ESTR2520: Optimization Methods **2023–24 Second Term**

Handout 9: Stochastic Gradient (SGD) Method and Its Analysis

Instructor: Hoi-To Wai Last updated: March 29, 2024

This note introduces the SGD method for stochastic optimization, and analyze its convergence.

1 Stochastic Gradient Method

Our quest is to solve the following **unconstrained**, **stochastic optimization** problem:

$$\min_{x\in\mathbb{R}^n} \; f(x) := \mathbb{E}_\xi[F(x;\xi)] \tag{1}$$

Given $x \in \mathbb{R}^n$, the objective function $f(x)$ is evaluated as the expectation of a random variable $F(x;\xi)$. In other words, let $\pi(\xi)$ be the pdf of ξ defined on the probability space Z, we have

$$f(x) = \int_{\mathsf{Z}} F(x;\xi)\pi(\xi)\,\mathrm{d}\xi.$$

Stochastic optimization problem such as (1) is motivated from applications when (A) ξ represents some nuance parameters that are random, or (B) ξ represents samples of data in a machine learning task and $\pi(\cdot)$ is the data distribution.

Example 1.1 *Consider a finite sum optimization problem (e.g., empirical risk in Handout 1):*

$$f(x) = \frac{1}{m}\sum_{i=1}^m F(x;\xi_i)$$

It is a special case of (1) *with an empirical distribution given by* $P(\xi = \xi_i) = 1/m$.

To this end, we shall also discuss when is (1) a *convex optimization* problem. The following fact gives a sufficient condition that is easy to check:

Fact 1 *If* $F(x;\xi)$ *is convex in* x *for any* $\xi \in \mathsf{Z}$*, then* $f(x)$ *is convex.*

Stochastic Gradient (SGD) Method

Input: x^0 - initialization, set the max no. of iteration as $T \geq 1$.

For $t = 0, 1, ..., T$,
  Draw $\xi^t \sim \pi(\cdot)$
  $x^{t+1} = x^t - \gamma_t \nabla F(x^t;\xi^t)$ where γ_t is a diminishing step size.
End

Output: $\bar{x}^T = \frac{\sum_{t=0}^T \gamma_t x^t}{\sum_{j=0}^T \gamma_j}$

Notice that the SGD method differs from the GD method by (A) it uses a **stochastic gradient** $\nabla F(x^t;\xi^t)$ for the update, (B) it outputs an *averaged* version of the iterates.

The SGD method was invented in 1960s in [3]. Yet it is only recently popularized with the boom in machine learning researches. For convex problems, the convergence of SGD is well known, with

$$\text{(convex)}\ \ \mathbb{E}[f(\bar{x}^T) - f^\star] = \mathcal{O}(1/\sqrt{T}) \quad \text{(strongly convex)}\ \ \mathbb{E}[\|x^T - x^\star\|^2] = \mathcal{O}(1/T)$$

Note that the convergence in the strongly convex case is that of the **last iterate**. In this lecture handout, we will present the proof for the convex case.

2 Convergence Analysis of SGD

Like in the last lectures, we assume:

Assumption 1 *The function $f(x)$ is convex and lower bounded, i.e.,* $\min_{x\in\mathbb{R}^n} f(x) > -\infty$.

Furthermore, the following conditions are needed for the stochastic gradients.

Assumption 2 *The stochastic gradient is unbiased, i.e.,*

$$\mathbb{E}_{\xi\sim\pi}[\nabla F(x;\xi)] = \nabla f(x),\ \forall\ x \in \mathbb{R}^n.$$

Assumption 3 *The stochastic gradient has a bounded variance, i.e., there exists σ such that*

$$\mathbb{E}_{\xi\sim\pi}[\|\nabla F(x;\xi) - \nabla f(x)\|^2] \le \sigma^2,\ \forall\ x \in \mathbb{R}^n.$$

The last condition is a strong assumption:

Assumption 4 *The gradient is always bounded, i.e., there exists G such that,*

$$\|\nabla f(x)\| \le G,\ \forall\ x \in \mathbb{R}^n.$$

However, we remark that it can be removed with a more sophisticated analysis.

Theorem 1: Convergence Rate of SGD Method

Under Assumption 1, 2, 3, 4. For any $T \ge 0$,

$$\mathbb{E}[f(\bar{x}^T) - f^\star] \le \frac{\|x^0 - x^\star\|^2 + (\sigma^2 + G^2)\sum_{t=0}^T \gamma_t^2}{2\sum_{t=0}^T \gamma_t}.$$

where the expectation is taken w.r.t. randomness in the SGD method.

Notice that by setting $\gamma_t = 1/\sqrt{t+1}$, we can achieve $\mathbb{E}[f(\bar{x}^T) - f^\star] = \mathcal{O}(\log T/\sqrt{T})$. In fact, the result implies that SGD converges to the optimal solution as long as one takes a step size sequence satisfying $\sum_{t=0}^\infty \gamma_t = \infty$, $\sum_{t=0}^\infty \gamma_t^2 < \infty$.

Proof In some way, the proof parallels that of the GD method in the convex objective case. Particularly, by the convexity of f, we observe that

$$f(x^\star) \ge f(x) + \nabla f(x)^\top (x^\star - x) \Longleftrightarrow \nabla f(x)^\top (x - x^\star) \ge f(x) - f(x^\star)$$

Thus,

$$\begin{aligned}\|x^{t+1}-x^\star\|^2 &= \|x^t-\gamma_t\nabla F(x^t;\xi^t)-x^\star\|^2\\ &= \|x^t-x^\star\|^2+\gamma_t^2\|\nabla F(x^t;\xi^t)\|^2-2\gamma_t\nabla F(x^t;\xi^t)^\top(x^t-x^\star)\end{aligned}$$

Taking the **conditional expectation** of $\mathbb{E}[\cdot|x^t]$, we yield

$$\mathbb{E}[\|x^{t+1}-x^\star\|^2|x^t]-\|x^t-x^\star\|^2=\mathbb{E}\left[\gamma_t^2\|\nabla F(x^t;\xi^t)\|^2-2\gamma_t\nabla F(x^t;\xi^t)^\top(x^t-x^\star)|x^t\right]$$

We have

$$\mathbb{E}[\nabla F(x^t;\xi^t)^\top(x^t-x^\star)|x^t] \overset{Ass.\ 2}{=} \nabla f(x^t)^\top(x^t-x^\star) \overset{Ass.\ 1}{\geq} f(x^t)-f^\star$$

and

$$\begin{aligned}\mathbb{E}[\|\nabla F(x^t;\xi^t)\|^2|x^t] &= \mathbb{E}[\|\nabla F(x^t;\xi^t)-\nabla f(x^t)+\nabla f(x^t)\|^2|x^t]\\ &\overset{Ass.\ 2}{=} \mathbb{E}[\|\nabla F(x^t;\xi^t)-\nabla f(x^t)\|^2|x^t]+\|\nabla f(x^t)\|^2\\ &\overset{Ass.\ 3,4}{\leq} \sigma^2+G^2.\end{aligned}$$

Thus the above yields

$$\mathbb{E}[\|x^{t+1}-x^\star\|^2|x^t]-\|x^t-x^\star\|^2\leq -2\gamma_t(f(x^t)-f^\star)+\gamma_t^2(\sigma^2+G^2)$$

Shuffling terms and taking the full expectation gives

$$2\gamma_t(f(x^t)-f^\star)\leq \mathbb{E}[\|x^t-x^\star\|^2-\|x^{t+1}-x^\star\|^2]+\gamma_t^2(\sigma^2+G^2)$$

Summing $t=0$ to $t=T$ gives

$$\sum_{t=0}^T\gamma_t(f(x^t)-f^\star)\leq\frac{1}{2}\left(\|x^0-x^\star\|^2+(\sigma^2+G^2)\sum_{t=0}^T\gamma_t^2\right)$$

Now, again applying the convexity of f yields

$$f(\bar{x}^T)-f^\star\leq\frac{1}{\sum_{j=0}^T\gamma_j}\sum_{t=0}^T\gamma_t(f(x^t)-f^\star)\leq\frac{1}{2\sum_{j=0}^T\gamma_j}\left(\|x^0-x^\star\|^2+(\sigma^2+G^2)\sum_{t=0}^T\gamma_t^2\right)$$

This concludes the proof. □

Extension Notice that the above analysis does not require the gradient to the Lipschitz continuous. In other words, it actually works for so-called **non-smooth** problems where the stochastic function is not differentiable. In the latter case, we can consider the **stochastic subgradient method**.

We recall that a vector $g\in\mathbb{R}^n$ is a subgradient at x for a convex function $f(x)$ is

$$f(y)-f(x)\geq g^\top(y-x),\ \forall\ y\in\mathbb{R}^n.$$

and we can denote $\partial f(x)$ to be the **set** of all subgradients at x.

The **stochastic subgradient method** can then be defined by replacing $\nabla F(x^t;\xi^t)$ by g^t such that $g^t\in\partial F(x^t;\xi^t)$. The convergence analysis follows exactly the same way we proved Theorem 1.

3 Perspectives

As a concluding remark, the SGD method is widely used in the literature and used in ML training. Especially given that the dataset nowadays are huge, e.g., with $m \approx 10^8$. It is an active research area with open questions such as studying other forms of convergence (e.g., with high probability, last iterate convergence), variance reduction techniques, distributed optimization, etc. For curiousity, you may check out the articles such as [2, 1].

References

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

[2] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

[3] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.