TRAINING-LOSS GUARANTEES FOR MUON WITH FINITE-STEP NEWTON-SCHULZ ORTHOGONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing convergence analyses of Muon assume exact orthogonalization and guarantee only stationarity, so it is unresolved what Muon's five tuned Newton-Schulz steps preserve and whether that suffices to reach a prescribed neural-network training loss. We establish a finite-time training guarantee that accounts for both momentum accumulation before orthogonalization and the tuned finite-step update. For full-batch training of a sufficiently wide two-layer ReLU network with fixed random output weights and a positive-definite limiting neural tangent kernel, we prove that Muon reaches any target empirical squared loss $\varepsilon > 0$ with high probability over initialization. For every momentum parameter $\mu \in [0, 1)$, a target-dependent constant learning rate proportional to $(1-\mu)\sqrt{\varepsilon}$ yields a hitting-time bound of $O((1-\mu)^{-1}\varepsilon^{-1/2})$, with other problem parameters fixed. The sufficient width is independent of both target accuracy and momentum. The analysis shows that the tuned Newton–Schulz map preserves alignment with the momentum buffer while bounding the update's spectral norm. Control of gradient variation near initialization transfers this alignment to the current gradient, ensuring descent until the target is reached without requiring exact orthogonalization. Numerical experiments support these mechanisms at widths below the sufficient theoretical threshold: gradient-update alignment remains above the analytical reference, and all 30 runs across six widths and five student initializations on a fixed teacher-student dataset reach the target loss while maintaining kernel positivity.

1 INTRODUCTION

How much can an optimizer alter the gradient and still provably fit the training data? Muon (*MomentUm Orthogonalized by Newton–Schulz*) makes this question concrete: it accumulates gradients in a momentum buffer and then reshapes the buffer's singular values through five tuned Newton–Schulz iterations before updating the weights (Jordan et al., 2024). This approximately orthogonalized update has proved effective in large-scale language-model training (Liu et al., 2025; Team et al., 2025), yet the transformation complicates the connection between the direction applied to the weights and the gradient of the current loss. Momentum carries information from earlier iterates, while finite orthogonalization changes how that information enters the update. Understanding why sufficient descent survives these operations is central to establishing a training-loss guarantee. We study this question for overparameterized two-layer ReLU networks and show that Muon can reach a prescribed loss while retaining both momentum and the tuned five-step Newton–Schulz transformation.

Existing analyses explain several aspects of Muon's update, but leave a gap between its practical construction and a neural-network training-loss guarantee. The exact polar factor admits a geometric interpretation as a steepest-descent direction under a spectral-norm constraint (Bernstein & Newhouse, 2024a;b). Analyses incorporating momentum establish stationarity guarantees for smooth objectives, while replacing the finite Newton–Schulz map by exact orthogonalization (Chen et al., 2025; Shen et al., 2025). Kim & Oh (2026) analyze finitely many Newton–Schulz iterations, but their guarantees concern classical polynomial iterations rather than the tuned coefficients used by Muon. Separately, overparameterized-network theory establishes conditions under which gradient descent reaches arbitrarily small training loss (Du et al., 2018b;a; Arora et al., 2019). These results do not transfer directly to Muon: a positive-definite tangent kernel controls the gradient through the training residual, but the actual loss decrease depends on the gradient's alignment with the transformed momentum buffer. The missing link is therefore a quantitative descent guarantee that accommodates

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.

both accumulated gradients and tuned, approximate orthogonalization. Table 1 (in the appendix) summarizes the closest results. For gradient descent on this model, the guarantees reach arbitrarily small training loss at a linear rate, with polynomial widths from n^6 (Du et al., 2018b) down to $\widetilde{\Omega}(n^2)$ (Arora et al., 2019). The two closest Muon analyses (Chen et al., 2025; Kim & Oh, 2026) guarantee only stationarity. Our result is a training-loss guarantee for Muon, with width $\Omega(n^5/(\lambda_0^4\delta^3))$.

We establish this link by identifying two properties of the finite Newton–Schulz map that suffice for a training-loss guarantee: alignment with its input and a bounded output spectral norm. For Muon's tuned five-step map, we prove both properties uniformly over nonzero momentum buffers, without requiring proximity to the exact polar factor. The remaining challenge is to transfer alignment from the buffer to the current gradient. We show that, within a controlled neighborhood of initialization, gradient variation is sufficiently small for this transfer to hold for every momentum parameter $\mu \in [0, 1)$, with a learning rate scaled by $1 - \mu$. A uniform neural tangent kernel argument controls the ReLU linearization error and ensures that the iterates remain in this neighborhood until the target loss is reached. Together, these estimates yield an explicit finite-time guarantee for full-batch training of a two-layer ReLU network with fixed random output weights, summarized below.

Theorem (Informal statement)**.** *Consider full-batch Muon with momentum $\mu \in [0, 1)$ and five tuned Newton–Schulz iterations, training the first layer of a width-m two-layer ReLU network with fixed random output weights. Let the n training inputs have unit norm, and suppose their limiting neural tangent kernel has smallest eigenvalue $\lambda_0 > 0$. Treat the output-weight scale and the average squared label magnitude as constants. For any target loss $\varepsilon > 0$ and failure probability $\delta \in (0, 1)$, take*

$$m = \Omega\left(\frac{n^5}{\lambda_0^4\delta^3}\right), \qquad \eta = \Theta\left(\frac{(1-\mu)\sqrt{\lambda_0\varepsilon}}{n}\right).$$

Then, with probability at least $1 - \delta$ over initialization, an iterate with empirical squared loss at most ε is reached within

$$T = \mathcal{O}\left(\frac{n^{3/2}}{(1-\mu)\lambda_0\sqrt{\delta\varepsilon}}\right)$$

steps. The precise assumptions and constants are given in Theorem 3.1.

The sufficient width is independent of both the target accuracy ε and momentum μ: a smaller target or larger momentum changes the prescribed learning rate and iteration budget without requiring a wider network. The result is a hitting-time guarantee with a target-dependent constant learning rate; it does not imply acceleration over gradient descent or a benefit from increasing momentum. By contrast, Wang et al. (2020) prove that heavy-ball momentum accelerates gradient descent for training a wide ReLU network with one hidden layer; our bound charges momentum the factor $(1 - \mu)^{-1}$ and shows no such gain. Our experiments directly examine the alignment and stability mechanisms underlying the proof. Across the tested momentum values, gradient-update alignment remains above the analytical reference. All 30 runs across six widths and five student initializations on a fixed dataset and teacher reach the target loss while maintaining kernel positivity, and kernel variation decreases with width. Comparisons on shared momentum buffers further quantify the alignment and norm properties of tuned Newton–Schulz updates relative to classical iterations and the exact polar factor. These findings support the relevance of the proof's mechanisms at widths many orders of magnitude below the sufficient theoretical threshold (Section 5), without establishing the necessity or tightness of the bounds.

2 Problem Setup and the Muon Update

For vectors, $\|\cdot\|_2$ denotes the Euclidean norm; for matrices, $\|\cdot\|_2$, $\|\cdot\|_F$, and $\|\cdot\|_*$ denote the spectral, Frobenius, and nuclear norms, respectively. We use the Frobenius inner product $\langle A, B\rangle = \mathrm{tr}(A^\top B)$ and column-wise vectorization.

2.1 Network and Training Objective

Let $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ be fixed training data with $\|x_i\|_2 = 1$. We consider the width-m two-layer ReLU network

$$f(x; W) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(w_r^\top x), \qquad W = (w_1, \ldots, w_m) \in \mathbb{R}^{d\times m}, \qquad \sigma(z) = \max\{0, z\}. \quad (1)$$

Algorithm 1 Full-batch Muon

Require: Initial weights W_0, learning rate $\eta > 0$, momentum $\mu \in [0, 1)$, target loss $\varepsilon > 0$, iteration budget T

$B_0 \leftarrow 0$
for $t = 1, \dots, T$ **do**
  if $L(W_{t-1}) \le \varepsilon$ **then**
    return W_{t-1}
  end if
  $G_t \leftarrow G(W_{t-1})$
  $B_t \leftarrow \mu B_{t-1} + G_t$
  $O_t \leftarrow \mathrm{NS}_5(B_t)$
  $W_t \leftarrow W_{t-1} - \eta O_t$
end for
return W_T

Only the first-layer weights W are trained. For a fixed $\alpha > 0$, the output weights are sampled independently as $a_r \sim \mathrm{Unif}\{\pm\alpha\}$ and held fixed. The first-layer weights are initialized independently as $w_r(0) \sim \mathcal{N}(0, I_d)$, independently of the output weights. All probability statements concern this random initialization, conditional on the training data.

Write $y = (y_1, \dots, y_n)^\top$, let $u(W) = (f(x_1; W), \dots, f(x_n; W))^\top$, and define the residual $e(W) = y - u(W)$. The training objective is the unnormalized empirical squared loss

$$L(W) = \frac{1}{2}\|u(W) - y\|_2^2 = \frac{1}{2}\|e(W)\|_2^2. \tag{2}$$

We adopt the derivative convention $\sigma'(z) = \mathbb{1}\{z \ge 0\}$, including at $z = 0$. The resulting selected gradient $G(W) \in \mathbb{R}^{d\times m}$ has columns

$$[G(W)]_{\cdot,r} = -\frac{a_r}{\sqrt{m}} \sum_{i=1}^{n} e(W)_i \, \mathbb{1}\{w_r^\top x_i \ge 0\} \, x_i. \tag{3}$$

Our objective is to bound the number of full-batch Muon updates required to reach $L(W) \le \varepsilon$ for a prescribed $\varepsilon > 0$.

2.2 Momentum and the Newton–Schulz Update

Muon constructs each update by applying a finite Newton–Schulz map to a momentum buffer. For a nonzero matrix $B \in \mathbb{R}^{d\times m}$, initialize $X_0 = B/\|B\|_F$ and perform

$$X_{k+1} = p_k(X_k X_k^\top) X_k, \qquad k = 0, \dots, K-1, \tag{4}$$

where p_k is a polynomial. Define $\mathrm{NS}_K(B) = X_K$ and $\mathrm{NS}_K(0) = 0$. Muon uses $K = 5$ with the same tuned polynomial at every step (Jordan et al., 2024):

$$p_k(s) = 3.4445 - 4.7750s + 2.0315s^2, \qquad k = 0, \dots, 4. \tag{5}$$

This map transforms the singular values of the normalized buffer while retaining its singular-vector directions. It does not, in general, return the exact polar factor. We formulate the analysis for general polynomial schedules and specialize the resulting bounds to Muon's tuned map.

Algorithm 1 specifies the full-batch trajectory. Momentum is accumulated *before* the Newton–Schulz transformation; orthogonalizing individual gradients and then averaging them would define a different update.

We write $u_t = u(W_t)$ and $e_t = e(W_t)$. Reaching the target loss is equivalent to $\|e_t\|_2 \le \sqrt{2\varepsilon}$. The analysis concerns the exact-arithmetic Newton–Schulz map; finite-precision effects and denominator regularization used in numerical implementations are outside its scope.

2.3 Kernel and Update Conditions

The network Jacobian $J(W) \in \mathbb{R}^{n\times dm}$, defined using the derivative convention above, has blocks

$$[J(W)]_{i,(r-1)d+1:rd} = \frac{a_r}{\sqrt{m}} \mathbb{1}\{w_r^\top x_i \ge 0\} \, x_i^\top. \tag{6}$$

Consequently, $\mathrm{vec}(G(W)) = -J(W)^\top e(W)$. Define the empirical neural tangent kernel as $H(W) = J(W)J(W)^\top$. Under the specified initialization, its expectation is $\mathbb{E}[H(W_0)] = \alpha^2 H^\infty$, where

$$H^\infty_{ij} = x_i^\top x_j \, \frac{\pi - \arccos\big(x_i^\top x_j\big)}{2\pi}. \tag{7}$$

We assume that this normalized limiting kernel is positive definite:

$$\lambda_0 := \lambda_{\min}(H^\infty) > 0. \tag{8}$$

Since $H^\infty_{ii} = 1/2$, necessarily $\lambda_0 \le 1/2$. Kernel positivity relates the gradient magnitude to the residual through $\|G(W)\|_F^2 = e(W)^\top H(W) e(W)$. For Muon, turning this relation into loss decrease also requires control of the transformed momentum update.

To describe that transformation, define $\phi_k(x) = x\, p_k(x^2)$. The Newton–Schulz recurrence acts on the normalized singular values through

$$\Phi = \phi_{K-1} \circ \cdots \circ \phi_0. \tag{9}$$

In particular, if $B = U\Sigma V^\top$ is a compact singular value decomposition of a nonzero buffer, then

$$\mathrm{NS}_K(B) = U\, \Phi\!\left(\frac{\Sigma}{\|B\|_F}\right) V^\top,$$

where Φ acts entrywise on the diagonal.

Assumption 1 (Newton–Schulz schedule)**.** *The composed scalar map satisfies*

$$\kappa := \inf_{0<x\le 1} \frac{\Phi(x)}{x} > 0, \qquad M := \max_{0\le x\le 1} \Phi(x). \tag{10}$$

The constant κ controls alignment with the input buffer, while M bounds the update's spectral norm. Indeed, for every nonzero B,

$$\langle B, \mathrm{NS}_K(B)\rangle \ge \kappa \|B\|_F, \qquad \|\mathrm{NS}_K(B)\|_2 \le M.$$

These conditions do not require the transformed singular values to be close to one. Proposition 1 establishes their validity for Muon's tuned five-step map. The next section uses these bounds together with kernel stability to transfer alignment from the momentum buffer to the current gradient and establish the training guarantee.

3 Finite-Time Training Guarantee

We establish a training-loss guarantee for any Newton–Schulz schedule satisfying Assumption 1, including Muon's tuned five-step map. Starting from $B_0 = 0$, consider

$$G_t = G(W_{t-1}), \quad B_t = \mu B_{t-1} + G_t, \quad O_t = \mathrm{NS}_K(B_t), \quad W_t = W_{t-1} - \eta O_t. \tag{11}$$

For $\delta \in (0,1)$, define the initial-residual bound

$$R_0 := \sqrt{\frac{3(\|y\|_2^2 + n\alpha^2/2)}{\delta}}. \tag{12}$$

Theorem 3.1 (Finite-time training guarantee)**.** *Consider the model and initialization of Section 2, and suppose Assumption 1 holds with constants κ and M. Fix $\varepsilon > 0$, $\delta \in (0,1)$, and $\mu \in [0,1)$. If*

$$m \ge 5 \cdot 10^{12} \left(\frac{M}{\kappa}\right)^8 \frac{n^4}{\lambda_0^4 \delta^3} \left(n + \frac{\|y\|_2^2}{\alpha^2}\right), \tag{13}$$

run (11) with

$$\eta = (1-\mu)\frac{\kappa\sqrt{\lambda_0 \varepsilon}}{8M^2 n\alpha}, \tag{14}$$

$$T = \left\lceil \frac{128 M^2 n R_0}{(1-\mu)\kappa^2 \lambda_0 \sqrt{\varepsilon}} \right\rceil. \tag{15}$$

Then, with probability at least $1-\delta$ over initialization,

$$\min_{0\le t\le T} L(W_t) \le \varepsilon. \tag{16}$$

Thus, stopping at the first iterate with loss at most ε requires no more than T updates.

For Muon's tuned coefficients, Proposition 1 permits $\kappa = 2/3$ and $M = 5/4$. The sufficient width is independent of both ε and μ, whereas the learning rate and iteration budget depend explicitly on them. The result is a target-dependent hitting-time guarantee; it does not assert convergence to zero under one fixed learning rate. The proof is given in Section 4.

Scope: what is specific to Muon. The theorem uses the Newton–Schulz map only through κ and M. It therefore also covers $K = 0$, where $\mathrm{NS}_0(B) = B/\|B\|_F$ and $\Phi(x) = x$: Frobenius-normalized momentum gradient descent satisfies Assumption 1 with $\kappa = M = 1$. What is specific to Muon is that its tuned five-step map satisfies the assumption, with $\kappa = 2/3$ and $M = 5/4$, without its transformed singular values being close to one, that is, without approximating the polar factor.

Why $\varepsilon^{-1/2}$. Since $\langle B_t, O_t\rangle \le \|B_t\|_F\|O_t\|_F$, the alignment bound gives $\|O_t\|_F \ge \kappa$ whenever $B_t \ne 0$: the update does not shrink with the gradient. With a fixed η, the descent inequality (20) then certifies progress only down to a residual of order η, because of its term $nM^2\alpha^2\eta^2$. Reaching loss ε therefore requires $\eta \propto \sqrt{\varepsilon}$, and the iteration count scales as $1/\eta$. Gradient descent differs because its steps shrink with the gradient, which gives a linear rate with a constant step size (Du et al., 2018b).

3.1 Stability Near Initialization

The analysis controls the trajectory within

$$\mathbb{B} := \left\{ W : \|W - W_0\|_F \le \frac{18M\sqrt{n}\,R_0}{\kappa\alpha\sqrt{\lambda_0}} \right\}. \tag{17}$$

Under (13), Lemma B.1 establishes an event $\mathcal{E}$ of probability at least $1 - \delta$ controlling the initial residual, kernel concentration, and the number of neuron–input pairs near an activation boundary. On this event, Lemma B.2 gives, uniformly for $W, V \in \mathbb{B}$,

$$\lambda_{\min}(H(W)) \ge \frac{\alpha^2\lambda_0}{2}, \qquad \|u(V) - u(W) - J(W)\,\mathrm{vec}(V - W)\|_2 \le \frac{\kappa\alpha\sqrt{\lambda_0}}{18M}\|V - W\|_2.$$

The full statements and proofs appear in Appendices B.1 and B.2. Kernel positivity relates the gradient magnitude to the residual, while the spectral-norm remainder bound controls the prediction error produced by an update. The trajectory containment argument uses the update norm and total iteration budget, without first assuming loss decrease.

3.2 From Momentum Alignment to Descent

By (3), every gradient and momentum buffer has its column space in $\mathrm{span}\{x_1, \dots, x_n\}$. The Newton–Schulz conditions therefore yield

$$\langle B_t, O_t\rangle \ge \kappa\|B_t\|_F, \qquad \|O_t\|_2 \le M, \qquad \|O_t\|_F \le M\sqrt{n}. \tag{18}$$

The Frobenius bound follows from the rank bound $\mathrm{rank}(B_t) \le n$; the complete statement is Lemma B.3 in Appendix B.3. These bounds control alignment with the buffer. The central step transfers that alignment to the current gradient.

Lemma 3.2 (Gradient-update alignment)**.** *Assume $\mathcal{E}$, let $W_0, \dots, W_{t-1} \in \mathbb{B}$, and suppose $\|e_{t-1}\|_2 > \sqrt{2\varepsilon}$. If*

$$\eta \le (1 - \mu)\frac{\kappa\sqrt{\lambda_0\varepsilon}}{8M^2n\alpha},$$

then

$$\langle G_t, O_t\rangle \ge \frac{\kappa}{2}\|G_t\|_F. \tag{19}$$

The proof controls the discrepancy between the normalized momentum buffer and the current gradient using the stability estimates within $\mathbb{B}$; see Appendix B.5. The learning-rate factor $1 - \mu$ compensates for the accumulation of older gradients.

Lemma 3.3 (One-step residual decrease)**.** *Assume $\mathcal{E}$, let $W_{t-1}, W_t \in \mathbb{B}$, and suppose (19) holds. Then*

$$\|e_t\|_2^2 \le \|e_{t-1}\|_2^2 - \frac{\kappa\alpha\eta\sqrt{\lambda_0}}{4}\|e_{t-1}\|_2 + nM^2\alpha^2\eta^2. \tag{20}$$

If, additionally,

$$\eta \le \frac{\kappa\sqrt{\lambda_0}\|e_{t-1}\|_2}{8M^2n\alpha}, \tag{21}$$

then

$$\|e_t\|_2 \le \|e_{t-1}\|_2 - \frac{\kappa\alpha\eta\sqrt{\lambda_0}}{16}. \tag{22}$$

The proof appears in Appendix B.4. The prescribed learning rate satisfies (21) whenever $L(W_{t-1}) > \varepsilon$. Thus, containment in $\mathbb{B}$ and gradient-update alignment yield a fixed decrease in residual norm until the target is reached; summing this decrease gives the budget in (15).

The exact polar factor maximizes buffer alignment per unit spectral norm (Proposition 3 in the appendix). Our guarantee requires only the bounds in (18), which the tuned finite-step map satisfies without needing to approximate the polar factor closely.

4 Proof of the Finite-Time Guarantee

The proof first establishes trajectory containment using only the update norm and iteration budget. This makes the uniform kernel bounds available without assuming loss decrease. Gradient-update alignment then yields a fixed residual decrease until the target is reached.

Proof of Theorem 3.1. By Lemma B.1, the event $\mathcal{E}$ has probability at least $1-\delta$. Fix an initialization in $\mathcal{E}$, so that $\|e_0\|_2 \le R_0$. If $R_0 \le \sqrt{2\varepsilon}$, the initial iterate already satisfies the conclusion. Hence assume $R_0 > \sqrt{2\varepsilon}$.

Trajectory containment. Every momentum buffer has its column space in $\mathrm{span}\{x_1, \dots, x_n\}$. Thus, Lemma B.3 gives

$$\|W_t - W_{t-1}\|_F = \eta\|O_t\|_F \le \eta M\sqrt{n}. \tag{23}$$

Using $\lceil z \rceil \le z+1$ and the prescribed learning rate and budget,

$$T\eta \le \frac{16R_0}{\kappa\alpha\sqrt{\lambda_0}} + \eta \le \frac{17R_0}{\kappa\alpha\sqrt{\lambda_0}}. \tag{24}$$

For the last inequality, observe that

$$\eta\kappa\alpha\sqrt{\lambda_0} = \frac{(1-\mu)\kappa^2\lambda_0\sqrt{\varepsilon}}{8M^2n} \le \frac{\sqrt{\varepsilon}}{16} < R_0,$$

where $\kappa \le M$, $\lambda_0 \le 1/2$, and $n \ge 1$. Consequently, for every $0 \le t \le T$,

$$\|W_t - W_0\|_F \le t\eta M\sqrt{n} \le \frac{17M\sqrt{n}\,R_0}{\kappa\alpha\sqrt{\lambda_0}} < \frac{18M\sqrt{n}\,R_0}{\kappa\alpha\sqrt{\lambda_0}}. \tag{25}$$

Thus $W_0, \dots, W_T \in \mathbb{B}$, and Lemma B.2 applies throughout the budget.

Descent until the target is reached. Suppose, for contradiction, that

$$\|e_t\|_2 > \sqrt{2\varepsilon} \qquad \text{for every } 0 \le t \le T. \tag{26}$$

For each $1 \le t \le T$, trajectory containment and the prescribed learning rate allow us to apply Lemma 3.2, yielding $\langle G_t, O_t\rangle \ge (\kappa/2)\|G_t\|_F$. Moreover,

$$\eta \le \frac{\kappa\sqrt{\lambda_0}\sqrt{\varepsilon}}{8M^2n\alpha} < \frac{\kappa\sqrt{\lambda_0}\|e_{t-1}\|_2}{8M^2n\alpha}.$$

Hence Lemma 3.3 gives

$$\|e_t\|_2 \le \|e_{t-1}\|_2 - \frac{\kappa\alpha\eta\sqrt{\lambda_0}}{16}. \tag{27}$$

Summing over the budget and substituting (14) yields

$$\begin{aligned}\|e_T\|_2 &\le R_0 - T\frac{\kappa\alpha\eta\sqrt{\lambda_0}}{16} \\ &= R_0 - T\frac{(1-\mu)\kappa^2\lambda_0\sqrt{\varepsilon}}{128M^2n} \le 0,\end{aligned} \tag{28}$$

where the last inequality follows from (15). This contradicts (26). Therefore some $t \le T$ satisfies $L(W_t) \le \varepsilon$. Since the argument holds on $\mathcal{E}$, the conclusion has probability at least $1-\delta$. □

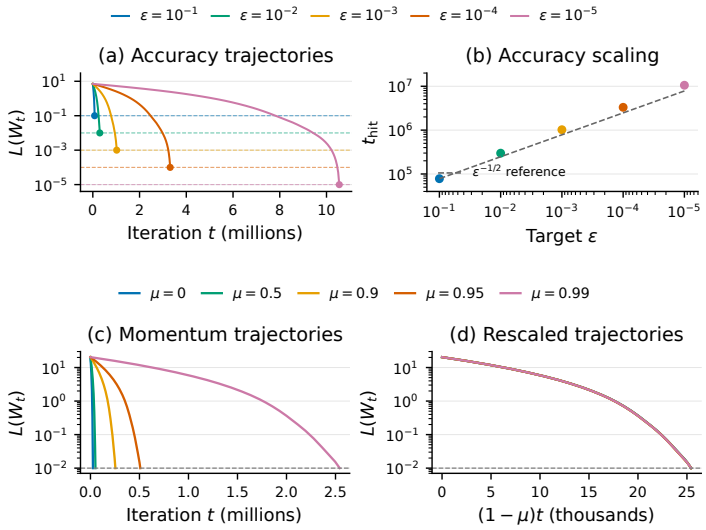

Figure 1: **Training loss across accuracy targets and momentum values.** (a) Loss curves for five target values of ε at $\mu = 0.95$; dots mark the first iteration reaching each target. (b) Hitting time versus ε, with an $\varepsilon^{-1/2}$ reference line. (c) Loss curves for five momentum values at target 10^{-2}, using $\eta = (1-\mu)\eta_0$ with $\eta_0 = 1.54 \times 10^{-4}$. (d) The curves in (c) plotted against $(1-\mu)t$.

5 Numerical Evaluation of the Theory

We examine the accuracy and momentum dependence of the training dynamics, gradient-update alignment, and kernel stability across widths. The experiments use widths far below the sufficient condition in Theorem 3.1: for our dataset, (13) requires $m \gtrsim 2 \times 10^{27}$ and (15) gives $T \approx 1.6 \times 10^8$ at $\mu = 0.95$ and $\varepsilon = 10^{-2}$, whereas the width sweep uses widths 64–2000 and its hitting times at the same μ and ε are below 7.3×10^5. The theoretical budget and stability bounds thus serve as references rather than certified guarantees for these runs. Detailed protocols are provided in Appendix E.

Setup. We train a two-layer ReLU student on unit-normalized Gaussian inputs with labels generated by an independent, fixed ReLU teacher. Only the student's first layer is trained, using full-batch Muon with five tuned Newton–Schulz steps evaluated in double precision. Unless otherwise stated, $n = 20$, $d = 100$, $m = 1000$, $\alpha = 1$, and $\delta = 0.05$. We compute λ_0 from the limiting kernel and R_0 from (12), then prescribe η and the reference budget T using (14) and (15). The accuracy panels use $\lambda_0 \approx 0.34$ and $R_0 \approx 26$; the saved runs used in the remaining diagnostics have $\lambda_0 = 0.335$ and $R_0 = 30.2$. We record the first target crossing, $t_{\text{hit}} = \min\{t : L(W_t) \le \varepsilon\}$.

Accuracy and momentum scaling. The accuracy sweep fixes $\mu = 0.95$ and uses $\eta \propto \sqrt{\varepsilon}$. The panels show target attainment before the reference budget and a trend consistent with its $\varepsilon^{-1/2}$ dependence (Figure 1a,b). This compares separate runs with different learning rates.

With $\eta = (1-\mu)\eta_0$, larger momentum requires more iterations, and plotting against $(1-\mu)t$ approximately collapses the trajectories (Figure 1c,d). This reflects the joint change in momentum and learning rate. A separate fixed-step comparison uses $\eta = \eta_0$ for 50,000 iterations: all three seeds at each displayed momentum reach 10^{-2} in 14,680–25,444 updates. Thus the factor $1-\mu$ remains a sufficient choice; these runs do not establish necessity.

Gradient-update alignment. We directly examine the mechanism of Lemma 3.2 by measuring gradient-update alignment and the discrepancy between the current gradient and normalized momentum:

$$A_t = \frac{\langle G_t, O_t \rangle}{\|G_t\|_F}, \qquad D_t = \frac{\|B_t/s_t - G_t\|_F}{\|G_t\|_F}, \qquad s_t = \frac{1-\mu^t}{1-\mu}. \tag{29}$$

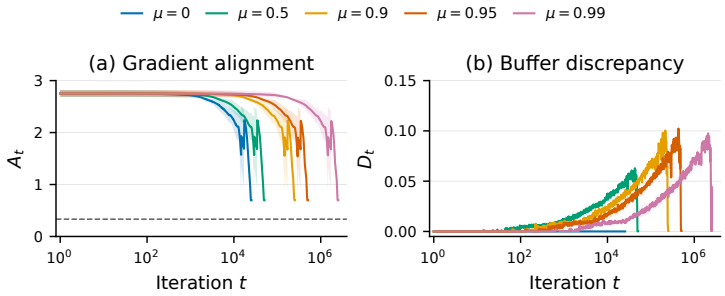

Figure 2: **Gradient alignment and momentum discrepancy.** Both panels plot iteration t (logarithmic axis) for five momentum values μ, each run until it first reaches the target 10^{-2}. (a) Alignment $A_t = \langle G_t, O_t \rangle / \|G_t\|_F$ between the gradient and the update; the dashed line is the reference $\kappa/2 = 1/3$. (b) Buffer discrepancy $D_t = \|B_t/s_t - G_t\|_F / \|G_t\|_F$, which is identically zero at $\mu = 0$.

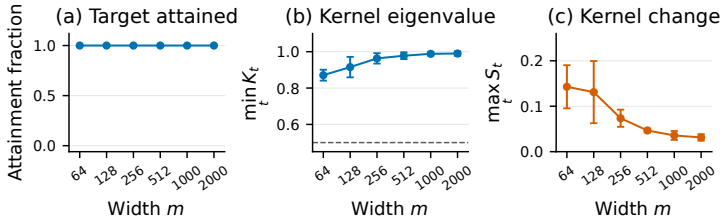

Figure 3: **Training success and kernel stability across widths.** (a) Target attainment within 1,414,000 updates; (b) minimum K_t, with reference $1/2$; (c) maximum S_t. Points and bars show mean ± 1 sample standard deviation over five seeds per width, measured through target attainment.

Here B_t/s_t is the weighted average of stored gradients. The theorem gives the reference $A_t \geq \kappa/2 = 1/3$ under its assumptions, while D_t measures the buffer discrepancy involved in transferring alignment to the current gradient. We record both quantities before target attainment for several momentum values.

Alignment remains about 2.7 for most of training and declines near the target; its smallest recorded value is 0.695, above $1/3$. The discrepancy is zero at $\mu = 0$ and has run-wise maxima in $[0.066, 0.073]$ at $\mu = 0.5$ and $[0.105, 0.124]$ for $\mu \geq 0.9$. The alignment decrease also occurs at $\mu = 0$, so it cannot be attributed solely to momentum discrepancy.

Width and kernel stability. We vary student width while keeping the dataset, teacher, target, momentum, and prescribed learning rate fixed. Across repeated student initializations, we record target attainment and monitor

$$K_t = \frac{\lambda_{\min}(H(W_t))}{\alpha^2 \lambda_0}, \qquad S_t = \frac{\|H(W_t) - H(W_0)\|_2}{\|H(W_0)\|_2}. \tag{30}$$

The reference $K_t = 1/2$ corresponds to the kernel lower bound used in the proof; S_t measures relative kernel variation. Figure 3 summarizes these quantities over each run. We also measure displacement relative to the analytical radius in (17) (Figure 5).

Every one of these 30 runs reaches 10^{-2}. The run-wise minimum of K_t has mean 0.871 ± 0.030 at width 64 and 0.990 ± 0.012 at width 2000, and the smallest value on any run is 0.826, above the reference $1/2$. The run-wise maximum of S_t falls from 0.143 ± 0.047 at width 64 to 0.031 ± 0.007 at width 2000. No run fails to reach the target. These measurements examine behavior below the sufficient width threshold; they do not establish that threshold's necessity or tightness.

Finite maps and parameter movement. Figure 4 compares the tuned five-step map, five classical cubic Newton–Schulz steps, and the exact polar factor on 59 shared training buffers. For a buffer B

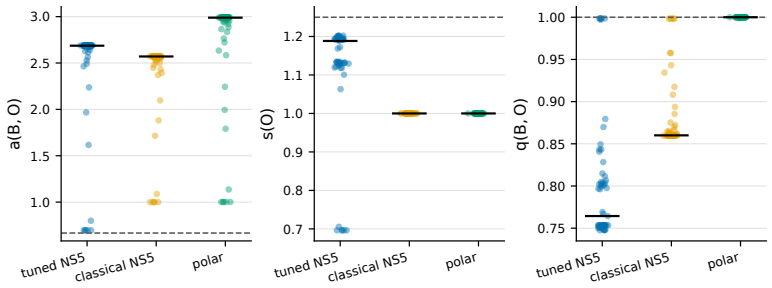

Figure 4: **Finite maps on shared training buffers.** Left to right: buffer alignment a, spectral norm s, and normalized alignment per spectral norm q. Points are 59 saved buffers from student seed 10; black ticks mark medians. Dashed references are $2/3$, $5/4$, and the polar optimum 1, respectively. Classical NS5 uses the cubic recurrence.

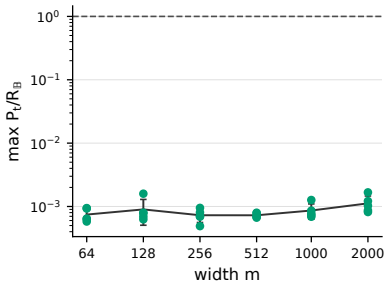

Figure 5: **Movement within the analytical radius.** Maximum $\|W_t - W_0\|_F / R_{\mathbb{B}}$ up to target attainment, with $R_{\mathbb{B}}$ from (17). Points show five seeds per width; the line and bars show their mean and ± 1 sample standard deviation. The dashed line is the ball boundary.

with update O, we record the alignment $a = \langle B, O\rangle / \|B\|_F$, the spectral norm $s = \|O\|_2$, and the ratio $q = \langle B, O\rangle / (\|B\|_* \|O\|_2)$, whose maximum 1 is attained by the polar factor. The tuned map satisfies $a \geq 2/3$ and $s \leq 5/4$ on every buffer, as Assumption 1 requires. The median q is 0.764 for the tuned map and 0.860 for the classical map. These are diagnostics of the maps; they do not compare training speed or distance to the polar factor.

In every width-sweep run, the target is reached with displacement at most $1.67 \times 10^{-3} R_{\mathbb{B}}$, where $R_{\mathbb{B}} = 7.88 \times 10^3$. The iterates thus remain well inside $\mathbb{B}$ and the kernel remains positive, consistent with the containment step of the proof. These measurements do not verify the uniform Jacobian and remainder bounds at these widths.

6 Conclusion

We proved a finite-time training-loss guarantee for full-batch Muon with momentum followed by five tuned Newton–Schulz iterations. For sufficiently wide two-layer ReLU networks with fixed random output weights, a learning rate proportional to $(1-\mu)\sqrt{\varepsilon}$ attains any prescribed loss ε within $\mathcal{O}\big((1-\mu)^{-1}\varepsilon^{-1/2}\big)$ iterations, with probability at least $1-\delta$ over initialization. The proof uses only two properties of the finite-step map, alignment with the momentum buffer and a bounded spectral norm. The tuned map has both without approximating the polar factor, so exact orthogonalization is unnecessary.

The experiments are consistent with these mechanisms at widths far below the sufficient threshold ($m \gtrsim 2 \times 10^{27}$ for our dataset): gradient-update alignment stays above $1/3$, the kernel ratio K_t above $1/2$, displacement below $1.7 \times 10^{-3} R_{\mathbb{B}}$, and all 30 width-sweep runs reach the target. They do not test necessity or tightness. On shared buffers, classical Newton–Schulz has a higher median alignment per spectral norm than the tuned map, which does not by itself imply faster training.

Future work. Four directions remain: (i) sharper width and momentum requirements, and whether momentum can accelerate Muon as it does gradient descent in Wang et al. (2020); (ii) analysis beyond the near-initialization regime, in which the features move; (iii) minibatch gradients and finite-precision arithmetic; and (iv) deeper networks and joint training of both layers.

REPRODUCIBILITY STATEMENT

The network, initialization, selected ReLU derivative, and full-batch loss are specified in Section 2. Algorithm 1 is the Muon update analyzed in the paper, and equations (4)–(5) fix all five Newton–Schulz steps; Assumption 1 states the only properties of the Newton–Schulz steps that the proof uses, and Appendix B.6 verifies them for the coefficients of Jordan et al. (2024). Equations (12)–(15) specify the width, momentum range, learning rate, and update budget. No empirical result is used in the proof.

AI USE STATEMENT

AI tools assisted with manuscript drafting, language editing, and LaTeX preparation. The authors remain responsible for the mathematical claims and reported experimental results.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/allen-zhu19a.html.

Sanjeev Arora, Simon Shaolei Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *ArXiv*, abs/1901.08584, 2019. URL https://api.semanticscholar.org/CorpusID:59222746.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *ArXiv*, abs/2409.20325, 2024a. URL https://api.semanticscholar.org/CorpusID:272987272.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. *ArXiv*, abs/2410.21265, 2024b. URL https://api.semanticscholar.org/CorpusID:273655185.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Trans. Mach. Learn. Res.*, 2026, 2025. URL https://api.semanticscholar.org/CorpusID:279447438.

Simon Shaolei Du, J. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *ArXiv*, abs/1811.03804, 2018a. URL https://api.semanticscholar.org/CorpusID:53250419.

Simon Shaolei Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *ArXiv*, abs/1810.02054, 2018b. URL https://api.semanticscholar.org/CorpusID:52920808.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

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

Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with Newton–Schulz. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2601.19156.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *ArXiv*, abs/1808.01204, 2018. URL https://api.semanticscholar.org/CorpusID:51920936.

Jingyuan Liu, Jianling Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yi-Bo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Meng Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training. *ArXiv*, abs/2502.16982, 2025. URL https://api.semanticscholar.org/CorpusID:276575212.

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

Jun-Kun Wang, Chi-Heng Lin, and Jacob D. Abernethy. A modular analysis of provable acceleration via polyak's momentum: Training a wide relu network and a deep linear network. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:235359199.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *Advances in neural information processing systems*, 32, 2019.

A Comparison with Prior Guarantees

Table 1: Non-asymptotic guarantees for wide-network training and Muon. The first eight rows follow the raw gradient: GD for the first five and SGD for the next three. Networks have depth 2 except in Du et al. (2018a); Zou & Gu (2019); Allen-Zhu et al. (2019); Cao & Gu (2019), which allow depth > 2. The last three rows use momentum followed by, respectively, the exact polar factor, classical Newton–Schulz (NS), and five tuned NS steps. The two prior Muon results concern general L-smooth objectives. Population-risk and stationarity guarantees are not directly comparable to training-loss guarantees. A dash denotes an unstated explicit bound. For our row, α and $\|y\|_2^2/n$ are treated as constants and $T = \mathcal{O}\big(n^{3/2}/((1-\mu)\lambda_0\sqrt{\delta\varepsilon})\big)$.

Reference	Guarantee	Width
Du et al. (2018b)	$L(W_T) \to 0$, linear	$\Omega(n^6/(\lambda_0^4\delta^3))$
Arora et al. (2019)	$L(W_T) \le \varepsilon$	$\widetilde{\Omega}(n^2/\lambda_0^2)$
Ji & Telgarsky (2019)	test error $\le \varepsilon$	$\mathrm{polylog}(1/\varepsilon)$
Du et al. (2018a)	$L(W_T) \to 0$, linear	– ($\mathrm{poly}(n)$)
Zou & Gu (2019)	$L(W_T) \to 0$, faster	< Allen-Zhu et al.
Li & Liang (2018)	pop. risk $\to 0$	margin-dependent
Allen-Zhu et al. (2019)	$L(W_T) \to 0$, poly time	$\Omega(n^{24})$
Cao & Gu (2019)	pop. risk $\le \widetilde{\mathcal{O}}(n^{-1/2})$	any width
Chen et al. (2025)	KKT score $\mathcal{O}(1/\sqrt{T})$	–
Kim & Oh (2026)	stationarity	–
This work	$\min_{0\le t\le T} L(W_t) \le \varepsilon$	$\Omega\big(n^5/(\lambda_0^4\delta^3)\big)$

B Supporting Lemmas and Proofs

B.1 Initialization

Lemma B.1 (Initialization event)**.** *Let*

$$\begin{aligned}
\mathcal{E}_1 &:= \{\|e_0\|_2 \le R_0\},\\
\mathcal{E}_2 &:= \left\{\|H(W_0) - \alpha^2 H^\infty\|_2 \le \frac{\alpha^2\lambda_0}{4}\right\},\\
\mathcal{E}_3 &:= \left\{\frac{1}{m}\sum_{i=1}^{n}\sum_{r=1}^{m} \mathbb{1}\Big\{|w_r(0)^\top x_i| \le \frac{\delta\kappa^2\lambda_0}{1620M^2n}\Big\} \le \frac{\kappa^2\lambda_0}{648M^2}\right\}.
\end{aligned} \tag{31}$$

Under the width condition (13), each complement has probability at most $\delta/3$. Hence $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ satisfies $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$.

Proof of Lemma B.1. We show that each of the three events fails with probability at most $\delta/3$; a union bound then gives $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$.

Part 1. For $Z \sim \mathcal{N}(0,1)$, $\mathbb{E}\sigma(Z)^2 = \mathbb{E}[Z^2\mathbb{1}\{Z \ge 0\}] = 1/2$. Since the a_r are independent of W_0 with $\mathbb{E}a_r = 0$ and $a_r^2 = \alpha^2$, and $w_r(0)^\top x_i \sim \mathcal{N}(0,1)$,

$$\mathbb{E}u(W_0)_i = 0, \qquad \mathbb{E}u(W_0)_i^2 = \frac{1}{m}\sum_{r=1}^{m}\alpha^2\,\mathbb{E}\sigma(w_r(0)^\top x_i)^2 = \frac{\alpha^2}{2}. \tag{32}$$

Then

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^c) &= \mathbb{P}(\|e_0\|_2^2 > R_0^2) \\
&\le \frac{\mathbb{E}\|e_0\|_2^2}{R_0^2} && \text{[Markov's inequality]} \\
&= \frac{\|y\|_2^2 + n\alpha^2/2}{R_0^2} && \text{[cross term vanishes by (32)]} \\
&= \frac{\delta}{3} && \text{[definition (12) of } R_0\text{]}. \qquad (33)
\end{aligned}$$

Part 2. For fixed i, j,

$$\frac{H(W_0)_{ij}}{\alpha^2} = \frac{1}{m}\sum_{r=1}^{m} Z_r^{ij}, \qquad Z_r^{ij} := x_i^\top x_j \, \mathbb{1}\{w_r(0)^\top x_i \ge 0,\ w_r(0)^\top x_j \ge 0\}. \qquad (34)$$

The Z_r^{ij} are independent over r, have mean H_{ij}^∞ by (7), and take values in an interval of length $|x_i^\top x_j| \le 1$. Since $\log z \le z$, $\kappa \le M$, $\lambda_0 \le \frac{1}{2}$ and $\delta < 1$, the width condition (13) gives $m \ge 48n^4/(\lambda_0^2\delta) \ge (8n^2/\lambda_0^2)\log\big(6n^2/\delta\big)$. Hence

$$\begin{aligned}
\mathbb{P}\Big(\max_{i,j}\Big|\frac{H(W_0)_{ij}}{\alpha^2} - H_{ij}^\infty\Big| > \frac{\lambda_0}{4n}\Big) &\le n^2 \cdot 2\exp\Big(-\frac{m\lambda_0^2}{8n^2}\Big) && \text{[union bound, Hoeffding's inequality]} \\
&\le \frac{\delta}{3} && [m \ge (8n^2/\lambda_0^2)\log\big(6n^2/\delta\big)]. \qquad (35)
\end{aligned}$$

Outside this event, $D := H(W_0)/\alpha^2 - H^\infty$ satisfies

$$\begin{aligned}
\|D\|_2 &\le \|D\|_F && \text{[operator norm} \le \text{Frobenius norm]} \\
&\le n \max_{i,j}|D_{ij}| && [D \text{ has } n^2 \text{ entries}] \\
&\le \frac{\lambda_0}{4}, && \qquad (36)
\end{aligned}$$

which is $\mathcal{E}_2$ after multiplying by α^2. Hence $\mathbb{P}(\mathcal{E}_2^c) \le \delta/3$.

Part 3. The density of $\mathcal{N}(0,1)$ is at most $1/\sqrt{2\pi}$, so $\mathbb{P}\big(|w_r(0)^\top x_i| \le \frac{\delta\kappa^2\lambda_0}{1620M^2n}\big) \le \sqrt{2/\pi}\,\frac{\delta\kappa^2\lambda_0}{1620M^2n}$. Let N be the double sum in $\mathcal{E}_3$. Then

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_3^c) &= \mathbb{P}\Big(\frac{N}{m} > \frac{\kappa^2\lambda_0}{648M^2}\Big) \\
&\le \frac{648M^2}{\kappa^2\lambda_0}\,\mathbb{E}\frac{N}{m} && \text{[Markov's inequality]} \\
&\le \frac{648M^2}{\kappa^2\lambda_0}\cdot n\sqrt{\frac{2}{\pi}}\,\frac{\delta\kappa^2\lambda_0}{1620M^2n} && \text{[linearity of expectation]} \\
&\le \frac{\delta}{3} && [\sqrt{2/\pi} \le \tfrac{5}{6}]. \qquad (37)
\end{aligned}$$

The union bound over Parts 1–3 completes the proof. □

B.2 Uniform control in the stability ball

Throughout this subsection, the *gate* of hidden unit r on input x_i at weights $Z = (z_1, \dots, z_m)$ is the indicator $\mathbb{1}\{z_r^\top x_i \ge 0\}$, and the gate *flips* between Z and Z' if $\mathbb{1}\{z_r^\top x_i \ge 0\} \ne \mathbb{1}\{z_r'^\top x_i \ge 0\}$. These indicators are exactly the ones appearing in (3) and (6).

Lemma B.2 (Uniform NTK and ReLU remainder)**.** *On $\mathcal{E}$, for every $W, V \in \mathbb{B}$ and every matrix O,*

$$\lambda_{\min}(H(W)) \geq \frac{\alpha^2 \lambda_0}{2}, \tag{38}$$

$$\|J(W) - J(W_0)\|_F \leq \frac{\kappa\alpha\sqrt{\lambda_0}}{18M}, \tag{39}$$

$$\|(J(W) - J(V))\,\mathrm{vec}(O)\|_2 \leq \frac{\kappa\alpha\sqrt{\lambda_0}}{18M}\,\|O\|_2, \tag{40}$$

$$\|u(V) - u(W) - J(W)\,\mathrm{vec}(V - W)\|_2 \leq \frac{\kappa\alpha\sqrt{\lambda_0}}{18M}\,\|V - W\|_2. \tag{41}$$

The ball is measured in Frobenius norm and the last two bounds in operator norm, because one step moves the weights by $\eta\|O_t\|_2 \leq \eta M$.

Proof of Lemma B.2. Fix an initialization in $\mathcal{E}$ and arbitrary $W = (w_1, \ldots, w_m)$ and $V = (v_1, \ldots, v_m)$ in $\mathbb{B}$. Part 1 counts the gates that can flip; Part 2 turns that count into the Jacobian bound (39) and the kernel bound (38); Parts 3 and 4 turn it into the remainder bound (41) and the bound (40).

Part 1. For $Z \in \{W, V\}$ and each i, define

$$\mathcal{S}_i := \Big\{r : |w_r(0)^\top x_i| \leq \frac{\delta\kappa^2\lambda_0}{1620M^2n}\Big\}, \qquad \mathcal{I}(Z) := \Big\{r : \|z_r - w_r(0)\|_2 > \frac{\delta\kappa^2\lambda_0}{1620M^2n}\Big\}. \tag{42}$$

If $r \notin \mathcal{S}_i \cup \mathcal{I}(Z)$, then, using $\|x_i\|_2 = 1$,

$$|(z_r - w_r(0))^\top x_i| \leq \|z_r - w_r(0)\|_2 \leq \frac{\delta\kappa^2\lambda_0}{1620M^2n} < |w_r(0)^\top x_i|, \tag{43}$$

so $z_r^\top x_i$ and $w_r(0)^\top x_i$ have the same nonzero sign and the gate of unit r on x_i does not flip between W_0 and Z. Hence the set

$$\mathcal{A}_i(W, V) := \{r : \text{the gate of } r \text{ on } x_i \text{ flips between } W_0 \text{ and } W, \text{ or between } W_0 \text{ and } V\} \tag{44}$$

satisfies $\mathcal{A}_i(W,V) \subseteq \mathcal{S}_i \cup \mathcal{I}(W) \cup \mathcal{I}(V)$. Each $r \in \mathcal{I}(Z)$ contributes more than $(\delta\kappa^2\lambda_0/(1620M^2n))^2$ to $\|Z - W_0\|_F^2$, and $Z \in \mathbb{B}$, so

$$|\mathcal{I}(Z)| \leq \Big(\frac{18M\sqrt{n}\,R_0}{\kappa\alpha\sqrt{\lambda_0}}\Big)^2\Big(\frac{1620M^2n}{\delta\kappa^2\lambda_0}\Big)^2 \leq \frac{8.5\cdot 10^8\, M^6 n^3 R_0^2}{\kappa^6\alpha^2\lambda_0^3\delta^2}. \tag{45}$$

Therefore

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^n |\mathcal{A}_i(W,V)| &\leq \frac{1}{m}\sum_{i=1}^n |\mathcal{S}_i| + \frac{n}{m}\big(|\mathcal{I}(W)| + |\mathcal{I}(V)|\big) && \text{[inclusion above]}\\
&\leq \frac{\kappa^2\lambda_0}{648M^2} + \frac{1.70\cdot 10^9\, M^6 n^4 R_0^2}{m\,\kappa^6\alpha^2\lambda_0^3\delta^2} && \text{[on } \mathcal{E}_3\text{; by (45)]}\\
&\leq \frac{\kappa^2\lambda_0}{324M^2} && \text{[width (13), } R_0^2 \leq \tfrac{3\alpha^2}{\delta}\big(n + \tfrac{\|y\|_2^2}{\alpha^2}\big)\text{]}.
\end{aligned} \tag{46}$$

Part 2. Take $V = W_0$. A block of $J(W) - J(W_0)$ is nonzero only when the gate flips, and then has squared norm α^2/m. Hence

$$\|J(W) - J(W_0)\|_F^2 \leq \frac{\alpha^2}{m}\sum_{i=1}^n |\mathcal{A}_i(W, W_0)| \leq \frac{\kappa^2\alpha^2\lambda_0}{324M^2} \tag{47}$$

by (46), which is (39). For every $q \in \mathbb{R}^n$,

$$\begin{aligned}
\|J(W)^\top q\|_2 &\geq \|J(W_0)^\top q\|_2 - \|J(W) - J(W_0)\|_2\|q\|_2 && \text{[reverse triangle inequality]}\\
&\geq \sqrt{q^\top H(W_0) q} - \frac{\kappa\alpha\sqrt{\lambda_0}}{18M}\|q\|_2 && \text{[by (47)]}\\
&\geq \Big(\frac{\sqrt{3}}{2} - \frac{1}{18}\Big)\alpha\sqrt{\lambda_0}\,\|q\|_2 && \text{[on } \mathcal{E}_2\text{; } \kappa \leq M\text{]}.
\end{aligned} \tag{48}$$

Here $\mathcal{E}_2$ gives $\lambda_{\min}(H(W_0)) \geq \frac{3}{4}\alpha^2\lambda_0$. Since $(\sqrt{3}/2 - 1/18)^2 > 1/2$, squaring (48) gives (38).

Part 3. Set

$$\xi := u(V) - u(W) - J(W)\,\mathrm{vec}(V - W), \qquad \xi_i = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r\, \rho\big(w_r^\top x_i, (v_r - w_r)^\top x_i\big), \tag{49}$$

where $\rho(z, h) := \sigma(z + h) - \sigma(z) - \mathbb{1}\{z \geq 0\}h$. The scalar remainder satisfies $|\rho(z, h)| \leq |h|$, and $\rho(z, h) = 0$ unless the gates at z and $z + h$ differ. If neither W nor V flips a gate relative to W_0, the gates at W and V agree, so only $r \in \mathcal{A}_i(W, V)$ contributes. Therefore

$$\begin{aligned}
|\xi_i| &\leq \frac{\alpha}{\sqrt{m}} \sum_{r \in \mathcal{A}_i(W,V)} |(v_r - w_r)^\top x_i| && [|a_r| = \alpha,\ |\rho(z,h)| \leq |h|] \\
&\leq \alpha \sqrt{\frac{|\mathcal{A}_i(W,V)|}{m}}\, \|(V - W)^\top x_i\|_2 && [\text{Cauchy–Schwarz over } r] \\
&\leq \alpha \sqrt{\frac{|\mathcal{A}_i(W,V)|}{m}}\, \|V - W\|_2 && [\|x_i\|_2 = 1].
\end{aligned} \tag{50}$$

Squaring and summing over i,

$$\begin{aligned}
\|\xi\|_2^2 &\leq \alpha^2 \|V - W\|_2^2 \cdot \frac{1}{m} \sum_{i=1}^{n} |\mathcal{A}_i(W,V)| \\
&\leq \frac{\kappa^2 \alpha^2 \lambda_0}{324 M^2}\, \|V - W\|_2^2 && [\text{by (46)}],
\end{aligned} \tag{51}$$

which is (41).

Part 4. Write o_r for the r-th column of O. The i-th entry of $(J(W) - J(V))\,\mathrm{vec}(O)$ is $\frac{1}{\sqrt{m}} \sum_r a_r \big(\mathbb{1}\{w_r^\top x_i \geq 0\} - \mathbb{1}\{v_r^\top x_i \geq 0\}\big) o_r^\top x_i$, and as in Part 3 only $r \in \mathcal{A}_i(W, V)$ contribute. Hence

$$\begin{aligned}
\|(J(W) - J(V))\,\mathrm{vec}(O)\|_2^2 &\leq \sum_{i=1}^{n} \alpha^2 \frac{|\mathcal{A}_i(W,V)|}{m}\, \|O^\top x_i\|_2^2 && [|a_r| = \alpha, \text{ Cauchy–Schwarz over } r] \\
&\leq \frac{\kappa^2 \alpha^2 \lambda_0}{324 M^2}\, \|O\|_2^2 && [\|x_i\|_2 = 1, (46)],
\end{aligned} \tag{52}$$

which is (40). □

B.3 Newton–Schulz steps

Lemma B.3 (Newton–Schulz matrix bounds)**.** *Under Assumption 1, every matrix B whose column space lies in* $\mathrm{span}\{x_1, \ldots, x_n\}$ *satisfies*

$$\langle B, \mathrm{NS}_K(B)\rangle \geq \kappa \|B\|_F, \qquad \|\mathrm{NS}_K(B)\|_2 \leq M, \qquad \|\mathrm{NS}_K(B)\|_F \leq M\sqrt{n}.$$

Proof of Lemma B.3. By the definition (10) of κ and M,

$$\kappa x \leq \Phi(x) \leq M, \qquad x \in [0, 1]. \tag{53}$$

The case $B = 0$ holds by the convention $\mathrm{NS}_K(0) = 0$. Otherwise let $B = U\Sigma V^\top$ be a compact SVD with positive singular values $s_1, \ldots, s_k$. Each step of (4) keeps U, V and applies ϕ_k to the diagonal, so

$$\mathrm{NS}_K(B) = U\,\mathrm{diag}\Big(\Phi\Big(\frac{s_1}{\|B\|_F}\Big), \ldots, \Phi\Big(\frac{s_k}{\|B\|_F}\Big)\Big) V^\top, \qquad \frac{s_j}{\|B\|_F} \in (0, 1]. \tag{54}$$

By (53) the diagonal entries lie in $[0, M]$, so $\|\mathrm{NS}_K(B)\|_2 \leq M$. The column space of B lies in $\mathrm{span}\{x_1, \ldots, x_n\}$, so $k \leq n$ and $\|\mathrm{NS}_K(B)\|_F \leq M\sqrt{k} \leq M\sqrt{n}$. Finally,

$$\begin{aligned}
\langle B, \mathrm{NS}_K(B)\rangle &= \sum_{j=1}^{k} s_j\, \Phi\Big(\frac{s_j}{\|B\|_F}\Big) && [\text{shared singular vectors in (54)}] \\
&\geq \frac{\kappa}{\|B\|_F} \sum_{j=1}^{k} s_j^2 && [\text{lower bound in (53)}] \\
&= \kappa \|B\|_F && [\textstyle\sum_j s_j^2 = \|B\|_F^2].
\end{aligned} \tag{55}$$

□

B.4 One-step descent

Proof of Lemma 3.3. Every buffer column lies in $\operatorname{span}\{x_1, \dots, x_n\}$, so Lemma B.3 applies to B_t, and $W_t - W_{t-1} = -\eta O_t$. Part 1 bounds how far the predictions move; Part 2 expands the new residual and proves (20); Part 3 converts it into (22).

Part 1. For each i,

$$\begin{aligned} |u(W_t)_i - u(W_{t-1})_i| &\le \frac{\alpha\eta}{\sqrt{m}} \sum_{r=1}^{m} |(O_t)_{\cdot,r}^\top x_i| && [\sigma \text{ is 1-Lipschitz}, |a_r| = \alpha] \\ &\le \alpha\eta \, \|O_t^\top x_i\|_2 && [\text{Cauchy–Schwarz over } r] \\ &\le M\alpha\eta && [\|x_i\|_2 = 1, \text{Lemma B.3}], \end{aligned} \tag{56}$$

so $\|u(W_t) - u(W_{t-1})\|_2 \le \sqrt{n}\, M\alpha\eta$. This bound holds at every step, whether or not the iterates lie in $\mathbb{B}$.

Part 2. By hypothesis (19) and the kernel bound (38),

$$\langle G_t, O_t \rangle \ge \frac{\kappa}{2} \|G_t\|_F = \frac{\kappa}{2} \big(e_{t-1}^\top H(W_{t-1}) e_{t-1} \big)^{1/2} \ge \frac{\kappa\alpha\sqrt{\lambda_0}}{2\sqrt{2}} \, \|e_{t-1}\|_2. \tag{57}$$

Let $\xi := u(W_t) - u(W_{t-1}) + \eta J(W_{t-1}) \operatorname{vec}(O_t)$ be the linearization remainder of the step. Both iterates lie in $\mathbb{B}$, so (41) and Lemma B.3 give $\|\xi\|_2 \le \frac{\kappa\alpha\sqrt{\lambda_0}}{18M} \cdot M\eta = \frac{\kappa\alpha\eta\sqrt{\lambda_0}}{18}$. Moreover,

$$-\eta \, e_{t-1}^\top J(W_{t-1}) \operatorname{vec}(O_t) = -\eta \, (J(W_{t-1})^\top e_{t-1})^\top \operatorname{vec}(O_t) = \eta \langle G_t, O_t \rangle. \tag{58}$$

Since $e_t = e_{t-1} - (u(W_t) - u(W_{t-1}))$,

$$\begin{aligned} \|e_t\|_2^2 &= \|e_{t-1}\|_2^2 - 2\eta\langle G_t, O_t\rangle - 2e_{t-1}^\top \xi + \|u(W_t) - u(W_{t-1})\|_2^2 && [\text{expand, (58)}] \\ &\le \|e_{t-1}\|_2^2 - \Big(\frac{1}{\sqrt{2}} - \frac{2}{18} \Big) \kappa\alpha\eta\sqrt{\lambda_0} \, \|e_{t-1}\|_2 + nM^2\alpha^2\eta^2 && [(57), \text{Part 1, Cauchy–Schwarz}] \\ &\le \|e_{t-1}\|_2^2 - \frac{\kappa\alpha\eta\sqrt{\lambda_0}}{4} \|e_{t-1}\|_2 + nM^2\alpha^2\eta^2 && [\tfrac{1}{\sqrt{2}} - \tfrac{2}{18} > \tfrac{1}{4}]. \end{aligned} \tag{59}$$

This is (20).

Part 3. Assume (21); since $\eta > 0$ it forces $\|e_{t-1}\|_2 > 0$. Multiplying (21) by $nM^2\alpha^2\eta$ gives $nM^2\alpha^2\eta^2 \le \frac{\kappa\alpha\eta\sqrt{\lambda_0}}{8} \|e_{t-1}\|_2$, so

$$\begin{aligned} \|e_t\|_2^2 &\le \|e_{t-1}\|_2^2 - \frac{\kappa\alpha\eta\sqrt{\lambda_0}}{8} \|e_{t-1}\|_2 && [\text{by (59)}] \\ &= \|e_{t-1}\|_2^2 \Big(1 - \frac{\kappa\alpha\eta\sqrt{\lambda_0}}{8\|e_{t-1}\|_2} \Big). \end{aligned} \tag{60}$$

The left side is nonnegative, so the bracket lies in $[0, 1]$, and

$$\begin{aligned} \|e_t\|_2 &\le \|e_{t-1}\|_2 \Big(1 - \frac{\kappa\alpha\eta\sqrt{\lambda_0}}{16\|e_{t-1}\|_2} \Big) && [\sqrt{1-z} \le 1 - z/2 \text{ on } [0,1]] \\ &= \|e_{t-1}\|_2 - \frac{\kappa\alpha\eta\sqrt{\lambda_0}}{16}, \end{aligned} \tag{61}$$

which is (22). □

B.5 Momentum

Proof of Lemma 3.2. Write $J_k := J(W_k)$. Unrolling $B_t = \mu B_{t-1} + G_t$ down to $B_0 = 0$ splits the buffer into the current gradient and the change of the older ones,

$$B_t = S\, G_t + D, \qquad S := \sum_{j=0}^{t-1} \mu^j \ge 1, \qquad D := \sum_{j=1}^{t-1} \mu^j \big(G_{t-j} - G_t \big). \tag{62}$$

Part 1 reduces the claim to a bound on D, Part 2 bounds one term of D, and Part 3 adds the terms up.

Part 1. The column space of B_t lies in $\operatorname{span}\{x_1, \dots, x_n\}$, so

$$\begin{aligned} S\langle G_t, O_t\rangle &= \langle B_t, O_t\rangle - \langle D, O_t\rangle && \text{[by (62)]} \\ &\geq \kappa\|B_t\|_F - |\langle D, O_t\rangle| && \text{[Lemma B.3]} \\ &\geq \kappa S\|G_t\|_F - \kappa\|D\|_F - |\langle D, O_t\rangle| && \text{[triangle inequality].} \end{aligned} \tag{63}$$

It therefore suffices to show

$$\kappa\|D\|_F + |\langle D, O_t\rangle| \leq \frac{\kappa}{2} S\|G_t\|_F. \tag{64}$$

Part 2. Fix $1 \leq j \leq t-1$ and put $s = t - j$. By the gradient identity, $\operatorname{vec}(G_s - G_t) = -J_{s-1}^{\top}(e_{s-1} - e_{t-1}) - (J_{s-1} - J_{t-1})^{\top} e_{t-1}$. Two facts control the first term. By Part 1 of the proof of Lemma 3.3, the residual moves by at most $\sqrt{n}\, M\alpha\eta$ per step, so $\|e_{s-1} - e_{t-1}\|_2 \leq j\sqrt{n}\, M\alpha\eta$. And for every W and every matrix O, each entry of $J(W)\operatorname{vec}(O)$ is at most $\alpha\|O^{\top}x_i\|_2 \leq \alpha\|O\|_2$ in magnitude, by Cauchy–Schwarz over the hidden units, so $\|J(W)\operatorname{vec}(O)\|_2 \leq \alpha\sqrt{n}\,\|O\|_2$ and $\|J(W)\|_2 \leq \alpha\sqrt{n}$. Since $W_{s-1}, W_{t-1} \in \mathbb{B}$,

$$\begin{aligned} \|G_s - G_t\|_F &\leq \|J_{s-1}\|_2\|e_{s-1} - e_{t-1}\|_2 + \|J_{s-1} - J_{t-1}\|_2\|e_{t-1}\|_2 && \text{[triangle inequality]} \\ &\leq jnM\alpha^2\eta + \frac{2\kappa\alpha\sqrt{\lambda_0}}{18M}\|e_{t-1}\|_2 && \text{[facts above, (39) twice],} \end{aligned} \tag{65}$$

$$\begin{aligned} |\langle G_s - G_t, O_t\rangle| &\leq \|e_{s-1} - e_{t-1}\|_2\|J_{s-1}\operatorname{vec}(O_t)\|_2 \\ &\quad + \|e_{t-1}\|_2\|(J_{s-1} - J_{t-1})\operatorname{vec}(O_t)\|_2 && \text{[Cauchy–Schwarz]} \\ &\leq jnM^2\alpha^2\eta + \frac{\kappa\alpha\sqrt{\lambda_0}}{18}\|e_{t-1}\|_2 && \text{[facts above, (40), } \|O_t\|_2 \leq M\text{].} \end{aligned} \tag{66}$$

Since $\kappa \leq M$, the two bounds combine to

$$\kappa\|G_s - G_t\|_F + |\langle G_s - G_t, O_t\rangle| \leq 2jnM^2\alpha^2\eta + \frac{3\kappa\alpha\sqrt{\lambda_0}}{18}\|e_{t-1}\|_2. \tag{67}$$

Part 3. The average of $j \in \{0, \dots, t-1\}$ under weights μ^j increases with t and tends to $\mu/(1-\mu)$, so $\sum_{j=1}^{t-1} j\mu^j \leq S\mu/(1-\mu)$; also $\sum_{j=1}^{t-1}\mu^j \leq S$. Summing (67) with weights μ^j,

$$\begin{aligned} \kappa\|D\|_F + |\langle D, O_t\rangle| &\leq S\Big(\frac{2\mu}{1-\mu}\, nM^2\alpha^2\eta + \frac{3\kappa\alpha\sqrt{\lambda_0}}{18}\|e_{t-1}\|_2\Big) && \text{[triangle inequality, weight sums]} \\ &\leq S\Big(\frac{\kappa}{4}\,\alpha\sqrt{\lambda_0\varepsilon} + \frac{3\kappa\alpha\sqrt{\lambda_0}}{18}\|e_{t-1}\|_2\Big) && \text{[bound on } \eta, \mu \leq 1\text{]} \\ &\leq \frac{\kappa}{2}\, S\|G_t\|_F && \text{[see below].} \end{aligned} \tag{68}$$

The last line uses the kernel bound (38) at W_{t-1} and $\|e_{t-1}\|_2 > \sqrt{2\varepsilon}$, which give $\|G_t\|_F \geq \alpha\sqrt{\lambda_0/2}\,\|e_{t-1}\|_2 > \alpha\sqrt{\lambda_0\varepsilon}$, together with $\frac{3}{18} \leq \frac{1}{4\sqrt{2}}$; each of the two terms is then at most $\frac{\kappa}{4}\|G_t\|_F$. This is (64), and with (63) it gives $\langle G_t, O_t\rangle \geq \kappa\|G_t\|_F - \frac{\kappa}{2}\|G_t\|_F$. For $t = 1$ the sums are empty, $D = 0$, and the claim is Lemma B.3 directly. □

B.6 Constants of Muon's coefficients and of the classical polynomials

The classical schedules used below are defined by the Taylor truncations

$$p^{(d)}(\lambda) = \sum_{s=0}^{d} \frac{(2s)!}{4^s(s!)^2}(1-\lambda)^s, \qquad \phi^{(d)}(x) = x p^{(d)}(x^2). \tag{69}$$

Here $d = 1$ gives the cubic scalar iteration and $d = 2$ the quintic. The tuned map instead uses the coefficients specified in the main text.

Proposition 1. *For $p(\lambda) = a + b\lambda + c\lambda^2$ with $(a, b, c) = (3.4445, -4.7750, 2.0315)$, $\phi(x) = x\,p(x^2)$ and $\Phi = \phi^{\circ 5}$,*

$$\Phi(x) \leq \frac{5}{4} \quad \textit{and} \quad \Phi(x) \geq \frac{2}{3}x \quad \textit{for } x \in [0, 1], \qquad \Phi(x) \geq 0.68 \quad \textit{for } x \in \Big[\frac{1}{400}, 1\Big]. \tag{70}$$

Proof. The proof uses four facts about the single quintic ϕ, each an evaluation of a polynomial. Let $x_a \approx 0.55453$ and $x_b \approx 1.05014$ be the positive roots of $\phi'(x) = a + 3bx^2 + 5cx^4$.

(F1) $p(z) \geq a - b^2/(4c) = 0.6386\ldots > 0$ for all z, and p decreases on $[0, -b/(2c)] = [0, 1.1752\ldots]$.

(F2) ϕ increases on $[0, x_a]$, decreases on $[x_a, x_b]$ and increases on $[x_b, \frac{5}{4}]$, with $\phi(x_a) = 1.2023\ldots$, $\phi(x_b) = 0.6818\ldots$ and $\phi(\frac{5}{4}) = 1.1790\ldots$.

(F3) $p(x_a^2) = 2.1682\ldots$.

(F4) The first five iterates of $\frac{1}{400}$ under ϕ are $0.0086\ldots$, $0.0296\ldots$, $0.1020\ldots$, $0.3464\ldots$ and $1.0048\ldots$.

By (F1) and (F2), ϕ maps $[0, \frac{5}{4}]$ into $[0, 1.2024]$ and maps $[x_a, \frac{5}{4}]$ into $[0.6818, 1.2024]$. Fix $x \in (0, 1]$ and let $x_k = \phi^{\circ k}(x)$, so every x_k lies in $[0, 1.2024]$ and in particular $\Phi(x) = x_5 \leq \frac{5}{4}$.

Part 1. Suppose $x_k \geq x_a$ for some $k \leq 4$. Then $x_k \in [x_a, \frac{5}{4}]$, every later iterate stays in $[0.6818, 1.2024] \subseteq [x_a, \frac{5}{4}]$, and so $\Phi(x) \geq 0.6818$. In particular $\Phi(x) \geq 0.68$ and $\Phi(x) \geq \frac{2}{3}x$, since $x \leq 1$.

Part 2. Suppose $x_k < x_a$ for every $k \leq 4$. Since p decreases on $[0, x_a^2]$,

$$\begin{aligned} \frac{\Phi(x)}{x} &= \prod_{k=0}^{4} p(x_k^2) && [\phi(x) = x\,p(x^2)] \\ &\geq p(x_a^2)^5 && [\text{(F1)}, x_k^2 < x_a^2] \\ &\geq 1 && [\text{(F3)}], \end{aligned} \tag{71}$$

so $\Phi(x) \geq \frac{2}{3}x$. If moreover $x \geq \frac{1}{400}$, then, since ϕ increases on $[0, x_a]$, induction on k gives $x_k \geq \phi^{\circ k}(\frac{1}{400})$ for $k \leq 4$, and so $\Phi(x) = \phi(x_4) \geq \phi\big(\phi^{\circ 4}(\frac{1}{400})\big) = 1.0048\ldots \geq 0.68$ by (F4).

Parts 1 and 2 cover all $x \in (0, 1]$ and prove (70); $x = 0$ is trivial. □

The first two claims give Assumption 1 for Muon's coefficients with $\kappa = \frac{2}{3}$ and $M = \frac{5}{4}$, the values used in Theorem 3.1. The third claim also yields a nuclear-norm alignment bound: if every nonzero singular value of B is at least $\|B\|_F/400$, then by (54) $\langle B, \mathrm{NS}_5(B)\rangle = \sum_j s_j \Phi(s_j/\|B\|_F) \geq 0.68\,\|B\|_*$.

Proposition 2 (Classical Newton–Schulz polynomials)**.** *Let Φ be the composition of K classical polynomials (69) of degrees $d_0, \ldots, d_{K-1} \geq 1$. For $x \in [0, 1]$,*

$$x \leq \Phi(x) \leq 1 \qquad \textit{and} \qquad 1 - \Phi(x)^2 \leq (1 - x^2)^{\prod_k (d_k+1)}. \tag{72}$$

Hence Assumption 1 holds with $\kappa = M = 1$ for every degree and every number of steps. For five steps of the same degree, $\Phi(x) \geq 0.68$ on $[\frac{1}{7}, 1]$ when $d = 1$ and on $[\frac{1}{19}, 1]$ when $d = 2$.

Proof. **Part 1.** For $\lambda \in (0, 1]$, the Taylor series $\lambda^{-1/2} = \sum_{s \geq 0} \frac{(2s)!}{4^s (s!)^2}(1 - \lambda)^s$ converges and has nonnegative terms, and $p^{(d)}$ is its truncation after the term $s = d$. Hence $1 \leq p^{(d)}(\lambda) \leq \lambda^{-1/2}$, and with $\lambda = x^2$,

$$x \leq x\,p^{(d)}(x^2) = \phi(x) \leq 1, \qquad x \in (0, 1]. \tag{73}$$

So each step maps $[0, 1]$ into itself without decreasing its argument, and composing K steps gives $x \leq \Phi(x) \leq 1$.

Part 2. For a classical polynomial of degree d and $u \in [0, 1]$, $1 - (1 - u)\,p^{(d)}(1 - u)^2 \leq u^{d+1}$ (Kim & Oh, 2026, Lemma 3). With $u = 1 - x^2$ this reads $1 - \phi(x)^2 \leq (1 - x^2)^{d+1}$. By Part 1 every iterate stays in $[0, 1]$, so applying this bound at each of the K steps gives the second claim of (72).

The bound $\Phi(x) \geq \big(1 - (1 - x^2)^{\prod_k (d_k+1)}\big)^{1/2}$ increases in x. For $K = 5$ and $d = 1$ the exponent is 32 and $(48/49)^{32} < 0.52$, so $\Phi \geq 0.69$ on $[\frac{1}{7}, 1]$. For $d = 2$ the exponent is 243 and $(360/361)^{243} < 0.51$, so $\Phi \geq 0.70$ on $[\frac{1}{19}, 1]$. □

C Derivation of the infinite-width kernel

Let $w \sim \mathcal{N}(0, I_d)$. The pair $(w^\top x_i, w^\top x_j)$ is a centered bivariate Gaussian with unit variances and correlation $x_i^\top x_j$. The two half-space conditions intersect in an angular wedge of length $\pi - \arccos\big(x_i^\top x_j\big)$, so

$$\mathbb{P}(w^\top x_i \ge 0,\ w^\top x_j \ge 0) = \frac{\pi - \arccos\big(x_i^\top x_j\big)}{2\pi}. \tag{74}$$

Multiplying by $x_i^\top x_j$ gives (7).

D Polar-Factor Optimality and Finite-Step Alignment

The exact polar factor maximizes alignment with a given buffer per unit spectral norm. This property concerns the buffer; transfer to the current gradient additionally requires Lemma 3.2.

Proposition 3 (Optimal buffer alignment per unit spectral norm)**.** *Let $B \neq 0$ have compact singular value decomposition $B = U\Sigma V^\top$. For every nonzero matrix O of the same dimensions,*

$$\frac{\langle B, O\rangle}{\|O\|_2} \le \|B\|_*, \tag{75}$$

with equality for $O = UV^\top$.

Proof. The trace inequality gives $\langle B, O\rangle \le \sum_j s_j(B)s_j(O) \le \|O\|_2 \sum_j s_j(B) = \|O\|_2\|B\|_*$. For $O = UV^\top$, $\|O\|_2 = 1$ and $\langle B, O\rangle = \mathrm{tr}(\Sigma) = \|B\|_*$. □

Under Assumption 1, the finite map satisfies

$$\frac{\langle B, \mathrm{NS}_K(B)\rangle}{\|\mathrm{NS}_K(B)\|_2} \ge \frac{\kappa}{M}\|B\|_F.$$

This bound suffices for the training analysis without establishing a uniform, rank-independent fraction of the polar optimum. If every nonzero singular value of B is at least $\|B\|_F/400$, then Proposition 1 additionally gives $\langle B, \mathrm{NS}_5(B)\rangle \ge 0.68\|B\|_*$. Together with $M \le 5/4$, this implies an alignment-per-norm ratio of at least $0.544\|B\|_*$. This stronger comparison requires the stated spectral condition and is not assumed in the training theorem.

E Experimental Details and Additional Diagnostics

This section collects the protocols and supporting diagnostics accompanying Section 5. The finite-map comparison is interpreted using Proposition 3.

E.1 Data, Initialization, and Numerical Settings

Inputs are sampled from a standard Gaussian distribution in $\mathbb{R}^d$ and normalized individually to unit norm. For the saved experiments, we use one dataset with seed 0 and an independent two-layer ReLU teacher with seed 1, width 1000, and output weights sampled from $\{\pm 1\}$. The teacher output uses the $1/\sqrt{1000}$ normalization in (1). The teacher and labels remain fixed across student initializations and widths. Unless otherwise specified, $n = 20$, $d = 100$, $m = 1000$, $\alpha = 1$, and $\delta = 0.05$. Student initialization and the unnormalized squared-loss convention follow Section 2. The alignment and fixed-step momentum experiments use student seeds 10, 11, and 12; the width sweep uses seeds 10–14 at each width.

For the saved dataset, we compute $\lambda_0 = 0.335$ from (7) and $R_0 = 30.2$ from (12). With $\kappa = 2/3$, $M = 5/4$, and target $\varepsilon = 10^{-2}$, the prescribed learning rate is $\eta = (1-\mu)\eta_0$, where $\eta_0 \approx 1.54 \times 10^{-4}$. The fixed-step comparison instead uses $\eta = \eta_0$ for every momentum value. The reference budget T is computed from (15); practical iteration caps are listed separately in Table 2 and are distinguished from T in Appendix E.2. Figure 1(c,d) uses the saved seed-10 trajectories from the alignment experiment.

Table 2: Experimental settings for the diagnostics in this appendix and in Section 5. Learning rates use the measured constants of the frozen draw, $\lambda_0 = 0.335$ and $R_0 = 30.2$. Horizons are the iteration caps actually used, not the theoretical budget.

Setting	Value or status
Default data and student	$n = 20$, $d = 100$, $m = 1000$, $\alpha = 1$
Budget parameter	$\delta = 0.05$
Accuracy panels	$\varepsilon \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$; $\mu = 0.95$; separate draw with $R_0 \approx 26$
Displayed momentum values	$\mu \in \{0, 0.5, 0.9, 0.95, 0.99\}$
Fixed-step comparison	Target 10^{-2}; shared step $\eta = \eta_0 = 1.54 \times 10^{-4}$; horizon 50,000
Teacher specification	Width 1000, output weights ± 1, seed 1, labels scaled by $m^{-1/2}$
Repetitions	One dataset (seed 0) and one teacher. Alignment and fixed-step: student seeds $10, 11, 12$. Width: seeds 10–14
Alignment study	Target 10^{-2}; $\mu \in \{0, 0.5, 0.9, 0.95, 0.99\}$; $\eta = (1-\mu)\eta_0$; horizons 70,700 to 7,070,000
Width study	$m \in \{64, 128, 256, 512, 1000, 2000\}$; target 10^{-2}; $\mu = 0.95$; horizon 1,414,000
Logging	A_t, D_t, C_t every step. Kernel and displacement every step in the width sweep. Buffers: 12 geometric steps plus the last, seed 10 only (59 buffers)
SVD computation	`numpy.linalg.svd`; keep $s > \max(d, m)\, \varepsilon_{\text{float64}}\, s_{\max}$. Discarded nuclear fraction $\le 8 \times 10^{-15}$
Implementation	Python 3.14.6, NumPy 2.5.3, float64. No denominator regularizer
Variability summaries	Each seed has equal weight. Bands and error bars are mean ± 1 standard deviation across seeds. Extrema are over every logged iteration

Figure 1(a,b) shows a separate accuracy sweep on a different draw of the data, with $\lambda_0 \approx 0.34$, $R_0 \approx 26$, and $\mu = 0.95$.

The experiments use Python 3.14.6 and NumPy 2.5.3. Newton–Schulz calculations use float64 arithmetic and Frobenius normalization without a denominator regularizer. Double precision reduces roundoff but does not make the computation identical to the exact-arithmetic map analyzed in the theorem. The SVD routine, numerical-rank threshold, logging schedule, and run horizons are specified in Table 2.

E.2 Stopping, Logging, and Alignment Conventions

For each run, we define $t_{\text{hit}} = \min\{t : L(W_t) \le \varepsilon\}$, with $t_{\text{hit}} = 0$ if the initial iterate already satisfies the target. Training terminates at target attainment or the practical horizon H listed in Table 2. A run that does not reach the target within H is recorded as unsuccessful within that horizon; it is not assigned a hitting time of H. The practical cap is distinct from the theorem's reference budget. For the saved dataset with $\mu = 0.95$ and $\varepsilon = 10^{-2}$, $T \approx 1.62 \times 10^8$, whereas the practical horizon is $H = 1{,}414{,}000$. The parameter δ specifies the theoretical reference calculation, not an empirical confidence interval or a guaranteed failure probability at the tested widths.

The alignment quantities A_t and D_t in (29) are recorded at every update, including the update that first reaches the target. The quantity A_t is not a cosine similarity because it is normalized only by the gradient norm. The normalization B_t/s_t, with $s_t = (1-\mu^t)/(1-\mu)$, removes the accumulation factor from the momentum buffer. Consequently, $D_t = 0$ whenever all stored gradients coincide; in particular, $D_t = 0$ for $\mu = 0$. Normalized diagnostics are defined only when their denominators are nonzero. A zero denominator is treated as an undefined measurement rather than assigned a numerical value.

Figure 2 uses student seeds 10, 11, and 12. At each iteration, the plotted mean includes only seeds that have not already terminated; each contributing seed receives equal weight. Shaded bands show one sample standard deviation above and below the mean when at least two seeds remain. Thus, the number of contributing seeds decreases after the first target crossing, and changes near the endpoints can reflect both within-run dynamics and changes in the set of contributing runs. The reported minimum alignment and ranges of run-wise maximum discrepancy are computed from individual runs, rather than from these averaged curves.

We also record buffer-update alignment at every update:

$$C_t := \frac{\langle B_t, O_t \rangle}{\|B_t\|_F}. \tag{76}$$

The reference $C_t \geq \kappa = 2/3$ follows from the tuned Newton–Schulz map in exact arithmetic, independently of network width. In contrast, the current-gradient bound $A_t \geq \kappa/2$ additionally requires the trajectory assumptions of Lemma 3.2. For the matrix-map comparison, buffers are saved from student seed 10 at geometrically spaced updates and at the final update, yielding 59 buffers across the tested momentum values. These measurements characterize the saved buffers; they do not constitute independent training runs.

The fixed-step comparison uses $\eta = \eta_0 \approx 1.54 \times 10^{-4}$, target 10^{-2}, and horizon $50{,}000$ for $\mu \in \{0, 0.5, 0.9, 0.95, 0.99\}$. All three student seeds at each momentum reach the target, with hitting times ranging from 14,680 to 25,444 updates. These runs show that the prescribed $(1-\mu)$ learning-rate scaling can be conservative for the tested problem; they do not establish a generally valid momentum-independent step size.

E.3 Finite Newton–Schulz Maps and the Exact Polar Factor

We compare three transformations on identical nonzero momentum buffers saved from tuned Muon trajectories: the tuned five-step Newton–Schulz map, five classical cubic Newton–Schulz iterations, and the exact polar factor evaluated numerically by SVD. Both iterative maps start from $X_0 = B/\|B\|_F$. The classical recurrence is

$$X_{k+1} = \frac{3}{2}X_k - \frac{1}{2}X_k X_k^\top X_k, \qquad k = 0, \ldots, 4, \tag{77}$$

corresponding to $d = 1$ in (69). Although Proposition 2 also covers higher-degree classical polynomials, the empirical comparison here uses only the cubic map. Evaluating each transformation on the same buffers isolates its effect from differences in training trajectories.

We compute the SVD using `numpy.linalg.svd` and retain singular values satisfying $s_j > \max(d, m)\epsilon_{\text{float64}} s_{\max}$, where $\epsilon_{\text{float64}}$ is machine precision. The numerical polar update is $U_r V_r^\top$, formed from the retained singular-vector pairs. Across the 59 saved buffers, the discarded singular values account for at most 8×10^{-15} of the nuclear norm. Thus, truncation has negligible effect on the alignment comparison at the reported precision.

For each buffer B and transformed update O, we evaluate

$$a(B, O) := \frac{\langle B, O \rangle}{\|B\|_F}, \qquad s(O) := \|O\|_2, \qquad q(B, O) := \frac{\langle B, O \rangle}{\|B\|_* \|O\|_2}. \tag{78}$$

The quantities a and s examine the alignment and norm bounds used in the training proof. The ratio q measures alignment per unit spectral norm relative to the optimum in Proposition 3; the exact polar factor attains $q = 1$. These quantities do not measure distance to the polar factor.

Figure 4 displays the measurements on the 59 buffers from student seed 10. Tuned NS5 has $a \in [0.697, 2.690]$, with median 2.687, and $s \in [0.696, 1.202]$. Thus, all sampled buffers satisfy the analytical references $a \geq 2/3$ and $s \leq 5/4$. Its normalized ratio lies in $q \in [0.748, 0.999]$, with median 0.764 and standard deviation 0.075 across buffers. Classical cubic NS5 has $s = 1$ to the reported precision and $q \in [0.860, 0.998]$, with median 0.860 and standard deviation 0.043. The numerical polar update has $s = 1$ and $q = 1$ to the reported precision.

The classical cubic map therefore achieves a higher median alignment-per-norm ratio on these sampled buffers, while the tuned map retains the properties sufficient for the training guarantee.

The medians and standard deviations summarize correlated buffers from the saved trajectories, not variability across independent datasets. This comparison establishes neither a ranking of training performance nor which map more closely approximates the polar factor.

E.4 Width, Displacement, and Kernel Diagnostics

We vary the student width over $m \in \{64, 128, 256, 512, 1000, 2000\}$ while holding the dataset, teacher, target loss, momentum, and prescribed learning rate fixed. Each width uses student seeds 10–14, giving 30 runs in total. We set $\varepsilon = 10^{-2}$, $\mu = 0.95$, and the practical horizon to $1{,}414{,}000$ updates. For this dataset, $\lambda_0 = 0.335$ and $R_0 = 30.2$. Kernel and displacement diagnostics are recorded at every iteration through target attainment or horizon exhaustion.

Alongside K_t and S_t in (30), we measure

$$P_t := \|W_t - W_0\|_F, \qquad R_{\mathbb{B}} := \frac{18M\sqrt{n}\,R_0}{\kappa\alpha\sqrt{\lambda_0}}, \qquad \widehat{P}_t := \frac{P_t}{R_{\mathbb{B}}}. \tag{79}$$

The radius is $R_{\mathbb{B}} \approx 7.88 \times 10^3$ and is shared across the width sweep. For each run, we compute the minimum K_t, maximum S_t, and maximum $\widehat{P}_t$ over the recorded trajectory. Summary points and error bars report the mean and one sample standard deviation of these run-wise extrema across the five student initializations, with equal weight assigned to each run.

All 30 runs reach the target loss within the practical horizon. The minimum normalized kernel eigenvalue has mean 0.871 ± 0.030 at width 64 and 0.990 ± 0.012 at width 2000. The smallest value recorded across all runs is 0.826, above the analytical reference $1/2$. The maximum relative kernel change decreases from 0.143 ± 0.047 at width 64 to 0.031 ± 0.007 at width 2000 (Figure 3).

The largest normalized displacement across all runs is 1.67×10^{-3}, attained at width 2000 (Figure 5). Thus, every observed trajectory remains far inside the analytical neighborhood while maintaining kernel positivity. The small relative displacement and reduced kernel variation at larger widths are consistent with the stability mechanism used in the proof.

These observations concern repeated student initializations on one fixed dataset and teacher. They do not establish the necessity or tightness of the sufficient width condition. In particular, containment within the analytical radius at these widths does not itself certify the uniform kernel bounds, and the measured kernel diagnostics do not verify every Jacobian-action and linearization estimate used in the analysis.