Revisiting the Last-Iterate Convergence of Stochastic Gradient Methods

Zijian Liu Zhengyuan Zhou

Stern School of Business, New York University

Basic Setup

$$\min_{x\in\mathcal{X}} F(x) = f(x) + h(x) \qquad \text{(OPT)}$$

- $f:\mathcal{X}\to\mathbb{R}$ and $h:\mathcal{X}\to\mathbb{R}$ are both convex.
- $\mathcal{X}\subseteq\mathbb{R}^d$ is nonempty closed convex.
- Given $x\in\mathcal{X}$, we can only access a stochastic gradient $\widehat{g}$ such that $\mathbb{E}\left[\widehat{g}\mid x\right]\in\partial f(x)$.

Proximal Stochastic Gradient Descent

Algorithm 1 Proximal Stochastic Gradient Descent

1: **Input:** initial point $x^1\in\mathcal{X}$, step size η_t.
2: **for** $t=1$ **to** T:
3: $x^{t+1} = \mathrm{argmin}_{x\in\mathcal{X}} h(x) + \|x-(x^t-\eta_t\widehat{g}^t)\|_2^2/(2\eta_t)$

The proximal version of stochastic gradient descent (SGD) is a popular method to solve (OPT).

- The convergence of the average iterate $x_{\text{avg}}^{T+1} = \sum_{t=1}^{T} x^{t+1}/T$ has been well-studied in different settings (e.g., Lipschitz/smooth f), see, for example, [1].
- However, in practice, people always use the last iterate as the output. Naturally, we want to know whether $F(x^{T+1})-F(x^*)$ converges? If it converges, how fast is it?

Related Work

All the previous works for the last iterate only consider $h=0$.

- f is Lipschitz under the 2-norm: [2-3] proved the high-probability rate $O\left(\sqrt{\log\frac{1}{\delta}/T}\right)$ on bounded domains. [4] showed the $O(1/\sqrt{T})$ expected rate for general domains.
- f is smooth under the 2-norm: The only result is [5], who established the $O(1/T^{1/3})$ rate in expectation.

Three Questions

There are three questions we want to ask:

- *Q1:Is it possible to prove the high-probability last-iterate convergence for Lipschitz convex functions without assuming compact domains?*
- *Q2:Does the last iterate of SGD provably converge in the rate of $O(1/\sqrt{T})$ for smooth and convex functions on a general domain?*
- *Q3:Is there a unified way to analyze the last-iterate convergence of stochastic gradient methods both in expectation and in high probability to accommodate general domains, composite objectives, non-Euclidean norms, Lipschitz conditions, smoothness, and (strong) convexity at once?*

In our work, we answer these three questions affirmatively.

Composite Stochastic Mirror Descent

Algorithm 2 Composite Stochastic Mirror Descent

1: **Input:** initial point $x^1\in\mathcal{X}$, step size η_t.
2: **for** $t=1$ **to** T:
3: $x^{t+1} = \mathrm{argmin}_{x\in\mathcal{X}} h(x) + \langle\widehat{g}^t, x-x^t\rangle + D_\psi(x,x^t)/\eta_t$

To accommodate a general norm $\|\cdot\|$, we consider the Composite Stochastic Mirror Descent (CSMD) algorithm, where $D_\psi(x,y)=\psi(x)-\psi(y)-\langle\nabla\psi(y),x-y\rangle$ and ψ is 1-strongly convex with respect to the norm $\|\cdot\|$ (i.e., $D_\psi(x,y)\ge\|x-y\|^2/2$).

Remark: When $\|\cdot\|=\|\cdot\|_2$, taking $\psi(x)=\|x\|^2/2$ to recover Proximal SGD.

The Central Assumption

(L,M)-smoothness assumption: $f(x)-f(y)-\langle g,x-y\rangle \le \frac{L\|x-y\|^2}{2} + M\|x-y\|^2, \forall x,y\in\mathcal{X}, g\in\partial f(y)$.

Remark: This function class contains all Lipschitz and smooth functions. It also includes Hölder smooth functions.

Remark: We do not require any compactness on $\mathcal{X}$.

New Last-iterate Results

High-Probability Convergence: Under sub-Gaussian noises (i.e., $\mathbb{E}\left[\exp\left(\|\widehat{g}-\mathbb{E}\left[\widehat{g}\mid x\right]\|_*^2/\sigma^2\right)\mid x\right]\le e$), for any $\delta\in(0,1)$, for properly picked η_t, with probability at least $1-\delta$, CSMD guarantees

$$F(x^{T+1})-F(x^*)\le\widetilde{O}\left(\frac{LD_\psi(x^1,x^*)}{T}+\frac{\left(M+\sigma\sqrt{\log\frac{1}{\delta}}\right)\sqrt{D_\psi(x^1,x^*)}}{\sqrt{T}}\right).$$

In-Expectation Convergence: Under the finite variance assumption (i.e., $\mathbb{E}\left[\|\widehat{g}-\mathbb{E}[\widehat{g}\mid x]\|_*^2\mid x\right]\le\sigma^2$), for properly picked η_t, CSMD guarantees

$$\mathbb{E}[F(x^{T+1})-F(x^*)]\le\widetilde{O}\left(\frac{LD_\psi(x^1,x^*)}{T}+\frac{(M+\sigma)\sqrt{D_\psi(x^1,x^*)}}{\sqrt{T}}\right).$$

For the strongly convex case, we refer the interested reader to our paper.

Proof Strategies and Extensions

- In the proof, we use a new auxiliary sequence z_t. Instead of bounding $F(x^{t+1})-F(x^*)$ in every step, we control $F(x^{t+1})-F(z^t)$ to finally obtain the rate for the last iterate.
- Our proof is unified and works for various assumptions at once.
- The proof technique provably extends to heavy-tailed noises, sub-Weibull noises, etc. We refer the interested reader to our paper for details.

References

[1] Guanghui Lan. *First-order and stochastic optimization methods for machine learning*. Springer, 2020.

[2] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for nonsmooth stochastic gradient descent. In *Conference on Learning Theory*, pp. 1579–1613. PMLR, 2019.

[3] Prateek Jain, Dheeraj M. Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. *SIAM Journal on Optimization*, 31(2):1108–1130, 2021.

[4] Francesco Orabona. Last iterate of sgd converges (even in unbounded domains). 2020.

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