SYS 6003: Optimization **Fall 2016**

Lecture 11&12

Instructor: Quanquan Gu *Date: Oct 5th*

Recall the gradient descent algorithm:

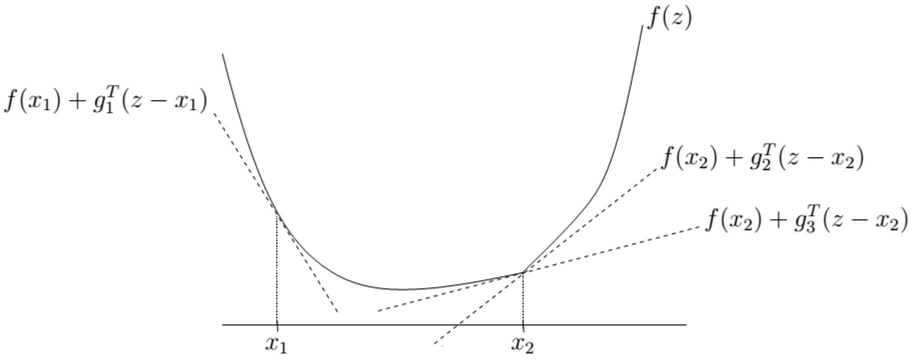
Algorithm 1 Gradient Descent

```
Input: $\eta_t, T$
Initialization: $\mathbf{x}_1$
for $t = 1$ to $T - 1$ do
    $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$
end for
```

What if f is not differentiable at $\mathbf{x}_t$ (i.e. $\nabla f(\mathbf{x}_t)$ does not exist)? Can we still apply this algorithm? In today's class, we are going to answer this question. We will introduce subgradient, which is a concept closely related to gradient. And for convex functions, we will show that even if the gradient does not exist, the subgradient always exists. Then, in convex optimization problems, when the function is nondifferential at certain point, we can use its subgradient as an alternative to the gradient.

First, we give the formal definition of subgradient.

Definition 1 (Subgradient) *A vector* $\mathbf{g} \in \mathbb{R}^d$ *is a subgradient of* $f : \mathbb{R}^d \to \mathbb{R}$ *at* $\mathbf{x} \in \mathbf{dom}\, f$*, if for any* $\mathbf{y} \in \mathbf{dom}\, f$*, we have* $f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x})$.


$f(z)$
$f(x_1) + g_1^T(z - x_1)$
$f(x_2) + g_2^T(z - x_2)$
$f(x_2) + g_3^T(z - x_2)$
x_1
x_2

Source: S. Boyd and L. Vandenberghe - Notes for EE364b, Stanford University, Winter 2006-07

Remark 1 *Clearly, if* f *is differentiable at* $\mathbf{x} \in \mathbf{dom}\, f$*, then its subgradient at* $\mathbf{x}$*,* $\mathbf{g}$*, is equal to* $\nabla f(\mathbf{x})$*. A function* f *is called subdifferentiable at* $\mathbf{x} \in \mathbf{dom}\, f$ *if there exists at least one subgradient at* $\mathbf{x}$*.*

Definition 2 (Subdifferential) *The set of subgradients of* f *at* $\mathbf{x} \in \mathbf{dom}\, f$ *is called the subdifferential of* f *at* $\mathbf{x}$*, denoted by* $\partial f(\mathbf{x})$*.*

Theorem 1 *The subdifferential of $f : \mathbb{R}^d \to \mathbb{R}$ at $\mathbf{x} \in \mathbf{dom}\, f$ is a closed and convex set.*

Proof: For any $\mathbf{g}_1, \mathbf{g}_2 \in \partial f(\mathbf{x})$ and any $\alpha \in [0, 1]$. We want to show $\alpha\mathbf{g}_1 + (1-\alpha)\mathbf{g}_2 \in \partial f(\mathbf{x})$. By definition of subgradients, we have

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \mathbf{g}_1^\top(\mathbf{y} - \mathbf{x}), \\ f(\mathbf{y}) &\geq f(\mathbf{x}) + \mathbf{g}_2^\top(\mathbf{y} - \mathbf{x}). \end{aligned}$$

Multiplying the first inequality by α and the second one by $(1 - \alpha)$, and then adding them together leads to

$$\begin{aligned} \alpha f(\mathbf{y}) + (1-\alpha)f(\mathbf{y}) &\geq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{x}) + [\alpha\mathbf{g}_1 + (1-\alpha)\mathbf{g}_2]^\top(\mathbf{y} - \mathbf{x}) \\ &= f(\mathbf{x}) + [\alpha\mathbf{g}_1 + (1-\alpha)\mathbf{g}_2]^\top(\mathbf{y} - \mathbf{x}) \end{aligned}$$

Thus, by definiton, we have $\alpha\mathbf{g}_1 + (1-\alpha)\mathbf{g}_2 \in \partial f(\mathbf{x})$ ■

We now turn to an important theorem, which shows that for a convex function, its subdifferential is always nonempty set.

Theorem 2 (Existence of subgradient for convex functions) *: If $f : \mathbb{R}^d \to \mathbb{R}$ is convex, and $\mathbf{x} \in \mathbf{int\ dom}\, f$, then $\partial f(\mathbf{x})$ is nonempty.*

The following theorem is very useful and is simple to proof.

Theorem 3 *A point $\mathbf{x}^*$ is a minimizer of a convex function $f : \mathbb{R}^d \to \mathbb{R}$, if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*)$*

Proof: If $\mathbf{x}^*$ is the minimizer of f, then for any $\mathbf{y} \in \mathbf{dom}\, f$, we have $f(\mathbf{y}) \geq f(\mathbf{x}^*)$, which can be rewritten as $f(\mathbf{y}) \geq f(\mathbf{x}^*) + \mathbf{0}^\top(\mathbf{y} - \mathbf{x}^*)$. Therefore, by definition, $\mathbf{0}$ is a subgradient of f at $\mathbf{x}^*$ and $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

Conversely, if $\mathbf{0} \in \partial f(\mathbf{x}^*)$, then for any $\mathbf{y} \in \mathbf{dom}\, f$, we have

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \mathbf{0}^\top(\mathbf{y} - \mathbf{x}^*) = f(\mathbf{x}^*).$$

Clearly, by definition, $\mathbf{x}^*$ is a minimizer of f. ■

Now let us see some examples of subgradients.

Example 1 *For $f(x) = |x|$,*

$$\partial f(x) = \begin{cases} \{1\} & \text{if } x > 0 \\ \{-1\} & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}.$$

Note that $f(x)$ is differentiable when $x > 0$ or $x < 0$, so the subgradient of f is equal to its gradient (1 and -1, respectively). At $x = 0$, for any $y \in \mathbb{R}$, its subgradient g should satisfy

$$\begin{aligned} |y| &\geq |0| + g(y - 0), \quad \textit{i.e.}, \\ g \cdot y &\leq |y|. \end{aligned}$$

Thus, $g \in [-1, 1]$.

Example 2 *For $f(\mathbf{x}) = \|\mathbf{x}\|_2, \mathbf{x} \in \mathbb{R}^d$,*

$$\partial f(\mathbf{x}) = \begin{cases} \{\mathbf{x}/\|\mathbf{x}\|_2\} & \text{if } \mathbf{x} \neq \mathbf{0} \\ \{\mathbf{g} : \|\mathbf{g}\|_2 \leq 1\} & \text{if } \mathbf{x} = \mathbf{0} \end{cases}.$$

When $\mathbf{x} \neq \mathbf{0}$, $f(\mathbf{x})$ is differentiable and its gradient can be easily obtained by applying the chain rule. At $\mathbf{x} = \mathbf{0}$, the subgradient $\mathbf{g}$ of $f(\mathbf{x})$ should satisfy

$$\|\mathbf{y}\|_2 \geq \|\mathbf{0}\|_2 + \mathbf{g}^\top(\mathbf{y} - \mathbf{0}) \quad \textit{for any } \mathbf{y} \in \mathbb{R}^d.$$

The inequality above can be written as

$$\mathbf{g}^\top \frac{\mathbf{y}}{\|\mathbf{y}\|_2} \leq 1.$$

We argue that $\|\mathbf{g}\|_2$ must be not greater than 1. If not, let $\mathbf{y} = \mathbf{g}$, then

$$\mathbf{g}^\top \frac{\mathbf{y}}{\|\mathbf{y}\|_2} = \frac{\mathbf{g}^\top \mathbf{g}}{\|\mathbf{g}\|_2} = \|\mathbf{g}\|_2 > 1,$$

which is a contradiction. Thus, $\partial f(\mathbf{0}) = \{\mathbf{g} : \|\mathbf{g}\|_2 \leq 1\}$.

A natural extension from the absolute value in the one dimensional case is the ℓ_1 norm of d dimensional vectors.

Example 3 *Let $f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i|$, and $\mathbf{g} = (g_1, \ldots, g_d)^\top$ be a subgradient of f at $\mathbf{x}$. Then we have*

$$g_i = \begin{cases} 1 & \text{if } x_i > 0 \\ -1 & \text{if } x_i < 0 \\ \in [-1, 1] & \text{if } x_i = 0 \end{cases}, i = 1, \ldots, d.$$

Finally, consider the subdifferential of point-wise maximum function.

Example 4 *Let $f(\mathbf{x}) = \max(f_1(\mathbf{x}), f_2(\mathbf{x}))$, where f_1 and f_2 are convex and differentiable. Then we have*

$$\partial f(\mathbf{x}) = \begin{cases} \{\nabla f_1(\mathbf{x})\} & \text{if } f_1(\mathbf{x}) > f_2(\mathbf{x}) \\ \{\nabla f_2(\mathbf{x})\} & \text{if } f_2(\mathbf{x}) > f_1(\mathbf{x}) \\ \mathbf{conv}\left(\{\nabla f_1(\mathbf{x}), \nabla f_2(\mathbf{x})\}\right) & \text{if } f_1(\mathbf{x}) = f_2(\mathbf{x}) \end{cases}.$$

Clearly, when $f_1(\mathbf{x}) \neq f_2(\mathbf{x})$, $f(\mathbf{x})$ is equal to either $f_1(\mathbf{x})$ or $f_2(\mathbf{x})$ and is therefore differentiable. When $f_1(\mathbf{x}) = f_2(\mathbf{x})$, both $\nabla f_1(\mathbf{x})$ and $\nabla f_2(\mathbf{x})$ are subgradients of $f(\mathbf{x})$. Since subdifferential is a convex set (see Theorem 1), $\partial f(\mathbf{x})$ should be a convex hull of $\nabla f_1(\mathbf{x})$ and $\nabla f_2(\mathbf{x})$.

Therefore, it is essential to calculate the subgradient for non-differential functions. For this purpose we have the following rules in calculating the subgradient.

1. (Non-negative Scaling) Suppose $f(\mathbf{x})$ is convex its domain **dom** f and that $\alpha > 0$ then

   $$\partial[\alpha f(\mathbf{x})] = \alpha \partial f(\mathbf{x}).$$

   Note that here "=" indicates set equality, i.e., for two sets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} = \mathcal{B}$ if and only if $\mathcal{A} \subseteq \mathcal{B}$ and $\mathcal{B} \subseteq \mathcal{A}$. This is the relevant interpretation used throughout the notes when appropriate.

2. (Summation) Suppose $f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$ where all of the f_i's are convex. Then

   $$\partial f(\mathbf{x}) = \sum_{i=1}^{n} \partial f_i(\mathbf{x}) = \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) + \ldots + \partial f_n(\mathbf{x}).$$

   Note that here "+" indicates set sum, i.e., for two sets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} + \mathcal{B} = \{(a+b) | a \in \mathcal{A}, b \in \mathcal{B}\}$.

3. (Composition with and Affine Function) Suppose $h : \mathbb{R}^d \to \mathbb{R}$ is defined as $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$, where $f : \mathbb{R}^m \to \mathbb{R}$ is convex, $\mathbf{A} \in \mathbb{R}^{m \times d}, \mathbf{b} \in \mathbb{R}^m$. Then

   $$\partial h(\mathbf{x}) = \mathbf{A}^\top \partial f(\mathbf{A}\mathbf{x} + \mathbf{b}).$$

4. (Pointwise Maximum) Let $f(\mathbf{x}) = \max\limits_{1 \leq i \leq n} f_i(\mathbf{x})$ where f_i is convex for each i then

   $$\partial f(\mathbf{x}) = \mathbf{conv} \left(\bigcup_{1 \leq i \leq n} \partial f_i(\mathbf{x}) \, \mathbb{1} \left\{ f_i(\mathbf{x}) = f(\mathbf{x}) \right\} \right).$$

   i.e., the convex hull of the union of the subdifferential of "active" functions.

In the following, we revisit the absolute value function, and use the pointwise maximum rule to calculate its subdifferential.

Example 5 *Consider the absolute value function*

$$f(x) = |x|.$$

Let $f_1(x) = x$ and $f_2(x) = -x$, both of which are differentiable. It is easy to show that $f(x) = \max\{f_1(x), f_2(x)\}$. At $x = 0$, we have $f(0) = f_1(0) = f_2(0)$. Thus,

$$\partial f_1(0) = \{\nabla f_1(0)\} = \{1\},$$

and

$$\partial f_2(0) = \{\nabla f_2(0)\} = \{-1\}.$$

By the pointwise maximum rule, we have

$$\partial f(0) = \mathbf{conv}\left(\partial f_1(0) \cup \partial f_2(0)\right) = \mathbf{conv}\left(\{1\} \cup \{-1\}\right) = [-1, 1].$$

Thus,

$$\partial f(x) = \begin{cases} \{1\} & x > 0 \\ \{-1\} & x < 0 \\ [-1, 1] & x = 0 \end{cases}.$$