CS2910 - Optimization Summer 2023

Lecture 5 — Stochastic Gradient Descent

Lecturer: Yunwei Ren *Scribed by Zhidan Li*

Contents

1 Overview

Now we introduce another variant of gradient descent method called *stochastic gradient descent*. Instead of walk against the gradient, at each update, we introduce some 'random' step size. We will analyze the convergence rate of stochastic gradient descent, and show the reason why we introduce this variant of gradient descent.

2 General Form of Stochastic Gradient Descent

Now we generalize the general form of the stochastic gradient descent.

Definition 1 (stochastic gradient descent)**.** *Given $f \in C^1(\mathbb{R}^d)$ and a step size $\eta > 0$, the **stochastic gradient descent** $\{x_k\}_{k\in\mathbb{N}}$ generated by f is defined as the update rule*

$$x_{k+1} = x_k - \eta \widetilde{\nabla} f(x_k)$$

where $\widetilde{\nabla} f(x_k)$ is a random variable with $\mathbb{E}\left[\widetilde{\nabla} f(x_k)\right] = \nabla f(x_k)$.

Remark 1. We will assume that the 'variance' is bounded along the SGD trajectory, i.e., $\mathbb{E}\left[\left\|\widetilde{\nabla} f(x_k)\right\|^2\right] \leq \sigma^2$ for all $k \in \mathbb{R}^d$. Also it's equivalent to use the 'real' variance $\mathbb{E}\left[\left\|\widetilde{\nabla} f(x_k) - \nabla f(x_k)\right\|^2\right]$ since usually the two terms have the same order.

Now we establish its convergence rate (under some regularity conditions)

Lemma 2 (SGD lemma)**.** *Under the assumptions of Definition 1, in addition if f is L-smooth, then*

$$f(x_{k+1}) \leq f(x_k) - \eta \left\langle \nabla f(x_k), \widetilde{\nabla} f(x_k) \right\rangle + \frac{L}{2}\eta^2 \left\| \widetilde{\nabla} f(x_k) \right\|^2.$$

Furthermore, assuming the bounded variance during SGD, it holds that

$$\mathbb{E}\left[f(x_{k+1})\right] \leq f(x_k) - \eta \|\nabla f(x_k)\|^2 + \frac{L}{2}\eta^2\sigma^2.$$

Proof. The proof is similar to what we do in the analysis of gradient descent. Since f is L-smooth, then

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 \\ &= f(x_k) - \eta \left\langle \nabla f(x_k), \widetilde{\nabla} f(x_k) \right\rangle + \frac{L}{2}\eta^2 \left\| \widetilde{\nabla} f(x_k) \right\|^2. \end{aligned}$$

Taking expectation on both sides we get the second equality. □

Lemma 3 (mirror descent for SGD)**.** *If f is convex, then for all $y \in \mathbb{R}^d$,*

$$f(x_k) \leq f(y) + \frac{1}{2\eta}\mathbb{E}\left[\|x_k - y\|^2 - \|x_{k+1} - y\|^2 + \|x_k - x_{k+1}\|^2\right].$$

Proof. By the linear lower bound, it holds that for all $y \in \mathbb{R}^d$,

$$\begin{aligned} f(y) &\geq f(x_k) + \langle \nabla f(x_k), y - x_k \rangle \\ &= f(x_k) + \frac{1}{\eta}\left\langle \mathbb{E}\left[x_k - x_{k+1}\right], y - x_k \right\rangle \\ &= f(x_k) + \frac{1}{\eta}\mathbb{E}\left[\langle y - x_k, x_k - x_{k+1} \rangle\right] \\ &= f(x_k) + \frac{1}{2\eta}\mathbb{E}\left[\|x_k - y\|^2 - \|x_{k+1} - y\|^2 + \|x_k - x_{k+1}\|^2\right]. \end{aligned}$$

□

By Lemma 3, we obtain the following convergence rate of SGD.

Proposition 4 (convergence rate of SGD)**.** *Under the settings of Definition 1 and given $T > 0$ be the total number of iterations, suppose that $\mathbb{E}\left[\left\|\widetilde{\nabla} f(x_k)\right\|^2\right] \leq \sigma^2$ for some $\sigma \geq 0$ and $k \leq T$. If we pick $\eta = \|x_0 - x_*\|/\sigma\sqrt{T}$, then*

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[f(x_k)\right] \leq f(x_*) + \frac{\sigma\|x_0 - x_*\|}{\sqrt{T}}.$$

Proof. From Lemma 3, for all $0 \le k \le T-1$, it holds that

$$\begin{aligned} f(x_k) &\le f(x_*) + \frac{1}{2\eta}\mathbb{E}\left[\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2 + \|x_k - x_{k+1}\|^2\right] \\ &\le f(x_*) + \frac{1}{2\eta}\mathbb{E}\left[\|x_k - y\|^2 - \|x_{k+1} - y\|^2 + \eta^2\sigma^2\right]. \end{aligned}$$

Summing over both sides, we obtain

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[f(x_k)\right] \le f(x_*) + \frac{\|x_0 - x_*\|^2}{2\eta T} + \frac{\eta\sigma^2}{2}.$$

Choose $\eta = \|x_0 - x_*\|/\sigma\sqrt{T}$, and we prove the proposition. □

Remark 2. Recall the convergence rate of gradient descent holds when $\eta = O\left(\frac{1}{T}\right)$, while in stochastic gradient descent we pick $\eta = O\left(\frac{1}{\sqrt{T}}\right)$. The reason is, in gradient descent, we have the gradient descent lemma bound $\sum_k \|\nabla f(x_k)\|^2$. But in stochastic gradient descent, even when x_k is near x_*, due to the randomness, we could not guarantee $\left\|\widetilde{\nabla} f(x)\right\|^2$ is small.

3 SGD and Empirical Risk Minimization (ERM)

Now we show an application of SGD. We aim to solve the *empirical risk minimization* (ERM) problem.

Given a dataset $\{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^{d+1}$, a model $h(x; w)$ with input x and parameter/weight w, and a regularizer $R(w)$, our goal is

$$\min_w L(w) = \frac{1}{n}\sum_{i=1}^n \ell(h(x_i; w), y_i) + R(w) \tag{1}$$

where the function ℓ is called per-sample loss, e.g., ℓ_2-norm in $\mathbb{R}^d$.

3.1 Mini-batch stochastic gradient descent

Now we show how to choose $\widetilde{\nabla} L(w_k)$. Consider a random multiset $B \subseteq [n]$ with size $b = |B|$. Then we set

$$\widetilde{\nabla} L(w) \triangleq \frac{1}{b}\sum_{i\in B}\nabla\ell(h(x_i; w), y_i) + \nabla R(w).$$

We simply claim that $\mathbb{E}\left[\widetilde{\nabla} L(w)\right] = \nabla L(w)$. Thus we can apply our stochastic gradient descent assuming some convexity.

Remark 3. We compare the gradient descent and the mini-batch stochastic gradient descent. Directly from our analysis, the gradient descent needs $O(1/\varepsilon)$ iterations and thus $O(n/\varepsilon)$ computations, while the mini-batch stochastic needs $O(1/\varepsilon^2)$ iterations and $O(b/\varepsilon^2)$ computations. When $n/b \le 1/\varepsilon$, the mini-batch stochastic gradient descent is more efficient.