Chapter 4 Algorithms for General Convex Problems

4.1 Introduction

In Chapter 3, we have addressed several special classes of convex problems that have found applications in engineering disciplines. There are however many problems in practice that do not fall into these classes. This chapter is devoted to studies of methods for general convex problems. These include Newton algorithms, proximal-point algorithms for composite convex functions, and alternating direction algorithms.

Several concepts and properties of convex functions that are needed in the development of these algorithms are introduced in Sec. 4.2. These include subgradients, convex functions with Lipschitz continuous gradients, strongly convex functions, conjugate functions, and proximal operators. Newton algorithms for unconstrained and constrained convex problems are addressed in Sec. 4.3. In Sec. 4.4, several algorithms for minimizing composite convex functions are studied. The l_1-l_2 minimization problem which finds applications in digital signal processing and machine learning is covered as a special case. Alternating direction methods have become increasingly important because of their ability to deal with large scale convex problems. In Sec. 4.5, we present two representative classes of alternating direction methods known as alternating direction methods of multipliers and alternating minimization algorithms.

4.2 Concepts and Properties of Convex Functions

The notion of convex functions and their elementary properties are addressed in Chapter 2. Here we introduce several additional concepts and properties of convex functions that are of use in the development of effective algorithms for convex problems.

4.2.1 Subgradients

Practical optimization problems involving non-differentiable objective functions and/or constraints are pervasive. The non-smoothness of the functions in an optimization problem may be encountered in several ways. For example, there are optimization problems with objective functions or constraints that are inherently non-differentiable. A simple example of the case is given by

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_1 \\ \text{subject to:} & \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2 \le \varepsilon \end{array}$$

where the l_1-norm of variable $\boldsymbol{x}$ is minimized subject to an l_2-norm constraint. Obviously, the objective function $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$ is continuous and convex, but *not differentiable*. There are also scenarios where the functions involved are differentiable, and it is the operation of these functions that yields non-smoothness. An example of the case is the objective function of the form

$$f(\boldsymbol{x}) = \max_{1\le j\le p}\{\varphi_j(\boldsymbol{x})\}$$

where $\varphi_j(\boldsymbol{x})$ are smooth convex functions, and the objective function is also convex but not necessarily differentiable, see Fig. 4.1 that depicts the objective function with three linear $\varphi_j(\boldsymbol{x})$.

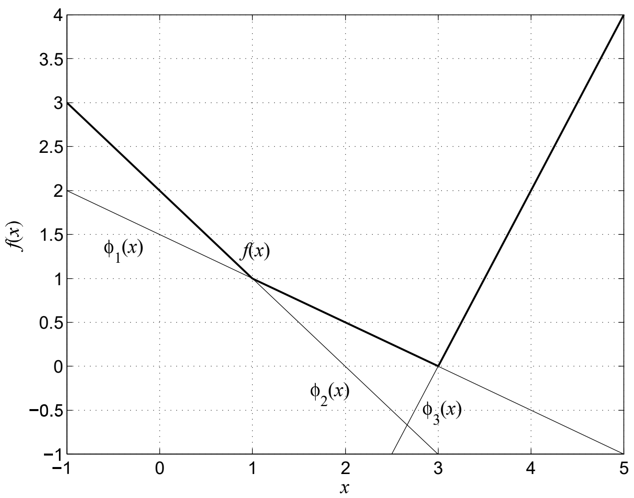

Figure 4.1. Pointwise maximum of three affine functions yields a piece-wise affine function which is convex but not differentiable.

Gradient plays an instrumental role in continuous optimization for functions that are differentiable. The concept of *subgradient* is a natural generalization of the concept of gradient that allows us to deal with optimization problems involving convex but non-differentiable functions [1]. Recall that the convexity of a differentiable function $f(\boldsymbol{x})$ can be characterized by

$$f(\tilde{\boldsymbol{x}}) \ge f(\boldsymbol{x}) + \nabla^T f(\boldsymbol{x})(\tilde{\boldsymbol{x}} - \boldsymbol{x}) \qquad \text{for} \;\; \boldsymbol{x}, \tilde{\boldsymbol{x}} \in \text{dom}(f) \tag{4.1}$$

where $\text{dom}(f)$ is the domain of function $f(\boldsymbol{x})$ that defines the set of points $\boldsymbol{x}$ where $f(\boldsymbol{x})$ assumes finite values. In geometric terms, Eq. (4.1) states that at any point $\boldsymbol{x}$ in the domain of a convex function $f(\boldsymbol{x})$, the tangent to the surface defined by $y = f(\boldsymbol{x})$ always lies below the surface.

Definition 4.1 If $f(\boldsymbol{x})$ is convex but not necessarily differentiable, then vector $\boldsymbol{g} \in R^n$ is said to be a *subgradient* of $f(\boldsymbol{x})$ at $\boldsymbol{x}$ if

$$f(\tilde{\boldsymbol{x}}) \ge f(\boldsymbol{x}) + \boldsymbol{g}^T (\tilde{\boldsymbol{x}} - \boldsymbol{x}) \quad \text{for} \;\; \tilde{\boldsymbol{x}} \in \text{dom}(f) \tag{4.2}$$

The subgradient of a convex function at point $\boldsymbol{x}$ where $f(\boldsymbol{x})$ is non-differentiable is *not* unique. The set of all subgradients at point $\boldsymbol{x}$ is called *subdifferential* of $f(\boldsymbol{x})$ and is denoted by $\partial f(\boldsymbol{x})$. ■

The right-hand side of Eq. (4.2) may be viewed as a linear lower bound of $f(\boldsymbol{x})$, and the subgradients at a point $\boldsymbol{x}$ where the convex function $f(\boldsymbol{x})$ is not differentiable correspond to different tangent lines at $\boldsymbol{x}$. This is illustrated in Fig. 4.2, where the two subgradients of $f(x)$ at x^* are given by $g_1 = \tan\theta_1$ and $g_2 = \tan\theta_2$. From the figure, it is obvious that any tangent line at x^* with a slop between g_2 and g_1 satisfies Eq. (4.2), therefore, any value $g \in [g_2, g_1]$ is a subgradient of $f(x)$ at x^*.

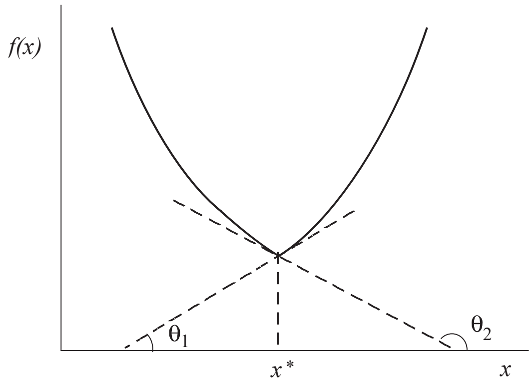

Figure 4.2. Two subgradients of $f(x)$ at x^* where $f(x)$ is not differentiable.

From Eq. (4.2), it follows that $f(\tilde{\boldsymbol{x}}) \geq f(\boldsymbol{x})$ as long as $\boldsymbol{g}^T(\tilde{\boldsymbol{x}} - \boldsymbol{x}) \geq 0$. Note that for a given point $\boldsymbol{x}$, $\boldsymbol{g}^T(\tilde{\boldsymbol{x}} - \boldsymbol{x}) = 0$ defines a hyperplane which passes through point $\boldsymbol{x}$ with $\boldsymbol{g}$ as its normal. This hyperplane divides space R^n into two parts with the hyperplane as boundary. In the part of the space where $\tilde{\boldsymbol{x}}$ satisfies $\boldsymbol{g}^T(\tilde{\boldsymbol{x}} - \boldsymbol{x}) > 0$, no minimizers exist because Eq. (4.2) in this case implies that $f(\tilde{\boldsymbol{x}}) > f(\boldsymbol{x})$. Consequently, a minimizer of $f(\boldsymbol{x})$ can only be found in the part of the space characterized by $\{\boldsymbol{x} : \boldsymbol{g}^T(\tilde{\boldsymbol{x}} - \boldsymbol{x}) \leq 0\}$. In this analysis, we see subgradient facilitates the construction of a cutting plane in the parameter space that reduces the search region significantly.

There are several important special cases in which the computation of a subgradient of a convex $f(\boldsymbol{x})$ can be readily carried out:

(a) If $f(\boldsymbol{x})$ is convex and differentiable at $\boldsymbol{x}$, then the subdifferential $\partial f(\boldsymbol{x})$ contains only one subgradient which is the same as the gradient $\nabla f(\boldsymbol{x})$;

(b) If $\alpha > 0$, a subgradient of $\alpha f(\boldsymbol{x})$ is given by $\alpha \boldsymbol{g}$ where $\boldsymbol{g}$ is a subgradient of $f(\boldsymbol{x})$;

(c) If $f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + \cdots + f_r(\boldsymbol{x})$ where $f_i(\boldsymbol{x})$ for $1 \leq i \leq r$ are convex, then $\boldsymbol{g} = \boldsymbol{g}_1 + \boldsymbol{g}_2 + \cdots + \boldsymbol{g}_r$ is a subgradient of $f(\boldsymbol{x})$ where $\boldsymbol{g}_i$ is a subgradient of $f_i(\boldsymbol{x})$;

(d) Define function $f(\boldsymbol{x}) = \max\left[f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_r(\boldsymbol{x})\right]$ where $f_i(\boldsymbol{x})$ for i = 1, 2, …, r are convex. At a given point $\boldsymbol{x}$, there exists at least one index, say i^*, such that $f(\boldsymbol{x}) = f_{i^*}(\boldsymbol{x})$. Then a subgradient of $f_{i^*}(\boldsymbol{x})$ is a subgradient of $f(\boldsymbol{x})$.

(e) If $h(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ where $f(\boldsymbol{x})$ is convex, then

$$\partial h(\boldsymbol{x}) = \boldsymbol{A}^T \partial f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$$

Example 4.1 Verify the formula of subdifferential of function $f(x) = |x|$

$$\partial |x| = \begin{cases} 1 & \text{for } x > 0 \\ \text{any } g \in [-1, 1] & \text{for } x = 0 \\ -1 & \text{for } x < 0 \end{cases} \tag{4.3}$$

Solution Function $f(x) = |x|$ is convex because for $0 \le \alpha \le 1$ we have

$$f(\alpha x_1 + (1-\alpha)x_2) = |\alpha x_1 + (1-\alpha)x_2| \le \alpha |x_1| + (1-\alpha)|x_2| = \alpha f(x_1) + (1-\alpha) f(x_2)$$

Consider a point x > 0, we have $f(x)$ = x which is obviously differentiable. Hence the differential of $f(x)$ is equal to the derivative of $f(x)$, i.e., $f'(x) = x' = 1$ which verifies the first line of the formula. Now consider a point x < 0, we have $f(x)$ = $-x$ which is differentiable. Hence the differential of $f(x)$ is equal to $f'(x) = (-x)' = -1$ which verifies the third line of the formula. At x = 0, Eq. (4.2) is reduced to $|\tilde{x}| \ge g\tilde{x}$ which holds for any g between −1 and 1. This verifies the second line of the formula in Eq. (4.3). ■

The next two theorems concern optimization problems where the functions involved are convex but not necessarily differentiable. These theorems may be regarded as extensions of the well-known first-order optimality condition and KKT conditions to their non-differentiable counterparts that are studied earlier in Chapter 2 and Chapter 10, respectively.

Theorem 4.1 ***First-order optimality condition for non-differentiable unconstrained convex problems*** Point $\boldsymbol{x}^*$ is a global solution of the minimization problem

$$\underset{\boldsymbol{x} \in \mathrm{dom}(f)}{\text{minimize}} \ f(\boldsymbol{x})$$

where $f(\boldsymbol{x})$ is convex, if and only if $\boldsymbol{0} \in \partial f(\boldsymbol{x}^*)$.

Proof Suppose $\mathbf{0} \in \partial f(\boldsymbol{x}^*)$. By letting $\boldsymbol{x} = \boldsymbol{x}^*$ and $\boldsymbol{g} = \mathbf{0}$ in Eq. (4.2), we obtain $f(\tilde{\boldsymbol{x}}) \ge f(\boldsymbol{x}^*)$ for all $\tilde{\boldsymbol{x}}$ in $\text{dom}(f)$, hence $\boldsymbol{x}^*$ is a global minimizer of $f(\boldsymbol{x})$. Conversely, if $\boldsymbol{x}^*$ is a global minimizer, then $f(\tilde{\boldsymbol{x}}) \ge f(\boldsymbol{x}^*)$ for all $\tilde{\boldsymbol{x}} \in \text{dom}(f)$, hence $f(\tilde{\boldsymbol{x}}) \ge f(\boldsymbol{x}^*) + \mathbf{0}^T(\tilde{\boldsymbol{x}} - \boldsymbol{x}^*)$ for all $\tilde{\boldsymbol{x}} \in \text{dom}(f)$, which implies that $\mathbf{0} \in \partial f(\boldsymbol{x}^*)$. ■

Theorem 4.2 ***KKT conditions for non-differentiable constrained convex problems*** Consider constrained convex problem

$$\begin{aligned} &\text{minimize} \quad f(\boldsymbol{x}) \\ &\text{subject to:} \quad \boldsymbol{a}_i^T \boldsymbol{x} = b_i \quad \text{for } 1 \le i \le p \\ &\qquad\qquad\quad c_j(\boldsymbol{x}) \le 0 \quad \text{for } 1 \le j \le q \end{aligned} \tag{4.4a-c}$$

where $f(\boldsymbol{x})$ and $c_j(\boldsymbol{x})$ are convex but not necessarily differentiable. A regular point $\boldsymbol{x}^*$ is a solution of the problem in Eq. (4.4) if and only if

(a) $\boldsymbol{a}_i^T \boldsymbol{x}^* = b_i$ for $1 \le i \le p$.

(b) $c_j(\boldsymbol{x}^*) \le 0$ for $1 \le j \le q$.

(c) there exist λ_i^* and μ_j^* such that $\mathbf{0} \in \partial f(\boldsymbol{x}^*) + \sum_{i=1}^{p} \lambda_i^* \boldsymbol{a}_i + \sum_{j=1}^{q} \mu_j^* \partial c_j(\boldsymbol{x}^*)$

(d) $\mu_j^* c_j(\boldsymbol{x}^*) = 0$ for $j = 1, 2, \ldots, q$.

(e) $\mu_j^* \ge 0$ for $j = 1, 2, \ldots, q$.

Proof Below we prove the sufficiency of these conditions for point $\boldsymbol{x}^*$ to be a global solution of the problem in Eq. (4.4), and leave the necessity part to the reader.

Suppose point $\boldsymbol{x}^*$ satisfies conditions (a) – (e). Conditions (a) and (b) imply that $\boldsymbol{x}^*$ is a feasible point of Eq. (4.4). Let $\boldsymbol{x}$ be an arbitrary feasible point for the problem in Eq. (4.4), below we show that $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})$.

Since both $\boldsymbol{x}$ and $\boldsymbol{x}^*$ are feasible, we can write

$$\sum_{i=1}^{p} \lambda_i^* \boldsymbol{a}_i^T (\boldsymbol{x} - \boldsymbol{x}^*) = \sum_{i=1}^{p} \lambda_i^* (\boldsymbol{a}_i^T \boldsymbol{x} - b_i) - \sum_{i=1}^{p} \lambda_i^* (\boldsymbol{a}_i^T \boldsymbol{x}^* - b_i) = 0 \tag{4.5}$$

Because $c_j(\boldsymbol{x}) \le 0$ and $\mu_j^* \ge 0$ for j = 1, 2, …, q, we have

$$\sum_{j=1}^{q} \mu_j^* c_j(\boldsymbol{x}) \le 0$$

which in conjunction with condition (d) implies that

$$\sum_{j=1}^{q} \mu_j^* \left(c_j(\boldsymbol{x}) - c_j(\boldsymbol{x}^*) \right) \le 0 \tag{4.6}$$

The convexity of functions $c_j(\boldsymbol{x})$ gives

$$c_j(\boldsymbol{x}) - c_j(\boldsymbol{x}^*) \ge \partial c_j(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*)$$

which leads Eq. (4.6) to

$$\sum_{j=1}^{q} \mu_j^* \partial c_j(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) \le 0 \tag{4.7}$$

From Eqs. (4.5), (4.7), and the convexity of $f(\boldsymbol{x})$, we deduce

$$\begin{aligned} & f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \\ & \ge \partial f(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) + \sum_{i=1}^{p} \lambda_i^* \boldsymbol{a}_i^T (\boldsymbol{x} - \boldsymbol{x}^*) + \sum_{j=1}^{q} \mu_j^* \partial c_j(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) \\ & = \left[\partial f(\boldsymbol{x}^*) + \sum_{i=1}^{p} \lambda_i^* \boldsymbol{a}_i + \sum_{j=1}^{q} \mu_j^* \partial c_j(\boldsymbol{x}^*) \right]^T (\boldsymbol{x} - \boldsymbol{x}^*) \end{aligned}$$

By condition (c), the expression in the last square bracket can be set to zero which leads to $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})$. ■

4.2.2 Convex functions with Lipschitz-continuous gradients

A continuously differentiable function $f(\boldsymbol{x})$ is said to have Lipschitz continuous gradient [2] if

$$\left\| \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \right\|_2 \le L \left\| \boldsymbol{x} - \boldsymbol{y} \right\|_2 \tag{4.8}$$

for any $\boldsymbol{x}$ and $\boldsymbol{y} \in \text{dom}(f)$, where $L > 0$ is called Lipschitz constant.

Example 4.2 Show that the gradient of $f(\boldsymbol{x}) = \frac{1}{2} \| \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \|_2^2$ is Lipschitz continuous.

Solution We write function $f(\boldsymbol{x})$ as

$$f(\boldsymbol{x}) = \tfrac{1}{2} \boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{b} + \tfrac{1}{2} \boldsymbol{b}^T \boldsymbol{b}$$

Hence the gradient of $f(\boldsymbol{x})$ is given by

$$\nabla f(\boldsymbol{x}) = \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x} - \boldsymbol{A}^T \boldsymbol{b}$$

and