IE 539: Convex Optimization **KAIST, Fall 2022**
Lecture #15: Analysis of SGD and variance reduction October 25, 2022
Lecturer: Dabeen Lee

1 Outline

In this lecture, we study

- Online binary classification,
- Stochastic optimization through the lens of OCO,
- Stochastic gradient descent.

2 Convergence of stochastic gradient descent

Recall that stochastic gradient descent (SGD) proceeds as the following.

Algorithm 1 Stochastic gradient descent (SGD)

Initialize $x_1 \in C$.
for $t = 1, \ldots, T$ **do**
  Obtain an estimator $\hat{g}_{x_t}$ of some $g_t \in \partial f(x_t)$.
  Update $x_{t+1} = \mathrm{Proj}_C \{x_t - \eta_t \hat{g}_{x_t}\}$ for a step size $\eta_t > 0$.
end for
Return $(1/T)\sum_{t=1}^T x_t$.

In this section, we analyze the convergence of SGD under the following assumption.

Assumption 1. Assume that $\hat{g}_x$ satisfies

$$\mathbb{E}[\hat{g}_x] = g_x \text{ for some } g_x \in \partial f(x), \quad \mathbb{E}\left[\|\hat{g}_x\|^2\right] \leq L^2.$$

This assumption is analogous to Lipschitz continuity. Under the assumption, let us analyze the performance of stochastic gradient descent given by Algorithm 1.

Theorem 15.1. *Algorithm 1 with step sizes $\eta_t = R/(L\sqrt{t})$ satisfies*

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^T x_t\right)\right] - f(x^*) \leq \frac{3LR}{2\sqrt{T}}$$

where the expectation is taken over the randomness in gradient estimation and $x^* \in \mathrm{argmin}_{x\in C} f(x)$.

2.1 Proof via online regret minimization

Suppose that $\mathbb{E}[\hat{g}_{x_t}] = g_t \in \partial f(x_t)$ for $t \geq 1$. First, let us observe the following.

$$\begin{aligned}
\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - f(x^*) &\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f(x_t)\right] - f(x^*) \\
&= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(f(x_t) - f(x^*))\right] \\
&\leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} g_t^\top (x_t - x^*)\right] \\
&= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\hat{g}_{x_t} | x_t\right]^\top (x_t - x^*)\right] \\
&= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \hat{g}_{x_t}^\top (x_t - x^*)\right]
\end{aligned}$$

where the inequalities are due to the convexity of f and the last equality is due to the tower rule. Now let us consider functions $f_1, \ldots, f_T$ given by

$$f_t(x) = \hat{g}_{x_t}^\top x.$$

Then

$$\begin{aligned}
\sum_{t=1}^{T} \hat{g}_{x_t}^\top (x_t - x^*) &= \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*) \\
&\leq \sum_{t=1}^{T} f_t(x_t) - \min_{x \in C} \sum_{t=1}^{T} f_t(x) \\
&\leq \frac{3}{2} LR\sqrt{T}
\end{aligned}$$

where the last inequality is from the convergence result of online gradient descent. Note that this upper bound holds regardless of any realization of $\hat{g}_{x_t}$'s. Therefore, the result follows.

2.2 Proof from the analysis of the subgradient method

Note that

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2 | x_t\right] &= \mathbb{E}\left[\|\mathrm{Proj}_C(x_t - \eta_t \hat{g}_{x_t}) - x^*\|_2^2 | x_t\right] \\
&\leq \mathbb{E}\left[\|x_t - \eta_t \hat{g}_{x_t} - x^*\|_2^2 | x_t\right] \\
&= \|x_t - x^*\|_2^2 + \eta_t^2 \mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2 | x_t\right] - 2\eta_t \mathbb{E}\left[\hat{g}_{x_t} | x_t\right]^\top (x_t - x^*) \\
&= \|x_t - x^*\|_2^2 + \eta_t^2 \mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2 | x_t\right] - 2\eta_t g_t^\top (x_t - x^*) \\
&\leq \|x_t - x^*\|_2^2 + \eta_t^2 \mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2 | x_t\right] - 2\eta_t (f(x_t) - f(x^*)).
\end{aligned}$$

Then, based on the tower rule,

$$\begin{aligned}\mathbb{E}\left[\|x_{t+1}-x^*\|_2^2\right] &\leq \mathbb{E}\left[\|x_t-x^*\|_2^2\right] + \eta_t^2\mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2\right] - 2\eta_t(\mathbb{E}\left[f(x_t)\right]-f(x^*)) \\ &\leq \mathbb{E}\left[\|x_t-x^*\|_2^2\right] + \eta_t^2 L^2 - 2\eta_t(\mathbb{E}\left[f(x_t)\right]-f(x^*)).\end{aligned}$$

Then it follows that

$$\mathbb{E}\left[f(x_t)\right]-f(x^*) \leq \frac{1}{2\eta_t}\left(\mathbb{E}\left[\|x_t-x^*\|_2^2\right]-\mathbb{E}\left[\|x_{t+1}-x^*\|_2^2\right]\right)+\frac{\eta_t}{2}L^2.$$

Summing up this for $t=1,\ldots,T$ and dividing each side by T, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f(x_t)\right]-f(x^*) &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|x_t-x^*\|_2^2\right]\left(\frac{1}{2\eta_t}-\frac{1}{2\eta_{t-1}}\right)+\frac{L^2}{2T}\sum_{t=1}^{T}\eta_t \\ &\leq \frac{R^2}{T}\sum_{t=1}^{T}\left(\frac{1}{2\eta_t}-\frac{1}{2\eta_{t-1}}\right)+\frac{L^2}{2T}\sum_{t=1}^{T}\eta_t \\ &\leq \frac{LR}{2\sqrt{T}}+\frac{LR}{\sqrt{T}}.\end{aligned}$$

By convexity,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f(x_t)\right] \geq \mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}x_t\right)\right],$$

and therefore, the result follows.

2.3 Strongly convex functions

For strongly convex functions, we have the following convergence result.

Theorem 15.2. *Assume the same conditions on $\hat{g}_x$ and that f is α-strongly convex with respect to the ℓ_2 norm for some $\alpha > 0$. Algorithm 1 with step sizes $\eta_t = 2/(\alpha(t+1))$ satisfies*

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

where the expectation is taken over the randomness in gradient estimation and $x^* \in \mathrm{argmin}_{x\in C} f(x)$.

Therefore, for Lipschitz continuous functions and functions that are stronngly convex and Lipschitz, we recover the same convergence rate as the subgradient method.

2.4 No self-tuning property due to variance

For gradient descent, smoothness does make difference due to the self-tuning property. For smooth functions, the convergence rate is $O(1/T)$ (we also saw the accelerated method achieving $O(1/T^2)$ rate). For smooth and strongly convex functions, we obtained $O(\gamma^T)$ rate for some $0 < \gamma < 1$. Is it the case for SGD as well? The answer is no.

The crucial property of smooth functions which we relied on in the convergence analysis was the self-tuning property. For a smooth function f, as we get close to an optimal solution $x^* \in$

$\mathrm{argmin}_{x\in\mathbb{R}^d} f(x)$, the size of the gradient $\|\nabla f(x)\|_2$ gets smaller. However, even if f is smooth and x goes to x^*, $\mathbb{E}\left[\|\hat{g}_x\|_2^2\right]$ does not converge to 0.

Let us consider the mean squared error minimization problem given by

$$\min_{\beta} \quad f(\beta) = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}(y_i - \beta^\top x_i)^2.$$

Here, f is smooth because

$$\begin{aligned}
\|\nabla f(\beta_1) - \nabla f(\beta_2)\|_2 &= \left\| \frac{1}{n}\sum_{i=1}^n (\beta_1 - \beta_2)^\top x_i x_i \right\|_2 \\
&\leq \frac{1}{n}\sum_{i=1}^n |(\beta_1 - \beta_2)^\top x_i| \, \|x_i\|_2 \\
&\leq \|\beta_1 - \beta_2\|_2 \left(\frac{1}{n}\sum_{i=1}^n \|x_i\|_2^2 \right) \\
&\leq M^2 \|\beta_1 - \beta_2\|_2
\end{aligned}$$

where $\max_{i\in[n]} \|x_i\| = M$.

Next take the optimal solution $\beta^* \in \mathrm{argmin}_\beta f(\beta)$ which satisfies $\nabla f(\beta^*) = 0$. Then sample a data point (x_i, y_i) to obtain an unbiased estimator

$$\hat{g}_{\beta^*} = (y_i - x_i^\top \beta^*)(-x_i).$$

Here, if the data point (x_i, y_i) is not on the line $y = \beta^\top x$ and x_i is nonzero, then $\hat{g}_{\beta^*} \neq 0$.

3 Mini-batch SGD

In this section, we consider the relationship between the variance in sampling stochastic gradients and the convergence rate of SGD.

Assumption 2. Assume that $\hat{g}_x$ satisfies

$$\mathbb{E}[\hat{g}_x] = g_x \text{ for some } g_x \in \partial f(x), \quad \mathrm{Var}(\hat{g}_x) \leq \sigma^2.$$

We further assume that

$$\|g_x\|_2 \leq L \quad \text{for all } g_x \in \partial f(x).$$

This is in contrast to assuming that $\mathbb{E}\left[\|\hat{g}_x\|_2^2\right] \leq L^2$ for all x. Basically, the objective function f is L-Lipschitz continuous, and we obtain stochastic estimates of its subgradients. Note that

$$\begin{aligned}
\mathrm{Var}(\hat{g}_x) &= \mathbb{E}\left[\|\hat{g}_x - \mathbb{E}[\hat{g}_x]\|_2^2\right] \\
&= \mathbb{E}\left[\|\hat{g}_x - g_x\|_2^2\right].
\end{aligned}$$

What does this imply in terms of the convergence of stochastic gradient descent? Note that

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2 \,|x_t\right] &= \mathbb{E}\left[\|\mathrm{Proj}_C(x_t - \eta_t \hat{g}_{x_t}) - x^*\|_2^2 \,|x_t\right] \\
&\leq \mathbb{E}\left[\|x_t - \eta_t \hat{g}_{x_t} - x^*\|_2^2 \,|x_t\right] \\
&= \|x_t - x^*\|_2^2 + \eta_t^2 \mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2 \,|x_t\right] - 2\eta_t \mathbb{E}\left[\hat{g}_{x_t}|x_t\right]^\top (x_t - x^*) \\
&= \|x_t - x^*\|_2^2 + \eta_t^2 \mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2 \,|x_t\right] - 2\eta_t g_t^\top (x_t - x^*) \\
&\leq \|x_t - x^*\|_2^2 + \eta_t^2 \mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2 \,|x_t\right] - 2\eta_t (f(x_t) - f(x^*)).
\end{aligned}$$

Here, we look at the term $\mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2 \,|x_t\right]$. Note that

$$\begin{aligned}
\mathbb{E}\left[\|\hat{g}_{x_t}\|_2^2 \,|x_t\right] &= \mathbb{E}\left[\|\hat{g}_{x_t} - g_t + g_t\|_2^2 \,|x_t\right] \\
&= \mathbb{E}\left[\|\hat{g}_{x_t} - g_t\|_2^2 \,|x_t\right] + \mathbb{E}\left[\|g_t\|_2^2 \,|x_t\right] + 2\mathbb{E}\left[g_t^\top (\hat{g}_{x_t} - g_t)|x_t\right] \\
&= \mathbb{E}\left[\|\hat{g}_{x_t} - g_t\|_2^2 \,|x_t\right] + \|g_t\|_2^2 + 2g_t^\top \mathbb{E}\left[\hat{g}_{x_t} - g_t|x_t\right] \\
&= \mathbb{E}\left[\|\hat{g}_{x_t} - g_t\|_2^2 \,|x_t\right] + \|g_t\|_2^2 \\
&\leq \sigma^2 + L^2.
\end{aligned}$$

Then, based on the tower rule,

$$\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2\right] \leq \mathbb{E}\left[\|x_t - x^*\|_2^2\right] + \eta_t^2(\sigma^2 + L^2) - 2\eta_t(\mathbb{E}\left[f(x_t)\right] - f(x^*))$$

Then it follows that

$$\mathbb{E}\left[f(x_t)\right] - f(x^*) \leq \frac{1}{2\eta_t}\left(\mathbb{E}\left[\|x_t - x^*\|_2^2\right] - \mathbb{E}\left[\|x_{t+1} - x^*\|_2^2\right]\right) + \frac{\eta_t}{2}(L^2 + \sigma^2).$$

Here, the last term in the right-hand side has $L^2 + \sigma^2$, instead of L^2. For the deterministic case, we had $\sigma = 0$, which recovers the analysis of the subgradient method. Hence, when the variance term σ^2 is large, the convergence rate gets worse. Therefore, one way to improve the convergence of SGD is to reduce the variance.

One way to reduce the variance is through sampling a batch of stochastic gradients, instead of a single one. Suppose that at $x \in C$, we sample $\hat{g}_x^1, \ldots, \hat{g}_x^B$ independently at random. Assuming

$$\mathbb{E}\left[\hat{g}_x^1\right] = \cdots = \mathbb{E}\left[\hat{g}_x^B\right] = g_x \text{ for some } g_x \in \partial f(x),$$

it follows that

$$\hat{g}_x = \frac{1}{B}\left(\hat{g}_x^1 + \cdots + \hat{g}_x^B\right)$$

is an unbiased estimator of g_x. Since $\hat{g}_x^1, \ldots, \hat{g}_x^B$ are independent,

$$\mathbb{E}\left[\|g_x - \hat{g}_x\|_2^2\right] = \mathbb{E}\left[\left\|g_x - \frac{1}{B}\sum_{i=1}^B \hat{g}_x^i\right\|_2^2\right] = \frac{1}{B^2}\sum_{i=1}^B \mathbb{E}\left[\|g_x - \hat{g}_x^i\|_2^2\right] \leq \frac{1}{B}\sigma^2.$$

Therefore, taking $\hat{g}_x$ as the average of a batch of the unbiased estimators $\hat{g}_x^1, \ldots, \hat{g}_x^B$ that are pairwise independent, we can reduce the variance from σ^2 to σ^2/B. Note that sampling or computing a gradient estimate $\hat{g}_x^i$ can be parallelizable. Stochastic gradient descent that uses the average of a batch of gradient estimates is often called mini-batch SGD.