SPECTRAL FLATTENING IS NOT ENOUGH: WHEN MUON IMPROVES LEARNING RATES AND CONVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon updates matrix-valued parameters by flattening the singular values of each gradient while preserving its singular directions. Despite its practical success, the mechanisms governing its learning-rate capacity and convergence remain insufficiently understood. We analyze deterministic, full-batch Muon through its exact polar update. First, we derive a descent-preserving learning-rate threshold and show that it is controlled by the block-size-weighted average singular scale of the gradient matrices. This characterization makes the role of gradient magnitude explicit and reveals how objective scaling affects the admissible step size. We then analyze Muon's convergence behavior and identify the spectral properties that govern its convergence rate. Our analysis shows that flattening the spectrum within individual parameter blocks is not sufficient to determine global convergence, because scale differences across blocks remain influential. Together, these results provide a theoretical account of how spectral flattening governs the learning-rate capacity and convergence behavior of Muon.

1 INTRODUCTION

Training modern neural networks is fundamentally an optimization problem over a highly nonconvex landscape. As models grow deeper and wider, the loss surface contains many flat regions, sharp directions, saddle points, and local structures that can slow or destabilize training (Liu et al., 2022). Consequently, the choice of optimizer is not a minor implementation detail: it often determines whether a model can be trained efficiently at all. This has motivated a long line of practical optimizers, from Adam (Kingma and Ba, 2015) to AdamW (Loshchilov and Hutter, 2019b), that improve stability and convergence by changing how gradients are scaled, accumulated, or regularized.

Muon has recently emerged as a promising optimizer for matrix-valued parameters in neural networks (Jordan, 2024). Before updating a weight matrix, Muon applies Newton–Schulz iterations to approximate the polar factor of its gradient. If $G = U\Sigma V^\top$, this transformation replaces G by approximately $UV^\top$: it preserves the singular directions while flattening the singular values. This simple modification has produced strong empirical results in settings ranging from language-model pretraining to grokking (Liu et al., 2025; Tveit et al., 2025; Wen et al., 2026). In particular, Muon is often reported to support larger learning rates and to converge faster than SGD. However, Muon is not uniformly better than SGD. Recent controlled evidence shows that although Muon's full spectral flattening can improve stability at aggressive learning rates, its advantage depends on the update family and tuning, and less aggressive spectral transformations can outperform full flattening once training is stable (Qi et al., 2026). This mixed evidence suggests that larger learning rates and faster convergence are not unconditional consequences of orthogonalization. It instead motivates the central questions of this work: **under what conditions does Muon admit a larger learning rate than SGD, and under what conditions does it converge faster?**

Existing work characterizes Muon as steepest descent under a matrix norm, a non-Euclidean trust-region method, or an optimizer with spectral-norm-constrained updates (Bernstein and Newhouse, 2024; Li and Hong, 2025; Kovalev, 2025; Chen et al., 2025; Davis and Drusvyatskiy, 2025). These perspectives clarify the geometry of Muon's update, but they do not identify sufficient conditions under which this geometry yields either of the two advantages above. We close this gap by comparing

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.

Muon and SGD through two complementary analyses: one-step descent thresholds and convergence-rate bounds.

We first compare the largest learning rates for which a common Hessian upper bound guarantees a one-step decrease in the loss. For SGD, this descent threshold is $\eta^{\max}_{t,\mathrm{SGD}} = 2/\lambda^{H}_{\max}$. For exact-polar Muon, the ratio between the two thresholds is

$$\frac{\eta^{\max}_{t,\mathrm{Muon}}}{\eta^{\max}_{t,\mathrm{SGD}}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m_i}\sigma_j(G_{t,i})}{\sum_{i=1}^{k} m_i} =: \bar{\sigma}_t.$$

Therefore, $\bar{\sigma}_t > 1$ is a sufficient condition for Muon to admit a larger descent-preserving learning rate than SGD. The learning-rate comparison is thus determined by a directly measurable property of the gradient spectrum: the block-size-weighted mean singular value.

For the convergence question, we view Muon as a gradient-dependent preconditioned method and compare its effective curvature ratio with that of GD under a multi-block Kronecker curvature model. We derive a sufficient spectral condition under which Muon achieves a larger curvature ratio and hence a tighter contraction bound. The condition requires the relative reduction of the upper spectral endpoints of the representation-gradient factors to be at least as large as that of their lower endpoints across all parameter blocks. This clarifies both the source and the limit of the potential acceleration: flattening the spectrum within individual blocks does not necessarily improve the overall curvature ratio, because scale differences across blocks also matter.

We complement the analysis with controlled and extensive experiments designed to test these conditions. We measure the spectral quantities appearing in the learning-rate and convergence results, vary them in controlled settings, and compare the resulting predictions with observed stability and optimization speed. Across the evaluated settings, the experiments show that the derived conditions track when Muon can use a larger learning rate and when its convergence advantage over SGD appears. They also expose regimes in which the conditions weaken or fail, reinforcing that Muon's advantage is conditional rather than universal.

In summary, our contributions are twofold:

- **Conditional theory for Muon versus SGD.** We derive learning-rate and convergence guarantees for both optimizers and identify sufficient spectral conditions under which Muon admits a larger descent-preserving learning rate or a faster contraction bound than SGD.
- **Direct empirical validation of the conditions.** Through controlled and extensive experiments, we evaluate the quantities in our theory and test whether they predict the observed learning-rate capacity and convergence behavior of Muon relative to SGD.

2 Related Work

Optimizers for deep learning. The practical success of deep learning is closely tied to stable and efficient first-order optimization. Adam (Kingma and Ba, 2015) and AdamW (Loshchilov and Hutter, 2019b) improve robustness through momentum, coordinate-wise normalization, and decoupled regularization, although their benefits can depend strongly on task and tuning protocol (Wilson et al., 2017). In contrast, Muon (Jordan, 2024) represents an emerging class of matrix-level optimizers: rather than rescaling individual coordinates, it transforms each gradient matrix according to its singular-value structure.

Understanding Muon. Operationally, Muon approximates the polar factor of each gradient matrix using Newton–Schulz iterations (Jordan, 2024). Recent empirical work has shown that Muon can scale to language-model pretraining and related settings (Liu et al., 2025; Shah et al., 2025), with reported applications to grokking, latent-attention and MoE transformers, finetuning, and quantized optimizer states (Tveit et al., 2025; Mehta et al., 2025; Page et al., 2025; Gupta et al., 2025). Systematic benchmarking further places Muon among the fastest matrix-based optimizers, while showing that its gains depend on scale and tuning protocol (Wen et al., 2026). This empirical progress has also motivated variants such as blockwise Muon, AdaMuon, MuonMax, and NAMO (Boreiko et al., 2025; Si et al., 2025; Crawshaw et al., 2025; Zhang et al., 2026).

Theoretical work has started to explain Muon's geometry. Existing analyses connect Muon to steepest descent under matrix norms (Bernstein and Newhouse, 2024; Li and Hong, 2025), non-Euclidean

trust-region optimization (Kovalev, 2025), spectral-norm constrained optimization (Chen et al., 2025), and conditions under which spectral updates outperform Euclidean ones (Davis and Drusvyatskiy, 2025). These works clarify important aspects of Muon's update geometry, but they do not identify when that geometry translates into a larger admissible learning rate or faster convergence than SGD. We address this gap by deriving sufficient spectral conditions for both advantages and validating these conditions through controlled and extensive experiments.

3 PRELIMINARIES

To simplify the theoretical analysis, we consider deterministic, full-batch versions of both optimizers without momentum.

3.1 MUON OPTIMIZER

Muon (Jordan, 2024) is a matrix-valued optimizer. At iteration t, given the current weight $\boldsymbol{W}_{t-1}$, it computes the gradient $\boldsymbol{G}_t = \nabla\mathcal{L}(\boldsymbol{W}_{t-1})$, obtains the orthogonalized update $\boldsymbol{O}_t = \mathrm{NewtonSchulz}(\boldsymbol{G}_t)$, and updates $\boldsymbol{W}_t = \boldsymbol{W}_{t-1} - \eta\boldsymbol{O}_t$. where η is a learning rate. Here we note that Newton-Schulz is used to iteratively approximate $\left(\boldsymbol{G}_t\boldsymbol{G}_t^\top\right)^{-1/2}\boldsymbol{G}_t$. Specifically, let $\boldsymbol{G}_t = U\Sigma V^\top$ be the SVD of $\boldsymbol{G}_t$, we then have $\left(\boldsymbol{G}_t\boldsymbol{G}_t^\top\right)^{-1/2}\boldsymbol{G}_t = UV^\top$. Geometrically, $UV^\top$ is the closest orthogonal matrix to $\boldsymbol{G}_t$ in Frobenius norm, effectively replacing the singular values with ones while preserving the singular directions.

3.2 GRADIENT DESCENT OPTIMIZER

At iteration t, given the current weight $\boldsymbol{W}_{t-1}$, gradient descent computes $\boldsymbol{G}_t = \nabla\mathcal{L}(\boldsymbol{W}_{t-1})$ and updates $\boldsymbol{W}_t = \boldsymbol{W}_{t-1} - \eta\boldsymbol{G}_t$, where η is the learning rate.

4 THEORETICAL ANALYSIS

We first introduce the notation and formulate simultaneous Muon updates in vectorized multi-block form. We then analyze Muon's learning-rate capacity and convergence rate in comparison with gradient descent.

4.1 NOTATION

Joint parameters and gradients. Let $W_i \in \mathbb{R}^{m_i\times n_i}$, $i = 1,\dots,k$, denote the matrix parameter blocks. We define their dimensions and the joint parameter vector as follows:

$$d_i = m_i n_i, \qquad d = \sum_{i=1}^{k} d_i, \qquad \boldsymbol{W} = \begin{bmatrix}\mathrm{vec}(W_1)\\ \vdots \\ \mathrm{vec}(W_k)\end{bmatrix} \in \mathbb{R}^d, \tag{1}$$

where vec stacks matrix columns. At iteration t, the corresponding blocks and joint vector are denoted by $W_{t,i}$ and $\boldsymbol{W}_t$. All block gradients are evaluated at the same joint iterate:

$$G_{t,i} = \nabla_{W_i}\mathcal{L}(\boldsymbol{W}_{t-1}), \qquad \boldsymbol{g}_{t,i} = \mathrm{vec}(G_{t,i}), \qquad \boldsymbol{g}_t = \begin{bmatrix}\boldsymbol{g}_{t,1}\\ \vdots \\ \boldsymbol{g}_{t,k}\end{bmatrix} = \nabla\mathcal{L}(\boldsymbol{W}_{t-1}). \tag{2}$$

In particular, the joint gradient norm satisfies: $\|\boldsymbol{g}_t\|_2^2 = \sum_{i=1}^{k}\|G_{t,i}\|_F^2$. Assume that the minimum of the joint loss is attained at $\boldsymbol{W}^*$. We write the optimal value and the objective gap as follows:

$$\mathcal{L}^* = \mathcal{L}(\boldsymbol{W}^*), \qquad \Delta_t = \mathcal{L}(\boldsymbol{W}_t) - \mathcal{L}^*. \tag{3}$$

Full curvature and metric notation. For Hessian-based arguments, let $\Omega \subseteq \mathbb{R}^d$ be a convex region containing the relevant iterates and $\boldsymbol{W}^*$. Assume that $\mathcal{L}$ is twice continuously differentiable on an open neighborhood of Ω and that the update segments under consideration remain in Ω. We use the full Hessian, partitioned according to the joint parameter vector:

$$H(\boldsymbol{z}) = \nabla^2 \mathcal{L}(\boldsymbol{z}) = \left[H_{ij}(\boldsymbol{z})\right]_{i,j=1}^{k}, \qquad H_{ij}(\boldsymbol{z}) \in \mathbb{R}^{d_i \times d_j}. \tag{4}$$

For any positive-definite matrix M, the squared metric norm is defined as follows:

$$\|\boldsymbol{u}\|_M^2 = \boldsymbol{u}^\top M \boldsymbol{u}. \tag{5}$$

Muon preconditioning. The convergence analysis considers deterministic Muon with exact polar updates. For its preconditioner, assume $m_i \leq n_i$ and that each $G_{t,i}$ has full row rank at the steps under consideration. Tall blocks can be represented by transposing both the parameter matrix and its gradient. We define the gradient Gram matrices, polar factors, and preconditioner as follows:

$$\begin{aligned} B_{t,i} &= G_{t,i} G_{t,i}^\top \succ 0, \\ O_{t,i} &= B_{t,i}^{-1/2} G_{t,i} \in \mathbb{R}^{m_i \times n_i}, \\ \boldsymbol{P}_{t,i} &= I_{n_i} \otimes B_{t,i}^{-1/2} \in \mathbb{R}^{d_i \times d_i}, \\ \boldsymbol{P}_t &= \mathrm{blkdiag}(\boldsymbol{P}_{t,1}, \ldots, \boldsymbol{P}_{t,k}) \succ 0. \end{aligned} \tag{6}$$

Here $\mathrm{blkdiag}$ forms a block-diagonal matrix from its arguments, and $O_{t,i}$ is the exact polar factor of $G_{t,i}$, satisfying $O_{t,i} O_{t,i}^\top = I_{m_i}$. Stacking the vectorized polar factors gives:

$$\boldsymbol{o}_t = \begin{bmatrix} \mathrm{vec}(O_{t,1}) \\ \vdots \\ \mathrm{vec}(O_{t,k}) \end{bmatrix} = \boldsymbol{P}_t \boldsymbol{g}_t. \tag{7}$$

Thus, the simultaneous Muon update is:

$$\boldsymbol{W}_t = \boldsymbol{W}_{t-1} - \eta_t \boldsymbol{o}_t = \boldsymbol{W}_{t-1} - \eta_t \boldsymbol{P}_t \boldsymbol{g}_t. \tag{8}$$

4.2 Analysis on Learning Rate

Assuming $\mathcal{L}$ is C^2. By a second-order Taylor expansion, we have the quadratic upper bound

$$\mathcal{L}\left(\boldsymbol{W}_{t-1} + \Delta \boldsymbol{W}\right) \leq \mathcal{L}\left(\boldsymbol{W}_{t-1}\right) + \nabla \mathcal{L}\left(\boldsymbol{W}_{t-1}\right)^\top \Delta \boldsymbol{W} + \frac{\lambda_{\max}^H}{2} \|\Delta \boldsymbol{W}\|_2^2. \tag{9}$$

where $\lambda_{\max}^H := \lambda_{\max}(H)$ denotes the largest eigenvalue of the Hessian. Detailed derivation could be found in Appendix A.1

4.2.1 Theoretical Analysis for Gradient Descent

For Gradient Descent, we have $\Delta \boldsymbol{W} = -\eta \boldsymbol{g}_t$, leading to

$$\mathcal{L}\left(\boldsymbol{W}_{t-1} + \Delta \boldsymbol{W}\right) \leq \mathcal{L}\left(\boldsymbol{W}_{t-1}\right) - \eta \nabla \mathcal{L}\left(\boldsymbol{W}_{t-1}\right)^\top \boldsymbol{g}_t + \frac{\lambda_{\max}^H}{2} \eta^2 \|\boldsymbol{g}_t\|_2^2. \tag{10}$$

The update is efficient (i.e., $\mathcal{L}\left(\boldsymbol{W}_t\right) < \mathcal{L}\left(\boldsymbol{W}_{t-1}\right)$) if we have

$$\eta_t < \eta_t^{\max} := \frac{2}{\lambda_{\max}^H} \frac{\nabla \mathcal{L}\left(\boldsymbol{W}_{t-1}\right)^\top \boldsymbol{g}_t}{\|\boldsymbol{g}_t\|_2^2}. \tag{11}$$

Theorem 1. *For gradient descent with $\boldsymbol{g}_t = \nabla \mathcal{L}(\boldsymbol{W}_{t-1})$, the maximal learning rate in* (11) *reduces to $\eta_t^{\max} = \frac{2}{\lambda_{\max}^H}$.*

The proof is deferred to Appendix A.2.

4.2.2 THEORETICAL ANALYSIS FOR MUON

For the Muon optimizer, we have $\Delta \boldsymbol{W} = -\eta \boldsymbol{P}_t \boldsymbol{g}_t = -\eta \boldsymbol{o}_t$. Applying the quadratic bound (9) and simplifying yields

$$\mathcal{L}\left(\boldsymbol{W}_{t-1} + \Delta \boldsymbol{W}\right) \leq \mathcal{L}\left(\boldsymbol{W}_{t-1}\right) - \eta \sum_{i=1}^{k} \operatorname{tr}\left(G_{t,i}^{\top} O_{t,i}\right) + \frac{\lambda_{\max}^{H}}{2} \eta^2 \sum_{i=1}^{k} m_i, \tag{12}$$

where m_i denotes the row dimension of parameter block $W_i \in \mathbb{R}^{m_i \times n_i}$ with $m_i \leq n_i$. A formal derivation could be found in Appendix A.3

The update is efficient (i.e., $\mathcal{L}(\boldsymbol{W}_t) < \mathcal{L}(\boldsymbol{W}_{t-1})$) if we have:

$$\eta_t \leq \eta_t^{\max} := \frac{2}{\lambda_{\max}^{H}} \frac{\sum_{i=1}^{k} \operatorname{tr}\left(G_{t,i}^{\top} O_{t,i}\right)}{\sum_{i=1}^{k} m_i}. \tag{13}$$

Theorem 2. *For the Muon optimizer with k parameter matrix blocks, the maximum learning rate in* (13) *satisfies:*

$$\eta_t^{\max} = \frac{2}{\lambda_{max}^{H}} \frac{\sum_{i=1}^{k} \sum_{j=1}^{m_i} \sigma_j(G_{t,i})}{\sum_{i=1}^{k} m_i}, \tag{14}$$

where $\sigma_j(G_{t,i})$ denotes the j-th singular value of the gradient matrix block $G_{t,i}$.

The proof is deferred to Appendix A.4.

4.2.3 IMPLICATIONS ON LEARNING-RATE CAPACITY

Condition for a Larger Muon Learning Rate. Comparing Theorem 2 with the corresponding gradient-descent threshold $\eta_{t,\mathrm{GD}}^{\max} = 2/\lambda_{\max}^{H}$ gives

$$R_t := \frac{\eta_{t,\mathrm{Muon}}^{\max}}{\eta_{t,\mathrm{GD}}^{\max}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{m_i} \sigma_j(G_{t,i})}{\sum_{i=1}^{k} m_i} = \bar{\sigma}_{\mathrm{avg}}(G_t). \tag{15}$$

Therefore, under the same objective and curvature bound, Muon has a larger sufficient learning-rate threshold precisely when $R_t > 1$.

Why $R_t > 1$ Can Occur in Practice. A matrix gradient accumulated over N_t tokens satisfies $G_{t,i}^{\mathrm{sum}} = N_t G_{t,i}^{\mathrm{mean}}$. Since singular values scale linearly, $R_t^{\mathrm{sum}} = N_t R_t^{\mathrm{mean}}$. Thus, $R_t^{\mathrm{sum}} > 1 \quad \Leftrightarrow \quad R_t^{\mathrm{mean}} > \frac{1}{N_t}$. Because modern training often uses effective batches containing thousands of tokens, even a small mean per-token gradient scale can yield $R_t > 1$ under the summed-objective convention. Meanwhile, Muon's polar update is invariant to this gradient scaling, which explains why it can tolerate a larger scalar learning rate than GD in this regime. Consistent with this prediction, the learning-rate experiments in Section B show that Muon remains stable at larger learning rates. Furthermore, Section 5.1 directly measures the spectral ratio R_t and finds $R_t > 1$ at 99.8% of the recorded steps for every tested Muon learning rate under the summed-objective convention.

4.3 ANALYSIS ON CONVERGENCE RATE

We compare full-gradient GD and simultaneous Muon updates using the joint notation introduced above. The GD and Muon convergence results concern the full loss and retain all cross-block interactions. The block-diagonal surrogate is used only in the comparison of curvature ratios.

4.3.1 THEORETICAL ANALYSIS FOR CONVERGENCE RATE OF GRADIENT DESCENT

We first analyze full-gradient GD as a deterministic baseline. GD updates all blocks simultaneously:

$$W_{t,i} = W_{t-1,i} - \eta_t G_{t,i}, \qquad i = 1, \ldots, k, \tag{16}$$

or, equivalently:

$$\boldsymbol{W}_t = \boldsymbol{W}_{t-1} - \eta_t \boldsymbol{g}_t. \tag{17}$$

We impose the following assumptions on the joint loss:

- **A1 (Smoothness):** The loss is differentiable and its gradient is β-Lipschitz, with $\beta > 0$:

$$\|\nabla\mathcal{L}(\boldsymbol{u}) - \nabla\mathcal{L}(\boldsymbol{v})\|_2 \le \beta\|\boldsymbol{u} - \boldsymbol{v}\|_2, \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d. \tag{18}$$

- **A2 (Polyak-Lojasiewicz condition):** There exists $\alpha \in (0, \beta]$ such that:

$$\frac{1}{2}\|\nabla\mathcal{L}(\boldsymbol{W})\|_2^2 \ge \alpha\big(\mathcal{L}(\boldsymbol{W}) - \mathcal{L}^*\big), \qquad \forall \boldsymbol{W} \in \mathbb{R}^d. \tag{19}$$

Both assumptions concern the full parameter vector and therefore account for interactions among all parameter blocks.

Theorem 3. *Suppose that **A1** and **A2** hold. Then simultaneous full-gradient GD with $\eta_t = 1/\beta$ satisfies, for all $t \ge 0$:*

$$\mathcal{L}(\boldsymbol{W}_t) - \mathcal{L}^* \le \left(1 - \frac{\alpha}{\beta}\right)^t \big(\mathcal{L}(\boldsymbol{W}_0) - \mathcal{L}^*\big). \tag{20}$$

The proof is deferred to Appendix A.7.

The contraction bound is governed by α/β: a larger ratio gives a smaller contraction factor. Using the full Hessian in (4), we can relate these constants to curvature. If $\mathcal{L}$ is twice continuously differentiable on $\mathbb{R}^d$, a sufficient condition for **A1** and **A2** is the uniform bound:

$$\alpha I_d \preceq H(\boldsymbol{W}) \preceq \beta I_d, \qquad \forall \boldsymbol{W} \in \mathbb{R}^d. \tag{21}$$

In particular, for a strongly convex quadratic loss with constant Hessian $H \succ 0$, we may choose:

$$\alpha = \lambda_{\min}(H), \qquad \beta = \lambda_{\max}(H), \qquad \frac{\alpha}{\beta} = \frac{\lambda_{\min}(H)}{\lambda_{\max}(H)}. \tag{22}$$

These are eigenvalues of the full Hessian, including its off-diagonal blocks. For a general nonlinear loss, **A1** and **A2** must hold uniformly as stated; eigenvalues evaluated at a single iterate do not by themselves establish these assumptions.

4.3.2 Theoretical Analysis for Convergence Rate of Muon

Using the preconditioner in (6), the simultaneous exact polar updates and their joint vector form are:

$$\begin{aligned} W_{t,i} &= W_{t-1,i} - \eta_t B_{t,i}^{-1/2} G_{t,i}, \qquad i = 1, \dots, k, \\ \boldsymbol{W}_t &= \boldsymbol{W}_{t-1} - \eta_t \boldsymbol{P}_t \boldsymbol{g}_t. \end{aligned} \tag{23}$$

The vector form follows from $\mathrm{vec}(AXB) = (B^\top \otimes A)\,\mathrm{vec}(X)$; the derivation is provided in Appendix A.5. We first state the metric smoothness and PL inequalities used in the convergence argument.

Lemma 1 (Smoothness and PL inequality in a fixed metric)**.** *Fix $\boldsymbol{W} \in \Omega$ and $\boldsymbol{P} \succ 0$. Let $H = \nabla^2\mathcal{L}(\boldsymbol{W}) \succ 0$ be the current full Hessian, and define:*

$$\tilde{\alpha} = \lambda_{\min}(\boldsymbol{P}H), \qquad \tilde{\beta} = \lambda_{\max}(\boldsymbol{P}H). \tag{24}$$

Assume that $\mathcal{L}$ is $\tilde{\beta}$-smooth and $\tilde{\alpha}$-strongly convex on Ω with respect to the fixed norm $\|\cdot\|_{\boldsymbol{P}^{-1}}$. Then the following inequalities hold.

(i) *For every $\boldsymbol{W}' \in \Omega$, we have:*

$$\mathcal{L}(\boldsymbol{W}') \le \mathcal{L}(\boldsymbol{W}) + \nabla\mathcal{L}(\boldsymbol{W})^\top(\boldsymbol{W}' - \boldsymbol{W}) + \frac{\tilde{\beta}}{2}\|\boldsymbol{W}' - \boldsymbol{W}\|_{\boldsymbol{P}^{-1}}^2. \tag{25}$$

(ii) *The gradient satisfies:*

$$\frac{1}{2}\|\nabla\mathcal{L}(\boldsymbol{W})\|_{\boldsymbol{P}}^2 \ge \tilde{\alpha}\big(\mathcal{L}(\boldsymbol{W}) - \mathcal{L}^*\big). \tag{26}$$

The proof is deferred to Appendix A.6. The fixed-metric assumptions are automatic for a strongly convex quadratic loss. For a general nonlinear loss, they are additional conditions; the eigenvalues of the current Hessian alone do not establish them.

Theorem 4 (Convergence of simultaneous Muon updates)**.** *Suppose the hypotheses of Lemma 1 hold at every iterate with $\boldsymbol{P} = \boldsymbol{P}_t$ and $H = H_t := \nabla^2\mathcal{L}(\boldsymbol{W}_{t-1})$, and all iterates remain in Ω. Write $\tilde{\alpha}_t = \lambda_{\min}(\boldsymbol{P}_t H_t)$ and $\tilde{\beta}_t = \lambda_{\max}(\boldsymbol{P}_t H_t)$. With $\eta_t = 1/\tilde{\beta}_t$, we have:*

$$\mathcal{L}(\boldsymbol{W}_T) - \mathcal{L}^* \le \left[\prod_{t=1}^{T}\left(1 - \frac{\tilde{\alpha}_t}{\tilde{\beta}_t}\right)\right]\big(\mathcal{L}(\boldsymbol{W}_0) - \mathcal{L}^*\big), \qquad T \ge 1. \tag{27}$$

In particular, define:

$$\tilde{\alpha} := \inf_{t\ge1}\tilde{\alpha}_t, \qquad \tilde{\beta} := \sup_{t\ge1}\tilde{\beta}_t. \tag{28}$$

If $0 < \tilde{\alpha} \le \tilde{\beta} < \infty$, then:

$$\mathcal{L}(\boldsymbol{W}_T) - \mathcal{L}^* \le \left(1 - \frac{\tilde{\alpha}}{\tilde{\beta}}\right)^T\big(\mathcal{L}(\boldsymbol{W}_0) - \mathcal{L}^*\big). \tag{29}$$

$$\mathcal{L}(\boldsymbol{W}_T) - \mathcal{L}^* \le \left(1 - \frac{\tilde{\alpha}}{\tilde{\beta}}\right)^T\big(\mathcal{L}(\boldsymbol{W}_0) - \mathcal{L}^*\big). \tag{30}$$

The proof is deferred to Appendix A.8.

To compare the curvature ratios governing the preceding convergence bounds, we adopt a block-diagonal curvature model with full within-block matrices. This choice is motivated by reported Hessian block structure (Zhang et al., 2024; Dong et al., 2025) and practical diagonal preconditioning: curvature estimates in Sophia (Liu et al., 2024) and gradient second moments in AdamW (Loshchilov and Hutter, 2019a). These observations motivate the model without implying that inter-block couplings generally vanish.

Kronecker curvature model. Fix an iterate and suppress the iteration index. For a matrix block $W_i \in \mathbb{R}^{m_i\times n_i}$, let $x_i^{(r)} \in \mathbb{R}^{n_i}$ denote its input activation and $\delta_i^{(r)} = \nabla_{s_i^{(r)}}\ell_r \in \mathbb{R}^{m_i}$ denote the derivative of the w.r.t per-example loss to its preactivation $s_i^{(r)} = W_i x_i^{(r)}$. For $\mathcal{L} = \frac{1}{N}\sum_{r=1}^{N}\ell_r$, define:

$$X_i = \frac{1}{\sqrt{N}}\begin{bmatrix} x_i^{(1)} & \cdots & x_i^{(N)}\end{bmatrix}, \qquad D_i = \frac{1}{\sqrt{N}}\begin{bmatrix}\delta_i^{(1)} & \cdots & \delta_i^{(N)}\end{bmatrix}. \tag{31}$$

Thus $G_i := \nabla_{W_i}\mathcal{L} = D_i X_i^\top$. Following the activation and backpropagated-gradient factorization underlying K-FAC (Martens and Grosse, 2015), we use the empirical second moments from this same objective and define:

$$A_i := X_i X_i^\top, \qquad C_i := D_i D_i^\top, \qquad B_i := G_i G_i^\top. \tag{32}$$

Here C_i is the representation-gradient factor, whereas B_i determines the Muon preconditioner. We adopt the Kronecker curvature model:

$$H_{ii}^{\mathrm{bd}} := A_i \otimes C_i, \qquad H_{\mathrm{bd}} := \mathrm{blkdiag}\big(H_{11}^{\mathrm{bd}}, \ldots, H_{kk}^{\mathrm{bd}}\big). \tag{33}$$

These equalities specify the analytical model, rather than an identity for the full network Hessian or the exact Fisher matrix. For the following spectral analysis, assume $A_i \succ 0$, $C_i \succ 0$, and $B_i \succ 0$ for every i. With column vectorization, the ideal Muon preconditioner is:

$$\boldsymbol{P}_i = I_{n_i} \otimes B_i^{-1/2}, \qquad \boldsymbol{P} = \mathrm{blkdiag}(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_k). \tag{34}$$

Define the symmetric preconditioned factors and full model:

$$\widetilde{C}_i := B_i^{-1/4} C_i B_i^{-1/4}, \qquad S_{\mathrm{bd}} := \boldsymbol{P}^{1/2} H_{\mathrm{bd}} \boldsymbol{P}^{1/2}. \tag{35}$$

The spectral constants are:

$$\begin{aligned} \alpha_{\mathrm{bd}} &:= \lambda_{\min}(H_{\mathrm{bd}}), \quad \beta_{\mathrm{bd}} := \lambda_{\max}(H_{\mathrm{bd}}), \\ \widetilde{\alpha}_{\mathrm{bd}} &:= \lambda_{\min}(S_{\mathrm{bd}}), \quad \widetilde{\beta}_{\mathrm{bd}} := \lambda_{\max}(S_{\mathrm{bd}}). \end{aligned} \tag{36}$$

Theorem 5 (Curvature ratios under a multi-block Kronecker model)**.** *Under the definitions and positive-definiteness assumptions above, the following identities hold.*

(i) *The GD curvature ratio is:*

$$\frac{\alpha_{\mathrm{bd}}}{\beta_{\mathrm{bd}}} = \frac{\min_{1\le i\le k}\{\lambda_{\min}(A_i)\lambda_{\min}(C_i)\}}{\max_{1\le i\le k}\{\lambda_{\max}(A_i)\lambda_{\max}(C_i)\}}. \tag{37}$$

(ii) *The Muon-preconditioned curvature ratio is:*

$$\frac{\widetilde{\alpha}_{\mathrm{bd}}}{\widetilde{\beta}_{\mathrm{bd}}} = \frac{\min_{1\le i\le k}\left\{\lambda_{\min}(A_i)\lambda_{\min}(\widetilde{C}_i)\right\}}{\max_{1\le i\le k}\left\{\lambda_{\max}(A_i)\lambda_{\max}(\widetilde{C}_i)\right\}}. \tag{38}$$

Equivalently, the eigenvalues of $\widetilde{C}_i$ can be replaced by those of $B_i^{-1/2}C_i$, since these two matrices are similar.

The proof is deferred to Appendix A.10. Theorem 5 separates the representation-gradient factor C_i from the weight-gradient factor B_i. Without an additional relation between them, Muon transforms C_i into $B_i^{-1/4}C_iB_i^{-1/4}$, which need not equal $C_i^{1/2}$. We further measure the corresponding GD and Muon curvature ratios in (37) and (38) and find that the Muon ratio is consistently larger throughout training. In Appendix A.9, we additionally discuss theoretical conditions under which the Muon ratio is guaranteed to exceed that of GD.

5 Experiments

We examine two spectral predictions of our analysis: whether the learning-rate ratio R_t exceeds one and whether Muon has a larger curvature ratio than GD. Both diagnostics are evaluated along Muon training trajectories, with GD quantities computed at the same parameter states. In Appendix B, we further verify that Muon remains stable at larger learning rates and converges faster than SGD in practice.

General Experimental Setup. We use a 12-layer GPT with RMSNorm, model width 768, head dimension 128, and vocabulary size 50,304, trained on FineWeb10B for 500 optimizer steps. Each step processes 65,536 tokens with sequence length 1,024 and 16 sequences per microbatch. All runs use seed 0, and no activation Newton–Schulz transformation is applied. The spectral diagnostics use gradients of the summed cross-entropy objective. Experiment-specific learning rates and measurement protocols are described below.

5.1 Empirical Evaluation of the Spectral Learning-Rate Ratio

We evaluate the condition $R_t > 1$ underlying the learning-rate comparison in (15). We sweep the Muon learning rate over $\{0.05, 0.1, 0.2, 0.3, 0.4\}$ and compute R_t at every optimizer step. Figure 1 reports the raw values, their 15-step moving averages, and the fraction of steps satisfying $R_t > 1$.

For every learning rate, the first recorded step has $R_t = 0$, while all remaining 499 steps satisfy $R_t > 1$, corresponding to a fraction of 99.8%. These measurements support the spectral condition associated with a larger sufficient learning-rate bound for Muon along the evaluated trajectories. Since R_t depends on objective scaling, this observation applies to the summed-objective convention used here; it does not directly measure the largest stable learning rate.

5.2 Empirical Comparison of Curvature Ratios

We train a 12-block transformer with model width 768 and head dimension 128 on FineWeb10B for 500 optimizer steps using Muon with learning rate 0.2. Every 50 steps, we evaluate the relevant factors over 72 linear matrices from the attention and MLP layers using a fixed validation batch of 1024 tokens. For each matrix i, we compute $A_i = X_iX_i^\top, C_i = D_iD_i^\top, B_i = G_iG_i^\top$, where $G_i = D_iX_i^\top$, together with the Muon-preconditioned factor $\widetilde{C}_i = B_i^{-1/4}C_iB_i^{-1/4}$.

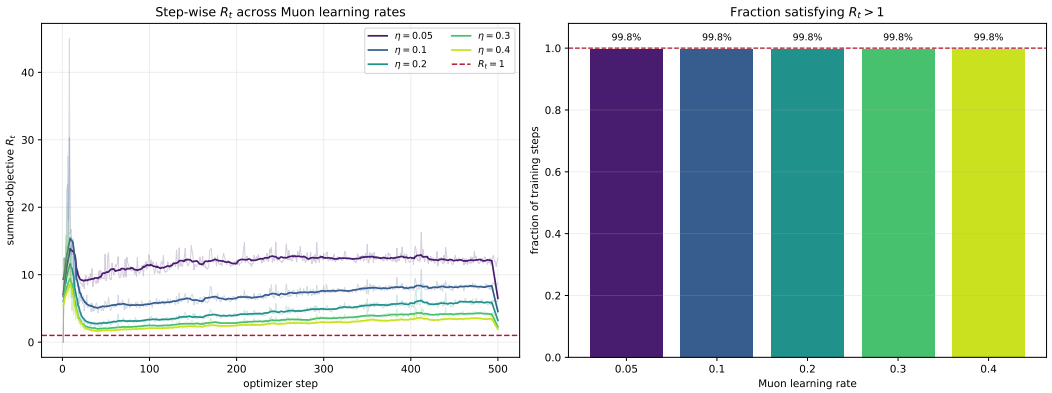

Figure 1: Spectral learning-rate ratio during GPT training. **Left:** Raw summed-objective $R_t = \sum_i \|G_{t,i}\|_* / \sum_i m_i$ (faint curves) and its 15-step moving average (solid curves) for five Muon learning rates; the dashed line marks $R_t = 1$. **Right:** Fraction of steps satisfying $R_t > 1$, equal to $499/500 = 99.8\%$ for every learning rate.

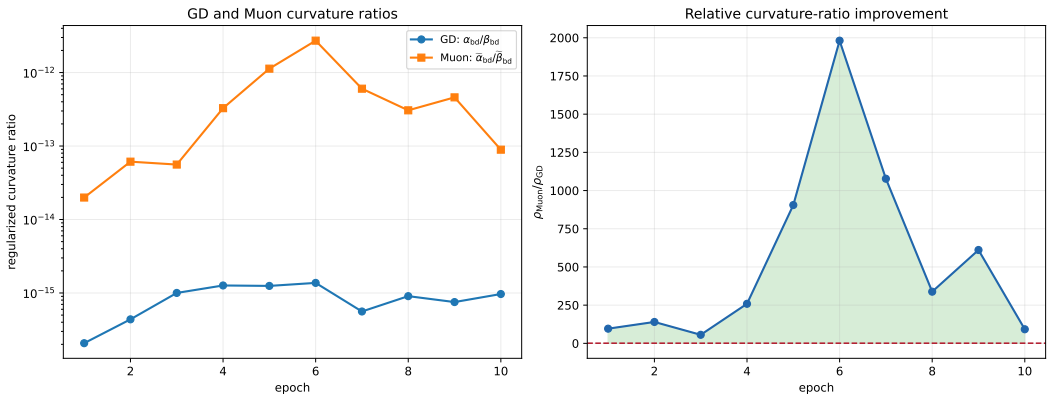

Figure 2: Regularized curvature ratios evaluated every 50 optimizer steps over 72 transformer matrices, using a relative eigenvalue floor of 10^{-6}. **Left:** GD and Muon-preconditioned ratios. **Right:** Their quotient, which exceeds one at all ten probes. The horizontal axis indexes probes, each separated by 50 optimizer steps.

We then measure the two curvature ratios

$$\rho_{\mathrm{GD}} = \frac{\min_i \lambda_{\min}(A_i)\lambda_{\min}(C_i)}{\max_i \lambda_{\max}(A_i)\lambda_{\max}(C_i)}, \qquad \rho_{\mathrm{Muon}} = \frac{\min_i \lambda_{\min}(A_i)\lambda_{\min}(\widetilde{C}_i)}{\max_i \lambda_{\max}(A_i)\lambda_{\max}(\widetilde{C}_i)}. \tag{39}$$

Figure 2 shows that ρ_{Muon} is consistently larger than ρ_{GD} at all 10 evaluation points. In particular, the ratio $\rho_{\mathrm{Muon}}/\rho_{\mathrm{GD}}$ remains above one throughout training and ranges from approximately 50 to 2000. This empirical behavior suggests that, in this training regime, Muon produces a substantially more favorable curvature ratio than GD. We emphasize that the exact positive-definiteness assumptions of our theoretical analysis are not satisfied numerically; hence, this experiment should be interpreted as a regularized empirical diagnostic rather than a direct verification of the theorem.

6 Conclusion

We analyze deterministic, full-batch Muon as a block-preconditioned gradient method. Our results show how gradient orthogonalization can increase the sufficient learning-rate bound and improve the curvature ratio governing convergence. Under explicit spectral conditions, these effects allow Muon to take larger steps and achieve a tighter contraction bound than gradient descent, connecting its update geometry to its optimization behavior.

AI USE STATEMENT

Generative AI tools were used to assist in checking theoretical arguments and revising portions of the manuscript, improving readability, and suggesting edits to experimental code. We take full responsibility for the content of this work, including all text, claims, proofs, code, and results.

REPRODUCIBILITY STATEMENT

We provide complete statements of the theoretical assumptions and results in the main text, with detailed proofs in the appendix. The experimental settings, model configurations, datasets, optimization hyperparameters, and evaluation procedures are documented in the paper and supplementary material. Upon acceptance, we will publicly release the complete source code required to reproduce all experiments and figures.

REFERENCES

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

Valentyn Boreiko, Zhiqi Bu, and Sheng Zha. Towards understanding orthogonalization in Muon. In *ICML Workshop on High-dimensional Learning Dynamics*, 2025.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

Michael Crawshaw, Chirag Modi, Mingrui Liu, and Robert M. Gower. An exploration of non-Euclidean gradient descent: Muon and its many variants. *arXiv preprint arXiv:2510.09827*, 2025.

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

Zhaorui Dong, Yushun Zhang, Jianfeng Yao, and Ruoyu Sun. Towards quantifying the hessian structure of neural networks, 2025. URL https://arxiv.org/abs/2505.02809.

Aman Gupta, Rafael Celente, Abhishek Shivanna, Daniel Thomas Braithwaite, Gregory Dexter, Shao Tang, Hiroto Udagawa, Daniel Silva, Rohan Ramanath, and Sathiya Keerthi. On quantizing the state of the Muon optimizer. *arXiv preprint arXiv:2509.23106*, 2025.

Keller Jordan. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024. Accessed: 2026-04-21.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2015.

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

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

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59: 85–116, 2022.

Hong Liu, Zhiyuan Li, David Leo Wright Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3xHDeA8Noi.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, 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*, 2019a. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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

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

Sushant Mehta, Raj Dandekar, Rajat Dandekar, and Sreedath Panat. Muon: Training and trade-offs with latent attention and MoE. *arXiv preprint arXiv:2509.24406*, 2025.

Saurabh Page, Advait Joshi, and S. S. Sonawane. Muonall: Muon variant for efficient finetuning of large language models. *arXiv preprint arXiv:2511.06086*, 2025.

Xianbiao Qi, Marco Chen, Jiaquan Ye, Yelin He, and Rong Xiao. Delving into muon and beyond: Deep analysis and extensions. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=xUQ0Gw11NL.

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.

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

Amund Tveit, Bjørn Remseth, and Arve Skogvold. Muon optimizer accelerates grokking. *arXiv preprint arXiv:2504.16041*, 2025.

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations*, 2026.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Minxin Zhang, Yuxuan Liu, and Hayden Schaeffer. Adam improves Muon: Adaptive moment estimation with orthogonalized momentum. *arXiv preprint arXiv:2602.17080*, 2026.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=X6rqEpbnj3.

A Mathematical Proofs

A.1 Derivation of Equation (9) (Quadratic Upper Bound).

Starting from a second-order Taylor expansion around $\boldsymbol{W}_{t-1}$ with perturbation $\Delta\boldsymbol{W}$:

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{W}_{t-1}+\Delta\boldsymbol{W}\right) &= \mathcal{L}\left(\boldsymbol{W}_{t-1}\right)+\nabla\mathcal{L}\left(\boldsymbol{W}_{t-1}\right)^{\top}\Delta\boldsymbol{W}+\frac{1}{2}\Delta\boldsymbol{W}^{\top}\nabla^{2}\mathcal{L}\left(\boldsymbol{W}_{t-1}\right)\Delta\boldsymbol{W}+o(\|\Delta\boldsymbol{W}\|_2^2)\\ &\approx \mathcal{L}\left(\boldsymbol{W}_{t-1}\right)+\nabla\mathcal{L}\left(\boldsymbol{W}_{t-1}\right)^{\top}\Delta\boldsymbol{W}+\frac{1}{2}\Delta\boldsymbol{W}^{\top}H\Delta\boldsymbol{W}\\ &\leq \mathcal{L}\left(\boldsymbol{W}_{t-1}\right)+\nabla\mathcal{L}\left(\boldsymbol{W}_{t-1}\right)^{\top}\Delta\boldsymbol{W}+\frac{\lambda_{\max}^{H}}{2}\|\Delta\boldsymbol{W}\|_2^2,\end{aligned}$$

where the first step is the exact second-order expansion, the second step approximates the Hessian by H, and the final inequality uses the Rayleigh bound $\Delta\boldsymbol{W}^{\top}H\Delta\boldsymbol{W}\leq\lambda_{\max}^{H}\|\Delta\boldsymbol{W}\|_2^2$.

A.2 Proof of Theorem 1

Proof. From (11), the maximal learning rate for gradient descent is

$$\eta_t^{\max}=\frac{2}{\lambda_{\max}^{H}}\frac{\nabla\mathcal{L}\left(\boldsymbol{W}_{t-1}\right)^{\top}\boldsymbol{g}_t}{\|\boldsymbol{g}_t\|_2^2}.$$

Since $\boldsymbol{g}_t=\nabla\mathcal{L}(\boldsymbol{W}_{t-1})$ by definition, the numerator simplifies to $\nabla\mathcal{L}(\boldsymbol{W}_{t-1})^{\top}\boldsymbol{g}_t=\|\boldsymbol{g}_t\|_2^2$, giving

$$\eta_t^{\max}=\frac{2}{\lambda_{\max}^{H}}\frac{\|\boldsymbol{g}_t\|_2^2}{\|\boldsymbol{g}_t\|_2^2}=\frac{2}{\lambda_{\max}^{H}}.$$

□

A.3 Derivation of Equation (12) (Muon Loss Bound).

Let $\Delta\boldsymbol{W}=-\eta\boldsymbol{o}_t$, where $\boldsymbol{o}_t=[\mathrm{vec}(O_{t,1})^{\top},\ldots,\mathrm{vec}(O_{t,k})^{\top}]^{\top}\in\mathbb{R}^{d}$. Substituting $\Delta\boldsymbol{W}$ into the quadratic upper bound yields:

$$\mathcal{L}\left(\boldsymbol{W}_{t-1}+\Delta\boldsymbol{W}\right)\leq\mathcal{L}\left(\boldsymbol{W}_{t-1}\right)-\eta\,\boldsymbol{g}_t^{\top}\boldsymbol{o}_t+\frac{\lambda_{\max}^{H}}{2}\eta^2\|\boldsymbol{o}_t\|_2^2, \tag{40}$$

where $\boldsymbol{g}_t=\nabla\mathcal{L}(\boldsymbol{W}_{t-1})=[\mathrm{vec}(G_{t,1})^{\top},\ldots,\mathrm{vec}(G_{t,k})^{\top}]^{\top}$.

We evaluate the linear and quadratic terms block-wise:

1. **Linear Term:** Using the Frobenius inner product identity $\mathrm{vec}(A)^{\top}\mathrm{vec}(B)=\mathrm{tr}(A^{\top}B)$,

$$\boldsymbol{g}_t^{\top}\boldsymbol{o}_t=\sum_{i=1}^{k}\mathrm{vec}(G_{t,i})^{\top}\mathrm{vec}(O_{t,i})=\sum_{i=1}^{k}\mathrm{tr}\left(G_{t,i}^{\top}O_{t,i}\right). \tag{41}$$

2. **Quadratic Term:** For each block i, $O_{t,i}\in\mathbb{R}^{m_i\times n_i}$ is an orthogonal polar factor satisfying $O_{t,i}O_{t,i}^{\top}=I_{m_i}$ (since $m_i\leq n_i$). Thus,

$$\|\boldsymbol{o}_t\|_2^2=\sum_{i=1}^{k}\|O_{t,i}\|_F^2=\sum_{i=1}^{k}\mathrm{tr}\left(O_{t,i}O_{t,i}^{\top}\right)=\sum_{i=1}^{k}\mathrm{tr}\left(I_{m_i}\right)=\sum_{i=1}^{k}m_i. \tag{42}$$

Combining the linear and quadratic terms yields Equation (12):

$$\mathcal{L}\left(\boldsymbol{W}_{t-1}+\Delta\boldsymbol{W}\right)\leq\mathcal{L}\left(\boldsymbol{W}_{t-1}\right)-\eta\sum_{i=1}^{k}\mathrm{tr}\left(G_{t,i}^{\top}O_{t,i}\right)+\frac{\lambda_{\max}^{H}}{2}\eta^2\sum_{i=1}^{k}m_i. \tag{43}$$

A.4 Proof of Theorem 2

Proof. Recall the maximum learning rate bound from Equation (13):

$$\eta_t^{\max} = \frac{2}{\lambda_{\max}^H} \frac{\sum_{i=1}^k \operatorname{tr}\left(G_{t,i}^\top O_{t,i}\right)}{\sum_{i=1}^k m_i}. \tag{44}$$

To evaluate $\operatorname{tr}(G_{t,i}^\top O_{t,i})$, let $G_{t,i} = U_i \Sigma_i V_i^\top$ be the thin Singular Value Decomposition (SVD) of block gradient $G_{t,i} \in \mathbb{R}^{m_i \times n_i}$, where $U_i \in \mathbb{R}^{m_i \times m_i}$ and $V_i \in \mathbb{R}^{n_i \times m_i}$ have orthonormal columns, and $\Sigma_i = \operatorname{diag}(\sigma_1(G_{t,i}), \ldots, \sigma_{m_i}(G_{t,i})) \in \mathbb{R}^{m_i \times m_i}$.

The exact polar update factor $O_{t,i}$ corresponds to the orthogonal factor in the polar decomposition of $G_{t,i}$:

$$O_{t,i} = B_{t,i}^{-1/2} G_{t,i} = \left(G_{t,i} G_{t,i}^\top\right)^{-1/2} G_{t,i} = U_i V_i^\top. \tag{45}$$

Substituting $G_{t,i} = U_i \Sigma_i V_i^\top$ and $O_{t,i} = U_i V_i^\top$ into the trace term:

$$\begin{aligned}
\operatorname{tr}\left(G_{t,i}^\top O_{t,i}\right) &= \operatorname{tr}\left((V_i \Sigma_i U_i^\top)(U_i V_i^\top)\right) && (46)\\
&= \operatorname{tr}\left(V_i \Sigma_i (U_i^\top U_i) V_i^\top\right) && (47)\\
&= \operatorname{tr}\left(\Sigma_i V_i^\top V_i\right) \quad \text{(by cyclic property of trace)} && (48)\\
&= \operatorname{tr}\left(\Sigma_i I_{m_i}\right) = \sum_{j=1}^{m_i} \sigma_j(G_{t,i}). && (49)
\end{aligned}$$

Substituting $\operatorname{tr}\left(G_{t,i}^\top O_{t,i}\right) = \sum_{j=1}^{m_i} \sigma_j(G_{t,i})$ into $\eta_t^{\max}$ yields:

$$\eta_t^{\max} = \frac{2}{\lambda_{\max}^H} \frac{\sum_{i=1}^k \sum_{j=1}^{m_i} \sigma_j(G_{t,i})}{\sum_{i=1}^k m_i}. \tag{50}$$

□

A.5 Explanation of the Vectorial Update for Muon

We have the update in the matrix form

$$\begin{aligned}
\boldsymbol{W}_t &= \boldsymbol{W}_{t-1} - \eta \left(G_t G_t^\top\right)^{-1/2} G_t \\
&= \boldsymbol{W}_{t-1} - \eta \left(G_t G_t^\top\right)^{-1/2} G_t \mathbb{I}_n.
\end{aligned}$$

Using the formula $\operatorname{vec}(ABC) = \left(C^\top \otimes A\right) \operatorname{vec}(B)$, we gain

$$\begin{aligned}
\boldsymbol{W}_t &= \boldsymbol{W}_{t-1} - \eta \left(\mathbb{I}_n \otimes \left(G_t G_t^\top\right)^{-1/2}\right) \operatorname{vec}(G_t) \\
&= \boldsymbol{W}_{t-1} - \eta \boldsymbol{P}_t \boldsymbol{g}_t.
\end{aligned}$$

A.6 Proof of Lemma 1

Proof. We now prove both (i) and (ii).

□

Proof. The spectral definitions are well posed because $\boldsymbol{P}H$ is similar to a symmetric positive definite matrix:

$$\boldsymbol{P}H = \boldsymbol{P}^{1/2}\left(\boldsymbol{P}^{1/2} H \boldsymbol{P}^{1/2}\right)\boldsymbol{P}^{-1/2}. \tag{51}$$

Hence $0 < \tilde{\alpha} \le \tilde{\beta} < \infty$.

The assumed smoothness and strong convexity imply, for any $\boldsymbol{U}, \boldsymbol{V} \in \Omega$:

$$\begin{aligned}
\frac{\tilde{\alpha}}{2}\|\boldsymbol{V} - \boldsymbol{U}\|_{\boldsymbol{P}^{-1}}^2 &\le \mathcal{L}(\boldsymbol{V}) - \mathcal{L}(\boldsymbol{U}) - \nabla\mathcal{L}(\boldsymbol{U})^\top(\boldsymbol{V} - \boldsymbol{U}) \\
&\le \frac{\tilde{\beta}}{2}\|\boldsymbol{V} - \boldsymbol{U}\|_{\boldsymbol{P}^{-1}}^2.
\end{aligned} \tag{52}$$

The upper inequality with $\boldsymbol{U} = \boldsymbol{W}$ and $\boldsymbol{V} = \boldsymbol{W}'$ proves (25).

For the PL inequality, set $\boldsymbol{g} = \nabla\mathcal{L}(\boldsymbol{W})$ and $\boldsymbol{v} = \boldsymbol{W} - \boldsymbol{W}^*$. The lower inequality with $\boldsymbol{U} = \boldsymbol{W}$ and $\boldsymbol{V} = \boldsymbol{W}^*$ gives:

$$\mathcal{L}(\boldsymbol{W}) - \mathcal{L}^* \leq \boldsymbol{g}^\top \boldsymbol{v} - \frac{\tilde{\alpha}}{2}\|\boldsymbol{v}\|^2_{\boldsymbol{P}^{-1}}. \tag{53}$$

Completing the square yields:

$$\begin{aligned} \boldsymbol{g}^\top \boldsymbol{v} - \frac{\tilde{\alpha}}{2}\|\boldsymbol{v}\|^2_{\boldsymbol{P}^{-1}} &= \frac{1}{2\tilde{\alpha}}\|\boldsymbol{g}\|^2_{\boldsymbol{P}} - \frac{\tilde{\alpha}}{2}\left\|\boldsymbol{v} - \frac{\boldsymbol{P}\boldsymbol{g}}{\tilde{\alpha}}\right\|^2_{\boldsymbol{P}^{-1}} \\ &\leq \frac{1}{2\tilde{\alpha}}\|\boldsymbol{g}\|^2_{\boldsymbol{P}}. \end{aligned} \tag{54}$$

Rearranging proves (26). □

A.7 Proof of Theorem 3

Proof. By **A1**,

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}_t) &\leq \mathcal{L}(\boldsymbol{w}_{t-1}) + \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle + \frac{\beta}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2^2 \\ &= \mathcal{L}(\boldsymbol{w}_{t-1}) - \eta_t\|\boldsymbol{g}_t\|_2^2 + \frac{\beta\eta_t^2}{2}\|\boldsymbol{g}_t\|_2^2 \\ &= \mathcal{L}(\boldsymbol{w}_{t-1}) - \eta_t\left(1 - \frac{\beta\eta_t}{2}\right)\|\boldsymbol{g}_t\|_2^2. \end{aligned}$$

Setting $\eta_t = 1/\beta$ yields

$$\mathcal{L}(\boldsymbol{w}_t) \leq \mathcal{L}(\boldsymbol{w}_{t-1}) - \frac{1}{2\beta}\|\boldsymbol{g}_t\|_2^2.$$

Since $\boldsymbol{g}_t = \nabla\mathcal{L}(\boldsymbol{w}_{t-1})$, **A2** implies

$$\mathcal{L}(\boldsymbol{w}_t) - \mathcal{L}^* \leq \left(1 - \frac{\alpha}{\beta}\right)\left(\mathcal{L}(\boldsymbol{w}_{t-1}) - \mathcal{L}^*\right).$$

Iterating this inequality proves

$$\mathcal{L}(\boldsymbol{w}_t) - \mathcal{L}^* \leq \left(1 - \frac{\alpha}{\beta}\right)^t\left(\mathcal{L}(\boldsymbol{w}_0) - \mathcal{L}^*\right).$$

□

A.8 Proof of Theorem 4

Proof. Let $\Delta_t = \mathcal{L}(\boldsymbol{W}_t) - \mathcal{L}^*$. The update rule gives:

$$\|\boldsymbol{W}_t - \boldsymbol{W}_{t-1}\|^2_{\boldsymbol{P}_t^{-1}} = \eta_t^2 \boldsymbol{g}_t^\top \boldsymbol{P}_t \boldsymbol{g}_t. \tag{55}$$

Applying Lemma 1(i) yields:

$$\mathcal{L}(\boldsymbol{W}_t) \leq \mathcal{L}(\boldsymbol{W}_{t-1}) - \eta_t\left(1 - \frac{\eta_t\tilde{\beta}_t}{2}\right)\|\boldsymbol{g}_t\|^2_{\boldsymbol{P}_t}. \tag{56}$$

Substituting $\eta_t = 1/\tilde{\beta}_t$ and applying Lemma 1(ii), we obtain:

$$\Delta_t \leq \Delta_{t-1} - \frac{1}{2\tilde{\beta}_t}\|\boldsymbol{g}_t\|^2_{\boldsymbol{P}_t} \leq \left(1 - \frac{\tilde{\alpha}_t}{\tilde{\beta}_t}\right)\Delta_{t-1}. \tag{57}$$

Iteration proves (27).

Under the uniform ratio bound, we have:

$$\prod_{t=1}^{T}\left(1 - \frac{\tilde{\alpha}_t}{\tilde{\beta}_t}\right) \leq \left(1 - \frac{\tilde{\alpha}}{\tilde{\beta}}\right)^T, \tag{58}$$

which proves (30). The matrices $\boldsymbol{P}_t$ and H_t may change between iterations; the metric is fixed within each one-step argument. □

A.9 SUFFICIENT CONDITION FOR CURVATURE-RATIO IMPROVEMENT

Corollary 1 (A sufficient condition for curvature-ratio improvement)**.** *Under the hypotheses of Theorem 5, suppose:*

$$\min_{1\le i\le k} \frac{\lambda_{\max}(C_i)}{\lambda_{\max}(\widetilde{C}_i)} \ge \max_{1\le i\le k} \frac{\lambda_{\min}(C_i)}{\lambda_{\min}(\widetilde{C}_i)}. \tag{59}$$

Then:

$$\frac{\widetilde{\alpha}_{\mathrm{bd}}}{\widetilde{\beta}_{\mathrm{bd}}} \ge \frac{\alpha_{\mathrm{bd}}}{\beta_{\mathrm{bd}}}, \tag{60}$$

with strict inequality whenever (59) *holds strictly. If, additionally, $C_i = B_i$ for every i, then* (59) *is equivalent to:*

$$\min_{1\le i\le k} \lambda_{\max}(G_i G_i^\top) \ge \max_{1\le i\le k} \lambda_{\min}(G_i G_i^\top). \tag{61}$$

The identity $C_i = B_i$ holds, in particular, if $X_i^\top X_i = I_N$ for every i. Strict inequality in (61) *is sufficient for strict improvement in this case.*

The proof is deferred to Appendix A.11.

Interpretation of the sufficient condition. The ratios in (59) measure the multiplicative changes in the largest and smallest eigenvalues of each representation-gradient factor after preconditioning. The condition requires the smallest reduction factor at the upper spectral endpoints to be no smaller than the largest reduction factor at the lower endpoints, even across different blocks. This ensures the global curvature ratio improves despite possible changes in the blocks attaining the extreme eigenvalues. When $C_i = B_i$, the condition reduces to a common intersection of the intervals $[\lambda_{\min}(B_i), \lambda_{\max}(B_i)]$. The condition is sufficient, rather than necessary.

Orthogonal-input special case. The condition $X_i^\top X_i = I_N$ is an orthogonality assumption on the scaled activation columns, not a consequence of normalizing individual activations. Together with $A_i = X_i X_i^\top \succ 0$, it forces $n_i = N$ and $A_i = I_{n_i}$, so X_i is square and orthogonal. In this special case, Theorem 5 gives the stronger identity:

$$\frac{\widetilde{\alpha}_{\mathrm{bd}}}{\widetilde{\beta}_{\mathrm{bd}}} = \sqrt{\frac{\alpha_{\mathrm{bd}}}{\beta_{\mathrm{bd}}}} \ge \frac{\alpha_{\mathrm{bd}}}{\beta_{\mathrm{bd}}}. \tag{62}$$

Thus no spectral-overlap assumption is needed in this special case; improvement is strict whenever $\alpha_{\mathrm{bd}} < \beta_{\mathrm{bd}}$. The derivation is included in Appendix A.11.

Implication for convergence bounds. For a quadratic objective with Hessian H_{bd} and fixed $\boldsymbol{P}$, the two ratios give valid contraction bounds with step sizes $1/\beta_{\mathrm{bd}}$ and $1/\widetilde{\beta}_{\mathrm{bd}}$, respectively. Consequently, strict improvement in (60) gives a smaller contraction upper bound for the simultaneous Muon-preconditioned step. For the original nonlinear loss, applying this comparison to Theorem 4 additionally requires the smoothness and PL conditions stated there to hold with the corresponding constants along the analyzed iterates. The spectral comparison alone does not establish those conditions.

A.10 PROOF OF THEOREM 5

Proof. Since A_i and C_i are symmetric positive definite, the extreme eigenvalues of their Kronecker product satisfy:

$$\begin{aligned} \lambda_{\min}(A_i \otimes C_i) &= \lambda_{\min}(A_i)\lambda_{\min}(C_i), \\ \lambda_{\max}(A_i \otimes C_i) &= \lambda_{\max}(A_i)\lambda_{\max}(C_i). \end{aligned} \tag{63}$$

The spectrum of a block-diagonal matrix is the union of its block spectra. Hence:

$$\begin{aligned} \alpha_{\mathrm{bd}} &= \min_i \left\{\lambda_{\min}(A_i)\lambda_{\min}(C_i)\right\}, \\ \beta_{\mathrm{bd}} &= \max_i \left\{\lambda_{\max}(A_i)\lambda_{\max}(C_i)\right\}. \end{aligned} \tag{64}$$

Taking the ratio proves (i). For (ii), the mixed-product identity gives:

$$\boldsymbol{P}_i H_{ii}^{\mathrm{bd}} = (I_{n_i} \otimes B_i^{-1/2})(A_i \otimes C_i) = A_i \otimes (B_i^{-1/2} C_i). \tag{65}$$

The factor $B_i^{-1/2} C_i$ need not be symmetric. It nevertheless has real positive eigenvalues because:

$$B_i^{-1/2} C_i = B_i^{-1/4} \widetilde{C}_i B_i^{1/4}, \qquad \widetilde{C}_i = B_i^{-1/4} C_i B_i^{-1/4} \succ 0. \tag{66}$$

Equivalently, applying the symmetric preconditioner gives:

$$\boldsymbol{P}_i^{1/2} H_{ii}^{\mathrm{bd}} \boldsymbol{P}_i^{1/2} = (I_{n_i} \otimes B_i^{-1/4})(A_i \otimes C_i)(I_{n_i} \otimes B_i^{-1/4}) = A_i \otimes \widetilde{C}_i. \tag{67}$$

Therefore:

$$S_{\mathrm{bd}} = \mathrm{blkdiag}\left(A_1 \otimes \widetilde{C}_1, \ldots, A_k \otimes \widetilde{C}_k\right). \tag{68}$$

Moreover, S_{bd} and $\boldsymbol{P} H_{\mathrm{bd}}$ have the same spectrum, since:

$$\boldsymbol{P} H_{\mathrm{bd}} = \boldsymbol{P}^{1/2} S_{\mathrm{bd}} \boldsymbol{P}^{-1/2}. \tag{69}$$

Applying the Kronecker and block-diagonal spectral identities again yields:

$$\begin{aligned} \widetilde{\alpha}_{\mathrm{bd}} &= \min_i \left\{ \lambda_{\min}(A_i) \lambda_{\min}(\widetilde{C}_i) \right\}, \\ \widetilde{\beta}_{\mathrm{bd}} &= \max_i \left\{ \lambda_{\max}(A_i) \lambda_{\max}(\widetilde{C}_i) \right\}. \end{aligned} \tag{70}$$

Taking the ratio proves (ii). No commutativity between B_i and C_i is required. □

A.11 Proof of Corollary 1

Proof. Choose indices attaining the extreme eigenvalues of the preconditioned model:

$$\begin{aligned} i^* &\in \arg\min_{1 \le i \le k} \left\{ \lambda_{\min}(A_i) \lambda_{\min}(\widetilde{C}_i) \right\}, \\ j^* &\in \arg\max_{1 \le j \le k} \left\{ \lambda_{\max}(A_j) \lambda_{\max}(\widetilde{C}_j) \right\}. \end{aligned} \tag{71}$$

By Theorem 5(i), we have:

$$\begin{aligned} \alpha_{\mathrm{bd}} &\le \lambda_{\min}(A_{i^*}) \lambda_{\min}(C_{i^*}), \\ \beta_{\mathrm{bd}} &\ge \lambda_{\max}(A_{j^*}) \lambda_{\max}(C_{j^*}). \end{aligned} \tag{72}$$

All quantities are positive, so:

$$\frac{\alpha_{\mathrm{bd}}}{\beta_{\mathrm{bd}}} \le \frac{\lambda_{\min}(A_{i^*}) \lambda_{\min}(C_{i^*})}{\lambda_{\max}(A_{j^*}) \lambda_{\max}(C_{j^*})} = \frac{\widetilde{\alpha}_{\mathrm{bd}}}{\widetilde{\beta}_{\mathrm{bd}}} \frac{\lambda_{\min}(C_{i^*})}{\lambda_{\min}(\widetilde{C}_{i^*})} \frac{\lambda_{\max}(\widetilde{C}_{j^*})}{\lambda_{\max}(C_{j^*})}. \tag{73}$$

Condition (59) implies:

$$\frac{\lambda_{\min}(C_{i^*})}{\lambda_{\min}(\widetilde{C}_{i^*})} \le \max_i \frac{\lambda_{\min}(C_i)}{\lambda_{\min}(\widetilde{C}_i)} \le \min_j \frac{\lambda_{\max}(C_j)}{\lambda_{\max}(\widetilde{C}_j)} \le \frac{\lambda_{\max}(C_{j^*})}{\lambda_{\max}(\widetilde{C}_{j^*})}. \tag{74}$$

Thus the product of the last two factors in (73) is at most one, which proves (60). If (59) is strict, that product is strictly less than one, so the ratio comparison is strict as well. Next, suppose $C_i = B_i$ for every i. Then:

$$\widetilde{C}_i = B_i^{-1/4} B_i B_i^{-1/4} = B_i^{1/2}. \tag{75}$$

The spectral mapping identity for the matrix square root gives:

$$\begin{aligned} \frac{\lambda_{\max}(C_i)}{\lambda_{\max}(\widetilde{C}_i)} &= \sqrt{\lambda_{\max}(B_i)}, \\ \frac{\lambda_{\min}(C_i)}{\lambda_{\min}(\widetilde{C}_i)} &= \sqrt{\lambda_{\min}(B_i)}. \end{aligned} \tag{76}$$

Consequently, condition (59) becomes:

$$\min_i \sqrt{\lambda_{\max}(B_i)} \ge \max_i \sqrt{\lambda_{\min}(B_i)}. \tag{77}$$

Since the square root is strictly increasing on positive numbers, this is equivalent to (61), with strict inequalities preserved. Finally, if $X_i^\top X_i = I_N$, then the exact gradient identity $G_i = D_i X_i^\top$ yields:

$$B_i = G_i G_i^\top = D_i X_i^\top X_i D_i^\top = D_i D_i^\top = C_i. \tag{78}$$

This verifies the stated sufficient input condition. □

Table 1: **NoNorm GPT learning-rate stability without momentum.** Diverged (3) means that the first non-finite loss or gradient was observed at optimizer step 3. Muon losses are reported after step 200.

	Muon		SGD	
Learning rate	Status	Validation loss	Status	Validation loss
3×10^{-4}	Complete	6.264	Diverged (3)	–
1×10^{-3}	Complete	6.061	Diverged (3)	–
3×10^{-3}	Complete	5.782	Diverged (3)	–
1×10^{-2}	Complete	5.551	Diverged (3)	–

Rank restriction and the orthogonal-input identity. For $X_i \in \mathbb{R}^{n_i \times N}$, the condition $X_i^\top X_i = I_N$ implies $\operatorname{rank}(X_i) = N \le n_i$. The standing assumption $A_i = X_i X_i^\top \succ 0$ implies $\operatorname{rank}(X_i) = n_i \le N$. Together they force $n_i = N$, so X_i is square and orthogonal and $A_i = I_{n_i}$. Using $C_i = B_i$, Theorem 5 then gives:

$$\frac{\alpha_{\mathrm{bd}}}{\beta_{\mathrm{bd}}} = \frac{\min_i \lambda_{\min}(B_i)}{\max_i \lambda_{\max}(B_i)}, \qquad \frac{\widetilde{\alpha}_{\mathrm{bd}}}{\widetilde{\beta}_{\mathrm{bd}}} = \frac{\min_i \sqrt{\lambda_{\min}(B_i)}}{\max_i \sqrt{\lambda_{\max}(B_i)}}. \tag{79}$$

The square root commutes with the finite minimum and maximum, which proves (62). Since $\sqrt{r} > r$ for $0 < r < 1$, the comparison is strict exactly when $\alpha_{\mathrm{bd}} < \beta_{\mathrm{bd}}$. If $n_i > N$ instead, A_i and H_{ii}^{bd} are singular; one cannot use the positive-definite theorem above on the full parameter space without modifying its statement.

B Learning-Rate Stability and Convergence Comparison

Model and data. We train a decoder-only GPT-style Transformer with 12 blocks, model dimension 768, six attention heads of dimension 128, MLP dimension 3072, and vocabulary size 50,304. Training examples are streamed from FineWeb10B with sequence length 1024. Each optimizer step contains 65,536 tokens (64 sequences) and is formed from four microbatches of 16 sequences. A fixed batch of 65,536 held-out tokens is used for validation. All models are trained in bfloat16 from the same initialization (seed 0) for at most 200 optimizer steps.

To expose optimizer instability directly, we use a *NoNorm* architecture: the two RMSNorm modules in every Transformer block and the final RMSNorm are replaced by the identity, and the query/key RMS normalization inside attention is disabled. Thus neither LayerNorm nor RMSNorm is active in this experiment. Projection matrices are initialized to zero, token embeddings are sampled from a unit normal distribution, and the remaining matrices use the repository's width-scaled initialization.

Optimizer isolation and loss convention. The target optimizer acts only on the 72 two-dimensional matrices inside the Transformer blocks:

$$12 \text{ blocks} \times \{W_Q, W_K, W_V, W_O, W_{\mathrm{up}}, W_{\mathrm{down}}\}.$$

The embedding table, vocabulary projection, and all vector-valued parameters are updated by the same reference AdamW optimizer in every run. Its learning rates are 0.3 for the embedding, $1/320$ for the vocabulary projection, and 0.01 for vector parameters, with $(\beta_1, \beta_2) = (0.8, 0.95)$, $\epsilon = 10^{-10}$, and zero weight decay.

For the target matrices, Muon and SGD both use zero momentum and zero weight decay. Muon uses 12 Newton–Schulz iterations and the standard rectangular matrix rescaling. We use no learning-rate schedule and no gradient clipping. We sweep

$$\eta \in \{3 \times 10^{-4},\ 10^{-3},\ 3 \times 10^{-3},\ 10^{-2}\}.$$

A run is declared divergent if the loss or target-gradient norm becomes non-finite, or if the reported mean loss exceeds 100.

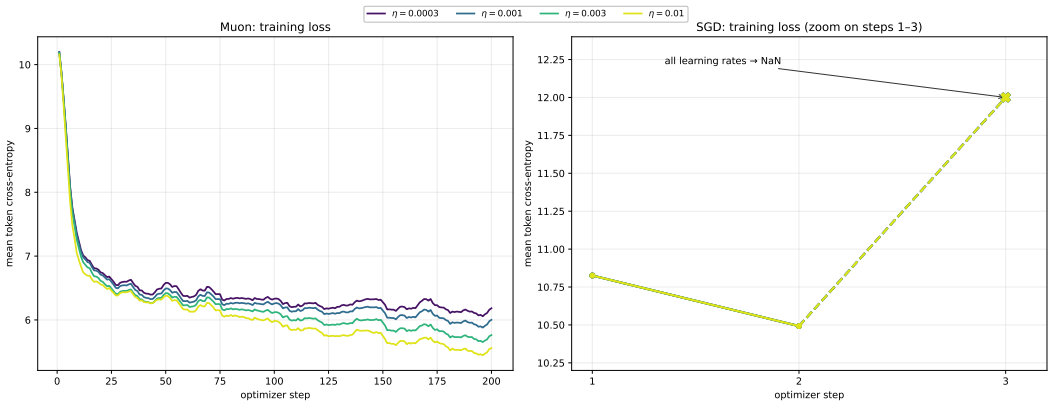

Figure 3: **Learning-rate stability for the NoNorm GPT experiment without momentum.** Left: Muon remains finite for all four learning rates and continues reducing the mean token cross-entropy through 200 steps. Right: the first three SGD steps are enlarged because every SGD run becomes non-finite at step 3. The dashed segment and cross denote the transition to NaN and are visual divergence markers rather than measured loss values. The four SGD curves coincide during the first two steps.

Results. Figure 3 and Table 1 show a clear separation between the two update rules. Muon completes all 200 steps at every tested learning rate, whereas SGD becomes non-finite at step 3 even at the smallest tested learning rate, 3×10^{-4}. Consequently, for this 200-step protocol the tested grid establishes

$$\eta_{\max}^{\mathrm{Muon}} \geq 10^{-2}, \qquad \eta_{\max}^{\mathrm{SGD}} < 3 \times 10^{-4},$$

and hence an observed stable-learning-rate separation greater than $10^{-2}/(3 \times 10^{-4}) \approx 33$. This is a lower bound on the separation, not an estimate of either exact stability boundary. As a longer-horizon check, Muon at $\eta = 3 \times 10^{-3}$ also completes 500 steps and reaches a validation loss of 5.251.

Within the stable Muon runs, increasing the learning rate improves optimization over the fixed budget: the final validation loss decreases monotonically from 6.264 at 3×10^{-4} to 5.551 at 10^{-2}. Thus the larger stable step sizes are useful rather than merely finite.

Normalized GPT Comparison at Small Learning Rates We additionally compare Muon and SGD using the repository's standard normalized GPT architecture. The model contains six Transformer blocks with model dimension 384, six attention heads of dimension 64, and an MLP dimension of 1536. We retain the architecture-standard RMSNorm layers, query/key RMS normalization, and the bounded-logit output map. Each optimizer step contains 8,192 tokens with sequence length 256, and every run is trained for 200 steps from the same initialization and with the same data stream.

The optimizer under evaluation updates only the two-dimensional attention and MLP weight matrices. A fixed AdamW reference optimizer updates the embedding, vocabulary projection, RMSNorm gains, biases, and all remaining parameters. Both Muon and SGD use zero momentum and zero weight decay on the target matrices. We use a constant learning rate, mean token cross-entropy, and no gradient clipping.

Figure 4 compares the three smallest learning rates in the sweep, $\eta \in \{0.03, 0.1, 0.3\}$. Muon and SGD are overlaid within each panel to allow a direct comparison under the same scalar learning rate.

All six runs complete the full 200-step training budget. At learning rates 0.03, 0.1, and 0.3, Muon obtains final validation losses of 6.866, 6.929, and 7.036, respectively. The corresponding SGD validation losses are 7.462, 7.426, and 7.390. Muon therefore improves the final validation loss by 0.596, 0.497, and 0.354 at the three learning rates.

The gap narrows as the shared learning rate increases, illustrating that an identical scalar learning rate does not imply an identical effective update magnitude for Muon and SGD. Moreover, because RMSNorm and the bounded-logit output map can keep the loss finite even when parameter scales

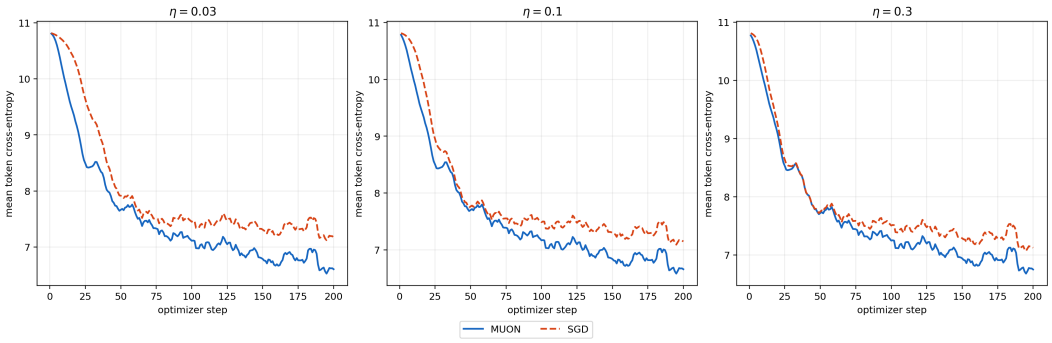

Figure 4: **Muon and SGD on a normalized GPT at the three smallest learning rates.** Training losses are shown as seven-step moving averages. Muon is represented by the solid blue curve and SGD by the dashed orange curve. Within each panel, both optimizers start from the same initialization and observe the same data stream. Both methods remain finite for all 200 steps, while Muon achieves a lower late-stage training loss at every displayed learning rate.

grow, these results should be interpreted as a comparison of finite-budget optimization quality rather than an estimate of the exact maximal stable learning rate.