(f^*)^* = \mathrm{cl} f$. In particular, if f is l.s.c., then $(f^*)^* = f$.*

Proof. The conjugate function of f^* at the point x is, by definition,

$$\sup_{d \in \mathbb{R}^n} [x^T d - f^*(d)] = \sup_{d \in \mathbb{R}^n, a \geq f^*(d)} [d^T x - a].$$

The second sup here is exactly the supremum of all affine minorants of f due to the origin of the Legendre–Fenchel transformation: $a \geq f^*(d)$ if and only if the affine form $d^T x - a$ is a minorant of f. The result follows since we already know that the upper bound of all affine minorants of f is the closure of f. ■

The Legendre–Fenchel transformation is a very powerful tool—this is a “global” transformation, so that *local* properties of f^* correspond to *global* properties of f.

- $d = 0$ belongs to the domain of f^* if and only if f is below bounded, and if it is the case, then $f^*(0) = -\inf f$;
- if f is proper and l.s.c., then the subgradients of f^* at $d = 0$ are exactly the minimizers of f on $\mathbb{R}^n$;
- $\mathrm{dom} f^*$ is the entire $\mathbb{R}^n$ if and only if $f(x)$ grows, as $\|x\|_2 \to \infty$, faster than $\|x\|_2$: there exists a function $r(t) \to \infty$, as $t \to \infty$ such that

$$f(x) \geq r(\|x\|_2) \quad \forall x.$$

Thus, whenever we can compute explicitly the Legendre–Fenchel transformation of f, we get a lot of “global” information on f.

Unfortunately, the more detailed investigation of the properties of Legendre–Fenchel transformation is beyond our scope. Below we simply list some facts and examples.

- From the definition of Legendre transformation,

$$f(x) + f^*(d) \geq x^T d \quad \forall x, d.$$

  Specifying here f and f^*, we get certain inequality, e.g., the following one: [Young’s Inequality] *if p and q are positive reals such that $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$\frac{|x|^p}{p} + \frac{|d|^q}{q} \geq xd \quad \forall x, d \in \mathbb{R}$$

- The Legendre–Fenchel transformation of the function

$$f(x) \equiv -a$$

  is the function which is equal to a at the origin and is $+\infty$ outside the origin; similarly, the Legendre–Fenchel transformation of an affine function $\bar{d}^T x - a$ is equal to a at $d = \bar{d}$ and is $+\infty$ when $d \neq \bar{d}$;
- The Legendre–Fenchel transformation of the strictly convex quadratic form

$$f(x) = \frac{1}{2} x^T A x$$

  ($A \succeq 0$) is the quadratic form

$$f^*(d) = \frac{1}{2} d^T A^{-1} d$$

- The Legendre–Fenchel transformation of the Euclidean norm

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

  is the function which is equal to 0 in the closed unit ball centered at the origin and is $+\infty$ outside the ball.

2.5 Exercises and Notes

Exercises.

1. *Determine whether the following sets are convex or not.*

   (a) $\{x \in \mathbb{R}^2 : x_1 + i^2 x_2 \le 1, i = 1, \ldots, 10\}$.
   (b) $\{x \in \mathbb{R}^2 : x_1^2 + 2ix_1x_2 + i^2x_2^2 \le 1, i = 1, \ldots, 10\}$.
   (c) $\{x \in \mathbb{R}^2 : x_1^2 + ix_1x_2 + i^2x_2^2 \le 1, i = 1, \ldots, 10\}$.
   (d) $\{x \in \mathbb{R}^2 : x_1^2 + 5x_1x_2 + 4x_2^2 \le 1\}$.
   (e) $\{x \in \mathbb{R}^2 : \exp\{x_1\} \le x_2\}$.
   (f) $\{x \in \mathbb{R}^2 : \exp\{x_1\} \ge x_2\}$.
   (g) $\{x \in \mathbb{R}^n : \sum_{i=1}^n x_i^2 = 1\}$.

2. Assume that $X = \{x_1, \ldots, x_k\}$ and $Y = \{y_1, \ldots, y_m\}$ are finite sets in $\mathbb{R}^n$, with $k + m \ge n + 2$, and all the points $x_1, \ldots, x_k, y_1, \ldots, y_m$ are distinct. Assume that for any subset $S \subset X \cup Y$ comprised of $n + 2$ points the convex hulls of the sets $X \cap S$ and $Y \cap S$ do not intersect. Then the convex hulls of X and Y also do not intersect. (Hint: Assume on contrary that the convex hulls of X and Y intersect, so that

$$\sum_{i=1}^k \lambda_i x_i = \sum_{j=1}^m \mu_j y_j$$

   for certain nonnegative λ_i, $\sum_i \lambda_i = 1$, and certain nonnegative μ_j, $\sum_j \mu_j = 1$, and look at the expression of this type with the minimum possible total number of nonzero coefficients λ_i, μ_j.)

3. Prove that the following functions are convex on the indicated domains:

   (a) $\frac{x^2}{y}$ on $\{(x,y) \in \mathbb{R}^2 \mid y > 0\}$,
   (b) $\ln(\exp\{x\} + \exp\{y\})$ on the 2D plane.

4. A function f defined on a convex set Q is called log-convex on Q, if it takes real positive values on Q and the function $\ln f$ is convex on Q. Prove that

   (a) a log-convex on Q function is convex on Q,
   (b) the sum (more generally, linear combination with positive coefficients) of two log-convex functions on Q also is log-convex on the set.

5. Show the following statements related to the computation of subgradients.

   (a) The subgradient of $f(x) = \sqrt{x}$ does not exist at $x = 0$.
   (b) The subdifferential of $f(x) = |x|$ is given by $[-1, 1]$.
   (c) Let u and v be given. What is the subdifferential of $f(w,b) = \max\{0, v(w^T u + b)\} + \rho \|w\|_2^2$ at w and b.
   (d) The subdifferential of $f(x) = \|x\|$ is given by $\partial f(0) = \{x \in \mathbb{R}^n | \|x\| \le 1\}$ and $\partial f(x) = \{x / \|x\|\}$ for $x \ne 0$.

6. Find the minimizer of a linear function

$$f(x) = c^T x$$

on the set

$$V_p = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n |x_i|^p \leq 1\},$$

where p, $1 < p < \infty$, is a parameter. What happens with the solution when the parameter becomes 0.5?

7. Let $a_1, \ldots, a_n > 0$, $\alpha, \beta > 0$. Solve the optimization problem

$$\min_x \left\{ \sum_{i=1}^n \frac{a_i}{x_i^\alpha} : x > 0, \sum_i x_i^\beta \leq 1 \right\}.$$

8. Consider the optimization problem

$$\max_{x,y} \{f(x,y) = ax + by + \ln(\ln y - x) + \ln(y) : (x,y) \in X = \{y > \exp\{x\}\}\},$$

where $a, b \in \mathbb{R}$ are parameters. Is the problem convex? What is the domain in space of parameters where the problem is solvable? What is the optimal value? Is it convex in the parameters?

9. Let $a_1, \ldots, a_n$ be positive reals, and let $0 < s < r$ be two reals. Find maximum and minimum of the function

$$\sum_{i=1}^n a_i |x_i|^r$$

on the surface

$$\sum_{i=1}^n |x_i|^s = 1.$$

Notes. Further readings on convex analysis and convex optimization theory can be found on the monographs [50, 118], classic textbooks [12, 16, 79, 97, 104, 107, 120], and online course materials [92].