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

Anonymous Authors[1]

Abstract

This paper studies the AdamW-style Shampoo optimizer, an effective implementation of the classical Shampoo that notably won the external tuning track of the AlgoPerf neural network training algorithm competition (Kasimbeg et al., 2025). Our analysis unifies one-sided and two-sided preconditioning and establishes the convergence rate $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \leq \mathcal{O}(\frac{\sqrt{m+n}C}{K^{1/4}})$ measured by nuclear norm, where K represents the iteration number, (m,n) denotes the size of matrix parameters, and C matches the constant in the optimal convergence rate of SGD. Theoretically, we have $\|\nabla f(\mathbf{X})\|_F \leq \|\nabla f(\mathbf{X})\|_* \leq \sqrt{\min\{m,n\}}\|\nabla f(\mathbf{X})\|_F$, supporting that our convergence rate can be considered to be analogous to the optimal $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_F\right] \leq \mathcal{O}(\frac{C}{K^{1/4}})$ 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.

1. Introduction

Adaptive gradient methods have become the predominant optimizer for training deep neural networks, especially in large language models. The development of adaptive gradient algorithms has followed two distinct lineages: diagonal preconditioning and non-diagonal preconditioning. The former has developed through the progression of AdaGrad (Duchi et al., 2011; McMahan & Streeter, 2010), RMSProp (Tieleman & Hinton, 2012), Adam (Kingma & Ba, 2015), and finally AdamW (Loshchilov & Hutter, 2019), which have served as the de facto optimizer for training deep networks over the past decade. The latter lineage, originating from full-matrix AdaGrad (Duchi et al., 2011) and advancing to methods such as Shampoo (Gupta et al., 2018), SOAP (Vyas et al., 2025), and Muon (Jordan et al., 2024), now demonstrates the potential to outperform its diagonal counterparts.

Diagonally preconditioned methods typically employ coordinate-wise scaling to the gradient. For instance, AdaGrad treats the network's parameters as a single high-dimensional vector and updates them according to the following procedure

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\Lambda_k^{-1/2}\mathbf{g}_k, \quad \Lambda_k = \operatorname{diag}\left(\sum_{t=1}^{k}\operatorname{diag}(\mathbf{g}_t\mathbf{g}_t^T)\right),$$

where Λ_k is a diagonal preconditioning matrix whose entries are the element-wise sum of squared historical gradients.

In contrast, non-diagonally preconditioned methods exploit the inherent matrix structure of parameters in deep networks. For example, the Shampoo optimizer operates according to the following formulation with two-sided preconditioning

$$\mathbf{X}_{k+1} = \mathbf{X}_k - \eta\mathbf{L}_k^{-1/4}\mathbf{G}_k\mathbf{R}_k^{-1/4},$$
$$\mathbf{L}_k = \sum_{t=1}^{k}\mathbf{G}_t\mathbf{G}_t^T, \quad \mathbf{R}_k = \sum_{t=1}^{k}\mathbf{G}_t^T\mathbf{G}_t,$$

where the gradient $\mathbf{G}_k \in \mathbb{R}^{m\times n}$ is a matrix and $\mathbf{L}_k \in \mathbb{R}^{m\times m}$ and $\mathbf{R}_k \in \mathbb{R}^{n\times n}$ are non-diagonal preconditioners consisting of sum of historical gradient outer products. It can be regarded as using the Kronecker product $\mathbf{R}_k^{1/2} \otimes \mathbf{L}_k^{1/2}$ to approximate the full-matrix AdaGrad preconditioner $\sum_{t=1}^{k}\mathbf{g}_t\mathbf{g}_t^T$, where $\mathbf{g}_t = \operatorname{vec}(\mathbf{G}_t)$.

A key advantage of non-diagonally preconditioned methods is their ability to capture the cross-parameter correlations within the gradient, thereby yielding a more informed search direction and potentially superior convergence compared to diagonal approaches. Recently, an implementation based on the distributed Shampoo (Anil et al., 2020; Shi et al., 2023) won the external tuning track of the AlgoPerf neural network training algorithm competition (Kasimbeg et al., 2025), demonstrating that non-diagonally preconditioned training algorithms can outperform currently popular diagonal preconditioning methods, such as Adam. The winner implementation achieved significantly accelerated training,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Algorithm 1 AdamW-style Shampoo

Hyper parameters: $\eta, \theta, \beta, \lambda, \varepsilon$, positive p, q with $\frac{1}{p}+\frac{1}{q}=1$, $\mathbf{L}_{k,\varepsilon}^{\pm\frac{1}{\infty}}=\mathbf{I}_m$, and $\mathbf{R}_{k,\varepsilon}^{\pm\frac{1}{\infty}}=\mathbf{I}_n$.
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**
$\mathbf{G}_k=\text{GradOracle}(\mathbf{X}_k)$
$\mathbf{M}_k=\theta\mathbf{M}_{k-1}+(1-\theta)\mathbf{G}_k$
$\mathbf{L}_k=\beta\mathbf{L}_{k-1}+(1-\beta)\mathbf{G}_k\mathbf{G}_k^T$
$\mathbf{R}_k=\beta\mathbf{R}_{k-1}+(1-\beta)\mathbf{G}_k^T\mathbf{G}_k$
$\mathbf{L}_{k,\varepsilon}=\mathbf{L}_k+\varepsilon\mathbf{I}_m, \quad \mathbf{R}_{k,\varepsilon}=\mathbf{R}_k+\varepsilon\mathbf{I}_n$
$\mathbf{X}_{k+1}=(1-\lambda\eta)\mathbf{X}_k-\eta\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}$
end for

with an average speedup of 28% than the NAdamW (Dozat, 2016) baseline across eight deep learning workloads. Algorithm 1 presents the core characteristics of the Shampoo implemented in (Anil et al., 2020; Shi et al., 2023) in a non-distributed manner, including the exponential moving average of the first and second moments, two-sided preconditioning with a tunable exponent, and decoupled weight decay.

Theoretically, convergence of diagonally preconditioned methods have been extensively studied (Défossez et al., 2022; Shi et al., 2020; Li et al., 2025b; Zhang et al., 2022; Hong & Lin, 2024; Li et al., 2023; 2025a). For non-diagonal methods, Muon represents the first method to receive a rigorous convergence analysis for nonconvex optimization (Li & Hong, 2025; Shen et al., 2025; Chen et al., 2025; Sato et al., 2025). Analyses of other optimizers in this class, such as Shampoo, have largely been confined to convex settings. For example, Gupta et al. (2018) established the regret bound of Shampoo within the online convex optimization framework, Xie et al. (2025a) provided a unified convergence analysis including full-matrix AdaGrad and *one-sided* variant of Shampoo for convex problems, where the update $\mathbf{L}_k^{-1/4}\mathbf{G}_k\mathbf{R}_k^{-1/4}$ in the original Shampoo is replaced by $\mathbf{L}_k^{-1/2}\mathbf{G}_k$, An et al. (2025) proposed ASGO, effectively equivalent to *one-sided* Shampoo, and studied its convergence for convex programming.

To the best of our knowledge, (Xie et al., 2025b) appears to be the only one to establish the convergence of Shampoo in nonconvex setting. However, their study is limited to the *one-sided* variant of Shampoo in the *AdaGrad-style* and *RMSProp-style*, and does not address the more complex, yet more commonly used, two-sided preconditioning. Furthermore, their analysis does not incorporate momentum or decoupled weight decay. While other works (Feinberg et al., 2023; Morwani et al., 2025; Eschenhagen et al., 2025; Lin et al., 2025) have explored Shampoo from different perspectives, none have provided a convergence guarantee for the nonconvex case.

In this paper, we study the AdamW-style Shampoo presented in Algorithm 1, which provides a unified treatment of two-sided ($p, q < +\infty$) and one-sided ($p=1, q=+\infty$ or $p=+\infty, q=1$) preconditioning. This formulation represents the most effective practical implementation of Shampoo (Anil et al., 2020; Shi et al., 2023; Kasimbeg et al., 2025).

1.1. Contributions

This paper establishes the following convergence rate of Algorithm 1 for nonconvex programming

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

while ensuring $\|\mathbf{X}_k\|_{op}<\frac{1}{\lambda}$ for all $k=1,2,\cdots,K$, where the notations can be found in Section 1.2. As a comparison, the classical convergence rate of SGD is (Bottou et al., 2018)

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

which matches the lower bound of nonconvex stochastic optimization (Arjevani et al., 2023). Since Frobenius norm and nuclear norm satisfies

$$\|\nabla f(\mathbf{X})\|_F \le \|\nabla f(\mathbf{X})\|_* \le \sqrt{\min\{m,n\}}\|\nabla f(\mathbf{X})\|_F,$$

our convergence rate also matches the same lower bound with respect to all the coefficients in the ideal case of $\|\nabla f(\mathbf{X})\|_* = \Theta(\sqrt{\min\{m,n\}})\|\nabla f(\mathbf{X})\|_F$ and balanced m and n, which is verified empirically on real training of GPT-2 in our experiment.

1.2. Problem Settings, Notations, and Assumptions

We study the following nonconvex problem with matrix parameter in this paper

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} f(\mathbf{X}), \quad (2)$$

where $f(\mathbf{X})=\mathbb{E}_{\zeta\in\mathcal{P}}[f(\mathbf{X};\zeta)]$ and ζ is the sample drawn from the data distribution $\mathcal{P}$.

We denote vectors by lowercase bold letters and matrices by uppercase bold letters. We use $\mathbf{I}_m$ for the identity matrix in $\mathbb{R}^{m\times m}$. For vectors, denote $\|\cdot\|$ as the ℓ_2 Euclidean norm. For matrices, denote $\|\cdot\|_F$, $\|\cdot\|_{op}$, and $\|\cdot\|_*$ as the Frobenius norm, spectral norm (largest singular value), and nuclear norm (sum of singular values), respectively. The trace of a square matrix is written as $\mathrm{tr}(\cdot)$. Denote the singular values of $\mathbf{A}\in\mathbb{R}^{m\times n}$ by $\sigma_1(\mathbf{A}),\cdots,\sigma_r(\mathbf{A})$ with

$r = \min\{m, n\}$. Denote $\mathcal{F}_k = \sigma(\zeta_1, \zeta_2, \cdots, \zeta_k)$ to be the sigma field of the stochastic samples up to k, denote $\mathbb{E}_{\mathcal{F}_k}[\cdot]$ as the expectation with respect to $\mathcal{F}_k$ and $\mathbb{E}_k[\cdot|\mathcal{F}_{k-1}]$ the conditional expectation with respect to ζ_k given $\mathcal{F}_{k-1}$. For the sake of brevity, $\mathbb{E}_{\mathcal{F}_K}[\cdot]$ will be denoted as $\mathbb{E}[\cdot]$. Finally, let f^* denote the lower bound of $f(\mathbf{X})$.

We make the following assumptions throughout this paper:

1. Smoothness: $\|\nabla f(\mathbf{Y}) - \nabla f(\mathbf{X})\|_F \leq L\|\mathbf{Y} - \mathbf{X}\|_F, \forall \mathbf{X}, \mathbf{Y}$,
2. Unbiased estimator: $\mathbb{E}_k\left[\mathbf{G}_k \middle| \mathcal{F}_{k-1}\right] = \nabla f(\mathbf{X}_k)$,
3. Bounded noise variance: $\mathbb{E}_k\left[(\mathbf{G}_k - \nabla f(\mathbf{X}_k))(\mathbf{G}_k - \nabla f(\mathbf{X}_k))^T \middle| \mathcal{F}_{k-1}\right] \preceq \Sigma_L$, $\mathbb{E}_k\left[(\mathbf{G}_k - \nabla f(\mathbf{X}_k))^T(\mathbf{G}_k - \nabla f(\mathbf{X}_k)) \middle| \mathcal{F}_{k-1}\right] \preceq \Sigma_R$ with symmetric positive semidefinite matrices Σ_L and Σ_R.

The first two assumptions are identical to the standard assumptions used in the analysis of SGD, while the third assumption is more restrictive than that in SGD analysis. In fact, from the third assumption, it readily follows that

$$\mathbb{E}_k\left[\|\mathbf{G}_k - \nabla f(\mathbf{X}_k)\|_F^2 \middle| \mathcal{F}_{k-1}\right] \leq \frac{\mathrm{tr}(\Sigma_L) + \mathrm{tr}(\Sigma_R)}{2} \equiv \sigma^2,$$

which is the standard bounded variance assumption in SGD analysis.

2. Convergence Rate of the AdamW-Style Shampoo

Based on Assumptions 1-3, we establish the convergence rate of Algorithm 1 in the following theorem. Due to the definitions of $\mathbf{L}_{k,\varepsilon}$ and $\mathbf{R}_{k,\varepsilon}$, condition (3) always holds for $\hat{\varepsilon} = \varepsilon$.

Theorem 2.1. *Suppose that Assumptions 1-3 and*

$$\mathbf{L}_{k,\varepsilon} \succeq \hat{\varepsilon}\mathbf{I}_m, \quad \mathbf{R}_{k,\varepsilon} \succeq \hat{\varepsilon}\mathbf{I}_n \tag{3}$$

hold for some $\hat{\varepsilon} \geq \varepsilon$. Let $\hat{\sigma}^2 = \max\left\{\sigma^2, \frac{L(f(\mathbf{X}_1) - f^*)}{K\gamma^2}\right\}$ with any $\gamma \in (0, 1]$, $\frac{1}{p} + \frac{1}{q} = 1$, $1 - \theta = \sqrt{\frac{L(f(\mathbf{X}_1) - f^*)}{K\hat{\sigma}^2}}$, $\theta \leq \beta \leq \sqrt{\theta}$, $\varepsilon = \frac{\tau\hat{\sigma}^2}{m+n}$ with any $\tau \leq 1$ being the hyperparameter to make ε small in practice, $\eta = \sqrt{\frac{\hat{\varepsilon}(f(\mathbf{X}_1) - f^*)}{4LK\hat{\sigma}^2}}$, $\lambda \leq \frac{1}{\sqrt{1152\hat{\varepsilon}}K^{3/4}}\sqrt[4]{\frac{L^3\hat{\sigma}^2}{f(\mathbf{X}_1) - f^*}}$, and $\|\mathbf{X}_1\|_{op} \leq \sqrt{\frac{\hat{\varepsilon}K(f(\mathbf{X}_1) - f^*)}{L\hat{\sigma}^2}}$. Then for Algorithm 1, we have $\|\mathbf{X}_k\|_{op} < \frac{1}{\lambda}$ for all $k = 1, 2, \cdots, K$ and

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

In the worst case, when $\hat{\varepsilon} = \varepsilon$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \leq 127\sqrt{\frac{m+n}{\tau}} \times \max\left\{\sqrt[4]{\frac{\sigma^2 L(f(\mathbf{X}_1) - f^*)}{K}}, \sqrt{\frac{L(f(\mathbf{X}_1) - f^*)}{K\gamma}}\right\}. \tag{5}$$

Furthermore, when $\tau = 1$, we achieve the best theoretical convergence rate

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \leq 127\sqrt{m+n} \times \max\left\{\sqrt[4]{\frac{\sigma^2 L(f(\mathbf{X}_1) - f^*)}{K}}, \sqrt{\frac{L(f(\mathbf{X}_1) - f^*)}{K\gamma}}\right\}. \tag{6}$$

Discussions on ε and $\hat{\varepsilon}$. In practice, the parameter ε is typically set to a very small value, for example, 10^{-12} as used in (Shi et al., 2023). On the other hand, in modern large language models, the dimensions of weight matrices optimized by non-diagonally preconditioned methods such as Shampoo/SOAP/Muon are not exceptionally large. For instance, in GPT-3 with 175 bilion parameters, the QKV projection matrices have dimensions $m = n = 12288$, while the weight matrices in the feed-forward network layer have dimensions $(m, n) = (12288, 49152)$ or $(49152, 12288)$. Consequently, the quantity $\frac{\hat{\sigma}^2}{m+n}$ is several orders of magnitude larger than the ε used in practice. To reconcile this discrepancy, we introduce the scaling factor τ into the setting of ε, aligning the analysis with practical configurations. This adjustment, however, leads to a slower convergence rate, as shown in (5), which depends explicitly on τ. The unpractical setting $\tau = 1$ yields the best convergence rate given in (6).

To bridge this gap between theory and practice, we further introduce condition (3). Informally, the preconditioners $\mathbf{L}_k$ and $\mathbf{R}_k$ can be regarded as approximating $\mathbb{E}\left[\mathbf{G}\mathbf{G}^T\right]$ and $\mathbb{E}\left[\mathbf{G}^T\mathbf{G}\right]$, respectively. In the training of modern large language models, empirical evidence indicates that the gradient norm remains $\mathcal{O}(1)$ (Wen et al., 2025, Figure 6). Consequently, Lemma 2.2 suggests that condition (3) is reasonable with $\hat{\varepsilon} = \mathcal{O}(\frac{1}{m+n})$. As illustrated in Figure 3 in Appendix D, this condition is empirically satisfied during GPT-2 training for a moderate value of $\hat{\varepsilon}$, which remains orders of magnitude larger than ε. With condition (3), our derived convergence rate (4) depends only on $\hat{\varepsilon}$, rather than ε or τ. When $\hat{\varepsilon} \geq \frac{\hat{\sigma}^2}{m+n}$, convergence rate (4) matches (6) , even for arbitrarily small τ.

Lemma 2.2. *When each element of $\mathbf{G} \in \mathbb{R}^{m\times n}$ is generated from Gaussian distribution with μ mean and ξ^2 vari-*

ance independently, we have

$$\mathbb{E}\left[\mathbf{G}\mathbf{G}^T\right] \succeq n\xi^2\mathbf{I}_m = \frac{\xi^2}{m(\xi^2+\mu^2)}\mathbb{E}\left[\|\mathbf{G}\|_F^2\right]\mathbf{I}_m,$$
$$\mathbb{E}\left[\mathbf{G}^T\mathbf{G}\right] \succeq m\xi^2\mathbf{I}_n = \frac{\xi^2}{n(\xi^2+\mu^2)}\mathbb{E}\left[\|\mathbf{G}\|_F^2\right]\mathbf{I}_n.$$

Optimality of our convergence rate. Comparing the optimal convergence rate (1) of SGD with our convergence rate (6), we observe that our result employs the nuclear norm and introduces an additional factor of $\sqrt{m+n}$. Denoting $\sigma_1, \sigma_2, \cdots, \sigma_r$ to be the singular values of $\nabla f(\mathbf{X})$ with $r = \min\{m, n\}$, the Frobenius norm and nuclear norm satisfies

$$\|\nabla f(\mathbf{X})\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2} \le \sum_{i=1}^r \sigma_i = \|\nabla f(\mathbf{X})\|_*,$$
$$\|\nabla f(\mathbf{X})\|_* = \sum_{i=1}^r \sigma_i \le \sqrt{r\sum_{i=1}^r \sigma_i^2} = \sqrt{r}\|\nabla f(\mathbf{X})\|_F,$$

which means that our convergence rate also matches the lower bound in nonconvex stochastic optimization (Arjevani et al., 2023) in the ideal case of $\|\nabla f(\mathbf{X})\|_* = \Theta(\sqrt{\min\{m,n\}})\|\nabla f(\mathbf{X})\|_F$ and balanced m and n, which is verified empirically on real training of GPT-2, as demonstrated in Figure 2 in Appendix D.

Unifying two-sided and one-sided preconditioning. When $p, q < +\infty$, Algorithm 1 employs two-sided preconditioning; for instance, setting $p = q = 2$ recovers the original Shampoo update proposed in (Gupta et al., 2018). If either p or q is infinite, the algorithm reduces to the one-sided preconditioning analyzed in (Xie et al., 2025a; An et al., 2025; Xie et al., 2025b). Our analysis framework permits any positive values p, q (including infinite ones) satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Empirically, Anil et al. (2020); Shi et al. (2023) observed that treating the exponents p and q as tunable hyperparameters can lead to improved performance.

AdamW-style Shampoo v.s. AdamW itself. Li et al. (2025a) established the convergence rate $\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|_1\right] \le \mathcal{O}\left(\frac{\sqrt{d}C}{K^{1/4}}\right)$ for AdamW while ensuring $\|\mathbf{x}_k\|_\infty < \frac{1}{\lambda}$ for all $k = 1, 2, \cdots, K$, where C matches the constant in the optimal convergence rate of SGD and d is the dimension of the variable. Comparing with (6), we observe that AdamW-style Shampoo employs the nuclear norm in its convergence rate and spectral norm in its implicit bias, those correspond to the ℓ_1 and ℓ_∞ norms applied to the singular values, respectively. Consequently, AdamW-style Shampoo can be interpreted as achieving theoretical behavior analogous to AdamW, but in the space of singular values.

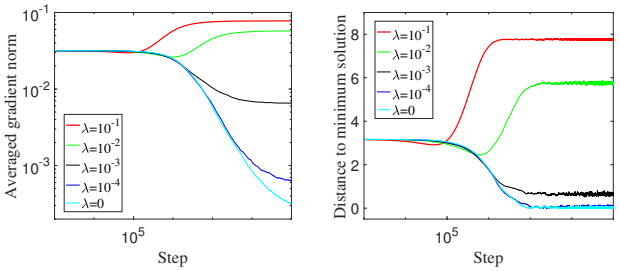

Figure 1. Illustrations of $\frac{1}{k}\sum_{t=1}^k \|\nabla f(\mathbf{X}_t)\|_F$ (left) and $\|\mathbf{X}_k - \mathbf{X}^*\|_F$ (right) over steps on the toy example (7).

Restricted weight decay parameter. Our theory requires the weight decay parameter λ to be sufficiently small. To demonstrate that such an upper bound is necessary for convergence, we follow (Li et al., 2025a) to consider a simple stochastic convex problem

$$f(\mathbf{X}) = \frac{\|\mathbf{X} - \mathbf{X}^*\|_F^2}{200} \quad \text{with} \quad \mathbf{X}^* = \begin{bmatrix} 4, & 4 \\ 4, & 4 \end{bmatrix}. \tag{7}$$

The stochastic gradient is given by

$$g(\mathbf{X}) = \begin{cases} \mathbf{X} - \mathbf{X}^* - \mathbf{A}, & \text{with probability } p = 0.1, \\ -\frac{1}{10}(\mathbf{X} - \mathbf{X}^* - \frac{10}{9}\mathbf{A}), & \text{with probability } 1-p. \end{cases}$$

We initialize $\mathbf{X}_1 = \mathbf{X}^* + \begin{bmatrix} -1, & -2 \\ 2, & 1 \end{bmatrix}$ and set $\mathbf{A} = \begin{bmatrix} 1, & 1 \\ 1, & 1 \end{bmatrix}$, $K = 10^9$, $\theta = 1 - \frac{1}{\sqrt{K}}$, $\beta = \sqrt{\theta}$, $\eta = \frac{1}{\sqrt{K}}$, $\varepsilon = 10^{-12}$, $\mathbf{M}_0 = \mathbf{0}$, $\mathbf{L}_0 = \mathbf{0}$, $\mathbf{R}_0 = \mathbf{0}$ for Algorithm 1. We test $\lambda = \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 0\}$ such that $\|\mathbf{X}^*\|_{op} = 8 < \frac{1}{\lambda}$ and $\|\mathbf{X}_1\|_{op} \le 7.7 < \frac{1}{\lambda}$. Figure 1 demonstrates that Algorithm 1 fails to converge to $\mathbf{X}^*$ when $\lambda = \{10^{-1}, 10^{-2}, 10^{-3}\}$. This indicates that, even for a simple convex problem, convergence to the minimum solution is not guaranteed if λ exceeds a certain threshold, although it is unclear whether our upper bound is tight.

Comparison with (Xie et al., 2025b). While Xie et al. (2025b) established convergence for adaptive optimizers under a nonconvex, adaptive smoothness framework, our work differs in several key aspects. Firstly, we provide a unified treatment of two-sided and one-sided Shampoo, addressing the more complex and widely used former variant, while Xie et al. (2025b) only studied the latter. Secondly, we incorporate practical components like momentum and decoupled weight decay which Xie et al. (2025b) omits. Thirdly, we utilize a stochastic expectation-based assumption, which is weaker than the deterministic condition $-\Sigma_L \preceq \mathbf{G}\mathbf{G}^T - \nabla f(\mathbf{X})\nabla f(\mathbf{X})^T \preceq \Sigma_L$ used in (Xie et al., 2025b). Finally, while both analyses achieve the $\mathcal{O}(\frac{C}{K^{1/4}})$ convergence rate, our constant C in (6) is simpler and matches that in the optimal convergence rate of SGD under the standard Euclidean smoothness assumption.

3. Proof of the Theorem

We present our techniques to address the challenging two-sided preconditioning in Sections 3.1 and 3.2, which constitute the primary technical contribution of this paper relative to existing literature. Section 3.3 then outlines the proof sketch of Theorem 2.1.

3.1. Bounding the Nuclear Norm of Gradient by Schatten-p Holder Inequality

In the analysis of AdamW, Li et al. (2025a) employs the following inequality

$$\left(\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_1\right]\right)^2 \leq\left(\sum_{k=1}^{K}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]\right)\left(\sum_{k=1}^{K}\sum_{i=1}^{d}\mathbb{E}\left[\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}\right]\right) \tag{8}$$

by Holder's inequality for some $\widetilde{\mathbf{v}}^k$ to approximate the second moment, where the superscript denotes the k-th iteration, and the subscript indicates the i-th element of a vector. To handle the more complex matrix case with two-sided preconditioning, we instead utilize the Schatten-p Holder inequality.

Definition 3.1. (Bhatia, 1997, (IV.31)) For any matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$ with $r=\min\{m,n\}$, the Schatten-p norm is defined as

$$\|\mathbf{A}\|_{S_p}=\begin{cases}\left(\sum_{i=1}^{r}\left(\sigma_i(\mathbf{A})\right)^p\right)^{1/p}, & p<\infty,\\ \sigma_1(\mathbf{A}), & p=\infty.\end{cases}$$

The following lemma extends Holder's inequality to the Schatten-p norm.

Lemma 3.2. *(Bhatia, 1997, Corollary IV.2.6, Exercise IV.2.7,(IV.33)) Let $\mathbf{A}_i\in\mathbb{R}^{m\times m}$ and p_i be positive real number $(i=1,2,\cdots,t)$ such that $\sum_{i=1}^{t}\frac{1}{p_i}=1$. Then*

$$\left\|\Pi_{i=1}^{t}\mathbf{A}_i\right\|_{S_1}\leq\Pi_{i=1}^{t}\|\mathbf{A}_i\|_{S_{p_i}}.$$

Note that we do not require $\mathbf{A}_i$ to be symmetric positive semidefinite. This lemma also holds for non-square matrices, because we can always obtain a square matrix by appending zero columns to the right and zero rows to the bottom of the original matrix, and these appended zeros do not affect the Schatten-p norm and matrix multiplication.

The following lemma is an analogue of (8).

Lemma 3.3. *Let $\frac{1}{p}+\frac{1}{q}=1$. Then for Algorithm 1, we have*

$$\sum_{k=1}^{K}\mathbb{E}[\|\nabla f(\mathbf{X}_k)\|_*]\leq\left(\sum_{k=1}^{K}\mathbb{E}\left[\mathrm{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)+\mathrm{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right]\right)^{1/2}\times\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\right]\right)^{1/2}.$$

Proof. From the definitions of the Schatten-p norm and nuclear norm, we have

$$\begin{aligned}&\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]=\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_{S_1}\right]\\ &=\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}\right\|_{S_1}\right]\\ &\overset{(1)}{\leq}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\right\|_{S_{2p}}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_{S_2}\left\|\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}\right\|_{S_{2q}}\right]\\ &\overset{(2)}{\leq}\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\right\|_{S_{2p}}^{2p}\right]\right)^{1/2p}\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}\right\|_{S_{2q}}^{2q}\right]\right)^{1/2q}\\ &\quad\times\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_{S_2}^2\right]\right)^{1/2},\end{aligned}$$

where we use $\frac{1}{2p}+\frac{1}{2q}+\frac{1}{2}=1$ and Lemma 3.2 of the Schatten-p Holder inequality in $\overset{(1)}{\leq}$, and Holder's inequality in $\overset{(2)}{\leq}$. Denote $\mathbf{A}=\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\in\mathbb{R}^{m\times n}$ with $r=\min\{m,n\}$. Then from Definition 3.1 of the Schatten-p norm, we have

$$\|\mathbf{A}\|_{S_2}^2=\sum_{i=1}^{r}\left(\sigma_i(\mathbf{A})\right)^2=\mathrm{tr}(\mathbf{A}^T\mathbf{A})=\|\mathbf{A}\|_F^2.$$

When $p,q<+\infty$, from Definition 3.1 and the fact that $\mathbf{L}_{k,\varepsilon}$ is symmetric and positive definite, we have

$$\left\|\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\right\|_{S_{2p}}^{2p}=\left(\sum_{i=1}^{m}\left(\sigma_i\left(\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\right)\right)^{2p}\right)^{\frac{1}{2p}\cdot 2p}=\mathrm{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right), \tag{9}$$

where we use $\sigma_i(\mathbf{B}^{\frac{1}{p}})=(\sigma_i(\mathbf{B}))^{\frac{1}{p}}$ for any symmetric positive definite matrix $\mathbf{B}$. The derivation for $\mathbf{R}_{k,\varepsilon}$ follows a similar approach. Using $a^{\frac{1}{2p}}b^{\frac{1}{2q}}\leq(a+b)^{\frac{1}{2p}+\frac{1}{2q}}=(a+b)^{1/2}$ for positive a,b, we have the conclusion.

When $p=+\infty$ and $q=1$, we have $\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}=\mathbf{I}_m$, $\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}=\mathbf{I}_m$, and $\left\|\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\right\|_{S_{2p}}=1$. So we have

$$\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\leq\left(\sum_{k=1}^{K}\mathbb{E}\left[\mathrm{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right]\right)^{1/2}\times\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\right]\right)^{1/2}.$$

The case when $q=+\infty$ and $p=1$ is similar. □

The essence of Lemma 3.3 lies in transforming $\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}$ and $\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4p}}$ into the more tractable $\mathbf{L}_{k,\varepsilon}^{1/2}$ and $\mathbf{R}_{k,\varepsilon}^{1/2}$, respectively, which can be handled by the following lemma.

Lemma 3.4. *Suppose that Assumptions 2-3 hold. Let $\beta < 1$. Then for Algorithm 1, we have*

$$\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)\right]$$

$$\leq K\operatorname{tr}\left(\Sigma_L^{1/2}\right)+Km\sqrt{\varepsilon}+\frac{2}{\sqrt{1-\beta}}\sum_{t=1}^{K}\mathbb{E}_{\mathcal{F}_{t-1}}[\|\nabla f(\mathbf{X}_t)\|_*],$$

$$\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right]$$

$$\leq K\operatorname{tr}\left(\Sigma_R^{1/2}\right)+Kn\sqrt{\varepsilon}+\frac{2}{\sqrt{1-\beta}}\sum_{t=1}^{K}\mathbb{E}_{\mathcal{F}_{t-1}}[\|\nabla f(\mathbf{X}_t)\|_*].$$

Proof. From the recursion of $\mathbf{L}_{k-t}$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}_{k-t}}\left[\operatorname{tr}\left(\left(\beta^t\mathbf{L}_{k-t}+(1-\beta^t)\Sigma_L+\varepsilon\mathbf{I}_m\right)^{1/2}\right)\right]\\
=&\mathbb{E}_{\mathcal{F}_{k-t-1}}\Big[\mathbb{E}_{k-t}\Big[\operatorname{tr}\Big(\big(\beta^{t+1}\mathbf{L}_{k-t-1}+\beta^t(1-\beta)\mathbf{G}_{k-t}\mathbf{G}_{k-t}^T\\
&+(1-\beta^t)\Sigma_L+\varepsilon\mathbf{I}_m\big)^{1/2}\Big)\,\Big|\mathcal{F}_{k-t-1}\Big]\Big]\\
\overset{(1)}{\leq}&\mathbb{E}_{\mathcal{F}_{k-t-1}}\left[\operatorname{tr}\left(\left(\beta^{t+1}\mathbf{L}_{k-t-1}+(1-\beta^t)\Sigma_L+\varepsilon\mathbf{I}_m\right.\right.\right.\\
&\left.\left.\left.+\beta^t(1-\beta)\mathbb{E}_{k-t}\left[\mathbf{G}_{k-t}\mathbf{G}_{k-t}^T\middle|\mathcal{F}_{k-t-1}\right]\right)^{1/2}\right)\right]\\
\overset{(2)}{\leq}&\mathbb{E}_{\mathcal{F}_{k-t-1}}\left[\operatorname{tr}\left(\left(\beta^{t+1}\mathbf{L}_{k-t-1}+(1-\beta^t)\Sigma_L+\varepsilon\mathbf{I}_m\right.\right.\right.\\
&\left.\left.\left.+\beta^t(1-\beta)\nabla f(\mathbf{X}_{k-t})\nabla f(\mathbf{X}_{k-t})^T+\beta^t(1-\beta)\Sigma_L\right)^{1/2}\right)\right]\\
=&\mathbb{E}_{\mathcal{F}_{k-t-1}}\left[\operatorname{tr}\left(\left(\beta^{t+1}\mathbf{L}_{k-t-1}+(1-\beta^{t+1})\Sigma_L+\varepsilon\mathbf{I}_m\right.\right.\right.\\
&\left.\left.\left.+\beta^t(1-\beta)\nabla f(\mathbf{X}_{k-t})\nabla f(\mathbf{X}_{k-t})^T\right)^{1/2}\right)\right]\\
\overset{(3)}{\leq}&\mathbb{E}_{\mathcal{F}_{k-t-1}}\left[\operatorname{tr}\left(\left(\beta^{t+1}\mathbf{L}_{k-t-1}+(1-\beta^{t+1})\Sigma_L+\varepsilon\mathbf{I}_m\right)^{1/2}\right)\right.\\
&\left.+\sqrt{\beta^t(1-\beta)}\operatorname{tr}\left(\left(\nabla f(\mathbf{X}_{k-t})\nabla f(\mathbf{X}_{k-t})^T\right)^{1/2}\right)\right]\\
=&\mathbb{E}_{\mathcal{F}_{k-t-1}}\left[\operatorname{tr}\left(\left(\beta^{t+1}\mathbf{L}_{k-t-1}+(1-\beta^{t+1})\Sigma_L+\varepsilon\mathbf{I}_m\right)^{1/2}\right)\right.\\
&\left.+\sqrt{\beta^t(1-\beta)}\left\|\nabla f(\mathbf{X}_{k-t})\right\|_*\right],
\end{aligned}$$

where we use the concavity of $\mathbf{X}^{1/2}$ presented in Lemma A.5 in $\overset{(1)}{\leq}$, Assumptions 2-3 and the monotonicity of $\mathbf{X}^{1/2}$ presented in Lemma A.4 in $\overset{(2)}{\leq}$, and the property $\operatorname{tr}\left((\mathbf{X}+\mathbf{Y})^{1/2}\right) \leq \operatorname{tr}(\mathbf{X}^{1/2})+\operatorname{tr}(\mathbf{Y}^{1/2})$ presented in Lemma A.3 in $\overset{(3)}{\leq}$. Applying the above inequality recursively for $t=0,1,2,\cdots,k-1$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}_k}\left[\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)\right]\\
&\leq\operatorname{tr}\left(\left(\beta^k\mathbf{L}_0+(1-\beta^k)\Sigma_L+\varepsilon\mathbf{I}_m\right)^{1/2}\right)\\
&\quad+\sqrt{1-\beta}\sum_{t=0}^{k-1}\sqrt{\beta^t}\mathbb{E}_{\mathcal{F}_{k-t-1}}\left[\|\nabla f(\mathbf{X}_{k-t})\|_*\right]\\
&\overset{(4)}{\leq}\operatorname{tr}\left(\Sigma_L^{1/2}+\sqrt{\varepsilon}\mathbf{I}_m\right)+\sqrt{1-\beta}\sum_{t=1}^{k}\sqrt{\beta^{k-t}}\mathbb{E}_{\mathcal{F}_{t-1}}[\|\nabla f(\mathbf{X}_t)\|_*]\\
&=\operatorname{tr}\left(\Sigma_L^{1/2}\right)+m\sqrt{\varepsilon}+\sqrt{1-\beta}\sum_{t=1}^{k}\sqrt{\beta^{k-t}}\mathbb{E}_{\mathcal{F}_{t-1}}[\|\nabla f(\mathbf{X}_t)\|_*],
\end{aligned}$$

where we use $\mathbf{L}_0=\mathbf{0}$ and $\operatorname{tr}\left((\mathbf{X}+\mathbf{Y})^{1/2}\right) \leq \operatorname{tr}(\mathbf{X}^{1/2})+\operatorname{tr}(\mathbf{Y}^{1/2})$ in $\overset{(4)}{\leq}$. Summing over $k=1,2,\cdots,K$, we have

$$\begin{aligned}
\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)\right] &\leq K\operatorname{tr}\left(\Sigma_L^{1/2}\right)+Km\sqrt{\varepsilon}\\
&\quad+\underbrace{\sqrt{1-\beta}\sum_{k=1}^{K}\sum_{t=1}^{k}\sqrt{\beta^{k-t}}\mathbb{E}_{\mathcal{F}_{t-1}}\left[\|\nabla f(\mathbf{X}_t)\|_*\right]}_{\text{term (a)}}
\end{aligned}$$

and term (a) can be addressed as follows

$$\begin{aligned}
\text{term (a)} =&\sqrt{1-\beta}\sum_{t=1}^{K}\sum_{k=t}^{K}\sqrt{\beta^{k-t}}\mathbb{E}_{\mathcal{F}_{t-1}}\left[\|\nabla f(\mathbf{X}_t)\|_*\right]\\
\leq&\frac{\sqrt{1-\beta}}{1-\sqrt{\beta}}\sum_{t=1}^{K}\mathbb{E}_{\mathcal{F}_{t-1}}\left[\|\nabla f(\mathbf{X}_t)\|_*\right]\\
\leq&\frac{2}{\sqrt{1-\beta}}\sum_{t=1}^{K}\mathbb{E}_{\mathcal{F}_{t-1}}\left[\|\nabla f(\mathbf{X}_t)\|_*\right].
\end{aligned}$$

Similarly, we also have the inequality for $\mathbf{R}_{k,\varepsilon}^{1/2}$. □

Combining Lemmas 3.3 and 3.4, we finally have

$$\begin{aligned}
\sum_{k=1}^{K}\mathbb{E}[\|\nabla f(\mathbf{X}_k)\|_*] \leq&\left(KC+\frac{4}{\sqrt{1-\beta}}\sum_{t=1}^{K}\mathbb{E}[\|\nabla f(\mathbf{X}_t)\|_*]\right)^{1/2}\\
&\times\underbrace{\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\right]\right)^{1/2}}_{\text{term (b)}},
\end{aligned}\tag{10}$$

where $C=\operatorname{tr}\left(\Sigma_L^{1/2}\right)+\operatorname{tr}\left(\Sigma_R^{1/2}\right)+(m+n)\sqrt{\varepsilon}$. So we only need to bound term (b).

3.2. Bounding the Spectral Norm of Update by Matrix Cauchy-Schwarz Inequality

In the analysis of AdamW (Li et al., 2025a), we can bound the update $\frac{|\mathbf{m}_i^k|}{\sqrt{\mathbf{v}_i^k}}$ coordinately. However, the matrix case is not as simple, especially with two-sided preconditioning. To address this challenge, we use the following matrix Cauchy-Schwarz inequality.

Lemma 3.5. *(Bhatia, 1997, Corollary IX.5.3) For $\mathbf{M}\in\mathbb{R}^{m\times n}$ and symmetric positive definite matrices $\mathbf{L}\in\mathbb{R}^{m\times m}$ and $\mathbf{R}\in\mathbb{R}^{n\times n}$, $0\leq\alpha\leq 1$, we have*

$$\|\mathbf{L}^{\alpha}\mathbf{M}\mathbf{R}^{1-\alpha}\|_{op}\leq\|\mathbf{L}\mathbf{M}\|_{op}^{\alpha}\|\mathbf{M}\mathbf{R}\|_{op}^{1-\alpha}.$$

Based on the above lemma, we can bound the update in Algorithm 1 measured by spectral norm.

Lemma 3.6. *Let $\theta \le \beta \le \sqrt{\theta} < 1$ and $\frac{1}{p}+\frac{1}{q}=1$. Then for Algorithm 1, we have*

$$\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\|_{op}\le 2.$$

Proof. From Lemma 3.5 and $\frac{1}{p}+\frac{1}{q}=1$, we have

$$\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\|_{op}\le\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2}}\mathbf{M}_k\right\|_{op}^{\frac{1}{p}}\left\|\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2}}\right\|_{op}^{\frac{1}{q}}.$$

So we only need to prove

$$\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2}}\mathbf{M}_k\right\|_{op}\le 2\quad\text{and}\quad\left\|\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2}}\right\|_{op}\le 2.$$

The two inequalities are analogous, so we prove only the first. Since $\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2}}\mathbf{M}_k\right\|_{op}^2=\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2}}\mathbf{M}_k\mathbf{M}_k^T\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2}}\right\|_{op}$, we only need to prove

$$\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2}}\mathbf{M}_k\mathbf{M}_k^T\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2}}\preceq 4\mathbf{I},$$

which is equivalent to

$$\mathbf{M}_k\mathbf{M}_k^T\preceq 4\mathbf{L}_{k,\varepsilon}\text{ and }\mathbf{y}^T\mathbf{M}_k\mathbf{M}_k^T\mathbf{y}\le 4\mathbf{y}^T\mathbf{L}_{k,\varepsilon}\mathbf{y},\forall\mathbf{y}\in\mathbb{R}^m.$$

From the recursions of $\mathbf{M}_k$ and $\mathbf{L}_k$, we have

$$\begin{aligned}\mathbf{y}^T\mathbf{M}_k&=(1-\theta)\sum_{t=1}^k\theta^{k-t}\mathbf{y}^T\mathbf{G}_t,\\ \mathbf{y}^T\mathbf{L}_k\mathbf{y}&=(1-\beta)\sum_{t=1}^k\beta^{k-t}\mathbf{y}^T\mathbf{G}_t\mathbf{G}_t^T\mathbf{y}\\&=(1-\beta)\sum_{t=1}^k\beta^{k-t}\left\|\mathbf{y}^T\mathbf{G}_t\right\|^2.\end{aligned}$$

From Holder's inequality, we have

$$\begin{aligned}\mathbf{y}^T\mathbf{M}_k\mathbf{M}_k^T\mathbf{y}&=(1-\theta)^2\left\|\sum_{t=1}^k\theta^{k-t}\mathbf{y}^T\mathbf{G}_t\right\|^2\\ &\le(1-\theta)^2\left(\sum_{t=1}^k\theta^{k-t}\left\|\mathbf{y}^T\mathbf{G}_t\right\|\right)^2\\ &\le(1-\theta)^2\left(\sum_{t=1}^k\beta^{k-t}\left\|\mathbf{y}^T\mathbf{G}_t\right\|^2\right)\left(\sum_{t=1}^k\left(\frac{\theta^2}{\beta}\right)^{k-t}\right)\\ &=\frac{(1-\theta)^2}{1-\beta}\mathbf{y}^T\mathbf{L}_k\mathbf{y}\left(\sum_{t=1}^k\left(\frac{\theta^2}{\beta}\right)^{k-t}\right)\\ &\le\frac{(1-\theta)^2}{1-\beta}\frac{1}{1-\frac{\theta^2}{\beta}}\mathbf{y}^T\mathbf{L}_k\mathbf{y}\overset{(1)}{\le}\frac{(1-\theta)^2}{(1-\beta)^2}\mathbf{y}^T\mathbf{L}_k\mathbf{y}\\ &\overset{(2)}{\le}\frac{(1-\sqrt{\theta})^2(1+\sqrt{\theta})^2}{(1-\sqrt{\theta})^2}\mathbf{y}^T\mathbf{L}_k\mathbf{y}\le 4\mathbf{y}^T\mathbf{L}_k\mathbf{y}\le 4\mathbf{y}^T\mathbf{L}_{k,\varepsilon}\mathbf{y},\end{aligned}$$

where we use $\theta\le\beta$ in $\overset{(1)}{\le}$ and $\beta\le\sqrt{\theta}$ in $\overset{(2)}{\le}$. □

By leveraging Lemma 3.6 and the distinct properties of decoupled weight decay, we ultimately establish that

$$\lambda\|\mathbf{X}_k\|_{op}\le\frac{3\sqrt{\nu}}{K^{1/4}}\tag{11}$$

holds for some constant ν and all $k=1,2,\cdots,K$ under appropriate parameter choices. The complete derivation is provided in Lemma C.1.

3.3. Proof Sketch of Theorem 2.1

Building upon the supporting lemmas in Sections 3.1 and 3.2, we can prove the main theorem following the framework in (Li et al., 2025a). We briefly outline the proof sketch in this section. From the Lipschitz smoothness and the update of $\mathbf{X}_{k+1}$, we have

$$\begin{aligned}f(\mathbf{X}_{k+1})-f(\mathbf{X}_k)&\le\langle\nabla f(\mathbf{X}_k),\mathbf{X}_{k+1}-\mathbf{X}_k\rangle+\frac{L}{2}\|\mathbf{X}_{k+1}-\mathbf{X}_k\|_F^2\\ &=-\eta\left\langle\nabla f(\mathbf{X}_k),\lambda\mathbf{X}_k+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\rangle\\ &\quad+\frac{L\eta^2}{2}\left\|\lambda\mathbf{X}_k+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\|_F^2\\ &=\underbrace{-\eta\left\langle\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}},\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\rangle}_{\text{term (c)}}\\ &\quad+\underbrace{\frac{L\eta^2}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\left(\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2}_{\text{term (d)}}.\end{aligned}$$

Decompose term (c) into

$$\begin{aligned}&-\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\\ &-\frac{\eta}{2}\left\|\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\\ &\underbrace{+\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}-\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}\right\|_F^2}_{\text{term (e)}}\end{aligned}$$

and relax term (e) to

$$\underbrace{\eta\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}(\nabla f(\mathbf{X}_k)-\mathbf{M}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2}_{\text{term (f)}}+\eta\lambda^2\underbrace{\left\|\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}\right\|_F^2}_{\text{term (g)}}.$$

From condition (3), we can relax terms (f) and (d) as follow

$$\text{term (f)}\le\frac{\eta}{\sqrt{\hat{\varepsilon}}}\left\|\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right\|_F^2,$$

$$\text{term (d)}\le\frac{L\eta^2}{2\sqrt{\hat{\varepsilon}}}\left\|\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2.$$

The handling of term (g) requires a high degree of skill:

$$\begin{aligned}\text{term (g)} &= \operatorname{tr}\left(\underbrace{\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{2q}}\mathbf{X}_k^T\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}}_{\mathbf{A}}\right) \overset{(1)}{=} \|\mathbf{A}\|_{S_1} \\ &\overset{(2)}{\leq} \left\|\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\right\|_{S_{2p}} \left\|\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{2q}}\mathbf{X}_k^T\right\|_{S_q} \left\|\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\right\|_{S_{2p}} \\ &\overset{(3)}{=} \left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)\right)^{\frac{1}{p}} \underbrace{\left\|\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{2q}}\mathbf{X}_k^T\right\|_{S_q}}_{\text{term (h)}},\end{aligned} \tag{12}$$

where we use Definition 3.1 of the Schatten-p norm and the fact that $\mathbf{A}$ is symmetric and positive definite in $\overset{(1)}{=}$, Lemma 3.2 of the Schatten-p Holder inequality in $\overset{(2)}{\leq}$, and (9) in $\overset{(3)}{=}$. For term (h), we have

$$\begin{aligned}&\text{term (h)} \\ =&\left(\sum_{i=1}^m \sigma_i\left(\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{2q}}\mathbf{X}^T\right)^q\right)^{\frac{1}{q}} \overset{(4)}{\leq} \left(\sum_{i=1}^n\left(\|\mathbf{X}_k\|_{op}^2\sigma_i\left(\mathbf{R}_{k,\varepsilon}^{\frac{1}{2q}}\right)\right)^q\right)^{\frac{1}{q}} \\ \overset{(5)}{=}&\left(\|\mathbf{X}_k\|_{op}^{2q}\sum_{i=1}^n\sigma_i\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right)^{\frac{1}{q}} = \|\mathbf{X}_k\|_{op}^2\left(\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right)^{\frac{1}{q}},\end{aligned}$$

where we use the properties $\sigma_i(\mathbf{AB}) \leq \sigma_i(\mathbf{A})\|\mathbf{B}\|_{op}$ and $\sigma_i(\mathbf{AB}) \leq \|\mathbf{A}\|_{op}\sigma_i(\mathbf{B})$ of singular values in $\overset{(4)}{\leq}$ and $\sigma_i(\mathbf{B}^{\frac{1}{q}}) = (\sigma_i(\mathbf{B}))^{\frac{1}{q}}$ for any symmetric positive definite matrix $\mathbf{B}$ in $\overset{(5)}{=}$. Plugging into (12), we have

$$\begin{aligned}\text{term (g)} \leq&\|\mathbf{X}_k\|_{op}^2\left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)\right)^{\frac{1}{p}}\left(\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right)^{\frac{1}{q}} \\ \overset{(6)}{\leq}&\|\mathbf{X}_k\|_{op}^2\left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)+\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right) \\ \overset{(7)}{\leq}&\frac{9\nu}{\lambda^2K^{1/2}}\left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)+\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right),\end{aligned}$$

where we use $a^{\frac{1}{p}}b^{\frac{1}{q}} \overset{(6)}{\leq} (a+b)^{\frac{1}{p}+\frac{1}{q}} = a+b$ for positive a, b in $\overset{(6)}{\leq}$, and (11) in $\overset{(7)}{\leq}$. Combining the above results and letting $\eta \leq \frac{\sqrt{\hat{\varepsilon}}}{2L}$, we have

$$\begin{aligned}f(\mathbf{X}_{k+1}) - f(\mathbf{X}_k) \leq& -\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 \\ &-\frac{\eta}{4}\left\|\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}} + \mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 \\ &+\frac{\eta}{\sqrt{\hat{\varepsilon}}}\|\nabla f(\mathbf{X}_k)-\mathbf{M}_k\|_F^2+\frac{9\eta\nu}{K^{1/2}}\left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)+\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right).\end{aligned} \tag{13}$$

Employing standard techniques in the analysis of momentum SGD, we can build a recursion (Lemma C.2) as follows

$$\begin{aligned}&\mathbb{E}_{\mathcal{F}_k}\left[\|\nabla f(\mathbf{X}_k) - \mathbf{M}_k\|_F^2\right] \\ \leq&\mathbb{E}_{\mathcal{F}_{k-1}}\left[\theta\|\nabla f(\mathbf{X}_{k-1}) - \mathbf{M}_{k-1}\|_F^2 + (1-\theta)^2\sigma^2\right. \\ &\left.+\frac{L^2\eta^2}{(1-\theta)\sqrt{\hat{\varepsilon}}}\left\|\lambda\mathbf{L}_{k-1,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_{k-1}\mathbf{R}_{k-1,\varepsilon}^{\frac{1}{4q}}+\mathbf{L}_{k-1,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_{k-1}\mathbf{R}_{k-1,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\right].\end{aligned} \tag{14}$$

Multiplying both sides of (14) by $\frac{\eta}{\sqrt{\hat{\varepsilon}}(1-\theta)}$, adding it to (13), letting $\eta^2 \leq \frac{\hat{\varepsilon}(1-\theta)^2}{4L^2}$, and arranging the terms, we have

$$\begin{aligned}\phi_{k+1} \leq& \phi_k - \frac{\eta}{2}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\right] \\ &+\frac{9\eta\nu}{K^{1/2}}\mathbb{E}_{\mathcal{F}_k}\left[\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)+\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right]+\frac{\eta(1-\theta)}{\sqrt{\hat{\varepsilon}}}\sigma^2,\end{aligned} \tag{15}$$

where $\phi_{k+1}=\mathbb{E}_{\mathcal{F}_k}\Big[f(\mathbf{X}_{k+1})-f^*+\frac{\eta\theta}{\sqrt{\hat{\varepsilon}}(1-\theta)}\|\nabla f(\mathbf{X}_k)-\mathbf{M}_k\|_F^2 + \frac{\eta}{4}\left\|\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}} + \mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\Big]$. Summing (15) over $k = 1, 2, \cdots, K$ and setting the parameters properly, we have

$$\begin{aligned}&\sum_{k=1}^K\mathbb{E}_{\mathcal{F}_k}\left[\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\right] \\ \leq&\frac{18\nu}{K^{1/2}}\sum_{k=1}^K\mathbb{E}_{\mathcal{F}_k}\left[\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)+\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right]+C',\end{aligned}$$

where $C' = 10\sqrt{\frac{K\hat{\sigma}^2L(f(\mathbf{X}_1)-f^*)}{\hat{\varepsilon}}}$. Plugging into (10) and using Lemma 3.4, we have

$$\begin{aligned}\sum_{k=1}^K\mathbb{E}[\|\nabla f(\mathbf{X}_k)\|_*] \leq& \left(KC+\frac{4}{\sqrt{1-\beta}}\sum_{t=1}^K\mathbb{E}[\|\nabla f(\mathbf{X}_t)\|_*]\right)^{\frac{1}{2}} \times \\ &\left(\frac{18\nu}{K^{1/2}}\frac{4}{\sqrt{1-\beta}}\sum_{t=1}^K\mathbb{E}[\|\nabla f(\mathbf{X}_t)\|_*]+18\nu K^{1/2}C+C'\right)^{\frac{1}{2}},\end{aligned}$$

Under proper parameter settings such that $\frac{288\nu}{K^{1/2}(1-\beta)} \leq \frac{1}{2}$, we finally have

$$\left(\sum_{k=1}^K\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2 \leq A\sum_{k=1}^K\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]+B$$

for some constant A and B. Solving this inequality, we have the conclusion. The derivation above is conducted for the case of $p, q < +\infty$. The conclusion remains valid when either p or q is infinite.

4. Conclusion

This paper studies the convergence properties of AdamW-style Shampoo with both one-sided and two-sided preconditioning. We establish the convergence rate $\frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \leq \mathcal{O}(\frac{\sqrt{m+n}C}{K^{1/4}})$ measured by nuclear norm. It can be considered to be analogous to the optimal $\frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_F\right] \leq \mathcal{O}(\frac{C}{K^{1/4}})$ 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.

Impact Statement

This study is primarily concerned with theoretical analysis and it does not yield direct negative societal impacts.

References

An, K., Liu, Y., Pan, R., Ren, Y., Ma, S., Goldfarb, D., and Zhang, T. ASGO: Adaptive structured gradient optimization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Anil, R., Gupta, V., Koren, T., Regan, K., and Singer, Y. Scalable second order optimization for deep learning. *arXiv:2002.09018*, 2020.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2023.

Bhatia, R. (ed.). *Matrix Analysis*. Springer, 1997.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Chen, L., Li, J., and Liu, Q. Muon optimizes under spectral norm constraints. *arXiv:2506.15054*, 2025.

Défossez, A., Bottou, L., Bach, F., and Usunier, N. A simple convergence proof of Adam and AdaGrad. *Transactions on Machine Learning Research*, 2022.

Dozat, T. Incorporating Nesterov momentum into Adam. In *Workshop of the International Conference on Learning Representations (ICLR Workshop)*, 2016.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

Eschenhagen, R., Defazio, A., Lee, T.-H., Turner, R. E., and Shi, H.-J. M. Purifying Shampoo: Investigating Shampoo's heuristics by decomposing its preconditioner. *arXiv:2506.03595*, 2025.

Feinberg, V., Chen, X., Sun, Y. J., Anil, R., and Hazan, E. Sketchy: Memory-efficient adaptive regularization with frequent directions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Gokaslan, A., Cohen, V., Pavlick, E., and Tellex, S. OpenWebText corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Gupta, V., Koren, T., and Singer, Y. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning (ICML)*, 2018.

Hong, Y. and Lin, J. On convergence of Adam for stochastic optimization under relaxed assumptions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Jordan, K., Jin, Y., Boza, V., You, J., Cesista, F., Newhouse, L., and Bernstein, J. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024.

Kasimbeg, P., Schneider, F., Eschenhagen, R., Bae, J., Sastry, C. S., Saroufim, M., Feng, B., Wright, L., Yang, E. Z., Nado, Z., Medapati, S., Hennig, P., Rabbat, M., and Dahl, G. E. Accelerating neural network training: An analysis of the AlgoPerf competition. In *International Conference on Learning Representations (ICLR)*, 2025.

Kingma, D. P. and Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Li, H., Rakhlin, A., and Jadbabaie, A. Convergence of Adam under relaxed assumptions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Li, H., Dong, Y., and Lin, Z. On the $O(\frac{\sqrt{d}}{K^{1/4}})$ convergence rate of AdamW measured by ℓ_1 norm. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025a.

Li, H., Dong, Y., and Lin, Z. On the $O(\frac{\sqrt{d}}{T^{1/4}})$ convergence rate of RMSProp and its momentum extension measured by ℓ_1 norm. *Journal of Machine Learning Research*, 26 (131):1–25, 2025b.

Li, J. and Hong, M. A note on the convergence of Muon. *arXiv:2502.02900*, 2025.

Lin, W., Lowe, S. C., Dangel, F., Eschenhagen, R., Xu, Z., and Grosse, R. B. Understanding and improving Shampoo and SOAP via Kullback-Leibler minimization. *arXiv: 2509.03378*, 2025.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

McMahan, H. B. and Streeter, M. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory (COLT)*, 2010.

Morwani, D., Shapira, I., Vyas, N., Malach, E., Kakade, S., and Janson, L. A new perspective on Shampoo's preconditioner. In *International Conference on Learning Representations (ICLR)*, 2025.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Sato, N., Naganuma, H., and Iiduka, H. Convergence bound and critical batch size of Muon optimizer. *arXiv:2507.01598*, 2025.

Shen, W., Huang, R., Huang, M., Shen, C., and Zhang, J. On the convergence analysis of Muon. *arXiv:2505.23737*, 2025.

Shi, H.-J. M., Lee, T.-H., Iwasaki, S., Gallego-Posada, J., Li, Z., Rangadurai, K., Mudigere, D., and Rabbat, M. A distributed data-parallel pytorch implementation of the distributed Shampoo optimizer for training neural networks at-scale. *arXiv:2309.06497*, 2023.

Shi, N., Li, D., Hong, M., and Sun, R. RMSProp converges with proper hyper-parameter. In *International Conference on Learning Representations (ICLR)*, 2020.

Tieleman, T. and Hinton, G. Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude. In *COURSERA: Neural Networks for Machine Learning*, 2012.

Vyas, N., Morwani, D., Zhao, R., Kwun, M., Shapira, I., Brandfonbrener, D., Janson, L., and Kakade, S. SOAP: Improving and stabilizing Shampoo using Adam. In *International Conference on Learning Representations (ICLR)*, 2025.

Wen, K., Hall, D., Ma, T., and Liang, P. Fantastic pretraining optimizers and where to find them. *arXiv: 2509.02046*, 2025.

Xie, S., Wang, T., reddi, S., Kumar, S., and Li, Z. Structured preconditioners in adaptive optimization: A unified analysis. In *International Conference on Machine Learning (ICML)*, 2025a.

Xie, S., Wang, T., Wu, B., and Li, Z. A tale of two geometries: Adaptive optimizers and non-Euclidean descent. *arXiv:2511.20584*, 2025b.

Zhang, Y., Chen, C., Shi, N., Sun, R., and Luo, Z. Adam can converge without any modification on update rules. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

A. Basic Properties in Matrix Analysis

We first introduce some basic properties in matrix analysis. For matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$, it holds that

$$\begin{aligned}\langle \mathbf{A}, \mathbf{B}\rangle &= \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{A}_{i,j}\mathbf{B}_{i,j} = \operatorname{tr}(\mathbf{A}^T\mathbf{B}),\\ \|\mathbf{A}\|_F^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{A}_{i,j}^2 = \operatorname{tr}(\mathbf{A}^T\mathbf{A}),\\ \operatorname{tr}(\mathbf{A}^T\mathbf{B}) &= \operatorname{tr}(\mathbf{B}\mathbf{A}^T).\end{aligned}$$

For symmetric positive definite matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times m}$ with $\mathbf{X} \preceq \mathbf{Y}$, it holds that

$$\operatorname{tr}(\mathbf{X}) \le \operatorname{tr}(\mathbf{Y}), \quad \mathbf{A}^T\mathbf{X}\mathbf{A} \preceq \mathbf{A}^T\mathbf{Y}\mathbf{A}, \quad \mathbf{X}^{-1} \succeq \mathbf{Y}^{-1}. \tag{16}$$

Fact A.1. *For symmetric positive semidefinite matrix $\mathbf{X} \in \mathbb{R}^{m\times m}$, the singular values coincide with the eigenvalues and thus $\operatorname{tr}(\mathbf{X}) = \sum_{i=1}^{m}\sigma_i(\mathbf{X})$.*

Definition A.2. For symmetric positive semidefinite matrix $\mathbf{X} \in \mathbb{R}^{m\times m}$, Let $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ be its eigenvalue decomposition with $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m)$, then $\mathbf{X}^{\frac{1}{p}}$ is defined to be $\mathbf{U}\operatorname{diag}(\lambda_1^{1/p}, \lambda_2^{1/p}, \cdots, \lambda_m^{1/p})\mathbf{U}^T$.

Lemma A.3. *(An et al., 2025, Lemma 3) For symmetric positive definite matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times m}$, it holds that*

$$\operatorname{tr}\left((\mathbf{X}+\mathbf{Y})^{1/2}\right) \le \operatorname{tr}\left(\mathbf{X}^{1/2}+\mathbf{Y}^{1/2}\right).$$

Since x^t with $t \in [0,1]$ is operator monotonic and operator concave on $[0,\infty)$, we have the following two properties (Bhatia, 1997, Theorem V.1.9, Theorem V.2.5):

Lemma A.4. *For symmetric positive definite matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times m}$ with $\mathbf{X} \preceq \mathbf{Y}$, it holds that $\mathbf{X}^t \preceq \mathbf{Y}^t$ with $t \in [0,1]$.*

Lemma A.5. *For symmetric positive definite matrix $\mathbf{X} \in \mathbb{R}^{m\times m}$, it holds that $\mathbb{E}[\mathbf{X}^t] \preceq (\mathbb{E}[\mathbf{X}])^t$ with $t \in [0,1]$.*

From Assumptions 2 and 3, we have

$$\begin{aligned}\mathbb{E}_k\left[\|\mathbf{G}_k - \nabla f(\mathbf{X}_k)\|_F^2 \,|\mathcal{F}_{k-1}\right] =& \mathbb{E}_k\left[\operatorname{tr}\left((\mathbf{G}_k - \nabla f(\mathbf{X}_k))(\mathbf{G}_k - \nabla f(\mathbf{X}_k))^T\right)|\mathcal{F}_{k-1}\right]\\ =& \mathbb{E}_k\left[\operatorname{tr}\left((\mathbf{G}_k - \nabla f(\mathbf{X}_k))^T(\mathbf{G}_k - \nabla f(\mathbf{X}_k))\right)|\mathcal{F}_{k-1}\right]\\ \le& \frac{\operatorname{tr}(\Sigma_L)+\operatorname{tr}(\Sigma_R)}{2} \equiv \sigma^2,\end{aligned} \tag{17}$$

and

$$\begin{aligned}\Sigma_L \succeq& \mathbb{E}_k\left[(\mathbf{G}_k - \nabla f(\mathbf{X}_k))(\mathbf{G}_k - \nabla f(\mathbf{X}_k))^T|\mathcal{F}_{k-1}\right]\\ =& \mathbb{E}_k\left[\mathbf{G}_k\mathbf{G}_k^T + \nabla f(\mathbf{X}_k)\nabla f(\mathbf{X}_k)^T - \mathbf{G}_k\nabla f(\mathbf{X}_k)^T - \nabla f(\mathbf{X}_k)\mathbf{G}_k^T|\mathcal{F}_{k-1}\right]\\ =& \mathbb{E}_k\left[\mathbf{G}_k\mathbf{G}_k^T|\mathcal{F}_{k-1}\right] - \nabla f(\mathbf{X}_k)\nabla f(\mathbf{X}_k)^T.\end{aligned} \tag{18}$$

B. Proof of Theorem 2.1

Proof. As the gradient is L-Lipschitz, we have

$$\begin{aligned}&\mathbb{E}_k\left[f(\mathbf{X}_{k+1})|\mathcal{F}_{k-1}\right] - f(\mathbf{X}_k)\\ \le& \mathbb{E}_k\left[\langle \nabla f(\mathbf{X}_k), \mathbf{X}_{k+1}-\mathbf{X}_k\rangle + \frac{L}{2}\|\mathbf{X}_{k+1}-\mathbf{X}_k\|_F^2|\mathcal{F}_{k-1}\right]\\ =& \mathbb{E}_k\left[-\eta\left\langle \nabla f(\mathbf{X}_k), \lambda\mathbf{X}_k + \mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\rangle + \frac{L\eta^2}{2}\left\|\lambda\mathbf{X}_k + \mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\|_F^2|\mathcal{F}_{k-1}\right]\\ =& \mathbb{E}_k\left[-\eta\left\langle \mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}, \lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}} + \mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\rangle + \frac{L\eta^2}{2}\left\|\lambda\mathbf{X}_k + \mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\|_F^2|\mathcal{F}_{k-1}\right]\end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}_k\left[-\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2-\frac{\eta}{2}\left\|\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\right.\\
&\left.+\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}-\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}\right\|_F^2+\frac{L\eta^2}{2}\left\|\lambda\mathbf{X}_k+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\|_F^2\Big|\mathcal{F}_{k-1}\right]\\
\le&\mathbb{E}_k\left[-\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2-\frac{\eta}{2}\left\|\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\right.\\
&\left.+\eta\underbrace{\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2}_{\text{term (a)}}+\eta\lambda^2\underbrace{\left\|\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}\right\|_F^2}_{\text{term (b)}}+\frac{L\eta^2}{2}\underbrace{\left\|\lambda\mathbf{X}_k+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\|_F^2}_{\text{term (c)}}\Big|\mathcal{F}_{k-1}\right].
\end{aligned}
\tag{19}
$$

From condition (3) and property (16), we have

$$
\mathbf{L}_{k,\varepsilon}^{-1}\preceq\frac{1}{\hat{\varepsilon}}\mathbf{I}_m,\quad \mathbf{R}_{k,\varepsilon}^{-1}\preceq\frac{1}{\hat{\varepsilon}}\mathbf{I}_n.
$$

For term (a), we have

$$
\begin{aligned}
\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2&=\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)^T\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right)\\
&\overset{(1)}{\le}\frac{1}{\hat{\varepsilon}^{\frac{1}{2p}}}\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)^T\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right)\\
&=\frac{1}{\hat{\varepsilon}^{\frac{1}{2p}}}\operatorname{tr}\left(\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)^T\right)\\
&\overset{(2)}{\le}\frac{1}{\hat{\varepsilon}^{\frac{1}{2p}+\frac{1}{2q}}}\operatorname{tr}\left(\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)\left(\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right)^T\right)\\
&\overset{(3)}{=}\frac{1}{\sqrt{\hat{\varepsilon}}}\left\|\nabla f(\mathbf{X}_k)-\mathbf{M}_k\right\|_F^2,
\end{aligned}
$$

where we use Lemma A.4 and (16) in $\overset{(1)}{\le}$ and $\overset{(2)}{\le}$ and $\frac{1}{p}+\frac{1}{q}=1$ in $\overset{(3)}{=}$. For term (b), from the analysis in Section 3.3, we have

$$
\left\|\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}\right\|_F^2\le\frac{9\nu}{\lambda^2K^{1/2}}\left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right)+\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right).
$$

For term (c), similar to the induction for term (a), we have

$$
\begin{aligned}
\left\|\lambda\mathbf{X}_k+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\|_F^2&=\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\underbrace{\left(\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right)}_{\mathbf{A}}\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\\
&=\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\mathbf{A}^T\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{A}\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right)\le\frac{1}{\hat{\varepsilon}^{\frac{1}{2p}}}\operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\mathbf{A}^T\mathbf{A}\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right)\\
&=\frac{1}{\hat{\varepsilon}^{\frac{1}{2p}}}\operatorname{tr}\left(\mathbf{A}\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\mathbf{A}^T\right)\le\frac{1}{\hat{\varepsilon}^{\frac{1}{2p}+\frac{1}{2q}}}\operatorname{tr}\left(\mathbf{A}\mathbf{A}^T\right)=\frac{1}{\sqrt{\hat{\varepsilon}}}\|\mathbf{A}\|_F^2\\
&=\frac{1}{\sqrt{\hat{\varepsilon}}}\left\|\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}}+\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2.
\end{aligned}
\tag{20}
$$

Plugging into (19) and using $\eta \leq \frac{\sqrt{\hat{\varepsilon}}}{2L}$, we have

$$\begin{aligned}
&\mathbb{E}_k\left[f(\mathbf{X}_{k+1})\middle|\mathcal{F}_{k-1}\right] - f(\mathbf{X}_k)\\
\leq&\mathbb{E}_k\left[-\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 - \frac{\eta}{2}\left\|\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}} + \mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 + \frac{\eta}{\sqrt{\hat{\varepsilon}}}\left\|\nabla f(\mathbf{X}_k) - \mathbf{M}_k\right\|_F^2\right.\\
&\left.+\frac{9\eta\nu}{K^{1/2}}\left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right) + \frac{L\eta^2}{2\sqrt{\hat{\varepsilon}}}\left\|\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}} + \mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\middle|\mathcal{F}_{k-1}\right] \qquad (21)\\
\leq&\mathbb{E}_k\left[-\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 - \frac{\eta}{4}\left\|\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}} + \mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\right.\\
&\left.+\frac{\eta}{\sqrt{\hat{\varepsilon}}}\left\|\nabla f(\mathbf{X}_k) - \mathbf{M}_k\right\|_F^2 + \frac{9\eta\nu}{K^{1/2}}\left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right)\middle|\mathcal{F}_{k-1}\right].
\end{aligned}$$

Multiplying both sides of (25) by $\frac{\eta}{\sqrt{\hat{\varepsilon}}(1-\theta)}$, adding it to (21), and arranging the terms, we have

$$\begin{aligned}
&\mathbb{E}_k\left[f(\mathbf{X}_{k+1}) - f^* + \frac{\eta}{4}\left\|\lambda\mathbf{L}_{k,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_k\mathbf{R}_{k,\varepsilon}^{\frac{1}{4q}} + \mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 + \frac{\eta\theta}{\sqrt{\hat{\varepsilon}}(1-\theta)}\left\|\nabla f(\mathbf{X}_k) - \mathbf{M}_k\right\|_F^2\middle|\mathcal{F}_{k-1}\right]\\
\leq&f(\mathbf{X}_k) - f^* + \mathbb{E}_k\left[-\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 + \frac{9\eta\nu}{K^{1/2}}\left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right)\middle|\mathcal{F}_{k-1}\right]\\
&+\frac{\eta\theta}{\sqrt{\hat{\varepsilon}}(1-\theta)}\left\|\nabla f(\mathbf{X}_{k-1}) - \mathbf{M}_{k-1}\right\|_F^2 + \frac{L^2\eta^3}{\hat{\varepsilon}(1-\theta)^2}\left\|\lambda\mathbf{L}_{k-1,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_{k-1}\mathbf{R}_{k-1,\varepsilon}^{\frac{1}{4q}} + \mathbf{L}_{k-1,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_{k-1}\mathbf{R}_{k-1,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 + \frac{\eta(1-\theta)}{\sqrt{\hat{\varepsilon}}}\sigma^2 \qquad (22)\\
\leq&f(\mathbf{X}_k) - f^* + \mathbb{E}_k\left[-\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 + \frac{9\eta\nu}{K^{1/2}}\left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right)\middle|\mathcal{F}_{k-1}\right]\\
&+\frac{\eta\theta}{\sqrt{\hat{\varepsilon}}(1-\theta)}\left\|\nabla f(\mathbf{X}_{k-1}) - \mathbf{M}_{k-1}\right\|_F^2 + \frac{\eta}{4}\left\|\lambda\mathbf{L}_{k-1,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_{k-1}\mathbf{R}_{k-1,\varepsilon}^{\frac{1}{4q}} + \mathbf{L}_{k-1,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_{k-1}\mathbf{R}_{k-1,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 + \frac{\eta(1-\theta)}{\sqrt{\hat{\varepsilon}}}\sigma^2,
\end{aligned}$$

where we let $\eta^2 \leq \frac{\hat{\varepsilon}(1-\theta)^2}{4L^2}$ in the last inequality. Taking expectation with respect to $\mathcal{F}_{k-1}$ and summing (21) with $k=1$ and (22) over $k = 2, \cdots, K$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}_K}\left[f(\mathbf{X}_{K+1}) - f^* + \frac{\eta}{4}\left\|\lambda\mathbf{L}_{K,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_K\mathbf{R}_{K,\varepsilon}^{\frac{1}{4q}} + \mathbf{L}_{K,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_K\mathbf{R}_{K,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 + \frac{\eta\theta}{\sqrt{\hat{\varepsilon}}(1-\theta)}\left\|\nabla f(\mathbf{X}_K) - \mathbf{M}_K\right\|_F^2\right]\\
\leq&f(\mathbf{X}_1) - f^* + \sum_{k=1}^K\mathbb{E}_{\mathcal{F}_k}\left[-\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 + \frac{9\eta\nu}{K^{1/2}}\left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right)\right]\\
&+\left(\frac{\eta\theta}{\sqrt{\hat{\varepsilon}}(1-\theta)} + \frac{\eta}{\sqrt{\hat{\varepsilon}}}\right)\mathbb{E}_{\mathcal{F}_1}\left[\left\|\nabla f(\mathbf{X}_1) - \mathbf{M}_1\right\|_F^2\right] + \frac{(K-1)\eta(1-\theta)}{\sqrt{\hat{\varepsilon}}}\sigma^2 \qquad (23)\\
=&f(\mathbf{X}_1) - f^* + \sum_{k=1}^K\mathbb{E}_{\mathcal{F}_k}\left[-\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 + \frac{9\eta\nu}{K^{1/2}}\left(\operatorname{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \operatorname{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right)\right]\\
&+\frac{\eta}{\sqrt{\hat{\varepsilon}}(1-\theta)}\mathbb{E}_{\mathcal{F}_1}\left[\left\|\nabla f(\mathbf{X}_1) - \mathbf{M}_1\right\|_F^2\right] + \frac{(K-1)\eta(1-\theta)}{\sqrt{\hat{\varepsilon}}}\sigma^2.
\end{aligned}$$

As the gradient is L-Lipschitz, we have

$$f^* \leq f\left(\mathbf{X} - \frac{1}{L}\nabla f(\mathbf{X})\right) \leq f(\mathbf{X}) - \frac{1}{L}\left\langle\nabla f(\mathbf{X}), \nabla f(\mathbf{X})\right\rangle + \frac{L}{2}\left\|\frac{1}{L}\nabla f(\mathbf{X})\right\|_F^2 = f(\mathbf{X}) - \frac{1}{2L}\left\|\nabla f(\mathbf{X})\right\|_F^2.$$

Using the recursion of $\mathbf{M}_1$ and $\mathbf{M}_0 = \mathbf{0}$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{F}_1}\left[\|\nabla f(\mathbf{X}_1) - \mathbf{M}_1\|_F^2\right] =& \mathbb{E}_{\mathcal{F}_1}\left[\|\theta\nabla f(\mathbf{X}_1) + (1-\theta)(\nabla f(\mathbf{X}_1) - \mathbf{G}_1)\|_F^2\right] \\ =& \theta^2\|\nabla f(\mathbf{X}_1)\|_F^2 + (1-\theta)^2\mathbb{E}_{\mathcal{F}_1}\left[\|\nabla f(\mathbf{X}_1) - \mathbf{G}_1\|_F^2\right] \\ \leq& 2L(f(\mathbf{X}_1) - f^*) + (1-\theta)^2\sigma^2.\end{aligned}$$

Plugging into (23), we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{F}_K}\left[f(\mathbf{X}_{K+1}) - f^* + \frac{\eta}{4}\left\|\lambda\mathbf{L}_{K,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_K\mathbf{R}_{K,\varepsilon}^{\frac{1}{4q}} + \mathbf{L}_{K,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_K\mathbf{R}_{K,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 + \frac{\eta\theta}{\sqrt{\hat{\varepsilon}}(1-\theta)}\|\nabla f(\mathbf{X}_K) - \mathbf{M}_K\|_F^2\right] \\ \leq& f(\mathbf{X}_1) - f^* + \sum_{k=1}^K\mathbb{E}_{\mathcal{F}_k}\left[-\frac{\eta}{2}\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2 + \frac{9\eta\nu}{K^{1/2}}\left(\mathrm{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \mathrm{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right)\right] \\ &+ \frac{2L\eta}{\sqrt{\hat{\varepsilon}}(1-\theta)}(f(\mathbf{X}_1) - f^*) + \frac{K\eta(1-\theta)}{\sqrt{\hat{\varepsilon}}}\sigma^2\end{aligned}$$

and

$$\begin{aligned}&\sum_{k=1}^K\mathbb{E}_{\mathcal{F}_k}\left[\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\right] \\ \leq& \frac{18\nu}{K^{1/2}}\sum_{k=1}^K\mathbb{E}_{\mathcal{F}_k}\left[\mathrm{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \mathrm{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right] + \frac{2(f(\mathbf{X}_1) - f^*)}{\eta} + \frac{4L}{\sqrt{\hat{\varepsilon}}(1-\theta)}(f(\mathbf{X}_1) - f^*) + \frac{2K(1-\theta)}{\sqrt{\hat{\varepsilon}}}\sigma^2 \\ \leq& \frac{18\nu}{K^{1/2}}\sum_{k=1}^K\mathbb{E}_{\mathcal{F}_k}\left[\mathrm{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \mathrm{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right] + \underbrace{\frac{2(f(\mathbf{X}_1) - f^*)}{\eta} + \frac{4L}{\sqrt{\hat{\varepsilon}}(1-\theta)}(f(\mathbf{X}_1) - f^*) + \frac{2K(1-\theta)}{\sqrt{\hat{\varepsilon}}}\hat{\sigma}^2}_{C_1},\end{aligned}$$

where we denote $\hat{\sigma}^2 = \max\left\{\sigma^2, \frac{L(f(\mathbf{X}_1)-f^*)}{K\gamma^2}\right\}$ with any $\gamma \in (0,1]$. From Lemma 3.4, we have

$$\sum_{k=1}^K\mathbb{E}\left[\mathrm{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \mathrm{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right] \leq K\left(\underbrace{\mathrm{tr}\left(\Sigma_L^{1/2}\right) + \mathrm{tr}\left(\Sigma_R^{1/2}\right) + (m+n)\sqrt{\varepsilon}}_{C_2}\right) + \frac{4}{\sqrt{1-\beta}}\sum_{t=1}^K\mathbb{E}\left[\|\nabla f(\mathbf{X}_t)\|_*\right].$$

From Lemma 3.3, we have

$$\begin{aligned}&\sum_{k=1}^K\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \\ \leq& \sqrt{\left(\sum_{k=1}^K\mathbb{E}\left[\mathrm{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \mathrm{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right]\right)\left(\sum_{k=1}^K\mathbb{E}\left[\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{4p}}\nabla f(\mathbf{X}_k)\mathbf{R}_{k,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2\right]\right)} \\ \leq& \sqrt{\left(\sum_{k=1}^K\mathbb{E}\left[\mathrm{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \mathrm{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right]\right)\left(\frac{18\nu}{K^{1/2}}\sum_{k=1}^K\mathbb{E}\left[\mathrm{tr}\left(\mathbf{L}_{k,\varepsilon}^{1/2}\right) + \mathrm{tr}\left(\mathbf{R}_{k,\varepsilon}^{1/2}\right)\right] + C_1\right)} \\ \leq& \sqrt{\left(\frac{4}{\sqrt{1-\beta}}\sum_{t=1}^K\mathbb{E}\left[\|\nabla f(\mathbf{X}_t)\|_*\right] + KC_2\right)\left(\frac{18\nu}{K^{1/2}}\frac{4}{\sqrt{1-\beta}}\sum_{t=1}^K\mathbb{E}\left[\|\nabla f(\mathbf{X}_t)\|_*\right] + 18\nu C_2K^{1/2} + C_1\right)}.\end{aligned}$$

So we have

$$\begin{aligned}&\left(\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2\\ \leq&\frac{288\nu}{K^{1/2}(1-\beta)}\left(\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2+\frac{4}{\sqrt{1-\beta}}\left(36\nu C_2K^{1/2}+C_1\right)\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]+18\nu C_2^2K^{3/2}+KC_1C_2.\end{aligned}\tag{24}$$

Next, we consider the constants. From Definition A.2 and Fact A.1, we have

$$\begin{aligned}\operatorname{tr}\left(\Sigma_L^{1/2}\right)+\operatorname{tr}\left(\Sigma_R^{1/2}\right)=&\sum_{i=1}^{m}\sqrt{\sigma_i(\Sigma_L)}+\sum_{i=1}^{n}\sqrt{\sigma_i(\Sigma_R)}\leq\sqrt{(m+n)\left(\sum_{i=1}^{m}\sigma_i(\Sigma_L)+\sum_{i=1}^{n}\sigma_i(\Sigma_R)\right)}\\ =&\sqrt{(m+n)\left(\operatorname{tr}\left(\Sigma_L\right)+\operatorname{tr}\left(\Sigma_R\right)\right)}=\sigma\sqrt{2(m+n)}\leq\hat{\sigma}\sqrt{2(m+n)}.\end{aligned}$$

Letting $\varepsilon=\frac{\tau\hat{\sigma}^2}{m+n}$ with any $\tau\leq 1$, we have $(m+n)\sqrt{\varepsilon}\leq\hat{\sigma}\sqrt{m+n}$ and

$$C_2\leq 2.5\hat{\sigma}\sqrt{m+n}.$$

Recall that we require the parameters satisfying the following relations in the above proof

$$\eta\leq\frac{\sqrt{\hat{\varepsilon}}}{2L},\quad \eta^2\leq\frac{\hat{\varepsilon}(1-\theta)^2}{4L^2}$$

and

$$\eta\lambda\leq\frac{\sqrt{\nu}}{2K^{5/4}},\quad \|\mathbf{X}_1\|_{op}\leq\frac{\sqrt{\nu}}{K^{1/4}\lambda},\quad \frac{\sqrt{\nu}}{K^{1/4}}\leq 1,\quad \theta\leq\beta\leq\sqrt{\theta}<1$$

in Lemma C.1. Letting

$$\begin{aligned}&1-\theta=\sqrt{\frac{L\left(f(\mathbf{X}_1)-f^*\right)}{K\hat{\sigma}^2}},\quad \eta=\sqrt{\frac{\hat{\varepsilon}\left(f(\mathbf{X}_1)-f^*\right)}{4LK\hat{\sigma}^2}},\quad \nu=\frac{1}{1152}\sqrt{\frac{L(f(\mathbf{X}_1)-f^*)}{\hat{\sigma}^2}},\\ &\lambda\leq\frac{1}{\sqrt{1152\hat{\varepsilon}}K^{3/4}}\sqrt[4]{\frac{L^3\hat{\sigma}^2}{f(\mathbf{X}_1)-f^*}},\quad \|\mathbf{X}_1\|_{op}\leq\sqrt{\frac{\hat{\varepsilon}K\left(f(\mathbf{X}_1)-f^*\right)}{L\hat{\sigma}^2}},\end{aligned}$$

the above requirements are satisfied by the definition of $\hat{\sigma}^2$. We also have

$$\begin{aligned}&\frac{1}{1-\beta}\leq\frac{1}{1-\sqrt{\theta}}\leq\frac{2}{1-\theta}=2\sqrt{\frac{K\hat{\sigma}^2}{L\left(f(\mathbf{X}_1)-f^*\right)}},\quad \frac{288\nu}{K^{1/2}(1-\beta)}\leq\frac{1}{2},\quad C_1\leq 10\sqrt{\frac{K\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)}{\hat{\varepsilon}}},\\ &\frac{C_1}{\sqrt{1-\beta}}\leq\frac{14.2\hat{\sigma}}{\sqrt{\hat{\varepsilon}}}\sqrt[4]{K^3\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)},\quad \frac{\nu C_2K^{1/2}}{\sqrt{1-\beta}}\leq\frac{\sqrt{m+n}}{288}\sqrt[4]{K^3\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)},\\ &\frac{4}{\sqrt{1-\beta}}\left(36\nu C_2K^{1/2}+C_1\right)\leq\left(\sqrt{m+n}+\frac{57\hat{\sigma}}{\sqrt{\hat{\varepsilon}}}\right)\sqrt[4]{K^3\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)},\\ &\nu C_2^2K^{3/2}\leq\frac{m+n}{144}\sqrt{K^3\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)},\quad KC_1C_2\leq 25\hat{\sigma}\sqrt{\frac{m+n}{\hat{\varepsilon}}}\sqrt{K^3\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)},\\ &18\nu C_2^2K^{3/2}+KC_1C_2\leq\left(m+n+25\hat{\sigma}\sqrt{\frac{m+n}{\hat{\varepsilon}}}\right)\sqrt{K^3\hat{\sigma}^2L(f(\mathbf{X}_1)-f^*)}\leq\left(\frac{27}{2}(m+n)+\frac{25}{2}\frac{\hat{\sigma}^2}{\hat{\varepsilon}}\right)\sqrt{K^3\hat{\sigma}^2L(f(\mathbf{X}_1)-f^*)}.\end{aligned}$$

So we have

$$\begin{aligned}\frac{1}{2}\left(\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2\leq&\left(\sqrt{m+n}+\frac{57\hat{\sigma}}{\sqrt{\hat{\varepsilon}}}\right)\sqrt[4]{K^3\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\\ &+\left(\frac{27}{2}(m+n)+\frac{25}{2}\frac{\hat{\sigma}^2}{\hat{\varepsilon}}\right)\sqrt{K^3\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)}.\end{aligned}$$

Solving inequality $x^2 - ax - b \le 0$, we have $x \le \frac{a+\sqrt{a^2+4b}}{2} \le a + \sqrt{b}$ and

$$\begin{aligned}\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] &\le \left(2\sqrt{m+n} + \frac{114\hat{\sigma}}{\sqrt{\hat{\varepsilon}}} + \sqrt{27(m+n) + 25\frac{\hat{\sigma}^2}{\hat{\varepsilon}}}\right) \sqrt[4]{K^3\hat{\sigma}^2 L\left(f(\mathbf{X}_1) - f^*\right)} \\ &\le \left(8\sqrt{m+n} + \frac{119\hat{\sigma}}{\sqrt{\hat{\varepsilon}}}\right) \sqrt[4]{K^3\hat{\sigma}^2 L\left(f(\mathbf{X}_1) - f^*\right)} \\ &= \left(8\sqrt{m+n} + \frac{119\hat{\sigma}}{\sqrt{\hat{\varepsilon}}}\right) \max\left\{\sqrt[4]{K^3\sigma^2 L\left(f(\mathbf{X}_1) - f^*\right)}, \sqrt{\frac{KL\left(f(\mathbf{X}_1) - f^*\right)}{\gamma}}\right\}.\end{aligned}$$

Dividing both sides by K, we have the conclusion. At last, Lemma C.1 guarantees

$$\lambda\|\mathbf{X}_k\|_{op} \le \frac{3\sqrt{\nu}}{K^{1/4}} = \frac{3}{\sqrt{1152}}\sqrt[4]{\frac{L\left(f(\mathbf{X}_1) - f^*\right)}{K\hat{\sigma}^2}} < 1$$

by the setting of $\hat{\sigma}^2$. □

C. Supporting Lemmas

The following lemma extends (Li et al., 2025a, Lemma 3) but replaces the infinite norm of vectors by spectral norm of matrices.

Lemma C.1. *Let $\eta\lambda \le \frac{\sqrt{\nu}}{2K^{5/4}}$, $\|\mathbf{X}_1\|_{op} \le \frac{\sqrt{\nu}}{K^{1/4}\lambda}$, $\frac{\sqrt{\nu}}{K^{1/4}} \le 1$, $\theta \le \beta \le \sqrt{\theta} < 1$, and $\frac{1}{p} + \frac{1}{q} = 1$. Then for Algorithm 1, we have*

$$\lambda\|\mathbf{X}_k\|_{op} \le \frac{3\sqrt{\nu}}{K^{1/4}}, \quad \forall k = 1, 2, \cdots, K.$$

Proof. From the update of $\mathbf{X}_{k+1}$, we have

$$\begin{aligned}\|\mathbf{X}_{k+1}\|_{op} - \frac{2}{\lambda} &= \left\|(1-\lambda\eta)\mathbf{X}_k - \eta\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\|_{op} - \frac{2}{\lambda} \\ &\le (1-\lambda\eta)\|\mathbf{X}_k\|_{op} + \eta\left\|\mathbf{L}_{k,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_k\mathbf{R}_{k,\varepsilon}^{-\frac{1}{2q}}\right\|_{op} - \frac{2}{\lambda} \\ &\overset{(1)}{\le} (1-\lambda\eta)\|\mathbf{X}_k\|_{op} + 2\eta - \frac{2}{\lambda} \\ &= (1-\lambda\eta)\left(\|\mathbf{X}_k\|_{op} - \frac{2}{\lambda}\right) \\ &\le (1-\lambda\eta)^k\left(\|\mathbf{X}_1\|_{op} - \frac{2}{\lambda}\right) \\ &= -\frac{1}{\lambda}(1-\lambda\eta)^k\left(2 - \frac{\sqrt{\nu}}{K^{1/4}}\right),\end{aligned}$$

where we use Lemma 3.6 in $\overset{(1)}{\le}$. Since $\ln x \le x - 1$ and $e^x \ge x + 1$ for any $x > 0$ and $\eta\lambda \le \frac{\sqrt{\nu}}{2K^{5/4}} \le \frac{1}{2}$, we have for any $k \le K$ that

$$k\ln(1-\eta\lambda) = -k\ln\frac{1}{1-\eta\lambda} \ge -K\left(\frac{1}{1-\eta\lambda} - 1\right) = -\frac{K\eta\lambda}{1-\eta\lambda} \ge -\frac{\sqrt{\nu}}{K^{1/4}},$$

$$(1-\eta\lambda)^k \ge e^{-\frac{\sqrt{\nu}}{K^{1/4}}} \ge 1 - \frac{\sqrt{\nu}}{K^{1/4}},$$

and

$$\|\mathbf{X}_{k+1}\|_{op} - \frac{2}{\lambda} \le -\frac{1}{\lambda}\left(1 - \frac{\sqrt{\nu}}{K^{1/4}}\right)\left(2 - \frac{\sqrt{\nu}}{K^{1/4}}\right) \le -\frac{2}{\lambda} + \frac{3}{\lambda}\frac{\sqrt{\nu}}{K^{1/4}}.$$

□

The following lemma is closely similar to (Li et al., 2025a, Lemma 4) and we list the proof here only for the sake of completeness.

Lemma C.2. *Suppose that Assumptions 1-3 and condition (3) hold and let* $\frac{1}{p}+\frac{1}{q}=1$. *Then for Algorithm 1, we have*

$$\begin{aligned}&\mathbb{E}_k\left[\|\nabla f(\mathbf{X}_k)-\mathbf{M}_k\|_F^2\left|\mathcal{F}_{k-1}\right.\right]\\ \leq&\theta\|\nabla f(\mathbf{X}_{k-1})-\mathbf{M}_{k-1}\|_F^2+\frac{L^2\eta^2}{(1-\theta)\sqrt{\hat{\varepsilon}}}\left\|\lambda\mathbf{L}_{k-1,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_{k-1}\mathbf{R}_{k-1,\varepsilon}^{\frac{1}{4q}}+\mathbf{L}_{k-1,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_{k-1}\mathbf{R}_{k-1,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2+(1-\theta)^2\sigma^2.\end{aligned}\tag{25}$$

Proof. Denoting $\Gamma_k=\mathbf{G}_k-\nabla f(\mathbf{X}_k)$, we have $\mathbb{E}_k\left[\Gamma_k\left|\mathcal{F}_{k-1}\right.\right]=0$, $\mathbb{E}_k\left[\|\Gamma_k\|_F^2\left|\mathcal{F}_{k-1}\right.\right]\leq\sigma^2$. From the update of $\mathbf{M}_k$, we have

$$\begin{aligned}\mathbf{M}_k-\nabla f(\mathbf{X}_k)=&\theta\mathbf{M}_{k-1}+(1-\theta)\mathbf{G}_k-\nabla f(\mathbf{X}_k)\\ =&\theta\left(\mathbf{M}_{k-1}-\nabla f(\mathbf{X}_{k-1})\right)+(1-\theta)\left(\nabla f(\mathbf{X}_k)+\Gamma_k\right)-\nabla f(\mathbf{X}_k)+\theta\nabla f(\mathbf{X}_{k-1})\\ =&\theta\left(\mathbf{M}_{k-1}-\nabla f(\mathbf{X}_{k-1})\right)+(1-\theta)\Gamma_k-\theta\left(\nabla f(\mathbf{X}_k)-\nabla f(\mathbf{X}_{k-1})\right)\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_k\left[\|\nabla f(\mathbf{X}_k)-\mathbf{M}_k\|_F^2\left|\mathcal{F}_{k-1}\right.\right]\\ =&\|\theta\left(\mathbf{M}_{k-1}-\nabla f(\mathbf{X}_{k-1})\right)-\theta\left(\nabla f(\mathbf{X}_k)-\nabla f(\mathbf{X}_{k-1})\right)\|_F^2+(1-\theta)^2\mathbb{E}_k\left[\|\Gamma_k\|_F^2\left|\mathcal{F}_{k-1}\right.\right]\\ \leq&\theta^2\left(1+\frac{1-\theta}{\theta}\right)\|\mathbf{M}_{k-1}-\nabla f(\mathbf{X}_{k-1})\|_F^2+\theta^2\left(1+\frac{\theta}{1-\theta}\right)\|\nabla f(\mathbf{X}_k)-\nabla f(\mathbf{X}_{k-1})\|_F^2+(1-\theta)^2\mathbb{E}_k\left[\|\Gamma_k\|_F^2\left|\mathcal{F}_{k-1}\right.\right]\\ \leq&\theta\|\mathbf{M}_{k-1}-\nabla f(\mathbf{X}_{k-1})\|_F^2+\frac{1}{1-\theta}\|\nabla f(\mathbf{X}_k)-\nabla f(\mathbf{X}_{k-1})\|_F^2+(1-\theta)^2\sigma^2\\ \leq&\theta\|\mathbf{M}_{k-1}-\nabla f(\mathbf{X}_{k-1})\|_F^2+\frac{L^2}{1-\theta}\|\mathbf{X}_k-\mathbf{X}_{k-1}\|_F^2+(1-\theta)^2\sigma^2\\ =&\theta\|\mathbf{M}_{k-1}-\nabla f(\mathbf{X}_{k-1})\|_F^2+\frac{L^2\eta^2}{1-\theta}\left\|\lambda\mathbf{X}_{k-1}+\mathbf{L}_{k-1,\varepsilon}^{-\frac{1}{2p}}\mathbf{M}_{k-1}\mathbf{R}_{k-1,\varepsilon}^{-\frac{1}{2q}}\right\|_F^2+(1-\theta)^2\sigma^2\\ \leq&\theta\|\mathbf{M}_{k-1}-\nabla f(\mathbf{X}_{k-1})\|_F^2+\frac{L^2\eta^2}{(1-\theta)\sqrt{\hat{\varepsilon}}}\left\|\lambda\mathbf{L}_{k-1,\varepsilon}^{\frac{1}{4p}}\mathbf{X}_{k-1}\mathbf{R}_{k-1,\varepsilon}^{\frac{1}{4q}}+\mathbf{L}_{k-1,\varepsilon}^{-\frac{1}{4p}}\mathbf{M}_{k-1}\mathbf{R}_{k-1,\varepsilon}^{-\frac{1}{4q}}\right\|_F^2+(1-\theta)^2\sigma^2,\end{aligned}$$

where we use (20) in the last inequality. □

Lemma C.3. *When each element of* $\mathbf{G}\in\mathbb{R}^{m\times n}$ *is generated from Gaussian distribution with* μ *mean and* ξ^2 *variance independently, we have*

$$\mathbb{E}\left[\mathbf{G}\mathbf{G}^T\right]\succeq\frac{\xi^2}{m(\xi^2+\mu^2)}\mathbb{E}\left[\|\mathbf{G}\|_F^2\right]\mathbf{I}_m,\quad\mathbb{E}\left[\mathbf{G}^T\mathbf{G}\right]\succeq\frac{\xi^2}{n(\xi^2+\mu^2)}\mathbb{E}\left[\|\mathbf{G}\|_F^2\right]\mathbf{I}_n.$$

Proof. When $\mathbf{G}_{i,j}\sim\mathcal{N}(\mu,\xi^2)$, we have

$$\mathbb{E}\left[\left(\mathbf{G}\mathbf{G}^T\right)_{p,q}\right]=\mathbb{E}\left[\sum_{j=1}^n\mathbf{G}_{p,j}\mathbf{G}_{q,j}\right]=\sum_{j=1}^n\mathbb{E}\left[\mathbf{G}_{p,j}\mathbf{G}_{q,j}\right]=\sum_{j=1}^n\mathbb{E}\left[\mathbf{G}_{p,j}\right]\mathbb{E}\left[\mathbf{G}_{q,j}\right]=n\mu^2\quad\text{if}\quad p\neq q$$

and

$$\mathbb{E}\left[\left(\mathbf{G}\mathbf{G}^T\right)_{p,q}\right]=\sum_{j=1}^n\mathbb{E}\left[\mathbf{G}_{p,j}^2\right]=\sum_{j=1}^n\left(\mathbb{E}\left[\left(\mathbf{G}_{p,j}-\mu\right)^2\right]+\mu^2\right)=n(\xi^2+\mu^2)\quad\text{if}\quad p=q.$$

So

$$\mathbb{E}\left[\mathbf{G}\mathbf{G}^T\right]=n\mu^2\mathbf{1}_m\mathbf{1}_m^T+n\xi^2\mathbf{I}_m\succeq n\xi^2\mathbf{I}_m,$$

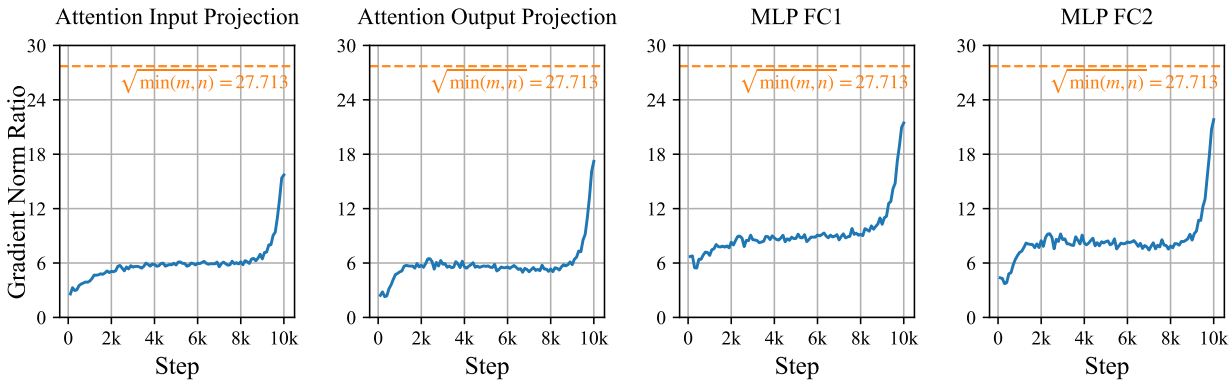

Figure 2. The gradient norm ratio $\|\nabla f(\mathbf{X})\|_*/\|\nabla f(\mathbf{X})\|_F$ during GPT-2 pretraining on OpenWebText for each class of Shampoo-handled parameters. The dashed horizontal line indicates the value of $\sqrt{\min(m,n)}$ for each matrix shape.

where $\mathbf{1}_m \in \mathbb{R}^m$ is the vector with all ones. We also have

$$\mathbb{E}\left[\|\mathbf{G}\|_F^2\right] = \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}\left[\mathbf{G}_{i,j}^2\right] = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mathbb{E}\left[(\mathbf{G}_{i,j}-\mu)^2\right]+\mu^2\right) = mn(\xi^2+\mu^2).$$

So we have

$$\mathbb{E}\left[\mathbf{G}\mathbf{G}^T\right] \succeq n\xi^2\mathbf{I}_m = \frac{\xi^2}{m(\xi^2+\mu^2)}\mathbb{E}\left[\|\mathbf{G}\|_F^2\right]\mathbf{I}_m.$$

Similarly, we also have

$$\mathbb{E}\left[\mathbf{G}^T\mathbf{G}\right] \succeq \frac{\xi^2}{n(\xi^2+\mu^2)}\mathbb{E}\left[\|\mathbf{G}\|_F^2\right]\mathbf{I}_n.$$

□

D. Experiments

In this section, we conduct experiments on real-world deep learning tasks to examine whether the theoretical claims developed in the main paper are reflected in practical training dynamics. While recent non-diagonal preconditioning optimizers, such as SOAP (Vyas et al., 2025) and Muon (Jordan et al., 2024), have primarily demonstrated their effectiveness through improved end-to-end performance on large language model (LLM) training, in light of this trend, we pretrain the GPT-2 (Radford et al., 2019) model from scratch on OpenWebText (Gokaslan et al., 2019) dataset.

One of the claims that requires numerical verification in our analysis is the relationship between the nuclear norm $\|\nabla f(\mathbf{X})\|_*$ and the Frobenius norm $\|\nabla f(\mathbf{X})\|_F$ of the gradient. Evaluating this relationship in practice involves computing the *full* gradient with respect to the training objective over the entire dataset. However, the OpenWebText dataset contains approximately 9 billion tokens, making exact full-gradient computation computationally intractable. As such, we adopt an approximation strategy where the stochastic gradients are accumulated over 100 consecutive mini-batches while keeping the model parameters frozen.

We apply this approximation to each Shampoo-handled 2D parameter in the GPT-2 model. These parameters fall into four categories: (i) attention input projection matrices, (ii) attention output projection matrices, (iii) the first layer of feed-forward network, and (iv) the second layer of feed-forward network. For clarity and brevity, we report the averaged results averaged within each parameter category.

Figure 2 reports the evolution of the ratio $\|\nabla f(\mathbf{X})\|_*/\|\nabla f(\mathbf{X})\|_F$ for the four classes of Shampoo-handled matrices. Across all parameter categories, we observe that this ratio remains close to the theoretical upper bound $\sqrt{\min(m,n)}$ throughout training, as indicated by the dashed reference lines. This behavior suggests that the relationship $\|\nabla f(\mathbf{X})\|_* =$

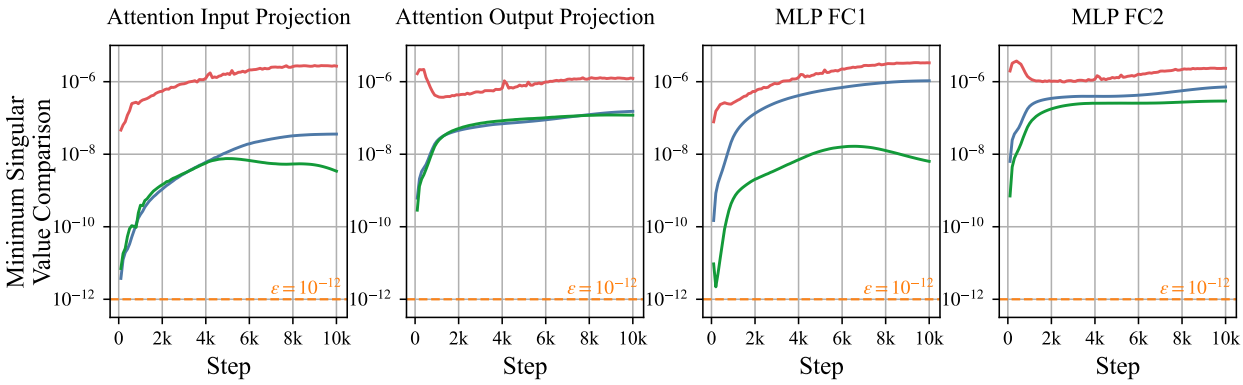

Figure 3. Comparison of the minimum eigenvalues with respect to the noise level during GPT-2 pretraining on OpenWebText. The red curve indicates the noise-dependent scale $\sigma^2/(m+n)$, while the blue and green curves correspond to the minimum eigenvalues of the regularized left and right preconditioner matrices $\mathbf{L}_{k,\varepsilon}$ and $\mathbf{R}_{k,\varepsilon}$, respectively. All quantities are plotted on a logarithmic scale.

$\Theta(\sqrt{\min\{m,n\}})\,\|\nabla f(\mathbf{X})\|_F$ indeed holds in practice, thereby supporting the claim in the main paper that nuclear-norm-based measures are comparable to Frobenius-norm-based rates in realistic LLM training settings.

Next, we turn to the empirical estimation of the gradient noise level and compare the relative magnitudes of $\sigma^2/(m+n)$, the parameter ε, and the minimum eigenvalues of the preconditioners $\mathbf{L}_{k,\varepsilon}$ and $\mathbf{R}_{k,\varepsilon}$, serving as estimates of $\hat{\epsilon}$. Similarly, we approximate σ^2 using the same large-batch gradient approximation described above, treating the accumulated gradient as a proxy for the true gradient $\nabla f(\mathbf{X})$. The parameter ε is fixed to 10^{-12} which is consistent with the experimental settings in (Shi et al., 2023). To estimate $\hat{\varepsilon}$, we record the minimum eigenvalues of the regularized left and right preconditioner matrices $\mathbf{L}_{k,\varepsilon}$ and $\mathbf{R}_{k,\varepsilon}$, respectively. These quantities are readily available during training, as they are produced as a byproduct of the eigendecomposition used to compute the matrix inverse roots in Shampoo.

Figure 3 compares the empirical noise scale $\sigma^2/(m+n)$ with the minimum eigenvalues of the preconditioner matrices, which serve as estimates of $\hat{\varepsilon}$. Across all four classes of Shampoo-handled parameters, we observe that the minimum eigenvalues of both $\mathbf{L}_{k,\varepsilon}$ and $\mathbf{R}_{k,\varepsilon}$ are consistently several orders of magnitude larger than the configured parameter $\varepsilon = 10^{-12}$. Moreover, the value of $\hat{\varepsilon}$ remains comparable to the noise-dependent scale $\sigma^2/(m+n)$ throughout training.