CHAPTER 8-1. MOMENTUM STOCHASTIC GRADIENT DESCENT

Yutong He   Kun Yuan

November 14, 2023

1 Problem formulation

This chapter considers the following stochastic optimization problem

$$\min_{x \in \mathbb{R}^d} \quad f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[F(x;\xi)] \tag{1}$$

where $\xi \sim \mathcal{D}$ denotes the random data sample and $\mathcal{D}$ denotes the data distribution. Since $\mathcal{D}$ is typically unknown in machine learning, the closed-form of $f(x)$ is also unknown.

Notation. We introduce the following notations:

- Let $x^\star := \arg\min_{x\in\mathbb{R}^d}\{f(x)\}$ be the optimal solution to problem (1).
- Let $f^\star := \min_{x\in\mathbb{R}^d}\{f(x)\}$ be the optimal function value.
- Let $\mathcal{F}_k = \{x_{k-1}, y_{k-1}, v_{k-1}, \xi_{k-1}, \cdots, x_0, y_0, v_0, \xi_0\}$ be the filtration containing all historical variables at and before iteration k. Note that ξ_k does not belong to $\mathcal{F}_k$.

2 Momentum stochastic gradient descent

In this section, we introduce the following stochastic gradient descent scheme with Nesterov's momentum:

$$y_k = (1-\theta_k)x_{k-1} + \theta_k v_{k-1}, \tag{2}$$
$$x_k = y_{k-1} - \gamma \nabla F(y_{k-1};\xi_k), \tag{3}$$
$$v_k = \theta_k^{-1} x_k + (1-\theta_k^{-1})x_{k-1}, \tag{4}$$

where γ is the learning rate, $\xi_k \sim \mathcal{D}$ is a random data sampled at iteration k, and θ_k is a parameter that controls the momentum acceleration. The initial state can be chosen at $y_0 = v_0 = x_0 \in \mathbb{R}^d$.

3 Convergence analysis

We use the same assumption on stochastic gradient oracle as in chap 6:

Assumption 3.1. Given the filtration $\mathcal{F}_k$, we assume

$$\begin{aligned} \mathbb{E}[\nabla F(y_{k-1};\xi_k) \mid \mathcal{F}_k] =& \nabla f(y_{k-1}), & (5)\\ \mathbb{E}[\|\nabla F(y_{k-1};\xi_k) - \nabla f(y_{k-1})\|_2^2 \mid \mathcal{F}_k] \le& \sigma^2. & (6)\end{aligned}$$

We have the following convergence result in the generally-convex and smooth scenario.

Theorem 3.2. Suppose $f(x)$ is $L-$smooth and convex, and Assumption 3.1 holds. If we choose $\theta_k = \frac{2}{k+1}$, and $0 < \gamma \le 1/L$, momentum SGD converges at the following rate:

$$\mathbb{E}[f(x_K) - f^\star] \le \frac{2L\Delta_0}{\gamma_0(K+1)^2} + \frac{(K+2)(2K+3)\gamma_0\sigma^2}{6L(K+1)}, \tag{7}$$

where $\Delta_0 := \|x_0 - x^\star\|_2^2$. If we further let

$$\gamma = \frac{1}{L\left(1 + \sqrt{\frac{(K+1)(K+2)(2K+3)\sigma^2}{12L^2\Delta_0}}\right)},$$

momentum SGD converges as

$$\mathbb{E}[f(x_K) - f^\star] \le \frac{2L\Delta_0}{(K+1)^2} + \sqrt{\frac{5\Delta_0\sigma^2}{K+1}}. \tag{8}$$

Proof. By iterations (2)(3)(4) and the unbiasedness of stochastic gradient oracle (5), we have

$$\begin{aligned} &\mathbb{E}[\|v_k - x^\star\|_2^2 \mid \mathcal{F}_k] \\ =&\mathbb{E}[\|\theta_k^{-1}x_k + (1-\theta_k^{-1})x_{k-1} - x^\star\|_2^2 \mid \mathcal{F}_k] \\ =&\mathbb{E}[\|\theta_k^{-1}y_{k-1} - \gamma\theta_k^{-1}\nabla F(y_{k-1};\xi_k) + (1-\theta_k^{-1})x_{k-1} - x^\star\|_2^2 \mid \mathcal{F}_k] \\ =&\mathbb{E}[\|v_{k-1} - x^\star - \gamma\theta_k^{-1}\nabla F(y_{k-1};\xi_k)\|_2^2 \mid \mathcal{F}_k] \\ =&\|v_{k-1} - x^\star\|_2^2 - 2\gamma\theta_k^{-1}\langle \nabla f(y_{k-1}), v_{k-1} - x^\star\rangle + \gamma^2\theta_k^{-2}\mathbb{E}[\|\nabla F(y_{k-1};\xi_k)\|_2^2 \mid \mathcal{F}_k], \end{aligned} \tag{9}$$

Applying (6) to (9), we obtain

$$\mathbb{E}[\|v_k - x^\star\|_2^2 \mid \mathcal{F}_k] \le \|v_{k-1} - x^\star\|_2^2 - \frac{2\gamma}{\theta_k}\langle \nabla f(y_{k-1}), v_{k-1} - x^\star\rangle + \frac{\gamma^2}{\theta_k^2}\|\nabla f(y_{k-1})\|_2^2 + \frac{\gamma^2\sigma^2}{\theta_k^2}. \tag{10}$$

By taking the expectations over the filtration $\mathcal{F}_k$, (10) becomes

$$\begin{aligned}\mathbb{E}[\langle \nabla f(y_{k-1}), v_{k-1} - x^\star \rangle] \le & -\frac{\theta_k}{2\gamma}\mathbb{E}[\|v_k - x^\star\|_2^2] + \frac{\theta_k}{2\gamma}\mathbb{E}[\|v_{k-1} - x^\star\|_2^2] + \frac{\gamma}{2\theta_k}\mathbb{E}[\|\nabla f(y_{k-1})\|_2^2] \\ & + \frac{\gamma\sigma^2}{2\theta_k}. \end{aligned} \tag{11}$$

By convexity and L-smoothness, we have

$$\begin{aligned} & \mathbb{E}[f(x_k) \mid \mathcal{F}_k] \\ \le & \mathbb{E}\left[f(y_{k-1}) + \langle \nabla f(y_{k-1}), x_k - y_{k-1} \rangle + \frac{L}{2}\|x_k - y_{k-1}\|_2^2 \;\middle|\; \mathcal{F}_k \right] \\ = & f(y_{k-1}) - \gamma \|\nabla f(y_{k-1})\|_2^2 + \frac{L\gamma^2}{2}\mathbb{E}[\|\nabla F(y_{k-1}; \xi_k)\|_2^2 \mid \mathcal{F}_k]. \end{aligned} \tag{12}$$

Applying (6) and $\gamma \le 1/L$ to (12), we obtain

$$\mathbb{E}[f(x_k) \mid \mathcal{F}_k] \le f(y_{k-1}) - \frac{\gamma}{2}\|\nabla f(y_{k-1})\|_2^2 + \frac{L\gamma^2\sigma^2}{2}. \tag{13}$$

By convexity, we have

$$f(y_{k-1}) \le f(x_{k-1}) - \langle \nabla f(y_{k-1}), x_{k-1} - y_{k-1} \rangle, \tag{14}$$

$$f(y_{k-1}) \le f^\star - \langle \nabla f(y_{k-1}), x^\star - y_{k-1} \rangle. \tag{15}$$

Consider (13)+$(1-\theta_k)\times$(14)+$\theta_k \times$(15), we obtain

$$\begin{aligned} \mathbb{E}[f(x_k) - f^\star \mid \mathcal{F}_k] \le & (1-\theta_k)[f(x_{k-1}) - f^\star] - \langle \nabla f(y_{k-1}), (1-\theta_k)x_{k-1} + \theta_k x^\star - y_{k-1} \rangle \\ & - \frac{\gamma}{2}\|\nabla f(y_{k-1})\|_2^2 + \frac{L\gamma^2\sigma^2}{2} \\ = & (1-\theta_k)[f(x_{k-1}) - f^\star] + \theta_k \langle \nabla f(y_{k-1}), v_{k-1} - x^\star \rangle - \frac{\gamma}{2}\|\nabla f(y_{k-1})\|_2^2 \\ & + \frac{L\gamma^2\sigma^2}{2}. \end{aligned} \tag{16}$$

By taking the expectations over the filtration $\mathcal{F}_k$ and applying (11), we obtain

$$\begin{aligned} \mathbb{E}[f(x_k) - f^\star] \le & (1-\theta_k)\mathbb{E}[f(x_{k-1}) - f^\star] - \frac{\theta_k^2}{2\gamma}\mathbb{E}[\|v_k - x^\star\|_2^2] + \frac{\theta_k^2}{2\gamma}\mathbb{E}[\|v_{k-1} - x^\star\|_2^2] \\ & + \frac{\gamma(1+L\gamma)\sigma^2}{2}. \end{aligned} \tag{17}$$

Apply the choice of θ_k and let $\gamma = \gamma_0/L$ (where $0 < \gamma_0 \le 1$) to (17), we obtain

$$\begin{aligned} \frac{\gamma_0(k+1)^2}{2L}\mathbb{E}[f(x_k) - f^\star] + \mathbb{E}[\|v_k - x^\star\|_2^2] \le & \frac{\gamma_0(k^2-1)}{2L}\mathbb{E}[f(x_{k-1}) - f^\star] + \mathbb{E}[\|v_{k-1} - x^\star\|_2^2] \\ & + \frac{(k+1)^2\gamma_0^2\sigma^2}{2L^2}. \end{aligned} \tag{18}$$

Summing up (18) from $k = 1$ to K, we obtain

$$\begin{aligned}
&\frac{\gamma_0 (K+1)^2}{2L}\mathbb{E}[f(x_K) - f^\star] \le \|x_0 - x^\star\|_2^2 + \frac{(K+1)(K+2)(2K+3)\gamma_0^2\sigma^2}{12L^2}, \\
\Rightarrow &\mathbb{E}[f(x_K) - f^\star] \le \frac{2L\|x_0 - x^\star\|_2^2}{\gamma_0(K+1)^2} + \frac{(K+2)(2K+3)\gamma_0\sigma^2}{6L(K+1)},
\end{aligned}$$

which is equivalent to (7). If we further choose

$$\gamma_0 = \left(\sqrt{\frac{(K+1)(K+2)(2K+3)\sigma^2}{12L^2\Delta_0}} + 1\right)^{-1},$$

the convergence rate of momentum SGD is given by

$$\mathbb{E}[f(x_K) - f^\star] \le \frac{2L\Delta_0}{(K+1)^2} + 2\sqrt{\frac{(K+2)(2K+3)\Delta_0\sigma^2}{3(K+1)^3}} \le \frac{2L\Delta_0}{(K+1)^2} + \sqrt{\frac{5\Delta_0\sigma^2}{K+1}}, \quad (19)$$

which is equivalent to (8). □

4 Convergence lower bound

In this section, we list the lower bound results under the generally-convex scenario.

Assumption 4.1. We consider algorithm class $\mathcal{A}$ of zero-respecting algorithms that, initialized with zero points and, for any $t \ge 1$ and $1 \le k \le d$, the following conditions are met:

1. If the algorithm queries the gradient oracle O at y_t with $[y_t]_k \ne 0$, then there exists some $1 \le s < t$ such that $[O(y_s; \xi_s)]_k \ne 0$.
2. If the output model x_t at time t satisfies $[x_t]_k \ne 0$, then there exists some $1 \le s \le t$ such that $[O(y_s; \xi_s)]_k \ne 0$.

Here, $[x]_k$ denotes the k-th entry of vector x, and $O(x; \xi)$ denotes the output of gradient oracle O given query point x and randomness ξ.

Remark. The zero-respecting properties can be generalized to another concept called *linear spanning*, which requires *1) query points* $y_t \in \operatorname{span}\{y_0, \hat{\nabla} f(y_0), \cdots, \hat{\nabla} f(y_{t-1})\}$ and *2) the output model* $x_t \in \operatorname{span}\{y_0, \hat{\nabla} f(y_0), \cdots, \hat{\nabla} f(y_{t-1})\}$, where $\hat{\nabla} f$ is the gradient oracle called to approximate ∇f. It's worth noting that the momentum SGD algorithm above, as well as most existing first-order stochastic algorithms are linear spanning, and thus zero-respecting.

Based on the zero-respecting property, we provide the lower bound results following [1]:

Proposition 4.2. For any $\Delta_0 > 0$, there exists a constant $c = \Theta(1)$, convex and L-smooth function f satisfying $\|x_0 - x^\star\|_2^2 \le \Delta_0$, stochastic gradient oracles O satisfying Assumption 3.1, such that the output $\hat{x}$ of any $A \in \mathcal{A}$ starting from x_0 requires

$$\Omega\left(\frac{\Delta_0 \sigma^2}{\epsilon} + \left(\frac{L\Delta_0}{\epsilon}\right)^{\frac{1}{2}}\right)$$

iterations to reach $\mathbb{E}[f(\hat{x})] - f^\star \le \epsilon$ for any $0 < \epsilon \le cL\Delta_0$.

5 Optimal convergence rates

Though we only give the convergence rate and lower bound of in the generally-convex case, optimal rate of stochastic first-order methods under strongly-convex or non-convex settings have already been constructed. Here we list the convergence complexity (number of iterations) for reaching an ϵ-optimal solution such that $\mathbb{E}[f(x_k) - f^\star] \le \epsilon$ (in convex cases) or $\mathbb{E}[\|\nabla f(x_k)\|_2^2] \le \epsilon$ (in non-convex cases),

Algorithm	non-convex	generally-convex	strongly-convex
SGD	$\mathcal{O}\left(\frac{L\sigma^2}{\epsilon^2} + \frac{L}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\sigma^2}{\epsilon^2} + \frac{L}{\epsilon}\right)$	$\tilde{\mathcal{O}}\left(\frac{L\sigma^2}{\mu^2\epsilon} \ln\left(\frac{1}{\epsilon}\right) + \frac{L}{\mu} \ln\left(\frac{1}{\epsilon}\right)\right)$
momentum SGD	$\mathcal{O}\left(\frac{L\sigma^2}{\epsilon^2} + \frac{L}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\sigma^2}{\epsilon^2} + \sqrt{\frac{L}{\epsilon}}\right)$	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu\epsilon} + \sqrt{\frac{L}{\mu}} \ln\left(\frac{1}{\epsilon}\right)\right)$
Lower Bound	$\Omega\left(\frac{L\sigma^2}{\epsilon^2} + \frac{L}{\epsilon}\right)$	$\Omega\left(\frac{\sigma^2}{\epsilon^2} + \sqrt{\frac{L}{\epsilon}}\right)$	$\tilde{\Omega}\left(\frac{\sigma^2}{\mu\epsilon} + \sqrt{\frac{L}{\mu}} \ln\left(\frac{1}{\epsilon}\right)\right)$

where $\tilde{\mathcal{O}}$, $\tilde{\Omega}$ hide logarithmic factors independent of ϵ.

References

[1] Y. He, X. Huang, Y. Chen, W. Yin, and K. Yuan, "Lower bounds and accelerated algorithms in distributed stochastic optimization with communication compression," *arXiv preprint arXiv:2305.07612*, 2023.