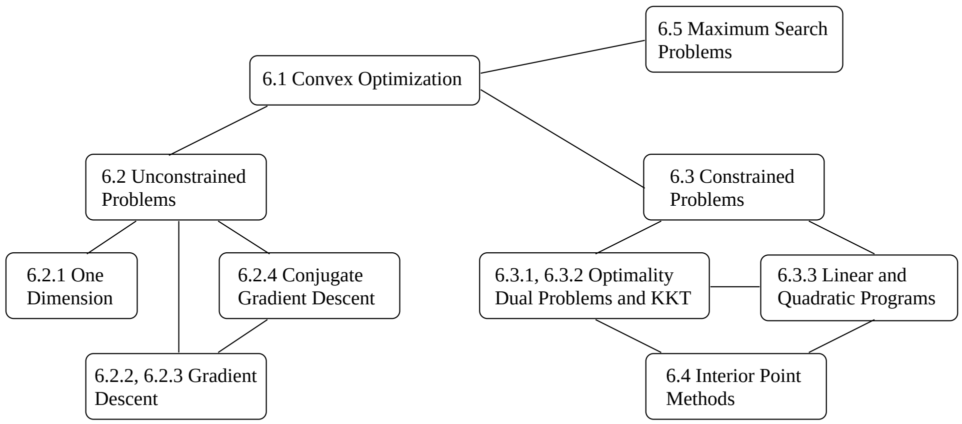


optimization problems (Section 6.2) are less common in this book and will only be required in the gradient descent methods of Section 10.6.1, and the Gaussian Process implementation methods of Section 16.4.

Prerequisites

The present chapter is intended as an introduction to the basic concepts of optimization. It is relatively self-contained, and requires only basic skills in linear algebra and multivariate calculus. Section 6.3 is somewhat more technical, Section 6.4 requires some additional knowledge of numerical analysis, and Section 6.5 assumes some knowledge of probability and statistics.

## 6.1 Convex Optimization

In the situations considered in this book, learning (or equivalently statistical estimation) implies the minimization of some risk functional such as $R_{\mathrm{emp}}[f]$ or $R_{\mathrm{reg}}[f]$ (cf. Chapter 4). While minimizing an arbitrary function on a (possibly not even compact) set of arguments can be a difficult task, and will most likely exhibit many local minima, minimization of a convex objective function on a convex set exhibits exactly one *global* minimum. We now prove this property.

Definition and Construction of Convex Sets and Functions

**Definition 6.1 (Convex Set)** *A set $X$ in a vector space is called convex if for any $x, x' \in X$ and any $\lambda \in [0,1]$, we have*

$$\lambda x + (1-\lambda)x' \in X. \tag{6.1}$$

**Definition 6.2 (Convex Function)** *A function $f$ defined on a set $X$ (note that $X$ need not be convex itself) is called convex if, for any $x, x' \in X$ and any $\lambda \in [0,1]$ such that $\lambda x + (1-\lambda)x' \in X$, we have*

$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda)f(x'). \tag{6.2}$$

*A function $f$ is called* strictly *convex if for $x \neq x'$ and $\lambda \in (0,1)$ (6.2) is a strict inequality.*

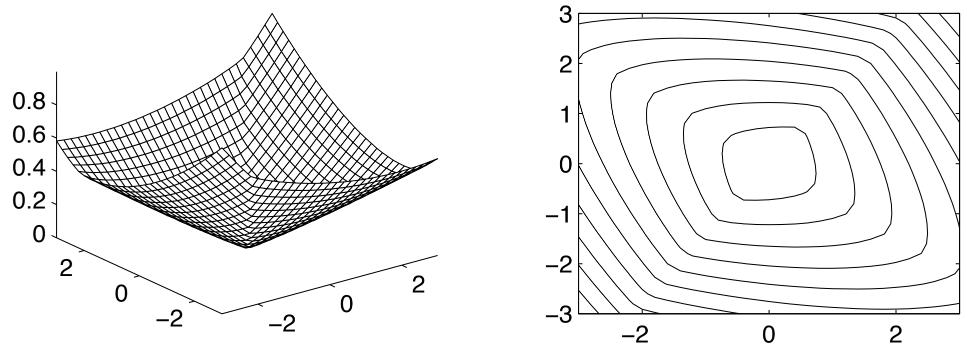


**Figure 6.1** Left: Convex Function in two variables. Right: the corresponding convex level sets $\{x|f(x) \leq c\}$, for different values of $c$.

There exist several ways to define convex sets. A convenient method is to define them via *below sets* of convex functions, such as the sets for which $f(x) \leq c$, for instance.

**Lemma 6.3 (Convex Sets as Below-Sets)** *Denote by $f : \mathcal{X} \to \mathbb{R}$ a convex function on a convex set $\mathcal{X}$. Then the set*

$$X := \{x|x \in \mathcal{X} \text{ and } f(x) \leq c\}, \text{ for all } c \in \mathbb{R}, \tag{6.3}$$

*is convex.*

***Proof*** We must show condition (6.1). For any $x, x' \in \mathcal{X}$, we have $f(x), f(x') \leq c$. Moreover, since $f$ is convex, we also have

$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x') \leq c \text{ for all } \lambda \in [0,1]. \tag{6.4}$$

Hence, for all $\lambda \in [0,1]$, we have $(\lambda x + (1-\lambda)x') \in X$, which proves the claim. Figure 6.1 depicts this situation graphically. ■

**Lemma 6.4 (Intersection of Convex Sets)** *Denote by $X, X' \subset \mathcal{X}$ two convex sets. Then $X \cap X'$ is also a convex set.*

Intersections

***Proof*** Given any $x, x' \in X \cap X'$, then for any $\lambda \in [0,1]$, the point $x_\lambda := \lambda x + (1-\lambda)x'$ satisfies $x_\lambda \in X$ and $x_\lambda \in X'$, hence also $x_\lambda \in X \cap X'$. ■

See also Figure 6.2. Now we have the tools to prove the central theorem of this section.

**Theorem 6.5 (Minima on Convex Sets)** *If the convex function $f : \mathcal{X} \to \mathbb{R}$ has a minimum on a convex set $X \subset \mathcal{X}$, then its arguments $x \in \mathcal{X}$, for which the minimum value is attained, form a convex set. Moreover, if $f$ is strictly convex, then this set will contain only one element.*

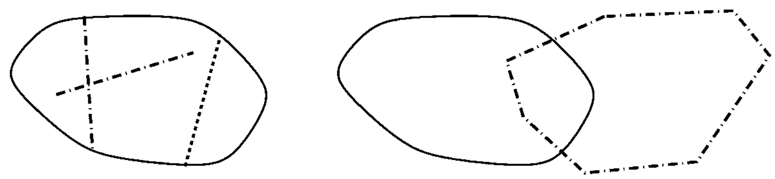

**Figure 6.2** Left: a convex set; observe that lines with points in the set are fully contained inside the set. Right: the intersection of two convex sets is also a convex set.

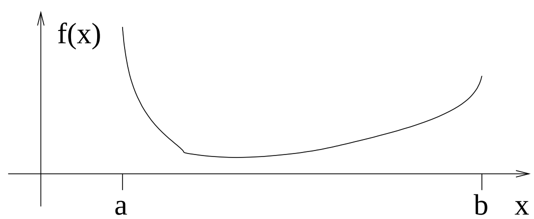


**Figure 6.3** Note that the maximum of a convex function is obtained at the ends of the interval $[a, b]$.

***Proof*** Denote by $c$ the minimum of $f$ on $X$. Then the set $X_m := \{x | x \in \mathcal{X} \text{ and } f(x) \leq c\}$ is clearly convex. In addition, $X_m \cap X$ is also convex, and $f(x) = c$ for all $x \in X_m \cap X$ (otherwise $c$ would not be the minimum).

If $f$ is strictly convex, then for any $x, x' \in X$, and in particular for any $x, x' \in X \cap X_m$, we have (for $x \neq x'$ and all $\lambda \in (0, 1)$),

$$f(\lambda x + (1-\lambda)x') < \lambda f(x) + (1-\lambda) f(x') = \lambda c + (1-\lambda)c = c. \tag{6.5}$$

This contradicts the assumption that $X_m \cap X$ contains more then one element. ■

Global Minima

A simple application of this theorem is in constrained convex minimization. Recall that the notation $[n]$, used below, is a shorthand for $\{1, \ldots, n\}$.

**Corollary 6.6 (Constrained Convex Minimization)** *Given the set of convex functions $f, c_1, \ldots, c_n$ on the convex set $\mathcal{X}$, the problem*

$$\begin{aligned} &\underset{x}{\text{minimize}} \quad f(x), \\ &\textit{subject to} \quad c_i(x) \leq 0 \textit{ for all } i \in [n], \end{aligned} \tag{6.6}$$

*has as its solution a convex set, if a solution exists. This solution is unique if $f$ is strictly convex.*

Many problems in Mathematical Programming or Support Vector Machines can be cast into this formulation. This means either that they all have unique solutions (if $f$ is strictly convex), or that all solutions are equally good and form a convex set (if $f$ is merely convex).

We might ask what can be said about convex *maximization*. Let us analyze a simple case first: convex maximization on an interval.

**Lemma 6.7 (Convex Maximization on an Interval)** *Denote by $f$ a convex function on $[a, b] \in \mathbb{R}$. Then the problem of maximizing $f$ on $[a, b]$ has $f(a)$ and $f(b)$ as solutions.*

Maxima on Extreme Points

***Proof*** Any $x \in [a, b]$ can be written as $\frac{b-x}{b-a}a + \left(1 - \frac{b-x}{b-a}\right) b$, and hence

$$f(x) \leq \frac{b-x}{b-a} f(a) + \left(1 - \frac{b-x}{b-a}\right) f(b) \leq \max(f(a), f(b)). \tag{6.7}$$

Therefore the maximum of $f$ on $[a, b]$ is obtained on one of the points $a, b$. ■

We will next show that the problem of convex *maximization* on a convex set is typically a hard problem, in the sense that the maximum can only be found at one of the extreme points of the constraining set. We must first introduce the notion of vertices of a set.

**Definition 6.8 (Vertex of a Set)** *A point $x \in X$ is a vertex of $X$ if, for all $x' \in X$ with $x' \neq x$, and for all $\lambda > 1$, the point $\lambda x + (1 - \lambda)x' \notin X$.*

This definition implies, for instance, that in the case of $X$ being an $\ell_2$ ball, the vertices of $X$ make up its surface. In the case of an $\ell_\infty$ ball, we have $2^n$ vertices in $n$ dimensions, and for an $\ell_1$ ball, we have only $2n$ of them. These differences will guide us in the choice of admissible sets of parameters for optimization problems (see, e.g., Section 14.4). In particular, there exists a connection between suprema on sets and their convex hulls. To state this link, however, we need to define the latter.

**Definition 6.9 (Convex Hull)** *Denote by $X$ a set in a vector space. Then the convex hull* $\text{co}\, X$ *is defined as*

$$\text{co}\, X := \left\{ \bar{x} \,\middle|\, \bar{x} = \sum_{i=1}^{n} \alpha_i x_i \text{ where } n \in \mathbb{N}, \alpha_i \geq 0 \text{ and } \sum_{i=1}^{n} \alpha_i = 1 \right\}. \tag{6.8}$$

**Theorem 6.10 (Suprema on Sets and their Convex Hulls)** *Denote by $X$ a set and by* $\text{co}\, X$ *its convex hull. Then for a convex function $f$*

$$\sup\{f(x) | x \in X\} = \sup\{f(x) | x \in \text{co}\, X\}. \tag{6.9}$$

Evaluating Convex Sets on Extreme Points

***Proof*** Recall that the below set of convex functions is convex (Lemma 6.3), and that the below set of $f$ with respect to $c = \sup\{f(x) | x \in X\}$ is by definition a superset of $X$. Moreover, due to its convexity, it is also a superset of $\text{co}\, X$. ■

This theorem can be used to replace search operations over sets $X$ by subsets $X' \subset X$, which are considerably smaller, if the convex hull of the latter generates $X$. In particular, the vertices of convex sets are sufficient to reconstruct the whole set.

**Theorem 6.11 (Vertices)** *A compact convex set is the convex hull of its vertices.*

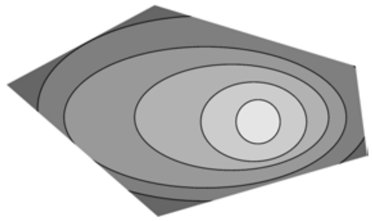

**Figure 6.4** A convex function on a convex polyhedral set. Note that the minimum of this function is unique, and that the maximum can be found at one of the vertices of the constraining domain.

Reconstructing Convex Sets from Vertices

The proof is slightly technical, and not central to the understanding of kernel methods. See Rockafellar [435, Chapter 18] for details, along with further theorems on convex functions. We now proceed to the second key theorem in this section.

**Theorem 6.12 (Maxima of Convex Functions on Convex Compact Sets)** *Denote by $X$ a compact convex set in $\mathcal{X}$, by $|X$ the vertices of $X$, and by $f$ a convex function on $X$. Then*

$$\sup\{f(x)|x \in X\} = \sup\{f(x)|x \in |X\}. \tag{6.10}$$

***Proof*** Application of Theorem 6.10 and Theorem 6.11 proves the claim, since under the assumptions made on $X$, we have $X = \mathrm{co}(|X)$. Figure 6.4 depicts the situation graphically. ■

## 6.2 Unconstrained Problems

After the characterization and uniqueness results (Theorem 6.5, Corollary 6.6, and Lemma 6.7) of the previous section, we will now study numerical techniques to obtain minima (or maxima) of convex optimization problems. While the choice of algorithms is motivated by applicability to kernel methods, the presentation here is not problem specific. For details on implementation, and descriptions of applications to learning problems, see Chapter 10.

### 6.2.1 Functions of One Variable

We begin with the easiest case, in which $f$ depends on only one variable. Some of the concepts explained here, such as the interval cutting algorithm and Newton's method, can be extended to the multivariate setting (see Problem 6.5). For the sake of simplicity, however, we limit ourselves to the univariate case.

Assume we want to minimize $f : \mathbb{R} \to \mathbb{R}$ on the interval $[a, b] \subset \mathbb{R}$. If we cannot make any further assumptions regarding $f$, then this problem, as simple as it may seem, cannot be solved numerically.

Continuous Differentiable Functions

If $f$ is *differentiable*, the problem can be reduced to finding $f'(x) = 0$ (see Problem 6.4 for the general case). If in addition to the previous assumptions, $f$ is convex, then $f'$ is nondecreasing, and we can find a fast, simple algorithm (Algorithm