EE C227C Convex Optimization and Approximation Lecture 24 - 04/14/2022

Lecture 24: Stochastic Gradient Method

Lecturer: Jiantao Jiao *Scribe: Luis Rangel DaCosta, Jisun Lee*

In this lecture, we study the convergence guarantee of the stochastic gradient descent (SGD) in two different settings; convex and strongly convex.

1 Stochastic Gradient Descent (SGD)

We will consider the following problem

$$\min_x f(x) = \min_x \mathbb{E}_\xi F(x, \xi)$$

where $f(x)$ is difficult to access, while $F(x, \xi)$ is not. Our strategy will be to use a first-order oracle to query x and draw independent sample $\xi \sim \mathbb{P}_\xi$, then return $g(x, \xi) := \frac{\partial F(x,\xi)}{\partial x}(= \nabla F(x, \xi))$. Our gradient estimator, $g(x, \xi)$, is an unbiased estimator, such that

$$\mathbb{E}_\xi \left[\nabla F(x, \xi)\right] = \nabla \mathbb{E}_\xi F(x, \xi) = \nabla f(x).$$

For example, we can consider the setting where $f(x)$ is the sum of a set of m functions, and where ξ is distributed uniformly on this set such that

$$f(x) = \frac{1}{m} \sum_{i=1}^m f_i(x) \;\Rightarrow\; F(x, \xi) = f_\xi(x),\; \xi \sim \text{Unif}[m].$$

Clearly, in expectation, first-order oracle access to $F(x, \xi)$ results in an unbiased estimate of the true gradient. Under such a framework, we can devise an alternative scheme to the gradient descent algorithm, where instead of utilizing the true gradient at each state, we use the stochastic estimator to update our iterates. Precisely, with an initial point $x_0 \in \mathbb{R}^n$, we iterate

$$x_{k+1} = x_k - h_k G_k, \quad \text{where } G_k := g(x_k, \xi_k).$$

As in previous lectures discussing different algorithms, we clarify here that this algorithm is not formally a descent method, in this case, due to its stochastic nature—in the literature and optimization community, however, it is referred to as such. Unlike the classical gradient descent algorithm, it is very difficult to accelerate this method in the primal view with techniques like estimate sequences because it is not a descent method. However, in the dual view, there always exists a descent method which could much more easily be accelerated. Nonetheless, we will find that SGD will prove to be quite a useful algorithm in settings like machine learning where full gradient computations are prohibitively expensive.

1.1 SGD for convex, Lipschitz f

We will first consider the most general convex setting and analyze the performance of SGD on convex, Lipschitz functions f. The Lipschitz property here bounds the gradient estimator such that

$$\sup_x \mathbb{E}_\xi |g(x, \xi)|^2 \le M^2. \tag{1}$$

Under these assumptions, we can derive a $1/\sqrt{t}$ convergence rate with iteration steps, just like in the classical gradient setting. However, with the stochastic estimator, we can obtain a much smaller overall computational burden.

Theorem 1. *Let* $\bar{x} = \frac{1}{T+1}\sum_{k=0}^{T} x_k$ *and* $h_k = h > 0$*. Then, given an initial point* x_0 *and applying the SGD algorithm to* f *for* T *steps, we obtain a bound on the expected suboptimality*

$$\mathbb{E}[f(\bar{x})] - f^* \leq \frac{\|x_0 - x_*\|^2}{2h(T+1)} + \frac{hM^2}{2}.$$

Further, if we choose $h = \frac{|x_0 - x_*|}{M\sqrt{T+1}}$*, then*

$$\mathbb{E}[f(\bar{x})] - f^* \leq \frac{\|x_0 - x_*\| M}{\sqrt{T+1}}.$$

Proof

$$\begin{aligned}
\mathbb{E}\|x_{k+1} - x_*\|^2 &= \mathbb{E}\|x_k - h_k G_k - x_*\|^2 \\
&= \mathbb{E}\|x_k - x_*\|^2 - 2h_k \mathbb{E}[\langle G_k, x_k - x_* \rangle] + h_k^2 \mathbb{E}\|G_k\|^2 \\
&\leq \mathbb{E}\|x_k - x_*\|^2 - 2h_k \mathbb{E}[\langle G_k, x_k - x_* \rangle] + h_k^2 M^2 \quad (\because (1)) \\
&= \mathbb{E}\|x_k - x_*\|^2 - 2h_k \mathbb{E}[\langle \nabla f(x_k), x_k - x_* \rangle] + h_k^2 M^2 \quad (\because \mathbb{E}[G_k | x_0, \xi_0, \ldots, \xi_{k-1}] = \nabla f(x_k)) \\
&\leq \mathbb{E}\|x_k - x_*\|^2 - 2h_k \mathbb{E}[f(x_k) - f^*] + h_k^2 M^2 \quad (\because f: \text{ convex})
\end{aligned}$$

Therefore,

$$2h_k(\mathbb{E}[f(x_k)] - f^*) \leq \mathbb{E}\|x_k - x_*\|^2 - \mathbb{E}\|x_{k+1} - x_*\| + h_k^2 M^2. \tag{2}$$

Summing (2) for $k = 0, \ldots, T$,

$$\sum_{k=0}^{T} 2h_k(\mathbb{E}[f(x_k)] - f^*) \leq \mathbb{E}\|x_0 - x_*\|^2 - \mathbb{E}\|x_{T+1} - x_*\| + \sum_{k=0}^{T} h_k^2 M^2 \leq \|x_0 - x_*\|^2 + \sum_{k=0}^{T} h_k^2 M^2.$$

Dividing both sides by $\sum_{k=0}^{T} 2h_k$,

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

As $h_k = h$, $\bar{x} = \frac{1}{T+1}\sum_{k=0}^{T} x_k = \frac{\sum_{k=0}^{T} h_k x_k}{\sum_{k=0}^{T} h_k}$. Then,

$$\mathbb{E}[f(\bar{x})] - f^* \leq \frac{\|x_0 - x_*\|^2 + \sum_{k=0}^{T} h_k^2 M^2}{\sum_{k=0}^{T} 2h_k} = \frac{\|x_0 - x_*\|^2}{2h(T+1)} + \frac{hM^2}{2}.$$

□

Note that to guarantee

$$\mathbb{E}[f(\bar{x})] - f^* \leq \epsilon,$$

we have $T = \Omega(\frac{1}{\epsilon^2})$. We note that even though it is a stochastic setting, the iteration number doesn't depend on m, the sample number.

1.2 Strongly convex, Lipschitz f

We can improve the performance with SGD further if we know that f is μ strongly convex. As a reminder, a function f is μ strongly convex if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2.$$

We will retain our previous Lipschitz assumption (1) in our analysis.

Theorem 2. *Suppose* $h_k = \frac{1}{\mu(k+1)}$. *Then,*

$$\begin{aligned} &(1) \quad \frac{1}{T+1}\sum_{k=0}^{T} \mathbb{E}[f(x_k) - f^*] \leq \frac{M^2}{2\mu(T+1)}(1 + \log(T+1)) \\ &(2) \quad \mathbb{E}\|x_k - x_*\|^2 \leq \frac{Q}{k+1}, \ \textit{where } Q = \max\left(\frac{M^2}{\mu^2}, \|x_0 - x_*\|^2\right) \end{aligned}$$

Proof (1) From the proof of Theorem 1, we have

$$\mathbb{E}\|x_{k+1} - x_*\|^2 \leq \mathbb{E}\|x_k - x_*\|^2 - 2h_k\mathbb{E}[\langle G_k, x_k - x_*\rangle] + h_k^2 M^2. \tag{3}$$

$$\begin{aligned} \mathbb{E}[f(x_k) - f^*] &\leq \mathbb{E}[\langle \nabla f(x_k), x_k - x_*\rangle] - \frac{\mu}{2}\cdot\mathbb{E}\|x_k - x_*\|^2 \quad (\because \text{ strongly convex}) \\ &\leq \frac{\mathbb{E}\|x_k - x_*\|^2 - \mathbb{E}\|x_{k+1} - x_*\|^2}{2h_k} + \frac{h_k M^2}{2} - \frac{\mu}{2}\cdot\mathbb{E}\|x_k - x_*\|^2 \quad (\because (3)) \\ &= \frac{\mu k}{2}\cdot\mathbb{E}\|x_k - x_*\|^* - \frac{\mu(k+1)}{2}\cdot\mathbb{E}\|x_{k+1} - x_*\|^2 + \frac{h_k M^2}{2}, \end{aligned}$$

where the last equation is obtained by inserting $h_k = \frac{1}{\mu(k+1)}$. Summing up both sides for $k = 0, \ldots, T$,

$$\sum_{k=0}^{T}(\mathbb{E}[f(x_k)] - f^*) \leq \sum_{k=0}^{T}\frac{h_k}{2}\cdot M^2 = \frac{M^2}{2\mu}\sum_{k=0}^{T}\frac{1}{k+1} \leq \frac{M^2}{2\mu}(1 + \log(T+1)).$$

(2) As f is μ-strongly convex, for $x, y \in \mathbb{R}^n$,

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2.$$

Using the above equation with (3), we have

$$\mathbb{E}\|x_{k+1} - x_*\|^2 \leq (1 - 2h_k\mu)\mathbb{E}\|x_k - x_*\|^2 + h_k^2 M^2.$$

Then, we finish the proof by induction. It is trivial to check that for $k = 0$,

$$\mathbb{E}\|x_0 - x_*\|^2 = \|x_0 - x_*\|^2 \leq \frac{Q}{0+1}.$$

Suppose that (2) is true for some k. Then,

$$\begin{aligned} \mathbb{E}\|x_{k+1} - x_*\|^2 &\leq (1 - 2h_k\mu)\mathbb{E}\|x_k - x_*\|^2 + h_k^2 M^2 \\ &\leq (1 - 2h_k\mu)\cdot\frac{Q}{k+1} + \frac{M^2}{\mu^2(k+1)^2} \\ &\leq (1 - \frac{2}{k+1})\cdot\frac{Q}{k+1} + \frac{Q}{(k+1)^2} \\ &\leq Q(\frac{1}{k+1}\cdot\frac{k}{k+1}) \\ &\leq \frac{Q}{k+2} \quad (\because (k+1)^2 \geq k(k+2)) \end{aligned}$$

□

We emphasize here the delicacy of this proof. In particular, this proof is very sensitive to the precise knowledge of the strong-convexity parameter μ, as well as to the algebraic operations in the final steps of the proof by induction. Without careful manipulation, this result does not hold and the proof cannot be performed in this manner.

2 Convergence Results of SGD for Non-Smooth Optimization [1]

In class, we only considered the convergence results of $\bar{x}$. In this paper, they show convergence guarantee for the individual iterate x_t for both strongly convex and general convex setting.

Theorem 3. *Suppose f is μ-strongly convex, and that $\mathbb{E}\|G_k\| \leq M^2$ for all k. Consider SGD with step sizes $h_k = 1/\mu t$. Then for any $T > 1$, it holds that*

$$\mathbb{E}\left[f(x_T) - f^*\right] \leq \frac{17M^2(1+\log T)}{\mu T}.$$

Theorem 4. *Suppose f is convex, and that for some constants D, M, it holds that $\mathbb{E}\|G_k\| \leq M^2$ for all t, and $\sup_{x,x' \in \mathcal{X}} \|x - x'\| \leq D$, where $\mathcal{X}$ is the feasible region of x. Consider SGD with step sizes $h_k = c/\sqrt{k}$ for some constant $c > 0$. Then for any $T > 1$, it holds that*

$$\mathbb{E}\left[f(x_T) - f^*\right] \leq \left(\frac{D^2}{c} + cM^2\right)\frac{2+\log T}{\sqrt{T}}.$$

The expected error bounds $\mathcal{O}(\log T/T)$ for strongly convex case and $\mathcal{O}(\log T/\sqrt{T})$ for convex case given above are close to minimax optimal rates, $\mathcal{O}(1/T)$ and $\mathcal{O}(1/\sqrt{T})$, respectively.

They also improve the expected error bound of the α-*suffix averaging scheme*[2]. The α-suffix averaging considers the average of the last αT iterates for some $\alpha \in (0,1)$ s.t. αT is an integer.

$$\bar{x}_T^{\alpha} = \frac{1}{\alpha T}\sum_{t=(1-\alpha)T+1}^{T} x_t.$$

In [2], the optimal error bound was given as $\mathcal{O}\left(\left(1+\log\left(\frac{1}{1-\alpha}\right)\right)/\alpha T\right)$, which is optimal in terms of T but increases rapidly as α gets small. In this paper, they show that for the μ-strongly convex f, tighter upper bound can be generated.

Theorem 5. *Under the same conditions of Theorem 3, and assuming αT is an integer, it holds that*

$$\mathbb{E}\left[f(\bar{x}_T^{\alpha}) - f^*\right] \leq \frac{17M^2\left(1+\log\left(\frac{1}{\min\{\alpha,(1+1/T)-\alpha\}}\right)\right)}{\mu T}.$$

For the general convex case, upper bound of $\mathcal{O}(\log(1/\alpha)/\sqrt{T})$ can be generated. A limitation of suffix averaging is that unless we can store all iterates, it requires us to guess the stopping time T in advance.

Thus, they propose a new and simple scheme, *polynomial-decay averaging*, which can be computed on-the-fly and gives an optimal rate. For a small constant $\eta \geq 0$, $\bar{x}_1^{\eta} = x_1$ and for any $k > 1$,

$$\bar{x}_k^{\eta} = \left(1 - \frac{\eta+1}{k+\eta}\right)\bar{x}_{k-1}^{\eta} + \frac{\eta+1}{k+\eta}x_k.$$

Theorem 6. *Suppose f is μ-strongly convex, and that $\mathbb{E}\|G_k\| \leq M^2$ for all k. Consider SGD initialized with x_1 and step sizes $h_k = 1/\mu k$. Also, let $\eta \geq 1$ be an integer. Then,*

$$\mathbb{E}\left[f(\bar{x}_T^\eta) - f^*\right] \leq 58\left(1 + \frac{\eta}{T}\right)\left(\eta(\eta+1) + \frac{(\eta+0.5)^3(1+\log T)}{T}\right)\frac{M^2}{\mu T}.$$

Note that η doesn't have to be an integer, and for a constant η, the bound is optimal. For the general convex f, the bound of $\mathcal{O}\left(\frac{\eta(D^2/c+cM^2)}{\sqrt{T}}\right)$ can be obtained.

3 Last-iterate and uniform convergence analysis of SGD [3]

An alternative approach to analyzing the performance of SGD on strongly-convex functions would be to perform a dynamic analyis of the evolution of the random variables x due to the gradient operator G and the underlying noise variables $\{\xi\}$. In short, by linearizing the recursion of random variables, bounding the moments of the noise variables and performing a concentration inequality on a stopped process, and transferring the concentration inequality from the stopped stochastic process to the original stochastic evolution so that the noise of $\{\xi\}$ can be eliminated in analysis, we can obtain probabilistic bounds on the evolution dynamics of SGD. Specifically, with this more robust probabilistic framework, we can obtain more informative bounds for both last-iterate convergence and the uniform convergence of all iterates.

Theorem 7. *Let f be a μ-strongly convex function and perform SGD on f for T steps. Then, for a bounded, unbiased gradient estimator with parameter M and step-size $h_k = \frac{1}{\mu k}$, we can guarantee last-iterate convergence for any $\delta > 0$ with probability*

$$Pr\left[\|x_T - x^*\| > \frac{1000M^2\log(1/\delta)}{\mu^2 T}\right] < \delta$$

for every $T \in \mathbb{N}$.
Further, we can guarantee strong uniform convergence of all the iterates with probability

$$Pr\left[\|x_t - x^*\| > \frac{1000M^2\left(\log(1/\delta) + \log\log(t+1)\right)}{\mu^2 t}\right] < \delta$$

for all $t \in \mathbb{N}$.

These bounds are thought to be relatively tight, but may potentially may be refined further with a multi-recursion analytical approach.

4 Analysis of Stochastic Subgradient Descent [4]

We can instead consider a stochastic implementation of subgradient descent. For a function $f(x)$ with domain $x \in G$, let $\psi(x,\xi)$ be an estimate of the subgradient of f at x distributed on ξ. For clarity, we present ξ as iteration independent, but formally, ξ_k are i.i.d. across iterates. Similar to the bound on the estimate of the gradient, we bound f in a quasi-Lipschitz fashion as

$$\sup_x \mathbb{E}_\xi \left|\psi(x,\xi)\right|^2 \leq L^2.$$

Our iterates in subgradient descent then progress as

$$x_{k+1} = \pi_G\left(x_k - \gamma_k \psi(x_k, \xi)\right)$$

where $\gamma_k \geq 0$ and π_G is the standard projector operator.

Finally, we define the k-th approximate solution as a sliding window average such that

$$x^k \equiv \left[\sum_{k/2 \le t \le k} \gamma_t \right]^{-1} \sum_{k/2 \le t \le k} \gamma_t x_t.$$

Theorem 8. *For the subgradient method with bounded subgradient estimates, and for any positive integer N, we can obtain a bound on the N-th approximate solution iterate*

$$\mathbb{E}\left|f(x^N) - f^*\right| \le \frac{D^2 + L^2 \sum_{N/2 \le k \le N} \gamma_k^2}{2 \sum_{N/2 \le k \le N} \gamma_k}$$

where D is the diameter of the domain G. Further, if we choose $\gamma_k = \frac{D}{L\sqrt{k}}$,

$$\mathbb{E}\left|f(x^N) - f^*\right| \le \mathcal{O}(1) \frac{LD}{\sqrt{N}}$$

for an appropriate constant $\mathcal{O}(1)$.

In fact, we can show that such $\frac{1}{\sqrt{N}}$ bounds are optimal for first-order stochastic optimization methods on for convex functions f.

Proposition 1. *For every stochastic oracle-based method M with first-order estimates bounded by L^2 and for every $L > 0$ and $D > 0$, there exists a function f such that*

$$\mathbb{E}\left|f(x^N) - f^*\right| \ge \mathcal{O}(1) \frac{LD}{\sqrt{N}}$$

where x^N is the result of the M as applied to f for N iterations.

Only for functions f which are strongly convex and have non-degenerate optima, i.e, their minima are attained at unique points x^*, can we obtain stronger bounds. Let f be strongly convex with a unique optimum x^* such that

$$f(x^*) + \frac{\theta}{2} \|x - x^*\|^2 \le f(x) \le f(x^*) + \frac{\Theta}{2} \|x - x^*\|^2$$

for positive θ, Θ.

Proposition 2. *Apply subgradient descent to the strongly convex f with stepsizes $\gamma_k = \frac{\gamma}{k}$ where $\gamma\theta \ge 1$. Then,*

$$\mathbb{E}\left|f(x^N) - f^*\right| \le c(\gamma\theta)\Theta \frac{D^2 + \gamma^2 L^2}{N}$$

where L is the bound on the first-order estimates, D is the diameter of the domain, and $c(\gamma\theta)$ is a problem independent function on $(1, \infty)$.

However, this bound is not robust to perturbations of γ, similar to the delicate bound obtained when applying SGD to strictly convex functions; it also heavily relies on the non-degeneracy of the optimal solution. For certain problems, inaccuracy in choosing the right γ can result in orders of magnitude more iterations for desired levels of ϵ accuracy in the iterates.

References

[1] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," in *International conference on machine learning.* PMLR, 2013, pp. 71–79.

[2] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," *arXiv preprint arXiv:1109.5647*, 2011.

[3] C.-N. Chou, J. S. Sandhu, M. B. Wang, and T. Yu, "A general framework for analyzing stochastic dynamics in learning algorithms," *arXiv preprint arXiv:2006.06171*, 2020.

[4] A. Nemirovski, "Information-based complexity of convex programming," *Lecture Notes*, vol. 834, 1995.