Momentum: Last Iterate Convergence and Variance Reduction

Yutong Dai

OptML Meeting 23' Spring

03/09/2023

Outline

Outline

Problem

Main Problem

$$\min_{x \in \mathcal{X}} \{ f(x) = \mathbb{E}_{\xi \sim \mathcal{P}} [\ell(x; \xi)] \}. \tag{1}$$

Assumption 1.1

- $\mathcal{X} \subset \mathbb{R}^n$ is a non-empty closed and convex set and $\|x\| \leq D_{\mathcal{X}}$ for all $x \in \mathcal{X}$.
- $\mathcal{P}$ is a probability distribution supported on $\Xi \subset \mathbb{R}^d$.
- For any $x \in \mathcal{X}$, $\mathbb{E}_{\xi \sim \mathcal{P}} [\ell(x; \xi)]$ is well-defined and finite.
- Convexity and bounded gradient:
  - f is convex over $\mathcal{X}$ and there exits a constant $G > 0$ such that $\|\nabla f(x)\| \leq G$ for all $x \in \mathcal{X}$.
  - For every $\xi \in \Xi, \ell(x, \xi)$ is convex over $\mathcal{X}$ and there exits a constant $G > 0$ such that $\|\nabla \ell(x, \xi)\| \leq G$ for all $x \in \mathcal{X}$.

Problem (Cont'd)

Question

Stochastic gradient descent with momentum (`SGD + M`) is very popular, yet its convergence property is not well understood. Only recently, the work[a] proved that stochastic heavy-ball's (`SGD + H`) convergence speed is not faster than that of `SGD` alone (convex case).

[a] Othmane Sebbouh et al. "Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball". In: *Conference on Learning Theory*. PMLR. 2021, pp. 3935–3971.

1. `SGD`: $x_{t+1} = \mathbf{Proj}_{\mathcal{X}}\left(x_t - \alpha_t \nabla \ell(x_t, \xi_t)\right)$.
2. `SGD + H`: $x_{t+1} = \mathbf{Proj}_{\mathcal{X}}\left(x_t - \alpha_t \nabla \ell(x_t, \xi_t) + \beta_t(x_t - x_{t-1})\right)$.
3. `SGD + M`: $x_{t+1} = \mathbf{Proj}_{\mathcal{X}}\left(x_t - \alpha_t m_t + \beta_t(x_t - x_{t-1})\right), m_{t+1} = \beta_t m_t + (1-\beta_t)\nabla \ell(x_{t+1}, \xi_{t+1})$,

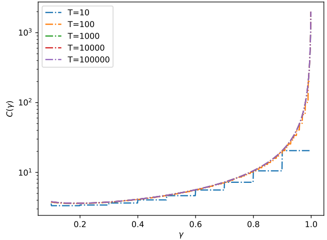
where $\{\alpha_t, \beta_t\}$ are some positive scalars and $\xi_t \sim \mathcal{P}$.

Outline

Algorithm Projected Stochastic Gradient Descent

1: **Initialization**: $x_0 \in \mathcal{X}$, total number of iterations T, and stepsize $\{\alpha_t\}_{t \leq T}$.
2: **for** $t = 1, \dots, T$ **do**
3: Draw $\xi_t \sim \mathcal{P}$ and compute $x_{t+1} = \mathbf{Proj}_{\mathcal{X}}(x_t - \alpha_t \nabla \ell(x_t, \xi_t))$
4: **end for**

- Averaged Iterates[1]: Define the averaged iterates $\bar{x}_{i:j} = \sum_{t=i}^{j} \nu_t^{i:j} x_t$ with $\nu_t^{i:j} = \frac{\alpha_t}{\sum_{\tau=i}^{j} \alpha_\tau}$.
    - When $\alpha_t = \frac{D_{\mathcal{X}}}{G\sqrt{T}}$, $\mathbb{E}[f(\bar{x}_{1:T})] - f(x^*) \leq \frac{GD_{\mathcal{X}}}{\sqrt{T}}$.
    - When $\alpha_t = \frac{D_{\mathcal{X}}}{G\sqrt{t}}$ and $\gamma \in (0,1)$, $\mathbb{E}\left[f(\bar{x}_{\lceil \gamma T \rceil : T})\right] - f(x^*) \leq \left(\frac{2T}{T - \lceil \gamma T \rceil + 1} + \frac{1}{2}\sqrt{\frac{T}{\lceil \gamma T \rceil}}\right) \frac{GD_{\mathcal{X}}}{\sqrt{T}}$.



[1] Arkadi Nemirovski et al. "Robust stochastic approximation approach to stochastic programming". In: *SIAM Journal on optimization* 19.4 (2009), pp. 1574–1609, Equation 2.21 and 2.26.

Algorithm Projected Stochastic Gradient Descent

1: **Initialization**: $x_0 \in \mathcal{X}$, total number of iterations T, and stepsize $\{\alpha_t\}_{t \le T}$.
2: **for** $t = 1, \dots, T$ **do**
3: Draw $\xi_t \sim \mathcal{P}$ and compute $x_{t+1} = \mathbf{Proj}_{\mathcal{X}}(x_t - \alpha_t \nabla \ell(x_t, \xi_t))$
4: **end for**

- Last iterate
  - diminishing stepsize strategy[1]: When $\alpha_t = \frac{D_{\mathcal{X}}}{G\sqrt{t}}$, $\mathbb{E}[f(x_T)] - f(x^*) \le \frac{2GD_{\mathcal{X}}(2+\log T)}{\sqrt{T}}$.
  - stage-wise stepsize strategy[2]: Let $k = \inf\{i : 2^i \ge T\}$. Define $T_i = T - \lceil \frac{T}{2^i} \rceil$ for all $0 \le i \le k$, and $T_{k+1} = T$. Let $C > 0$ be some constant

$$\alpha_t = \frac{C \cdot 2^{-i}}{\sqrt{T}} \text{ for } T_i < t \le T_{i+1} \text{ and } 0 \le i \le k.$$

Under Assumption 1.1, if $C = D_{\mathcal{X}}/G$ and $T > 4$, then

$$\mathbb{E}[f(x_T)] - f(x^*) \le \frac{15GD_{\mathcal{X}}}{\sqrt{T}}.$$

[1] Ohad Shamir and Tong 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, Theorem 2.

[2] Prateek Jain et al. "Making the last iterate of sgd information theoretically optimal". In: *Conference on Learning Theory*. PMLR. 2019, pp. 1752–1755, Theorem 1.

Reformulation: momentum and heavy ball methods are connected to iterates averaging

When $\mathcal{X} = \mathbb{R}^n$ (violates the Assumption 1.1)

SGD + M

Let $c_1 \in (0, 1)$ and $x_0 = z_0$. For all $t \geq 1$, when $c_{t+1} = \beta_t \frac{\alpha_t}{\alpha_{t-1}} \frac{c_t}{1-c_t}$ and $\eta_t = \frac{\alpha_t}{c_{t+1}}(1 - \beta_t)$. The following two forms are equivalent.

$$
\begin{array}{lll}
\text{(original)}^a & m_{t+1} = \beta_t m_t + (1 - \beta_t)\nabla \ell(x_t, \xi_t) & x_{t+1} = x_t - \alpha_t m_{t+1} \\
\text{(iterate averaging)} & z_{t+1} = z_t - \eta_t \nabla \ell(x_t, \xi_t) & x_{t+1} = (1 - c_{t+1})x_t + c_{t+1} z_{t+1}.
\end{array}
$$

[a] The indices are changed compared with previous slides for the presentation's purpose.

The result is adapted from[3] with correction on the edge case.

[3] Aaron Defazio and Robert M Gower. "The Power of Factorial Powers: New Parameter settings for (Stochastic) Optimization". In: *Asian Conference on Machine Learning*. PMLR. 2021, pp. 49–64, Theorem 1.

Reformulation: momentum and heavy ball methods are connected to iterates averaging

When $\mathcal{X} = \mathbb{R}^n$ (violates the Assumption 1.1)

SGD + H

Let $c_t \subset (0,1)$, $\eta_t > 0$ for all $t \geq 0$, $x_0 = z_0$, and $x_1 = x_0 - c_1 \eta_0 \nabla \ell(x_0, \xi_0)$. For all $t \geq 0$, when $\alpha_t = \eta_t c_{t+1}$ and $\beta_t = \frac{c_{t+1}}{c_t} - c_{t+1}$. The following two forms are equivalent.

$$
\begin{array}{lll}
\text{(original)} & x_{t+1} = x_t - \alpha_t \nabla \ell(x_t, \xi_t) + \beta_t (x_t - x_{t-1}) & \\
\text{(iterate averaging)} & z_{t+1} = z_t - \eta_t \nabla \ell(x_t, \xi_t) & x_{t+1} = (1 - c_{t+1}) x_t + c_{t+1} z_{t+1}.
\end{array}
$$

The result is adapted from[3] with correction on the edge case.

[3] Othmane Sebbouh et al. "Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball". In: *Conference on Learning Theory*. PMLR. 2021, pp. 3935–3971, Proposition 1.6.

Last Iterate convergence of SGD + M

Consider the SGD + M in the iterates averaging form with $x_0 = z_0 \in \mathcal{X}$ and

$$z_{t+1} = \mathbf{Proj}_{\mathcal{X}}\left(z_t - \eta_t \nabla \ell(x_t, \xi_t)\right) \text{ and } x_{t+1} = (1 - c_{t+1})x_t + c_{t+1} z_{t+1}.$$

Theorem 1

For all $t \geq 0$, *let* $\eta_t = \frac{\eta}{(t+1/2)^{\overline{1/2}}}$ *with* $\eta = \frac{\sqrt{1/2} D_{\mathcal{X}}}{G}$ *and* $c_{t+1} = \frac{1}{t+1}$. *Then* $\mathbb{E}\left[f(x_t)\right] - f^* \leq \frac{\sqrt{2} D_{\mathcal{X}} G}{(t+3/2)^{\overline{1/2}}}$.[a]

[a] Aaron Defazio and Robert M Gower. "The Power of Factorial Powers: New Parameter settings for (Stochastic) Optimization". In: *Asian Conference on Machine Learning*. PMLR. 2021, pp. 49–64, Theorem 2.

Remark

- Factorial power: For $k + r > 0$ and $k > 0$, define $k^{\bar{r}} = \frac{\Gamma(k+r)}{\Gamma(k)}$.
  - $k^{\bar{r}} = \prod_{i=1}^{r}(k + i - 1)$ for positive integers k and r.
  - $k^{\overline{-r}} = \frac{1}{(k-r)^{\bar{r}}}$ for $k > r$ and $k \geq 1$.
  - $\sqrt{(k - 1/2)} \leq k^{\overline{1/2}} \leq \sqrt{k}$ and $\frac{1}{\sqrt{k-1/2}} < k^{\overline{-1/2}} < \frac{1}{\sqrt{k-1}}$ for $k > 0$.

Last Iterate convergence of SGD + M

Consider the SGD + M in the iterates averaging form with $x_0 = z_0 \in \mathcal{X}$ and

$$z_{t+1} = \mathbf{Proj}_{\mathcal{X}}\left(z_t - \eta_t \nabla \ell(x_t, \xi_t)\right) \text{ and } x_{t+1} = (1 - c_{t+1})x_t + c_{t+1} z_{t+1}.$$

Theorem 1

For all $t \geq 0$, *let* $\eta_t = \frac{\eta}{(t+1/2)^{1/2}}$ *with* $\eta = \frac{\sqrt{1/2}D_{\mathcal{X}}}{G}$ *and* $c_{t+1} = \frac{1}{t+1}$. *Then* $\mathbb{E}\left[f(x_t)\right] - f^* \leq \frac{\sqrt{2}D_{\mathcal{X}} G}{(t+3/2)^{1/2}}$.[a]

[a] Aaron Defazio and Robert M Gower. "The Power of Factorial Powers: New Parameter settings for (Stochastic) Optimization". In: *Asian Conference on Machine Learning*. PMLR. 2021, pp. 49–64, Theorem 2.

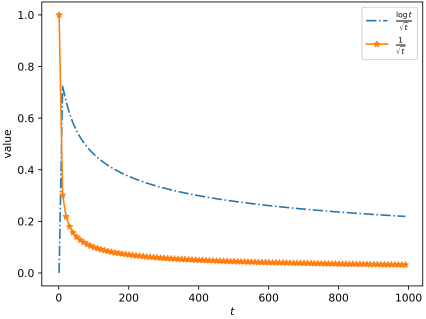
Remark

- The convergence rate is strictly better than the previously known $\mathcal{O}(1/\sqrt{t})$ for SGD + M, i.e., with $\eta_t = \frac{D_{\mathcal{X}}}{G\sqrt{2(t+1)}}$ and $c_{t+1} = \frac{1}{t+2}$, then $\mathbb{E}\left[f(x_t)\right] - f^* \leq \frac{\sqrt{2}D_{\mathcal{X}} G}{(t+1)^{1/2}}$.
- SGD + M has a clear advantage over SGD in the convex case in terms of the last iterate convergence. Recall that for SGD, if $\alpha_t = \frac{D_{\mathcal{X}}}{G\sqrt{t}}$, then $\mathbb{E}\left[f(x_t)\right] - f(x^*) \leq \frac{2GD_{\mathcal{X}}(2+\log t)}{t^{1/2}}$.

Deep Dive

Questions

1. Despite the factorial power trick, what is the exact contribution of momentum in improving the last iterate convergence compared with SGD? ($\mathcal{O}(\log t/\sqrt{t}) \to \mathcal{O}(1/\sqrt{t})$)
2. How does the factorial power trick work?



Deep Dive

Questions

1. Despite the factorial power trick, what is the exact contribution of momentum in improving the last iterate convergence compared with SGD? ($\mathcal{O}(\log t/\sqrt{t}) \to \mathcal{O}(1/\sqrt{t})$)
2. How does the factorial power trick work?

- "... summation and difference operations applied to $k^{\bar{r}}$ result in other factorial powers (instead of polynomials) ... It is this closure under summation and differencing that allows us to derive improved convergence rates when choosing step-sizes and momentum parameters based on factorial powers."[4]

[4] Aaron Defazio and Robert M Gower. "The Power of Factorial Powers: New Parameter settings for (Stochastic) Optimization". In: *Asian Conference on Machine Learning*. PMLR. 2021, pp. 49–64.

Deep Dive: Proof for SGD (I)

Recall $x_{t+1} = \textbf{Proj}_{\mathcal{X}}\left(x_t - \eta_t \nabla \ell(x_t, \xi_t)\right)$.

Step 1: Bound the distance between $\|x_{t+1} - x^*\|^2$, where x^* is an optimal solution.

For any $t \geq 0$, it holds almost surely that,

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|\textbf{Proj}_{\mathcal{X}}\left(x_t - \eta_t \nabla \ell(x_t, \xi_t)\right) - x^*\|^2 \\
&\leq \|(x_t - x^*) - \eta_t \nabla \ell(x_t, \xi_t)\|^2 \\
&\leq \|x_t - x^*\|^2 + \eta_t^2 G^2 - 2\eta_t \nabla \ell(x_t, \xi_t)^T (x_t - x^*) \\
&[\nabla \ell(x_t, \xi_t)^T (x^* - x_t)) \leq \ell(x^*, \xi_t) - \ell(x_t, \xi_t) \text{ due to the convexity of } \ell(\cdot, \xi_t)] \\
&\leq \|x_t - x^*\|^2 + \eta_t^2 G^2 - 2\eta_t (\ell(x_t, \xi_t) - \ell(x^*, \xi_t)).
\end{aligned}$$

Let $\eta_t = \frac{\eta}{\sqrt{t+1}}$ and multiply $\sqrt{t+1}$ on both sides of the above inequality

$$\sqrt{t+1}\,\|x_{t+1} - x^*\|^2 \leq \sqrt{t+1}\,\|x_t - x^*\|^2 + \frac{\eta}{\sqrt{t+1}} G^2 - 2\eta(\ell(x_t, \xi_t) - \ell(x^*, \xi_t)).$$

Deep Dive: Proof for SGD (II)

Step 2: Telescoping based on

$$\sqrt{t+1}\,\|x_{t+1}-x^*\|^2 \leq \sqrt{t+1}\,\|x_t-x^*\|^2 + \frac{\eta}{\sqrt{t+1}}G^2 - 2\eta(\ell(x_t,\xi_t)-\ell(x^*,\xi_t)).$$

- When $t=0$, $\mathbb{E}_{\xi_0}\left[\|x_1-x^*\|^2\right] \leq \|x_0-x^*\|^2 + \eta G^2 - 2\eta(f(x_0)-f(x^*))$.
- When $t \geq 1$, denote $\mathcal{F}_t = \sigma(\{\xi_0,\cdots,\xi_{t-1}\})$.

$$\begin{aligned}\sqrt{t+1}\mathbb{E}\left[\|x_{t+1}-x^*\|^2 \,|\mathcal{F}_t\right] &\leq \sqrt{t+1}\,\|x_t-x^*\|^2 + \frac{\eta}{\sqrt{t+1}}G^2 - 2\eta(f(x_t)-f(x^*)).\\ &[\text{using } \sqrt{t+1} \leq \sqrt{t} + \frac{1}{2\sqrt{t}}]\\ &\leq \left\{\sqrt{t}\,\|x_t-x^*\|^2 + \frac{1}{2\sqrt{t}}D_{\mathcal{X}}^2\right\} + \frac{\eta}{\sqrt{t+1}}G^2 - 2\eta(f(x_t)-f(x^*)).\end{aligned}$$

Taking the total expectation and telescoping, for all $t \geq 1$, one has

$$\sqrt{t+1}\mathbb{E}\left[\|x_{t+1}-x^*\|^2\right] \leq \|x_0-x^*\|^2 + \frac{D_{\mathcal{X}}^2}{2}\sum_{i=1}^{t}\frac{1}{\sqrt{i}} + \eta^2 G^2\sum_{i=0}^{t}\frac{1}{\sqrt{i+1}} - 2\eta\sum_{i=0}^{t}\mathbb{E}\left[f(x_i)-f(x^*)\right].$$

Deep Dive: Proof for SGD (III)

Step 3: Get the complexity bound the for averaged iterate.

$$\mathbb{E}\left[f\left(\frac{1}{t+1}\sum_{i=0}^{t} x_i\right)\right] - f(x^*) \le \frac{1}{\sqrt{t+1}}\left(\frac{1}{2\eta}D_{\mathcal{X}}^2 + \eta G^2\right).$$

$$\sqrt{t+1}\,\mathbb{E}\left[\|x_{t+1} - x^*\|^2\right] \le \|x_0 - x^*\|^2 + \frac{D_{\mathcal{X}}^2}{2}\sum_{i=1}^{t}\frac{1}{\sqrt{i}} + \eta^2 G^2 \sum_{i=0}^{t}\frac{1}{\sqrt{i+1}} - 2\eta\sum_{i=0}^{t}\mathbb{E}\left[f(x_i) - f(x^*)\right].$$

$$\sum_{i=1}^{t}\frac{1}{\sqrt{i}} \le 2(\sqrt{t} - 1)$$

$$\sum_{i=0}^{t}\frac{1}{\sqrt{i+1}} \le 2\sqrt{t+1}$$

$$f\left(\frac{1}{t+1}\sum_{i=0}^{t} x_i\right) - f(x^*) \le \frac{1}{t+1}\sum_{i=0}^{t}\left(f(x_i) - f(x^*)\right)$$

Deep Dive: Proof for SGD (IV)

Step 4: Get the complexity bound the last iterate.
Establish the inequality $\mathbb{E}\left[f(x_t) - f(x^*)\right] \leq \frac{1}{t+1}\sum_{i=0}^{t}\mathbb{E}\left[(f(x_i) - f(x^*))\right] + \text{sth}.$

For any $t \geq 0$ and $x \in \mathcal{X}$, it holds almost surely that,[5]

$$\|x_{t+1} - x\|^2 = \|\mathbf{Proj}_{\mathcal{X}}\left(x_t - \eta_t \nabla \ell(x_t, \xi_t)\right) - x\|^2 \leq \|(x_t - x) - \eta_t \nabla \ell(x_t, \xi_t)\|^2 \leq \|x_t - x\|^2 + \eta_t^2 G^2 - 2\eta_t \nabla \ell(x_t, \xi_t)^T (x_t - x)$$

Again using the convexity of $\ell(\cdot, \xi_t)$ with the above inequality

$$2\eta_t(\ell(x_t, \xi_t) - \ell(x, \xi_t)) \leq 2\eta_t \nabla \ell(x_t, \xi_t)(x_t - x) \leq \|x_t - x\|^2 - \|x_{t+1} - x\|^2 + \eta_t^2 G^2$$

Telescoping in t and taking the total expectation gives, for all $0 \leq k \leq t$

$$\sum_{i=t-k}^{t} \mathbb{E}\left[f(x_i) - f(x)\right] \leq \frac{\mathbb{E}\left[\|x_{t-k} - x\|^2\right]}{2\eta_{t-k}} + \sum_{i=t-k+1}^{t} \frac{\mathbb{E}\left[\|x_i - x\|^2\right]}{2}\left(\frac{1}{\eta_i} - \frac{1}{\eta_{i-1}}\right) + \frac{G^2}{2}\sum_{i=t-k}^{k} \eta_i.$$

Set $x = x_{t-k}$. Since $\eta_t = \frac{\eta}{\sqrt{t+1}}$ and $\sum_{i=t-k}^{k} \frac{1}{\sqrt{i+1}} \leq 2(\sqrt{t+1} - \sqrt{t-k})$, then , one obtains

$$\sum_{i=t-k}^{t} \mathbb{E}\left[f(x_i) - f(x_{t-k})\right] \leq \frac{D_{\mathcal{X}}^2}{2\eta}(\sqrt{t+1} - \sqrt{t-k+1}) + \frac{\eta G^2}{2} 2(\sqrt{t+1} - \sqrt{t-k}) \leq \left(\frac{D_{\mathcal{X}}^2}{2\eta} + \eta G^2\right)\frac{k+1}{\sqrt{t+1}}.$$

[5] Adapt from (Ohad Shamir and Tong 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, Theorem 2)

Deep Dive: Proof for SGD (V)

Step 4: Get the complexity bound on the last iterate.
Establish the inequality $\mathbb{E}\left[f(x_t) - f(x^*)\right] \leq \frac{1}{t+1}\sum_{i=0}^{t}\mathbb{E}\left[(f(x_i) - f(x^*))\right] + \text{sth}.$

From last page...

$$\sum_{i=t-k}^{t}\mathbb{E}\left[f(x_i) - f(x_{t-k})\right] \leq \frac{D_{\mathcal{X}^2}}{2\eta}(\sqrt{t+1} - \sqrt{t-k+1}) + \frac{\eta G^2}{2}2(\sqrt{t+1} - \sqrt{t-k}) \leq \left(\frac{D_{\mathcal{X}}^2}{2\eta} + \eta G^2\right)\frac{k+1}{\sqrt{t+1}}.$$

Denote $S_k^t = \frac{1}{k+1}\sum_{i=t-k}^{t}\mathbb{E}\left[f(x_i)\right]$. Then from the above inequality $S_k^t - \mathbb{E}\left[f(t_k)\right] \leq \left(\frac{D_{\mathcal{X}}^2}{2\eta} + \eta G^2\right)\frac{1}{\sqrt{t+1}}$. Together with the fact that $kS_{k-1}^t = (k+1)S_k^t - \mathbb{E}\left[f(t_k)\right]$, one has $S_{k-1}^t \leq S_k^t + \left(\frac{D_{\mathcal{X}}^2}{2\eta} + \eta G^2\right)\frac{1}{k\sqrt{t+1}}$. By telescoping on k, one obtains

$$\mathbb{E}\left[f(x_t)\right] = S_0^t \leq S_t^t + \left(\frac{D_{\mathcal{X}}^2}{2\eta} + \eta G^2\right)\frac{1}{\sqrt{t+1}}\sum_{i=1}^{t}\frac{1}{i} \leq \frac{1}{t+1}\sum_{i=0}^{t}\mathbb{E}\left[f(x_i)\right] + \left(\frac{D_{\mathcal{X}}^2}{2\eta} + \eta G^2\right)\frac{1}{\sqrt{t+1}}(1 + \log(t+1)).$$

Subtracting $f(x^*)$ on both sides gives the claimed $\mathcal{O}(\log t/t)$ convergence rate.

How to improve the last iterate bound?

Recall in the proof of SGD, from

$$\begin{aligned}\|x_{t+1} - x^*\|^2 = \|\mathbf{Proj}_{\mathcal{X}}(x_t - \eta_t \nabla \ell(x_t, \xi_t)) - x^*\|^2 &\leq \|(x_t - x^*) - \eta_t \nabla \ell(x_t, \xi_t)\|^2 \\ &\leq \|x_t - x^*\|^2 + \eta_t^2 G^2 - 2\eta_t \nabla \ell(x_t, \xi_t)^T (x_t - x^*) \\ &\leq \|x_t - x^*\|^2 + \eta_t^2 G^2 - 2\eta_t (\ell(x_t, \xi_t) - \ell(x^*, \xi_t)).\end{aligned}$$

we established

$$\sqrt{t+1}\mathbb{E}\left[\|x_{t+1} - x^*\|^2\right] \leq \|x_0 - x^*\|^2 + \frac{D_{\mathcal{X}}^2}{2}\sum_{i=1}^{t}\frac{1}{\sqrt{i}} + \eta^2 G^2 \sum_{i=0}^{t}\frac{1}{\sqrt{i+1}} - 2\eta \sum_{i=0}^{t}\mathbb{E}\left[f(x_i) - f(x^*)\right].$$

It is the $\sum_{i=0}^{t}(f(x_i) - f(x^*))$ that "slows" the last iterate convergence of SGD, which arises from the $\nabla \ell(x_t, \xi_t)^T (x_t - x^*)$. If we can create something like

$$-(i+1)(f(x_i) - f(x^*)) + i(f(x_{i-1}) - f(x^*))$$

from the $\nabla \ell(x_t, \xi_t)^T (z_t - x^*)$, then the telescoping will cancel the annoying $\sum_{i=0}^{t}(f(x_i) - f(x^*))$. This requires $z_t - x^*$ (as a replacement of $x_t - x^*$) to encode both information from x_t and x_{t-1}.

Deep Dive: Proof for SGD + M without the Factorial Trick

Recall SGD + M: $z_{t+1} = z_t - \eta_t \nabla \ell(x_t, \xi_t), x_{t+1} = (1 - c_{t+1})x_t + c_{t+1} z_{t+1}$.
Setup: $\eta_t = \frac{\eta}{\sqrt{t+1}}$ **and** $c_t = \frac{1}{t+1}$ **for all** $t \geq 0$.
When $c_t \leq 1$ for all t, using the same analysis as is done in SGD,

$$\|z_{t+1} - x_*\|^2 \leq \|z_t - x_*\|^2 + \eta_t^2 G^2 - 2\frac{1}{c_t}\eta_t \left[\ell\left(x_t, \xi_t\right) - \ell\left(x_*, \xi_t\right)\right] + 2\left(\frac{1}{c_t} - 1\right)\eta_t \left[\ell\left(x_{t-1}, \xi_t\right) - \ell\left(x_*, \xi_t\right)\right],$$

$$\sqrt{t+1}\mathbb{E}\left[\|z_{t+1} - x^*\|^2 \mid \mathcal{F}_t\right] \leq \|z_t - x^*\|^2 + \frac{D_{\mathcal{X}}^2}{2}\frac{1}{\sqrt{t}} + \eta^2 G^2 \frac{1}{\sqrt{t+1}} - 2\eta(t+1)(f(x_t) - f(x^*)) + 2\eta t(f(x_{t-1}) - f(x^*))$$

Then taking the total expectation and telescoping, for all $t \geq 1$, one has

$$\sqrt{t+1}\mathbb{E}\left[\|z_{t+1} - x^*\|^2\right] \leq \|z_0 - x^*\|^2 + \frac{D_{\mathcal{X}}^2}{2}\sum_{i=1}^{t}\frac{1}{\sqrt{i}} + \eta^2 G^2 \sum_{i=0}^{t}\frac{1}{\sqrt{i+1}} - 2\eta(t+1)\mathbb{E}\left[f(x_t) - f(x^*)\right].$$

The rest of proof is exactly the same as the one for SGD.

Deep Dive: What does the factorial power buy us? (I)

New setup: $\eta_t = \eta(t+1)^{\overline{-1/2}}$ **and** $c_t = \frac{1}{t+1}$ **for all** $t \geq 0$.

$$\frac{1}{(t+1)^{\overline{-1/2}}}\mathbb{E}\left[\|z_{t+1} - x^*\|^2\right] \leq \frac{1}{1^{\overline{-1/2}}}\|z_0 - x^*\|^2$$
$$+ \frac{D_{\mathcal{X}}^2}{2}\sum_{i=1}^{t}(i+1/2)^{\overline{-1/2}}$$
$$+ \eta^2 G^2 \sum_{i=0}^{t}(i+1)^{\overline{-1/2}}$$
$$- 2\eta(t+1)\mathbb{E}\left[f(x_t) - f(x^*)\right].$$

$$\sum_{i=1}^{t}(i+1/2)^{\overline{-1/2}} \leq 2(t+1)^{\overline{1/2}}$$

$$\sum_{i=0}^{t}(i+1)^{\overline{-1/2}} \leq 2(t+1)^{\overline{1/2}}$$

For $a + r > 0$ and $b \geq a \geq 1$, $\sum_{i=a}^{b} i^{\overline{r}} = \frac{1}{r+1} b^{\overline{r+1}} - \frac{1}{r+1} a^{\overline{r+1}}$.

Old setup: $\eta_t = \eta(t+1)^{-1/2}$ **and** $c_t = \frac{1}{t+1}$ **for all** $t \geq 0$.

$$\sqrt{t+1}\,\mathbb{E}\left[\|z_{t+1} - x^*\|^2\right] \leq \|z_0 - x^*\|^2$$
$$+ \frac{D_{\mathcal{X}}^2}{2}\sum_{i=1}^{t}\frac{1}{\sqrt{i}}$$
$$+ \eta^2 G^2 \sum_{i=0}^{t}\frac{1}{\sqrt{i+1}}$$
$$- 2\eta(t+1)\mathbb{E}\left[f(x_t) - f(x^*)\right].$$

$$\sum_{i=1}^{t}\frac{1}{\sqrt{i}} \leq 2(\sqrt{t} - 1)$$

$$\sum_{i=0}^{t}\frac{1}{\sqrt{i+1}} \leq 2\sqrt{t+1}$$

Deep Dive: What does the factorial power buy us? (II)

$$\mathbb{E}\left[f(x_t)\right] - f(x^*) \leq \frac{(t+1)^{\overline{1/2}}}{t+1}\left(\frac{1}{2\eta}D_{\mathcal{X}}^2 + \eta G^2\right).$$

Together with the ratio property, i.e., for $k+r>0$, $k+r+q>0$ and $k \geq 1$.

$$\frac{k^{\overline{r+q}}}{k^{\overline{r}}} = (k+r)^{\overline{q}}$$

we have

$$\mathbb{E}\left[f(x_t)\right] - f(x^*) \leq (t+2)^{\overline{-1/2}}\left(\frac{1}{2\eta}D_{\mathcal{X}}^2 + \eta G^2\right).$$

Extension: Beyond projected `SGD + M`.

[6] established a varaint of `SGD + M` has $\mathcal{O}(1/\sqrt{t})$ without bounded domain constriants.

Algorithm 1 FTRL-based SGDM

Input: A sequence $\alpha_1, ..., \alpha_T$, with $\alpha_1 > 0$. Non-increasing sequence $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{T-1}$. $\boldsymbol{m}_0 = 0$. $\boldsymbol{x}_1 \in \mathbb{R}^d$.
for $t = 1, \ldots, T$ **do**
  Get $\boldsymbol{g}_t$ at $\boldsymbol{x}_t$ such that $\mathbb{E}_t[\boldsymbol{g}_t] = \nabla f(\boldsymbol{x}_t)$
  $\beta_t = \frac{\sum_{i=1}^{t-1} \alpha_i}{\sum_{i=1}^{t} \alpha_i}$ (Define $\sum_{i=1}^{0} \alpha_i = 0$)
  $\boldsymbol{m}_t = \beta_t \boldsymbol{m}_{t-1} + (1 - \beta_t)\boldsymbol{g}_t$
  $\boldsymbol{\eta}_t = \frac{\alpha_{t+1} \sum_{i=1}^{t} \alpha_i}{\sum_{i=1}^{t+1} \alpha_i} \boldsymbol{\gamma}_t$
  $\boldsymbol{x}_{t+1} = \frac{\sum_{i=1}^{t} \alpha_i}{\sum_{i=1}^{t+1} \alpha_i} \boldsymbol{x}_t + \frac{\alpha_{t+1}}{\sum_{i=1}^{t+1} \alpha_i} \boldsymbol{x}_1 - \boldsymbol{\eta}_t \boldsymbol{m}_t$
end for

- **(H1)** f is L-smooth, that is, f is continuously differentiable and its gradient is L-Lipschitz, i.e., $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \le L\|\boldsymbol{x} - \boldsymbol{y}\|$.

We also use one or more of the following assumptions on the stochastic gradients $\boldsymbol{g}_t$.

- **(H2)** bounded variance: $\mathbb{E}_t\|\boldsymbol{g}_t - \nabla f(\boldsymbol{x}_t)\|^2 \le \sigma^2$.
- **(H3)** bounded in expectation: $\mathbb{E}\|\boldsymbol{g}_t\|^2 \le G^2$.
- **(H3')** ℓ_2 bounded: $\|\boldsymbol{g}_t\| \le G$.
- **(H3")** ℓ_∞ bounded: $\|\boldsymbol{g}_t\|_\infty \le G_\infty$.

Corollary 1. *Assume **(H3)** and set $\alpha_t = 1$ for all t. Algorithm 1 with either $\boldsymbol{\gamma}_{t-1} = \frac{c}{G\sqrt{t}} \cdot \mathbf{1}$ or $\boldsymbol{\gamma}_{t-1} = \frac{c}{G\sqrt{T}} \cdot \mathbf{1}$ guarantees*

$$\mathbb{E}[f(\boldsymbol{x}_T)] - f^\star \le \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 G}{c\sqrt{T}} + \frac{2cG}{\sqrt{T}} .$$

[6] Xiaoyu Li et al. "On the Last Iterate Convergence of Momentum Methods". In: *International Conference on Algorithmic Learning Theory*. PMLR. 2022, pp. 699–717.

Outline

Informal: In the long run the noise is the stochastic gradient vanishes (I)

Consider solving $\min_{x\in\mathbb{R}^n} f(x) = \mathbb{E}_{\xi\sim\mathcal{P}}[\ell(x,\xi)]$ with `SGD + M` using its standard form

$$m_{t+1} = \beta_t m_t + (1-\beta_t)\nabla\ell(x_t,\xi_t), x_{t+1} = x_t - \alpha_t m_{t+1}.$$

Lemma 2 (variance recursion)

When ∇f is L_f Lipschitz continuous and $\mathbb{E}_{\xi\sim\mathcal{P}}[\nabla\ell(x,\xi)] = \nabla f(x)$ for all $(x,\xi)\in\mathbb{R}^n\times\Xi$, then for all $t\geq 1$,

$$\mathbb{E}\left[\|m_{t+1}-\nabla f(x_t)\|^2 \mid \mathcal{F}_t\right] \leq \beta_t\|m_t - \nabla f(x_{t-1})\|^2 + 2(1-\beta_t)^2\mathbb{E}\left[\|\nabla\ell(x_t,\xi_t)-\nabla f(x_t)\|^2 \mid \mathcal{F}_t\right] + \frac{L_f^2\|x_t - x_{t-1}\|^2}{1-\beta_t}.$$

The lemma is adapted from [7] with the change of notations for consistency.

Proof. Define $v_t = \beta_t(\nabla f(x_t) - \nabla f(x_{t-1}))$, then it follows from the definition $m_{t+1} = \beta_t m_t + (1-\beta_t)\nabla\ell(x_t,\xi_t)$, smooth and, unbiasedness assumptions,

$$\begin{aligned}\mathbb{E}\left[\|m_{t+1}-\nabla f(x_t)+v_t\|^2 \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\|\beta_t(m_t - \nabla f(x_{t-1})) + (1-\beta_t)(\nabla\ell(x_t,\xi_t)-\nabla f(x_t))\|^2 \mid \mathcal{F}_t\right]\\ &= \beta_t^2\|m_t - \nabla f(x_{t-1})\|^2 + (1-\beta_t)^2\mathbb{E}\left[\|\nabla\ell(x_t,\xi_t)-\nabla f(x_t)\|^2 \mid \mathcal{F}_t\right] + 0 \qquad (2)\end{aligned}$$

[7] (Zhishuai Guo et al. "A novel convergence analysis for algorithms of the adam family and beyond". In: *arXiv preprint arXiv:2104.14840* [2021], Lemma 4)

Informal: In the long run the noise is the stochastic gradient vanishes (II)

Utilizing the fact $\|a+b\|^2 \leq (1+\epsilon)\|a\|^2 + (1+1/\epsilon)\|b\|^2$ for any $\epsilon > 0$ and $(a,b) \in \mathbb{R}^{2n}$, one has

$$\|m_{t+1} - \nabla f(x_t)\|^2 \leq (1+1-\beta_t)\|m_{t+1} - \nabla f(x_t) + v_t\|^2 + (1+1/(1-\beta_t))\|v_t\|^2 . \tag{3}$$

Combining (2) and (3) and taking the conditional expectation, we get

$$\begin{aligned}
&\mathbb{E}\left[\|m_{t+1} - \nabla f(x_t)\|^2 \mid \mathcal{F}_t\right] \\
&\leq (1+1-\beta)\beta_t^2 \|m_t - \nabla f(x_{t-1})\|^2 + (1-\beta_t)^2(1+1-\beta_t)\mathbb{E}\left[\|\nabla \ell(x_t,\xi_t) - \nabla f(x_t)\|^2 \mid \mathcal{F}_t\right] + \frac{(1+1-\beta_t)}{1-\beta_t}\|v_t\|^2 \\
&\overset{1-\beta_t = r_t}{\leq} (1-r_t^2)(1-r_t)\|m_t - \nabla f(x_{t-1})\|^2 + 2(1-\beta_t)^2\mathbb{E}\left[\|\nabla \ell(x_t,\xi_t) - \nabla f(x_t)\|^2 \mid \mathcal{F}_t\right] \\
&\quad + \frac{(1+r_t)(1-r_t)}{r_t} L_f^2 \|x_t - x_{t-1}\|^2 \\
&= \beta_t \|m_t - \nabla f(x_{t-1})\|^2 + 2(1-\beta_t)^2\mathbb{E}\left[\|\nabla \ell(x_t,\xi_t) - \nabla f(x_t)\|^2 \mid \mathcal{F}_t\right] + \frac{L_f^2\|x_t - x_{t-1}\|^2}{1-\beta_t}.
\end{aligned}$$

Informal: In the long run the noise is the stochastic gradient vanishes (III)

Now we further assume that

- There exits $(\sigma, C) \in \mathbb{R}_+$ such that $\mathbb{E}_{\xi\sim\mathcal{P}}\left[\|\nabla\ell(x) - \nabla f(x)\|^2\right] \le \sigma^2(1 + C\|\nabla f(x)\|^2)$ for all $x \in \mathbb{R}^n$.
- For any given $\epsilon > 0$, let $\beta_t := \beta \le \frac{\epsilon^2}{12\sigma^2}$, $\alpha_t = \alpha = \min\{\frac{\beta}{2L_f}, \frac{1}{\sqrt{2}L_f}\}$.

Then for $T > \max\{\frac{6\mathbb{E}[\Delta_0]}{\beta\epsilon^2}, \frac{12(f(x_0) - f_{\min})}{\alpha\epsilon^2}\}$, it follows from the variance recursion, above assumptions, and $\Delta_t = \frac{\Delta_{t+1} - (1-\beta)\Delta_t}{\beta} - \frac{\Delta_{t+1} - \Delta_t}{\beta}$, one can establish

$$\boxed{\mathbb{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\Delta_t\right] \le \frac{2\mathbb{E}\left[\Delta_0\right]}{\beta T} + 4\beta\sigma^2 + \mathbb{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla f(x_t)\|^2\right] \le 2\epsilon^2.}$$

Q&A

Thank you and Questions?

Contact: yud319@lehigh.edu