6FMAI19 Nonlinear Optimization Spring, 2022

Lecture 8 — 09 March, 2022

Lecturer: Yura Malitsky *Scribe: Ahmet Kaplan*

1 Last Lecture: Stochastic Subgradient Method

During the last lecture, we introduced the stochastic subgradient method for the following problem

$$\underset{\mathbf{x}}{\text{minimize}} \quad f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}). \tag{1}$$

We analyzed the convergence by assuming the subgradient of f_{i_k} at x_k, $\mathbf{g}_k \in \partial f_{i_k}(\mathbf{x}_k)$, satisfies $E[||\mathbf{g}_k||] \leq G$. Now, we provide the convergence analysis of the stochastic gradient method when the $f(\cdot)$ is L-smooth.

2 Stochastic Gradient with Lipschitz Smoothness

Assumption 1. *We assume that $f(\cdot)$ is L-smooth.*

We cannot guarantee our previous assumption: $E[||\mathbf{g}_k||] \leq G$ due to the Assumption 1. Our new assumption is given as

Assumption 2. *We assume that $E[||\Delta f_\xi(\mathbf{x})||^2] \leq A + B||\Delta f(\mathbf{x})||^2$. We can rewrite our assumption as follows*

$$\frac{1}{n}||\nabla f_1(\mathbf{x})||^2 + \ldots + \frac{1}{n}||\nabla f_n(\mathbf{x})||^2 \leq A + \frac{B}{n^2}||\nabla f_1(\mathbf{x}) + \ldots + \nabla f_n(\mathbf{x})||^2. \tag{2}$$

Here, when $B = 0$, we return to our previous assumption.

Assumption 3. *We assume that $1 - \frac{\alpha_k LB}{2} \geq \frac{1}{2} \iff \alpha_k \leq \frac{1}{LB}$.*

Assumption 4. *We assume that $f(\mathbf{x})$ is lower bounded as $f(\mathbf{x}) \geq f_{low}$.*

In addition, there is no information about the convexity of $f(\cdot)$. For the stochastic gradient method, we apply Algorithm 1.

Algorithm 1 Stochastic Gradient Method

Require: $\mathbf{x}_0$
Ensure: $\mathbf{x}_\tau$
  for $k = 0, \cdots, K-1$ **do**
    Sample $\xi_k \in \{1, 2, \ldots, n\}$ uniformly
    Calculate $\mathbf{x}_{k+1} := \mathbf{x}_k - \alpha_k \nabla f_{\xi_k}(\mathbf{x}_k)$
  end for
  Sample τ from p.m.f. $P\{\tau = t\} = \frac{\alpha_t}{\sum_{i=0}^{K}\alpha_i}$
  Return $\mathbf{x}_\tau$ ▷ We have a concrete point.

2.1 Convergence Analysis

Definition 5. *(Descent Lemma)*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2 \tag{3}$$

We use descent lemma in our calculations. Let $\mathbf{y}, \mathbf{x}$ replaced by $\mathbf{x}_{k+1}, \mathbf{x}_k$. Then, we can write the following inequality

$$\begin{aligned} f\left(\mathbf{x}_{k+1}\right) &\leq f\left(\mathbf{x}_k\right) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} ||\mathbf{x}_{k+1} - \mathbf{x}_k||^2 \\ &= f\left(\mathbf{x}_k\right) - \alpha_k \langle \nabla f(\mathbf{x}_k), \nabla f_{\xi_k}(\mathbf{x}_k) \rangle + \frac{\alpha_k^2 L}{2} ||\nabla f_{\xi_k}(\mathbf{x}_k)||^2, \end{aligned} \tag{4}$$

where we use the equation, $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f_{\xi_k}(\mathbf{x}_k)$, to transform RHS as in the second line. We calculate the conditional expectation of $\nabla f_{\xi_k}(\mathbf{x}_k)$ as follows

$$E_k[\nabla f_{\xi_k}(\mathbf{x}_k)] = \frac{1}{n} \nabla f_1(\mathbf{x}_k) + \ldots + \frac{1}{n} \nabla f_n(\mathbf{x}_k) = \nabla f(\mathbf{x}_k). \tag{5}$$

Now, we take the conditional expectation of Eq. 4. w.r.t. k as follows

$$E_k[f\left(\mathbf{x}_{k+1}\right) - f\left(\mathbf{x}_k\right)] \leq -\alpha_k ||\nabla f(\mathbf{x}_k)||^2 + \frac{\alpha_k^2 L}{2} E_k[||\nabla f_{\xi_k}(\mathbf{x}_k)||^2]. \tag{6}$$

By using Assumption 2, Eq. 6 can be written as

$$\begin{aligned} E_k[f\left(\mathbf{x}_{k+1}\right) - f\left(\mathbf{x}_k\right)] &\leq -\alpha_k ||\nabla f(\mathbf{x}_k)||^2 + \frac{\alpha_k^2 L}{2} (A + B||\nabla f(\mathbf{x}_k)||^2) \\ \alpha_k \left(1 - \frac{\alpha_k L B}{2}\right) ||\nabla f(\mathbf{x}_k)||^2 &\leq E_k[f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})] + \frac{\alpha_k^2 L A}{2} \end{aligned} \tag{7}$$

By using Assumption 3, Eq. 7 can be written as

$$\frac{\alpha_k}{2} ||\nabla f(\mathbf{x}_k)||^2 \leq E_k[f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})] + \frac{\alpha_k^2 A L}{2} \tag{8}$$

In Eq. 8, when we calculate the following expectation $E_k[f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})]$ w.r.t. k, the term $f(\mathbf{x}_{k+1})$ does not disappear due to $k+1$. That is why we take expectation of Eq. 8 by assuming x_k is also random as follows

$$\frac{\alpha_k}{2} E_{\xi_1, \ldots, \xi_{K-1}} \left[||\nabla f(\mathbf{x}_k)||^2 \right] \leq E_{\xi_1, \ldots, \xi_{K-1}} \left[f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \right] + \frac{\alpha_k^2 A L}{2}. \tag{9}$$

Now, similar to the previous lectures it is possible to sum up as follows

$$\sum_{i=0}^{K} \alpha_i E_{\xi_1, \ldots, \xi_{K-1}} \left[||\nabla f(\mathbf{x}_i)||^2 \right] \leq 2 E_{\xi_1, \ldots, \xi_{K-1}} [f(\mathbf{x}_0) - f(\mathbf{x}_{K+1})] + A L \sum_{i=0}^{K} \alpha_i^2. \tag{10}$$

By using Assumption 4, Eq. 10 can be written as

$$\sum_{i=0}^{K} \alpha_i E_{\xi_1, \ldots, \xi_{K-1}} \left[||\nabla f(\mathbf{x}_i)||^2 \right] \leq 2 \left(f(\mathbf{x}_0) - f_{low} \right) + A L \sum_{i=0}^{K} \alpha_i^2. \tag{11}$$

By using Eq. 11, we can show that

$$\begin{aligned} E_{\tau, \xi_1, \ldots, \xi_{K-1}} \left[||\nabla f(\mathbf{x}_\tau)||^2 \right] &\leq \frac{2 \left(f(\mathbf{x}_0) - f_{low} \right) + A L \sum_{i=0}^{K} \alpha_i^2}{\sum_{i=0}^{K} \alpha_i} \\ \sum_{t=0}^{K} E_{\xi_1, \ldots, \xi_{K-1}} \left[||\nabla f(\mathbf{x}_t)||^2 \right] \frac{\alpha_t}{\sum_{i=0}^{K} \alpha_i} &\leq \frac{2 \left(f(\mathbf{x}_0) - f_{low} \right) + A L \sum_{i=0}^{K} \alpha_i^2}{\sum_{i=0}^{K} \alpha_i}, \end{aligned} \tag{12}$$

where $\frac{\alpha_t}{\sum_{i=0}^{K} \alpha_i}$ stands for the probability mass function (pmf). Here, for $\alpha \sim \frac{1}{\sqrt{K}}$, $\sum_{i=0}^{K} \alpha_i^2$ goes to infinity slower than $\sum_{i=0}^{K} \alpha_i$.

3 μ-Strongly Convex Functions

Definition 6. *(Convex function). Let $C \subset \mathbb{R}^n$ be a convex set. The function $f : C \to \mathbb{R}$ is convex if for all $x, y \in C, \alpha \in [0,1]$, it follows that*

$$\alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) \geq f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}). \tag{13}$$

One of the simplest convex functions is $f : \mathbb{R}^n \to \mathbb{R}, f(\mathbf{x}) = \frac{1}{2}||\mathbf{x}||^2$.

Definition 7. *(μ-strongly convex function). Let $C \subset \mathbb{R}^n$ be a convex set. The function $f : C \to \mathbb{R}$ is μ-strongly convex if for all $x, y \in C, \alpha \in [0,1]$, it follows that*

$$\alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) \geq f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) + \frac{\mu}{2}\alpha(1-\alpha)||\mathbf{y} - \mathbf{x}||^2. \tag{14}$$

Algorithms usually converge faster with these functions. The equivalent form of Definition 7 for μ-strongly convex function is given as

$$\alpha \left(f(\mathbf{x}) - \frac{\mu}{2}||x||^2 \right) + (1-\alpha) \left(f(\mathbf{y}) - \frac{\mu}{2}||y||^2 \right) \geq f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) - \frac{\mu}{2}||\alpha \mathbf{x} + (1-\alpha)\mathbf{y}||^2. \tag{15}$$

The other definition is that f is μ-strongly convex function if and only if $F(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}||\mathbf{x}||^2$ is convex.

If f is differentiable and μ-strongly convex, then we can apply the inequality for convexity as follows

$$\begin{aligned} F(\mathbf{y}) &\geq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\ f(\mathbf{y}) - \frac{\mu}{2}||\mathbf{y}||^2 &\geq f(\mathbf{x}) - \frac{\mu}{2}||x||^2 + \langle \nabla f(\mathbf{x}) - \mu \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle. \end{aligned} \tag{16}$$

We can write Eq. 16 in more compact form as follows

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}||\mathbf{y} - \mathbf{x}||^2. \tag{17}$$

If f is two times differentiable, then the following holds

$$f \text{ is two times differentiable, } \mu\text{-strongly convex} \Longleftrightarrow \nabla^2 f(\mathbf{x}) \geq \mu \mathbf{I}, \tag{18}$$

where $\mathbf{I}$ is an identity matrix.

4 Gradient Descent for μ-Strongly Convex Functions

We assume that f is L-smooth and μ-strongly convex function. For our analyses, we will apply gradient descent method as follows

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k). \tag{19}$$

Bu using Definition 5 and replacing $\mathbf{y}, \mathbf{x}$ by $\mathbf{x}_{k+1}, \mathbf{x}_k$, we can write

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}||\mathbf{x}_{k+1} - \mathbf{x}_k||^2 \\ &= f(\mathbf{x}_k) - \alpha \langle \nabla f(\mathbf{x}_k), \nabla f(\mathbf{x}_k) \rangle + \frac{L}{2}|| - \alpha \nabla f(\mathbf{x}_k)||^2 \\ &= f(\mathbf{x}_k) - \alpha ||\nabla f(\mathbf{x}_k)||^2 + \frac{\alpha^2 L}{2}||f(\mathbf{x}_k)||^2 \\ &= f(\mathbf{x}_k) - \alpha \left(1 - \frac{\alpha L}{2}\right) ||f(\mathbf{x}_k)||^2. \end{aligned} \tag{20}$$

Let $\alpha = 1/L$, then we can write Eq. 20 as follows

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2L}||f(\mathbf{x}_k)||^2. \tag{21}$$

By using Eq. 17, one can prove that

$$\frac{1}{2\mu}||\nabla f(\mathbf{x})||^2 \geq f(\mathbf{x}) - f_* \quad \forall \mathbf{x}. \tag{22}$$

Here, f is μ-strongly convex, so it has a unique minimum f_*. By substituting Eq. 22 into Eq. 21, we can write the following inequality

$$\begin{aligned}
f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) - \frac{\mu}{L}(f(\mathbf{x}_k) - f_*) \\
f(\mathbf{x}_{k+1}) - f_* &\leq \left(1 - \frac{\mu}{L}\right)(f(\mathbf{x}_k) - f_*) \\
f(\mathbf{x}_{k+1}) - f_* &\leq \left(1 - \frac{\mu}{L}\right)^2 (f(\mathbf{x}_{k-1}) - f_*) \\
&\vdots \\
f(\mathbf{x}_{k+1}) - f_* &\leq \left(1 - \frac{\mu}{L}\right)^{k+1} (f(\mathbf{x}_0) - f_*),
\end{aligned} \tag{23}$$

where $\frac{\mu}{L}$ is called the condition number. If it has a small value, it is difficult to optimize.