Convergence Rate Analysis of the AdamW-Style Shampoo: Unifying One-Sided and Two-Sided Preconditioning

Huan Li
Nankai University

Yiming Dong
Peking University

Zhouchen Lin
Peking University

Introduction

Consider nonconvex optimization problem

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} f(\mathbf{X})$$

where $\mathbf{X}$ is the matrix parameter with dimension $m \times n$

AdamW-style Shampoo

- An effective implementation of the classical Shampoo
- Won the AlgoPerf neural network training algorithm competition
- However, its convergence behavior is not well-understood

Algorithm AdamW-style Shampoo

Hyper parameters: $\eta, \theta, \beta, \lambda, \varepsilon, \mathbf{L}_{k,\varepsilon}^{\pm\frac{1}{\infty}} = \mathbf{I}_m, \mathbf{R}_{k,\varepsilon}^{\pm\frac{1}{\infty}} = \mathbf{I}_n$
positive p, q with $\frac{1}{p} + \frac{1}{q} = 1$

Initialize: $\mathbf{X}_1, \mathbf{M}_0 = \mathbf{0}, \mathbf{L}_0 = \mathbf{0}, \mathbf{R}_0 = \mathbf{0}$

for $k = 1, 2, \cdots, K$ **do**

$$\mathbf{G}_k = \mathrm{GradOracle}(\mathbf{X}_k)$$
$$\mathbf{M}_k = \theta\mathbf{M}_{k-1} + (1-\theta)\mathbf{G}_k$$
$$\mathbf{L}_k = \beta\mathbf{L}_{k-1} + (1-\beta)\mathbf{G}_k\mathbf{G}_k^T$$
$$\mathbf{R}_k = \beta\mathbf{R}_{k-1} + (1-\beta)\mathbf{G}_k^T\mathbf{G}_k$$
$$\mathbf{X}_{k+1} = (1-\lambda\eta)\mathbf{X}_k - \eta\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}$$

endfor

Contributions

We prove the following convergence rate for AdamW-style Shampoo measured by nuclear norm

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \leq O\left(\sqrt{m+n}\right) \times \max\left\{\sqrt[4]{\frac{\sigma^2 L(f(\mathbf{X}_1)-f^*)}{K}}, \sqrt{\frac{L(f(\mathbf{X}_1)-f^*)}{K}}\right\}$$

and $\|\mathbf{X}_k\|_{op} < \frac{1}{\lambda}$ for all iterates ($\|\cdot\|_{op}$ denotes the spectral norm). It can be considered to be analogous to the following optimal convergence rate of SGD

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_F\right] \leq O\left(\sqrt[4]{\frac{\sigma^2 L(f(\mathbf{X}_1)-f^*)}{K}}\right)$$

in the ideal case of $\|\nabla f(\mathbf{X})\|_* = \Theta(\sqrt{\min\{m,n\}})\|\nabla f(\mathbf{X})\|_F$ and balanced m and n .

Assumptions

- Smoothness: $\|\nabla f(\mathbf{Y}) - \nabla f(\mathbf{X})\| \leq L\|\mathbf{Y} - \mathbf{X}\|$

- Unbiased estimator: $\mathbb{E}\left[\mathbf{G}_k\right] = \nabla f(\mathbf{X}_k)$

- Bounded row-wise and column-wise second central moment matrices:

$$\mathbb{E}\left[\left(\mathbf{G}_k - \nabla f(\mathbf{X}_k)\right)\left(\mathbf{G}_k - \nabla f(\mathbf{X}_k)\right)^T\right] \preceq \Sigma_L$$

$$\mathbb{E}\left[\left(\mathbf{G}_k - \nabla f(\mathbf{X}_k)\right)^T\left(\mathbf{G}_k - \nabla f(\mathbf{X}_k)\right)\right] \preceq \Sigma_R$$

From the third assumption, it readily follows that

$$\mathbb{E}\left[\left\|\mathbf{G}_k - \nabla f(\mathbf{X}_k)\right\|_F^2\right] \leq \frac{\mathrm{tr}(\Sigma_L)+\mathrm{tr}(\Sigma_R)}{2} \equiv \sigma^2$$

Theorem

Suppose that the three assumptions and conditions $\mathbf{L}_{k,\varepsilon} \succeq \hat{\varepsilon}\mathbf{I}_m$ and $\mathbf{R}_{k,\varepsilon} \succeq \hat{\varepsilon}\mathbf{I}_n$ hold for some $\hat{\varepsilon} \geq \varepsilon$. Define $\hat{\sigma}^2 = \max\left\{\sigma^2, \frac{L(f(\mathbf{X}_1)-f^*)}{K\gamma^2}\right\}$ with any $\gamma \in (0, 1]$. Let

$$\frac{1}{p} + \frac{1}{q} = 1, \quad 1 - \theta = \sqrt{\frac{L(f(\mathbf{X}_1)-f^*)}{K\hat{\sigma}^2}}, \quad \theta \leq \beta \leq \sqrt{\theta},$$

$\varepsilon = \frac{\tau\hat{\sigma}^2}{m+n}$ with any $\tau \leq 1$ being the hyperparameter to make ε small in practice.

$$\eta = \sqrt{\frac{\hat{\varepsilon}(f(\mathbf{X}_1)-f^*)}{4KL\hat{\sigma}^2}}, \quad \lambda \leq \frac{1}{\sqrt{1152\hat{\varepsilon}}K^{3/4}}\sqrt[4]{\frac{L^3\hat{\sigma}^2}{f(\mathbf{X}_1)-f^*}}, \quad \|\mathbf{X}_1\|_\infty \leq \sqrt{\frac{\hat{\varepsilon}K(f(\mathbf{X}_1)-f^*)}{L\hat{\sigma}^2}}.$$

Then for AdamW-style Shampoo, we have $\|\mathbf{X}_k\|_{op} < \frac{1}{\lambda}$ for all $k = 1, 2, \cdots, K$ and

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \leq \left(8\sqrt{m+n} + \frac{119\hat{\sigma}}{\sqrt{\hat{\varepsilon}}}\right) \times \max\left\{\sqrt[4]{\frac{\sigma_s^2 L(f(\mathbf{X}_1) - f^*)}{K}}, \sqrt{\frac{L(f(\mathbf{X}_1) - f^*)}{K}}\right\}.$$

In the worst case, when $\hat{\varepsilon} = \varepsilon$, we have $\left(8\sqrt{m+n} + \frac{119\hat{\sigma}}{\sqrt{\hat{\varepsilon}}}\right) = 127\sqrt{\frac{m+n}{\tau}}$.

Furthermore, when $\tau = 1$, we have $\left(8\sqrt{m+n} + \frac{119\hat{\sigma}}{\sqrt{\hat{\varepsilon}}}\right) = 127\sqrt{m+n}$.

Discussions

Optimality of Our Convergence Rate

- Optimal with respect to $K, \sigma, L, f(\mathbf{X}_1) - f^*$.
- Analogous to the optimal convergence rate of SGD in the ideal case of $\|\nabla f(\mathbf{X})\|_* = \Theta(\sqrt{\min\{m,n\}})\|\nabla f(\mathbf{X})\|_F$ and balanced m and n.

Unifying two-sided and one-sided preconditioning

- A unified treatment of two-sided ($p, q < +\infty$) and one-sided ($p = 1, q = +\infty$ or $q = 1, p = +\infty$) preconditioning.
- Intuitively, the left preconditioning $\mathbf{L}^{-\frac{1}{2}}\mathbf{M}$ captures correlations within each column of $\mathbf{M}$, while the right preconditioning $\mathbf{M}\mathbf{R}^{-\frac{1}{2}}$ captures correlations within each row of $\mathbf{M}$. Two-sided preconditioning $\mathbf{L}^{-\frac{1}{2p}}\mathbf{M}\mathbf{L}^{-\frac{1}{2q}}$ combines these advantages and captures correlations within both rows and columns of $\mathbf{M}$.

Discussions

AdamW-style Shampoo v.s. AdamW

- Convergence rate of AdamW:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_k)\right\|_1\right] \leq O\left(\frac{\sqrt{d}}{K^{1/4}}\sqrt[4]{\sigma^2 L(f(\mathbf{x}_1)-f^*)}+\sqrt{\frac{dL(f(\mathbf{x}_1)-f^*)}{K}}\right)$$

  while ensuring $\left\|\mathbf{x}_k\right\|_\infty < \frac{1}{\lambda}$, where d is the dimension.

- AdamW-style Shampoo can be interpreted as achieving theoretical behavior analogous to AdamW, but in the space of singular values.

Discussions

AdamW-style Shampoo v.s. Muon

- Let $\mathbf{U}_k \Sigma_k \mathbf{V}_k^T$ be the compact SVD of $\mathbf{M}_k$. From the identities

$$(\mathbf{M}_k \mathbf{M}_k^T)^{\frac{1}{2p}} = \mathbf{U}_k \Sigma_k^{\frac{1}{p}} \mathbf{U}_k^T \quad \text{and} \quad (\mathbf{M}_k^T \mathbf{M}_k)^{\frac{1}{2p}} = \mathbf{V}_k \Sigma_k^{\frac{1}{q}} \mathbf{V}_k^T$$

the update $\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}} \mathbf{M}_k \mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}$ can be written equivalently as

$$\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}} \left(\mathbf{M}_k \mathbf{M}_k^T\right)^{\frac{1}{2p}} \mathbf{U}_k \mathbf{V}_k^T \left(\mathbf{M}_k^T \mathbf{M}_k\right)^{\frac{1}{2q}} \mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}$$

- In an informal sense, $\mathbf{M}_k$, $\mathbf{L}_k$, and $\mathbf{R}_k$ can be interpreted as approximations to the first moment matrix $\mathbb{E}[\mathbf{G}]$, the row-wise second raw moment matrix $\mathbb{E}\left[\mathbf{G}\mathbf{G}^T\right]$ and the column-wise second raw moment matrix $\mathbb{E}\left[\mathbf{G}^T\mathbf{G}\right]$, respectively.
- The quantities $\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}} \left(\mathbf{M}_k \mathbf{M}_k^T\right)^{\frac{1}{2p}}$ and $\left(\mathbf{M}_k^T \mathbf{M}_k\right)^{\frac{1}{2q}} \mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}$ can be therefore regarded as the row-wise and column-wise signal-to-total-energy ratio (STR) matrices, respectively.
- AdamW-style Shampoo can be viewed as an STR preconditioned variant of Muon. This relationship is analogous to that between AdamW and SignSGD

Thanks for reading!