CHAPTER 6. STOCHASTIC GRADIENT DESCENT

Kun Yuan

April 11, 2024

1 Problem formulation

This chapter considers the following stochastic optimization problem

$$\min_{x \in \mathbb{R}^d} \quad f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[F(x; \xi)] \tag{1}$$

where $\xi \sim \mathcal{D}$ denotes the random data sample and $\mathcal{D}$ denotes the data distribution. Since $\mathcal{D}$ is typically unknown in machine learning, the closed-form of $f(x)$ is also unknown.

Notation. We introduce the following notations:

- Let $x^\star := \arg\min_{x \in \mathbb{R}^d} \{f(x)\}$ be the optimal solution to problem (1).
- Let $f^\star := \min_{x \in \mathbb{R}^d} \{f(x)\}$ be the optimal function value.
- Let $\mathcal{F}_k = \{x_k, \xi_{k-1}, x_{k-1}, \cdots, \xi_0\}$ be the filtration containing all historical variables at and before iteration k. Note that ξ_k does not belong to $\mathcal{F}_k$.

2 Stochastic gradient descent

Since $f(x)$ does not have a closed-form, we cannot access its gradient. However, since $F(x; \xi)$ is known, we can use $\nabla_x F(x; \xi)$ to approximate the true gradient $\nabla f(x)$. Throughout this lecture, we let $\nabla F(x; \xi) = \nabla_x F(x; \xi)$ for notation simplicity. Given any arbitrary initialization variable x_0, stochastic gradient descent (SGD) iterates as follows

$$x_{k+1} = x_k - \gamma \nabla F(x_k; \xi_k), \quad \forall\, k = 0, 1, 2, \cdots \tag{2}$$

where γ is the learning rate, and $\xi_k \sim \mathcal{D}$ is a random data sampled at iteration k. Since ξ_k is a random variable for any $k = 0, 1, \cdots$, each variable x_k is also a random variable for $k = 1, 2, \cdots$.

3 Convergence analysis

To facilitate convergence analysis, we introduce the following assumption:

Assumption 3.1. Given the filtration $\mathcal{F}_k$, we assume

$$\mathbb{E}[\nabla F(x_k; \xi_k) | \mathcal{F}_k] = \nabla f(x_k) \tag{3}$$

$$\mathbb{E}[\|\nabla F(x_k; \xi_k) - \nabla f(x_k)\|^2 | \mathcal{F}_k] \le \sigma^2 \tag{4}$$

The above assumption indicates that, conditioned on the filtration $\mathcal{F}_k$, the stochastic gradient $\nabla F(x_k; \xi_k)$ is an unbiased estimate on $\nabla f(x_k)$, and the variance is bounded by σ^2. Under the above assumption, it is easy to verify that

$$\begin{aligned}
\mathbb{E}[\|\nabla F(x_k; \xi_k)\|^2 | \mathcal{F}_k] &= \mathbb{E}[\|\nabla F(x_k; \xi_k) - \nabla f(x_k) + \nabla f(x_k)\|^2 | \mathcal{F}_k] \\
&= \|\nabla f(x_k)\|^2 + \mathbb{E}[\|\nabla F(x_k; \xi_k) - \nabla f(x_k)\|^2 | \mathcal{F}_k] \\
&\le \|\nabla f(x_k)\|^2 + \sigma^2 \qquad (5)
\end{aligned}$$

where the second equality holds due to (3) and the last inequality holds due to (4).

3.1 Smooth and non-convex problem

Theorem 3.2. Suppose $f(x)$ is L-smooth and Assumption 3.1 holds. If $\gamma \le 1/L$, SGD will converge at the following rate

$$\frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}[\|\nabla f(x_k)\|^2] \le \frac{2\Delta_0}{\gamma(K+1)} + \gamma L \sigma^2, \tag{6}$$

where $\Delta_0 = f(x_0) - f^\star$. If we further choose $\gamma = \left[\left(\frac{2\Delta_0}{(K+1)L\sigma^2}\right)^{-\frac{1}{2}} + L\right]^{-1}$, SGD converges as

$$\frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}[\|\nabla f(x_k)\|^2] \le \sqrt{\frac{8L\Delta_0\sigma^2}{K+1}} + \frac{2L\Delta_0}{K+1}. \tag{7}$$

Remark. If $\sigma^2 = 0$, the stochastic gradient reduces to the true gradient, and hence, SGD reduces to GD. Substituting $\sigma^2 = 0$ to SGD rate (7), we recover the rate $O(L/K)$ for GD. In other words, our convergence rate for SGD is consistent with GD.

Proof. Since $f(x)$ is L-smooth, we have

$$
\begin{aligned}
\mathbb{E}[f(x_{k+1})|\mathcal{F}_k] &\le f(x_k) + \mathbb{E}[\langle \nabla f(x_k), x_{k+1} - x_k\rangle|\mathcal{F}_k] + \frac{L}{2}\mathbb{E}[\|x_{k+1} - x_k\|^2|\mathcal{F}_k] \\
&= f(x_k) - \gamma\mathbb{E}[\langle \nabla f(x_k), \nabla F(x_k;\xi_k)\rangle|\mathcal{F}_k] + \frac{L\gamma^2}{2}\mathbb{E}[\|\nabla F(x_k;\xi_k)\|^2|\mathcal{F}_k] \\
&\overset{(a)}{\le} f(x_k) - \gamma(1 - \frac{L\gamma}{2})\|\nabla f(x_k)\|^2 + \frac{L\gamma^2\sigma^2}{2} \\
&\overset{(b)}{\le} f(x_k) - \frac{\gamma}{2}\|\nabla f(x_k)\|^2 + \frac{L\gamma^2\sigma^2}{2} \qquad (8)
\end{aligned}
$$

where inequality (a) holds due to Assumption 3.1, and inequality (b) holds if $\gamma \le 1/L$. By taking expectations over the filtration $\mathcal{F}_k$, we have

$$
\mathbb{E}[f(x_{k+1})] \le \mathbb{E}[f(x_k)] - \frac{\gamma}{2}\mathbb{E}[\|\nabla f(x_k)\|^2] + \frac{L\gamma^2\sigma^2}{2} \qquad (9)
$$

which is equivalent to

$$
\mathbb{E}[\|\nabla f(x_k)\|^2] \le \frac{2}{\gamma}\left(\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]\right) + \gamma L\sigma^2 \qquad (10)
$$

Taking averaging over $k = 0, 1, \cdots, K$, we have

$$
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(x_k)\|^2] \le \frac{2(f(x_0) - f^\star)}{\gamma(K+1)} + \gamma L\sigma^2. \qquad (11)
$$

Defining $\Delta_0 := f(x_0) - f^\star$, if we set

$$
\gamma = \left[\left(\frac{2\Delta_0}{(K+1)L\sigma^2}\right)^{-\frac{1}{2}} + L\right]^{-1}, \qquad (12)
$$

it then holds that

$$
\gamma \le \min\left\{\frac{1}{L}, \gamma_1\right\}, \quad \text{where} \quad \gamma_1 = \left(\frac{2\Delta_0}{(K+1)L\sigma^2}\right)^{\frac{1}{2}}. \qquad (13)
$$

Substituting (12) and (13) into (11), we have

$$
\begin{aligned}
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(x_k)\|^2] &\le \frac{2(f(x_0) - f^\star)}{\gamma(K+1)} + \gamma_1 L\sigma^2 \\
&= 2\sqrt{\frac{2L\Delta_0\sigma^2}{K+1}} + \frac{2L\Delta_0}{K+1}. \qquad (14)
\end{aligned}
$$

□