

DGA-MUON: DECOUPLED GEOMETRY-ALIGNED ADAPTIVE SCALING FOR MUON

Wenpeng Zhang*

Independent Reasearcher

zhangwenpeng0@gmail.com

Runsheng Yu*

Independent Reasearcher

runshengyu@gmail.com

ABSTRACT

While NorMuon has achieved strong empirical performance in large-scale pre-training by enhancing Muon with row-wise adaptive scaling, its underlying adaptive mechanism remains largely heuristic and poorly understood. In this work, we provide the first systematic theoretical analysis of NorMuon’s adaptivity, revealing that it originates primarily from orthogonalization-induced geometry rather than genuine optimization-relevant information. Under exact orthogonalization, the adaptive scaling factors degenerate into a single global scalar for square and wide matrices, while for tall matrices their variation arises from the non-uniform distribution of row energy after orthogonalization. Under approximate orthogonalization, the resulting orthogonality residual introduces additional variation into the adaptive scaling factors, giving rise to a counterintuitive *Orthogonalization–Adaptivity Paradox*: more accurate orthogonalization weakens adaptivity, which vanishes entirely for square and wide matrices in the exact limit. We further show that NorMuon’s rigid row-wise scaling is geometrically misaligned with the one-sided orthogonal structure of tall matrices, i.e., it distorts the extracted orthogonal column directions. Based on the analysis of these limitations, we propose two core design principles that a desirable adaptive mechanism for Muon should satisfy. First, adaptive scaling should be decoupled from orthogonalization, with the scaling factors computed directly from raw gradients. Second, adaptive scaling should be aligned with the shape-dependent orthogonal structure of the polar factor, using row-wise scaling for wide matrices and column-wise scaling for tall matrices. We provide a formal proof that this geometry-aligned scaling preserves the orthogonal structure of the update. By additionally incorporating several other design considerations, including sum-based second-moment estimates, bias correction, and adaptive clipping of the scaling factors, we obtain the Decoupled Geometry-Aligned Muon (DGA-Muon) optimizer. We establish convergence guarantees for DGA-Muon and empirically validate both our theoretical characterization of NorMuon’s scaling degeneration and the superiority of DGA-Muon over NorMuon.

1 INTRODUCTION

The Muon (MomentUm Orthogonalized by Newton–Schulz) optimizer (Jordan, 2024), specifically designed for optimizing two-dimensional parameters in neural networks, first gained traction through its remarkable empirical success in the Modded-NanoGPT speedrun (Jordan et al., 2024), where it demonstrated substantially faster convergence than AdamW (Loshchilov & Hutter, 2019). Unlike AdamW, which treats weight matrices as flattened one-dimensional vectors, Muon preserves their inherent two-dimensional structure and applies an orthogonalization operation to equalize their singular values. Following this initial success, subsequent studies (Liu et al., 2025; Shah et al., 2025) further validated Muon’s advantages over AdamW in the pretraining of multi-billion-parameter language models. To date, Muon has been widely adopted for training frontier foundation models, including open-weight models such as Kimi (Team et al., 2025; 2026), GLM (Zeng et al., 2025; 2026), and DeepSeek (Xu et al., 2026a).

The remarkable success of Muon has sparked a surge of research interest. One line of research has sought to understand the mechanisms underlying Muon’s superior performance (Nguyen et al., 2026; Wang et al., 2026a;b) and to establish theoretical guarantees for its convergence (Li & Hong, 2025;

Kim & Oh, 2026). An even larger body of research is devoted to further enhancing the practical performance of Muon. One prominent branch of these efforts focuses on developing new methods that improve the accuracy or efficiency of the orthogonalization procedure (Amsel et al., 2026; Zhang et al., 2026a; Liu et al., 2026). Another promising direction explores adaptive preconditioning for Muon to better account for the geometry of the loss landscape. Conceptually, this mirrors how adaptive optimizers such as AdaGrad (Duchi et al., 2011) and Adam (Kingma & Ba, 2015) improved upon vanilla SGD. More fundamentally, a theoretical study (Su, 2025) suggests that the perfectly uniform spectral flattening induced by Muon may not be optimal, providing a principled foundation for incorporating adaptive preconditioning into Muon. While several initial attempts have been made, most of these methods either incur prohibitive computational or memory costs (e.g., requiring expensive matrix decompositions or maintaining full-matrix preconditioners) (Xu et al., 2026b; Zhang et al., 2026c; Song et al., 2026) or degrade to a simplistic single global scaling factor (Zhang et al., 2025a). Striking an effective balance between computational efficiency and adaptive expressiveness, NorMuon (Li et al., 2026) has emerged as the most widely adopted adaptive Muon variant. By employing a lightweight yet fine-grained neuron-wise (i.e., row-wise) scaling strategy, NorMuon consistently outperforms vanilla Muon. Notably, it achieves state-of-the-art performance on two leading speedruns: Modded-NanoGPT and NanoChat (Karpathy, 2025).

A distinctive characteristic of NorMuon is that its adaptive scaling factors are computed from the orthogonalized momentum, rather than directly from historical gradients as in conventional adaptive optimizers such as AdaGrad (Duchi et al., 2011), Adam (Kingma & Ba, 2015), and Shampoo (Gupta et al., 2018). This design tightly couples adaptivity with the orthogonalization procedure, constituting a fundamentally different paradigm for adaptive optimization. However, this design choice is largely heuristic, and the rationale for using orthogonalized momentum as the source of adaptivity lacks theoretical justification. Thus far, neither theoretical convergence guarantees for NorMuon nor a principled characterization of its adaptive behavior have been established, leaving this unconventional adaptive mechanism poorly understood.

In this work, we provide the first systematic theoretical analysis of NorMuon’s adaptivity. To understand its nature, we analyze the variation of NorMuon’s adaptive scaling factors, asking what information is encoded in their differences and what drives this variation. Surprisingly, we find that NorMuon’s adaptivity does not originate from conventional optimization dynamics, but instead arises primarily from the geometry induced by orthogonalization. To see this clearly, recall that NorMuon’s scaling factors are derived from the squared ℓ_2 -norms of the rows of the orthogonalized matrix. Although orthogonalization equalizes singular values, it does not necessarily imply uniform row-wise norms. In fact, under exact orthogonalization, this row-wise non-uniformity constitutes the primary source of variation in the scaling factors. However, such variation arises exclusively for tall matrices, where orthogonalization enforces only one-sided orthogonality, namely column orthonormality, but not row orthonormality. Consequently, while the total squared ℓ_2 -norm across all rows is fixed, the squared ℓ_2 -norm of each individual row remains unconstrained, allowing the row-wise scaling factors to vary (see Subsection 3.3 for a full derivation). In contrast, square and wide matrices satisfy row orthonormality, which ensures that the squared ℓ_2 -norm of every row is exactly one. Consequently, the row-wise adaptive scaling factors become identical across rows, and the adaptive scaling completely degenerates into a single global factor, thereby nullifying the intended row-wise adaptivity (see Subsection 3.1 for detailed analysis). In practical implementations, orthogonalization is typically approximate. The resulting orthogonality residual introduces an additional source of variation into the scaling factors. Notably, this gives rise to a rather counterintuitive and undesirable phenomenon, which we term the Orthogonalization–Adaptivity Paradox: the more accurate the orthogonalization, the weaker the resulting adaptivity, which vanishes entirely under exact orthogonalization for square and wide matrices; conversely, less accurate orthogonalization permits greater variation in the scaling factors and thus stronger adaptivity (see Subsections 3.2 and 3.3 for details). Together, these observations reveal a fundamental limitation of NorMuon: its adaptive scaling primarily acts as a compensatory mechanism that offsets geometric non-uniformity induced by orthogonalization, rather than capturing genuine optimization-relevant information.

The preceding analysis of NorMuon’s inherent limitation suggests that a desirable adaptive mechanism for Muon should preserve NorMuon’s computational efficiency while deriving adaptivity from optimization-relevant signals rather than orthogonalization-induced geometry. At its core, NorMuon’s limitation arises from computing scaling factors from the orthogonalized momentum, which inevitably causes the resulting adaptivity to inherit geometric distortions induced by orthogonaliza-

tion. To break this coupling, we introduce a decoupled design principle: computing adaptive scaling factors directly from raw gradients, independent of orthogonalization, and subsequently applying them to the orthogonalized matrix, i.e., the polar factor (Higham, 2008; 2020) of the momentum. Another design deficiency of NorMuon lies in its reliance on rigid row-wise scaling, which overlooks a fundamental geometric reality: for non-square matrices, the polar factor inherently exhibits only one-sided orthogonality. Specifically, for tall matrices, the polar factor is column-orthogonal, with its columns defining the orthogonal directions. Row-wise scaling, however, is a one-sided operation applied exclusively along the row dimension, which is misaligned with the column-wise orthogonal structure. As a result, this geometric mismatch distorts the column geometry and breaks the original orthogonality induced by the polar factor. To prevent this structural distortion, we introduce our second design principle: geometry-aligned scaling, which aligns the scaling with the one-sided orthogonal structure. Specifically, the scaling axis is chosen according to the shape-dependent orthogonality structure of the polar factor: row-wise scaling is applied for wide matrices, whereas column-wise scaling is applied for tall matrices (or either for square matrices), thereby preserving the orthogonal structure obtained through orthogonalization. We further provide a formal theoretical justification for the orthogonality-preserving property of the proposed adaptive scaling mechanism (see Subsection 4.2 for the proof).

Beyond the two core design principles, we incorporate several additional design considerations to better regulate the dynamics of the adaptive scaling factors. First, we construct the second-moment estimate for each adaptive unit (i.e., row or column) from the *sum*, rather than the *mean*, of the squared gradient entries, thereby making the scaling of each unit depend on its total accumulated gradient energy. This differs from NorMuon, which normalizes the total row-wise gradient energy by the row dimension, making its adaptive scaling depend on average per-coordinate energy. Second, we apply standard bias correction to the second-moment estimates to mitigate the initialization bias of the exponential moving average (EMA) during the early stages of optimization. Third, to curb the excessive anisotropy that adaptive scaling can introduce across different orthogonal directions, we employ an adaptive clipping mechanism that bounds the dynamic range of the scaling factors. This preserves Muon’s favorable ability to support large step sizes while maintaining training stability. Combining all these ingredients, we propose Decoupled Geometry-Aligned Muon (DGA-Muon; see Algorithm 1), which equips Muon with geometry-aligned adaptive scaling based on optimization-relevant gradient signals while achieving computational efficiency comparable to that of NorMuon.

We establish convergence guarantees for DGA-Muon under standard assumptions in the non-convex setting. We empirically verify our theoretical characterization of NorMuon, showing that removing its adaptive scaling on the square Q/K/V/O projection matrices in NanoChat leaves the validation performance essentially unchanged. We further demonstrate that DGA-Muon substantially outperforms NorMuon in an end-to-end NanoChat evaluation, validating the effectiveness of our proposed adaptive scaling mechanism. Finally, we note that our analysis and algorithmic design provide broader insights into general matrix-aware optimizer design beyond merely enhancing Muon or NorMuon. Our proposed principles suggest that adaptive mechanisms should decouple the extraction of optimization-relevant signals from matrix transformations and align adaptive scaling with the inherent geometric structure of matrix directions. Furthermore, our analysis highlights the intrinsic necessity of row-wise/column-wise adaptivity as a natural mechanism for adapting to the directional structure of matrices (characterized by either rows or columns) rather than viewing it merely as an approximation to element-wise adaptation.

2 PRELIMINARIES: MUON AND NORMUON

In this section, we briefly review Muon and NorMuon, providing the basis for our subsequent theoretical analysis.

Muon. Muon is an optimizer tailored for two-dimensional matrix parameters. It maintains a momentum buffer M_t , updated according to $M_t = \beta_1 M_{t-1} + (1 - \beta_1) G_t$, where G_t denotes the gradient at iteration t , $\beta_1 \in [0, 1)$ is the momentum coefficient, and $M_0 = \mathbf{0} \in \mathbb{R}^{m \times n}$. However, rather than updating the parameters directly along the raw momentum direction, Muon applies the update along the direction of the *orthogonalized* momentum. Specifically, given the current weight matrix $W_{t-1} \in \mathbb{R}^{m \times n}$ and the momentum $M_t \in \mathbb{R}^{m \times n}$, the update rule of Muon is given as follows

$$W_t = W_{t-1} - \eta \text{Orth}(M_t)$$

where η is the learning rate, and $\text{Orth}(M_t)$ denotes the orthogonalization of M_t , i.e., $\text{Orth}(M_t) = UV^\top$ with $M_t = U\Sigma V^\top$ being the (thin) singular value decomposition (SVD). The orthogonalization procedure discards all singular value information by setting every singular value to one, while preserving the directional structure induced by the singular vectors. The resulting orthogonalized matrix $UV^\top$ is commonly referred to as the *polar factor* (Higham, 2008; 2020) of M_t . In practice, however, Muon avoids the high cost of computing the full SVD by approximating the orthogonalization using efficient Newton–Schulz iterations (Jordan, 2024; Amsel et al., 2026).

NorMuon. NorMuon extends Muon by introducing row-wise adaptive scaling. In addition to the momentum buffer $M_t \in \mathbb{R}^{m \times n}$, it maintains a second-moment estimate $v_t \in \mathbb{R}^m$. Specifically, v_t tracks the row-wise mean square of the orthogonalized momentum. Rather than directly updating the parameters with the orthogonalized momentum, NorMuon first scales it by the inverse square root of v_t before applying the update. Formally, let $O_t = \text{Orth}(M_t)$ and $\text{mean}_{\text{cols}}(\cdot)$ denote the row-wise mean, i.e., the mean across columns for each row. Given the decay rate $\beta_2 \in [0, 1)$ and the initialization $v_0 = \mathbf{0} \in \mathbb{R}^m$, the second-moment estimate v_t is computed via an exponential moving average (EMA)

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) \text{mean}_{\text{cols}}(O_t \odot O_t),$$

where $\odot$ denotes the Hadamard (element-wise) product. The final parameter update then uses O_t scaled by the inverse square root of v_t

$$W_t = W_{t-1} - \eta_t \frac{O_t}{\sqrt{v_t} + \delta}.$$

Here, the division is broadcast row-wise, η_t is the learning rate, and $\delta > 0$ is a small constant for numerical stability. For clarity, we omit the weight decay and learning rate scaling used in the original NorMuon, as neither affects the subsequent analysis.

3 ON THE ORIGIN AND LIMITATIONS OF NORMUON’S ADAPTIVITY

Although NorMuon’s row-wise adaptive scaling has demonstrated strong empirical performance, its underlying theoretical foundation remains largely unexplored.¹ In this section, we theoretically analyze the origin of NorMuon’s adaptivity and uncover its inherent limitations. Specifically, we show that this adaptivity primarily arises from orthogonalization-induced geometry rather than optimization dynamics, and further demonstrate that such adaptivity can degenerate under certain conditions.

3.1 EXACT ORTHOGONALIZATION

We first consider the $m \leq n$ regime, where the matrix is square ($m = n$) or wide ($m < n$). To facilitate the subsequent analysis, we express the row-wise mean of the element-wise square $O_t \odot O_t$ in terms of the Gram matrix of O_t

$$\text{mean}_{\text{cols}}(O_t \odot O_t) = \frac{1}{n} \text{diag}(O_t O_t^\top),$$

where $\text{diag}(\cdot)$ denotes the operator that extracts the diagonal entries of a square matrix into a vector. Recall that the thin SVD of M_t is given by $M_t = U\Sigma V^\top$. For $m \leq n$, assuming that M_t has full row rank, we have $U \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{m \times m}$, and $V \in \mathbb{R}^{n \times m}$. By definition of the SVD, U is orthogonal ($UU^\top = U^\top U = I_m$) and V is column-orthonormal ($V^\top V = I_m$). Then, under exact orthogonalization, the identity $O_t = UV^\top$ holds exactly. Therefore, we obtain

$$O_t O_t^\top = (UV^\top)(UV^\top)^\top = U(V^\top V)U^\top = UU^\top = I_m,$$

i.e., O_t is row-orthonormal, and hence

$$\text{mean}_{\text{cols}}(O_t \odot O_t) = \frac{1}{n} \text{diag}(O_t O_t^\top) = \frac{1}{n} \mathbf{1}_m,$$

where $\mathbf{1}_m$ denotes the all-ones vector in $\mathbb{R}^m$. Substituting this identity into the EMA update rule for v_t , unrolling the recursion, evaluating the resulting geometric series, and using $v_0 = \mathbf{0}$, we obtain

$$v_t = \beta_2^t v_0 + (1 - \beta_2^t) \frac{1}{n} \mathbf{1}_m = (1 - \beta_2^t) \frac{1}{n} \mathbf{1}_m,$$

¹To the best of our knowledge, no theoretical convergence guarantee has yet been established for NorMuon.

which is uniform across rows, with only a global scalar varying over time. Furthermore, if the standard bias correction omitted in NorMuon were applied, it would cancel the time-dependent factor, yielding exactly $\frac{1}{n}\mathbf{1}_m$, which is functionally equivalent to a fixed global scalar throughout training. The resulting scaling vector $\frac{1}{\sqrt{v_t+\delta}}$ therefore assigns the same scaling coefficient to each row. Hence, the originally intended neuron-wise adaptive scaling completely collapses to a trivial global scalar, thereby rendering the adaptive mechanism fundamentally ineffective.

3.2 APPROXIMATE ORTHOGONALIZATION

The preceding analysis assumes exact orthogonalization. In practice, however, NorMuon employs approximate orthogonalization via a finite number of Newton–Schulz iterations (Li et al., 2026). We now analyze the approximate orthogonalization case. Assume that the approximately orthogonalized matrix O_t satisfies

$$O_t O_t^\top = I_m + E_t,$$

where $E_t \in \mathbb{R}^{m \times m}$ denotes the orthogonality residual induced by approximate orthogonalization. Then we have

$$\text{mean}_{\text{cols}}(O_t \odot O_t) = \frac{1}{n} \text{diag}(O_t O_t^\top) = \frac{1}{n} (\mathbf{1}_m + \text{diag}(E_t)).$$

Substituting this expression into the EMA update of v_t , unrolling the recursion, evaluating the resulting series, and using $v_0 = \mathbf{0}$, yields

$$v_t = (1 - \beta_2^t) \frac{1}{n} \mathbf{1}_m + \frac{1 - \beta_2}{n} \sum_{i=1}^t \beta_2^{t-i} \text{diag}(E_i).$$

For convenience, define

$$R_t = \frac{1 - \beta_2}{n} \sum_{i=1}^t \beta_2^{t-i} \text{diag}(E_i).$$

The term $(1 - \beta_2^t) \frac{1}{n} \mathbf{1}_m$ coincides exactly with the row-invariant scaling term derived in the exact orthogonalization case, whereas R_t captures the exponentially weighted accumulation of the orthogonality residuals and constitutes the only source of row-wise variation in v_t . To further characterize the magnitude of R_t , we assume that the orthogonality residual is uniformly bounded in spectral norm by some small constant $\epsilon > 0$, i.e., $\|E_t\|_2 \leq \epsilon$ for all t . This assumption is consistent with practical Newton–Schulz implementations, where the orthogonalization error is controlled within a prescribed tolerance. The spectral norm bound directly implies a bound on the diagonal entries of E_t . Specifically, for each $j \in \{1, \dots, m\}$, by the Cauchy–Schwarz inequality, the definition of the induced operator norm, and the fact that $\|\mathbf{e}_j\|_2 = 1$, we have

$$|(E_t)_{jj}| = |\mathbf{e}_j^\top E_t \mathbf{e}_j| \leq \|\mathbf{e}_j\|_2 \|E_t \mathbf{e}_j\|_2 \leq \|E_t\|_2 \leq \epsilon.$$

Consequently, we can bound R_t in the infinity norm as

$$\|R_t\|_\infty \leq \frac{1 - \beta_2}{n} \sum_{i=1}^t \beta_2^{t-i} \epsilon = (1 - \beta_2^t) \frac{\epsilon}{n} \leq \frac{\epsilon}{n}.$$

It immediately follows that the difference between any two entries of v_t is entirely determined by R_t . That is, for any $j, k \in \{1, \dots, m\}$, we have

$$|(v_t)_j - (v_t)_k| = |(R_t)_j - (R_t)_k| \leq 2\|R_t\|_\infty \leq \frac{2\epsilon}{n}.$$

This bounded variation in v_t directly translates into a bounded variation in the final adaptive scaling factors $\frac{1}{\sqrt{v_t+\delta}}$. Specifically, by the Mean Value Theorem, there exists some constant $C > 0$ such that, for any $j, k \in \{1, \dots, m\}$, we have

$$\left| \frac{1}{\sqrt{(v_t)_j + \delta}} - \frac{1}{\sqrt{(v_t)_k + \delta}} \right| \leq \frac{C\epsilon}{n}.$$

In practice, Newton–Schulz-based orthogonalization rapidly reduces the orthogonality residual as iterations proceed, typically resulting in a very small ϵ . Consequently, the adaptive scaling factors

exhibit only negligible variation across rows, rendering the row-wise adaptivity practically ineffective.

In short, the above analysis reveals that, in the $m \leq n$ regime, the apparent row-wise adaptivity in NorMuon is not driven by the underlying gradient statistics or curvature information that adaptive optimizers are typically designed to capture. Instead, the remaining row-wise variation in the adaptive scaling factors is solely induced by the orthogonalization residual. In other words, the observed adaptivity of NorMuon is merely a byproduct of inexact orthogonalization. As $\epsilon \rightarrow 0$, this residual-induced adaptivity vanishes, and the behavior of approximate orthogonalization converges to that of exact orthogonalization, recovering uniform, non-adaptive row-wise scaling.

Remark. A counterintuitive implication of the preceding analysis, which we term the *Orthogonalization–Adaptivity Paradox*, is that the adaptivity of NorMuon is inversely coupled with the quality of orthogonalization. The more accurate the orthogonalization, the weaker the adaptivity becomes; in the limit of exact orthogonalization, the adaptivity disappears entirely. Conversely, less accurate orthogonalization permits greater variation in the row-wise scaling factors and thus allows for stronger adaptivity. This suggests that NorMuon’s adaptive mechanism essentially serves to compensate for imperfections in orthogonalization, rather than providing genuine optimization-driven adaptivity. This stands in stark contrast to the design philosophy of Muon, where accurate orthogonalization is the cornerstone of the optimizer. Consequently, NorMuon’s adaptive mechanism is intrinsically incompatible with ongoing research efforts to improve Muon through increasingly accurate orthogonalization (Grishina et al., 2025; He et al., 2025; Amsel et al., 2026).

Note that the above analysis relies on the full row rank assumption. In fact, this already covers the generic case encountered in practice. Rank deficiency is an undesirable edge case that can hinder effective training and impair the model’s representational capacity, and is thus generally avoided in practice. Since it rarely occurs, we do not treat it separately here or in the subsequent analysis. We do emphasize, however, that even under rank deficiency, the row-wise scaling of NorMuon remains entirely determined by the orthogonalization procedure rather than by the optimization dynamics. A more detailed discussion of this case will be provided in an extended version of this work.

3.3 THE $m > n$ REGIME

In the $m > n$ regime, where the matrix is tall, the situation is fundamentally different. The row-orthonormality property $O_t O_t^\top = I_m$ no longer holds due to the dimensional constraint $m > n$. Instead, under exact orthogonalization, the column-orthonormality property $O_t^\top O_t = I_n$ holds. This implies that every column of O_t has unit norm, i.e., $\|(O_t)_{:,j}\|_2^2 = 1$ for each $j \in \{1, \dots, n\}$. In contrast, since $O_t O_t^\top \neq I_m$, the squared norm of each row of O_t is no longer constrained to 1. Nevertheless, the aggregate row energy is constrained to be fixed. Specifically, we have

$$\sum_{i=1}^m \|(O_t)_{i,:}\|_2^2 = \sum_{i=1}^m \sum_{j=1}^n (O_t)_{ij}^2 = \sum_{j=1}^n \sum_{i=1}^m (O_t)_{ij}^2 = \sum_{j=1}^n \|(O_t)_{:,j}\|_2^2 = n,$$

where the last equality follows from the column-orthonormality of O_t . Since there is no row-orthonormality constraint, this fixed aggregate energy is no longer forced to be uniformly distributed across the m rows. Consequently, the squared norm of each row can deviate substantially from the uniform allocation n/m , inducing row-wise variation in $\text{mean}_{\text{cols}}(O_t \odot O_t)$ and hence in the resulting adaptive scaling factors. This indicates that, in the $m > n$ regime, NorMuon’s adaptivity is not driven by conventional optimization dynamics underlying most existing adaptive optimizers, but instead stems from the intrinsic uneven row-energy distribution induced by tall-matrix orthogonalization.

The above analysis focuses on exact orthogonalization. For approximate orthogonalization, the analysis largely follows that of the $m \leq n$ regime, where deviations from exact orthogonality further introduce residual-induced variation in the adaptive scaling factors. This residual effect also diminishes as the orthogonalization accuracy improves, following the same Orthogonalization–Adaptivity Paradox.

Summary. Our analysis in the three preceding subsections reveals that NorMuon’s adaptivity is not optimization-driven in the conventional sense. Instead, it primarily serves as a compensation

Algorithm 1 Decoupled Geometry-Aligned Adaptive Muon (**DGA-Muon**)

-
- 1: **Input:** Step size $\eta > 0$, decay rates $\beta_1, \beta_2 \in [0, 1]$, stabilizer $\delta > 0$, clip ratios $c \in (0, 1], d \geq 1$
 - 2: **Initialize:** $W_1 \in \mathbb{R}^{m \times n}$, $M_0 = \mathbf{0} \in \mathbb{R}^{m \times n}$, $v_0 = \mathbf{0} \in \mathbb{R}^{\min(m, n)}$
 - 3: **for** $t = 1, \dots, T$ **do**
 - 4: Compute stochastic gradient G_t
 - 5: Update first moment estimate $M_t = \beta_1 M_{t-1} + (1 - \beta_1) G_t$
 - 6: Update second moment estimate from raw gradients

$$v_t = \begin{cases} \beta_2 v_{t-1} + (1 - \beta_2) \text{diag}(G_t G_t^\top), & m \leq n, \\ \beta_2 v_{t-1} + (1 - \beta_2) \text{diag}(G_t^\top G_t), & m > n. \end{cases}$$
 - 7: Apply bias correction $\hat{v}_t = v_t / (1 - \beta_2^t)$
 - 8: Compute adaptive scaling factors $s_t = 1 / (\sqrt{\hat{v}_t} + \delta)$
 - 9: Compute median of scaling factors $s_{\text{med}} = \text{median}(s_t)$
 - 10: Clip scaling factors $s_t \leftarrow \text{clip}(s_t, c \cdot s_{\text{med}}, d \cdot s_{\text{med}})$
 - 11: Orthogonalize first moment $O_t = \text{Orth}(M_t)$
 - 12: Apply geometry-aligned adaptive scaling

$$\tilde{O}_t = \begin{cases} \text{Diag}(s_t) O_t, & m \leq n, \\ O_t \text{Diag}(s_t), & m > n. \end{cases}$$
 - 13: Update parameters $W_{t+1} = W_t - \eta \tilde{O}_t$
 - 14: **end for**
-

mechanism for the non-uniformity induced by orthogonalization, including residual errors and uneven row-energy distribution. Although this compensation can provide empirical benefits, the resulting adaptivity is largely an orthogonalization-induced byproduct rather than a principled adaptive mechanism. Furthermore, approximate orthogonalization gives rise to the Orthogonalization-Adaptivity Paradox, revealing an inherent limitation of this form of adaptation. These findings motivate the design of a more principled adaptive mechanism tailored specifically to Muon.

4 DECOUPLED GEOMETRY-ALIGNED ADAPTIVE SCALING

In this section, we present a novel adaptive scaling mechanism specifically designed for Muon. We first outline the core design principles and then formally justify the orthogonality-preserving property of the resulting scaling strategy. We further discuss additional design considerations for constructing the adaptive scaling factors. The complete optimization procedure is summarized in Algorithm 1².

4.1 DESIGN PRINCIPLES: DECOUPLED ADAPTIVITY AND ORTHOGONALITY PRESERVATION

We begin by examining the root cause of the limitations of NorMuon’s adaptivity identified in the preceding section. Specifically, our analysis reveals that these limitations stem from the inherent coupling between orthogonalization and adaptive scaling. Because the scaling factors are computed from the orthogonalized momentum, the resulting adaptivity is inherently dictated by orthogonalization-induced geometry rather than direct gradient information. This analysis naturally motivates a core design principle: the computation of adaptive scaling factors should be decoupled from the orthogonalization procedure. Guided by this principle, we propose a decoupled adaptive mechanism that computes the adaptive scaling factors directly from the raw gradients, rather than from the orthogonalized momentum, while applying these factors to the orthogonalized update, i.e., the polar factor. In this way, the adaptive scaling is no longer dictated by orthogonalization-induced geometry, but instead reflects genuine optimization-relevant information—a defining characteristic of standard adaptive optimizers.

²We omit the weight decay term here for generality, as it is independent of the core algorithm and various variants can be seamlessly incorporated (e.g., standard or cautious weight decay (Chen et al., 2026)).

While decoupling the computation of adaptive scaling factors resolves the fundamental coupling between adaptivity and orthogonalization, a closer examination of the geometric structure of the orthogonalized update reveals another design deficiency in NorMuon. Specifically, in the $m > n$ regime, the polar factor has orthonormal columns, which define the orthogonal directions of the update, whereas NorMuon applies adaptive scaling along the row dimension. This dimensional mismatch inevitably distorts these orthogonal directions and destroys the orthogonality that the polar factor is designed to preserve. This observation motivates a second design principle: adaptive scaling should respect the intrinsic geometry of the orthogonalized update. Intuitively, since orthogonalization is designed to extract a set of orthogonal update directions, any subsequent adaptive scaling should preserve this directional structure rather than distort it. Following this principle, we introduce a geometry-aligned scaling mechanism that applies row-wise scaling for $m < n$ and column-wise scaling for $m > n$, while either choice can be used when $m = n$, thereby preserving the orthogonality of the update in all regimes.

4.2 FORMAL JUSTIFICATION OF THE ORTHOGONALITY-PRESERVING PROPERTY

We now provide a formal justification of the orthogonality-preserving property for the proposed geometry-aligned scaling mechanism. We first consider the $m > n$ regime under exact orthogonalization, where the polar factor $O_t \in \mathbb{R}^{m \times n}$ satisfies $O_t^\top O_t = I_n$. In this setting, the columns of O_t form a set of mutually orthogonal directions. Let $v_t \in \mathbb{R}^n$ denote the column-wise second-moment estimate, where each entry tracks the EMA of the squared ℓ_2 -norm of the corresponding column of the raw gradients, i.e.,

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) \text{diag}(G_t^\top G_t).$$

Note that the Gram matrix here is introduced only for notational convenience; in practice, only its diagonal entries are needed, so the full Gram matrix is never explicitly formed. The associated diagonal scaling matrix is constructed as $H_t = \text{Diag}(1/(\sqrt{v_t} + \delta)) \in \mathbb{R}^{n \times n}$, where $\text{Diag}(\cdot)$ denotes the operator that constructs a diagonal matrix from a vector. Applying column-wise adaptive scaling to O_t is equivalent to right-multiplying O_t by H_t , yielding the scaled update $\tilde{O}_t = O_t H_t$. The Gram matrix of the scaled update satisfies

$$\tilde{O}_t^\top \tilde{O}_t = H_t^\top O_t^\top O_t H_t = H_t^\top I_n H_t = H_t^2,$$

which remains diagonal. This implies that the columns of $\tilde{O}_t$ remain mutually orthogonal, i.e.,

$$\langle \tilde{O}_{:,i}, \tilde{O}_{:,j} \rangle = 0, \quad i \neq j,$$

while their squared norms satisfy

$$\|\tilde{O}_{:,i}\|_2^2 = (H_t^2)_{ii},$$

indicating that the magnitude of each orthogonal direction is independently scaled by its corresponding adaptive scaling factor. Thus, column-wise adaptive scaling transforms the polar factor with unit column norms into an update with orthogonal columns but heterogeneous column norms. It preserves the orthogonal directions extracted by orthogonalization while assigning an adaptive step size to each direction. In contrast, applying row-wise adaptive scaling to O_t is equivalent to left-multiplying O_t by the corresponding diagonal scaling matrix $D_t \in \mathbb{R}^{m \times m}$, constructed analogously from the row-wise second-moment estimate, yielding the scaled update $\tilde{O}_t = D_t O_t$. The resulting Gram matrix becomes

$$\tilde{O}_t^\top \tilde{O}_t = O_t^\top D_t^\top D_t O_t = O_t^\top D_t^2 O_t,$$

which is generally non-diagonal. This indicates that row-wise adaptive scaling reintroduces coupling between the orthogonal column directions extracted by orthogonalization, and hence the orthogonal structure established by orthogonalization is no longer preserved in the $m > n$ regime.

For the $m < n$ regime, the analysis proceeds analogously, with the roles of rows and columns reversed. In this case, we construct the adaptive scaling factors using the row-wise second-moment estimate $v_t \in \mathbb{R}^m$, where each entry tracks the EMA of the squared ℓ_2 -norm of the corresponding row of the raw gradients, i.e.,

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) \text{diag}(G_t G_t^\top).$$

The polar factor O_t satisfies $O_t O_t^\top = I_m$, implying that the rows of O_t form a set of mutually orthogonal directions. Applying row-wise adaptive scaling to O_t is equivalent to left-multiplying

O_t by the corresponding diagonal scaling matrix, thereby preserving the row-wise orthogonal structure. In contrast, applying column-wise adaptive scaling, equivalently right-multiplying O_t by the corresponding diagonal scaling matrix, generally distorts the row-wise orthogonal structure. For the $m = n$ regime, the polar factor O_t is a square orthogonal matrix, satisfying both $O_t O_t^\top = I_m$ and $O_t^\top O_t = I_m$. Hence, both the rows and columns of O_t form sets of mutually orthogonal directions, and either row-wise or column-wise adaptive scaling preserves the corresponding orthogonal directions, adjusting only their magnitudes.

Remark: Broader Insights for Matrix-Aware Optimizer Design. The design principles and geometric justifications presented above provide deeper insights into the general paradigm of matrix-aware optimization.

1. **A Generalized Design Philosophy.** Our decoupled, direction-aware adaptive mechanism transcends being a mere enhancement for Muon; it constitutes a more general design philosophy for matrix-aware optimizers. Notably, the directions utilized need not be restricted to the orthogonal directions used by Muon, but can naturally extend to broader classes of directions that characterize matrices.
2. **Intrinsic Necessity of Row-Wise/Column-Wise Adaptivity.** This perspective formally establishes why row-wise or column-wise adaptivity is fundamentally necessary: it provides a natural mechanism for adapting to the one-sided orthogonal directions of matrices. Rather than viewing row-wise/column-wise scaling as a structural approximation to Adam-like element-wise scaling (as in Adam-mini (Zhang et al., 2025b) and other related work), our approach grounds it in the intrinsic geometry of matrices (Zhang et al., 2026b). Furthermore, although methods such as NorMuon also employ row-wise/column-wise adaptation, our motivation and design arise from a distinct geometric alignment principle rather than heuristic normalization.

4.3 ADDITIONAL DESIGN CONSIDERATIONS

Beyond the two aforementioned design principles, we further incorporate several additional considerations when constructing the adaptive scaling factors to better regulate their dynamics during optimization. First, the second-moment estimate is constructed from the *sum* rather than the *mean* of the squared entries. While this is already implicit in the construction of v_t above, we make this design choice more explicit here. For a row or column gradient vector, the corresponding second-moment term is given by its squared ℓ_2 -norm, without normalization by the vector dimension. This reflects our design philosophy of treating each row (or column) gradient vector as a single adaptive unit, with its scaling determined by the *total* accumulated gradient energy of the vector. Consequently, units with larger accumulated gradient energy receive smaller scaling factors, while units with lower accumulated energy are assigned larger ones. This contrasts with NorMuon, which normalizes the row-wise gradient energy by the row dimension, making its adaptive scaling depend on the average per-coordinate energy rather than the total energy of each row.

Since the second-moment estimate v_t is computed directly from raw gradients and is decoupled from orthogonalization, it suffers from the standard initialization bias of EMA estimators. In particular, during the early optimization steps, v_t tends to underestimate the underlying second moment due to the zero initialization of the accumulator. We therefore apply the standard bias-correction technique to obtain a bias-corrected estimate $\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$, which is then used to compute the adaptive scaling factor $s_t = \frac{1}{\sqrt{\hat{v}_t + \delta}}$. Here v_t , $\hat{v}_t$, and s_t are generic notations that can represent vectors in either $\mathbb{R}^m$ or $\mathbb{R}^n$. Note that the first moment M_t requires no bias correction, owing to the scale-invariance of the subsequent orthogonalization operation (Pethick et al., 2025).

While introducing adaptive scaling into Muon provides more fine-grained control over different adaptive units, it may also lead to excessive anisotropic scaling, resulting in overly imbalanced updates across different orthogonal directions. Such excessive anisotropy necessitates the use of smaller step sizes to maintain stable optimization. To preserve Muon’s favorable property of supporting large step sizes, we introduce a clipping operation that bounds the degree of adaptivity. Specifically, we apply the following element-wise clipping operation

$$\text{clip}(s_t, s_{t,\min}, s_{t,\max}) = \min(\max(s_t, s_{t,\min}), s_{t,\max}).$$

We determine the clipping bounds $s_{t,\min}$ and $s_{t,\max}$ adaptively based on the median of the scaling factors, denoted by $s_{\text{med}} = \text{median}(s_t)$. Specifically, we set $s_{t,\min} = c \cdot s_{\text{med}}$ and $s_{t,\max} =$

$d \cdot s_{\text{med}}$, where $c \in (0, 1]$ and $d \geq 1$ control the allowable dynamic range. One can also use a single hyperparameter by setting $d = 1/c$. The clipping operation effectively bounds the condition number of the induced diagonal scaling matrix and prevents disproportionately large or small updates along specific orthogonal directions. When $s_{t,\min} = s_{t,\max}$, the clipping removes all relative adaptivity, reducing the update to a uniformly scaled Muon update.

5 THEORETICAL ANALYSIS

In this section, we establish the convergence guarantee of DGA-Muon under exact orthogonalization. We first introduce the assumptions required for the analysis.

Assumptions. We make the following assumptions throughout the analysis.

(A1) (L -smoothness) $\|\nabla f(X) - \nabla f(Y)\|_F \leq L\|X - Y\|_F$, for all X, Y .

(A2) (Unbiased stochastic gradients with bounded variance) The stochastic gradients satisfy

$$\mathbb{E}[G_t \mid \mathcal{F}_{t-1}] = \nabla f(W_t), \quad \mathbb{E}\|G_t - \nabla f(W_t)\|_F^2 \leq \sigma^2.$$

(A3) (Bounded stochastic gradients) There exists a constant $G_{\max} > 0$ such that

$$\|G_t\|_F \leq G_{\max} \text{ for all } t.$$

We now proceed with the convergence analysis. By the L -smoothness of f and the update rule $W_{t+1} = W_t - \eta \tilde{O}_t$, we have

$$\begin{aligned} f(W_{t+1}) &\leq f(W_t) + \langle \nabla f(W_t), W_{t+1} - W_t \rangle_F + \frac{L}{2} \|W_{t+1} - W_t\|_F^2 \\ &= f(W_t) - \eta \langle \nabla f(W_t), \tilde{O}_t \rangle_F + \frac{\eta^2 L}{2} \|\tilde{O}_t\|_F^2. \end{aligned}$$

We first analyze the quadratic term $\|\tilde{O}_t\|_F^2$. For the $m \leq n$ regime, the geometry-aligned adaptive update $\tilde{O}_t$ is given by $\tilde{O}_t = \text{Diag}(s_t) O_t$. Since $O_t = U_t V_t^\top$ is the polar factor of M_t , we have $O_t O_t^\top = U_t V_t^\top V_t U_t^\top = U_t U_t^\top = I_m$. Then using the definition of the Frobenius norm and the cyclic property of the trace, we have

$$\|\tilde{O}_t\|_F^2 = \text{tr}(\tilde{O}_t^\top \tilde{O}_t) = \text{tr}(O_t^\top \text{Diag}(s_t)^2 O_t) = \text{tr}(\text{Diag}(s_t)^2 O_t O_t^\top) = \text{tr}(\text{Diag}(s_t)^2) = \sum_{i=1}^m s_{t,i}^2.$$

Let $s_{t,\max} = \max_i s_{t,i}$ denote the maximum entry of the scaling vector s_t , then we have

$$\|\tilde{O}_t\|_F^2 = \sum_{i=1}^m s_{t,i}^2 \leq m s_{t,\max}^2.$$

For the $m > n$ regime, the adaptive update is given by $\tilde{O}_t = O_t \text{Diag}(s_t)$. By definition of the polar factor, we have $O_t^\top O_t = I_n$. Applying an analogous derivation, we obtain $\|\tilde{O}_t\|_F^2 = \sum_{j=1}^n s_{t,j}^2 \leq n s_{t,\max}^2$. By introducing $\ell = \min(m, n)$, both regimes can be unified and compactly written as $\|\tilde{O}_t\|_F^2 \leq \ell s_{t,\max}^2$. Consequently, the quadratic term in the smoothness inequality satisfies

$$\frac{\eta^2 L}{2} \|\tilde{O}_t\|_F^2 = \frac{\eta^2 L}{2} \sum_{k=1}^{\ell} s_{t,k}^2 \leq \frac{\eta^2 \ell L s_{t,\max}^2}{2}.$$

We now analyze the Frobenius inner product term $\langle \nabla f(W_t), \tilde{O}_t \rangle_F$. By decomposing $\nabla f(W_t)$ into M_t and the residual $\nabla f(W_t) - M_t$, we obtain

$$\langle \nabla f(W_t), \tilde{O}_t \rangle_F = \langle M_t, \tilde{O}_t \rangle_F + \langle \nabla f(W_t) - M_t, \tilde{O}_t \rangle_F.$$

The first term characterizes the alignment of the adaptively scaled orthogonal direction with the momentum, and the second term captures the momentum tracking error due to the discrepancy between M_t and the current gradient. We handle these two terms separately in the following. To begin, we address the first term, for which we have the following lemma.

Lemma 1 (Adaptive Polar Alignment Lower Bound). *Let M_t , O_t , $\tilde{O}_t$ and s_t be defined as in Algorithm 1. Let $s_{t,\min} = \min_i s_{t,i}$ denote the minimum entry of the scaling vector s_t . Then, the following lower bound on the adaptive polar alignment holds*

$$\langle M_t, \tilde{O}_t \rangle_F \geq s_{t,\min} \|M_t\|_F.$$

Proof. We first consider the $m \leq n$ regime, where the adaptive scaling is applied via left multiplication: $\tilde{O}_t = \text{Diag}(s_t)O_t$. Substituting $M_t = U_t \Sigma_t V_t^\top$ and $\tilde{O}_t = \text{Diag}(s_t)U_t V_t^\top$ into the Frobenius inner product $\langle M_t, \tilde{O}_t \rangle_F = \text{tr}(M_t^\top \tilde{O}_t)$, we obtain

$$\langle M_t, \tilde{O}_t \rangle_F = \text{tr}((U_t \Sigma_t V_t^\top)^\top (\text{Diag}(s_t)U_t V_t^\top)) = \text{tr}(V_t \Sigma_t U_t^\top \text{Diag}(s_t)U_t V_t^\top).$$

Using the cyclic property of the trace and the fact that V_t has orthonormal columns (i.e., $V_t^\top V_t = I_m$), and expanding the trace, we have

$$\langle M_t, \tilde{O}_t \rangle_F = \text{tr}(\Sigma_t U_t^\top \text{Diag}(s_t)U_t) = \sum_{j=1}^m \sigma_j (U_t^\top \text{Diag}(s_t)U_t)_{jj},$$

where $\{\sigma_j\}_{j=1}^m$ are the singular values of M_t . By definition, the diagonal entries of $U_t^\top \text{Diag}(s_t)U_t$ are given by

$$(U_t^\top \text{Diag}(s_t)U_t)_{jj} = \sum_{i=1}^m s_{t,i} U_{t,ij}^2,$$

where $\{s_{t,i}\}_{i=1}^m$ are the entries of s_t . Recall that $s_{t,\min}$ denotes the minimum scaling factor of s_t , i.e., $s_{t,\min} = \min_i s_{t,i}$. Then, for each j , we have

$$(U_t^\top \text{Diag}(s_t)U_t)_{jj} = \sum_{i=1}^m s_{t,i} U_{t,ij}^2 \geq \sum_{i=1}^m s_{t,\min} U_{t,ij}^2 = s_{t,\min} \sum_{i=1}^m U_{t,ij}^2 = s_{t,\min},$$

where the last equality follows from the column orthonormality of U_t , i.e., $U_t^\top U_t = I_m$. Therefore, we obtain a lower bound for $\langle M_t, \tilde{O}_t \rangle_F$

$$\langle M_t, \tilde{O}_t \rangle_F = \sum_{j=1}^m \sigma_j (U_t^\top \text{Diag}(s_t)U_t)_{jj} \geq s_{t,\min} \sum_{j=1}^m \sigma_j = s_{t,\min} \|M_t\|_*,$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix. Finally, since the norm inequality $\|A\|_* \geq \|A\|_F$ holds for any matrix A (Horn & Johnson, 2013), it follows that

$$\langle M_t, \tilde{O}_t \rangle_F \geq s_{t,\min} \|M_t\|_F.$$

The $m > n$ regime is analogous, where the adaptive scaling is applied via right multiplication as $\tilde{O}_t = O_t \text{Diag}(s_t)$. Utilizing the column orthonormality of both U_t and V_t , and following an analogous argument, we obtain the same bound. Combining the two regimes, we prove the claimed bound in the lemma. $\square$

Now, by decomposing M_t into the sum of the gradient $\nabla f(W_t)$ and the momentum tracking error $M_t - \nabla f(W_t)$, and applying the reverse triangle inequality for the Frobenius norm, we have

$$\|M_t\|_F = \|\nabla f(W_t) + (M_t - \nabla f(W_t))\|_F \geq \|\nabla f(W_t)\|_F - \|M_t - \nabla f(W_t)\|_F.$$

Combining this inequality with Lemma 1, we obtain

$$\langle M_t, \tilde{O}_t \rangle_F \geq s_{t,\min} \|M_t\|_F \geq s_{t,\min} \|\nabla f(W_t)\|_F - s_{t,\min} \|M_t - \nabla f(W_t)\|_F.$$

We now proceed to address the term $\langle \nabla f(W_t) - M_t, \tilde{O}_t \rangle_F$. We first consider its negation. By the Cauchy-Schwarz inequality, we have

$$\langle M_t - \nabla f(W_t), \tilde{O}_t \rangle_F \leq \|M_t - \nabla f(W_t)\|_F \|\tilde{O}_t\|_F.$$

As established earlier, we have $\|\tilde{O}_t\|_F^2 \leq \ell s_{t,\max}^2$ with $\ell = \min(m, n)$. Consequently, we obtain

$$\langle M_t - \nabla f(W_t), \tilde{O}_t \rangle_F \leq \sqrt{\ell} \|M_t - \nabla f(W_t)\|_F s_{t,\max}.$$

Therefore, by negating the above inequality, we obtain

$$\langle \nabla f(W_t) - M_t, \tilde{O}_t \rangle_F \geq -\sqrt{\ell} \|M_t - \nabla f(W_t)\|_F s_{t,\max}.$$

Substituting the above bounds for the alignment and tracking error terms into the smoothness inequality, we obtain the following descent lemma for DGA-Muon.

Lemma 2 (Per-step Descent Inequality). *Let W_t , M_t and s_t be defined as in Algorithm 1. Let $s_{t,\min}$ and $s_{t,\max}$ denote the minimum and maximum entries of s_t , respectively, and let $\ell = \min(m, n)$. Assume that f is L -smooth. Then, the updates of DGA-Muon satisfy the following per-step descent inequality*

$$f(W_{t+1}) \leq f(W_t) - \eta s_{t,\min} \|\nabla f(W_t)\|_F + \eta (s_{t,\min} + \sqrt{\ell} s_{t,\max}) \|M_t - \nabla f(W_t)\|_F + \frac{\eta^2 \ell L s_{t,\max}^2}{2}.$$

Next, we analyze $\|M_t - \nabla f(W_t)\|_F$, which quantifies the deviation between the stochastic momentum and the true gradient. Recall that $M_t = \beta_1 M_{t-1} + (1 - \beta_1) G_t$. Subtracting the true gradient $\nabla f(W_t)$, we have

$$M_t - \nabla f(W_t) = \beta_1 M_{t-1} + (1 - \beta_1) G_t - \nabla f(W_t).$$

To isolate the previous-step error, we add and subtract $\beta_1 \nabla f(W_{t-1})$

$$M_t - \nabla f(W_t) = \beta_1 (M_{t-1} - \nabla f(W_{t-1})) + (1 - \beta_1) (G_t - \nabla f(W_t)) + \beta_1 (\nabla f(W_{t-1}) - \nabla f(W_t)).$$

Unrolling recursively from t down to 1 gives

$$\begin{aligned} M_t - \nabla f(W_t) &= \beta_1^{t-1} (M_1 - \nabla f(W_1)) + (1 - \beta_1) \sum_{s=2}^t \beta_1^{t-s} (G_s - \nabla f(W_s)) \\ &\quad + \beta_1 \sum_{s=2}^t \beta_1^{t-s} (\nabla f(W_{s-1}) - \nabla f(W_s)). \end{aligned}$$

Taking the Frobenius norm and applying the triangle inequality, we obtain

$$\begin{aligned} \|M_t - \nabla f(W_t)\|_F &\leq \beta_1^{t-1} \|M_1 - \nabla f(W_1)\|_F + \left\| (1 - \beta_1) \sum_{s=2}^t \beta_1^{t-s} (G_s - \nabla f(W_s)) \right\|_F \\ &\quad + \beta_1 \sum_{s=2}^t \beta_1^{t-s} \|\nabla f(W_{s-1}) - \nabla f(W_s)\|_F. \end{aligned}$$

The right-hand side naturally decomposes into three contributions: the initialization term, the stochastic noise term, and the gradient drift term. In the following, we analyze each of them separately.

For the initialization term, since $M_1 = (1 - \beta_1) G_1$, we decompose

$$M_1 - \nabla f(W_1) = (1 - \beta_1) (G_1 - \nabla f(W_1)) - \beta_1 \nabla f(W_1).$$

Therefore,

$$\mathbb{E} \|\beta_1^{t-1} (M_1 - \nabla f(W_1))\|_F \leq \beta_1^{t-1} (1 - \beta_1) \mathbb{E} \|G_1 - \nabla f(W_1)\|_F + \beta_1^t \|\nabla f(W_1)\|_F$$

Using the unbiasedness of G_1 , Jensen's inequality and the gradient bound assumption, we have

$$\|\nabla f(W_1)\|_F = \|\mathbb{E}[G_1 | \mathcal{F}_0]\|_F \leq \mathbb{E}[\|G_1\|_F | \mathcal{F}_0] \leq G_{\max}.$$

Then applying Jensen's inequality, $\mathbb{E}\|X\|_F \leq \sqrt{\mathbb{E}\|X\|_F^2}$, and using the bounded-variance assumption $\mathbb{E}\|G_1 - \nabla f(W_1)\|_F^2 \leq \sigma^2$, we obtain

$$\mathbb{E} [\|\beta_1^{t-1} (M_1 - \nabla f(W_1))\|_F] \leq \beta_1^{t-1} (1 - \beta_1) \sigma + \beta_1^t G_{\max}.$$

For the stochastic noise term, we first bound its second moment. For notational convenience, we introduce the shorthand $\mathcal{N}_t = (1 - \beta_1) \sum_{s=2}^t \beta_1^{t-s} (G_s - \nabla f(W_s))$. Expanding the squared Frobenius norm, taking the expectation, and using the linearity of expectation, we obtain

$$\begin{aligned} \mathbb{E} \|\mathcal{N}_t\|_F^2 &= (1 - \beta_1)^2 \mathbb{E} \left\langle \sum_{s=2}^t \beta_1^{t-s} (G_s - \nabla f(W_s)), \sum_{r=2}^t \beta_1^{t-r} (G_r - \nabla f(W_r)) \right\rangle \\ &= (1 - \beta_1)^2 \sum_{s=2}^t \sum_{r=2}^t \beta_1^{2t-s-r} \mathbb{E} \langle G_s - \nabla f(W_s), G_r - \nabla f(W_r) \rangle. \end{aligned}$$

By assumption (A2), the stochastic gradient noise $\xi_t = G_t - \nabla f(W_t)$ forms a martingale difference sequence, i.e., $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1}$ denotes the filtration representing all information up to time $t - 1$. Hence, the cross terms ($s \neq r$) vanish in expectation, leaving only the diagonal terms

$$\mathbb{E}\|\mathcal{N}_t\|_F^2 = (1 - \beta_1)^2 \sum_{s=2}^t \beta_1^{2(t-s)} \mathbb{E}\|G_s - \nabla f(W_s)\|_F^2 \leq (1 - \beta_1)^2 \sigma^2 \sum_{k=0}^{t-2} \beta_1^{2k}.$$

Evaluating the geometric sum gives

$$\mathbb{E}\|\mathcal{N}_t\|_F^2 \leq (1 - \beta_1)^2 \sigma^2 \frac{1 - \beta_1^{2(t-1)}}{1 - \beta_1^2} = \sigma^2 \frac{(1 - \beta_1)(1 - \beta_1^{2(t-1)})}{1 + \beta_1} \leq \sigma^2 \frac{1 - \beta_1}{1 + \beta_1}.$$

Then applying Jensen's inequality and using the bounded-variance assumption, we obtain

$$\mathbb{E}\|\mathcal{N}_t\|_F \leq \sqrt{\mathbb{E}\|\mathcal{N}_t\|_F^2} \leq \frac{\sqrt{1 - \beta_1}}{\sqrt{1 + \beta_1}} \sigma.$$

We now proceed to handle the gradient drift term, which, unlike the stochastic noise term, can be bounded deterministically. By the L -smoothness of f , the update rule $W_s = W_{s-1} - \eta \tilde{O}_{s-1}$, and the bound $\|\tilde{O}_t\|_F \leq \sqrt{\ell} s_{t,\max}$, we have

$$\begin{aligned} \beta_1 \sum_{s=2}^t \beta_1^{t-s} \|\nabla f(W_{s-1}) - \nabla f(W_s)\|_F &\leq \beta_1 \sum_{s=2}^t \beta_1^{t-s} L \|W_{s-1} - W_s\|_F \\ &= \beta_1 \eta L \sum_{s=2}^t \beta_1^{t-s} \|\tilde{O}_{s-1}\|_F \leq \beta_1 \eta \sqrt{\ell} L \sum_{s=2}^t \beta_1^{t-s} s_{s-1,\max}. \end{aligned}$$

Let $s_{\max}$ be the maximum of all $s_{t,\max}$, then we obtain

$$\beta_1 \sum_{s=2}^t \beta_1^{t-s} \|\nabla f(W_{s-1}) - \nabla f(W_s)\|_F \leq \beta_1 \eta \sqrt{\ell} L s_{\max} \sum_{s=2}^t \beta_1^{t-s}$$

For the geometric sum, we have

$$\sum_{s=2}^t \beta_1^{t-s} = \sum_{k=0}^{t-2} \beta_1^k = \frac{1 - \beta_1^{t-1}}{1 - \beta_1} \leq \frac{1}{1 - \beta_1}.$$

Therefore, the gradient drift term admits the following deterministic bound

$$\beta_1 \sum_{s=2}^t \beta_1^{t-s} \|\nabla f(W_{s-1}) - \nabla f(W_s)\|_F \leq \frac{\beta_1 \eta \sqrt{\ell} L s_{\max}}{1 - \beta_1}.$$

Combining the bounds for these three terms yields the following bound on the expected momentum error.

Lemma 3 (Expected Momentum Error Bound). *Let M_t and W_t be defined as in Algorithm 1. Let $s_{\max}$ denote a uniform upper bound on the adaptive scaling factors satisfying $(s_t)_i \leq s_{\max} < \infty$ for all t and i , and let $\ell = \min(m, n)$. Assume that f is L -smooth. Then, under the bounded-variance and bounded-gradient assumptions, the expected momentum error at each step satisfies*

$$\mathbb{E}\|M_t - \nabla f(W_t)\|_F \leq \beta_1^{t-1} (1 - \beta_1) \sigma + \beta_1^t G_{\max} + \frac{\sqrt{1 - \beta_1}}{\sqrt{1 + \beta_1}} \sigma + \frac{\beta_1 \eta \sqrt{\ell} L s_{\max}}{1 - \beta_1}.$$

Recall that Lemma 2 establishes the following one-step descent inequality for the updates of Algorithm 1

$$f(W_{t+1}) \leq f(W_t) - \eta s_{t,\min} \|\nabla f(W_t)\|_F + \eta (s_{t,\min} + \sqrt{\ell} s_{t,\max}) \|M_t - \nabla f(W_t)\|_F + \frac{\eta^2 \ell L s_{t,\max}^2}{2}.$$

To simplify the subsequent analysis, we replace the time-dependent quantities by uniform bounds over the optimization horizon. Let $s_{\min}$ denote a uniform lower bound on the adaptive scaling factors satisfying $0 < s_{\min} \leq (s_t)_i$ for all t and i . Since

$$s_{\min} \leq s_{t,\min} \leq s_{t,\max} \leq s_{\max},$$

we have

$$-\eta s_{t,\min} \|\nabla f(W_t)\|_F \leq -\eta s_{\min} \|\nabla f(W_t)\|_F,$$

and

$$s_{t,\min} + \sqrt{\ell} s_{t,\max} \leq (1 + \sqrt{\ell}) s_{\max}.$$

Moreover, $s_{t,\max}^2 \leq s_{\max}^2$. Therefore, the one-step descent inequality can be uniformly bounded as

$$f(W_{t+1}) \leq f(W_t) - \eta s_{\min} \|\nabla f(W_t)\|_F + \eta(1 + \sqrt{\ell}) s_{\max} \|M_t - \nabla f(W_t)\|_F + \frac{\eta^2 \ell L s_{\max}^2}{2}.$$

Rearranging to isolate the gradient term and taking expectations on both sides yields

$$\begin{aligned} \eta s_{\min} \mathbb{E}[\|\nabla f(W_t)\|_F] &\leq \mathbb{E}[f(W_t) - f(W_{t+1})] + \eta(1 + \sqrt{\ell}) s_{\max} \mathbb{E}[\|M_t - \nabla f(W_t)\|_F] \\ &\quad + \frac{\eta^2 \ell L s_{\max}^2}{2}. \end{aligned}$$

Then, summing over $t = 1, \dots, T$, using the telescoping sum $\sum_{t=1}^T \mathbb{E}[f(W_t) - f(W_{t+1})] = \mathbb{E}[f(W_1) - f(W_{T+1})]$, and subsequently dividing by $\eta s_{\min}$, we obtain

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}[\|\nabla f(W_t)\|_F] &\leq \frac{\mathbb{E}[f(W_1) - f(W_{T+1})]}{\eta s_{\min}} + \frac{(1 + \sqrt{\ell}) s_{\max}}{s_{\min}} \sum_{t=1}^T \mathbb{E}[\|M_t - \nabla f(W_t)\|_F] \\ &\quad + \frac{\eta \ell L s_{\max}^2 T}{2 s_{\min}}. \end{aligned}$$

Summing the momentum error inequality given by Lemma 3 from $t = 1$ to T yields

$$\sum_{t=1}^T \mathbb{E}\|M_t - \nabla f(W_t)\|_F \leq (1 - \beta_1) \sigma \sum_{t=1}^T \beta_1^{t-1} + G_{\max} \sum_{t=1}^T \beta_1^t + T \frac{\sqrt{1 - \beta_1}}{\sqrt{1 + \beta_1}} \sigma + T \frac{\beta_1 \eta \sqrt{\ell} L s_{\max}}{1 - \beta_1}.$$

Evaluating the geometric sums $\sum_{t=1}^T \beta_1^{t-1} \leq \frac{1}{1 - \beta_1}$ and $\sum_{t=1}^T \beta_1^t \leq \frac{\beta_1}{1 - \beta_1}$, we obtain

$$\sum_{t=1}^T \mathbb{E}\|M_t - \nabla f(W_t)\|_F \leq \sigma + \frac{\beta_1 G_{\max}}{1 - \beta_1} + T \frac{\sqrt{1 - \beta_1}}{\sqrt{1 + \beta_1}} \sigma + T \frac{\beta_1 \eta \sqrt{\ell} L s_{\max}}{1 - \beta_1}.$$

Denote $\Delta = f(W_1) - f^*$, where $f^* = \inf_W f(W)$. Since $f(W_{T+1}) \geq f^*$, we have

$$\mathbb{E}[f(W_1) - f(W_{T+1})] \leq \Delta.$$

Substituting the bound for $\sum_{t=1}^T \mathbb{E}\|M_t - \nabla f(W_t)\|_F$ and $\mathbb{E}[f(W_1) - f(W_{T+1})] \leq \Delta$ into the previous inequality, we obtain the final bound for the cumulative expected gradient norm:

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}[\|\nabla f(W_t)\|_F] &\leq \frac{\Delta}{\eta s_{\min}} + \frac{(1 + \sqrt{\ell}) s_{\max}}{s_{\min}} \left(\sigma + \frac{\beta_1 G_{\max}}{1 - \beta_1} \right) \\ &\quad + T \frac{(1 + \sqrt{\ell}) s_{\max}}{s_{\min}} \frac{\sqrt{1 - \beta_1}}{\sqrt{1 + \beta_1}} \sigma \\ &\quad + T \frac{\beta_1 \eta \sqrt{\ell} L (1 + \sqrt{\ell}) s_{\max}^2}{(1 - \beta_1) s_{\min}} + T \frac{\eta \ell L s_{\max}^2}{2 s_{\min}}. \end{aligned}$$

Dividing both sides by T immediately yields a bound on the average expected gradient norm. Since the minimum is always upper bounded by the average, i.e., $\min_{1 \leq t \leq T} \mathbb{E}[\|\nabla f(W_t)\|_F] \leq \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla f(W_t)\|_F]$, we obtain the final convergence bound.

Theorem 1 (Convergence Guarantee of DGA-Muon). *Let W_t be defined as in Algorithm 1. Let $s_{\min}$ and $s_{\max}$ denote uniform lower and upper bounds on the adaptive scaling factors satisfying $0 < s_{\min} \leq (s_t)_i \leq s_{\max} < \infty$ for all t and i , and let $\ell = \min(m, n)$. Denote $\Delta = f(W_1) - f^*$. Assume that f is L -smooth and that the stochastic gradients are unbiased with bounded variance*

and bounded norm. Then, for a fixed step size η , DGA-Muon satisfies the following convergence guarantee

$$\begin{aligned} \min_{1 \leq t \leq T} \mathbb{E}[\|\nabla f(W_t)\|_F] &\leq \frac{\Delta}{\eta s_{\min} T} + \frac{(1 + \sqrt{\ell})s_{\max}}{s_{\min} T} \left(\sigma + \frac{\beta_1 G_{\max}}{1 - \beta_1} \right) \\ &\quad + \frac{(1 + \sqrt{\ell})s_{\max}}{s_{\min}} \frac{\sqrt{1 - \beta_1}}{\sqrt{1 + \beta_1}} \sigma \\ &\quad + \frac{\beta_1 \eta \sqrt{\ell} L (1 + \sqrt{\ell}) s_{\max}^2}{(1 - \beta_1) s_{\min}} + \frac{\eta \ell L s_{\max}^2}{2 s_{\min}}. \end{aligned}$$

In particular, this bound can be summarized as

$$\min_{1 \leq t \leq T} \mathbb{E}[\|\nabla f(W_t)\|_F] = O\left(\frac{1}{T} + \sigma + \eta\right).$$

Therefore, for a fixed step size η , DGA-Muon achieves a convergence rate of $\mathcal{O}(1/T + \sigma + \eta)$. The adaptive scaling primarily influences the constant factors within the convergence bound. This result reveals the standard bias-variance tradeoff inherent in constant step-size stochastic optimization: the transient optimization error decays at an $\mathcal{O}(1/T)$ rate, while the stochastic gradient variance and the finite learning rate introduce an irreducible steady-state error. Consequently, DGA-Muon converges to a neighborhood of a stationary point, characterized by a bounded expected gradient norm whose radius depends on the stochastic variance σ and the step size η . Similar to classical stochastic optimization methods, a diminishing step-size schedule is expected to gradually eliminate this steady-state error, although extending the present analysis to this setting is beyond the scope of this work. Furthermore, the current analysis establishes a foundational convergence guarantee for DGA-Muon. A rigorous theoretical characterization of when the geometry-aligned adaptive scaling yields provable improvements over the original Muon update remains an open question and is left for future work.

6 EXPERIMENTS

In this section, we report our experimental results. Note that, due to limited computational resources, some experiments were conducted only at a small scale and are not included in the current version. We will continue to expand the experimental evaluation as additional results become available.

6.1 EXPERIMENTAL SETUP

We conduct our experiments using NanoChat (Karpathy, 2025), a highly optimized transformer pretraining framework that can serve as a testbed for evaluating optimization methods. Both the model architecture and training recipe have undergone extensive engineering and empirical tuning in the development of NanoChat, including multiple rounds of automated optimization with *Autoresearch* (Karpathy, 2026). As a result, NanoChat provides a strong and carefully tuned baseline for optimizer evaluation. We use the default NanoChat model architecture without modification and refer readers to the repository for full implementation details. For our experiments, we use a model that has 24 layers and 1.38B parameters in total, with full attention used across all layers. All experiments are conducted on a single node with eight NVIDIA A100 (80GB) GPUs.

We next describe the optimizer configuration used in NanoChat, which incorporates a range of advanced and carefully tuned optimization techniques. NanoChat employs a modified version of NorMuon (Li et al., 2026). As in conventional Muon-based training, the optimizer is applied only to matrix-valued parameters, including the Q/K/V/O and FFN weight matrices, while the remaining parameters are optimized with AdamW (Loshchilov & Hutter, 2019). Unlike standard NorMuon, which applies row-wise scaling to all matrix parameters, NanoChat selects the scaling direction according to the shape of each matrix: row-wise scaling is used for tall and square matrices, whereas column-wise scaling is used for wide matrices. After the adaptive scaling, NanoChat further normalizes the resulting gradient to maintain a consistent update norm across optimization steps (see the implementation for more details). For the matrix-valued parameters, NanoChat employs the recently

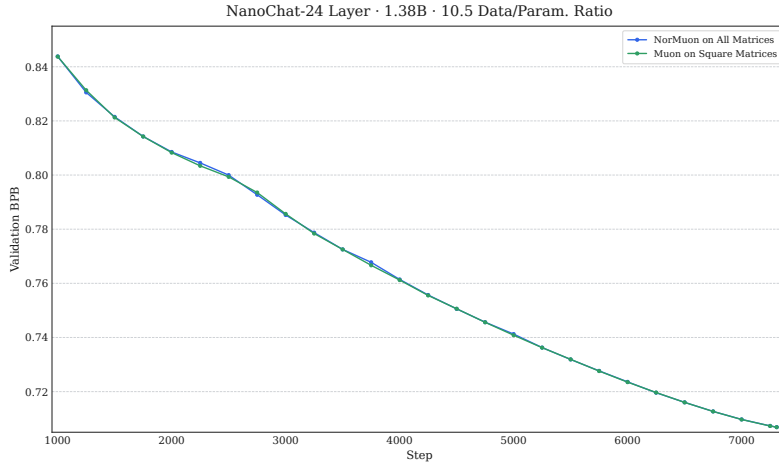


Figure 1: Empirical verification of NorMuon’s adaptive scaling degeneration on square matrices in NanoChat. Removing NorMuon’s adaptive scaling on the square Q/K/V/O projection matrices, effectively reducing NorMuon to standard Muon, leaves the validation BPB essentially unchanged.

proposed cautious weight decay (Chen et al., 2026) in place of conventional decoupled weight decay. We also ablate this design choice by replacing cautious weight decay with standard decoupled weight decay and find a clear degradation in training performance.

6.2 EMPIRICAL VERIFICATION OF NORMUON’S SCALING DEGENERATION

Recall that in Section 3, we theoretically showed that NorMuon’s adaptive scaling degenerates to an approximately constant factor on square matrices. In this experiment, we empirically verify this theoretical characterization by removing the adaptive scaling on the Q/K/V/O matrices, in which case NorMuon effectively reduces to standard Muon. In the current NanoChat implementation, the Q/K/V/O projection matrices are optimized as full matrices over all attention heads rather than separately on a per-head basis, and hence are square. We also tested per-head optimization of the Q/K/V matrices, but observed no improvement over joint optimization.

Following the NanoChat setup, the total number of training steps is 7,308, and validation is performed every 250 steps, with an additional validation run at the end of training. We use validation bits per byte (BPB) as the primary evaluation metric, which measures the average number of bits required to encode each byte of the validation data and is invariant to the tokenizer vocabulary size. Lower BPB indicates better language-modeling performance.

As shown in Figure 1, removing NorMuon’s adaptive scaling on the square Q/K/V/O matrices leaves the validation BPB essentially unchanged, with the curves of the two optimizers nearly overlapping. For the training loss, the two curves are virtually indistinguishable, and we therefore report them in Appendix A. This result is consistent with our theoretical analysis: for square matrices, the corresponding adaptive scaling of NorMuon degenerates to an approximately constant factor, effectively eliminating its intended adaptive effect. Consequently, removing this scaling has negligible impact on optimization performance, providing empirical support for our theoretical characterization.

6.3 DGA-MUON OUTPERFORMS NORMUON

We now evaluate the proposed DGA-Muon on NanoChat, using NorMuon as the baseline. We retain the same 24-layer, 1.38B-parameter Transformer architecture as in the ablation study above. For this end-to-end evaluation, we increase the data-to-parameter ratio from 10 to 18, which is also closer to the Chinchilla-style scaling regime (Hoffmann et al., 2022).

Note that DGA-Muon and NorMuon have different optimization characteristics due to their distinct adaptive scaling mechanisms, and therefore require different hyperparameter configurations. We first explain the rationale behind our hyperparameter choices for DGA-Muon. Recall that DGA-

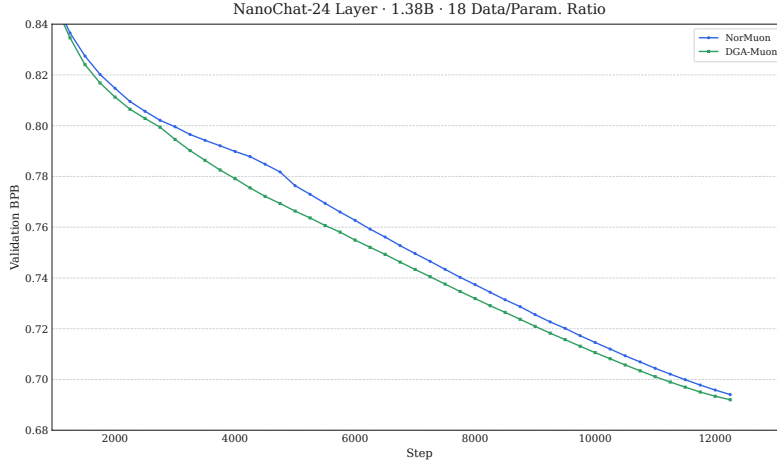


Figure 2: Comparison of DGA-Muon and NorMuon in NanoChat. DGA-Muon consistently outperforms NorMuon, achieving lower validation BPB throughout training.

Muon constructs its adaptive scaling signal from the raw gradients, whereas NorMuon derives it from the orthogonalized gradients. Orthogonalization provides Muon with favorable isotropic spectral properties, but also removes information encoded in the relative magnitudes of different singular directions. This information becomes increasingly valuable in the later stages of training, as the overall update magnitude decreases and finer-grained differences across singular directions become more important for effective optimization. In contrast, the raw-gradient signal retains this anisotropic information, while exhibiting greater step-to-step variation. These characteristics motivate a more gradual learning-rate schedule for DGA-Muon. Specifically, we extend the warmup phase, increase the warmdown ratio, and reduce the final learning rate fraction, while slightly lowering the learning rate for matrix-valued parameters. The complete hyperparameter configurations are provided in Table 1 in the Appendix.

As shown in Figure 2, DGA-Muon significantly outperforms NorMuon in the end-to-end evaluation, achieving a final validation BPB of 0.6910, compared with 0.6924 for NorMuon. The difference in training loss is more pronounced, with DGA-Muon reaching a final loss of 2.2708 compared with 2.2799 for NorMuon; the corresponding training-loss curves are provided in Figure 4 in the Appendix. We further evaluate the downstream performance of the two optimizers on the DCLM CORE benchmark, the standard aggregate evaluation metric used in NanoChat. DGA-Muon achieves a CORE score of 0.3089, while NorMuon obtains 0.2942, yielding an absolute improvement of 0.0147 and a relative gain of 5.0%. Full downstream results are provided in Table 2 in the Appendix. These results validate the effectiveness of our proposed adaptive-scaling design. By constructing adaptive scaling signals directly from raw gradients and aligning the scaling with the underlying geometry of matrices, DGA-Muon achieves more effective adaptation than NorMuon, leading to lower training loss and validation BPB, as well as better downstream performance.

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 (ICLR)*, 2026.
- Lizhang Chen, Jonathan Li, Kaizhao Liang, Baiyu Su, Cong Xie, Chen Liang, Ni Lao, and Qiang Liu. Cautious weight decay. In *International Conference on Learning Representations*, 2026.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 2011.
- Ekaterina Grishina, Matvey Smirnov, and Maxim Rakhuba. Accelerating newton-schulz iteration for orthogonalization via chebyshev-type polynomials. *arXiv preprint arXiv:2506.10935*, 2025.

- Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning (ICML)*, 2018.
- Wei He, Kai Han, Hang Zhou, Hanting Chen, Zhicheng Liu, Xinghao Chen, and Yunhe Wang. Root: Robust orthogonalized optimizer for neural network training. *arXiv preprint arXiv:2511.20626*, 2025.
- Nicholas J Higham. *Functions of matrices: theory and computation*. Society for Industrial and Applied Mathematics (SIAM), 2008.
- Nicholas J. Higham. What is the polar decomposition, 2020. URL <https://nhigham.com/2020/07/28/what-is-the-polar-decomposition/>.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 2nd edition, 2013.
- Keller Jordan. Muon: An optimizer for hidden layers in neural networks, 2024. URL <https://kellerjordan.github.io/posts/muon/>.
- Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, Boza Vlado, Jiacheng You, Franz Cesista, Braden Koszarsky, et al. Modded-NanoGPT: Speedrunning the NanoGPT baseline, 2024. URL <https://github.com/KellerJordan/modded-nanogpt>.
- Andrej Karpathy. NanoChat: The best ChatGPT that \$100 can buy, 2025. URL <https://github.com/karpathy/nanochat>.
- Andrej Karpathy. Autoresearch, 2026. URL <https://github.com/karpathy/autoresearch>.
- Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations (ICLR)*, 2026.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon. *arXiv preprint arXiv:2502.02900*, 2025.
- Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. NorMuon: Making Muon more efficient and scalable. In *International Conference on Machine Learning (ICML)*, 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.
- Ziyue Liu, Ruijie Zhang, Zhengyang Wang, Yequan Zhao, Yupeng Su, Zi Yang, and Zheng Zhang. Muon²: Boosting Muon via adaptive second-moment preconditioning. *arXiv preprint arXiv:2604.09967*, 2026.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.
- Tien-Phat Nguyen, Truong Nguyen, Minh-Phuc Truong, Tuc Nguyen, James Bailey, and Trung Le. Spectral flattening is all Muon needs: How orthogonalization controls learning rate and convergence. *arXiv preprint arXiv:2605.13079*, 2026.

- 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 (ICML)*, 2025.
- Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvvaraju, 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.
- Zitao Song, Cedar Site Bai, Zhe Zhang, Brian Bullins, and David F Gleich. Decoupling variance and scale-invariant updates in adaptive gradient descent for unified vector and matrix optimization. *arXiv preprint arXiv:2602.06880*, 2026.
- Weijie Su. Isotropic curvature model for understanding deep learning optimization: Is gradient orthogonalization optimal? *arXiv preprint arXiv:2511.00674*, 2025.
- 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, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.
- Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why Muon outperforms Adam: A curvature perspective. *arXiv preprint arXiv:2606.04662*, 2026a.
- 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 (ICLR)*, 2026b.
- 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*, 2026a.
- Chenrui Xu, Wenjing Yan, and Ying-Jun Angela Zhang. Fismo: Fisher-structured momentum-orthogonalized optimizer. *arXiv preprint arXiv:2601.21750*, 2026b.
- Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. GLM-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 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.
- Jack Zhang, Noah Amsel, Berlin Chen, and Tri Dao. Gram Newton–Schulz, 2026a. URL <https://dao-ailab.github.io/blog/2026/gram-newton-schulz/>.
- Minxin Zhang, Yuxuan Liu, and Hayden Schaeffer. AdaGrad meets Muon: Adaptive stepsizes for orthogonal updates. In *Workshop on Optimization for Machine Learning in NeurIPS*, 2025a.
- Wenpeng Zhang, Runsheng Yu, and Peilin Zhao. Matrix adagrad: Row-wise and column-wise adaptive subgradient methods, 2026b. URL <https://arxiv.org/abs/2609.21815>.
- Yechen Zhang, Shuhao Xing, Junhao Huang, Kai Lv, Yunhua Zhou, Xipeng Qiu, Qipeng Guo, and Kai Chen. Mousse: Rectifying the geometry of Muon with curvature-aware preconditioning. *arXiv preprint arXiv:2603.09697*, 2026c.
- Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik Durk Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. In *International Conference on Learning Representations (ICLR)*, 2025b.

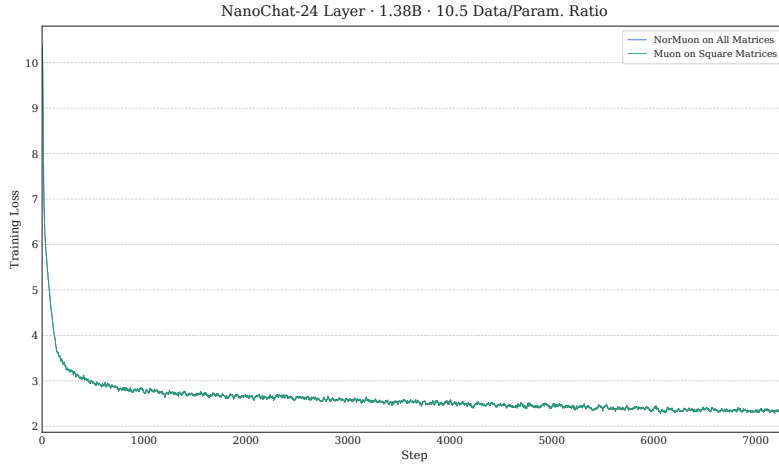


Figure 3: Empirical verification of NorMuon’s adaptive scaling degeneration on square matrices in NanoChat. Removing NorMuon’s adaptive scaling on the square Q/K/V/O projection matrices, effectively reducing NorMuon to standard Muon, yields virtually indistinguishable training losses.

Table 1: Hyperparameter configurations for the NorMuon baseline and DGA-Muon in NanoChat.

Hyperparameter	NorMuon Baseline	DGA-Muon
Device batch size	32	32
Total batch size	Auto	Auto
Weight decay	0.28	0.28
Embedding LR	0.3	0.3
Unembedding LR	0.008	0.008
Scalar LR	0.5	0.5
Matrix LR	0.020	0.019
Warmup steps	40	200
Warmdown ratio	0.65	0.78
Final LR fraction	0.05	0.01
Lower scaling clip	—	0.001
Upper scaling clip	—	20

APPENDIX

A ADDITIONAL EXPERIMENTAL DETAILS

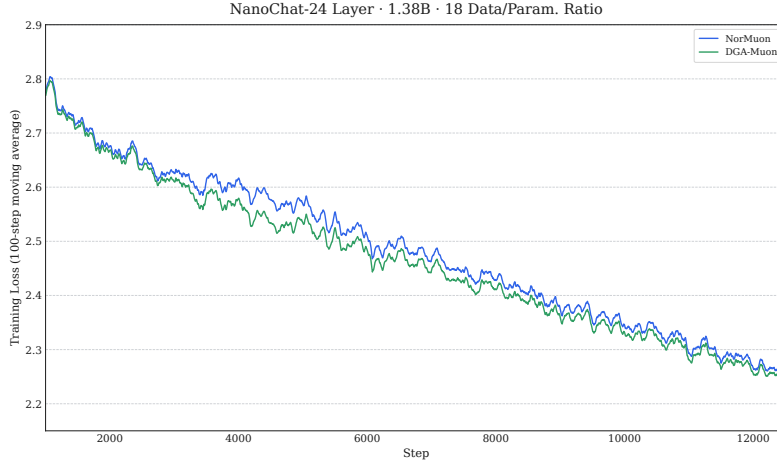


Figure 4: Comparison of DGA-Muon and NorMuon in NanoChat. DGA-Muon consistently outperforms NorMuon, achieving lower training loss throughout training.

Table 2: Downstream task performance of NorMuon and DGA-Muon on the DCLM CORE benchmark. Higher scores indicate better performance.

Task	NorMuon	DGA-Muon	Diff.
HellaSwag ZeroShot	0.6080	0.6100	+0.0020
Jeopardy	0.1680	0.1840	+0.0160
BIG-Bench QA Wikidata	0.5300	0.5560	+0.0260
ARC-Easy	0.7460	0.7420	−0.0040
ARC-Challenge	0.4760	0.4580	−0.0180
COPA	0.6400	0.7100	+0.0700
CommonsenseQA	0.2300	0.2560	+0.0260
PIQA	0.7800	0.7840	+0.0040
OpenBookQA	0.4320	0.4080	−0.0240
LAMBADA	0.4880	0.4860	−0.0020
HellaSwag	0.6180	0.6260	+0.0080
Winograd	0.7143	0.6996	−0.0147
Winogrande	0.5760	0.5660	−0.0100
BIG-Bench Dyck Languages	0.0660	0.0940	+0.0280
AGI-Eval LSAT-AR	0.2478	0.3130	+0.0652
BIG-Bench CS Algorithms	0.3940	0.4220	+0.0280
BIG-Bench Operators	0.2238	0.1857	−0.0381
BIG-Bench Repeat Copy Logic	0.0312	0.0312	0.0000
SQuAD	0.5100	0.4740	−0.0360
CoQA	0.3220	0.3100	−0.0120
BoolQ	0.5660	0.6280	+0.0620
BIG-Bench Language Identification	0.2640	0.2500	−0.0140
CORE	0.2942	0.3089	+0.0147