Columbia University in the City of New York
Optimization Methods for Machine Learning
Instructors: Satyen Kale
Authors: Gurpreet Singh
Email: gurpreet.s@columbia.edu

SCRIBE

12

Convergence of SGD - Continued

1. Recap

In this section, we do a quick recap of what Stochastic Gradient Descent (SGD) looks like and some preliminary results we derived in the previous lecture. The optimization objective is given as

$$\min_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x}) \;=\; \min_{\mathbf{x}\in\mathcal{K}} \mathop{\mathbb{E}}_{\xi\sim\mathcal{D}} \left[\, g(\mathbf{x},\xi) \,\right]$$

The assumption is that g is a convex function and, as asual, $\mathcal{K}$ is a convex set. The SGD algorithm for the above problem is given in algorithm 1.

Algorithm 1: Stochastic Gradient Descent

1. Start with an arbitrary initial point $\mathbf{x}_0 \in \mathcal{K}$
2. For $t = 1, 2 \ldots T$
   - (a) Draw $\xi_t \sim \mathcal{D}$
   - (b) Set $\mathbf{y}_t = \mathbf{x}_{t-1} - \eta \nabla g(\mathbf{x}_{t-1}, \xi_t)$
   - (c) Update $\mathbf{x}_t = \Pi_{\mathcal{K}}(\mathbf{y})$
3. Output the final estimate as some combination of $\{\mathbf{x}_0, \mathbf{x}_1 \ldots \mathbf{x}_T\}$

We will only look at the unconstrained case, i.e. $K = \mathbb{R}^d$, so no projections are necessary. In this setting, for the case when g is smooth with the smoothness coefficient β, we have the following result which was shown in last lecture in the convergence analysis of SGD with smooth functions.

Result 12.0.1 For a convex and smooth function $f : \mathbb{R}^d \to \mathbb{R}$ with the smoothness coefficient β, we have

$$\mathbb{E}\left[\, f(\mathbf{x}_{t+1} \,|\, \mathbf{x}_t) \,\right] \;\leq\; f(\mathbf{x}_t) - \frac{\eta}{2} \left\| \nabla f(\mathbf{x}_t) \right\|^2 + \frac{\eta}{2}\sigma^2$$

where $\eta \leq \frac{1}{\beta}$ and σ^2 is the bound for the variance of $\nabla g(\mathbf{x},\xi) - \nabla f(\mathbf{x})$ over $\xi \sim \mathcal{D}$ for all $\mathbf{x} \in \mathcal{K}$

where $\underset{\xi\sim\mathcal{D}}{\mathbb{E}}\left[\, g(\mathbf{x},\xi)\,\right] = f(\mathbf{x})$, *i.e.* $\underset{\xi\sim\mathcal{D}}{\mathbb{E}}\left[\, \|\, \nabla g(\mathbf{x},\xi) - \nabla f(\mathbf{x})\, \|^2\,\right] \leq \sigma^2 \;\forall\; \mathbf{x}\in\mathcal{K}$.

We will assume, for the following sections, that f is convex and the stochastic function g satisfies for all $\mathbf{x}\in\mathcal{K}$

$$\underset{\xi\sim\mathcal{D}}{\mathbb{E}}\left[\, \|\, \nabla g(\mathbf{x},\xi) - \nabla f(\mathbf{x})\, \|^2\,\right] \;\leq\; \sigma^2$$

2. SGD for β-Smooth and α-Strongly Convex Functions

Suppose we have that the function f is both smooth (with coefficient β) as well as strongly convex (with coefficient α)

Repeating from the previous lectures, we have

$$\begin{aligned}
\mathbb{E}\left[\, \|\, \mathbf{x}_{t+1} - \mathbf{x}^*\, \|^2 \,\middle|\, \mathbf{x}_t\right] \;&=\; \mathbb{E}\left[\, \|\, \mathbf{x}_{t+1} - \mathbf{x}_t\, \|^2 \,\middle|\, \mathbf{x}_t\right] + \|\, \mathbf{x}_t - \mathbf{x}^*\, \|^2 + 2\mathbb{E}\left[\, (\mathbf{x}_{t+1} - \mathbf{x}_t)^{\mathrm{T}}(\mathbf{x}_t - \mathbf{x}^*) \,\middle|\, \mathbf{x}_t\right] \\
&=\; \underset{\xi}{\mathbb{E}}\left[\, \eta^2\, \|\, \nabla g(\mathbf{x}_t,\xi)\, \|^2 \,\middle|\, \mathbf{x}_t\right] + \|\, \mathbf{x}_t - \mathbf{x}^*\, \|^2 - 2\eta\; \mathbb{E}\left[\, (\mathbf{x}_t - \mathbf{x}^*)^{\mathrm{T}}\nabla g(\mathbf{x}_t,\xi) \,\middle|\, \mathbf{x}_t\right] \\
&=\; \eta^2\; \underset{\xi}{\mathbb{E}}\left[\, \|\, \nabla g(\mathbf{x}_t,\xi) - \nabla f(\mathbf{x}_t)\, \|^2 + \|\, \nabla f(\mathbf{x}_t)\, \|^2 \,\middle|\, \mathbf{x}_t\right] + \|\, \mathbf{x}_t - \mathbf{x}^*\, \|^2 \\
&\qquad\qquad - 2\eta\; \mathbb{E}\left[\, (\mathbf{x}_t - \mathbf{x}^*)^{\mathrm{T}}\nabla g(\mathbf{x}_t,\xi) \,\middle|\, \mathbf{x}_t\right] \\
&\leq\; \eta^2\sigma^2 + \eta^2\, \|\, \nabla f(\mathbf{x}_t)\, \|^2 + \|\, \mathbf{x}_t - \mathbf{x}^*\, \|^2 - 2\eta(\mathbf{x}_t - \mathbf{x}^*)^{\mathrm{T}}\mathbb{E}\left[\, \nabla g(\mathbf{x}_t,\xi) \,\middle|\, \mathbf{x}_t\right] \\
&=\; \eta^2\sigma^2 + \eta^2\, \|\, \nabla f(\mathbf{x}_t)\, \|^2 + \|\, \mathbf{x}_t - \mathbf{x}^*\, \|^2 - 2\eta\nabla f(\mathbf{x}_t)(\mathbf{x}_t - \mathbf{x}^*) \qquad (1)
\end{aligned}$$

From strong convexity of f, we have

$$f(\mathbf{x}^*) \;\geq\; f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^{\mathrm{T}}(\mathbf{x}^* - \mathbf{x}_t) + \frac{\alpha}{2}\, \|\, \mathbf{x}^* - \mathbf{x}_t\, \|^2 \qquad (2)$$

Using the above two results (1 and 2), we can write

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \;\leq\; \frac{1}{2\eta}(1-\eta\alpha)\, \|\, \mathbf{x}_t - \mathbf{x}^*\, \|^2 + \frac{\eta}{2}\, \|\, \nabla f(\mathbf{x}_t)\, \|^2 + \frac{\eta}{2}\sigma^2 - \frac{1}{2\eta}\mathbb{E}\left[\, \|\, \mathbf{x}_{t+1} - \mathbf{x}^*\, \|^2 \,\middle|\, \mathbf{x}_t\right] \qquad (3)$$

Looping in result 12.0.1, we can write

$$\frac{\eta}{2}\, \|\, \nabla f(\mathbf{x}_t)\, \|^2 \;\leq\; f(\mathbf{x}_t) - \mathbb{E}\left[\, f(\mathbf{x}_{t+1}) \,\middle|\, \mathbf{x}_t\right] + \frac{\eta}{2}\sigma^2$$

Using the above equation and equation 3, we have

$$\mathbb{E}\left[\, f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \,\middle|\, \mathbf{x}_t\right] \;\leq\; \frac{1}{2\eta}(1-\eta\alpha)\, \|\, \mathbf{x}_t - \mathbf{x}^*\, \|^2 + \eta\sigma^2 - \frac{1}{2\eta}\mathbb{E}\left[\, \|\, \mathbf{x}_{t+1} - \mathbf{x}^*\, \|^2 \,\middle|\, \mathbf{x}_t\right]$$

Since the LHS is always non-negative, we have

$$\mathbb{E}\left[\, \|\, \mathbf{x}_{t+1} - \mathbf{x}^*\, \|^2 \,\middle|\, \mathbf{x}_t\right] \;\leq\; (1-\eta\alpha)\, \|\, \mathbf{x}_t - \mathbf{x}^*\, \|^2 + 2\eta^2\sigma^2$$

Taking expectation on both sides, we have

$$\mathbb{E}\left[\, \|\, \mathbf{x}_{t+1} - \mathbf{x}^*\, \|^2\,\right] \;\leq\; (1-\eta\alpha)\mathbb{E}\left[\, \|\, \mathbf{x}_t - \mathbf{x}^*\, \|^2\,\right] + 2\eta^2\sigma^2$$

Applying this inequality recursively, we get

$$\begin{aligned}
\mathbb{E}\left[\,\|\, \mathbf{x}_{T+1} - \mathbf{x}^* \,\|^2\right] &\leq (1-\eta\alpha)\mathbb{E}\left[\,\|\, \mathbf{x}_{T} - \mathbf{x}^* \,\|^2\right] + 2\eta^2\sigma^2 \\
&\leq (1-\eta\alpha)^2\mathbb{E}\left[\,\|\, \mathbf{x}_{T-1} - \mathbf{x}^* \,\|^2\right] + 2\eta^2\sigma^2 + 2\eta^2(1-\eta\alpha)\sigma^2 \\
&\leq \ldots \\
&\leq (1-\eta\alpha)^T\mathbb{E}\left[\,\|\, \mathbf{x}_{0} - \mathbf{x}^* \,\|^2\right] + 2\eta^2\sigma^2\sum_{i=0}^{T-1}(1-\eta\alpha)^i \\
&= (1-\eta\alpha)^T \,\|\, \mathbf{x}_{0} - \mathbf{x}^* \,\|^2 + 2\eta^2\sigma^2\frac{1-(1-\eta\alpha)^T}{1-(1-\eta\alpha)} \\
&\leq (1-\eta\alpha)^T \,\|\, \mathbf{x}_{0} - \mathbf{x}^* \,\|^2 + \frac{2\eta}{\alpha}\sigma^2
\end{aligned}$$

Suppose if the point $\mathbf{x}^*$ is a local (or global) minima, then we have $\nabla f(\mathbf{x}^*) = \mathbf{0}$ since we are in the unconstrained setting. Therefore, using smoothness, we can write $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2}\,\|\, \mathbf{x}_T - \mathbf{x}^* \,\|^2$. Therefore, we have

$$\mathbb{E}\left[\, f(\mathbf{x}_T) - f(\mathbf{x}^*) \,\right] \leq \frac{\beta}{2}(1-\eta\alpha)^T \,\|\, \mathbf{x}_0 - \mathbf{x}^* \,\|^2 + \frac{\beta\eta}{\alpha}\sigma^2$$

Suppose we set $\eta = \frac{\log(()T)}{\alpha T}$. Since we require η to be less than $\frac{1}{\beta}$ for the result 12.0.1 to hold, we will assume that T is large enough so that $\eta \leq \frac{1}{\beta}$. With this setting of ηs, we get the following bound

$$\mathbb{E}\left[\, f(\mathbf{x}_T) \,\right] - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\,\|\, \mathbf{x}_0 - \mathbf{x}^* \,\|^2 + \frac{\beta\sigma^2 \log T}{\alpha^2 T}$$

3. SGD for Strongly Convex and Lipschitz Functions

We can write the same result as equation 1 since no assumptions (sucha as smoothness, strong convexity, etc.) are required on the function f for that inequality to be true. Moreover, since f is strongly convex, we can directly use the result in equation 3. Therefore, we have

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{1}{2\eta_t}(1-\eta_t\alpha)\,\|\, \mathbf{x}_t - \mathbf{x}^* \,\|^2 + \frac{\eta_t}{2}\,\|\, \nabla f(\mathbf{x}_t) \,\|^2 + \frac{\eta_t}{2}\sigma^2 - \frac{1}{2\eta_t}\mathbb{E}\left[\,\|\, \mathbf{x}_{t+1} - \mathbf{x}^* \,\|^2 \,\middle|\, \mathbf{x}_t\right]$$

Since the function f is assumed to be L-Lipschitz, we know $\|\, \nabla f(\mathbf{x}) \,\| \leq L \ \forall\ \mathbf{x} \in \mathcal{K}$. Therefore,

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{1}{2\eta_t}(1-\eta_t\alpha)\,\|\, \mathbf{x}_t - \mathbf{x}^* \,\|^2 + \frac{\eta_t}{2}\left(\sigma^2 + L^2\right) - \frac{1}{2\eta_t}\mathbb{E}\left[\,\|\, \mathbf{x}_{t+1} - \mathbf{x}^* \,\|^2 \,\middle|\, \mathbf{x}_t\right]$$

Summing over $t = 0$ to $T-1$ and taking expectation on both sides, we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] &\leq \frac{1}{2}\left(\sigma^2 + L^2\right)\sum_{t=0}^{T-1}\eta_t + \frac{1}{2}\left(\frac{1}{\eta_0} - \alpha\right)\|\, \mathbf{x}_0 - \mathbf{x}^* \,\|^2 \\
&\quad - \frac{1}{2\eta_{T-1}}\mathbb{E}\left[\,\|\, \mathbf{x}_T - \mathbf{x}^* \,\|^2\right] + \sum_{t=1}^{T-1}\left(\frac{1}{2}\left(\frac{1}{\eta_t} - \alpha\right) - \frac{1}{2\eta_{t-1}}\right)\|x_t - x^*\|^2
\end{aligned}$$

Setting $\eta_t = \frac{1}{\alpha(t+1)}$, then we have

$$\mathbb{E}\left[\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \;\leq\; \frac{1}{2}\left(\sigma^2 + L^2\right)\sum_{t=1}^{T}\frac{1}{\alpha t}$$

Using the fact that $\sum_{i=1}^{n}\frac{1}{i} \leq \ln(n) + 1$, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \;\leq\; \frac{\left(\sigma^2 + L^2\right)}{2\alpha}\frac{\ln(T) + 1}{T}$$

Using the convexity of f, we can say

$$\mathbb{E}\left[\, f(\bar{\mathbf{x}}) - f(\mathbf{x}^*)\,\right] \;\leq\; \frac{\left(\sigma^2 + L^2\right)}{2\alpha}\frac{\ln(T) + 1}{T}$$

Therefore, we ouput $\bar{\mathbf{x}}$ which crudely observes a $\mathcal{O}\left(\frac{1}{T}\right)$ bound.