CONVERGENCE OF PRACTICAL MUON

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon is emerging as a promising alternative to AdamW for large-scale neural network training, yet theoretical understanding of its practical implementation remains incomplete, as existing analyses often simplify or omit two key components: (i) practical Newton–Schulz iterations with empirically tuned polynomial coefficients $(3.4445, -4.7750, 2.0315)$; and (ii) decoupled weight decay for regularization. In this paper, we provide an optimization interpretation and establish convergence for practical Muon, jointly accounting for both components. Specifically, we interpret practical Muon as right-preconditioned optimization of the original loss with a dynamic weighted ℓ_2 regularizer that vanishes as stationarity is approached, so that the optimization target remains the original objective. We then establish, to our best knowledge, the first convergence guarantee for practical Muon in the stochastic nonconvex setting, with an $\mathcal{O}(T^{-1/4})$ convergence rate in terms of the expected Frobenius norm of the gradient, improving the dimension dependence of the best known AdamW's convergence rate by a factor of $\sqrt{d}$, where T is the iteration horizon and d is the parameter dimension. Experiments further support the theoretical convergence results.

1 INTRODUCTION

By exploiting the matrix structure of model parameters, Muon (Jordan et al., 2024) has been shown the potential to become a dominant optimizer for large-scale neural network training, displacing adaptive methods such as AdaGrad (Duchi et al., 2011), RMSProp (Tieleman, 2012), Adam (Kingma & Ba, 2015), and AdamW (Loshchilov & Hutter, 2019), which have dominated the field over the past decade, as well as classical SGD with momentum (Robbins & Monro, 1951; Sutskever et al., 2013). Remarkable empirical success has been reported for Muon and its variants in language modeling (Liu et al., 2025; Shah et al., 2025; Team et al., 2025; Xu et al., 2026), multimodal and agentic systems (Team et al., 2026; Zeng et al., 2026; Hong et al., 2026), and embodied intelligence (Fan et al., 2026). In particular, Liu et al. (2025) report roughly twice AdamW's computational efficiency on 399M–1.5B dense models, while DeepSeek-V4-Pro demonstrates Muon's scalability to 1.6 trillion parameters and 33 trillion training tokens (Xu et al., 2026).

Despite the empirical success, convergence remains an open problem for practical Muon, which combines two key ingredients: (i) practical Newton–Schulz (NS) iterations that use matrix polynomials with empirically tuned coefficients $(3.4445, -4.7750, 2.0315)$ to produce approximately orthogonalized momentum matrices as update directions; and (ii) decoupled weight decay (WD) to provide regularization and support stable training at scale (Liu et al., 2025). The convergence analysis faces two main challenges. First, practical NS is difficult to analyze directly, and existing studies often replace it with exact SVD-based orthogonalization (Shen et al., 2026), or standard NS using different polynomial coefficients (Kim & Oh, 2026; Choudhury et al., 2026) which can yield poorer orthogonalization of ill-conditioned momentum matrices than practical NS (Jordan et al., 2024). To our best knowledge, neither alternative has been publicly validated in large-scale commercial model training (Liu et al., 2025; Team et al., 2025; Xu et al., 2026; Team et al., 2026; Zeng et al., 2026; Hong et al., 2026). As for practical NS without WD, Do et al. (2026) assume strong convexity and establish convergence to a neighborhood of the minimizer, while Qian & Richtárik (2026) establish an $\mathcal{O}(\sqrt{r}/T^{1/4})$ convergence rate for nuclear norm stationarity in the stochastic nonconvex setting, where r is the smaller dimension of the matrix-valued parameter and T is the iteration horizon.

Second, WD further complicates the convergence analysis through its interaction with NS. Existing studies that account for WD typically replace NS with exact SVD-based orthogonalization, and show

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.

Table 1: Comparison of existing convergence results for Muon. The rate column uses $\mathbb{E}\|\nabla f\|_{\mathrm{F}}$ unless a KKT stationarity measure or expected objective gap (obj.) is indicated. †: The nuclear norm rates established in the original papers are converted to Frobenius norm rates under our assumptions. Std. and Pract. denote standard and practical NS, respectively.

Reference	WD	NS	Nonconvex	Convergence rate
Shen et al. (2026)	×	Exact	✓	$\mathcal{O}(\sqrt{r}/T^{1/4})$
Kim & Oh (2026)	×	Standard	✓	$\mathcal{O}(\sqrt{r}/T^{1/4})^{\dagger}$
Choudhury et al. (2026)	×	Standard	✓	$\mathcal{O}(r/T^{1/4})$
Do et al. (2026)	×	Practical	Strongly convex	neighborhood (obj.)
Qian & Richtárik (2026)	×	Std./Pract.	✓	$\mathcal{O}(r/T^{1/4})^{\dagger}$
Chen et al. (2026)	✓	Exact	✓	neighborhood (KKT)
Pethick et al. (2025)	✓	Exact	✓	$\mathcal{O}(r/T^{1/4})$ (KKT)
Qian & Richtárik (2026)	✓	Std./Pract.	Star-convex	$\mathcal{O}(\sqrt{r}(\log T/T)^{1/3})$ (obj.)
Ours	✓/×	Std./Pract.	✓	$\boldsymbol{\mathcal{O}(1/T^{1/4})}$

that Muon is equivalent to solving the following constrained optimization problem (Pethick et al., 2025; Chen et al., 2026):

$$\min_{W\in\mathbb{R}^{m\times n}} f(W) \quad \text{s.t.} \quad \|W\|_{\mathrm{op}} \le 1/\lambda,$$

where f denotes the original loss function, W is the matrix-valued parameter, $\lambda > 0$ is the WD coefficient, and $\|\cdot\|_{\mathrm{op}}$ denotes the spectral norm. For stochastic nonconvex objectives, Pethick et al. (2025) establish an $\mathcal{O}(r/T^{1/4})$ convergence rate in terms of a KKT stationarity measure for the above constrained problem. When both NS and WD are considered, Qian & Richtárik (2026) provide the only convergence guarantee that we are aware of. Under star-convexity and with parameter choices that depend on the norm of an unknown global minimizer, they prove convergence to an optimal solution of the original problem.

These observations naturally raise two questions about Muon with practical NS and WD in the stochastic nonconvex setting:

Q1) What does Muon actually optimize when practical NS and WD are combined?
Q2) Does it converge? If so, at what rate, and can the convergence rate be sharper than AdamW's?

Controbutions. We address these questions and make the following contributions:

- We interpret Muon with NS and WD as right-preconditioned optimization of f with a dynamic weighted ℓ_2 regularizer that vanishes as $\|\nabla f\|_F \to 0$, indictating that Muon still targets the original objective. While this interpretation is exact when momentum in Muon and noise in stochastic gradient are omitted, our convergence guarantees remain valid in their presence. Thus we address Q1.
- To the best of our knowledge, we establish the first convergence guarantee for Muon with practical NS and WD in the stochastic nonconvex setting, as summarized in Table 1, achieving
$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\,\|\nabla f(W_t)\|_F = \mathcal{O}\Big(\frac{\sqrt[4]{L(f(W_0)-f_{\inf})\sigma^2/b_{\mathrm{mb}}}}{T^{1/4}}\Big).$$
Notably, the rate above has sharper dimension dependence than the best known AdamW's rate (Li et al., 2026) by a factor of $\sqrt{d}=\sqrt{mn}$ in the Frobenius norm, with AdamW satisfying
$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_1 = \mathcal{O}\Big(\frac{\sqrt{d}\sqrt[4]{L(f(W_0)-f_{\inf})\sigma^2/b_{\mathrm{mb}}}}{T^{1/4}}\Big), \tag{1}$$
where $\|A\|_1 = \sum_{i,j}|A_{ij}|$ denotes the entrywise ℓ_1 norm, and the Frobenius norm comparison follows from $\|A\|_F \le \|A\|_1$. Our rate also matches that of SGD in the Frobenius norm (Ghadimi & Lan, 2013) and that of AdamW in the entrywise ℓ_1 norm. Thus we address Q2.

Outline. The rest of the paper is organized as follows. Section 2 presents the preliminaries. Then Sections 3–5 provide an optimization interpretation of practical Muon, its convergence results, and the corresponding proofs, respectively. Section 6 illustrates practical NS's superiority over standard NS. Section 7 empirically supports Muon's theoretical advantage. Finally, Section 8 concludes the paper, with additional related work in Appendix A and further details in the remaining appendices.

2 PRELIMINARIES

We consider the following stochastic nonconvex matrix optimization problem:

$$\min_{W\in\mathbb{R}^{m\times n}} f(W) = \mathbb{E}_{\xi\sim\mathcal{D}}[F(W,\xi)], \tag{2}$$

where $\xi \in \Xi$ is a data sample drawn from the distribution $\mathcal{D}$, and $F : \mathbb{R}^{m\times n} \times \Xi \to \mathbb{R}$ is a differentiable sample loss with stochastic gradient $\nabla_W F(W,\xi)$.

2.1 PRACTICAL MUON ALGORITHM

To solve (2), we consider Muon (Algorithm 1), which combines Nesterov-style momentum, Newton–Schulz (NS) iterations, and decoupled weight decay (WD). The NS iterations approximately orthogonalize the momentum matrices through matrix polynomials, avoiding an explicit SVD. We consider two choices of the polynomial p in Definition 1, corresponding to practical and standard NS, and refer to Muon with practical NS as practical Muon.

Algorithm 1 Muon (with Newton–Schulz iterations and decoupled weight decay)

1: **Input:** Iteration horizon $T \in \mathbb{N}$, mini-batch size $b_{\mathrm{mb}} \in \mathbb{N}$, momentum coefficient $\beta \in [0,1)$, normalization floor $\varepsilon > 0$, NS polynomial p, stepsize $\eta > 0$, and WD coefficient $\lambda \geq 0$
2: **Initialization:** parameter $W_0 \in \mathbb{R}^{m\times n}$, and momentum buffer $B_{-1} = 0$
3: **for** $t = 0, \ldots, T-1$ **do**
4: Independently draw $\xi_{t,1}, \ldots, \xi_{t,b_{\mathrm{mb}}} \sim \mathcal{D}$, and compute $G_t \leftarrow b_{\mathrm{mb}}^{-1} \sum_{i=1}^{b_{\mathrm{mb}}} \nabla_W F(W_t, \xi_{t,i})$
5: $B_t \leftarrow \beta B_{t-1} + (1-\beta) G_t$
6: $M_t \leftarrow \beta B_t + (1-\beta) G_t$ ▷ Nesterov-style momentum
7: $\tau_t \leftarrow \max\{\|M_t\|_F, \varepsilon\}$[1], $\quad X_{t,0} \leftarrow M_t/\tau_t$
8: **for** $j = 0, \ldots, 4$ **do**
9: $X_{t,j+1} \leftarrow p(X_{t,j}X_{t,j}^\top)X_{t,j}$ ▷ Newton–Schulz iterations (Line 7–11)
10: **end for**
11: $D_t \leftarrow X_{t,5}$,
12: $W_{t+1} \leftarrow (1-\eta\lambda)W_t - \eta D_t = W_t - \eta H_t$ ▷ Decoupled weight decay
13: **end for**

Definition 1 (Practical and standard Newton–Schulz polynomials). The practical NS polynomial is

$$p_{\mathrm{P}}(x) = a_{\mathrm{P}} + b_{\mathrm{P}}x + c_{\mathrm{P}}x^2, \qquad (a_{\mathrm{P}}, b_{\mathrm{P}}, c_{\mathrm{P}}) = (3.4445, -4.7750, 2.0315).$$

For $k \in \mathbb{N}$, the degree-k standard NS polynomial is the Taylor polynomial of $x^{-1/2}$ at $x = 1$:

$$p_{\mathrm{S}}^{(k)}(x) = \sum\nolimits_{\ell=0}^{k} \frac{(2\ell)!}{4^\ell(\ell!)^2}(1-x)^\ell.$$

In particular, $p_{\mathrm{S}}^{(2)}(x) = a_{\mathrm{S}} + b_{\mathrm{S}}x + c_{\mathrm{S}}x^2$ with $(a_{\mathrm{S}}, b_{\mathrm{S}}, c_{\mathrm{S}}) = (15/8, -5/4, 3/8)$.

In Algorithm 1, the NS polynomial is chosen as either p_{P} or $p_{\mathrm{S}}^{(k)}$. For notational simplicity, we denote either choice by p, write a, b, c for its coefficients of $1, x, x^2$, respectively. Moreover, define $q(s) = sp(s^2)$, $q_0(s) = s$, $q_{j+1}(s) = q(q_j(s))$, $j = 0, \ldots, 4$, and write $(q_{\mathrm{P}}, q_{\mathrm{P},j})$ and $(q_{\mathrm{S}}^{(k)}, q_{\mathrm{S},j}^{(k)})$ for (q, q_j) under practical and standard NS, respectively. The following lemma characterizes how these polynomials are used in NS iterations to approximate matrix orthogonalization.

Lemma 1 (Approximate orthogonalization via NS; Kim & Oh, 2026, Proposition 1 and Lemma 1). ***(i) NS output.*** *In Algorithm 1, write the compact SVD $M_t = U_t \operatorname{diag}(\sigma_{1,t}, \ldots, \sigma_{r_t,t})V_t^\top$, with $r_t = \operatorname{rank}(M_t)$, and $s_{i,t} = \sigma_{i,t}/\tau_t \in (0,1]$. The five NS steps preserve the left and right singular-vector matrices U_t and V_t, and yield*

$$D_t = X_{t,5} = U_t \operatorname{diag}\big(q_5(s_{1,t}), \ldots, q_5(s_{r_t,t})\big)V_t^\top.$$

[1] Another common normalization uses $\|M_t\|_F + \varepsilon$. The two denominators differ by $\min\{\|M_t\|_F, \varepsilon\} \leq \varepsilon$. We adopt the maximum form, consistent with the PyTorch implementation.

(ii) Approximation error. *The singular-value map q_{P} of one practical NS step is nonmonotone on $[0,1]$ and does not fix 1. Specifically,*

$$q'_{\mathrm{P}}(0) = 3.4445 > 0, \qquad q'_{\mathrm{P}}(1) = -0.7230 < 0, \qquad q_{\mathrm{P}}(1) = 0.7010 \neq 1.$$

In contrast, for every $k \geq 1$, the singular-value map $q_{\mathrm{S}}^{(k)}$ of one standard NS step is nondecreasing on $[0,1]$ and fixes 1. These properties, together with the Taylor construction, yield

$$\|X_{t,j} - U_t V_t^\top\|_{\mathrm{op}} \leq (1 - \min_{1\leq i\leq r_t} s_{i,t}^2)^{(k+1)^j}, \qquad j = 0, \ldots, 5. \tag{3}$$

Thus, each standard NS step approximates exact orthogonalization with convergence order $k+1$.

Challenges in analyzing practical Muon. The role of NS in Muon is to approximate the orthogonalized momentum matrix $U_t V_t^\top$ without explicitly computing an SVD. This amounts to preserving the singular vectors while transforming each positive singular value toward 1. The normalization in Algorithm 1 places the initial singular values in $(0, 1]$. For standard NS, monotonicity and the fixed point at 1 yield the superlinear error bound in (3). Practical NS lacks both properties, making its approximation error hard to control and Muon's convergence difficult to analyze.

2.2 Assumptions

Our analysis is based on the following assumptions, which are standard in stochastic nonconvex optimization (e.g., Ghadimi & Lan, 2013; Li et al., 2026; Shen et al., 2026; Choudhury et al., 2026).

Assumption 1 (Low boundedness). The objective satisfies $f_{\inf} = \inf_{W\in\mathbb{R}^{m\times n}} f(W) > -\infty$.

Assumption 2 (Smoothness). There exists a constant $L > 0$ such that, for all $X, Y \in \mathbb{R}^{m\times n}$,

$$\|\nabla f(X) - \nabla f(Y)\|_F \leq L\|X - Y\|_F.$$

Assumption 3 (Stochastic gradient oracle). There exists a constant $\sigma \geq 0$ such that, for all $W\in\mathbb{R}^{m\times n}$,

$$\mathbb{E}_\xi[\nabla F_W(W,\xi)] = \nabla f(W), \qquad \mathbb{E}_\xi\big[\|\nabla F_W(W,\xi) - \nabla f(W)\|_F^2\big] \leq \sigma^2.$$

For any mini-batch of size b_{mb}, the averaged gradient has variance at most σ^2/b_{mb}.

3 Optimization interpretation

Main viewpoint. We show that Muon with NS and WD, using either practical or standard NS, admits a unified interpretation as right-preconditioned optimization on a dynamic objective consisting of f and a weighted ℓ_2 regularizer, with exact equivalence when the momentum is omitted.

Dynamic objective. We seek an objective for which the Muon update provides a descent direction. A natural idea is to examine the fixed point of Algorithm 1 (with $m = n = 1$ for illustration , allowing us to write $D_t = q_5(M_t/\tau_t)$ without an SVD representation), where $M_t = B_t = G_t$ and

$$D_t + \lambda W_t = q_5(G_t/\tau_t) + \lambda W_t = 0.$$

Inverting q_5 would allow us to isolate the gradient and identify a corresponding objective. For practical NS, however, both q_{P} and $q_{\mathrm{P},5}$ are nonmonotone, so the latter admits no global inverse, obstructing this direct approach. Motivated by the preconditioning interpretation of Muon with exact orthogonalization (Ma et al., 2026), beyond fixed points, we seek a right preconditioner P_t such that

$$D_t = M_t P_t, \;\; \text{i.e., } \; U_t \operatorname{diag}\big(q_5(\sigma_{1,t}/\tau_t), \ldots, q_5(\sigma_{r_t,t}/\tau_t)\big)V_t^\top = U_t \operatorname{diag}\big(\sigma_{1,t}, \ldots, \sigma_{r_t,t}\big)V_t^\top P_t.$$

This determines P_t on the span of V_t, while the WD term λW_t suggests a P_t^{-1} weighted ℓ_2 regularizer, which requires P_t to be invertible. Thus, we extend P_t positively to the nullspace of M_t and define

$$P_t = V_t \operatorname{diag}\Big(\frac{\chi_5(s_{1,t})}{\tau_t}, \ldots, \frac{\chi_5(s_{r_t,t})}{\tau_t}\Big)V_t^\top + \frac{\chi_5(0)}{\tau_t}(I_n - V_t V_t^\top), \tag{4}$$

where $s_{i,t} \in (0,1]$, and $\chi_5(s) = q_5(s)/s > 0$ for $s \in (0,1]$, with the continuous extension $\chi_5(0) = a^5$. The resulting dynamic objective is

$$\Phi_t(W) = f(W) + \mathcal{R}_t(W), \qquad \mathcal{R}_t(W) = \frac{\lambda}{2}\operatorname{tr}(W P_t^{-1} W^\top). \tag{5}$$

Here, P_t is treated as a fixed matrix when differentiating Φ_t in W. The following proposition establishes the exact equivalence between the Muon update and right-preconditioned optimization of the dynamic objective Φ_t.

Proposition 1. ***(i) Properties of the right preconditioner and regularizer.*** *For both practical and standard NS, the preconditioner P_t in* (4) *satisfies*

$$P_t \succ 0, \qquad D_t = M_t P_t, \qquad \langle M_t, D_t \rangle_F = \|M_t P_t^{1/2}\|_F^2 \geq 0.$$

Thus, D_t is positively aligned with the momentum direction M_t whenever $M_t \neq 0$. Moreover, for $\lambda > 0$, the regularizer $\mathcal{R}_t$ in (5) *is $\lambda\tau_t/C_1$-strongly convex, where $C_1 = \max_{s\in[0,1]} \chi_5(s) > 0$.*

(ii) Exact right-preconditioned equivalence. *Ignore the momentum by setting $\beta = 0$, Muon update in Algorithm 1 is exactly one right-preconditioned stochastic gradient descent step on the dynamic objective* (5)*, i.e.,*

$$W_{t+1} = W_t - \eta \widehat{\nabla}\Phi_t(W_t) P_t, \qquad \widehat{\nabla}\Phi_t(W_t) = G_t + \lambda W_t P_t^{-1}.$$

If Assumption 3 futher holds, we have $\mathbb{E}[\widehat{\nabla}\Phi_t(W_t) - \nabla\Phi_t(W_t)] = 0$.

Proof. Recall that $q_0(s) = s$ and $q_{j+1}(s) = q(q_j(s)) = q_j(s)p(q_j(s)^2)$ for $j = 0, \ldots, 4$. For practical NS, $p_{\mathrm{P}}(x) > 0$ for all $x \in \mathbb{R}$, since $c_{\mathrm{P}} > 0$ and $b_{\mathrm{P}}^2 - 4a_{\mathrm{P}}c_{\mathrm{P}} < 0$. For standard NS, whenever $s \in [0, 1]$, Lemma 1 gives $q_{\mathrm{S}}^{(k)}(s) \in [0, 1]$, $p_{\mathrm{S}}^{(k)}(s) = 1 + \sum_{\ell=1}^{k} \frac{(2\ell)!}{4^\ell(\ell!)^2}(1-s)^\ell > 0$, and hence $p(q_{\mathrm{S}}^{(k)}(s)) > 0$. As a result, for either NS choice, applying the recurrence over all five NS steps gives

$$\chi_5(s) = \prod\nolimits_{j=0}^{4} p(q_j(s)^2) > 0, \qquad s \in (0, 1].$$

The eigenvalues of P_t are either $\chi_5(s_{i,t})/\tau_t$ with $s_{i,t} \in (0, 1]$, or $\chi_5(0)/\tau_t$, giving

$$\frac{C_2}{\tau_t} I_n \preceq P_t \preceq \frac{C_1}{\tau_t} I_n, \qquad C_2 = \min_{s\in[0,1]} \chi_5(s) > 0. \tag{6}$$

Thus $P_t \succ 0$, and the strong convexity of $\mathcal{R}_t$ follows from $P_t^{-1} \succeq (\tau_t/C_1)I_n$. By Lemma 1 and $V_t^\top(I_n - V_t V_t^\top) = 0$, we have

$$D_t = U_t \operatorname{diag}(q_5(s_{1,t}), \ldots, q_5(s_{r_t,t}))\, V_t^\top = M_t P_t,$$

which implies $\langle M_t, D_t \rangle_F = \|M_t P_t^{1/2}\|_F^2 \geq 0$. When $\beta = 0$, we have $M_t = G_t$, and

$$\widehat{\nabla}\Phi_t(W_t) P_t = G_t P_t + \lambda W_t = D_t + \lambda W_t = H_t,$$

proving the update identities. Holding P_t fixed when differentiating gives $\nabla\mathcal{R}_t(W) = \lambda W P_t^{-1}$, and Assumption 3 directly implies $\mathbb{E}\left[\widehat{\nabla}\Phi_t(W_t) - \nabla\Phi_t(W_t)\right] = 0$. □

Recovery of the original problem. Consider the deterministic setting where $G_t = \nabla f(W_t)$. By the definition of $\mathcal{R}_t$ in (5) and the bounds (6), we have

$$0 \leq \mathcal{R}_t(W) \leq \frac{\lambda\tau_t}{2C_2} \|W\|_F^2, \qquad \|\nabla\mathcal{R}_t(W)\|_F \leq \frac{\lambda\tau_t}{C_2} \|W\|_F. \tag{7}$$

If $\|\nabla f(W_t)\|_F \to 0$, then $\|M_t\|_F \to 0$ and hence $\tau_t = \max\{\|M_t\|_F, \varepsilon\} \to \varepsilon$, where ε is chosen to be very small in practice, e.g., 10^{-7} in the official PyTorch implementation. Together with the boundedness of $\{W_t\}$, which WD naturally promotes, inequalities (7) show that both $\mathcal{R}_t(W_t)$ and $\|\nabla\mathcal{R}_t(W_t)\|_F$ are $\mathcal{O}(10^{-7})$, and the dynamic objective reduces to the original objective up to numerical precision. Independently of this interpretation, our main theorem in the next section allows for stochastic setting and directly establishes stationarity for the original stochastic nonconvex problem.

4 Main results

For Muon with practical or standard NS and WD, we establish the following convergence guarantees in the stochastic nonconvex setting.

Theorem 1 (NS with WD). *Under Assumptions 1–3, consider the iterates of Algorithm 1 with*

$$\eta = \frac{c_1}{T^{3/4}},\ 1 - \beta = \frac{c_2}{T^{1/2}},\ \varepsilon \in \left(0, \frac{c_3}{T^{1/4}}\right],\ \lambda = \frac{C_2}{2\left(\|W_0\|_F + C_3 c_1 T^{1/4}\right)},\ T \geq \max\{1, c_2^2\}, \tag{8}$$

where c_1, c_2, c_3 are $\mathcal{O}(1)$ constants defined in (20)*, and C_3 is an $\mathcal{O}(1)$ constant defined in* (11)*. Then*

$$\frac{1}{T}\sum\nolimits_{t=0}^{T-1} \mathbb{E}\,\|\nabla f(W_t)\|_F = \mathcal{O}\Big(\frac{\sqrt[4]{L(f(W_0) - f_{\inf})\sigma^2/b_{\mathrm{mb}}}}{T^{1/4}}\Big). \tag{9}$$

Discussion: practical NS with WD.

- ***Convergence guarantee.*** To the best of our knowledge, Theorem 1 provides the first convergence guarantee for practical Muon in the stochastic nonconvex setting, based on a direct analysis of the practical NS iteration and its interaction with WD. While Qian & Richtárik (2026) also cover practical NS with WD, their guarantees rely on star-convexity and parameter choices that depend on the unknown norm $\|W^*\|$ of a global minimizer W^*. Other analyses either use exact orthogonalization when accounting for WD (Pethick et al., 2025; Chen et al., 2026), or omit WD while studying practical NS (Do et al., 2026; Qian & Richtárik, 2026), standard NS (Kim & Oh, 2026; Choudhury et al., 2026), or exact orthogonalization (Shen et al., 2026).
- ***Comparison with AdamW and SGD.*** The theoretical convergence rate in (9) improves dimension dependence by a factor of $\sqrt{d}$ over the Frobenius norm guarantee implied by the best-known AdamW result given by (1) (Li et al., 2026) in the stochastic nonconvex setting. Specifically, AdamW's entrywise ℓ_1 guarantee contains an additional $\sqrt{d}$ factor that persists after conversion to the Frobenius norm. Moreover, our rate matches that of SGD in the Frobenius norm (Ghadimi & Lan, 2013), and that of AdamW in the entrywise ℓ_1 norm.
- ***Parameter choices.*** The choice $\lambda = \mathcal{O}(T^{-1/4})$ is reasonable for practical training horizons. For example, Moonlight pretraining with Muon takes approximately 1.76×10^5 updates (Liu et al., 2025), giving $T^{-1/4} \approx 0.05$. This mild dependence on T permits a non-negligible WD coefficient that retains its regularization role. The theoretical condition on ε is also mild, since commonly used values, such as the PyTorch default 10^{-7}, are far below the corresponding scale $T^{-1/4}$. Moreover, the condition on T imposes only an $\mathcal{O}(1)$ lower threshold. Small stepsize η and momentum coefficient β close to one are common in stochastic nonconvex convergence analyses (Ghadimi & Lan, 2013; Li et al., 2026). Finally, the problem dependent choices of c_1, c_2, c_3 only improve the multiplicative factor in the convergence rate, and any positive values retain the $\mathcal{O}(T^{-1/4})$ rate.

Discussion: standard NS with WD. The comparisons with AdamW and SGD and the discussion of parameter choices above also apply to standard NS with WD.

- ***Convergence guarantee.*** To the best of our knowledge, Theorem 1 provides the first convergence guarantee for Muon with standard NS and WD in the stochastic nonconvex setting. In contrast, the existing analysis requires both star-convexity and parameter choices that depend on the norm $\|W^*\|$ of an unknown global minimizer (Qian & Richtárik, 2026), while the proof of Proposition 3.3 in Sato et al. (2025) contains an error identified by Apte (2026). The counterexample in Apte's Proposition 3.2 also invalidates Sato et al.'s Proposition 3.3 when standard NS is used.
- ***Convergence rate.*** By analyzing practical and standard NS in a unified framework, we obtain the rate in (9) independent of the orthogonalization error of standard NS, which can lead to a very large multiplicative factor in existing convergence rates (Kim & Oh, 2026; Choudhury et al., 2026). For example, Kim & Oh (2026) obtain the factor $(1-\epsilon_5)^{-1}$ for standard NS without WD, where

$$\epsilon_{t,5} = \|X_{t,5} - U_t V_t^\top\|_{\mathrm{op}} = 1 - q_5\big(\min_{1\le i\le r_t} s_{i,t}\big), \qquad \epsilon_5 = \max_{0\le t<T} \epsilon_{t,5}.$$

  Despite the superlinear decay in (3), this factor can be large for small normalized singular values. For a quantitative illustration, Wu et al. (2026, Figure 1(c)) show a Muon momentum spectrum in a LLaMA attention layer spanning roughly eight or more orders of magnitude. For such an NS input, five degree-2 standard NS steps would give $(1-\epsilon_5)^{-1} > 10^6$.

The following corollary gives the corresponding convergence guarantee without WD.

Corollary 1 (NS without WD). *Under Assumptions 1–3, consider the iterates of Algorithm 1 with*

$$\eta = \frac{c_1}{T^{3/4}}, \quad 1-\beta = \frac{c_2}{T^{1/2}}, \quad \varepsilon \in \Big(0, \frac{c_3}{T^{1/4}}\Big], \quad \lambda = 0, \quad T \ge \max\{1, c_2^2\},$$

where c_1, c_2, c_3 are defined in (20) *with $(C_2/2 + C_3)$ replaced by C_3. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\,\|\nabla f(W_t)\|_F = \mathcal{O}\Big(\frac{\sqrt[4]{L(f(W_0) - f_{\inf})\sigma^2/b_{\mathrm{mb}}}}{T^{1/4}}\Big). \tag{10}$$

Discussion: practical and standard NS without WD. Corollary 1 yields a convergence rate similar to that with WD. The Muon rate (10) improves dimension dependence by a factor of $\sqrt{d}$ over the AdamW Frobenius norm rate implied by the best-known result (1), and does not depend on the worst-case orthogonalization error. Regarding the choice of NS polynomial, the corollary does not reveal a clear difference in convergence behavior between practical and standard NS.

5 Convergence analysis

We first state the key insight for our analysis: quantifying the alignment of NS directions, particularly those produced by practical NS, with the momentum direction, which supports their use as effective descent directions. We then prove convergence, with a few routine details deferred to Appendix B.

5.1 Key insight

The favorable alignment of practical and standard NS directions with the momentum is key to Muon's convergence, and is quantified in the following lemma.

Lemma 2 (NS direction alignment). *For both practical and standard NS, the momentum M_t and the resulting direction D_t in Algorithm 1 satisfy*

$$\langle M_t, D_t\rangle_F \geq C_2 \|M_t\|_F - C_2\varepsilon/4, \qquad \|D_t\|_F \leq C_3,$$

where $C_2 = \min_{s\in[0,1]} \chi_5(s) > 0$, $C_1 = \max_{s\in[0,1]} \chi_5(s) > 0$, recalled here for convenience, and

$$C_3 = \min\{C_1, q_{\max}\sqrt{r}\}, \qquad q_{\max} = \max_{s\in[0,1]} q_5(s), \qquad r = \min\{m, n\}. \tag{11}$$

In particular,

$$(C_2, C_1, q_{\max}) = \begin{cases} (0.6964\ldots,\ 484.8762\ldots,\ 1.2023\ldots), & \textit{practical NS}, \\ (1,\ a^5,\ 1), & \textit{standard NS}. \end{cases} \tag{12}$$

Proof. The proof of Proposition 1 ensures that χ_5, C_1, C_2, and $q_{\max}$ are well defined. By definition,

$$C_2 s \leq q_5(s) \leq C_1 s, \qquad 0 \leq q_5(s) \leq q_{\max}, \qquad s \in [0,1]. \tag{13}$$

We then prove the bounds in Lemma 2, with sharp constants C_1, C_2, and $q_{\max}$ determined in Appendix B. By Lemma 1, $D_t = U\,\mathrm{diag}(q_5(s_{i,t}), \ldots, q_5(s_{r_t,t}))V^\top$. Then (13) yields

$$\langle M_t, D_t\rangle_F = \tau_t \textstyle\sum_i s_{i,t} q_5(s_{i,t}) \geq C_2\tau_t \sum_i s_{i,t}^2 = C_2 \|M_t\|_F^2 / \tau_t,$$

where $\|M_t\|_F^2/\tau_t \geq \|M_t\|_F - \varepsilon/4$, since for $\|M_t\|_F \leq \varepsilon$ the inequality reduces to $(\|M_t\|_F - \varepsilon/2)^2/\varepsilon \geq 0$, whereas for $\|M_t\|_F \geq \varepsilon$, the left-hand side equals $\|M_t\|_F$. Moreover, (13) together with $\sum_i s_{i,t}^2 = \|M_t\|_F^2/\tau_t^2 \leq 1$ gives

$$\|D_t\|_F = \sqrt{\textstyle\sum_i q_5(s_{i,t})^2} \leq \min\left\{C_1\sqrt{\textstyle\sum_i s_{i,t}^2}, q_{\max}\sqrt{r}\right\} \leq C_3. \qquad \square$$

Discussion. Lemma 2 establishes favorable alignment of both practical and standard NS directions with the momentum, up to a small $O(\varepsilon)$ term, with $\varepsilon = 10^{-7}$ by default in PyTorch. This estimate is independent of the orthogonalization error, which is particularly useful for practical NS since its singular-value map is nonmonotone and does not fix 1, precluding asymptotically exact orthogonalization, let alone superlinear convergence of standard NS. Moreover, the lemma establishes a uniform bound on the Frobenius norm of the NS direction, $\|D_t\|_F \leq C_3 \leq C_1 = \mathcal{O}(1)$, preventing unbounded growth with r and limiting the rank dependence of the resulting convergence rate. Together, alignment and boundedness form the basis of our descent analysis of the original objective.

5.2 Convergence proof

We now prove Theorem 1, deferring a few routine details of the momentum analysis to Appendix B. Corollary 1 follows from the same argument by omitting the WD term, i.e., setting $\lambda = 0$.

Proof. The update $W_{t+1} = W_t - \eta H_t = W_t - \eta(D_t + \lambda W_t)$ and Assumption 2 gives

$$f(W_{t+1}) \leq f(W_t) - \eta\langle \nabla f(W_t), H_t\rangle_F + L\eta^2/2\,\|H_t\|_F^2\,. \tag{14}$$

Let $e_t = M_t - \nabla f(W_t)$. The linear term decomposes as

$$\langle \nabla f(W_t), H_t\rangle_F = \langle M_t, D_t\rangle_F - \langle e_t, D_t\rangle_F + \lambda\langle \nabla f(W_t), W_t\rangle_F. \tag{15}$$

Step 1. Alignment of the NS directions with momentum. We use Lemma 2 to control the alignment $\langle M_t, D_t\rangle_F$ and the norm $\|D_t\|_F$.

Step 2. Bounds on λW_t and H_t From (8), (11), and (12), $0 < \eta\lambda \le C_2/(2C_3T) < 1$. Lemma 2 and the update in Algorithm 1 therefore imply

$$\|W_{t+1}\|_F \le (1-\eta\lambda)\|W_t\|_F + \eta C_3 \le \|W_t\|_F + \eta C_3.$$

For $0 \le t \le T$, (8) gives

$$\lambda\|W_t\|_F \le \lambda\big(\|W_0\|_F + C_3\eta t\big) \le C_2/2, \qquad \|H_t\|_F \le C_4 = C_2/2 + C_3. \tag{16}$$

We can now bound the linear term (15). Lemma 2, (16), and $\|M_t\|_F \ge \|\nabla f(W_t)\|_F - \|e_t\|_F$ yield

$$\begin{aligned}\langle \nabla f(W_t), H_t\rangle_F &\ge C_2\|M_t\|_F - C_2\varepsilon/4 - C_3\|e_t\|_F - \lambda\|W_t\|_F\|\nabla f(W_t)\|_F \\ &\ge \big(C_2 - \lambda\|W_t\|_F\big)\|\nabla f(W_t)\|_F - (C_2+C_3)\|e_t\|_F - C_2\varepsilon/4 \\ &\ge \frac{C_2}{2}\|\nabla f(W_t)\|_F - (C_2+C_3)\|e_t\|_F - \frac{C_2\varepsilon}{4}. \end{aligned}\tag{17}$$

Substituting (17) and (16) into (14) gives

$$f(W_{t+1}) \le f(W_t) - \frac{C_2\eta}{2}\|\nabla f(W_t)\|_F + \eta(C_2+C_3)\|e_t\|_F + \frac{C_2\eta\varepsilon}{4} + \frac{L\eta^2C_4^2}{2}. \tag{18}$$

Step 3. Tracking error of the Nesterov-style momentum.

Lemma 3 (Momentum tracking). *Under the assumptions and parameter choices of Theorem 1,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|M_t - \nabla f(W_t)\|_F \le \frac{\|B_{-1}-\nabla f(W_0)\|_F}{c_2}T^{-1/2} + \Big(\frac{Lc_1C_4}{c_2} + \sqrt{2c_2}\,\frac{\sigma}{\sqrt{b_{\mathrm{mb}}}}\Big)T^{-1/4}.$$

Summing (18), taking expectations, and using Lemma 3, (8), and $f(W_t) \ge f_{\inf}$, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_F \le{}& \Big[\frac{2(f(W_0)-f_{\inf})}{C_2c_1} + \frac{2(C_2+C_3)LC_4c_1}{C_2c_2} + \frac{2\sqrt{2}(C_2+C_3)\sqrt{c_2}\sigma}{C_2\sqrt{b_{\mathrm{mb}}}} + \frac{c_3}{2}\Big]T^{-1/4} \\ &+ \frac{2(C_2+C_3)}{C_2c_2}\|B_{-1}-\nabla f(W_0)\|_F\,T^{-1/2} + \frac{Lc_1C_4^2}{C_2}T^{-3/4}. \end{aligned}\tag{19}$$

For fixed c_2, the first two terms in the bracket are minimized by $c_1^2 = (f(W_0)-f_{\inf})c_2/[(C_2+C_3)LC_4]$. Then optimizing over c_2 and choosing c_3 to match the resulting statistical scale gives

$$c_1 = \frac{2^{1/4}(f(W_0)-f_{\inf})^{3/4}b_{\mathrm{mb}}^{1/4}}{(C_2+C_3)^{3/4}(C_2/2+C_3)^{1/4}(L\sigma^2)^{1/4}},$$

$$c_2 = \Big(\frac{2(C_2/2+C_3)L(f(W_0)-f_{\inf})b_{\mathrm{mb}}}{(C_2+C_3)\sigma^2}\Big)^{1/2},\ c_3 = \Big(\frac{L(f(W_0)-f_{\inf})\sigma^2}{b_{\mathrm{mb}}}\Big)^{1/4}. \tag{20}$$

For these choices, the bracket in (19) equals

$$\Big[2^{11/4}\Big(1+\frac{C_3}{C_2}\Big)^{3/4}\Big(\frac{1}{2}+\frac{C_3}{C_2}\Big)^{1/4}+\frac{1}{2}\Big]\Big(\frac{L(f(W_0)-f_{\inf})\sigma^2}{b_{\mathrm{mb}}}\Big)^{1/4}. \qquad \square$$

6 Advantage of practical NS

Our convergence results do not reveal an advantage of practical NS over standard NS. As discussed for standard NS with WD, the NS iterations may leave very small singular values far from 1, potentially slowing convergence. Practical NS was designed to amplify them faster by increasing the singular-value map's slope at zero (Jordan et al., 2024), and the following ill-conditioned quadratic illustrates its optimization advantage. See Appendix C for full trajectory analysis and simulations.

Proposition 2 (Ill-conditioned quadratic). *Let $h, \ell, \kappa \in \mathbb{N}$, with $\kappa \gg 1$, $\ell = h\kappa$, $r = h+\ell$. Define*

$$Q = \mathrm{diag}(\kappa I_h, I_\ell), \quad f(W) = \tfrac{1}{2}\mathrm{tr}(W^\top QW), \quad W_0 = I_r.$$

The high-curvature block accounts for most of the gradient energy, while each block contributes half of the objective value. Consider one deterministic Muon update without momentum or WD, using five NS steps, exact Frobenius normalization, and an optimal stepsize. For fixed h and $\kappa \to \infty$,

$$\frac{f(W_1)}{f(W_0)} = \frac{1}{2} - \frac{a^5}{\kappa\sqrt{h}\,q_5(1/\sqrt{h})} + \mathcal{O}(\kappa^{-2}).$$

The leading improvement depends on a^5, about $20\times$ larger for practical than standard NS ($k=2$).

7 EXPERIMENTS

As full gradient evaluation is impractical for large language models, we examine Muon's $\mathcal{O}(T^{-1/4})$ convergence rate using ResNet-32 (He et al., 2016) with 118M parameters (Figure 1, left), where its Frobenius gradient norm follows an overall trend qualitatively similar to the $T^{-1/4}$ reference curve. We further examine Muon's theoretical $\sqrt{d}$ improvement over AdamW in the Frobenius norm using ResNet-32 of varying dimensions (Figure 1, right), where the normalized Muon-to-AdamW gradient norm ratios, denoted by $\sqrt{d}\|\nabla f(W_{\text{Muon}})\|_F / \|\nabla f(W_{\text{AdamW}})\|_F$, remain broadly stable across dimensions despite fluctuations, suggesting a growing relative advantage with dimension.

We also show on a 124M-parameter GPT-2-style model (Radford et al., 2019) that a large factor $(1-\epsilon_5)^{-1}$ arising from the worst-case NS orthogonalization error does not necessarily imply slow convergence (Figure 2, left), even when it is on the order of 10^7 for standard NS and 10^5 for practical NS (Figure 2, right). See Appendices D and E for further details, including the setup, larger-model results, and hyperparameter sweeps.

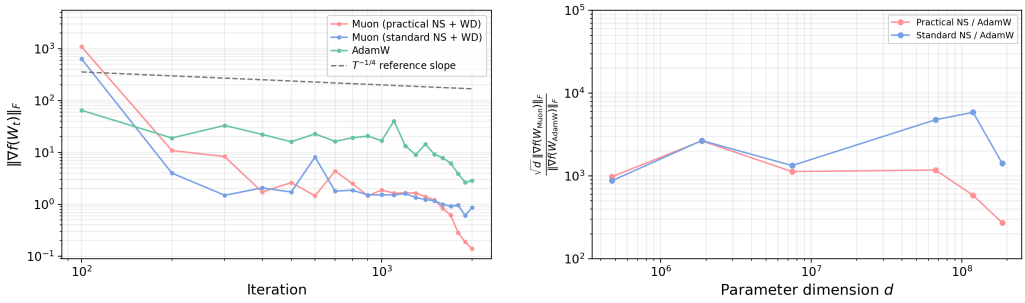

Figure 1: Empirical $\mathcal{O}(T^{-1/4})$ convergence and $\sqrt{d}$ improvement over AdamW on ResNet-32.

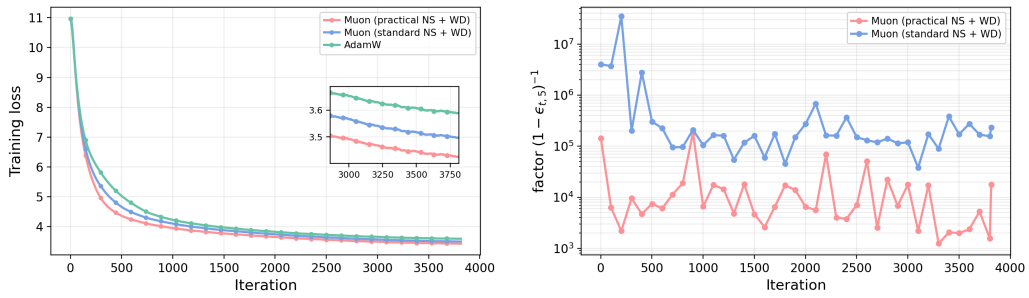

Figure 2: Training loss and orthogonalization error factor $(1-\epsilon_{t,5})^{-1}$ on the 124M GPT-2-style model.

8 CONCLUSION

In this paper, we provided an optimization interpretation and established convergence for Muon with practical Newton–Schulz iterations and decoupled weight decay. Our interpretation reveals a dynamic regularization effect that vanishes as stationarity is approached, preserving the original objective as the convergence target. On this basis, we established an $\mathcal{O}(T^{-1/4})$ convergence rate in the stochastic nonconvex setting, with a $\sqrt{d}$ improvement in dimension dependence over the best known AdamW's convergence rate in terms of the expected Frobenius norm of the gradient. Future work of particular interest includes understanding Muon's advantages and improving its performance.

AI USE STATEMENT

In this work, we used generative AI tools to improve the grammar of the manuscript. We have not used generative AI tools for research ideation, theoretical analysis, or the interpretation of ex-

perimental results. We have reviewed all AI-assisted work to ensure that it preserves the intended meaning and technical content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the Muon algorithm. In *International Conference on Learning Representations*, pp. 138323–138360, 2026.

Anuj Apte. Scale weight decay and train better. *arXiv preprint arXiv:2607.23777*, 2026.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *NeurIPS Workshop on Optimization for Machine Learning*, 2024.

Åke Björck and Clazett Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971.

Lizhang Chen, Jonathan Li, and qiang liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026.

Sayantan Choudhury, Xiaoran Cheng, Martin Takáč, Sen Na, and Mladen Kolar. Muon with Nesterov momentum: Heavy-tailed noise and (randomized) inexact polar decomposition. *arXiv preprint arXiv:2605.06884*, 2026.

Thang Do, Steffen Dereich, and Arnulf Jentzen. On MUON optimization: From non-convergence to an error analysis with Polar Express and the Newton-Schulz polynomial from implementations. *arXiv preprint arXiv:2608.04607*, 2026.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Chongyu Fan, Gaowen Liu, Mingyi Hong, Ramana Rao Kompella, and Sijia Liu. Rethinking Muon beyond pretraining: Spectral failures and high-pass remedies for VLA and RLVR. *arXiv preprint arXiv:2605.19282*, 2026.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Ekaterina Grishina, Matvey Smirnov, and Maxim Rakhuba. Accelerating Newton-Schulz iteration for orthogonalization via Chebyshev-type polynomials. *arXiv preprint arXiv:2506.10935*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Nicholas J Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008.

Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, et al. GLM-5V-Turbo: Toward a native foundation model for multimodal agents. *arXiv preprint arXiv:2604.26752*, 2026.

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/`.

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations*, pp. 45672–45718, 2026.

Jueun Kim, Baekrok Shin, Jihun Yun, Beomhan Baek, Minhak Song, and Chulhee Yun. Amuse: Anytime Muon with stable gradient evaluation. *arXiv preprint arXiv:2605.22432*, 2026.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Tim Tsz-Kit Lau, Qi Long, and Weijie Su. Polargrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2026.

Adrian S Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1):173–183, 1995.

Huan Li, Yiming Dong, and Zhouchen Lin. On the $O(\frac{\sqrt{d}}{K^{1/4}})$ convergence rate of AdamW measured by ℓ_1 norm. In *Advances in Neural Information Processing Systems*, pp. 132360–132387, 2026.

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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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

Elliot Paquette, Noah Marshall, Lucas Benigni, Guangyuan Wang, Atish Agarwala, and Courtney Paquette. Phases of Muon: When Muon eclipses SignSGD. *arXiv preprint arXiv:2605.09552*, 2026.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, pp. 30811–30849, 2024.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *International Conference on Machine Learning*, pp. 49069–49104, 2025.

Xun Qian and Peter Richtárik. Convergence analysis of Muon-type methods with inexact LMO in the degenerate case. *arXiv preprint arXiv:2606.21581*, 2026.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Walter Rudin. Principles of Mathematical Analysis. 2021.

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

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and Muons: Optimization via stochastic Frank–Wolfe under heavy-tailed noise. In *International Conference on Machine Learning*, 2026.

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. *Transactions on Machine Learning Research*, 2026.

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

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pp. 1139–1147, 2013.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Tijmen Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *Coursera: Neural Networks for Machine Learning*, 4(2):26, 2012.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. In *International Conference on Learning Representations*, pp. 142276–142311, 2026.

Kaifei Wang, Binghui Li, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws. In *International Conference on Machine Learning*, 2026a.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why Muon outperforms Adam: A curvature perspective. *arXiv preprint arXiv:2606.04662*, 2026b.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Tan. Muon outperforms Adam in tail-end associative memory learning. In *International Conference on Learning Representations*, pp. 37382–37419, 2026c.

Huaijin Wu, Bingrui Li, Yebin Yang, Yi Tu, Zhanpeng Zhou, Jianfei Chen, and Junchi Yan. Achieving low-bit Muon through subspace preservation and grid quantization. In *International Conference on Learning Representations*, 2026.

Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: ℓ_∞-norm constrained optimization. In *International Conference on Machine Learning*, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. GLM-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*, pp. 28386–28399, 2022.

Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. Towards understanding convergence and generalization of AdamW. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:6486–6493, 2024.

A ADDITIONAL RELATED WORK

Convergence of AdamW. Although the convergence of Adam and its variants has been extensively studied (Reddi et al., 2018; Zhang et al., 2022), decoupled WD complicates both the optimization interpretation and convergence analysis of AdamW. Xie & Li (2024) show that, if deterministic AdamW converges, its limit is a KKT point of an ℓ_∞-constrained problem. Zhou et al. (2024) provide the first convergence analysis of AdamW through a surrogate objective combining the original loss with a dynamic weighted ℓ_2 regularizer. However, their analysis requires an exponentially decreasing of the WD coefficient, progressively weakening regularization. More recently, Li et al. (2026) establish the best known convergence rate (1) for stationarity of the original objective with a WD coefficient held fixed across iterations. Inspired by these works, we investigate the role of WD in Muon while additionally accounting for its interaction with practical NS, thereby providing an optimization interpretation and convergence analysis for practical Muon.

Newton–Schulz iterations. NS iterations approximately orthogonalize matrices through repeated matrix polynomial updates (Björck & Bowie, 1971; Higham, 2008). Standard Taylor-based polynomials are designed to progressively improve the orthogonalization accuracy, whereas Muon's practical polynomial is tuned to move more rapidly toward orthogonalization within a small fixed number of iterations while allowing deviations from exact orthogonalization (Jordan et al., 2024). Recent methods such as Polar Express (Amsel et al., 2026) and CANS (Grishina et al., 2025) accelerate orthogonalization using a sequence of NS polynomials, each optimized for the singular-value range produced by the previous iteration.

Optimization interpretations of Muon. Muon's update direction is often interpreted as a steepest descent direction for a surrogate objective under a suitable norm constraint. Examples include spectral-norm steepest descent (Bernstein & Newhouse, 2024) and spectral-ball LMO formulations (Pethick et al., 2025) for Muon with exact orthogonalization, as well as Frank–Wolfe interpretations (Pethick et al., 2025; Sfyraki & Wang, 2026) and the Lion-$\mathcal{K}$ framework (Chen et al., 2026), which additionally account for WD. Beyond explaining the update direction, the Frank–Wolfe and Lion-$\mathcal{K}$ interpretations also recast Muon with exact orthogonalization and WD as a method for minimizing the original objective under a spectral-norm constraint on the parameters. A complementary perspective studies Muon through matrix preconditioning. Bernstein & Newhouse (2024) interpret Muon's exact orthogonalization as Shampoo preconditioning without accumulation, while Lau et al. (2026); Ma et al. (2026) interpret orthogonalization as preconditioning with inverse square roots of gradient Gram matrices, which can offer benefits such as reduced gradient anisotropy and faster convergence on certain structured matrix problems. Motivated by these perspectives, we also interpret Muon's update direction as a descent direction for an underlying objective and provide a corresponding preconditioning interpretation.

Understanding Muon's advantages over Adam and AdamW. Beyond the success discussed in the introduction, existing work investigates the mechanisms underlying Muon's advantages over Adam and AdamW from the perspectives of long-tail learning, directional curvature, and convergence complexity. For long-tail learning, these advantages are linked to Muon's more uniform update singular values, which can reduce the dominance of frequent patterns and promote more balanced learning across frequencies (Wang et al., 2026c; Vasudeva et al., 2026; Wang et al., 2026a). Regarding curvature, the preconditioning interpretation discussed above provides one theoretical perspective. Su (2025) offers another through an isotropic curvature model for a single optimization step, in which gradient orthogonalization is optimal when the curvature penalty has a sufficiently sharp kink. Empirical studies complement these theoretical perspectives by showing that Muon exhibits lower normalized directional curvature than Adam at matched validation loss (Wang et al., 2026b) and allocates a larger fraction of its update to low-curvature Hessian subspaces than AdamW (Kim et al., 2026). In terms of convergence complexity, Ma et al. (2026) and Paquette et al. (2026) compare Muon with exact orthogonalization against SignGD and SignSGD, respectively, while Kim & Oh (2026) compare Muon with standard NS against SGD with momentum, obtaining sharper rank dependence in convergence rates for nuclear norm stationarity. We complement these perspectives by directly comparing the convergence rate we establish for practical Muon in Theorem 1 with the best known AdamW's rate (1), showing Muon's sharper dimension dependence when stationarity is measured in the Frobenius norm.

B SUPPORTING LEMMAS

This appendix provides the complete proofs of Lemmas 2 and 3. We begin with Lemma 2 , deriving the sharp constants omitted from the main text.

Lemma 2 (NS direction alignment). *For both practical and standard NS, the momentum M_t and the resulting direction D_t in Algorithm 1 satisfy*

$$\langle M_t, D_t\rangle_F \geq C_2 \|M_t\|_F - C_2\varepsilon/4, \qquad \|D_t\|_F \leq C_3,$$

where $C_2 = \min_{s\in[0,1]} \chi_5(s) > 0$, $C_1 = \max_{s\in[0,1]} \chi_5(s) > 0$, as defined above, and

$$C_3 = \min\{C_1, q_{\max}\sqrt{r}\}, \qquad q_{\max} = \max_{s\in[0,1]} q_5(s), \qquad r = \min\{m, n\}.$$

In particular,

$$(C_2, C_1, q_{\max}) = \begin{cases} (0.6964\ldots,\ 484.8762\ldots,\ 1.2023\ldots), & \textit{practical NS}, \\ (1,\ a^5,\ 1), & \textit{standard NS}. \end{cases}$$

Proof. **(i)** The proof of Proposition 1 ensures that χ_5, C_1, C_2, and $q_{\max}$ are well defined. Their definitions directly give

$$C_2 s \leq q_5(s) \leq C_1 s, \qquad 0 \leq q_5(s) \leq q_{\max}, \qquad s \in [0,1].$$

We first establish the bounds in Lemma 2 and then determine the sharp constants C_1, C_2, and $q_{\max}$ for both practical and standard NS. By Lemma 1, $D_t = U\,\mathrm{diag}(q_5(s_{i,t}),\ldots,q_5(s_{r_t,t})V^\top$. Then (13) yields

$$\langle M_t, D_t\rangle_F = \tau_t \sum\nolimits_i s_{i,t} q_5(s_{i,t}) \geq C_2\tau_t \sum\nolimits_i s_{i,t}^2 = C_2 \|M_t\|_F^2 / \tau_t,$$

where $\|M_t\|_F^2/\tau_t \geq \|M_t\|_F - \varepsilon/4$, since for $\|M_t\|_F \leq \varepsilon$ the inequality reduces to $(\|M_t\|_F - \varepsilon/2)^2/\varepsilon \geq 0$, whereas for $\|M_t\|_F \geq \varepsilon$, the left-hand side equals $\|M_t\|_F$.

Moreover, (13) together with $\sum_i s_{i,t}^2 = \|M_t\|_F^2/\tau_t^2 \leq 1$ gives

$$\|D_t\|_F = \sqrt{\sum\nolimits_i q_5(s_{i,t})^2} \leq \min\left\{C_1\sqrt{\sum\nolimits_i s_{i,t}^2},\ q_{\max}\sqrt{r}\right\} \leq C_3.$$

(ii) We then determine the sharp constants C_1, C_2, and $q_{\max}$, which are illustrated in Figure 3. For practical and standard NS, respectively, we denote χ_5 by $\chi_{\mathrm{P},5}$ and $\chi_{\mathrm{S},5}^{(k)}$, and $q_{\max}$ by $q_{\mathrm{P},\max}$ and $q_{\mathrm{S},\max}^{(k)}$. We first consider the practical NS and establish the following bounds with both constants being sharp:

$$q_{\mathrm{P},5}(1)s \leq q_{\mathrm{P},5}(s) \leq a_{\mathrm{P}}^5 s, \qquad \forall s \in [0,1].$$

Hence $C_2 = q_{\mathrm{P},5}(1) = 0.6964\ldots$ and $C_1 = a_{\mathrm{P}}^5 = 484.8762\ldots$.

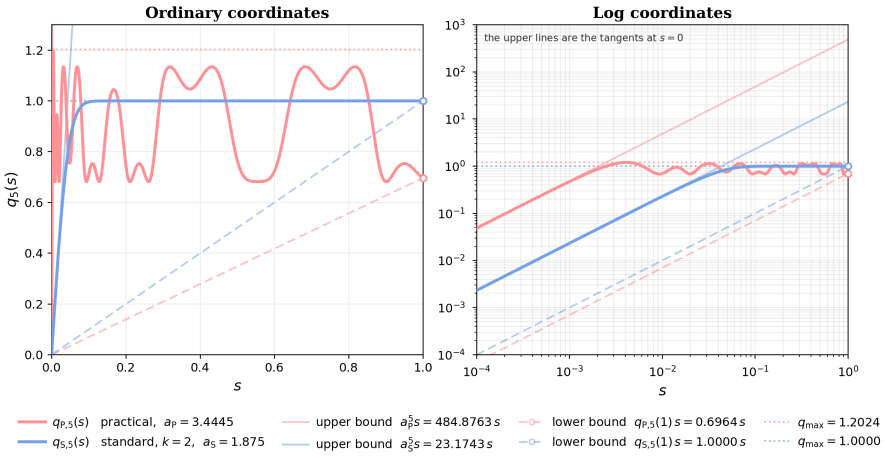

Figure 3: Sharp bounds for five-step Newton–Schulz maps.

The two positive roots of $q'_{\mathrm{P}}(s) = a_{\mathrm{P}} + 3b_{\mathrm{P}}s^2 + 5c_{\mathrm{P}}s^4$ are

$$s_- = \sqrt{\frac{-3b_{\mathrm{P}} - \sqrt{9b_{\mathrm{P}}^2 - 20a_{\mathrm{P}}c_{\mathrm{P}}}}{10c_{\mathrm{P}}}} = 0.5545\ldots,\ s_+ = \sqrt{\frac{-3b_{\mathrm{P}} + \sqrt{9b_{\mathrm{P}}^2 - 20a_{\mathrm{P}}c_{\mathrm{P}}}}{10c_{\mathrm{P}}}} = 1.0501\ldots.$$

We have $\bar{q}_{\mathrm{P}} = \max_{s\in[0,1]} q_{\mathrm{P}}(s) = q_{\mathrm{P}}(s_-) = 1.2023\ldots$, and denote $\underline{q}_{\mathrm{P}} = q_{\mathrm{P}}(s_+) = 0.6818\ldots$. The monotonicity intervals of q_{P}, together with $q_{\mathrm{P}}(\bar{q}_{\mathrm{P}}) = 0.9465\ldots < \bar{q}_{\mathrm{P}}$, gives

$$q_{\mathrm{P}}([0,\bar{q}_{\mathrm{P}}]) \subset [0,\bar{q}_{\mathrm{P}}], \qquad q_{\mathrm{P}}([\underline{q}_{\mathrm{P}},\bar{q}_{\mathrm{P}}]) \subset [\underline{q}_{\mathrm{P}},\bar{q}_{\mathrm{P}}].$$

Since q_{P} is continuous, $q_{\mathrm{P}}(0) = 0$, and $q_{\mathrm{P}}(s_-) = \bar{q}_{\mathrm{P}}$ with $s_- \in [0,1] \subset [0,\bar{q}_{\mathrm{P}}]$, the intermediate value theorem (Rudin, 2021) and the invariance of $[0,\bar{q}_{\mathrm{P}}]$ give

$$q_{\mathrm{P}}([0,1]) = q_{\mathrm{P}}([0,\bar{q}_{\mathrm{P}}]) = [0,\bar{q}_{\mathrm{P}}].$$

By induction, $q_{\mathrm{P},j}([0,1]) = [0,\bar{q}_{\mathrm{P}}]$ for $j = 1,\ldots,5$. Therefore,

$$q_{\mathrm{P,max}} = \max_{s\in[0,1]} q_{\mathrm{P},5}(s) = \bar{q}_{\mathrm{P}} = 1.2023\ldots.$$

Recall that $p_{\mathrm{P}}(s) = a_{\mathrm{P}} + b_{\mathrm{P}}s + c_{\mathrm{P}}s^2$ and $q_{\mathrm{P}}(s) = sp_{\mathrm{P}}(s^2)$. A direct calculation shows that p_{P} is decreasing on $[0,\underline{q}_{\mathrm{P}}^2]$ and $p_{\mathrm{P}}(\underline{q}_{\mathrm{P}}^2) = 1.6636\ldots > 1$, so $q_{\mathrm{P}}(s) \geq s$ for $s \in [0,\underline{q}_{\mathrm{P}}]$. Consequently, considering the process $q_{\mathrm{P},0}(s) = s$ and $q_{\mathrm{P},j+1}(s) = q_{\mathrm{P}}(q_{\mathrm{P},j}(s)) = q_{\mathrm{P},j}(s)p(q_{\mathrm{P},j}(s)^2)$ for $j = 0,\ldots,4$, every orbit starting from $s \in [0,1]$ either remains below $\underline{q}_{\mathrm{P}}$ and is nondecreasing, or enters the invariant interval $[\underline{q}_{\mathrm{P}},\bar{q}_{\mathrm{P}}]$. Hence

$$q_{\mathrm{P},5}(s) \geq \min\{s,\underline{q}_{\mathrm{P}}\}, \qquad 0 \leq s \leq 1.$$

Set $s_\star = \underline{q}_{\mathrm{P}}/q_{\mathrm{P},5}(1) = 0.9790\ldots$. For $s \in [0,s_\star]$, $q_{\mathrm{P},5}(s) \geq \min\{s,\underline{q}_{\mathrm{P}}\} \geq q_{\mathrm{P},5}(1)s$. For $s \in [s_\star,1]$, we first analyse the monotonicity of $q_{\mathrm{P},5}$. The following array tracks $q_{\mathrm{P},j}(s)$ after j NS steps, with $q_{\mathrm{P},0}(s) = s$.

NS steps j	Iterate	Enclosing interval I_j	$I_j \subset$
0	s	[0.9790, 1.0000]	(s_-,s_+)
1	$q_{\mathrm{P},1}(s)$	[0.7010, 0.7187]	(s_-,s_+)
2	$q_{\mathrm{P},2}(s)$	[1.0924, 1.1137]	$(s_+,+\infty)$
3	$q_{\mathrm{P},3}(s)$	[0.6983, 0.7209]	(s_-,s_+)
4	$q_{\mathrm{P},4}(s)$	[1.0897, 1.1167]	$(s_+,+\infty)$

These enclosures follow successively from $[s_\star,1] \subseteq I_0$ and $q_{\mathrm{P}}(I_j) \subseteq I_{j+1}$ for $j = 0,\ldots,3$, verified using the endpoint values and the monotonicity of q_{P} on each I_j. We know that $q'_{\mathrm{P}}(q_{\mathrm{P},j}(s)) < 0$ for $j = 0,1,3$ and $q'_{\mathrm{P}}(q_{\mathrm{P},j}(s)) > 0$ for $j = 2,4$. Thus the chain-rule product contains three negative factors and two positive factors, and $q'_{\mathrm{P},5}(s) = \prod_{j=0}^{4} q'_{\mathrm{P}}(q_{\mathrm{P},j}(s)) < 0$. Therefore, for $s \in [0,s_\star]$, $q_{\mathrm{P},5}(s) \geq q_{\mathrm{P},5}(1) \geq q_{\mathrm{P},5}(1)s$. Together, we obtain

$$q_{\mathrm{P},5}(1)s \leq q_{\mathrm{P},5}(s), \qquad 0 \leq s \leq 1.$$

Since equality holds at $x = 1$, $C_2 = q_{\mathrm{P},5}(1)$ is sharp.

Note that $\bar{q}_{\mathrm{P}}^2 < -b_{\mathrm{P}}/c_{\mathrm{P}}$, invariance of $[0,\bar{q}_{\mathrm{P}}]$ gives $p_{\mathrm{P}}(q_{\mathrm{P},j}(s)^2) \leq p_{\mathrm{P}}(0) = a_{\mathrm{P}}$. Thus

$$q_{\mathrm{P},5}(s) = s\prod\nolimits_{j=0}^{4} p_{\mathrm{P}}(q_{\mathrm{P},j}(s)^2) \leq a_{\mathrm{P}}^5 s,$$

and $C_1 = \max_{s\in[0,1]} \chi_{\mathrm{P},5}(s) = \chi_{\mathrm{P},5}(0) = a_{\mathrm{P}}^5 = 484.8762\ldots$.

(iii) We next consider the standard NS. The monotonicity of $q_{\mathrm{S}}^{(k)}$, together with $q_{\mathrm{S}}^{(k)}(0) = 0$, and $q_{\mathrm{S}}^{(k)}(1) = 1$ in Lemma 1 gives $q_{\mathrm{S}}^{(k)}([0,1]) = [0,1]$. By induction, $q_{\mathrm{S},j}^{(k)}([0,1]) = [0,1]$ and $q_{\mathrm{S,max}}^{(k)} = 1$. Moreover, the definition $p_{\mathrm{S}}^{(k)}(s) = 1 + \sum_{\ell=1}^{k} \frac{(2\ell)!}{4^\ell(\ell!)^2}(1-s)^\ell$ yields

$$1 \leq p_{\mathrm{S}}(s) \leq p_{\mathrm{S}}(0) = a, \qquad s \in [0,1].$$

Thus,

$$s \leq q_{\mathrm{S},5}(s) = s\prod\nolimits_{j=0}^{4} p_{\mathrm{S}}\big(q_{\mathrm{S},j}(s)^2\big) \leq \min\{a^5 s, 1\}, \qquad s \in [0,1], \tag{21}$$

which implies $C_2 = 1$ since $q_{\mathrm{S},5}(1) = 1$, and $C_1 = \max_{s\in[0,1]} \chi_{\mathrm{S},5}^{(k)}(s) = \chi_{\mathrm{S},5}^{(k)}(0) = a^5$. □

We next prove Lemma 3, which bounds the tracking error of the Nesterov-style momentum.

Lemma 3 (Momentum tracking). *Under the assumptions and parameter choices of Theorem 1,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|M_t-\nabla f(W_t)\right\|_F \le \frac{\left\|B_{-1}-\nabla f(W_0)\right\|_F}{c_2}T^{-1/2}+\Big(\frac{Lc_1C_4}{c_2}+\sqrt{2c_2}\,\frac{\sigma}{\sqrt{b_{\mathrm{mb}}}}\Big)T^{-1/4}.$$

Proof. Unrolling the two momentum recursions defining B_t and M_t in Algorithm 1 gives

$$\begin{aligned} B_t &= \beta^{t+1}B_{-1}+(1-\beta)\sum_{j=0}^{t}\beta^{t-j}G_j,\\ M_t &= \beta^{t+2}B_{-1}+(1-\beta)\sum_{j=0}^{t-1}\beta^{t-j+1}G_j+(1-\beta^2)G_t. \end{aligned} \tag{22}$$

Subtracting $\nabla f(W_t)$ from (22), using $G_j=\nabla f(W_j)+G_j-\nabla f(W_j)$, and the reindexing identity

$$\begin{aligned} &(1-\beta)\sum_{j=0}^{t-1}\beta^{t-j+1}\nabla f(W_j)-\beta^2\nabla f(W_t)\\ &=\sum_{j=0}^{t-1}\beta^{t-j+1}\nabla f(W_j)-\sum_{j=0}^{t-1}\beta^{t-j+2}\nabla f(W_j)-\beta^2\nabla f(W_t)\\ &=\sum_{j=1}^{t}\beta^{t-j+2}\nabla f(W_{j-1})-\beta^{t+2}\nabla f(W_0)-\sum_{j=1}^{t}\beta^{t-j+2}\nabla f(W_j)\\ &=-\beta^{t+2}\nabla f(W_0)-\sum_{j=1}^{t}\beta^{t-j+2}\big(\nabla f(W_j)-\nabla f(W_{j-1})\big), \end{aligned}$$

we obtain

$$\begin{aligned} e_t &= \beta^{t+2}\big(B_{-1}-\nabla f(W_0)\big)-\sum_{j=1}^{t}\beta^{t-j+2}\big(\nabla f(W_j)-\nabla f(W_{j-1})\big)+N_t, \qquad (23)\\ N_t &= (1-\beta^2)\big(G_t-\nabla f(W_t)\big)+(1-\beta)\sum_{j=0}^{t-1}\beta^{t-j+1}\big(G_j-\nabla f(W_j)\big). \end{aligned}$$

Assumption 2 and (16) imply $\left\|\nabla f(W_j)-\nabla f(W_{j-1})\right\|_F \le L\eta C_4$. Thus the triangle inequality and the geometric-series bounds in (23) give

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|e_t\right\|_F \le \frac{\left\|B_{-1}-\nabla f(W_0)\right\|_F}{T(1-\beta)}+\frac{L\eta C_4}{1-\beta}+\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|N_t\right\|_F. \tag{24}$$

By independent sampling and Assumption 3, we have

$$\mathbb{E}\|N_t\|_F^2 \le \frac{\sigma^2}{b_{\mathrm{mb}}}\Big[(1-\beta^2)^2+(1-\beta)^2\sum_{j=0}^{t-1}\beta^{2(t-j+1)}\Big] \le \frac{\sigma^2}{b_{\mathrm{mb}}}\,\frac{(1-\beta)(1+2\beta-2\beta^3)}{1+\beta} \le \frac{2(1-\beta)\sigma^2}{b_{\mathrm{mb}}}.$$

By the Cauchy–Schwarz inequality,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|N_t\right\|_F \le \Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|N_t\right\|_F^2\Big)^{1/2} \le \frac{\sigma}{\sqrt{b_{\mathrm{mb}}}}\sqrt{2(1-\beta)}.$$

Substituting this bound, $\eta=c_1T^{-3/4}$, and $(1-\beta)=c_2T^{-1/2}$ into (24) yields the desired result. □

C Details on advantage of practical NS

We provide a more detailed convergence analysis of the ill-conditioned quadratic example to illustrate the advantage of practical NS over standard NS in the following extension of Proposition 2.

Proposition 2 (Ill-conditioned quadratic). *Let $h,\ell,\kappa\in\mathbb{N}$, with $\kappa\gg 1$, $\ell=h\kappa$, and $r=h+\ell$. Define*

$$Q=\mathrm{diag}(\kappa I_h, I_\ell), \qquad f(W)=\frac{1}{2}\,\mathrm{tr}(W^\top QW), \qquad W_0=I_r.$$

The high-curvature block accounts for most of the gradient energy, while each block contributes half of the objective value. We consider practical NS or standard NS with $k=2$, and normalize without an ε-stabilizer, stopping when the gradient vanishes.

(i) One-step contraction. *Denote* $\tau_0 = \|\nabla f(W_0)\|_F = \sqrt{h\kappa(\kappa+1)},\ u = q_5(\kappa/\tau_0),\ v = q_5(1/\tau_0)$. *For deterministic Muon without momentum or WD, the respective optimal stepsizes minimizing* $f(W_1)$ *for the two NS choices yield*

$$\frac{f(W_1)}{f(W_0)} = \frac{h\kappa\ell(u-v)^2}{(h\kappa+\ell)(h\kappa u^2+\ell v^2)} = \frac{(u-v)^2}{2(u^2+v^2)} \le \left(\frac{u-v}{u+v}\right)^2.$$

Consider the example with $h = 1$ *and* $\ell = \kappa = 100$*, five practical and standard NS iterations give* $(u, v) \approx (0.7020, 0.6968)$ *and* $(1, 0.2274)$*, respectively, leaving* 0.0014% *and* 28.3789% *of* $f(W_0)$ *after the one-step update.*

For fixed h *and* $\kappa \to \infty$*, substituting* $u = q_5(1/\sqrt{h}) + \mathcal{O}(\kappa^{-1})$ *and* $v = a^5/(\kappa\sqrt{h}) + \mathcal{O}(\kappa^{-2})$ *gives*

$$\frac{f(W_1)}{f(W_0)} = \frac{1}{2} - \frac{a^5}{\kappa\sqrt{h}\, q_5(1/\sqrt{h})} + \mathcal{O}(\kappa^{-2}).$$

When $q_5(1/\sqrt{h})$ *is close to one for both NS choices, the leading-order improvement is mainly governed by* a^5*, which is approximately* 484.8762 *for practical NS and* 23.1743 *for standard NS. Thus, practical NS yields a substantially stronger one-step contraction.*

(ii) Full optimization trajectory. *Consider the example with* $h = 1$ *and* $\ell = \kappa = 100$*. Let* W_t^{P} *and* W_t^{S} *denote the Muon iterates generated by practical and standard NS, respectively, using the optimal stepsizes at each iteration* t *that minimize* $f(W_{t+1}^{\mathrm{P}})$ *and* $f(W_{t+1}^{\mathrm{S}})$*. Then*

$$\frac{f(W_t^{\mathrm{P}})}{f(W_0)} \le (6.4514 \times 10^{-5})^t, \qquad \frac{f(W_t^{\mathrm{S}})}{f(W_0)} \ge (0.1406)^t.$$

Further comparisons for different values of h *and* κ *are provided in Table 2.*

Table 2: One-step and full trajectory comparisons for the ill-conditioned quadratic. Panel (a) reports $f(W_1^{\mathrm{P}})/f(W_1^{\mathrm{S}})$; panel (b) reports $T_{\mathrm{P}}/T_{\mathrm{S}}$, where T_{P} and T_{S} are the numbers of iterations required by practical and standard NS, respectively, to reduce the objective to 10^{-6} of its initial value.

(a) One-step contraction

$h\backslash\kappa$	100	10^3	10^4	10^5	10^6
1	4.9749×10^{-5}	0.0762	0.8655	0.9865	0.9987
10	5.4271×10^{-7}	0.7462	0.9744	0.9974	0.9997
100	0.0839	0.8685	0.9868	0.9987	0.9999

(b) Full trajectory (numerical simulations)

$h\backslash\kappa$	100	10^3	10^4	10^5	10^6
1	0.1250	0.0376	0.0176	0.0112	0.0280
10	0.0208	0.0440	0.0033	0.0012	0.0011
100	0.0298	0.0157	0.0175	0.0263	0.0288

Proof. **(i) One-step contraction.** The normalized gradient and the NS direction are

$$\frac{\nabla f(W_0)}{\tau_0} = \mathrm{diag}\left(\frac{\kappa}{\tau_0} I_h, \frac{1}{\tau_0} I_\ell\right), \qquad D_0 = \mathrm{diag}(uI_h, vI_\ell).$$

Hence,

$$f(W_0 - \eta D_0) = \frac{1}{2}\big(h\kappa(1-\eta u)^2 + \ell(1-\eta v)^2\big),$$

and the optimal stepsize minimizing $f(W_0 - \eta D_0)$ is

$$\eta_0^\star = \frac{h\kappa u + \ell v}{h\kappa u^2 + \ell v^2}.$$

With $\eta = \eta_0^\star$, we have

$$\frac{f(W_1)}{f(W_0)} = 1 - \frac{(h\kappa u+\ell v)^2}{(h\kappa+\ell)(h\kappa u^2+\ell v^2)} = \frac{h\kappa\ell(u-v)^2}{(h\kappa+\ell)(h\kappa u^2+\ell v^2)} = \frac{(u-v)^2}{2(u^2+v^2)} \le \left(\frac{u-v}{u+v}\right)^2,$$

where the last two steps use $\ell = h\kappa$ and $(u+v)^2 \le 2(u^2+v^2)$.

For $h = 1$ and $\ell = \kappa = 100$, direct evaluation of the five NS iterations gives

$$(u_{\mathrm{P}}, v_{\mathrm{P}}) \approx (0.7020\ldots, 0.6967\ldots), \qquad (u_{\mathrm{S}}, v_{\mathrm{S}}) \approx (1, 0.2273\ldots).$$

Hence, the values of $f(W_1)/f(W_0)$ are approximately 1.4118×10^{-5} and $0.2837\ldots$ for practical and standard NS, respectively, which gives the stated percentages.

Next, fix h and let $\kappa \to \infty$ with $\ell = h\kappa \in \mathbb{N}$. We have

$$\frac{\kappa}{\tau_0} = \frac{1}{\sqrt{h}}(1+\kappa^{-1})^{-1/2} = \frac{1}{\sqrt{h}} + \mathcal{O}(\kappa^{-1}), \qquad \frac{1}{\tau_0} = \frac{1}{\kappa\sqrt{h}}(1+\kappa^{-1})^{-1/2} = \frac{1}{\kappa\sqrt{h}} + \mathcal{O}(\kappa^{-2}).$$

Since $q(s) = as + bs^3 + cs^5$, its fivefold composition satisfies $q_5'(0) = a^5$. Consequently,

$$u = q_5(1/\sqrt{h}) + \mathcal{O}(\kappa^{-1}), \qquad v = \frac{a^5}{\kappa\sqrt{h}} + \mathcal{O}(\kappa^{-2}),$$

and, because $u > 0$,

$$\frac{v}{u} = \frac{a^5}{\kappa\sqrt{h}\, q_5(1/\sqrt{h})} + \mathcal{O}(\kappa^{-2}).$$

Using $(1-z)/(1+z) = 1 - 2z + \mathcal{O}(z^2)$ as $z \to 0$ yields

$$\frac{f(W_1)}{f(W_0)} \le \left(1 - \frac{2a^5}{\kappa\sqrt{h}\, q_5(1/\sqrt{h})}\right)^2 + \mathcal{O}(\kappa^{-2}).$$

In fact, the exact contraction formula also gives

$$\frac{f(W_1)}{f(W_0)} = \frac{1}{2} - \frac{v/u}{1+(v/u)^2} = \frac{1}{2} - \frac{a^5}{\kappa\sqrt{h}\, q_5(1/\sqrt{h})} + \mathcal{O}(\kappa^{-2}).$$

When $q_5(1/\sqrt{h})$ is close to one, the leading-order comparison is governed by a^5, with $a_{\mathrm{P}}^5 \approx 484.8762$ and $a_{\mathrm{S}}^5 \approx 23.1743$.

(ii) Full optimization trajectory. Set $h = 1$ and $\ell = \kappa = 100$. The block structure is preserved throughout the iterations, so write

$$W_t = \mathrm{diag}(x_t, y_t I_{100}), \qquad x_0 = y_0 = 1.$$

Then

$$f(W_t) = 50(x_t^2 + y_t^2), \qquad \nabla f(W_t) = \mathrm{diag}(100x_t, y_t I_{100}).$$

Whenever $x_t, y_t \neq 0$, define

$$z_t = \frac{|x_t|}{|y_t|}, \quad u(z) = q_5\left(\frac{100z}{\sqrt{10^4 z^2 + 100}}\right), \quad v(z) = q_5\left(\frac{1}{\sqrt{10^4 z^2 + 100}}\right), \quad \Psi(z) = \frac{v(z)}{u(z)}.$$

With $u_t = u(z_t)$ and $v_t = v(z_t)$, the NS direction is

$$D_t = \mathrm{diag}\big(\mathrm{sgn}(x_t)u_t, \mathrm{sgn}(y_t)v_t I_{100}\big),$$

where $\mathrm{sgn}(\cdot)$ denotes the sign function. Since $x_t = \mathrm{sgn}(x_t)|x_t|$ and $y_t = \mathrm{sgn}(y_t)|y_t|$, we have

$$f(W_t - \eta D_t) = 50\big((|x_t| - \eta u_t)^2 + (|y_t| - \eta v_t)^2\big)$$

Differentiating with respect to η gives

$$\frac{\mathrm{d}}{\mathrm{d}\eta} f(W_t - \eta D_t) = 100\big(\eta(u_t^2 + v_t^2) - |x_t|u_t - |y_t|v_t\big), \ \eta_t^\star = \frac{|x_t|u_t + |y_t|v_t}{u_t^2 + v_t^2}.$$

Substitution into the update gives

$$x_{t+1} = \mathrm{sgn}(x_t)\frac{v_t(|x_t|v_t - |y_t|u_t)}{u_t^2 + v_t^2},$$
$$y_{t+1} = \mathrm{sgn}(y_t)\frac{u_t(|y_t|u_t - |x_t|v_t)}{u_t^2 + v_t^2}.$$

If $|x_t|v_t - |y_t|u_t = 0$, both new blocks vanish and the algorithm reaches the minimizer. Otherwise,

$$\begin{aligned} z_{t+1} &= \frac{|x_{t+1}|}{|y_{t+1}|} = \frac{v_t}{u_t} = \Psi(z_t), \\ \frac{f(W_{t+1})}{f(W_t)} &= \frac{x_{t+1}^2 + y_{t+1}^2}{x_t^2 + y_t^2} = \frac{(u_t^2 + v_t^2)(|x_t|v_t - |y_t|u_t)^2}{(x_t^2 + y_t^2)(u_t^2 + v_t^2)^2} \\ &= \frac{(|x_t|v_t - |y_t|u_t)^2}{(x_t^2 + y_t^2)(u_t^2 + v_t^2)} = \frac{|y_t|^2 u_t^2 (z_t \frac{v_t}{u_t} - 1)^2}{|y_t|^2(1 + z_t^2)\, u_t^2 (1 + (\frac{v_t}{u_t})^2)} \\ &= \frac{(1 - z_t \Psi(z_t))^2}{(1 + z_t^2)(1 + \Psi(z_t)^2)}. \end{aligned} \tag{25}$$

We next find an interval J containing $z_0 = 1$ such that $\Psi(J) \subset J$. We also bound $z\Psi(z)$ on J to derive the convergence bounds for the objective value.

1) Practical NS. We show that $z_t \in [0.9920, 1]$ for all t, which keeps the numerator in (25) small.

Suppose $z \in [0.9920, 1]$. The normalized singular values satisfy

$$\frac{100z}{\sqrt{10^4 z^2 + 100}} \in [0.9949\ldots, 0.9950\ldots], \qquad \frac{1}{\sqrt{10^4 z^2 + 100}} \in [0.0099\ldots, 0.0100\ldots].$$

On these two intervals, $q_{\mathrm{P},5}$ is decreasing and increasing, respectively, as verified by direct interval propagation through the five NS steps, following the same argument as in the proof of Lemma 2. A direct calculation gives

$$0.7020\ldots \le u(z) \le 0.7021\ldots, \qquad 0.6967\ldots \le v(z) \le 0.7002\ldots.$$

Consequently,

$$0.9920 < \frac{0.6967\ldots}{0.7021\ldots} \le \Psi(z) \le \frac{0.7002\ldots}{0.7020\ldots} < 1.$$

Since $z_0 = 1$ and $z_{t+1} = \Psi(z_t)$, induction gives $z_t \in [0.9920, 1]$ for every t. In particular,

$$0.9920^2 \le z_t \Psi(z_t) < 1.$$

Applying (25), we obtain

$$\frac{f(W_{t+1}^{\mathrm{P}})}{f(W_t^{\mathrm{P}})} \le \frac{(1 - 0.9920^2)^2}{(1 + 0.9920^2)^2} \le 6.4514 \times 10^{-5}.$$

Multiplying the above inequality yields the desired result.

2) Standard NS. We show that $z_t \in J_{\mathrm{S}} = [1/5, 1]$ for all t. Suppose $z \in [1/5, 1]$. The normalized singular values satisfy $1/\sqrt{10^4 z^2 + 100} \ge s_0 = 1/\sqrt{10100}$. Recall that $a_{\mathrm{S}} = 15/8$. By (21),

$$s \le q_{\mathrm{S},5}^{(2)}(s) \le \min\{a_{\mathrm{S}}^5 s, 1\}, \qquad s \in [0, 1].$$

For the first four NS step ($j = 0, \ldots, 4$), we have

$$q_{\mathrm{S},j}^{(2)}(s_0) \le a_{\mathrm{S}}^j s_0 < \frac{a_{\mathrm{S}}^4}{100} < \frac{1}{8}.$$

Thus, at each of the five NS steps, we can apply

$$q_{\mathrm{S}}^{(2)}(s) = \frac{15}{8}s - \frac{5}{4}s^3 + \frac{3}{8}s^5 \ge \left(\frac{15}{8} - \frac{5}{256}\right) s = \frac{475}{256}s, \qquad s \in [0, 1/8],$$

which gives

$$q_{\mathrm{S},5}^{(2)}(s_0) \ge (\frac{475}{256})^5 s_0 > \frac{1}{101}(\frac{475}{256})^5 > \frac{1}{5}.$$

For $z \in J_{\mathrm{S}}$, the normalized singular values satisfy $s_0 \le 1/\sqrt{10^4 z^2 + 100} \le 100z/\sqrt{10^4 z^2 + 100} \le 1$. Monotonicity of $q_{\mathrm{S},5}^{(2)}$ therefore yields

$$\frac{1}{5} < v(z) \le u(z) \le 1, \qquad \frac{1}{5} < \Psi(z) = \frac{v(z)}{u(z)} \le 1,$$

which proves the claimed inclusion. Moreover,

$$z\Psi(z) \le z\frac{a_{\mathrm{S}}^5/\sqrt{10^4z^2+100}}{100z/\sqrt{10^4z^2+100}} = \frac{a_{\mathrm{S}}^5}{100} < \frac{1}{4}.$$

Starting from $z_0 = 1$, (25) thus gives,

$$\frac{f(W_{t+1}^{\mathrm{S}})}{f(W_t^{\mathrm{S}})} = \frac{(1-z_t\Psi(z_t))^2}{(1+z_t^2)(1+\Psi(z_t)^2)} \ge \frac{(1-1/4)^2}{4} = \frac{9}{64} = 0.140625.$$

Multiplying the above inequality yields the desired result. □

D Experimental details on ResNet-32

This appendix details the ResNet-32 experiments used to examine Muon's convergence rate and its advantage over AdamW in dimension dependence, including the experimental setup, gradient norm convergence rates, accuracy and loss results, and hyperparameter sweeps.

D.1 Experimental setting

Models and data. We use ResNet-32 models of varying dimensions measured by the number of model parameters. By changing the base channel width $w \in \{16, 32, 64, 192, 256, 320\}$, we obtain models with $\{0.47, 1.86, 7.41, 66.46, 118.11, 184.50\}$ million parameters, respectively. For each choice of w, the model consists of an initial 3×3 convolution followed by BatchNorm and ReLU, and three residual stages with channel widths w, $2w$, and $4w$, respectively. Each stage contains five residual blocks, each comprising two 3×3 convolutions with BatchNorm, while the first block of the second and third stages downsamples the feature map. The resulting features are then globally average pooled and passed through a linear classifier to produce logits for 100 classes. All models are trained from random initialization.

We train the models on CIFAR-100 (Krizhevsky, 2009), using the standard training set containing 50,000 images and test set containing 10,000 images. All images are normalized per channel, with no data augmentation.

Compared optimizers and hyperparameters. We compare three optimizers: Muon with practical NS and WD, Muon with standard NS ($k = 2$) and WD, and AdamW.

Muon is applied to convolutional and fully connected weight matrices, with convolutional kernels reshaped into two-dimensional matrices, while an auxiliary AdamW optimizer updates biases and normalization parameters with a fixed stepsize of 10^{-3} and zero WD. Both Muon variants use five NS steps, momentum coefficient $\beta = 0.95$, and normalization floor $\varepsilon = 10^{-7}$. For the AdamW baseline, AdamW updates all parameters, with biases and normalization parameters excluded from WD. It uses first- and second-moment coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$ and numerical stability constant $\hat{\varepsilon} = 10^{-8}$. The selected stepsizes and WD coefficients are reported in Table 3, and the stepsize schedule consists of a 5% linear warmup followed by cosine decay to zero. Specifically, for each dimension and optimizer, we jointly sweep the stepsize and WD coefficient, training each setting for 2,000 iterations with a batch size of 128 and a fixed random seed, and select the setting with the lowest final validation loss. The resulting sweep landscapes are shown in Appendix D.4 (Figure 6). The final experiments use three random seeds. The forward and backward passes do not use BF16 autocast, while model parameters and optimizer states are kept in FP32, with CUDA TF32 operations enabled. Neither gradient clipping nor shape dependent Muon scaling is used.

Table 3: Selected stepsize and WD pairs for ResNet-32.

Dimension (M)	Muon (Practical NS + WD)	Muon (Standard NS + WD)	AdamW
0.47	$(0.1, 0.0625)$	$(0.1, 0.0625)$	$(0.01, 0.0625)$
1.86	$(0.05, 0.1)$	$(0.05, 0.1)$	$(0.01, 0.15)$
7.41	$(0.02, 0.3)$	$(0.05, 0.1)$	$(0.001, 0.04)$
66.46	$(0.05, 0.15)$	$(0.0075, 0.35)$	$(0.001, 0.0625)$
118.11	$(0.01, 0.6)$	$(0.1, 0.0625)$	$(0.0003, 0.1)$
184.50	$(0.02, 0.4)$	$(0.01, 0.35)$	$(0.0003, 0.2)$

Evaluation metric. We measure stationarity at every iteration using the Frobenius norm of the gradient computed over the entire training set, $\|\nabla f(W)\|_F$. To compare the dimension dependence of Muon and AdamW, at iteration 2,000, we report $\sqrt{d}\|\nabla f(W_{\text{Muon}})\|_F / \|\nabla f(W_{\text{AdamW}})\|_F$, where W_{Muon} and W_{AdamW} denote the parameters obtained by Muon and AdamW, respectively. If this ratio remains approximately constant as d increases, the results support Muon's $\sqrt{d}$ improvement in dimension dependence over AdamW established by our theoretical comparison. Training and test losses are recorded every 100 iterations, with final top-1 accuracy evaluated on the full test set.

D.2 Gradient norm convergence rates

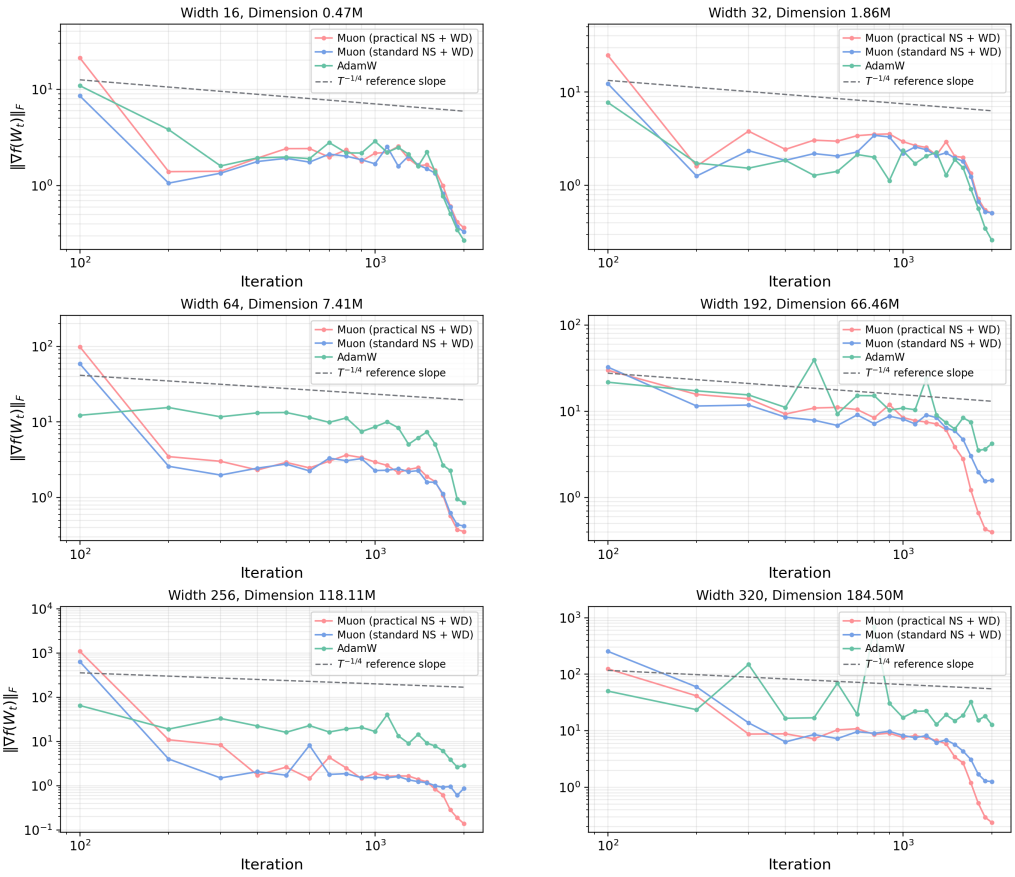

Figure 4: Frobenius gradient norm convergence rates for ResNet-32 models of varying dimensions. The gradient norms exhibit an overall $T^{-1/4}$ trend, with Muon generally converging faster than AdamW.

D.3 Accuracy and loss results

Table 4: CIFAR-100 test top-1 accuracy (over three seeds) for ResNet-32 models of varying dimensions after 2,000 iterations. Muon consistently achieves higher accuracy than AdamW.

Algorithm \ Dimension (M)	0.47	1.86	7.41	66.46	118.11	184.50
Muon (practical + WD)	59.64 ± 0.21	66.35 ± 0.38	70.22 ± 0.25	73.66 ± 0.18	72.74 ± 0.62	74.57 ± 0.04
Muon (standard + WD)	58.71 ± 0.36	65.59 ± 0.31	68.97 ± 0.38	72.28 ± 0.20	70.88 ± 0.40	72.92 ± 0.15
AdamW	45.95 ± 1.15	52.29 ± 1.88	57.03 ± 0.55	59.79 ± 1.61	62.69 ± 0.37	63.17 ± 0.62

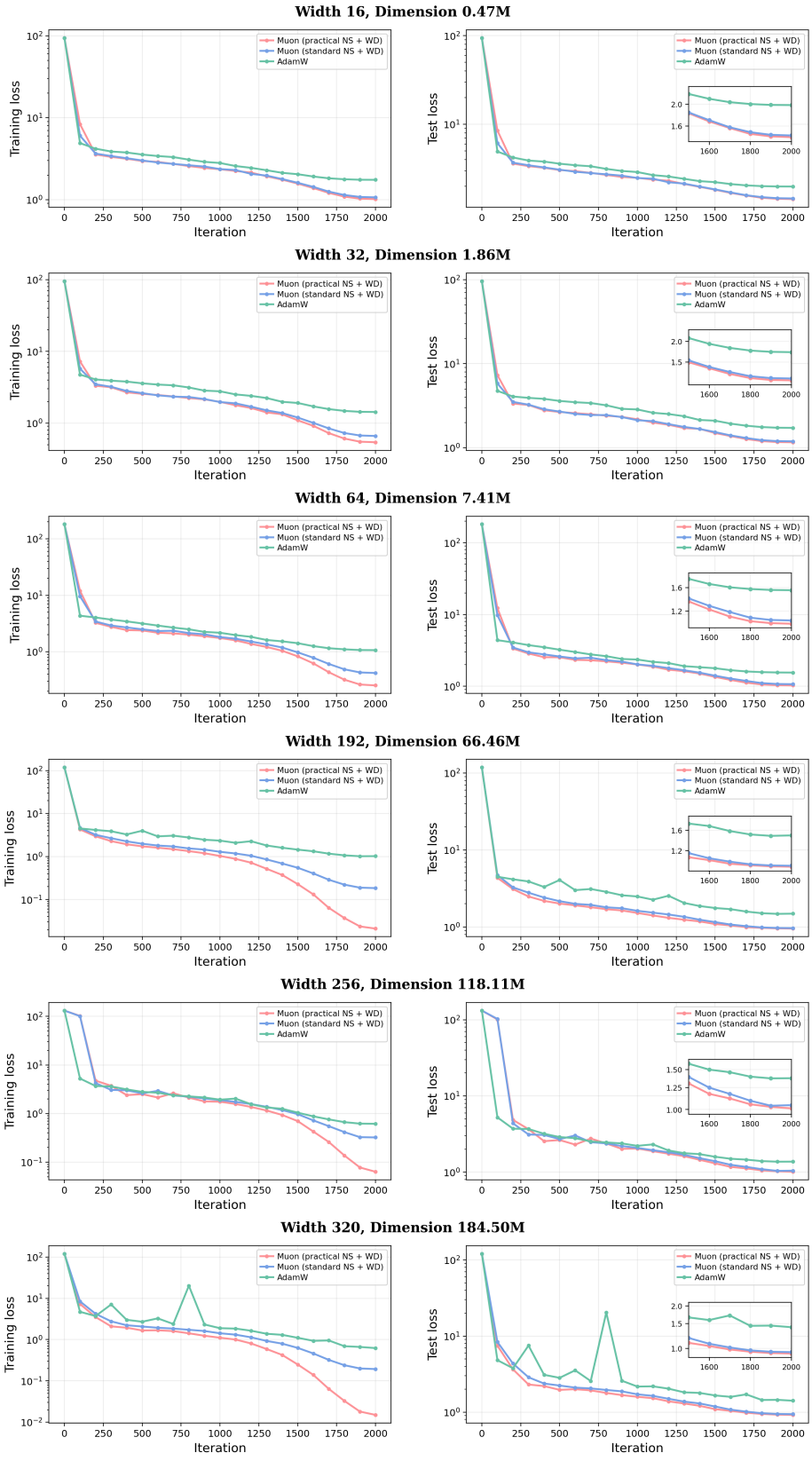

Figure 5: Loss curves for ResNet-32 models of varying dimensions. Muon consistently achieves lower training and test losses than AdamW.

D.4 HYPERPARAMETER SWEEP

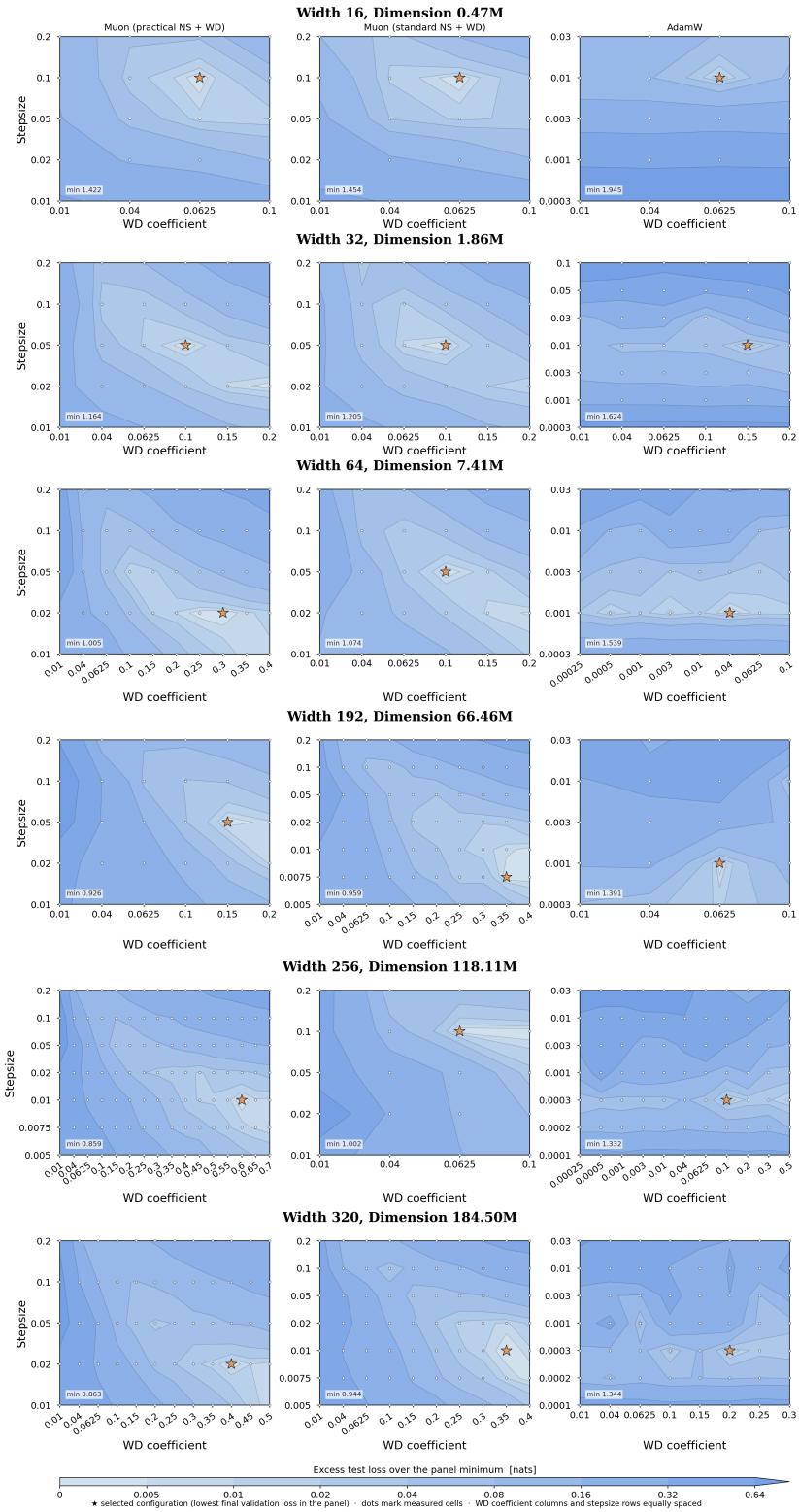

Figure 6: Joint sweep over stepsize and WD for ResNet-32 models of varying dimensions.

E EXPERIMENTAL DETAILS ON LANGUAGE MODELS

This appendix details the experiments on the 124M GPT-2-style model and presents additional results on the larger 355M GPT-2-style and 0.6B Qwen3-style models (Yang et al., 2025), further showing that large factors arising from worst-case NS orthogonalization errors need not slow Muon's convergence. Specifically, we provide the experimental setup, loss curves, singular value distributions, maximum orthogonalization errors and the resulting factors, and hyperparameter sweeps.

E.1 EXPERIMENTAL SETTING

Models and data. We use three models: 124M- and 355M-parameter GPT-2-style models, and a 0.6B-parameter Qwen3-style model. The 124M model has 12 layers, hidden size 768, and 12 attention heads. The 355M model has 24 layers, hidden size 1,024, and 16 attention heads. The 0.6B model has 28 layers, hidden size 1,024, 16 query heads, 8 key-value heads, and an intermediate size of 3,072. All models are trained from random initialization.

We train the 124M, 355M, and 0.6B models for 500M, 200M, and 300M tokens, respectively, and evaluate them on a separate validation set containing 5M tokens. Training data come from the FineWeb-Edu sample-10BT subset (Penedo et al., 2024). The 124M and 355M GPT-2-style models use the GPT-2 tokenizer, while the 0.6B Qwen3-style model uses the Qwen3 tokenizer. All models use a sequence length of 512 tokens, and a batch size of 131,072 tokens per optimizer update.

Task and metrics. We consider autoregressive language modeling with a next-token prediction objective and report training and validation cross-entropy losses. To illustrate that the worst-case orthogonalization error does not necessarily imply slow convergence, we track the singular-value distributions and orthogonalization errors $\epsilon_{t,5}$ of selected momentum matrices, along with the factors $(1-\epsilon_{t,5})^{-1}$, motivated by the factor $(1-\epsilon_5)^{-1}$ in the convergence analysis of standard NS without WD by Kim & Oh (2026).

Training loss is recorded at every iteration, and validation loss is evaluated every 200 iterations for the 124M model and every 100 iterations for the 355M and 0.6B models. Each evaluation is performed using 20 batches for the 124M and 355M GPT-2-style models and 10 batches for the 0.6B Qwen3-style model.

The singular values of momentum matrices for representative attention and MLP weights in the first, middle, and last transformer blocks are computed using FP64 SVDs without affecting training and recorded at every iteration. Let $s_{i,t}^{l}$ denote the i-th normalized singular value of momentum matrix l at iteration t. The orthogonalization error $\epsilon_{t,5}$ is computed as follows:

$$\epsilon_{t,5}^{(\hat{a})} = \max_{l} \max_{i} \left| q_{\hat{a},5}(s_{i,t}^{l}) - 1 \right|, \qquad \hat{a} \in \{\mathrm{P}, \mathrm{S}\}.$$

Compared optimizers and hyperparameters. We use the three optimizers as in Appendix D. Muon is applied to two-dimensional hidden-layer matrices, while embeddings, output heads, biases, and normalization parameters are updated by an auxiliary AdamW optimizer. The auxiliary AdamW uses a fixed stepsize of 6×10^{-4} for the 124M model and 3×10^{-4} for the 355M and 0.6B models. It applies the selected WD of the corresponding Muon variant to embeddings and output-head matrices, while biases and normalization parameters use zero WD. For the AdamW baseline, all parameters are updated by AdamW, with the selected WD applied to matrix parameters and zero WD applied to biases and normalization parameters.

The optimizer settings follow Appendix D, except where otherwise specified. AdamW uses $\beta_2 = 0.95$. The selected stepsizes and WD coefficients are reported in Table 5, and the stepsize schedule consists of a 2% linear warmup followed by cosine decay to 10% of the selected value. Specifically, for each model and optimizer, we jointly sweep the stepsize and WD coefficient, training each setting with a fixed random seed for 30% of the corresponding training schedule, i.e., 1,145, 458, and 687 iterations for the 124M, 355M, and 0.6B models, respectively, and select the setting with the lowest final validation loss. The resulting sweep landscapes are shown in Appendix E.4 (Figure 10). The final experiments use three seeds for the 124M and 355M models and one seed for the 0.6B model. The forward and backward passes use BF16 autocast.

Table 5: Selected stepsize and WD pairs for language model training.

Model	Muon (Practical NS + WD)	Muon (Standard NS + WD)	AdamW
124M GPT-2-style	$(0.02, 0.000625)$	$(0.02, 0.0025)$	$(0.001, 0.01)$
355M GPT-2-style	$(0.02, 0.02)$	$(0.02, 0.0025)$	$(0.0003, 0.00125)$
0.6B Qwen3-style	$(0.04, 0.0025)$	$(0.02, 0.0025)$	$(0.0015, 0.000625)$

E.2 Loss curves

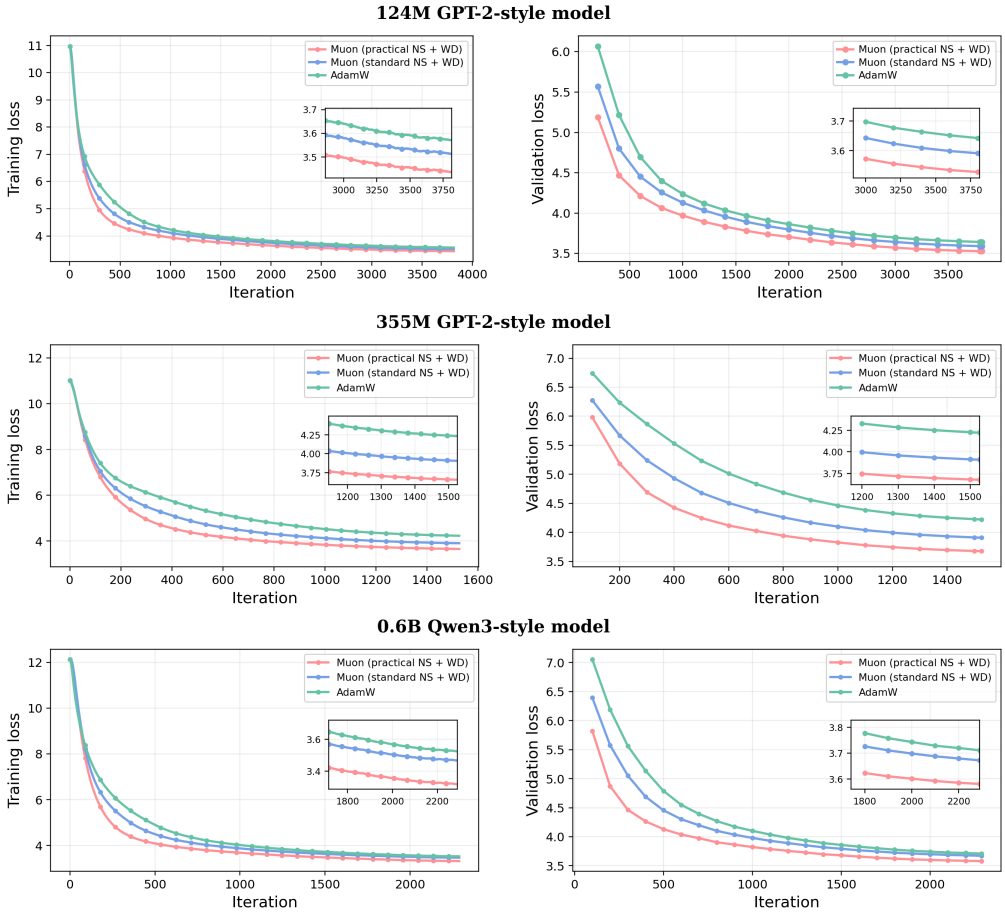

Figure 7: Loss curves across the three language models. Muon consistently achieves lower training and validation losses than AdamW.

E.3 ORTHOGONALIZATION ERROR

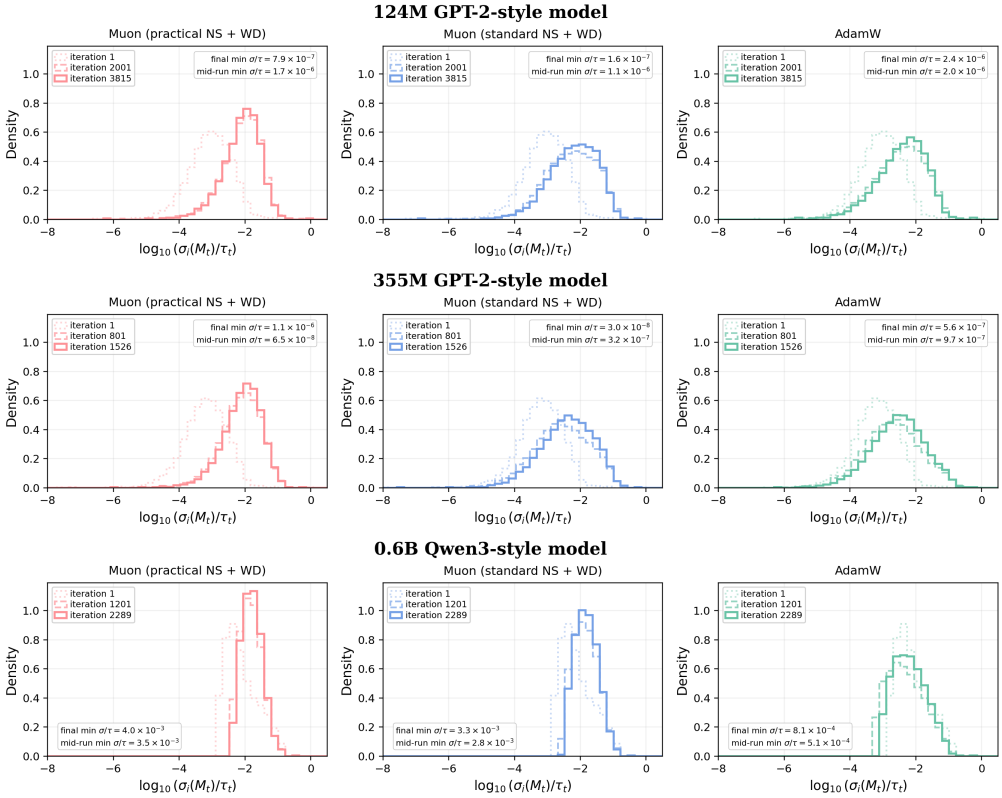

Figure 8: Momentum singular-value distributions for the query projection in the middle transformer block of the three language models. The distributions span a broad range and include very small normalized singular values.

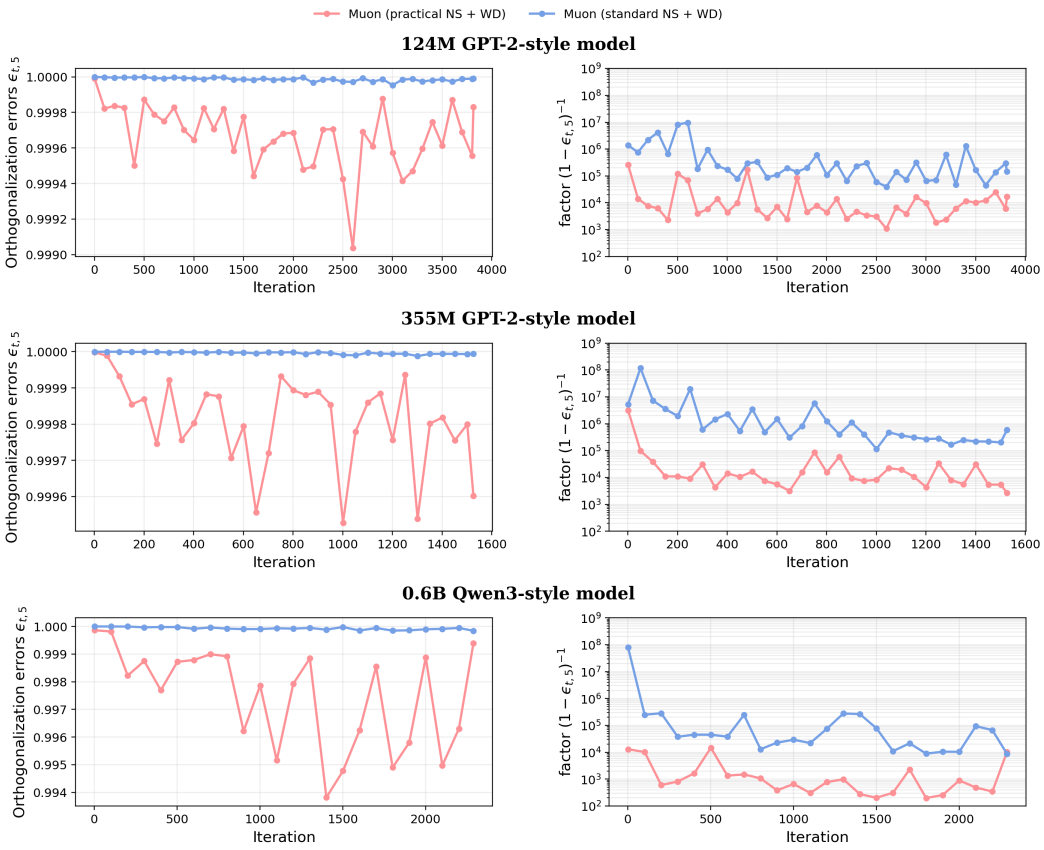

Figure 9: Orthogonalization errors $\epsilon_{t,5}$ and the corresponding factors $(1-\epsilon_{t,5})^{-1}$ for practical and standard NS across the three language models. Practical NS consistently exhibits smaller errors and factors, while the large factor values need not imply slow convergence.

E.4 HYPERPARAMETER SWEEP

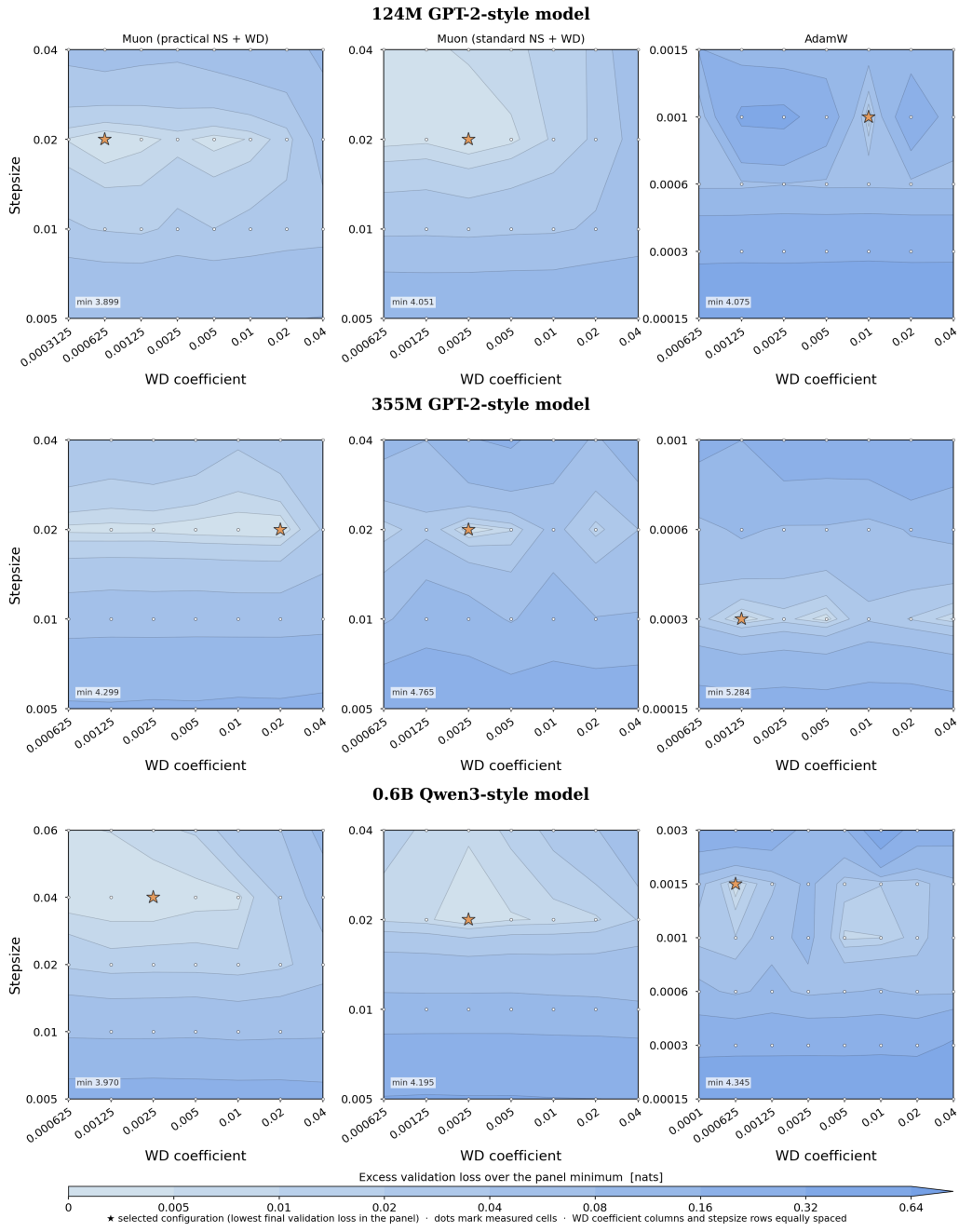

Figure 10: Joint sweep over stepsize and WD for language model training.

F INTERPRETING STANDARD NS WITH WD

As discussed in Section 3, Muon with standard NS admits a more direct optimization interpretation through its fixed-point condition, since $q_{\mathrm{S},5}$ is strictly increasing on $[0,1]$ and hence invertible.

Dynamic objective. Recall the scalar fixed-point condition $q_{\mathrm{S},5}(G_t/\tau_t)+\lambda W_t = 0$ from Section 3, where $q_{\mathrm{S},5}$ denotes $q_{\mathrm{S},5}^{(k)}$ for arbitrary $k \in \mathbb{N}$. Applying the odd inverse $q_{\mathrm{S},5}^{-1}$ gives

$$G_t + \tau_t q_{\mathrm{S},5}^{-1}(\lambda W_t) = 0.$$

We obtain a regularizer by integrating the second term on the left-hand side with respect to W. Extending this construction to matrices yields

$$\widetilde{\Phi}_t(W) = f(W) + \widetilde{\mathcal{R}}_t(W), \qquad \widetilde{\mathcal{R}}_t(W) = \frac{\tau_t}{\lambda} \sum_{i=1}^{r} \int_0^{\lambda\sigma_i(W)} \tilde{q}_{\mathrm{S},5}^{-1}(s)\,\mathrm{d}s. \tag{26}$$

Here, τ_t is held fixed when differentiating in W, and $\sigma_i(W)$ denotes the i-th singular value of W. The extended inverse is defined by

$$\tilde{q}_{\mathrm{S},5}^{-1}(s) = \begin{cases} q_{\mathrm{S},5}^{-1}(s), & 0 \le s \le 1, \\ s, & s > 1, \end{cases}$$

where $q_{\mathrm{S},5}^{-1} : [0,1] \to [0,1]$ denotes the inverse of $q_{\mathrm{S},5}$ restricted to $[0,1]$. For negative arguments, we use the odd extension. Alternatively, one may set $\tilde{q}_{\mathrm{S},5}^{-1}(s) = +\infty$ for $s > 1$, but we omit discussion of this case for brevity. The following proposition establishes the equivalence between the Muon update using standard NS and optimization of the dynamic objective $\widetilde{\Phi}_t(W)$.

Proposition 3. ***(i) Properties of the regularizer.***

1) The regularizer $\widetilde{\mathcal{R}}_t$ in (26) *is continuously differentiable and $\lambda\tau_t/C_1$-strongly convex.*
2) For an SVD $W = U_W \operatorname{diag}(\sigma_1(W), \dots, \sigma_r(W)) V_W^\top$,

$$\nabla\widetilde{\mathcal{R}}_t(W) = \tau_t U_W \operatorname{diag}\Big(\tilde{q}_{\mathrm{S},5}^{-1}(\lambda\sigma_1(W)), \dots, \tilde{q}_{\mathrm{S},5}^{-1}(\lambda\sigma_r(W))\Big) V_W^\top.$$

3) The following bounds hold:

$$\frac{\lambda\tau_t}{2C_1}\|W\|_F^2 \le \widetilde{\mathcal{R}}_t(W) \le \frac{\lambda\tau_t}{2}\|W\|_F^2, \qquad \|\nabla\widetilde{\mathcal{R}}_t(W)\|_F \le \lambda\tau_t\|W\|_F.$$

(ii) Fixed-point equivalence and descent alignment. *Ignoring the momentum by setting $\beta = 0$, Muon update in Algorithm 1 using standard NS has the following properties:*

1) The update direction and the stochastic gradient vanish simultaneously, i.e.,

$$H_t = 0 \quad \Longleftrightarrow \quad \widehat{\nabla}\widetilde{\Phi}_t(W_t) = 0, \qquad \widehat{\nabla}\widetilde{\Phi}_t(W_t) = G_t + \nabla\widetilde{\mathcal{R}}_t(W_t).$$

2) The update direction H_t is positively aligned with the stochastic gradient $\widehat{\nabla}\widetilde{\Phi}_t(W_t)$, i.e.,

$$\langle \widehat{\nabla}\widetilde{\Phi}_t(W_t), H_t \rangle_F \ge \frac{\tau_t}{C_1}\|H_t\|_F^2.$$

(iii) Exact generalized Frank–Wolfe equivalence. *With $\beta = 0$, and $\gamma = \eta\lambda \in (0,1]$, Muon update in Algorithm 1 using standard NS is exactly a stochastic generalized Frank–Wolfe step on the dynamic objective* (26)*, with stepsize γ.*

Specifically, Muon selects its update direction by seeking the greatest decrease in the regularized linear model, whose unique minimizer is

$$\operatorname*{arg\,min}_{W \in \mathbb{R}^{m\times n}} \Big\{ \langle G_t, W\rangle_F + \widetilde{\mathcal{R}}_t(W) \Big\} = -\frac{D_t}{\lambda}. \tag{27}$$

The Muon update therefore takes the equivalent form

$$W_{t+1} = (1-\gamma)W_t + \gamma\big(-\frac{D_t}{\lambda}\big), \tag{28}$$

which moves a fraction γ of the way from W_t to the unique minimizer $-D_t/\lambda$ of the regularized linear subproblem.

(iv) Recovery of the original objective. *Consider the deterministic setting where $G_t = \nabla f(W_t)$. If $\|\nabla f(W_t)\|_F \to 0$, then $\tau_t \to \varepsilon$. Together with the boundedness of $\{W_t\}$, which WD naturally promotes, the bounds in (i) imply that both $\widetilde{\mathcal{R}}_t(W_t)$ and $\|\nabla\widetilde{\mathcal{R}}_t(W_t)\|_F$ are asymptotically $\mathcal{O}(\varepsilon)$. Thus, $\widetilde{\Phi}_t$ approximates the original objective near stationarity.*

Proof. We first establish bounds on $\tilde{q}_{\mathrm{S},5}^{-1}$. The standard NS map satisfies

$$q_{\mathrm{S}}'(s) = q_{\mathrm{S}}'(0)(1-s^2)^k.$$

Indeed, let $\omega_\ell = (2\ell)!/[4^\ell(\ell!)^2]$. Differentiating $q_{\mathrm{S}}(s) = s\sum_{\ell=0}^k \omega_\ell(1-s^2)^\ell$ and using $(2\ell+1)\omega_\ell = 2(\ell+1)\omega_{\ell+1}$ gives

$$\begin{aligned} q_{\mathrm{S}}'(s) &= \sum\nolimits_{\ell=0}^k \omega_\ell(1-s^2)^\ell + \big(2(1-s^2)-2\big)\sum\nolimits_{\ell=0}^k \omega_\ell\ell(1-s^2)^{\ell-1} \\ &= \sum\nolimits_{\ell=0}^k (2\ell+1)\omega_\ell(1-s^2)^\ell - \sum\nolimits_{\ell=0}^{k-1} 2(\ell+1)\omega_{\ell+1}(1-s^2)^\ell \\ &= (2k+1)\omega_k(1-s^2)^k. \end{aligned}$$

Evaluating at $s=0$ yields $q_{\mathrm{S}}'(0) = (2k+1)\omega_k$, proving the identity. The chain rule then gives

$$0 < q_{\mathrm{S},5}'(s) = \prod\nolimits_{j=0}^4 q_{\mathrm{S}}'(q_{\mathrm{S},j}(s)) \le \big(q_{\mathrm{S}}'(0)\big)^5 = C_1, \qquad s\in[0,1),$$

where the last equality follows from $C_1 = a^5$ and, by direct differentiation of $q_{\mathrm{S}}(s) = s\,p_{\mathrm{S}}(s^2)$, $q_{\mathrm{S}}'(0) = p_{\mathrm{S}}(0) = a$. Thus, $q_{\mathrm{S},5}$ is strictly increasing and C_1-Lipschitz on $[0,1]$. Since $q_{\mathrm{S},5}(1) = 1$ and $C_1 \ge 1$ ($p_{\mathrm{S}}(0) = 1 + \sum_{\ell=1}^k \frac{(2\ell)!}{4^\ell(\ell!)^2} > 1$), its extension $\tilde{q}_{\mathrm{S},5}(s)$ preserves both properties on $[0,\infty)$, where $\tilde{q}_{\mathrm{S},5}(s) = q_{\mathrm{S},5}(s)$ for $s\in[0,1]$ and $\tilde{q}_{\mathrm{S},5}(s) = s$ for $s>1$. The inverse $\tilde{q}_{\mathrm{S},5}^{-1}$ is therefore continuous and satisfies $\tilde{q}_{\mathrm{S},5}^{-1}(0) = 0$. Recalling from (21) that $s \le q_{\mathrm{S},5}(s) \le C_1 s$ for $s\in[0,1]$, we have

$$\frac{u}{C_1} \le \tilde{q}_{\mathrm{S},5}^{-1}(u) \le u, \qquad u \ge 0. \tag{29}$$

For $u \ge v \ge 0$, we have

$$u - v = \tilde{q}_{\mathrm{S},5}(\tilde{q}_{\mathrm{S},5}^{-1}(u)) - \tilde{q}_{\mathrm{S},5}(\tilde{q}_{\mathrm{S},5}^{-1}(v)) \le C_1\big(\tilde{q}_{\mathrm{S},5}^{-1}(u) - \tilde{q}_{\mathrm{S},5}^{-1}(v)\big),$$

which implies

$$\tilde{q}_{\mathrm{S},5}^{-1}(u) - \tilde{q}_{\mathrm{S},5}^{-1}(v) \ge \frac{u-v}{C_1}. \tag{30}$$

(i) Properties of the regularizer. By definition,

$$\widetilde{\mathcal{R}}_t(W) - \frac{\lambda\tau_t}{2C_1}\|W\|_F^2 = \frac{\tau_t}{\lambda}\sum_{i=1}^r \int_0^{\lambda\sigma_i(W)} \left(\tilde{q}_{\mathrm{S},5}^{-1}(u) - \frac{u}{C_1}\right) du.$$

(29) and (30) imply that the integrand is nonnegative and nondecreasing, so the integral is nondecreasing and convex. By the convexity rule for singular-value functions, the displayed matrix function is convex. Hence $\widetilde{\mathcal{R}}_t$ is $(\lambda\tau_t/C_1)$-strongly convex.

Since $\tilde{q}_{\mathrm{S},5}^{-1}$ is continuous and $\tilde{q}_{\mathrm{S},5}^{-1}(0) = 0$, the differentiation rule for singular-value functions implies that $\widetilde{\mathcal{R}}_t$ is continuously differentiable. Let u_i and v_i denote the i-th columns of U_W and V_W, respectively. Holding τ_t fixed, the differentiation formula for singular-value functions (Lewis, 1995, Theorem 3.1) gives

$$\begin{aligned} \mathrm{d}\widetilde{\mathcal{R}}_t(W) &= \tau_t\sum_{i=1}^r \tilde{q}_{\mathrm{S},5}^{-1}(\lambda\sigma_i(W))\, u_i^\top(\mathrm{d}W)v_i \\ &= \Big\langle \tau_t\sum_{i=1}^r \tilde{q}_{\mathrm{S},5}^{-1}(\lambda\sigma_i(W))\, u_i v_i^\top,\, \mathrm{d}W\Big\rangle_F. \end{aligned}$$

This formula also holds at repeated or zero singular values. Therefore,

$$\nabla\widetilde{\mathcal{R}}_t(W) = \tau_t U_W \operatorname{diag}\Big(\tilde{q}_{\mathrm{S},5}^{-1}(\lambda\sigma_1(W)), \dots, \tilde{q}_{\mathrm{S},5}^{-1}(\lambda\sigma_r(W))\Big) V_W^\top.$$

Integrating (29) gives

$$\frac{\lambda^2\sigma_i(W)^2}{2C_1} \le \int_0^{\lambda\sigma_i(W)} \tilde{q}_{\mathrm{S},5}^{-1}(u)\, du \le \frac{\lambda^2\sigma_i(W)^2}{2}.$$

Multiplying by τ_t/λ and summing over i yields

$$\frac{\lambda\tau_t}{2C_1}\|W\|_F^2 \le \widetilde{\mathcal{R}}_t(W) \le \frac{\lambda\tau_t}{2}\|W\|_F^2.$$

Finally, the gradient formula and $\tilde{q}_{\mathrm{S},5}^{-1}(u) \le u$ imply

$$\|\nabla\widetilde{\mathcal{R}}_t(W)\|_F^2 = \tau_t^2 \sum_{i=1}^r \big[\tilde{q}_{\mathrm{S},5}^{-1}(\lambda\sigma_i(W))\big]^2 \le \lambda^2\tau_t^2\|W\|_F^2.$$

Taking square roots completes part (i).

(ii) Fixed-point equivalence and descent alignment. When $\beta = 0$, we have $M_t = G_t$. Recall that $s_{i,t} = \sigma_{i,t}/\tau_t \in (0, 1]$, and

$$G_t = U_t \operatorname{diag}(\sigma_{1,t}, \dots, \sigma_{r,t}) V_t^\top, \qquad D_t = U_t \operatorname{diag}\Big(q_{\mathrm{S},5}(s_{1,t}), \dots, q_{\mathrm{S},5}(s_{r,t})\Big) V_t^\top.$$

Since $\tilde{q}_{\mathrm{S},5}^{-1}(q_{\mathrm{S},5}(s)) = s$ for $s \in [0, 1]$, the gradient formula in (i) yields

$$\nabla\widetilde{\mathcal{R}}_t(-D_t/\lambda) = -\tau_t U_t \operatorname{diag}(s_{1,t}, \dots, s_{r,t}) V_t^\top = -U_t \operatorname{diag}(\sigma_{1,t}, \dots, \sigma_{r,t}) V_t^\top = -G_t. \tag{31}$$

Consequently,

$$\widehat{\nabla}\widetilde{\Phi}_t(W_t) = G_t + \nabla\widetilde{\mathcal{R}}_t(W_t) = \nabla\widetilde{\mathcal{R}}_t(W_t) - \nabla\widetilde{\mathcal{R}}_t(-D_t/\lambda).$$

By the $(\lambda\tau_t/C_1)$-strong convexity of $\widetilde{\mathcal{R}}_t$ a+nd $H_t = D_t + \lambda W_t$, we have

$$\begin{aligned}\big\langle \widehat{\nabla}\widetilde{\Phi}_t(W_t), H_t\big\rangle_F &= \lambda\big\langle \nabla\widetilde{\mathcal{R}}_t(W_t) - \nabla\widetilde{\mathcal{R}}_t(-D_t/\lambda), W_t + D_t/\lambda\big\rangle_F \\ &\ge \frac{\lambda^2\tau_t}{C_1}\|W_t + D_t/\lambda\|_F^2 = \frac{\tau_t}{C_1}\|H_t\|_F^2.\end{aligned}$$

If $H_t = 0$, then $W_t = -D_t/\lambda$, and hence

$$\widehat{\nabla}\widetilde{\Phi}_t(W_t) = \nabla\widetilde{\mathcal{R}}_t(W_t) - \nabla\widetilde{\mathcal{R}}_t(-D_t/\lambda) = 0.$$

Conversely, if $\widehat{\nabla}\widetilde{\Phi}_t(W_t) = 0$, then

$$0 = \big\langle \widehat{\nabla}\widetilde{\Phi}_t(W_t), H_t\big\rangle_F \ge \frac{\tau_t}{C_1}\|H_t\|_F^2.$$

Since $\tau_t/C_1 > 0$, this implies $H_t = 0$. Thus,

$$H_t = 0 \quad\Longleftrightarrow\quad \widehat{\nabla}\widetilde{\Phi}_t(W_t) = 0.$$

(iii) Exact generalized Frank–Wolfe equivalence. Fix an iteration t and a realization of G_t. Throughout this step, τ_t is held fixed. Since $\beta = 0$, (31) still holds. Then, by the $(\lambda\tau_t/C_1)$-strong convexity of $\widetilde{\mathcal{R}}_t$, for every $W \in \mathbb{R}^{m\times n}$,

$$\widetilde{\mathcal{R}}_t(W) \ge \widetilde{\mathcal{R}}_t\Big(\frac{-D_t}{\lambda}\Big) - \Big\langle G_t, W + \frac{D_t}{\lambda}\Big\rangle_F + \frac{\lambda\tau_t}{2C_1}\Big\|W + \frac{D_t}{\lambda}\Big\|_F^2.$$

Adding $\langle G_t, W\rangle_F$ to both sides gives

$$\langle G_t, W\rangle_F + \widetilde{\mathcal{R}}_t(W) \ge \Big\langle G_t, -\frac{D_t}{\lambda}\Big\rangle_F + \widetilde{\mathcal{R}}_t\Big(\frac{-D_t}{\lambda}\Big) + \frac{\lambda\tau_t}{2C_1}\Big\|W + \frac{D_t}{\lambda}\Big\|_F^2.$$

Since $\lambda\tau_t/C_1 > 0$, the last term is strictly positive whenever $W \ne -D_t/\lambda$. Thus $-D_t/\lambda$ is the unique minimizer, proving (27).

Finally, substituting $\eta = \gamma/\lambda$ into the Muon update gives

$$W_{t+1} = W_t - \eta(D_t + \lambda W_t) = W_t - \frac{\gamma}{\lambda}(D_t + \lambda W_t) = (1-\gamma)W_t + \gamma\big(-D_t/\lambda\big). \qquad \square$$