Optimization Theory and Algorithm	**Lecture 10 - 05/28/2021**

Lecture 10

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

1 Subgradient Descent

In the last subsection, we have shown that how to use gradient descent algorithms to solve smooth and convex objective function.

Q: How about non-smooth objective function?

Example 1 *Least Absolute Deviation Regression (LAD Regression), it is similar to the Least Squares problems with the optimization formulation as:*

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_1. \tag{1}$$

We need a way to measure stationarity in the non-smooth case. For convex functions, a natural notion is that of the subgradient/subdifferential.

1.1 Subgradient and Subdifferential

Definition 1 *A subgradient of a convex possible non-smooth function $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{x} \in \mathbb{R}^n$ is a vector $\mathbf{g} \in \mathbb{R}^n$*

$$f(\mathbf{y}) \geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x}\rangle + f(\mathbf{x})$$

for all $\mathbf{y}$.

Definition 2 *The subdifferential of f at $\mathbf{x}$ is the set of all subgradients, denoted $\partial f(\mathbf{x})$. Equivalently*

$$\partial f(\mathbf{x}) := \{\mathbf{g} \in \mathbb{R}^n : f(\mathbf{y}) \geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x}\rangle + f(\mathbf{x}) \text{ for all } \mathbf{y}\}.$$

Theorem 1 *$\mathbf{x}^*$ is a global minimal point of the convex possible non-smooth function f if and only if $0 \in \partial f(\mathbf{x}^*)$.*

Remark 1 *Geometric Interpretation of Subgradient: Assume that $(\mathbf{y}, t) \in epi(f)$, then $f(\mathbf{y}) \leq t$. Thus, $t \geq f(\mathbf{y}) \geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x}\rangle + f(\mathbf{x})$. This implies*

$$\langle \begin{pmatrix} \mathbf{g} \\ -1 \end{pmatrix}, \begin{pmatrix} \mathbf{y} \\ t \end{pmatrix} - \begin{pmatrix} \mathbf{x} \\ f(\mathbf{x}) \end{pmatrix} \rangle \leq 0. \tag{2}$$

Theorem 2 *Suppose that $f(\mathbf{x})$ is convex and differentiable at point $\mathbf{x}_0$, then $\partial f(\mathbf{x}_0) = \{\nabla f(\mathbf{x}_0)\}$.*

Proof 1 *Obviously, $\nabla f(\mathbf{x}_0) \in \partial f(\mathbf{x}_0)$. Assume that $\mathbf{g} \in \partial f(\mathbf{x}_0)$ but $\mathbf{g} \neq \nabla f(\mathbf{x}_0)$. For any $\mathbf{d} \in \mathbb{R}^n, \mathbf{d} \neq 0$, and exist $t > 0$ such that $\mathbf{x}_0 + t\mathbf{d} \in (f)$. So, $f(\mathbf{x}^0) + t\mathbf{d}) \geq f(\mathbf{x}^0) + t\langle \mathbf{g}, \mathbf{d}\rangle$. Let $\mathbf{d} = g - \nabla f(\mathbf{x}^0) \neq 0$, then*

$$\frac{f(\mathbf{x}^0 + t\mathbf{d}) - f(\mathbf{x}^0) - t\langle \nabla f(\mathbf{x}^0), \mathbf{d}\rangle}{t\|\mathbf{d}\|} \geq \frac{\langle \mathbf{g} - \nabla f(\mathbf{x}^0), \mathbf{d}\rangle}{\|\mathbf{d}\|} = \|\mathbf{d}\| > 0. \tag{3}$$

However, as $t \to 0$, Eq.(3) should be goes to zero. Thus, it is controversial.

Theorem 3 *Suppose that f is a convex function, if $\mathbf{x} \in int(f)$ then $\partial f(\mathbf{x}) \neq \emptyset$.*

Proof 2 *For any $\mathbf{x} \in dom(f)$ and $(\mathbf{x}, f(\mathbf{x})) \in epi(f)$, it has $epi(f)$ is convex due to the convexity of f. Based on Supporting Hyperplan Theorem, there exists $\mathbf{a}, \mathbf{b}$ such that*

$$\langle \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix}, \begin{pmatrix} \mathbf{y} \\ t \end{pmatrix} - \begin{pmatrix} \mathbf{x} \\ f(\mathbf{x}) \end{pmatrix} \rangle \leq 0, \forall (\mathbf{y}, t) \in epi(f). \tag{4}$$

So, $\langle \mathbf{a}, \mathbf{y} - \mathbf{x} \rangle \leq b(f(\mathbf{x}) - t), \forall (\mathbf{y}, t) \in epi(f)$. Consider $t \to \infty$, then b should be $b \leq 0$. In addition, b is not zero, because $\langle \mathbf{a}, \mathbf{y} - \mathbf{x} \rangle \leq 0$ is not corrected for all $\mathbf{y}$. Then $b < 0$. Let $\mathbf{g} = -\frac{\mathbf{a}}{b}$, then

$$\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle = \langle -\frac{\mathbf{a}}{b}, \mathbf{y} - \mathbf{x} \rangle \leq t - f(\mathbf{x}). \tag{5}$$

Take $t = f(\mathbf{y})$, then $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle$. So, $\mathbf{g} \in \partial f(\mathbf{x}) \neq \emptyset$.

Theorem 4 *(Monotonicity) Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex, and $\mathbf{x}, \mathbf{y} \in dom(f)$, then*

$$\langle \mathbf{a} - \mathbf{b}, f(\mathbf{x}) - f(\mathbf{y}) \rangle \geq 0 \tag{6}$$

where $\mathbf{a} \in \partial f(\mathbf{x})$ and $\mathbf{b} \in \partial f(\mathbf{y})$.

Example 2 *Let us show some examples of deriving subgradient and subdifferential.*

- *$f(x) = |x|$, Then*

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

- *$f(x) = \max(x, 0)$ is called ReLU which is widely used in Deep Learning models. You can compute the subdifferential of it by yourself.*

- *$f(\mathbf{x}) = \|\mathbf{x}\|_2, \mathbf{x} \in \mathbb{R}^n$.*

$$\partial f(x) = \begin{cases} \left\{ \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \right\}, & \mathbf{x} \neq 0 \\ \{\mathbf{g} : \ \|\mathbf{g}\|_2 \leq 1\}, & \mathbf{x} = 0. \end{cases} \tag{7}$$

Computational Rules of Subgradients: See Page 68-75.

(1) f_1 and f_2 are convex, and $int(f_1) \cap int(f_2) \neq \emptyset$, then for any $\mathbf{x} \in int(f_1) \cap int(f_2)$ and $f(\mathbf{x}) = \alpha_1 f_1 + \alpha_2 f_2, \alpha_1 > 0, \alpha_2 > 0$, we have

$$\partial f(\mathbf{x}) = \alpha_1 \partial f_1 + \alpha_2 \partial f_2. \tag{8}$$

(2) Assume that h is convex, and $f(\mathbf{x}) = h(A\mathbf{x} + \mathbf{b})$, then

$$\partial f(\mathbf{x}) = A^\top \partial h(A\mathbf{x} + \mathbf{b}). \tag{9}$$

(3) Suppose that $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ are convex, let $f = \max\{f_1, \ldots, f_m\}$, then for any $\mathbf{x}^0 \in \cap_{i=1}^m intdom(f_i)$, denote $I(\mathbf{x}^0) = \{i : f_i(\mathbf{x}^0) = f(\mathbf{x}^0)\}$ then

$$\partial f(\mathbf{x}^0) = conv(\cup_{i \in I(\mathbf{x}^0)} \partial f_i(\mathbf{x}^0)) \tag{10}$$

The usefulness of the rules can be found in Example 2.16, 2.17, and 2.18 at Page 71 and 72.

1.1.1 Subgradient Descent

Subgradient descent algorithm should be

$$\mathbf{x}^{t+1} = \mathbf{x}^t - s_t \mathbf{g}^t \tag{11}$$

where $\mathbf{g}^t \in \partial f(\mathbf{x}^t)$.

Compared with the standard gradient descent algorithm, we need to consider the following problems:

- How to select $\mathbf{g}^t \in \partial f(\mathbf{x}^t)$?
- How to choice the step size s_t?
- How to stop the algorithm?

We will answer these questions for the specific non-smooth objective function which is a Lipschitz continuous function.

Definition 3 *Function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz function with respect to a constant $G > 0$ if for any $\mathbf{x}, \mathbf{y} \in dom(f)$*

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq G\|\mathbf{x} - \mathbf{y}\|_2, \tag{12}$$

where G is referred as to Lipschitz constant of f.

Example 3
- *$f(\mathbf{x}) = \|\mathbf{x}\|$ is 1-Lip.*
- *$f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$ is $\|a\|$-Lip.*

Theorem 5 *f is convex, then f is a G-Lip function if and only if $\|\mathbf{g}\| \leq G$, for any $\mathbf{g} \in \partial f(\mathbf{x}), \mathbf{x} \in dom(f)$.*

Proof 3 Part 1: *If f is a convex, G-Lip function, and there exists $\mathbf{g} \in \partial f(\mathbf{x})$ such that $\|\mathbf{g}\| > G$. Let $\mathbf{y} = \mathbf{x} + \frac{\mathbf{g}}{\|\mathbf{g}\|}$. Then by the definition of G-Lip, we have*

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq G\|\mathbf{y} - \mathbf{x}\| < \|\mathbf{g}\|. \tag{13}$$

However, according to the definition of subgradient, we have

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle = \|\mathbf{g}\|. \tag{14}$$

These two inequalities are controversial.

Part 2: *Assume that f is convex and for any $\mathbf{g} \in \partial f(\mathbf{x}), \|\mathbf{g}\| \leq G$. Then for any $\mathbf{x}, \mathbf{y} \in (f)$, we have*

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \mathbf{g}_\mathbf{x}, \mathbf{y} - \mathbf{x} \rangle \geq -\|\mathbf{g}_\mathbf{x}\|\|\mathbf{x} - \mathbf{y}\| \geq -G\|\mathbf{x} - \mathbf{y}\|, \tag{15}$$

$$f(\mathbf{y}) - f(\mathbf{x}) \leq \langle \mathbf{g}_\mathbf{y}, \mathbf{y} - \mathbf{x} \rangle \leq \|\mathbf{g}_\mathbf{y}\|\|\mathbf{x} - \mathbf{y}\| \leq G\|\mathbf{x} - \mathbf{y}\|. \tag{16}$$

These indicate the results.

Theorem 6 *Assume that f is a convex and G-Lip function, $\mathbf{x}^* = \arg\min f(\mathbf{x}), f^* = f(\mathbf{x}^*) > -\infty$, then $\{\mathbf{x}^t\}_{t=0}^{\infty}$ is generated form the subgradient descent algorithm, then for any $T > 0$, it has*

$$f(\mathbf{x}^{t^*}) - f^* \leq \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + G^2 \sum_{t=0}^T s_t^2}{2\sum_{t=0}^T s_t}, \tag{17}$$

where $t^* = \arg\min_{0 \leq t \leq T} f(\mathbf{x}^t)$.

Proof 4

$$\begin{aligned}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}^t - s_t\mathbf{g}_t - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}^t - \mathbf{x}^*\|^2 - 2s_t\langle \mathbf{g}_t, \mathbf{x}^t - \mathbf{x}^*\rangle + s_t^2\|\mathbf{g}_t\|^2 \\ &\le \|\mathbf{x}^t - \mathbf{x}^*\|^2 - 2s_t(f(\mathbf{x}^t) - f^*) + s_t^2G^2,\end{aligned}$$

where the last inequality by the convexity of f. So, it can be derived as

$$2s_t(f(\mathbf{x}^t) - f^*) \le \|\mathbf{x}^t - \mathbf{x}^*\|^2 - \|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 + s_t^2G^2.$$

Thus,

$$\begin{aligned}2(f(\mathbf{x}^{t^*}) - f^*)\sum_{t=0}^{T} s_t \le 2\sum_{t=0}^{T} s_t(f(\mathbf{x}^t) - f^*) \\ &\le \|\mathbf{x}^0 - \mathbf{x}^*\|^2 - \|\mathbf{x}^T - \mathbf{x}^*\|^2 + G^2\sum_{t=0}^{T} s_t^2 \\ &\le \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + G^2\sum_{t=0}^{T} s_t^2.\end{aligned}$$

Finally,

$$f(\mathbf{x}^{t^*}) - f^* \le \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + G^2\sum_{t=0}^{T} s_t^2}{2\sum_{t=0}^{T} s_t}.$$

Let us discuss the above theorem.

(1) $f(\mathbf{x}^t) - f(\mathbf{x}^*)$ may be not decreasing!

(2) Let $\|\mathbf{x}^0 - \mathbf{x}^*\|^2 = R^2, s_t = s$, then

$$f(\mathbf{x}^{t^*}) - f^* \le \frac{R^2}{2Ts} + \frac{sTG^2}{2} := \Phi(s). \tag{18}$$

Obvisouly, if $s = \frac{R}{G\sqrt{T}}$, then $\min \Phi(s) = \frac{GR}{\sqrt{T}}$. Thus,

$$f(\mathbf{x}^{t^*}) - f^* \le \inf_s \Phi(s) = \frac{GR}{\sqrt{T}}.$$

This indicates that the convergence speed is the same with the only β-smooth objective function.

(3) To $f(\mathbf{x}^{t^*}) - f^* \to 0$, it should be $\sum_{t=1}^{\infty} s_t = +\infty$ and $\sum_{t=1}^{\infty} s_t^2 \le M$, where M is a constant.
Q: Could you please give us an example of $\{s_t\}_{t=0}^{\infty}$.

References