ON THE CONVERGENCE ANALYSIS OF THE ADAMUON OPTIMIZER

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, Muon has demonstrated excellent experimental performance in neural networks. This success has given rise to numerous Muon variants, among which AdaMuon introduces a dynamic adaptive mechanism for large-scale neural network training. However, despite its strong empirical performance, the convergence guarantees hold only under some specific assumptions and are not fully established. To to further narrow the theoretical gap, this paper provides a systematic convergence analysis of AdaMuon. Specifically, we establish convergence analysis under three distinct settings. For nonconvex case under and without Frobenius norm Lipschitz smoothness, we prove the convergence rate is $\mathcal{O}(T^{-1/4})$. For star-convex case under Frobenius norm Lipschitz smoothness, we prove the convergence rate to reach the precision ϵ is $\tilde{\mathcal{O}}(\epsilon^{-3}\log(\epsilon^{-1}))$. These results extend the theoretical scope of the original analysis, offering a more complete characterization of AdaMuon across different optimization assumptions.

1 INTRODUCTION

1.1 BACKGROUND

In recent years, large models have been widely applied in machine learning, and thus the efficient training of large models has received a great deal of attention. Optimizers are a fundamental building block in modern deep learning engineering (Wang & Choromanska, 2025), which are the key component that turns a static parameterized network architecture into a working system capable of generalization. The Muon optimizer (Jordan et al., 2024) is an optimizer designed for matrix parameters in neural networks, which is a representative of the new generation of optimizers that can better adapt to modern large models. Many existing studies and empirical results show that Muon has better performance in optimizing neural networks (Liu et al., 2025)(Shah et al., 2025). Several studies, including Kim & Oh (2026)Shen et al. (2025)Nagashima & Iiduka (2026)Chen et al. (2025), have already conducted theoretical analyses of the convergence properties of Muon, and the theoretical understanding of Muon is now relatively comprehensive.

To meet the requirement of better adapting to modern large model training, many Muon variants have emerged. However, the theoretical convergence analysis of these Muon variants has often not been fully understood, and there remains a certain gap between their practical superior performance and theoretical guarantees. AdaMuon is a Muon variant that not only preserves the main properties of Muon which are well-conditioned and geometry-aware updates, but also combines the efficiency of first-order methods with the robustness of second-order adaptivity (Si et al., 2025).

This paper studies the convergence analysis theory of the AdaMuon optimizer in neural networks. The goal is to obtain the convergence rate of Adamuon under commonly used optimization assumptions in modern practice.

1.2 RELATED WORK

There have been massive studies on the design and convergence analysis of Muon variants.

Convergence analysis of Muon: Shen et al. (2025) presented a comprehensive convergence rate analysis of Muon with polar decomposition and its comparison with Gradient Descent under various smoothness assumptions, including Frobenius and nuclear norm Lipshchitz smoothness and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

star-convex setting. Kim & Oh (2026) proved that Muon with Newton–Schulz iteration converges to a stationary point at the same rate as idealization of polar decomposition. Chen et al. (2025) provided a theoretical analysis of Muon by placing it within the Lion-$\mathcal{K}$ family of optimizers, showing that Muon corresponds to Lion-$\mathcal{K}$ (Chen et al., 2024) when equipped with the nuclear norm, and proposed the theoretical results of Lion-$\mathcal{K}$ to establish that Muon with decoupled weight decay implicitly solves an optimization problem that enforces a constraint on the spectral norm of weight matrix. Nagashima & Iiduka (2026) established sharper convergence guarantees for the Muon optimizer through a direct and simplified analysis that does not rely on restrictive assumptions about the update rule, covering a broader class of problem settings. They summarized the existing results of convergence rate of Muon in different indicators, learning rates, batch sizes and momentum parameters. Sato et al. (2025) provided the convergence analysis for Muon in four practical settings, systematically examining its behavior with and without the inclusion of Nesterov momentum and weight decay and emphasized the importance of addition of weight decay, which ensures almost-sure boundedness of the parameter and gradient norms without relying on the commonly imposed bounded-gradient assumption.

Studies on Muon variants: Chang et al. (2025) analyzed two momentum-based variance-reduced Muon variants inspired by methods related to normalized SGD with momentum variants. Huang et al. (2025) proposed Limuon optimizer based on momentum-based variance reduction technique and randomized singular value decomposition. They showed that Limuon has lower memory and lower sample complexity for finding an ϵ-stationary solution of non-convex stochastic optimization under the generalized smoothness condition. Si et al. (2025) improved Muon by adding an element-wise second momentum estimator and sign-stabilized orthogonal update for large-scale model training. Zhang et al. (2025) proposed AdaGO, which combines a norm-based AdaGrad-type step size (Duchi et al., 2011) with an orthogonal update direction, preserving the orthogonality of the update direction while adapting the step sizes to the optimization landscape.

1.3 Contributions

The main contributions of this paper are summarized as follow:

- We find that the cosine similarity $\cos\theta_t$ between momentum matrix M_t and update matrix $\gamma_t\hat{O}_t$ has a positive global minimum during training, and the cosine value in each iteration increases monotonically as training proceeds. The per-step minimum is an extreme-order statistic rather than a representative one.
- We provide the solid convergence analysis for the AdaMuon optimizer, and prove that the convergence rate of nonconvex function under and without Frobenius norm Lipschitz smoothness assumption is $\mathcal{O}(T^{-1/4})$, and for star-convex function under Frobenius norm Lipschitz smoothness assumption to reach the precision $\mathbb{E}[f(W_T)] - f^* < \epsilon$, the complexity is $\mathcal{O}(\epsilon^{-3}\log(\epsilon^{-1}))$.

2 Preliminaries

2.1 Main problem and assumptions

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

$$\min_{W\in\mathbb{R}^{m\times n}} f(W), \text{ where } f(W) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(W;\xi)],$$

where $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ denotes the loss function of training, W denotes the weight matrix, and ξ denotes a random variable drawn from an unknown distribution $\mathcal{D}$. Without loss of generality, we discuss the case where $m \geq n$ in this paper.

Assumption 2.1 (Frobenius norm Lipschitz smoothness)**.** *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*

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

Assumption 2.1 is a variant of standard Lipschitz smoothness, which serves as a commonly adopted smoothness assumption in optimization problems involving matrix parameters.

Assumption 2.2 (Unbiased estimator with bounded variance)**.** *We assume $\nabla f(W;\xi)$ is an unbiased stochastic estimator and has a bounded variance, i.e.*

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

Definition 2.3 (ϵ-nuclear norm stationary point)**.** *We say W is an ϵ-nuclear norm stationary point of f if we have $\|\nabla f(W)\|_* \le \epsilon$.*

2.2 AdaMuon

In this subsection, we mainly introduce the AdaMuon optimizer and outline our ideas on how to analyze convergence. AdaMuon is a novel optimizer that combines element-wise adaptivity with orthogonal updates for large-scale neural networks (Si et al., 2025). It retains Muon's core idea of using the Newton–Schulz iteration to approximate the polar decomposition of the momentum M_t (Chen & Chow, 2014)(Kovarik, 1970)(Bernstein & Newhouse, 2024), and addresses the following two practical challenges that Muon encounters. First, the root-mean-square (RMS) magnitude of the weight matrix from Muon can become very huge. Second, Muon is applied only to two-dimensional parameters, while one-dimensional parameters are still optimized with Adam (Kingma & Ba, 2014), this requires different learning rates in existing trainings. Adamuon controls the RMS of weights by weight decay to solve the first challenge. For the second challenge, Adamuon maintains compatibility with Adam's learning rate schedules by scaling the RMS norm of update matrix $\hat{O}_t$ to match Adam's empirical RMS value of ≈ 0.2 (Si et al., 2025), then obtains

$$\gamma_t = \frac{0.2}{\mathrm{RMS}(\hat{O}_t)} = \frac{0.2\sqrt{mn}}{\|\hat{O}_t\|_F}. \tag{1}$$

In the following analysis, we denote Adam's empirical RMS value as $c_{rms} = 0.2$.

Algorithm 1 The AdaMuon Optimizer

Input: Initial 2D-weights $W_0 \in \mathbb{R}^{m\times n}$, learning rates η, weight decay λ, momentum β, small constant ε, initial sample $\xi_0 \sim \mathcal{D}$.
Output: Updated weights W_T.
1: Initialize first momentum $M_{-1} \leftarrow 0$, second momentum $V_{-1} \leftarrow 0$.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Draw a sample $\xi_t \sim \mathcal{D}$.
4: $G_t \leftarrow \nabla f(W_t;\xi_t)$.
5: $M_t \leftarrow \beta M_{t-1} + G_t$.
6: $O_t \leftarrow$ Newton-Schulz(Sign(M_t)).
7: $V_t \leftarrow \beta V_{t-1} + (1-\beta) O_t \odot O_t$.
8: $\hat{O}_t \leftarrow O_t \oslash (\sqrt{V_t} + \varepsilon I)$.
9: RMS-aligned: $\gamma_t = 0.2\sqrt{mn}/\|\hat{O}_t\|_F$.
10: $W_{t+1} = W_t - \eta(\gamma_t \hat{O}_t + \lambda W_t)$.
11: **end for**

In line 6 from Algorithm 1, AdaMuon use Newton-Schulz iteration on Sign(M_t) (element-wise sign operation) instead of momentum M_t to approximate the orthogonal matrix, the polar decomposition of (Sign(M_t)). There are two main reasons of applying element-wise sign operation to momentum M_t. First, in the early to middle stages of training, the element-wise fluctuations of the momentum M_t are huge, and the approximate orthogonal matrix Newton-Schulz(M_t) also undergoes huge changes accordingly. Second, if one directly uses Newton-Schulz(M_t) to compute the second-order momentum, the estimated variance is also unstable, failing to provide effective normalization and even amplifying noise. Si et al. (2025) prove in order to stabilize element-wise fluctuations while preserving the global scale invariance of polar decomposition, the only element-wise transformation that satisfies the conditions is the sign function.

In lines 7-8 from Algorithm 1, Adamuon accumulates the second-momentum term on $\hat{O}_t$ to make element-wise variance adaptation. $\odot$ denotes Hadamard product, $\oslash$ denotes element-wise division, $\sqrt{V_t}$ represents the element-wise square root of the variance estimates V_t.

Assumption 2.4. *We assume there exists a constant $\kappa > 0$, for $t = 0, 1, \ldots, T-1$,*

$$\cos\theta_t = \frac{\langle M_t, \gamma_t \hat{O}_t \rangle}{\|M_t\|_{\mathrm{F}} \|\gamma_t \hat{O}_t\|_{\mathrm{F}}} \geq \kappa,$$

where we call $\cos\theta_t$ the cosine similarity between M_t and $\gamma_t \hat{O}_t$ with respect to the Frobenius inner product.

Assumption 2.4 describes a fundamental property of AdaMuon, which is that the update matrix $\gamma_t \hat{O}_t$ remains positively correlated with the momentum M_t throughout training. We note that this assumption is not an artificially imposed requirement, but rather an inherent consequence of the algorithm design. We are going to support it from both theoretical intuition and experimental results.

For theoretical intuition, the element-wise sign operation preserves momentum direction and suppresses noise. The Newton-Schulz iteration $O_t = \mathrm{NewtonSchulz}(\mathrm{Sign}(M_t))$ ensures the approximate consistency of main direction between $\mathrm{Sign}(M_t)$ and O_t. Element-wise variance normalization introduces only bounded perturbations, and RMS rescaling does not alter direction.

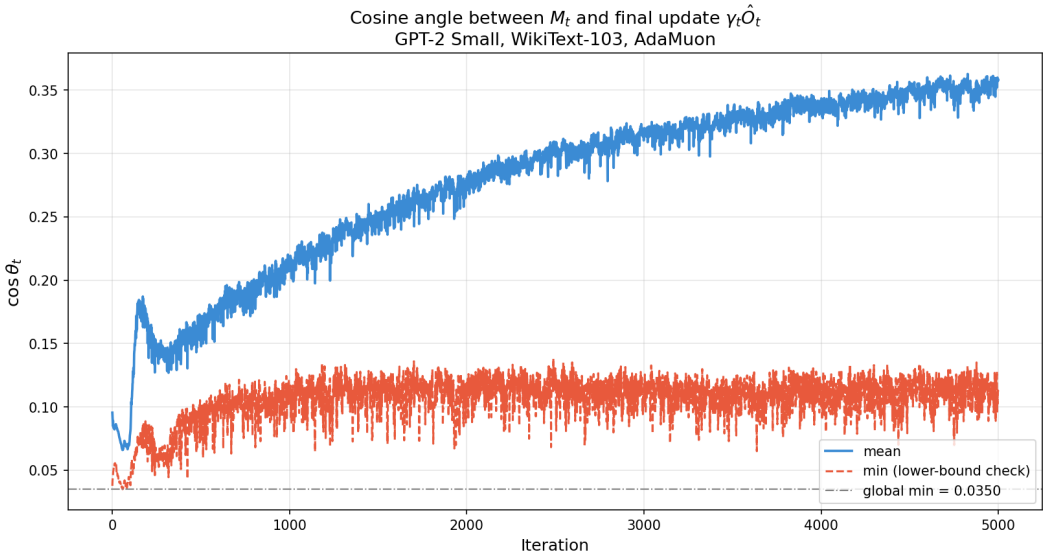

Figure 1: Mean and min cosine values

In the numerical experiment, each training iteration produces 48 cosine values between M_t and $\gamma_t \hat{O}_t$. Detailed experiment background will be given in section 4. Figure 1 shows the mean and minimum cosine similarity between the momentum M_t and the update direction $\gamma_t \hat{O}_t$ over 5,000 training iterations on GPT-2 Small (WikiText-103). For each weight matrix $W_t^{(l)}, l = 1, 2, \ldots, 48$, we compute $\cos\theta_t^{(l)}$ at each step and report the mean and minimum over all 48 weight matrices in the network. Two observations stand out. First, the mean cosine (blue) increases monotonically from approximately 0.09 at the start of training to 0.35 at iteration 5,000, indicating that the update direction aligns better with the momentum direction as training proceeds. Second, the per-step minimum (red) stabilizes in the range of 0.1 to 0.13 after a warm-up phase of roughly 200 iterations. The global minimum of 0.0350 occurs only during this early warm-up period and is never reached again, providing empirical evidence that a positive lower bound on $\cos\theta_t^{(l)}$ exists throughout the entire training.

Figure 2 presents the per-matrix cosine scatter plot, where each point corresponds to one cosine value at one iteration. Figure 2 reveals a key property: the per-step minimum is an extreme-order statistic rather than a representative one. At any given iteration t , the cosine values of the 48 matrices span a wide band, and the vast majority are concentrated near the mean. The minimum is reached by only a small number of matrices at each step, and these matrices vary across iterations.

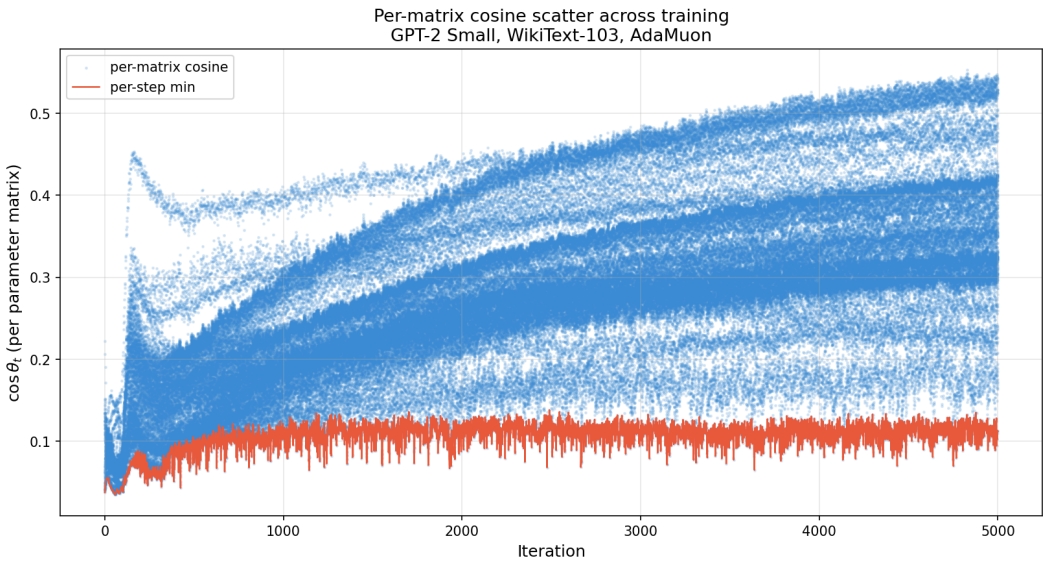

Figure 2: Per-matrix cosine scatter plot

For the later derivation of the convergence analysis, we propose the following lemma.

Lemma 2.5. *We denote $N_t \triangleq (1-\beta)M_t$ for $t = 0, 1, \ldots, T-1$, then by assumption 2.4 we can obtain*

$$\langle N_t, \gamma_t \hat{O}_t \rangle \geq c_0 \|N_t\|_*,$$

where $c_0 \triangleq c_{rms}\kappa\sqrt{m}$.

For convenience, we omit the weight decay, which means setting $\lambda = 0$ to focus on the theoretical analysis of AdaMuon. Adding $\lambda > 0$ modifies constants in a routine way and will not change the convergence rate (Sato et al., 2025). The Newton-Schulz iteration approximates the process of polar decomposition. So, for theoretical analysis we apply Algorithm 2 to analyze the convergence in the rest of the paper.

Algorithm 2 The simplified AdaMuon Optimizer

Input: $W_0 \in \mathbb{R}^{m\times n}$, η, β, ε, $\xi_0 \sim \mathcal{D}$.
Output: W_T.
Initialize $M_{-1} \leftarrow 0$, $V_{-1} \leftarrow 0$.
for $t = 0, 1, \ldots, T-1$ **do**
  Draw a sample $\xi_t \sim \mathcal{D}$.
  $G_t \leftarrow \nabla f(W_t; \xi_t)$.
  $M_t \leftarrow \beta M_{t-1} + G_t$.
  $O_t \leftarrow \text{polar}(\text{Sign}(M_t))$.
  $V_t \leftarrow \beta V_{t-1} + (1-\beta) O_t \odot O_t$.
  $\hat{O}_t \leftarrow O_t \oslash (\sqrt{V_t} + \varepsilon I)$.
  RMS-aligned: $\gamma_t = 0.2\sqrt{mn}/\|\hat{O}_t\|_F$.
  $W_{t+1} = W_t - \eta\gamma_t \hat{O}_t$.
end for

3 CONVERGENCE ANALYSIS

3.1 NONCONVEX CASE UNDER FROBENIUS NORM LIPSCHITZ SMOOTHNESS

Lemma 3.1. *Under Assumption 2.1, 2.2, we can apply Algorithm 2 and then obtain the following estimation for $t = 0, 1, \ldots, T-1$:*

$$\mathbb{E}[\|\nabla f(W_t) - N_t\|_{\mathrm{F}}] \leq \beta^{t+1}\sqrt{2L\Delta} + \frac{c_{crms}\beta L\eta\sqrt{mn}}{1-\beta} + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}},$$

where $c_{rms} = 0.2$ is constant, $\Delta \triangleq \mathbb{E}[f(W_0)] - f^*$, in the rest of the paper we denote $f^* \triangleq \inf_W f(W)$.

Theorem 3.2. *Under Assumption 2.1, 2.2 and Lemma 2.5, 3.1, we can apply Algorithm 2 and then obtain the following estimation:*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \leq & \frac{1}{\kappa}\Bigg(\frac{\Delta}{c_{rms}\eta T\sqrt{m}} + \frac{2\beta\sqrt{2nL\Delta}}{(1-\beta)T} + \frac{2c_{rms}n\beta L\eta\sqrt{m}}{1-\beta} \\ & + \frac{2\sigma\sqrt{n(1-\beta)}}{\sqrt{1+\beta}} + \frac{c_{rms}nL\eta\sqrt{m}}{2}\Bigg),\end{aligned} \tag{2}$$

where $\kappa > 0, c_{rms} = 0.2$ are constants.

Remark. If we set $\eta = \sqrt{\frac{(1-\beta)\Delta}{nmLT}}$ and $1-\beta = \min\left\{\sqrt{\frac{L\Delta}{\sigma^2 T}}, 1\right\}$, in inequality 2, we can obtain the following estimation:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] &\leq \mathcal{O}\left(\sqrt[4]{\frac{n^2\sigma^2 L\Delta}{T}} + \sqrt{\frac{n\sigma^2}{T}} + \sqrt[4]{\frac{n^2L^3\Delta^3}{\sigma^2T^3}}\right) \\ &= \mathcal{O}\left(\frac{1}{T^{1/4}} + \frac{1}{T^{1/2}} + \frac{1}{T^{3/4}}\right).\end{aligned} \tag{3}$$

From inequality 3, we can obtain an ϵ-nuclear norm stationary point of f, and the complexity of this process is $\mathcal{O}(n^2L\sigma^2\Delta\epsilon^{-4})$.

Theorem 3.2 shows that under Frobenius norm Lipschitz smoothness, the AdaMuon optimizer can reach the complexity of $\mathcal{O}(T^{-1/4})$ with a particular selection of learning rate η and momentum β.

Proof Sketch. We briefly show the main ideas of proof. Firstly, we apply the descent lemma (Lemma A.2) for $t = 0, 1, \ldots, T-1$ and relax the inequality by the triangle inequality and the property of update matrix, norm and inner product. Secondly, the main term of the relaxed inequality mentioned above is $\mathbb{E}[\|\nabla f(W_t) - N_t\|_{\mathrm{F}}]$, where $N_t = (1-\beta)M_t$ for $t = 0, 1, \ldots, T-1$. We apply the triangle inequality

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

where

$$C_t = \beta C_{t-1} + (1-\beta)\nabla f(W_t),\ t = 1, 2, \ldots, T-1,\ C_0 = (1-\beta)\nabla f(W_0).$$

We obtain the estimated result in Lemma 3.1 with main terms $t, 1-\beta$ we are going to analyze. Finally, we sum from 0 to $T-1$ and the average on both sides of the main relaxed inequality and choose particular learning rate η and momentum β to obtain the convergence rate. The full details are shown in Appendix B.

3.2 NONCONVEX CASE WITHOUT FROBENIUS NORM LIPSCHITZ SMOOTHNESS

In this subsection, we mainly analyze the estimation of the mean of the expected iterative gradient in the nonconvex setting without Frobenius norm Lipschitz smoothness. Unlike section 3.1, we will no longer adopt smoothness assumption. However, to analyze the convergence deeper, we impose the following weaker assumption.

Definition 3.3. *We denote $f_v : \mathbb{R}^{mn} \to \mathbb{R}$, where $f_v(w) = f(W)$ and $w \triangleq \mathrm{vec}(W)$ for any $W \in \mathbb{R}^{m\times n}$. Especially, we denote the Hessian matrix of f at W_t as $H_t \triangleq \nabla^2 f_v(w_t)$, where $w_t \triangleq \mathrm{vec}(W_t)$ for $t = 0, 1, \ldots, T-1$.*

Definition 3.4. *For any third-order tensor $\mathbf{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$, we denote its operator norm as*

$$\|\mathbf{X}\|_2 = \max_{\|u\|_2=\|v\|_2=\|w\|_2=1} \left| \sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\sum_{k=1}^{d_3} \mathbf{X}_{ijk} u_i v_j w_k \right|,$$

where $u \in \mathbb{R}^{d_1}, v \in \mathbb{R}^{d_2}, w \in \mathbb{R}^{d_3}$.

Assumption 3.5. *We assume for any $w \in \mathbb{R}^{mn}$, we have $\|\nabla^3 f_v(w)\|_2 \le s$.*

Lemma 3.6. *Under Assumption 2.2, 3.5, we can apply Algorithm 2 and then obtain the following estimation for $t = 0, 1, \ldots, T-1$:*

$$\mathbb{E}[\|\nabla f(W_t) - N_t\|_{\mathrm{F}}] \le \beta^{t+1}\mathbb{E}[\|\nabla f(W_0)\|_{\mathrm{F}}] + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}} + \eta\sum_{i=0}^{t-1}\beta^{t-i}\hat{J}_i + \frac{sc_{crms}^2 mn^{3/2}\beta\eta^2}{2(1-\beta)},$$

where $c_{rms} = 0.2$ is constant, $\hat{J}_t \triangleq \mathrm{vec}(U_t V_t^\top)^\top H_t \,\mathrm{vec}(\gamma_t \hat{O}_t)$, and $U_t \in \mathbb{R}^{m\times n}, V_t \in \mathbb{R}^{n\times n}$ are the left and right singular matrices of $\nabla f(W_t) - \nabla f(W_{t+1})$ for $t = 0, 1, \ldots, T-1$, where the columns of U_t are orthogonal and V_t is orthogonal.

Lemma 3.7. *For $t = 0, 1, \ldots, T-1$, we have the following estimation of $\nabla f(W_t) - \nabla f(W_{t+1})$:*

$$\|\nabla f(W_t) - \nabla f(W_{t+1})\|_{\mathrm{F}} \le \eta\hat{J}_t + \frac{sc_{rms}^2 mn^{3/2}\eta^2}{2}. \tag{4}$$

Theorem 3.8. *Under Assumption 2.2, 3.5 and Lemma 2.5, 3.6, we can apply algorithm 2 and then obtain the following estimation:*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \le \frac{1}{\kappa}\Bigg(&\frac{\Delta}{c_{rms}\eta T\sqrt{m}} + \frac{2\beta\sqrt{n}\mathbb{E}[\|\nabla f(W_0)\|_{\mathrm{F}}]}{(1-\beta)T} + \frac{2\sigma\sqrt{n(1-\beta)}}{\sqrt{1+\beta}} + \frac{2\beta J^*\eta\sqrt{n}}{1-\beta} \\ &+ \frac{sc_{rms}^2 mn^2\beta\eta^2}{(1-\beta)T} + \frac{J^*\eta}{2c_{rms}\sqrt{m}} + \frac{sc_{rms}^2 mn^{3/2}\eta^2}{6}\Bigg),\end{aligned} \tag{5}$$

where $\kappa > 0, c_{rms} = 0.2$ are constants, $J^* \triangleq \max\{\mathbb{E}[\hat{J}], \mathbb{E}[J]\}$, $\hat{J} \triangleq \frac{1}{T}\sum_{t=0}^{T-2}\hat{J}_t$, $J \triangleq \frac{1}{T}\sum_{t=0}^{T-1} J_t$, $J_t \triangleq \mathrm{vec}(\gamma_t\hat{O}_t)^\top H_t \,\mathrm{vec}(\gamma_t\hat{O}_t)$ for $t = 0, 1, \ldots, T-1$.

Remark. If we set $\eta = \sqrt{\frac{(1-\beta)\Delta}{m^{1/2}n^{1/2}J^*T}}$ and $1-\beta = \min\left\{\sqrt{\frac{J^*\Delta}{m^{1/2}n^{1/2}\sigma^2 T}}, 1\right\}$ in inequality 5, we can obtain the following estimation:

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \\ &\le \mathcal{O}\Bigg(\sqrt[4]{\frac{n^{3/2}J^*\sigma^2\Delta}{m^{1/2}T}} + \sqrt{\frac{m^{1/2}n^{3/2}\sigma^2(\mathbb{E}[\|\nabla f(W_0)\|_{\mathrm{F}}])^2}{J^*T\Delta}} + \sqrt[4]{\frac{J^{*3}\Delta^3}{m^{7/2}n^{3/2}\sigma^2T^3}} \\ &\quad + \sqrt{\frac{s^2m^{1/2}n^{3/2}\Delta^3}{J^*\sigma^2T^3}} + \frac{sm^{1/2}n^{3/2}\Delta}{J^*T^2}\Bigg) \\ &= \mathcal{O}\left(\frac{1}{T^{1/4}} + \frac{1}{T^{1/2}} + \frac{1}{T^{3/4}} + \frac{1}{T^{3/2}} + \frac{1}{T^2}\right).\end{aligned} \tag{6}$$

From inequality 6, we can obtain an ϵ-nuclear norm stationary point of f, and the complexity of this process is $\mathcal{O}(m^{-1/2}n^{3/2}\sigma^2J^*\Delta\epsilon^{-4})$.

Theorem 3.8 shows that without Frobenius norm Lipschitz smoothness, the AdaMuon optimizer can reach the complexity of $\mathcal{O}(T^{-1/4})$ with a particular selection of learning rate η and momentum β.

Although the main convergence rate is equal to Frobenius norm Lipschitz smoothness case mentioned in Theorem 3.2, the constant of $T^{-1/4}$ is different. Theoretically, the bound in Theorem 3.8 is larger. J is considered as a scale of Hessian, and Shen et al. (2025) discuss how the property of Hessian influences the convergence rate in detail.

Proof Sketch. We briefly show the main ideas of proof. Due to losing the Frobenius norm Lipschitz smoothness, we can't apply the descent lemma straightly. However, we can take the Taylor expansion of $f(W_t)$ for $t = 0, 1, \ldots, T-1$ and relax the inequality by the triangle inequality and the property of update matrix, norm and inner product, where the technique is similar to the first step of proof sketch of Theorem 3.2. For the estimation of terms involving $\|\nabla^3 f_v(w)\|_2$, we apply the Assumption 3.5. Similarly, the main term we need to estimate is

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

Secondly, we need to obtain the estimation of $\|\nabla f(W_t) - \nabla f(W_{t+1})\|_{\mathrm{F}}$ due to losing the Frobenius norm Lipschitz smoothness and the result is mentioned in Lemma 3.7. Finally, we sum from 0 to $T-1$ and the average on both sides of the main relaxed inequality and choose particular learning rate η and momentum β to obtain the convergence rate. The full details are shown in Appendix C.

3.3 Star-convex case under Frobenius norm Lipschitz smoothness

Star-convex functions (Nesterov & Polyak, 2006) arise as loss functions in non-convex machine learning contexts, including, for data points X_i, parameter vector θ, and any real exponent p, the loss function

$$h_{\theta,X}(\hat{\theta}) = \left(\sum_i \left| (\hat{\theta} - \theta) X_i \right|^p \right)^{1/p},$$

which is called the L_p-norm loss (Lee & Valiant, 2016). Note that the star-convex functions are nonconvex functions, including convex functions as a special case. Some empirical evidence suggests that the optimization path of neural networks can satisfy the star-convexity (Zhou et al., 2019) and the star-convexity can be used as the geometric constraint for reshaping loss landscapes (Zhang et al., 2024).

Definition 3.9 (Star-convex functions)**.** *A function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is star-convex if there is an optimal point $W^* \in \arg\min_W f(W)$ such that for all $\alpha \in [0,1]$ and $W \in \mathbb{R}^{m\times n}$,*

$$f(\alpha W^* + (1-\alpha)W) \leq \alpha f(W^*) + (1-\alpha) f(W).$$

Lemma 3.10. *For any $W \in \mathbb{R}^{m\times n}$, the following inequality holds if f is star-convex,*

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

To ensure the convergence of AdaMuon in the star-convex case, we adopt the following reasonable assumption.

Assumption 3.11. *For W_t, $t = 0, 1, \ldots, T$, generated by Algorithm 2, we assume $\|W_t - W^*\|_2 \leq D$.*

Note that the same assumption is also used in Gupta et al. (2018), Shen et al. (2025) and An et al. (2026) for their convergence analysis in similar settings.

Theorem 3.12. *Under Assumption 2.1, 2.2, 3.11 and Lemma 2.5 3.11 we can apply Algorithm 2 and then obtain the following estimation:*

$$\begin{aligned} &\mathbb{E}[f(W_T)] - f^* \\ &\leq \Delta \left(1 - \frac{\eta c_0}{D}\right)^T + \frac{2 c_{rms} \beta \eta \sqrt{nmL}}{1-\beta} \left(\frac{D\sqrt{nL}}{\kappa} + \sqrt{2\Delta} \right) + \frac{2\sigma D \sqrt{n(1-\beta)}}{\kappa \sqrt{1+\beta}} + \frac{c_{rms} n D L \eta \sqrt{m}}{2\kappa}, \end{aligned} \tag{7}$$

where $c_0 = c_{rms}\kappa\sqrt{m}$, $\kappa > 0$ is a constant.

Remark. If we want to reach the precision $\mathbb{E}[f(W_T)] - f^* < \epsilon$ in inequality 7, we can set $\eta = \frac{\kappa \epsilon^3}{n m^{1/2} L D}$, then the complexity of reaching the above precision is $\mathcal{O}(nLD^2\kappa^{-2}\epsilon^{-3}\log(\Delta\epsilon^{-1}))$.

Theorem 3.12 shows that in star-convex case under Frobenius norm Lipschitz smoothnes, the AdaMuon optimizer can reach the complexity of $\mathcal{O}(\epsilon^{-1/3}\log(\epsilon^{-1}))$ to reach the precision ϵ with a particular selection of learning rate η .

Proof Sketch. We briefly show the main ideas of proof. Firstly, we apply the descent lemma (Lemma A.2) for $t = 0, 1, \ldots, T-1$ and obtain the same result as the case with Frobenius norm Lipschitz smoothness. Moreover, we can obtain the estimation of $\mathbb{E}[\|\nabla f(W_t)\|_*]$ due to the star-convex assumption, which is

$$\mathbb{E}[\|\nabla f(W_t)\|_*] \geq \frac{\mathbb{E}[f(W_t)] - f^*}{D}.$$

Therefore, the indicator can be $\mathbb{E}[f(W_T)] - f^*$, which is different from the indicator $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*]$ in nonconvex case with and without Frobenius norm Lipschitz smoothness. The full details are shown in Appendix D.

4 Numerical experiment setup

This experiment is conducted on GPT-2 Small, a 12-layer Transformer with hidden dimension 768, 12 attention heads, and approximately 124M parameters. The training dataset is WikiText-103 (approximately 103M BPE tokens), tokenized using the GPT-2 tokenizer. Each training step processes a batch of 4 sequences of length 512, corresponding to 2048 tokens per step. The optimizer configuration follows the original AdaMuon repository: AdaMuon updates the two-dimensional weight matrices inside the Transformer blocks, while AdamW handles the token embedding, language model head, and LayerNorm parameters. Both optimizers share a learning rate of 6×10^{-4} with a linear warm-up over the first 1,000 steps. AdaMuon is configured with momentum $\beta = 0.95$, Newton-Schulz iteration steps 5, second-moment stabilizer $\varepsilon = 10^{-8}$, and weight decay $\lambda = 0.1$. AdamW is configured with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay $\lambda = 0.1$.

Each Transformer block contains four 2D weight matrices: the attention input projection, the attention output projection, the MLP expansion matrix, and the MLP compression matrix, which yield $12 \times 4 = 48$ weight matrices in total. At each training iteration, AdaMuon updates each of these 48 matrices, and the update matrices are composed through the sequential forward pass of the Transformer to produce the model output at the next iteration.

AI use statement

In this work, we used generative AI tools for assistive writing and related work searching. We have not used generative AI tools for research design, theorem analysis, or experiment results. Additionally, we used generative AI tools for improving clarity of expression, correcting grammatical errors, polishing word choice and searching part of related work. We checked LLM-generated list of related work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. Asgo: Adaptive structured gradient optimization. *Advances in Neural Information Processing Systems*, 38:126775–126814, 2026.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. *arXiv preprint arXiv:2410.21265*, 2024.

Da Chang, Yongxiang Liu, and Ganzhao Yuan. On the convergence of muon and beyond. *arXiv preprint arXiv:2509.15816*, 2025.

Jie Chen and Edmond Chow. A stable scaling of newton-schulz for improving the sign function computation of a hermitian matrix. *Preprint]. ANL/MCS-P5059-0114*, 2014.

Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves a constrained optimization: As lyapunov predicts. In *International Conference on Learning Representations*, volume 2024, pp. 35404–35439, 2024.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 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.

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

Feihu Huang, Yuning Luo, and Songcan Chen. Limuon: Light and fast muon optimizer for large models. *arXiv preprint arXiv:2509.14562*, 2025.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan.github.io/posts/muon*, 6(3):4, 2024.

Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with newton-schulz. *arXiv preprint arXiv:2601.19156*, 2026.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970.

Jasper CH Lee and Paul Valiant. Optimizing star-convex functions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 603–614. IEEE, 2016.

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*, 2025.

Shuntaro Nagashima and Hideaki Iiduka. Improved convergence rates of muon optimizer for nonconvex optimization. *arXiv preprint arXiv:2601.19400*, 2026.

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

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of muon optimizer. *arXiv preprint arXiv:2507.01598*, 2025.

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of muon. *arXiv preprint arXiv:2505.23737*, 2025.

Chongjie Si, Debing Zhang, and Wei Shen. Adamuon: Adaptive muon optimizer. *arXiv preprint arXiv:2507.11005*, 2025.

Jing Wang and Anna Choromanska. A survey of optimization methods for training dl models: Theoretical perspective on convergence and generalization. *arXiv preprint arXiv:2501.14458*, 2025.

Minxin Zhang, Yuxuan Liu, and Hayden Schaeffer. Adagrad meets muon: Adaptive stepsizes for orthogonal updates. *arXiv preprint arXiv:2509.02981*, 2025.

Ziming Zhang, Yuping Shao, Yiqing Zhang, Fangzhou Lin, Haichong Zhang, and Elke Rundensteiner. Deep loss convexification for learning iterative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1501–1513, 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.

A Appendix

A.1 Proofs of lemma and remarks

Remark of Assumption 2.2 (Unbiased estimator and with bounded variance): throughout the paper, all expectations involving W_t and ξ_t are understood to be conditional on the history up to t i.e., conditional on $\mathcal{F}_t = \sigma(\xi_0, \xi_1, \ldots, \xi_{t-1})$. Thus, we have

$$\mathbb{E}[\nabla f(W_t; \xi_t) \mid \mathcal{F}_t] = \nabla f(W_t), \quad \mathbb{E}\left[\|\nabla f(W; \xi) - \nabla f(W)\|_{\mathrm{F}}^2 \mid \mathcal{F}_t\right] \leq \sigma^2. \tag{8}$$

Lemma A.1. *For $i = 0, 1, \ldots, t$, we denote $\zeta_i = \nabla f(W_i) - \nabla f(W_i; \xi_i)$, which is the stochastic variant, then*

$$\mathbb{E}\left[\left\|\sum_{i=0}^{t} \zeta_i\right\|_{\mathrm{F}}^2\right] = \sum_{i=0}^{t} \mathbb{E}\left\|\zeta_i\right\|_{\mathrm{F}}^2,$$

which means $\zeta_0, \zeta_1, \ldots, \zeta_t$ are i.i.d.

Proof. For any $j < k \in \{0, 1, \ldots, t\}$, we have

$$\mathbb{E}[\langle \zeta_j, \zeta_k \rangle] = \mathbb{E}[\mathbb{E}[\langle \zeta_j, \zeta_k \rangle] \mid \mathcal{F}_k] = \mathbb{E}[\langle \zeta_j, \mathbb{E}[\zeta_k \mid \mathcal{F}_k] \rangle] = 0,$$

where the second equation above holds because ζ_j is deterministic conditional on $\mathcal{F}_k$ for $j < k$, the third equation above holds due to equation 8. □

Lemma A.2. *If $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ is L Frobenius norm Lipschitz smooth, then for any $W, W' \in \mathbb{R}^{m \times n}$,*

$$f(W') \leq f(W) + \langle \nabla f(W), W' - W \rangle + \frac{L}{2} \|W' - W\|_{\mathrm{F}}^2.$$

Proof. We denote $g(t) = f(W + t(W' - W)), t \in [0, 1]$. By the fundamental theorem of calculus, we have

$$f(W') - f(W) = g(1) - g(0) = \int_0^1 g'(t) \, \mathrm{d}t = \int_0^1 \langle \nabla f(W + t(W' - W)), W' - W \rangle \, \mathrm{d}t.$$

Thus, we have

$$f(W') - f(W) = \langle \nabla f(W), W' - W \rangle + \int_0^1 \langle \nabla f(W + t(W' - W)) - \nabla f(W), W' - W \rangle \, \mathrm{d}t.$$

Thus, we have

$$\begin{aligned} f(W') - f(W) - \langle \nabla f(W), W' - W \rangle &= \int_0^1 \langle \nabla f(W + t(W' - W)) - \nabla f(W), W' - W \rangle \, \mathrm{d}t \\ &\leq \int_0^1 \|\nabla f(W + t(W' - W)) - \nabla f(W)\|_{\mathrm{F}} \|W' - W\|_{\mathrm{F}} \, \mathrm{d}t \\ &\leq \int_0^1 tL \|W' - W\|_{\mathrm{F}}^2 \, \mathrm{d}t \\ &= \frac{L}{2} \|W' - W\|_{\mathrm{F}}^2. \end{aligned} \tag{9}$$

Then we obtain the result by simple transposition of inequality 9 □

Lemma A.3. *According to Lemma A.2, for any $W \in \mathbb{R}^{m \times n}$, we have*

$$\|\nabla f(W)\|_{\mathrm{F}} \leq \sqrt{2L(f(W) - f^*)}.$$

Proof. In Lemma A.1, let $W' = W - \frac{1}{L}\nabla f(W)$, then we have

$$\begin{aligned} f\left(W - \frac{1}{L}\nabla f(W)\right) &\leq f(W) + \left\langle \nabla f(W), W - \frac{1}{L}\nabla f(W) - W \right\rangle + \frac{L}{2}\left\| W - \frac{1}{L}\nabla f(W) - W \right\|_{\mathrm{F}}^2 \\ &= f(W) - \frac{1}{L}\|\nabla f(W)\|_{\mathrm{F}}^2 + \frac{1}{2L}\|\nabla f(W)\|_{\mathrm{F}}^2 \\ &= f(W) - \frac{1}{2L}\|\nabla f(W)\|_{\mathrm{F}}^2. \end{aligned}$$

Since $f^* \leq f\left(W - \frac{1}{L}\nabla f(W)\right)$, we have

$$\|\nabla f(W)\|_{\mathrm{F}}^2 \leq 2L(f(W) - f^*).$$

By taking square roots on both sides of inequality above we can yield the result. □

Proof of Lemma 2.5:

Proof. By Assumption 2.4, we have

$\langle N_t, \gamma_t \hat{O}_t \rangle = \langle (1-\beta)M_t, \gamma_t \hat{O}_t \rangle \geq \kappa \|(1-\beta)M_t\|_{\mathrm{F}} \|\gamma_t \hat{O}_t\|_{\mathrm{F}} = c_{rms}\kappa\sqrt{mn}\|N_t\|_{\mathrm{F}} \geq c_{rms}\kappa\sqrt{m}\|N_t\|_*$,
where we denote $c_0 = c_{rms}\kappa\sqrt{m}$ □

Proof of Lemma 3.10:

Proof. By Definition 3.9, for any $\alpha \in (0, 1]$ we have

$$f(W + \alpha(W^* - W)) - f(W) \leq \alpha(f(W^*) - f(W)).$$

Divide both sides by α, then

$$\frac{f(W + \alpha(W^* - W)) - f(W)}{\alpha} \leq f(W^*) - f(W).$$

Take the limit on both sides as $\alpha \to 0^+$, then

$$\lim_{\alpha \to 0^+} \frac{f(W + \alpha(W^* - W)) - f(W)}{\alpha} = \langle \nabla f(W), W^* - W \rangle \leq f(W^*) - f(W).$$

Thus, we have $\langle \nabla f(W), W - W^* \rangle \geq f(W) - f^*$. □

B Nonconvex case under Frobenius norm Lipschitz smoothness

B.1 Proof of Lemma 3.1

Proof. We denote $C_t \triangleq \beta C_{t-1} + (1-\beta)\nabla f(W_t)$, for $t = 1, 2, \ldots, T-1$, and $C_0 = (1-\beta)\nabla f(W_0)$, thus by induction we have $C_t = (1-\beta)\sum_{i=0}^{t} \beta^{t-i} \nabla f(W_i)$. By the triangle inequality, we have

$$\mathbb{E}[\|\nabla f(W_t) - N_t\|_{\mathrm{F}}] \leq \underbrace{\mathbb{E}[\|\nabla f(W_t) - C_t\|_{\mathrm{F}}]}_{(A)} + \underbrace{\mathbb{E}[\|C_t - N_t\|_{\mathrm{F}}]}_{(B)}. \tag{10}$$

For term (A),

$$\begin{aligned} \mathbb{E}[\|\nabla f(W_t) - C_t\|_{\mathrm{F}}] &= \mathbb{E}[\|\nabla f(W_t) - \beta C_{t-1} - (1-\beta)\|\nabla f(W_t)\|_{\mathrm{F}}] \\ &= \beta \mathbb{E}[\|\nabla f(W_t) - C_{t-1}\|_{\mathrm{F}}] \\ &\leq \beta \mathbb{E}[\|\nabla f(W_{t-1}) - C_{t-1}\|_{\mathrm{F}}] + \beta \mathbb{E}[\|\nabla f(W_t) - \nabla f(W_{t-1})\|_{\mathrm{F}}] \\ &\overset{\text{(i)}}{\leq} \beta \mathbb{E}[\|\nabla f(W_{t-1}) - C_{t-1}\|_{\mathrm{F}}] + \beta L \mathbb{E}[\|W_t - W_{t-1}\|_{\mathrm{F}}] \\ &= \beta \mathbb{E}[\|\nabla f(W_{t-1}) - C_{t-1}\|_{\mathrm{F}}] + \beta L \mathbb{E}[\|\eta \gamma_t \hat{O}_t\|_{\mathrm{F}}] \\ &= \beta \mathbb{E}[\|\nabla f(W_{t-1}) - C_{t-1}\|_{\mathrm{F}}] + c_{rms}\beta L \eta \sqrt{mn} \\ &\overset{\text{(ii)}}{\leq} \beta^t \mathbb{E}[\|\nabla f(W_0) - C_0\|_{\mathrm{F}}] + \sum_{i=1}^{t} \beta^i c_{rms} L \eta \sqrt{mn} \\ &\overset{\text{(iii)}}{\leq} \beta^{t+1}\sqrt{2L\Delta} + \frac{c_{rms}\beta L \eta \sqrt{mn}}{1-\beta}, \end{aligned} \tag{11}$$

where inequality (i) above holds due to Assumption 2.1, inequality (ii) above holds due to induction, inequality (iii) above holds due to Lemma A.3 by replacing W with W_0.

For term (B),

$$\begin{aligned}
\mathbb{E}[\|C_t - N_t\|_{\mathrm{F}}] &= (1-\beta)\mathbb{E}\left[\left\|\sum_{i=0}^{t}\beta^{t-i}(\nabla f(W_i) - G_i)\right\|_{\mathrm{F}}\right] \\
&\overset{\text{(i)}}{\leq} (1-\beta)\sqrt{\mathbb{E}\left[\left\|\sum_{i=0}^{t}\beta^{t-i}(\nabla f(W_i) - \nabla f(W_i;\xi_i))\right\|_{\mathrm{F}}^2\right]} \\
&\overset{\text{(ii)}}{=} (1-\beta)\sqrt{\sum_{i=0}^{t}\beta^{2(t-i)}\mathbb{E}\left[\|\nabla f(W_i) - \nabla f(W_i;\xi_i)\|_{\mathrm{F}}^2\right]} \\
&\overset{\text{(iii)}}{\leq} (1-\beta)\sqrt{\frac{(1-\beta^{2t+2})\sigma^2}{1-\beta^2}} \\
&\leq \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}}, \qquad (12)
\end{aligned}$$

where inequality (i) above holds due to Jensen inequality, the above equation (ii) above holds due to Lemma A.1, inequality (iii) above holds due to equation 8.

Thus by inequality 11, 12, inequality 10 can be written as

$$\mathbb{E}[\|\nabla f(W_t) - N_t\|_{\mathrm{F}}] \leq \beta^{t+1}\sqrt{2L\Delta} + \frac{c_{crms}\beta L\eta\sqrt{mn}}{1-\beta} + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}}.$$

□

B.2 Proof of Theorem 3.2

Proof. By Lemma A.2, for $t = 0, 1, \ldots, T-1$ we have

$$\begin{aligned}
\mathbb{E}[f(W_{t+1}) - f(W_t)] &\leq \mathbb{E}\left[\langle \nabla f(W_t), W_{t+1} - W_t\rangle + \frac{L}{2}\|W_{t+1} - W_t\|_{\mathrm{F}}^2\right] \\
&\leq \mathbb{E}\left[-\eta\langle \nabla f(W_t), \gamma_t \hat{O}_t\rangle + \frac{L}{2}\|\gamma_t \hat{O}_t\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[-\eta\langle N_t, \gamma_t \hat{O}_t\rangle + \eta\langle N_t - \nabla f(W_t), \gamma_t \hat{O}_t\rangle + \frac{c_{rms}^2 mnL\eta^2}{2}\right] \\
&\overset{\text{(i)}}{\leq} \mathbb{E}\left[-\eta c_0\|N_t\|_* + \eta\|N_t - \nabla f(W_t)\|_{\mathrm{F}}\|\gamma_t \hat{O}_t\|_{\mathrm{F}} + \frac{c_{rms}^2 mnL\eta^2}{2}\right] \\
&\leq \mathbb{E}\left[-\eta c_0\left(\|\nabla f(W_t)\|_* - \|\nabla f(W_t) - N_t\|_*\right) + c_{rms}\eta\sqrt{mn}\|N_t - \nabla f(W_t)\|_{\mathrm{F}} + \frac{c_{rms}^2 mnL\eta^2}{2}\right] \\
&\leq \mathbb{E}\left[-\eta c_0\|\nabla f(W_t)\|_* + \left(c_{rms}\eta\kappa\sqrt{mn} + c_{rms}\eta\sqrt{mn}\right)\|\nabla f(W_t) - N_t\|_{\mathrm{F}} + \frac{c_{rms}^2 mnL\eta^2}{2}\right] \\
&\leq \mathbb{E}\left[-\eta c_0\|\nabla f(W_t)\|_* + 2c_{rms}\eta\sqrt{mn}\|\nabla f(W_t) - N_t\|_{\mathrm{F}} + \frac{c_{rms}^2 mnL\eta^2}{2}\right], \qquad (13)
\end{aligned}$$

where inequality (i) above holds due to Lemma 2.5. By Lemma 3.1 we have

$$\begin{aligned}
&\mathbb{E}[f(W_{t+1}) - f(W_t)] \\
&\leq -\eta c_0\mathbb{E}[\|\nabla f(W_t)\|_*] + 2c_{rms}\eta\sqrt{mn}\left(\beta^{t+1}\sqrt{2L\Delta} + \frac{c_{rms}\beta L\eta\sqrt{mn}}{1-\beta} + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}}\right) + \frac{c_{rms}^2 mnL\eta^2}{2}. \qquad (14)
\end{aligned}$$

Then sum from 0 to $T-1$ and take the average on both sides of inequality 14, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \\
&\leq \frac{\mathbb{E}[\nabla f(W_0)-\nabla f(W_T)]}{c_0\eta T}+\frac{2c_{rms}\sqrt{mn}}{c_0}\left(\frac{\sum_{t=0}^{T-1}\beta^{t+1}\sqrt{2L\Delta}}{T}+\frac{c_{rms}\beta L\eta\sqrt{mn}}{1-\beta}+\sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}}\right)+\frac{c_{rms}^2 mnL\eta^2}{2c_0} \\
&\leq \frac{1}{\kappa}\left(\frac{\Delta}{c_{rms}\eta T\sqrt{m}}+\frac{2\beta\sqrt{2nL\Delta}}{(1-\beta)T}+\frac{2c_{rms}n\beta L\eta\sqrt{m}}{1-\beta}+\frac{2\sigma\sqrt{n(1-\beta)}}{\sqrt{1+\beta}}+\frac{c_{rms}nL\eta\sqrt{m}}{2}\right).
\end{aligned}$$

□

By Theorem 3.2, if we set $\eta=\sqrt{\frac{(1-\beta)\Delta}{nmLT}}$ in inequality 2, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*]\leq\mathcal{O}\left(\sqrt{\frac{nL\Delta}{(1-\beta)T}}+\frac{\sqrt{nL\Delta}}{(1-\beta)T}+\sigma\sqrt{n(1-\beta)}+\sqrt{\frac{nL\Delta(1-\beta)}{T}}\right). \tag{15}$$

Then we set $1-\beta=\min\left\{\sqrt{\frac{L\Delta}{\sigma^2T}},1\right\}$ in inequality 15, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*]\leq\mathcal{O}\left(\sqrt[4]{\frac{n^2\sigma^2L\Delta}{T}}+\sqrt{\frac{n\sigma^2}{T}}+\sqrt[4]{\frac{n^2L^3\Delta^3}{\sigma^2T^3}}\right). \tag{16}$$

Thus, we can obtain an ϵ-nuclear norm stationary point of f from inequality 16, and the complexity of this process is $\mathcal{O}(n^2L\sigma^2\Delta\epsilon^{-4})$.

C Nonconvex case without Frobenius norm Lipschitz smoothness

C.1 Proof of Lemma 3.7

Proof. Suppose that the SVD of $\nabla f(W_t)-\nabla f(W_{t+1})$ is $U_t\Sigma_tV_t^\top$, where $U_t\in\mathbb{R}^{m\times n}$, $V_t\in\mathbb{R}^{n\times n}$, $\Sigma_t\in\mathbb{R}^{n\times n}$, then we have

$$\begin{aligned}
\|\nabla f(W_t)-\nabla f(W_{t+1})\|_{\mathrm{F}} &\leq \|\nabla f(W_t)-\nabla f(W_{t+1})\|_* \\
&= \|\Sigma_t\|_* \\
&= \mathrm{tr}(\Sigma_t) \\
&= \mathrm{tr}\left(U_t^\top(\nabla f(W_t)-\nabla f(W_{t+1}))V_t\right) \\
&= \mathrm{tr}\left(V_tU_t^\top(\nabla f(W_1)-\nabla f(W_{t+1}))\right) \\
&= \mathrm{vec}\left(\left(V_tU_t^\top\right)^\top\right)^\top\mathrm{vec}(\nabla f(W_t)-\nabla f(W_{t+1})).
\end{aligned} \tag{17}$$

Take the Taylor expansion of the i-th element of $\mathrm{vec}(\nabla f(W_t))$ for $i=0,1,\ldots,T-1$, which is

$$\begin{aligned}
&\mathrm{vec}(\nabla f(W_{t+1}))_i-\mathrm{vec}(\nabla f(W_t))_i \\
&=\sum_{j=1}^{mn}\left[\nabla^2 f_v(w_t)\right]_{ij}\mathrm{vec}(W_{t+1}-W_t)_j+\frac{1}{2}\sum_{j,k=1}^{mn}\left[\nabla^3 f_v(\tilde{\theta}_t)\right]_{ijk}\mathrm{vec}(W_{t+1}-W_t)_j\mathrm{vec}(W_{t+1}-W_t)_k \\
&=\eta\sum_{j=1}^{mn}(H_t)_{ij}\,\mathrm{vec}(\gamma_t\hat{O}_t)_j+\frac{\eta^2}{2}\sum_{j,k=1}^{mn}\left[\nabla^3 f_v(w_t)\right]_{ijk}\mathrm{vec}(\gamma_t\hat{O}_t)_j\mathrm{vec}(\gamma_t\hat{O}_t)_k,
\end{aligned} \tag{18}$$

where $\tilde{\theta}_t \in \mathbb{R}^{mn}$ is some vector between w_t and w_{t+1}. Then combining equation 18 with inequality 17, we have

$$
\begin{aligned}
&\|\nabla f(W_{t-1}) - \nabla f(W_t)\|_{\mathrm{F}} \\
&\leq \eta \sum_{i,j=1}^{mn} \mathrm{vec}(U_t V_t^\top)_i \, (H_t)_{ij} \, \mathrm{vec}(\gamma_t \hat{O}_t)_j + \frac{\eta^2}{2} \sum_{i,j,k=1}^{mn} \left[\nabla^3 f_v(\tilde{\theta}_t)\right]_{ijk} \mathrm{vec}(U_t V_t^\top)_i \mathrm{vec}(\gamma_t \hat{O}_t)_j \mathrm{vec}(\gamma_t \hat{O}_t)_k \\
&\leq \eta \hat{J}_t + \frac{\eta^2}{2} \|\nabla^3 f_v(\tilde{\theta}_t)\|_2 \|\mathrm{vec}(U_t V_t^\top)\|_2 \|\mathrm{vec}(\gamma_t \hat{O}_t)\|_2^2 \\
&\leq \eta \hat{J}_t + \frac{s c_{rms}^2 m n^{3/2} \eta^2}{2}.
\end{aligned}
$$

where the last inequality above holds due to Assumption 3.5. □

C.2 Proof of Lemma 3.6

Proof. By the triangle inequality, we have

$$
\mathbb{E}[\|\nabla f(W_t) - N_t\|_{\mathrm{F}}] \leq \underbrace{\mathbb{E}[\|\nabla f(W_t) - C_t\|_{\mathrm{F}}]}_{(A)} + \underbrace{\mathbb{E}[\|C_t - N_t\|_{\mathrm{F}}]}_{(B)}. \tag{19}
$$

For term (A),

$$
\begin{aligned}
\mathbb{E}[\|\nabla f(W_t) - C_t\|_{\mathrm{F}}] &= \mathbb{E}[\|\nabla f(W_t) - \beta C_{t-1} - (1-\beta)\nabla f(W_t)\|_{\mathrm{F}}] \\
&= \beta \mathbb{E}[\|\nabla f(W_t) - C_{t-1}\|_{\mathrm{F}}] \\
&\leq \beta \mathbb{E}[\|\nabla f(W_{t-1}) - C_{t-1}\|_{\mathrm{F}}] + \beta \mathbb{E}[\|\nabla f(W_{t-1}) - \nabla f(W_t)\|_{\mathrm{F}}] \\
&\overset{\text{(i)}}{\leq} \beta \mathbb{E}\left[\|\nabla f(W_{t-1}) - C_{t-1}\|_{\mathrm{F}}] + \eta \hat{J}_{t-1} + \frac{s c_{rms}^2 m n^{3/2} \eta^2}{2}\right] \\
&\overset{\text{(ii)}}{\leq} \beta^t \mathbb{E}[\|\nabla f(W_0) - C_0\|_{\mathrm{F}}] + \eta \sum_{i=0}^{t-1} \beta^{t-i} \hat{J}_i + \frac{s c_{rms}^2 m n^{3/2} \eta^2 \sum_{i=0}^{t-1} \beta^{t-i}}{2} \\
&\leq \beta^{t+1} \mathbb{E}[\|\nabla f(W_0)\|_{\mathrm{F}}] + \eta \sum_{i=0}^{t-1} \beta^{t-i} \hat{J}_i + \frac{s c_{rms}^2 m n^{3/2} \eta^2 \beta}{2(1-\beta)},
\end{aligned} \tag{20}
$$

where inequality (i) above holds due to Lemma 3.7, inequality (ii) above holds due to induction.

For term (B), the derivation is same as the derivation of inequality 12 in subsection B.1, thus we have

$$
\mathbb{E}[\|C_t - N_t\|_{\mathrm{F}}] \leq \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}}. \tag{21}
$$

Thus by inequality 20, 21, inequality 19 can be written as

$$
\mathbb{E}[\|\nabla f(W_t) - N_t\|_{\mathrm{F}}] \leq \beta^{t+1} \mathbb{E}[\|\nabla f(W_0)\|_{\mathrm{F}}] + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}} + \eta \sum_{i=0}^{t-1} \beta^{t-i} \hat{J}_i + \frac{s c_{crms}^2 m n^{3/2} \beta \eta^2}{2(1-\beta)}.
$$

□

C.3 Proof of Theorem 3.8

Proof. Take the Taylor expansion of $f(W_t)$ for $t = 0, 1, \ldots, T-1$,

$$\begin{aligned}
\mathbb{E}[f(W_{t+1}) - f(W_t)] &= \mathbb{E}\Bigg[\langle \nabla f(W_t), W_{t+1} - W_t\rangle + \frac{1}{2}\mathrm{vec}(W_{t+1} - W_t)^\top \nabla^2 f_v(w_t)\mathrm{vec}(W_{t+1} - W_t) \\
&\quad + \frac{1}{6}\sum_{i,j,k=1}^{mn}\Big[\nabla^3 f_v(\tilde{\theta}_t)\Big]_{ijk}\mathrm{vec}(W_{t+1} - W_t)_i \mathrm{vec}(W_{t+1} - W_t)_j \mathrm{vec}(W_{t+1} - W_t)_k\Bigg] \\
&\overset{\text{(i)}}{\leq} \mathbb{E}\left[-\eta\langle \nabla f(W_t), \gamma_t \hat{O}_t\rangle + \frac{\eta^2}{2}J_t + \frac{s\eta^3\|\mathrm{vec}(\gamma_t\hat{O}_t)\|_2^3}{6}\right] \\
&\leq \mathbb{E}\left[-\eta\langle N_t, \gamma_t\hat{O}_t\rangle + \eta\langle N_t - \nabla f(W_t), \gamma_t\hat{O}_t\rangle + \frac{\eta^2}{2}J_t + \frac{sc_{rms}^3(mn)^{3/2}\eta^3}{6}\right] \\
&\overset{\text{(ii)}}{\leq} \mathbb{E}\left[-\eta c_0\|N_t\|_* + \eta\|N_t - \nabla f(W_t)\|_{\mathrm{F}}\|\gamma_t\hat{O}_t\|_{\mathrm{F}} + \frac{\eta^2}{2}J_t + \frac{sc_{rms}^3(mn)^{3/2}\eta^3}{6}\right] \\
&\leq \mathbb{E}\Bigg[-\eta c_0\left(\|\nabla f(W_t)\|_* - \|\nabla f(W_t) - N_t\|_*\right) + c_{rms}\eta\sqrt{mn}\|N_t - \nabla f(W_t)\|_{\mathrm{F}} \\
&\quad + \frac{\eta^2}{2}J_t + \frac{sc_{rms}^3(mn)^{3/2}\eta^3}{6}\Bigg] \\
&\leq \mathbb{E}\left[-\eta c_0\|\nabla f(W_t)\|_* + c_{rms}\eta\sqrt{mn}(\kappa+1)\|\nabla f(W_t) - N_t\|_{\mathrm{F}} + \frac{\eta^2}{2}J_t + \frac{sc_{rms}^3(mn)^{3/2}\eta^3}{6}\right] \\
&\leq \mathbb{E}\left[-\eta c_0\|\nabla f(W_t)\|_* + 2c_{rms}\eta\sqrt{mn}\|\nabla f(W_t) - N_t\|_{\mathrm{F}} + \frac{\eta^2}{2}J_t + \frac{sc_{rms}^3(mn)^{3/2}\eta^3}{6}\right],
\end{aligned} \tag{22}$$

where $J_t = \mathrm{vec}(\gamma_t\hat{O}_t)^\top H_t\,\mathrm{vec}(\gamma_t\hat{O}_t)$ for $t = 0, 1, \ldots, T-1$, $\tilde{\theta}_t \in \mathbb{R}^{mn}$ is some vector between w_t and w_{t+1}, inequality (i) above holds due to Assumption 3.5, inequality (ii) above holds due to Lemma 2.5. By Lemma 3.6 we have

$$\begin{aligned}
\mathbb{E}[f(W_{t+1}) - f(W_t)] \leq -\eta c_0\mathbb{E}[\|\nabla f(W_t)\|_*] + 2c_{rms}\eta\sqrt{mn}\Bigg(&\beta^{t+1}\mathbb{E}[\|\nabla f(W_0)\|_{\mathrm{F}}] + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}} \\
&+ \eta\sum_{i=0}^{t-1}\beta^{t-i}\hat{J}_i + \frac{sc_{crms}^2 mn^{3/2}\beta\eta^2}{2(1-\beta)}\Bigg) + \frac{\eta^2}{2}J_t + \frac{sc_{rms}^3(mn)^{3/2}\eta^3}{6}
\end{aligned} \tag{23}$$

Denote $\hat{J} = \frac{1}{T}\sum_{t=0}^{T-2}\hat{J}_t$, $J = \frac{1}{T}\sum_{t=0}^{T-1}J_t$, then sum from 0 to $T-1$ and take the average on both sides of inequality 23, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] &\leq \frac{\mathbb{E}[\nabla f(W_0) - \nabla f(W_T)]}{c_0\eta T} + \frac{2c_{rms}\sqrt{mn}}{c_0}\Bigg(\frac{\sum_{t=0}^{T-1}\beta^{t+1}\mathbb{E}[\|\nabla f(W_0)\|_{\mathrm{F}}]}{T} + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}} \\
&\quad + \frac{sc_{crms}^2 mn^{3/2}\beta\eta^2}{2(1-\beta)} + \frac{\eta\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\beta^{t-i}\mathbb{E}[\hat{J}_i]}{T}\Bigg) + \frac{\eta\mathbb{E}[J]}{2c_0} + \frac{sc_{rms}^3(mn)^{3/2}\eta^2}{6c_0} \\
&\leq \frac{1}{\kappa}\Bigg(\frac{\Delta}{c_{rms}\eta T\sqrt{m}} + \frac{2\beta\sqrt{n}\mathbb{E}[\|\nabla f(W_0)\|_{\mathrm{F}}]}{(1-\beta)T} + \frac{2\sigma\sqrt{n(1-\beta)}}{\sqrt{1+\beta}} + \frac{2\beta J^*\eta\sqrt{n}}{1-\beta} \\
&\quad + \frac{sc_{rms}^2 mn^2\beta\eta^2}{(1-\beta)T} + \frac{J^*\eta}{2c_{rms}\sqrt{m}} + \frac{sc_{rms}^2 mn^{3/2}\eta^2}{6}\Bigg),
\end{aligned}$$

where $J^* = \max\{\mathbb{E}[J], \mathbb{E}[\hat{J}]\}$, the second inequality above holds due to

$$\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\beta^{t-i}\mathbb{E}[\hat{J}_i] = \sum_{i=0}^{T-2}\mathbb{E}[\hat{J}_i]\sum_{t=i+1}^{T-1}\beta^{t-i} < \frac{\beta\mathbb{E}[\hat{J}]}{1-\beta}.$$

□

By Theorem 3.8, if we set $\eta = \sqrt{\frac{(1-\beta)\Delta}{m^{1/2}n^{1/2}J^*T}}$ in inequality 5, we have:

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*]\\ &\leq \mathcal{O}\left(\sqrt{\frac{n^{1/2}J^*\Delta}{(1-\beta)m^{1/2}T}} + \sigma\sqrt{n(1-\beta)} + \frac{\sqrt{n}\|\nabla f(W_0)\|_{\mathrm{F}}}{(1-\beta)T} + \frac{sm^{1/2}n^{3/2}\Delta}{J^*T^2} + \sqrt{\frac{J^*\Delta(1-\beta)}{m^{3/2}n^{1/2}T}} + \frac{sm^{1/2}n\Delta(1-\beta)}{J^*T}\right).\end{aligned} \tag{24}$$

Then we set $1-\beta = \left\{\sqrt{\frac{J^*\Delta}{m^{1/2}n^{1/2}\sigma^2T}}, 1\right\}$ in inequality 24, we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*]\\ &\leq \mathcal{O}\left(\sqrt[4]{\frac{n^{3/2}J^*\sigma^2\Delta}{m^{1/2}T}} + \sqrt{\frac{m^{1/2}n^{3/2}\sigma^2(\mathbb{E}[\|\nabla f(W_0)\|_{\mathrm{F}}])^2}{J^*T\Delta}} + \sqrt[4]{\frac{J^{*3}\Delta^3}{m^{7/2}n^{3/2}\sigma^2T^3}} + \sqrt{\frac{s^2m^{1/2}n^{3/2}\Delta^3}{J^*\sigma^2T^3}} + \frac{sm^{1/2}n^{3/2}\Delta}{J^*T^2}\right)\end{aligned} \tag{25}$$

From inequality 25, we can obtain an ϵ-nuclear norm stationary point of f, and the complexity of this process is $\mathcal{O}(m^{-1/2}n^{3/2}\sigma^2J^*\Delta\epsilon^{-4})$.

D Star-convex case under Frobenius norm Lipschitz smoothness

D.1 Proof of Theorem 3.12

Proof. By Lemma A.2, the estimation of $\mathbb{E}[f(W_{t+1}) - f(W_t)]$ is same as the derivation of inequality 13, which is

$$\begin{aligned}\mathbb{E}[f(W_{t+1}) - f(W_t)] &\leq \mathbb{E}\left[\langle\nabla f(W_t), W_{t+1} - W_t\rangle + \frac{L}{2}\|W_{t+1} - W_t\|_{\mathrm{F}}^2\right]\\ &\leq -\eta c_0\underbrace{\mathbb{E}[\|\nabla f(W_t)\|_*]}_{(A)} + 2c_{rms}\eta\sqrt{mn}\underbrace{\mathbb{E}[\|\nabla f(W_t) - N_t\|_{\mathrm{F}}]}_{(B)} + \frac{c_{rms}^2mnL\eta^2}{2}.\end{aligned} \tag{26}$$

For term (A), by Lemma 3.10 we have

$$\mathbb{E}[\|\nabla f(W_t)\|_*] \geq \mathbb{E}\left[\frac{[\|\nabla f(W_t)\|_*\|W_t - W^*\|_2]}{D}\right] \geq \frac{\mathbb{E}[\langle\nabla f(W_t), W_t - W^*\rangle]}{D} \geq \frac{\mathbb{E}[f(W_t) - f^*]}{D}, \tag{27}$$

where the first inequality above holds due to Assumption 3.11.

For term (B), the derivation is same as the derivation of Lemma 3.1, which is

$$\mathbb{E}[\|\nabla f(W_t) - N_t\|_{\mathrm{F}}] \leq \beta^{t+1}\sqrt{2L\Delta} + \frac{c_{crms}\beta L\eta\sqrt{mn}}{1-\beta} + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}}. \tag{28}$$

Then combining inequality 26 with inequality 27 and inequality 28, we have

$$\begin{aligned}&\mathbb{E}[f(W_{t+1}) - f(W_t)]\\ &\leq -\frac{\eta c_0}{D}\mathbb{E}[f(W_t) - f^*] + 2c_{rms}\eta\sqrt{mn}\left(\beta^{t+1}\sqrt{2L\Delta} + \frac{c_{crms}\beta L\eta\sqrt{mn}}{1-\beta} + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}}\right) + \frac{c_{rms}^2mnL\eta^2}{2},\end{aligned}$$

then replacing $t+1$ with T, we have

$$
\begin{aligned}
&\mathbb{E}[f(W_T)] - f^* \\
&\leq \left(1 - \frac{\eta c_0}{D}\right) \mathbb{E}[f(W_{T-1}) - f^*] + 2c_{rms}\eta\sqrt{mn}\left(\beta^T\sqrt{2L\Delta} + \frac{c_{crms}\beta L\eta\sqrt{mn}}{1-\beta} + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}}\right) + \frac{c_{rms}^2 mnL\eta^2}{2} \\
&\overset{\text{(i)}}{\leq} \left(1 - \frac{\eta c_0}{D}\right)^T \mathbb{E}[f(W_0) - f^*] + \sum_{t=0}^{T-1}\left(1 - \frac{\eta c_0}{D}\right)^{T-t-1}\left[2c_{rms}\eta\sqrt{mn}\left(\frac{c_{crms}\beta L\eta\sqrt{mn}}{1-\beta} + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}}\right)\right. \\
&\quad \left. + \frac{c_{rms}^2 mnL\eta^2}{2}\right] + 2c_{rms}\eta\sqrt{2mnL\Delta}\sum_{t=0}^{T-1}\beta^{t+1}\left(1 - \frac{\eta c_0}{D}\right)^{T-t-1} \\
&\overset{\text{(ii)}}{\leq} \left(1 - \frac{\eta c_0}{D}\right)^T \mathbb{E}[f(W_0) - f^*] + \frac{D}{\eta c_0}\left[2c_{rms}\eta\sqrt{mn}\left(\frac{c_{crms}\beta L\eta\sqrt{mn}}{1-\beta} + \sqrt{\frac{(1-\beta)\sigma^2}{1+\beta}}\right) + \frac{c_{rms}^2 mnL\eta^2}{2}\right] \\
&\quad + \frac{2c_{rms}\beta\eta\sqrt{2mnL\Delta}}{1-\beta} \\
&\leq \Delta\left(1 - \frac{\eta c_0}{D}\right)^T + \frac{2c_{rms}\beta\eta\sqrt{nmL}}{1-\beta}\left(\frac{D\sqrt{nL}}{\kappa} + \sqrt{2\Delta}\right) + \frac{2\sigma D\sqrt{n(1-\beta)}}{\kappa\sqrt{1+\beta}} + \frac{c_{rms}nDL\eta\sqrt{m}}{2\kappa},
\end{aligned}
$$

where inequality (i) above holds due to induction, inequality (ii) above holds due to $(1 - \eta c_0/D)^{T-t-1} \leq 1$ for $t = 0, 1, \ldots, T-1$ and

$$
\sum_{t=0}^{T-1}\left(1 - \frac{\eta c_0}{D}\right)^{T-t-1} < \sum_{t=0}^{+\infty}\left(1 - \frac{\eta c_0}{D}\right)^t = \frac{1}{1 - (1 - \frac{\eta c_0}{D})} = \frac{D}{\eta c_0}.
$$

Note that we will choose the appropriate η to make the derivation above reasonable. □

By Theorem 3.12, if we want to let $\mathbb{E}[f(W_T)] - f^* < \epsilon$ in inequality 7, we need to let the following estimation first:

$$
\frac{2c_{rms}\beta\eta\sqrt{nmL}}{1-\beta}\left(\frac{D\sqrt{nL}}{\kappa} + \sqrt{2\Delta}\right) + \frac{2\sigma D\sqrt{n(1-\beta)}}{\kappa\sqrt{1+\beta}} + \frac{c_{rms}nDL\eta\sqrt{m}}{2\kappa} < \epsilon. \tag{29}
$$

From inequality 29, we can obtain $\eta = \frac{\kappa\epsilon^3}{nm^{1/2}LD}$ which is a possible value to make inequality 29 valid. Thus, to estimate the complexity of reaching the precision $\mathbb{E}[f(W_T)] - f^* < \epsilon$, we set $\eta = \frac{\kappa\epsilon^3}{nm^{1/2}LD}$ in the first term of the right side of inequality 7 and we have

$$
\Delta\left(1 - \frac{\eta c_0}{D}\right)^T = \Delta\left(1 - \frac{c_0\kappa\epsilon^3}{nm^{1/2}LD^2}\right)^T = \Delta\left(1 - \frac{c_{rms}\kappa^2\epsilon^3}{nLD^2}\right)^T < \epsilon. \tag{30}
$$

From inequality 30 we can obtain the complexity $T = \mathcal{O}(nLD^2\kappa^{-2}\epsilon^{-3}\log(\Delta\epsilon^{-1}))$.