On the Convergence Analysis of Muon

Anonymous authors
Paper under double-blind review

Abstract

The majority of parameters in neural networks are naturally represented as matrices. However, most commonly used optimizers treat these matrix parameters as flattened vectors during optimization, potentially overlooking their inherent structural properties. Recently, an optimizer called Muon has been proposed, specifically designed to optimize matrix-structured parameters. Extensive empirical evidence shows that Muon can significantly outperform traditional optimizers when training neural networks. Nonetheless, the theoretical understanding of Muon's convergence behavior and the reasons behind its superior performance remain limited. In this work, we present a comprehensive convergence rate analysis of Muon and its comparison with Gradient Descent (GD). We characterize the conditions under which Muon can outperform GD. Our theoretical results reveal that Muon can benefit from the low-rank structure of Hessian matrices, a phenomenon widely observed in practical neural network training. Our experimental results support and corroborate the theoretical findings.

1 Introduction

Modern neural networks, such as large language models (LLMs) (Adler et al., 2024), typically consist of a massive amount of parameters and require significant computational resources for training. As a result, designing optimizers that can efficiently train such models has become an important and valuable research problem. Note that most of the parameters in neural networks are naturally represented as matrices, for example, weight matrices in fully connected layers, or the query, key, and value matrices in attention mechanisms. However, most widely adopted optimization algorithms such as Stochastic Gradient Descent (SGD), Adam (Kingma & Ba, 2015), and their variants often treat these matrices as flattened vectors, potentially overlooking their inherent structural properties.

Recently, Muon, an optimizer specifically designed for matrix-structured parameters, was proposed in Jordan et al. (2024). The key idea is to update the matrix parameters along an orthogonalized version of their gradients at each iteration (Algorithm 1). Empirical results across various studies and network scales consistently demonstrate the superior performance of Muon in optimizing neural networks with matrix parameters (Jordan et al., 2024; Liu et al., 2025a; An et al., 2025; Liu et al., 2025b; Ahn & Xu, 2025). However, the convergence behavior of Muon and the underlying mechanisms that contribute to its advantage over traditional gradient descent-based optimizers, such as GD and SGD, are not yet fully understood. Therefore, in this work, we aim to bridge this gap by providing a comprehensive theoretical analysis of Muon's convergence properties, along with detailed comparisons to GD-based algorithms.

In classical optimization theory, the stepsize of GD typically depends on the Lipschitz smoothness constant L of the Frobenius norm, namely the maximum singular value of the Hessian. However, this quantity can be extremely large or vary rapidly during neural network training, making the stepsize of GD difficult to tune in practice. In this work, we analyze the convergence behavior of Muon not only under the standard Lipschitz smoothness constant L, but also under the spectral norm Lipschitz smoothness constant L_*, and even without assuming uniform Lipschitz smoothness. We show that the convergence of Muon can depend on J, which can be viewed as an average of the global Hessian information throughout training. Compared with L, which is determined by the maximum singular value of the Hessian, J can be interpreted as averaging across both iterations and singular values. Such a smoothing effect provides potential advantages for Muon

over GD-based algorithms, especially when the Hessians exhibits low-rank structures, a phenomenon widely observed in practical neural network training. Our experimental results on quadratic regressions and neural network trainings validate these theoretical findings.

2 Related Work

Muon was originally proposed in Jordan et al. (2024), which demonstrated the superior performance of Muon on various models and datasets. Subsequently, Liu et al. (2025a) improved the Muon by incorporating techniques such as weight decay and adjusting the per-parameter update scale, and showed that Muon can outperform Adam when optimizing large language models (LLMs). One possible explanation for Muon's superior performance is provided by Bernstein & Newhouse (2024), which showed that Muon's update direction corresponds to the steepest descent under the spectral norm constraint. Moreover, we note that there are some recent works (Li & Hong, 2025; Pethick et al., 2025; An et al., 2025; Kovalev, 2025) that also analyze the convergence behavior of Muon. For example, Li & Hong (2025) analyzed the Frobenius norm convergence (Definition 3.4) of Muon under Frobenius norm Lipschitz smooth (Assumption 3.1). An et al. (2025) analyzed the convergence of the Simplified Muon (Algorithm 2) with Assumption 4.11. Pethick et al. (2025) viewed Muon as one of the linear minimization oracle (lmo) based family algorithms and Kovalev (2025) viewed Muon as a special case of first-order trust-region optimization methods. Both Pethick et al. (2025); Kovalev (2025) provided the convergence analysis of Muon with Assumption 3.2.

Recently, several works have also investigated the superiority of Muon from theoretical perspectives. Wang et al. (2025) demonstrated that Muon can outperform Adam in tail-end associative memory learning, showing that spectral orthogonalization enables more balanced updates across singular modes. Su (2025) introduced an isotropic curvature model to analyze matrix-structured optimization updates. Davis & Drusvyatskiy (2025) analyzed the squared nuclear-to-Frobenius ratio of the gradient relative to the stable rank of the incoming activations to characterize when Muon outperforms GD, and show that this condition holds in random-feature models, deep networks, and transformer training. Ma et al. (2026) analyzed matrix factorization and in-context learning of linear transformers, showing that for both problems Muon achieves iteration complexity independent of the condition number, thereby outperforming GD and Adam.

Compared to these prior works, our work provides a more comprehensive convergence analysis of Muon under various smoothness assumptions (Assumption 3.1, 3.2, 4.11), including scenarios without uniform Lipschitz smoothness (Assumption 4.13). Additionally, we offer detailed comparisons with GD and characterize the conditions under which Muon outperforms GD, and validate these conditions through experiments. Therefore, our work presents distinct contributions and insights. We believe that all these works collectively contribute to a deeper understanding of Muon.

Low-rank and approximate blockwise diagonal structure of Hessian. Extensive prior works (Collobert, 2004; Zhang et al., 2024b;a; Dong et al., 2025) have shown, both theoretically and empirically, that the Hessian of a neural network tends to exhibit a blockwise diagonal structure with each block corresponding to an individual neuron. Recently, Zhang et al. (2024b;a) numerically confirmed this property in small Transformers, and Dong et al. (2025) provided theoretical explanations for why such a blockwise diagonal structure emerges in neural network Hessians. In addition, many studies (Sagun et al., 2016; 2017; Wu et al., 2020; Papyan, 2020; Yao et al., 2020) have also observed that neural network Hessians are typically low-rank, i.e., their effective ranks, which can be measured by the ratio $\|H\|_*/\|H\|_{\mathrm{op}}$, can be significantly smaller than their dimensionality.

Other structured optimization methods. Although optimizers commonly used for training deep neural networks, such as SGD, Adam (Kingma & Ba, 2015), and AdamW (Loshchilov & Hutter, 2017), typically treat structured parameters (e.g., matrices) as flattened vectors, in recent years, there has been growing interest in designing structured optimizers that explicitly leverage the inherent structure of parameters. For instance, Adafactor (Shazeer & Stern, 2018), LAMB (You et al., 2019), and Adam-mini (Zhang et al., 2024b) incorporate matrix- or layer-level structure to reduce memory consumption. KFAC (Martens & Grosse, 2015) and TNT (Ren & Goldfarb, 2021) approximate the Fisher matrix to implement natural gradient methods. Shampoo (Gupta et al., 2018) is specifically designed for matrix or tensor parameters and can be viewed as an approximation to the full-matrix preconditioner in AdaGrad (Duchi et al., 2011). Morwani et al. (2024)

provided additional theoretical analyses and modifications to Shampoo. Recently, SOAP (Vyas et al., 2024) was introduced, which combines the ideas of Adam and Shampoo. Galore (Zhao et al., 2024) was proposed to exploit the low-rank structure of gradients for memory efficiency. Additionally, Liu et al. (2025b) proposed COSMOS, which can be seen as a combination of SOAP and Muon. More recently, Xie et al. (2025) and An et al. (2025) introduced ASGO (One-Sided Shampoo), which only applies a one-sided preconditioner within the Shampoo. Xie et al. (2025); An et al. (2025) showed that ASGO can achieve better convergence rate than the original Shampoo.

3 Preliminaries

Notation. For a vector v, we denote its l_2 norm as $\|v\|_2$. For a matrix $A \in \mathbb{R}^{m\times n}$, we denote its nuclear norm as $\|A\|_*$, spectral norm as $\|A\|_{\mathrm{op}}$, Frobenius norm as $\|A\|_{\mathrm{F}}$, and Λ norm as $\|A\|_\Lambda = \sqrt{\mathrm{tr}\,(A\Lambda A^\top)}$ where $\Lambda \in \mathbb{R}^{n\times n}$ is a real symmetric positive definite matrix. For $A \in \mathbb{R}^{m\times n}$ with rows $a_1, \ldots, a_m$, we define $\overline{\mathrm{vec}}(A) = (a_1\, a_2 \cdots a_m)^\top \in \mathbb{R}^{mn}$ as the vectorization of A. We denote I_d as the identity matrix with dimension d. For matrix $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m'\times n'}$, we denote their Kronecker product as $A \otimes B \in \mathbb{R}^{mm'\times nn'}$. We define the operator norm of a third-order tensor $H \in \mathbb{R}^{d_1\times d_2\times d_3}$ as $\|H\|_{\mathrm{op}} = \max_{\|u\|_2=\|v\|_2=\|w\|_2=1} |\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\sum_{k=1}^{d_3} H_{ijk}u_iv_jw_k|$, where $u \in \mathbb{R}^{d_1}, v \in \mathbb{R}^{d_2}, w \in \mathbb{R}^{d_3}$.

3.1 Optimization problem

In this paper, we consider the following stochastic optimization problem:

$$\min_{W\in\mathbb{R}^{m\times n}} f(W) = \mathbb{E}_\xi[f(W;\xi)]. \tag{1}$$

We denote $f^* = \inf_W f(W)$, $r = \min\{m, n\}$, $f_v(w) = f(W)$, and $w = \overline{\mathrm{vec}}(W) \in \mathbb{R}^{mn}$. We assume $f^* > -\infty$. For initial point W_0, we denote $\Delta = f(W_0) - f^*$, and $D_F = \|W_0 - W^*\|_F$. We consider the following smoothness assumptions in this paper.

Assumption 3.1 (Frobenius norm Lipschitz smooth)**.** *We say $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is L Frobenius norm Lipschitz smooth if for any $W, W' \in \mathbb{R}^{m\times n}$, we have*

$$\|\nabla f(W) - \nabla f(W')\|_{\mathrm{F}} \le L\|W - W'\|_{\mathrm{F}}.$$

Assumption 3.2 (Spectral norm Lipschitz smooth)**.** *We say $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is L_* spectral norm Lipschitz smooth if for any $W, W' \in \mathbb{R}^{m\times n}$, we have*

$$\|\nabla f(W) - \nabla f(W')\|_* \le L_*\|W - W'\|_{\mathrm{op}}.$$

Assumption 3.1 is the standard Frobenius norm-based smoothness, which serves as a natural extension of the conventional l_2-norm smoothness for functions with vector parameters to functions with matrix parameters. Assumption 3.2 is a spectral norm smoothness, which accounts for the distinct structure of matrix parameters. In this work, we establish the convergence of Muon under both assumptions.

For a stochastic setting, we also use the following standard bounded variance assumption.

Assumption 3.3 (Bounded variance)**.** *We assume $\nabla f(W;\xi)$ is an unbiased stochastic estimator of the true gradient $\nabla f(W)$ and have a bounded variance, i.e.*

$$\mathbb{E}[\nabla f(W;\xi)] = \nabla f(W), \qquad \mathbb{E}\|\nabla f(W;\xi) - \nabla f(W)\|_{\mathrm{F}}^2] \le \sigma^2.$$

Definition 3.4. We say $\widehat{W}$ is an ϵ-Frobenius norm stationary point of f if $\|\nabla f(\widehat{W})\|_{\mathrm{F}} \le \epsilon$.

Definition 3.5. We say $\widehat{W}$ is an ϵ-nuclear norm stationary point of f if $\|\nabla f(\widehat{W})\|_* \le \epsilon$.

For a matrix A with rank r, there is a well-known relationship between its Frobenius norm and nuclear norm: $\|A\|_{\mathrm{F}} \le \|A\|_* \le \sqrt{r}\|A\|_{\mathrm{F}}$. Thus, we have the following proposition.

Proposition 3.6. *If $\widehat{W} \in \mathbb{R}^{m\times n}$ is an ϵ-nuclear norm stationary point of f, then it is also an ϵ-Frobenius norm stationary point of f. If $\widehat{W} \in \mathbb{R}^{m\times n}$ is an ϵ-Frobenius norm stationary point of f, then it is also an $\sqrt{r}\epsilon$-nuclear norm stationary point of f, where $r = \min\{m, n\}$.*

We note that for functions with matrix parameters, the nuclear norm stationarity is tighter than the ordinary Frobenius norm stationarity.

3.2 Muon

Algorithm 1 Muon

1: **Input**: Initial weights W_0, learning rate schedule $\{\eta_t\}$, $\beta \in [0, 1)$, batch size B
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Sample batch $\{\xi_{t,i}\}_{i=1}^{B}$ uniformly
4: $G_t = \frac{1}{B}\sum_{i=1}^{B} \nabla f(W_t; \xi_{t,i})$ (or $G_t = \nabla f(W_t)$ in the deterministic setting)
5: If $t > 0$, $M_t = \beta M_{t-1} + (1-\beta)G_t$. If $t = 0$, $M_0 = G_0$.
6: $(U_t, S_t, V_t) = \mathrm{SVD}(M_t)$
7: $W_{t+1} = W_t - \eta_t U_t V_t^\top$
8: **end for**

Algorithm 2 Simplified Muon

1: **Input**: Initial weights W_0, learning rate schedule $\{\eta_t\}$
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: $G_t = \nabla f(W_t)$
4: $(U_t, S_t, V_t) = \mathrm{SVD}(G_t)$
5: $W_{t+1} = W_t - \eta_t U_t V_t^\top$
6: **end for**

In this subsection, we introduce the Muon algorithms. The original idea of Muon in Jordan et al. (2024) is to orthogonalize the update matrix. For example, suppose the original update direction is G_t with rank r_t and the singular value decomposition (SVD) of G_t is $U_t S_t V_t^\top$, where $S_t \in \mathbb{R}^{r_t\times r_t}$ is the diagonal matrix of the singular values of G_t, $U_t \in \mathbb{R}^{m\times r_t}$ and $V_t \in \mathbb{R}^{n\times r_t}$ are the left and right singular vector matrices of G_t, respectively. Then, in Muon, the update matrix will be $U_t V_t^\top$, which is the nearest semi-orthogonal matrix to the original G_t. Following this main idea, we have Algorithm 2, which can be viewed as the simplest form of Muon. Additionally, if we add momentum and conduct orthogonalization over the momentum direction, we have Algorithm 1, which is the main algorithm we will analyze in the stochastic setting. In practice, computing the SVD is very costly. Therefore, a common practical implementation of Muon uses Newton–Schulz iterations (Bernstein & Newhouse, 2024; Jordan et al., 2024) to approximate the orthogonalization process. In this paper, we primarily analyze the theoretical properties of Algorithm 1 and 2 using SVD.

4 Convergence Analysis of Muon

In this section, we analyze the convergence behaviors of Muon under various smoothness assumptions and compare them with GD. The proofs of theorems in this section can be found in Appendix B and Appendix C.

4.1 Frobenius norm Lipschitz smooth

We first analyze the convergence of Muon with the conventional Frobenius norm Lipschitz smoothness assumption (Assumption 3.1).

Theorem 4.1. *Under Assumptions 3.1 and 3.3, if we apply Algorithm 1 with $\eta_t = \eta$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \le \frac{\mathbb{E}[f(W_0) - f(W_T)]}{T\eta} + \frac{Lr\eta}{2} + \frac{2\sigma\sqrt{r(1-\beta)}}{\sqrt{B(1+\beta)}} + \frac{2\beta\sigma\sqrt{r}}{(1-\beta)T\sqrt{B}} + \frac{2r\eta\beta L}{1-\beta}.$$

By choosing the parameters in Theorem 4.1, we have following corollary.

Corollary 4.2. *Under the same assumptions as Theorem 4.1, we have the following convergence guarantees for Algorithm 1:*

1. ***Stochastic setting:*** *If we set* $B = 1$, $\eta = \sqrt{\frac{(1-\beta)\Delta}{rTL}}$, *and* $1-\beta = \min\left\{\frac{\sqrt{L\Delta}}{\sigma\sqrt{T}}, 1\right\}$, *then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq O\left(\sqrt[4]{\frac{r^2L\Delta\sigma^2}{T}} + \sqrt{\frac{rL\Delta}{T}} + \frac{\sqrt{r}\sigma^2}{\sqrt{L\Delta T}}\right).$$

2. ***Deterministic setting*** *(*$\sigma = 0$*):* *If we set* $\eta = \sqrt{\frac{\Delta}{rTL}}$, *and* $\beta = O(1)$, *then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(W_t)\|_* \leq O\left(\sqrt{\frac{rL\Delta}{T}}\right).$$

Therefore, the complexity to find an ϵ*-nuclear norm stationary point of* f *is* $O(r^2L\sigma^2\Delta\epsilon^{-4})$ *in the stochastic setting and* $O(rL\Delta\epsilon^{-2})$ *in the deterministic setting.*

Moreover, to study the convergence of function value of Muon, we also consider the star convex condition.

Assumption 4.3 (Star convex)**.** *For any* $W \in \mathbb{R}^{m\times n}$, *we assume the following inequality holds*

$$\langle \nabla f(W), W - W^* \rangle \geq f(W) - f^*.$$

where $W^* \in \arg\min_W f(W)$ *is the optimal point.*

Note that the star convex condition (Nesterov & Polyak, 2006) is weaker than the convex condition, and empirical evidence (Kleinberg et al., 2018; Zhou et al., 2019) suggests that, in some cases, the optimization path of neural networks can satisfy this condition.

To ensure the convergence of Muon in the star convex case, we adopt the following mild assumption. Note that the same assumption is also used in Gupta et al. (2018); An et al. (2025) for their convergence analysis of their algorithms in similar settings.

Assumption 4.4. *For* $W_t, t = 0, \ldots, T$, *generated by Algorithm 2, we assume* $\|W_t - W^*\|_{\text{op}} \leq D_{\text{op}}$.

In the experiment (Figure 1(b)), we can observe that when we use Muon on a quadratic function, $\|W_t - W^*\|$ is bounded, which satisfies Assumption 4.4.

With these assumptions, we have the following theorem.

Theorem 4.5. *Under Assumptions 4.3, 3.1, and 4.4, if we apply Algorithm 2 with* $\eta_t = \eta$, *then*

$$f(W_T) - f^* \leq \left(1 - \frac{\eta}{D_{\text{op}}}\right)^T (f(W_0) - f^*) + \frac{rLD_{\text{op}}\eta}{2}.$$

Thus, to reach the precision $f(W_T) - f^* \leq \epsilon$, *we can take* $\eta = O(\frac{\epsilon}{rLD_{\text{op}}})$, *and the complexity is* $O(rLD_{\text{op}}^2\epsilon^{-1}log\frac{\Delta}{\epsilon})$.

Theorem 4.6. *Under the same assumptions as Theorem 4.5, if we apply Algorithm 2 with adaptive learning rates* $\eta_t = \frac{\|\nabla f(W_t)\|_*}{rL}$, *we have*

$$f(W_t) - f^* \leq \frac{2rL(f(W_0) - f^*)D_{\text{op}}^2}{2rLD_{\text{op}}^2 + t(f(W_0) - f^*)}.$$

Thus, to reach the precision $f(W_T) - f^* \leq \epsilon$, *the complexity is* $O(rLD_{\text{op}}^2\epsilon^{-1})$.

4.1.1 Comparison with GD

Note that in the nonconvex case, the (S)GD's complexity of finding ϵ-Frobenius norm stationary points of f is $O(L\sigma^2\Delta\epsilon^{-4})$ in the stochastic setting and $O(L\Delta\epsilon^{-2})$ in the deterministic setting (Garrigos & Gower, 2023), which means (S)GD can also find ϵ-nuclear norm stationary points of f with $O(r^2L\sigma^2\Delta\epsilon^{-4})$ complexity in the stochastic setting and $O(rL\Delta\epsilon^{-2})$ in the deterministic setting according to Proposition 3.6. Moreover, to reach the precision $f(W_T)-f^* \leq \epsilon$ in the star convex case, the complexity of GD is $O(LD_F^2\epsilon^{-1})$ (Garrigos & Gower, 2023). We have $D_F = \|W_0 - W^*\|_F \leq \|W_0 - W^*\|_* \leq D_{\text{op}}$. Comparing the complexity results in Theorem 4.1 and Theorem 4.5, we can observe that under the conventional Frobenius norm Lipschitz smoothness assumption, the convergence rate of Muon is no better than that of (S)GD, and may even be worse.

4.2 Spectral norm Lipschitz smooth

Considering the matrix structure of the parameters and the properties of the Muon optimizer, the spectral norm Lipschitz smoothness assumption is natural in practice. In this subsection, we adopt Assumption 3.2 and establish the following theorems and corollaries.

Theorem 4.7. *Under Assumptions 3.2 and 3.3, if we apply Algorithm 1 with $\eta_t = \eta$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq \frac{\mathbb{E}[f(W_0)-f(W_T)]}{T\eta} + \frac{L_*\eta}{2} + \frac{2\sigma\sqrt{r(1-\beta)}}{\sqrt{(1+\beta)B}} + \frac{2\beta\sigma\sqrt{r}}{(1-\beta)T\sqrt{B}} + \frac{2\eta\beta L_*}{1-\beta}.$$

Corollary 4.8. *Under the same assumptions as Theorem 4.7, we have the following convergence guarantees for Algorithm 1:*

1. ***Stochastic setting:*** *If we set $B = 1$, $\eta = \sqrt{\frac{(1-\beta)\Delta}{TL_*}}$, and $1-\beta = \min\left\{\frac{\sqrt{L_*\Delta}}{\sigma\sqrt{rT}}, 1\right\}$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq O\left(\sqrt[4]{\frac{rL_*\Delta\sigma^2}{T}} + \sqrt{\frac{L_*\Delta}{T}} + \frac{r\sigma^2}{\sqrt{L_*\Delta T}}\right).$$

2. ***Deterministic setting ($\sigma = 0$):*** *If we set $\eta = \sqrt{\frac{\Delta}{TL_*}}$, and $\beta = O(1)$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(W_t)\|_* \leq O\left(\sqrt{\frac{L_*\Delta}{T}}\right).$$

Consequently, achieving an ϵ-nuclear norm stationary point of f requires a complexity of $O(rL_*\sigma^2\Delta\epsilon^{-4})$in stochastic setting and $O(L_*\Delta\epsilon^{-2})$ in deterministic setting.

Theorem 4.9. *Under Assumption 4.3, 3.2 and 4.4, if we apply Algorithm 2 with $\eta_t = \eta$, then*

$$f(W_T) - f^* \leq \left(1 - \frac{\eta}{D_{\text{op}}}\right)^T (f(W_0) - f^*) + \frac{L_* D_{\text{op}}\eta}{2}.$$

Thus, to reach the precision $f(W_T) - f^* \leq \epsilon$, we can take $\eta = O(\frac{\epsilon}{L_* D_{\text{op}}})$, and the complexity is $O(L_* D_{\text{op}}^2\epsilon^{-1} log\frac{\Delta}{\epsilon})$.

Theorem 4.10. *Under the same assumptions as Theorem 4.9, if we apply Algorithm 2 with adaptive learning rates $\eta_t = \frac{\|\nabla f(W_t)\|_*}{L_*}$, we have*

$$f(W_t) - f^* \leq \frac{2L_*(f(W_0) - f^*)D_{\text{op}}^2}{2L_* D_{\text{op}}^2 + t(f(W_0) - f^*)}.$$

Thus, to reach the precision $f(W_T) - f^* \leq \epsilon$, the complexity is $O(L_* D_{\text{op}}^2\epsilon^{-1})$.

Though we have derived the convergence theorems in terms of L_*, it remains unclear how to interpret them. For the Frobenius norm Lipschitz smoothness constant L, we know it is the largest singular value of the Hessian. The question, therefore, is what the relationship between L and L_* is. For a function f, its L and L_* are related to its Hessian $\nabla^2 f$. In neural network optimization, a widely observed phenomenon is that Hessians are typically low-rank (Sagun et al., 2016; 2017; Wu et al., 2020; Papyan, 2020; Yao et al., 2020) and are approximately blockwise diagonal (Dong et al., 2025; Zhang et al., 2024b;a; Collobert, 2004) (e.g., See Figure 2 in Zhang et al. (2024a) and Figure 1-3 in Dong et al. (2025)). Thus, we adopt Assumption 4.11 from An et al. (2025) to better illustrate the connections and differences between L and L_* in neural networks.

Assumption 4.11 (Λ-norm Lipschitz smooth)**.** *We say $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is 1-Λ-norm Lipschitz smooth with a positive definite matrix $\Lambda \in \mathbb{R}^{n\times n}$, if for any $W, W' \in \mathbb{R}^{m\times n}$,*

$$\|\nabla f(W) - \nabla f(W')\|_{\Lambda^{-1}} \le \|W - W'\|_{\Lambda}.$$

Assumption 4.11 is closely related to the blockwise diagonal structure of the Hessian. Actually, when f is twice continuously differentiable, Assumption 4.11 is equivalent to the following assumption of Hessian:

$$-I_m \otimes \Lambda = -\begin{bmatrix} \Lambda & & & \\ & \Lambda & & \\ & & \dots & \\ & & & \Lambda \end{bmatrix} \preceq \nabla^2 f_v(w) \preceq \begin{bmatrix} \Lambda & & & \\ & \Lambda & & \\ & & \dots & \\ & & & \Lambda \end{bmatrix} = I_m \otimes \Lambda \in \mathbb{R}^{mn\times mn}, \tag{2}$$

where $f_v(w) = f(W)$ and $w = \overline{\mathrm{vec}}(W) \in \mathbb{R}^{mn}$. The proof of the equivalence of these two assumptions can be found in Lemma A.3. Assumption 4.11 has also been used in An et al. (2025) to analyze matrix-structured optimization. Moreover, we have the following lemma to establish the relationship between Assumptions 3.1, 3.2, and 4.11, whose proof can be found in Appendix A.

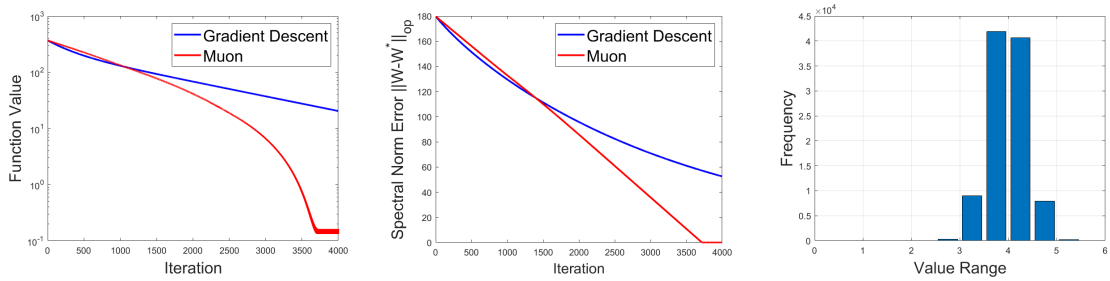
Lemma 4.12. *If $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is 1-Λ-norm Lipschitz smooth with a positive definite matrix $\Lambda \in \mathbb{R}^{n\times n}$, then f is also $\|\Lambda\|_{\mathrm{op}}$ Frobenius norm Lipschitz smooth and $\|\Lambda\|_*$ spectral norm Lipschitz smooth.*

4.2.1 Comparison with GD

In this section, we first argue that Muon inherently exploits blockwise diagonal structures and relatively low-rank properties. For example, if the Hessians of f have a blockwise diagonal structure and are relatively "low rank", i.e. if it satisfies Assumption 4.11 with Λ such that $\|\Lambda\|_{\mathrm{op}} \approx \|\Lambda\|_* \ll r\|\Lambda\|_{\mathrm{op}}$, then the convergence rate of Muon for finding the nuclear norm stationary points can be better than that of (S)GD in the nonconvex case, i.e. $O(\|\Lambda\|_*\Delta\epsilon^{-2}) \ll O(r\|\Lambda\|_{\mathrm{op}}\Delta\epsilon^{-2})$ and $O(r\|\Lambda\|_*\sigma^2\Delta\epsilon^{-4}) \ll O(r^2\|\Lambda\|_{\mathrm{op}}\sigma^2\Delta\epsilon^{-4})$.

In the star convex case, recall that the complexity of GD is $O(LD_{\mathrm{F}}^2\epsilon^{-1})$. Hence, to compare the convergence rate of Muon and GD, we need to compare LD_{F}^2 with $L_*D_{\mathrm{op}}^2$. Similarly, if f satisfies Assumption 4.11 with Λ such that $\|\Lambda\|_{\mathrm{op}} \approx \|\Lambda\|_* \ll r\|\Lambda\|_{\mathrm{op}}$, then $L_*D_{\mathrm{op}}^2 \le \|\Lambda\|_*D_{\mathrm{op}}^2 \ll r\|\Lambda\|_{\mathrm{op}}D_{\mathrm{op}}^2$. In the experiments, we usually find that $\|W_t - W^*\|_{\mathrm{op}}$ decreases as t increases (Figure 1(b)). Thus, we assume $D_{\mathrm{op}} \approx \|W_0 - W^*\|_{\mathrm{op}}$. Despite the success of LoRA (Hu et al., 2022) in fine-tuning foundation models indicating a relatively low rank of $W_0 - W^*$ is enough for fine-tuning, many empirical evidences (Lialin et al., 2024; Jiang et al., 2024; Huang et al., 2025) demonstrate that a relatively high rank of $W_0 - W^*$ is needed for pretraining and some complex fine-tuning tasks for large foundation models, and a low rank of $W_0 - W^*$ can lead to degraded performance in those scenarios. Thus, we assume $W_0 - W^*$ is relatively "high rank". Then, we can expect $r\|W_0 - W^*\|_{\mathrm{op}}^2 \approx \|W_0 - W^*\|_{\mathrm{F}}^2$ and $L_*D_{\mathrm{op}}^2 \ll r\|\Lambda\|_{\mathrm{op}}D_{\mathrm{op}}^2 \approx LD_{\mathrm{F}}^2$, which means Muon can have a better performance than GD.

To further evaluate this argument, we conduct experiments comparing Muon and GD on a quadratic function $f(W) = \mathrm{tr}\left((W - W^*)^\top Q(W - W^*)\right)$, where Q is a positive definite matrix with a poor condition number and is relatively "low rank", i.e. $\|Q\|_* \approx \|Q\|_{\mathrm{op}}$. Note that the Hessian of f is $I \otimes Q$, which satisfies Assumption 4.11. We randomly choose W^*, and set $W_0 = 0$. We report the results in Figure 1. We can see that Muon outperforms GD and most samples of ratio $\frac{D_{\mathrm{F}}^2 L}{D_{\mathrm{op}}^2 L_*}$ are larger than 3, which validates the theoretical findings.



(a) Function Value (b) $\|W_t - W^*\|_{\mathrm{op}}$ (c) $D_{\mathrm{F}}^2 L/(D_{\mathrm{op}}^2 L_*)$

Figure 1: Experiments on a quadratic function $f(W) = \mathrm{tr}\left((W - W^*)^\top Q(W - W^*)\right)$. Detailed settings can be found in Appendix E.

Next, we consider a more realistic example to illustrate how the low-rank structure of the Hessian arises and how Muon can benefit from it. Specifically, we consider a multi-class classification problem with dataset $\{(x_i, y_i)\}_{i=1}^B$, where $x_i \in \mathbb{R}^d$ is the feature data, $y_i \in \mathbb{R}^c$ is the corresponding c-dimensional one-hot vector of the label, c is the number of classes and B is the number of samples. We first consider a linear model $g(W; x) = Wx \in \mathbb{R}^c$, where $W \in \mathbb{R}^{c\times d}$ is the parameter, and its mean-square (MSE) loss. Denote $X = (x_1, \ldots, x_B) \in \mathbb{R}^{d\times B}$, $Y = (y_1, \ldots, y_B) \in \mathbb{R}^{c\times B}$. The MSE loss can be defined as follows:

$$f(W) = \frac{1}{2B}\|WX - Y\|_{\mathrm{F}}^2. \tag{3}$$

We can calculate the Hessian of this loss as

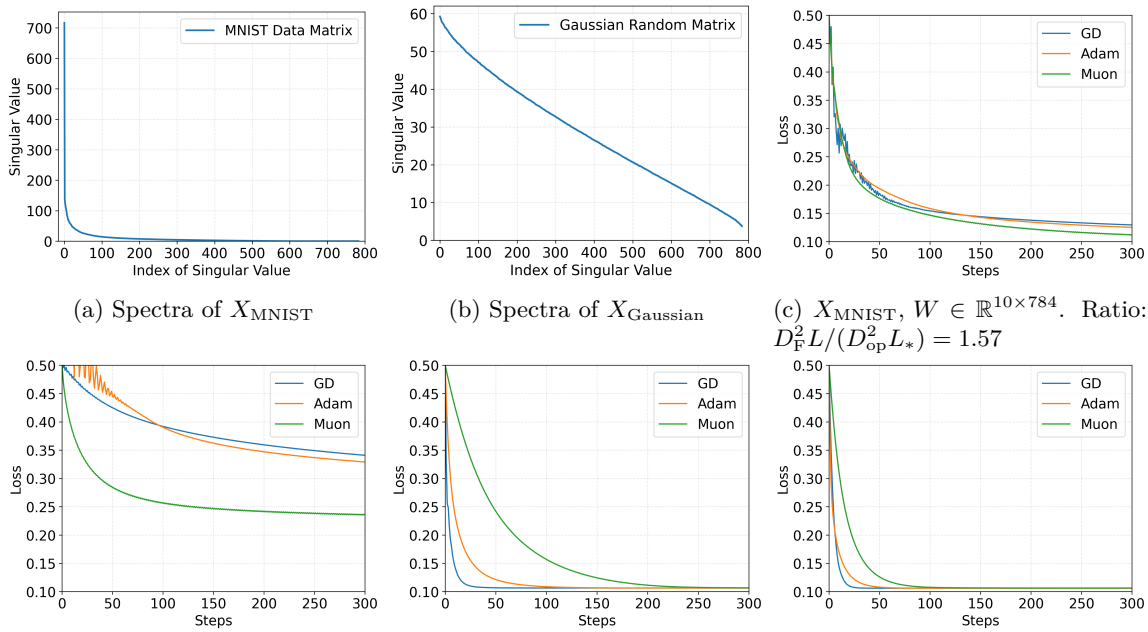
$$\nabla^2 f_v(\overline{\mathrm{vec}}(W)) = I_c \otimes XX^\top. \tag{4}$$

We note that this Hessian is block-diagonal with c blocks and satisfy Assumption 4.11 with $\Lambda = XX^\top$. We can calculate $L = \|XX^\top\|_{\mathrm{op}} = \|X\|_{\mathrm{op}}^2$ and $L_* = \|XX^\top\|_* = \|X\|_{\mathrm{F}}^2$.

Actually, we find that the feature matrix X in practical machine learning problem is typically very "low rank", or more precisely, their singular values are highly concentrated (See Figure 2(a) and additional examples in Appendix D) such that $\|X\|_{\mathrm{op}} \approx \|X\|_{\mathrm{F}}$. For example, we randomly select 1000 samples from training datasets of MNIST (LeCun et al., 1998) and construct $X_{\mathrm{MNIST}} \in \mathbb{R}^{784\times 1000}$. Our calculation shows that $\|X_{\mathrm{MNIST}}\|_{\mathrm{F}}^2/\|X_{\mathrm{MNIST}}\|_{\mathrm{op}}^2 = 1.41$, which means $L_* \approx L$, the Hessian of this loss is extremely low-rank. Note that this phenomenon is not limited to MNIST; similar low-rank behavior can also be observed in other real-world datasets (e.g., CIFAR-10 and text data), see Appendix D.

According to the theory, Muon can benefit from this low-rank structure and potentially converge faster than GD, thus, to verify this, we conduct the following experiments. We first use X_{MNIST} and their corresponding 1000 labels y from the training dataset as $Y \in \mathbb{R}^{10\times 1000}$ to optimize $W \in \mathbb{R}^{10\times 784}$ under the MSE loss in (3). We compare GD (with/without Nesterov momentum), Adam, and Muon, tuning all hyperparameters for each method. Results in Figure 2(c) show that Muon outperforms GD and Adam. Then, we compute the Frobenius norm and operator norm of $W_0 - W^*$ (where W^* is selected as the final converged point), denoted as D_F and D_{op}, respectively, and calculate the ratio $\frac{D_{\mathrm{F}}^2 L}{D_{\mathrm{op}}^2 L_*}$. This ratio equals 1.57, exceeding 1 and thus supporting Muon's advantage, though it is not large, explaining the modest empirical gain in Figure 2(c). One reason is that the intrinsic dimension of $W \in \mathbb{R}^{10\times 784}$ limits its rank to be relatively low, which in turn leads to a relatively small value of $D_{\mathrm{F}}^2/D_{\mathrm{op}}^2$. To amplify this effect, we replace Y with 1000 random one-hot vectors over 100 classes, yielding $W \in \mathbb{R}^{100\times 784}$. We then conduct the same experiment comparing GD, Adam, and Muon. The results are shown in Figure 2(d). We can observe that the ratio is now larger, reaching 3.52. Correspondingly, Muon more significantly outperforms GD and Adam in the figure, which is consistent with our theoretical predictions.

While, if we construct $X_{\mathrm{Gaussian}} \in \mathbb{R}^{784\times 1000}$ as a Gaussian random matrix, then we can find that the singular values of this random matrix are more evenly distributed (See Figure 2(b)) and $\|X_{\mathrm{Gaussian}}\|_{\mathrm{F}}^2/\|X_{\mathrm{Gaussian}}\|_{\mathrm{op}}^2 =$



(a) Spectra of X_{MNIST}

(b) Spectra of X_{Gaussian}

(c) X_{MNIST}, $W \in \mathbb{R}^{10\times 784}$. Ratio: $D_{\text{F}}^2 L/(D_{\text{op}}^2 L_*) = 1.57$

(d) X_{MNIST}, $W \in \mathbb{R}^{100\times 784}$. Ratio: $D_{\text{F}}^2 L/(D_{\text{op}}^2 L_*) = 3.52$

(e) X_{Gaussian}, $W \in \mathbb{R}^{10\times 784}$. Ratio: $D_{\text{F}}^2 L/(D_{\text{op}}^2 L_*) = 0.0122$

(f) X_{Gaussian}, $W \in \mathbb{R}^{100\times 784}$. Ratio: $D_{\text{F}}^2 L/(D_{\text{op}}^2 L_*) = 0.0232$

Figure 2: (a, b): Spectra of X_{MNIST} and X_{Gaussian}. (c, d, e, f): Optimizing (3) with different X and Y.

224.06, which means its L_* is significantly larger than L and indicating Muon will outperform worser than GD. Our experiments (Figure 2(e, f)) validated our theoretical predictions. Notably, when we increase the dimension of y from 10 to 100, the corresponding ratio rises from 0.0122 to 0.0232, and the performance gap between Muon and GD (Adam) becomes smaller. As shown in Figure 2(c, d, e, f), Muon performs better relative to GD when this ratio is larger, which is consistent with our theoretical analysis. Meanwhile, our experiments also suggest that the low-rank property of the Hessian in practical machine learning problems may partially stem from the fact that real-world data are not Gaussian random, but instead exhibit some inherent low-rank structure.

4.3 Without uniform Lipschitz smoothness assumptions

Though the previous subsections offer some insight into how Muon can outperform GD, Assumption 4.11 appears overly restrictive. To further investigate how the structure and dynamics of the Hessian matrix affect the convergence of Muon and GD in a more fine-grained manner, we consider the following assumption and analysis.

Assumption 4.13. *We assume the norm of $\nabla^3 f$ is bounded, i.e., we assume for any $W \in \mathbb{R}^{m\times n}$, $\|\nabla^3 f_v(w)\|_{\text{op}} \le s$. Here $f_v(w) = f(W)$ and $w = \overline{\text{vec}}(W) \in \mathbb{R}^{mn}$.*

Theorem 4.14. *Under Assumption 4.13, if we apply Algorithm 2 with $\eta_t = \eta$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(W_t)\|_* \le \frac{f(W_0) - f(W_T)}{T\eta} + \frac{\eta J}{2} + \frac{s\eta^2 r^{3/2}}{6}$$

where $J = \frac{1}{T}\sum_{t=0}^{T-1} J_t$, $J_t = \overline{\text{vec}}(U_t V_t^\top)^\top H_t \overline{\text{vec}}(U_t V_t^\top)$, $H_t = \nabla^2 f_v(\overline{\text{vec}}(W_t))$, and $f_v(\overline{\text{vec}}(W_t)) = f(W_t)$.

If $J > 0$[1], and $\eta = O\left(\sqrt{\frac{\Delta}{JT}}\right)$, then

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(W_t)\|_* \leq \left(\sqrt{\frac{J\Delta}{T}} + \frac{sr^{3/2}\Delta}{JT}\right).$$

Thus, it is possible that Muon can find an ϵ-nuclear norm stationary point of f with a complexity of $O(J\Delta\epsilon^{-2})$.

Similarly, for the star convex case, we have

Theorem 4.15. *Under Assumptions 4.3, 4.13 and 4.4, if we apply Algorithm 2 with $\eta_t = \eta$, we have*

$$f(W_T) - f^* \leq \left(1 - \frac{\eta}{D_{\text{op}}}\right)^T \Delta + \frac{\eta^2 \tilde{J} T}{2} + \frac{s\eta^2 D_{\text{op}} r^{3/2}}{6}.$$

where $\tilde{J} = \frac{1}{T}\sum_{t=0}^{T-1}\left(1 - \frac{\eta}{D_{\text{op}}}\right)^{T-1-t} J_t$, $J_t = \overline{\text{vec}}(U_t V_t^\top)^\top H_t \overline{\text{vec}}(U_t V_t^\top)$, $H_t = \nabla^2 f_v(\overline{\text{vec}}(W_t))$, *and* $f_v(\overline{\text{vec}}(W_t)) = f(W_t)$.

Note that $\tilde{J}$ can be viewed as a weighted average of J_t and $\tilde{J} \leq \frac{1}{T}\sum_{t=0}^{T-1}|J_t|$. If $\tilde{J} > 0$[1] and $\eta = \min\left\{\frac{D_{\text{op}}}{T}\log\left(\frac{T\Delta}{D_{\text{op}}^2\tilde{J}}\right), D_{\text{op}}\right\}$, then

$$\begin{aligned} f(W_T) - f^* \leq & \frac{D_{\text{op}}^2\tilde{J}}{T} + \frac{D_{\text{op}}^2\tilde{J}}{2T}\left[\log\left(\frac{T\Delta}{D_{\text{op}}^2\tilde{J}}\right)\right]^2 + \frac{sr^{3/2}D_{\text{op}}^3}{6T^2}\left[\log\left(\frac{T\Delta}{D_{\text{op}}^2\tilde{J}}\right)\right]^2 \\ \leq & \tilde{O}\left(\frac{D_{\text{op}}^2\tilde{J}}{T}\right). \end{aligned}$$

Thus, it is possible that Muon can reach the precision $f(W_T) - f^* \leq \epsilon$ with a complexity of $\tilde{O}(\tilde{J}D_{\text{op}}^2\epsilon^{-1})$.

4.3.1 Comparison with GD

Therefore, the key to comparing Muon and GD lies in comparing the relationship between J (or $\tilde{J}$) and L. Note that J is the average of J_t, while $\tilde{J}$ is bounded above by the average of $|J_t|$. Hence, we may first analyze J_t. Note that

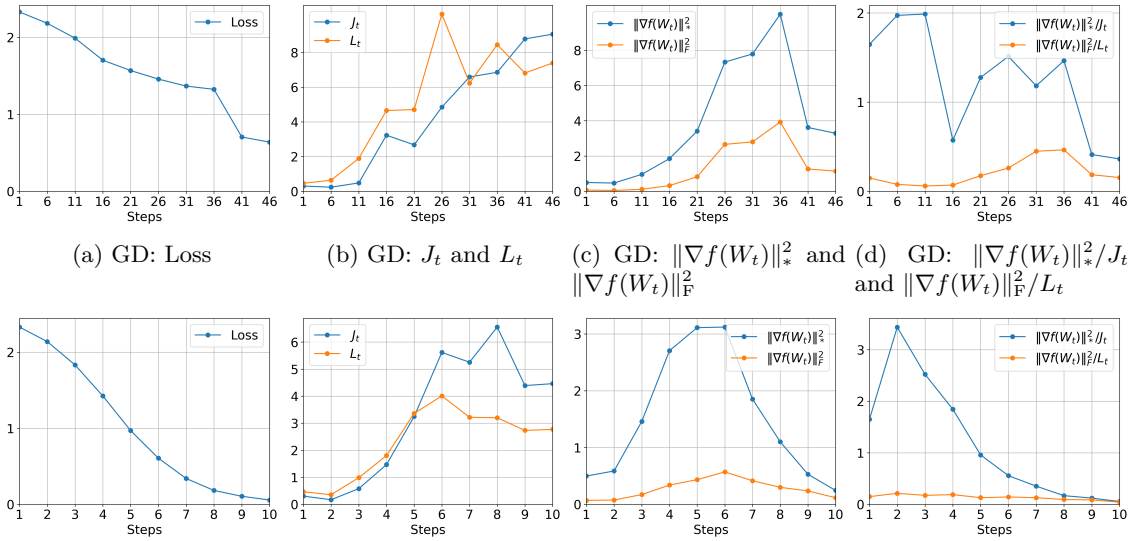
$$\begin{aligned} J_t &= \overline{\text{vec}}(U_t V_t^\top)^\top H_t \overline{\text{vec}}(U_t V_t^\top) \\ &= \overline{\text{vec}}(I_{r_t})^\top (U_t^\top \otimes V_t^\top) H_t (U_t \otimes V_t)\overline{\text{vec}}(I_{r_t}) \\ &= \sum_{i,j\in[r_t]} [(U_t^\top \otimes V_t^\top) H_t (U_t \otimes V_t)]_{(i-1)*r_t+i,(j-1)*r_t+j}. \end{aligned} \tag{5}$$

Thus, J_t is the sum of r_t^2 elements of $A_t \triangleq (U_t^\top \otimes V_t^\top)H_t(U_t \otimes V_t)$, where $A_t \in \mathbb{R}^{r_t^2 \times r_t^2}$ can be viewed as a representation of H_t under a certain congruence transformation depending on U_t and V_t.

To better illustrate the relationship between J_t and L, we assume that H_t can be expressed as a Kronecker product, i.e., $H_t = P_t \otimes Q_t$, where $P_t \in \mathbb{R}^{m\times m}$ and $Q_t \in \mathbb{R}^{n\times n}$. Let $\sigma_{p,t,1} \geq \sigma_{p,t,2} \geq \cdots \geq \sigma_{p,t,m}$ denote the singular values of P_t, and $\sigma_{q,t,1} \geq \sigma_{q,t,2} \geq \cdots \geq \sigma_{q,t,n}$ denote the singular values of Q_t. Then, we have $A_t = (U_t^\top P_t U_t) \otimes (V_t^\top Q_t V_t)$ and $J_t = \langle U_t^\top P_t U_t, V_t^\top Q_t V_t\rangle \leq \sum_{i=1}^{r_t}\sigma_{p,t,i}\sigma_{q,t,i}$. Here, we use the Von Neumann's trace inequality.

Note that in the nonconvex case, the complexity of GD for ϵ-nuclear norm stationary point is $O(rL\Delta\epsilon^{-2})$, where L is at least larger than $\max_t L_t$ with $L_t \triangleq \|H_t\|_{\text{op}} = \|P_t\|_{\text{op}}\|Q_t\|_{\text{op}}$. Thus, when H_t can be represented

[1] We find that usually $J_t > 0$ in experiments. Thus, we mainly discuss the situations when $J > 0$ ($\tilde{J} > 0$) in the main paper. When $J \leq 0$ ($\tilde{J} \leq 0$), there is a better convergence rate regarding T; see discussions in Appendix B.3 and C.5.



(a) GD: Loss (b) GD: J_t and L_t (c) GD: $\|\nabla f(W_t)\|_*^2$ and $\|\nabla f(W_t)\|_{\mathrm{F}}^2$ (d) GD: $\|\nabla f(W_t)\|_*^2/J_t$ and $\|\nabla f(W_t)\|_{\mathrm{F}}^2/L_t$

(e) Muon: Loss (f) Muon: J_t and L_t (g) Muon: $\|\nabla f(W_t)\|_*^2$ and $\|\nabla f(W_t)\|_{\mathrm{F}}^2$ (h) Muon: $\|\nabla f(W_t)\|_*^2/J_t$ and $\|\nabla f(W_t)\|_{\mathrm{F}}^2/L_t$

Figure 3: Comparison of J_t, L_t, $\|\nabla f(W_t)\|_*^2$ and $\|\nabla f(W_t)\|_{\mathrm{F}}^2$ over the training process of GD and Muon (Algorithm 2) on a MLP model with three matrix parameters $W^1 \in \mathbb{R}^{128\times 784}, W^2 \in \mathbb{R}^{64\times 128}, W^3 \in \mathbb{R}^{10\times 64}$. We show the gradients and Hessians with respect to W^2 in this Figure. Detailed settings can be found in Appendix E.

by $P_t \otimes Q_t$ and Q_t, P_t are relatively low-rank such that $\sum_{i=1}^{r_t} \sigma_{p,t,i}\sigma_{q,t,i} \ll r\sigma_{p,t,1}\sigma_{q,t,1}$, then $J = \frac{1}{T}\sum_t J_t \leq \frac{1}{T}\sum_t \sum_{i=1}^{r_t} \sigma_{p,t,i}\sigma_{q,t,i} \ll \frac{1}{T}\sum_t rL_t \leq rL$, and the convergence rate of Muon can be better than GD's. (Or one of Q_t, P_t is relatively low-rank, i.e. if Q_t is relatively low-rank such that $\|Q_t\|_{\mathrm{op}} \approx \|Q_t\| \ll r\|Q_t\|_{\mathrm{op}}$, we can also have $J = \frac{1}{T}\sum_t J_t \leq \frac{1}{T}\sum_t \|P_t\|_{\mathrm{op}}\|Q_t\|_* \ll r\max_t L_t \leq rL$.)

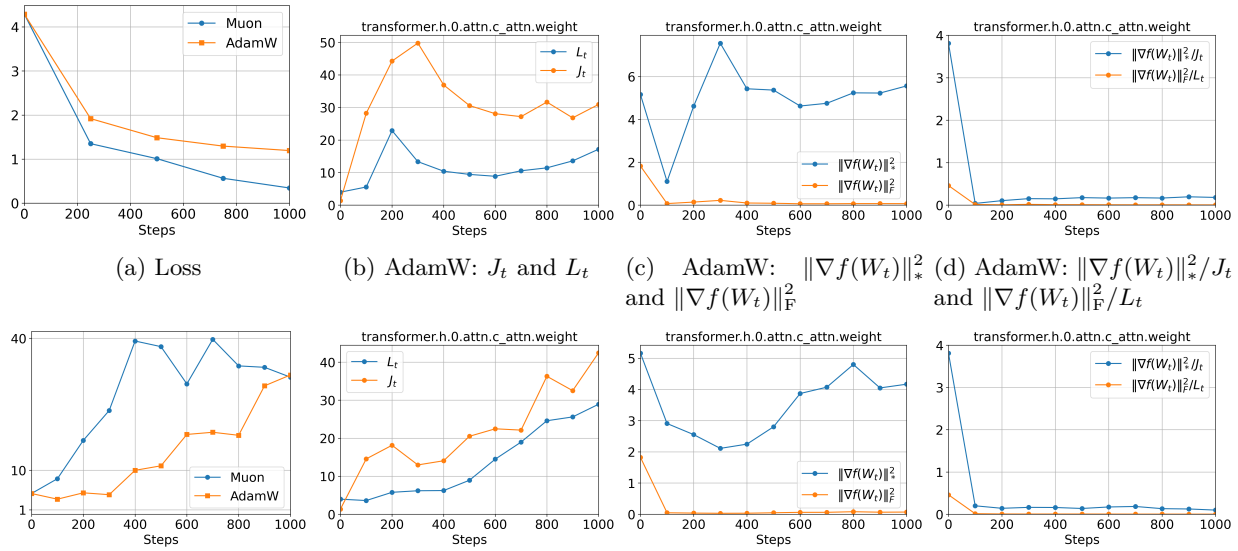
Similar to the previous discussions, in the star convex case, if $W_0 - W^*$ is relatively "high rank" such that $rD_{\mathrm{op}}^2 \approx \|W_0 - W^*\|_{\mathrm{F}}^2$ and Q_t, P_t are relatively low-rank such that $\tilde{J} \ll rL$, then we have $\tilde{J}D_{\mathrm{op}}^2 \ll LD_{\mathrm{F}}^2$ and Muon can have a better performance than GD. For stochastic case and Algorithm 1, we have similar conclusions, see Appendix B.4 for more detailed discussions.

We introduce some examples to illustrate the relationship between L and J. Similar to the last subsection, we consider the same multi-class classification dataset $\{(x_i, y_i)\}_{i=1}^B$ with a linear model $g(W; x) = Wx \in \mathbb{R}^c$, where $W \in \mathbb{R}^{c\times d}$. We have discussed its MSE loss (3) in the last subsection, whose Hessian is given by $\nabla^2 f_v(\overline{\mathrm{vec}}(W)) = I_c \otimes XX^\top$, which has a Kronecker product structure. Thus, we have $J(\text{or } \tilde{J}) \leq \|I_c\|_{\mathrm{op}}\|XX^\top\|_* = \|X\|_{\mathrm{F}}^2 = L_*$ and the conclusion is similar to the last subsection. Here, we consider a more general case; for example, when the loss function with respect to W can be represented as $f(W) = \frac{1}{B}\sum_{i=1}^B l(Wh(x_i))$, where $h : \mathbb{R}^d \to \mathbb{R}^n$, $W \in \mathbb{R}^{m\times n}$, $l : \mathbb{R}^m \to \mathbb{R}$. Then, we have

$$\nabla^2 f_v(\overline{\mathrm{vec}}(W)) = \frac{1}{B}\sum_{i=1}^B \nabla^2 l(Wh(x_i)) \otimes (h(x_i)h(x_i)^\top). \tag{6}$$

and $J(\text{or } \tilde{J}) \leq \frac{1}{T}\sum_{t=1}^T \|\nabla^2 l(W_t h(x))\|_{\mathrm{op}}\|h(x)\|_2^2 \leq \max_t L_t \leq L$. Note that when $B = 1$, we have $\nabla^2 f_v(\overline{\mathrm{vec}}(W_t)) = \nabla^2 l(W_t h(x_1)) \otimes (h(x_1)h(x_1)^\top)$, and this Hessian can be represented as a Kronecker product. Thus, we have $L \geq \max_t L_t = \|\nabla^2 l(W_t h(x_1))\|_{\mathrm{op}}\|h(x_1)h(x_1)^\top\|_{\mathrm{op}} = \max_t \|\nabla^2 l(W_t h(x_1))\|_{\mathrm{op}}\|h(x_1)\|_2^2$ and $J(\text{or } \tilde{J}) \leq \frac{1}{T}\sum_{t=1}^T \|\nabla^2 l(W_t h(x_1))\|_{\mathrm{op}}\|h(x_1)h(x_1)^\top\|_* = \frac{1}{T}\sum_{t=1}^T \|\nabla^2 l(W_t h(x_1))\|_{\mathrm{op}}\|h(x_1)\|_2^2$. Thus, in single sample case, $J(\text{or } \tilde{J}) \leq L$, demonstrating the significant advantage of Muon compared to GD.

For multi sample case $B > 1$ and more general nonconvex loss functions, theoretical analyses of L and J are difficult. Here, we conduct experiments to calculate the L_t and J_t alongside the optimization trajectory. Moreover, various studies (Gur-Ari et al., 2018; Zhao et al., 2021; 2024; Cosson et al., 2023) have shown that the gradients during neural network optimization are also typically low rank. Thus, when we convert the



(a) Loss (b) AdamW: J_t and L_t (c) AdamW: $\|\nabla f(W_t)\|_*^2$ and $\|\nabla f(W_t)\|_{\mathrm{F}}^2$ (d) AdamW: $\|\nabla f(W_t)\|_*^2/J_t$ and $\|\nabla f(W_t)\|_{\mathrm{F}}^2/L_t$

(e) Average ratio (f) Muon: J_t and L_t (g) Muon: $\|\nabla f(W_t)\|_*^2$ and $\|\nabla f(W_t)\|_{\mathrm{F}}^2$ (h) Muon: $\|\nabla f(W_t)\|_*^2/J_t$ and $\|\nabla f(W_t)\|_{\mathrm{F}}^2/L_t$

Figure 4: Comparison of J_t, L_t, $\|\nabla f(W_t)\|_*^2$, $\|\nabla f(W_t)\|_{\mathrm{F}}^2$ and the average ratio over the training process of a six-layer GPT-2 style model with AdamW and Muon. (e): The average ratio ($\|\nabla f(W_t)\|_*^2 L_t/(\|\nabla f(W_t)\|_{\mathrm{F}}^2 J_t)$) of all matrix parameters optimized by Muon. In fact, not only does the average ratio of all matrix parameters satisfy Equation (7), but every matrix parameter optimized by Muon also satisfies Equation (7) at the steps tested in our experiments. Due to space limitations, in (b, c, d, f, g, h) we present records for the parameter "transformer.h.0.attn.c_attn.weight," which is the attention matrix of the first attention layer. Detailed settings can be found in Appendix E.

stationary point of Frobenius norm to nuclear norm, the additional coefficient can be much smaller than $\sqrt{r}$. Therefore, to compare the convergence rates of Muon and GD more precisely and more fairly, one should examine the ratios between the nuclear norm and Frobenius norm of their gradients as well as J_t and L_t. For example, we can examine the following inequality:

$$\frac{J_t}{L_t} \leq \frac{\|\nabla f(W_t)\|_*^2}{\|\nabla f(W_t)\|_{\mathrm{F}}^2}, \tag{7}$$

If this inequality consistently holds alongside the optimization trajectory, Muon is expected to converge faster than GD.

We investigate and validate (7) during the optimization of a three-layer Multi-Layer Perceptron (MLP) trained on the MNIST dataset, as well as a six-layer GPT-2 style model trained on the Shakespeare character-level corpus. We optimize the MLP model using GD and Muon (Algorithm 2), and the GPT-2 model using AdamW and Muon. During training, we record the loss, J_t, L_t, $\|\nabla f(W_t)\|_*^2$, $\|\nabla f(W_t)\|_{\mathrm{F}}^2$ and calculate the ratio (7). The results are shown in Figure 3 and 4, with detailed settings provided in Appendix E. In both experiments (Figure 3 and Figure 4), we observe that Muon's ratio ($\|\nabla f(W_t)\|_*^2/J_t$) are consistently larger than GD's ($\|\nabla f(W_t)\|_{\mathrm{F}}^2/L_t$), demonstrating the advantage of Muon.

5 Conclusions and future directions

In this work, we have presented a comprehensive convergence rate analysis of Muon under various assumptions. We provided detailed comparisons between Muon and GD, and established the conditions under which Muon can outperform GD with different assumptions. We showed that Muon can exploit the low-rank structure of Hessian matrices, a property commonly observed during neural network training. Our experiments on neural networks and quadratic regressions supported and corroborated the theoretical findings. Promising directions for future research include investigating the structure of Hessian matrices from a theoretical perspective and leveraging these structural properties to develop new or more effective optimization methods.

References

Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, et al. Gpt-4 technical report, 2024. *URL https://arxiv. org/abs/2303.08774*, 2:6, 2024.

Kwangjun Ahn and Byron Xu. Dion: A communication-efficient optimizer for large models. *arXiv preprint arXiv:2504.05295*, 2025.

Kang An, Yuxing Liu, Rui Pan, Shiqian Ma, Donald Goldfarb, and Tong Zhang. Asgo: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Ronan Collobert. Large scale machine learning. 2004.

Romain Cosson, Ali Jadbabaie, Anuran Makur, Amirhossein Reisizadeh, and Devavrat Shah. Low-rank gradient descent. *IEEE Open Journal of Control Systems*, 2:380–395, 2023.

Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025.

Zhaorui Dong, Yushun Zhang, Zhi-Quan Luo, Jianfeng Yao, and Ruoyu Sun. Towards quantifying the hessian structure of neural networks. *arXiv preprint arXiv:2505.02809*, 2025.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Qiushi Huang, Tom Ko, Zhan Zhuang, Lilian Tang, and Yu Zhang. HiRA: Parameter-efficient hadamard high-rank adaptation for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ting Jiang, Shaohan Huang, Shengyue Luo, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, et al. Mora: High-rank updating for parameter-efficient fine-tuning. *arXiv preprint arXiv:2405.12130*, 2024.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *The 3rd International Conference on Learning Representations*, 2015.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International conference on machine learning*, pp. 2698–2707. PMLR, 2018.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jiaxiang Li and Mingyi Hong. A note on the convergence of muon and further. *arXiv preprint arXiv:2502.02900*, 2025.

Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. ReLoRA: High-rank training through low-rank updates. In *The Twelfth International Conference on Learning Representations*, 2024.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025a.

Liming Liu, Zhenghao Xu, Zixuan Zhang, Hao Kang, Zichong Li, Chen Liang, Weizhu Chen, and Tuo Zhao. Cosmos: A hybrid adaptive optimizer for memory-efficient training of llms. *arXiv preprint arXiv:2502.17410*, 2025b.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in muon. *arXiv preprint arXiv:2601.13474*, 2026.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

Depen Morwani, Itai Shapira, Nikhil Vyas, Eran Malach, Sham Kakade, and Lucas Janson. A new perspective on shampoo's preconditioner. *arXiv preprint arXiv:2406.17748*, 2024.

Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical programming*, 108(1):177–205, 2006.

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.

Yi Ren and Donald Goldfarb. Tensor normal training for deep learning models. *Advances in Neural Information Processing Systems*, 34:26040–26052, 2021.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018.

Weijie Su. Isotropic curvature model for understanding deep learning optimization: Is gradient orthogonalization optimal? *arXiv preprint arXiv:2511.00674*, 2025.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent YF Tan. Muon outperforms adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

Yikai Wu, Xingyu Zhu, Chenwei Wu, Annie Wang, and Rong Ge. Dissecting hessian: Understanding common structure of hessian in neural networks. *arXiv preprint arXiv:2010.04261*, 2020.

Shuo Xie, Tianhao Wang, Sashank Reddi, Sanjiv Kumar, and Zhiyuan Li. Structured preconditioners in adaptive optimization: A unified analysis. *arXiv preprint arXiv:2503.10537*, 2025.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pp. 581–590. IEEE, 2020.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhiquan Luo. Why transformers need adam: A hessian perspective. *Advances in Neural Information Processing Systems*, 37:131786–131823, 2024a.

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. *arXiv preprint arXiv:2406.16793*, 2024b.

Jiawei Zhao, Florian Tobias Schaefer, and Anima Anandkumar. Zero initialization: Initializing residual networks with only zeros and ones. *arXiv preprint arXiv:2110.12661*, 2021.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. Sgd converges to global minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*, 2019.

Appendix

The Appendix is organized as follows. In Appendix A, we introduce some lemmas that will be utilized in the subsequent proofs, and we also give the proofs of Lemma 4.12 and Lemma A.3. In Appendix B, we present the proofs of theorems in the nonconvex setting. In Appendix C, we present the proofs of theorems in the star convex setting. In Appendix D, we present additional examples that illustrate the low-rank structure of real-world data matrices. In Appendix E, we provide the detailed settings of our experiments.

A Lemmas

Lemma A.1 (Lemma 8 in An et al. (2025))**.** *For a symmetric positive definite matrix $\Lambda \in \mathbb{R}^{n\times n}$ and matrix $A \in \mathbb{R}^{m\times n}$, it holds that*

$$\|A\|_* \le \sqrt{\|\Lambda\|_*}\|A\|_{\Lambda^{-1}}.$$

Lemma A.2. *For a symmetric positive definite matrix $\Lambda \in \mathbb{R}^{n\times n}$ and matrix $A \in \mathbb{R}^{m\times n}$, it holds that*

$$\|A\|_{\mathrm{F}} \le \sqrt{\|\Lambda\|_{\mathrm{op}}}\|A\|_{\Lambda^{-1}}, \tag{8}$$

$$\|A\|_{\Lambda} \le \sqrt{\|\Lambda\|_{\mathrm{op}}}\|A\|_{\mathrm{F}}, \tag{9}$$

$$\|A\|_{\Lambda} \le \sqrt{\|\Lambda\|_*}\,\|A\|_{\mathrm{op}}. \tag{10}$$

Proof. We first prove Equation (8) and Equation (9).

Suppose the spectral decomposition of Λ is $\Lambda = U\Sigma U^\top$, where U is an orthogonal matrix, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n$ being the ordered singular values of Λ. Denote $AU = B = (b_1\, b_2\, \ldots\, b_n)$, where b_i is the i-th column of B. Then, we have

$$\|A\|_\Lambda^2 = \mathrm{tr}\left(A\Lambda A^\top\right) = \mathrm{tr}\left(AU\Sigma U^\top A^\top\right) = \mathrm{tr}\left(\Sigma B^\top B\right) = \sum_{i=1}^n \sigma_i \|b_i\|_2^2.$$

Note that

$$\|A\|_{\mathrm{F}}^2 = \sum_{i=1}^n \|b_i\|_2^2.$$

Thus, we have

$$\|A\|_\Lambda \le \sqrt{\sigma_1}\|A\|_{\mathrm{F}} = \sqrt{\|\Lambda\|_{\mathrm{op}}}\|A\|_{\mathrm{F}},$$

and

$$\|A\|_\Lambda \ge \sqrt{\sigma_n}\|A\|_{\mathrm{F}}.$$

Note that $\sigma_n^{-1} = \left\|\Lambda^{-1}\right\|_{\mathrm{op}}$. Thus,

$$\|A\|_{\mathrm{F}} \le \sqrt{\sigma_n^{-1}}\|A\|_\Lambda = \left\|\Lambda^{-1}\right\|_{\mathrm{op}}\|A\|_\Lambda. \tag{11}$$

Since Equation (11) holds for any Λ, we can define $\Lambda' = \Lambda^{-1}$ and get

$$\|A\|_{\mathrm{F}} \le \|\Lambda'\|_{\mathrm{op}}\,\|A\|_{\Lambda'^{-1}}. \tag{12}$$

Thus, we have proved Equation (8) and Equation (9). For Equation (10), we can prove it with the following inequalities

$$\|A\|_\Lambda = \sqrt{\mathrm{tr}\left(A\Lambda A^\top\right)} = \sqrt{\mathrm{tr}\left(A^\top A\Lambda\right)} \le \sqrt{\|A^\top A\|_{\mathrm{op}}\,\|\Lambda\|_*} = \sqrt{\|\Lambda\|_*}\,\|A\|_{\mathrm{op}}.$$

□

Lemma A.3. *Let $\Lambda \in \mathbb{R}^{n\times n}$ be positive definite. For a twice continuously differentiable function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, define $f_v(w) = f(W)$ with $w = \overline{\mathrm{vec}}(W) \in \mathbb{R}^{mn}$. Then the following are equivalent:*

1. *f satisfies Assumption 4.11, i.e. for all $W, W' \in \mathbb{R}^{m\times n}$,*
$$\|\nabla f(W) - \nabla f(W')\|_{\Lambda^{-1}} \le \|W - W'\|_{\Lambda}.$$
2. *For all $w \in \mathbb{R}^{mn}$,*
$$-I_m \otimes \Lambda \preceq \nabla^2 f_v(w) \preceq I_m \otimes \Lambda.$$

Proof. Let $M := I_m \otimes \Lambda$. For any $A \in \mathbb{R}^{m\times n}$, we have
$$\mathrm{tr}\left(A\Lambda A^\top\right) = \overline{\mathrm{vec}}(A)^\top (I_m \otimes \Lambda)\overline{\mathrm{vec}}(A) = \overline{\mathrm{vec}}(A)^\top M \overline{\mathrm{vec}}(A),$$
and similarly
$$\mathrm{tr}\left(G\Lambda^{-1}G^\top\right) = \overline{\mathrm{vec}}(G)^\top M^{-1}\overline{\mathrm{vec}}(G).$$
Moreover,
$$\nabla f_v(w) = \overline{\mathrm{vec}}(\nabla f(W)).$$
Hence the first condition is equivalent to
$$\|\nabla f_v(w) - \nabla f_v(w')\|_{M^{-1}} \le \|w - w'\|_M, \qquad \forall w, w' \in \mathbb{R}^{mn}.$$
Here, for a vector $v \in \mathbb{R}^d$, its norm $\|v\|_A$ is defined as $\|v\|_A := \sqrt{v^\top A v} = \|A^{1/2}v\|_2$, where $A \in \mathbb{R}^{d\times d}$ is positive definite.

We first prove (2) $\Rightarrow$ (1). For any $w, w' \in \mathbb{R}^{mn}$, we have
$$\nabla f_v(w) - \nabla f_v(w') = \int_0^1 \nabla^2 f_v(w' + t(w - w'))\,(w - w')\,dt.$$
Therefore,
$$M^{-1/2}(\nabla f_v(w) - \nabla f_v(w')) = \int_0^1 \left(M^{-1/2}\nabla^2 f_v(w' + t(w - w'))M^{-1/2}\right)M^{1/2}(w - w')\,dt.$$
Under (2), we have
$$-I \preceq M^{-1/2}\nabla^2 f_v(\cdot)M^{-1/2} \preceq I,$$
so its operator norm is at most 1. Thus
$$\begin{aligned}\|M^{-1/2}(\nabla f_v(w) - \nabla f_v(w'))\|_2 =& \left\|\int_0^1 \left(M^{-1/2}\nabla^2 f_v(w' + t(w - w'))M^{-1/2}\right)M^{1/2}(w - w')\,dt\right\|_2 \\ \le& \int_0^1 \left\|M^{-1/2}\nabla^2 f_v(w' + t(w - w'))M^{-1/2}\right\|_{\mathrm{op}} \|M^{1/2}(w - w')\|_2\,dt \\ \le& \|M^{1/2}(w - w')\|_2,\end{aligned}$$
which means
$$\|\nabla f_v(w) - \nabla f_v(w')\|_{M^{-1}} \le \|w - w'\|_M.$$
Next we prove (1) $\Rightarrow$ (2). Fix any w and any direction $u \in \mathbb{R}^{mn}$. Applying (1) to $w + tu$ and w gives
$$\|\nabla f_v(w + tu) - \nabla f_v(w)\|_{M^{-1}} \le |t|\,\|u\|_M.$$
Dividing by $|t|$ and letting $t \to 0$ yields
$$\|\nabla^2 f_v(w)u\|_{M^{-1}} \le \|u\|_M, \qquad \forall u.$$

Let $v := M^{1/2}u$, we have $u = M^{-1/2}v$ and

$$\|M^{-1/2}\nabla^2 f_v(w)M^{-1/2}v\|_2 \leq \|v\|_2,$$

which implies

$$\|M^{-1/2}\nabla^2 f_v(w)M^{-1/2}\|_{\mathrm{op}} \leq 1.$$

Since $\nabla^2 f_v(w)$ is symmetric, this is equivalent to

$$-I \preceq M^{-1/2}\nabla^2 f_v(w)M^{-1/2} \preceq I.$$

Multiplying both sides by $M^{1/2}$ proves

$$-M \preceq \nabla^2 f_v(w) \preceq M,$$

i.e.,

$$-I_m \otimes \Lambda \preceq \nabla^2 f_v(w) \preceq I_m \otimes \Lambda.$$

This completes the proof. □

Proof of Lemma 4.12

Proof. If $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is 1-Λ-norm Lipschitz smooth with a positive definite matrix $\Lambda \in \mathbb{R}^{n\times n}$, then, for any $W, W' \in \mathbb{R}^{m\times n}$, we have

$$\begin{aligned}\|\nabla f(W) - \nabla f(W')\|_{\mathrm{F}} \leq& \sqrt{\|\Lambda\|_{\mathrm{op}}}\|\nabla f(W) - \nabla f(W')\|_{\Lambda^{-1}}\\ \leq& \sqrt{\|\Lambda\|_{\mathrm{op}}}\|W - W'\|_{\Lambda}\\ \leq& \|\Lambda\|_{\mathrm{op}}\|W - W'\|_{\mathrm{F}}\end{aligned}$$

where the first inequality is due to Equation (8) in Lemma A.2, the second inequality is the definition of the Λ-smoothness, and the third inequality is due to Equation (9) in Lemma A.2. Thus, f is also $\|\Lambda\|_{\mathrm{op}}$ Frobenius norm Lipschitz smooth.

Moreover, we have

$$\begin{aligned}\|\nabla f(W) - \nabla f(W')\|_{*} \leq& \sqrt{\|\Lambda\|_{*}}\|\nabla f(W) - \nabla f(W')\|_{\Lambda^{-1}}\\ \leq& \sqrt{\|\Lambda\|_{*}}\|W - W'\|_{\Lambda}\\ \leq& \|\Lambda\|_{*}\|W - W'\|_{\mathrm{op}}\end{aligned}$$

where the first inequality is due to Lemma A.1, the second inequality is the definition of the Λ-smoothness, and the third inequality is due to Equation (10) in Lemma A.2. Thus, f is also $\|\Lambda\|_{*}$ spectral norm Lipschitz smooth. □

Lemma A.4. *For $t = 0, 1, \ldots, T$, M_t and W_t generated by Algorithm 1, defining $C_0 = \nabla f(W_0)$, and when $t > 0$, $C_t = \beta C_{t-1} + (1-\beta)\nabla f(W_t) = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\nabla f(W_i) + \beta^t \nabla f(W_0)$, we have*

$$\mathbb{E}[\|C_t - M_t\|_{\mathrm{F}}] \leq \sqrt{\frac{1-\beta}{1+\beta}}\frac{\sigma}{\sqrt{B}} + \beta^t\frac{\sigma}{\sqrt{B}}.$$

Proof. According to Assumption 3.3, we have the following relationship about G_t and $\nabla f(W_t)$.

$$\begin{aligned}\mathbb{E}[\|G_t - \nabla f(W_t)\|_{\mathrm{F}}^2] =& \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B}\nabla f(W_t;\xi_{t,i}) - \nabla f(W_t)\right\|_{\mathrm{F}}^2\right]\\ =& \frac{1}{B^2}\sum_{i=1}^{B}\mathbb{E}[\|\nabla f(W_t;\xi_{t,i}) - \nabla f(W_t)\|_{\mathrm{F}}^2]\end{aligned}$$

$$\leq \frac{\sigma^2}{B} \tag{13}$$

where the second equality is due to $\mathbb{E}[\nabla f(W_t;\xi_{t,i})] = \nabla f(W_t)$ and the last inequality is due to $\mathbb{E}\|\nabla f(W;\xi_{t,i}) - \nabla f(W)\|_{\mathrm{F}}^2] \leq \sigma^2$. Moreover, according to the Cauchy-Schwarz inequality, we have

$$\mathbb{E}[\|G_t - \nabla f(W_t)\|_{\mathrm{F}}] \leq \sqrt{\mathbb{E}[\|G_t - \nabla f(W_t)\|_{\mathrm{F}}^2]} = \frac{\sigma}{\sqrt{B}}. \tag{14}$$

Note that $M_0 = G_0$, $M_t = \beta M_{t-1} + (1-\beta)G_t = (1-\beta)\sum_{i=1}^t \beta^{t-i}G_i + \beta^t G_0$. Thus, we have

$$\begin{aligned}
\mathbb{E}[\|C_t - M_t\|_{\mathrm{F}}] \leq& (1-\beta)\mathbb{E}[\|\sum_{i=1}^t \beta^{t-i}(G_i - \nabla f(W_i))\|_{\mathrm{F}}] + \beta^t \mathbb{E}[\|G_0 - \nabla f(W_0)\|_{\mathrm{F}}] \\
\leq& \sqrt{(1-\beta)^2\mathbb{E}[\|\sum_{i=1}^t \beta^{t-i}(G_i - \nabla f(W_i))\|_{\mathrm{F}}^2]} + \beta^t \frac{\sigma}{\sqrt{B}} \\
\leq& \sqrt{(1-\beta)^2 \sum_{i=1}^t \beta^{t-i}\frac{\sigma^2}{B}} + \beta^t \frac{\sigma}{\sqrt{B}} \\
\leq& \sqrt{\frac{1-\beta}{1+\beta}}\frac{\sigma}{\sqrt{B}} + \beta^t \frac{\sigma}{\sqrt{B}}
\end{aligned}$$

where the second inequality is due to the Cauchy-Schwarz inequality and Equation (14), the third inequality is due to Equation (13). □

B Nonconvex

B.1 Proof of Theorem 4.1

Proof. Set $\eta_t = \eta$. Since f is L Frobenius norm Lipschitz smooth, we have

$$\begin{aligned}
\mathbb{E}[f(W_t) - f(W_{t+1})] \geq& \mathbb{E}[\eta\langle \nabla f(W_t), U_t V_t^\top\rangle - \frac{L}{2}\eta^2\|U_t V_t^\top\|_{\mathrm{F}}^2] \\
\geq& \mathbb{E}[\eta\langle M_t, U_t V_t^\top\rangle - \frac{rL}{2}\eta^2 - \eta\langle \nabla f(W_t) - M_t, U_t V_t^\top\rangle] \\
\geq& \mathbb{E}[\eta\langle M_t, U_t V_t^\top\rangle - \frac{rL}{2}\eta^2 - \eta\|\nabla f(W_t) - M_t\|_{\mathrm{F}}\|U_t V_t^\top\|_{\mathrm{F}}] \\
\geq& \mathbb{E}[\eta\|M_t\|_* - \frac{rL}{2}\eta^2 - \eta\sqrt{r}\|\nabla f(W_t) - M_t\|_{\mathrm{F}}] \\
\geq& \mathbb{E}[\eta\|\nabla f(W_t)\|_* - \frac{rL}{2}\eta^2 - \eta(\sqrt{r}+1)\|\nabla f(W_t) - M_t\|_{\mathrm{F}}] \\
\geq& \mathbb{E}[\eta\|\nabla f(W_t)\|_* - \frac{rL}{2}\eta^2 - 2\eta\sqrt{r}\|\nabla f(W_t) - M_t\|_{\mathrm{F}}].
\end{aligned}$$

Then, we need to analyze and bound the error $\|\nabla f(W_t) - M_t\|_{\mathrm{F}}$. Note that $M_0 = G_0$, $M_t = \beta M_{t-1} + (1-\beta)G_t = (1-\beta)\sum_{i=1}^t \beta^{t-i}G_i + \beta^t G_0$. We can define $C_0 = \nabla f(W_0)$, and when $t > 0$, $C_t = \beta C_{t-1} + (1-\beta)\nabla f(W_t) = (1-\beta)\sum_{i=1}^t \beta^{t-i}\nabla f(W_i) + \beta^t \nabla f(W_0)$. Then, according to Lemma A.4, we have

$$\mathbb{E}[\|C_t - M_t\|_{\mathrm{F}}] \leq \sqrt{\frac{1-\beta}{1+\beta}}\frac{\sigma}{\sqrt{B}} + \beta^t \frac{\sigma}{\sqrt{B}}.$$

Moreover, when $t > 0$, we note that

$$\mathbb{E}[\|\nabla f(W_t) - C_t\|_{\mathrm{F}}]$$

$$
\begin{aligned}
&=\mathbb{E}[\|\nabla f(W_t) - (\beta C_{t-1} + (1-\beta)\nabla f(W_t))\|_{\mathrm{F}}] \\
&=\mathbb{E}[\beta\|\nabla f(W_t) - C_{t-1}\|_{\mathrm{F}}] \\
&\leq\mathbb{E}[\beta\|\nabla f(W_{t-1}) - C_{t-1}\|_{\mathrm{F}} + \beta\|\nabla f(W_{t-1}) - \nabla f(W_t)\|_{\mathrm{F}}] \\
&\leq\mathbb{E}[\beta\|\nabla f(W_{t-1}) - C_{t-1}\|_{\mathrm{F}} + \beta L\|W_{t-1} - W_t\|_{\mathrm{F}}] \\
&=\mathbb{E}[\beta\|\nabla f(W_{t-1}) - C_{t-1}\|_{\mathrm{F}} + \beta L\eta\|U_{t-1}V_{t-1}^\top\|_{\mathrm{F}}] \\
&\leq\mathbb{E}[\beta\|\nabla f(W_{t-1}) - C_{t-1}\|_{\mathrm{F}} + \beta L\eta\sqrt{r}] \\
&\leq\beta^t\|\nabla f(W_0) - C_0\|_{\mathrm{F}} + \sum_{i=1}^{t}\beta^i L\eta\sqrt{r} \\
&\leq\frac{\sqrt{r}\beta L\eta}{1-\beta}
\end{aligned}
$$

where the second inequality is due to Assumption 3.1.

Thus, we have

$$
\mathbb{E}[\|\nabla f(W_t) - M_t\|_{\mathrm{F}}] \leq \mathbb{E}[\|C_t - M_t\|_{\mathrm{F}} + \|\nabla f(W_t) - C_t\|_{\mathrm{F}}] \leq \sqrt{\frac{1-\beta}{1+\beta}}\frac{\sigma}{\sqrt{B}} + \beta^t\frac{\sigma}{\sqrt{B}} + \frac{\sqrt{r}\beta L\eta}{1-\beta},
$$

and

$$
\begin{aligned}
\mathbb{E}[f(W_t) - f(W_{t+1})] \geq&\mathbb{E}[\eta\|\nabla f(W_t)\|_* - \frac{rL}{2}\eta^2 - 2\eta\sqrt{r}\|\nabla f(W_t) - M_t\|_{\mathrm{F}}] \\
\geq&\mathbb{E}[\eta\|\nabla f(W_t)\|_* - \frac{rL}{2}\eta^2 - 2\eta\sqrt{\frac{1-\beta}{1+\beta}}\frac{\sigma\sqrt{r}}{\sqrt{B}} - 2\eta\beta^t\frac{\sigma\sqrt{r}}{\sqrt{B}} - \frac{2r\eta^2\beta L}{1-\beta}].
\end{aligned}
$$

Summing over $t = 0, 1, \ldots, T-1$, we can get

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq \frac{1}{T\eta}\mathbb{E}[f(W_0) - f(W_T)] + \frac{Lr\eta}{2} + \frac{2\sigma\sqrt{r(1-\beta)}}{\sqrt{(1+\beta)B}} + \frac{2\beta\sigma\sqrt{r}}{(1-\beta)T\sqrt{B}} + \frac{2r\eta\beta L}{1-\beta}.
$$

Set $\eta = \sqrt{\frac{(1-\beta)\Delta}{rTL}}$, we can get

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq \frac{\Delta}{T\eta} + \frac{Lr\eta}{2} + \frac{2\sigma\sqrt{r(1-\beta)}}{\sqrt{(1+\beta)B}} + \frac{2\beta\sigma\sqrt{r}}{(1-\beta)T\sqrt{B}} + \frac{2r\eta\beta L}{1-\beta}.
$$

When $B = 1$, we can set $1-\beta = \min\{\frac{\sqrt{L\Delta}}{\sigma\sqrt{T}}, 1\}$ and get

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq O(1)\left(\sqrt[4]{\frac{r^2L\Delta\sigma^2}{T}} + \sqrt{\frac{rL\Delta}{T}} + \frac{\sqrt{r}\sigma^2}{\sqrt{L\Delta T}}\right).
$$

Thus, we can find an ϵ-nuclear norm stationary point of f with a complexity of $O(r^2L\sigma^2\Delta\epsilon^{-4})$. □

B.2 Proof of Theorem 4.7

Proof. Set $\eta_t = \eta$. Since f is L_* spectral norm Lipschitz smooth, we have

$$
\begin{aligned}
\mathbb{E}[f(W_t) - f(W_{t+1})] \geq&\mathbb{E}[\eta\langle\nabla f(W_t), U_tV_t^\top\rangle - \frac{L_*}{2}\eta^2\|U_tV_t^\top\|_{\mathrm{op}}^2] \\
\geq&\mathbb{E}[\eta\langle M_t, U_tV_t^\top\rangle - \frac{L_*}{2}\eta^2 - \eta\langle\nabla f(W_t) - M_t, U_tV_t^\top\rangle] \\
\geq&\mathbb{E}[\eta\langle M_t, U_tV_t^\top\rangle - \frac{L_*}{2}\eta^2 - \eta\|\nabla f(W_t) - M_t\|_*\|U_tV_t^\top\|_{\mathrm{op}}]
\end{aligned}
$$

$$
\begin{aligned}
&\geq \mathbb{E}[\eta\|M_t\|_* - \frac{L_*}{2}\eta^2 - \eta\|\nabla f(W_t) - M_t\|_*] \\
&\geq \mathbb{E}[\eta\|\nabla f(W_t)\|_* - \frac{L_*}{2}\eta^2 - 2\eta\|\nabla f(W_t) - M_t\|_*].
\end{aligned}
$$

Then, we need to analyze and bound the error $\|\nabla f(W_t) - M_t\|_*$. Note that $M_0 = G_0$, $M_t = \beta M_{t-1} + (1-\beta)G_t = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}G_i + \beta^t G_0$. We can define $C_0 = \nabla f(W_0)$, and when $t > 0$, $C_t = \beta C_{t-1} + (1-\beta)\nabla f(W_t) = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\nabla f(W_i) + \beta^t \nabla f(W_0)$. Then, according to Lemma A.4, we have

$$
\mathbb{E}[\|C_t - M_t\|_*] \leq \sqrt{r}\mathbb{E}[\|C_t - M_t\|_{\mathrm{F}}] \leq \sqrt{\frac{1-\beta}{1+\beta}}\frac{\sigma\sqrt{r}}{\sqrt{B}} + \beta^t\frac{\sigma\sqrt{r}}{\sqrt{B}}.
$$

Moreover, when $t > 0$, we can note that

$$
\begin{aligned}
&\mathbb{E}[\|\nabla f(W_t) - C_t\|_*] \\
=&\mathbb{E}[\|\nabla f(W_t) - (\beta C_{t-1} + (1-\beta)\nabla f(W_t))\|_*] \\
=&\mathbb{E}[\beta\|\nabla f(W_t) - C_{t-1}\|_*] \\
\leq&\mathbb{E}[\beta\|\nabla f(W_{t-1}) - C_{t-1}\|_* + \beta\|\nabla f(W_{t-1}) - \nabla f(W_t)\|_*] \\
\leq&\mathbb{E}[\beta\|\nabla f(W_{t-1}) - C_{t-1}\|_* + \beta L_*\|W_{t-1} - W_t\|_{\mathrm{op}}] \\
=&\mathbb{E}[\beta\|\nabla f(W_{t-1}) - C_{t-1}\|_* + \beta L_*\eta\|U_{t-1}V_{t-1}^\top\|_{\mathrm{op}}] \\
\leq&\mathbb{E}[\beta\|\nabla f(W_{t-1}) - C_{t-1}\|_* + \beta L_*\eta] \\
\leq&\beta^t\|\nabla f(W_0) - C_0\|_{\mathrm{F}} + \sum_{i=1}^{t}\beta^i L_*\eta \\
\leq&\frac{\beta L_*\eta}{1-\beta}
\end{aligned}
$$

where the second inequality is due to Assumption 3.2.

Thus, we have

$$
\mathbb{E}[\|\nabla f(W_t) - M_t\|_*] \leq \mathbb{E}[\|C_t - M_t\|_* + \|\nabla f(W_t) - C_t\|_*] \leq \sqrt{\frac{1-\beta}{1+\beta}}\frac{\sigma\sqrt{r}}{\sqrt{B}} + \beta^t\frac{\sigma\sqrt{r}}{\sqrt{B}} + \frac{\beta L_*\eta}{1-\beta},
$$

and

$$
\begin{aligned}
\mathbb{E}[f(W_t) - f(W_{t+1})] \geq&\mathbb{E}[\eta\|\nabla f(W_t)\|_* - \frac{L_*}{2}\eta^2 - 2\eta\|\nabla f(W_t) - M_t\|_*] \\
\geq&\mathbb{E}[\eta\|\nabla f(W_t)\|_* - \frac{L_*}{2}\eta^2 - 2\eta\sqrt{\frac{1-\beta}{1+\beta}}\frac{\sigma\sqrt{r}}{\sqrt{B}} - 2\eta\beta^t\frac{\sigma\sqrt{r}}{\sqrt{B}} - \frac{2\eta^2\beta L_*}{1-\beta}].
\end{aligned}
$$

Summing over $t = 0, 1, \ldots, T-1$, we can get

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq \frac{1}{T\eta}\mathbb{E}[f(W_0) - f(W_T)] + \frac{L_*\eta}{2} + \frac{2\sigma\sqrt{r(1-\beta)}}{\sqrt{(1+\beta)B}} + \frac{2\beta\sigma\sqrt{r}}{(1-\beta)T\sqrt{B}} + \frac{2\eta\beta L_*}{1-\beta}.
$$

Set $\eta = \sqrt{\frac{(1-\beta)\Delta}{TL_*}}$, we can get

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq \frac{\Delta}{T\eta} + \frac{L_*\eta}{2} + \frac{2\sigma\sqrt{r(1-\beta)}}{\sqrt{(1+\beta)B}} + \frac{2\beta\sigma\sqrt{r}}{(1-\beta)T\sqrt{B}} + \frac{2\eta\beta L_*}{1-\beta}.
$$

When $B = 1$, we can set $1 - \beta = \min\{\frac{\sqrt{L_*\Delta}}{\sigma\sqrt{rT}}, 1\}$ and get

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq O(1)\left(\sqrt[4]{\frac{rL_*\Delta\sigma^2}{T}} + \sqrt{\frac{L_*\Delta}{T}} + \frac{r\sigma^2}{\sqrt{L_*\Delta T}}\right).
$$

Thus, we can find an ϵ-nuclear norm stationary point of f with a complexity of $O(rL_*\sigma^2\Delta\epsilon^{-4})$. □

B.3 Proof of Theorem 4.14

Proof. Denote $J_t = \overline{\text{vec}}(U_t V_t^\top)^\top H_t \overline{\text{vec}}(U_t V_t^\top)$, $H_t = \nabla^2 f_v(\overline{\text{vec}}(W_t))$, and $f_v(\overline{\text{vec}}(W_t)) = f(W_t)$. Set $\eta_t = \eta$. Taking the Taylor expansion at W_t, we have

$$\begin{aligned}
f(W_{t+1}) - f(W_t) =& \langle \nabla f(W_t), W_{t+1} - W_t \rangle + \frac{1}{2}\overline{\text{vec}}(W_{t+1} - W_t)^\top \nabla^2 f_v(\overline{\text{vec}}(W_t))\overline{\text{vec}}(W_{t+1} - W_t) \\
&+ \frac{1}{6}\sum_{i,j,k=1}^{d} [\nabla^3 f_v(\theta)]_{ijk} \overline{\text{vec}}(W_{t+1} - W_t)_i \overline{\text{vec}}(W_{t+1} - W_t)_j \overline{\text{vec}}(W_{t+1} - W_t)_k \\
\leq& -\eta \langle \nabla f(W_t), U_t V_t^\top \rangle + \frac{\eta^2 J_t}{2} + \frac{s\eta^3 \|\overline{\text{vec}}(U_t V_t^\top)\|_{\text{F}}^3}{6} \\
\leq& -\eta \langle \nabla f(W_t), U_t V_t^\top \rangle + \frac{\eta^2 J_t}{2} + \frac{s\eta^3 r^{3/2}}{6} \\
=& -\eta \|\nabla f(W_t)\|_* + \frac{\eta^2 J_t}{2} + \frac{s\eta^3 r^{3/2}}{6}
\end{aligned} \tag{15}$$

where $\theta \in \mathbb{R}^{mn}$ can be some vectors between $\overline{\text{vec}}(W_t)$ and $\overline{\text{vec}}(W_{t+1})$, and the first inequality is due to Assumption 4.13.

Denote $J = \frac{1}{T}\sum_{t=0}^{T-1} J_t$. Summing over $t = 0, 1, \ldots, T-1$, we can get

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(W_t)\|_* \leq \frac{f(W_0) - f(W_T)}{T\eta} + \frac{\eta J}{2} + \frac{s\eta^2 r^{3/2}}{6}.$$

□

In experiments, we found that usually $J > 0$. When $J > 0$ and $\eta = \sqrt{\frac{2\Delta}{JT}}$, we can get

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(W_t)\|_* \leq \sqrt{\frac{2J\Delta}{T}} + \frac{sr^{3/2}\Delta}{3JT}.$$

Thus, it is possible that Muon can find an ϵ-nuclear norm stationary point of f with a complexity of $O(J\Delta\epsilon^{-2})$.

When $J \leq 0$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(W_t)\|_* \leq \frac{\Delta}{T\eta} + \frac{s\eta^2 r^{3/2}}{6}.$$

We can set $\eta = \sqrt[3]{\frac{\Delta}{sTr^{3/2}}}$ and get

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(W_t)\|_* \leq O\left(\frac{\Delta^{2/3} s^{1/3} r^{1/2}}{T^{2/3}}\right).$$

Thus, when $J \leq 0$, Muon can find an ϵ-nuclear norm stationary point of f with a complexity of $O(s^{1/2} r^{3/4} \Delta \epsilon^{-3/2})$.

B.4 Convergence analysis of Algorithm 1 with Assumption 4.13 in the stochastic setting

Theorem B.1. *Under Assumptions 4.13 and 3.3, if we apply Algorithm 1 with $\eta_t = \eta$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(W_t)\|_*] \leq \frac{\Delta}{T\eta} + \frac{J^*\eta}{2} + \frac{2\sigma\sqrt{r(1-\beta)}}{\sqrt{(1+\beta)B}} + \frac{2\beta\sigma\sqrt{r}}{(1-\beta)T\sqrt{B}} + \frac{2\eta\beta J^*}{1-\beta} + O\left(\frac{sr^{3/2}\eta^2}{1-\beta}\right).$$

where $J^* = \max\{\mathbb{E}[\hat{J}], \mathbb{E}[J]\}$, $\hat{J} = \frac{1}{T}\sum_{t=0}^{T-2}\hat{J}_t$, $J = \frac{1}{T}\sum_{t=0}^{T-1} J_t$, $J_t = \overline{\text{vec}}(U_t V_t^\top)^\top \nabla^2 f_v(\overline{\text{vec}}(W_t))\overline{\text{vec}}(U_t V_t^\top)$, $\hat{J}_t = \overline{\text{vec}}(U_{g,t}V_{g,t}^\top)^\top \nabla^2 f_v(\overline{\text{vec}}(W_t))\overline{\text{vec}}(U_t V_t^\top)$, $H_t = \nabla^2 f_v(\overline{\text{vec}}(W_t))$, $f_v(\overline{\text{vec}}(W_t)) = f(W_t)$, *and* $U_{g,t}, V_{g,t}$ *are the the matrices of left and right singular vectors of* $\nabla f(W_t) - \nabla f(W_{t+1})$, *i.e.* $\nabla f(W_t) - \nabla f(W_{t+1}) = U_{g,t}S_{g,t}V_{g,t}^\top$.

Proof. Denote $d = mn$, $J_t = \overline{\text{vec}}(U_t V_t^\top)^\top H_t \overline{\text{vec}}(U_t V_t^\top)$, $H_t = \nabla^2 f_v(\overline{\text{vec}}(W_t))$, and $f_v(\overline{\text{vec}}(W_t)) = f(W_t)$. Set $\eta_t = \eta$. Taking the Taylor expansion at W_t, we have

$$
\begin{aligned}
&\mathbb{E}[f(W_{t+1}) - f(W_t)] \\
=&\mathbb{E}[\langle \nabla f(W_t), W_{t+1} - W_t\rangle + \frac{1}{2}\overline{\text{vec}}(W_{t+1} - W_t)^\top \nabla^2 f_v(\overline{\text{vec}}(W_t))\overline{\text{vec}}(W_{t+1} - W_t) \\
&+ \frac{1}{6}\sum_{i,j,k=1}^{d}[\nabla^3 f_v(\theta)]_{ijk}\overline{\text{vec}}(W_{t+1} - W_t)_i \overline{\text{vec}}(W_{t+1} - W_t)_j \overline{\text{vec}}(W_{t+1} - W_t)_k] \\
\leq&\mathbb{E}\left[-\eta\langle \nabla f(W_t), U_t V_t^\top\rangle + \frac{\eta^2 J_t}{2} + \frac{s\eta^3\|\overline{\text{vec}}(U_t V_t^\top)\|_F^3}{6}\right] \\
\leq&\mathbb{E}\left[-\eta\langle M_t, U_t V_t^\top\rangle + \eta\langle M_t - \nabla f(W_t), U_t V_t^\top\rangle + \frac{\eta^2 J_t}{2}\right] + \frac{s\eta^3 r^{3/2}}{6} \\
\leq&\mathbb{E}\left[-\eta\|M_t\|_* + \eta\|M_t - \nabla f(W_t)\|_*\|U_t V_t^\top\|_{\text{op}} + \frac{\eta^2 J_t}{2}\right] + \frac{s\eta^3 r^{3/2}}{6} \\
\leq&\mathbb{E}\left[-\eta\|\nabla f(W_t)\|_* + 2\eta\|M_t - \nabla f(W_t)\|_* + \frac{\eta^2 J_t}{2}\right] + \frac{s\eta^3 r^{3/2}}{6}
\end{aligned}
$$

where $\theta \in \mathbb{R}^d$ can be some vectors between $\overline{\text{vec}}(W_t)$ and $\overline{\text{vec}}(W_{t+1})$, the first inequality is due to Assumption 4.13.

Then, we need to analyze and bound the error $\|\nabla f(W_t) - M_t\|_*$. Note that $M_0 = G_0$, $M_t = \beta M_{t-1} + (1-\beta)G_t = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}G_i + \beta^t G_0$. We can define $C_0 = \nabla f(W_0)$, and when $t > 0$, $C_t = \beta C_{t-1} + (1-\beta)\nabla f(W_t) = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\nabla f(W_i) + \beta^t \nabla f(W_0)$. Then, according to Lemma A.4, we have

$$
\mathbb{E}[\|C_t - M_t\|_*] \leq \sqrt{r}\mathbb{E}[\|C_t - M_t\|_F] \leq \sqrt{\frac{1-\beta}{1+\beta}}\frac{\sigma\sqrt{r}}{\sqrt{B}} + \beta^t\frac{\sigma\sqrt{r}}{\sqrt{B}}
$$

Suppose the SVD of $\nabla f(W_t) - \nabla f(W_{t+1})$ is $\nabla f(W_t) - \nabla f(W_{t+1}) = U_{g,t}S_{g,t}V_{g,t}^\top$. Then, we have

$$
\begin{aligned}
&\|\nabla f(W_t) - \nabla f(W_{t+1})\|_* \\
=&\|U_{g,t}^\top[\nabla f(W_t) - \nabla f(W_{t+1})]V_{g,t}\|_* \\
=&\text{tr}\left(U_{g,t}^\top[\nabla f(W_t) - \nabla f(W_{t+1})]V_{g,t}\right) \\
=&\text{tr}\left((U_{g,t}V_{g,t}^\top)^\top[\nabla f(W_t) - \nabla f(W_{t+1})]\right)
\end{aligned}
$$

For any $i \in [d]$, taking the Taylor expansion of $\overline{\text{vec}}(\nabla f(W_{t+1}))_i$ at W_t, we have

$$
\begin{aligned}
&[\overline{\text{vec}}(\nabla f(W_t) - \nabla f(W_{t+1}))]_i \\
=&\sum_{j=1}^{d}[\nabla^2 f_v(\overline{\text{vec}}(W_t))]_{ij}\overline{\text{vec}}(W_t - W_{t+})_j + \frac{1}{2}\sum_{j,k=1}^{d}[\nabla^3 f_v(\theta_i)]_{ijk}\overline{\text{vec}}(W_t - W_{t+1})_j\overline{\text{vec}}(W_t - W_{t+1})_k \\
=&\eta\sum_{j=1}^{d}[\nabla^2 f_v(\overline{\text{vec}}(W_t))]_{ij}\overline{\text{vec}}(U_t V_t^\top)_j + \frac{\eta^2}{2}\sum_{j,k=1}^{d}[\nabla^3 f_v(\theta_i)]_{ijk}\overline{\text{vec}}(U_t V_t^\top)_j\overline{\text{vec}}(U_t V_t^\top)_k
\end{aligned}
$$

where $\theta_i \in \mathbb{R}^d$ can be some vectors between $\overline{\text{vec}}(W_t)$ and $\overline{\text{vec}}(W_{t+1})$.

Denote $\hat{J}_t = \overline{\text{vec}}(U_{g,t}V_{g,t}^\top)^\top \nabla^2 f_v(\overline{\text{vec}}(W_t))\overline{\text{vec}}(U_t V_t^\top)$, we have

$$\begin{aligned}
&\|\nabla f(W_{t+1}) - \nabla f(W_t)\|_* \\
=&\text{tr}\left((U_{g,t}V_{g,t}^\top)^\top [\nabla f(W_t) - \nabla f(W_{t+1})]\right) \\
=&\eta \sum_{i,j=1}^{d} \overline{\text{vec}}(U_{g,t}V_{g,t}^\top)_i \nabla^2 f_v(\overline{\text{vec}}(W_t))]_{ij} \overline{\text{vec}}(U_t V_t^\top)_j \\
&+ \frac{\eta^2}{2} \sum_{i,j,k=1}^{d} [\nabla^3 f_v(\theta_i)]_{ijk} \overline{\text{vec}}(U_{g,t}V_{g,t}^\top)_i \overline{\text{vec}}(U_t V_t^\top)_j \overline{\text{vec}}(U_t V_t^\top)_k \\
\leq&\eta \hat{J}_t + \frac{s\eta^2 r^{3/2}}{2}
\end{aligned}$$

where the last inequality is due to Assumption 4.13.

Then, for $t > 0$, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla f(W_t) - C_t\|_*] \\
=&\mathbb{E}[\|\nabla f(W_t) - (\beta C_{t-1} + (1-\beta)\nabla f(W_t))\|_*] \\
=&\mathbb{E}[\beta\|\nabla f(W_t) - C_{t-1}\|_*] \\
\leq&\mathbb{E}[\beta\|\nabla f(W_{t-1}) - C_{t-1}\|_* + \beta\|\nabla f(W_{t-1}) - \nabla f(W_t)\|_*] \\
\leq&\mathbb{E}[\beta\|\nabla f(W_{t-1}) - C_{t-1}\|_* + \beta\eta\hat{J}_{t-1} + \frac{s\eta^2 r^{3/2}\beta}{2}] \\
\leq&\sum_{i=0}^{t-1} \beta^{t-i}\eta\mathbb{E}[\hat{J}_i] + \frac{s\eta^2 r^{3/2}\beta}{2(1-\beta)}
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\mathbb{E}[\|\nabla f(W_t) - M_t\|_*] \leq&\mathbb{E}[\|C_t - M_t\|_* + \|\nabla f(W_t) - C_t\|_*] \\
\leq&\sqrt{\frac{1-\beta}{1+\beta}}\frac{\sigma\sqrt{r}}{\sqrt{B}} + \beta^t\frac{\sigma\sqrt{r}}{\sqrt{B}} + \sum_{i=0}^{t-1}\beta^{t-i}\eta\mathbb{E}[\hat{J}_i] + \frac{s\eta^2 r^{3/2}\beta}{2(1-\beta)},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[f(W_t) - f(W_{t+1})] \\
\geq&\mathbb{E}[\eta\|\nabla f(W_t)\|_* - \frac{J_t}{2}\eta^2 - 2\eta\|\nabla f(W_t) - M_t\|_*] - \frac{s\eta^3 r^{3/2}}{6} \\
\geq&\mathbb{E}[\eta\|\nabla f(W_t)\|_* - \frac{J_t}{2}\eta^2 - \sum_{i=0}^{t-1} 2\beta^{t-i}\eta^2\hat{J}_i] - \frac{s\eta^3 r^{3/2}}{6} - 2\eta\sqrt{\frac{1-\beta}{1+\beta}}\frac{\sigma\sqrt{r}}{\sqrt{B}} - 2\eta\beta^t\frac{\sigma\sqrt{r}}{\sqrt{B}} - \frac{s\eta^3 r^{3/2}\beta}{1-\beta}.
\end{aligned}$$

Denote $\hat{J} = \frac{1}{T}\sum_{t=0}^{T-2}\hat{J}_t$, $J = \frac{1}{T}\sum_{t=0}^{T-1} J_t$ Summing over $t = 0, 1, \ldots, T-1$, we can get

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq&\frac{1}{T\eta}\mathbb{E}[f(W_0) - f(W_T)] + \frac{\eta\sum_{t=0}^{T-1}\mathbb{E}[J_t]}{2T} + \frac{2\sigma\sqrt{r(1-\beta)}}{\sqrt{(1+\beta)B}} + \frac{2\beta\sigma\sqrt{r}}{(1-\beta)T\sqrt{B}} \\
&+ 2\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=0}^{t-1} 2\beta^{t-i}\eta\mathbb{E}[\hat{J}_i] + O(sr^{3/2}\eta^2/(1-\beta)) \\
\leq&\frac{1}{T\eta}\mathbb{E}[f(W_0) - f(W_T)] + \frac{\eta\mathbb{E}[J]}{2} + \frac{2\sigma\sqrt{r(1-\beta)}}{\sqrt{(1+\beta)B}} + \frac{2\beta\sigma\sqrt{r}}{(1-\beta)T\sqrt{B}} \\
&+ \frac{2\eta\beta\mathbb{E}[\hat{J}]}{1-\beta} + O(sr^{3/2}\eta^2/(1-\beta))
\end{aligned}$$

Denote $J^* = \max\{\mathbb{E}[\hat{J}], \mathbb{E}[J]\}$, we can get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq \frac{\Delta}{T\eta} + \frac{J^*\eta}{2} + \frac{2\sigma\sqrt{r(1-\beta)}}{\sqrt{(1+\beta)B}} + \frac{2\beta\sigma\sqrt{r}}{(1-\beta)T\sqrt{B}} + \frac{2\eta\beta J^*}{1-\beta} + O\left(\frac{sr^{3/2}\eta^2}{1-\beta}\right).$$

□

When $B = 1$, $J^* > 0$ and $\eta = \sqrt{\frac{(1-\beta)\Delta}{TJ^*}}$, $1-\beta = \min\{\frac{\sqrt{J^*\Delta}}{\sigma\sqrt{rT}}, 1\}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq O\left(\sqrt[4]{\frac{rJ^*\Delta\sigma^2}{T}} + \sqrt{\frac{J^*\Delta}{T}} + \frac{r\sigma^2}{\sqrt{J^*\Delta T}} + \frac{sr^{3/2}\Delta}{TJ^*}\right).$$

Thus, it is possible that Muon can find an ϵ-nuclear norm stationary point of f with a complexity of $O(rJ^*\sigma^2\Delta\epsilon^{-4})$.

Note that $\hat{J}$ satisfies properties analogous to those of J discussed in Section 4.3.

For example, if we make the same assumptions as those in Section 4.3, i.e. H_t can be expressed as a Kronecker product $H_t = P_t \otimes Q_t$, where $P_t \in \mathbb{R}^{m\times m}$ and $Q_t \in \mathbb{R}^{n\times n}$. Let $\sigma_{p,t,1} \geq \sigma_{p,t,2} \geq \cdots \geq \sigma_{p,t,m}$ denote the singular values of P_t, and $\sigma_{q,t,1} \geq \sigma_{q,t,2} \geq \cdots \geq \sigma_{q,t,n}$ denote the singular values of Q_t. Then, we have $\hat{J}_t = \langle U_{g,t}^\top P_t U_t, V_{g,t}^\top Q_t V_t\rangle \leq \sum_{i=1}^{r_t}\sigma_{p,t,i}\sigma_{q,t,i}$ (Von Neumann's trace inequality).

Note that in the stochastic setting, the complexity of SGD for ϵ-nuclear norm stationary point is $O(r^2L\sigma^2\Delta\epsilon^{-4})$. Thus, Similar to the discussions in Section 4.3, when H_t can be represented by $P_t \otimes Q_t$ and Q_t, P_t are relatively low-rank such that $\sum_{i=1}^{r_t}\sigma_{p,t,i}\sigma_{q,t,i} \ll r\sigma_{p,t,1}\sigma_{q,t,1}$, then $J^* \leq \max\{\hat{J}, J\} \leq \max_t\sum_{i=1}^{r_t}\sigma_{p,t,i}\sigma_{q,t,i} \ll r\max_t L_t \leq rL$, and the convergence rate of Muon can be better than GD's. (Or one of Q_t, P_t is relatively low-rank, i.e. if Q_t is relatively low-rank such that $\|Q_t\|_{\mathrm{op}} \approx \|Q_t\| \ll r\|Q_t\|_{\mathrm{op}}$, we can also have $J^* \leq \max_t\|P_t\|_{\mathrm{op}}\|Q_t\|_* \ll r\max_t L_t \leq rL$.)

C Star-Convex

C.1 Proof of Theorem 4.5

Proof. **Constant stepsize**

First, by the L Frobenius norm Lipschitz smoothness of f, we have

$$\begin{aligned} f(W_{t+1}) &\leq f(W_t) + \langle\nabla f(W_t), W_{t+1} - W_t\rangle + \frac{L}{2}\|W_{t+1} - W_t\|_{\mathrm{F}}^2 \\ &\leq f(W_t) - \eta_t\|\nabla f(W_t)\|_* + \frac{rL\eta_t^2}{2} \end{aligned} \tag{16}$$

Then, according to (20), we have

$$-\|\nabla f(W_t)\|_* \leq -\frac{f(W_t) - f^*}{D_{\mathrm{op}}}$$

Set $\eta_t = \eta$ and combine (16) and (20). We have

$$f(W_{t+1}) - f^* \leq \left(1 - \frac{\eta}{D_{\mathrm{op}}}\right)(f(W_t) - f^*) + \frac{rL\eta^2}{2},$$

and

$$f(W_T) - f^* \leq \left(1 - \frac{\eta}{D_{\mathrm{op}}}\right)^T(f(W_0) - f^*) + \sum_{t=0}^{T-1}\left(1 - \frac{\eta}{D_{\mathrm{op}}}\right)^{T-1-t}\frac{rL\eta^2}{2}$$

$$\leq \left(1 - \frac{\eta}{D_{\text{op}}}\right)^T (f(W_0) - f^*) + \frac{rLD_{\text{op}}\eta}{2}.$$

Thus, to reach the precision $f(W_T) - f^* \leq \epsilon$, we can take $\eta = O(\frac{\epsilon}{rLD_{\text{op}}})$, and the complexity is $O(rLD_{\text{op}}^2\epsilon^{-1}\log\frac{\Delta}{\epsilon})$, where $\Delta = f(W_0) - f^*$. □

C.2 Proof of Theorem 4.6

Proof. **Adaptive stepsize**

By the L spectral norm Lipschitz smoothness of f and setting $\eta_t = \frac{\|\nabla f(W_t)\|_*}{rL}$, we have

$$\begin{aligned} f(W_{t+1}) &\leq f(W_t) + \langle \nabla f(W_t), W_{t+1} - W_t \rangle + \frac{L}{2}\|W_{t+1} - W_t\|_{\text{F}}^2 \\ &\leq f(W_t) - \eta_t \|\nabla f(W_t)\|_* + \frac{rL\eta_t^2}{2} \\ &= f(W_t) - \frac{\|\nabla f(W_t)\|_*^2}{2rL}. \end{aligned} \tag{17}$$

From the star-convex property, we have

$$\Delta_t := f(W_t) - f^* \leq \langle \nabla f(W_t), W_t - W^* \rangle \leq \|\nabla f(W_t)\|_* \|W_t - W^*\|_{\text{op}}. \tag{18}$$

Combining Equation (17) and Equation (18), we obtain

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{2rL\|W_t - W^*\|_{\text{op}}^2}\Delta_t^2.$$

Then we have

$$\frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{\Delta_t}{2rL\|W_t - W^*\|_{\text{op}}^2 \Delta_{t+1}}.$$

Using $\|W_t - W^*\|_{\text{op}} \leq D_{\text{op}}$, we can sum above inequality to obtain

$$\frac{1}{\Delta_t} \geq \frac{1}{\Delta_0} + \sum_{i=1}^{t} \frac{\Delta_{t-1}}{2rLD_{\text{op}}^2 \Delta_t} \geq \frac{1}{\Delta_0} + \frac{t}{2rLD_{\text{op}}^2},$$

where we use $\frac{\Delta_t}{\Delta_{t-1}} \geq 1$.

After rearranging terms, we obtain

$$f(W_t) - f^* \leq \frac{2rL(f(W_0) - f^*)D_{\text{op}}^2}{2rLD_{\text{op}}^2 + t(f(W_0) - f^*)}.$$

□

C.3 Proof of Theorem 4.9

Proof. **Constant stepsize**

First, by the L_* spectral norm Lipschitz smoothness of f, we have

$$\begin{aligned} f(W_{t+1}) &\leq f(W_t) + \langle \nabla f(W_t), W_{t+1} - W_t \rangle + \frac{L_*}{2}\|W_{t+1} - W_t\|_{\text{op}}^2 \\ &= f(W_t) - \eta_t \|\nabla f(W_t)\|_* + \frac{L_* \eta_t^2}{2} \end{aligned} \tag{19}$$

Then, from the star-convex condition, we obtain

$$\begin{aligned} f(W_t) &\leq f^* + \langle \nabla f(W_t), W_t - W^* \rangle \\ &\leq f^* + \|\nabla f(W_t)\|_* \|W_t - W^*\|_{\text{op}} \\ &\leq f^* + \|\nabla f(W_t)\|_* D_{\text{op}} \\ -\|\nabla f(W_t)\|_* &\leq -\frac{f(W_t) - f^*}{D_{\text{op}}} \end{aligned} \tag{20}$$

where we apply the Cauchy–Schwarz inequality in the second inequality, and in the third inequality we use Assumption 4.4.

Set $\eta_t = \eta$ and combine (19) and (20). We have

$$f(W_{t+1}) - f^* \leq \left(1 - \frac{\eta}{D_{\text{op}}}\right)(f(W_t) - f^*) + \frac{\eta^2 L_*}{2},$$

and

$$\begin{aligned} f(W_T) - f^* &\leq \left(1 - \frac{\eta}{D_{\text{op}}}\right)^T (f(W_0) - f^*) + \sum_{t=0}^{T-1} \left(1 - \frac{\eta}{D_{\text{op}}}\right)^{T-1-t} \frac{\eta^2 L_*}{2} \\ &\leq \left(1 - \frac{\eta}{D_{\text{op}}}\right)^T (f(W_0) - f^*) + \frac{\eta L_* D_{\text{op}}}{2}. \end{aligned}$$

Thus, to reach the precision $f(W_T) - f^* \leq \epsilon$, we can take $\eta = O(\frac{\epsilon}{L_* D_{\text{op}}})$, and the complexity is $O(L_* D_{\text{op}}^2 \epsilon^{-1} \log \frac{\Delta}{\epsilon})$, where $\Delta = f(W_0) - f^*$. □

C.4 Proof of Theorem 4.10

Proof. **Adaptive stepsize**

By the L_* spectral norm Lipschitz smoothness of f and setting $\eta_t = \frac{\|\nabla f(W_t)\|_*}{L_*}$, we have

$$\begin{aligned} f(W_{t+1}) &\leq f(W_t) + \langle \nabla f(W_t), W_{t+1} - W_t \rangle + \frac{L_*}{2} \|W_{t+1} - W_t\|_{\text{op}}^2 \\ &= f(W_t) - \eta_t \|\nabla f(W_t)\|_* + \frac{L_* \eta_t^2}{2} \\ &= f(W_t) - \frac{\|\nabla f(W_t)\|_*^2}{2L_*}. \end{aligned} \tag{21}$$

From the star-convex property, we have

$$\Delta_t := f(W_t) - f^* \leq \langle \nabla f(W_t), W_t - W^* \rangle \leq \|\nabla f(W_t)\|_* \|W_t - W^*\|_{\text{op}}. \tag{22}$$

Combining Equation (21) and Equation (22), we obtain

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{2L_* \|W_t - W^*\|_{\text{op}}^2} \Delta_t^2.$$

Then we have

$$\frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{\Delta_t}{2L_* \|W_t - W^*\|_{\text{op}}^2 \Delta_{t+1}}.$$

Using $\|W_t - W^*\|_{\text{op}} \leq D_{\text{op}}$, then we can sum above inequality to obtain

$$\frac{1}{\Delta_t} \geq \frac{1}{\Delta_0} + \sum_{i=1}^{t} \frac{\Delta_{t-1}}{2L_* D_{\text{op}}^2 \Delta_t} \geq \frac{1}{\Delta_0} + \frac{t}{2L_* D_{\text{op}}^2},$$

where we use $\frac{\Delta_t}{\Delta_{t-1}} \geq 1$.

After rearranging terms, we obtain

$$f(W_t) - f^* \leq \frac{2L_*(f(W_0) - f^*)D_{\text{op}}^2}{2L_* D_{\text{op}}^2 + t(f(W_0) - f^*)}.$$

□

C.5 Proof of Theorem 4.15

Proof. According to Equation (15), we have

$$f(W_{t+1}) - f(W_t) \leq -\eta \|\nabla f(W_t)\|_* + \frac{\eta^2 J_t}{2} + \frac{s\eta^3 r^{3/2}}{6} \tag{23}$$

Moreover, according to (20), we have

$$-\|\nabla f(W_t)\|_* \leq -\frac{f(W_t) - f^*}{D_{\text{op}}}$$

Combine (23) and (20), we can obtain

$$f(W_{t+1}) - f^* \leq \left(1 - \frac{\eta}{D_{\text{op}}}\right)(f(W_t) - f^*) + \frac{\eta^2 J_t}{2} + \frac{s\eta^3 r^{3/2}}{6}.$$

Then, we have

$$\begin{aligned} f(W_T) - f^* &\leq \left(1 - \frac{\eta}{D_{\text{op}}}\right)^T (f(W_0) - f^*) + \sum_{t=0}^{T-1} \left(1 - \frac{\eta}{D_{\text{op}}}\right)^{T-1-t} \frac{\eta^2 J_t}{2} + \sum_{t=0}^{T-1} \left(1 - \frac{\eta}{D_{\text{op}}}\right)^{T-1-t} \frac{s\eta^3 r^{3/2}}{6} \\ &\leq \left(1 - \frac{\eta}{D_{\text{op}}}\right)^T (f(W_0) - f^*) + \sum_{t=0}^{T-1} \left(1 - \frac{\eta}{D_{\text{op}}}\right)^{T-1-t} \frac{\eta^2 J_t}{2} + \frac{s\eta^2 D_{\text{op}} r^{3/2}}{6}. \end{aligned}$$

Denote $\tilde{J} = \frac{1}{T}\sum_{t=0}^{T-1} \left(1 - \frac{\eta}{D_{\text{op}}}\right)^{T-1-t} J_t$. We have

$$f(W_T) - f^* \leq \left(1 - \frac{\eta}{D_{\text{op}}}\right)^T \Delta + \frac{\eta^2 \tilde{J} T}{2} + \frac{s\eta^2 D_{\text{op}} r^{3/2}}{6}.$$

□

Thus, the convergence complexity of Muon can be depending on $\tilde{J}$, which can be viewed as a weighted average of J_t.

In experiments, we found that usually $J_t > 0$. Note that if f is strict convex and if $U_t V_t^\top \neq 0$, then J_t is strictly larger than 0.

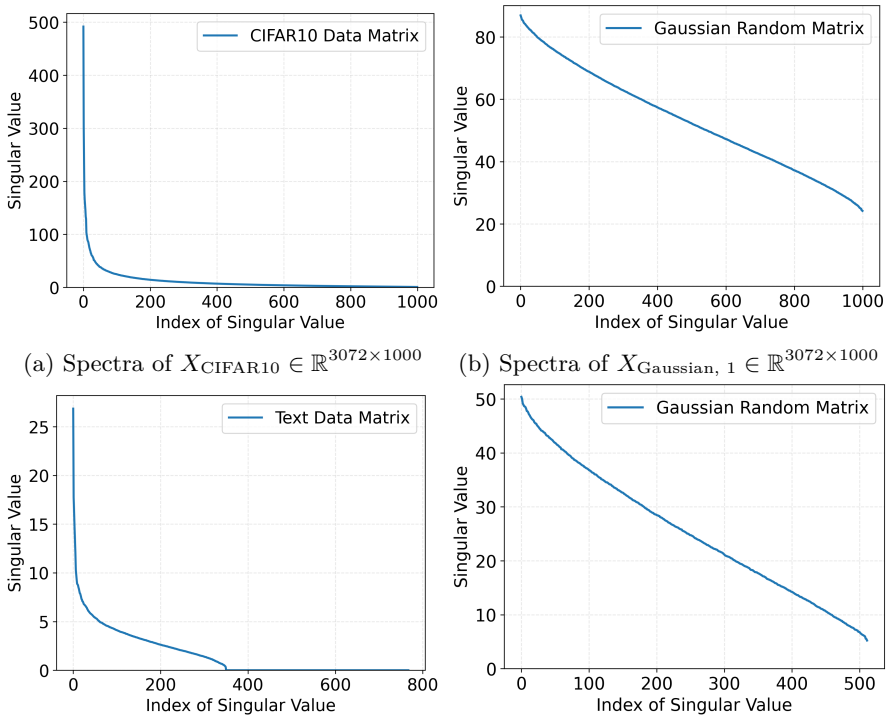
If $\tilde{J} > 0$ and $\eta = \min\left\{\frac{D_{\text{op}}}{T}\log\left(\frac{T\Delta}{D_{\text{op}}^2 \tilde{J}}\right), D_{\text{op}}\right\}$. We have

$$\begin{aligned} f(W_T) - f^* \leq& \frac{D_{\text{op}}^2 \tilde{J}}{T} + \frac{D_{\text{op}}^2 \tilde{J}}{2T}\left[\log\left(\frac{T\Delta}{D_{\text{op}}^2 \tilde{J}}\right)\right]^2 + \frac{s r^{3/2} D_{\text{op}}^3}{6T^2}\left[\log\left(\frac{T\Delta}{D_{\text{op}}^2 \tilde{J}}\right)\right]^2 \\ \leq& \tilde{O}\left(\frac{D_{\text{op}}^2 \tilde{J}}{T}\right). \end{aligned}$$

When $\tilde{J} \leq 0$, we can set $\eta = \min\left\{\frac{D_{\text{op}}}{T}\log\left(\frac{T^2\Delta}{D_{\text{op}}^3 sr^{3/2}}\right), D_{\text{op}}\right\}$. We have

$$\begin{aligned} f(W_T) - f^* \leq & \frac{sr^{3/2}D_{\text{op}}^3}{T^2} + \frac{sr^{3/2}D_{\text{op}}^3}{6T^2}\left[\log\left(\frac{T^2\Delta}{D_{\text{op}}^3 sr^{3/2}}\right)\right]^2 \\ \leq & \tilde{O}\left(\frac{sr^{3/2}D_{\text{op}}^3}{T^2}\right). \end{aligned}$$

D Low-rank Structure of Real-World Data Matrices



(a) Spectra of $X_{\text{CIFAR10}} \in \mathbb{R}^{3072\times 1000}$ (b) Spectra of $X_{\text{Gaussian, 1}} \in \mathbb{R}^{3072\times 1000}$

(c) Spectra of $X_{\text{Text}} \in \mathbb{R}^{768\times 836}$ (d) Spectra of $X_{\text{Gaussian, 2}} \in \mathbb{R}^{768\times 836}$

Figure 5: Spectra of X_{CIFAR10}, X_{Text}, $X_{\text{Gaussian, 1}}$ and $X_{\text{Gaussian, 2}}$.

In addition to demonstrating the low-rank property of MNIST in Figure 2, we also test and verify similar low-rank structures on the CIFAR-10 (Krizhevsky et al., 2009) and Shakespeare datasets[2].

Following a procedure similar to that in Section 4.2.1, for CIFAR-10 we randomly sample 1000 samples from training dataset and construct the matrix $X_{\text{CIFAR10}} \in \mathbb{R}^{3072\times 1000}$. We then compute its singular values and compare them with those of a Gaussian random matrix of the same size, $X_{\text{Gaussian, 1}} \in \mathbb{R}^{3072\times 1000}$. The results are shown in Figure 5(a)(b), where we observe that the singular values of X_{CIFAR10} are significantly more concentrated than $X_{\text{Gaussian, 1}}$.

For the Shakespeare dataset, we take the first 3000 characters and use the RoBERTa (Liu et al., 2019) tokenizer and embedding model to convert the text into an embedding matrix $X_{\text{Text}} \in \mathbb{R}^{768\times 836}$, where 768 is the embedding dimension and 836 is the token length. We compute its singular values and compare them with those of a Gaussian random matrix of the same size, $X_{\text{Gaussian, 2}} \in \mathbb{R}^{768\times 836}$. As shown in Figure 5(c)(d), similar to the image datasets (MNIST and CIFAR-10), the text embedding matrix also exhibits a pronounced low-rank structure.

[2]https://raw.githubusercontent.com/karpathy/char-rnn/master/data/tinyshakespeare/input.txt

E Experimental Settings

Experiments of Figure 2, Figure 3, and Figure 4 are conducted on NVIDIA RTX 6000 Ada GPUs. Experiments of Figure 1 are conducted on Intel(R) Core(TM) i9-14900HX.

In Figure 1, we consider the quadratic function $f(W) = \frac{1}{2}\mathrm{tr}\left((W - W_{\mathrm{opt}})^\top Q(W - W_{\mathrm{opt}})\right)$, $W \in \mathbb{R}^{15\times 20}, Q \in \mathbb{R}^{15\times 15}$, where W_{opt} is randomly chosen and Q is chosen with an ill condition number. Then, we apply GD and Muon to optimize this function, both start from the $W_0 = 0$ with 4000 iterations. We choose the optimal constant stepsize $\frac{1}{L}$ for GD and choose the stepsize for Muon such that Muon can converge in 4000 iterations with the best function value. For each iteration of both algorithms, we record the difference in function value from the optimum and the spectral norm error to the optimal point. The results are shown in Figure 1(a) and (b). In Figure 1(c), we compare the key constant in convergence analysis under quadratic function with ill-conditioned Q. We choose $W_{\mathrm{opt}} \sim U(-50, 50)$, which means each element of W_{opt} is i.i.d. random variables drawn from a continuous uniform distribution. Then, we estimate the key constants $D_{\mathrm{op}}^2 L_*$ and $D_{\mathrm{F}}^2 L$, where we choose the initial point as $W_0 = 0$. We calculate the ratios $\frac{D_{\mathrm{F}}^2 L}{D_{\mathrm{op}}^2 L_*}$ over 10^4 random samples. The results are shown in Figure 1(c).

In Figure 3, we randomly select a fixed subset of 120 samples from MNIST. We then train a fully connected neural network (Table 1) with three matrix parameters ($W^1 \in \mathbb{R}^{128\times 784}$, $W^2 \in \mathbb{R}^{64\times 128}$, $W^3 \in \mathbb{R}^{10\times 64}$) on this subset using GD and Muon (Algorithm 2). We tuned the learning rates for GD and Muon to their optimal values. During the optimization procedure, we record the loss, J_t, L_t, $\|\nabla f(W_t)\|_*^2$, and $\|\nabla f(W_t)\|_{\mathrm{F}}^2$. These quantities are computed with respect to W^2, i.e., $\nabla f = \nabla_{W^2} f(W_t^1, W_t^2, W_t^3)$ and $\nabla^2 f = \nabla^2_{\overline{\mathrm{vec}}(W^2)\overline{\mathrm{vec}}(W^2)} f_v(\overline{\mathrm{vec}}(W_t^1), \overline{\mathrm{vec}}(W_t^2), \overline{\mathrm{vec}}(W_t^3))$, where $f_v(\overline{\mathrm{vec}}(W_t^1), \overline{\mathrm{vec}}(W_t^2), \overline{\mathrm{vec}}(W_t^3)) = f(W_t^1, W_t^2, W_t^3)$.

Table 1: MLP model used in Figure 3. For every Linear module, the bias term is set as false, so this model contain only matrix parameters.

Layer Type	Matrix Shape ($m \times n$)
Fully Connected + ReLU	128×784
Fully Connected + ReLU	64×128
Fully Connected	10×64

In Figure 4, we adopted the code from nanoGPT[3] and trained a 6-layer GPT-2–style model on the Shakespeare character dataset. Model architectural and training hyperparameters can be found in Table 2. We trained the model using AdamW and Muon, respectively. When using Muon, we applied the tricks proposed in Liu et al. (2025a), i.e., using weight decay and adjusting the per-parameter update scale. Muon was used only to optimize the matrix parameters in the non-embedding layers, while the embedding layers and vector parameters were optimized with AdamW.

Table 2: Model architectural and training hyperparameters for the experiments in Figure 4.

Hyperparameter	Value
Number of Layers	6
Number of Heads	6
Hidden Dimension	384
Dropout	0
Batch Size	64
Training Steps	1000

[3]https://github.com/karpathy/nanoGPT