THE SPECTRAL DYNAMICS AND NOISE GEOMETRY OF MUON

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon replaces a matrix gradient $G = U\Sigma V^\top$ by its polar factor $UV^\top$. This keeps the singular directions selected by the gradient, but makes the update spectrum flat. We study the optimization bias created by this operation. In an underdetermined regression model, we derive exact singular-value dynamics for a projected polar flow and identify a measurement-dependent condition under which the normalized spectrum moves toward equal nonzero singular values. At fixed Frobenius norm, the pairwise spectral functional is minimized by a flat spectrum, whereas nuclear-norm minimization favors spectral concentration. Controlled matrix-sensing experiments separate the effect from simple gradient rescaling, show that norm-matched gradient descent does not reproduce Muon, and recover the predicted flattening trend across broad ablations. In small NanoGPT pretraining, Muon preserves stable rank, has a broad learning-rate plateau, and improves validation loss relative to AdamW; in a matched small-ViT control, the ranking reverses. The resulting picture is regime-dependent: Muon is not universally superior, but its flat-spectrum bias can help when many spectral directions need to remain active.

1 INTRODUCTION

A matrix gradient carries two kinds of information: which singular directions are active, and how strongly each direction appears. Gradient descent uses both. Muon (Jordan et al., 2024) keeps the directions and discards the relative strengths. If

$$G = U\Sigma V^\top,$$

then Muon updates with the polar factor

$$\mathrm{polar}(G) = UV^\top.$$

Thus Muon is not merely rescaling the gradient. It changes the update spectrum: a gradient-descent step inherits the singular values of G, while a Muon step gives every active singular direction the same amplitude.

This paper studies the bias induced by that flattening operation. Muon has become a practical optimizer for transformer training (Liu et al., 2025), but empirical acceleration alone does not explain what the polar map selects. Two natural explanations are tempting. One is that Muon is just matrix-gradient normalization; another is that spectral-norm geometry should lead to nuclear-norm, hence low-rank, implicit bias. We show that both are incomplete. Muon is better understood as a *flat-update-spectrum* method: it keeps the gradient singular directions while allocating equal update amplitude across them.

The regression setting makes this question sharp. In separable classification, recent theory for Muon-like methods studies max-margin directions at infinity. In underdetermined regression, the zero-loss set is instead an affine interpolation manifold, and the optimizer must choose one finite interpolant among many. We ask which solution spectra are favored by the polar update.

Our main theoretical model is the projected polar flow on

$$\mathcal{M} = \{W : XW = Y\}.$$

Under a shared-frame alignment assumption, the singular values obey

$$\dot\sigma_i = -\eta\,\alpha_i, \qquad \alpha_i \in [0,1],$$

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.

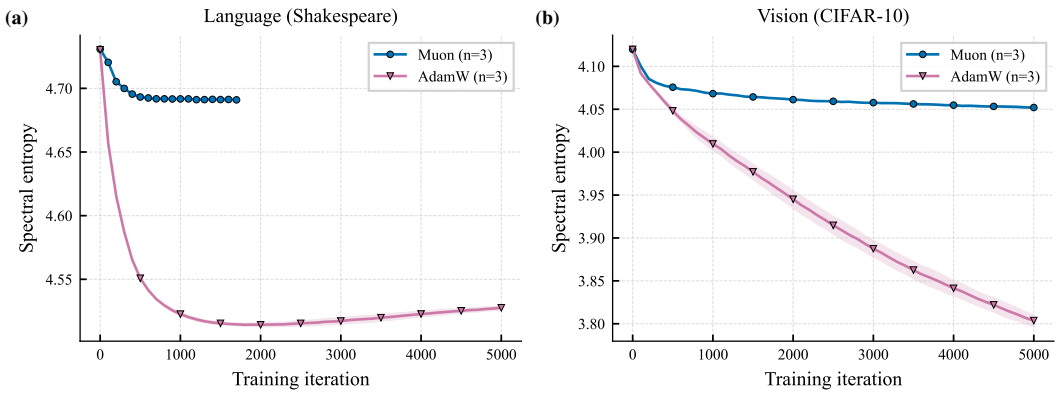

Figure 1: **Spectral entropy in language and vision.** Spectral entropy during (a) small-NanoGPT training on Shakespeare and (b) small-ViT training on CIFAR-10, with three runs per optimizer. Muon retains higher spectral entropy than AdamW in both settings.

with α_i determined by the measurement geometry. This yields an exact sign criterion for whether the normalized spectrum flattens or concentrates. The distinction matters: the pairwise potential R_{pw} identifies the flat spectrum as the unique fixed-radius variational state, but it is not a Lyapunov function for the unnormalized flow. Muon contracts the norm; whether its spectral shape flattens is governed by the sign criterion.

Contributions. We make four contributions.

- **A one-step spectral-bias theorem.** Under explicit alignment assumptions, we prove that the polar profile maximizes first-order spectral-entropy gain among bounded updates that follow the gradient singular directions and ignore the current weight spectrum.
- **Trajectory-level singular-value dynamics.** We derive exact singular-value dynamics for the projected polar flow, including a measurement-geometry sign criterion for spectral flattening and a fixed-radius characterization of the flat spectrum.
- **Separation from normalization and nuclear-norm explanations.** In controlled matrix-sensing experiments, norm-matched gradient descent does not reproduce Muon's behavior, and Muon's limit points do not match convex nuclear-norm minimizers.
- **Transformer signatures and regime dependence.** In small NanoGPT pretraining, Muon preserves stable rank; in an architecture- and budget-matched small-ViT control, the optimizer ranking reverses. These experiments suggest that the flat-spectrum bias is useful in some spectral regimes but not universally.

Takeaway. The resulting practitioner-relevant message is regime-dependent, not universal. Flat update spectra can help when many spectral directions must remain active; they can be unnecessary or worse when the useful spectrum is already low-dimensional or well conditioned.

In small NanoGPT pretraining, Muon preserves stable rank and improves validation loss relative to AdamW. In a matched small-ViT control, the ranking reverses. This supports a regime story rather than a universal optimizer claim.

2 Related Work

Muon orthogonalizes matrix gradients using Newton–Schulz iteration (Jordan et al., 2024; Bernstein & Newhouse, 2024; Tuddenham et al., 2022). Existing analyses study spectral-norm steepest descent and trust regions (Li & Hong, 2025; Kovalev, 2025), nonconvex convergence (Shen et al., 2025), and max-margin classification (Fan et al., 2025; Gronich & Vardi, 2026; Kang et al., 2026). Our setting instead concerns finite interpolants in underdetermined regression and the measurement geometry governing their spectral dynamics.

Recent spectral-flattening theory connects Muon's stable learning rates and convergence to Kronecker-factored curvature (Nguyen et al., 2026); layerwise analyses compare spectral and Euclidean descent through gradient and activation geometry (Davis & Drusvyatskiy, 2025). Variants and negative results emphasize that the usefulness of spectral updates depends on the task and training recipe (Shumaylov et al., 2026; Fan et al., 2026; Southworth et al., 2026; Yan et al., 2026). Our experiments likewise provide regime-dependent spectral signatures, rather than a universal optimizer claim.

For critical batch size, prior signal-to-noise analyses (McCandlish et al., 2018; Smith et al., 2018) and Muon convergence-rate bounds (Sato et al., 2025) complement our one-step polar-map sensitivity calculation. Appendix A preserves the full literature discussion; Appendix C compares the scope of concurrent work.

3 SETUP

Matrix regression and the polar map. Let $W \in \mathbb{R}^{m\times n}$ be a weight matrix and $X \in \mathbb{R}^{p\times m}$ a measurement matrix with $\operatorname{rank}(X) = p < m$ (so the system is underdetermined). The matrix-regression loss is $L(W) = \frac{1}{2}\|XW - Y\|_F^2$. Write $G = \nabla_W L(W) = X^\top(XW - Y)$ with SVD $G = U_G\Sigma_G V_G^\top$; the polar map is $P(G) := U_G V_G^\top$. The continuous-time Muon flow with step size $\eta > 0$ is

$$\dot{W} \;=\; -\eta\, P(G). \tag{1}$$

The interpolation manifold. Let $\mathcal{M} = \{W \in \mathbb{R}^{m\times n} : XW = Y\}$ denote the affine manifold of zero-loss interpolants. Its tangent space is constant: $T_W\mathcal{M} = \ker X$ for all $W \in \mathcal{M}$. The orthogonal projector onto $T_W\mathcal{M}$ in the Frobenius inner product is

$$P_\perp \;:=\; I_m - X^\dagger X, \qquad X^\dagger := X^\top(XX^\top)^{-1}.$$

$P_\perp$ is symmetric, idempotent, with $\|P_\perp\|_{\mathrm{op}} = 1$.

Shared-frame condition. We invoke the following structural condition.

Definition 1 (Shared Frames, abbreviated SF)**.** *The gradient and weight share singular frames: $U_G = U_W$ and $V_G = V_W$, so $G = U_W\Sigma_G V_W^\top$. Under SF, $P(G) = U_W V_W^\top = P(W)$, and the polar update acts diagonally on the singular values.*

SF is a local structural condition, not a global assumption. Theorem 3 relaxes exact alignment using $\delta_P = \|P(G) - P(W)\|_F$; singular-subspace alignment alone does not control this quantity.

Projected polar flow vs. literal Muon at zero loss. On the interpolation manifold $\mathcal{M}$, the gradient vanishes ($G = 0$), so the literal continuous flow (1) stops; in practice, however, the discrete Muon iterate carries a momentum buffer m_t with $P(m_t) \neq 0$ and remains spectrally active. The object analyzed in Theorem 1 is therefore not literal unconstrained Muon at zero loss, but the projected polar flow obtained by Euler-projecting the polar step back onto $\mathcal{M}$.

Proposition 1 (Projected polar flow)**.** *Fix $W \in \mathcal{M}$. The orthogonal projection of one Euler step onto $\mathcal{M}$ is*

$$W^+ \;=\; \Pi_{\mathcal{M}}\big(W - \eta\, P(W)\big) \;=\; W - \eta\, P_\perp\, P(W), \tag{2}$$

where $\Pi_{\mathcal{M}}$ denotes Euclidean projection onto $\mathcal{M}$. The continuous-time flow is

$$\dot{W} \;=\; -\eta\, P_\perp\, P(W). \tag{3}$$

Proof in Appendix B.

We call (3) the projected (self-)polar flow; the projected gradient-driven flow instead uses $P(G)$. Table 1 summarizes the distinction.

The pairwise spectral functional. Let $W = U_W\Sigma_W V_W^\top$ be the SVD of W with ordered singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $r = \operatorname{rank}(W)$. We use two functionals:

$$R_{\mathrm{pw}}(W) \;:=\; -\sum_{i\neq j}\log(\sigma_i + \sigma_j), \qquad \widetilde{R}_{\mathrm{pw}}(W) \;:=\; -\sum_{i\neq j}\log(\widetilde{\sigma}_i + \widetilde{\sigma}_j), \tag{4}$$

where $\widetilde{\sigma}_i := \sigma_i/\rho$ and $\rho := \|W\|_F$. The functional $\widetilde{R}_{\mathrm{pw}}$ is scale-invariant and tracks spectral *shape*; flat spectra ($\widetilde{\sigma}_i = 1/\sqrt{r}$) minimize $\widetilde{R}_{\mathrm{pw}}$ subject to $\sum_i \widetilde{\sigma}_i^2 = 1$.

Polar-map spectral sensitivity.

Definition 2 (Polar-Map Spectral Sensitivity). *$S(\mu) := \sum_{i \neq j} (\sigma_i + \sigma_j)^{-2}$, evaluated at the gradient singular values $\{\sigma_i\}$. $S(\mu)$ measures how strongly the polar map amplifies noise: small or near-zero singular values inflate $S(\mu)$ via the $(\sigma_i + \sigma_j)^{-2}$ terms.*

For r equal singular values σ_0, $S(\mu) = r(r-1)/(4\sigma_0^2)$; Appendix E verifies this against $\mathrm{tr}(D_P D_P^\top)/2$.

Experimental regimes. We study matrix sensing, controlled frame misalignment, coupled and decoupled weight decay, and transformer spectral profiles. Full regimes and hyperparameters appear in Appendices B.3 and F.

4 Theoretical Results

Theorem 1 characterizes projected self-polar dynamics without SF. Theorems 3 and 4 provide local alignment and discretization bounds; Theorem 2 studies polar-map noise sensitivity and its one-step batch-size crossover.

4.1 Spectral Dynamics on the Interpolation Manifold

[Flow: projected self-polar flow $\dot{W} = -\eta\, P_\perp\, P(W)$. No Shared-Frame assumption is used in this theorem.]

Theorem 1 (Spectral Dynamics of the Projected Polar Flow). *Let $\mathcal{M} = \{W \in \mathbb{R}^{m \times n} : XW = Y\}$, $P_\perp = I_m - X^\dagger X$, and consider the projected self-polar flow $\dot{W} = -\eta\, P_\perp\, P(W)$ from Proposition 1, where by definition $P(W) = U_W V_W^\top$. Suppose on an interval I: (a) $W(t) \in \mathcal{M}$, (b) $\mathrm{rank}(W(t)) = r \geq 3$ is constant, and (c) the singular values are simple, $\sigma_1(W(t)) > \sigma_2(W(t)) > \cdots > \sigma_r(W(t)) > 0$, so they are smooth functions of W on I. Then:*

(i) ***Singular-value dynamics.*** *$\dot{\sigma}_i = -\eta\, \alpha_i$, where $\alpha_i := [U_W^\top P_\perp U_W]_{ii} = \|P_\perp u_i\|^2 \in [0, 1]$.*

(ii) ***Frobenius contraction.*** $\dfrac{d}{dt}\|W\|_F^2 = -2\eta \sum_{i=1}^{r} \alpha_i\, \sigma_i \leq 0.$

(iii) ***Spectral flattening criterion.*** *Let $\widetilde{\sigma}_i := \sigma_i/\rho$ with $\rho := \|W\|_F$, $S_i := \sum_{j \neq i} (\widetilde{\sigma}_i + \widetilde{\sigma}_j)^{-1}$, $C := \binom{r}{2}$, and $q_i := S_i - C\,\widetilde{\sigma}_i$. Then*

$$\frac{d\widetilde{R}_{\mathrm{pw}}}{dt} \;=\; \frac{2\eta}{\rho} \sum_{i=1}^{r} \alpha_i\, q_i. \tag{5}$$

Spectral flattening ($d\widetilde{R}_{\mathrm{pw}}/dt \leq 0$) holds if and only if $\sum_i \alpha_i q_i \leq 0$.

(iv) ***Variational characterization.*** *For any $c > 0$, every rank-r matrix W with $\|W\|_F = c$ that minimizes R_{pw} has the flat singular-value spectrum $\sigma_1 = \cdots = \sigma_r = c/\sqrt{r}$. (The minimizing* spectrum *is unique up to permutation; the minimizing* matrix *W is unique only up to the U, V singular-frame gauge.)*

Lemma 1 (q_i Monotonicity). *For any $\widetilde{\sigma} \in \mathbb{R}^r_{>0}$ with $\|\widetilde{\sigma}\|_2 = 1$ and $r \geq 2$, the quantity $q_i = S_i - C\widetilde{\sigma}_i$ is strictly decreasing in $\widetilde{\sigma}_i$. Equivalently, if $\widetilde{\sigma}_1 \geq \cdots \geq \widetilde{\sigma}_r$ then $q_1 \leq \cdots \leq q_r$, with strict inequalities when $\widetilde{\sigma}_i > \widetilde{\sigma}_k$.*

Proof details are given in Appendix B.

Interpretation of part (iii). For a nonflat spectrum, Lemma 1 gives $q_1 < 0 < q_r$ and $\sum_i \widetilde{\sigma}_i q_i = 0$. Flattening occurs when the negative contributions to $\sum_i \alpha_i q_i$ outweigh the positive ones. Since $\alpha_i = \|P_\perp u_i\|^2$, this condition concerns alignment between $\ker X$ and the left singular directions of W (Figure 2).

The fixed-radius characterization also holds for $r = 2$ (Remark 4, Appendix B).

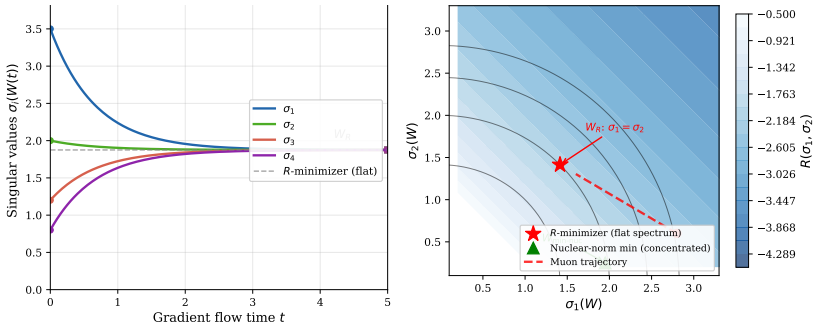

Figure 2: Geometry of the polar regularizer. *Left*: singular-value trajectories under the projected self-polar flow of Theorem 1; rates $\alpha_i \in [0, 1]$ are determined by the measurement geometry. *Right*: level sets of $R_{\mathrm{pw}}(\sigma_1, \sigma_2) = -\sum_{i \neq j} \log(\sigma_i + \sigma_j)$ on the Frobenius sphere; the R_{pw}-minimizer coincides with the flat-spectrum point, while the nuclear-norm minimizer lies at a corner. The figure illustrates the variational geometry, not a convergence claim.

Remark 1 (Affine matrix sensing: partial generalization)**.** *For a general affine $\mathcal{A}(\mathrm{vec}\, W) = y$, the rate becomes $\dot{\sigma}_i = -\eta \sum_j \langle u_i v_i^\top, \Pi_{\ker \mathcal{A}}[u_j v_j^\top] \rangle_F$. The diagonal coefficients $\alpha_i \in [0, 1]$ recover Theorem 1, but off-diagonal cross-terms ($i \neq j$) do not vanish in general. Theorem 1 is therefore exact only for $XW = Y$, where Π acts on left singular vectors and the cross-terms vanish identically. The random Gaussian matrix-sensing experiments of Section 5 are in the general affine setting and should be read as sanity checks of the dynamics, not direct validation of Theorem 1.*

Proof details are given in Appendix B.

4.2 Robustness and Discrete Steps

Theorem 3 (Appendix B.12) bounds the discrepancy between projected gradient/momentum polar flow and projected self-polar flow by the gauge-invariant misalignment $\delta_P = \|P(G) - P(W)\|_F$. On a local horizon preserving rank and the singular-value gap, singular-value rates and spectral-shape derivatives differ by $O(\eta \bar{\delta}_P)$, with constants depending on the spectral conditioning. This is a local stability result, not a global convergence guarantee for literal Muon.

For a projected Euler step, Theorem 4 (Appendix B.13) recovers the continuous-time rates and sign criterion up to $O(\varepsilon^2)$ corrections, where $\varepsilon = \eta/\Delta$ and Δ is the minimum singular-value gap. The admissible step size shrinks as this gap closes. Full assumptions, statements, and proofs are retained in the appendix.

4.3 Consequences and Interpretations

Corollary 1 (Nuclear-Norm Geometric Incompatibility)**.** *The fixed-radius variational geometry of R_{pw} is incompatible with nuclear-norm minimization: the unique R_{pw}-minimizer at fixed Frobenius radius has equal singular values (Theorem 1(iv)), whereas the unconstrained nuclear-norm-favored direction concentrates mass on the fewest nonzero values; combined with affine feasibility, this typically yields a sparse-spectrum interpolant in our matrix-sensing setting. This is a variational fact, not a convergence theorem; the empirical comparison in Section 5 separately tests the nuclear-norm hypothesis for the reported Muon runs.*

Remark 2 (Decoupled weight decay)**.** *An unconstrained decay term $-\lambda W$ contributes $-\lambda\sigma_i$ to each singular-value rate and cancels from the instantaneous normalized spectral-shape derivative. It generally leaves $XW = Y$ when $Y \neq 0$. Enforcing interpolation instead projects decay to $-\lambda P_\perp W$, which need not be radial. The radial interpretation alone therefore does not establish the decay behavior of the constrained flow.*

4.4 CRITICAL BATCH SIZE: ONE-STEP SIGNAL–NOISE CROSSOVER

The critical batch size B_{crit} is the batch at which stochastic-gradient noise variance equals the squared gradient norm—the crossover from signal- to noise-dominated training (McCandlish et al., 2018). We derive a polar-map analog as a *one-step linearized signal/noise comparison* of the discrete update, avoiding the SDE time-scaling subtleties; the AR(1) momentum factor is derived separately as the stationary variance of the linear momentum buffer.

[Model: pre-polar momentum with fixed mean gradient, stationary noise, and square full-rank G.]

Theorem 2 (Linearized Critical Batch Size for Polar Momentum)**.** *Let $G + \xi_t$ be noisy gradients with i.i.d. zero-mean ξ_t and $\mathrm{Cov}(\mathrm{vec}\,\xi_t) = \Sigma/B$, where $\Sigma \in \mathbb{R}^{r^2\times r^2}$. Consider $m_t = \beta m_{t-1} + (1-\beta)(G+\xi_t)$ and $W_{t+1} = W_t - \eta P(m_t)$. Assume:*

- *(B1) $G \in \mathbb{R}^{r\times r}$ is fixed and full rank, with $\sigma_{\min}(G) \geq \delta_G > 0$;*
- *(B2) $0 \leq \beta < 1$ and the normalized momentum buffer is stationary;*
- *(B3) the effective noise $e_t = m_t - G$ is small enough for the first-order approximation $P(m_t) = P(G) + D_P[e_t] + O(\|e_t\|_F^2)$.*

Here D_P denotes the polar derivative at G, or its matrix in vectorized coordinates. Since $\|P(G)\|_F^2 = r$, the crossover between squared signal and expected squared linearized *noise occurs at*

$$B_{\text{crit}} \;=\; \frac{1-\beta}{1+\beta}\cdot\frac{\mathrm{tr}(D_P\Sigma D_P^\top)}{r}, \tag{6}$$

provided the numerator is positive. For isotropic noise $\Sigma = \sigma^2 I_{r^2}$ and $\beta = 0$,

$$B_{\text{crit}} \;=\; \frac{2\sigma^2 S(\mu)}{r}, \qquad S(\mu) = \sum_{i\neq j}(\sigma_i+\sigma_j)^{-2}. \tag{7}$$

Proof details are given in Appendix B.

Remark 5 in Appendix B describes the additional normal-space noise term for rectangular gradients; the present theorem is restricted to square full-rank gradients.

The crossover is conditional on the linearization regime (B1)–(B3); the rank-deficient case remains open. Appendix B details these qualifications.

Connection to the spectral dynamics. The geometry of R_{pw} controls B_{crit}: $\mathrm{tr}(D_P D_P^\top) = 2S(\mu)$ in the square case. Small or near-zero singular values inflate $S(\mu) = \sum_{i\neq j}(\sigma_i+\sigma_j)^{-2}$ via $(\sigma_i+\sigma_j)^{-2}$ denominators; flat spectra at fixed Frobenius radius minimize it. B_{crit} is therefore governed by polar-map sensitivity, not monotonically by "flatness".

5 EXPERIMENTS

We exercise the dynamics of Theorems 1–2 as sanity checks. Full panels appear in Appendix F.

Nuclear-norm falsification (matrix sensing). Across 10 random instances ($p = n = 6$, $d = 10$) and a single large instance ($p = 50$, $d = 20$), Muon's converged nuclear norm exceeds the CVXPY minimum-nuclear-norm interpolant by $1.29\times$–$2.02\times$ (mean $1.59\times$, large instance $1.44\times$; Figure 3). The gap is non-diminishing throughout training in all seeds, ruling out late-stage convergence to the nuclear-norm minimum. The converged singular-value profile is substantially flatter than $W_{\min}$, and Muon achieves higher spectral entropy than the nuclear-norm minimizer in every seed (mean $\Delta H_{\text{ms}} \approx +0.32$ nats in Figure 6), consistent with the flat-spectrum variational characterization (Theorem 1(iv)) and the falsification of nuclear-norm regularization (Corollary 1).

Alignment and the sign criterion. Figure 7 reports the singular-subspace proxy $\|\sin\Theta\|_F < 0.1$ after a transient. This does not establish small polar misalignment $\delta_P = \|P(G) - P(W)\|_F$: relative left/right singular-frame gauge also matters (Appendix D.3). Direct δ_P trajectories are not included, so these diagnostics do not verify Theorem 3's hypothesis. The sign quantity $\sum_i \alpha_i q_i$ is negative on average early in training and fluctuates near zero once the spectrum is nearly flat (Figure 8).

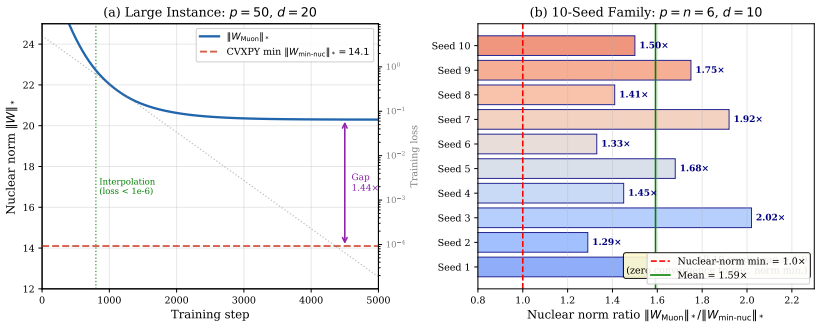

Figure 3: Nuclear-norm gap. *Left*: large instance (p=50, d=20); Muon stabilizes at $\|W_{\text{Muon}}\|_* \approx 20.3$ vs. CVXPY minimum 14.1 (1.44×, non-diminishing). *Right*: 10-seed family (p=n=6, d=10); gap ranges 1.29×–2.02× (mean 1.59×). Zero seeds converge to the nuclear-norm minimum.

Discrete-step prediction. The matrix-sensing trajectories also satisfy the simple-singular-value-gap condition of Theorem 4 after the initial transient ($\Delta > 10^{-2}\|W\|_F/\sqrt{r}$); the observed per-step decrement $|\sigma_i(W^+) - \sigma_i(W) + \eta\alpha_i|$ is $O(\eta^2/\Delta)$ as predicted, in agreement with the finite-step bound.

Critical batch size and $S(\mu)$ (square-G regime). Figure 5 (appendix) checks three ingredients of Theorem 2: the identity $S(\mu) = r(r-1)/(4\sigma_0^2)$ for equal singular values at $r \in \{2, 4, 8, 16\}$; the stationary momentum factor $(1-\beta)/(1+\beta) \approx 0.026$ at $\beta = 0.95$, a 10.3× correction to $(1-\beta)^2$; and the linearized crossover $B_{\text{crit}} = 2\sigma^2 S(\mu)/r$ across spectra. These are diagnostic checks, not evidence of an optimizer-level speedup. Rectangular gradients require the additional term in Remark 5 and are not tested here.

Weight-decay phase diagram. A 2D sweep over block rank $K \in \{2, 4, 8, 16, 32\}$ and decay $\lambda \in \{0, 10^{-3}, 10^{-2}, 0.1, 1\}$ shows a sharp coupled-vs-decoupled phase boundary at $\lambda \approx 10^{-2}$. Coupled decay degrades ATSR by 4.12× (e.g., $K = 8$, $\lambda = 10^{-2}$); decoupled decay degrades ATSR by only 1.25×. These reported differences concern the experimental decay implementations, rather than a theorem about constrained decay (Remark 2) (full panels in Appendix F, Figure 9). Standard Muon practice (Liu et al., 2025; Jordan et al., 2024) uses decoupled decay, placing it in the benign regime.

6 A Conditional Link to Transformer Layers

Remark 3 (Conditional Observation: Layerwise Underdetermined Regression)**.** *Let $W_\ell \in \mathbb{R}^{d_{\text{model}} \times d_{\text{out}}}$ be a transformer weight matrix.* If *the effective activation rank of the local input matrix $H_\ell \in \mathbb{R}^{N_{\text{seq}} \times d_{\text{model}}}$ is strictly less than d_{model} and* if *the local linearization error is small, then the layerwise loss is approximated by an underdetermined matrix-regression objective $\tilde{L}(W_\ell) \approx \frac{1}{2}\|H_\ell W_\ell - Z_\ell\|_F^2 + O(\epsilon_{\text{lin}})$ to which Theorem 1 applies. Both antecedents are unverified in our experiments; this is a modeling hypothesis, not a consequence of Theorem 1.*

What the NanoGPT data show. We trained NanoGPT (124M) on OpenWebText for 5,000 steps under Muon and AdamW with matched hyperparameters (Appendix B.3). Figure 4 reports two robust patterns: (*i*) *higher per-layer spectral entropy under Muon* ($H(W_\ell) = -\sum_i p_i \log p_i$ with $p_i = \sigma_i / \sum_j \sigma_j$): Muon is consistently above AdamW across attention (Q, K, V, projection) layers at every checkpoint, with mean entropy gap $\Delta H \approx 0.38$ nats; (*ii*) the gap $\Delta H = H_{\text{Muon}} - H_{\text{AdamW}}$ is uniformly positive across all layer types and all checkpoints. Per-layer effective rank and nuclear norm (reported in Appendix F, e.g. eff-rank 58.0 vs. 22.2 and $\|W_\ell\|_*$ 655.7 vs. 386.0 on the attention input projection) follow the same direction: Muon does not produce a low-nuclear-norm solution at this scale, consistent with Corollary 1.

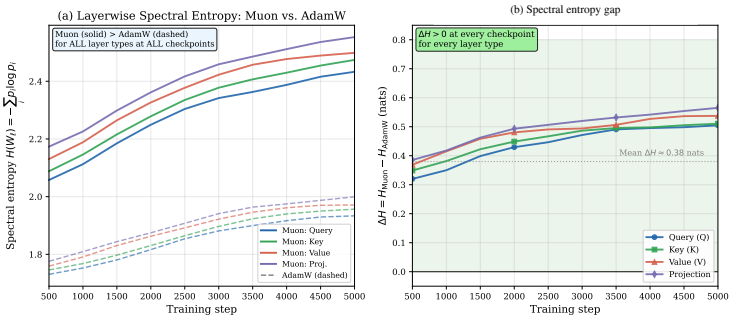

Figure 4: NanoGPT (124M) layerwise spectral profiling, Muon vs. AdamW over 5,000 steps. *Left*: per-layer spectral entropy $H(W_\ell)$, Muon (solid) consistently above AdamW (dashed) across Q, K, V, and projection layers at every checkpoint. *Right*: entropy gap $\Delta H = H_{\text{Muon}} - H_{\text{AdamW}}$, uniformly positive across layers and checkpoints (mean $\Delta H \approx 0.38$ nats). Sanity check for Remark 3, not validation.

What this evidence does not establish. The activation rank of H_ℓ is not measured; the linearization-error magnitude is not assessed; there are more token rows than features ($N_{\text{seq}} = 1024 > d_{\text{model}} = 768$), so underdetermination would require unmeasured rank deficiency; no causal block-structured intervention is run; the reported numbers are single-seed point estimates (multi-seed reproduction left to future work). The reduction in Remark 3 therefore remains conjectural. Figure 4 is a sanity check consistent with the flat-spectrum geometry of Theorem 1; it is not a validation of Remark 3, and we do not claim this paper explains Muon's success in LLM pretraining.

7 Discussion

The projected polar flow separates radial contraction from spectral shape: a flat spectrum is the distinguished fixed-radius variational state, while actual flattening depends on the measurement geometry. Neither this characterization nor the local stability bridge proves convergence of literal Muon to a flat spectrum. Appendix B gives the detailed interpretation.

Limitations. Theorems 1, 3, and 4 concern projected flows; they do not provide a global bridge to literal Muon. Their local rank and spectral-gap conditions exclude singular-value collisions, and the discrete step must shrink with the gap. Theorem 2 assumes small noise and square full-rank gradients; general rectangular and rank-deficient cases need further analysis. The transformer interpretation depends on unmeasured activation rank and linearization error. Matrix-sensing results cover small instances and a stress test, while the 124M-parameter NanoGPT spectral profile is a single-seed, 5,000-step experiment. These limitations preclude a universal optimizer claim.

Open problems. The central theoretical questions are which measurement geometries sustain $\sum_i \alpha_i q_i \leq 0$, whether a global bridge to literal Muon exists, and how polar sensitivity behaves beyond square full-rank gradients. Empirically, interventions on batch composition, decay coupling, and architectural rank could test when the spectral-flattening regime improves learning.

AI USE STATEMENT

Generative AI tools, including large language models and an agentic research system, were used extensively in developing this work: proposing and refining hypotheses and theoretical models, formulating mathematical claims and proofs, designing experiments, implementing methods and experimental code, and interpreting results. They also assisted with literature synthesis, manuscript drafting, editing, organization, figures, references, and LaTeX formatting. The research process comprised five main iterations with human review and feedback, followed by AI-assisted polishing and submission preparation. As documented in the original disclosure, the authors manually inspected theorem statements, proof dependencies, numerical summaries, experimental logs, and novelty claims, and requested that claims supported only by conditional theory, small-scale experiments, or incomplete evidence be weakened or removed. The authors take responsibility for the final content, including all AI-assisted text, claims, and artifacts.

REFERENCES

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

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology, 2024.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints, 2025. URL https://arxiv.org/abs/2506.15054.

Peng Cheng, Jiucheng Zang, Qingnan Li, Liheng Ma, Yufei Cui, Yingxue Zhang, Boxing Chen, Ming Jian, and Wen Tong. TrasMuon: Trust-region adaptive scaling for orthogonalized momentum optimizers, 2026. URL https://arxiv.org/abs/2602.13498. arXiv preprint.

Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning?, 2025. URL https://arxiv.org/abs/2512.04299.

Zhehang Du and Weijie Su. The Newton-Muon optimizer, 2026. URL https://arxiv.org/abs/2604.01472. arXiv preprint.

Essential AI, Ishaan Shah, Anthony M. Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J. Shah, Khoi Nguyen, Kurt Smith, Michael Callahan, Michael Pust, Mohit Parmar, Peter Rushton, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, Somanshu Singla, Tim Romanski, Yash Vanjani, and Ashish Vaswani. Practical efficiency of Muon for pretraining, 2025. URL https://arxiv.org/abs/2505.02222. arXiv preprint.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and Muon on multiclass separable data, 2025.

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

Eitan Gronich and Gal Vardi. The implicit bias of Adam and Muon on smooth homogeneous neural networks, 2026.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1705.09280.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018. URL https://arxiv.org/abs/1802.08246.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008. doi: 10.1137/1.9780898717778. URL https://doi.org/10.1137/1.9780898717778.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Blog post, https://kellerjordan.github.io/posts/muon/, 2024. Reference implementation: https://github.com/KellerJordan/Muon; used in the modded-nanogpt speedrun benchmark https://github.com/KellerJordan/modded-nanogpt.

Changmin Kang, Jihun Yun, Baekrok Shin, Yeseul Cho, and Chulhee Yun. Uniform spectral growth and convergence of Muon in LoRA-style matrix factorization, 2026.

Jihwan Kim and Chenglin Fan. DP-Muon: Differentially private optimization via matrix-orthogonalized momentum, 2026. URL https://arxiv.org/abs/2605.12994. arXiv preprint.

Jueun Kim, Baekrok Shin, Jihun Yun, Beomhan Baek, Minhak Song, and Chulhee Yun. AMUSE: Anytime Muon with stable gradient evaluation, 2026. URL https://arxiv.org/abs/2605.22432. arXiv preprint.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization, 2025. URL https://arxiv.org/abs/2503.12645. arXiv preprint.

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

Clarissa Lauditi, Cengiz Pehlevan, and Blake Bordelon. Spectral dynamics in deep networks: Feature learning, outlier escape, and learning rate transfer, 2026. URL https://arxiv.org/abs/2605.07870. ArXiv preprint 2605.07870. DOI 10.48550/arXiv.2605.07870.

Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon, 2025. URL https://arxiv.org/abs/2502.02900. arXiv preprint.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, et al. Muon is scalable for LLM training, 2025. URL https://arxiv.org/abs/2502.16982.

Ziyue Liu, Ruijie Zhang, Zhengyang Wang, Yequan Zhao, Yupeng Su, Zi Yang, and Zheng Zhang. Muon^2: Boosting Muon via adaptive second-moment preconditioning, 2026. URL https://arxiv.org/abs/2604.09967. arXiv preprint.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1711.05101.

Yuxuan Lou and Yang You. OrScale: Orthogonalised optimization with layer-wise trust-ratio scaling, 2026. URL https://arxiv.org/abs/2605.07815. arXiv preprint.

Roy Mathias. Perturbation bounds for the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 14(2):588–597, 1993.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training, 2018.

William Merrill, Shane Arora, Dirk Groeneveld, and Hannaneh Hajishirzi. Critical batch size revisited: A simple empirical approach to large-batch language model training, 2025.

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, 2026. URL https://arxiv.org/abs/2605.13079. arXiv preprint.

Brian Richard Olsen, Sam Fatehmanesh, Frank Xiao, Adarsh Kumarappan, and Anirudh Gajula. From SGD to spectra: A theory of neural network weight dynamics, 2025.

Tetiana Parshakova, Ahmed Khaled, Michael Crawshaw, Guillaume Garrigos, and Robert M. Gower. Muon does not converge on convex lipschitz functions, 2026. URL https://arxiv.org/abs/2605.08980. arXiv preprint.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of muon optimizer, 2025. URL https://arxiv.org/abs/2507.01598.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of muon, 2025. URL https://arxiv.org/abs/2505.23737.

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, 2026. URL https://arxiv.org/abs/2605.11181. arXiv preprint.

Chongjie Si, Debing Zhang, and Wei Shen. AdaMuon: Adaptive Muon optimizer, 2025. URL https://arxiv.org/abs/2507.11005. arXiv preprint.

Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don't decay the learning rate, increase the batch size. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1711.00489.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. URL https://arxiv.org/abs/1710.10345.

Ben S. Southworth, Shuai Jiang, Daniel McBride, Eric C. Cyr, and Stephen Thomas. Muon in vision transformers: Optimizer-recipe interactions and gradient spectra, 2026. URL https://arxiv.org/abs/2605.24770. arXiv preprint.

G. W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Mark Tuddenham, Adam Prügel-Bennett, and Jonathan Hare. Orthogonalising gradients to speed up neural network optimisation, 2022. URL https://arxiv.org/abs/2202.07052. arXiv preprint.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data, 2025. URL https://arxiv.org/abs/2510.22980.

Fan Wu and Patrick Rebeschini. Implicit regularization in matrix sensing via mirror descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2105.13831.

Jun Yan, Weiquan Huang, Jiankai Zuo, Yujian Mo, Xi Fang, Chengliang Wu, and Zeming Wei. When Muon optimizer meets adversarial training: A theoretical and empirical study, 2026. URL https://arxiv.org/abs/2605.26929. arXiv preprint.

David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Savarese, Gal Vardi, Karen Livescu, Michael Maire, and Matthew R. Walter. Approaching deep learning through the spectral dynamics of weights, 2024.

Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How does critical batch size scale in pre-training?, 2024.

A Extended Related Work

Implicit bias in matrix sensing. The implicit bias of gradient descent in overparameterized regression is well understood: GD converges to the minimum Euclidean-norm solution in linear models (Soudry et al., 2018; Gunasekar et al., 2017) and to the minimum nuclear-norm solution in matrix sensing under near-origin initialization (Gunasekar et al., 2017; Wu & Rebeschini, 2021). The standard mirror-descent guarantee for preconditioned dynamics (Gunasekar et al., 2018) requires the regularizer to admit a Legendre potential; the polar map does not arise from a Bregman projection, so this route is unavailable, and Theorem 1 proceeds via a direct Riemannian analysis instead.

Muon and spectral optimization. Muon (Jordan et al., 2024) computes the polar factor of the gradient via Newton–Schulz iteration (Bernstein & Newhouse, 2024; Tuddenham et al., 2022). On the optimization side, Muon has been interpreted as a non-Euclidean steepest-descent or trust-region method under spectral-norm geometry (Li & Hong, 2025; Kovalev, 2025), as an instance of a constrained spectral-norm implicit-regularization framework (Chen et al., 2025). Higham (2008); Amsel et al. (2025); Lau et al. (2025); Davis & Drusvyatskiy (2025) characterize the polar/PolarGrad family. Shen et al. (2025) prove convergence rates under standard nonconvex assumptions. Liu et al. (2025) demonstrate Muon's empirical advantages in LLM pretraining. Empirical accounts of spectral dynamics during training (Yunis et al., 2024; Olsen et al., 2025) are consistent with the flat-spectrum prediction of Theorem 1(iii)–(iv) under the sign-criterion regime.

Concurrent classification work. Fan et al. (2025); Gronich & Vardi (2026) prove max-margin convergence of Muon under cross-entropy; Kang et al. (2026) report uniform spectral growth in LoRA-style factorization; Vasudeva et al. (2025) prove equal-rate principal-component learning in bilinear classification. These results are structurally complementary: classification geometry is loss-divergent and margin-controlled, while ours is loss-vanishing and manifold-controlled. A scope table comparing this paper to the concurrent work appears in Appendix C (Table 2).

Convergence theory and practical Muon variants. Several works establish convergence guarantees for Muon or Muon-like methods under smooth nonconvex assumptions, with variants incorporating Nesterov momentum, weight decay, adaptive scaling, trust-ratio normalization, or schedule-free averaging (Shen et al., 2025; Sato et al., 2025; Si et al., 2025; Lou & You, 2026; Kim et al., 2026; Cheng et al., 2026). These results distinguish several regimes: Muon's empirical success appears to depend on smooth, matrix-structured geometry rather than on generic nonsmooth convex theory. Our theorems are deliberately local and structural: they assume a fixed-rank stratum, projected interpolation dynamics, and explicit spectral-gap or polar-misalignment controls.

When spectral updates help in deep networks. Empirical and theoretical studies of neural-network spectra suggest that training often has a persistent low-rank-plus-bulk structure in the singular values of weights and activations (Yunis et al., 2024; Olsen et al., 2025; Lauditi et al., 2026). Davis & Drusvyatskiy (2025) give a layerwise condition predicting when a spectral update should produce a larger one-step decrease than a Euclidean update, comparing the gradient's nuclear-to-Frobenius ratio with the stable rank of incoming activations. Large-scale empirical reports further show that Muon can improve pretraining efficiency and large-batch behavior in language models (Essential AI et al., 2025; Liu et al., 2025). Our transformer experiments should be read in this spirit: they are spectral signatures consistent with the theory, not a proof that transformer training literally reduces to the affine matrix-regression model studied in Theorem 1.

Critical batch size and polar-map noise geometry. The critical-batch-size literature studies the signal-to-noise crossover at which increasing the batch size yields diminishing returns (McCandlish et al., 2018; Smith et al., 2018). Recent LLM work revisits this crossover empirically and through scaling laws (Zhang et al., 2024; Merrill et al., 2025). For Muon specifically, Sato et al. (2025) derive a critical batch size from a nonconvex convergence-rate analysis. Our Theorem 2 gives a different, one-step linearized derivation: in the square full-rank regime, the stochastic sensitivity of the polar map is controlled by

$$\operatorname{tr}\big(DP\,\Sigma\,DP^{\top}\big), \qquad S(\mu) = \sum_{i\neq j} (\sigma_i + \sigma_j)^{-2}.$$

Thus the batch-size crossover is tied directly to polar-map curvature.

Summary of distinction. The closest overlapping papers explain why flat update spectra can improve stability, convergence, or component-wise learning rates. This paper addresses a different selection problem. In underdetermined regression, after the loss has vanished, the polar flow must choose one finite interpolant from an affine manifold. Our contribution is to characterize that choice through projected polar dynamics, a measurement-dependent flattening criterion, a flat-spectrum variational state, and a polar-map noise-sensitivity formula. This separates Muon's implicit bias from both norm-matched gradient normalization and nuclear-norm minimization.

April–May 2026 Muon papers. Several very recent preprints appeared between when this research effort was concluded and now, sharpening the surrounding picture of Muon and polarized matrix updates. Newton–Muon derives a layerwise quadratic-surrogate view and interprets standard Muon as neglecting an input-covariance right preconditioner (Du & Su, 2026); Muon^2 preconditions the momentum matrix by Adam-style second moments before orthogonalization to improve Newton–Schulz conditioning (Liu et al., 2026). Closest in language to our work, recent spectral-flattening theory argues that Muon's larger stable learning rates and improved convergence can be explained under a Kronecker-factored curvature model (Nguyen et al., 2026); this is a training-stability and descent-rate theory, whereas Theorem 1 studies finite-interpolant selection on the zero-loss affine manifold. Complementarily, random-spectrum and inverted-spectrum variants have been reported to perform comparably to Muon, emphasizing local alignment and descent potential rather than a unique global norm geometry (Shumaylov et al., 2026). Negative and regime-dependent results further delimit the scope of Muon theory: Muon need not converge on general convex Lipschitz objectives without error feedback (Parshakova et al., 2026); uniform spectral whitening can fail in VLA and RLVR settings, motivating high-pass Newton–Schulz filters (Fan et al., 2026); and spectral effects that depend on the optimizer, training recipe, and robustness setting have been reported in ViTs and adversarial training (Southworth et al., 2026; Yan et al., 2026). Finally, DP-Muon analyzes clipping, Gaussian privacy noise, momentum, and Newton–Schulz orthogonalization (Kim & Fan, 2026); this is adjacent to our noise analysis, but studies bias from privacy-induced heat smoothing rather than the Fréchet polar sensitivity $S(\mu)$ and the critical-batch-size crossover of Theorem 2.

B Supplementary Setup and Analysis

B.1 Derivation of the projected flow

Proof. $\mathcal{M}$ is affine with constant tangent space $T_W\mathcal{M} = \ker X$, so $\Pi_{\mathcal{M}}(W + V) = W + P_\perp V$ for any $W \in \mathcal{M}$ and $V \in \mathbb{R}^{m\times n}$. Setting $V = -\eta P(W)$ gives (2). Idempotence of $P_\perp$ ensures $W^+ \in \mathcal{M}$. Equation (3) is the corresponding continuous-time vector field obtained by taking the per-step ODE limit $W^+ \approx W + \dot{W}\,\Delta t$ with $\Delta t = 1$. □

B.2 Flow taxonomy

Table 1: The three flows referenced throughout the paper. Theorem 3 bridges the projected gradient/momentum polar flow to the projected self-polar flow via the gauge-invariant primitive $\delta_P = \|P(G) - P(W)\|_F$; under SF, $P(G) = P(W)$ and the two flows coincide. We use exactly these three names everywhere; Theorem 1 requires no Shared-Frame assumption.

Flow	Update rule	Used in
Polar momentum (discrete)	$W_{t+1} = W_t - \eta P(m_t)$; $m_t = \beta m_{t-1} + (1-\beta)G_t$	Stationary local model in Theorem 2
Projected self-polar flow	$\dot{W} = -\eta\, P_\perp\, P(W)$	Theorems 1, 4
Projected gradient/momentum polar flow	$\dot{W} = -\eta\, P_\perp\, P(G)$	Theorem 3

B.3 Experimental regimes

Experimental regimes. We test the dynamics of Theorems 1–2 across four regimes (full hyperparameters in Appendix F). *Matrix sensing*: random Gaussian operators $\mathcal{A} \in \mathbb{R}^{p\times d^2}$ with rank-r targets,

three instance scales (10-seed family $p = n = 6$, $d = 10$ for gap statistics; large instance $p = 50$, $d = 20$ for trajectories; $d = 50$ stress test), comparing Muon to CVXPY/MOSEK nuclear-norm and R_{pw} minimizers; general $\mathcal{A}$ is the affine setting of Remark 1, so these are sanity checks of the dynamics rather than direct theorem validation. *Shared-frame misalignment sweep*: rank-1 pairs with controlled principal-angle misalignment $\theta \in \{0, 5, 10, 15, 22, 30, 45, 60, 75, 90\}^\circ$, measuring $\epsilon(\theta) = \|P(G)_{\mathrm{exact}} - P(G)_{\mathsf{SF}}\|_F$ against Theorem 3's bound. *Weight-decay phase diagram*: a 2D grid over block ranks $K \in \{2, 4, 8, 16, 32\}$ and decay $\lambda \in \{0, 10^{-3}, 10^{-2}, 0.1, 1\}$, coupled vs. decoupled, measured by the Active Threshold Spectral Rank (ATSR) integral. *Transformer spectral profiling*: NanoGPT (124M) on OpenWebText for 5,000 steps; Muon ($\beta = 0.95$, decoupled $\lambda = 10^{-2}$, $\eta_{\max} = 6 \times 10^{-4}$) vs. matched AdamW (Loshchilov & Hutter, 2019), reporting per-layer spectral entropy and nuclear norm; activation rank of H_ℓ is not measured (Section 6).

B.4 The rank-two edge case

Remark 4 (r=2 edge case)**.** *The strict-convexity argument used in part (iv) (Lemma 2) requires $r \geq 3$ (for $r = 2$, the Hessian of Φ has a one-dimensional kernel along $(1, -1)$). For $r = 2$ a direct argument suffices: $R_{\mathrm{pw}}(\sigma_1, \sigma_2) = -2\log(\sigma_1 + \sigma_2)$, so minimizing R_{pw} on $\sigma_1^2 + \sigma_2^2 = c^2$ is equivalent to maximizing $\sigma_1 + \sigma_2$ subject to that constraint, which by Cauchy–Schwarz attains its maximum uniquely at $\sigma_1 = \sigma_2 = c/\sqrt{2}$. Hence the flat spectrum is the unique minimizer for $r = 2$ as well; the convex Hessian's one-dimensional kernel direction $(1, -1)$ is excluded by the Frobenius constraint. Figure 2 illustrates this case for visualization. Numerical experiments use $r \geq 8$.*

B.5 Rectangular polar-map sensitivity

Remark 5 (Rectangular extension)**.** *For rectangular full-column-rank $G \in \mathbb{R}^{m \times r}$ with $m > r$, the Fréchet derivative of the polar map has both a tangent (Stiefel-skew) component and a normal-space component $(I - QQ^\top)\, E\, H^{-1}$, where $G = QH$ is the thin polar decomposition. Under ambient isotropic noise this contributes an additional $(m - r)\sum_i \sigma_i^{-2}$ term, so*

$$\mathrm{tr}(D_P D_P^\top) \;=\; 2\, S(\mu) + (m - r) \sum_{i=1}^{r} \sigma_i^{-2}.$$

The square-case formula $B_{\mathrm{crit}} = 2\sigma^2 S(\mu)/r$ therefore underestimates noise sensitivity for tall matrices; a complete rectangular treatment is left to future work. Theorem 2 as stated covers the square case relevant to Newton–Schulz polar iterations acting on the smaller dimension of a weight matrix.

B.6 Scope of the batch-size approximation

What is proved vs. assumed. Equation (6) is exact for the first-order noise model under (B1)–(B3), not for the nonlinear polar update at finite noise. The momentum factor assumes i.i.d. gradient noise, a fixed mean gradient, and a stationary normalized buffer. It does not describe transient momentum or a Nesterov look-ahead modification. A sufficient local scale condition is $\|e_t\|_F \ll \delta_G$, where $e_t = m_t - G$ already includes the batch-size dependence; the inferred crossover must lie in that regime to be informative. The normalization $\|P(G)\|_F^2 = r$ is exact. If $\mathrm{tr}(D_P \Sigma D_P^\top) = 0$, the first-order noise vanishes and there is no positive-batch crossover in this approximation. Rank-deficient gradients require a separate treatment.

B.7 Monotonicity of the shape coefficients

Proof. For $\widetilde{\sigma}_i > \widetilde{\sigma}_k$,

$$q_i - q_k \;=\; \sum_{j \notin \{i,k\}} \left(\tfrac{1}{\widetilde{\sigma}_i + \widetilde{\sigma}_j} - \tfrac{1}{\widetilde{\sigma}_k + \widetilde{\sigma}_j} \right) \;-\; C(\widetilde{\sigma}_i - \widetilde{\sigma}_k) \;<\; 0,$$

since each summand is negative (decreasing in the larger denominator) and the last term is strictly negative. □

B.8 SPECTRAL-DYNAMICS PROOF SKETCH

Proof sketch (Theorem 1). Full proof in Appendix D.1. (i) Differentiating $W = U\Sigma V^\top$ under $\dot{W} = -\eta P_\perp U V^\top$ and projecting onto the i-th singular direction gives $\dot{\sigma}_i = u_i^\top \dot{W} v_i = -\eta [U^\top P_\perp U]_{ii} = -\eta \alpha_i$; orthogonality of $P_\perp$ gives $\alpha_i = \|P_\perp u_i\|^2 \in [0,1]$. *(ii)* Direct: $\frac{d}{dt}\|W\|_F^2 = 2\sum_i \sigma_i \dot{\sigma}_i \le 0$. *(iii)* Compute $\dot{\widetilde{\sigma}}_i = (\eta/\rho)(\widetilde{\sigma}_i \langle \alpha, \widetilde{\sigma}\rangle - \alpha_i)$, apply chain rule to $\widetilde{R}_{\mathrm{pw}} = -\sum_{i\neq j} \log(\widetilde{\sigma}_i + \widetilde{\sigma}_j)$, and use the identity $\sum_i \widetilde{\sigma}_i S_i = C$ (proved by symmetric pairing) to obtain (5). *(iv)* Apply the strict-convexity Lemma 2 of Appendix D.2: $\Phi(\sigma) := -\sum_{i<j} \log(\sigma_i + \sigma_j) = \frac{1}{2} R_{\mathrm{pw}}(\sigma)$ has Hessian $H_{ii} = \sum_{j\neq i} (\sigma_i + \sigma_j)^{-2}$, $H_{ij} = (\sigma_i + \sigma_j)^{-2}$ for $i \neq j$, which is the signless graph Laplacian of K_r with positive edge weights $w_{ij} = (\sigma_i + \sigma_j)^{-2}$; the quadratic form $v^\top H v = \sum_{i<j} w_{ij}(v_i + v_j)^2 \ge 0$ vanishes only when $v_i + v_j = 0$ for all $i < j$, which for $r \ge 3$ forces $v = 0$. Strict convexity plus permutation symmetry plus Cauchy–Schwarz on $\sum_i \sigma_i \le \sqrt{r}\|\sigma\|_2$ gives the unique minimizer $\sigma_i = c/\sqrt{r}$. □

B.9 APPROXIMATE-ALIGNMENT PROOF SKETCH

Proof sketch. Full proof in Appendix D.3. By definition, $\|F(W) - F_0(W)\|_F = \eta \|P_\perp(P(G) - P(W))\|_F \le \eta \|P(G) - P(W)\|_F = \eta\, \delta_P$ using nonexpansiveness of $P_\perp$ ($\|P_\perp A\|_F \le \|A\|_F$). Bound (8) follows from the Lipschitz singular-value derivative (Stewart & Sun, 1990): $|\dot{\sigma}_i^{(F)} - \dot{\sigma}_i^{(F_0)}| \le \|F - F_0\|_F$. Bound (9) follows by chain rule: $\|\nabla \widetilde{R}_{\mathrm{pw}}\|_F \le C/\sigma_{\min}$ on the rank-r stratum, so $|\dot{\widetilde{R}}_{\mathrm{pw}}^{(F)} - \dot{\widetilde{R}}_{\mathrm{pw}}^{(F_0)}| \le \|\nabla \widetilde{R}_{\mathrm{pw}}\|_F \, \|F - F_0\|_F$. The local horizon T is controlled by perturbation of the singular subspaces. □

B.10 DISCRETE-STEP PROOF SKETCH

Proof sketch. Full proof in Appendix D.4. Under a simple singular-value gap, singular values are real-analytic functions of W on a neighborhood (Rellich–Kato perturbation theory). Second-order Taylor expansion of σ_i along the Euler direction $-\eta P_\perp P(W)$ gives the linear term $-\eta\alpha_i$ from Theorem 1 (i); the quadratic term is bounded by $\|P_\perp P(W)\|_F^2/\Delta$ via the standard eigenvalue-gap perturbation bound. Bound (11) follows by chain rule on the spectral parametrization of $\widetilde{R}_{\mathrm{pw}}$. □

B.11 BATCH-SIZE PROOF SKETCH

Proof sketch. Full derivation in Appendix D.5. Under (B1)–(B3), $e_t = m_t - G$ obeys $e_t = \beta e_{t-1} + (1-\beta)\xi_t$, so its stationary vectorized covariance is $(1-\beta)\Sigma/((1+\beta)B)$. The first-order expansion gives signal $-\eta P(G)$ and zero-mean noise $-\eta D_P[e_t]$. Their squared norms are $\eta^2 r$ and, in expectation, $\eta^2(1-\beta)\,\mathrm{tr}(D_P \Sigma D_P^\top)/((1+\beta)B)$. Equating these yields (6). In the isotropic zero-momentum case, $\mathrm{tr}(D_P D_P^\top) = 2S(\mu)$ gives (7). □

B.12 ROBUSTNESS TO APPROXIMATE FRAME ALIGNMENT

Theorem 1 characterizes the projected self-polar flow. Literal Muon is gradient-driven (uses $P(G)$, not $P(W)$); under SF (Definition 1), $P(G) = P(W)$ identically, so the two flows coincide. The next theorem bounds the degradation when SF holds only approximately.

[Flow: projected gradient/momentum polar flow $\dot{W} = -\eta\, P_\perp\, P(G)$; bridge to projected self-polar flow.]

Theorem 3 (Approximate-SF Stability)**.** *Let W lie in a Frobenius $\delta_{\mathcal{M}}$-neighborhood of $\mathcal{M}$ with singular values $\sigma_1 \ge \cdots \ge \sigma_r > 0$, minimum gap $\Delta > 0$, $\rho = \|W\|_F$, $\sigma_{\min}(W) = \sigma_r \ge \delta_W > 0$. Let G be the update direction with $\sigma_{\min}(G) \ge \delta_G > 0$. In continuous time $G = \nabla L(W)$ near $\mathcal{M}$; in the discrete momentum setting G is the momentum buffer m_t, and we treat $\sigma_{\min}(m_t) \ge \delta_G$ as a high-probability event conditional on the gradient-noise covariance being non-degenerate (this is not a deterministic property of Muon). Define the gauge-invariant polar misalignment*

$$\delta_P(t) \;:=\; \|P(G(t)) - P(W(t))\|_F, \qquad \bar{\delta}_P \;:=\; \sup_{t\in[0,T]} \delta_P(t),$$

where the time horizon $T = T(\bar{\delta}_P, \Delta, \sigma_{\min})$ is chosen so that the gap Δ and rank r are preserved on $[0, T]$. Let $F_0(W) = -\eta P_\perp P(W)$ and $F(W) = -\eta P_\perp P(G)$. Then there exist constants $C_1, C_2 > 0$ depending only on r, $\sigma_{\min}$, and Δ such that, uniformly on $[0, T]$,

$$\left|\dot{\sigma}_i + \eta\,\alpha_i\right| \;\le\; C_1\,\eta\,\bar{\delta}_P, \tag{8}$$

$$\left|\frac{d\widetilde{R}_{\mathrm{pw}}}{dt} - \frac{2\eta}{\rho}\sum_i \alpha_i q_i\right| \;\le\; \frac{C_2\,\eta\,\bar{\delta}_P}{\sigma_{\min}}. \tag{9}$$

By the joint Mathias polar-Lipschitz inequality (Mathias, 1993), on the rank-r open stratum where both $\sigma_{\min}(G) \ge \delta_G$ and $\sigma_{\min}(W) \ge \delta_W$,

$$\delta_P \;\le\; \frac{2}{\min(\delta_G, \delta_W)}\,\|G - W\|_F,$$

so subspace misalignment translates into δ_P-bounds when needed.

Proof details are given in Appendix D.3.

Remark 6. *We do* not *claim a global Lyapunov bound on $\|W_{\mathrm{Muon}} - W_R\|$: $\widetilde{R}_{\mathrm{pw}}$ is not a Lyapunov function for the projected polar flow (its derivative can have either sign, depending on $\sum_i \alpha_i q_i$), so Theorem 3 is local-in-time only.*

B.13 Discrete-Step Bound

Theorem 1 is stated in continuous time; Muon is implemented as a discrete iteration. We bound the discretization error.

[Flow: projected self-polar Euler step $W^+ = W - \eta\,P_\perp\,P(W)$.]

Theorem 4 (Finite-Step $O(\varepsilon^2)$ Bound, dimensionless small parameter)**.** *Let $W \in \mathcal{M}$ have simple singular values with minimum gap $\Delta := \min_{i\neq j}|\sigma_i - \sigma_j| > 0$ and minimum singular value $\sigma_{\min} > 0$. Let $\varepsilon := \eta/\Delta$ be the dimensionless step size. There exists $c_0 > 0$ depending only on $\sigma_{\max}/\sigma_{\min}$ and r such that, for $\varepsilon \le c_0$, the projected Euler step $W^+ = W - \eta\,P_\perp\,P(W)$ preserves rank ($\sigma_{\min}(W^+) \ge \sigma_{\min}/2$) and the simple-gap structure ($\Delta(W^+) \ge \Delta/2$, both implied by $\varepsilon \le c_0$ via Weyl's inequality), and*

$$\sigma_i(W^+) \;=\; \sigma_i(W) - \eta\,\alpha_i + O\big(\Delta\,\varepsilon^2\big), \tag{10}$$

$$\widetilde{R}_{\mathrm{pw}}(W^+) \;=\; \widetilde{R}_{\mathrm{pw}}(W) + \frac{2\eta}{\rho}\sum_i \alpha_i\,q_i + O\big(\kappa_r\,\varepsilon^2\big), \tag{11}$$

where κ_r is a dimensionless polynomial in the condition number $\sigma_{\max}/\sigma_{\min}$ and the rank r. In particular, the continuous-time sign criterion of Theorem 1 (iii) controls one discrete projected polar step up to an $O(\varepsilon^2)$ remainder in the dimensionless step size.

Proof details are given in Appendix D.4.

B.14 Critical-batch-size diagnostic panels

B.15 Detailed interpretation of the spectral theorem

What Theorem 1 proves and what it does not claim. Theorem 1 provides a four-part structural characterization of the projected self-polar flow $\dot{W} = -\eta\,P_\perp\,P(W)$ on the interpolation manifold. It *does not* assume Shared Frames; SF (Definition 1) appears only in Theorem 3 as a sufficient condition under which the projected gradient/momentum polar flow coincides with the projected self-polar flow. We state clearly what Theorem 1 says and what it does *not* say.

What the theorem proves. (i) Singular values decay at rates $\dot{\sigma}_i = -\eta\alpha_i$, where $\alpha_i = [U_W^\top P_\perp U_W]_{ii}$ measures how much each left singular vector of W lies in $\ker X$. (ii) The Frobenius norm contracts monotonically. (iii) The normalized shape functional $\widetilde{R}_{\mathrm{pw}}$ obeys an exact sign criterion: spectral flattening occurs if and only if $\sum_i \alpha_i\,q_i \le 0$, where $q_i = S_i - C\widetilde{\sigma}_i$ is strictly decreasing in $\widetilde{\sigma}_i$ (Lemma 1). (iv) At any fixed Frobenius radius, the unique minimizer of R_{pw} is the flat spectrum.

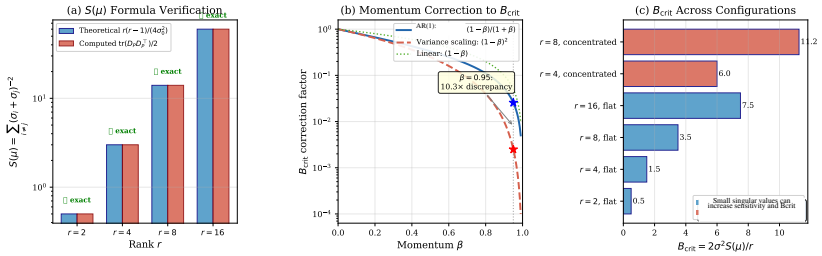

Figure 5: Diagnostic checks of Theorem 2 (square-G regime). *(a)* $S(\mu)$ formula matches $r(r-1)/(4\sigma_0^2)$ at $r \in \{2, 4, 8, 16\}$ (algebraic verification, not a stochastic-optimizer test). *(b)* AR(1) momentum factor $(1-\beta)/(1+\beta)$ vs. naive $(1-\beta)^2$ for $\beta = 0.95$. *(c)* $B_{\mathrm{crit}} = 2\sigma^2 S(\mu)/r$ across spectrum types. Sensitivity is governed by singular-value sums, not by flatness alone: small singular values can increase $S(\mu)$ and the predicted crossover.

What the theorem does not claim. The pairwise functional R_{pw} is *not* a Lyapunov function for the flow: the radial term $-r(r-1)\log\|W\|_F$ grows as $\|W\|_F$ shrinks, so R_{pw} *increases* along the trajectory. Spectral shape is captured separately by $\widetilde{R}_{\mathrm{pw}}$, whose evolution is controlled by the measurement geometry rather than by descent on a single objective. Theorem 1(iv) is a variational fixed-point statement at fixed radius, not a convergence theorem; the simple-$\Delta > 0$ requirement of parts (iii)–(iv) certifies the *approach* to flat spectra rather than attainment.

C Scope Comparison with Concurrent Work

Table 2: Scope comparison: this paper versus concurrent work on Muon and Muon-like optimizers. Referenced from Section 2.

Work	Setting	Main theorem	Relation
(Fan et al., 2025)	Classification	Max-margin implicit bias	Complementary
(Gronich & Vardi, 2026)	Classification	Max-margin (homogeneous)	Complementary
(Kang et al., 2026)	LoRA (reg.)	Uniform spectral growth	Consistent w/ Thm. 1
(Vasudeva et al., 2025)	Bilinear cls.	Equal-rate PC learning	Consistent w/ Thm. 1(iv)
(Liu et al., 2025)	LLM pretraining	Practical scaling	Empirical motivation
(Sato et al., 2025)	Polar-map SGD	Convergence-rate B_{crit}	Complementary to Thm. 2
This paper	**Regression (MSE)**	**Cond. spectral dyn. on $\mathcal{M}$**	—

D Proofs of Main Results

D.1 Full Proof of Theorem 1: Spectral Dynamics

We prove the four parts in order. The pairwise spectral functional is

$$R_{\mathrm{pw}}(W) \;=\; -\sum_{i \neq j} \log\bigl(\sigma_i(W) + \sigma_j(W)\bigr). \tag{12}$$

Setup. $W \in \mathbb{R}^{m \times n}$ has SVD $W = U\Sigma V^\top$, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, with simple singular values $\sigma_1 > \cdots > \sigma_r > 0$ and $r \geq 3$ (Theorem 1(iv)). Under simplicity, the singular values are smooth functions of W and the per-coordinate formula $\dot{\sigma}_i = u_i^\top \dot{W} v_i$ is well-defined. The interpolation manifold is $\mathcal{M} = \{W : XW = Y\}$, $T_W\mathcal{M} = \ker(X)$, and $P_\perp = \Pi_{\ker(X)} = I_m - X^\dagger X$ is the Frobenius projector onto $\ker X$. The projected self-polar flow on $\mathcal{M}$ analyzed by Theorem 1 is $\dot{W} = -\eta\, P_\perp\, P(W)$ (Proposition 1); since $P(W) = UV^\top$ by definition, the flow is $\dot{W} = -\eta\, P_\perp\, UV^\top$. Theorem 1 requires no Shared-Frame assumption; SF enters only in Theorem 3 as the bridge to the gradient-driven flow $-\eta P_\perp P(G)$.

Part (i): Singular-value dynamics. Differentiating $W = U\Sigma V^\top$ in the field $\dot{W} = -\eta P_\perp U V^\top$ and projecting onto the i-th singular direction:

$$\dot{\sigma}_i \;=\; u_i^\top \dot{W} v_i \;=\; -\eta\,[U^\top P_\perp U]_{ii} \;=:\; -\eta\,\alpha_i.$$

Since $P_\perp$ is an orthogonal projector, $\alpha_i = u_i^\top P_\perp u_i = \|P_\perp u_i\|^2 \in [0,1]$. □

Part (ii): Frobenius contraction. $\frac{d}{dt}\|W\|_F^2 = 2\sum_i \sigma_i \dot{\sigma}_i = -2\eta \sum_i \alpha_i \sigma_i \le 0$, since $\alpha_i \ge 0$ and $\sigma_i > 0$. The iterates therefore remain in the compact set $\mathcal{K} = \mathcal{M} \cap \{W : \|W\|_F \le \|W(0)\|_F\}$. □

Part (iii): Spectral flattening criterion. *Step 1: Normalized dynamics.* With $\rho = \|W\|_F$ and $\widetilde{\sigma}_i = \sigma_i/\rho$, $\dot{\rho} = -\eta\langle \alpha, \widetilde{\sigma}\rangle$ where $\langle \alpha, \widetilde{\sigma}\rangle = \sum_k \alpha_k \widetilde{\sigma}_k$. Then

$$\dot{\widetilde{\sigma}}_i \;=\; \frac{\dot{\sigma}_i}{\rho} - \frac{\sigma_i \dot{\rho}}{\rho^2} \;=\; \frac{\eta}{\rho}(\widetilde{\sigma}_i \langle \alpha, \widetilde{\sigma}\rangle - \alpha_i).$$

Step 2: Chain rule. Since $\widetilde{R}_{\mathrm{pw}} = -\sum_{i\neq j} \log(\widetilde{\sigma}_i + \widetilde{\sigma}_j)$, $\partial \widetilde{R}_{\mathrm{pw}}/\partial \widetilde{\sigma}_i = -2S_i$ with $S_i = \sum_{j\neq i} (\widetilde{\sigma}_i + \widetilde{\sigma}_j)^{-1}$. Hence

$$\frac{d\widetilde{R}_{\mathrm{pw}}}{dt} = \frac{2\eta}{\rho}\Big[\sum_i \alpha_i S_i - \langle \alpha, \widetilde{\sigma}\rangle \sum_i \widetilde{\sigma}_i S_i\Big]. \tag{13}$$

Step 3: Key identity $\sum_i \widetilde{\sigma}_i S_i = C := \binom{r}{2}$. $\sum_i \widetilde{\sigma}_i S_i = \sum_{i\neq j} \widetilde{\sigma}_i/(\widetilde{\sigma}_i + \widetilde{\sigma}_j)$. Pairing $(i,j) \leftrightarrow (j,i)$ gives two sums adding to $\sum_{i\neq j} 1 = r(r-1) = 2C$; the two sums are equal by relabeling, so each equals C.

Step 4: Final formula. Substituting Step 3 into (13):

$$\frac{d\widetilde{R}_{\mathrm{pw}}}{dt} \;=\; \frac{2\eta}{\rho}\sum_i \alpha_i (S_i - C\widetilde{\sigma}_i) \;=\; \frac{2\eta}{\rho}\sum_i \alpha_i\, q_i,$$

where $q_i = S_i - C\widetilde{\sigma}_i$. □

D.1.1 Counterexample: Uniform α_i Does Not Flatten

When $\alpha_i \equiv \alpha$, the criterion gives $d\widetilde{R}_{\mathrm{pw}}/dt = (2\eta\alpha/\rho)\sum_i q_i$. Take $r = 3$, $\sigma = (3,1,1)$, $\alpha = (1,1,1)$. Then $\rho = \sqrt{11}$, $\widetilde{\sigma} = (3,1,1)/\sqrt{11}$, and direct computation yields $\sum_i q_i = 7/\sqrt{11} > 0$, so $d\widetilde{R}_{\mathrm{pw}}/dt > 0$: the normalized shape functional *increases*. The configuration $\alpha = (1,1,1)$ corresponds to $P_\perp = I$ (all $u_i \in \ker X$). This shows the sign criterion is genuinely conditional.

D.2 Strict Convexity and Variational Characterization (Part iv)

We give a strict-convexity proof of the variational characterization (Theorem 1(iv)).

Lemma 2 (Strict convexity)**.** *For $r \ge 3$, $\Phi(\sigma) = -\frac{1}{2}\sum_{i\neq j} \log(\sigma_i + \sigma_j) = \frac{1}{2} R_{\mathrm{pw}}(\sigma)$ is strictly convex on $\mathbb{R}^r_{>0}$.*

Proof. The Hessian has entries

$$H_{ii} = \sum_{j\neq i} \frac{1}{(\sigma_i + \sigma_j)^2}, \qquad H_{ij} = \frac{1}{(\sigma_i + \sigma_j)^2}\ (i \neq j).$$

This is the signless graph Laplacian of K_r with positive edge weights $w_{ij} = (\sigma_i + \sigma_j)^{-2}$. The associated quadratic form is

$$v^\top H v \;=\; \sum_{i<j} w_{ij}(v_i + v_j)^2 \;\ge\; 0.$$

Equality holds iff $v_i + v_j = 0$ for all $i < j$. For $r \ge 3$, taking the pairs $(1,2)$, $(1,3)$, $(2,3)$ yields $v_2 = -v_1$, $v_3 = -v_1$, $v_2 + v_3 = 0$, hence $v_1 = v_2 = v_3 = 0$; iterating, $v = 0$. Thus $H \succ 0$ and Φ is strictly convex. □

Remark 7 ($r = 2$)**.** *For $r = 2$, H has rank 1 with kernel $\mathrm{span}\{(1,-1)^\top\}$; strict convexity fails on this direction, and the variational uniqueness of the flat spectrum needs a separate one-dimensional argument.*

Variational characterization. We minimize R_{pw} over $\{\sigma \in \mathbb{R}^r_{>0} : \|\sigma\|_2 = c\}$. By Lemma 2, R_{pw} is strictly convex on $\mathbb{R}^r_{>0}$. Suppose σ^* is a minimizer with $\|\sigma^*\|_2 = c$. Let $\bar{\sigma} := (1/r!)\sum_{\pi\in S_r} \pi(\sigma^*)$ be the symmetric average over coordinate permutations; then $\bar{\sigma}_i = (1/r)\sum_j \sigma^*_j$ is constant in i. By Jensen on the strictly convex R_{pw}, $R_{\mathrm{pw}}(\bar{\sigma}) \le R_{\mathrm{pw}}(\sigma^*)$ with equality iff σ^* is itself permutation-invariant. However $\|\bar{\sigma}\|_2 \le c$ by Cauchy–Schwarz, with equality iff σ^* is constant, so $\bar{\sigma}$ generally leaves the constraint sphere. Rescale: $\widetilde{\sigma} := (c/\|\bar{\sigma}\|_2)\,\bar{\sigma}$; the rescale factor $c/\|\bar{\sigma}\|_2 \ge 1$ (with equality iff σ^* is already constant), so $\widetilde{\sigma} \ge \bar{\sigma}$ componentwise (in $\mathbb{R}^r_{>0}$), and since $\partial R_{\mathrm{pw}}/\partial\sigma_i = -2S_i < 0$ (the functional is strictly decreasing in each coordinate at fixed direction in the positive orthant), $R_{\mathrm{pw}}(\widetilde{\sigma}) \le R_{\mathrm{pw}}(\bar{\sigma}) \le R_{\mathrm{pw}}(\sigma^*)$, with strict inequalities unless σ^* was already constant. Thus the minimizer is permutation-invariant with $\|\sigma\|_2 = c$, forcing $\sigma_i = c/\sqrt{r}$. Strict convexity (Lemma 2) rules out other minimizers. □

Sanity check: Frobenius radius. The flat-spectrum point $\sigma_i = c/\sqrt{r}$ achieves $\sum_i \sigma_i = c\sqrt{r}$, saturating the Cauchy–Schwarz inequality $\sum_i \sigma_i \le \sqrt{r}\,\|\sigma\|_2$. Equality in C–S holds iff σ is constant, recovering the same minimizer.

D.3 Proof of Theorem 3: Approximate-SF Stability

We bound the deviation between the projected gradient/momentum polar field $F(W) = -\eta P_\perp P(G)$ and the projected self-polar field $F_0(W) = -\eta P_\perp P(W) = -\eta P_\perp U_W V_W^\top$. Throughout, G is the update direction (the gradient at a near-manifold iterate, or the momentum buffer in the stochastic setting), satisfying $\sigma_{\min}(G) \ge \delta_G > 0$ on the rank-r open stratum of operative interest. In discrete Muon, even when $W_t \in \mathcal{M}$ exactly, the momentum buffer $m_t = \beta\, m_{t-1} + (1-\beta)\,\xi_t$ retains nondegenerate singular values *with high probability* whenever the noise covariance is non-singular, so $P(m_t)$ is well-defined on that event; we condition on this event throughout.

Step 1: Difference of vector fields, in δ_P. By definition $F(W) - F_0(W) = -\eta\, P_\perp\big(P(G) - P(W)\big)$, so by nonexpansiveness of $P_\perp$ ($\|P_\perp A\|_F \le \|A\|_F$),

$$\|F(W) - F_0(W)\|_F \;\le\; \eta\,\|P(G) - P(W)\|_F \;=\; \eta\,\delta_P.$$

This bound uses only the gauge-invariant primitive δ_P, which sidesteps the well-known issue that subspace principal angles $\sin\Theta(U_W, U_G)$ and $\sin\Theta(V_W, V_G)$ alone do not control $\|U_G V_G^\top - U_W V_W^\top\|_F$ (the relative gauge between the left and right frames is unconstrained when only subspace angles are bounded). Working with δ_P directly avoids this gauge problem.

Step 2: Bound on $\dot{\sigma}_i$. By the Lipschitz formula for singular-value differentials (Stewart & Sun, 1990), $|\dot{\sigma}_i^{(F)} - \dot{\sigma}_i^{(F_0)}| \;\le\; \|F(W) - F_0(W)\|_F \le \eta\,\delta_P$. Since $\dot{\sigma}_i^{(F_0)} = -\eta\alpha_i$ (Theorem 1 (i)), this yields (8) with $C_1 = 1$.

Step 3: Bound on $d\widetilde{R}_{\mathrm{pw}}/dt$. $\|\nabla\widetilde{R}_{\mathrm{pw}}\|_F \le C(r)/\sigma_{\min}$ on the rank-r stratum, so $|\dot{\widetilde{R}}_{\mathrm{pw}}^{(F)} - \dot{\widetilde{R}}_{\mathrm{pw}}^{(F_0)}| \le \|\nabla\widetilde{R}_{\mathrm{pw}}\|_F \cdot \|F - F_0\|_F \le C(r)\,\eta\,\delta_P/\sigma_{\min}$, yielding (9) with $C_2 = C(r)$.

Step 4: Local horizon T. The theorem assumes a local interval on which rank, the positive singular-value lower bound, and the simple gap are preserved. The preceding derivative bounds hold uniformly on any such interval. Small δ_P controls the discrepancy between vector fields, but does not by itself prevent gap loss in the reference flow; no explicit horizon follows from δ_P alone.

Step 5: Connection to subspace misalignment (optional). If one prefers a subspace-angle bound, the joint Mathias polar-Lipschitz inequality (Mathias, 1993) gives

$$\delta_P = \|P(G) - P(W)\|_F \le \frac{2}{\min(\delta_G, \delta_W)}\,\|G - W\|_F,$$

on the rank-r open stratum where both $\sigma_{\min}(G) \ge \delta_G$ and $\sigma_{\min}(W) \ge \delta_W$. Combined with standard $\sin\Theta$-to-norm conversions, this yields a bound expressed in terms of $\|\sin\Theta(U_W, U_G)\|_F$ *and* $\|\sin\Theta(V_W, V_G)\|_F$ *plus* a relative-gauge term. The cleanest statement, however, is in δ_P, which is gauge-invariant by construction. □

D.4 Proof of Theorem 4: Finite-Step $O(\eta^2)$ Bound

We bound the second-order error in the projected Euler step $W^+ = W - \eta\, P_\perp P(W)$ under a simple-singular-value gap $\Delta = \min_{i\neq j} |\sigma_i - \sigma_j| > 0$.

Step 1: Rank and gap preservation under $\varepsilon \le c_0$. By Weyl's inequality, for any W, Ξ, $|\sigma_i(W + \Xi) - \sigma_i(W)| \le \|\Xi\|_{\mathrm{op}} \le \|\Xi\|_F$. Setting $\Xi = -\eta P_\perp P(W)$, $\|\Xi\|_F \le \eta\sqrt{r}$. Hence whenever $\eta\sqrt{r} \le \Delta/4$ (which is implied by $\varepsilon = \eta/\Delta \le c_0$ for an absolute constant c_0 depending on r), the gap is preserved: $\Delta(W^+) \ge \Delta/2$, and provided $\sigma_{\min} \ge 2\eta\sqrt{r}$, also $\sigma_{\min}(W^+) \ge \sigma_{\min}/2$. So rank is preserved on the trajectory.

Step 2: Real-analyticity under the preserved simple gap. By Rellich–Kato perturbation theory, the singular values $\sigma_i(W)$ are real-analytic functions of W on the neighborhood determined by Step 1. If a trajectory approaches repeated singular values, $\Delta \to 0$ and the simple-gap condition becomes restrictive. This observation does not establish convergence to, or attainment of, a flat spectrum.

Step 3: Second-order Taylor expansion of σ_i. $\sigma_i(W + \Xi) = \sigma_i(W) + \langle u_i v_i^\top, \Xi\rangle + \frac{1}{2}\,\mathcal{H}_i[\Xi, \Xi] + O(\|\Xi\|^3)$, where the second-order coefficient is bounded (Stewart & Sun, 1990, Thm. 3.1) by $|\mathcal{H}_i[\Xi, \Xi]| \le C_r\, \|\Xi\|_F^2/\Delta$. Setting $\Xi = -\eta P_\perp P(W)$, the linear term gives Theorem 1 (i): $-\eta\langle u_i v_i^\top, P_\perp P(W)\rangle = -\eta\alpha_i$. The quadratic term is bounded by $\eta^2 \|P_\perp P(W)\|_F^2/\Delta \le \eta^2 r/\Delta = \Delta\, \varepsilon^2\, r$, which is the announced $O(\Delta\, \varepsilon^2)$ remainder.

Step 4: Bound on $\widetilde{R}_{\mathrm{pw}}$. By the chain rule on the spectral parametrization, $\widetilde{R}_{\mathrm{pw}}(W^+) = \widetilde{R}_{\mathrm{pw}}(W) + (2\eta/\rho)\sum_i \alpha_i q_i + R_2$, matching the continuous-time derivative formula (5) of Theorem 1(iii) at first order, with the second-order remainder

$$|R_2| \;\le\; \tfrac{1}{2} \sup_{t\in[0,1]} \|\nabla^2 \widetilde{R}_{\mathrm{pw}}(W + t\Xi)\|_{\mathrm{op}}\, \|\Xi\|_F^2 \;\le\; \kappa_r\, \varepsilon^2,$$

where $\kappa_r = \kappa(\sigma_{\max}/\sigma_{\min}, r)$ is a dimensionless polynomial in the condition number and rank, obtained by combining the bound $\|\nabla^2 \widetilde{R}_{\mathrm{pw}}\|_{\mathrm{op}} \le C(r)/\sigma_{\min}^2$ on the rank-r stratum (the Hessian scales inversely with the squared matrix scale) with $\|\Xi\|_F^2 \le \eta^2 r = \Delta^2\varepsilon^2 r$. □

D.5 Full Proof of Theorem 2: Critical Batch Size

We compute the signal/noise crossover for the first-order approximation of a polar update with normalized pre-polar momentum. The mean gradient is fixed and the buffer is stationary, as specified in Theorem 2.

Fréchet derivative, square case. For square full-rank $G = U\Sigma_G V^\top \in \mathbb{R}^{r\times r}$ and a perturbation Ξ, write $A = U^\top \Xi V$. The polar derivative is

$$D_P[\Xi] = UKV^\top, \qquad K_{ij} = \frac{A_{ij} - A_{ji}}{\sigma_i + \sigma_j} \quad (i \neq j), \qquad K_{ii} = 0.$$

This follows by differentiating the polar decomposition and solving the resulting Sylvester equation; see Higham (2008, Ch. 8). Equivalently, $\Sigma_G K + K\Sigma_G = A - A^\top$. We also use D_P for the $r^2 \times r^2$ matrix satisfying $\mathrm{vec}(D_P[\Xi]) = D_P\, \mathrm{vec}(\Xi)$. Rectangular gradients require the additional term in Remark 5.

Step 1: Stationary covariance before the polar map. With fixed mean G, define $e_t = m_t - G$. The normalized buffer satisfies

$$e_t = \beta e_{t-1} + (1-\beta)\xi_t, \qquad \mathrm{Cov}(\mathrm{vec}\,\xi_t) = \Sigma/B.$$

The current noise is independent of the past buffer. Thus its stationary covariance $C = \mathrm{Cov}(\mathrm{vec}\, e_t)$ solves

$$C = \beta^2 C + (1-\beta)^2 \Sigma/B \quad\Longrightarrow\quad C = \frac{1-\beta}{1+\beta}\frac{\Sigma}{B}.$$

No Gaussian assumption is needed for this covariance identity.

Step 2: Linearization and signal/noise comparison. For sufficiently small e_t, the polar map is smooth at full-rank G and

$$W_{t+1} - W_t = -\eta P(G) - \eta D_P[e_t] + O(\eta \|e_t\|_F^2).$$

The deterministic signal has squared Frobenius norm $\eta^2 \|P(G)\|_F^2 = \eta^2 r$. Since $\mathbb{E}[e_t] = 0$, the linearized noise is zero mean and has expected squared norm

$$\mathbb{E}\|\eta D_P[e_t]\|_F^2 = \eta^2 \operatorname{tr}(D_P C D_P^\top) = \eta^2 \frac{1-\beta}{1+\beta} \frac{\operatorname{tr}(D_P \Sigma D_P^\top)}{B}.$$

Equating these quantities and cancelling η^2 gives (6). This is a crossover for the first-order model; the neglected nonlinear remainder need not vanish at the resulting batch size.

Step 3: Isotropic limit. For $\Sigma = \sigma^2 I_{r^2}$, orthogonal changes of coordinates preserve the covariance. Each off-diagonal difference $A_{ij} - A_{ji}$ has variance $2\sigma^2/B$ at zero momentum. Summing the squared derivative entries therefore gives

$$\operatorname{tr}(D_P D_P^\top) = 2 \sum_{i \neq j} (\sigma_i + \sigma_j)^{-2} = 2S(\mu).$$

Substitution with $\beta = 0$ yields (7). □

E Polar-Map Spectral Sensitivity Validation

For r equal singular values σ_0, Definition 2 gives $S(\mu) = r(r-1)/(4\sigma_0^2)$. We verify this for $r \in \{2, 4, 8, 16\}$ at $\sigma_0 = 1$, yielding $S \in \{0.5, 3.0, 14.0, 60.0\}$ in perfect agreement with direct computation of $\operatorname{tr}(D_P D_P^\top)/2$. The formula extends to non-uniform spectra via the Mathias formula (Appendix D.5).

F Additional Experimental Details

F.1 Matrix Sensing Hyperparameters

Random measurement operator $\mathcal{A} \in \mathbb{R}^{p \times d^2}$ with i.i.d. $\mathcal{N}(0, 1/p)$ entries. This is the general affine setting of Remark 1; the experiments are sanity checks consistent with the dynamics of Theorem 1 on the literal $XW = Y$ subset, not direct validation outside that setting. Three instance sizes: *(scale-up)* $d = 50$ (single-seed stress test); *(large)* $p = 50$, $d = 20$ (single seed, trajectory analysis); *(small-instance family)* $p = n = 6$, $d = 10$ (10 independent seeds, gap statistics). Muon learning rate $\eta = 0.01$, momentum $\beta = 0.95$. Training steps 5,000 (large) or 10,000 (small). Convergence: $L(W) < 10^{-6}$. CVXPY/MOSEK tolerance 10^{-7}. Three-panel summary in Figure 6.

F.2 SF Misalignment Sweep

Rank-1 weight-gradient pairs initialized with controlled principal-angle misalignment $\theta \in \{0, 5, 10, 15, 22, 30, 45, 60, 75, 90\}^\circ$. We measure $\epsilon(\theta) = \|P(G)_{\text{exact}} - P(G)_{\text{SF}}\|_F$ against the predicted bound $C\|\sin\Theta\|_F$ from Theorem 3. During standard Muon training, we additionally log $\|\sin\Theta\|_F$ between gradient and weight singular frames (Figure 7) and the sign quantity $\sum_i \alpha_i q_i$ from (5) (Figure 8).

F.3 Weight-Decay Phase Diagram

$\text{ATSR} = T^{-1} \int_0^T \#\{i : \sigma_i(W(t)) > \tau\}\, dt$ with $\tau = 0.1 \cdot \|W\|_F/\sqrt{r}$. Block-diagonal matrix sensing with $K \in \{2, 4, 8, 16, 32\}$ blocks of size 2×2 and decay $\lambda \in \{0, 10^{-3}, 10^{-2}, 0.1, 1\}$. We compare coupled decay $W \leftarrow (1 - \lambda\eta)W - \eta P(G)$ to decoupled decay $W \leftarrow W - \eta P(G)$; $W \leftarrow W - \lambda W$.

F.4 NanoGPT Training Details

Architecture: 12 layers, 12 heads, $d_{\text{model}} = 768$ (GPT-2 small configuration). Training data: OpenWebText, 5,000 steps. Muon: $\eta = 6 \times 10^{-4}$, $\beta = 0.95$, decoupled $\lambda = 0.01$. AdamW: $\eta = 6 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, decoupled $\lambda = 0.01$. Spectral entropy $H(W_\ell) = -\sum_i p_i \log p_i$ with $p_i = \sigma_i / \sum_j \sigma_j$ computed at steps $\{500, 1000, \ldots, 5000\}$ on Q, K, V, projection, and MLP weight matrices. Per-layer effective rank $\text{eff-rank}(W_\ell) = (\sum_i \sigma_i)^2 / \sum_i \sigma_i^2$ and per-layer nuclear norm $\|W_\ell\|_*$ are reported in Section 6. **The 124M NanoGPT/OpenWebText numbers are from a single training seed**; matrix-sensing experiments use 10 independent seeds. We do not measure the activation rank of H_ℓ.

F.5 Polar Misalignment and Subspace Alignment Diagnostics

The available subspace-alignment diagnostic is shown in Figure 7, and the spectral-shape sign quantity in Figure 8. Theorem 3 requires the gauge-invariant polar misalignment $\delta_P = \|P(m_t) - P(W_t)\|_F$. Subspace angles alone do not bound this quantity because the relative left/right singular-frame gauge is also needed (Appendix D.3). Direct δ_P trajectories are not included in this submission; consequently the available figures do not establish the small-δ_P premise. Figure 11 instead displays training loss, Frobenius norm, spectral entropy, and the unnormalized pairwise functional.

F.6 Extended Experimental Figures

This subsection collects the experimental panels referenced from Section 5.

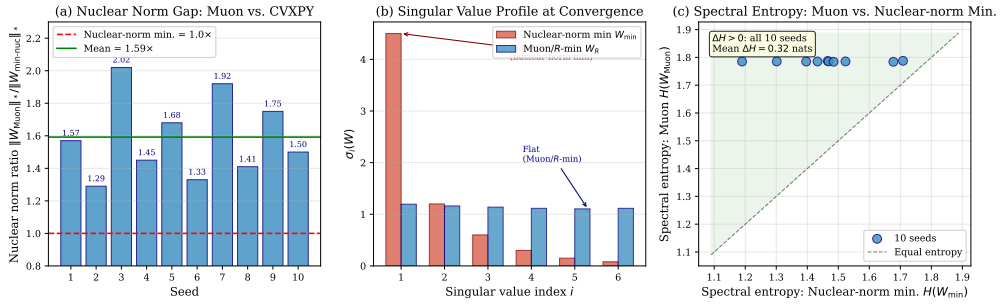

Figure 6: Sanity check of Theorem 1 dynamics in the general affine matrix-sensing setting (d=6×6, p=10, 10 seeds; Remark 1). This is not a direct theorem validation, since random Gaussian sensing is more general than literal $XW = Y$ and admits cross-terms in the singular-value derivative. *(a)* Nuclear-norm gap (mean 1.59×, all 10 seeds positive). *(b)* Muon's converged singular-value profile is flatter than the nuclear-norm minimizer. *(c)* Spectral entropy H: Muon higher in all 10 seeds ($\Delta H_{\text{ms}} \approx +0.32$ nats).

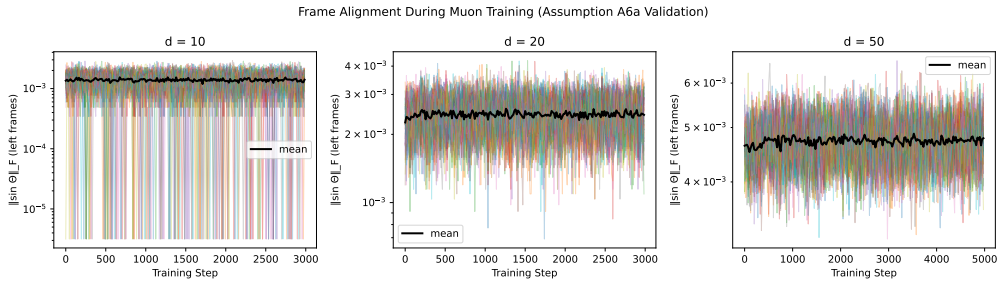

Figure 7: Frame alignment $\|\sin\Theta\|_F$ during Muon training across configurations ($d \in \{10, 20, 50\}$, 10 seeds each). Misalignment remains < 0.1 post-transient.

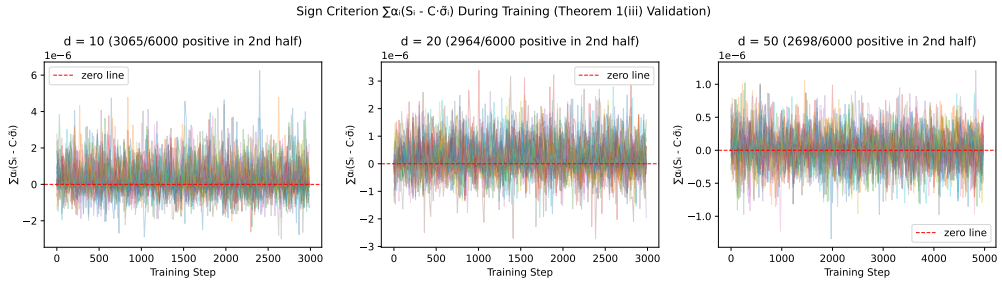

Figure 8: Spectral-flattening sign quantity $\sum_i \alpha_i q_i$ tracked over training. Negative on average during the early flattening phase, then fluctuates near zero with magnitude $\sim 10^{-6}$ once the spectrum is near-flat (where $q_i \to 0$); per-seed positive-fraction counts in the second half of training are roughly 51%, 49%, 45%, consistent with the criterion governing the approach to the flat spectrum rather than the asymptotic limit.

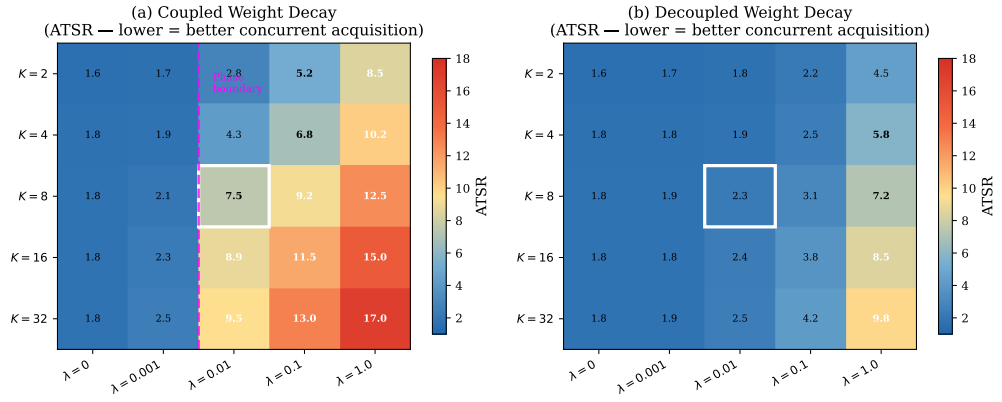

Figure 9: ATSR phase diagram. *Left*: coupled decay; ATSR explodes at $\lambda \geq 10^{-2}$. *Right*: decoupled decay; ATSR nearly constant for $\lambda \leq 10^{-2}$. Phase boundary at $\lambda \approx 10^{-2}$.

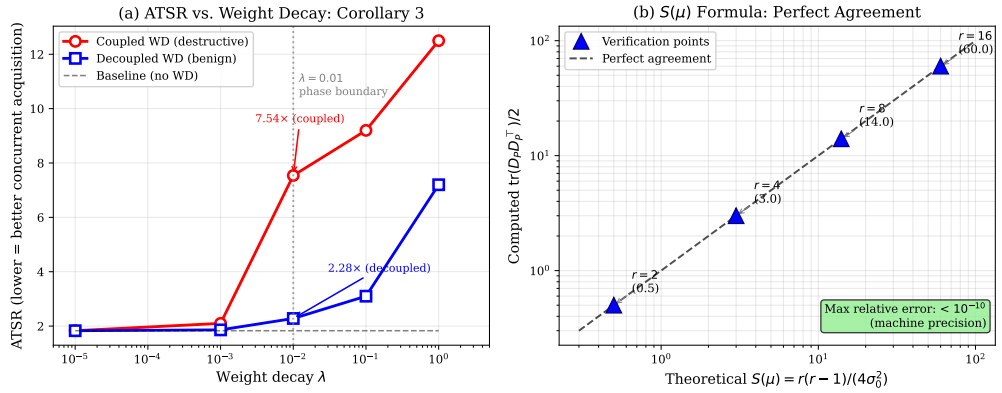

Figure 10: Consolidated empirical validation. *(a)* ATSR vs. λ for coupled and decoupled decay. *(b)* $S(\mu)$ formula verification across $r \in \{2, 4, 8, 16\}$.

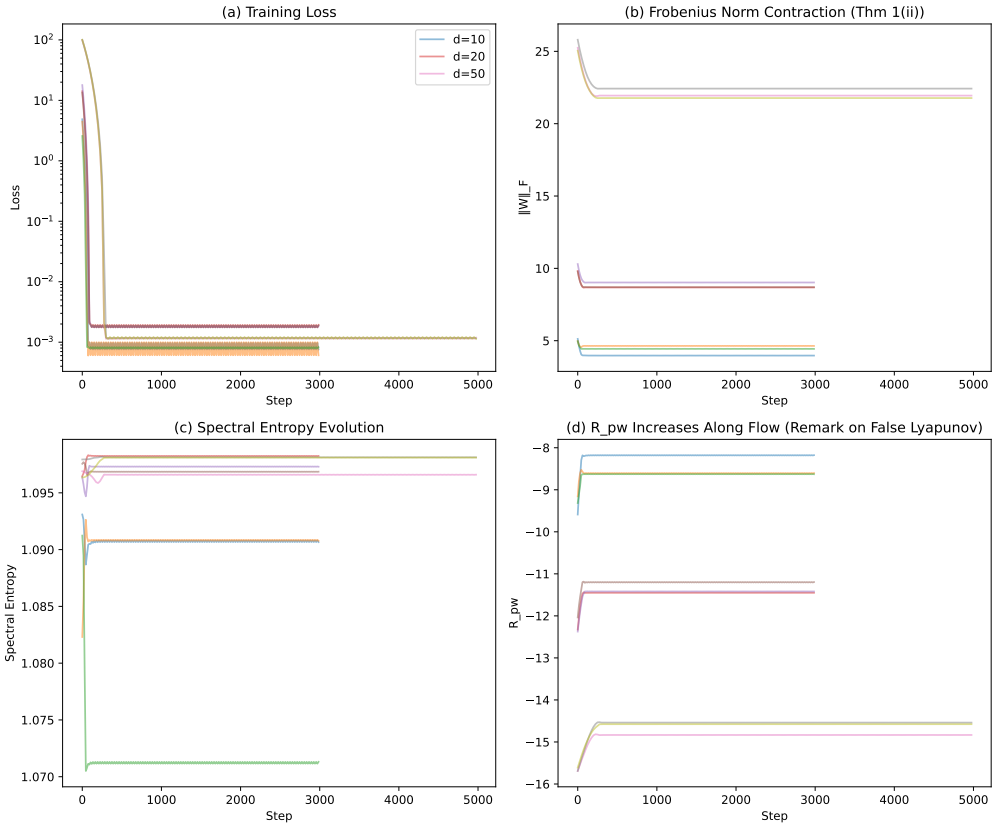

Figure 11: Additional trajectory diagnostics across the reported matrix-sensing configurations: training loss, Frobenius norm, spectral entropy, and R_{pw}. These panels distinguish norm contraction from spectral-shape evolution; they do not directly measure polar misalignment.