MUSE: REPRESENTATION GEOMETRY OF MUON BEYOND NORMALIZED MOMENTUM

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon applies a spectral transformation to matrix momentum, while normalized SGD with momentum (nSGDM) normalizes its vectorized form. What does the matrix representation contribute beyond vector normalization? We introduce Muse to investigate this question by varying the matrix representation used to compute the update. In 130M and 600M language models, order-preserving reshaping retains lower validation loss than nSGDM. However, fixed-shape entry shuffling removes Muon's advantage over nSGDM in a controlled row-block problem and 600M language-model training. To illustrate how representation affects optimization, we first show that single-step parameter-error progress depends on both nuclear support and update alignment. We then analyze convergence under heterogeneous row-block curvature. Muon's spectral constraint limits the update energy in each row, yielding bounds that depend on average row curvature, whereas the corresponding nSGDM bounds depend on the maximum. We establish stationarity guarantees for exact polar and fixed-iteration Newton–Schulz updates. Under heterogeneous native row-block curvature, the optimized bound is tighter for exact polar than for nSGDM; we also identify conditions under which this advantage extends to the five-step update.

1 INTRODUCTION

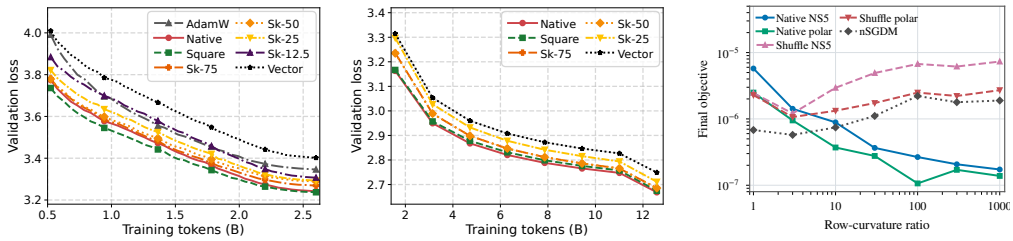
Muon-style optimizers have achieved performance comparable to AdamW in language-model training, with greater training efficiency in some settings (Liu et al., 2025; Shah et al., 2025; Wen et al., 2026b). Existing analyses have studied Muon through the geometry of matrix updates under spectral-norm constraints (Bernstein & Newhouse, 2024; Chen et al., 2025; Pethick et al., 2025), the spectral effects of orthogonalization (Nguyen et al., 2026), its convergence properties (Shen et al., 2025; Chang et al., 2025), and finite-step Newton–Schulz and inexact polar updates (Kim & Oh, 2026; Shulgin et al., 2025; Choudhury et al., 2026). Here we study how Muon's update depends on the matrix representation of momentum—the matrix shape and the assignment of entries to rows and columns. Native Muon computes its spectral update using the parameter matrix's original layout. Applying the same operation to a single-row representation instead yields the direction of per-matrix normalized SGD with momentum (nSGDM)[1]. This matrix-to-vector connection motivates our question: **which aspects of matrix representation contribute to Muon's advantage over normalized momentum?**

In this work, we first test whether the native shape of a parameter matrix is necessary for this advantage. To do so, we introduce **Muse**, which changes the representation used to compute the spectral update while keeping parameters and momentum in their original coordinates (Section 2). Ordered reshaping changes the row–column partition without changing entry order. We compare the native representation (Native), a near-square representation (Square), narrower representations (Skinny), and a single-row representation (Vector). At the Vector endpoint, exact polar decomposition recovers nSGDM; the five-step Newton–Schulz update (NS5) yields the same direction up to a fixed factor that can be absorbed into the learning rate. With learning rates selected separately, a non-native or-

[1]Here nSGDM normalizes the momentum of each parameter matrix separately, giving the direction $\mathbf{M}_t/\|\mathbf{M}_t\|_F$. Cutkosky & Mehta (2020) analyze normalized momentum for a single vector parameter, although their experiments implement normalization per layer.

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.

dered representation achieves lower validation loss than Vector in both 130M and 600M language models (Figure 1(a,b)). Thus, this advantage does not require the native matrix shape.



(a) LLaMA2-130M. (b) LLaMA2-600M. (c) Fixed-shape shuffling.

Figure 1: **Order-preserving reshaping and fixed-shape entry shuffling.** (a,b) LLaMA2 validation loss versus training tokens: (a) 130M to 2.60B tokens; (b) 600M to 12.58B tokens. (c) Final row-block objective versus native-row curvature ratio (log scales), averaged over three seeds; objective and matrix dimensions are fixed at each ratio. Learning rates are selected separately. Square and Skinny finish below Vector, while shuffling removes Native's advantage at high heterogeneity for both NS5 and exact polar.

We next ask whether entry organization matters when matrix shape is fixed. We rearrange entries without changing matrix dimensions, thereby changing which coordinates share a row or column. This intervention leaves Frobenius-normalized momentum updates unchanged but can change Muon's spectral update direction. On a controlled objective with row-block curvature, fixed-shape entry shuffling removes Native's advantage (Figure 1(c) and Figure 15). In 600M language-model training, fixed shuffling yields higher validation loss than both Native and nSGDM (Table 2). These comparisons point to the relationship between the rows induced by a representation and the curvature of the objective. Prior work has linked heterogeneity across Hessian blocks to the role of block-specific scalar learning rates (Zhang et al., 2024; 2025b). We therefore study how a matrix spectral update allocates energy across rows and how that allocation interacts with curvature blocks.

To analyze this interaction, we characterize the spectral constraint induced by each representation (Section 3.1). At a fixed state, its nuclear-norm support and the alignment achieved by the update determine the first-order effect. An exact identity separates these contributions from the quadratic step-size cost in single-step parameter-error progress (Section 3.2). Under the isotropic setting used in the local analysis, a Marchenko–Pastur model calibrates how support changes with matrix shape (Marčenko & Pastur, 1967). Under native row-block curvature, controlling the spectral energy of each update row replaces maximum-row curvature terms in the stationarity bounds with terms involving averages of the row-curvature bounds and their squares (Section 3.3). We also derive explicit spectral and alignment constants for a fixed-coefficient NS map across fixed representations, extending earlier finite-step analysis of a Taylor-truncated polynomial (Kim & Oh, 2026). Our contributions are as follows:

- **Representation-dependent optimization behavior. Muse** separates changes in matrix shape from changes in entry organization and recovers nSGDM at the Vector endpoint. Shared-state comparisons, controlled curvature interventions, and language-model training show that the advantage can persist after ordered reshaping but can be lost under fixed-shape entry shuffling.

- **Geometry and convergence analysis.** We derive an exact identity for single-step parameter-error progress, calibrate shape-dependent nuclear-norm support, and establish stationarity bounds for exact polar and fixed-coefficient NS updates. Under native row-block curvature, heterogeneity aligned with the native rows can make the optimized theoretical bound favor Native Muon over nSGDM (Table 3).

Thus, matrix shape affects spectral support, while entry organization affects the spectral update's interaction with curvature. Appendix A relates this view to coordinate-adaptive methods (Duchi et al., 2011; Kingma & Ba, 2015; Loshchilov & Hutter, 2019; Shazeer & Stern, 2018), matrix preconditioning (Martens & Grosse, 2015; Gupta et al., 2018; Vyas et al., 2024), and Muon analyses (Chen et al., 2025; Pethick et al., 2025).

2 A REPRESENTATION FRAMEWORK FOR MUON

Before presenting our analysis, we introduce notation and the problem setup. Throughout, $\langle \mathbf{A}, \mathbf{B} \rangle_F = \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$ denotes the Frobenius inner product, while $\|\cdot\|_F$, $\|\cdot\|_2$, and $\|\cdot\|_*$ denote the Frobenius, spectral, and nuclear norms. A fixed vectorization identifies $\mathbb{R}^{m\times n}$ with $\mathbb{R}^d$, where $d = mn$. For a compact SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ retaining nonzero singular values, we define $\mathrm{Orth}(\mathbf{A}) = \mathbf{U}\mathbf{V}^\top$, with $\mathrm{Orth}(\mathbf{0}) = \mathbf{0}$.

Under this identification, a matrix representation is a Frobenius-isometric linear bijection $T : \mathbb{R}^d \to \mathbb{R}^{p_T \times q_T}$ that rearranges and reshapes entries, with $p_T q_T = d,\ p_T \le q_T$. For $\mathbf{Z} \in \mathbb{R}^{m\times n}$, its spectral update direction is

$$\mathbf{D}_T(\mathbf{Z}) = T^{-1}(\mathrm{Orth}(T(\mathbf{Z}))).$$

Consider the possibly nonconvex stochastic objective

$$\min_{\mathbf{W}\in\mathbb{R}^{m\times n}} f(\mathbf{W}), \qquad f(\mathbf{W}) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(\mathbf{W};\xi)], \tag{1}$$

Let $\mathcal{F}_{t-1}$ denote the history before sampling ξ_t, and write $\mathbf{G}_t = \nabla f(\mathbf{W}_t)$ and $\widehat{\mathbf{G}}_t = \nabla f(\mathbf{W}_t;\xi_t)$ for the population and stochastic gradients. Algorithm 1 computes updates in represented coordinates and maintains momentum in original coordinates.

2.1 REPRESENTATIONS FROM NATIVE MUON TO NSGDM

Muon computes spectral updates in the native parameter-matrix layout (Jordan et al., 2024), which also underlies matrix-based spectral-norm analyses (Chen et al., 2025; Pethick et al., 2025). To vary this layout, we use order-preserving reshaping to arrange row-major entries into a $p \times (d/p)$ matrix, with $p \mid d$, $p \le d/p$, and native short side $s = \min(m, n)$. We compare four representations.

Algorithm 1 Muse: **MU**on under re**S**haping and r**E**ordering

1: **Input:** Initial parameters $\mathbf{W}_1 \in \mathbb{R}^{m\times n}$, fixed representation T, total iterations K, learning rates $\eta_t > 0$, and momentum coefficient $\alpha \in [0, 1)$.
2: Initialize $\mathbf{M}_0 = \mathbf{0}$.
3: **for** $t = 1$ **to** K **do**
4: Draw $\xi_t \sim \mathcal{D}$ and compute $\widehat{\mathbf{G}}_t = \nabla f(\mathbf{W}_t;\xi_t)$.
5: $\mathbf{M}_t = \alpha\mathbf{M}_{t-1} + (1-\alpha)\widehat{\mathbf{G}}_t$.
6: $\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \mathbf{D}_T(\mathbf{M}_t)$.
7: **end for**

Native retains the original matrix layout, transposed when needed to place the short side in the rows. **Square** uses the largest divisor $p \le \sqrt{d}$, giving the near-square shape. **Skinny** chooses $p \in \arg\min_{a\mid d,\, a\le s} |a - \theta s|$ for $\theta \in \{1/8, 1/4, 1/2, 3/4\}$. Plot legends abbreviate Skinny as Sk. **Vector** takes $p = 1$, placing all entries in one row. For nonzero momentum, $\mathbf{D}_T(\mathbf{M}) = \mathbf{M}/\|\mathbf{M}\|_F$, recovering nSGDM. NS5 gives the same direction up to a fixed factor absorbed into the learning rate.

2.2 ENTRY SHUFFLING AT FIXED SHAPE

Next, we vary which entries share rows or columns at fixed shape. For an entry permutation Π and base representation T_0, set $T = \Pi \circ T_0$. **Fixed shuffling** reuses one sampled Π throughout training; **stepwise shuffling** samples Π_t each update and uses $T_t = \Pi_t \circ T_0$. After inverse mapping, general entry permutations can change spectral directions but preserve nSGDM; row and column permutations preserve both exact-polar and NS5 directions in real arithmetic (Appendix B.1).

3 REPRESENTATION GEOMETRY AND OPTIMIZATION PROGRESS

We begin with a local, single-step analysis of how matrix representation affects Muon's progress relative to normalized momentum. We first characterize the induced spectral constraint and first-order decrease, then compare teacher–student parameter-error progress to separate nuclear support, update alignment, and finite-step cost. Building on this analysis, Section 3.3 establishes multistep stationarity guarantees under heterogeneous row-block curvature.

3.1 SPECTRAL-BALL GEOMETRY INDUCED BY REPRESENTATION

To understand why representation changes the update direction, we first examine its constraint in original parameter coordinates. For fixed T, the represented unit spectral-norm constraint induces $\mathcal{K}_T = \{\Delta\mathbf{W} : \|T(\Delta\mathbf{W})\|_2 \le 1\}$. For a population-gradient update, exact polar gives the steepest descent direction under this constraint, with maximal first-order decrease equal to nuclear support.

The following theorem combines this geometry with smoothness to bound single-step descent, extending Muon's norm-constrained interpretation across representations (Bhatia, 2009; Bernstein & Newhouse, 2024; Chen et al., 2025; Pethick et al., 2025; Shen et al., 2025).

Assumption 3.1. *The population objective f is differentiable and L-smooth on $\mathbb{R}^{m\times n}$: for all $\mathbf{W}, \widetilde{\mathbf{W}} \in \mathbb{R}^{m\times n}$, $\|\nabla f(\widetilde{\mathbf{W}}) - \nabla f(\mathbf{W})\|_F \le L\|\widetilde{\mathbf{W}} - \mathbf{W}\|_F$.*

Theorem 3.1. *Let T be a fixed Frobenius-isometric representation from Section 2, with $\nabla f(\mathbf{W}) \neq \mathbf{0}$. Then $\langle \nabla f(\mathbf{W}), \mathbf{D}_T(\nabla f(\mathbf{W}))\rangle_F = \|T(\nabla f(\mathbf{W}))\|_*$, and $-\mathbf{D}_T(\nabla f(\mathbf{W})) \in \arg\min_{\|T(\Delta\mathbf{W})\|_2\le 1}\langle \nabla f(\mathbf{W}), \Delta\mathbf{W}\rangle_F$. Under Assumption 3.1, for every $\eta > 0$,*

$$f(\mathbf{W} - \eta \mathbf{D}_T(\nabla f(\mathbf{W}))) \le f(\mathbf{W}) - \eta\|T(\nabla f(\mathbf{W}))\|_* + \frac{L\eta^2}{2}\operatorname{rank}(T(\nabla f(\mathbf{W}))). \tag{2}$$

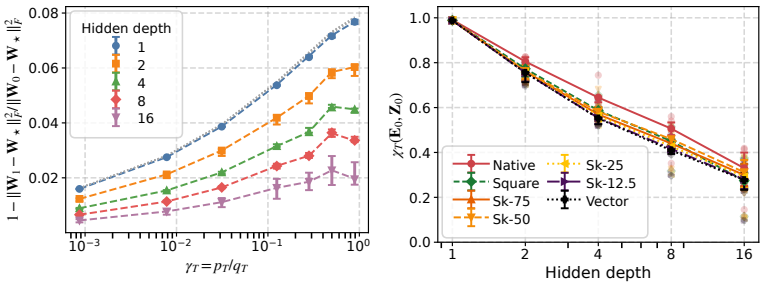
See Appendix B for the proof.

Remark 3.1. *Thus representation affects first-order decrease through its spectral ball, while finite steps incur a rank-dependent quadratic cost. At the Vector endpoint, $\mathcal{K}_T$ becomes the Frobenius unit ball, placing matrix spectral updates and normalized momentum in one framework.*

The theorem uses the population gradient, whereas the algorithm transforms momentum. We therefore quantify the realized first-order decrease. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{A} \neq \mathbf{0}$, be the reference population gradient or parameter error, and $\mathbf{Z} \in \mathbb{R}^{m\times n}$ the gradient or momentum defining the direction. Define update alignment by

$$\chi_T(\mathbf{A}, \mathbf{Z}) = \frac{\langle \mathbf{A}, \mathbf{D}_T(\mathbf{Z})\rangle_F}{\|T(\mathbf{A})\|_*}.$$

For exact polar, $|\chi_T(\mathbf{A}, \mathbf{Z})| \le 1$. The denominator is the maximal inner product with $\mathbf{A}$ under the constraint; the numerator is the achieved value, so their ratio records relative magnitude and sign. With $\mathbf{A} = \nabla f(\mathbf{W})$, $\mathbf{Z} = \mathbf{M}$, first-order action equals nuclear support times update alignment. We thereby separate available support from its realized fraction; an isotropic spectral model will quantify shape's effect on support.



(a) Matrix shape and parameter-error progress.

(b) Network depth and update alignment.

Figure 2: **MP predictions agree with measured progress after accounting for update alignment.** (a) Predictions using measured alignment (colored dashed) match first-layer progress (points) across shapes and depths; alignment 1 (gray dotted) overestimates progress in deeper networks. (b) Alignment declines with depth, explaining this discrepancy. Points and error bars show medians and interquartile ranges over six shared states.

3.2 Single-Step Parameter-Error Progress through Nuclear Support and Update Alignment

We next compare how matrix spectral updates and Vector's normalized momentum approach a known solution. For a common measurable reference, we use a multilayer ReLU teacher–student network: the teacher defines a zero-loss solution, and the student starts near its parameters. In this local setting, single-step parameter-error progress measures the decrease in squared distance to the teacher: positive values indicate approach, negative values departure. Specifically, for a hidden-layer block with teacher parameters $\mathbf{W}_\star$, let $\mathbf{E}_0 = \mathbf{W}_0 - \mathbf{W}_\star \neq \mathbf{0}$. Given gradient or momentum $\mathbf{Z}_0 \in \mathbb{R}^{m\times n}$, set $\mathbf{W}_1 = \mathbf{W}_0 - \eta\mathbf{D}_T(\mathbf{Z}_0)$, and write $\chi_T = \chi_T(\mathbf{E}_0, \mathbf{Z}_0)$ as in Section 3.1. Normalizing the reduction in squared error by step size and initial squared error gives the following identity.

Theorem 3.2. *For every fixed Frobenius-isometric representation T, every $\mathbf{W}_0 \neq \mathbf{W}_\star$, every $\mathbf{Z}_0 \in \mathbb{R}^{m\times n}$, and every $\eta > 0$, we have $|\chi_T| \leq 1$ and*

$$\frac{\|\mathbf{W}_0 - \mathbf{W}_\star\|_F^2 - \|\mathbf{W}_1 - \mathbf{W}_\star\|_F^2}{\eta\|\mathbf{W}_0 - \mathbf{W}_\star\|_F^2} = 2\chi_T \frac{\|T(\mathbf{E}_0)\|_*}{\|T(\mathbf{E}_0)\|_F^2} - \eta\frac{\operatorname{rank}(T(\mathbf{Z}_0))}{\|T(\mathbf{E}_0)\|_F^2},$$

where the finite-step cost involves rank at most p_T. See Appendices G and H for the proof.

This separates first-order action toward the teacher, determined by nuclear support and update alignment, from the finite-step cost determined by rank and step size. Using Theorem 3.2, we first quantify how shape affects nuclear support. For isotropic teacher–student parameter perturbations, Marchenko–Pastur (MP) theory (Marčenko & Pastur, 1967) gives the shape dependence in Corollary F.1 (Appendix F.2): at equal parameter dimension and perturbation scale, Square's leading support factor is $\Theta(d^{1/4})$ times Vector's, with progress depending on update alignment. At full row rank, matching Frobenius update lengths cancels this dimension factor; spectral constants and alignment retain shape dependence. Figure 2 compares MP predictions with measured progress: measured alignment yields agreement across shapes and depths.

These equal-step comparisons reflect both direction and length. Next, we optimize each representation's scalar step to identify the remaining directional differences. For nonzero $\mathbf{D} = \mathbf{D}_T(\mathbf{Z}_0)$, this gives

$$\max_{\eta\geq 0} \frac{\|\mathbf{E}_0\|_F^2 - \|\mathbf{E}_0 - \eta\mathbf{D}\|_F^2}{\|\mathbf{E}_0\|_F^2} = \frac{\max\{0, \langle\mathbf{E}_0, \mathbf{D}\rangle_F\}^2}{\|\mathbf{E}_0\|_F^2\|\mathbf{D}\|_F^2}. \tag{3}$$

Optimal relative progress equals the square of the positive part of normalized directional alignment. This identity holds for every nonzero direction, including NS5. Combining the decomposition, MP analysis, step optimization, and paired experiments, we obtain the following conclusion.

Claim 1. After separate scalar-step optimization, normalized directional alignment determines single-step parameter-error progress. In our teacher–student experiments at depths $2, 4, 8$, matrix representations retain a directional advantage over nSGDM (Appendix J.1).

3.3 Matrix Geometry under Heterogeneous Row-Block Curvature

Next, we study how these directional differences affect repeated updates on general stochastic objectives, measuring progress through gradient norms. Directions affect both first-order decrease and curvature costs along update paths; we therefore study their relationship to row-block curvature. Specifically, we first establish a unified stationarity bound, then compare matrix spectral updates with normalized momentum under row-block structure. We use Assumption 3.1 and standard stochastic-oracle conditions (Cutkosky & Orabona, 2019; Fang et al., 2018; Zhou et al., 2020; Li et al., 2023; Wang et al., 2023; Li & Hong, 2025; Shen et al., 2025; Chang et al., 2025; Nagashima & Iiduka, 2026; Zhang & Lin, 2026; Duchi et al., 2011; Kingma & Ba, 2015; Shazeer & Stern, 2018; Loshchilov & Hutter, 2019; Gupta et al., 2018).

Assumption 3.2. *At iteration t, ξ_t is sampled independently of $\mathcal{F}_{t-1}$, and the stochastic gradient is conditionally unbiased: $\mathbb{E}[\widehat{\mathbf{G}}_t \mid \mathcal{F}_{t-1}] = \mathbf{G}_t$.*

Assumption 3.3. *There exists $\sigma^2 < \infty$ such that, for every t, $\mathbb{E}[\|\widehat{\mathbf{G}}_t - \mathbf{G}_t\|_F^2 \mid \mathcal{F}_{t-1}] \leq \sigma^2$.*

Newton–Schulz map and five-step bounds. To cover finite-step spectral updates used in practice, we also analyze a fixed-coefficient NS map. For fixed integer $J \geq 1$ and $\mathbf{A} \neq \mathbf{0}$, define in real arithmetic

$$\begin{aligned} \mathbf{X}_0 &= \mathbf{A}/\|\mathbf{A}\|_F, \\ \mathbf{X}_{j+1} &= a_0\mathbf{X}_j + a_1(\mathbf{X}_j\mathbf{X}_j^\top)\mathbf{X}_j + a_2(\mathbf{X}_j\mathbf{X}_j^\top)^2\mathbf{X}_j, \\ \mathcal{N}_J(\mathbf{A}) &= \mathbf{X}_J, \qquad \mathcal{N}_J(\mathbf{0}) = \mathbf{0}, \end{aligned} \tag{4}$$

where $j = 0, \ldots, J-1$. Throughout, we fix $(a_0, a_1, a_2) = (3.4445, -4.7750, 2.0315)$, following Jordan et al. (2024); Liu et al. (2025). The direction mapped back to original coordinates is $\widetilde{\mathbf{D}}_{T,J}(\mathbf{M}) = T^{-1}(\mathcal{N}_J(T(\mathbf{M})))$. Here J counts polynomial iterations and K training updates.

Theorem 3.3. *Let T be a fixed Frobenius-isometric representation with short side p_T. Fix $\eta > 0$, $0 < \alpha \leq 1$, and set $\eta_t = \eta t^{-3/4}$, $\alpha_t = 1 - \alpha t^{-1/2}$. Run Algorithm 1 with these schedules and either $\mathbf{D}_T$ or, for fixed $J \geq 1$, $\widetilde{\mathbf{D}}_{T,J}$. For $\mathbf{D}_T$, set $(c, B) = (1, 1)$; for $\widetilde{\mathbf{D}}_{T,J}$, set $(c, B) = (1/(2\sqrt{p_T}), 5/4)$. For $J = 5$, the stronger constant $c = \max\{1/(2\sqrt{p_T}), 972/(p_T + 1458)\}$ is*

valid with $B = 5/4$. Suppose Assumptions 3.1, 3.2, and 3.3 hold, f has lower bound $f^\star$, and $\mathbb{E}[f(\mathbf{W}_1)] < \infty$, $\mathbb{E}\|\nabla f(\mathbf{W}_1)\|_F^2 < \infty$. Then, for every $K \geq 2$,

$$\min_{1\leq t\leq K} \mathbb{E}\|T(\nabla f(\mathbf{W}_t))\|_* \leq \frac{1}{c\eta K^{1/4}}\Bigg[\mathbb{E}[f(\mathbf{W}_1)] - f^\star + (1-\alpha)^2\mathbb{E}\|\nabla f(\mathbf{W}_1)\|_F^2 + \alpha^2\sigma^2 + (1+\ln K)\left(\frac{2\sqrt{2}L^2B^2p_T\eta^2}{\alpha} + \alpha^2\sigma^2 + \frac{(c+B)^2p_T\eta^2}{2\alpha} + \frac{LB^2p_T\eta^2}{2}\right)\Bigg]. \tag{5}$$

Remark 3.2. *The theorem covers Native, Vector, and other fixed Muse representations, with exact polar or finite-step NS. Its rate is $\widetilde{O}(K^{-1/4})$, where $\widetilde{O}$ suppresses logarithmic factors. For $s = \min(m, n)$, it also controls stationarity in original coordinates:*

$$\min_{1\leq t\leq K} \mathbb{E}\|\nabla f(\mathbf{W}_t)\|_* \leq \sqrt{s} \min_{1\leq t\leq K} \mathbb{E}\|T(\nabla f(\mathbf{W}_t))\|_* \leq \widetilde{O}(K^{-1/4}).$$

For NS5, we establish stationarity under the fixed coefficients in (4)*. Starting from Frobenius-normalized inputs, we bound the spectral norm above and update alignment below to control update length and first-order decrease, then combine momentum tracking with Assumptions 3.1, 3.2, and 3.3. The constants depend on the short side, allowing different fixed Frobenius-isometric representations. Appendix E proves these results in real arithmetic.*

We next examine matrix spectral updates' advantage over normalized momentum. Under common smoothness, noise bounds, and schedules, the exact-polar dimension-dependent costs in (5) increase with p_T, favoring Vector ($p_T = 1$). Yet Native and Square achieve lower final validation loss than Vector in our 130M and 600M experiments with separately selected learning rates (Section 4.3). To investigate this gap, we shuffle entries at fixed shape, changing row–column groupings. This removes Native's advantage over normalized momentum on the controlled row-block objective; fixed shuffling also yields higher 600M validation loss than Native and nSGDM (Figure 15; Table 2). We therefore study how spectrally constrained row update energy interacts with curvature to affect progress. Prior work links optimizer behavior to heterogeneous Hessian-block spectra and weak coupling (Zhang et al., 2024; 2025b; Dong et al., 2025), consistent with our MLP Hessian and attention-head responses (Figure 3). Accordingly, we group native rows (e.g., single rows or attention-head output rows) and allow nonzero coupling as follows.

Assumption 3.4. *For $\mathbf{W} \in \mathbb{R}^{p\times q}$, $p \leq q$, let f be twice continuously differentiable. Fix a partition of the native rows into nonempty groups $I_1, \ldots, I_k$. In coordinates ordered by these groups, let $\mathbf{H} = \nabla^2 f(\mathbf{W})$, with $\mathbf{B} = \operatorname{blkdiag}_j \mathbf{H}_{jj}$ retaining its principal group blocks. Along every update segment of the compared methods,*

$$\|\mathbf{H}_{jj}\|_2 \leq L_j, \qquad \|\mathbf{H} - \mathbf{B}\|_2 \leq \varepsilon,$$

where L_j, ε are common nonnegative bounds.

Under this assumption, we compare energy allocation across row groups. Muon's spectral constraint bounds each group's energy proportionally to its row count (Lemma E.6), giving group-size-weighted average curvature costs. To quantify improvement over maximum curvature, for $\max_j L_j > 0$, define the maximum-to-weighted-average ratios

$$h_r = \frac{\max_j L_j^r}{p^{-1}\sum_{j=1}^k |I_j| L_j^r}, \qquad r \in \{1, 2\}, \qquad 1 \leq h_r \leq p. \tag{6}$$

The corresponding costs are $\max_j L_j/h_1$ and $\max_j L_j/\sqrt{h_2}$, whereas Frobenius normalization uses $\max_j L_j$. Singleton groups recover the native-row ratios. Applying these row-block curvature bounds to the unified stationarity analysis and optimizing each base step size lets us compare Native and nSGDM using the same original-coordinate stationarity measure:

Proposition 3.4. *Suppose the assumptions of Theorem 3.3 and Assumption 3.4 hold, with $\max_j L_j > 0$. Use the theorem's schedules and zero initial momentum. Index NS5-Native, exact-polar Native, and nSGDM by $i = 1, 2, 3$; their directions are $\mathcal{N}_5(\mathbf{M}_t)/\sqrt{p}$, $\operatorname{Orth}(\mathbf{M}_t)/\sqrt{p}$, and $\mathbf{M}_t/\|\mathbf{M}_t\|_F$, respectively, with zero directions at zero momentum. Set $c_5 =$*

$\max\{1/(2\sqrt{p}), 972/(p+1458)\}$, $R = \mathbb{E}[f(\mathbf{W}_1)] - f^\star + (1-\alpha)^2\mathbb{E}\|\nabla f(\mathbf{W}_1)\|_F^2 + 2\alpha^2\sigma^2$, *and*

$$C_1 = \frac{25\sqrt{2}}{8\alpha c_5^2}\left(\frac{\max_j L_j}{\sqrt{h_2}} + \varepsilon\right)^2 + \frac{(c_5+5/4)^2}{2\alpha c_5^2} + \frac{25}{32c_5^2}\left(\frac{\max_j L_j}{h_1} + \varepsilon\right),$$

$$C_2 = \frac{2\sqrt{2}}{\alpha}\left(\frac{\max_j L_j}{\sqrt{h_2}} + \varepsilon\right)^2 + \frac{2}{\alpha} + \frac{1}{2}\left(\frac{\max_j L_j}{h_1} + \varepsilon\right),$$

$$C_3 = \frac{2\sqrt{2}}{\alpha}(\max_j L_j + \varepsilon)^2 + \frac{2}{\alpha} + \frac{\max_j L_j + \varepsilon}{2}.$$

If $R > 0$, separately optimizing each method's base step size gives, for every $K \geq 2$,

$$\min_{1\leq t\leq K} \mathbb{E}\|\nabla f(\mathbf{W}_t)\|_* \leq \frac{2(1+\ln K)\sqrt{pRC_i}}{K^{1/4}}, \qquad i \in \{1,2,3\}. \tag{7}$$

If the L_j are not all equal, then $C_2 < C_3$, so exact polar has a tighter optimized bound than nSGDM. NS5-Native also has a tighter bound when curvature heterogeneity reduces curvature costs enough to offset the finite-step NS constants. The precise condition is

$$C_1 < C_3. \tag{8}$$

See Appendix E.5 for the proof.

Singleton groups with $\varepsilon = 0$ recover Corollary E.1 and Table 3. From these stationarity bounds and coefficient comparisons, we conclude:

Claim 2. Muon's spectral constraint limits energy within native-row groups. At fixed maximum curvature and coupling, larger h_1, h_2 reduce its curvature costs; nSGDM uses maxima and the same coupling bound. Under heterogeneous native row-block curvature, exact polar has a tighter optimized bound than nSGDM. NS5 shares this advantage when heterogeneity offsets its finite-step constants, as specified by (8).

4 Representation Geometry in Controlled and Language-Model Experiments

In this section, we use toy and LLM experiments to provide evidence for our analysis of single-step progress, row-block curvature, and entry organization.

Experimental setup. Our toy experiments use teacher–student networks and row-block quadratics. The former compare six paired states at each depth in $1, 2, 4, 8, 16$, sharing teachers, initializations, and gradient batches across representations; fixed-step comparisons use exact polar with coefficient 0.002 (Appendix J). The latter jointly optimize four 16×32 matrices with one row-block curvature bound 20 and others $20/\rho$, for $\rho \in \{1, 3, 10, 30, 100, 300, 1000\}$. Methods follow theoretical schedules for 1000 updates from three paired starts, selecting rates independently per criterion from 17 common candidates (Appendix L). Our stationarity measure sums original-coordinate gradient nuclear norms across matrices, averages across seeds, then minimizes over iterations.

For LLMs, we train LLaMA2-style 130M and 600M models on FineWeb and C4, respectively, for next-token prediction (Touvron et al., 2023). Representations share momentum and NS5 with separate rate selection; Appendix I specifies training and both rate-selection protocols.

4.1 Single-Step Parameter-Error Progress at Shared States

We first test the decomposition in Theorem 3.2 and MP prediction in Corollary F.1. For joint progress across layers, median relative prediction error is 0.14% with measured alignment versus 157.4% with alignment 1 (Appendix J), supporting the joint role of nuclear support and update alignment in single-step progress.

Next, we optimize each representation's common step across layers using (3). With exact polar and NS5, Native, Square, and Skinny-50% outperform Vector at every paired state of depths $2, 4, 8$; Vector leads at depth 1. Thus directional differences persist after step optimization and vary with depth; Appendix J.1 gives further controls.

4.2 Heterogeneous Row-Block Curvature and Entry Organization

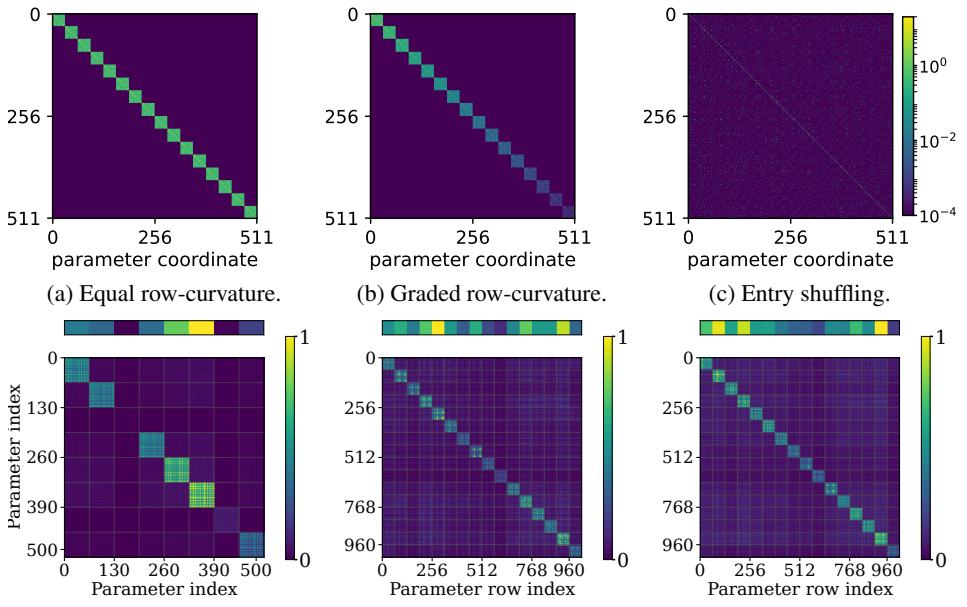
We next test the curvature analysis in Proposition 3.4. The measured Hessian structure in Figure 3 supports row-block modeling; Appendix M compares stages and layers.

First, we consider uncoupled single rows. Increasing ρ raises $h_r = 16/(1 + 15\rho^{-r})$, $r \in \{1, 2\}$, reducing Native's average row-curvature costs. In Figure 4(a–c), nSGDM has lower final loss at $\rho = 1$; at $\rho = 1000$, NS5-Native, exact-polar Native, and nSGDM reach 1.73×10^{-7}, 1.38×10^{-7}, and 1.89×10^{-6}, with both Native variants also attaining lower stationarity measures. This agrees with the curvature analysis: exact polar has a tighter bound for $\rho > 1$, and NS5 under (8).

Next, we add within-group and cross-group coupling with $1, 5, 5, 5$ rows per group, preserving h_r. Figure 4(d–f) shows the same trend, supporting the row-block analysis with coupling.

To further test energy allocation, we compare row-normalized momentum, RMNP (Deng et al., 2026). Relative to nSGDM, it typically lowers shared-state directional curvature (Appendix L, Figure 17); across eight fresh starts, it lowers mean final loss for $\rho \geq 10$, while nSGDM leads at $\rho = 1$. This supports row energy's role in curvature costs.

Finally, we shuffle entries at fixed objective, initialization, and shape to change row grouping. At $\rho = 1000$, NS5's final loss rises to 7.32×10^{-6}, above nSGDM (Figure 15), supporting the role of correspondence between entry organization and row-block curvature. Appendix L gives coupling changes and exact-polar controls.



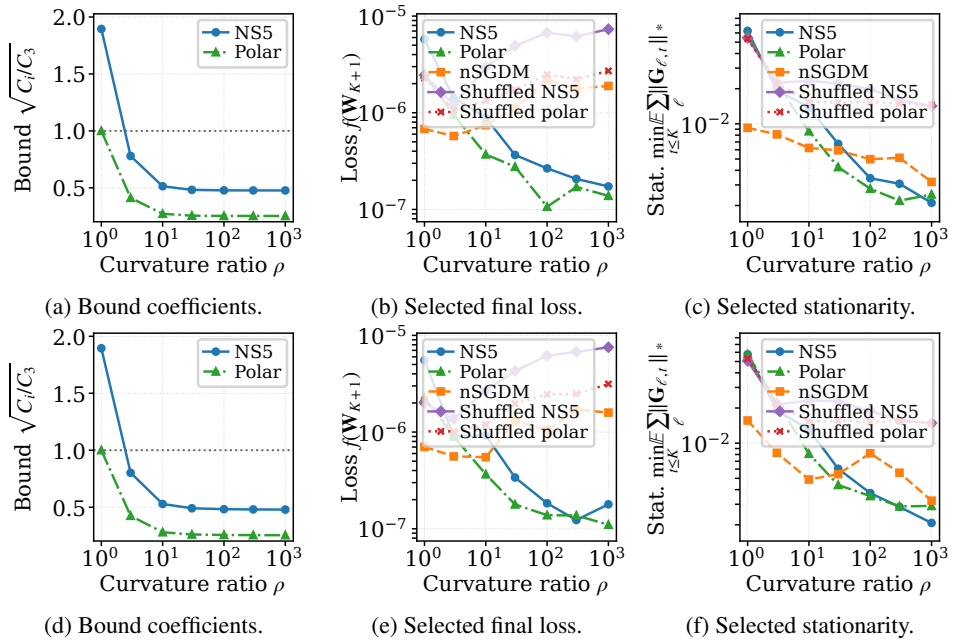
(a) Equal row-curvature. (b) Graded row-curvature. (c) Entry shuffling.

(d) MLP Hessian. (e) LLaMA query matrix. (f) LLaMA key matrix.

Figure 3: **Measured curvature exhibits block heterogeneity and weak coupling.** (a–c) Hessians with row-curvature bounds equal to 20, graded from 20 to 0.02, and fixed shuffling of the latter; shared logarithmic scale. (d) Replicated eight-neuron MLP Zhang et al. (2024, Appendix D.3): maximum-to-average block-norm and squared-norm ratios 2.61, 3.96. (e,f) LLaMA-600M layer-16 query/key Hessian responses: maximum-to-average attention-head response ratios 7.76, 5.59. Off-block/within-block strength ratios for (d–f) are 0.008, 0.128, 0.111. Strips show normalized block spectral norms (d) or attention-head responses (e,f); square-root color is capped at the within-block 99th percentile.

4.3 Order-Preserving Reshaping in Language-Model Training

Building on the controlled experiments, we examine representation throughout language-model training. Figure 1(a,b) and Tables 1, 2 show lower final validation loss for Native and Square than Vector at both scales, with Skinny intermediate: the advantage survives ordered reshaping.



(a) Bound coefficients. (b) Selected final loss. (c) Selected stationarity.

(d) Bound coefficients. (e) Selected final loss. (f) Selected stationarity.

Figure 4: **Curvature coefficients and optimization exhibit consistent trends.** Uncoupled rows (a–c) and coupled groups (d–f) share h_r. Polar denotes exact polar. Columns show optimized bound-coefficient ratios to nSGDM, final loss, and the stationarity measure. Coefficients use $4C_i$ from (21), matching Table 3 (Appendix E.5); dotted lines mark equality, and corresponding columns share axes.

Table 1: **LLaMA2-130M on FineWeb.** Validation loss at 2.60B tokens.

Method	Mean ± SD
AdamW	3.3453 ± 0.0012
Native	3.2413 ± 0.0021
Square	**3.2363** ± 0.0021
Skinny-75%	3.2687 ± 0.0025
Skinny-50%	3.2900 ± 0.0010
Skinny-25%	3.2933 ± 0.0006
Skinny-12.5%	3.3067 ± 0.0006
Vector	3.4017 ± 0.0035
Fixed shuffle	3.4910 ± 0.0334

Table 2: **LLaMA2-600M on C4.** Validation loss at 12.58B tokens.

Method	Mean ± SD
Native	**2.6684** ± 0.0014
Square	2.6745 ± 0.0013
Skinny-75%	2.6856 ± 0.0012
Skinny-50%	2.6857 ± 0.0013
Skinny-25%	2.7111 ± 0.0026
Vector	2.7491 ± 0.0020
Fixed shuffle	2.8911 ± 0.0305
Stepwise shuffle	2.8882 ± 0.0301

4.4 Entry Organization at Fixed Matrix Shape

Finally, we compare fixed and stepwise shuffling at fixed matrix shape (Section 2.2; Appendix I.5). At 600M, both yield higher final validation loss than Native and Vector (Table 2). The Native and Square advantages over Vector and the shuffling disadvantage persist when only matrix learning rates are tuned, with the auxiliary AdamW rate fixed (Appendix I.2; Table 6). Consistent with the toy experiments, this supports the dependence of matrix spectral updates' advantage on entry organization.

5 Discussion

We introduced Muse to isolate matrix shape and entry organization in Muon's spectral update. Our analysis separates nuclear support, update alignment, and finite-step cost. Under native row-block curvature, spectral control replaces maximum-row curvature costs with averages in stationarity bounds. Ordered reshaping retains lower validation loss than nSGDM, while fixed-shape shuffling removes Native's advantage in the controlled row-block problem and 600M training. These results connect Muon's gains to how representations organize updates relative to objective curvature.

REFERENCES

Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates. *arXiv preprint arXiv:2504.05295*, 2025.

Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. Asgo: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

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

Rajendra Bhatia. Positive definite matrices. In *Positive Definite Matrices*. Princeton university press, 2009.

Thibaut Boissin, Thomas Massena, Franck Mamalet, and Mathieu Serrurier. Turbo-muon: Accelerating orthogonality-based optimization with pre-conditioning. *arXiv preprint arXiv:2512.04632*, 2025.

Valentyn Boreiko, Zhiqi Bu, and Sheng Zha. Towards understanding of orthogonalization in muon. In *High-dimensional Learning Dynamics 2025*, 2025.

Da Chang and Ganzhao Yuan. Mgup: A momentum-gradient alignment update policy for stochastic optimization. *Advances in Neural Information Processing Systems*, 38:20488–20537, 2026.

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

Da Chang, Qiankun Shi, Lvgang Zhang, Yu Li, and Ruijie Zhang. A note on stability for orthogonalized matrix momentum with client sampling. *arXiv preprint arXiv:2606.01720*, 2026a.

Da Chang, Qiankun Shi, Lvgang Zhang, Yu Li, Ruijie Zhang, Yao Lu, Yongxiang Liu, and Ganzhao Yuan. Muoneq: Balancing before orthogonalization with lightweight equilibration. *arXiv preprint arXiv:2603.28254*, 2026b.

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

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

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. *arXiv preprint arXiv:2002.03305*, 2020.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *ArXiv*, abs/1905.10018, 2019. URL https://api.semanticscholar.org/CorpusID:165163984.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Shenyang Deng, Zhuoli Ouyang, Tianyu Pang, Zihang Liu, Ruochen Jin, Shuhua Yu, and Yaoqing Yang. Rmnp: Row-momentum normalized preconditioning for scalable matrix-based optimization. *arXiv preprint arXiv:2603.20527*, 2026.

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

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Runa Eschenhagen, Alexander Immer, Richard Turner, Frank Schneider, and Philipp Hennig. Kronecker-factored approximate curvature for modern neural network architectures. *Advances in Neural Information Processing Systems*, 36:33624–33655, 2023.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *ArXiv*, abs/1807.01695, 2018. URL https://api.semanticscholar.org/CorpusID:49572205.

Vladimir Feinberg, Xinyi Chen, Y Jennifer Sun, Rohan Anil, and Elad Hazan. Sketchy: Memory-efficient adaptive regularization with frequent directions. *Advances in Neural Information Processing Systems*, 36:75911–75924, 2023.

Kevin Frans, Sergey Levine, and Pieter Abbeel. A stable whitening optimizer for efficient neural network training. *arXiv preprint arXiv:2506.07254*, 2025.

Ekaterina Grishina, Matvey Smirnov, and Maxim Rakhuba. Accelerating newton-schulz iteration for orthogonalization via chebyshev-type polynomials. *arXiv preprint arXiv:2506.10935*, 2025.

Kaja Gruntkowska, Alexander Gaponov, Zhirayr Tovmasyan, and Peter Richtárik. Error feedback for muon and friends. *arXiv preprint arXiv:2510.00643*, 2025.

Aman Gupta, Rafael Celente, Abhishek Shivanna, DT Braithwaite, Gregory Dexter, Shao Tang, Hiroto Udagawa, Daniel Silva, Rohan Ramanath, and S Sathiya Keerthi. Effective quantization of muon optimizer states. *arXiv preprint arXiv:2509.23106*, 2025.

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

Di He, Songjun Tu, Ajay Jaiswal, Li Shen, Ganzhao Yuan, Shiwei Liu, and Lu Yin. Alphadecay: Module-wise weight decay for heavy-tailed balancing in llms. *Advances in Neural Information Processing Systems*, 38:29745–29768, 2026a.

Di He, Songjun Tu, Keyu Wang, Lu Yin, and Shiwei Liu. One lr doesn't fit all: Heavy-tail guided layerwise learning rates for llms. *arXiv preprint arXiv:2605.22297*, 2026b.

Feihu Huang, Junyi Li, and Heng Huang. Super-adam: Faster and universal framework of adaptive gradients. *ArXiv*, abs/2106.08208, 2021. URL https://api.semanticscholar.org/CorpusID:235436027.

Florian Hübler, Thomas Pethick, and Suvrit Sra. Free heavy-tailed lunch for muon: A theoretical justification of empirical success. *arXiv preprint arXiv:2606.14560*, 2026.

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

Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization. *arXiv preprint arXiv:2510.16981*, 2025.

Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with newton-schulz. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lJSfxtLpLm.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015*, 2015. URL http://arxiv.org/abs/1412.6980.

Tim Tsz-Kit Lau, Qi Long, and Weijie Su. Polargrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a3cc50126338b175e56bb3cad134db0b-Abstract-Conference.html.

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

Mingyi Li and Taira Tsuchiya. Muon with finite newton-schulz: The smoothing benefit in non-smooth nonconvex optimization. *arXiv preprint arXiv:2608.26288*, 2026. URL https://arxiv.org/abs/2608.26288.

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.

Zihan Liu, Ruiheng Zheng, Shaobo Zhang, Changxin Tian, Kunlong Chen, Zhiqiang Zhang, and Lei Wu. Effective learning rate governs loss dynamics in language model pretraining. *arXiv preprint arXiv:2608.24814*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Chao Ma, Wenbo Gong, Meyer Scetbon, and Edward Meeds. Swan: Sgd with normalization and whitening enables stateless llm training. *ArXiv*, abs/2412.13148, 2024. URL https://api.semanticscholar.org/CorpusID:274788976.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

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

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

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.

Weronika Ormaniec, Felix Dangel, and Sidak Pal Singh. What does it mean to be a transformer? insights from a theoretical hessian analysis. In *International Conference on Learning Representations*, volume 2025, pp. 17046–17081, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2b2bf329be5da02422a1d15ce4a81fdb-Abstract-Conference.html.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.

Zhaonan Qu, Wenzhi Gao, Oliver Hinder, Yinyu Ye, and Zhengyuan Zhou. Optimal diagonal preconditioning. *Operations Research*, 73(3):1479–1495, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Daniel Ruiz. A scaling algorithm to equilibrate both rows and columns norms in matrices. Technical report, CM-P00040415, 2001.

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

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and muons: Optimization via stochastic frank-wolfe. *arXiv preprint arXiv:2506.04192*, 2025.

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

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. *ArXiv*, abs/1804.04235, 2018. URL https://api.semanticscholar.org/CorpusID:4786918.

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

Hao-Jun Michael Shi, Tsung-Hsien Lee, Shintaro Iwasaki, Jose Gallego-Posada, Zhijing Li, Kaushik Rangadurai, Dheevatsa Mudigere, and Michael Rabbat. A distributed data-parallel pytorch implementation of the distributed shampoo optimizer for training neural networks at-scale. *arXiv preprint arXiv:2309.06497*, 2023.

Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the ideal: Analyzing the inexact muon update. *arXiv preprint arXiv:2510.19933*, 2025.

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

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

Jianlin Su. Newton-schulz iteration of the msign operator (part 1), May 2025. URL https://spaces.ac.cn/archives/10922.

Akiyoshi Tomihari and Issei Sato. Understanding transformer optimization via gradient heterogeneity, 2025. URL https://api.semanticscholar.org/CorpusID:276094438.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melissa Hall Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. URL https://api.semanticscholar.org/CorpusID:259950998.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

Bohan Wang, Jingwen Fu, Huishuai Zhang, Nanning Zheng, and Wei Chen. Closing the gap between the upper bound and lower bound of adam's iteration complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/7ac19fdcdf4f311f3e3ef2e7ef4784d7-Abstract-Conference.html.

Jinbo Wang, Mingze Wang, Zhanpeng Zhou, Junchi Yan, Weinan E, and Lei Wu. The sharpness disparity principle in transformers for accelerating language model pre-training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine*

Learning Research, pp. 64859–64879, 2025. URL https://proceedings.mlr.press/v267/wang25dl.html.

Kaiyue Wen, Xingyu Dang, Kaifeng Lyu, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them II: Hyperball optimization. *arXiv preprint arXiv:2606.16899*, 2026a.

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations*, volume 2026, pp. 144731–144838, 2026b.

Ziqing Wen, Jiahuan Wang, Ping Luo, Dongsheng Li, and Tao Sun. Foam: Blocked state folding for memory-efficient llm training. *arXiv preprint arXiv:2512.07112*, 2025.

Yihao Xiao, Jialong Sun, Zitian Gao, Zeming Wei, Chutian Wang, Ran Tao, Jiaye Teng, and Bryan Dai. Hyperball may not be a free lunch. *arXiv preprint arXiv:2607.22444*, 2026.

Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9508–9520, 2024.

Ruihan Xu, Jiajin Li, and Yiping Lu. On the width scaling of neural optimizers under matrix operator norms i: Row/column normalization and hyperparameter transfer. *arXiv preprint arXiv:2603.09952*, 2026.

Huizhuo Yuan, Yifeng Liu, Shuang Wu, Xun Zhou, and Quanquan Gu. Mars: Unleashing the power of variance reduction for training large models. *arXiv preprint arXiv:2411.10438*, 2024.

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

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

Yubo Zhang and Junhong Lin. On convergence of muon for nonconvex stochastic optimization under generalized smoothness. *Authorea Preprints*, 2026.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ee0e45ff4de76cbfdf07015a7839f339-Paper-Conference.pdf.

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*, volume 2025, pp. 28033–28063, 2025b.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *Journal of machine learning research*, 21(103):1–63, 2020.

APPENDIX

A MORE RELATED WORK

► **Adaptive and structured preconditioning.** Coordinate-wise adaptive methods use gradient-history statistics (Duchi et al., 2011; Kingma & Ba, 2015; Loshchilov & Hutter, 2019; Shazeer & Stern, 2018). Natural-gradient, K-FAC, Shampoo, SOAP, and related tensor methods organize curvature at the matrix or layer scale (Martens & Grosse, 2015; Gupta et al., 2018; Shi et al., 2023; Eschenhagen et al., 2023; Vyas et al., 2024). Extensions refine momentum and variance control, sketching, whitening, and module-wise calibration (Xie et al., 2024; Yuan et al., 2024; Huang et al., 2021; Chang & Yuan, 2026; Feinberg et al., 2023; An et al., 2025; Wen et al., 2025; Ma et al., 2024; Frans et al., 2025; He et al., 2026b;a).

► **Muon convergence and inexact polar updates.** Muon approximates the polar factor of matrix momentum through Newton–Schulz iterations (Jordan et al., 2024; Bernstein & Newhouse, 2024; Liu et al., 2025; Shah et al., 2025; DeepSeek-AI, 2026). Analyses interpret polar updates through spectral-norm descent, norm-constrained linear minimization, stochastic Frank–Wolfe, orthogonalization, and matrix-gradient preconditioning (Chen et al., 2025; Pethick et al., 2025; Sfyraki & Wang, 2025; Boreiko et al., 2025; Lau et al., 2025). Stochastic nonconvex convergence has been studied under several assumptions (Li & Hong, 2025; Shen et al., 2025; Sato et al., 2025; Chang et al., 2025; Nagashima & Iiduka, 2026; Zhang & Lin, 2026); spectral flattening and inexact-polar error have separate analyses (Nguyen et al., 2026; Shulgin et al., 2025). Kim & Oh (2026) establish finite-step NS convergence for a Taylor-truncated polynomial. Choudhury et al. (2026) analyze inexact polar steps with Nesterov momentum and heavy-tailed noise, while Li & Tsuchiya (2026) study finite-step NS smoothing in nonsmooth nonconvex optimization. Our analysis treats a fixed-coefficient NS map across fixed representations. Other variants alter polar computation or optimizer components through communication, quantization, adaptation, balancing, hybridization, and width scaling (Su, 2025; Amsel et al., 2025; Boissin et al., 2025; Grishina et al., 2025; Ahn et al., 2025; Khaled et al., 2025; Gruntkowska et al., 2025; Gupta et al., 2025; Si et al., 2025; Zhang et al., 2026; 2025a; Chang et al., 2026b; Xu et al., 2026; Chang et al., 2026a).

► **Spectral targets and matrix representation.** Spectral-flattening analyses examine orthogonalization's effects on learning rate and convergence (Nguyen et al., 2026), while random or inverted target spectra probe the role of singular-value targets (Shumaylov et al., 2026). Spectral–nuclear norm duality and polar-factor identities describe the resulting matrix geometry (Bhatia, 2009); norm-constrained linear minimization connects it to steepest descent (Pethick et al., 2025). Classical random-matrix theory provides shape-dependent spectral baselines (Marčenko & Pastur, 1967), which we use to calibrate nuclear support across matrix representations. Matrix balancing and diagonal equilibration rescale rows, columns, or coordinates (Ruiz, 2001; Qu et al., 2025). **Muse** instead rearranges entries without changing their values or Frobenius norm, separating matrix shape from entry organization.

► **Hessian structure and curvature heterogeneity.** Hessian-block heterogeneity describes spectral differences among principal blocks under a specified parameter partition (Zhang et al., 2024). Adammini uses dense Hessian sub-blocks to assign block-specific scalar learning rates, so several blocks may lie within one parameter matrix (Zhang et al., 2025b). Near-block-diagonality describes off-diagonal coupling, studied in small networks by Dong et al. (2025). Transformer components also differ in sharpness (Wang et al., 2025) and self-attention Hessian dependencies (Ormaniec et al., 2025), while gradient heterogeneity provides a related optimization perspective (Tomihari & Sato, 2025). Assumption 3.4 groups native rows and bounds cross-group coupling. Our comparison relates spectral update energy to size-weighted averages of row-curvature bounds; singleton groups recover the row-curvature ratios h_1, h_2.

B PROOF OF THEOREM 3.1

Proof. Let $\mathbf{A} = T(\mathbf{G})$ and let $\mathbf{P} = \mathrm{Orth}(\mathbf{A})$. Since T is a bijection and $\mathbf{G} \neq \mathbf{0}$, $\mathbf{A} \neq \mathbf{0}$. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be a compact SVD. Then $\mathbf{P} = \mathbf{U}\mathbf{V}^\top$ and $\|\mathbf{P}\|_2 = 1$. Set $\mathbf{D} = T^{-1}(\mathbf{P}) = \mathbf{D}_T(\mathbf{G})$. Frobenius isometry gives, for every $\Delta\mathbf{W}$,

$$\langle \mathbf{G}, \Delta\mathbf{W} \rangle_F = \langle T(\mathbf{G}), T(\Delta\mathbf{W}) \rangle_F = \langle \mathbf{A}, T(\Delta\mathbf{W}) \rangle_F. \tag{9}$$

Consequently,

$$\langle \mathbf{G}, \mathbf{D} \rangle_F = \langle \mathbf{A}, \mathbf{P} \rangle_F = \operatorname{tr}(\mathbf{\Sigma}) = \|\mathbf{A}\|_* = \|T(\mathbf{G})\|_*. \tag{10}$$

Let $\|T(\Delta\mathbf{W})\|_2 \le 1$ and set $\mathbf{H} = T(\Delta\mathbf{W})$. By (9) and spectral–nuclear duality,

$$\langle \mathbf{G}, \Delta\mathbf{W} \rangle_F = \langle \mathbf{A}, \mathbf{H} \rangle_F \ge -\|\mathbf{A}\|_* \|\mathbf{H}\|_2 \ge -\|\mathbf{A}\|_*. \tag{11}$$

The choice $\Delta\mathbf{W} = -\mathbf{D}$ satisfies $\|T(\Delta\mathbf{W})\|_2 = \|\mathbf{P}\|_2 = 1$ and attains equality in (11) by (10). This proves the steepest-descent characterization over the unit spectral ball.

The smooth descent bound follows from Assumption 3.1. Since $\operatorname{Orth}(\mathbf{A})$ has Frobenius norm squared equal to $\operatorname{rank}(\mathbf{A})$, Frobenius isometry gives

$$\|\mathbf{D}\|_F^2 = \|\operatorname{Orth}(T(\mathbf{G}))\|_F^2 = \operatorname{rank}(T(\mathbf{G})).$$

Substituting $\Delta\mathbf{W} = -\eta\mathbf{D}$ into the smoothness upper bound proves (2).

For the related spectral quadratic model in Appendix F, the same duality argument also determines the optimal step length. Define

$$\mathcal{Q}_{T,\lambda}(\Delta\mathbf{W}) = f(\mathbf{W}) + \langle \mathbf{G}, \Delta\mathbf{W} \rangle_F + \frac{\lambda}{2}\|T(\Delta\mathbf{W})\|_2^2.$$

For arbitrary $\Delta\mathbf{W}$, set $\mathbf{H} = T(\Delta\mathbf{W})$. Using (9),

$$\mathcal{Q}_{T,\lambda}(\Delta\mathbf{W}) - f(\mathbf{W}) = \langle \mathbf{A}, \mathbf{H} \rangle_F + \frac{\lambda}{2}\|\mathbf{H}\|_2^2 \ge -\|\mathbf{A}\|_* \|\mathbf{H}\|_2 + \frac{\lambda}{2}\|\mathbf{H}\|_2^2.$$

The scalar lower bound is minimized at $\|\mathbf{H}\|_2 = \|\mathbf{A}\|_*/\lambda$, with value $-\|\mathbf{A}\|_*^2/(2\lambda)$. Equality is attained by $\mathbf{H}^\star = -(\|\mathbf{A}\|_*/\lambda)\mathbf{P}$. Thus

$$T^{-1}(\mathbf{H}^\star) = -\frac{\|T(\mathbf{G})\|_*}{\lambda}\mathbf{D} = -\frac{\|T(\mathbf{G})\|_*}{\lambda}\mathbf{D}_T(\mathbf{G}) \in \arg\min_{\Delta\mathbf{W}} \mathcal{Q}_{T,\lambda}(\Delta\mathbf{W}),$$
$$\min_{\Delta\mathbf{W}} \mathcal{Q}_{T,\lambda}(\Delta\mathbf{W}) = f(\mathbf{W}) - \frac{\|T(\mathbf{G})\|_*^2}{2\lambda}.$$

□

B.1 Permutation equivariance

Let $\mathbf{P}$ and $\mathbf{Q}$ be row and column permutation matrices. If $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is a compact SVD, then $(\mathbf{P}\mathbf{U})\mathbf{\Sigma}(\mathbf{Q}\mathbf{V})^\top$ is one for $\mathbf{P}\mathbf{A}\mathbf{Q}^\top$, so $\operatorname{Orth}(\mathbf{P}\mathbf{A}\mathbf{Q}^\top) = \mathbf{P}\operatorname{Orth}(\mathbf{A})\mathbf{Q}^\top$. For the finite-step map, Frobenius invariance gives the transformed initialization $\mathbf{P}\mathbf{X}_0\mathbf{Q}^\top$. Each polynomial term in (4) obeys

$$\left[(\mathbf{P}\mathbf{X}_j\mathbf{Q}^\top)(\mathbf{P}\mathbf{X}_j\mathbf{Q}^\top)^\top\right]^k(\mathbf{P}\mathbf{X}_j\mathbf{Q}^\top) = \mathbf{P}(\mathbf{X}_j\mathbf{X}_j^\top)^k\mathbf{X}_j\mathbf{Q}^\top, \qquad k \in \{0,1,2\}.$$

Induction therefore yields $\mathcal{N}_J(\mathbf{P}\mathbf{A}\mathbf{Q}^\top) = \mathbf{P}\mathcal{N}_J(\mathbf{A})\mathbf{Q}^\top$, including the zero-input convention. Applying the inverse permutations leaves the update in the original parameter coordinates unchanged. General entry permutations change the grouping of coordinates in these matrix products, giving the entry-organization interventions studied in the experiments.

C Momentum Alignment

Proposition C.1. *Let T be a fixed Frobenius-isometric representation with short side p_T. For any current gradient $\mathbf{G}$ and momentum state $\mathbf{M}$,*

$$|\langle \mathbf{G}, \mathbf{D}_T(\mathbf{M}) \rangle_F - \|T(\mathbf{M})\|_*| \le \|T(\mathbf{M} - \mathbf{G})\|_* \le \sqrt{p_T}\|\mathbf{M} - \mathbf{G}\|_F.$$

Consequently,

$$\langle \mathbf{G}, \mathbf{D}_T(\mathbf{M}) \rangle_F \ge \|T(\mathbf{G})\|_* - 2\sqrt{p_T}\|\mathbf{M} - \mathbf{G}\|_F.$$

Proof. By Frobenius isometry and the definition of $\mathbf{D}_T$,

$$\begin{aligned}\langle \mathbf{G}, \mathbf{D}_T(\mathbf{M})\rangle_F &= \langle T(\mathbf{G}), \mathrm{Orth}(T(\mathbf{M}))\rangle_F \\ &= \langle T(\mathbf{M}), \mathrm{Orth}(T(\mathbf{M}))\rangle_F - \langle T(\mathbf{M}-\mathbf{G}), \mathrm{Orth}(T(\mathbf{M}))\rangle_F \\ &= \|T(\mathbf{M})\|_* - \langle T(\mathbf{M}-\mathbf{G}), \mathrm{Orth}(T(\mathbf{M}))\rangle_F.\end{aligned}$$

Therefore

$$\begin{aligned}|\langle \mathbf{G}, \mathbf{D}_T(\mathbf{M})\rangle_F - \|T(\mathbf{M})\|_*| &= |\langle T(\mathbf{M}-\mathbf{G}), \mathrm{Orth}(T(\mathbf{M}))\rangle_F| \\ &\overset{(\circ)}{\le} \|T(\mathbf{M}-\mathbf{G})\|_* \\ &\overset{(\star)}{\le} \sqrt{p_T}\|\mathbf{M}-\mathbf{G}\|_F.\end{aligned}$$

$(\circ)$ uses spectral–nuclear duality and $\|\mathrm{Orth}(T(\mathbf{M}))\|_2 \le 1$; $(\star)$ uses the short-side bound and Frobenius isometry. The lower alignment bound follows from

$$\begin{aligned}\langle \mathbf{G}, \mathbf{D}_T(\mathbf{M})\rangle_F &\ge \|T(\mathbf{M})\|_* - \|T(\mathbf{M}-\mathbf{G})\|_* \\ &\overset{(\circ)}{\ge} \|T(\mathbf{G})\|_* - 2\|T(\mathbf{M}-\mathbf{G})\|_* \\ &\overset{(\star)}{\ge} \|T(\mathbf{G})\|_* - 2\sqrt{p_T}\|\mathbf{M}-\mathbf{G}\|_F.\end{aligned}$$

$(\circ)$ uses the reverse triangle inequality for the nuclear norm, and $(\star)$ uses $\|T(\mathbf{M}-\mathbf{G})\|_* \le \sqrt{p_T}\|\mathbf{M}-\mathbf{G}\|_F$. This estimate quantifies the loss of descent due to the discrepancy between momentum and the current gradient. Lemma E.3 extends it to the finite-step Newton–Schulz directions used in Theorem 3.3. □

D Supplementary Setup and Curvature Details

Action and curvature. For an objective twice continuously differentiable along the update segment, first-order action and curvature cost satisfy

$$f(\mathbf{W}) - f(\mathbf{W}-\eta\mathbf{D}) = \eta\langle \mathbf{G}, \mathbf{D}\rangle_F - \eta^2 \int_0^1 (1-s)\left\langle \mathbf{D}, \nabla^2 f(\mathbf{W}-s\eta\mathbf{D})\mathbf{D}\right\rangle_F ds.$$

For the exact-polar direction $\mathbf{D} = \mathbf{D}_T(\mathbf{M})$, the first term is $\eta\chi_T(\mathbf{G},\mathbf{M})\|T(\mathbf{G})\|_*$; the second depends on how the update distributes energy across parameter coordinates. This connects representation geometry to the objective's curvature.

NS5 bounds.

Lemma D.1. *For every $\mathbf{A} \in \mathbb{R}^{p_T \times q_T}$, the map in* (4) *with $J = 5$ satisfies*

$$\|\mathcal{N}_5(\mathbf{A})\|_2 \le \frac{5}{4}, \qquad \langle \mathbf{A}, \mathcal{N}_5(\mathbf{A})\rangle_F \ge \max\left\{\frac{1}{2\sqrt{p_T}}, \frac{972}{p_T + 1458}\right\}\|\mathbf{A}\|_*. \tag{12}$$

The proof is given in Appendix E.2.

At $p_T = 1024$, the five-step constant improves the uniform lower bound from 0.015625 to approximately 0.392.

One-matrix curvature bounds. When $\mathbf{H} = \mathrm{blkdiag}_i \mathbf{H}_i$ with $\|\mathbf{H}_i\|_2 \le L_i$, the scaled NS5 direction $\mathbf{D} = \mathcal{N}_5(\mathbf{M})/\sqrt{p}$ satisfies

$$\|\mathbf{D}_{i,:}\|_2^2 \le \frac{25}{16p}, \qquad |\langle \mathbf{D}, \mathbf{H}\mathbf{D}\rangle_F| \le \frac{25\max_i L_i}{16h_1}, \qquad \|\mathbf{H}\mathbf{D}\|_F^2 \le \frac{25\max_i L_i^2}{16h_2}.$$

The corresponding nSGDM bounds are $\max_i L_i$ and $\max_i L_i^2$. At fixed maxima, larger h_1, h_2 therefore reduce these NS5 curvature bounds. Equation (21) gives the multi-matrix comparison.

Interpretation of the comparison coefficient. For singleton groups and zero coupling, Native curvature terms involve row averages. Since $p^{-1}\sum_i L_i^r = \max_i L_i^r/h_r$ for $r \in \{1,2\}$, their exact coefficient comparison with nSGDM is

$$\frac{25\sqrt{2}}{8\alpha c_5^2}\frac{\max_i L_i^2}{h_2} + \frac{(c_5+5/4)^2}{2\alpha c_5^2} + \frac{25}{32c_5^2}\frac{\max_i L_i}{h_1} < \frac{2\sqrt{2}}{\alpha}\max_i L_i^2 + \frac{2}{\alpha} + \frac{\max_i L_i}{2}.$$

The left side is C_1, the right side is C_3, and their optimized coefficient ratio is $\sqrt{C_1/C_3}$. At fixed $\max_i L_i$, increasing h_1 or h_2 reduces the left-side curvature terms.

Remark D.1. *For one matrix, Native uses the $1/\sqrt{p}$ scaling in Proposition 3.4. With $p = 1024$ and one nonzero row-curvature bound L, the optimized NS5-to-nSGDM coefficient ratio tends to $5/(4c_5\sqrt{p}) \approx 0.100$ as $L \to \infty$. Thus NS5 has the smaller optimized bound coefficient for sufficiently large L.*

E Proof of Theorem 3.3

The proof connects representation geometry to objective progress through three estimates: spectral control of the direction, first-order action after accounting for momentum error, and accumulated curvature cost. The fixed quintic map satisfies the spectral and alignment estimates at every fixed iteration count. Related convergence analyses include (Chang et al., 2025; Kim & Oh, 2026; Hübler et al., 2026).

Throughout this section, T is fixed with short side p_T, and (c, B) is the pair specified in Theorem 3.3.

E.1 Finite-step Newton–Schulz bounds

Lemma E.1. *Let*

$$\phi(s) = a_0 s + a_1 s^3 + a_2 s^5, \qquad \psi_J = \phi^{\circ J}.$$

With the coefficients fixed as in (4)*, for every $s \in [0,1]$ and integer $J \geq 1$,*

$$0 \leq \psi_J(s) \leq \frac{5}{4}, \qquad \psi_J(s) \geq \min\{s, 1/2\} \geq \frac{s}{2}. \tag{13}$$

Moreover, $\phi(s) > 0$ for $s > 0$.

Proof. Recall that $(a_0, a_1, a_2) = (3.4445, -4.7750, 2.0315)$. Write $\phi(s) = sh(s^2)$, where

$$h(u) = a_0 + a_1 u + a_2 u^2 = a_2\left(u + \frac{a_1}{2a_2}\right)^2 + a_0 - \frac{a_1^2}{4a_2}.$$

The fixed coefficients satisfy $a_2 > 0$ and $a_0 - a_1^2/(4a_2) > 5/8$, so $h(u) > 5/8$ and $\phi(s) > 0$ for $s > 0$. Moreover,

$$\phi'(s) = a_0 + 3a_1 s^2 + 5a_2 s^4.$$

Substitution gives $\phi'(0.554) > 0$, $\phi'(0.555) < 0$, and $\phi'(5/4) > 0$. Since ϕ' is quadratic in s^2, it has exactly two positive zeros in $(0, 5/4)$. The smaller zero, $s_0 \in (0.554, 0.555)$, is a local maximum of ϕ; the other is a local minimum. The maximum on $[0, 5/4]$ therefore occurs at s_0 or $5/4$. For the fixed coefficients, $\phi(5/4) < 5/4$, and $a_0, a_2 > 0 > a_1$ gives

$$\phi(s_0) \leq a_0(0.555) + a_1(0.554)^3 + a_2(0.555)^5 < \frac{5}{4}.$$

Together with positivity, this proves $\phi([0, 5/4]) \subseteq [0, 5/4]$.

Finally, $h'(u) = a_1 + 2a_2 u < 0$ on $[0, 16/25]$ and $h(16/25) > 1$, so $\phi(s) \geq s$ on $[0, 4/5]$. For $s \in [4/5, 5/4]$, the bound $h(u) > 5/8$ gives

$$\phi(s) = sh(s^2) > \frac{4}{5}\cdot\frac{5}{8} = \frac{1}{2}.$$

It follows that $\phi(s) \geq \min\{s, 1/2\}$ on $[0, 5/4]$. Induction within this invariant interval yields $\psi_J(s) \geq \min\{s, 1/2\}$, which is at least $s/2$ on $[0, 1]$. Thus every polynomial iterate stays in $[0, 5/4]$ and retains at least half its input singular value, as claimed in (13). □

Lemma E.2. *For every $\mathbf{A} \in \mathbb{R}^{p_T \times q_T}$ and integer $J \geq 1$,*

$$\begin{aligned} \|\mathcal{N}_J(\mathbf{A})\|_2 \leq \frac{5}{4}, \qquad \|\mathcal{N}_J(\mathbf{A})\|_F^2 \leq \frac{25 p_T}{16}, \\ \langle \mathbf{A}, \mathcal{N}_J(\mathbf{A}) \rangle_F \geq \frac{1}{2}\|\mathbf{A}\|_F \geq \frac{1}{2\sqrt{p_T}}\|\mathbf{A}\|_*. \end{aligned} \tag{14}$$

Proof. The claims hold at zero by definition. Otherwise, write the compact SVD $\mathbf{A} = \mathbf{U}_r \operatorname{diag}(\sigma_1, \ldots, \sigma_r)\mathbf{V}_r^\top$ using its nonzero singular values, and set $s_i = \sigma_i / \|\mathbf{A}\|_F$. Direct multiplication in (4) gives

$$\mathcal{N}_J(\mathbf{A}) = \mathbf{U}_r \operatorname{diag}(\psi_J(s_1), \ldots, \psi_J(s_r))\mathbf{V}_r^\top.$$

Each $s_i \in (0, 1]$, so Lemma E.1 bounds every output singular value by $5/4$. This gives the operator bound and, since $r \leq p_T$, the Frobenius bound. The lower scalar bound gives

$$\langle \mathbf{A}, \mathcal{N}_J(\mathbf{A}) \rangle_F = \sum_{i=1}^{r} \sigma_i \psi_J(s_i) \geq \frac{\sum_{i=1}^{r} \sigma_i^2}{2\|\mathbf{A}\|_F} = \frac{1}{2}\|\mathbf{A}\|_F.$$

Finally, $\|\mathbf{A}\|_* \leq \sqrt{p_T}\|\mathbf{A}\|_F$. Hence every fixed $J \geq 1$ has spectral constant $B = 5/4$ and update-alignment constant $c = 1/(2\sqrt{p_T})$, the constants used for finite-step NS in Theorem 3.3. □

E.2 The five-step improvement

Proof of Lemma D.1. We strengthen the scalar estimate in Lemma E.1. For its polynomial $\phi(s) = 3.4445s - 4.7750s^3 + 2.0315s^5$,

$$\phi(s) \geq \min\{3s, 2/3\}, \qquad 0 \leq s \leq 5/4.$$

Indeed, for $0 \leq s \leq 2/9$, $\phi(s) - 3s \geq s(0.4445 - 4.7750(2/9)^2) \geq 0$. On $[2/9, 5/4]$, the derivative has two positive roots, as established in Lemma E.1; only the larger is a local minimum. Denote it by u. The signs $\phi'(1.050136) < 0 < \phi'(1.050137)$ and the quadratic form of ϕ' in s^2 give $1.050136 < u < 1.050137$, and hence

$$\phi(u) \geq 3.4445(1.050136) - 4.7750(1.050137)^3 + 2.0315(1.050136)^5 > 0.6818 > 2/3.$$

The endpoint values satisfy $\phi(2/9) > 0.714$ and $\phi(5/4) > 1.179$. Together with the invariant interval from Lemma E.1, induction yields $\psi_5(s) \geq \min\{243s, 2/3\}$ for $0 \leq s \leq 1$.

For nonzero $\mathbf{A}$, let $s_i = \sigma_i(\mathbf{A})/\|\mathbf{A}\|_F$, padded with zeros to p_T terms. Thus $\sum_i s_i^2 = 1$. Write $S = \sum_i s_i$ and $Z = \sum_i s_i \psi_5(s_i)$. Since $\min\{243s, 2/3\} \geq (2/3)s$ on $[0, 1]$, we have $Z \geq 2/3$. Also,

$$243s^2 - \frac{2}{3}s = 243\left(s - \frac{1}{729}\right)^2 - \frac{1}{2187} \geq -\frac{1}{2187},$$

so $Z \geq (2/3)S - p_T/2187$. It follows that

$$\frac{2}{3}S \leq Z + \frac{p_T}{2187} \leq Z\left(1 + \frac{p_T}{1458}\right), \qquad \frac{Z}{S} \geq \frac{972}{p_T + 1458}.$$

The singular-value formula in the proof of Lemma E.2 identifies Z/S with $\langle \mathbf{A}, \mathcal{N}_5(\mathbf{A}) \rangle_F / \|\mathbf{A}\|_*$. Thus the fixed five-step map has alignment at least $972/(p_T + 1458)$ times the nuclear norm. Combining this with the all-J bound $1/(2\sqrt{p_T})$ and the spectral bound $5/4$ gives both inequalities in (12); the zero case follows by definition. □

E.3 Alignment and momentum tracking

Write the chosen update as $\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \mathbf{D}_t$. The exact polar identities and Lemmas E.2 and D.1 give

$$\|T(\mathbf{D}_t)\|_2 \leq B, \qquad \langle \mathbf{M}_t, \mathbf{D}_t \rangle_F \geq c\|T(\mathbf{M}_t)\|_*. \tag{15}$$

Lemma E.3. *For either update,*

$$\begin{aligned} \|\mathbf{D}_t\|_F^2 \leq B^2 p_T, \\ \langle \mathbf{G}_t, \mathbf{D}_t \rangle_F \geq c\|T(\mathbf{G}_t)\|_* - (c + B)\sqrt{p_T}\|\mathbf{M}_t - \mathbf{G}_t\|_F. \end{aligned} \tag{16}$$

Proof. Frobenius isometry and the short-side rank bound give the first assertion. For the second, spectral–nuclear duality and the reverse triangle inequality give

$$\begin{aligned}\langle \mathbf{G}_t, \mathbf{D}_t\rangle_F &= \langle \mathbf{M}_t, \mathbf{D}_t\rangle_F - \langle T(\mathbf{M}_t - \mathbf{G}_t), T(\mathbf{D}_t)\rangle_F \\ &\geq c\|T(\mathbf{M}_t)\|_* - B\|T(\mathbf{M}_t - \mathbf{G}_t)\|_* \\ &\geq c\|T(\mathbf{G}_t)\|_* - (c+B)\|T(\mathbf{M}_t - \mathbf{G}_t)\|_* \\ &\geq c\|T(\mathbf{G}_t)\|_* - (c+B)\sqrt{p_T}\|\mathbf{M}_t - \mathbf{G}_t\|_F.\end{aligned}$$

For $c = B = 1$, this recovers Proposition C.1. □

Let $\mathcal{F}_t = \sigma(\mathcal{F}_0, \xi_1, \ldots, \xi_t)$, with $\mathbf{W}_1$ measurable with respect to $\mathcal{F}_0$. Both spectral direction maps are Borel measurable. Thus $\mathbf{W}_t$ and $\mathbf{G}_t$ are $\mathcal{F}_{t-1}$-measurable, while $\mathbf{M}_t, \mathbf{D}_t, \mathbf{W}_{t+1}$ are $\mathcal{F}_t$-measurable. The deterministic displacement bound

$$\|\mathbf{W}_t - \mathbf{W}_1\|_F \leq B\sqrt{p_T}\sum_{s=1}^{t-1}\eta_s$$

and smoothness imply finite second moments of $\mathbf{G}_t$ at every finite t. The oracle assumptions and momentum recursion then give finite second moments of $\widehat{\mathbf{G}}_t$ and $\mathbf{M}_t$. The smoothness estimate

$$|f(\mathbf{W}_t) - f(\mathbf{W}_1)| \leq \|\mathbf{G}_1\|_F\|\mathbf{W}_t - \mathbf{W}_1\|_F + \frac{L}{2}\|\mathbf{W}_t - \mathbf{W}_1\|_F^2$$

also gives integrability of $f(\mathbf{W}_t)$. Hence all expectations below are well defined.

Lemma E.4. *Suppose nonnegative sequences satisfy*

$$e_{t+1} \leq (1 - r_{t+1})e_t + a_{t+1}, \qquad r_t = rt^{-p}, \quad r \in (0, 1], \quad p \in (0, 1].$$

Then $r_t e_t \leq 2(e_t - e_{t+1} + a_{t+1})$.

Proof. The recursion implies

$$r_t e_t - 2(e_t - e_{t+1} + a_{t+1}) \leq (r_t - 2r_{t+1})e_t \leq 0,$$

since $r_t/r_{t+1} = ((t+1)/t)^p \leq 2$. □

Lemma E.5. *Under the assumptions of Theorem 3.3, set*

$$\mathbf{S}_t = \mathbf{M}_t - \mathbf{G}_t, \qquad e_t = \mathbb{E}\|\mathbf{S}_t\|_F^2, \qquad r_t = 1 - \alpha_t = \alpha t^{-1/2}.$$

For $1 \leq t < K$,

$$e_{t+1} \leq (1 - r_{t+1})e_t + \left(\frac{2\sqrt{2}L^2B^2p_T\eta^2}{\alpha} + \alpha^2\sigma^2\right)(t+1)^{-1}. \tag{17}$$

Consequently, for $K \geq 2$,

$$\sum_{t=1}^{K} r_t e_t \leq 2e_1 + 2\left(\frac{2\sqrt{2}L^2B^2p_T\eta^2}{\alpha} + \alpha^2\sigma^2\right)\sum_{t=2}^{K}\frac{1}{t}, \tag{18}$$

where $e_1 \leq (1-\alpha)^2\mathbb{E}\|\mathbf{G}_1\|_F^2 + \alpha^2\sigma^2$.

Proof. Since $\mathbf{M}_1 = \alpha\widehat{\mathbf{G}}_1$, conditioning on $\mathcal{F}_0$ and using unbiasedness eliminates the cross term in

$$e_1 = \mathbb{E}\|\alpha(\widehat{\mathbf{G}}_1 - \mathbf{G}_1) - (1-\alpha)\mathbf{G}_1\|_F^2,$$

giving the initialization bound, including at $\alpha = 1$. For $1 \leq t < K$, the error recursion is

$$\mathbf{S}_{t+1} = (1 - r_{t+1})(\mathbf{S}_t + \mathbf{G}_t - \mathbf{G}_{t+1}) + r_{t+1}(\widehat{\mathbf{G}}_{t+1} - \mathbf{G}_{t+1}).$$

The first term is $\mathcal{F}_t$-measurable, so its cross term with the new oracle noise vanishes after conditioning on $\mathcal{F}_t$. For $r \in (0, 1]$, the nonnegative-square identity

$$(1-r)\|\mathbf{U}\|_F^2 + \frac{(1-r)^2}{r}\|\mathbf{V}\|_F^2 - (1-r)^2\|\mathbf{U}+\mathbf{V}\|_F^2 = r(1-r)\left\|\mathbf{U} - \frac{1-r}{r}\mathbf{V}\right\|_F^2$$

therefore yields

$$\begin{aligned} e_{t+1} &\le (1-r_{t+1})e_t + \frac{(1-r_{t+1})^2}{r_{t+1}}\mathbb{E}\|\mathbf{G}_t - \mathbf{G}_{t+1}\|_F^2 + r_{t+1}^2\sigma^2 \\ &\le (1-r_{t+1})e_t + \frac{L^2B^2p_T\eta_t^2}{r_{t+1}} + r_{t+1}^2\sigma^2. \end{aligned}$$

The second line uses smoothness and $\|\mathbf{W}_{t+1} - \mathbf{W}_t\|_F^2 \le B^2p_T\eta_t^2$. Substituting the schedules and

$$\frac{\eta_t^2}{r_{t+1}} = \frac{\eta^2}{\alpha}t^{-3/2}(t+1)^{1/2} \le \frac{2\sqrt{2}\eta^2}{\alpha(t+1)}$$

proves (17). Apply Lemma E.4 for $t = 1, \dots, K-1$, then add r_Ke_K:

$$\sum_{t=1}^K r_te_t \le 2e_1 + 2\left(\frac{2\sqrt{2}L^2B^2p_T\eta^2}{\alpha} + \alpha^2\sigma^2\right)\sum_{t=2}^K\frac{1}{t} - (2-r_K)e_K.$$

Dropping the last nonpositive term proves (18). □

E.4 The common stationarity bound

Proof of Theorem 3.3. For either direction, smoothness gives

$$f(\mathbf{W}_{t+1}) \le f(\mathbf{W}_t) - \eta_t\langle\mathbf{G}_t, \mathbf{D}_t\rangle_F + \frac{L\eta_t^2}{2}\|\mathbf{D}_t\|_F^2. \tag{19}$$

Lemma E.3 therefore implies

$$\begin{aligned} f(\mathbf{W}_{t+1}) \le f(\mathbf{W}_t) &- c\eta_t\|T(\mathbf{G}_t)\|_* \\ &+ (c+B)\sqrt{p_T}\eta_t\|\mathbf{S}_t\|_F + \frac{LB^2p_T}{2}\eta_t^2. \end{aligned} \tag{20}$$

Young's inequality bounds the tracking term by

$$(c+B)\sqrt{p_T}\eta_t\|\mathbf{S}_t\|_F \le \frac{r_t}{2}\|\mathbf{S}_t\|_F^2 + \frac{(c+B)^2p_T\eta_t^2}{2r_t}.$$

Taking expectations, summing, and using $f \ge f^\star$ gives

$$\begin{aligned} c\sum_{t=1}^K\eta_t\mathbb{E}\|T(\mathbf{G}_t)\|_* \le \mathbb{E}[f(\mathbf{W}_1)] - f^\star + \frac{1}{2}\sum_{t=1}^K r_te_t \\ + \frac{(c+B)^2p_T\eta^2}{2\alpha}\sum_{t=1}^K t^{-1} + \frac{LB^2p_T\eta^2}{2}\sum_{t=1}^K t^{-3/2}. \end{aligned}$$

By Lemma E.5 and $\sum_{t=1}^K t^{-3/2} \le \sum_{t=1}^K t^{-1} \le 1 + \ln K$,

$$\begin{aligned} c\sum_{t=1}^K\eta_t\mathbb{E}\|T(\mathbf{G}_t)\|_* \le \mathbb{E}[f(\mathbf{W}_1)] - f^\star + (1-\alpha)^2\mathbb{E}\|\mathbf{G}_1\|_F^2 + \alpha^2\sigma^2 \\ + (1+\ln K)\left(\frac{2\sqrt{2}L^2B^2p_T\eta^2}{\alpha} + \alpha^2\sigma^2 + \frac{(c+B)^2p_T\eta^2}{2\alpha} + \frac{LB^2p_T\eta^2}{2}\right). \end{aligned}$$

Finally, $\sum_{t=1}^{K} \eta_t \geq \eta K^{1/4}$, and

$$\min_{1\leq t\leq K} \mathbb{E}\|T(\nabla f(\mathbf{W}_t))\|_* \leq \frac{\sum_{t=1}^{K} \eta_t \mathbb{E}\|T(\mathbf{G}_t)\|_*}{\sum_{t=1}^{K} \eta_t}.$$

Dividing the preceding weighted bound by $c\eta K^{1/4}$ gives exactly the stationarity inequality (5). The exact polar update uses $(c, B) = (1, 1)$; every fixed-iteration NS update uses $(1/(2\sqrt{p_T}), 5/4)$; NS5 also permits $c = \max\{1/(2\sqrt{p_T}), 972/(p_T + 1458)\}$. Thus the displayed inequality applies to each of these updates and directly yields the stated $\widetilde{O}(K^{-1/4})$ rate for the nuclear norm of the represented gradient. □

E.5 Native row-block curvature and a common stationarity measure

We first bound Hessian action on groups of rows, then derive the three stationarity coefficients. Hessians act on row-stacked parameters; inner products and norms use the equivalent matrix coordinates. The native-row specialization and its product-space extension follow below.

Lemma E.6 (Direction bounds under native row-block curvature)**.** *Under the pointwise Hessian bounds in Assumption 3.4, with* $\max_j L_j > 0$*, a direction satisfying* $\|\mathbf{U}\|_2 \leq s/\sqrt{p}$*,* $s \geq 0$*, obeys*

$$\|\mathbf{HU}\|_F^2 \leq s^2 \left(\frac{\max_j L_j}{\sqrt{h_2}} + \varepsilon\right)^2, \qquad |\langle \mathbf{U}, \mathbf{HU}\rangle_F| \leq s^2 \left(\frac{\max_j L_j}{h_1} + \varepsilon\right).$$

NS5-Native permits $s = 5/4$*, and exact-polar Native permits* $s = 1$*. For nSGDM, the corresponding bounds are* $(\max_j L_j + \varepsilon)^2$ *and* $\max_j L_j + \varepsilon$*.*

Proof. Since the groups partition p rows,

$$\frac{\max_j L_j^r}{p} \leq \frac{1}{p}\sum_j |I_j| L_j^r \leq \max_j L_j^r.$$

Thus $1 \leq h_r \leq p$, and $p^{-1}\sum_j |I_j| L_j^r = \max_j L_j^r / h_r$. For each group, summing squared row norms gives

$$\|\mathbf{U}_{I_j,:}\|_F^2 = \sum_{a\in I_j} \|\mathbf{U}_{a,:}\|_2^2 \leq |I_j| \|\mathbf{U}\|_2^2 \leq \frac{s^2 |I_j|}{p}, \qquad \|\mathbf{U}\|_F^2 \leq s^2.$$

Since $\mathbf{B}$ is block diagonal,

$$\|\mathbf{BU}\|_F^2 = \sum_j \|\mathbf{H}_{jj}\mathbf{U}_{I_j,:}\|_F^2 \leq \sum_j L_j^2 \|\mathbf{U}_{I_j,:}\|_F^2 \leq s^2 \frac{\max_j L_j^2}{h_2},$$
$$|\langle \mathbf{U}, \mathbf{BU}\rangle_F| \leq \sum_j L_j \|\mathbf{U}_{I_j,:}\|_F^2 \leq s^2 \frac{\max_j L_j}{h_1}.$$

The residual satisfies $\|(\mathbf{H} - \mathbf{B})\mathbf{U}\|_F \leq \varepsilon \|\mathbf{U}\|_F$ and $|\langle \mathbf{U}, (\mathbf{H} - \mathbf{B})\mathbf{U}\rangle_F| \leq \varepsilon \|\mathbf{U}\|_F^2$. Thus the triangle inequality gives

$$\|\mathbf{HU}\|_F \leq s\frac{\max_j L_j}{\sqrt{h_2}} + s\varepsilon, \qquad |\langle \mathbf{U}, \mathbf{HU}\rangle_F| \leq s^2 \frac{\max_j L_j}{h_1} + s^2\varepsilon.$$

Squaring the first inequality proves the claim. Lemma D.1 and $\|\operatorname{Orth}(\mathbf{M})\|_2 \leq 1$ give the stated values of s. For nSGDM, $\|\mathbf{U}\|_F \leq 1$ and $\|\mathbf{H}\|_2 \leq \|\mathbf{B}\|_2 + \varepsilon \leq \max_j L_j + \varepsilon$; applying the operator-norm bound to $\mathbf{HU}$ and its inner product with $\mathbf{U}$ gives both assertions. □

Proof of Proposition 3.4. Write $\mathbf{G}_t = \nabla f(\mathbf{W}_t)$, $\mathbf{S}_t = \mathbf{M}_t - \mathbf{G}_t$, $e_t = \mathbb{E}\|\mathbf{S}_t\|_F^2$, and $r_t = \alpha t^{-1/2}$. We track four constants: γ, d in the alignment inequality and V, Q bounding squared Hessian action

and the absolute quadratic form. Their values, in method order $i = 1, 2, 3$, are

$$(\gamma, d, V, Q) = \left(c_5, c_5 + \tfrac{5}{4}, \tfrac{25}{16} \left(\frac{\max_j L_j}{\sqrt{h_2}} + \varepsilon \right)^2, \tfrac{25}{16} \left(\frac{\max_j L_j}{h_1} + \varepsilon \right) \right),$$
$$(\gamma, d, V, Q) = \left(1, 2, \left(\frac{\max_j L_j}{\sqrt{h_2}} + \varepsilon \right)^2, \frac{\max_j L_j}{h_1} + \varepsilon \right),$$
$$(\gamma, d, V, Q) = \left(1, 2, (\max_j L_j + \varepsilon)^2, \max_j L_j + \varepsilon \right).$$

For the two spectral directions, Lemma E.3, divided by $\sqrt{p}$, gives

$$\langle \mathbf{G}_t, \mathbf{U}_t \rangle_F \geq \frac{\gamma}{\sqrt{p}} \|\mathbf{G}_t\|_* - d\|\mathbf{S}_t\|_F.$$

For nonzero nSGDM momentum,

$$\langle \mathbf{G}_t, \mathbf{U}_t \rangle_F \geq \|\mathbf{M}_t\|_F - \|\mathbf{S}_t\|_F \geq \|\mathbf{G}_t\|_F - 2\|\mathbf{S}_t\|_F \geq \frac{\|\mathbf{G}_t\|_*}{\sqrt{p}} - 2\|\mathbf{S}_t\|_F.$$

At zero momentum the right side is nonpositive, so the inequality still holds. Lemma E.6 gives the stated V, Q.

Gradient drift and descent. The fundamental theorem of calculus and Jensen's inequality give

$$\mathbf{G}_{t+1} - \mathbf{G}_t = -\eta_t \int_0^1 \nabla^2 f(\mathbf{W}_t - u\eta_t \mathbf{U}_t)\mathbf{U}_t \, du,$$
$$\|\mathbf{G}_{t+1} - \mathbf{G}_t\|_F^2 \leq \eta_t^2 \int_0^1 \|\nabla^2 f(\mathbf{W}_t - u\eta_t \mathbf{U}_t)\mathbf{U}_t\|_F^2 du \leq \eta_t^2 V.$$

The integral Taylor formula gives

$$\begin{aligned} f(\mathbf{W}_{t+1}) &= f(\mathbf{W}_t) - \eta_t \langle \mathbf{G}_t, \mathbf{U}_t \rangle_F \\ &\quad + \eta_t^2 \int_0^1 (1-u) \langle \mathbf{U}_t, \nabla^2 f(\mathbf{W}_t - u\eta_t \mathbf{U}_t)\mathbf{U}_t \rangle_F du \\ &\leq f(\mathbf{W}_t) - \frac{\gamma \eta_t}{\sqrt{p}} \|\mathbf{G}_t\|_* + d\eta_t \|\mathbf{S}_t\|_F + \frac{Q\eta_t^2}{2}. \end{aligned}$$

Momentum tracking. The initialization and error recursion are

$$e_1 = \mathbb{E}\|\alpha(\widehat{\mathbf{G}}_1 - \mathbf{G}_1) - (1-\alpha)\mathbf{G}_1\|_F^2 \leq (1-\alpha)^2 \mathbb{E}\|\mathbf{G}_1\|_F^2 + \alpha^2\sigma^2,$$
$$\mathbf{S}_{t+1} = (1 - r_{t+1})(\mathbf{S}_t + \mathbf{G}_t - \mathbf{G}_{t+1}) + r_{t+1}(\widehat{\mathbf{G}}_{t+1} - \mathbf{G}_{t+1}).$$

Conditional unbiasedness eliminates the noise cross terms. The weighted-square inequality proved in Lemma E.5 therefore yields

$$\begin{aligned} e_{t+1} &\leq (1 - r_{t+1})e_t + \frac{(1 - r_{t+1})^2}{r_{t+1}} \mathbb{E}\|\mathbf{G}_t - \mathbf{G}_{t+1}\|_F^2 + r_{t+1}^2 \sigma^2 \\ &\leq (1 - r_{t+1})e_t + \left(\frac{2\sqrt{2}\eta^2 V}{\alpha} + \alpha^2 \sigma^2 \right) (t+1)^{-1}, \end{aligned}$$

where $(1 - r_{t+1})^2 \leq 1$ and $\eta_t^2 / r_{t+1} = (\eta^2/\alpha) t^{-3/2} (t+1)^{1/2} \leq 2\sqrt{2}\eta^2/[\alpha(t+1)]$. Summing Lemma E.4 from 1 to $K-1$ and adding $r_K e_K$ gives

$$\sum_{t=1}^{K} r_t e_t \leq 2e_1 + 2 \left(\frac{2\sqrt{2}\eta^2 V}{\alpha} + \alpha^2 \sigma^2 \right) \sum_{t=2}^{K} \frac{1}{t} - (2 - r_K) e_K.$$

The final term is nonpositive and can be dropped.

Summation and step-size choice. Young's inequality bounds the alignment error by

$$d\eta_t\|\mathbf{S}_t\|_F \le \frac{r_t}{2}\|\mathbf{S}_t\|_F^2 + \frac{d^2\eta_t^2}{2r_t}.$$

Taking expectations in the descent inequality, summing, and using $f \ge f^\star$ yields

$$\begin{aligned}\frac{\gamma}{\sqrt{p}}\sum_{t=1}^{K}\eta_t\mathbb{E}\|\mathbf{G}_t\|_* &\le \mathbb{E}[f(\mathbf{W}_1)] - f^\star + \frac{1}{2}\sum_{t=1}^{K} r_t e_t \\ &\quad + \frac{d^2\eta^2}{2\alpha}\sum_{t=1}^{K} t^{-1} + \frac{Q\eta^2}{2}\sum_{t=1}^{K} t^{-3/2} \\ &\le \mathbb{E}[f(\mathbf{W}_1)] - f^\star + e_1 \\ &\quad + (1+\ln K)\left[\alpha^2\sigma^2 + \eta^2\left(\frac{2\sqrt{2}V}{\alpha} + \frac{d^2}{2\alpha} + \frac{Q}{2}\right)\right].\end{aligned}$$

Here $\sum t^{-3/2} \le \sum t^{-1} \le 1 + \ln K$. Substituting the three sets of method constants gives

$$\frac{2\sqrt{2}V}{\alpha} + \frac{d^2}{2\alpha} + \frac{Q}{2} = \gamma^2 C_i.$$

The initialization bound and $1 + \ln K \ge 1$ thus bound the preceding right side by $(1 + \ln K)(R + \eta^2\gamma^2 C_i)$. Since $\sum_{t=1}^{K}\eta_t = \eta\sum_{t=1}^{K} t^{-3/4} \ge \eta K^{1/4}$, taking a weighted minimum gives

$$\min_{t\le K}\mathbb{E}\|\mathbf{G}_t\|_* \le \frac{\sqrt{p}(1+\ln K)}{K^{1/4}}\left(\frac{R}{\gamma\eta} + \gamma\eta C_i\right).$$

For $R > 0$, the derivative of the parenthesis is $-R/(\gamma\eta^2) + \gamma C_i$. It vanishes at $\eta = \sqrt{R}/(\gamma\sqrt{C_i})$, where the second derivative is positive and the value is $2\sqrt{RC_i}$. This proves (7). Finally, nonempty groups and unequal L_j give $h_1, h_2 > 1$; hence both curvature terms in C_2 are strictly smaller than in C_3, proving $C_2 < C_3$. The common factors in the optimized bounds are positive, so the NS5 bound is smaller exactly when $C_1 < C_3$. □

Corollary E.1 (Native-row curvature)**.** *Under Proposition 3.4, let $I_j = \{j\}$ and $\varepsilon = 0$. Then $\mathbf{H} = \mathrm{blkdiag}_j \mathbf{H}_{jj}$, and $h_r = \max_j L_j^r/(p^{-1}\sum_j L_j^r)$. The coefficients in* (7) *reduce to Table 3.*

Proof. Singleton groups give $p^{-1}\sum_j L_j^r = \max_j L_j^r/h_r$. Setting $\varepsilon = 0$ in C_1, C_2, C_3 leaves curvature terms $\max_j L_j^2/h_2$, $\max_j L_j/h_1$, and the corresponding maxima for nSGDM. Substitution gives the table. □

Table 3: One-matrix stationarity bounds on $\min_{t\le K}\mathbb{E}\|\nabla f(\mathbf{W}_t)\|_*$ for singleton groups and zero cross-group coupling; $\widetilde{O}$ hides logarithmic factors. Proposition 3.4 defines R and c_5; Equation (6) defines h_1, h_2. Heterogeneous native row-block curvature gives $C_2 < C_3$ for exact polar; NS5 requires $C_1 < C_3$. Increasing h_1, h_2 at fixed p, α, and maximum row curvature lowers C_1 and the optimized coefficient ratio $\sqrt{C_1/C_3}$ at unchanged C_3.

Method	Optimized upper bound	Curvature dependence
nSGDM	$\widetilde{O}\left(\frac{\sqrt{pRC_3}}{K^{1/4}}\right)$	$C_3 = \frac{2\sqrt{2}}{\alpha}\max_i L_i^2 + \frac{2}{\alpha} + \frac{\max_i L_i}{2}$
Muon (exact polar)	$\widetilde{O}\left(\frac{\sqrt{pRC_2}}{K^{1/4}}\right)$	$C_2 = \frac{2\sqrt{2}}{\alpha}\frac{\max_i L_i^2}{h_2} + \frac{2}{\alpha} + \frac{\max_i L_i}{2h_1}$
Muon (NS5)	$\widetilde{O}\left(\frac{\sqrt{pRC_1}}{K^{1/4}}\right)$	$C_1 = \frac{25\sqrt{2}}{8\alpha}\frac{\max_i L_i^2}{c_5^2 h_2} + \frac{(c_5+5/4)^2}{2\alpha c_5^2} + \frac{25\max_i L_i}{32 c_5^2 h_1}$

Product-space extension. For b jointly optimized matrices with $p_\ell \le q_\ell$, retain the oracle, smoothness, and initialization hypotheses of Proposition 3.4. Along all update segments, assume the full Hessian is $\mathrm{blkdiag}_{\ell,j}\mathbf{H}_{\ell j}$, with one block per native row and common bounds $\|\mathbf{H}_{\ell j}\|_2 \le L_{\ell j}$. Set

$P = \max_\ell p_\ell$ and $c = \max\{1/(2\sqrt{P}), 972/(P + 1458)\}$. Scale each NS5-Native or exact-polar Native direction by $1/\sqrt{P}$; normalize each nSGDM momentum matrix by its own Frobenius norm. All matrices use the same schedules and base step size. Let R be the initialization and noise expression in the proposition for the joint objective, with squared Frobenius norms summed over matrices, and define

$$\Gamma_1 = \frac{1}{c^2}\left(\frac{25\sqrt{2}}{8\alpha P}\sum_{\ell,j} L_{\ell j}^2 + \frac{(c+5/4)^2 b}{2\alpha} + \frac{25}{32P}\sum_{\ell,j} L_{\ell j}\right),$$

$$\Gamma_2 = \frac{2\sqrt{2}}{\alpha P}\sum_{\ell,j} L_{\ell j}^2 + \frac{2b}{\alpha} + \frac{1}{2P}\sum_{\ell,j} L_{\ell j},$$

$$\Gamma_3 = \frac{2\sqrt{2}}{\alpha}\sum_{\ell} \max_j L_{\ell j}^2 + \frac{2b}{\alpha} + \frac{1}{2}\sum_{\ell} \max_j L_{\ell j}.$$

For $R > 0$, separate optimization of the base step size yields

$$\min_{1\le t\le K} \mathbb{E}\sum_{\ell=1}^{b} \|\nabla_{\mathbf{W}_\ell} f(\mathbf{W}_t)\|_* \le \frac{2(1+\ln K)\sqrt{PR\Gamma_i}}{K^{1/4}}, \qquad i \in \{1,2,3\}. \tag{21}$$

To derive these coefficients, row-energy bounds and block diagonality give, for NS5,

$$\|\mathbf{HU}\|_F^2 \le \sum_{\ell,j} L_{\ell j}^2 \|\mathbf{U}_{\ell,j,:}\|_2^2 \le \frac{25}{16P}\sum_{\ell,j} L_{\ell j}^2, \quad |\langle \mathbf{U}, \mathbf{HU}\rangle_F| \le \frac{25}{16P}\sum_{\ell,j} L_{\ell j}.$$

Exact polar replaces $25/16$ by 1. For nSGDM, summing $\sum_j L_{\ell j}^r \|\mathbf{U}_{\ell,j,:}\|_2^2 \le \max_j L_{\ell j}^r$, $r = 1, 2$, gives the corresponding sums of maxima. Summing the per-matrix alignment inequalities and using $\sum_\ell \|\mathbf{S}_{\ell,t}\|_F \le \sqrt{b}\|\mathbf{S}_t\|_F$ gives

$$\langle \mathbf{G}_t, \mathbf{U}_t\rangle_F \ge \frac{\gamma}{\sqrt{P}}\sum_\ell \|\mathbf{G}_{\ell,t}\|_* - d\|\mathbf{S}_t\|_F.$$

For the three methods, the constants used in the preceding proof are

$$(\gamma, d, V, Q) = \left(c, (c+\tfrac{5}{4})\sqrt{b}, \tfrac{25}{16P}\sum_{\ell,j} L_{\ell j}^2, \tfrac{25}{16P}\sum_{\ell,j} L_{\ell j}\right),$$

$$(\gamma, d, V, Q) = \left(1, 2\sqrt{b}, \tfrac{1}{P}\sum_{\ell,j} L_{\ell j}^2, \tfrac{1}{P}\sum_{\ell,j} L_{\ell j}\right),$$

$$(\gamma, d, V, Q) = \left(1, 2\sqrt{b}, \sum_{\ell} \max_j L_{\ell j}^2, \sum_{\ell} \max_j L_{\ell j}\right).$$

Joint gradient drift and descent obey the same integral estimates; conditional unbiasedness and the joint variance bound give the same tracking recursion. Substitution gives $2\sqrt{2}V/\alpha + d^2/(2\alpha) + Q/2 = \gamma^2\Gamma_i$, so summation yields

$$\min_{t\le K} \mathbb{E}\sum_\ell \|\mathbf{G}_{\ell,t}\|_* \le \frac{\sqrt{P}(1+\ln K)}{K^{1/4}}\left(\frac{R}{\gamma\eta} + \gamma\eta\Gamma_i\right).$$

Choosing $\eta = \sqrt{R}/(\gamma\sqrt{\Gamma_i})$ proves (21). If every matrix has the same p and row bounds L_j, then $P = p$, the NS5 constant is c_5, and each curvature sum and squared alignment-error coefficient is b times its one-matrix counterpart. Thus $\Gamma_i = bC_i$ in the singleton, zero-coupling case, giving $4C_i$ for the four-matrix experiment and coefficient ratio $\sqrt{C_1/C_3}$.

For independent matrices with equal p, identical group sizes, bounds L_j, and coupling bound ε, the same product-space calculation gives $\Gamma_i = bC_i$ for the row-block coefficients in Proposition 3.4.

Indeed, Lemma E.6 applies to each matrix, so squared Hessian actions and quadratic forms sum to bV and bQ. The alignment error is bounded by $d\sum_\ell \|\mathbf{S}_{\ell,t}\|_F \le d\sqrt{b}\|\mathbf{S}_t\|_F$. Hence

$$\frac{1}{\gamma^2}\left(\frac{2\sqrt{2}bV}{\alpha}+\frac{bd^2}{2\alpha}+\frac{bQ}{2}\right)=bC_i,$$

and the displayed product-space bound holds with $P=p$. This covers both rows of Figure 4.

F Supplementary Models and Single-Step Parameter-Error Progress

F.1 A spectral-norm quadratic model

For $\lambda>0$, spectral–nuclear duality gives a quadratic-model interpretation: the model $f(\mathbf{W})+\langle\mathbf{G},\Delta\mathbf{W}\rangle_F+\frac{\lambda}{2}\|T(\Delta\mathbf{W})\|_2^2$ over $\Delta\mathbf{W}$ is minimized at $-\lambda^{-1}\|T(\mathbf{G})\|_*\mathbf{D}_T(\mathbf{G})$, and the corresponding model decrease is $\|T(\mathbf{G})\|_*^2/(2\lambda)$.

F.2 MP Moment Conditions and Shape Factors

For isotropic teacher–student parameter perturbations, we model $T(\mathbf{E}_0)$ through Marchenko–Pastur (MP) theory (Marčenko & Pastur, 1967), with entry scale $\sigma_0>0$ and aspect ratio $\gamma_T=p_T/q_T$. Let μ_γ be the limiting singular-value distribution normalized by $\sigma_0\sqrt{q_T}$, with first moment $\beta(\gamma)=\int x\,d\mu_\gamma(x)$.

Assumption F.1. *Along each sequence of nonzero parameter errors $\mathbf{E}_0$ and representations T with $d=p_Tq_T\to\infty$, fix $\sigma_0>0$. Assume*

$$\frac{\sum_i\sigma_i(T(\mathbf{E}_0))}{p_T\sigma_0\sqrt{q_T}}-\beta(\gamma_T)\to0,\qquad\frac{\sum_i\sigma_i(T(\mathbf{E}_0))^2}{p_T\sigma_0^2q_T}\to1,$$

with $\beta(\gamma_T)$ uniformly bounded away from zero.

Corollary F.1. *Under Assumption F.1 and fixed $\eta>0$, $\|T(\mathbf{E}_0)\|_*/\|T(\mathbf{E}_0)\|_F^2\sim\beta(\gamma_T)/(\sigma_0\sqrt{q_T})=\beta(\gamma_T)\gamma_T^{1/4}/(\sigma_0d^{1/4})$, and*

$$\frac{\|\mathbf{W}_0-\mathbf{W}_\star\|_F^2-\|\mathbf{W}_1-\mathbf{W}_\star\|_F^2}{\eta\|\mathbf{W}_0-\mathbf{W}_\star\|_F^2}=\frac{2\chi_T\beta(\gamma_T)\gamma_T^{1/4}}{\sigma_0d^{1/4}}+o(d^{-1/4})$$

Remark F.1. *The leading support factors for Square and Vector are $\beta(1)/(\sigma_0d^{1/4})$ and $1/(\sigma_0\sqrt{d})$, respectively; their ratio is $\beta(1)d^{1/4}=\Theta(d^{1/4})$, with $\beta(1)=8/(3\pi)$. At full row rank, matching Frobenius update lengths instead gives the leading factor $2\chi_T\beta(\gamma_T)/(\sigma_0\sqrt{d})$.*

At full row rank, $\|\mathbf{D}_T(\mathbf{Z}_0)\|_F^2=p_T$. Normalizing the update length in the preceding support calibration gives

$$\frac{\|T(\mathbf{E}_0)\|_*}{\|\mathbf{E}_0\|_F\|\mathbf{D}_T(\mathbf{Z}_0)\|_F}\sim\beta(\gamma_T).$$

Appendices H and F give the moment ratios, finite-step remainder, and learning-rate compensation.

F.3 Learning-Rate Compensation under MP Calibration

Using the error and alignment notation from Section 3.2, MP calibration relates the step coefficient to nuclear support and signed update alignment. Equation (3) gives the exact optimum along a realized direction; the comparison below describes the leading-order compensation between representations under the spectral model.

Corollary F.2 (Learning-rate compensation)**.** *Under Assumption F.1, let $\chi_{T_1},\chi_{T_2}>0$. For each representation and its chosen step size, assume*

$$\eta_T\,\mathrm{rank}(T(\mathbf{Z}_0))=o(\chi_T\|T(\mathbf{E}_0)\|_*)\,.$$

If two representations with the same d, σ_0, and step size attain the same leading-order relative decrease, then $\chi_{T_1}\beta(\gamma_{T_1})\gamma_{T_1}^{1/4} \sim \chi_{T_2}\beta(\gamma_{T_2})\gamma_{T_2}^{1/4}$. Thus a larger gain in nuclear support permits a proportionally smaller update alignment. Allowing representation-dependent step sizes, equal leading-order single-step parameter-error progress under MP calibration requires

$$\eta_{T_1}\chi_{T_1}\frac{\beta(\gamma_{T_1})}{\sqrt{q_{T_1}}} \sim \eta_{T_2}\chi_{T_2}\frac{\beta(\gamma_{T_2})}{\sqrt{q_{T_2}}}.$$

Here η_T includes any scalar multiplier applied to the exact-polar direction.

The proof is given in Appendix H.

F.4 Local Learning-Rate Compensation Across Depth

Norm control and effective-learning-rate matching relate nominal learning rates to relative parameter motion (Wen et al., 2026a; Xiao et al., 2026; Liu et al., 2026). Near the reference optimum, the remaining distance $\|\mathbf{E}_0\|_F$ provides the corresponding local scale: $\eta/\|\mathbf{E}_0\|_F$ is the error-normalized step coefficient, and $\eta\|\mathbf{D}_T(\mathbf{Z}_0)\|_F/\|\mathbf{E}_0\|_F$ is the relative update length. The leading fractional single-step parameter-error progress factors as

$$2\eta\chi_T\frac{\|T(\mathbf{E}_0)\|_*}{\|\mathbf{E}_0\|_F^2} = 2\frac{\eta}{\|\mathbf{E}_0\|_F}\chi_T\frac{\|T(\mathbf{E}_0)\|_*}{\|\mathbf{E}_0\|_F}.$$

Appendix J.1 evaluates this compensation for single-step parameter-error progress at common parameter states.

G Supporting Lemmas for Single-Step Parameter-Error Progress and MP Calibration

The exact identity for single-step parameter-error progress in Theorem 3.2 follows from Frobenius isometry and the compact-polar identities below. Corollaries F.1 and F.2 evaluate its support term under MP spectral calibration. This establishes the passage from the single-step parameter-error progress decomposition to its shape-dependent support model.

G.1 Compact-polar identities

Lemma G.1. *Let $\mathbf{P} = \mathrm{Orth}(\mathbf{G})$ for $\mathbf{G} \in \mathbb{R}^{p\times q}$, with $\mathrm{Orth}(\mathbf{0}) = \mathbf{0}$. Then*

$$\|\mathbf{P}\|_2 \le 1, \qquad \|\mathbf{P}\|_F^2 = \mathrm{rank}(\mathbf{G}), \qquad |\langle \mathbf{A}, \mathbf{P}\rangle_F| \le \|\mathbf{A}\|_*$$

for every conformable matrix $\mathbf{A}$.

Proof. Take a compact singular-value decomposition $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. If $\mathbf{G} \neq \mathbf{0}$, then $\mathbf{P} = \mathbf{U}\mathbf{V}^\top$, whose nonzero singular values are all one. Hence $\|\mathbf{P}\|_2 = 1$ and $\|\mathbf{P}\|_F^2 = \mathrm{rank}(\mathbf{G})$. The zero case follows from the convention above. Finally, spectral–nuclear duality gives $|\langle \mathbf{A}, \mathbf{P}\rangle_F| \le \|\mathbf{A}\|_*\|\mathbf{P}\|_2 \le \|\mathbf{A}\|_*$. □

G.2 MP spectral calibration

Lemma G.2. *Under Assumption F.1,*

$$\frac{\|T(\mathbf{E}_0)\|_*}{\|T(\mathbf{E}_0)\|_F^2} = \frac{\beta(\gamma_T)}{\sigma_0\sqrt{q_T}}\big(1 + o(1)\big) = \frac{\beta(\gamma_T)\gamma_T^{1/4}}{\sigma_0 d^{1/4}}\big(1 + o(1)\big).$$

For fixed η, the rank correction in Theorem 3.2 is lower order than the MP support scale $\beta(\gamma_T)/(\sigma_0\sqrt{q_T})$, and is $o(d^{-1/4})$.

Proof. By the two moment limits in Assumption F.1,

$$\|T(\mathbf{E}_0)\|_* = p_T\sigma_0\sqrt{q_T}\,\beta(\gamma_T)\big(1 + o(1)\big), \qquad \|T(\mathbf{E}_0)\|_F^2 = p_T\sigma_0^2 q_T\big(1 + o(1)\big).$$

Their ratio is $\beta(\gamma_T)/(\sigma_0\sqrt{q_T})(1+o(1))$; the additive first-moment assumption gives a relative error because $\beta(\gamma_T)$ is bounded away from zero. The MP singular-value law has unit second moment, so Cauchy–Schwarz gives $\beta(\gamma_T) \le 1$. Since $p_T q_T = d$ and $\gamma_T = p_T/q_T$, one has $q_T = \sqrt{d/\gamma_T}$, which gives the second expression. Lemma G.1, $p_T \le \sqrt{d}$, and $\|T(\mathbf{E}_0)\|_F^2 = \sigma_0^2 d(1+o(1))$ further give

$$\eta \frac{\operatorname{rank}(T(\mathbf{Z}_0))}{\|T(\mathbf{E}_0)\|_F^2} \le \eta \frac{p_T}{\sigma_0^2 d(1+o(1))} = O(d^{-1/2}) = o(d^{-1/4}).$$

Relative to the MP support scale, the same bound is at most

$$\frac{\eta p_T}{\|T(\mathbf{E}_0)\|_F^2} \frac{\sigma_0\sqrt{q_T}}{\beta(\gamma_T)} \le \frac{\eta}{\sigma_0 \beta(\gamma_T)\sqrt{q_T}}(1+o(1)) \longrightarrow 0,$$

since $q_T \ge \sqrt{d}$ and $\beta(\gamma_T)$ is bounded away from zero. □

H Proofs for Single-Step Parameter-Error Progress and MP Calibration

Proof of Theorem 3.2. Set $\mathbf{E}_1 := \mathbf{W}_1 - \mathbf{W}_\star$ and

$$\mathbf{P} := \operatorname{Orth}(T(\mathbf{Z}_0)), \qquad \mathbf{D} := T^{-1}(\mathbf{P}) = \mathbf{D}_T(\mathbf{Z}_0).$$

Since $\mathbf{E}_0 \neq \mathbf{0}$ and T is a Frobenius isometry, $\|T(\mathbf{E}_0)\|_* > 0$. The alignment identity follows from

$$\begin{aligned} \langle \mathbf{E}_0, \mathbf{D} \rangle_F &\overset{(\circ)}{=} \langle T(\mathbf{E}_0), \mathbf{P} \rangle_F \\ &\overset{(\star)}{=} \chi_T \|T(\mathbf{E}_0)\|_*. \end{aligned}$$

Here, $(\circ)$ uses Frobenius isometry and the definition of $\mathbf{D}$, while $(\star)$ is the definition of update alignment. Moreover,

$$\begin{aligned} |\chi_T| &\overset{(\circ)}{=} \frac{|\langle T(\mathbf{E}_0), \mathbf{P} \rangle_F|}{\|T(\mathbf{E}_0)\|_*} \\ &\overset{(\star)}{\le} \frac{\|T(\mathbf{E}_0)\|_* \|\mathbf{P}\|_2}{\|T(\mathbf{E}_0)\|_*} \\ &\overset{(\bullet)}{\le} 1. \end{aligned}$$

Here, $(\circ)$ uses the preceding isometric form of the numerator; $(\star)$ is spectral–nuclear duality; and $(\bullet)$ uses $\|\operatorname{Orth}(\cdot)\|_2 \le 1$, including the convention $\operatorname{Orth}(\mathbf{0}) = \mathbf{0}$, as in Appendix B.

For the single-step parameter-error progress expansion, linearity of T and the definition of $\mathbf{D}$ give

$$\begin{aligned} \|T(\mathbf{E}_1)\|_F^2 - \|T(\mathbf{E}_0)\|_F^2 &\overset{(\circ)}{=} \|T(\mathbf{E}_0) - \eta \mathbf{P}\|_F^2 - \|T(\mathbf{E}_0)\|_F^2 \\ &\overset{(\star)}{=} -2\eta \langle T(\mathbf{E}_0), \mathbf{P} \rangle_F + \eta^2 \|\mathbf{P}\|_F^2 \\ &\overset{(\bullet)}{=} -2\eta \chi_T \|T(\mathbf{E}_0)\|_* + \eta^2 \|\mathbf{D}\|_F^2. \end{aligned}$$

Here, $(\circ)$ substitutes the update and uses $T(\mathbf{D}) = \mathbf{P}$; $(\star)$ expands the squared Frobenius norm; and $(\bullet)$ uses the alignment identity above together with Frobenius isometry. Dividing by $\eta \|T(\mathbf{E}_0)\|_F^2 > 0$ and using $\|T(\mathbf{E}_i)\|_F = \|\mathbf{W}_i - \mathbf{W}_\star\|_F$ proves the exact identity for single-step parameter-error progress.

The finite-step term uses the same compact-polar rank identity as the proof of Theorem 3.1:

$$\begin{aligned} \|\mathbf{D}\|_F^2 &\overset{(\circ)}{=} \|\mathbf{P}\|_F^2 \\ &\overset{(\star)}{=} \operatorname{rank}(T(\mathbf{Z}_0)) \\ &\overset{(\bullet)}{\le} p_T. \end{aligned}$$

Here, $(\circ)$ is Frobenius isometry; $(\star)$ uses that the nonzero singular values of a compact polar factor are all one, with the zero-input case covered by $\operatorname{Orth}(\mathbf{0}) = \mathbf{0}$; and $(\bullet)$ uses the represented shape $p_T \times q_T$, with $p_T \le q_T$. This completes the proof of the exact law. □

Proof of Corollary F.1. Assumption F.1 and Lemma G.2 give

$$\frac{\|T(\mathbf{E}_0)\|_*}{\|T(\mathbf{E}_0)\|_F^2} = \frac{\beta(\gamma_T)\gamma_T^{1/4}}{\sigma_0 d^{1/4}}\big(1 + o(1)\big).$$

Substitution into the update coefficient yields

$$\begin{aligned}\frac{\langle \mathbf{E}_0, \mathbf{D}_T(\mathbf{Z}_0)\rangle_F}{\|T(\mathbf{E}_0)\|_F^2} &\overset{(\circ)}{=} \chi_T \frac{\|T(\mathbf{E}_0)\|_*}{\|T(\mathbf{E}_0)\|_F^2} \\ &\overset{(\star)}{=} \frac{\beta(\gamma_T)\gamma_T^{1/4}}{\sigma_0 d^{1/4}}\big(\chi_T + o(1)\big).\end{aligned}$$

Here, $(\circ)$ is the alignment identity, while $(\star)$ uses the MP spectral evaluation and $|\chi_T| \le 1$, which keeps the multiplied remainder at order $o(1)$. This yields the claimed $\chi_T \beta(\gamma_T)\gamma_T^{1/4}$ first-order scaling. The exact law now gives

$$\begin{aligned}\frac{\|\mathbf{W}_0 - \mathbf{W}_\star\|_F^2 - \|\mathbf{W}_1 - \mathbf{W}_\star\|_F^2}{\eta\|\mathbf{W}_0 - \mathbf{W}_\star\|_F^2} &= \frac{2\beta(\gamma_T)\gamma_T^{1/4}}{\sigma_0 d^{1/4}}\big(\chi_T + o(1)\big) \\ &\quad - \eta\frac{\operatorname{rank}(T(\mathbf{Z}_0))}{\|T(\mathbf{E}_0)\|_F^2}.\end{aligned}$$

Because $|\chi_T| \le 1$, the first remainder remains $o(d^{-1/4})$; the second is $O(d^{-1/2})$. Combining the two proves the displayed MP law. For the square endpoint, substitute $p_T = q_T = \sqrt{d}$. At the vector endpoint, the nuclear and Frobenius norms coincide, so the second moment assumption directly gives $\|T(\mathbf{E}_0)\|_*/\|T(\mathbf{E}_0)\|_F^2 = 1/(\sigma_0\sqrt{d})(1 + o(1))$. □

Proof of Corollary F.2. The exact identity in Theorem 3.2, multiplied by η_T, gives the relative decrease

$$2\eta_T\chi_T\frac{\|T(\mathbf{E}_0)\|_*}{\|T(\mathbf{E}_0)\|_F^2}\left(1 - \frac{\eta_T \operatorname{rank}(T(\mathbf{Z}_0))}{2\chi_T\|T(\mathbf{E}_0)\|_*}\right).$$

The condition in the corollary makes the parenthesized correction $1 + o(1)$. MP calibration then makes this decrease

$$2\eta_T\chi_T\frac{\beta(\gamma_T)}{\sigma_0\sqrt{q_T}}(1 + o(1)).$$

For two representations with the same d, σ_0, and η, equality of their leading decreases is therefore equivalent to

$$\chi_{T_1}\beta(\gamma_{T_1})\gamma_{T_1}^{1/4} \sim \chi_{T_2}\beta(\gamma_{T_2})\gamma_{T_2}^{1/4}.$$

Thus update alignment may fall in inverse proportion to the gain in nuclear support without changing leading-order progress. With representation-dependent step sizes, equality of the preceding positive leading terms instead gives

$$\eta_{T_1}\chi_{T_1}\frac{\beta(\gamma_{T_1})}{\sqrt{q_{T_1}}} \sim \eta_{T_2}\chi_{T_2}\frac{\beta(\gamma_{T_2})}{\sqrt{q_{T_2}}}.$$

□

I Experimental Details

We first specify language-model training, parameter groups, and evaluation, then distinguish the fixed-AdamW comparisons in Table 6 from the coupled-rate trajectories in Figure 1. Within each setting, representations share architecture, data, momentum, NS5, and token budget. Hessian measurements and controlled experiments are specified separately below.

I.1 Language-Model Architectures, Data, and Training Budgets

We train LLaMA2-style decoder-only Transformers (Touvron et al., 2023) with RMSNorm, tied token embeddings, no bias parameters, and no dropout. Table 4 summarizes the two settings. The 130M model has 123.59M trainable parameters, RMSNorm $\epsilon = 10^{-6}$, and vocabulary size 50,304. Its FineWeb10B stream (Penedo et al., 2024) contains 10,255,324,043 training tokens and a separate 100,000,000-token validation stream; the training budget uses 2.60B tokens. The 600M model uses C4 with T5-base tokenization (Raffel et al., 2020). We compare representations within each model–dataset setting.

Table 4: Language-model configurations. Global batch size counts sequences; token budgets equal sequence length times global batch size times training steps.

Setting	LLaMA2-130M	LLaMA2-600M
Training data	FineWeb10B	C4
Tokenizer	GPT-2	T5-base
Layers / hidden width	12 / 768	32 / 1024
Attention heads / head dimension	12 / 64	16 / 64
MLP hidden width	2048	4096
Sequence length	4096	4096
Global batch size	128	320
Training steps	4960	9600
Training tokens	2,600,468,480	12,582,912,000
Learning-rate schedule	Cosine	Warmup–stable–decay
Warmup steps	496	800
Final decay phase	—	960 steps
Weight decay	0.1	0
Gradient clipping norm	1.0	1.0
Precision	bfloat16	bfloat16
Hardware	RTX PRO 6000 Blackwell	Ascend 910C

Parameter groups and update scaling. Muon acts on the attention and MLP matrices in Table 5, using momentum 0.95, no Nesterov correction, and five Newton–Schulz iterations (Jordan et al., 2024; Liu et al., 2025). The 600M model contains 224 such blocks, totaling 536.87M parameters. Token embeddings, the tied language-model head, and normalization parameters use auxiliary AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019), with $(\beta_1, \beta_2) = (0.9, 0.95)$ at both scales; the 130M implementation uses bias correction and $\epsilon = 10^{-8}$. Following the Moonshot Muon scaling convention (Liu et al., 2025), matrix base rates are multiplied by $0.2\sqrt{\max(m, n)}$, using each block's native dimensions m, n in the selected comparisons. The theoretical step coefficient η includes this multiplier. Reshaping and shuffling therefore retain the same blockwise scalar multiplier. The 130M all-parameter baseline uses AdamW throughout.

Training and evaluation. The 130M runs use four 96GB GPUs, with 16 sequences per device and two gradient-accumulation steps per update. Their data order is held fixed across runs. Validation is performed every 100 steps and at the terminal step 4960, using one evaluation batch per evaluation. The 600M runs evaluate every 1200 steps, including step 9600; the ordered-representation runs use 16 Ascend 910C NPUs, and shuffling device counts are specified in Appendix I.5. Its linear warmup–stable–decay schedule warms up for 800 steps, remains at the base rate until the final 960 steps, and decays to 0.1 of that rate. Each parameter group follows its model's schedule. A fixed AdamW rate below means a common base rate across representations and matrix-rate candidates, with the scheduled variation retained. Learning-rate selection and reporting use terminal validation loss. Every reported SD is the sample standard deviation across three seeds; means include the seed or seeds used for selection.

I.2 Matrix Learning-Rate Selection with Fixed Auxiliary AdamW

The comparisons in Table 6 hold the auxiliary AdamW base rate at 0.008 for 130M and 0.002 for 600M. For each representation, we vary only the matrix base rate, retaining the native-dimension multiplier and all other training settings. We select the candidate with the lowest single-seed terminal validation loss, then repeat that matrix rate with two additional seeds. Table 6 reports the resulting

Module group	Count	Native shape	Nearest-square shape
LLaMA2-130M (12 layers)			
Attention QKV	12	2304×768	1152×1536
Attention output	12	768×768	768×768
MLP $\mathbf{W}_1$	12	2048×768	1024×1536
MLP $\mathbf{W}_2$	12	2048×768	1024×1536
MLP projection	12	768×2048	1024×1536
LLaMA2-600M (32 layers)			
Attention Q, K, V	96	1024×1024	1024×1024
Attention output	32	1024×1024	1024×1024
MLP $\mathbf{W}_1$	32	4096×1024	2048×2048
MLP $\mathbf{W}_2$	32	4096×1024	2048×2048
MLP projection	32	1024×4096	2048×2048

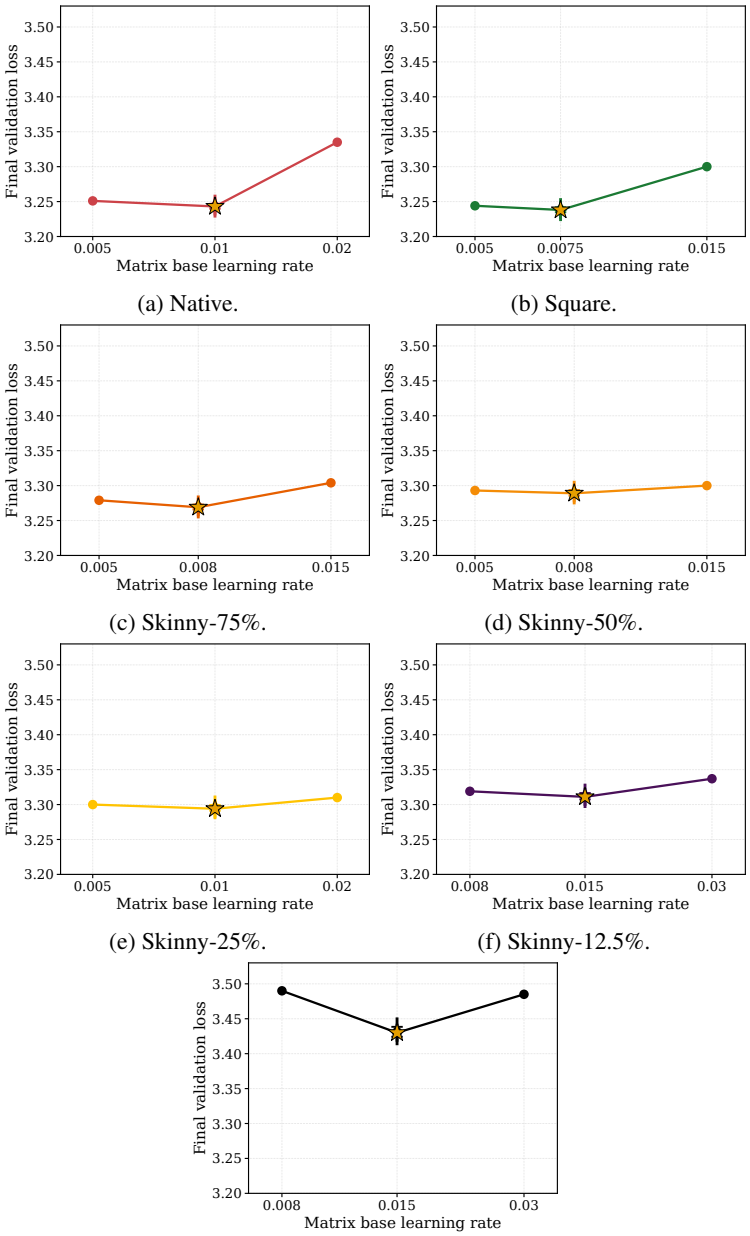
Table 5: Matrix-module shapes for the LLaMA2-130M and LLaMA2-600M experiments. Shapes are those of the optimizer blocks: QKV is concatenated in 130M and split into separate query, key, and value matrices in 600M. Token embeddings, the language-model head, normalization parameters, and other non-matrix parameters use AdamW.

endpoints. The 130M comparison includes Native, Square, four Skinny variants, and Vector; the 600M comparison includes Native, Square, Vector, and both fixed-shape shuffling controls.

Table 6: Individual terminal validation losses with fixed auxiliary AdamW. The matrix column gives the selected base rate before native-dimension scaling. Each setting uses one matrix rate for all three seeds.

Method	Fixed AdamW LR	Selected matrix LR	Seed 1	Seed 2	Seed 3	Mean $\pm$ SD
LLaMA2-130M / FineWeb, 2.60*B tokens*						
Native	0.008	0.01	3.243	3.247	3.240	3.2433 ± 0.0035
Square	0.008	0.0075	3.238	3.242	3.235	3.2383 ± 0.0035
Skinny-75%	0.008	0.008	3.269	3.273	3.266	3.2693 ± 0.0035
Skinny-50%	0.008	0.008	3.289	3.294	3.286	3.2897 ± 0.0040
Skinny-25%	0.008	0.01	3.294	3.300	3.292	3.2953 ± 0.0042
Skinny-12.5%	0.008	0.015	3.311	3.317	3.308	3.3120 ± 0.0046
Vector	0.008	0.015	3.430	3.439	3.425	3.4313 ± 0.0071
LLaMA2-600M / C4, 12.58*B tokens*						
Native	0.002	0.002	2.668	2.667	2.670	2.6683 ± 0.0015
Square	0.002	0.002	2.674	2.674	2.676	2.6747 ± 0.0012
Vector	0.002	0.02	2.790	2.800	2.795	2.7950 ± 0.0050
Fixed shuffle	0.002	0.0005	2.915	2.945	2.900	2.9200 ± 0.0229
Stepwise shuffle	0.002	0.0005	2.905	2.940	2.895	2.9133 ± 0.0236

Figures 5 and 6 include all 38 candidates: three per method at 130M; four for Native and Square and three for each other method at 600M. Each selected rate is interior to its evaluated grid. These sweeps compare representations after matrix-rate tuning with common auxiliary AdamW settings.



(a) Native.

(b) Square.

(c) Skinny-75%.

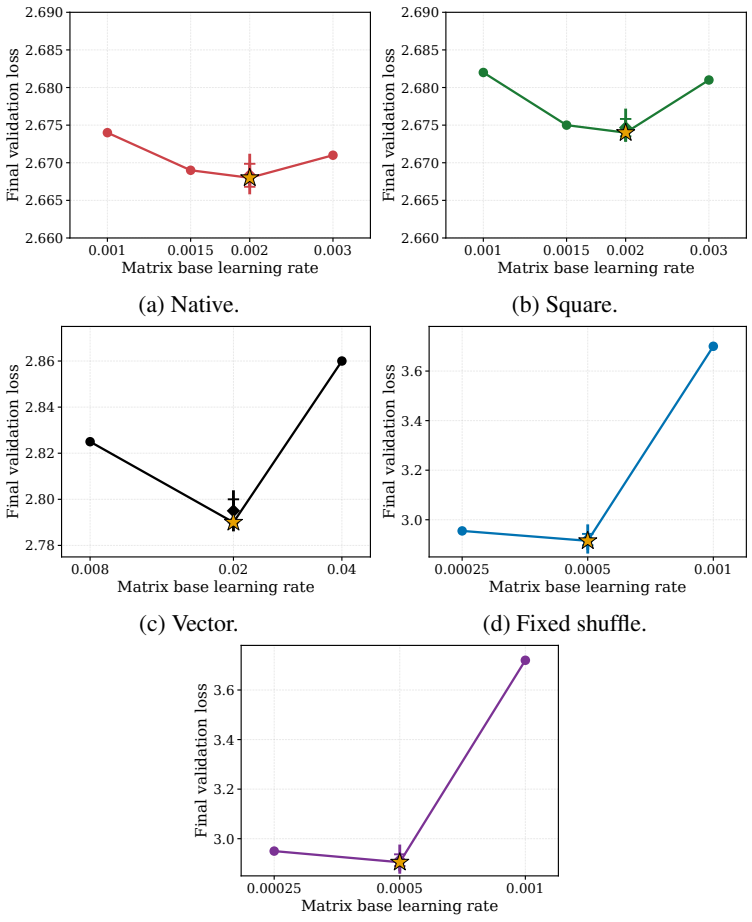
(d) Skinny-50%.

(e) Skinny-25%.

(f) Skinny-12.5%.

(g) Vector.

Figure 5: **Complete fixed-AdamW matrix-rate sweeps at 130M.** All 21 single-seed endpoints after 2.60B tokens, with AdamW base rate 0.008. Lines connect measured candidates; stars mark single-seed minima. At each selected rate, vertical ticks show all three seed losses and a diamond with error bars shows their mean ± sample SD. Panels share loss limits and label every candidate rate.



(a) Native.

(b) Square.

(c) Vector.

(d) Fixed shuffle.

(e) Stepwise shuffle.

Figure 6: **Complete fixed-AdamW matrix-rate sweeps at 600M.** All 17 single-seed endpoints after 12.58B tokens, with AdamW base rate 0.002. Markers follow Figure 5. Native and Square share loss limits, as do the two shuffling panels; Vector uses a separate range. Both shuffled representations retain native dimensions and have higher selected loss than Native and Vector.

I.3 COUPLED LEARNING-RATE PROTOCOL

The trajectories in Figure 1(a,b) use a single base learning rate to control both parameter groups. Following the Muon scaling convention (Liu et al., 2025), the matrix update multiplies this rate by $0.2\sqrt{\max(m,n)}$, while auxiliary AdamW parameters use the base rate directly. Here m, n are the native dimensions of each parameter block in the selected comparisons. The selected base rate jointly determines the matrix and auxiliary AdamW rates. The additional represented-dimension and unrecorded-scaling runs are distinguished in Figure 8.

At 130M, each method selects by single-seed terminal loss over the full grid in Figure 7; the fixed-shuffle control uses Figure 10. Two additional seeds then repeat that rate. Table 7 reports all three seeds. The AdamW and Skinny-12.5% selections are at their upper grid boundaries.

130M method	Coupled base-rate candidates
AdamW	$0.003, 0.005, 0.008$
Native	$0.002, 0.003, 0.005, 0.0075, 0.008, 0.01, 0.02$
Square	$0.002, 0.003, 0.005, 0.0075, 0.008, 0.02$
Each Skinny variant	$0.003, 0.005, 0.0075, 0.008, 0.01, 0.015$
Vector	$0.005, 0.008, 0.01, 0.015, 0.02$
Fixed shuffle	$0.001, 0.002, 0.003$

Method	AdamW base LR	Matrix base LR	Seed 1	Seed 2	Seed 3	Mean ± SD
AdamW	0.008	—	3.346	3.344	3.346	3.3453 ± 0.0012
Native	0.01	0.01	3.242	3.239	3.243	3.2413 ± 0.0021
Square	0.0075	0.0075	3.237	3.234	3.238	3.2363 ± 0.0021
Skinny-75%	0.008	0.008	3.269	3.271	3.266	3.2687 ± 0.0025
Skinny-50%	0.008	0.008	3.289	3.290	3.291	3.2900 ± 0.0010
Skinny-25%	0.01	0.01	3.293	3.294	3.293	3.2933 ± 0.0006
Skinny-12.5%	0.015	0.015	3.307	3.307	3.306	3.3067 ± 0.0006
Vector	0.015	0.015	3.402	3.398	3.405	3.4017 ± 0.0035
Fixed shuffle	0.001	0.001	3.455	3.521	3.497	3.4910 ± 0.0334

Table 7: LLaMA2-130M terminal validation losses under the coupled-rate protocol at 2.60B tokens. Each selected base rate is shared across parameter groups and seeds; matrix updates additionally receive native-dimension scaling. The all-parameter AdamW baseline has no separate matrix rate. SD is the sample standard deviation.

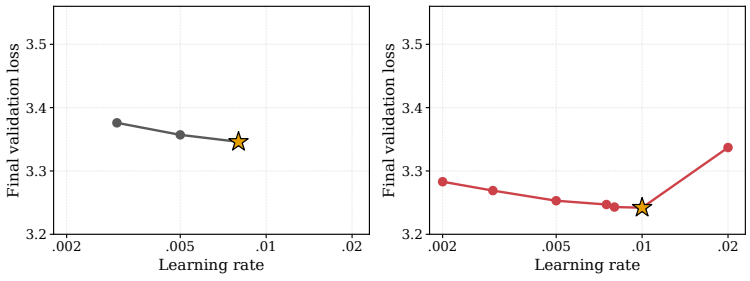
I.4 COUPLED-RATE SELECTION AT 600M

For each ordered representation at 600M, we select the learning rate with the lowest mean final validation loss over the candidate grid evaluated with native-dimension scaling for three seeds. The candidate rates are $\{0.001, 0.002, 0.005, 0.01\}$ for Native, $\{0.002, 0.005, 0.01\}$ for Square, $\{0.001\}$ for each Skinny variant, and $\{0.003, 0.008, 0.02\}$ for Vector. Each Skinny variant uses a single evaluated rate in the three-seed comparison. Native and Square both select 0.002 from their three-seed grids; their single-seed sweeps attain minima at 0.0015. Vector selects its upper grid boundary, 0.02. Figure 8 separately presents broader one- and two-seed sweeps and alternative scaling conventions.

At learning rate 0.002, the paired Square-minus-Native final-loss differences are $0.00542, 0.00693, 0.00608$ across three seeds, with mean 0.00614 and sample SD 0.00076. For Vector, the three-seed final losses at 0.008 and 0.02 are 2.80194 ± 0.00439 and 2.74913 ± 0.00204, respectively, compared with 3.07797 ± 0.00383 at 0.003. Increasing the learning rate reduces the validation-loss difference between Vector and Native; the selected 0.02 rate is used throughout the loss comparisons.

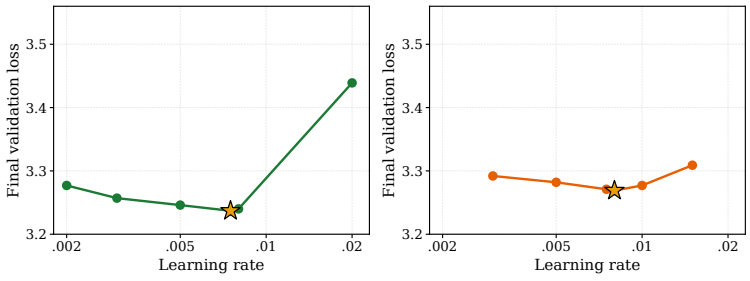
Method	AdamW base LR	Matrix base LR	Seed 1	Seed 2	Seed 3	Mean $\pm$ SD
Native	0.002	0.002	2.66813	2.66712	2.66995	2.66840 ± 0.00144
Square	0.002	0.002	2.67354	2.67405	2.67603	2.67454 ± 0.00132
Skinny-75%	0.001	0.001	2.68492	2.68492	2.68703	2.68562 ± 0.00122
Skinny-50%	0.001	0.001	2.68454	2.68542	2.68712	2.68569 ± 0.00131
Skinny-25%	0.001	0.001	2.70980	2.70946	2.71415	2.71114 ± 0.00262
Vector	0.020	0.020	2.74701	2.74931	2.75107	2.74913 ± 0.00204
Fixed shuffle	0.0005	0.0005	2.85974	2.89282	2.92059	2.89105 ± 0.03046
Stepwise shuffle	0.0005	0.0005	2.85406	2.89966	2.91082	2.88818 ± 0.03007

Table 8: LLaMA2-600M terminal validation losses under the coupled learning-rate protocol at 12.58B tokens. The two base rates are equal within each run; the matrix rate additionally receives the native-dimension multiplier. Ordered representations select by three-seed mean over their eligible grids; shuffling selects using a single seed and repeats that rate with two additional seeds. SD is the sample standard deviation. The reported means include the seeds used for selection.



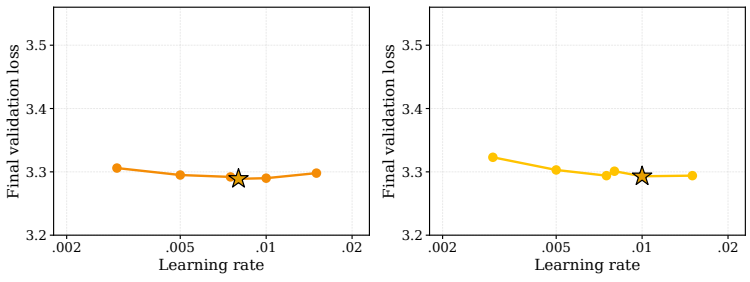
(a) AdamW.

(b) Native.



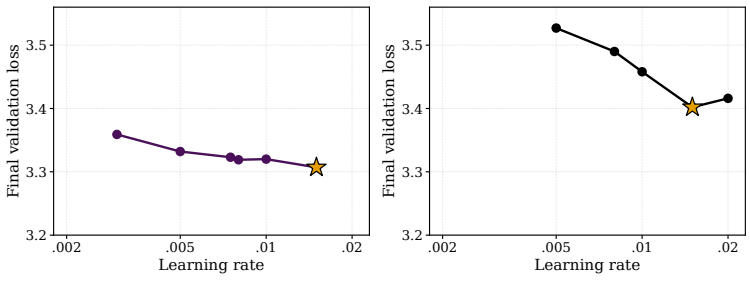
(c) Square.

(d) Skinny-75%.



(e) Skinny-50%.

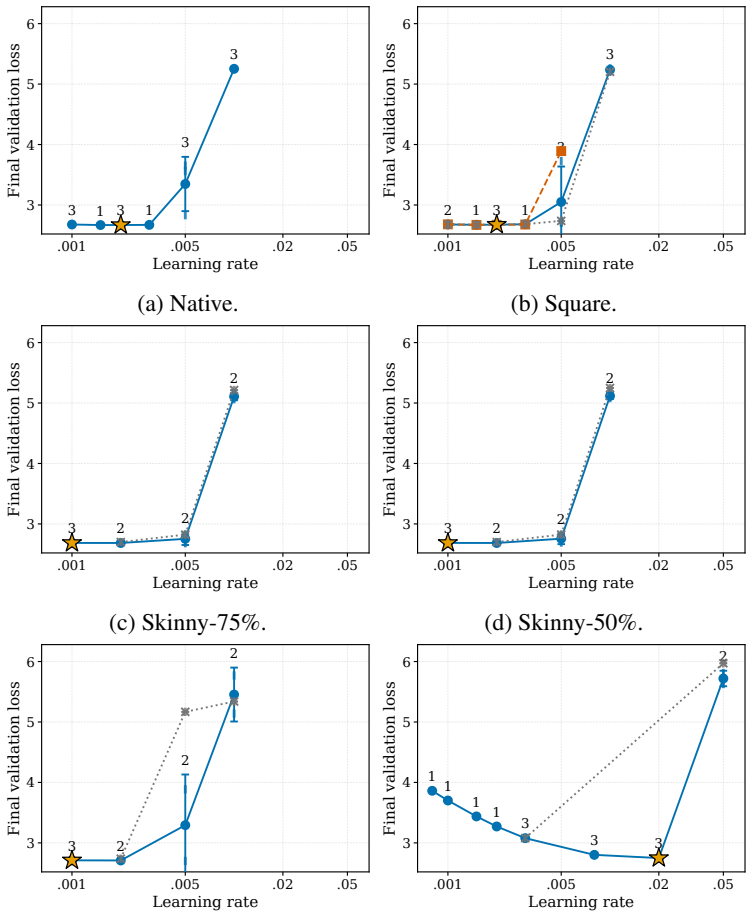
(f) Skinny-25%.



(g) Skinny-12.5%.

(h) Vector.

Figure 7: **Complete LLaMA2-130M coupled-rate sweeps at** 2.60**B tokens.** All 45 completed single-seed runs across the eight unshuffled methods in Table 7 are shown on common axes. Stars mark the selected rates. Figure 10 gives the fixed-shuffle control.



(a) Native. (b) Square.

(c) Skinny-75%. (d) Skinny-50%.

(e) Skinny-25%. (f) Vector.

Figure 8: **Complete LLaMA2-600M learning-rate sweeps at** 12.58**B tokens.** All 91 completed runs are shown across 54 representation–scaling–learning-rate combinations. Lines and markers show means; error bars show one sample standard deviation when at least two seeds are available. Vertical ticks show individual seed losses at their exact learning rates. Blue circles denote native-dimension scaling, orange squares denote represented-dimension scaling, and gray crosses denote unknown scaling. Annotations give the number of seeds for native-dimension scaling; orange and gray results each use a single seed. Gray results have unrecorded scaling and are reported separately from both known conventions. Stars mark selections from the three-seed grids in Table 8; one- and two-seed runs extend sweep coverage. All panels use common axes.

I.5 Fixed-Shape Shuffling in Language-Model Training

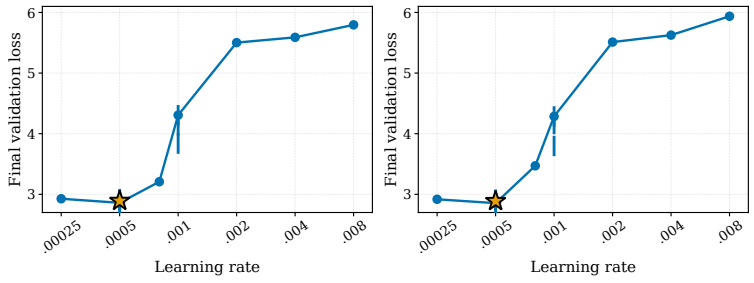
The 600M shuffling controls use the architecture, data, global batch size 320, sequence length 4096, 9600-step budget, momentum, and NS5 specified above. At the selected rate, the three fixed-shuffle runs use $8, 16, 16$ Ascend 910C NPUs, respectively; the three stepwise-shuffle runs use $16, 40, 40$. Gradient accumulation preserves the global batch size. Seeds vary initialization and permutation streams. We evaluate both learning-rate protocols: the fixed-AdamW experiment retains auxiliary rate 0.002, while the coupled experiment varies both base rates together. Both retain native-dimension matrix multipliers.

For each of the 224 matrix blocks, a uniform permutation reorders the entries of the prospective momentum before reshaping to its native dimensions. The inverse permutation returns the computed direction to parameter coordinates. Momentum itself remains in original coordinates. Fixed shuffling retains this permutation throughout training; stepwise shuffling draws a fresh permutation independently for each block and update. A private random generator for each block separates permutation sampling from training randomness. Single-seed learning-rate candidates share the same permutation or permutation sequence. General entry permutations change entry organization, whereas row-only and column-only permutations preserve the NS5 direction in original coordinates in real arithmetic (Appendix B.1).

Under fixed AdamW, each mode selects matrix rate 0.0005 from $\{0.00025, 0.0005, 0.001\}$ using a single seed (Figure 6); Table 6 reports the three-seed endpoints. Under coupled rates, Table 9 reports all seven single-seed candidates. Both modes again select 0.0005, with three-seed results in Table 8. Figure 9 also includes the additional seeds at 0.001.

Learning rate	Fixed shuffle	Stepwise shuffle
0.00025	2.92682	2.91797
0.0005	2.85974	2.85406
0.0008	3.20740	3.47052
0.001	4.30592	4.28497
0.002	5.50252	5.51147
0.004	5.58947	5.62607
0.008	5.79555	5.93756

Table 9: Complete single-seed shuffling sweep under coupled rates: LLaMA2-600M final validation loss after 9600 steps (12.58B tokens). Both groups use the listed base rate; matrix updates also receive native-dimension scaling. Both modes select 0.0005.



(a) Fixed shuffle. (b) Stepwise shuffle.

Figure 9: **LLaMA2-600M shuffling sweeps under coupled rates.** Each panel shows all seven single-seed candidates and two additional seeds at 0.001 and 0.0005. Vertical ticks show individual losses at both replicated rates; stars and error bars show the three-seed mean and sample standard deviation at the selected rate 0.0005. Both panels share axes.

I.6 Fixed Shuffling at 130M

At 130M, fixed shuffling retains each block's native dimensions and uses one fixed entry permutation throughout training. The inverse permutation returns the NS5 update to parameter coordinates, and momentum remains in its original coordinates. The four single-seed candidates use coupled base rates $0.0005, 0.001, 0.002, 0.003$ and attain terminal losses $3.492, 3.455, 3.521, 5.402$, respectively, at 2.60B tokens. Repeating the selected rate 0.001 with two additional seeds gives $3.521, 3.497,$

yielding 3.4910 ± 0.0334 across three seeds (Table 7). The selected rate lies in the interior of the evaluated grid.

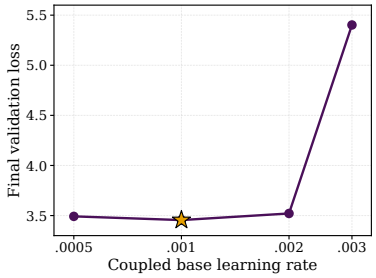


Figure 10: **LLaMA2-130M fixed-shuffle sweep.** All four completed single-seed candidates at 2.60B tokens, using a single base rate for the matrix and auxiliary AdamW groups. The star marks the lowest observed terminal validation loss, at base rate 0.001.

J Single-Step Parameter-Error Progress and Learning-Rate Compensation

Figure 11 uses a fixed-output teacher–student network with hidden width 24, input dimension 48, class dimension 4096, 4096 fresh Gaussian samples, step size 2×10^{-3}, hidden depths $1, 2, 4, 8, 16$, and six paired seeds. The first hidden layer has shape 24×48, subsequent hidden layers are 24×24, and every student layer is initialized by adding an independent Gaussian perturbation of scale $0.05/\sqrt{n_\ell}$ to its teacher counterpart, where n_ℓ is that layer's input dimension. The output Gram matrix is generated from a fixed near-orthogonal teacher output with perturbation scale 0.20. For aggregate single-step parameter-error progress, all hidden layers receive one full-batch polar update, and the layerwise terms are weighted by their initial squared errors. Figure 11 measures the first-layer normalized nuclear support $d^{-1/4}\|T(\mathbf{E}_0)\|_*/\|\mathbf{E}_0\|_F$, using the initial errors in the progress decomposition. All seven representations share the same teacher, student initialization, and gradient batch within each depth–seed pair.

Single-step parameter-error progress and MP prediction at shared parameter states. For layer ℓ, write $S_{T,\ell} = \|T(\mathbf{E}_{0,\ell})\|_*/\|\mathbf{E}_{0,\ell}\|_F^2$, $\chi_{T,\ell}$ for its measured update alignment, and $r_{T,\ell} = \mathrm{rank}(T(\mathbf{G}_\ell))$. With error-energy weights $w_\ell = \|\mathbf{E}_{0,\ell}\|_F^2 / \sum_j \|\mathbf{E}_{0,j}\|_F^2$, the normalized joint single-step parameter-error progress is

$$R_T := \frac{\sum_\ell (\|\mathbf{E}_{0,\ell}\|_F^2 - \|\mathbf{E}_{0,\ell} - \eta \mathbf{D}_T(\mathbf{G}_\ell)\|_F^2)}{\eta \sum_\ell \|\mathbf{E}_{0,\ell}\|_F^2} = 2 \sum_\ell w_\ell \chi_{T,\ell} S_{T,\ell} - \eta \frac{\sum_\ell r_{T,\ell}}{\sum_\ell \|\mathbf{E}_{0,\ell}\|_F^2}. \tag{22}$$

We compare the shape-only and alignment-adjusted MP predictions, respectively:

$$\begin{aligned} &2 \sum_\ell w_\ell \frac{\beta(\gamma_{T,\ell})\sqrt{p_{T,\ell}}}{\|\mathbf{E}_{0,\ell}\|_F} - \eta \frac{\sum_\ell p_{T,\ell}}{\sum_\ell \|\mathbf{E}_{0,\ell}\|_F^2}, \\ &2 \sum_\ell w_\ell \chi_{T,\ell} \frac{\beta(\gamma_{T,\ell})\sqrt{p_{T,\ell}}}{\|\mathbf{E}_{0,\ell}\|_F} - \eta \frac{\sum_\ell r_{T,\ell}}{\sum_\ell \|\mathbf{E}_{0,\ell}\|_F^2}. \end{aligned} \tag{23}$$

For either prediction $\widehat{R}_T$, the relative error is $|\widehat{R}_T - R_T|/|R_T|$. The two median relative errors across the 210 depth–seed–representation states are 157.4% and 0.14%, respectively. The second prediction evaluates the MP approximation using the measured update alignment and root-mean-square parameter error. Using observed ranks in both predictions leaves these median errors unchanged to the reported precision. Substituting measured nuclear support in (22) recovers the measured single-step parameter-error progress to 1.5×10^{-13} absolute error.

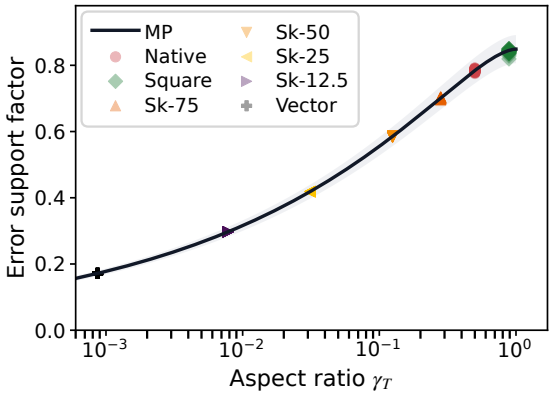


Figure 11: **Calibration of nuclear support for parameter error.** The first-layer normalized nuclear support $d^{-1/4}\|T(\mathbf{E}_0)\|_*/\|\mathbf{E}_0\|_F$ is compared with the MP curve and a $\pm 5\%$ reference band at the same parameter states as the single-step parameter-error progress experiment.

J.1 Learning-Rate Compensation for Single-Step Parameter-Error Progress

The scalar-compensation experiment uses hidden depths $1, 2, 4, 8, 16$, hidden width 24, input dimension 48, and 4096 output coordinates. Teacher hidden weights have orthonormal rows in the first

layer and are orthogonal in subsequent layers. The fixed output matrix follows the near-orthogonal construction above, with perturbation scale 0.20. For each depth and each of six seeds, we construct a zero-loss teacher solution and a perturbation direction. Independent Gaussian batches of 4096 samples each are used to compute gradients and evaluate output loss. The loss is squared output matching, divided by the output dimension; this common scalar leaves both normalized update maps unchanged.

Each hidden weight receives an independent Gaussian perturbation scaled by the inverse square root of its input dimension. The concatenated perturbation is normalized and then rescaled to the prescribed distance from the teacher. All representations share this perturbed state and use $\mathbf{M}_0 = \mathbf{0}$, momentum coefficient 0.95, and one full-batch update. Native, Square, Skinny-50%, and Vector have first-layer shapes 24×48, 32×36, 12×96, and 1×1152, respectively. In subsequent layers these become 24×24, 24×24, 12×48, and 1×576. Single-step parameter-error progress is evaluated both jointly and by layer; Native and Square have distinct shapes only in the first layer.

The zero-loss teacher provides a common reference for comparing optimized single-step parameter-error progress. Equation (3) relates this optimum to normalized directional alignment. The output-loss measurements additionally evaluate how the resulting parameter displacements act on the network objective.

Scalar compensation. For a realized direction $\mathbf{D}$, concatenated over hidden blocks, we evaluate 161 logarithmically spaced values of $\eta/\|\mathbf{E}_0\|_F$ between 10^{-4} and 2, and 161 uniformly spaced relative update lengths $\eta\|\mathbf{D}\|_F/\|\mathbf{E}_0\|_F$ between 0 and 2. We compute the exact error-minimizing step and optimal single-step parameter-error progress from (3). All directions are nonzero, and their optimal step lengths lie in the interior of the evaluated intervals. Optimizing the distance to the teacher along each direction distinguishes directional effects from scalar step-length effects. Separately optimizing each layer's coefficient tests compensation beyond a single global rate (Figure 13).

Perturbation size and activation regions. The main experiment uses $\|\mathbf{E}_0\|_F/\|\mathbf{W}_\star\|_F = 10^{-3}$, with 10^{-2} as an additional perturbation level. At 10^{-3}, the median fraction of ReLU units whose activation status changes ranges from 0.033% to 0.077% across depths. We also restrict the perturbations to a neighborhood in which the activation patterns on both batches are unchanged. A sufficient radius follows by propagating perturbation bounds through the teacher layers using spectral norms and activation margins, with identically zero activations treated separately. Every evaluated update and its segment from the teacher lie within this neighborhood. The resulting relative initial distances range from 1.42×10^{-15} to 8.02×10^{-9}. To avoid cancellation at these distances, network output differences are evaluated recursively in parameter-error coordinates. Reducing these perturbations by a further factor of four changes the optimized remaining-error fraction by at most 8.35×10^{-11}, after rescaling the learning rate accordingly.

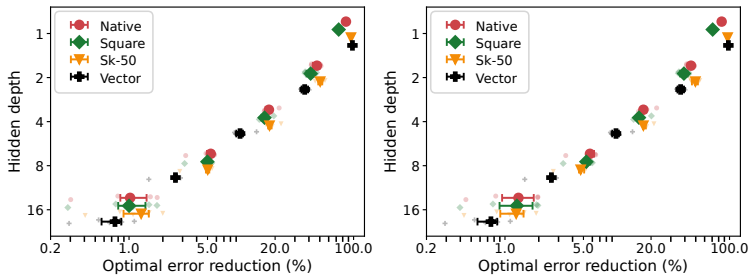
Optimized single-step parameter-error progress and numerical accuracy. Table 10 reports optimized single-step parameter-error progress. All three matrix representations exceed Vector at every paired state of depths $2, 4, 8$ for both update rules. At depth 16, Native and Skinny exceed Vector in all six seeds; Square does so in five seeds for exact polar and all six for NS5. At depth one, Vector exceeds every matrix representation in all seeds. The same median ordering relative to Vector holds within the fixed activation regions and at relative distance 10^{-2}. Independent per-layer step-size optimization also retains the larger median single-step parameter-error progress for the matrix representations at depths $4, 8, 16$. Thus the local compensation response depends on depth as well as representation.

The preceding experiment measures single-step parameter-error progress; we additionally evaluate held-out output loss at the error-minimizing steps. At these steps, the median held-out loss ratio for Vector exceeds one at depths 8 and 16 within the fixed activation regions. Direct error updates agree with the quadratic expansion to 6.83×10^{-13} absolute error. Direct network loss evaluation at relative distance 10^{-2} agrees with residual propagation to 2.31×10^{-15} relative error, and central finite differences agree with the gradients to within 2.01×10^{-4} relative error over all 30 depth–seed pairs. Computations use double precision and the same fixed quintic coefficients as (4) for NS5.

Parameter error and output loss. To compare parameter error with output loss over multiple steps, we perform 100 NS5 updates with momentum 0.95, zero initial momentum, initial relative teacher

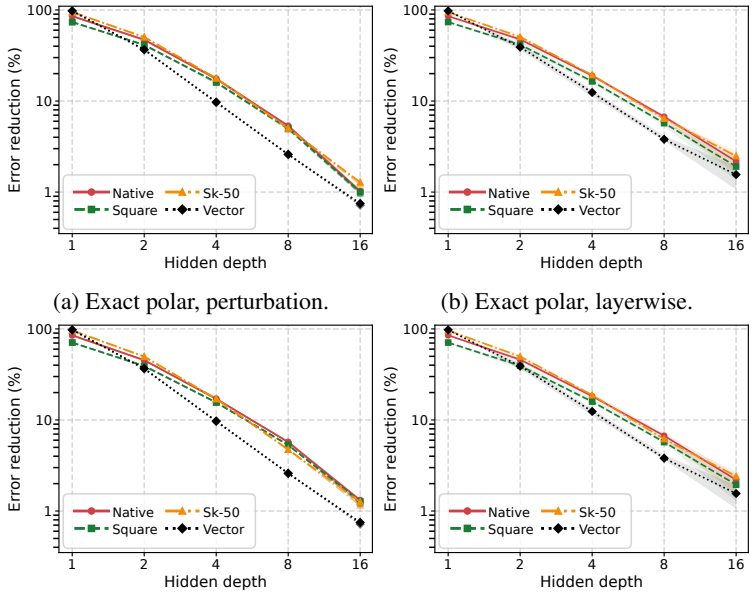
Update and representation	Depth 1	Depth 2	Depth 4	Depth 8	Depth 16
Exact polar: Native	85.53	47.15	17.63	5.34	1.02
Exact polar: Square	73.68	41.44	16.09	4.99	1.00
Exact polar: Skinny-50%	94.77	50.42	17.83	5.02	1.28
Exact polar: Vector	97.68	36.68	9.75	2.59	0.75
NS5: Native	85.10	45.38	17.09	5.71	1.32
NS5: Square	70.78	39.24	15.60	5.30	1.28
NS5: Skinny-50%	96.25	49.67	17.09	4.72	1.26
NS5: Vector	97.68	36.68	9.75	2.59	0.75

Table 10: Median optimized single-step parameter-error progress (%), measured as the reduction in squared parameter error over six paired seeds, at relative initial distance 10^{-3} from the teacher solution. Each update uses its own scalar error-minimizing learning rate.



(a) Exact polar. (b) NS5.

Figure 12: **Optimized single-step parameter-error progress across depths and update rules.** For each seed and representation, the learning rate maximizes the fractional reduction in joint squared parameter error along the realized direction. Large markers and horizontal intervals indicate medians and interquartile ranges, respectively; faint points show individual results for all six seeds. All representations share the perturbed teacher state within each depth–seed pair.



(a) Exact polar, perturbation. (b) Exact polar, layerwise.

(c) NS5, perturbation. (d) NS5, layerwise.

Figure 13: **Optimized single-step parameter-error progress under different perturbation sizes and per-layer step coefficients.** Panels (a,b) use exact polar; (c,d) use NS5. **(a,c)** Faint, solid-opacity, and intermediate-opacity curves use perturbations confined to a fixed activation region, relative distance 10^{-3}, and relative distance 10^{-2}, respectively, with the same perturbation direction. **(b,d)** Each hidden layer receives its own error-minimizing scalar coefficient at relative distance 10^{-3}. Lines show medians over six paired seeds; bands in (b,d) show interquartile ranges.

error 10^{-3}, and relative update length $\eta_t\|\widehat{\mathbf{D}}_t\|_F/\|\mathbf{E}_t\|_F = 0.05$. Six paired states at each depth share fixed gradient and independent evaluation batches of 4096 samples. Table 11 reports all five depths. At depth sixteen, Vector has lower remaining parameter error but higher output loss than Native and Square. The two measurements distinguish contraction toward the teacher from progress on the output objective.

Depth	Representation	Remaining squared error (%)	Remaining loss (%)
1	Native	0.174	0.173
1	Skinny-50%	0.125	0.125
1	Square	0.236	0.236
1	Vector	0.080	0.080
2	Native	5.011	0.531
2	Skinny-50%	3.066	0.420
2	Square	4.800	0.574
2	Vector	2.651	0.394
4	Native	13.884	0.749
4	Skinny-50%	10.868	0.667
4	Square	13.841	0.755
4	Vector	11.536	0.786
8	Native	35.628	1.194
8	Skinny-50%	29.581	2.085
8	Square	36.228	1.245
8	Vector	38.704	11.588
16	Native	79.719	3.407
16	Skinny-50%	71.528	8.891
16	Square	80.443	5.709
16	Vector	77.191	11.952

Table 11: Median remaining fractions after 100 NS5 updates over six paired seeds, normalized by each state's initial error or independent loss.

K Nuclear Support and Update Alignment in Language Models

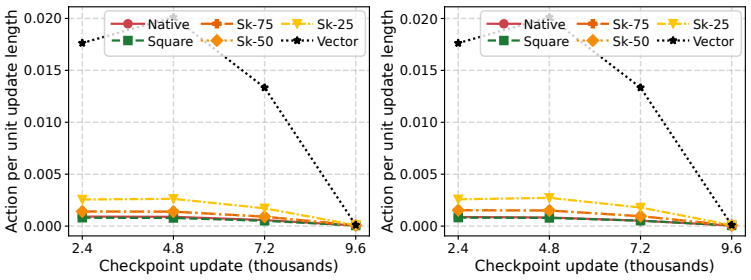
We evaluate the factorization into nuclear support and update alignment at fixed parameter states along a Native 600M training trajectory (single seed, learning rate 0.0015). At iterations $2400, 4800, 7200, 9600$, all representations are applied to the same 224 matrix blocks and momentum inputs. The momentum is updated using 32 training sequences of length 4096, and the reference gradient is estimated from 64 independent held-out sequences, weighted by token count. Newton–Schulz iterations use bfloat16 arithmetic, gradients are accumulated in single precision, and exact polar factors are computed by SVD in double precision. The learning-rate multipliers $\tau_\ell = 0.2\sqrt{\max(m_\ell, n_\ell)}$ depend on the native block dimensions and are identical across representations.

Table 12 reports nuclear support $\sum_\ell \tau_\ell \|T(\mathbf{G}_\ell)\|_*$, action $\sum_\ell \tau_\ell \langle \mathbf{G}_\ell, \widehat{\mathbf{D}}_{T,\ell}\rangle_F$, and aggregate update alignment, defined as first-order action divided by nuclear support. For exact polar, this is a weighted average of per-matrix update alignment, bounded by one in magnitude. For NS5, it also includes the approximate direction's spectral scale and is bounded in magnitude by the NS5 spectral constant in real arithmetic. The four checkpoints provide shared-state comparisons of local directions along one training trajectory.

At the three intermediate states, Native's greater nuclear support compensates for its lower update alignment relative to Vector, while Square increases nuclear support at similar update alignment to Native. Skinny attains larger first-order action at these states, whereas Native and Square attain lower final validation loss in full training. At iteration 9600, Vector's action has opposite signs on the two disjoint held-out subsets. At the nominal learning rate of each intermediate state, a single Native update increases validation loss and a Vector update decreases it. These measurements distinguish nuclear support, realized first-order action, and finite-step objective change. They connect the direction computed in represented coordinates to the parameter displacement applied in training.

Update	Representation	Nuclear support	Update alignment	First-order action
2400	Native	189.11	0.01855	3.5083
2400	Square	228.81	0.01886	4.3155
2400	Skinny-75%	207.52	0.02852	5.9188
2400	Skinny-50%	207.52	0.02852	5.9188
2400	Skinny-25%	176.75	0.04161	7.3541
2400	Vector	18.39	0.13882	2.5532
4800	Native	156.96	0.02152	3.3784
4800	Square	189.89	0.02181	4.1416
4800	Skinny-75%	175.16	0.03351	5.8694
4800	Skinny-50%	175.16	0.03351	5.8694
4800	Skinny-25%	154.21	0.04885	7.5339
4800	Vector	19.47	0.14972	2.9145
7200	Native	146.47	0.01461	2.1392
7200	Square	177.27	0.01547	2.7425
7200	Skinny-75%	164.39	0.02303	3.7860
7200	Skinny-50%	164.39	0.02303	3.7860
7200	Skinny-25%	146.18	0.03383	4.9446
7200	Vector	19.55	0.09881	1.9315
9600	Native	134.33	0.00182	0.2446
9600	Square	162.99	0.00210	0.3430
9600	Skinny-75%	150.44	0.00178	0.2680
9600	Skinny-50%	150.44	0.00178	0.2680
9600	Skinny-25%	132.73	0.00202	0.2679
9600	Vector	16.98	0.00088	0.0149

Table 12: Nuclear support, update alignment, and first-order action at four fixed language-model states. Each row aggregates all 224 matrix blocks using common objective gradients, momentum inputs, NS5 approximations, and native-dimension learning-rate multipliers. Update alignment is the ratio of first-order action to nuclear support; displayed values are rounded.



(a) Exact polar. (b) NS5.

Figure 14: First-order action per unit joint update length. The update retains the native-dimension learning-rate multipliers before normalization by its joint Frobenius norm. Results include all four parameter states.

L Optimization under Row-Block Curvature

We jointly optimize four independent matrices with the objective $f = \frac{1}{2}\sum_{\ell,i} \lambda_i \mathbf{e}_{\ell i}^\top \mathbf{C}_\ell \mathbf{e}_{\ell i}$, where $\mathbf{e}_{\ell i} \in \mathbb{R}^{32}$ is the transposed parameter-error row and $\mathbf{C}_\ell$ is a fixed dense positive-definite matrix with geometrically spaced eigenvalues in $[1/4, 1]$. We set $\lambda_1 = 20$ and $\lambda_i = 20/\rho$ for $i > 1$. Thus the full Hessian is exactly block diagonal with blocks $\lambda_i \mathbf{C}_\ell$, and its row-block spectral bounds equal λ_i. This construction varies the maximum-to-average ratios h_1, h_2 in (6) through $h_r = 16/(1+15\rho^{-r})$, while keeping the maximum row-block curvature fixed. Each objective is shared by all methods, so the entry-organization comparison is made within a curvature condition.

Three Gaussian initial errors are normalized to unit joint Frobenius norm and shared across methods and curvature conditions. Gradients are evaluated analytically without sampling noise. Native directions are divided by $\sqrt{16}$; nSGDM normalizes each momentum matrix by its Frobenius norm. Momentum starts at zero and uses coefficient $1 - t^{-1/2}$; the step size is $\eta t^{-3/4}$. We search 17 geometrically spaced values of η between 0.001 and 3. All selected values are interior to the grid. The original sweeps use the same three initializations for learning-rate selection and reporting; Fig-

ure 17(c) evaluates eight additional paired initializations without retuning. The stationarity criterion averages across seeds before minimizing over states $1, \ldots, 1000$; final loss is evaluated after 1000 updates.

Fixed shuffling uses one permutation per matrix, shared across curvature conditions, seeds, and polar approximations. Momentum remains in original coordinates. The permutation is applied before the NS5 or exact-polar map, followed by its inverse. Per-matrix nSGDM is invariant to this operation. The fraction of squared Hessian Frobenius norm outside represented diagonal row blocks is zero for Native and approximately 0.126 after shuffling, averaged over the four matrices. We evaluate the row-block coefficients for the unshuffled Native representation using (21); all four matrices have the same row bounds $L_i = \lambda_i$. Shuffling changes the correspondence between spectral row constraints and curvature blocks. The exact-polar comparison measures this intervention with an exact spectral map. Row and column permutations preserve both update maps, as shown in Appendix B.1.

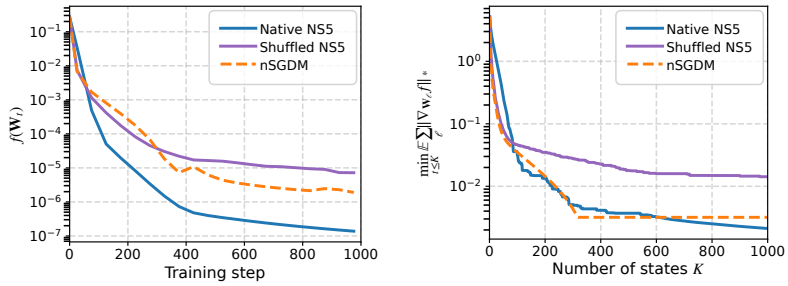
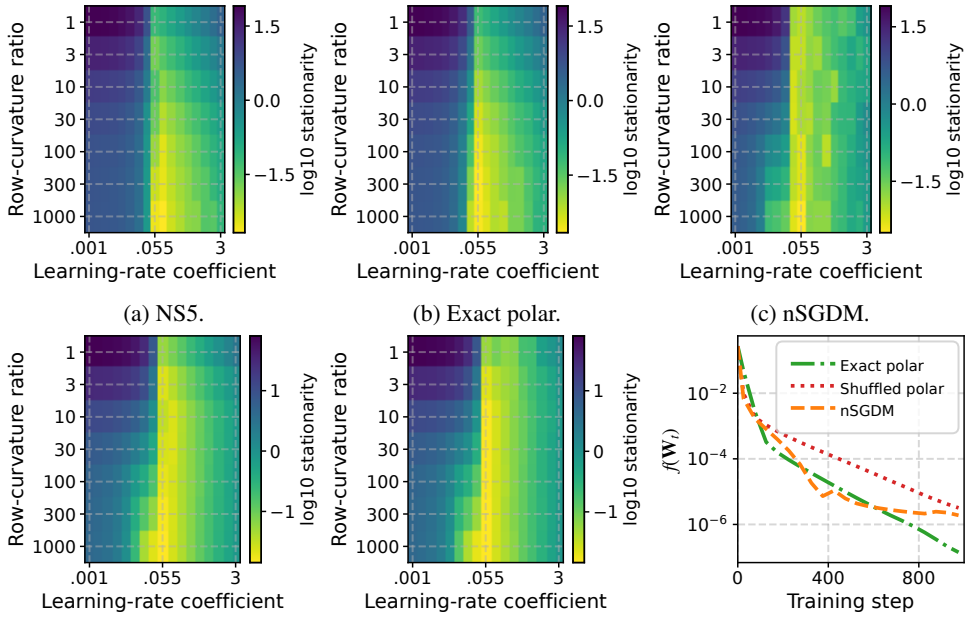


Figure 15: **Fixed-shape shuffling removes NS5-Native's advantage.** At $\rho = 1000$, compare NS5-Native, shuffled NS5, and nSGDM. Left: 50-step-averaged loss; right: unsmoothed stationarity. Rates are selected separately by final mean loss and stationarity.



(a) NS5. (b) Exact polar. (c) nSGDM.

(d) Shuffled NS5. (e) Shuffled exact polar. (f) Exact-polar control, $\rho = 1000$.

Figure 16: **Learning-rate sensitivity and exact-polar shuffling control.** (a–e) The complete stationarity grids: colors show base-ten logarithms of the criterion and rows cover all seven curvature conditions. (f) Original and shuffled exact-polar updates compared with nSGDM at $\rho = 1000$, after independent learning-rate selection by final mean loss. Curves use nonoverlapping 50-step means. This control separates representation effects from the NS5 approximation.

Native row-block curvature with coupling. The second row of Figure 4 retains the four matrices, initial errors, permutations, schedules, and learning-rate grid. Group sizes are $1, 5, 5, 5$, with bounds

$L_1 = 20$, $L_j = 20/\rho$ for $j > 1$, preserving $h_r = 16/(1+15\rho^{-r})$. For each matrix ℓ, set $\mathbf{B}_{\ell 1} = \mathbf{C}_\ell$; for $j > 1$, normalize $3\mathbf{I}_5 \otimes \mathbf{C}_\ell + \mathbf{A}_j \otimes \mathbf{D}_{\ell j}$ to spectral norm one to obtain $\mathbf{B}_{\ell j}$. The fixed positive-definite matrices $\mathbf{A}_j \in \mathbb{R}^{5\times 5}$, $\mathbf{D}_{\ell j}, \mathbf{F}_\ell \in \mathbb{R}^{32\times 32}$ have independent random orthogonal eigenbases and geometrically spaced eigenvalues from $1/4$ to 1. Let $\mathbf{S}$ be diagonal with entry $\sqrt{L_j}$ on rows in I_j, and $\mathbf{K}$ be symmetric with zero diagonal group blocks and $\|\mathbf{K}\|_2 = 1$. Define

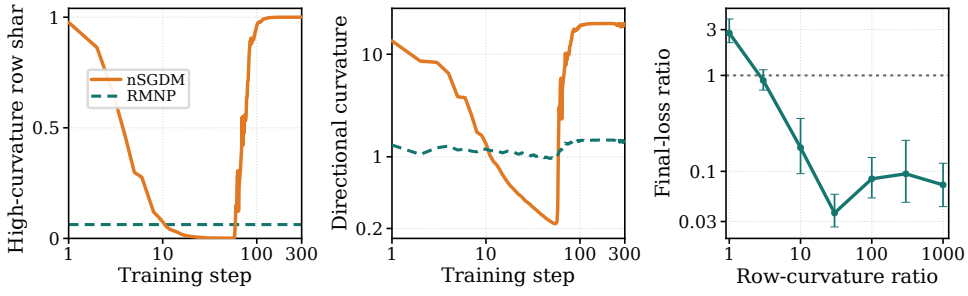
$$\kappa = 0.2 \min_{\ell,j} \lambda_{\min}(\mathbf{B}_{\ell j}), \qquad \mathbf{H}_\ell = \mathrm{blkdiag}_j(L_j \mathbf{B}_{\ell j}) + \kappa(\mathbf{SKS}) \otimes \mathbf{F}_\ell, \qquad f = \frac{1}{2}\sum_\ell \mathbf{e}_\ell^\top \mathbf{H}_\ell \mathbf{e}_\ell,$$

where $\mathbf{e}_\ell$ stacks rows of the parameter-error matrix. Congruence by $\mathbf{S}^{-1} \otimes \mathbf{I}$ gives diagonal blocks with eigenvalues at least 5κ and a perturbation of norm at most κ; hence $\mathbf{H}_\ell \succ 0$. Its principal blocks have norms L_j, and $\varepsilon = \kappa\|\mathbf{SKS}\|_2$. Distinct right factors give noncommuting column-contracted row operators. The full grid contains 1785 trajectories. Coefficient ratios use $\Gamma_i = 4C_i$ from Appendix E.5.

Rank-one identity and RMNP baseline. For an implementation check at $\rho = 1000$, we retain only the first native row of each original initial-error matrix and renormalize the joint Frobenius norm to one. Row-block diagonality keeps subsequent gradients and momenta on that row. Each nonzero momentum matrix is then rank one, so its exact polar direction is its Frobenius-normalized direction, and NS5 gives $\psi_5(1) = 0.696436409\ldots$ times that direction. With an nSGDM step coefficient $\eta = 0.0547723$, exact polar uses $\sqrt{16}\,\eta$, and NS5 uses $\sqrt{16}\,\eta/\psi_5(1)$. The three paired 200-step parameter trajectories agree to a maximum entrywise difference of 3.23×10^{-14}, verifying the spectral-map implementations and step scaling in this rank-one case. Each 16×32 array evolves as a single 32-dimensional vector.

For the original isotropic initializations, RMNP uses row-normalized momentum (Deng et al., 2026), rescaled by $1/\sqrt{16}$: $\mathbf{U}_{\ell,i,:} = \mathbf{M}_{\ell,i,:}/(\sqrt{16}\|\mathbf{M}_{\ell,i,:}\|_2)$ for nonzero rows and zero otherwise. It uses the same momentum and step schedules, 17 candidate rates, three initializations, and separate selection criteria as the other methods. At $\rho = 1000$, both criteria select rate 0.0547723, yielding final mean loss 5.05×10^{-8} and minimum seed-mean sum of native gradient nuclear norms 1.31×10^{-3}. Across the full seven-condition sweep, nSGDM has lower selected final loss at $\rho = 1$, whereas RMNP has lower selected final loss than NS5-Native, exact-polar Native, and nSGDM for every $\rho \geq 10$. This direction caps row energy through independent row normalization.

For Figure 17(c), we evaluate nSGDM and RMNP at their original loss-selected rates on eight further paired Gaussian initializations. The plotted ratios compare seed-mean final losses; the 95% percentile intervals use 20,000 paired bootstrap resamples. Panels (a,b) evaluate both directions from the same momentum at each of the first 300 pre-update states of the loss-selected nSGDM trajectory at $\rho = 30$. Directional curvature is $\langle \mathbf{D}, \mathbf{HD}\rangle/\|\mathbf{D}\|_F^2$. The row-normalized direction assigns $1/16$ of each matrix's squared update length to every nonzero row; the nSGDM shares vary over the trajectory. At $\rho = 300$, RMNP lowers mean final loss and improves six of eight paired starts.



(a) Row update energy. (b) Directional curvature. (c) Final loss from fresh starts.

Figure 17: **Equal row update energy can lower directional curvature and final loss.** (a,b) Three-seed means along nSGDM trajectories at $\rho = 30$: RMNP assigns $1/16$ energy to each nonzero row at unit Frobenius update length; directional curvature is normalized by squared update length. (c) Mean final-loss ratios (RMNP/nSGDM) over eight fresh starts, retaining rates selected on the original three. Bars: 95% paired-bootstrap intervals; dotted line: equality.

For a nonzero gradient matrix, the effective rank $(\|\mathbf{G}\|_*/\|\mathbf{G}\|_F)^2 \in [1, 16]$ gives the squared nuclear-to-Frobenius norm ratio. At shared states 1 and 100 of the loss-selected NS5-Native path, the means across four matrices and three seeds are 1.031 and 1.271, respectively. Figure 18 shows the rank-one identity check, the RMNP baseline with isotropic starts, and direction terms at the shared first state. Let $\mathbf{H}$ denote the full constant Hessian, and let $\mathbf{G}$ and $\mathbf{D}$ collect the four gradient and direction blocks. Table 13 reports the direction terms at the first state. For this quadratic objective, the exact one-step change is $f(\mathbf{W} - \eta\mathbf{D}) - f(\mathbf{W}) = -\eta A + \eta^2 Q/2$.

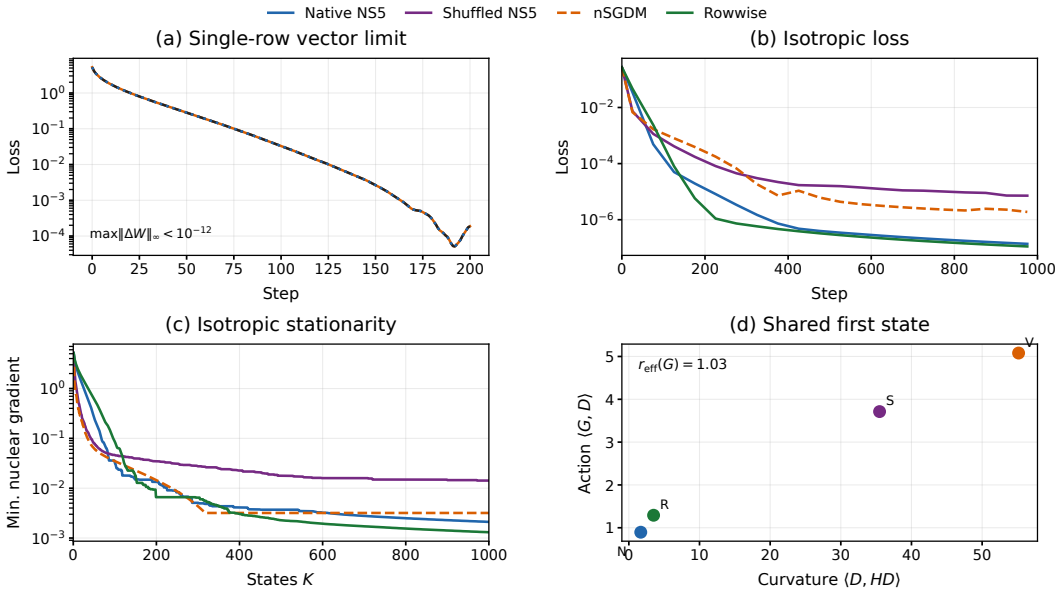


Figure 18: **Vector-limit check and RMNP diagnostics at $\rho = 1000$.** (a) With only one active row per matrix, the mapped exact-polar, NS5, and nSGDM trajectories coincide. (b,c) Selected-rate loss and stationarity for the original three paired isotropic initializations, adding RMNP (Rowwise in the legend) to NS5-Native, shuffled NS5, and nSGDM. (d) Mean first-order action and directional curvature at the shared first state; effective rank averages across matrices and seeds.

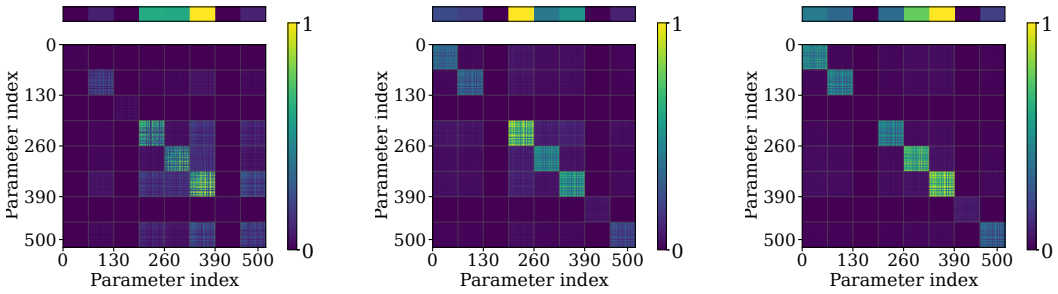
Direction	$\|\mathbf{D}\|_F^2$	A	Q	$\|\mathbf{HD}\|_F^2$
NS5-Native	1.143	0.896	1.686	25.233
Exact-polar Native	4.000	1.289	3.493	52.020
RMNP	4.000	1.292	3.500	52.044
nSGDM	4.000	5.080	55.160	832.672
Shuffled NS5	3.014	3.713	35.485	532.783

Table 13: Actual terms at the shared first state for $\rho = 1000$, averaged over three paired initializations. The joint quantities are $A = \langle \mathbf{G}, \mathbf{D} \rangle_F$, $Q = \langle \mathbf{D}, \mathbf{HD} \rangle_F$, and $\|\mathbf{HD}\|_F^2$. Directions use the existing Native scaling $1/\sqrt{16}$ and each method's normalization; no learning-rate coefficient is applied. The mean per-matrix gradient effective rank is 1.031.

M Measured Hessian Block Structure

The MLP replication follows the synthetic binary-classification setup of Zhang et al. (2024, Appendix D.3): 100 examples, 64 input features, eight tanh hidden neurons, and 1000 Adam steps at learning rate 10^{-4}. We calculate the exact 520×520 principal Hessian of hidden weights and biases, ordered into eight blocks of 65 parameters: one row of 64 input weights and its bias. The off-block to within-block Frobenius norm ratio declines from 0.153 at step 10 to 0.031 at step 500 and 0.008 at step 1000.

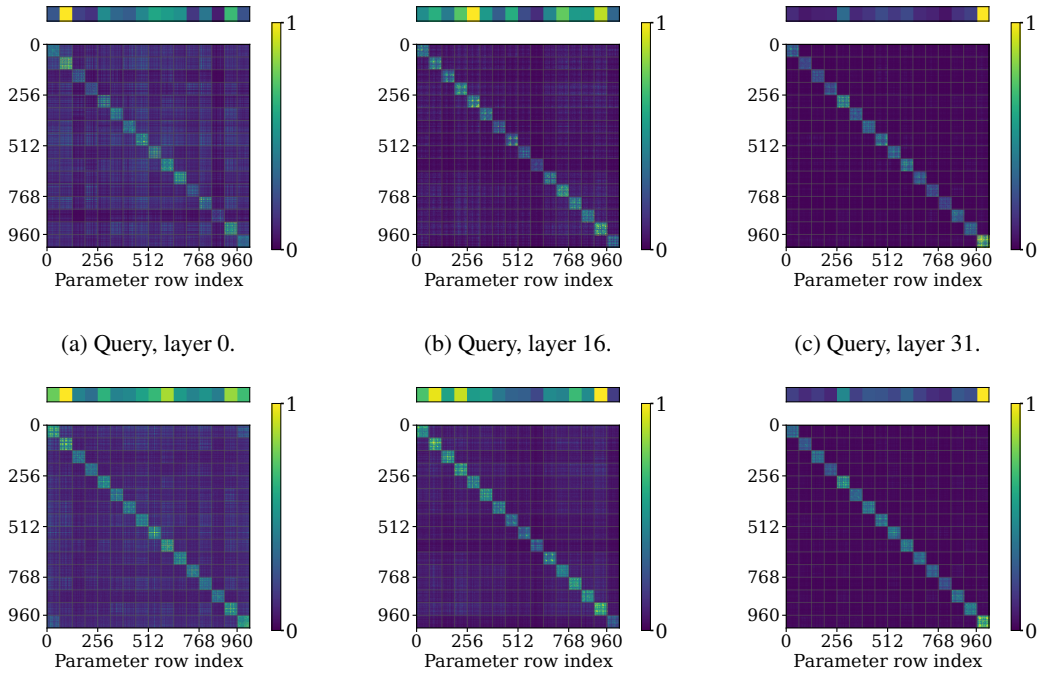
For the LLaMA-600M checkpoint at step 9600, each 1024×1024 query or key weight matrix contains 16 attention heads of 64 output rows. Each head thus defines a group of native rows of the type considered in Proposition 3.4. On a fixed validation sequence of length 128, we apply



(a) Step 10. (b) Step 500. (c) Step 1000.

Figure 19: **Near-block-diagonal structure throughout MLP training.** Each panel shows absolute entries of the exact hidden-layer Hessian, with parameters ordered into eight neuron blocks. The strip shows their spectral norms normalized by the largest block in that panel. The square-root color scale places the within-block 99th percentile at its maximum.

two isotropic unit probes to each source row and record the norm of each target-row response to the exact FP32 Hessian-vector product. We average the squared responses over probes and across the two row directions before taking square roots. Taking the square root of the sum of squared row responses over each 64×64 head pair gives the head-block strengths shown above the maps in Figure 20. A separate head-supported probe measurement on a fixed training sequence gives last-layer across-head to within-head response ratios 0.022 for query and 0.018 for key.



(a) Query, layer 0. (b) Query, layer 16. (c) Query, layer 31.

(d) Key, layer 0. (e) Key, layer 16. (f) Key, layer 31.

Figure 20: **Attention-head Hessian response across LLaMA layers.** Columns show layers 0, 16, and 31; rows show query and key projections. Each panel resolves 64 parameter rows per head, and the strip shows diagonal head strengths normalized by their panel maximum. Across-head to within-head response ratios are 0.388, 0.128, 0.013 for query and 0.308, 0.111, 0.014 for key. The square-root color scale places the within-head 99th percentile at its maximum.

N AI USE STATEMENT

We used AI tools for language polishing and improving the clarity of the manuscript. The authors take full responsibility for its content.