18

Stochastic Gradient Descent

In this lecture we introduce **Stochastic Gradient Descent** (SGD), a probabilistic version of gradient descent that has been around since the 1950s, and that has become popular in the context of data science and machine learning. To motivate the algorithm, consider a set of functions $\mathcal{H} = \{h_{\boldsymbol{w}} \colon \boldsymbol{w} \in \mathbb{R}^d\}$, where each such function depends on d parameters. Also consider a smooth loss functions L, or a smooth approximation of a loss function. Given samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$, define the functions

$$f_i(\boldsymbol{w}) = L(h_{\boldsymbol{w}}(\boldsymbol{x}_i), \boldsymbol{y}_i), \quad i \in \{1, \dots, n\}.$$

The problem of finding functions that minimize the empirical risk is

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{w}).$$

The f_i are often assumed to be convex and smooth. In addition one often considers a regularization term $R(\boldsymbol{w})$. In what follows, we abstract from the machine learning context and consider purely the associated optimization problem.

Stochastic Gradient Descent

We consider an objective function of the form

$$f(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{w}). \tag{18.1}$$

In what follows we assume that the functions f_i are convex and differentiable. If n is large, then computing the gradient can be very expensive. However, and considering the machine learning context, where $f(\boldsymbol{w})$ is an estimator of the generalization risk $\mathbb{E}_{\boldsymbol{\xi}}[f_{\boldsymbol{\xi}}(\boldsymbol{w})]$ of a family of functions $f_{\boldsymbol{\xi}}$ parametrized by a random vector $\boldsymbol{\xi}$, we can shift the focus to finding an **unbiased estimator** of the gradient. Quite trivially, choosing

an index j uniformly at random and computing the gradient of $f_j(\boldsymbol{w})$ gives such an unbiased estimator *by definition*:

$$\mathbb{E}_U[f_U(\boldsymbol{w})] = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w}),$$

where $\mathbb{P}\{U = j\} = 1/n$ for $j \in [n] = \{1, \ldots, n\}$. The Stochastic Gradient Descent (SGD) algorithm proceeds as follows. Begin with $\boldsymbol{w}^0 \in \mathbb{R}^d$. At each step k, compute an unbiased estimator $\boldsymbol{g}^k$ of the gradient at $\boldsymbol{w}^k$:

$$\mathbb{E}[\boldsymbol{g}^k \mid \boldsymbol{w}^k] = \nabla f(\boldsymbol{w}^k).$$

Next, take a step in direction $\boldsymbol{g}^k$:

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \eta_k \boldsymbol{g}^k,$$

where η_k is a step length (or **learning rate** in machine learning jargon).

While there are many variants of stochastic gradient descent, we consider the simplest version in which $\boldsymbol{g}^k$ is chosen by picking one of the gradients $\nabla f_i(\boldsymbol{w})$ uniformly at random, and we refer to this as SGD *with uniform sampling*. A commonly used generalization is **mini-batch** sampling, where one chooses a small set of indices $I \subset \{1, \ldots, n\}$ at random, instead of only one. We also restrict to the smooth setting without a regularization term; in the non-smooth setting one would apply a proximal operator. Since SGD involves random choices, convergence results are stated in terms of the expected value. Let U be a random variable with distribution $\mathbb{P}\{U = i\} = 1/n$ for $i \in [n]$. Then

$$\mathbb{E}_U[\nabla f_U(\boldsymbol{w})] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{w}) = \nabla f(\boldsymbol{w}),$$

so that ∇f_U is an unbiased estimator of ∇f. Assuming that f has a unique minimizer $\boldsymbol{w}^*$, we define the empirical **variance** at the optimal point $\boldsymbol{w}^*$ as

$$\sigma^2 = \mathbb{E}_U[\|\nabla f_U(\boldsymbol{w}^*)\|^2] = \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\boldsymbol{w}^*)\|^2. \tag{18.2}$$

We can now state the convergence result for stochastic gradient descent.

Theorem 18.1. *Assume the function f is α-strongly convex and that the f_i are convex and β-smooth for $i \in [n]$ and $4\beta > \alpha$. Assume f has a unique minimizer $\boldsymbol{w}^*$ and define the variance as in* (18.2)*. Then for any staring point $\boldsymbol{w}^0$, the sequence of iterates $\{\boldsymbol{w}^k\}$ generated by SGD with uniform sampling and step length $\eta = 1/(2\beta)$ satisfies*

$$\mathbb{E}[\|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2] \leq \left(1 - \frac{\alpha}{4\beta}\right)^k \|\boldsymbol{w}^0 - \boldsymbol{w}^*\|^2 + \frac{2\sigma^2}{\alpha\beta}.$$

Proof. As in the analysis of gradient descent, we get

$$\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|^2 = \|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2 - 2\eta \nabla f_U(\boldsymbol{w}^k)^\top (\boldsymbol{w}^k - \boldsymbol{w}^*) + \eta^2 \|\nabla f_U(\boldsymbol{w}^k)\|^2.$$

Taking the expectation conditional on $\boldsymbol{w}^k$, we get

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|^2 \mid \boldsymbol{w}^k] = \|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2 - 2\eta \nabla f(\boldsymbol{w}^k)^\top (\boldsymbol{w}^k - \boldsymbol{w}^*) \\ + \eta^2 \,\mathbb{E}[\|\nabla f_U(\boldsymbol{w}^k)\|^2 \mid \boldsymbol{w}^k],\end{aligned} \tag{18.3}$$

where we used the fact that the expectation satisfies $\mathbb{E}[\nabla f_U(\boldsymbol{w})] = \nabla f(\boldsymbol{w})$. For the last term we use the bound

$$\begin{aligned}\mathbb{E}[\|\nabla f_U(\boldsymbol{w}^k)\|^2 \mid \boldsymbol{w}^k] &= \mathbb{E}[\|\nabla f_U(\boldsymbol{w}^k) - \nabla f_U(\boldsymbol{w}^*) + \nabla f_U(\boldsymbol{w}^*)\|^2] \\ &\le 2\,\mathbb{E}[\|\nabla f_U(\boldsymbol{w}^k) - \nabla f_U(\boldsymbol{w}^*)\|^2] + 2\,\mathbb{E}[\|\nabla f_U(\boldsymbol{w}^*)\|^2] \\ &= 2\,\mathbb{E}[\|\nabla f_U(\boldsymbol{w}^k) - \nabla f_U(\boldsymbol{w}^*)\|^2] + 2\sigma^2.\end{aligned}$$

Using the characterization of β smoothness from the previous chapter, namely

$$\frac{1}{\beta}\|\nabla f_i(\boldsymbol{w}) - \nabla f_i(\boldsymbol{v})\|^2 \le (\nabla f_i(\boldsymbol{w}) - \nabla f_i(\boldsymbol{v}))^\top (\boldsymbol{w} - \boldsymbol{v}), \tag{18.4}$$

we get that

$$\begin{aligned}\mathbb{E}[\|\nabla f_U(\boldsymbol{w}^k) - \nabla f_U(\boldsymbol{w}^*)\|^2 \mid \boldsymbol{w}^k] &= \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{w}^k) - \nabla f_i(\boldsymbol{w}^*)\|^2 \\ &\overset{(18.4)}{\le} \frac{\beta}{n}\sum_{i=1}^n (\nabla f_i(\boldsymbol{w}^k) - \nabla f_i(\boldsymbol{w}^*))^\top (\boldsymbol{w}^k - \boldsymbol{w}^*) \\ &= \frac{\beta}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{w}^k)^\top (\boldsymbol{w}^k - \boldsymbol{w}^*) \\ &\qquad - \frac{\beta}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{w}^*)^\top (\boldsymbol{w}^k - \boldsymbol{w}^*) \\ &= \beta \nabla f(\boldsymbol{w})^\top (\boldsymbol{w}^k - \boldsymbol{w}^*),\end{aligned}$$

where we used that $\nabla f(\boldsymbol{w}^*) = \boldsymbol{0}$ for the last equality. Hence, we get the bound

$$\mathbb{E}[\|\nabla f_U(\boldsymbol{w}^k)\|^2 \mid \boldsymbol{w}^k] \le 2\beta \nabla f(\boldsymbol{w})^\top (\boldsymbol{w}^k - \boldsymbol{w}^*) + 2\sigma^2.$$

Plugging this into (18.3), we get

$$\mathbb{E}[\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|^2 \mid \boldsymbol{w}^k] \le \|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2 - (2\eta - 2\eta^2\beta)\nabla f(\boldsymbol{w}^k)^\top (\boldsymbol{w}^k - \boldsymbol{w}^*) + 2\eta^2\sigma^2.$$

Using the α-strong convexity, we get the bound

$$\nabla f(\boldsymbol{w}^k)^\top \boldsymbol{w}^k - \boldsymbol{w}^* \ge f(\boldsymbol{w}^k) - f(\boldsymbol{w}^*) + \frac{\alpha}{2}\|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2 \ge \frac{\alpha}{2}\|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2,$$

since $f(\boldsymbol{w}^k) \geq f(\boldsymbol{w}^*)$, so that we get

$$\mathbb{E}[\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|^2 \mid \boldsymbol{w}^k] \leq (1 - \eta\alpha(1 - \eta\beta))\|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2 + 2\eta^2\sigma^2.$$

With step length $\eta = 1/(2\beta)$, and taking the expected value over all previous iterates, we get

$$\mathbb{E}[\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|^2] \leq \left(1 - \frac{\alpha}{4\beta}\right) \mathbb{E}[\|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2] + \frac{\sigma^2}{2\beta^2}.$$

Applying this bound recursively (and moving the index down), we get

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2] &\leq \left(1 - \frac{\alpha}{4\beta}\right)^k \|\boldsymbol{w}^0 - \boldsymbol{w}^*\|^2 + \frac{\sigma^2}{2\beta^2} \sum_{j=0}^{k-1} \left(1 - \frac{\alpha}{4\beta}\right)^j \\ &\leq \left(1 - \frac{\alpha}{4\beta}\right)^k \|\boldsymbol{w}^0 - \boldsymbol{w}^*\|^2 + \frac{2\sigma^2}{\alpha\beta},\end{aligned}$$

where we used that $4\beta > \alpha$. □

Example 18.2. Consider the problem of **logistic regression**, where the aim is to minimize the objective function

$$f(\boldsymbol{w}) = \sum_{i=1}^{n} \left(\log\left(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{w})\right) - y_i \boldsymbol{x}_i^T \boldsymbol{w}\right)$$

over a vector of weights $\boldsymbol{w}$. This problem arises in the context of a binary classification problem with data pairs $(\boldsymbol{x}_i, y_i)$ and $y_i \in \{0, 1\}$. Setting

$$p := \frac{\mathrm{e}^{\boldsymbol{x}^T \boldsymbol{w}}}{1 + \mathrm{e}^{\boldsymbol{x}^T \boldsymbol{w}}},$$

the resulting classifier is the function

$$h(\boldsymbol{w}) = \begin{cases} 1 & p > 1/2 \\ 0 & p \leq 1/2 \end{cases}$$

The function f is convex, and the gradient is

$$\nabla f(\boldsymbol{w}) = -\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{p}(\boldsymbol{w})),$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is the matrix with the $\boldsymbol{x}_i^T$ as rows, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, and $\boldsymbol{p}(\boldsymbol{w}) \in \mathbb{R}^n$ has coordinates

$$p_i(\boldsymbol{w}) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{w})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{w})}, \quad 1 \leq i \leq n.$$

We can apply different versions of gradient descent to this problem. Figure 1 shows the typical paths of gradient descent and of stochastic gradient descent for a problem with 100 data points. Note that using a naive approach to computing the gradient, one would need to compute 100 gradients at each step. Stochastic gradient descent, on the other hand, fails to converge due to the variance of the gradient estimator (see Figure 2).

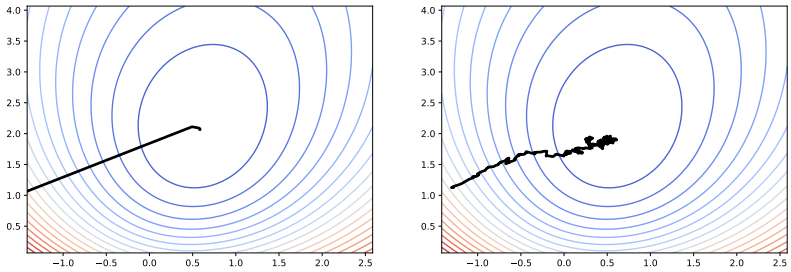

Figure 18.1: The path of gradient descent and of stochastic gradient descent with constant step length.

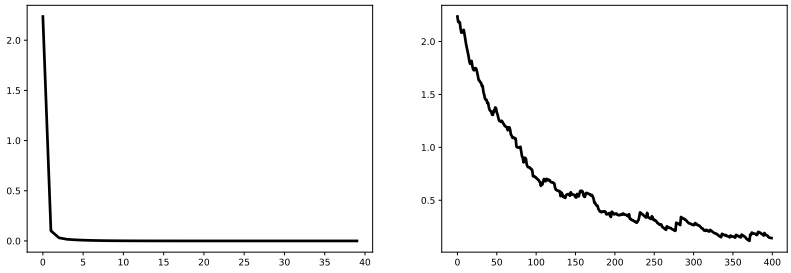

Figure 18.2: Convergence of gradient descent and SGD.

Extensions

The version of SGD described here is the most basic one. There are many possible extensions to the methods. The include considering different sampling schemes, including **mini-batching** and **importance sampling**. These sampling strategies have the effect of reducing the variance σ^2. In addition, improvements can be made in the step length selection and when dealing with non-smooth function, where the proximal operator comes into play.

Notes

The origins of stochastic gradient descent go back to the work of Robbins and Monro in 1951 [4]. The algorithm has been rediscovered many times, and gained popularity due to its effectiveness in training deep neural networks on large data sets, where gradient computations are very expensive. Despite its simplicity, a systematic and rigorous analysis has not been available until recently. The presentation in this chapter is based loosely on the papers [3] and [1]. A more general and systematic analysis of SGD that includes non-smooth objectives is given in [2]. These works also discuss general sampling techniques, not just uniform sampling.

[1] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[2] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. *arXiv preprint arXiv:1905.11261*, 2019.

[3] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. *arXiv preprint arXiv:1901.09401*, 2019.

[4] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.