BEYOND HESSIAN SHARPNESS: TWO LOCAL STABILITY EDGES IN PRACTICAL MUON

Anonymous authors
Paper under double-blind review

ABSTRACT

For gradient descent, a Hessian direction is stable when $0 < \alpha\lambda < 2$, and full batch training can run near this Edge of Stability. We ask what determines local stability for practical Muon with its five Newton–Schulz steps and stored momentum. Linearizing one update in both weights and momentum gives two multipliers for each eigenvalue of $\alpha\mathbf{RH}$, where $\mathbf{H}$ is the batch Hessian and $\mathbf{R}$ is the derivative of Muon's matrix map. A real mode is stable between a *lower edge*, beyond which it grows without changing sign, and an *upper edge*, beyond which it grows while flipping sign. Gradient descent with weight decay has the same two edges, but positive semidefinite curvature cannot take it past the lower one. Muon's finite matrix map can reverse a perturbation and can reach that region even under positive semidefinite curvature. Rescaling stored momentum changes stability while preserving the next weight. In 58M to 1.25B language models, all 240 sampled matrix–checkpoint pairs have negative derivative eigenvalues. Gauss–Newton tests find modes beyond the lower edge in all 14 model-state–batch cases, and momentum interventions move selected modes across both edges. Along a 3.33M training continuation with changing batches, resolved lower modes lie beyond their edge as validation loss falls. When one text batch is repeated, resolved upper modes lie beyond their edge at sampled states. Yet measured perturbations grow only 13 to 15 times over 63 updates, while repeatedly applying the first Jacobian predicts more than 10^7. Thus a boundary for one update does not determine growth along a changing training trajectory.

1 INTRODUCTION

For gradient descent on a fixed batch, a perturbation along a Hessian eigenvector is multiplied by $1 - \alpha\lambda$ at each step. It shrinks when $0 < \alpha\lambda < 2$. In full batch neural network training, the largest Hessian eigenvalue often approaches or slightly exceeds $2/\alpha$ while loss falls over longer intervals (Cohen et al., 2021). This is the Edge of Stability. Curvature change and oscillation help explain this behavior (Ahn et al., 2022; Arora et al., 2022; Damian et al., 2023; Chen & Bruna, 2023). We ask what replaces Hessian sharpness for Muon, whether training stays near the resulting edges, and whether local expansion predicts perturbation growth over multiple updates.

Muon turns each matrix's momentum input into an approximately orthogonal update using five Newton–Schulz steps (Jordan et al., 2024). The steps use matrix multiplication instead of an SVD. Their polynomial brings small singular values toward one quickly, but the output is not monotone in the input singular value. A smaller input can therefore produce a larger output in some directions. That reversal matters when a small change in weights changes the next update. We retain the original coefficients, Nesterov momentum, shape factor, and the two normalization rules used by the reference and PyTorch implementations.[1]

The intuition is simple. Muon forms a matrix $\mathbf{P}$, the *proposal*, from the gradient and stored momentum. The value $\Psi(\mathbf{P})$ sets the weight step. Its derivative tells us how the step responds when the proposal changes. A weight perturbation changes the gradient through the Hessian, and the stored momentum carries part of this change into the next step. In some directions, the finite matrix map reverses the proposal perturbation. The resulting weight step can amplify the original perturbation without

[1] Code snapshots: `Muon reference code` and `torch.optim.Muon`.

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.

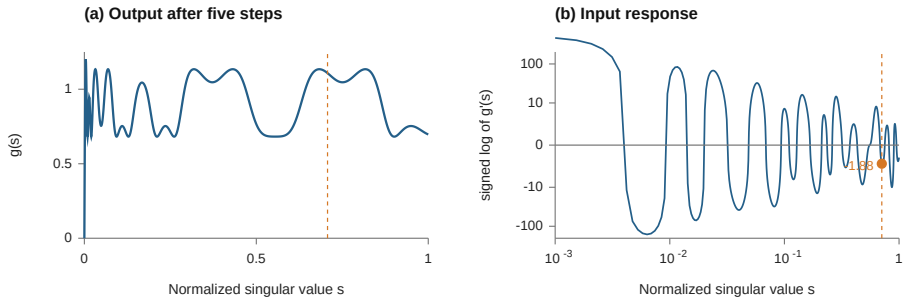


Figure 1: $g(s)$ is the output of Muon's five polynomial steps for a normalized singular value s. Its derivative changes sign even though $g(s) > 0$ for $s > 0$. Panel (b) uses a logarithmic horizontal axis from 10^{-3} to 1 and displays $\operatorname{sgn}(g') \log_{10}(1 + |g'|)$ vertically. The orange marker shows $g'(1/\sqrt{2}) \approx -1.88$, used in Section 3.

flipping its sign, even when curvature is positive. We call this calculation *local* because it examines small perturbations near a specified weight and momentum state on a fixed batch. During training the state changes, so we also study successive Jacobians.

Contributions. Our main contributions are:

- **Two stability edges.** We derive the joint weight and momentum stability equation (Theorem 1). A real eigenvalue κ of $\alpha \mathrm{D}\Psi(\mathbf{P})\mathbf{H}$ is stable between a lower edge, beyond which a perturbation grows without changing sign, and an upper edge, beyond which it grows while alternating sign. Gradient descent with weight decay has the same two edges. Positive semidefinite curvature cannot take it or the exact polar map past the lower edge, while practical Muon can.
- **The next weight does not determine stability.** Rescaling stored momentum keeps the next weight fixed while moving a mode across either edge (Theorem 3). At a fixed state, a smaller learning rate cannot restore a mode beyond the lower edge. Two strongly convex quadratic examples show that crossing an edge can create attracting fixed points or an attracting orbit.
- **Measurements in language models.** GGN and output-matched controls isolate the matrix map's role in negative modes; momentum interventions verify both crossings. Training measurements show that a local boundary need not attract the trajectory. Paired-run responses instead agree with successive Jacobians, exposing the limits of extrapolating one local multiplier.

Related work. Existing work studies several aspects of Muon's stability and dynamics. Generalized sharpness describes training near an Edge of Stability for non-Euclidean gradient descent, including spectral descent and Muon without momentum (Islamov et al., 2026). Exact polar models study trajectories and sensitivity to perturbations (Rhee et al., 2026; Gonon et al., 2026; Beneventano et al., 2026). Other work bounds optimization error from finite polynomial updates or changes the spectral map (Do et al., 2026; Li & Tsuchiya, 2026; Liu et al., 2026a). Here the object is the one-step Jacobian of the original five-step Muon update with its stored momentum. Appendix A compares the formulas and assumptions in Table 2.

2 Practical Muon

Consider a matrix weight $\mathbf{W} \in \mathbb{R}^{m \times n}$ and its stored momentum $\mathbf{M}$. We hold the batch $\mathcal{B}$ and all random choices fixed when differentiating an update. Write

$$\mathbf{G}(\mathbf{W}) = \nabla_{\mathbf{W}} \mathcal{L}_{\mathcal{B}}(\mathbf{W}), \qquad \mathbf{H} = \mathrm{D}\mathbf{G}(\mathbf{W}).$$

Let $\alpha > 0$ be the learning rate, $0 \le \beta < 1$ the momentum coefficient, and $\lambda_{\mathrm{wd}} \ge 0$ the decoupled weight decay coefficient (Loshchilov & Hutter, 2019). The decay per step is $d = \alpha \lambda_{\mathrm{wd}}$.

Algorithm 1: One practical Muon update for a matrix block

Gradient: $\mathbf{G} \leftarrow \nabla_{\mathbf{W}} \mathcal{L}_{\mathcal{B}}(\mathbf{W})$
Updated momentum: $\mathbf{M}^+ \leftarrow \beta\mathbf{M} + (1-\beta)\mathbf{G}$
Nesterov proposal: $\mathbf{P} \leftarrow \beta\mathbf{M}^+ + (1-\beta)\mathbf{G}$
Matrix transform: $\mathbf{\Delta} \leftarrow \Psi(\mathbf{P})$ `// Normalization and five Newton--Schulz steps`
Weight decay and update: $\mathbf{W}^+ \leftarrow (1-d)\mathbf{W} - \alpha\mathbf{\Delta}$

We call $\mathbf{P} = \beta^2\mathbf{M} + (1-\beta^2)\mathbf{G}(\mathbf{W})$ the *proposal*. It combines the current gradient with the stored momentum. For a full rank proposal with thin singular value decomposition $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, the exact polar factor $\mathbf{U}\mathbf{V}^\top$ replaces its singular values by one (Higham, 2008, Ch. 8). It is the nearest matrix with orthonormal columns or rows in Frobenius norm (Higham, 1986). Applied to a gradient, its negative gives the steepest linearized decrease among steps of unit spectral norm (Bernstein & Newhouse, 2024).

Practical Muon normalizes $\mathbf{P}$ and applies five polynomial iterations. We use the reference and PyTorch configurations linked in the Introduction: Nesterov momentum, coefficients $(\gamma_1, \gamma_3, \gamma_5) = (3.4445, -4.7750, 2.0315)$, and shape factor $\chi_{m,n} = \sqrt{\max\{1, m/n\}}$. The shape factor adjusts update RMS for rectangular matrices. The mathematical analysis uses the smooth exact-arithmetic map.

The two implementations use different normalization rules:

$$\nu_\delta(\mathbf{P}) = \|\mathbf{P}\|_{\mathrm{F}} + \delta, \qquad \nu_{\mathrm{cl}}(\mathbf{P}) = \max\{\|\mathbf{P}\|_{\mathrm{F}}, \epsilon_{\mathrm{cl}}\}, \qquad \epsilon_{\mathrm{cl}} = 10^{-7}. \tag{1}$$

The reference implementation uses $\delta = 10^{-7}$, and the PyTorch snapshot uses the clamp rule. For either normalization ν, define

$$\begin{aligned} \mathbf{X}_0 &= \mathbf{P}/\nu(\mathbf{P}), \\ \mathbf{X}_{j+1} &= \gamma_1\mathbf{X}_j + \gamma_3(\mathbf{X}_j\mathbf{X}_j^\top)\mathbf{X}_j + \gamma_5(\mathbf{X}_j\mathbf{X}_j^\top)^2\mathbf{X}_j, \qquad j = 0, \dots, 4, \\ \Psi_\nu(\mathbf{P}) &= \chi_{m,n}\mathbf{X}_5. \end{aligned} \tag{2}$$

For $m > n$, the implementation applies the recurrence to $\mathbf{P}^\top$ and transposes the result. We write Ψ_δ and Ψ_{cl} for the two transformations. Setting $\delta = 0$ defines Ψ_0 at nonzero proposals. Table 3 in Appendix B lists which map each result uses. Write $\phi(x) = \gamma_1 x + \gamma_3 x^3 + \gamma_5 x^5$ and $g = \phi^{\circ 5}$. The five iterations map each normalized singular value s to $g(s)$, as derived in Appendix C. Primes denote scalar derivatives, and D denotes Fréchet derivatives.

Throughout the theory, vectors, matrices, and linear operators are bold, scalar entries are not, and $\mathbf{I}$ is the identity. The appendix opens with a notation table. Muon carries $\mathbf{M}^+$ into the next step and computes $\mathbf{P}$ within the current step. We therefore differentiate $(\mathbf{W}, \mathbf{M}) \mapsto (\mathbf{W}^+, \mathbf{M}^+)$. Define

$$\mathbf{R} = \mathrm{D}\Psi(\mathbf{P}), \qquad \mathbf{K} = \alpha\mathbf{R}\mathbf{H}.$$

A weight perturbation changes the gradient through $\mathbf{H}$ and the transformed proposal through $\mathbf{R}$. For several Muon matrices, $\mathbf{W}$ and $\mathbf{M}$ collect all their weights and momentum buffers. The derivative $\mathbf{R}$ is block diagonal, while $\mathbf{H}$ includes curvature within and across these matrices. If the blocks share α, β, and λ_{wd}, Theorem 1 applies to the combined operator.

3 Local Stability of Practical Muon

The momentum rule converts each eigenvalue of $\mathbf{K}$ into two state multipliers. We first derive this relation, then show how finite Newton–Schulz steps create negative modes with positive definite curvature. Changing stored momentum can preserve the next weight while moving these modes. Finally, two quadratic examples describe the nonlinear behavior after a crossing.

3.1 Two Stability Edges for Weights and Momentum

Let $\mathbf{A}$ be the Jacobian of one update on $(\mathbf{W}, \mathbf{M})$. A perturbation in an eigendirection is multiplied to first order by the corresponding eigenvalue μ, which we call a *multiplier*. The Jacobian is Schur stable when all multipliers lie inside the unit circle. At a fixed point on a fixed batch, this implies local

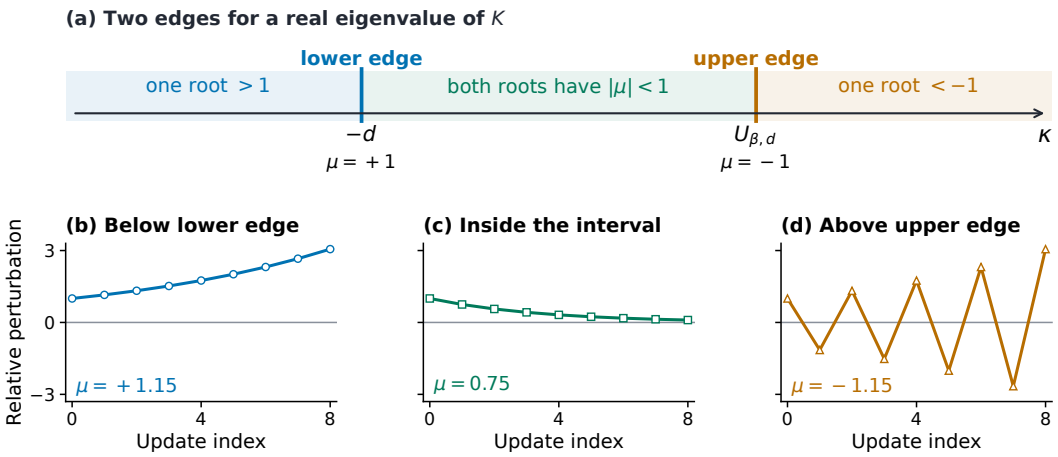


Figure 2: What the two edges mean. (a) The stable interval for a real eigenvalue κ of $\mathbf{K}$ is shown schematically. Both associated multipliers must have modulus below one. (b–d) For $\beta = 0.5$ and $d = 0.1$, these examples plot μ^t along one eigenvector of a fixed Jacobian. Below the lower edge, a mode grows without alternating its sign. Inside the interval, it decays. Above the upper edge, a mode grows while alternating its sign.

exponential stability (Appendix B). Along training, perturbations follow the product of Jacobians at successive states.

For perturbations of the current weight and stored momentum, Algorithm 1 gives, to first order,

$$\begin{aligned}\Delta\mathbf{M}^+ &= \beta\Delta\mathbf{M} + (1-\beta)\mathbf{H}\Delta\mathbf{W},\\ \Delta\mathbf{P} &= \beta\Delta\mathbf{M}^+ + (1-\beta)\mathbf{H}\Delta\mathbf{W} = \beta^2\Delta\mathbf{M} + (1-\beta^2)\mathbf{H}\Delta\mathbf{W},\\ \Delta\mathbf{W}^+ &= (1-d)\Delta\mathbf{W} - \alpha\mathbf{R}\Delta\mathbf{P}.\end{aligned}$$

These relations give the state Jacobian $\mathbf{A}$. Appendix D.1 writes its blocks and derives the following equation.

Theorem 1 (Stability boundary for weights and momentum)**.** *Fix a batch and a state where* $\mathbf{G}$ *and* Ψ *are differentiable. Assume* $\alpha > 0$*,* $0 \le \beta < 1$*, and* $0 \le d < 2$*. Each eigenvalue* κ *of* $\mathbf{K}$*, counted with algebraic multiplicity, gives two eigenvalues of* $\mathbf{A}$ *through*

$$(\mu - \beta)(\mu - 1 + d) + (1-\beta)((1+\beta)\mu - \beta)\kappa = 0. \tag{3}$$

For real κ*, both roots have modulus below one exactly when*

$$-d < \kappa < U_{\beta,d}, \qquad U_{\beta,d} = \frac{(1+\beta)(2-d)}{(1-\beta)(1+2\beta)}. \tag{4}$$

We call $\kappa = -d$ *the* lower edge*, where one multiplier is* $+1$*. We call* $\kappa = U_{\beta,d}$ *the* upper edge*, where one multiplier is* -1*.*

An eigenvalue κ of $\mathbf{K}$ and its associated perturbation define a *mode*. A real mode lies beyond the lower edge when $\kappa < -d$, and beyond the upper edge when $\kappa > U_{\beta,d}$.

At the lower edge, one update leaves the critical perturbation unchanged. At the upper edge, one update flips its sign. Figure 2 shows growth and decay on either side. The complete Jacobian is Schur stable when every eigenvalue of $\mathbf{K}$ gives two roots inside the unit disk. Complex eigenvalues also obey Equation 3, and Appendix D gives their stable region. The language model searches below select real modes, to which the two edges apply directly.

For gradient descent with the same decay rule, a Hessian eigenvalue λ has multiplier $1 - d - \alpha\lambda$. Its stable interval is $-d < \alpha\lambda < 2 - d$. Gradient descent with weight decay therefore also has two edges. If $\mathbf{H} \succeq 0$, its modes cannot lie beyond the lower one. The exact polar derivative has the same property under positive semidefinite curvature (Lemma 7). Muon's finite polynomial can reverse a proposal perturbation and produce a negative mode. With $d = 0$ and $\mathbf{R} = \mathbf{I}$, Theorem 1 also gives the known exponential moving average Nesterov threshold $\alpha\lambda < 2(1+\beta)/((1-\beta)(1+2\beta))$ (Lee & Lee, 2026).

3.2 How the Muon Map Reaches the Lower Edge

Positive definite curvature can give a negative eigenvalue of $\mathbf{K}$ when the Muon's matrix map reverses a proposal perturbation. An isotropic proposal lets us see this directly. Appendix C derives the full derivative from the singular value map formula (Noferini, 2017) and the normalization derivative.

A 2×2 example shows the sign change directly. Let $\mathbf{P} = \sigma\mathbf{I}$ with $\sigma > 0$ lie above the clamp threshold, and take $\mathbf{E} = \mathrm{diag}(1, -1)$. This perturbation increases one diagonal entry and decreases the other. For small t,

$$\frac{\sigma\mathbf{I} + t\mathbf{E}}{\|\sigma\mathbf{I} + t\mathbf{E}\|_{\mathrm{F}}} = \frac{\mathbf{I}}{\sqrt{2}} + \frac{t\mathbf{E}}{\sqrt{2}\sigma} + O(t^2), \qquad \mathbf{R}[\mathbf{E}] = \frac{g'(1/\sqrt{2})}{\sqrt{2}\sigma}\mathbf{E}.$$

The five Newton–Schulz steps satisfy

$$g(1/\sqrt{2}) > 0, \qquad g'(1/\sqrt{2}) \approx -1.88.$$

The update still points along the proposal because $g(1/\sqrt{2}) > 0$, but its derivative sends this perturbation in the opposite direction. At this isotropic proposal $\mathbf{R}$ is self-adjoint. With $\mathbf{H} \succ 0$, $\alpha\mathbf{RH}$ has as many negative eigenvalues as $\mathbf{R}$ (Corollary 5).

Proposition 2 (When positive curvature can reach the lower edge)**.** *Let $\mathbf{R}$ be a real linear map and $\alpha > 0$. A positive semidefinite $\mathbf{H}$ can make $\alpha\mathbf{RH}$ have a negative real eigenvalue if and only if some $\mathbf{E}$ satisfies $\langle \mathbf{E}, \mathbf{R}[\mathbf{E}] \rangle_{\mathrm{F}} < 0$. If the quadratic form is nonnegative in every direction, all eigenvalues of $\alpha\mathbf{RH}$ have nonnegative real part for every $\mathbf{H} \succeq 0$.*

The criterion concerns the response to a perturbation, not just the eigenvalues of $\mathbf{R}$. A negative eigenvalue of $\mathbf{R}$ gives a negative quadratic direction, but the measured curvature must couple to that direction. At a general proposal $\mathbf{R}$ need not be self-adjoint. The exact polar derivative is self-adjoint and positive semidefinite (Lemma 7), so it cannot produce this lower-edge mechanism under positive semidefinite curvature. Appendix C.6 proves the criterion and explains how normalization can interact with a nonlinear polynomial even when the polynomial is monotone. Figure 1(b) shows that the implemented polynomial also has negative scalar derivatives at normalized singular values near those in trained matrices. Section 4.1 tests the combined derivative under positive semidefinite Gauss–Newton curvature.

3.3 The Same Weight Step Can Have Different Stability

A map is *scale invariant* if $\Psi(c\mathbf{Q}) = \Psi(\mathbf{Q})$ for every $\mathbf{Q} \neq 0$ and $c > 0$. The same transformed proposal then gives the same weight step. A perturbation of fixed size has less effect when the proposal is larger. Differentiating the identity gives

$$\mathrm{D}\Psi(c\mathbf{P}) = c^{-1}\mathrm{D}\Psi(\mathbf{P}).$$

Theorem 3 (Proposal scaling preserves the next weight)**.** *Fix a batch, $\mathbf{W}$, a nonzero proposal $\mathbf{P}$, $\alpha > 0$, $0 < \beta < 1$, and $0 \leq d < 2$. Suppose $\mathbf{G}$ is differentiable at $\mathbf{W}$ and Ψ is differentiable at every nonzero matrix and satisfies $\Psi(c\mathbf{Q}) = \Psi(\mathbf{Q})$ for all $c > 0$ and $\mathbf{Q} \neq 0$. Define*

$$\mathbf{M}(c) = \frac{c\mathbf{P} - (1 - \beta^2)\mathbf{G}(\mathbf{W})}{\beta^2}, \qquad c > 0. \tag{5}$$

All states $(\mathbf{W}, \mathbf{M}(c))$ have the same loss, gradient, Hessian, weight step, and next weight, while $\mathbf{K}(c) = \mathbf{K}(1)/c$. If $\mathbf{K}(1)$ has a nonzero eigenvalue, $\mathbf{A}(c)$ has an eigenvalue outside the unit circle for all sufficiently small c. If $0 < d < 2$, $\mathbf{A}(c)$ is Schur stable for all sufficiently large c.

The transformation Ψ_0 is scale invariant. The clamp rule has this property above its threshold. Scale invariance also gives $\mathbf{R}[\mathbf{P}] = 0$, so $\mathbf{K}$ has a zero eigenvalue. Without weight decay, that mode has a multiplier $+1$ and strict stability of the full state is impossible. The additive offset 10^{-7} only approximately preserves scale invariance when $\|\mathbf{P}\|_{\mathrm{F}}$ is much larger; Section 4 checks its effect. Appendix E proves the scaling statement.

At a fixed state and batch, $\mathbf{R}$ and $\mathbf{H}$ do not depend on α. Dividing the stable interval by α gives $-\lambda_{\mathrm{wd}} < \kappa/\alpha < U_{\beta,d}/\alpha$. A smaller learning rate raises the upper edge in these units, as for gradient descent. It moves a mode beyond the lower edge closer to multiplier $+1$ but cannot return it inside.

3.4 Nonlinear Dynamics Beyond the Two Edges

Section 4.3 finds training states beyond the local edges. An edge describes the linearized update at one state, so we also ask what the full update does after a crossing. Muon's normalized matrix step makes its optimizer map nonlinear even on a quadratic loss. Two 2×2 examples have attracting states after crossing, with eight weight and momentum coordinates in the full state.

At a fixed point, $\mathbf{M}_\star = \mathbf{P}_\star = \mathbf{G}(\mathbf{W}_\star)$. The fixed-point equation and its derivative are

$$\lambda_{\mathrm{wd}}\mathbf{W}_\star + \Psi(\mathbf{G}(\mathbf{W}_\star)) = 0, \qquad \mathrm{D}\big(\lambda_{\mathrm{wd}}\mathbf{W} + \Psi(\mathbf{G}(\mathbf{W}))\big) = (d\mathbf{I} + \mathbf{K}_\star)/\alpha.$$

At the lower edge this derivative becomes singular. The same 2×2 isotropic proposal used above then gives a pitchfork: one fixed point gives way to two attracting ones (Theorem 9). A second quadratic gives a flip, or period doubling, to an attracting orbit of period two at the upper edge (Theorem 11). Both statements establish stability in every weight and momentum direction. The lower construction uses the scale invariant part of clamp normalization. The upper construction uses $\|\mathbf{P}\|_{\mathrm{F}} < 10^{-7}$, where the clamp denominator is constant. Appendices F and G prove the results, and Appendix J.8 checks both constructions numerically, including the additive normalization rule.

The lower example needs a negative mode under positive definite curvature, which the exact polar map cannot supply. The attracting states persist under sufficiently small smooth changes to the loss (Proposition 12).

Sophia and Lion. Appendix I derives the Sophia and Lion comparisons in Table 2. Lion also has an attracting orbit of period two while its single step multipliers lie inside the unit disk. Thus oscillation alone does not establish an upper edge crossing.

4 Experiments

We test three claims in language models trained on FineWeb-Edu: the matrix map can create negative modes under positive semidefinite curvature, momentum can change stability at the same next weight, and local multipliers need not predict trajectory growth (Penedo et al., 2024; Lozhkov et al., 2024). Original training uses BF16 forward operations; derivative measurements and controlled continuations use the smooth FP64 update.[2] A *conditional Muon* calculation fixes non-Muon parameters; a *resolved* mode passes the numerical checks in Appendix J.

4.1 Negative Modes with Positive Semidefinite Curvature

To isolate the matrix map, we replace the Hessian with positive semidefinite generalized Gauss–Newton (GGN) curvature by linearizing the logits at saved weights $\mathbf{W}_0$:

$$\widetilde{\mathcal{L}}(\mathbf{W}_0 + \mathbf{\Delta}) = \ell(f(\mathbf{W}_0) + \mathbf{J}_0\mathbf{\Delta}), \qquad \mathbf{J}_0 = \mathrm{D}f(\mathbf{W}_0). \tag{6}$$

This preserves the gradient, proposal, and next optimizer state (Martens, 2020, Sec. 8). All 14 model-state–batch cases have a mode beyond the lower edge: 12 at 3.33M and 7.25M (Table 7), and two at 58M. The latter give $\kappa = -0.196, -0.071$ and measured multipliers $1.085, 1.042$ (Appendix J.5).

An output-matched control replaces the five-step map with

$$\Psi_{\mathrm{ctrl}}(\mathbf{Q}) = \Psi_{10^{-7}}(\mathbf{P}_0) + N(\mathbf{Q}) - N(\mathbf{P}_0), \qquad N(\mathbf{Q}) = \frac{\chi_{m,n}\mathbf{Q}}{\|\mathbf{Q}\|_{\mathrm{F}} + 10^{-7}}. \tag{7}$$

At the reference proposal $\mathbf{P}_0$, this preserves both next weights and momentum while replacing the derivative by $\mathrm{D}N$. Its positive quadratic form rules out negative GGN modes (Proposition 2). Figure 3 shows response alignment changing from negative to positive on the same tested directions. Thus the finite matrix map can produce lower-edge modes without negative loss curvature.

[2] The 3.33M and 7.25M training runs store parameters and moments in FP32; the 58M–1.25B runs use FP32 accumulation. Across the 240 matrix–checkpoint pairs in Appendix J.4, BF16 and FP64 forward outputs differ by 1.9%–11.7% in relative norm. Native BF16 finite differences in Appendix J.11 fail the two-amplitude criterion, so the smooth responses do not establish native-precision perturbation dynamics.

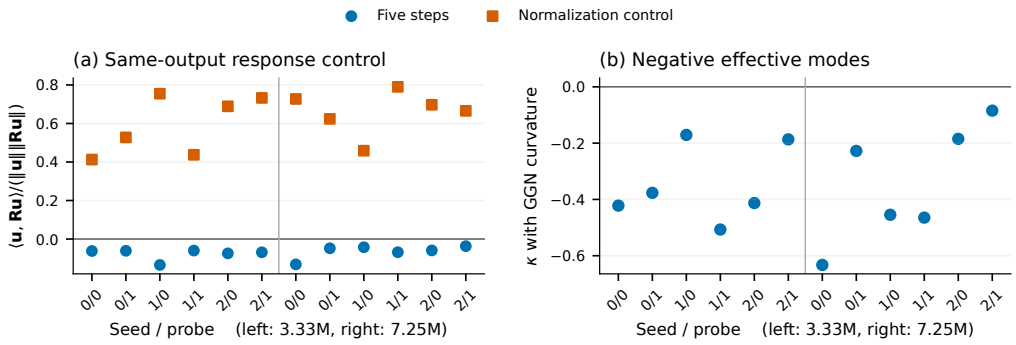


Figure 3: The matrix map reverses perturbations despite positive semidefinite curvature. (a) Response alignment on the same directions is negative for the five-step map and positive for the output-matched normalization control. (b) All twelve seed–probe cases have negative κ under GGN curvature. Each model uses three seeds and two probes per seed; Appendix J.6 gives the checks.

The mechanism is present across model sizes. A survey of 58M–1.25B models finds negative derivative eigenvalues in all 240 sampled matrix–checkpoint pairs (Figure 4). At step 2000, negative eigenvalues account for about 7–8% of the spectrum pooled across 16 matrices per model. The survey establishes the presence of negative derivative directions; the GGN tests establish their coupling to positive semidefinite curvature.

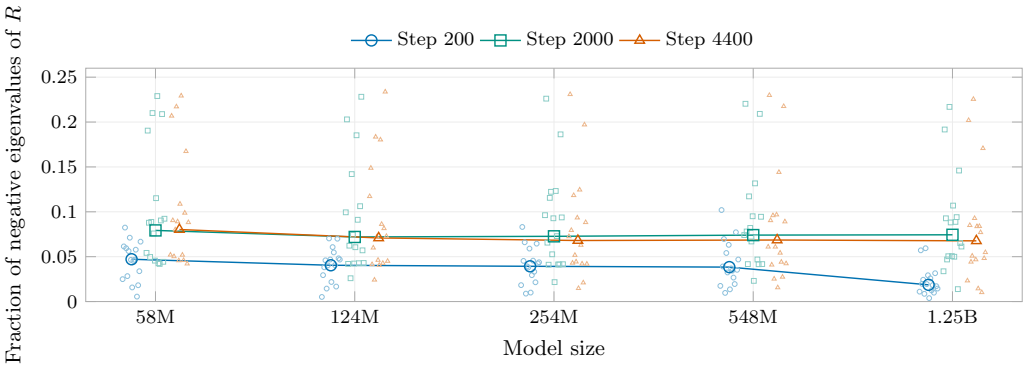


Figure 4: Negative eigenvalues of the smooth Muon derivative $\mathbf{R}$. Small markers give the fraction for each of 16 matrices at a checkpoint. Large markers divide the total number of negative eigenvalues by the total number of matrix coordinates. Each model size has one training run.

4.2 Crossing Both Edges with the Same Weight Step

With weights and batch fixed, we adjust stored momentum so that $\mathbf{P}(c) = c\mathbf{P}(1)$. Under Ψ_0, this leaves the next weight unchanged and gives $\mathbf{K}(c) = \mathbf{K}(1)/c$. For selected negative and positive real modes, the predicted crossing scales are

$$c_- = -\kappa_-(1)/d, \qquad c_+ = \kappa_+(1)/U_{\beta,d}.$$

We test $c/c_\pm = 0.8, 1, 1.2$, comparing the modal prediction with an independently differentiated optimizer step and paired finite differences (Figure 5). Checks confirm 12/12 lower and 11/12 upper crossings (Table 7); these concern selected modes, not full-state stability. Appendix J.6 gives the resolution criteria and failed case.

With the additive offset used in training, proposal scales $c = 0.5, 1, 2$ at 58M change the weight update by less than 10^{-5} in relative norm. The selected upper mode moves inside its edge at $c = 2$, while Hessian sharpness stays 12.45 (Appendix J.7). Nearly identical weight updates can therefore have different local stability.

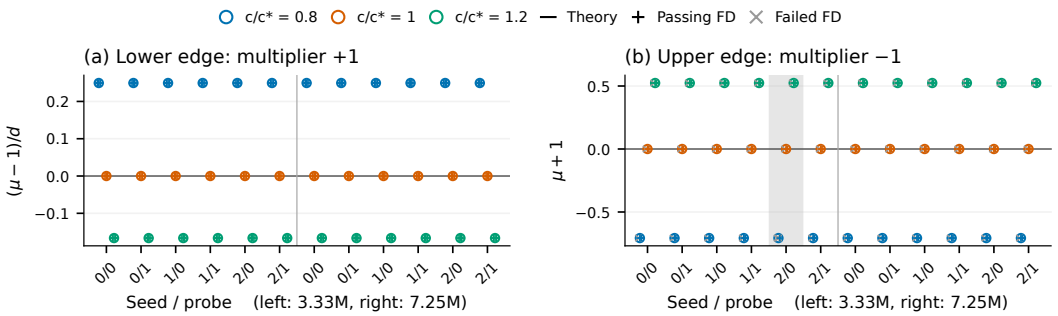


Figure 5: Pointwise multiplier checks for the small-model momentum interventions. Each seed and text batch contributes three proposal scales. Ticks show modal predictions, circles independent derivatives, and plus signs passing finite differences. Gray crosses show failed finite difference amplitudes. One upper crossing does not pass the full test and remains in the denominator.

4.3 Where Training Sits Relative to the Edges

In a 2048-update 3.33M continuation with changing batches and both optimizers active, validation loss falls from 6.26 to 5.67 (Figure 6). Across 64 sampled states, the upper ratio $\kappa_+/U_{\beta,d}$ ranges from 0.59 to 1.76, with median 0.99. Only 13/64 values lie within 5% of the edge, so the run does not show sustained residence near it. All 24 resolved lower modes lie beyond their edge; 40 searches remain unresolved.

With one batch repeated, three seeds are each continued with Muon alone or both optimizers active. Across these six windows, upper-ratio medians range from 2.56 to 3.15, every resolved upper ratio exceeds 1.62, and every resolved lower ratio satisfies $-\kappa_-/d > 6.4 \times 10^4$ (Appendix J.9). These conditional Muon spectra cover all trainable coordinates in Muon-only runs. They use the full loss Hessian, so negative curvature can contribute to lower modes.

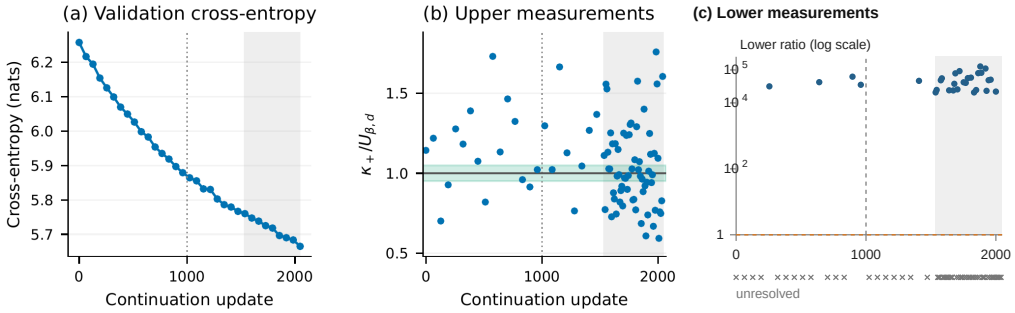


Figure 6: Loss falls while local modes cross or remain beyond the edges. Conditional Muon spectra are measured along a full Muon and AdamW continuation. Shading marks the primary 64-state window; 24 earlier states are also shown. The upper ratio crosses one; resolved lower ratios lie far above one. Crosses mark unresolved searches. State and batch both change between measurements.

4.4 How Perturbations Grow During Training

In a 64-update 3.33M continuation of the full Muon and AdamW state, paired perturbed runs give all-weight gains of 13 and 15 from update 1 to 64 (Figure 7). Successive Jacobians match these responses; repeating the initial Jacobian predicts more than 10^7. The displayed directions are random and AdamW-state-only perturbations, not identified eigenmodes. Across all measured directions and amplitudes, 6/12 combinations pass every step of the all-weight response check (Appendix J.11).

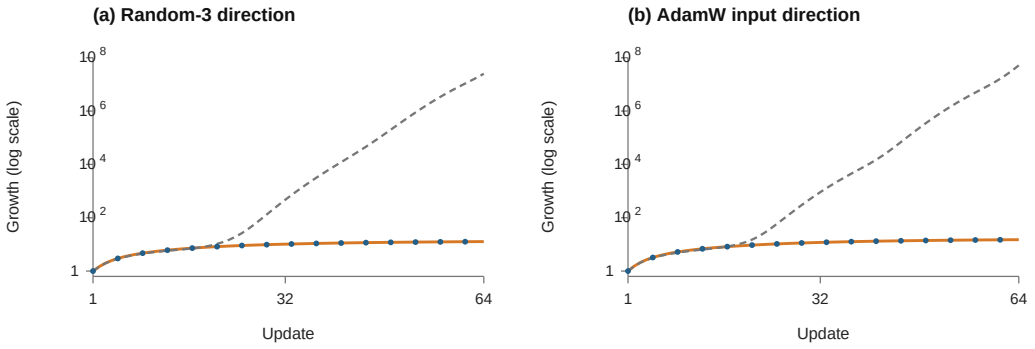


Figure 7: Successive Jacobians track paired-run responses; the initial Jacobian overpredicts growth. Both directions use amplitude 10^{-7} and all-weight norms normalized at update 1. Blue dots: paired runs; orange: successive Jacobians; gray dashes: repeated initial Jacobian.

The result also holds for identified eigenmodes. A 58M replay starts in selected lower and upper eigendirections and evolves all Muon weights and momentum for 16 updates on one fixed batch. Gains use a fixed scaled weight–momentum norm relative to the initial perturbation (Appendix J.10). Table 1 shows that successive Jacobians match both paired-run responses, whereas the initial Jacobian overpredicts growth by orders of magnitude. Local expansion alone therefore does not predict growth along training.

Table 1: Joint perturbation gains after 16 conditional Muon updates at amplitude 10^{-6}. All columns use the same fixed scaled norm.

Initial mode	Frozen derivative	Ordered derivatives	Paired runs
Lower	4.286×10^{15}	442.6	442.8
Upper	3689	5.220	5.220

Scope of the trajectory checks. Agreement requires relative vector error at most 10^{-3} at every tested update, not just matching the final response norm. The full-state replay also requires nonnegative Adam second moments. Six additional direction–amplitude combinations, corresponding to two requested real eigenmodes, remain unresolved. For the 58M lower mode, agreement persists for 2, 7, and 16 updates at amplitudes 10^{-4}, 10^{-5}, and 10^{-6}; the upper mode passes all 16 updates at all three amplitudes. Thus the interval over which finite perturbations follow the linear response depends on direction and amplitude. These checks establish responses along the tested trajectories, not uniform stability over all perturbations.

5 Conclusion

Practical Muon's finite matrix map and stored momentum determine two local stability edges through $\alpha\mathbf{RH}$, where $\mathbf{H}$ is the batch Hessian and $\mathbf{R}$ is the derivative of Muon's matrix map. GGN controls isolate negative modes created by the map, and momentum interventions cross either edge without changing the next weight. Training can improve beyond these local boundaries, while successive Jacobians explain measured perturbations far better than a fixed initial Jacobian. The quadratic constructions also exhibit attracting behavior after a crossing.

For a Muon variant, the derivative's quadratic form tests whether positive curvature can create negative modes; products of Jacobians test whether local expansion persists along training. Applying the GGN control at successive states would further separate the map's contribution from negative loss curvature during training.

AI USE STATEMENT

We used GPT-6 to improve the manuscript's language and to assist with the proofs beyond the two stability edges (Appendices F and G). We used Codex to assist with the experiments. The authors checked all AI-assisted text, proofs, code, and experimental outputs and take full responsibility for the paper.

REFERENCES

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 247–257, 2022.

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The Polar Express: Optimal Matrix Sign Methods and Their Application to the Muon Algorithm. In *The Fourteenth International Conference on Learning Representations*, 2026.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of Stochastic Stability: Revisiting the Edge of Stability for SGD. *arXiv preprint arXiv:2412.20553*, 2024.

Arseniy Andreyev, Advikar Ananthkumar, Marc Walden, Tomaso Poggio, and Pierfrancesco Beneventano. Momentum Further Constrains Sharpness at the Edge of Stochastic Stability. *arXiv preprint arXiv:2604.14108*, 2026.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding Gradient Descent on the Edge of Stability in Deep Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 948–1024, 2022.

Pierfrancesco Beneventano, Mahmoud Abdelmoneum, and Tomaso Poggio. The Spectral Dynamics and Noise Geometry of Muon. *arXiv preprint arXiv:2606.08388*, 2026.

Jeremy Bernstein and Laker Newhouse. Old Optimizer, New Norm: An Anthology. In *OPT 2024: Optimization for Machine Learning*, 2024.

Jeremy Bernstein and Laker Newhouse. Modular Duality in Deep Learning. In *Forty-second International Conference on Machine Learning*, volume 267, pp. 3920–3930, 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*, 2026.

Lei Chen and Joan Bruna. Beyond the Edge of Stability via Two-Step Gradient Updates. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4330–4391, 2023.

Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves a constrained optimization: As Lyapunov predicts. In *International Conference on Learning Representations*, 2024.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon Optimizes Under Spectral Norm Constraints. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Jeremy M. Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive Gradient Methods at the Edge of Stability. *arXiv preprint arXiv:2207.14484*, 2022.

Jeremy M. Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding Optimization in Deep Learning with Central Flows. In *International Conference on Learning Representations*, 2025.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability. In *International Conference on Learning Representations*, 2023.

Chao Ding, Defeng Sun, Jie Sun, and Kim-Chuan Toh. Spectral Operators of Matrices. *Mathematical Programming*, 168(1–2):509–531, 2018.

Thang Do, Steffen Dereich, and Arnulf Jentzen. On MUON Optimization: From Non-Convergence to an Error Analysis with Polar Express and the Newton–Schulz Polynomial from Implementations. *arXiv preprint arXiv:2608.04607*, 2026.

Zhehang Du and Weijie Su. The Newton-Muon Optimizer. *arXiv preprint arXiv:2604.01472*, 2026.

Antoine Gonon, Andreea-Alexandra Muşat, and Nicolas Boumal. Insights on Muon from Simple Quadratics. *arXiv preprint arXiv:2602.11948*, 2026.

Nicholas J. Higham. Computing the Polar Decomposition—with Applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2 edition, 2012.

Rustem Islamov, Michael Crawshaw, Jeremy Cohen, and Robert M. Gower. Non-Euclidean Gradient Descent Operates at the Edge of Stability. In *Forty-third International Conference on Machine Learning*, 2026.

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

Eliahu I. Jury. *Theory and Application of the Z-Transform Method*. John Wiley & Sons, New York, 1964.

Gyu Yeol Kim and Min hwan Oh. Convergence of Muon with Newton–Schulz. In *The Fourteenth International Conference on Learning Representations*, 2026.

Kimi Team. Kimi K2: Open Agentic Intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Yuri A. Kuznetsov. *Elements of Applied Bifurcation Theory*, volume 112 of *Applied Mathematical Sciences*. Springer, New York, 2 edition, 1998.

Jaerin Lee and Kyoung Mu Lee. The Road Taken: The Role of Optimizers at the Edge of Stability. *arXiv preprint arXiv:2608.18415*, 2026.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, 2 edition, 2013.

Mingyi Li and Taira Tsuchiya. Muon with Finite Newton–Schulz: The Smoothing Benefit in Nonsmooth Nonconvex Optimization. *arXiv preprint arXiv:2608.26288*, 2026.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *The Twelfth International Conference on Learning Representations*, 2024.

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.

Zhuanghua Liu, Menglian Wang, and Luo Luo. Musec: MomentUm SpEctral Clipping for Stable Muon-type Training. *arXiv preprint arXiv:2609.11655*, 2026a.

Ziyue Liu, Ruijie Zhang, Zhengyang Wang, Yequan Zhao, Yupeng Su, Zi Yang, and Zheng Zhang. Muon^2: Boosting Muon via Adaptive Second-Moment Preconditioning. *arXiv preprint arXiv:2604.09967*, 2026b.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. FineWeb-Edu: The Finest Collection of Educational Content, 2024. URL `https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu`.

James Martens. New Insights and Perspectives on the Natural Gradient Method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Vanni Noferini. A Formula for the Fréchet Derivative of a Generalized Matrix Function. *SIAM Journal on Matrix Analysis and Applications*, 38(2):434–457, 2017.

Tetiana Parshakova, Ahmed Khaled, Michael Crawshaw, Guillaume Garrigos, and Robert M. Gower. Muon Does Not Converge on Convex Lipschitz Functions. *arXiv preprint arXiv:2605.08980*, 2026.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL `https://openreview.net/forum?id=n6SCkn2QaG`.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training Deep Learning Models with Norm-Constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49069–49104, 2025.

Mark Rhee, Jamie Simon, and Dhruva Karkada. Muon learns balanced solutions in matrix factorization without slow saddle-to-saddle dynamics. *arXiv preprint arXiv:2606.30509*, 2026.

Yousef Saad. *Numerical methods for large eigenvalue problems: revised edition*. SIAM, 2011.

Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the Ideal: Analyzing the Inexact Muon Update. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 5167–5175, 2026.

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.

Ziyuan Tang, Tianshi Xu, Yousef Saad, and Yuanzhe Xi. Hierarchical Muon: Tiled Newton–Schulz Updates for Efficient Muon Optimization. *arXiv preprint arXiv:2606.27216*, 2026.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why Muon Outperforms Adam: A Curvature Perspective. *arXiv preprint arXiv:2606.04662*, 2026.

Xiaodong Wu, Wenyi Yu, Chao Zhang, and Philip Woodland. Spectral Allocation: Why Muon Outperforms Adam, and How to Improve Muon. *arXiv preprint arXiv:2608.25990*, 2026.

Yechen Zhang, Shuhao Xing, Junhao Huang, Kai Lv, Yunhua Zhou, Xipeng Qiu, Qipeng Guo, and Kai Chen. Mousse: Rectifying the Geometry of Muon with Curvature-Aware Preconditioning. *arXiv preprint arXiv:2603.09697*, 2026.

GUIDE TO THE APPENDIX

Appendix A reviews the background comparisons in Table 2 and nearby Muon analyses. Readers new to local stability can start with Appendix B, which introduces optimizer state, matrix derivatives, and state multipliers. Appendix C derives the Muon transformation derivative. Appendices D and E prove the stability boundary and proposal scaling law.

Appendices F and G prove the two nonlinear examples, and Appendix H studies small changes to them. Appendix I derives the Sophia and Lion rows. The experimental appendix gives the shared protocols first. Appendices J.6 and J.5 test the negative-mode mechanism under GGN curvature; Appendix J.6 also checks both crossings across seeds. Appendix J.4 surveys map derivatives across model sizes. Appendices J.9, J.10, and J.11 connect local modes to training trajectories. Appendix J.7 checks proposal scaling with the additive offset, and Appendix J.8 verifies the nonlinear constructions.

NOTATION

Symbol	Meaning	Defined in
$\mathbf{W}, \mathbf{M}, \mathbf{P}$	Weight, stored momentum, proposal	§2
$\mathbf{G}, \mathbf{H}$	Batch gradient, Hessian	§2
$\alpha, \beta, \lambda_{\mathrm{wd}}, d$	Learning rate, momentum, weight decay, decay per step	§2
$\Psi_\delta, \Psi_{\mathrm{cl}}, \Psi_0$	Muon maps with additive, clamp, or scale invariant normalization	§2
ϕ, g, s	One scalar step, five steps, normalized singular value	§2
$\mathbf{R}, \mathbf{K}$	Map derivative, $\alpha\mathbf{R}\mathbf{H}$	§2
$\mathbf{A}, \mu, \kappa$	State Jacobian, multiplier, eigenvalue of $\mathbf{K}$	§3
$U_{\beta,d}$	Upper edge in κ units	Thm. 1
c	Proposal scale	Thm. 3
$p, \chi_{m,n}$	$\min(m,n)$, rectangular shape factor	§2

A RELATED WORK IN DETAIL

Gradient Descent at the Edge of Stability. During full batch gradient descent on neural networks, the largest Hessian eigenvalue often rises toward $2/\alpha$. The loss can then oscillate between consecutive steps while decreasing over longer intervals (Cohen et al., 2021). The reference value comes from a fixed quadratic loss: a perturbation along a Hessian eigenvector with eigenvalue λ is multiplied by $1 - \alpha\lambda$. For $\lambda > 0$, this perturbation contracts exactly when $0 < \alpha\lambda < 2$, as shown in Table 2. Directional smoothness, motion along a manifold of minima, and higher derivative feedback describe several ways optimization can continue near or beyond this threshold (Ahn et al., 2022; Arora et al., 2022; Damian et al., 2023). Two-step analyses describe period two motion in specific models (Chen & Bruna, 2023). Central flow models explain averaged motion near an edge and extend the picture to adaptive methods (Cohen et al., 2025). We study Muon's local boundary and give separate examples of nonlinear behavior after a crossing in Section 3.

Mini-batches and Momentum at the Edge of Stability. For stochastic gradient descent, batch sharpness measures curvature in stochastic update directions and can approach an edge even when the full batch top Hessian eigenvalue does not (Andreyev & Beneventano, 2024). Momentum changes this picture and can produce different batch-size regimes (Andreyev et al., 2026). Our Jacobian describes one Muon update on a specified batch. Batch sharpness and its eigenvalues measure different responses and need not reach their thresholds together.

Table 2: Local stability formulas. We derive the practical Muon row. Appendix I derives the Sophia and Lion comparisons. The spectral descent row reports a quantity measured during training.

Method	Setting	Quantity	Formula and meaning
Background comparisons			
GD (Cohen et al., 2021)	Fixed quadratic	$\lambda \in \operatorname{spec}(\mathbf{H})$	$\mu = 1 - \alpha\lambda$ Stable iff $0 < \alpha\lambda < 2$.
Preconditioned GD (Cohen et al., 2022)	Fixed quadratic and $\mathbf{B} \succ 0$, no momentum	Eigenvalue λ of $\mathbf{B}^{1/2}\mathbf{H}\mathbf{B}^{1/2}$	$\mu = 1 - \alpha\lambda$ Stable iff $0 < \alpha\lambda < 2$.
Normalized spectral descent (Islamov et al., 2026)	Exact polar direction, no momentum	$S_{\mathrm{op}}/\lVert\mathbf{G}\rVert_{\mathrm{nuc}}$, $\mathbf{G} \neq 0$	$\alpha S_{\mathrm{op}}/\lVert\mathbf{G}\rVert_{\mathrm{nuc}} \approx 2$ Empirical sharpness scale.
Filtered updates (Lee & Lee, 2026)	Fixed quadratic and $\mathbf{B} \succ 0$, stable causal filter Q	Eigenvalue $\lambda > 0$ of $\mathbf{B}^{1/2}\mathbf{H}\mathbf{B}^{1/2}$, $Q(1) = 1$	$\mu - 1 + \alpha\lambda Q(\mu) = 0$ First positive $\alpha\lambda$ with $\lvert\mu\rvert = 1$.
Results derived in this work			
Practical Muon	Fixed batch and state, five polynomial steps, Nesterov momentum, decay	$\mathbf{K} = \alpha\mathbf{R}\mathbf{H}$, $\mathbf{R} = \mathrm{D}\Psi(\mathbf{P})$, $\kappa \in \operatorname{spec}(\mathbf{K})$	Real modes are stable iff $-d < \kappa < \dfrac{(1+\beta)(2-d)}{(1-\beta)(1+2\beta)}$. Endpoint multipliers: $+1, -1$. All modes obey Equation 3.
Sophia (Liu et al., 2024)	Fixed positive denominator, away from clipping thresholds	$\mathbf{K}_{\mathrm{S}} = \alpha\mathbf{B}\mathbf{H}$, $\kappa \in \operatorname{spec}(\mathbf{K}_{\mathrm{S}})$	Real modes are stable iff $-d < \kappa < \dfrac{(1+\beta)(2-d)}{1-\beta}$. Endpoint multipliers: $+1, -1$.
Lion (Chen et al., 2023)	Every proposal coordinate is nonzero	Joint weight and momentum Jacobian $\mathbf{A}_{\mathrm{L}}$	$\operatorname{spec}(\mathbf{A}_{\mathrm{L}}) = \{1 - d, \beta_2\}$ Stable iff $0 < d < 2$.

$\mathbf{B}$ is a fixed positive preconditioner in the GD and filter rows. For Sophia it is the coordinate derivative defined below. For one matrix block, $S_{\mathrm{op}} = \max_{\lVert\mathbf{E}\rVert_{\mathrm{op}}=1} \langle \mathbf{E}, \mathbf{H}[\mathbf{E}] \rangle_{\mathrm{F}}$. The spectral norm $\lVert\mathbf{E}\rVert_{\mathrm{op}}$ is the largest singular value of $\mathbf{E}$. The nuclear norm $\lVert\mathbf{G}\rVert_{\mathrm{nuc}}$ is the sum of the singular values of $\mathbf{G}$. For Sophia, $\mathbf{B} = \operatorname{diag}(\mathbf{1}\{\lvert m_i^+\rvert < q_i\}/q_i)$ with fixed $q_i > 0$ and $0 \leq \beta < 1$. Additional gradient norm clipping is absent or inactive. For Lion, $\beta_2 \in [0, 1)$ is the stored momentum coefficient. The Muon conditions are those of Theorem 1. The Sophia interval assumes $0 \leq d < 2$. Appendix A reviews the background comparisons. Appendix I derives the Sophia and Lion formulas.

Fixed Preconditioners and Momentum. Preconditioned sharpness extends the GD comparison to adaptive methods (Cohen et al., 2022). For the frozen update $\mathbf{W}^+ = \mathbf{W} - \alpha\mathbf{B}\mathbf{G}(\mathbf{W})$, where $\mathbf{B} \succ 0$ is constant, the product $\mathbf{B}\mathbf{H}$ is similar to a symmetric matrix because

$$\mathbf{B}^{-1/2}(\mathbf{B}\mathbf{H})\mathbf{B}^{1/2} = \mathbf{B}^{1/2}\mathbf{H}\mathbf{B}^{1/2}.$$

The GD multiplier formula therefore applies to the eigenvalues of $\mathbf{B}^{1/2}\mathbf{H}\mathbf{B}^{1/2}$. Here $\mathbf{B}$ is the matrix that multiplies the gradient. It equals the inverse of the preconditioner denoted by $\mathbf{P}$ in the cited paper. This explains the preconditioned GD row of Table 2. The adaptive optimizer experiments compare training with such frozen references while the preconditioner itself continues to change. Stored momentum introduces additional variables, so its boundary also depends on the momentum rule.

Linear filtering describes how an optimizer combines current and past gradients (Lee & Lee, 2026). To define the Q in Table 2, write the filtered gradient as $\sum_{j \geq 0} q_j \mathbf{g}_{t-j}$ and set $Q(z) = \sum_{j \geq 0} q_j z^{-j}$. The quadratic analysis fixes $\mathbf{H}$, $\mathbf{B}$, and the filter. The filter is real, causal, and rational, its poles lie inside the unit circle, and $Q(1) = 1$. For a positive eigenvalue λ of $\mathbf{B}^{1/2}\mathbf{H}\mathbf{B}^{1/2}$, a state multiplier satisfies

$$\mu - 1 + \alpha\lambda Q(\mu) = 0.$$

The boundary is the first positive gain $\alpha\lambda$ at which a root reaches $|\mu| = 1$. Its location depends on the filter, so the first crossing need not occur at $\mu = -1$. This analysis includes heavy ball and Nesterov filters, and the experiments also examine the direction of the executed update. In practical Muon, the gradient and momentum then pass through a nonlinear matrix transformation. Its derivative enters the local calculation in place of a fixed positive preconditioner and can have negative eigenvalues.

Lion and Sophia. Lion uses the sign of a gradient and momentum proposal for the weight update, with a separate coefficient for stored momentum (Chen et al., 2023). Lyapunov analyses relate Lion and its extensions to constrained or regularized optimization (Chen et al., 2024). Sophia divides a moving average of gradients by a curvature estimate and clips the resulting coordinates (Liu et al., 2024). Appendix I derives their state responses. Lion's sign map is locally constant when every proposal coordinate is nonzero. An unsaturated Sophia coordinate has a positive response to its numerator. The derivations also track Sophia's curvature refresh and gradient norm clipping.

Exact Polar Updates and Spectral Norm Constraints. The exact polar direction has a simple optimization interpretation. If $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, then $-\mathbf{U}\mathbf{V}^\top$ minimizes $\langle \mathbf{G}, \mathbf{D} \rangle_{\mathrm{F}}$ over $\|\mathbf{D}\|_{\mathrm{op}} \leq 1$. Norm duality and its extension across network modules give a geometric view (Bernstein & Newhouse, 2024; 2025). Linear minimization over norm balls also gives stochastic methods for constrained and unconstrained training (Pethick et al., 2025). Exact polar Muon with momentum fits the Lion-$\mathcal{K}$ framework. With decoupled weight decay, this connection gives convergence results for a spectral norm constrained problem (Chen et al., 2026). These results concern the optimization problem and its stationary conditions. Our local boundary concerns the derivative of one update.

Generalized Sharpness for Spectral Descent. Generalized sharpness extends the Edge of Stability comparison to non-Euclidean gradient descent by maximizing Hessian curvature over directions of unit norm in the chosen geometry (Islamov et al., 2026). For one matrix block with the spectral norm, this quantity is S_{op} in Table 2. For the normalized polar update and a nonzero gradient, the empirical comparison uses $S_{\mathrm{op}}/\|\mathbf{G}\|_{\mathrm{nuc}}$, where the nuclear norm is the sum of the singular values. For several blocks with the maximum block spectral norm, the dual norm is the sum of the block nuclear norms. The quadratic convergence theorems in the cited work concern the unnormalized update, which includes an additional dual gradient norm factor. The normalized row in our table therefore reports the measured sharpness scale.

Exact Polar Muon: Oscillations and Noise. Exact polar Muon admits explicit trajectory analyses on structured objectives. In aligned matrix factorization, the factor dynamics reduce to scalar sign updates. This gives balancing results and describes oscillation traps near the origin at a fixed step size (Rhee et al., 2026). That analyzed update omits momentum, weight decay, and Newton–Schulz approximation. On structured quadratic objectives, exact polar updates can confine a trajectory to a lattice, which restricts the points it can approach at a fixed step size (Gonon et al., 2026). The latter work also treats momentum and Nesterov momentum applied before the polar projection in its appendix. It studies how approximate sign updates and stochastic noise alter these trajectories and compares finite Newton–Schulz updates in quadratic experiments. These examples show why bounded oscillation and convergence to a fixed parameter value are different outcomes, even for a quadratic loss.

Continuous polar flows and discrete matrix updates have been studied through their singular value dynamics in specified projected models (Beneventano et al., 2026). Under alignment assumptions, a local comparison connects a projected self polar flow to a gradient based polar update. The same work differentiates the polar factor to quantify sensitivity to small matrix perturbations and separately derives the stationary noise variance of a linear momentum buffer. This provides a direct precedent for studying a transformation through its derivative. Our analysis evaluates the derivative of the normalized five-step polynomial and combines it with the loss Hessian in the joint weight and momentum Jacobian.

Finite Newton–Schulz Iterations and Inexact Updates. Polar approximation methods compute the matrix map without an SVD. The classical cubic scalar step $x \mapsto (3x - x^3)/2$ is increasing on $[0, 1]$, while the derivative of Muon's five-step quintic changes sign there (Figure 1). Monotonicity of a scalar step alone does not settle the sign of the complete normalized matrix derivative (Proposition 2). Polynomial iterations use matrix multiplication (Higham, 2008, Chap. 8). Polar Express constructs

iterations with approximation guarantees on specified singular value intervals (Amsel et al., 2026). Taylor Newton–Schulz polynomials preserve the unit spectral ball and give stationarity rates for momentum under smoothness and bounded stochastic variance (Kim & hwan Oh, 2026). The latter update uses a momentum buffer without the additional Nesterov proposal. Approximation error bounds concern proximity to the polar direction. Local stability also depends on the derivative of the approximation.

General normalized polynomial analyses include the original five-step recurrence and Polar Express, with error bounds and examples that fail to converge on specified stochastic objectives (Do et al., 2026). Finite Newton–Schulz iterations can also help optimization by smoothing the polar update (Li & Tsuchiya, 2026). The latter stationarity analysis uses a fixed normalization scale, classical Taylor coefficients, and an iteration depth chosen for the target accuracy. These results show why the polynomial, normalization, and iteration count should be specified when comparing Muon analyses.

An approximate update can be modeled by its additive error from the exact linear minimization direction. Deterministic and stochastic convergence bounds then expose how this error interacts with the learning rate and momentum (Shulgin et al., 2026). Further results explicitly include Nesterov momentum, inexact polar computation, and heavy tailed gradient noise (Choudhury et al., 2026). Their approximation conditions control alignment with the input matrix and the operator norm of the output. Under these conditions, the analysis bounds the number of iterations needed to reach an approximate stationary point. Our local boundaries are determined by the eigenvalues of $\alpha \mathrm{D}\Psi(\mathbf{P})\mathbf{H}$ and the momentum rule.

Exact polar Muon with a fixed momentum coefficient can also fail to converge on convex Lipschitz objectives under nonincreasing stepsizes chosen in advance (Parshakova et al., 2026). These nonsmooth counterexamples apply the polar map directly to the momentum buffer, without an additional Nesterov proposal.

Loss Curvature Along Muon Updates. Gradient alignment and directional curvature describe loss changes along Muon updates, with supporting analysis on structured matrix quadratics (Wang et al., 2026). Local quadratic loss profiles on held-out batches can also probe individual singular directions of the momentum (Wu et al., 2026). Those measurements motivate SAMuon, which reallocates update strength across singular components. These quantities describe the loss along a chosen update direction. The operator $\mathrm{D}\Psi(\mathbf{P})\mathbf{H}$ additionally describes how the update itself changes when the current state is perturbed.

Sensitivity to Changes in Training Data. Algorithmic stability measures how training changes when one sample in the dataset is replaced. This sensitivity gives generalization bounds for Muon, together with a Lipschitz result for a finite Newton–Schulz map with positive normalization offset (Chang et al., 2026). It compares training runs on neighboring datasets. Our local stability calculation fixes the batch and perturbs the weights and momentum at one state. Along a changing training trajectory, these local derivatives form an ordered product, as explained in Appendix B.

Scaling, Clipping, and Preconditioning. Practical variants change several parts of the update. Large language model implementations study weight decay, scaling across matrix shapes, and distributed execution (Liu et al., 2025). Kimi K2 combines Muon with QK-Clip, which rescales query and key weights after an update when attention logits exceed a threshold (Kimi Team, 2025). Hierarchical Muon applies Newton–Schulz independently to matrix tiles (Tang et al., 2026). Each change alters the local derivative, so its stability calculation must use the corresponding update map.

Musec replaces spectral flattening with spectral clipping to address Muon training instability (Liu et al., 2026a). Its hard clipping algorithm has stationarity guarantees for weakly convex stochastic objectives under Lipschitz and bounded variance assumptions, with prescribed parameter schedules. Soft Musec uses smooth spectral saturation approximated by finite coupled Newton–Schulz iterations and is evaluated for training stability.

Random and inverted update spectra compare alignment and directional loss reduction with Muon (Shumaylov et al., 2026). Muon^2 applies coordinatewise second moment scaling before Newton–Schulz (Liu et al., 2026b). Newton–Muon uses layer input second moments to precondition the matrix before taking its polar factor (Du & Su, 2026). Mousse combines matrix preconditioning, Newton–Schulz, and a transformation back to the original coordinates (Zhang et al., 2026). Their

derivatives and stored variables differ from those in Section 2; the state Jacobian must be formed from each complete update.

B Background on Optimizers and Local Stability

B.1 Gradient Descent and the Classical Stability Boundary

Let $\boldsymbol{\theta}_t$ denote the trainable parameters, and let

$$\mathbf{g}_t = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{\theta}_t)$$

be the gradient evaluated on the realized batch $\mathcal{B}_t$. Throughout a local calculation, we hold this batch and every other random choice fixed. Gradient descent uses

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \mathbf{g}_t.$$

For a fixed learning rate α, let $\mathbf{H}$ be the batch Hessian at $\boldsymbol{\theta}_t$. A perturbation in an eigendirection $\mathbf{v}$, with $\mathbf{H}\mathbf{v} = \lambda \mathbf{v}$, changes to first order as

$$\mathbf{v}^+ = (1 - \alpha\lambda)\mathbf{v}.$$

The scalar $1 - \alpha\lambda$ is the multiplier of this mode. For $\lambda > 0$, the mode is stable exactly when $0 < \alpha\lambda < 2$. It reaches the alternating multiplier -1 at $\alpha\lambda = 2$. With decoupled weight decay $d = \alpha\lambda_{\mathrm{wd}}$, the multiplier becomes $1 - d - \alpha\lambda$, and its stable interval is $-d < \alpha\lambda < 2 - d$.

Appendix A reviews how this local threshold relates to the Edge of Stability observed during training.

B.2 Why the State Includes Momentum

On a fixed batch, the current weight determines the gradient. The next update also uses stored optimizer variables. In practical Muon, the stored momentum and the proposal satisfy

$$\mathbf{M}^+ = \beta\mathbf{M} + (1 - \beta)\mathbf{G}(\mathbf{W}), \qquad \mathbf{P} = \beta\mathbf{M}^+ + (1 - \beta)\mathbf{G}(\mathbf{W}). \tag{8}$$

Thus $\mathbf{P} = \beta^2\mathbf{M} + (1 - \beta^2)\mathbf{G}(\mathbf{W})$. At the same $\mathbf{W}$, two momentum buffers separated by $\Delta\mathbf{M}$ give proposals separated by $\beta^2\Delta\mathbf{M}$. To first order, their next weights differ by

$$\Delta\mathbf{W}^+ = -\alpha\beta^2 \mathbf{R}[\Delta\mathbf{M}], \qquad \mathbf{R} = \mathrm{D}\Psi(\mathbf{P}).$$

The state $(\mathbf{W}, \mathbf{M})$ determines the next update and lets us follow a perturbation across updates. In indexed notation, $\mathbf{M} = \mathbf{M}_t$ and $\mathbf{M}^+ = \mathbf{M}_{t+1}$. Both denote the same persistent buffer at consecutive steps. The proposal $\mathbf{P}$ is computed from the current weights and buffer during the update.

Adam provides a familiar example with two stored variables (Kingma & Ba, 2015, Alg. 1):

$$\mathbf{m}_{t+1} = \beta_1 \mathbf{m}_t + (1 - \beta_1)\mathbf{g}_t,$$
$$\mathbf{v}_{t+1} = \beta_2 \mathbf{v}_t + (1 - \beta_2)\mathbf{g}_t^{\odot 2}.$$

Here the square is coordinatewise. Adam uses both moments to compute the parameter update. AdamW keeps these recursions and adds decoupled weight decay (Loshchilov & Hutter, 2019). At a given step, their local Jacobian acts on $(\boldsymbol{\theta}, \mathbf{m}, \mathbf{v})$ wherever the update is differentiable, with the step counter held fixed. The bias correction factors are evaluated separately at each step. For each optimizer, the state contains the parameters and the variables carried into the next update.

B.3 Practical Muon for Matrix Parameters

For each matrix parameter considered here, Muon normalizes the proposal in Equation 8 and applies five iterations of a quintic Newton–Schulz polynomial (Section 2). We now describe how this matrix transformation differs from the exact polar factor.

If a full-rank matrix has thin singular value decomposition

$$\mathbf{P} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top,$$

its rectangular polar factor is $\mathbf{U}\mathbf{V}^\top$ (Higham, 2008, Thm. 8.1). The polar factor preserves the singular vectors and replaces every singular value by one. The five Newton–Schulz iterations also preserve the singular vectors. They take each normalized singular value s to $g(s) = \phi^{\circ 5}(s)$. The derivative depends on the polynomial, the normalization, and the current proposal.

The quadratic constructions use the clamp map above or below its threshold, the language model measurements use the additive offset, and the exact scaling tests use Ψ_0. Lemma 7 gives the exact polar comparison.

Table 3: Muon transformations used in the analysis.

Transformation	Definition	Use
Training code	Cited floating-point normalization, five Newton–Schulz steps, and shape factor	Training runs
Ψ_{cl}	Clamp normalization, five Newton–Schulz steps, and shape factor	Quadratic examples at both edges
Ψ_δ	Division by $\|\mathbf{P}\|_{\mathrm{F}} + \delta$, five Newton–Schulz steps, and shape factor	Derivative formulas and training state calculations
Ψ_0	Division by $\|\mathbf{P}\|_{\mathrm{F}}$, followed by the same five steps and shape factor	Exact proposal scaling tests
Exact polar	$\chi_{m,n}\mathbf{U}\mathbf{V}^\top$ for $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$	Derivative comparison

The clamp transformation is smooth when $\|\mathbf{P}\|_{\mathrm{F}} > \epsilon_{\mathrm{cl}}$ and when $\|\mathbf{P}\|_{\mathrm{F}} < \epsilon_{\mathrm{cl}}$, but not at equality. The normalization in Ψ_δ is smooth at every nonzero proposal. When $\delta = 0$, $\Psi_0(c\mathbf{P}) = \Psi_0(\mathbf{P})$ for every $c > 0$. The clamp transformation has the same scaling property whenever $c\|\mathbf{P}\|_{\mathrm{F}} > \epsilon_{\mathrm{cl}}$ throughout the interval under study.

B.4 How Small Perturbations Change an Update

Let $\mathbf{z}$ collect the parameters and stored optimizer variables, so $\mathbf{z} = (\mathbf{W}, \mathbf{M})$ for Muon. With the batch and all other random choices fixed, write one optimizer step as $\mathbf{z}^+ = \mathcal{T}_{\mathcal{B}}(\mathbf{z})$. An autonomous update applies the same function at every step. A fixed point is a state $\mathbf{z}_\star$ satisfying $\mathcal{T}_{\mathcal{B}}(\mathbf{z}_\star) = \mathbf{z}_\star$. At a state $\mathbf{z}_t$, a small state perturbation $\Delta\mathbf{z}_t$ satisfies

$$\Delta\mathbf{z}_{t+1} = \mathbf{A}_t\Delta\mathbf{z}_t + o(\|\Delta\mathbf{z}_t\|), \qquad \mathbf{A}_t = \mathrm{D}\mathcal{T}_{\mathcal{B}_t}(\mathbf{z}_t).$$

The eigenvalues of $\mathbf{A}_t$ are the multipliers at this state. Repeating $\mathbf{A}_t$ gives a Schur stable recurrence exactly when every multiplier lies in the open unit disk (Jury, 1964).

The same condition gives local exponential stability at a fixed point $\mathbf{z}_\star$ of a continuously differentiable autonomous update. The following direct argument also covers a singular Jacobian. Write $\mathbf{A}_\star = \mathrm{D}\mathcal{T}(\mathbf{z}_\star)$ and suppose its spectral radius is below one. Then

$$\mathbf{Q} = \sum_{j=0}^{\infty} (\mathbf{A}_\star^\top)^j \mathbf{A}_\star^j \succ 0, \qquad \mathbf{A}_\star^\top \mathbf{Q} \mathbf{A}_\star = \mathbf{Q} - \mathbf{I}.$$

Thus $\mathbf{A}_\star$ strictly contracts the norm $\|\mathbf{x}\|_{\mathbf{Q}} = (\mathbf{x}^\top\mathbf{Q}\mathbf{x})^{1/2}$. Continuity of $\mathrm{D}\mathcal{T}$ gives a small ball around $\mathbf{z}_\star$ on which every derivative has norm at most some $q < 1$. Since $\mathcal{T}(\mathbf{z}_\star) = \mathbf{z}_\star$, the mean value inequality gives

$$\|\mathcal{T}(\mathbf{z}) - \mathbf{z}_\star\|_{\mathbf{Q}} \le q\|\mathbf{z} - \mathbf{z}_\star\|_{\mathbf{Q}}.$$

The ball maps into itself, and iteration gives geometric decay. This is local exponential stability.

A multiplier at $+1$ leaves its direction unchanged to first order, while a multiplier at -1 reverses it. Higher order terms determine nearby states after a crossing. Our lower example is a pitchfork: two stable fixed points appear after a crossing at $+1$. An attracting orbit of period two can appear after a flip at -1 (Kuznetsov, 1998, Sec. 4.5). The quadratic constructions solve the resulting fixed point or orbit equations and check their stability with Theorem 1.

Along a trajectory with changing batches or other random inputs, fix the realized input sequence and write the update at step j as $\mathcal{T}_j$, with $\mathbf{A}_j = \mathrm{D}\mathcal{T}_j(\mathbf{z}_j)$. If each update is differentiable near the reference states, the chain rule gives

$$\mathbf{z}_{t+k}^{(\varepsilon)} - \mathbf{z}_{t+k} = \varepsilon \mathbf{A}_{t+k-1} \cdots \mathbf{A}_t \mathbf{v} + o(\varepsilon).$$

Here the perturbed run starts at $\mathbf{z}_t + \varepsilon\mathbf{v}$ and uses the same input sequence. The result holds for every fixed finite k by the chain rule. It needs no distributional assumption on batch noise. The product of changing Jacobians describes perturbation growth along the trajectory. Repeating one Jacobian answers a different question.

For practical Muon, a weight perturbation first changes the batch gradient through the Hessian $\mathbf{H}$ and then passes through $\mathbf{R} = \mathrm{D}\Psi(\mathbf{P})$. Their composition is $\mathbf{K} = \alpha\mathbf{R}\mathbf{H}$. Theorem 1 converts every eigenvalue of $\mathbf{K}$ into two multipliers of the joint weight and momentum state. When $\beta = d = 0$ and $\Psi(\mathbf{P}) = \mathbf{P}$, we have $\mathbf{R} = \mathbf{I}$, $\mathbf{K} = \alpha\mathbf{H}$, and recover the gradient descent multiplier $1 - \alpha\lambda$.

B.5 Matrix Derivatives Used in the Proofs

For matrices $\mathbf{E}_1, \mathbf{E}_2 \in \mathbb{R}^{m\times n}$, the Frobenius inner product and norm are

$$\langle \mathbf{E}_1, \mathbf{E}_2 \rangle_{\mathrm{F}} = \mathrm{tr}(\mathbf{E}_1^\top \mathbf{E}_2), \qquad \|\mathbf{E}\|_{\mathrm{F}} = \sqrt{\langle \mathbf{E}, \mathbf{E} \rangle_{\mathrm{F}}}.$$

A linear map $\mathbf{T}$ on this matrix space is self-adjoint when

$$\langle \mathbf{E}_1, \mathbf{T}[\mathbf{E}_2] \rangle_{\mathrm{F}} = \langle \mathbf{T}[\mathbf{E}_1], \mathbf{E}_2 \rangle_{\mathrm{F}}.$$

In an orthonormal basis, a self-adjoint map is represented by a symmetric matrix, so its eigenvalues are real. It is positive semidefinite when $\langle \mathbf{E}, \mathbf{T}[\mathbf{E}] \rangle_{\mathrm{F}} \geq 0$ for every $\mathbf{E}$.

The Fréchet derivative of a matrix map F at $\mathbf{P}$ is the linear map $\mathrm{D}F(\mathbf{P})$ such that

$$F(\mathbf{P} + \mathbf{E}) = F(\mathbf{P}) + \mathrm{D}F(\mathbf{P})[\mathbf{E}] + o(\|\mathbf{E}\|_{\mathrm{F}}).$$

This is the standard definition (Noferini, 2017, Sec. 2.3). The value $F(\mathbf{P})$ gives the transformed update at $\mathbf{P}$, while $\mathrm{D}F(\mathbf{P})$ describes its first-order change under a perturbation. Two proposals can therefore produce the same transformed update and have different derivatives.

Let $p = \min(m, n)$ and let $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be a thin singular value decomposition. For an odd polynomial $g(x) = \sum_{k=0}^{N} a_k x^{2k+1}$, its matrix map is

$$F_g(\mathbf{P}) = \mathbf{U} g(\mathbf{\Sigma}) \mathbf{V}^\top = \sum_{k=0}^{N} a_k \mathbf{P} (\mathbf{P}^\top \mathbf{P})^k.$$

The polynomial expression is smooth in the matrix entries, including at repeated and zero singular values. Its derivative separates singular-value changes, rotations of singular-vector pairs, and unmatched rectangular directions. General formulas for rectangular singular value maps are available (Noferini, 2017; Ding et al., 2018). Appendix C applies them to Muon and then differentiates the normalization.

C Derivative of the Practical Muon Transformation

Let $\mathbf{P} = \mathbf{U}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_p)\mathbf{V}^\top$ have rank $p = \min(m, n)$. Let ν be one of the rules in Equation 1, and set $s_i = \sigma_i / \nu(\mathbf{P})$. For the clamp rule, assume $\|\mathbf{P}\|_{\mathrm{F}} \neq \epsilon_{\mathrm{cl}}$.

In singular-vector coordinates, diagonal changes have the form $\mathbf{U}\,\mathrm{diag}(\mathbf{z})\mathbf{V}^\top$. A symmetric pair changes entries (i, j) and (j, i) equally, while a rotational pair changes them with opposite signs. For a rectangular matrix, perturbations can also involve a left or right direction outside the columns of $\mathbf{U}$ or $\mathbf{V}$. A derivative gain is the scalar by which the derivative multiplies a perturbation in an eigendirection. The derivative formula for real rectangular matrix functions describes these changes before normalization (Noferini, 2017, Cor. 3.10 and Thm. 4.5). We first differentiate the singular value map, then the normalization, and combine them by the chain rule.

Theorem 4 (Eigenvalues of the Muon derivative)**.** *For $\mathbf{R} = \mathrm{D}\Psi_\nu(\mathbf{P})$, the eigenvalues in the symmetric and rotational (i, j) directions, $i < j$, are, respectively,*

$$\frac{\chi_{m,n}}{\nu(\mathbf{P})}\frac{g(s_i) - g(s_j)}{s_i - s_j}, \qquad \frac{\chi_{m,n}}{\nu(\mathbf{P})}\frac{g(s_i) + g(s_j)}{s_i + s_j}. \tag{9}$$

Repeated singular values are handled by continuity. Each unmatched rectangular direction associated with s_i has eigenvalue

$$\frac{\chi_{m,n}}{\nu(\mathbf{P})}\frac{g(s_i)}{s_i}.$$

For Ψ_δ with $\delta > 0$, the positive and negative counts on the diagonal subspace equal the corresponding counts among $g'(s_1), \ldots, g'(s_p)$. For Ψ_{cl} below the clamp threshold, the diagonal eigenvalues are $\chi_{m,n}g'(s_i)/\epsilon_{\mathrm{cl}}$. For Ψ_0 and for Ψ_{cl} above the clamp threshold, the radial eigenvalue is zero. The remaining diagonal eigenvalues are $\chi_{m,n}/\|\mathbf{P}\|_{\mathrm{F}}$ times the eigenvalues of the compression of $\mathrm{diag}(g'(s_i))$ to

$$\left\{\mathbf{z} \in \mathbb{R}^p : \sum_{i=1}^{p} \sigma_i z_i = 0\right\}.$$

Here compression means applying $\mathrm{diag}(g'(s_i))$ and then projecting orthogonally back onto this subspace.

The derivative formula also gives a simple counting bound for the matrix survey. The diagonal subspace has dimension p, and symmetric pairs have dimension $p(p-1)/2$. Since $g(s) > 0$ for $s > 0$, the rotational and unmatched rectangular gains are positive. Thus no more than $p(p+1)/2$ eigenvalues of $\mathbf{R}$ can be negative. Writing the model width as D, the four surveyed shapes at each depth are $D \times 3D$, $D \times D$, $D \times 4D$, and $4D \times D$. They have $12D^2$ coordinates in total, and the bound is $2D(D+1)$ negative directions, a fraction $(D+1)/(6D)$. At the 58M width $D = 512$, this is about 16.7%. The observed 7–8% also depends on the singular values of each proposal and cannot be inferred from this dimension count alone.

Corollary 5 (Signs of the Muon derivative at an isotropic proposal)**.** *Let $p = \min(m, n) \geq 2$, let $\mathbf{P}$ have p equal singular values $\sigma > 0$, and set $s = \sigma/\nu(\mathbf{P})$. Assume $g'(s) < 0$ and $g(s) > 0$. For Ψ_δ with $\delta > 0$, and for Ψ_{cl} below the clamp threshold, $\mathbf{R}$ has $p(p+1)/2$ negative eigenvalues and $p(p-1)/2 + p|m-n|$ positive eigenvalues. For Ψ_0, and for Ψ_{cl} above the clamp threshold, $\mathbf{R}$ has $p(p+1)/2 - 1$ negative eigenvalues, $p(p-1)/2 + p|m-n|$ positive eigenvalues, and one zero eigenvalue. In all four cases, $\mathbf{R}$ is self-adjoint. If $\mathbf{H} \succ 0$, then $\mathbf{K} = \alpha\mathbf{R}\mathbf{H}$ has the same numbers of positive, negative, and zero eigenvalues as $\mathbf{R}$. If $\mathbf{H} \succeq 0$, every nonzero eigenvalue of $\mathbf{K}$ is real, and both signs occur exactly when the quadratic form of $\mathbf{R}$ takes both signs on* $\mathrm{range}(\mathbf{H})$.

C.1 How the Polynomial Changes Singular Values

For $\mathbf{X}_0 = \mathbf{U}\,\mathrm{diag}(s_1, \ldots, s_p)\mathbf{V}^\top$, orthonormality of the columns of $\mathbf{U}$ and $\mathbf{V}$ gives

$$(\mathbf{X}_0\mathbf{X}_0^\top)\mathbf{X}_0 = \mathbf{U}\,\mathrm{diag}(s_1^3, \ldots, s_p^3)\mathbf{V}^\top,$$
$$(\mathbf{X}_0\mathbf{X}_0^\top)^2\mathbf{X}_0 = \mathbf{U}\,\mathrm{diag}(s_1^5, \ldots, s_p^5)\mathbf{V}^\top.$$

Thus one Newton–Schulz step applies ϕ to each singular value while preserving $\mathbf{U}$ and $\mathbf{V}$. Repeating it five times gives

$$\mathbf{X}_5 = \mathbf{U}\,\mathrm{diag}(g(s_1), \ldots, g(s_p))\mathbf{V}^\top.$$

For the implementation coefficients, $\gamma_1 > 0$, $\gamma_5 > 0$, and $\gamma_3^2 - 4\gamma_1\gamma_5 < 0$. Hence $\gamma_1 + \gamma_3 y + \gamma_5 y^2 > 0$ for every real y, and $\phi(x) > 0$ for $x > 0$. Rotation and unmatched rectangular directions therefore have positive derivative eigenvalue. Diagonal and symmetric directions depend on scalar derivatives and divided differences, which can be negative.

Lemma 6 (Derivative of an odd singular value map)**.** *Let g be an odd polynomial and let $p = \min(m, n)$. Write $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{m\times p}$, $\mathbf{V} \in \mathbb{R}^{n\times p}$, and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$. For a perturbation $\mathbf{E}$, set $\mathbf{Z} = \mathbf{U}^\top\mathbf{E}\mathbf{V}$. The matrix $\mathbf{U}^\top\mathrm{D}F_g(\mathbf{P})[\mathbf{E}]\mathbf{V} \in \mathbb{R}^{p\times p}$ has diagonal entries $g'(\sigma_i)Z_{ii}$ and off-diagonal entries*

$$\frac{g(\sigma_i) - g(\sigma_j)}{\sigma_i - \sigma_j}\frac{Z_{ij} + Z_{ji}}{2} + \frac{g(\sigma_i) + g(\sigma_j)}{\sigma_i + \sigma_j}\frac{Z_{ij} - Z_{ji}}{2}. \tag{10}$$

If $m > p$, choose an orthonormal complement $\mathbf{U}_\perp$. Then

$$\mathbf{U}_\perp^\top \mathrm{D}F_g(\mathbf{P})[\mathbf{E}]\mathbf{V} = (\mathbf{U}_\perp^\top \mathbf{E}\mathbf{V})\,\mathrm{diag}\left(\frac{g(\sigma_i)}{\sigma_i}\right).$$

If $n > p$, choose an orthonormal complement $\mathbf{V}_\perp$. Then

$$\mathbf{U}^\top \mathrm{D}F_g(\mathbf{P})[\mathbf{E}]\mathbf{V}_\perp = \mathrm{diag}\left(\frac{g(\sigma_i)}{\sigma_i}\right)(\mathbf{U}^\top \mathbf{E}\mathbf{V}_\perp).$$

Every quotient is interpreted by its continuous limit. In particular, $g(\sigma)/\sigma$ tends to $g'(0)$, and a divided difference at a repeated singular value tends to the corresponding derivative. These coordinate subspaces are orthogonal in the Frobenius inner product, so $\mathrm{D}F_g(\mathbf{P})$ is self-adjoint.

Proof. Form the symmetric dilation

$$\mathrm{dil}(\mathbf{P}) = \begin{pmatrix} 0 & \mathbf{P} \\ \mathbf{P}^\top & 0 \end{pmatrix}.$$

For odd g, the upper-right block of $g(\mathrm{dil}(\mathbf{P}))$ is $F_g(\mathbf{P})$. The eigenvalues paired with $\sigma_i > 0$ are σ_i and $-\sigma_i$. The derivative of a polynomial on a symmetric matrix multiplies a perturbation between eigenspaces with eigenvalues a and b by $(g(a) - g(b))/(a - b)$ when $a \neq b$, and by $g'(a)$ when $a = b$ (Noferini, 2017, Thm. 2.11). Pairs (σ_i, σ_j) and $(\sigma_i, -\sigma_j)$ give the two divided differences in Equation 10. Pairs between a nonzero singular subspace and a zero eigenspace give $g(\sigma_i)/\sigma_i$. Pairs inside a zero singular subspace have the limiting gain $g'(0)$, so the result is independent of the singular vectors chosen in that subspace. Because g is an odd polynomial, $g(0) = 0$ and the limits above exist. The derivative is therefore well defined at repeated and zero singular values. These gains agree with the real rectangular generalized Daleckii–Krein formula (Noferini, 2017, Cor. 3.10 and Thm. 4.5). □

C.2 Normalization and the Muon Derivative

First consider

$$N_\delta(\mathbf{P}) = \frac{\mathbf{P}}{\|\mathbf{P}\|_{\mathrm{F}} + \delta}.$$

Let $r = \|\mathbf{P}\|_{\mathrm{F}} > 0$ and $\mathbf{u} = \mathbf{P}/r$. For a perturbation $\mathbf{E}$,

$$\mathrm{D}N_\delta(\mathbf{P})[\mathbf{E}] = \frac{\mathbf{E}}{r + \delta} - \frac{\mathbf{P}\langle \mathbf{P}, \mathbf{E}\rangle_{\mathrm{F}}}{r(r + \delta)^2}.$$

Equivalently,

$$\mathrm{D}N_\delta(\mathbf{P}) = \frac{1}{r + \delta}(\mathbf{I} - \mathbf{u} \otimes \mathbf{u}) + \frac{\delta}{(r + \delta)^2}(\mathbf{u} \otimes \mathbf{u}), \tag{11}$$

where $(\mathbf{u} \otimes \mathbf{u})[\mathbf{E}] = \mathbf{u}\langle \mathbf{u}, \mathbf{E}\rangle_{\mathrm{F}}$. The eigenvalue is $1/(r + \delta)$ on directions orthogonal to $\mathbf{P}$ and $\delta/(r + \delta)^2$ in the radial direction.

When $\|\mathbf{P}\|_{\mathrm{F}} > \epsilon_{\mathrm{cl}}$, clamp normalization has derivative

$$\mathrm{D}\left(\frac{\mathbf{P}}{\|\mathbf{P}\|_{\mathrm{F}}}\right) = \frac{1}{r}(\mathbf{I} - \mathbf{u} \otimes \mathbf{u}).$$

This is Equation 11 with $\delta = 0$. When $\|\mathbf{P}\|_{\mathrm{F}} < \epsilon_{\mathrm{cl}}$, its derivative is $\mathbf{I}/\epsilon_{\mathrm{cl}}$. These formulas exclude the clamp boundary.

If N denotes the chosen normalization and $g = \phi^{\circ 5}$, the chain rule gives

$$\mathbf{R} = \mathrm{D}\Psi(\mathbf{P}) = \chi_{m,n}\mathrm{D}F_g(N(\mathbf{P}))\,\mathrm{D}N(\mathbf{P}).$$

A proposal perturbation therefore passes first through $\mathrm{D}N(\mathbf{P})$ and then through $\mathrm{D}F_g(N(\mathbf{P}))$.

C.3 Proof of Theorem 4

Proof. Let $\mathbf{e}_i$ be the ith coordinate vector. The symmetric and rotational directions

$$\frac{1}{\sqrt{2}}\mathbf{U}(\mathbf{e}_i\mathbf{e}_j^\top \pm \mathbf{e}_j\mathbf{e}_i^\top)\mathbf{V}^\top$$

are orthogonal to $\mathbf{P}$. Under either normalization rule, the normalization derivative multiplies each of these directions by $1/\nu(\mathbf{P})$. Lemma 6 then gives the two gains in Equation 9. Every unmatched rectangular direction is also orthogonal to $\mathbf{P}$, so the same argument gives its stated gain.

It remains to study diagonal changes. In singular-vector coordinates, let

$$\mathbf{D}_g = \operatorname{diag}(g'(s_1), \dots, g'(s_p)).$$

For N_δ with $\delta > 0$, the diagonal restriction $\mathbf{B}_\mathrm{d}$ of the normalization derivative is positive definite. The product $\mathbf{D}_g\mathbf{B}_\mathrm{d}$ is similar to the symmetric matrix $\mathbf{B}_\mathrm{d}^{1/2}\mathbf{D}_g\mathbf{B}_\mathrm{d}^{1/2}$, which is congruent to $\mathbf{D}_g$. Sylvester's law of inertia gives the stated sign counts (Horn & Johnson, 2012, Thm. 4.5.8).

When $\|\mathbf{P}\|_\mathrm{F} < \epsilon_\mathrm{cl}$, $\mathbf{B}_\mathrm{d} = \mathbf{I}/\epsilon_\mathrm{cl}$. The diagonal gains are therefore $\chi_{m,n}g'(s_i)/\epsilon_\mathrm{cl}$.

When $\|\mathbf{P}\|_\mathrm{F} > \epsilon_\mathrm{cl}$, let $\mathbf{u}_\mathrm{d} = (\sigma_1, \dots, \sigma_p)^\top/\|\mathbf{P}\|_\mathrm{F}$ and $\mathbf{\Pi}_\mathrm{d} = \mathbf{I} - \mathbf{u}_\mathrm{d}\mathbf{u}_\mathrm{d}^\top$. The diagonal normalization derivative is $\mathbf{\Pi}_\mathrm{d}/\|\mathbf{P}\|_\mathrm{F}$. In an orthonormal basis beginning with $\mathbf{u}_\mathrm{d}$, the product $\mathbf{D}_g\mathbf{\Pi}_\mathrm{d}$ has the block form

$$\begin{pmatrix} 0 & * \\ 0 & (\mathbf{\Pi}_\mathrm{d}\mathbf{D}_g\mathbf{\Pi}_\mathrm{d})|_{\mathbf{u}_\mathrm{d}^\perp} \end{pmatrix}.$$

Its eigenvalues are the radial zero and the eigenvalues of the symmetric compression, counted with algebraic multiplicity. Multiplying by $\chi_{m,n}/\|\mathbf{P}\|_\mathrm{F}$ gives the stated formula. The same calculation applies to additive normalization at $\delta = 0$. □

C.4 Proof of Corollary 5

Proof. Write $\mathbf{P} = \sigma\mathbf{U}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{m\times p}$ and $\mathbf{V} \in \mathbb{R}^{n\times p}$ have orthonormal columns. Let $r = \|\mathbf{P}\|_\mathrm{F} = \sigma\sqrt{p}$ and $s = \sigma/\nu(\mathbf{P})$. In singular-vector coordinates, the matrix space splits orthogonally into the radial direction $\operatorname{span}\{\mathbf{P}\}$, symmetric trace-free directions, skew-symmetric directions, and unmatched rectangular directions.

The derivative of the five-step scalar transformation has gain $g'(s)$ on the radial and symmetric trace-free directions. It has gain $g(s)/s$ on the skew-symmetric and unmatched directions. The multiplicities of these three groups are

$$1, \qquad \frac{p(p+1)}{2} - 1, \qquad \frac{p(p-1)}{2} + p|m-n|.$$

For Ψ_δ with $\delta > 0$, the normalization gains are $\delta/(r+\delta)^2$ in the radial direction and $1/(r+\delta)$ in every orthogonal direction. For Ψ_cl below the clamp threshold, every normalization gain is $1/\epsilon_\mathrm{cl}$. Hence $g'(s) < 0$ gives $p(p+1)/2$ negative eigenvalues, while $g(s) > 0$ gives $p(p-1)/2 + p|m-n|$ positive eigenvalues.

For Ψ_0, and for Ψ_cl above the clamp threshold, the normalization derivative is $(\mathbf{I} - \mathbf{u}\otimes\mathbf{u})/r$, where $\mathbf{u} = \mathbf{P}/r$. The radial gain is zero. The symmetric trace-free gains have the sign of $g'(s)$, and the remaining gains have the sign of $g(s)$. This gives the second set of counts. In every case, the singular value derivative and the normalization derivative preserve the same orthogonal decomposition, so $\mathbf{R}$ is self-adjoint.

If $\mathbf{H} \succ 0$, then $\mathbf{RH}$ is similar to $\mathbf{H}^{1/2}\mathbf{R}\mathbf{H}^{1/2}$. This symmetric matrix is congruent to $\mathbf{R}$, so Sylvester's law of inertia gives the same sign counts (Horn & Johnson, 2012, Thm. 4.5.8). If $\mathbf{H} \succeq 0$, the nonzero eigenvalues of $\mathbf{RH}$ equal those of $\mathbf{H}^{1/2}\mathbf{R}\mathbf{H}^{1/2}$ by the **AB–BA** identity (Horn & Johnson, 2012, Thm. 1.3.22). They are real. Both signs occur exactly when the quadratic form of $\mathbf{R}$ takes both signs on $\operatorname{range}(\mathbf{H}^{1/2}) = \operatorname{range}(\mathbf{H})$. □

C.5 The Exact Polar Map

Lemma 7 (Derivative of the exact polar factor)**.** *Let* $\mathbf{P}$ *have full rank on its smaller dimension, let* $\alpha > 0$*, and define*

$$\Psi_{\mathrm{polar}}(\mathbf{P}) = \chi_{m,n}\,\mathrm{polar}(\mathbf{P}).$$

Its Fréchet derivative is self-adjoint and positive semidefinite in the Frobenius inner product. If $\mathbf{H}$ *is positive semidefinite, every nonzero eigenvalue of*

$$\alpha \mathrm{D}\Psi_{\mathrm{polar}}(\mathbf{P})\mathbf{H}$$

is real and positive.

Proof. For $m \geq n$, the derivative of the full-rank polar factor has zero gain on symmetric directions within the square singular-vector block, rotational gains $2/(\sigma_i + \sigma_j)$, and unmatched rectangular gains $1/\sigma_i$ (Noferini, 2017, Cor. 3.12). The case $m < n$ follows by applying the result to $\mathbf{P}^\top$ and transposing. Since $\chi_{m,n} > 0$, every gain of $\mathrm{D}\Psi_{\mathrm{polar}}(\mathbf{P})$ is nonnegative. The derivative is therefore self-adjoint and positive semidefinite.

If $\mathbf{H} \succeq 0$, the nonzero eigenvalues of $\mathrm{D}\Psi_{\mathrm{polar}}(\mathbf{P})\mathbf{H}$ equal those of

$$\mathbf{H}^{1/2}\mathrm{D}\Psi_{\mathrm{polar}}(\mathbf{P})\mathbf{H}^{1/2}$$

by the $\mathbf{AB}$–$\mathbf{BA}$ identity (Horn & Johnson, 2012, Thm. 1.3.22). This matrix is symmetric and positive semidefinite. Its nonzero eigenvalues are positive, and multiplication by $\alpha > 0$ preserves their signs. □

C.6 When Positive Curvature Can Give a Negative Mode

Proof of Proposition 2. Write $\mathbf{S} = (\mathbf{R} + \mathbf{R}^*)/2$ for the self-adjoint part of $\mathbf{R}$. If $\langle \mathbf{E}, \mathbf{R}[\mathbf{E}]\rangle_{\mathrm{F}} \geq 0$ for every real $\mathbf{E}$, then $\mathbf{S} \succeq 0$. For any $\mathbf{H} \succeq 0$, the nonzero eigenvalues of $\mathbf{RH}$ equal those of $\mathbf{H}^{1/2}\mathbf{R}\mathbf{H}^{1/2}$ by the $\mathbf{AB}$–$\mathbf{BA}$ identity (Horn & Johnson, 2012, Thm. 1.3.22). If $\mathbf{z}$ is a complex eigenvector of the latter matrix with eigenvalue λ, then

$$\mathrm{Re}\,\lambda\,\|\mathbf{z}\|^2 = \mathbf{z}^*\mathbf{H}^{1/2}\mathbf{S}\mathbf{H}^{1/2}\mathbf{z} \geq 0.$$

Conversely, take a unit $\mathbf{E}$ with $q = \langle \mathbf{E}, \mathbf{R}[\mathbf{E}]\rangle_{\mathrm{F}} < 0$ and set $\mathbf{H} = t(\mathbf{E} \otimes \mathbf{E})$ for $t > 0$. Then $\mathbf{RH}[\mathbf{R}[\mathbf{E}]] = tq\,\mathbf{R}[\mathbf{E}]$, so αtq is a negative real eigenvalue of $\alpha\mathbf{RH}$. Choosing $t > d/(-\alpha q)$ places it beyond the lower edge. A sufficiently small positive ridge can make $\mathbf{H}$ positive definite while preserving this negative eigenvalue by continuity. □

The distinction between scalar monotonicity and the quadratic form is visible in a 2×2 example. Normalize a positive diagonal proposal by its Frobenius norm and write its normalized singular values as $s_1, s_2 > 0$ with $s_1^2 + s_2^2 = 1$. Let g be five iterations of the classical cubic $x \mapsto (3x - x^3)/2$. It is increasing on $(0, 1)$, so $d_i = g'(s_i) > 0$. On diagonal perturbations, the derivative is a positive multiple of $\mathrm{diag}(d_1, d_2)(\mathbf{I} - \mathbf{s}\mathbf{s}^\top)$. Its self-adjoint part has determinant

$$-\frac{1}{4}s_1^2 s_2^2 (d_1 - d_2)^2,$$

which is negative whenever $d_1 \neq d_2$. Thus normalization coupled to even a monotone nonlinear scalar map can give a negative quadratic direction. For the implemented quintic, negative scalar derivatives provide an additional direct source, as the isotropic example in Section 3 shows.

D Proof of Theorem 1

D.1 Proof of Theorem 1

Collecting the coefficients of $(\Delta\mathbf{W}, \Delta\mathbf{M})$ yields

$$\mathbf{A} = \begin{pmatrix} (1-d)\mathbf{I} - (1-\beta^2)\mathbf{K} & -\alpha\beta^2\mathbf{R} \\ (1-\beta)\mathbf{H} & \beta\mathbf{I} \end{pmatrix}, \qquad \mathbf{K} = \alpha\mathbf{R}\mathbf{H}.$$

The Jacobian satisfies the characteristic identity

$$\det(\mu\mathbf{I} - \mathbf{A}) = \det((\mu - \beta)(\mu - 1 + d)\mathbf{I} + (1 - \beta)((1 + \beta)\mu - \beta)\mathbf{K})\,.$$

Proof. Let n_W be the dimension of the vectorized weight space. For $\mu \neq \beta$, the Schur complement of the lower right block of $\mu\mathbf{I} - \mathbf{A}$ gives

$$\begin{aligned}\det(\mu\mathbf{I} - \mathbf{A}) &= (\mu - \beta)^{nw} \det\left((\mu - 1 + d)\mathbf{I} + (1 - \beta^2)\mathbf{K} + \frac{\beta^2(1-\beta)}{\mu - \beta}\mathbf{K}\right) \\ &= \det((\mu - \beta)(\mu - 1 + d)\mathbf{I} + (1 - \beta)((1 + \beta)\mu - \beta)\mathbf{K}) \,.\end{aligned}$$

The characteristic determinant on the left and the final determinant are polynomials in μ. They agree for $\mu \neq \beta$, so they agree at $\mu = \beta$ as well. A Schur triangularization of $\mathbf{K}$ (Horn & Johnson, 2012, Thm. 2.3.1) then factors the determinant into one copy of Equation 3 for each eigenvalue κ of $\mathbf{K}$, counted with algebraic multiplicity.

Write the modal polynomial as $f_\kappa(\mu) = \mu^2 + c_1\mu + c_0$, where

$$\begin{aligned}c_1 &= d - 1 - \beta + (1 - \beta^2)\kappa, \\ c_0 &= \beta(1 - d) - \beta(1 - \beta)\kappa.\end{aligned}$$

For this real quadratic, set $\mu = (1 + \omega)/(1 - \omega)$, which maps the open left half-plane to the open unit disk. Then

$$(1 - \omega)^2 f_\kappa\left(\frac{1 + \omega}{1 - \omega}\right) = (1 - c_1 + c_0)\omega^2 + 2(1 - c_0)\omega + (1 + c_1 + c_0).$$

If both roots of f_κ lie in the disk, then $f_\kappa(-1) = 1 - c_1 + c_0 > 0$. A real quadratic with positive leading coefficient has both roots in the open left half-plane exactly when its other two coefficients are positive. We obtain the three Jury conditions (Jury, 1964). Substitution gives

$$\begin{aligned}(1 - \beta)(d + \kappa) &> 0, \\ (1 + \beta)(2 - d) - (1 - \beta)(1 + 2\beta)\kappa &> 0, \\ 1 - \beta + \beta d + \beta(1 - \beta)\kappa &> 0.\end{aligned}$$

The first two inequalities give Equation 4. Under $\kappa > -d$, the third left-hand side is at least $1 - \beta + \beta^2 d > 0$. At $\kappa = -d$ and $\kappa = U_{\beta,d}$, substitution gives the unit roots $+1$ and -1. Their companion roots are, respectively,

$$\beta(1 - \beta d), \qquad \frac{\beta(1 + \beta d)}{1 + 2\beta}.$$

Both have modulus below one, so each unit root is simple. Differentiating the modal equation gives

$$\left.\frac{\mathrm{d}\mu}{\mathrm{d}\kappa}\right|_{\kappa=-d} = -\frac{1 - \beta}{1 - \beta + \beta^2 d} < 0, \qquad \left.\frac{\mathrm{d}\mu}{\mathrm{d}\kappa}\right|_{\kappa=U_{\beta,d}} = -\frac{(1 - \beta)(1 + 2\beta)^2}{1 + 3\beta + \beta^2 d} < 0.$$

Thus both real modes cross the unit circle with nonzero slope. □

D.2 Boundary for Complex Eigenvalues

Proposition 8 (Boundary for complex eigenvalues)**.** *Assume $0 \leq \beta < 1$ and $0 \leq d < 2$. Define the stable region as the set of $\kappa \in \mathbb{C}$ for which both roots of Equation 3 lie in the open unit disk. Its boundary is*

$$\kappa_{\beta,d}(\theta) = -\frac{(e^{i\theta} - \beta)(e^{i\theta} - 1 + d)}{(1 - \beta)((1 + \beta)e^{i\theta} - \beta)}, \qquad 0 \leq \theta < 2\pi. \tag{12}$$

The only real points on this boundary are $-d$ and $U_{\beta,d}$.

Proof. Let $z = e^{i\theta}$ be one modal root. The product of the two roots is $\beta(1 - d) - \beta(1 - \beta)\kappa$, so the other root is

$$\mu_2(z) = \frac{\beta(1 - d) - \beta(1 - \beta)\kappa_{\beta,d}(\theta)}{z} = \frac{\beta(z - \beta d)}{(1 + \beta)z - \beta}.$$

For $t = \operatorname{Re} z \in [-1, 1]$,

$$|\mu_2(z)|^2 = \frac{\beta^2(1 + \beta^2 d^2 - 2\beta d t)}{1 + 2\beta + 2\beta^2 - 2\beta(1 + \beta)t}.$$

The denominator is $|(1+\beta)z-\beta|^2$ and is strictly positive on the unit circle. Both the numerator and denominator are affine in t, and the denominator is positive. The derivative of their ratio has constant sign or is zero, so the maximum on $[-1,1]$ occurs at an endpoint. At $z=1$ and $z=-1$,

$$|\mu_2(1)| = \beta|1-\beta d| < 1, \qquad |\mu_2(-1)| = \frac{\beta(1+\beta d)}{1+2\beta} < 1$$

for $0<\beta<1$ and $0 \le d < 2$. When $\beta = 0$, the second root is zero. Thus the second root remains inside the unit disk along the curve.

For small ε, replace z by $(1+\varepsilon)z$ in the right-hand side of Equation 12 and call the result κ_ε. Then $(1+\varepsilon)z$ is one modal root at κ_ε. The other root varies continuously and remains inside the unit disk. For $\varepsilon < 0$, both roots lie inside. For $\varepsilon > 0$, one lies outside. As $\varepsilon \to 0$, κ_ε reaches the displayed curve, so every point on it is a boundary point. Conversely, if neither root has unit modulus, root continuity gives a neighborhood with the same number of roots inside the disk (Horn & Johnson, 2012, Appendix D). Thus Equation 12 traces the full boundary.

If κ is real, the modal polynomial has real coefficients. A nonreal unit root would have its conjugate as the other root, contradicting the strict bound on the second root proved above. Hence a real boundary point has unit root $+1$ or -1. Substituting these two values into Equation 3 gives

$$\kappa = -d \qquad \text{or} \qquad \kappa = U_{\beta,d}.$$

Both values occur on the boundary, so they are its only real points. □

E Proof of Theorem 3

Proof. For a fixed $c > 0$, scale invariance holds for every nonzero $\mathbf{Q}$ near $\mathbf{P}$. Differentiating $\Psi(c\mathbf{Q}) = \Psi(\mathbf{Q})$ at $\mathbf{Q} = \mathbf{P}$ in an arbitrary direction $\mathbf{E}$ gives $\mathrm{D}\Psi(c\mathbf{P})[c\mathbf{E}] = \mathrm{D}\Psi(\mathbf{P})[\mathbf{E}]$. Hence $\mathrm{D}\Psi(c\mathbf{P}) = c^{-1}\mathrm{D}\Psi(\mathbf{P})$ and $\mathbf{K}(c) = c^{-1}\mathbf{K}(1)$. Substituting Equation 5 into the Nesterov proposal gives $c\mathbf{P}$. The loss, gradient, Hessian, transformed update, weight step, and next weight therefore remain fixed. The next momentum buffer changes with c.

The frozen Jacobian along this family is

$$\mathbf{A}_c = \begin{pmatrix} (1-d)\mathbf{I} - (1-\beta^2)\mathbf{K}(1)/c & -\alpha\beta^2\mathrm{D}\Psi(\mathbf{P})/c \\ (1-\beta)\mathbf{H} & \beta\mathbf{I} \end{pmatrix}.$$

If $\mathbf{K}(1)$ has a nonzero eigenvalue κ, the sum of the two associated roots of Equation 3 contains the term $-(1-\beta^2)\kappa/c$ and diverges as $c \to 0$. The two roots therefore cannot both remain inside the unit disk.

When $0 < d < 2$, $\mathbf{A}_c$ converges as $c \to \infty$ to a block triangular matrix whose eigenvalues are $1-d$ and β. Both lie inside the unit disk. Continuity of matrix eigenvalues then gives Schur stability for all sufficiently large c (Horn & Johnson, 2012, Appendix D). □

Learning rate at a fixed state. With the state and batch fixed, $\mathbf{R}$ and $\mathbf{H}$ do not change with α. Since $d = \alpha\lambda_{\mathrm{wd}}$, the lower inequality is

$$-d < \kappa \quad \Longleftrightarrow \quad -\lambda_{\mathrm{wd}} < \kappa/\alpha.$$

During training, changing the learning rate also changes future weights and momentum, and therefore the later operators.

F Two Stable Fixed Points at the Lower Edge

The idea is to choose a quadratic whose curvature along the direction $\mathbf{E}_c$ can be changed by a scalar ζ. At $\zeta = 1$, the corresponding mode reaches the lower edge. We solve the fixed point equation on an invariant plane and check all eight state multipliers with Theorem 1.

At a fixed point, the momentum and proposal both equal $\mathbf{G}(\mathbf{W}_\star)$. To relate the fixed-point equation to the full state Jacobian, solve $(\mathbf{A}_\star - \mathbf{I})(\mathbf{v}, \mathbf{m}) = 0$. Its momentum row gives

$$(1-\beta)\mathbf{H}_\star\mathbf{v} + (\beta-1)\mathbf{m} = 0, \qquad \mathbf{m} = \mathbf{H}_\star\mathbf{v}.$$

Substitution into the weight row gives

$$-d\mathbf{v} - \alpha(1-\beta^2)\mathbf{R}_\star\mathbf{H}_\star\mathbf{v} - \alpha\beta^2\mathbf{R}_\star\mathbf{H}_\star\mathbf{v} = -(d\mathbf{I} + \mathbf{K}_\star)\mathbf{v} = 0.$$

Thus a nonzero $\mathbf{v}$ lies in the kernel of the linearized fixed-point equation exactly when $(\mathbf{v}, \mathbf{H}_\star\mathbf{v})$ is a $+1$ state eigenvector. We now solve for two new fixed points at this crossing and apply Theorem 1 to prove their stability.

Theorem 9 (Stable fixed points beyond the lower edge)**.** *Fix $0 \le \beta < 1$, $\alpha > 0$, and $\lambda_{\mathrm{wd}} > 0$ with $0 < d = \alpha\lambda_{\mathrm{wd}} < 2$. Consider the clamp-normalized transformation Ψ_{cl} with the Newton–Schulz coefficients and five iterations in Section 2. There exist a family of globally strongly convex quadratic losses $\mathcal{L}_\zeta$ on 2×2 matrices and a fixed point $\mathbf{z}_\star = (\mathbf{W}_\star, \mathbf{M}_\star)$ whose proposal satisfies $\|\mathbf{P}_\star\|_{\mathrm{F}} > \epsilon_{\mathrm{cl}}$. At $\zeta = 1$, the weight and momentum Jacobian has one simple multiplier $+1$, while every other multiplier lies inside the unit disk. For $\zeta < 1$ close to one, $\mathbf{z}_\star$ is locally exponentially stable. For $\zeta > 1$ close to one, $\mathbf{z}_\star$ is unstable and two locally exponentially stable fixed points $\mathbf{z}_\pm(\zeta)$ exist. In the coordinate x along $\mathbf{E}_c$,*

$$x_\pm(\zeta) = \pm C_{\mathrm{pf}}\sqrt{\zeta - 1} + O((\zeta-1)^{3/2}), \qquad C_{\mathrm{pf}} > 0.$$

The proposals at both new fixed points also satisfy $\|\mathbf{P}\|_{\mathrm{F}} > \epsilon_{\mathrm{cl}}$.

F.1 The Quadratic Loss and Its Critical Multiplier

Use the exact rational form of the implementation coefficients,

$$\phi(x) = \frac{6889}{2000}x - \frac{191}{40}x^3 + \frac{4063}{2000}x^5, \qquad g = \phi^{\circ 5}.$$

Define the Frobenius orthonormal basis

$$\mathbf{E}_r = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, \quad \mathbf{E}_c = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix},$$
$$\mathbf{E}_s = \frac{1}{\sqrt{2}}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \quad \mathbf{E}_a = \frac{1}{\sqrt{2}}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}.$$

Choose $r_0 > \epsilon_{\mathrm{cl}}$, set $\mathbf{P}_\star = r_0\mathbf{E}_r$ and $s = 1/\sqrt{2}$, and define

$$a_1 = g'(s), \qquad b_r = -\frac{1}{2}g'(s) + \frac{1}{2\sqrt{2}}g''(s), \qquad a_3 = -\frac{1}{2}g'(s) - \frac{1}{2\sqrt{2}}g''(s) + \frac{1}{12}g'''(s). \tag{13}$$

Lemma 10 below proves $g(s) > 0$, $a_1 < 0$, $b_r < 0$, and $a_3 > 0$. It also gives the derivative eigenvalues at $\mathbf{P}_\star$: zero on $\mathbf{E}_r$, a_1/r_0 on $\mathbf{E}_c$ and $\mathbf{E}_s$, and $\sqrt{2}g(s)/r_0$ on $\mathbf{E}_a$.

Let

$$h_0 = -\frac{\lambda_{\mathrm{wd}} r_0}{a_1} > 0, \qquad 0 < h_s < h_0, \qquad h_r > 0,$$

and choose $h_a > 0$ so that $\alpha\sqrt{2}g(s)h_a/r_0 < U_{\beta,d}$. Here h_0 is the curvature that places the $\mathbf{E}_c$ mode on the lower edge at $\zeta = 1$. For $\zeta > 0$, define

$$\mathbf{H}_\zeta = \zeta h_0\mathbf{E}_c \otimes \mathbf{E}_c + h_r\mathbf{E}_r \otimes \mathbf{E}_r + h_s\mathbf{E}_s \otimes \mathbf{E}_s + h_a\mathbf{E}_a \otimes \mathbf{E}_a.$$

Set

$$\mathbf{W}_\star = -\frac{\sqrt{2}g(s)}{\lambda_{\mathrm{wd}}}\mathbf{E}_r, \qquad \mathbf{W}_{\min} = \mathbf{W}_\star - \frac{r_0}{h_r}\mathbf{E}_r,$$

and use the quadratic loss

$$\mathcal{L}_\zeta(\mathbf{W}) = \frac{1}{2}\langle \mathbf{W} - \mathbf{W}_{\min}, \mathbf{H}_\zeta(\mathbf{W} - \mathbf{W}_{\min})\rangle_{\mathrm{F}}, \qquad \mathbf{G}_\zeta(\mathbf{W}) = \mathbf{H}_\zeta(\mathbf{W} - \mathbf{W}_{\min}).$$

This loss is globally strongly convex. Since $\mathbf{G}_\zeta(\mathbf{W}_\star) = \mathbf{P}_\star$ and $\lambda_{\mathrm{wd}}\mathbf{W}_\star + \Psi_{\mathrm{cl}}(\mathbf{P}_\star) = 0$, the state $\mathbf{z}_\star = (\mathbf{W}_\star, \mathbf{P}_\star)$ is fixed for every $\zeta > 0$.

At this state, the four eigenvalues of $\mathbf{K}$ are

$$\kappa_c(\zeta) = -\zeta d, \qquad \kappa_s = -d\frac{h_s}{h_0} \in (-d, 0), \qquad \kappa_r = 0, \qquad \kappa_a = \frac{\alpha\sqrt{2}g(s)}{r_0}h_a \in (0, U_{\beta,d}).$$

Only κ_c reaches an edge at $\zeta = 1$. Theorem 1 gives one simple multiplier $+1$ and seven multipliers inside the unit disk. It gives stability for $\zeta < 1$ near one and instability for $\zeta > 1$ near one. The derivative of the critical multiplier is

$$a_\mu = \left.\frac{\mathrm{d}\mu_c}{\mathrm{d}\zeta}\right|_{\zeta=1} = \frac{d(1-\beta)}{1 - \beta + \beta^2 d} > 0.$$

F.2 Solving for the Two New Fixed Points

The diagonal weight and momentum subspace is invariant. Write

$$\mathbf{W} = \mathbf{W}_\star + x\mathbf{E}_c + y\mathbf{E}_r.$$

At a fixed point, $\mathbf{M} = \mathbf{G}_\zeta(\mathbf{W})$ and

$$\mathbf{P} = \mathbf{G}_\zeta(\mathbf{W}) = (r_0 + h_r y)\mathbf{E}_r + \zeta h_0 x \mathbf{E}_c.$$

For (x, y) near the origin, define

$$t = \frac{\zeta h_0 x}{r_0 + h_r y}, \qquad s_\pm(t) = \frac{1 \pm t}{\sqrt{2(1+t^2)}},$$

and

$$\mathcal{E}(t) = \frac{g(s_+(t)) - g(s_-(t))}{\sqrt{2}}, \qquad \mathcal{Q}(t) = \frac{g(s_+(t)) + g(s_-(t))}{\sqrt{2}}.$$

Here $\mathcal{E}$ is odd and $\mathcal{Q}$ is even. Lemma 10 gives

$$\mathcal{E}(t) = a_1 t + a_3 t^3 + O(t^5), \qquad \mathcal{Q}(t) = \sqrt{2}g(s) + b_r t^2 + O(t^4). \tag{14}$$

The two coordinates of the fixed-point equation are

$$F_c(x, y, \zeta) = \lambda_{\mathrm{wd}} x + \mathcal{E}(t) = 0, \qquad F_r(x, y, \zeta) = \lambda_{\mathrm{wd}} y + \mathcal{Q}(t) - \mathcal{Q}(0) = 0.$$

Since $\partial_y F_r(0, 0, \zeta) = \lambda_{\mathrm{wd}} > 0$, the implicit function theorem gives a locally unique smooth solution $y = Y(x, \zeta)$ (Lee, 2013, Thm. C.40). Because F_r is even in x, uniqueness makes Y even as well. Its expansion is

$$Y(x, \zeta) = -\frac{b_r}{\lambda_{\mathrm{wd}}} \left(\frac{\zeta h_0}{r_0}\right)^2 x^2 + O(x^4).$$

Let $\tau = \zeta - 1$ and $k = h_0/r_0 = -\lambda_{\mathrm{wd}}/a_1$. Substitution into F_c gives

$$\widehat{F}_c(x, \zeta) := F_c(x, Y(x, \zeta), \zeta) = x\left[-\lambda_{\mathrm{wd}}\tau + C_\star x^2 + O(|\tau|x^2 + x^4)\right],$$

where

$$C_\star = k^3 \left(a_3 + \frac{a_1 h_r b_r}{\lambda_{\mathrm{wd}} r_0}\right) > 0.$$

The signs $a_3 > 0$ and $a_1 b_r > 0$ give $C_\star > 0$. Both $\mathcal{Q}(t)$ and $\mathcal{E}(t)/t$ are smooth functions of t^2 near zero, so the first implicit function argument also makes Y smooth in x^2. After division by x, the reduced equation extends smoothly to $x = 0$ and has derivative $C_\star > 0$ with respect to x^2 at $(x, \tau) = (0, 0)$. The implicit function theorem then gives, for small $\tau > 0$,

$$x_\pm = \pm\sqrt{\frac{\lambda_{\mathrm{wd}}}{C_\star}}\sqrt{\tau} + O(\tau^{3/2}), \qquad y_\pm = -\frac{b_r k^2}{C_\star}\tau + O(\tau^2).$$

These give fixed states

$$\mathbf{W}_\pm = \mathbf{W}_\star + x_\pm \mathbf{E}_c + y_\pm \mathbf{E}_r, \qquad \mathbf{M}_\pm = \mathbf{G}_\zeta(\mathbf{W}_\pm).$$

Thus $C_{\mathrm{pf}} = \sqrt{\lambda_{\mathrm{wd}}/C_\star}$. Since $b_r < 0$, both $y_\pm$ are positive, and

$$\|\mathbf{G}_\zeta(\mathbf{W}_\pm)\|_{\mathrm{F}} = \sqrt{(r_0 + h_r y_\pm)^2 + (\zeta h_0 x_\pm)^2} > r_0 > \epsilon_{\mathrm{cl}}.$$

Both fixed points therefore remain in the smooth region above the clamp threshold.

F.3 Stability in All Weight and Momentum Directions

At either new fixed point, let

$$\mathbf{B}_\pm = \mathrm{D}_{(x,y)}(F_c, F_r)\big|_{(x_\pm, y_\pm, \zeta)}.$$

Differentiating $F_r(x, Y(x, \zeta), \zeta) = 0$ gives

$$\partial_x Y = -\frac{\partial_x F_r}{\partial_y F_r}, \qquad \partial_x \widehat{F}_c = \partial_x F_c - \frac{\partial_y F_c \, \partial_x F_r}{\partial_y F_r}.$$

Multiplying by $\partial_y F_r$ gives the exact identity

$$\det \mathbf{B}_\pm = \partial_y F_r(x_\pm, y_\pm, \zeta)\, \partial_x \widehat{F}_c(x_\pm, \zeta).$$

The expansions above imply

$$\partial_y F_r = \lambda_{\mathrm{wd}} + O(\tau), \qquad \partial_x \widehat{F}_c(x_\pm, \zeta) = 2\lambda_{\mathrm{wd}}\tau + O(\tau^2), \qquad \det \mathbf{B}_\pm = 2\lambda_{\mathrm{wd}}^2\tau + O(\tau^2).$$

At $\tau = 0$, $\mathbf{B}_\pm$ tends to $\mathrm{diag}(0, \lambda_{\mathrm{wd}})$. Its diagonal entries change by $O(\tau)$ and its off-diagonal entries are $O(\sqrt{\tau})$, so its trace is $\lambda_{\mathrm{wd}} + O(\tau)$. Its two eigenvalues remain real and distinct for small τ, since the limiting eigenvalues are distinct. One is $\lambda_{\mathrm{wd}} + O(\tau)$, and the determinant identity gives the other as

$$\lambda_{\mathrm{small}}(\mathbf{B}_\pm) = 2\lambda_{\mathrm{wd}}\tau + O(\tau^2).$$

The diagonal matrix subspace is invariant under the derivative, and $\alpha\mathbf{B}_\pm = d\mathbf{I} + \mathbf{K}_{\mathrm{diag},\pm}$. Hence the eigenvalue of $\mathbf{K}$ near the lower edge is

$$\kappa_\pm = -d + 2d\tau + O(\tau^2).$$

It lies strictly between the two edges for small $\tau > 0$. The modal equation gives its critical multiplier as

$$\mu_\pm = 1 - 2a_\mu\tau + O(\tau^2),$$

which lies inside the unit disk. The other seven state multipliers converge to the seven noncritical multipliers at $\mathbf{z}_\star$ and remain inside the unit disk by continuity. Both new fixed points are therefore locally exponentially stable. This argument includes $\beta = 0$.

F.4 Exact Bounds for the Scalar Coefficients

Lemma 10 (Bounds for the lower-edge coefficients)**.** *The coefficients in Equation 13 satisfy*

$$g(s) > 0, \quad -1.881 < a_1 < -1.880, \quad -11.792 < b_r < -11.791, \quad 363.65 < a_3 < 363.66.$$

Moreover,

$$\begin{aligned} \mathrm{D}\Psi_{\mathrm{cl}}(\mathbf{P}_\star)[\mathbf{E}_r] &= 0, \\ \mathrm{D}\Psi_{\mathrm{cl}}(\mathbf{P}_\star)[\mathbf{E}_c] &= \frac{a_1}{r_0}\mathbf{E}_c, \\ \mathrm{D}\Psi_{\mathrm{cl}}(\mathbf{P}_\star)[\mathbf{E}_s] &= \frac{a_1}{r_0}\mathbf{E}_s, \\ \mathrm{D}\Psi_{\mathrm{cl}}(\mathbf{P}_\star)[\mathbf{E}_a] &= \frac{\sqrt{2}g(s)}{r_0}\mathbf{E}_a. \end{aligned} \tag{15}$$

Proof. Substituting $s_\pm(t)$ into g gives Equation 14. To bound its coefficients using rational arithmetic, set

$$z_{\pm,0}(t) = \frac{1 \pm t}{2\sqrt{1+t^2}}, \qquad \widehat{\phi}(z) = \frac{6889}{2000}z - \frac{191}{20}z^3 + \frac{4063}{500}z^5,$$

and define

$$z_{\pm,j+1}(t) = \widehat{\phi}(z_{\pm,j}(t)), \qquad j = 0, \ldots, 4.$$

The identity $\phi(\sqrt{2}z) = \sqrt{2}\widehat{\phi}(z)$ gives

$$g(s_\pm(t)) = \sqrt{2}z_{\pm,5}(t).$$

Moreover, $z_{-,j}(t) = z_{+,j}(-t)$ for every j. Write

$$z_{+,j}(t) = u_j + v_j t + w_j t^2 + q_j t^3 + O(t^4).$$

The initial coefficients are

$$(u_0, v_0, w_0, q_0) = \left(\frac{1}{2}, \frac{1}{2}, -\frac{1}{4}, -\frac{1}{4}\right).$$

Writing $\widehat{\phi}$, Taylor composition gives

$$\begin{aligned} u_{j+1} &= p(u_j), \\ v_{j+1} &= p'(u_j)v_j, \\ w_{j+1} &= p'(u_j)w_j + \frac{1}{2}p''(u_j)v_j^2, \\ q_{j+1} &= p'(u_j)q_j + p''(u_j)v_jw_j + \frac{1}{6}p'''(u_j)v_j^3. \end{aligned}$$

All quantities in this recurrence are rational. After five exact iterations, the symmetry between $z_{+,5}$ and $z_{-,5}$ gives

$$\begin{aligned} \sqrt{2}g(s) = 2u_5 \in (1.567, 1.568), \qquad & a_1 = 2v_5 \in (-1.881, -1.880), \\ b_r = 2w_5 \in (-11.792, -11.791), \quad & a_3 = 2q_5 \in (363.655, 363.656). \end{aligned}$$

Each decimal endpoint represents a rational number with denominator 10^3. Comparing integers after clearing positive denominators proves these inclusions and the bounds stated in the lemma.

The outer normalization has zero derivative in the radial direction and gain $1/r_0$ in the three orthogonal basis directions. Lemma 6 then gives Equation 15. □

G A Stable Orbit of Period Two at the Upper Edge

We choose curvature so that one multiplier reaches -1 at $\zeta = 1$. The invariant plane then gives an explicit orbit of period two, and the full return Jacobian checks attraction. This construction uses proposals below the clamp threshold. Appendix J.8 tests the same loss family with additive normalization.

We construct the orbit by finding nonzero solutions of $\mathcal{T}_\zeta(\mathbf{z}) = -\mathbf{z}$. The update is odd, so each such solution gives an orbit with two opposite states. Theorem 1 then proves its stability in the full weight and momentum space.

Theorem 11 (A stable period-two orbit beyond the upper edge)**.** *Fix $0 \le \beta < 1$, $\alpha > 0$, and $0 \le d < 2$. Consider the clamp-normalized transformation Ψ_{cl} with the Newton–Schulz coefficients and five iterations in Section 2. There is a family of globally strongly convex quadratic losses $\mathcal{L}_\zeta$ on 2×2 matrices for which the origin is a fixed point. A neighborhood of the origin can be chosen so that every proposal in this neighborhood satisfies $\|\mathbf{P}\|_{\mathrm{F}} < \epsilon_{\mathrm{cl}}$. At $\zeta = 1$, the weight and momentum Jacobian has one simple multiplier -1, while every other multiplier lies inside the unit disk. For $\zeta < 1$ close to one, the origin is locally exponentially stable. For $\zeta > 1$ close to one, the origin is unstable and a locally exponentially stable period-two orbit exists. In the weight coordinate w, its two points satisfy*

$$w_\pm(\zeta) = \pm C_{\mathrm{flip}}\sqrt{\zeta - 1} + O((\zeta - 1)^{3/2}), \qquad C_{\mathrm{flip}} > 0.$$

G.1 The Quadratic Loss and Its Critical Multiplier

The coefficients in Equation 2 give the expansion

$$g(x) = g_1 x + g_3 x^3 + O(x^5), \qquad g_1 = \gamma_1^5 > 0, \qquad g_3 = \gamma_3\gamma_1^4\sum_{j=0}^{4}\gamma_1^{2j} < 0.$$

Indeed, the linear and cubic coefficients of the iterates satisfy $u_{k+1} = \gamma_1 u_k$ and $v_{k+1} = \gamma_1 v_k + \gamma_3 u_k^3$, starting from $u_0 = 1$ and $v_0 = 0$.

The square block has shape factor one. Below the clamp threshold, normalization divides by the constant ϵ_{cl}, so

$$\mathrm{D}\Psi_{\mathrm{cl}}(0) = \frac{g_1}{\epsilon_{\mathrm{cl}}}\mathbf{I}.$$

Let $\mathbf{E} = \mathbf{e}_1\mathbf{e}_1^\top$ and set

$$\kappa_\star = U_{\beta,d}, \qquad h_0 = \frac{\epsilon_{\mathrm{cl}}\kappa_\star}{\alpha g_1} > 0.$$

For $\zeta > 0$, use the loss

$$\mathcal{L}_\zeta(\mathbf{W}) = \frac{\zeta h_0}{2} W_{11}^2 + \frac{h_0}{4}\left(W_{12}^2 + W_{21}^2 + W_{22}^2\right).$$

It is globally strongly convex. Its gradient is linear, and the optimizer update $\mathcal{T}_\zeta$ is odd. The origin is a fixed point, with

$$\operatorname{spec}(\mathbf{K}) = \{\zeta\kappa_\star, \kappa_\star/2, \kappa_\star/2, \kappa_\star/2\}.$$

At $\zeta = 1$, the critical eigenvalue gives the multiplier -1 and the partner

$$\frac{\beta(1+\beta d)}{1+2\beta} \in [0, 1).$$

The other three eigenvalues of $\mathbf{K}$ lie strictly between the edges. Theorem 1 therefore gives one simple multiplier -1 and seven multipliers inside the unit disk. It also gives stability of the origin for $\zeta < 1$ near one and instability for $\zeta > 1$ near one. Differentiating the modal equation gives

$$\mu_c(\zeta) = -1 - a_{\mathrm{up}}(\zeta - 1) + O((\zeta-1)^2), \qquad a_{\mathrm{up}} = \frac{(1-\beta)(1+2\beta)^2\kappa_\star}{1+3\beta+\beta^2 d} > 0.$$

G.2 Solving for the Two Orbit Points

The plane $\mathbf{W} = w\mathbf{E}$, $\mathbf{M} = m\mathbf{E}$ is invariant. Write the proposal as $\mathbf{P} = \epsilon_{\mathrm{cl}} u\mathbf{E}$, so the clamp condition is $|u| < 1$. To solve $\mathcal{T}_\zeta(w\mathbf{E}, m\mathbf{E}) = (-w\mathbf{E}, -m\mathbf{E})$, the momentum equation gives

$$m = -\frac{(1-\beta)\zeta h_0}{1+\beta} w.$$

Substitution into the proposal, followed by the definition of h_0, gives

$$u = \frac{\zeta h_0 (1-\beta)(1+2\beta)}{(1+\beta)\epsilon_{\mathrm{cl}}} w = \frac{\zeta(2-d)}{\alpha g_1} w.$$

The weight equation is $\alpha g(u) = (2-d)w$. Eliminating w yields

$$\zeta g(u) = g_1 u. \tag{16}$$

Let $\tau = \zeta - 1$. For $u \neq 0$, divide by u and expand:

$$0 = g_1\tau + \zeta g_3 u^2 + O(u^4).$$

The left side is smooth in u^2 and τ, with derivative $g_3 < 0$ with respect to u^2 at the origin. The implicit function theorem (Lee, 2013, Thm. C.40) gives

$$u_\pm = \pm\sqrt{-\frac{g_1}{g_3}}\sqrt{\tau} + O(\tau^{3/2}).$$

The two states are $\mathbf{z}_\pm = (w_\pm\mathbf{E}, m_\pm\mathbf{E})$, where

$$w_\pm = \frac{\alpha g_1}{\zeta(2-d)} u_\pm, \qquad m_\pm = -\frac{\epsilon_{\mathrm{cl}}}{1+2\beta} u_\pm.$$

Thus $C_{\mathrm{flip}} = \alpha g_1\sqrt{-g_1/g_3}/(2-d)$. The two states are opposite, and Equation 16 gives $\mathcal{T}_\zeta(\mathbf{z}_+) = \mathbf{z}_-$ and $\mathcal{T}_\zeta(\mathbf{z}_-) = \mathbf{z}_+$. Since $u_+ \neq 0$, the orbit has period two. Since $u_\pm = O(\sqrt{\tau})$, both proposals remain below the clamp threshold for sufficiently small $\tau > 0$.

G.3 Stability in All Weight and Momentum Directions

At either orbit point, $\mathbf{P} = \epsilon_{\mathrm{cl}} u\mathbf{E}$. Lemma 6, with $g(0) = 0$ and $g'(0) = g_1$, gives the four derivative eigenvalues:

direction	$\mathbf{e}_1\mathbf{e}_1^\top$	$\mathbf{e}_1\mathbf{e}_2^\top$	$\mathbf{e}_2\mathbf{e}_1^\top$	$\mathbf{e}_2\mathbf{e}_2^\top$
gain of $\mathrm{D}\Psi_{\mathrm{cl}}(\mathbf{P})$	$g'(u)/\epsilon_{\mathrm{cl}}$	$g(u)/(\epsilon_{\mathrm{cl}}u)$	$g(u)/(\epsilon_{\mathrm{cl}}u)$	$g_1/\epsilon_{\mathrm{cl}}$.

The Hessian is diagonal in the same basis, and Equation 16 gives $g(u)/u = g_1/\zeta$. Thus the four eigenvalues of $\mathbf{K}$ are

$$\kappa_{\|} = \frac{\zeta\kappa_\star}{g_1} g'(u), \qquad \frac{\kappa_\star}{2\zeta}, \qquad \frac{\kappa_\star}{2\zeta}, \qquad \frac{\kappa_\star}{2}.$$

The expansion of $u_\pm$ gives, successively,

$$u^2 = -\frac{g_1}{g_3}\tau + O(\tau^2), \qquad g'(u) = g_1(1-3\tau) + O(\tau^2), \qquad \kappa_{\|} = \kappa_\star(1-2\tau) + O(\tau^2).$$

All four eigenvalues lie in $(-d, U_{\beta,d})$ for sufficiently small $\tau > 0$. By Theorem 1, all eight eigenvalues of $\mathbf{A}_+ = \mathrm{D}\mathcal{T}_\zeta(\mathbf{z}_+)$ lie inside the unit disk.

The derivative of an odd map is even, so $\mathrm{D}\mathcal{T}_\zeta(\mathbf{z}_-) = \mathbf{A}_+$. The Jacobian over one period is

$$\mathrm{D}(\mathcal{T}_\zeta^2)(\mathbf{z}_+) = \mathbf{A}_+^2.$$

Its eigenvalues are the squares of those of $\mathbf{A}_+$ and all have modulus below one. This proves local exponential stability of the full orbit. The critical multiplier over one period is $1 - 4a_{\mathrm{up}}\tau + O(\tau^2)$. The argument applies throughout $0 \le \beta < 1$ and $0 \le d < 2$, including $\beta = 0$ and $d = 0$.

H Attracting States under Small Changes to the Update

The attracting states in the two constructions persist under small smooth changes to the loss. The following result fixes a parameter beyond the crossing, where the relevant multipliers are strictly inside the unit disk.

Proposition 12 (Persistence under loss perturbations)**.** *Fix a parameter $\zeta > 1$ for which Theorem 9 or Theorem 11 gives its attracting states, and write the corresponding quadratic loss as $\mathcal{L}_0$. Let*

$$\mathcal{L}_\eta(\mathbf{W}) = \mathcal{L}_0(\mathbf{W}) + \eta r(\mathbf{W}), \qquad r \in C^2.$$

For sufficiently small $|\eta|$, the two attracting fixed points in the lower construction persist as two distinct locally exponentially stable fixed points. The attracting orbit in the upper construction persists with minimal period two and remains locally exponentially stable. All these states remain on their original smooth side of the clamp threshold. If $\mathcal{L}_0$ is globally $\nabla^2\mathcal{L}_0 \succeq \underline{h}\mathbf{I}$ and $\|\nabla^2 r\| \le C$ globally, then $|\eta|C < \underline{h}$ also preserves global strong convexity.

Proof. Write the perturbed update as $\mathcal{T}_\eta$. The strict clamp inequalities and $r \in C^2$ make this map jointly C^1 in state and η near the original states. For an attracting fixed point $\mathbf{z}_0$, choose a norm in which $\mathrm{D}\mathcal{T}_0(\mathbf{z}_0)$ has norm below one, as in Appendix B. By continuity, a closed ball of radius r_{b} around $\mathbf{z}_0$ and a number $q < 1$ can be chosen so that $\|\mathrm{D}\mathcal{T}_\eta\| \le q$ throughout the ball for all sufficiently small $|\eta|$. Reduce $|\eta|$ until

$$\|\mathcal{T}_\eta(\mathbf{z}_0) - \mathbf{z}_0\| < (1-q)r_{\mathrm{b}}.$$

The mean value inequality then makes the ball invariant and the map a contraction. Consecutive iterates differ by at most q^j times the first difference, so they converge to a unique fixed point $\mathbf{z}_\eta$. Moreover,

$$\|\mathbf{z}_\eta - \mathbf{z}_0\| \le \frac{\|\mathcal{T}_\eta(\mathbf{z}_0) - \mathbf{z}_0\|}{1-q} \longrightarrow 0.$$

The same contraction gives local exponential stability. Apply this argument in disjoint balls around the two lower fixed points.

For the upper orbit, apply the argument to $\mathcal{T}_\eta^2$ near one orbit point. Its original Jacobian is Schur stable. The resulting fixed point $\mathbf{z}_\eta$ and $\mathbf{y}_\eta = \mathcal{T}_\eta(\mathbf{z}_\eta)$ remain near the two distinct original points, so their minimal period is two. The return derivative is

$$\mathrm{D}(\mathcal{T}_\eta^2)(\mathbf{z}_\eta) = \mathrm{D}\mathcal{T}_\eta(\mathbf{y}_\eta)\,\mathrm{D}\mathcal{T}_\eta(\mathbf{z}_\eta).$$

Contraction of the return map gives exponential convergence at even steps. The local Lipschitz bound for the intervening update gives it at odd steps. The neighborhoods can be chosen within the original clamp regions. Finally, $\nabla^2\mathcal{L}_\eta \succeq (\underline{h} - |\eta|C)\mathbf{I}$, proving the strong convexity claim. □

For example, $r(\mathbf{W}) = \sin(\langle \mathbf{a}, \mathrm{vec}\,\mathbf{W}\rangle + \varphi)$ with $\|\mathbf{a}\| = 1$ has $\|\nabla^2 r\| \le 1$. It gives an explicit nonquadratic family. A general perturbation can break the symmetry of the two fixed points or make the two orbit points cease to be opposites. At a fixed $\zeta > 1$ covered by the proposition, the nearby states remain locally exponentially stable.

Small changes to the update. The same contraction argument preserves these fixed points and period two orbits under sufficiently small C^1 changes to the complete update on their neighborhoods. For the lower construction, fix a parameter at which the two new fixed points are strictly attracting. Their proposals have norm greater than $r_0 > \epsilon_{\mathrm{cl}}$, so small closed neighborhoods can be chosen with $\|\mathbf{P}\|_{\mathrm{F}} \geq r_0$. On these neighborhoods the clamp rule equals $N_0(\mathbf{P}) = \mathbf{P}/\|\mathbf{P}\|_{\mathrm{F}}$. For $N_\delta(\mathbf{P}) = \mathbf{P}/(\|\mathbf{P}\|_{\mathrm{F}} + \delta)$ and $\delta \geq 0$,

$$\|N_\delta(\mathbf{P}) - N_0(\mathbf{P})\|_{\mathrm{F}} \leq \frac{\delta}{r_0}, \qquad \|\mathrm{D}N_\delta(\mathbf{P}) - \mathrm{D}N_0(\mathbf{P})\|_{\mathrm{F}\to\mathrm{F}} \leq \frac{\delta}{r_0^2}.$$

The second bound follows by separating the radial direction from its orthogonal complement. Composing with the finite matrix polynomial and the fixed momentum rule gives convergence of the complete updates in C^1 on these neighborhoods. Thus, for the same loss and hyperparameters, both attracting fixed points persist for every sufficiently small positive additive offset. The admissible offset depends on the chosen states and their stability margins. The inner upper construction uses a constant normalization denominator and requires a separate argument for additive normalization.

I Local Stability of Lion and Sophia

We derive the Sophia and Lion formulas from their published updates (Chen et al., 2023; Liu et al., 2024). The following table gives the conditional results before the full derivations. Each fixes the batch and realized random choices.

Method and condition	Local result derived here
Sophia with fixed positive denominator and inactive additional gradient clipping	For a real eigenvalue κ of $\alpha\mathbf{BH}$, stability requires $-d < \kappa < (1+\beta)(2-d)/(1-\beta)$.
Lion away from sign changes	$\mathrm{spec}(\mathbf{A}_{\mathrm{L}}) = \{1-d, \beta_2\}$, so the local state is stable exactly when $0 < d < 2$.

Sophia's curvature refresh and gradient norm clipping change this fixed-denominator calculation. Lion gives an attracting period two orbit even when its single step multipliers lie inside the unit disk.

Lion and sign changes. Lion uses $0 \leq \beta_1, \beta_2 < 1$ and the update

$$\mathbf{c} = \beta_1\mathbf{m} + (1-\beta_1)\mathbf{G}(\mathbf{w}),$$
$$\mathbf{w}^+ = (1-d)\mathbf{w} - \alpha\,\mathrm{sign}(\mathbf{c}), \quad \mathbf{m}^+ = \beta_2\mathbf{m} + (1-\beta_2)\mathbf{G}(\mathbf{w}).$$

The sign acts coordinatewise, with $\mathrm{sign}(0) = 0$. If every $c_i \neq 0$, the signs are constant in a neighborhood, so

$$\mathbf{A}_{\mathrm{L}} = \begin{pmatrix} (1-d)\mathbf{I} & 0 \\ (1-\beta_2)\mathbf{H} & \beta_2\mathbf{I} \end{pmatrix}.$$

The eigenvalues are $1-d$ and β_2, each repeated once per weight coordinate. This matrix is Schur stable exactly when $0 < d < 2$. The Hessian still affects coupling and transient growth. If $\beta_1 > 0$ and $c_i = 0$, varying m_i to either side gives limiting next weights separated by 2α. No ordinary Jacobian exists there. Away from these surfaces, ordered Jacobian products describe sufficiently small perturbations that preserve the finite sequence of signs.

For Lion, sign changes also give an explicit attracting orbit of period two. For $L(w) = hw^2/2$, $h > 0$, $0 < d < 2$, and $2\beta_1 < 1 + \beta_2$, the two states

$$w_\pm = \pm\frac{\alpha}{2-d}, \qquad m_\pm = \mp\frac{1-\beta_2}{1+\beta_2}\frac{h\alpha}{2-d}$$

form an attracting orbit. Indeed, $c_+ = hw_+(1+\beta_2-2\beta_1)/(1+\beta_2) > 0$, and substitution gives $(w^+, m^+) = (-w_+, -m_+)$. Odd symmetry gives the return step. Both points have the same local derivative $\mathbf{A}_{\mathrm{L}}$, so the return multipliers are $(1-d)^2$ and β_2^2. Their strict bounds give local exponential

stability by the argument in Appendix B. For $0 < d < 1$, both single step multipliers lie in $[0, 1)$. This orbit does not arise from a fixed point with a smooth multiplier crossing at -1: the origin lies on a discontinuity. At $d = 0$, the return map is locally affine with a unit eigenvalue, and the orbit is not an isolated attractor.

Sophia with its denominator fixed. Fix positive coordinate denominators q_i and use the raw loss gradient, with any additional gradient clipping absent or locally inactive. Sophia updates

$$\mathbf{m}^+ = \beta\mathbf{m} + (1-\beta)\mathbf{G}(\mathbf{w}), \qquad \mathbf{w}^+ = (1-d)\mathbf{w} - \alpha\,\mathrm{clip}(\mathbf{m}^+/\mathbf{q}, -1, 1), \qquad 0 \le \beta < 1.$$

Division and clipping act coordinatewise. Away from $|m_i^+| = q_i$, define

$$\mathbf{B} = \mathrm{diag}\left(\frac{\mathbf{1}\{|m_i^+| < q_i\}}{q_i}\right), \qquad \mathbf{K}_{\mathrm{S}} = \alpha\mathbf{B}\mathbf{H}.$$

Differentiation and a Schur complement give

$$\mathbf{A}_{\mathrm{S}} = \begin{pmatrix} (1-d)\mathbf{I} - \alpha(1-\beta)\mathbf{B}\mathbf{H} & -\alpha\beta\mathbf{B} \\ (1-\beta)\mathbf{H} & \beta\mathbf{I} \end{pmatrix},$$
$$\det(\mu\mathbf{I} - \mathbf{A}_{\mathrm{S}}) = \det((\mu-\beta)(\mu-1+d)\mathbf{I} + (1-\beta)\mu\mathbf{K}_{\mathrm{S}})\,.$$

The determinant identity extends to $\mu = \beta$ by polynomial equality, so it allows singular $\mathbf{B}$ and does not require diagonalizable $\mathbf{K}_{\mathrm{S}}$. Each eigenvalue κ gives two roots through

$$(\mu-\beta)(\mu-1+d) + (1-\beta)\mu\kappa = 0.$$

For real κ, the quadratic criterion in Appendix D gives

$$(1-\beta)(d+\kappa) > 0, \qquad (1+\beta)(2-d) - (1-\beta)\kappa > 0, \qquad 1 - \beta(1-d) > 0.$$

The last inequality is automatic. Thus both roots lie in the open unit disk exactly when

$$-d < \kappa < \frac{(1+\beta)(2-d)}{1-\beta}.$$

The endpoint roots are $+1$ and -1. Since $\mathbf{B} \succeq 0$, its quadratic form is nonnegative. Proposition 2 therefore rules out a negative real mode under positive semidefinite $\mathbf{H}$. For $d = 0$ and $\kappa = 0$, a multiplier remains at $+1$.

When the curvature estimate changes. If Sophia refreshes its curvature estimate during an update, that estimate belongs in the optimizer state. A refresh before the weight update and a refresh afterward have different complete Jacobians. Neither generally follows the fixed-denominator two-root law above. The published algorithm refreshes before the weight update, while the official Sophia-G training script refreshes afterward.[3] The derivative also includes any gradient clipping applied while estimating curvature.

Additional gradient clipping. The same script clips the gradient norm before updating momentum. If this operation has derivative $\mathbf{S}$, the conditional operator becomes $\alpha\mathbf{B}\mathbf{S}\mathbf{H}$. In the state Jacobians above, this replaces $\mathbf{H}$ by $\mathbf{S}\mathbf{H}$ in the momentum derivatives. Three positive semidefinite factors need not give a nonnegative spectrum. For example,

$$\mathbf{B} = \mathrm{diag}(1, 4), \quad \mathbf{S} = \tfrac{1}{4}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} 4.1 & -2 \\ -2 & 1.1 \end{pmatrix} \quad \Longrightarrow \quad \mathrm{spec}(\mathbf{B}\mathbf{S}\mathbf{H}) = \{0, -3/8\}.$$

Here $\mathbf{H}$ has eigenvalues 0.1 and 5.1, and $\mathbf{S}$ is the derivative of ideal norm clipping at gradient $(\sqrt{2}, -\sqrt{2})$ with threshold one. Positive denominators $(1, 1/4)$ and $\mathbf{m}^+ = (0.1, 0.1)$ realize the unsaturated Sophia branch. This example illustrates Proposition 2: the product $\mathbf{B}\mathbf{S}$ is not self-adjoint and has a negative quadratic direction, although each factor is positive semidefinite.

J Experimental Details

This appendix specifies the training, spectral searches, and perturbation checks behind Section 4. Counts include unresolved requests where indicated.

[3] Sophia code.

J.1 Curvature Operators and Arnoldi Estimates

For a fixed batch, write $\mathcal{L}(\mathbf{W}) = \ell(f(\mathbf{W}))$, $\mathbf{J} = \mathrm{D}f(\mathbf{W})$, and $\mathbf{C} = \nabla_f^2 \ell$. The exact Hessian and GGN are

$$\mathbf{H}^{\mathrm{ex}} = \mathbf{J}^\top \mathbf{C}\mathbf{J} + \sum_i [\nabla_f \ell]_i \nabla_{\mathbf{W}}^2 f_i, \qquad \mathbf{H}^{\mathrm{GGN}} = \mathbf{J}^\top \mathbf{C}\mathbf{J} \succeq 0.$$

The GGN is positive semidefinite for convex output losses such as cross-entropy (Martens, 2020, Sec. 8). We use $\mathbf{H}^{\mathrm{ex}}$ for the original update and $\mathbf{H}^{\mathrm{GGN}}$ for the linearized-logit control in Equation 6.

We apply $\mathbf{K} = \alpha \mathbf{R}\mathbf{H}$ without forming it: first compute $\mathbf{H}\mathbf{v}$, then differentiate Ψ to obtain $\mathbf{R}(\mathbf{H}\mathbf{v})$. Arnoldi estimates selected eigenvalues by projecting onto a Krylov subspace (Saad, 2011, Sec. 6.2.1). We also search $\mathbf{K}^\top$ and compare dimensions. For a candidate $(\widehat{\kappa}, \mathbf{v})$, the right relative residual is

$$r_{\mathrm{R}} = \frac{\|\mathbf{K}\mathbf{v} - \widehat{\kappa}\mathbf{v}\|}{\|\mathbf{K}\mathbf{v}\| + |\widehat{\kappa}|\|\mathbf{v}\|},$$

with r_{L} defined analogously for $\mathbf{K}^\top$. The GGN and small-model crossing tests require both residuals $\leq 10^{-5}$ and left/right and cross-dimension eigenvalue agreement within 10^{-3}. Sign distances must exceed ten times the empirical numerical resolution; experiment-specific searches and checks follow below.

J.2 Training the 3.33M and 7.25M Language Models

Both autoregressive Transformers use learned positional embeddings, shared input and output embeddings, pre-LayerNorm, GELU, vocabulary size 50,304, and no dropout. Table 4 specifies their architectures and state dimensions. Muon updates matrix coordinates, and AdamW updates the remaining parameters.

Table 4: Models used for the repeated mechanism tests. The complete state includes weights, Muon momentum, and both AdamW moments.

	3.33M model	7.25M model
Total parameters	3,327,744	7,248,640
Layers / width / attention heads	2 / 64 / 1	4 / 128 / 2
Feedforward width	256	512
Muon weight coordinates	98,304	786,432
Complete state coordinates	9,884,928	20,959,488

Each model trains for 2000 updates with 256 sequences of length 128 per update, processing 65,536,000 target tokens from FineWeb-Edu `sample/10BT`, revision v1.0.0, with the GPT-2 tokenizer. Seeds 0, 1, and 2 use distinct initializations and permutations of the same token blocks. Muon uses learning rate 0.01, momentum 0.95, weight decay 0.01, five Newton–Schulz steps, and additive offset 10^{-7}. AdamW uses learning rate 3×10^{-4}, moment coefficients $(0.9, 0.95)$, offset 10^{-8}, and weight decay 0.01. A 100-step linear warmup is followed by a constant learning rate. The validation pack contains 128 sequences. Seeds share the underlying training and validation corpora.

Local mechanism tests use step 1000 and two fixed probes, each with four sequences of length 128, starting at token offsets 70,000,000 and 71,000,000. The Hessian includes couplings across all Muon matrices. Other parameter coordinates are fixed for these conditional derivatives. Complete replays include every parameter and moment, with the realized batch sequence and discrete optimizer counters specified at each step.

J.3 Language Model Training from 58M to 1.25B Parameters

The larger models use the same dataset revision, tokenizer, vocabulary, embedding scheme, pre-LayerNorm, GELU, and zero dropout as the small models (Appendix J.2). Context length is 1024. Table 5 gives the architectures. Each uses seed 0 and 5000 updates with 512 sequences per update, totaling 2,621,440,000 target tokens.

The Muon learning rate rises linearly to 0.01 over 200 updates, stays constant until step 4500, and decreases linearly to 0.001. Muon uses $\beta = 0.95$, weight decay 0.01, five Newton–Schulz steps, coefficients $(3.4445, -4.7750, 2.0315)$, additive normalization offset 10^{-7}, fused QKV matrices, and shape factor $\sqrt{\max\{1, m/n\}}$. AdamW updates the remaining parameters with peak learning rate 3×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and weight decay 0.01.

Table 5: Language models used for the derivative survey. Each row is one training run with the same token budget. The last three columns divide the number of negative derivative eigenvalues by the total number of coordinates in the 16 selected matrices.

Parameters	Layers	Width	Heads	Step 200	Step 2000	Step 4400
58M	10	512	8	4.712%	7.926%	8.027%
124M	12	768	12	4.043%	7.197%	7.088%
254M	16	1024	16	3.928%	7.268%	6.801%
548M	20	1408	22	3.837%	7.404%	6.855%
1.25B	26	1920	30	1.863%	7.438%	6.764%

At each saved state $(\mathbf{W}_t, \mathbf{M}_t)$, we recompute the probe gradient and proposal $\mathbf{P}_{t,\mathcal{B}} = \beta^2 \mathbf{M}_t + (1-\beta^2)\nabla_{\mathbf{W}} \mathcal{L}_{\mathcal{B}}(\mathbf{W}_t)$. Conditional derivatives include all Muon matrices and their cross-couplings, with other parameters fixed. Unless stated otherwise, the probe has four sequences of length 1024 starting at token offset 1,048,576,000. Local interventions and replays specify their learning rate and normalization below.

In the intervention experiments, independent checks of the complete conditional update use relative tolerances 10^{-7} for automatic differentiation and 10^{-3} for finite differences. Appendices J.5 and J.6 specify the spectral searches and additional checks.

J.4 Measuring Negative Derivatives Across Model Sizes

At steps 200, 2000, and 4400, we sample the fused QKV, attention output, and two feedforward matrices at four depths: layers $(3, 5, 8, 10)$ for 58M, $(3, 6, 9, 12)$ for 124M, $(4, 8, 12, 16)$ for 254M, $(5, 10, 15, 20)$ for 548M, and $(7, 13, 20, 26)$ for 1.25B. This gives 16 matrices per checkpoint and 240 matrix–checkpoint pairs across five training runs.

Each pair has negative derivative eigenvalues in the diagonal or symmetric off-diagonal families. For each matrix, an eigenvalue is counted as negative when it is below $-10^{-6} \max\{1, \max_i |\lambda_i(\mathbf{R})|\}$. The fractions in Table 5 divide the negative count by the total matrix dimension, pooling all 16 matrices. Larger matrices therefore contribute more directions to this fraction.

The 58M, 124M, and 548M measurements use saved proposals. For 254M and 1.25B, we recompute proposals from saved states and the original training batches. The 96 reconstructed records check gradient, momentum, and proposal RMS values against the training logs, with a largest relative difference of 5.54×10^{-8}. The reconstruction check compares tensor-level scalar summaries.

J.5 Gauss–Newton Tests at 58M

Both GGN probes use the 58M step 2000 weights $\mathbf{W}_0$, all 40 Muon matrices, and saved momentum $\mathbf{M}_0$. Each has four sequences of length 1024; $\mathcal{B}_0$ and $\mathcal{B}_1$ start at token offsets 1,048,576,000 and 1,572,864,000. We use $\alpha = 0.01$, $\beta = 0.95$, $d = 10^{-4}$, and additive offset 10^{-7}. For each probe, $\mathbf{G}_0$ and $\mathbf{P}_0 = \beta^2 \mathbf{M}_0 + (1-\beta^2)\mathbf{G}_0$ are recomputed at the saved state.

A convex loss with the reference GGN. For token i, let $\mathbf{J}_i$ be the reference logit Jacobian and $\boldsymbol{\pi}_i$ the softmax probability vector. With mean cross-entropy over T targets, Equation 6 satisfies

$$\nabla \widetilde{\mathcal{L}}(\mathbf{W}_0) = \mathbf{G}_0, \qquad \nabla^2 \widetilde{\mathcal{L}}(\mathbf{W}_0) = \frac{1}{T} \sum_i \mathbf{J}_i^\top \big(\operatorname{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i^\top\big) \mathbf{J}_i = \mathbf{H}_0^{\mathrm{GGN}}.$$

The reference logits and $\mathbf{J}_i$ stay fixed; the softmax probabilities and loss Hessian can vary away from $\mathbf{W}_0$. Weight decay acts on the physical weight $\mathbf{W}_0 + \mathbf{\Delta}$.

Two implementations check the GGN action. A cached calculation uses the reference probabilities as auxiliary cross-entropy targets: its output gradient vanishes at the reference, leaving the GGN as its weight Hessian. An independent calculation applies $\mathbf{J}_i$, the softmax Hessian, and $\mathbf{J}_i^\top$ directly. Relative action differences are below 1.3×10^{-15}; the gradient of $\widetilde{\mathcal{L}}$ matches the original within 2.2×10^{-15}.

Changing the response while preserving the reference update. We apply the output-matched control in Equation 7 blockwise, with $\delta = 10^{-7}$. With $\mathbf{P}_0$ fixed, $\Psi_{\mathrm{ctrl}}(\mathbf{P}_0) = \Psi_{10^{-7}}(\mathbf{P}_0)$ and $\mathrm{D}\Psi_{\mathrm{ctrl}}(\mathbf{P}_0) = \mathrm{D}N(\mathbf{P}_0)$. The control therefore preserves $\mathbf{W}^+$ and $\mathbf{M}^+$ while changing the transformation derivative.

To see the derivative signs, vectorize a nonzero proposal as $\mathbf{p}$, with $r = \|\mathbf{p}\|$. Then

$$\mathrm{D}N(\mathbf{p}) = \chi_{m,n}\left(\frac{\mathbf{I}}{r+\delta} - \frac{\mathbf{p}\mathbf{p}^\top}{r(r+\delta)^2}\right).$$

The radial eigenvalue is $\chi_{m,n}\delta/(r+\delta)^2$, and every orthogonal eigenvalue is $\chi_{m,n}/(r+\delta)$. At zero, the derivative is $\chi_{m,n}\mathbf{I}/\delta$. Thus $\mathbf{R}_{\mathrm{ctrl}} \succ 0$, and Proposition 2 rules out negative real modes for $\mathbf{R}_{\mathrm{ctrl}}\mathbf{H}_0^{\mathrm{GGN}}$.

The measured differences between the two reference weight updates, divided by the update norm, are 1.25×10^{-14} and 1.24×10^{-14} for the two probes. The stored next momentum agrees exactly. Independent differentiation and all three central difference amplitudes also validate the control against its own response operator.

Negative modes and their measured multipliers. We apply Arnoldi to $\mathbf{K}_{\mathrm{GGN}} = \alpha\mathrm{D}\Psi_{10^{-7}}(\mathbf{P}_0)\mathbf{H}_0^{\mathrm{GGN}}$ and its transpose. We compare Krylov dimensions 64 and 128, with a prescribed extension to 128 and 256, and stop once a negative mode passes the tests below. On $\mathcal{B}_0$, we initialize the right search with a candidate direction from a calculation at dimension 64. Probe $\mathcal{B}_0$ resolves a negative mode at dimension 128 and $\mathcal{B}_1$ at 256. These searches test the existence of a negative mode.

Eigenvalue differences use the normalization $\max(1, |\kappa_1|, |\kappa_2|)$ in the agreement tests of Appendix J.1. The imaginary part must be at most $10^{-8}\max(1, |\operatorname{Re}\kappa|)$. The empirical sensitivity indicator combines cross-dimension differences and residuals divided by left/right overlap. Table 6 reports the resolved modes; their distances to zero and $-d$ exceed ten times this indicator.

We independently differentiate the optimizer step for $\widetilde{\mathcal{L}}$. For each right mode $\mathbf{v}$, we measure its action on $\mathrm{span}\{(\mathbf{v}, 0), (0, \mathbf{H}_0^{\mathrm{GGN}}\mathbf{v})\}$, using separate weight and momentum scales and measuring the component outside that subspace. The eigenvalues of this measured 2×2 restriction agree with Equation 3 (Table 6). Central differences at relative amplitudes 10^{-6}, 10^{-7}, and 10^{-8} all pass.

Table 6: Resolved 58M GGN modes at step 2000. Independent multipliers agree with Equation 3 within 2.5×10^{-15}; residuals and empirical sensitivity quantify numerical resolution.

Probe	κ	Measured μ	r_{R}	r_{L}	Sensitivity
$\mathcal{B}_0$	-0.1962	1.085	2.12×10^{-14}	6.08×10^{-10}	2.39×10^{-6}
$\mathcal{B}_1$	-0.07082	1.042	8.23×10^{-13}	1.08×10^{-12}	7.41×10^{-8}

J.6 GGN Controls and Edge Crossings Across Seeds

The GGN tests use Equation 6, additive offset 10^{-7}, and the output-matched control from Appendix J.5. Crossings use the original cross-entropy objective.

All six cases in each model have a negative GGN mode that passes the numerical checks below. Measured κ ranges from -0.507 to -0.171 in the 3.33M model and from -0.633 to -0.0846 in the

Table 7: Cases passing the GGN and crossing tests (three seeds, two probes per seed). One 3.33M upper crossing fails the requirement that at least two finite-difference amplitudes pass.

Model	Negative GGN mode	Lower crossing	Upper crossing
3.33M	6/6	6/6	5/6
7.25M	6/6	6/6	6/6

7.25M model. For the same selected direction $\mathbf{v}$, set $\mathbf{u} = \mathbf{H}^{\mathrm{GGN}}\mathbf{v}$ and compare the original map with the normalization control through

$$\frac{\langle \mathbf{u}, \mathbf{R}\mathbf{u} \rangle}{\|\mathbf{u}\| \, \|\mathbf{R}\mathbf{u}\|}$$

Figure 3 plots this alignment: it is negative for the five-step map and positive for the control in all twelve cases.

The crossing tests use Ψ_0, $\alpha = 0.01$, $\beta = 0.95$, and $d = 10^{-4}$. For selected negative and positive real modes, the critical proposal scales are $c_- = -\kappa_-(1)/d$ and $c_+ = \kappa_+(1)/U_{\beta,d}$, respectively. At $c = 1$, the selected positive modes are inside their edge and the selected negative modes are beyond theirs. For each selected real mode, we measure the baseline and the three scales $c/c_\pm = 0.8, 1, 1.2$. The lower critical scales are 4.05×10^4–6.68×10^4 in the 3.33M model and 6.98×10^4–1.55×10^5 in the 7.25M model. The upper scales are 0.516–0.804 and 0.655–0.862, respectively. Within each family, we check the transformed update, the $1/c$ derivative scaling, and the independent joint state response.

Arnoldi compares dimensions $32/64$, $64/128$, and $128/256$ within the operator budget, using the residual and agreement tests in Appendix J.1. Left/right overlap must be at least 10^{-12}, and the spectral gap must exceed ten times the empirical resolution. Real directions require zero computed eigenvalue imaginary parts and imaginary vector norm $\leq 10^{-7}$. Independent differentiation must agree within relative vector error 10^{-7}. At least two amplitudes among $10^{-6}, 10^{-7}, 10^{-8}$ must pass both the 10^{-3} finite-difference vector test and the edge-distance test.

The upper case at seed 2, probe 0 of the 3.33M model is unresolved. At $c/c_+ = 0.8$, the requested multiplier errors are 1.37×10^{-4}, 1.37×10^{-6}, and 6.78×10^{-9} across the three amplitudes. Only the smallest passes its edge test. The upper crossing count is therefore 11 of 12.

J.7 Proposal Scaling with the Additive Normalization Offset

At 58M step 2000, we hold weights, the four-sequence probe, and hyperparameters fixed and set $c = 0.5, 1, 2$ through stored momentum. With additive offset 10^{-7}, the weight update changes by relative 7.234×10^{-6} and 3.617×10^{-6} at $c = 0.5$ and 2. The selected upper mode moves inside its edge only at $c = 2$; the selected lower mode stays beyond its edge at all three scales.

Hessian sharpness stays 12.45, while momentum changes by 4.881 and 9.762 times its original norm. Independent differentiation and central differences use the same two reference directions across scales; spectra are recomputed at each state.

J.8 Numerical Checks Beyond the Two Edges

The quadratic constructions. We evaluate the exact loss families in Appendices F and G, with $\alpha = 1$ and clamp threshold 10^{-7}. The lower construction uses $r_0 = 100\epsilon_{\mathrm{cl}}$, $h_r = h_s = h_0/2$, $\kappa_a = U_{\beta,d}/2$, $d = 0.1$, and $\beta \in \{0, 0.5, 0.95\}$. The upper construction uses the same three momentum coefficients with $d \in \{0, 0.1\}$. Each configuration uses

$$\tau \in \{10^{-8}, 3 \times 10^{-8}, 10^{-7}, 3 \times 10^{-7}, 10^{-6}, 3 \times 10^{-6}, 10^{-5}, 3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}\}.$$

This gives 30 lower and 60 upper parameter points. We solve the fixed point and orbit equations at 80 and 120 decimal digits, reconstruct the full states, and check that each state satisfies its defining equation. Independent derivatives, all eight finite difference directions, and the full state multipliers test stability. For period two we also differentiate the actual return map. The leading amplitude coefficients come from the proofs.

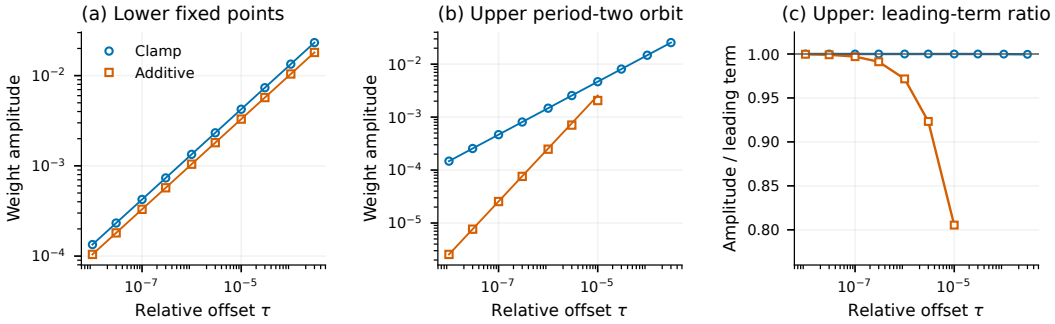


Figure 8: Numerical solutions of the two quadratic constructions. Markers solve the full state equations and lines give leading amplitude terms. The normalization rules differ near the upper crossing. Full eight-dimensional return maps determine stability.

All 90 points pass the equation and derivative checks. The first eigenvector test resolves 80 points. Ten upper points with $\beta = d = 0$ remain unresolved under that test because their zero modes have ill-conditioned eigenvectors. A subsequent Schur and Lyapunov calculation classifies all original matrices and scale checks without requiring diagonalizability: 406 matrices are stable and 94 states at the original fixed point are unstable. This matches the predicted pattern at all 90 grid points and four scale checks. At $\beta = 0.5, d = 0.1$, the measured amplitude differs from its leading prediction by at most 0.040% over the grid.

Changing the normalization. We next replace clamp normalization by the additive rule with the same offset 10^{-7} and keep the quadratic loss family fixed. In the lower construction, we solve again for the central fixed point and its critical parameter. The critical value moves from $\zeta = 1$ to approximately 1.2374. Across three momentum settings and ten relative offsets $\tau = \zeta/\zeta_\star - 1$, all 30 cases have two closed, numerically stable fixed points. Their amplitudes follow the independently computed square-root leading term, with maximum relative deviation 0.0434%. The comparison uses each rule's own critical parameter.

For the upper construction, the derivative at the origin remains $g_1\mathbf{I}/\epsilon_{\mathrm{cl}}$, so the critical parameter remains one. The nonzero orbit changes its scale. To see why, write $\mathbf{P} = \epsilon_{\mathrm{cl}} u\mathbf{E}$ on the invariant plane and take the positive phase. The additive transformed value is $g(u/(1+u))$. Eliminating the momentum and weight as in Appendix G gives, successively,

$$\begin{aligned}(1+\tau)g\left(\frac{u}{1+u}\right) &= g_1 u,\\ 0 &= g_1\tau - g_1 u + O(\tau u + u^2),\\ u &= \tau + O(\tau^2),\\ |w| &= \frac{\alpha g_1}{2-d}\tau + O(\tau^2).\end{aligned}$$

The second line follows by dividing by $u > 0$. Its derivative in u at $(u, \tau) = (0, 0)$ is $-g_1$, so the one-sided solution follows from the implicit function theorem (Lee, 2013). The opposite phase follows from oddness. The additive map is continuously differentiable at the origin, but its normalization contributes a term proportional to $u|u|$. Thus the linear amplitude here differs from the square-root amplitude under the locally constant clamp denominator.

We test all six upper momentum/decay configurations at seven offsets from 10^{-8} to 10^{-5}. All 42 cases have closed period-two orbits and numerically stable full return maps. At $\beta = 0.5, d = 0.1$, the amplitude divided by the linear leading term is 0.9997 at $\tau = 10^{-8}$ and 0.8054 at 10^{-5}. The full grid retains this finite-offset deviation. Figure 8 compares both normalization rules.

J.9 Boundary Measurements during Continued Training

The changing-batch continuation starts from the 3.33M seed 0, step 1000 checkpoint and evolves the complete optimizer for 2048 updates. Each update uses 256 new sequences of length 128, with Muon

and AdamW learning rates 0.01 and 3×10^{-4}. After update 1000, data extend beyond the originally saved training segment. Validation cross-entropy falls from 6.258 to 5.665 (Figure 6).

We measure conditional Muon modes on each state's training batch: 64 primary states at updates $1536, 1544, \ldots, 2040$, plus 24 earlier states at $0, 64, \ldots, 1472$. The earlier upper ratios have median 1.14. Our residence criterion requires at least 52 primary points within $\pm 5\%$ of the edge and at least 58 resolved points, accounting for numerical uncertainty and retaining unresolved points in the denominator. All 88 upper measurements resolve. Primary ratios span 0.593–1.757 (median 0.988), with 30 above one and 34 below; the 5%, 10%, and 20% bands contain 13, 20, and 36 of 64 points. These measurements fail the residence criterion and do not separate batch variation from state changes.

The lower search resolves 24/64 primary points and 29/88 overall. The other 59 exhaust their Arnoldi budgets and are not assigned to either side of the edge. Resolved primary ratios $-\kappa_-/d$ range from 2.08×10^4 to 1.28×10^5.

For the repeated-batch control, seeds 3, 4, and 5 each run with Muon alone and with both optimizers for 2048 updates on one fixed 16,384-token target, at Muon learning rate 0.01. All 528 spectral requests are attempted. Five primary windows resolve 64 upper points; one resolves 63. In seed order, upper-ratio medians are $(2.56, 3.15, 3.02)$ for Muon alone and $(2.92, 2.84, 2.79)$ for both optimizers. Every resolved upper ratio exceeds 1.62, outside the 5% band. Lower searches resolve $(61, 60, 59)$ and $(58, 61, 58)$ of 64 points, respectively, all with $-\kappa_-/d > 6.4 \times 10^4$. The spectra remain conditional on Muon coordinates.

J.10 Following Two Muon Modes over 16 Updates

This experiment follows perturbations initialized in the two identified modes of the 58M model at step 2000. It uses the original cross-entropy loss on a fixed probe of four sequences of length 1024. All 40 Muon matrices and their momentum evolve with $\Psi_{10^{-7}}$, $\alpha = 0.01$, $\beta = 0.95$, and $d = 10^{-4}$. The other parameters stay fixed. The initial lower and upper modes have $\kappa \approx -82.54$ and 33.27, with state multipliers approximately 9.484 and -1.671.

We scale each joint eigenvector $\mathbf{v} = (\mathbf{v}_W, \mathbf{v}_M)$ by one scalar so that the larger of $\|\mathbf{v}_W\|_{\mathrm{F}}/\|\mathbf{W}_0\|_{\mathrm{F}}$ and $\|\mathbf{v}_M\|_{\mathrm{F}}/\|\mathbf{M}_0\|_{\mathrm{F}}$ is one. Paired runs start from $(\mathbf{W}_0, \mathbf{M}_0) \pm \varepsilon \mathbf{v}$ at amplitudes $10^{-4}, 10^{-5}, 10^{-6}$. We divide weight and momentum changes by their initial norms, approximately 411.7 and 0.01418, and use the Euclidean norm of the resulting joint vector. These scales stay fixed. Gain is the response norm divided by the initial perturbation norm.

Frozen propagation repeats the initial Jacobian, while ordered propagation uses the Jacobian at each reference state. Central differences of the paired nonlinear runs provide the comparison. Both initial modes pass independent derivative and finite difference checks. An additional derivative check at step eight has relative error below 3.2×10^{-15} in both state components.

The relative vector error of ordered propagation stays below 10^{-3} through steps 2, 7, and 16 for the lower direction at the three decreasing amplitudes. The upper direction passes through step 16 at every amplitude. At 10^{-6} and step 16, the errors are 6.22×10^{-4} and 4.83×10^{-7}. All 102 direction–amplitude–state records are finite.

J.11 Complete Muon and AdamW Response over 64 Updates

The streamed replay starts from seed 1 of the 3.33M model at step 1000. It uses 64 actual training batches, each with 256 sequences of length 128, the original data order, Muon learning rate 0.01, and AdamW learning rate 3×10^{-4}. The continuous state has 9,884,928 coordinates. The reference update and its differentiable implementation agree to a largest relative state-group error of 1.47×10^{-16} along the reference path.

We compare central differences of paired perturbed runs with the chain rule and the frozen initial derivative. For a projection $\mathbf{\Pi}$ onto a reported state group, the relative vector error is

$$\frac{\left\| \mathbf{\Pi} \left[(\mathbf{z}_k^+ - \mathbf{z}_k^-)/(2\varepsilon) - \mathbf{A}_{k-1} \cdots \mathbf{A}_0 \mathbf{v} \right] \right\|}{\| \mathbf{\Pi} \mathbf{A}_{k-1} \cdots \mathbf{A}_0 \mathbf{v} \|}.$$

Here the superscripts on $\mathbf{z}_k^{\pm}$ identify trajectories started at $\mathbf{z}_0 \pm \varepsilon \mathbf{v}$. The frozen comparison replaces the product by $\mathbf{A}_0^k$. Group errors use the physical coordinates and the prediction norm as denominator. A zero denominator remains undefined. Initial coordinates divide each of the Muon weights, AdamW weights, Muon momentum, and Adam first moment by its initial Euclidean norm, with floor 10^{-30}. Adam second-moment coordinates divide elementwise by the maximum of the initial and reference next-step second moments. These scales remain fixed throughout the replay. We normalize three fixed Gaussian directions in these coordinates. A fourth Gaussian direction is restricted to the AdamW coordinates by zeroing its Muon weight and momentum components, and then normalized. In the error formula, $\mathbf{v}$ denotes the resulting physical direction.

Three random directions and one Adam-input direction are measured at relative amplitudes $10^{-6}, 10^{-7}, 10^{-8}$. A combination passes through step k when the initial perturbation is represented accurately, the error stays at most 10^{-3} at every step up to k, the predicted response is nonzero, and the Adam second moments remain nonnegative in both perturbed runs. Once a step fails, later smaller errors do not make the full sequence pass. Two requested real eigenmodes of the complete Jacobian remain unresolved. This gives twelve measured and six unresolved direction–amplitude combinations.

Table 8: Complete streamed replay: combinations passing every step. Each denominator is four measured directions times three amplitudes in one window. Six additional modal combinations are unresolved.

State group	Ordered, 16 steps	Ordered, 64 steps	Frozen, 64 steps
All weights	6/12	6/12	0/12
Muon weights	10/12	10/12	0/12
AdamW weights	6/12	6/12	0/12
Muon momentum	10/12	10/12	0/12
Adam first moment	10/12	10/12	0/12
Adam second moment	10/12	10/12	0/12

The random-3 and Adam-input directions pass at both 10^{-7} and 10^{-8} in every group. Random-1 and random-2 pass in all-weight coordinates only at 10^{-8}. Their 10^{-6} perturbations already leave the nonnegative variance domain at initialization. Four other all-weight combinations exceed the error threshold at the first update. Repeating the initial Jacobian passes the error criterion only through the first step in each group.

The 64-step replay extends the same window selected for the initial 16-step check. It measures linear response along one realized trajectory; the direction and amplitude checks delimit which finite perturbations follow it.