NORMALIZATION RIGIDITY AND RECOVERY IN MATRIX MOMENTUM

Anonymous authors
Paper under double-blind review

ABSTRACT

Matrix optimizers such as Muon normalize a momentum matrix and then change its singular values before updating the model. We study how this transformation responds to gradient noise. Symmetric noise can cancel when the true gradient is zero, while the average update still responds in the wrong direction when a gradient is added. We show that preventing this behavior for every symmetric noise distribution is equivalent to monotonicity of the update map. For methods that apply the same smooth function independently to Frobenius-normalized singular values, this requirement forces the function to be linear once there are at least three singular values. We explain how the shared normalization causes the problem and which gradient directions are affected. We also construct smooth convex tasks on which actual momentum updates keep the gradient away from zero over a rigorously verified time window. Convex spectral potentials provide another design choice: they preserve alignment with the gradient while still bringing singular values closer together. A smoothed nuclear norm additionally controls sensitivity to numerical errors. In our transformer experiments, however, all 600 task-response estimates remain positive, and matched training favors Schatten over the smoothed map. These results identify a mathematical failure mechanism and its limits as an explanation of training behavior.

1 INTRODUCTION

Matrix optimizers change more than the length of a gradient. Muon, for example, applies a short Newton–Schulz iteration to a normalized momentum matrix (Jordan et al., 2024; Liu et al., 2025). This operation can bring singular values closer together, so directions associated with small singular values receive more weight relative to dominant directions. Such transformations raise a basic issue when gradients are noisy: the average of the transformed gradients may behave differently from a transformation of the true gradient.

In this paper, we study whether the average update still follows the true gradient. Let g be a fixed gradient and Z be gradient noise. Ordinary stochastic gradient descent uses $g + Z$, whose expectation is g when $\mathbb{E}Z = 0$. A nonlinear map Ψ instead uses $\Psi(g + Z)$. For a small learning rate, the quantity $\langle g, \mathbb{E}\Psi(g + Z)\rangle$ determines whether this direction favors a decrease in the loss, because the algorithm subtracts the update from the parameters.

Symmetric noise gives one useful guarantee. If Ψ is odd, meaning $\Psi(-P) = -\Psi(P)$, then the average update at $g = 0$ is zero. This cancellation does not settle what happens when a small nonzero gradient is added. The average update can change in the opposite direction. We use the term *recovery* for this response to the true gradient, and study the conditions that make it point in the right direction.

The normalization step plays an essential role. Dividing by a matrix's own Frobenius norm makes every normalized singular value depend on all the others. Increasing one singular value also increases the common denominator and can reduce the remaining normalized values. As a result, a function that is increasing in each scalar input can still give an unfavorable matrix response. The relevant property is monotonicity of the whole map, which requires a change in its output to align with the change in its input.

Our analysis proceeds in three steps. First, we characterize the maps that preserve alignment under every bounded symmetric noise distribution. Within the family that applies a smooth function

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.

independently to Frobenius-normalized singular values, only a linear function has this guarantee when there are at least three singular values. We then identify the specific task directions that produce a negative response. Some noise distributions, including isotropic Gaussian noise, average out this problem, so the map alone does not determine whether it occurs.

Second, we connect the response calculation to actual momentum updates. On constructed smooth convex tasks, we prove that the gradient stays a positive distance from zero for an explicit number of updates. The proofs include the momentum history and start from zero momentum. Third, we study updates obtained from convex spectral potentials. These maps can preserve alignment while still flattening singular-value ratios. We compare Schatten norms with a smoothed nuclear norm to understand the resulting tradeoff between flattening, recovery strength and sensitivity to numerical errors.

Finally, we test small transformers. All 600 estimated task responses remain positive after correcting gradient-estimation bias, and matched training favors Schatten over the smoothed map. Understanding the practical importance of the constructed failures therefore requires evidence about the noise and gradients encountered during training.

2 Setup and the response to a gradient

We work with matrices in $\mathbb{H} = \mathbb{R}^{m\times n}$ and use the Frobenius inner product $\langle A, B\rangle = \operatorname{tr}(A^\top B)$. Let $r = \min(m, n)$ and write a singular value decomposition as $P = U \operatorname{diag}(\sigma_i(P))V^\top$.

The matrix update. We first divide the singular values by the input norm and then apply a scalar function f to each of them:

$$\Psi_f(P) = U \operatorname{diag}\Big(f\big(\sigma_i(P)/\|P\|_F\big)\Big)V^\top, \qquad P \neq 0, \qquad \Psi_f(0) = 0. \tag{1}$$

Here f is continuous and $f(0) = 0$. For an odd polynomial $f(s) = \sum_{j=0}^{J} a_j s^{2j+1}$, the same map is $\sum_j a_j (PP^\top)^j P/\|P\|_F^{2j+1}$ and can be computed with matrix products. It is odd and unchanged when P is multiplied by a positive scalar; the latter property is called zero homogeneity. The reference implementation composes $p(u) = 3.4445u - 4.7750u^3 + 2.0315u^5$ a total of K times (Jordan et al., 2024), giving $f_K = p^{\circ K}$ and the map Ψ_K. We analyze this particular finite iteration. Since $p(1) = .701$, it does not converge to the polar factor as do some other coefficient choices (Kim & Oh, 2026).

Momentum and noise. Our trajectory analysis uses plain exponential moving average (EMA) momentum:

$$\bar{\xi}_t = \frac{1}{B}\sum_{b=1}^{B} \xi_{t,b}, \qquad M_t = \beta M_{t-1} + (1-\beta)(\nabla F(W_t) + \bar{\xi}_t), \qquad W_{t+1} = W_t - \eta\Psi(M_t). \tag{2}$$

The batch size is $B \geq 1$, the momentum coefficient is $0 \leq \beta < 1$, and the learning rate is $\eta > 0$. We start with $M_{-1} = 0$ and update the parameters at every step. The theorems use real arithmetic and no denominator regularizer ($\varepsilon = 0$). The neural experiments separately examine the implemented Nesterov update and numerical precision.

Holding the gradient fixed. To isolate the effect of the map, first hold a gradient g fixed and average over noise $Z_w = \sum_i w_i \xi_i$, where $w_i \geq 0$ and $\sum_i w_i = 1$. Define

$$h_w(g) = \mathbb{E}\Psi(g + Z_w), \qquad b_w = h_w(0), \qquad A_w = Dh_w(0). \tag{3}$$

The function h_w is the average update at each fixed gradient, often called the *frozen field*. Its value b_w describes the update when there is no signal; its derivative A_w describes the response to a small signal. For centrally symmetric noise (Z_w and $-Z_w$ have the same distribution), oddness gives $b_w = 0$. This cancellation describes the value at zero. It does not tell us how the average update changes when a nonzero gradient is added. For a direction E, a negative value of $\langle E, A_w E\rangle$ means that adding a small positive gradient along E changes the average update against that direction.

For stationary EMA weights $w_{j,b} = (1-\beta)\beta^j/B$, the squared-weight sum is $v = \sum_i w_i^2 = (1-\beta)/(B(1+\beta))$. The bounded-noise results take $\mathbb{E}\xi_i = 0$ and $\|\xi_i\|_F \leq 1$; the examples specify their own noise scale. For independent samples with covariance Σ, averaging gives covariance $v\Sigma$.

When a derivative is unavailable, we measure the response by the finite difference $\mathcal{R}_w(g; E, h) = \langle E, h_w(g + hE) - h_w(g - hE)\rangle/(2h)$, with $\|E\|_F = 1$. This fixed-gradient calculation is the starting point. A training trajectory also changes g over time and carries the previous gradients in its momentum.

3 What normalization allows

We begin with a guarantee that does not depend on a particular task: the average update should align with every fixed gradient under every bounded symmetric noise distribution. The following result identifies exactly when an odd map has this property.

Proposition 1 (Symmetric alignment equals monotonicity)**.** *For a bounded, Borel measurable, odd Ψ, the following are equivalent:*

$$\langle g, \mathbb{E}\Psi(g + Z)\rangle \geq 0 \qquad \text{for all } g \text{ and all bounded centrally symmetric } Z;$$
$$\langle A - B, \Psi(A) - \Psi(B)\rangle \geq 0 \quad \text{for all } A, B.$$

For every $\varrho \geq 0$,

$$\inf_{Z \overset{d}{=} -Z,\ \|Z\|_F \leq \varrho} \langle g, \mathbb{E}\Psi(g + Z)\rangle = \inf_{\|z\|_F \leq \varrho} \tfrac{1}{2}\langle g, \Psi(z + g) - \Psi(z - g)\rangle.$$

Symmetric two-point laws approach the infimum; attainment is not required.

The second inequality is the definition of *monotonicity* for a matrix map. It says that an output change cannot point against the corresponding input change. To see the connection with noise, pair Z with $-Z$. Oddness turns their average contribution into $[\Psi(Z + g) - \Psi(Z - g)]/2$, so the same inequality applies. Conversely, any pair of inputs A, B can be represented by $g = (A - B)/2$ and noise taking the values $\pm(A + B)/2$. Thus two opposite noise values already suffice to test the guarantee.

Why an increasing scalar function is not enough. Consider singular values $(4, 1, 1)$. Increasing the first value to 5 changes the smaller normalized values from $1/\sqrt{18}$ to $1/\sqrt{27}$. They decrease because the common denominator has increased. This interaction remains even if f is increasing in each scalar input. The next theorem shows how restrictive it is.

Theorem 2 (Rigidity of normalized pointwise spectral maps)**.** *Let $r \geq 3$, $f \in C^1([0, 1])$ and $f(0) = 0$. Then Ψ_f of* (1) *is monotone if and only if $f(s) = cs$ on $[0, 1]$ for some $c \geq 0$.*

In other words, this family can give the universal guarantee only by reducing to scaled Frobenius-normalized descent. The requirement of three singular values is necessary: at $r = 2$, the nonlinear function $f(u) = u - u^3/3 + u^5/5$ is a monotone example (Proposition 11).

The proof explains the restriction. Restrict the map to three positive diagonal entries, normalized to $s^\top s = 1$. Its Jacobian is $A = D(I - ss^\top)$, where $D = \operatorname{diag}(f'(s_i))$. The factor $I - ss^\top$ comes from the normalization. A change parallel to s only rescales the input, so $As = 0$. Monotonicity then forces the symmetric part of A to vanish on s as well, which implies that all positive coordinates have the same scalar derivative. With three coordinates, we can link any two amplitudes through a common smaller coordinate; hence f' is constant. The full argument is in Appendix C. The general connection between zero-homogeneous monotone maps and convex potentials is classical (Kushnir & Krishnamoorthy, 2022).

Changing the normalizing norm. Other norms allow other responses. For the Schatten norm $\|P\|_q = (\sum_i \sigma_i(P)^q)^{1/q}$, $1 < q < \infty$, the matching rule is

$$P \mapsto U \operatorname{diag}[f(\sigma_i(P)/\|P\|_q)]V^\top \text{ is monotone} \quad \Longleftrightarrow \quad f(s) = cs^{q-1}, \quad c \geq 0. \tag{4}$$

Here $r \geq 3$, f is continuous on $[0, 1]$, continuously differentiable on $(0, 1)$, $f(0) = 0$, and the map is zero at the origin. Theorem 16 proves this extension. It explains why Schatten gradients are a natural alternative: their spectral response is matched to the norm used in their denominator.

Measuring the failure. The size of the negative response depends on how much the scalar slopes differ. Set $\bar{d} = \sum_i s_i^2 f'(s_i)$, $V_f(s) = \sum_i s_i^2[f'(s_i) - \bar{d}]^2$ and $M = \|D\|_{\mathrm{op}}$. For $S = (A + A^\top)/2$, Proposition 12 gives

$$\lambda_{\min}(S) \leq -\frac{V_f(s)}{2(\sqrt{M^2 + V_f(s)} + M)} < 0 \qquad \text{if } V_f(s) > 0. \tag{5}$$

This bound guarantees a negative direction even when every scalar slope is positive. The direction remains unfavorable under sufficiently small perturbations of the noise (Corollary 13). It still matters whether the task gradient points in that direction, which we examine next.

4 When the task gradient is affected

A map may have a direction with negative response while responding positively to the actual task gradient. To distinguish these situations, we separate the effect of changing an input's magnitude from the effect of changing its direction.

Separating magnitude and direction. At $P \neq 0$, let $\rho = \|P\|_F$ and $X = P/\rho$. Write a gradient direction as $G = aX + H$, where $a = \langle X, G\rangle$ and $H \perp X$. The term aX changes the magnitude of P; H changes its direction while leaving its norm unchanged to first order. Let $\mathcal{B}$ be the derivative of the spectral map before normalization, evaluated at X, and let Π remove the component parallel to X.

Proposition 3 (Task response and normalization coupling)**.** *Set $v = \Pi\mathcal{B}X$ and $T = \Pi\mathcal{B}\Pi|_{X^\perp}$. Then*

$$\rho\langle G, D\Psi_f(P)[G]\rangle = \underbrace{\langle G, \mathcal{B}G\rangle}_{\textit{fixed denominator}} - \underbrace{\langle X, G\rangle\langle G, \mathcal{B}X\rangle}_{\textit{normalization coupling}} = \langle H, TH\rangle + a\langle v, H\rangle. \tag{6}$$

If $T \succ 0$ and $V_f(s) > 0$, the full symmetric Jacobian has exactly one negative eigenvalue. A task direction has negative response exactly when $a\langle v, H\rangle < -\langle H, TH\rangle$. Purely tangential directions are nonnegative. If $V_f(s) = 0$ and $T \succ 0$, the symmetric Jacobian is positive semidefinite. The condition $T \succ 0$ holds for $f' > 0$ on $[0, 1]$ and for the classical cubic compositions at full-rank inputs of rank at least two.

The first equality compares two calculations: one holds the denominator fixed, while the other differentiates it with the input. The second equality shows where a negative response can arise. A pure rescaling gives no response, because normalization removes it. For the increasing functions covered by the proposition, a purely tangential change has positive response. A negative response requires both changes to occur together, with the cross term outweighing the positive tangential term. Theorem 17 describes the resulting region of directions explicitly; Figure 1(a) shows one section of it. The tuned reference polynomial can have negative scalar slopes, so it may have several negative directions; the decomposition itself still applies.

The noise distribution also matters. Noise averaging can remove the negative response. Suppose the noise density depends only on its Frobenius norm and does not increase farther from zero. For any bounded spectral map with nonnegative output values, the average update then aligns with every gradient (Proposition 18). This includes uniform noise in a matrix ball and isotropic Gaussian noise. The proof pairs singular frames that differ in the sign of one singular vector: the frame better aligned with g receives at least as much probability as its opposite. No increasing-slope assumption is needed.

These statements require symmetry in the full matrix space. Noise confined to a chosen line or plane need not satisfy them, and we make no such assumption about transformer noise. The uniform-sphere case is treated in Proposition 19.

Estimating the response from noisy gradients. At a fixed input, put $S = \operatorname{sym} D\Psi(P)$. In an experiment we estimate g from batches. If $G_1, \dots, G_n$ are independent unbiased estimates with covariance Σ, simply inserting their average into the quadratic form adds a bias:

$$\mathbb{E}\langle \bar{G}, S\bar{G}\rangle = \langle g, Sg\rangle + \frac{\operatorname{tr}(S\Sigma)}{n}, \qquad \widehat{q}_U = \frac{1}{n(n-1)}\sum_{i\neq j}\langle G_i, SG_j\rangle, \quad \mathbb{E}\widehat{q}_U = \langle g, Sg\rangle. \tag{7}$$

The estimation noise has mean zero, but its quadratic contribution need not. Comparing two independently estimated gradients removes this contribution, giving the cross-estimate $\widehat{q}_U$. We keep the map input separate from these batches and sample all of them from the same evaluation pool. Proposition 20 gives the exact variance. We assess the sign from this numerator.

Connecting the response to loss. A response tells us how an update changes when a gradient is added. The actual update already contains momentum from earlier steps, and a finite step also encounters loss curvature. Let $\mathcal{A}$ be its expected alignment with the current gradient, conditional on

the past, and let V be its expected squared norm. For an L_F-smooth loss, the expected loss change differs from $-\eta\mathcal{A}$ by at most $L_F\eta^2V/2$ (Proposition 21). Thus predicting loss requires the whole update, including its history, as well as the local response measured above.

5 From a response calculation to a momentum trajectory

We now ask whether a negative response can affect the actual optimization process. This requires more than holding a gradient fixed: the parameters move, and momentum contains earlier gradients. We construct tasks on which these effects can be controlled throughout a specified number of updates.

An example caused by normalization. Take the increasing function $f_+(s) = (s + 3s^3)/4$ and the unit matrices $S_3 = \mathrm{diag}(4, 1, 1)/\sqrt{18}$ and $E_3 = \mathrm{diag}(3, 2, 2)/\sqrt{17}$. Every scalar slope is positive. Using the decomposition in (6), however, gives

$$\langle E_3, D\Psi_{f_+}(S_3)[E_3]\rangle = \underbrace{\frac{93}{68}}_{\text{fixed denominator}} - \underbrace{\frac{76}{51}}_{\text{change in denominator}} = -\frac{25}{204}. \tag{8}$$

The first term is restoring, but the effect of changing the common denominator is larger. Let the loss be $F(W) = \langle E_3, W\rangle^2/2$ and let the noise take the two values $\pm\sigma S_3$ with equal probability, independently at each step. The loss is a smooth convex quadratic, and the noisy gradient is unbiased. Its true gradient is always parallel to E_3, so this task uses the direction with the unfavorable response above. The following result concerns the actual recursion (2), including its zero-initialized momentum.

Corollary 4 (A normalization-only trajectory obstruction)**.** *For this task, $W_0 = E_3/4$, $M_{-1} = 0$, $\sigma = 1/16$, $\beta = 0.1$, $B = 1$ and $\eta = 10^{-5}$, the actual EMA obeys, with probability at least 0.95,*

$$1.86767 \times 10^{-4} \le \|\nabla F(W_t)\|_F^2 \le 4.16481 \times 10^{-3} \qquad \textit{for every } 295417 \le t \le 10^6.$$

The same guarantee holds after adding independent smooth centrally symmetric noise of Frobenius norm at most $\sigma/8000$ in the full matrix space.

The positive lower bound means that the gradient does not approach zero during this window. The statement holds simultaneously at every update in the window, rather than only at its endpoint. Theorem 29 gives the more general conditions. If we keep the same scalar function but divide by the fixed scale σ, the response becomes positive (65). This provides a control for the effect of the input-dependent denominator.

An example that also flattens singular values. The first function expands singular-value ratios. To check that this is not necessary, consider the classical Newton–Schulz cubic $f_{\mathrm{cub}}(u) = (3u - u^3)/2$. It is increasing on $[0, 1]$ and brings every positive ratio closer to one. At the same S_3, the direction $E = \mathrm{diag}(7, 1, 1)/\sqrt{51}$ has response $-1/34$. With noise $\pm\sigma S_3$, $\sigma = 1/16$, $W_0 = E/4$, $\beta = .3$ and $\eta = 10^{-6}$, Theorem 45 proves, with probability .95,

$$\frac{4913}{61440000} \le \|\nabla F(W_t)\|_F^2 \le \frac{1230137}{245760000} \qquad (666776 \le t \le 10^7). \tag{9}$$

A further full-matrix noise perturbation of norm at most $\sigma/60000$ is allowed. Thus the normalization problem can occur even when the scalar function is increasing and the output spectrum is flatter than the input spectrum.

What makes the proof possible. At these momentum values, the newest random sign dominates the combined contribution of the older signs. The noise EMA therefore stays away from zero. We can bound the expected update uniformly over the relevant states, then account for the delay in the signal momentum and the random fluctuations along the path. All numerical inequalities are checked with exact rational arithmetic, using 43 and 217 interval boxes for the two examples. The independent checker is described in Appendix X. The guarantee is limited to the stated time window and step size.

Higher momentum and other examples. At $\beta = .95$, a rank-128 example with five classical cubic iterations maintains a nonzero gradient level in simulation. Rank-512 examples with eight iterations show the same behavior at $\beta = .9$ and $.95$, again with noise in a chosen matrix direction. The weaker example needs steps near 10^{-8}, and the longest trajectory has 2.97×10^8 updates. The

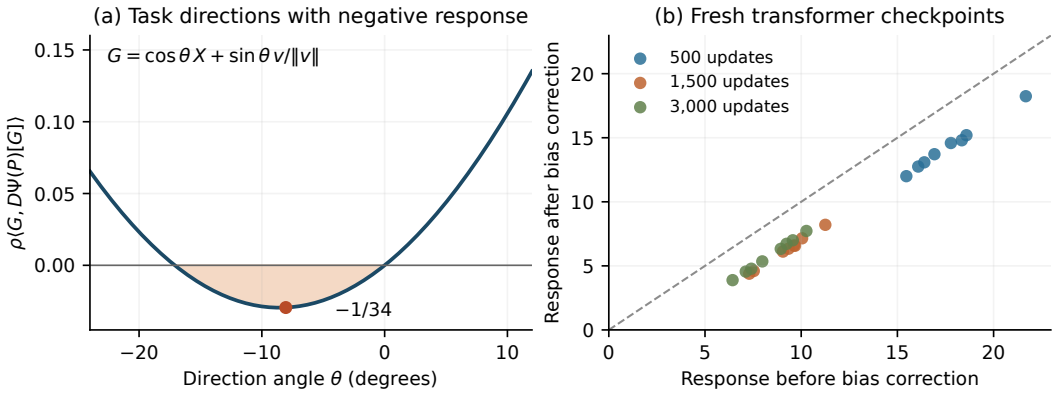


Figure 1: A negative direction need not be the task gradient. (a) The shaded region contains negative-response directions for the increasing classical cubic at S_3. The marked $\mathrm{diag}(7,1,1)/\sqrt{51}$ has response $-1/34$. (b) At all 24 fresh transformer checkpoints, correcting the gradient-estimation bias reduces the response estimate but leaves it positive. Checkpoints are repeated measurements within eight training seeds.

statistical bounds certify the averaged field used in these calculations, not the trajectories themselves (Appendix J.2).

A separate example for the tuned reference quintic places noise orthogonally to the signal and studies a negative scalar slope (Appendix B). Its trajectory theorem has a feasible instance at $T = 10^{12}$, with entry bound about 1.77×10^{11} and step size 1.62×10^{-11}. The shorter simulations are outside this guarantee. Auxiliary forecasts approximate their gradient levels, but a complete first-order nonlinear correction still requires unproved $o(h)$ error bounds (Appendix K). These examples establish that the mechanism is possible; they do not establish how important it is at ordinary training budgets.

6 Designing maps that preserve alignment

The preceding results suggest changing how singular values interact. A smooth nonlinear function applied independently after Frobenius normalization cannot give the universal guarantee. We can instead choose a convex function of the whole matrix and use its gradient as the update. This allows the singular-value gains to depend on one another.

Why a convex potential helps. Let Φ be an even potential, meaning $\Phi(-P) = \Phi(P)$. Convexity makes its gradient monotone. If it is also positively one-homogeneous, meaning $\Phi(cP) = c\Phi(P)$ for $c > 0$, the gradient is unchanged by positive rescaling, just like a normalized update. The next proposition also allows subgradients where the potential is not differentiable.

Proposition 5 (Alignment from even convex potentials)**.** *Let Φ be finite, convex, even and positively one-homogeneous. For any Borel-measurable selection $\widetilde{\Psi}(P) \in \partial\Phi(P)$, the map is bounded and satisfies $\langle g, \mathbb{E}\widetilde{\Psi}(g+Z)\rangle \geq 0$ for every g and bounded centrally symmetric Z. The minimum-norm selection is odd and zero at the origin; an arbitrary selection need not be.*

The proof follows directly from the subgradient inequality: $\langle g, \widetilde{\Psi}(g+z)\rangle \geq \Phi(g+z) - \Phi(z)$. Averaging over symmetric noise makes the right side nonnegative by convexity and evenness. Appendix M gives the details and shows why evenness is needed in this class.

Schatten norms as a baseline. A standard choice is the Schatten norm. For $\varsigma \geq 1$, its canonical update is

$$\Phi_{(\varsigma)}(P) = \Big(\sum_i \sigma_i(P)^{\varsigma}\Big)^{1/\varsigma}, \qquad \Psi_{(\varsigma)}(P) = U\,\mathrm{diag}[(\sigma_i/\|P\|_{\varsigma})^{\varsigma-1}]V^{\top}. \tag{10}$$

We set the map to zero at $P = 0$ and map zero singular values to zero at $\varsigma = 1$. For $1 < \varsigma < 2$, an input ratio $r_0 > 1$ becomes $r_0^{\varsigma-1}$. Choosing ς closer to one therefore makes the output singular values

more similar. It also weakens recovery on the transverse example: after matching the output norm, the slope at zero signal is $\varsigma - 1$. Proposition 54 gives the exact response and its sharp lower bound. At the polar endpoint $\varsigma = 1$, this local slope is zero.

A smooth alternative. We next study a norm that smooths the small singular values. Let $r = \min(m, n)$, $\tau = \kappa/r$ and $X = P/\|P\|_F$. Define

$$\Phi^{\mathrm{sn}}_\kappa(P) = \sum_{i=1}^{r} \sqrt{\sigma_i(P)^2 + \tau\|P\|_F^2}, \qquad \Psi^{\mathrm{sn}}_\kappa(P) = [Y + \tau \operatorname{tr}(Y) I]X, \tag{11}$$

where $Y = (XX^\top + \tau I)^{-1/2}$ after transposing to the smaller matrix shape. The parameter $\kappa > 0$ controls the smoothing. Each square root uses the norm of the whole matrix, so changing one singular value affects all summands. Their derivatives produce the shared trace term in the update. Omitting that term changes the map and loses the justification from this convex potential.

Proposition 6 (Regular convex spectral flattening)**.** *For every $\kappa > 0$, $\Phi^{\mathrm{sn}}_\kappa$ is a norm, smooth away from zero, and $\Psi^{\mathrm{sn}}_\kappa = \nabla\Phi^{\mathrm{sn}}_\kappa$ there, with value zero at zero. It has universal symmetric-noise alignment and strict singular-value ratio compression. Writing $\mu_\kappa = \kappa\sqrt{r}/\sqrt{1+\kappa}$ and $L_\kappa = \sqrt{r}(1+\kappa)/\sqrt{\kappa}$, for every $P \neq 0$,*

$$\frac{\mu_\kappa}{\|P\|_F}\|\Pi_P^\perp H\|_F^2 \le \langle H, D\Psi^{\mathrm{sn}}_\kappa(P)[H]\rangle \le \frac{L_\kappa}{\|P\|_F}\|\Pi_P^\perp H\|_F^2. \tag{12}$$

If $P^{[k]}$ consists of k diagonal copies of $P/\sqrt{k}$, the output consists of k copies of $\Psi^{\mathrm{sn}}_\kappa(P)$. Thus its spectral ratios are invariant under replication.

Here $\Pi_P^\perp H$ is the part of H orthogonal to P. The lower bound guarantees a positive response to this change; the upper bound limits how much the output can change. Both remain valid when some singular values are zero, provided $P \neq 0$. This smoothness is the distinction from Schatten maps with exponent below two. The ratio property under replication is shared by both families.

Both maps can be computed from the smaller of $PP^\top$ and $P^\top P$, called the Gram matrix, using one eigendecomposition. Their actual speed and numerical accuracy still depend on shape and precision (Appendix O). A cheaper quartic potential uses only two matrix products and also preserves alignment, but its flattening weakens as the spectrum is replicated (Theorem 50).

What the error bound means. Suppose we compare two nearby inputs $P \pm hE$, where $\|E\|_F = 1$ and $0 < h < \rho = \|P\|_F$. If their computed outputs have errors at most e_+, e_-, then

$$\frac{\langle E, \widehat{\Psi}(P + hE) - \widehat{\Psi}(P - hE)\rangle}{2h} \ge \frac{\mu_\kappa \rho^2(1 - a^2)}{(\rho + h)^3} - \frac{e_+ + e_-}{2h}, \qquad a = \langle E, P/\rho\rangle. \tag{13}$$

The first term is the restoring response guaranteed by the exact map. The second is the amount that output errors can remove. If the right side is positive, the computed response is still positive (Corollary 22). Such a guarantee needs bounds on the full output error, not just the sensitivity of individual coefficients.

For example, if both errors are bounded by e, setting $h = \rho/2$ keeps the response positive whenever $e < 4\mu_\kappa(1 - a^2)/27$. The factor $1 - a^2$ makes the role of direction clear: an orthogonal change has a positive margin, while pure rescaling has none. This gives a concrete accuracy target for an implementation once its input and tested direction are specified.

Smoothness does not imply a better optimizer. At matched zero-signal output norm, the smoothed map's response slope on the transverse example is $1 - (1 + \kappa)^{-2}$, compared with $\varsigma - 1$ for Schatten (Proposition 60). At $\kappa = .001$ and $\varsigma = 1.05$, these are .001997 and .05. Thus a smooth, strongly flattening map can respond less strongly to a gradient. We test the loss and computation tradeoffs separately in the next section.

7 Experiments

We first test the constructed failures, then examine task gradients during transformer training, and finally compare the spectral maps as optimizers. These experiments address different questions, so we interpret their outcomes separately.

Table 1: Task response on eight seeds, averaged over three checkpoints per seed. Brackets give nominal 95% intervals; cross-estimates remove gradient covariance bias.

Component	Plug-in	Cross-estimate
Net response	11.751 [11.216, 12.287]	8.778 [8.247, 9.309]
Fixed denominator	11.794 [11.208, 12.380]	8.821 [8.243, 9.399]
Normalization term	0.043 [−0.079, 0.165]	0.043 [−0.078, 0.164]
Positive contribution	13.295 [12.721, 13.869]	10.059 [9.489, 10.630]
Negative magnitude	1.544 [1.481, 1.606]	1.281 [1.219, 1.344]

Checking the constructed failure. For the settings in Corollary 4, all 128 independent trajectories enter and stay in the certified band by update 34,935, before the proved entry limit 295,417. Replacing the input norm by a fixed scale reduces the final squared gradient by about four orders of magnitude. Across three learning rates, the normalized method keeps a nonzero level while the fixed-scale control improves as the step gets smaller. The classical-cubic example gives the same qualitative result (Appendices F and J.1). These are tests of the specified convex examples; the high-momentum simulations have the separate limits described above.

Diagnosing a trained reference model. We train eight new 10.9M-parameter byte-level transformers on a fixed WikiText-103 corpus, using the reference map, momentum .95, learning rate .008 and batch size 32. At updates 500, 1500 and 3000, we record the actual inputs to all 25 matrix-map calls. The task gradient for this diagnostic is the gradient of the held-out crop distribution. We estimate it with 16 independent batches of 64 sequences per checkpoint.

We use the cross-estimate in (7) to remove the covariance bias and sum the matrix responses with the optimizer's positive shape factors. Each seed first averages its three checkpoints; confidence intervals then use the eight independent seeds. The protocol and code were fixed before the runs. After each probe, we restore the complete model, both optimizers and random states, so the diagnostic does not change the continuing training run (Appendix U).

What the diagnostic shows. Every one of the 600 inputs has a direction with negative response. However, all 600 cross-estimates along the evaluation gradient are positive, as are all 24 checkpoint aggregates. The mean is 8.778 [8.247, 9.309] across seeds. The correction removes 25.3% of the original estimate without changing its sign. Table 1 explains why: the positive spectral contribution is 10.059, while the negative magnitude is 1.281. Each direction contributes its squared gradient projection multiplied by its eigenvalue. A substantial gradient overlap with negative directions can therefore be outweighed by stronger positive directions. The normalization term's interval includes zero. These are estimates; the layer counts are not simultaneous guarantees for all population response signs.

The actual one-step loss measurement is less conclusive. On common evaluation crops, its mean change is −.000381, with 95% interval [−.000865, +.000103] across seeds. The interval crosses zero. This measurement includes auxiliary parameter updates and loss curvature, so its sign cannot be read directly from the matrix response. An earlier, separate 600-input fork study also found positive task responses despite 21.3% average gradient energy in negative eigenspaces; replacing the map increased validation loss (Appendix S). Together, the studies do not show that the adverse response explains loss behavior in these transformers.

Comparing training losses. We separately train 10.9M and 25.6M transformers for 2500 updates with each map. Every arm receives the same five learning rates, two tuning seeds and eight paired test seeds. Within a seed, initial parameters and batches are shared. Test text is disjoint, and layer gains are fixed from the first common gradient. All maps use float64. Table 2 reports direct paired loss differences. Schatten-1.05 has lower loss than smoothed nuclear at both sizes. At the smaller size, a reference run with 2850 updates approximates the smoothed map's time budget and also has lower loss.

Both designs flatten the spectrum, but this does not determine their loss ordering. Timing also depends on the matrix shape: the Gram-based arms take 1.14–1.15 times the reference's time at the smaller size and about .73 times at the wider size. We have no comparison at matched wall time for the wider

Table 2: Paired test-loss differences on eight seeds per scale. All arms use 2500 updates except the 2850-update reference; negative differences favour the first method.

Model	Comparison	Δ loss	95% CI
10.9M	Schatten − smoothed	-0.0041	$[-0.0069, -0.0014]$
25.6M	Schatten − smoothed	-0.0050	$[-0.0066, -0.0034]$
10.9M	Longer reference − smoothed	-0.0165	$[-0.0179, -0.0150]$

model. These measurements therefore describe the tested implementations, not general production throughput. The numerical cutoff and output-error checks are in Appendices O and W.

Additional prediction tests. Direct evaluation of paired losses is more accurate than the midpoint-gradient formula (NRMSE 0.042 versus 0.110) and cheaper per pair. Frozen diagnostics do not reliably improve 10/50-step loss predictions. For the separate mean-shift forecast, median error is 0.6% in 10 local cases and 2.1% across all 16 cases. These results, including the failed cases, are in Appendices V and K.6.

8 Related work

Failures of normalized optimizers can have several causes. For sign-based methods, Karimireddy et al. (2019) give counterexamples with skewed noise, nonsmooth objectives and symmetric state-dependent noise, and study error feedback. For matrix methods, Parshakova et al. (2026) analyze a deterministic nonsmooth example, while Do et al. (2026) give a stochastic finite-iteration nonconvergence result. Our trajectory examples use smooth convex losses and fixed symmetric additive noise. We also characterize which normalized spectral maps preserve alignment and explain how their response depends on the task direction. Appendix Q compares the assumptions and quantifiers.

Work on finite Newton–Schulz iterations studies polynomial approximation (Amsel et al., 2025) and convergence to the polar factor under other coefficient choices (Kim & Oh, 2026). The normalization in Li & Tsuchiya (2026) uses a fixed external scale. Our rigidity result concerns the norm of the input matrix itself, which is why it also covers increasing scalar functions. Spectral emphasis and high-pass interventions (Beneventano et al., 2026; Fan et al., 2026) study related aspects of matrix updates. Convex potentials and norm-based updates are established tools (Bernstein & Newhouse, 2024; Kushnir & Krishnamoorthy, 2022); our design results quantify their flattening, recovery and numerical-sensitivity tradeoffs. Existing smooth convergence results (Sato et al., 2025; Chang et al., 2025; Choudhury et al., 2026; Cutkosky & Mehta, 2020) use other algorithmic conditions or limiting regimes, and do not contradict the finite-window examples studied here.

9 Discussion

Normalizing a matrix by its own norm couples changes in its singular values. We showed that this can produce a response against the gradient even when the scalar transformation is increasing. The structural theorem identifies which independent spectral functions avoid the problem, and the task-response calculation identifies which directions are affected. On selected smooth convex tasks, the effect persists along actual momentum updates for a rigorously checked time window.

The transformer results give a different outcome. Correcting gradient-estimation bias still leaves the measured responses positive, and the normalization term has an interval containing zero. To establish practical training harm, one would need to observe the adverse response on task gradients and connect it to the subsequent loss. The existence of a negative direction alone does not provide this evidence.

Convex spectral potentials preserve alignment while allowing flattening, but their useful properties come with tradeoffs. Smoothing controls sensitivity to numerical errors and can weaken recovery; in our matched experiments, the implemented Schatten baseline trains better. Extending the trajectory guarantees to common high-momentum budgets and realistic noise distributions, and determining when smoothness improves training outcomes, remain open questions. The diagnostic gives a direct way to check for negative local responses along task gradients.

REPRODUCIBILITY STATEMENT

The supplementary proof archive provides exact rational interval boxes, an independent standard-library checker, selected reference-polynomial certificates, checking outputs and file digests. It verifies the computational premises of the two normalization-only finite-window theorems without GPU access; instructions and scope are in Appendix X. The experiment sources, frozen protocols, complete-seed raw statistics, map-input fixture and build commands accompany the manuscript. All new task-response runs restore the full executed training state after probing. The numerical high-momentum trajectories and finite-step forecasts are not presented as certified paths.

ETHICS STATEMENT

This study concerns stochastic matrix optimization, using synthetic tasks and small transformers trained on public text. It involves no human subjects or deployed system. Its neural experiments measure local geometry and controlled short interventions; they do not establish general training harm from Muon or superiority of a replacement optimizer.

AI USE STATEMENT

Generative AI tools helped develop the framework, formulate mathematical claims and proof arguments, design synthetic experiments, implement methods, analyze results, review related literature and draft the manuscript. The authors are responsible for all mathematical claims, empirical results and manuscript content.

REFERENCES

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The polar express: Optimal matrix sign methods and their application to the Muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025. URL https://arxiv.org/abs/2505.16932.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2023. URL https://arxiv.org/abs/1912.02365. Also arXiv:1912.02365.

Pierfrancesco Beneventano, Mahmoud Abdelmoneum, and Tomaso Poggio. The spectral dynamics and noise geometry of Muon. *arXiv preprint arXiv:2606.08388*, 2026. URL https://arxiv.org/abs/2606.08388.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. *arXiv preprint arXiv:2410.21265*, 2024. URL https://arxiv.org/abs/2410.21265.

Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, 1997.

Da Chang, Yongxiang Liu, and Ganzhao Yuan. On the convergence of Muon and beyond. *arXiv preprint arXiv:2509.15816*, 2025. URL https://arxiv.org/abs/2509.15816.

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

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2260–2268, 2020. URL https://proceedings.mlr.press/v119/cutkosky20b.html.

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. URL https://arxiv.org/abs/2608.04607v1. Version 1, August 5, 2026.

Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. *Mathematical Programming*, 187:57–79, 2021. URL https://arxiv.org/abs/1910.01845v3. Theorem 7; also arXiv:1910.01845v3.

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

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, December 2024. URL https://kellerjordan.github.io/posts/muon/. Blog post, 8 December 2024; reference implementation at https://github.com/KellerJordan/Muon.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261, 2019. URL https://proceedings.mlr.press/v97/karimireddy19a.html.

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. *arXiv preprint arXiv:2601.19156*, 2026. URL https://arxiv.org/abs/2601.19156.

Alexey Kushnir and Vinod Krishnamoorthy. A simple characterization of supply correspondences. *arXiv preprint arXiv:2205.10472*, 2022. URL https://arxiv.org/abs/2205.10472.

Mingyi Li and Taira Tsuchiya. Muon with finite Newton–Schulz: The smoothing benefit in nonsmooth nonconvex optimization. *arXiv preprint arXiv:2608.26288*, 2026. URL https://arxiv.org/abs/2608.26288.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, et al. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025. URL https://arxiv.org/abs/2502.16982.

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. URL https://arxiv.org/abs/2605.08980v1. Version 1, May 9, 2026.

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

Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: a review. *Statistics Surveys*, 8:45–114, 2014. URL https://arxiv.org/abs/1404.5886.

Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the ideal: Analyzing the inexact Muon update. *arXiv preprint arXiv:2510.19933*, 2025. URL https://arxiv.org/abs/2510.19933.

Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of SIGNSGD under weaker assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33077–33099, 2023. URL https://proceedings.mlr.press/v202/sun23l.html.

Hongyi Tao, Dingzhi Yu, and Lijun Zhang. When and why SignSGD outperforms SGD: A theoretical study based on ℓ_1-norm lower bounds. *arXiv preprint arXiv:2605.06615*, 2026. URL https://arxiv.org/abs/2605.06615.

A ASSUMPTIONS AND SCOPE

Some results in this paper describe a whole class of update maps, while others concern a particular task and training budget. The tables below state what each result guarantees and the conditions needed to use them. They also distinguish proved statements from measurements and approximations.

Table 3: Principal results and the conditions needed to use them.

Result	Statement	Essential scope
Prop. 1; Thms. 2, 16	Symmetric-noise alignment is equivalent to monotonicity. Frobenius normalization forces a linear response; Schatten-q normalization forces cs^{q-1}.	Bounded odd maps for the equivalence. Rigidity needs at least three singular values and the stated scalar regularity; the rank-two exception is explicit.
Prop. 3; Thm. 17	Exact fixed-scale/normalization and radial/tangential decompositions; the negative cone and exactly one negative eigenvalue when $T \succ 0$ and $V_f(s) > 0$.	A local map statement. The tuned reference can have indefinite T and several negative modes.
Prop. 20; Appendix U	Independent gradient cross-products remove quadratic covariance bias. All 600 fresh layer point estimates remain positive; the seed-level mean has a positive interval.	Same gradient distribution, independent of the captured input. Eight complete seeds; no simultaneous claim about all population layer signs.
Prop. 21	Actual conditional alignment is history alignment plus integrated current-signal response; smoothness gives a finite-step loss interval.	Conditional expectations and stated segment regularity. Negative local response alone need not imply ascent.
Cor. 4; Thm. 45	Actual zero-initialized EMA has certified nonzero-gradient residence at $\beta = .1$ and .3.	Fixed smooth convex tasks, bounded Rademacher matrix noise and small full-matrix perturbations; explicit finite horizons and step budgets.
Props. 5, 6; Cor. 22	Convex coupling preserves alignment and flattening. Smoothed nuclear has bounded gain, nonzero-input smoothness and a finite-response error margin.	Exact maps unless endpoint forward errors are bounded. No training-loss ranking; replication-invariant ratios are shared with Schatten.
Appendix W	Direct paired losses favour implemented Gram Schatten-1.05 over smoothed nuclear at both scales.	One corpus, two scales of one model family, float64 maps. Timing reverses across shapes; only the smaller scale has an approximate equal-time reference.

The frozen-field worst-case scaling takes no infimum over step sizes and is not a minimax trajectory risk. The quartic repair contracts on its specified transverse task; an uncoupled ablation can repair that task while violating universal monotonicity. A small finite-step locality flag is not an error certificate. The computational proof artifacts verify only the premises explicitly listed in Appendix X.

B THE TRANSVERSE TASK AND ITS EXACT REDUCTION

We define a small matrix task in which the gradient and the noise occupy orthogonal directions. The noisy gradient therefore reveals the signal component exactly. The example tests whether transforming the whole matrix can nevertheless change the update along that signal. The matrices below make the calculation explicit.

Definition 7 (The fixed transverse task)**.** Let $E = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$, $C_0 = I/\sqrt{2}$ and $C_1 = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, which are Frobenius-orthonormal with $C_0^\top C_1$ nonzero and antisymmetric. Fix $0 < \sigma \le 1/16$ and take per-sample noise $\xi = \frac{\sigma}{\sqrt{2}}(UC_0 + VC_1)$ with U, V uniform on $[-1, 1]$ and independent of each other, across steps and within a batch, so that $\|\xi\|_F \le \sigma$, $\mathbb{E}\xi = 0$, $\mathbb{E}\|\xi\|_F^2 = \sigma^2/3$ and $\xi \overset{d}{=} -\xi$. Take $F(W) = \phi(\langle E, W\rangle)$ with ϕ the Huber function, $\phi(x) = x^2/2$ for $|x| \le 1$ and $|x| - 1/2$ otherwise,

Table 4: Complementary constructions and approximations. Their guarantees are separate from the principal finite-window examples.

Result	Guarantee and limitation
High momentum, Appendix J.2	Classical cubic constructions at rank 128/depth five and rank 512/depth eight, $\beta = .9, .95$, show numerical persistence. Statistical averaged-field brackets do not certify trajectories. The weakest regime uses steps near 10^{-8} and up to 2.97×10^8 updates.
Transverse task, Thm. 31	The actual-EMA theorem covers a fixed Huber ridge and uniform transverse noise at $K \in \{1, 5\}$. Its displayed feasible instance has $T = 10^{12}$ and a very small step; the shorter reported simulations are outside that certificate.
Averaging and mean shift, Appendix K	The averaging/CLT is for fixed B, β as the step tends to zero on a positive compact branch. The colored-noise filter is exact as a linear model. A nonlinear first-order mean correction still requires unproved $o(h)$ covariance, variance and curvature errors. The sealed median is .6% in ten local cells and 2.1% over all sixteen.
Local loss diagnostics, Appendix V	Midpoint Taylor prediction applies to every predetermined pair. Direct paired evaluation is more accurate and cheaper per pair here; direction ablations do not establish unique recovery information. Frozen 10/50-step incremental gains have intervals crossing zero.

and $W_0 = x_0 E$ with $x_0 = 1/4$. Then F is convex and 1-smooth, the population gradient has norm at most one, each sample gradient at most $\sqrt{1+\sigma^2}$, the initial gap is $1/32$, and nothing here depends on β, B, η or T.

Because the gradient always points along E, writing $x_t = \langle E, W_t \rangle$ and $M_t = a_t E + Z_t$ with $Z_t \in \operatorname{span}\{C_0, C_1\}$ splits (2) into $a_t = \beta a_{t-1} + (1-\beta)\phi'(x_t)$ and $Z_t = \beta Z_{t-1} + (1-\beta)\bar{\xi}_t$, exactly and not on average.

Lemma 8 (Exact reduction). *For any $Z \in \operatorname{span}\{C_0, C_1\}$ with $\|Z\|_F = \rho > 0$ and any $a \in \mathbb{R}$,*

$$\langle E, \Psi_K(aE + Z) \rangle = q_K(a/\rho), \qquad q_K(t) = \frac{1}{\sqrt{2}} \left[f_K\Big(\tfrac{1+t}{\sqrt{2(1+t^2)}}\Big) - f_K\Big(\tfrac{1-t}{\sqrt{2(1+t^2)}}\Big) \right], \quad (14)$$

so the projection depends on Z only through its norm, never its direction, and $x_{t+1} = x_t - \eta q_K(a_t/\rho_t)$ with $\rho_t = \|Z_t\|_F$. At $\rho_t = 0$, the projected update is $\sqrt{2} f_K(1/\sqrt{2}) \operatorname{sign}(a_t)$ if $a_t \neq 0$ and zero if $a_t = 0$.

The proof factors $aE + Z$ into a rotation and a symmetric matrix with eigenvalues proportional to $1 \pm t$ (Appendix C). At depth one $q_1(t) = t(-4.7145 + 9.671t^2 + 6.2595t^4)/(4(1+t^2)^{5/2})$, whose positive root $t_\star = 0.62399$ satisfies

$$q_K(t_\star) = 0 \qquad \text{for every } K \geq 1, \quad (15)$$

because the first application sends two distinct positive singular values to one value. Near the origin $q_K'(0) = f_K'(1/\sqrt{2})$, with values -1.1786, -1.3352, $+1.6123$, $+1.6063$, -1.8804 for $K \leq 5$ and

$$\operatorname{sign} f_K'(1/\sqrt{2}) = (-, -, +, +, -, -, +, +, \dots) \quad \text{for } 1 \leq K \leq 31, \quad (16)$$

first breaking at $K = 32$ (Appendix C). Depth does not order the recovery signs; Section 5 shows these local signs transfer to the frozen field for every (B, β).

C Spectral map algebra and the rigidity theorem

We first write the spectral update as a matrix polynomial and derive its Jacobian. These identities explain why differentiating the normalization matters. We then prove the alignment characterization and rigidity theorem from Section 3. The final part derives the scalar response for the transverse task in Appendix B and checks the constants used at each iteration depth.

C.1 THE MAP, AND WHY THE NORMALIZATION MATTERS

For an odd polynomial $f(s) = \sum_{j=0}^{J} a_j s^{2j+1}$, the singular-value formula in (1) equals the matrix-product expression given after it, since $(PP^\top)^j P = U\Sigma^{2j+1}V^\top$ for $P = U\Sigma V^\top$, so Ψ_f is defined without reference to a singular value decomposition and requires neither full rank nor a gap. It is odd, zero-homogeneous away from the origin, and smooth there. Zero-homogeneity is used repeatedly below and has one consequence worth isolating.

Lemma 9 (Euler relation and the radial null direction)**.** *For $P \neq 0$, $D\Psi_f(P)[P] = 0$ and $\|D\Psi_f(P)\|_{\mathrm{op}} \cdot \|P\|_F$ is scale invariant. On the subspace of matrices sharing the singular frames of P, with $\varsigma_i = \sigma_i/\|P\|_F$, the Jacobian acts as $A = D(I - \varsigma\varsigma^\top)$ where $D = \mathrm{diag}(f'(\varsigma_1), \ldots, f'(\varsigma_r))$, after multiplication by $\|P\|_F$.*

Proof. Differentiating $\Psi_f(cP) = \Psi_f(P)$ at $c = 1$ gives $D\Psi_f(P)[P] = 0$; applying the same identity to the operator norm gives the scale invariance. For the second claim, write $X = P/\|P\|_F$ and perturb P within the frame-preserving subspace, so only the singular values move. Then $\sigma_i \mapsto \sigma_i + h_i$ and $\|P\|_F \mapsto \|P\|_F + \langle \varsigma, h\rangle + O(h^2)$, hence $\varsigma_i \mapsto \varsigma_i + (h_i - \varsigma_i\langle \varsigma, h\rangle)/\|P\|_F + O(h^2)$, and the response of the ith output singular value is $f'(\varsigma_i)$ times that. In vector form this is $D(I - \varsigma\varsigma^\top)h/\|P\|_F$, and absorbing the scale gives A as stated. □

The factor $I - \varsigma\varsigma^\top$ is the derivative of the *endogenous* normalization and is what makes Theorem 2 true. With an externally fixed scale $c > 0$, the frame-preserving Jacobian is $\mathrm{diag}(f'(\sigma_i/c))/c$, with no projection factor. Algorithm 1 of Li & Tsuchiya (2026) uses precisely such a fixed input bound G_{op}; it does not divide by the current momentum's operator norm.

C.2 AN EXACT DIAGNOSTIC ON THE FULL MATRIX SPACE

The same calculation extends to all perturbations. The following spectral differentiation formula makes it possible to measure both an adverse direction and the actual gradient's overlap with adverse directions without forming an $mn \times mn$ Jacobian or truncating P. It is a computational consequence of the spectral map, rather than a separate convergence guarantee.

Proposition 10 (Full symmetric-Jacobian decomposition)**.** *Let f be an odd polynomial, $P \neq 0$, $R = \|P\|_F$, and $s_i = \sigma_i(P)/R$, $1 \leq i \leq r = \min(m, n)$, including zero singular values. Put $D = \mathrm{diag}(f'(s_i))$ and $S = \frac{1}{2}[D(I - ss^\top) + (I - ss^\top)D]$. The spectrum of $R\mathcal{A}(P)$, counting multiplicities, is the union of the spectrum of S, one copy each of*

$$\frac{f(s_i) - f(s_j)}{s_i - s_j}, \qquad \frac{f(s_i) + f(s_j)}{s_i + s_j} \quad (i < j), \tag{17}$$

and $|m - n|$ copies of $f(s_i)/s_i$ for each i. A zero denominator is evaluated by its continuous limit: $f'(s_i)$ for the first quotient and $f'(0)$ for the others. The same statement holds for $f \in C^1([0, 1])$, $f(0) = 0$, interpreted by odd extension.

Proof. Orthogonal equivariance permits $P = R\,\mathrm{diag}(s)$, padded with zeros. In these coordinates the Frobenius-orthogonal decomposition of a perturbation consists of its diagonal entries, the symmetric pairs $(E_{ij} + E_{ji})/\sqrt{2}$, the antisymmetric pairs $(E_{ij} - E_{ji})/\sqrt{2}$, and the rectangular complement. Differentiating $F_f(X) = \sum_j a_j(XX^\top)^j X$ on these blocks gives, respectively, D, the two divided differences in (17), and $f(s_i)/s_i$. The derivative of P/R is $(H - X\langle X, H\rangle)/R$, so its extra projection acts only on the diagonal block. That block becomes $D(I - ss^\top)/R$; taking its symmetric part proves the formula. The other blocks are already self-adjoint. Their dimensions sum to $r + 2\binom{r}{2} + |m - n|r = mn$. The polynomial formulas have the stated limits at repeated and zero singular values. For a C^1 response, approximate its even continuous derivative uniformly on $[-1, 1]$ by even polynomials and integrate from zero. The resulting odd polynomials converge in C^1, and their divided differences converge uniformly to those displayed above. Thus the derivative formula extends to this case as well, without a singular-value gap. □

For a reference gradient $G \neq 0$, expand $G/\|G\|_F$ in these orthonormal blocks, and diagonalize only the $r \times r$ matrix S. Summing squared coefficients on negative eigenvalues gives the full-space negative-energy fraction; weighting the same coefficients by the eigenvalues gives $\langle G, \mathcal{A}(P)[G]\rangle/\|G\|_F^2$. The

former can be positive while the latter is positive too. Their distinction is essential: adverse overlap need not outweigh the restoring directions. For the coupled quartic, the same block decomposition applies with $y_i = s_i(\alpha - 4\lambda s_i^2)$, $\alpha = 1 + 3\lambda \sum_i s_i^4$, and the diagonal derivative before normalization

$$B = \operatorname{diag}(\alpha - 12\lambda s_i^2) + 12\lambda s (s^{\odot 3})^\top.$$

Replace D by B in the diagonal block and $f(s_i)$ by y_i in the other blocks. The implementation evaluates all singular values and checks this formula against a separately assembled dense automatic-differentiation Jacobian, including rectangular, rank-deficient and repeated-spectrum cases (Appendix R).

C.3 Proposition 1

Proof. Let Ψ be bounded, Borel measurable and odd, so every expectation below exists. If $Z \overset{d}{=} -Z$ then replacing Z by $-Z$ and averaging,

$$\langle g, \mathbb{E}\Psi(g + Z)\rangle = \tfrac{1}{2}\mathbb{E}\langle g, \Psi(Z + g) - \Psi(Z - g)\rangle. \tag{18}$$

If Ψ is monotone, each realization of the bracket is nonnegative with $A = Z + g$, $B = Z - g$ and $A - B = 2g$, so the left side is nonnegative. Conversely, given A, B, let Z take the values $\pm(A+B)/2$ with probability $1/2$ each, which is bounded and centrally symmetric, and set $g = (A - B)/2$. Then $Z + g$ and $Z - g$ equal A and B on the first atom and $-B$ and $-A$ on the second, so by oddness (18) evaluates to $\frac{1}{2}\langle (A - B)/2, \Psi(A) - \Psi(B)\rangle$, and nonnegativity gives the secant inequality up to the positive factor $1/4$, which does not affect its sign and so closes the equivalence. The formula for the worst alignment follows because (18) exhibits it as an average of the deterministic quantity $\frac{1}{2}\langle g, \Psi(z + g) - \Psi(z - g)\rangle$ over z in the ball, and the infimum of an average over a family of laws is the infimum of the integrand. Taking z along a sequence approaching that infimum and using the two-point law $\pm z$ shows symmetric two-point laws approach it. We do not claim the infimum is attained: Ψ is assumed only bounded and measurable, so $z \mapsto \langle g, \Psi(z + g) - \Psi(z - g)\rangle$ need not be lower semicontinuous and a minimizing z need not exist. Nothing below uses attainment, and the equivalence itself does not need it. □

No derivative, density, nondegeneracy, or spectral gap enters. The equivalence also explains why central symmetry cannot be the relevant hypothesis: it only certifies $b_w = h_w(0) = 0$, whereas the question of whether the field restores toward the origin is the secant condition.

C.4 Theorem 2

Proof. Sufficiency: if $f(s) = cs$ with $c \geq 0$ then $\Psi_f(P) = cP/\|P\|_F$, which is c times an element of $\partial\|P\|_F$, and the subdifferential of a convex function is monotone.

Necessity. Suppose Ψ_f is monotone with $f \in C^1([0, 1])$, $f(0) = 0$ and $r \geq 3$. Monotone maps that are differentiable have positive semidefinite symmetric Jacobian part, so $S = (A + A^\top)/2 \succeq 0$ at every $P \neq 0$. Take $P = \operatorname{diag}(s_1, s_2, s_3)$ with $s_i > 0$ and $\|s\|_2 = 1$, padded with zero rows or columns to the ambient shape; a matrix with distinct frames is not needed, and padding does not change the nonzero singular values. Restrict attention to the frame-preserving subspace, where Lemma 9 gives $A = D(I - ss^\top)$ with $D = \operatorname{diag}(f'(s_1), f'(s_2), f'(s_3))$.

Since $As = D(s - s\|s\|^2) = 0$, we get $s^\top Ss = s^\top As = 0$. A positive semidefinite matrix with $s^\top Ss = 0$ annihilates s, because for any H and $\theta \in \mathbb{R}$, $0 \leq (s + \theta H)^\top S(s + \theta H) = 2\theta H^\top Ss + \theta^2 H^\top SH$, and letting $\theta \to 0$ with either sign forces $H^\top Ss = 0$ for all H. Hence $Ss = 0$, that is $As + A^\top s = 0$, and as $As = 0$ this gives $A^\top s = 0$, i.e. $(I - ss^\top)Ds = 0$. So Ds is parallel to s: there is μ with $f'(s_i)s_i = \mu s_i$ for each i, and since every $s_i > 0$,

$$f'(s_1) = f'(s_2) = f'(s_3) = \mu. \tag{19}$$

Now let $a, b \in (0, 1)$ be arbitrary. Choose $c_0 > 0$ small enough that $a^2 + c_0^2 < 1$ and $b^2 + c_0^2 < 1$, and apply (19) to the normalized triples containing (a, c_0) and (b, c_0) respectively, filling the third coordinate so the vector has unit norm and all entries positive. This is possible precisely because $r \geq 3$: with two singular values the vector (a, c_0) normalized has no free coordinate left, and the argument gives no conclusion. Both applications yield $f'(a) = f'(c_0)$ and $f'(b) = f'(c_0)$, hence

$f'(a) = f'(b)$. So f' is constant on $(0,1)$, and continuity extends it to $[0,1]$; with $f(0) = 0$ this gives $f(s) = cs$. Monotonicity of $s \mapsto cs$ under Ψ_f requires $c \geq 0$, since $c < 0$ makes $\Psi_f = -|c|P/\|P\|_F$ antimonotone. □

Proposition 11 (The dimension threshold is sharp)**.** *Let* $\min(m,n) = 2$*,* $f \in C^1([0,1])$*, and* $f(0) = 0$*. Then* Ψ_f *is monotone if and only if*

$$f'(u) = f'\big(\sqrt{1-u^2}\big) \geq 0 \qquad (0 < u < 1). \tag{20}$$

In particular, $f(u) = u - u^3/3 + u^5/5$ *is nonlinear and gives a monotone map in this dimension.*

Proof. At a positive unit spectrum $s = (u, \sqrt{1-u^2})$, the radial-null argument in Theorem 2 forces the two derivatives to agree. The diagonal block is then $d(I - ss^\top)$, whose other eigenvalue is d, so monotonicity also forces $d \geq 0$. Conversely, (20) makes this block positive semidefinite. Since $f' \geq 0$ and $f(0) = 0$, all the divided differences and rectangular coefficients in Proposition 10 are nonnegative. Thus $\mathcal{A}(P) \succeq 0$ for every $P \neq 0$, including boundary spectra by continuity. Integrating along any segment avoiding zero gives the secant inequality. A segment crossing zero has opposite collinear endpoints; oddness, zero-homogeneity and $\langle P, \Psi_f(P)\rangle = R\sum_i s_i f(s_i) \geq 0$ give the inequality there too. Finally the example has derivative $1 - u^2 + u^4 \geq 3/4$, invariant under $u \mapsto \sqrt{1-u^2}$, so it satisfies (20). □

Thus three singular values are necessary for the general rigidity conclusion. This does not exempt the reference polynomial in two dimensions: its derivatives fail the condition, as the negative recovery slope on our 2×2 task explicitly demonstrates.

Proposition 12 (A quantitative adverse direction)**.** *At a unit spectrum* s *with positive entries, put* $d_i = f'(s_i)$*,* $D = \operatorname{diag}(d_i)$*,* $\bar{d} = \sum_i s_i^2 d_i$ *and*

$$V_f(s) = \sum_i s_i^2 (d_i - \bar{d})^2 = \sum_{i<j} s_i^2 s_j^2 (d_i - d_j)^2 = \|b\|^2, \qquad b = (I - ss^\top)Ds. \tag{21}$$

If $V_f(s) > 0$*, let* $d_b = b^\top Db / V_f(s)$ *and* $M = \|D\|_{\mathrm{op}}$*. Then, for* $S = (D(I-ss^\top) + (I-ss^\top)D)/2$*,*

$$\lambda_{\min}(S) \leq -\gamma_f(s), \qquad \gamma_f(s) = \frac{\sqrt{d_b^2 + V_f(s)} - d_b}{2} \geq \frac{V_f(s)}{2\big(\sqrt{M^2 + V_f(s)} + M\big)} > 0. \tag{22}$$

Conversely, for any $\tau \geq 0$*,*

$$S \succeq -\tau I \quad \Longrightarrow \quad V_f(s) \leq 4\tau(M + \tau). \tag{23}$$

The unit direction attaining $-\gamma_f(s)$ *on* $\operatorname{span}\{s, b\}$ *is proportional to* $s - 2\gamma_f(s)b/V_f(s)$*. A simpler explicit direction* $H = s - \delta b$ *satisfies* $H^\top AH \leq -\frac{\delta}{2}\|b\|^2 < 0$ *for every* $0 < \delta \leq [2(1 + \|D\|_{\mathrm{op}})]^{-1}$*, and* $\|H\|_F^2 = 1 + \delta^2\|b\|^2$ *since* $\langle s, b\rangle = 0$*.*

Proof. Since $\sum_i s_i^2 = 1$, $b_i = s_i(d_i - \bar{d})$, which proves the first and last forms in (21); expanding the pair sum proves the middle one. Write $u = b/\sqrt{V_f(s)}$. The vectors s, u are orthonormal, $As = 0$, and $Au = Du$. Therefore the compression of S to their span is exactly

$$\begin{pmatrix} s^\top Ss & s^\top Su \\ u^\top Ss & u^\top Su \end{pmatrix} = \begin{pmatrix} 0 & \sqrt{V_f(s)}/2 \\ \sqrt{V_f(s)}/2 & d_b \end{pmatrix}.$$

Its smaller eigenvalue is $-\gamma_f(s)$, so the variational principle gives the first bound in (22) and its eigenvector gives the stated direction. The function $d \mapsto (\sqrt{d^2 + V} - d)/2$ decreases in d, and $d_b \leq M$, which gives the lower bound on $\gamma_f(s)$ after rationalizing. If $S + \tau I$ is positive semidefinite then so is this compression plus τI; its determinant gives $V_f(s)/4 \leq \tau(d_b + \tau) \leq \tau(M + \tau)$, proving (23).

For the simpler direction, using $As = 0$ and $A = D(I - ss^\top)$, $H^\top AH = (s - \delta b)^\top A(s - \delta b) = -\delta s^\top Ab + \delta^2 b^\top Ab$. Now $Ab = D(I - ss^\top)b = Db$ because $\langle s, b\rangle = \langle s, (I - ss^\top)Ds\rangle = 0$, and $s^\top Db = \langle Ds, b\rangle = \langle (I - ss^\top)Ds + s\langle s, D\rangle s, b\rangle = \|b\|^2$. Hence $H^\top AH = -\delta\|b\|^2 + \delta^2 b^\top Db$. Since $|b^\top Db| \leq \|D\|_{\mathrm{op}}\|b\|^2 \leq (1 + \|D\|_{\mathrm{op}})\|b\|^2$, the stated range of δ gives $\delta^2|b^\top Db| \leq \frac{\delta}{2}\|b\|^2$ and the claim follows. Choosing δ this way rather than by comparison with $\|b\|^2/|b^\top Db|$ avoids dividing by a quantity that vanishes whenever $b^\top Db = 0$, which is possible here. □

The variance $V_f(s)$ measures heterogeneity of the spectral slopes at the observed spectrum; it is zero even for a nonlinear f at spectra where those slopes happen to agree. Thus a nonlinear map need not have a negative direction at every input. Equation (22) gives a sufficient, quantitative certificate at a given input, while (23) says that approximate monotonicity requires small weighted slope variation. Both concern the frame-preserving subspace; other matrix directions can be more negative. For P of norm different from one the response eigenvalues scale as $\|P\|_F^{-1}$.

For $f \in C^1$, an immediate finite-scale witness is $Z = \pm P$, $g = hH$, since symmetric pairing gives $\langle g, \mathbb{E}\Psi_f(g+Z)\rangle = h^2\langle H, D\Psi_f(P)[H]\rangle + o(h^2) < 0$ for small enough $h > 0$. For the smooth polynomial maps used here, the following result gives an explicit radius and allows a full-dimensional smooth noise density.

Corollary 13 (Finite-signal and noise perturbation margins)**.** *Let Ψ be odd, continuously differentiable on the balls of radius r_0 around $\pm P$, and suppose its Jacobian on the positive ball satisfies $\|D\Psi(Q) - D\Psi(P)\|_{\mathrm{op}} \le L\|Q - P\|_F$. Let $\|H\|_F = 1$ and $\langle H, D\Psi(P)[H]\rangle \le -\gamma < 0$. For any random U with $\|U\|_F \le \nu$, independent Rademacher sign ϵ, and $Z = \epsilon(P + U)$, if*

$$h > 0, \qquad \nu + h < r_0, \qquad L(\nu + h/2) \le \gamma/2,$$

then

$$\langle hH, \mathbb{E}\Psi(hH + Z)\rangle \le -\tfrac{1}{2}\gamma h^2. \tag{24}$$

In particular U may have a centered, smooth density supported on the radius-ν ball, giving bounded centrally symmetric noise with a smooth density and positive definite covariance. For $F(W) = \frac{1}{2}\langle H, W\rangle^2$, a fresh stochastic step from $W = hH$ satisfies

$$\mathbb{E}F\big(W - \eta\Psi(\nabla F(W) + Z)\big) - F(W) \ge \tfrac{1}{2}\eta\gamma h^2 \qquad \textit{for every } \eta > 0. \tag{25}$$

Proof. Conditioning on U and averaging the two signs, oddness gives

$$\frac{\langle hH, \mathbb{E}\Psi(hH + Z)\rangle}{h^2} = \frac{1}{2}\mathbb{E}\int_{-1}^{1} \langle H, D\Psi(P + U + thH)[H]\rangle\, dt.$$

All points of integration lie in the positive ball. The integrand is at most $-\gamma + L(\nu + |t|h)$, so the right side is at most $-\gamma + L(\nu + h/2) \le -\gamma/2$. This proves (24); no independence among the entries of U is required. A smooth radial bump density for U has positive definite covariance; the symmetric mixture of its translates about P and $-P$ retains that property. Finally write $Y = \langle H, \Psi(hH + Z)\rangle$. The exact quadratic identity $\mathbb{E}F(W - \eta\Psi(hH + Z)) - F(W) = -\eta h\mathbb{E}Y + \eta^2\mathbb{E}Y^2/2$ and (24) give (25). □

The local constant can be bounded explicitly without spectral gaps. If $f(s) = \sum_{j=0}^{J} a_j s^{2j+1}$, $\|P\|_F = 1$ and $0 < r_0 < 1$, a valid choice is

$$L = \frac{C_2 + 6C_1}{(1 - r_0)^2}, \qquad C_1 = \sum_{j=0}^{J} |a_j|(2j+1), \quad C_2 = \sum_{j=0}^{J} |a_j|(2j+1)(2j). \tag{26}$$

To see this, write $\Psi_f = F_f \circ N$ with $N(Q) = Q/\|Q\|_F$ and $F_f(X) = \sum_j a_j (XX^\top)^j X$. Product differentiation and $\|X\|_{\mathrm{op}} \le \|X\|_F = 1$ give $\|DF_f(X)\|_{\mathrm{op}} \le C_1$ and $\|D^2F_f(X)\|_{\mathrm{op}} \le C_2$. Direct differentiation of N gives $\|DN(Q)\|_{\mathrm{op}} \le \|Q\|_F^{-1}$ and $\|D^2N(Q)\|_{\mathrm{op}} \le 6\|Q\|_F^{-2}$. The second-order chain rule, followed by integration along the line from P to Q, proves (26). A tighter local bound can enlarge the permitted perturbation radius; the stated coefficient bound is only a sufficient one. These are finite-signal and one-step consequences for the specified noise laws, not a residence theorem for correlated EMA noise or a prediction about unmeasured training gradients.

Two map-class consequences follow at once. For any fixed nonlinear finite Newton–Schulz polynomial the conclusion applies whatever the depth and whether or not the scalar response increases on $[0, 1]$; and for $\vartheta < 1$ the mixture $(1 - \vartheta)f(s) + \vartheta s$ is still a nonlinear spectral response, so mixing with normalized descent does not restore monotonicity.

C.5 LEMMA 8, AND THE DEPTH CONSTANTS

Proof. Let $Z \in \operatorname{span}\{C_0, C_1\}$ with $\|Z\|_F = \rho > 0$, and write $Z = \rho(\cos\theta\, C_0 + \sin\theta\, C_1)$, which is possible since C_0, C_1 are Frobenius-orthonormal. Put $R_\theta = \sqrt{2}(\cos\theta\, C_0 + \sin\theta\, C_1)$. Explicitly $R_\theta = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$, a rotation, so $R_\theta^\top R_\theta = I$. Also $\sqrt{2}E$ is symmetric orthogonal, hence $S := R_\theta^\top \sqrt{2}E$ is orthogonal, and it is symmetric because $S^\top = \sqrt{2}ER_\theta$ and a direct computation gives $\sqrt{2}ER_\theta = R_\theta^\top \sqrt{2}E$ for the two-dimensional reflection and rotation above. An orthogonal symmetric matrix has eigenvalues ± 1, and $\operatorname{tr} S = 0$ here, so its spectrum is exactly $\{+1, -1\}$. With $t = a/\rho$,

$$aE + Z = \rho\big(tR_\theta S/\sqrt{2} + R_\theta/\sqrt{2}\big) = \frac{\rho}{\sqrt{2}} R_\theta (I + tS).$$

Because S is symmetric with eigenvalues ± 1 and R_θ is orthogonal, this is a polar form: the singular values of $aE + Z$ are $\frac{\rho}{\sqrt{2}}|1 \pm t|$, its Frobenius norm is $\rho\sqrt{1+t^2}$, and the eigenvectors of S give the singular frames. Applying an odd spectral polynomial to the normalized singular values $|1 \pm t|/\sqrt{2(1+t^2)}$ and projecting onto $E = R_\theta S/\sqrt{2}$ retains the eigenvalue signs, giving

$$\langle E, \Psi_K(aE + Z)\rangle = \frac{1}{\sqrt{2}}\Big[f_K\big(\tfrac{1+t}{\sqrt{2(1+t^2)}}\big) - f_K\big(\tfrac{1-t}{\sqrt{2(1+t^2)}}\big)\Big],$$

which is (14). Oddness of f_K makes the second argument's sign immaterial, so the formula covers $|t| > 1$ as well, and the right side does not involve θ. □

Nothing in the proof needs θ to be uniformly distributed, so the identity applies to the correlated Z_t produced by the EMA recursion and not only to a single fresh sample. The same computation with f replaced by any odd spectral function gives the corresponding profile, which is how the entries of Table 5 are obtained.

At depth one, expanding (14) with the reference coefficients gives $q_1(t) = t(-4.7145 + 9.671t^2 + 6.2595t^4)/(4(1+t^2)^{5/2})$, whose numerator has the unique positive root in t^2

$$t_\star^2 = \frac{-9.671 + \sqrt{9.671^2 + 4(6.2595)(4.7145)}}{2(6.2595)} = 0.3893635522184,$$
$$t_\star = 0.6239900257363.$$

Since $q_1(t_\star) = 0$ means f_1 maps the two distinct normalized singular values $(1 \pm t_\star)/\sqrt{2(1+t_\star^2)}$ to the same value, every further application of the same function keeps them equal, so $q_K(t_\star) = 0$ for every $K \geq 1$, which is (15). Passing $t_\star$ does not make the response positive, and no statement of that form is available at all depths: at $K = 5$ the certificate of Appendix P encloses $q_5(0.7) \in [-0.2240492, -0.2240491]$, strictly negative although $0.7 > t_\star$. The region where every depth is positive is $t \geq 1$.

Lemma 14 (All-depth positive region)**.** *With $\mu = 2594691/4063000$ one has $p(u) \geq \mu u$ for $u \geq 0$, hence $f_K(u) \geq \mu^K u$ on $u \geq 0$, and $q_K(t) \geq \mu^K t/\sqrt{1+t^2} > 0$ for every $t \geq 1$ and $K \geq 1$. At the shared zero $t_\star$ one has the identity*

$$q_K'(t_\star) = f_{K-1}'(v)\, q_1'(t_\star), \qquad q_1'(t_\star) = 1.2445504 > 0,$$
$$v = p\big(\tfrac{1+t_\star}{\sqrt{2(1+t_\star^2)}}\big) = 0.7233506. \tag{27}$$

so whenever $f_{K-1}'(v) \neq 0$ the zero is simple and the sign of q_K immediately above $t_\star$ is the sign of $f_{K-1}'(v)$.

Proof. The exact factorization $p(u) - \mu u = u(4063u^2 - 4775)^2/8{,}126{,}000 \geq 0$ gives the first claim, and μ is sharp, being the exact minimum of $p(u)/u$ over $u > 0$, attained at $u = 5\sqrt{776033}/4063 = 1.0841$. Iterating needs no monotonicity of p, which would be false: $p' < 0$ on $(0.5545, 1.0501)$, an interval containing $1/\sqrt{2}$, so p is strictly *decreasing* where the orbit starts. What the induction uses is that the envelope holds at every nonnegative argument, giving $f_K(u) = p(f_{K-1}(u)) \geq \mu f_{K-1}(u) \geq \mu^K u$ with $f_{K-1}(u) \geq 0$ by induction. For the same reason the sharper constant $p(1) = 0.701$ on $[0, 1]$ is unusable: $p(1/\sqrt{2}) = 1.1065$ leaves the unit interval at the first step. For $t \geq 1$ both

arguments of (14) are nonnegative, because $1-t \le 0$ makes the second argument $(1-t)/\sqrt{2(1+t^2)}$ nonpositive and oddness of f_K turns $-f_K$ of it into $+f_K$ of its modulus. So the two signed spectral contributions *add* rather than cancel, and

$$q_K(t) = \frac{1}{\sqrt{2}}\Big[f_K\big(\tfrac{1+t}{\sqrt{2(1+t^2)}}\big) + f_K\big(\tfrac{t-1}{\sqrt{2(1+t^2)}}\big)\Big] \;\ge\; \frac{\mu^K}{\sqrt{2}} \cdot \frac{(1+t)+(t-1)}{\sqrt{2(1+t^2)}} = \frac{\mu^K t}{\sqrt{1+t^2}},$$

which is strictly positive. The bound is used at $K = 5$ in Appendix P to certify the outer barrier. For (27), the collapse that proves (15) sends both arguments to the common value $v = p(u_\pm(t_\star))$ after one application, so $f_K = p^{\circ(K-1)} \circ p$ and the chain rule at $t_\star$ gives $q_K'(t_\star) = f_{K-1}'(v)\, q_1'(t_\star)$, the derivative of the outer iterate being evaluated at the single point v because the two arguments have already merged there. Since $q_1'(t_\star) > 0$, a nonzero $f_{K-1}'(v)$ makes the zero simple with the stated sign. Numerical evaluation for $K = 1, \dots, 10$ gives the signs $(+, -, -, +, +, -, -, +, +, -)$; this finite check does not assert simplicity or this sign pattern at every depth. In particular q_K is negative immediately above $t_\star$ at $K = 2, 3, 6, 7$, with $q_2'(t_\star) = -1.5805505$. This is a stronger refutation of the withdrawn positivity claim than the point value $q_5(0.7)$, since it fails arbitrarily close to the root. □

Near the origin the two arguments of (14) are $1/\sqrt{2} \pm t/\sqrt{2} + O(t^2)$ with the $O(t^2)$ terms equal, so $q_K'(0) = f_K'(1/\sqrt{2})$ and the chain rule gives the product

$$f_K'(1/\sqrt{2}) = \prod_{j=0}^{K-1} p'\big(p^{\circ j}(1/\sqrt{2})\big), \tag{28}$$

so the sign is decided by how many negative factors the orbit of $1/\sqrt{2}$ collects. That is a question about the orbit, and the answer is different at small and large K.

Why the sign alternates in fours, and only for a while. The map p has a two-cycle $\{u_-, u_+\}$ with $u_- = 0.70926$ and $u_+ = 1.10398$, and $p'(u_-) < 0 < p'(u_+)$, so a full lap contributes one negative factor and the sign of (28) changes every second step: that is the period-four pattern. These two values and the multiplier below are computed in floating point and are *not* part of the interval certificate, which covers the four-cycle boxes; nothing downstream depends on them beyond the qualitative repulsion. The orbit starts at $1/\sqrt{2} = 0.70711$, within 2.2×10^{-3} of u_-, which is why the pattern is visible at all. But the two-cycle is *repelling*: its multiplier is $p'(u_-)p'(u_+) = -1.279$, of modulus above one, so the orbit is expelled at a geometric rate and the sign bookkeeping that assumes the orbit stays near it must eventually fail. It does, at $K = 32$, where $f_{32}'(1/\sqrt{2}) = -0.0455$ while a period-four reading predicts a positive value. The orbit converges instead to an attracting four-cycle $\{0.68192, 1.13427, 0.75292, 1.04691\}$, whose multiplier is -0.1428. Two consequences follow, and the step that carries them is that p' has a *constant sign* on each of the four certified boxes, namely $(-, +, -, -)$, with p' enclosed in $[-1.0204, -1.0203]$, $[1.8275, 1.8278]$, $[-1.4120, -1.4119]$ and $[-0.0547, -0.0532]$ there. Once the orbit is inside the boxes it stays inside them, so from that depth on each lap of (28) contributes the same three negative factors and one positive factor. The multiplier is negative, so a single lap flips the sign and the asymptotic sign pattern has period *eight*, not four. And its modulus is below one, so $|f_K'(1/\sqrt{2})| \to 0$ geometrically, at a factor 0.1428 per four steps, after peaking at $K = 23$ with the value 14.13; by $K = 80$ it is 3×10^{-12}. So the correct statement is the finite-range one recorded in (16), and the assertions that the pattern continues for all K or that the magnitude grows with depth are both false. What the argument does establish is the load-bearing point: the sign is not monotone in depth and is not eventually one sign, so depth is neither uniformly harmful nor uniformly helpful on this task, and a claim that added iterations keep failing is false as stated. The depths used by Theorem 31, $K \in \{1, 5\}$, sit far inside the range where (16) holds and are in any case verified individually.

C.6 Proposition 30

Proof. By the polar form in the proof of Lemma 8, the two singular values of $aE + Z$ are $\frac{\rho}{\sqrt{2}}(1+t)$ and $\frac{\rho}{\sqrt{2}}|1-t|$ with $t = a/\rho > 0$, and the corresponding signed spectral components along the eigenvectors of S are proportional to $1+t$ and $1-t$. The reduced polar factor sends a nonzero singular value to 1 and a zero one to 0. For $0 < t < 1$ both singular values are strictly positive

Table 5: Response profiles on the task of Definition 7, all obtained from Lemma 8, and the local recovery coefficient $\mathcal{R}'(0)$ each implies. The last column is the sign of the frozen-field slope, which is $\mathcal{R}'(0)\,\mathbb{E}R^{-1}$ and therefore has the same sign at every (B, β).

Map	Profile at unit noise radius	$\mathcal{R}'(0)$	Sign
Ψ_1	$t(-4.7145 + 9.671t^2 + 6.2595t^4)/(4(1+t^2)^{5/2})$	$-9429/8000$	negative
Ψ_2	(composition)	-1.3352	negative
Ψ_3	(composition)	$+1.6123$	positive
Ψ_4	(composition)	$+1.6063$	positive
Ψ_5	(composition)	$\in (-1.881, -1.880)$	negative
polar	0 on $t < 1$; $1/\sqrt{2}$ at $t = 1$; $\sqrt{2}$ on $t > 1$	0	dead zone
$P/\|P\|_F$	$t/\sqrt{1+t^2}$	1	positive
$\widetilde{\Psi}_{1/8}$	$t(7 + 34t^2 + 15t^4)/(16(1+t^2)^{5/2})$	$7/16$	positive
Ψ_{unc} at $\lambda = \frac{1}{8}$	$t(1+3t^2)/(4(1+t^2)^{3/2})$	$1/4$	positive
$\Psi_{(\varsigma)}$, $\varsigma > 1$	(138)	$(\varsigma - 1)2^{1/\varsigma - 1/2}$	positive
$\Psi_{(\varsigma)}$, $\varsigma = 1$	polar, as above	0	dead zone

with signs $+$ and $-$, so both map to unit magnitude with opposite signs and their contributions to $\langle E, \mathrm{polar}(aE+Z)\rangle = (\mathrm{sign}(1+t) - \mathrm{sign}(1-t))/\sqrt{2}$ cancel exactly, giving 0. At $t = 1$ the second singular value is zero, so the reduced factor drops it and only the first contributes, giving $1/\sqrt{2}$. For $t > 1$ the second signed component is negative, so $\mathrm{sign}(1-t) = -1$ and the two contributions add to $2/\sqrt{2} = \sqrt{2}$. □

The dead zone is a property of the reduced convention, which is what an implementation that thresholds small singular values performs. On the full-rank interval $0 < t < 1$ it is not a convention at all: the input has two strictly positive singular values, so its polar factor is unique and no completion rule can change the cancellation. Completion choices act only on rank-deficient inputs, which here means $t = 1$. What averaging does change is the mean field.

Proposition 15 (Polar frozen field under a continuous radius)**.** *On the task of Definition 7, for $x > 0$ the frozen field of the reduced polar map is*

$$h_{\mathrm{polar}}(x) = \sqrt{2}\,\mathbb{P}(R < x) + \tfrac{1}{\sqrt{2}}\mathbb{P}(R = x) = \sqrt{2}\,\mathbb{P}(R < x) > 0, \tag{29}$$

the second equality because the stationary radius is continuous. If the two-dimensional stationary density ϱ_∞ is continuous at the origin then $h_{\mathrm{polar}}(x) = \sqrt{2}\pi\varrho_\infty(0)x^2 + o(x^2)$ as $x \downarrow 0$, and the density bound of Lemma 32 gives $h_{\mathrm{polar}}(x) \le \sqrt{2}\pi(x/s)^2$.

Proof. Integrate the three cases of Proposition 30 against the law of R: the response is 0 on $\{R > x\}$, $1/\sqrt{2}$ on $\{R = x\}$ and $\sqrt{2}$ on $\{R < x\}$, since $t = x/R$ crosses one exactly when R crosses x. The coordinates of Z_∞ are symmetric log-concave convolutions of nondegenerate uniforms with the origin interior to their support, so R has a density and $\mathbb{P}(R = x) = 0$, while $\mathbb{P}(R < x) > 0$ for every $x > 0$. Continuity of ϱ_∞ at the origin gives $\mathbb{P}(R < x) = \pi\varrho_\infty(0)x^2 + o(x^2)$, and integrating $\|\varrho_\infty\|_\infty \le s^{-2}$ over a disc of radius x gives the upper bound. □

Thus the per-input dead zone disappears after averaging over this continuous radius, but the restoring field is only quadratic near zero. This result alone provides neither a convergence rate nor a nonzero residence band for an actual EMA trajectory with exact polar.

D When an adverse direction is activated by a task

The structural theorem quantifies over tasks and noise laws. A fixed training task has a particular gradient, and the current momentum contains history. This section gives exact conditions connecting these objects, and an unbiased estimator of the local task response. All inner products in this section are Frobenius inner products. A Jacobian of a matrix map is a linear operator on the full matrix space, including changes of singular vectors.

D.1 Compatibility of the spectral response and its normalizer

The rigidity principle identifies a mismatch between the normalizer and the spectral response; it does not preclude every nonlinear response under every normalization.

Theorem 16 (Schatten normalization fixes the response)**.** *Let $r = \min(m, n) \geq 3$, $1 < q < \infty$, and $f \in C([0, 1]) \cap C^1((0, 1))$ with $f(0) = 0$. Define*

$$\Psi_{f,q}(P) = U \operatorname{diag}[f(\sigma_i(P)/\|P\|_q)]V^\top \quad (P \neq 0), \qquad \Psi_{f,q}(0) = 0.$$

Then $\Psi_{f,q}$ is monotone if and only if $f(s) = cs^{q-1}$ for $c \geq 0$. The admissible map is $c\nabla\|P\|_q$ away from zero.

Proof. Restrict a monotone map to three positive diagonal coordinates, and normalize them to $\sum_i s_i^q = 1$. Let $w_i = s_i^{q-1}$, so $w^\top s = 1$, and let $D = \operatorname{diag}(f'(s_i))$. Up to the positive norm factor its Jacobian is $A = D(I - sw^\top)$ and $As = 0$. Monotonicity requires $S = (A + A^\top)/2 \succeq 0$. Since $s^\top Ss = 0$, positive semidefiniteness implies $Ss = 0$, hence $A^\top s = 0$. Therefore

$$Ds = w(s^\top Ds), \qquad \frac{f'(s_i)}{s_i^{q-2}} = s^\top Ds \quad (i = 1, 2, 3).$$

For arbitrary $u, v \in (0, 1)$, choose $t > 0$ sufficiently small that $u^q + t^q < 1$ and $v^q + t^q < 1$. Applying the displayed equality to $(u, t, (1 - u^q - t^q)^{1/q})$ and $(v, t, (1 - v^q - t^q)^{1/q})$ shows that $f'(s)/s^{q-2}$ is constant on $(0, 1)$. Integrating gives $f(s) = cs^{q-1} + C$; continuity and $f(0) = 0$ give $C = 0$. The secant against zero gives $c \geq 0$. Zero-padding embeds the restriction in every larger shape. Conversely, the stated map is a selection of the subgradient of the convex norm $c\|P\|_q$, including its zero selection at the origin, and is monotone. □

For $1 < q < 2$, the compatible scalar derivative diverges at zero, which is why only interior C^1 regularity is assumed here. The resulting matrix gradient is continuous at rank deficiency away from the origin but need not be Lipschitz. The theorem does not cover $q = 1$ or $q = \infty$, where the normalizer or canonical spectral response is nonsmooth.

D.2 Radial and tangential components of the normalization obstruction

Let $\mathcal{F}_f(X) = U \operatorname{diag}(f(\sigma_i(X)))V^\top$ be the spectral map before normalization. For the odd polynomials in this paper this is a smooth matrix polynomial. More generally, the following result applies wherever $\mathcal{F}_f$ is continuously differentiable in a neighborhood of X, with the usual continuously differentiable odd extension at zero. At $P \neq 0$, set $\rho = \|P\|_F$, $X = P/\rho$ and $\mathcal{B} = D\mathcal{F}_f(X)$. The operator $\mathcal{B}$ is self-adjoint: $\mathcal{F}_f$ is the gradient of $\sum_i \int_0^{\sigma_i(X)} f(u)\, du$. Let

$$\Pi H = H - \langle X, H\rangle X, \qquad v = \Pi\mathcal{B}X, \qquad T = \Pi\mathcal{B}\Pi\big|_{X^\perp}.$$

For $s = \sigma(X)$, $d_i = f'(s_i)$ and $\bar{d} = \sum_i s_i^2 d_i$, this gives $\|v\|_F^2 = \sum_i s_i^2(d_i - \bar{d})^2 = V_f(s)$.

Theorem 17 (Exact task cone and inertia)**.** *Write an arbitrary task direction as $G = aX + H$, where $a = \langle X, G\rangle$ and $H \perp X$. Then*

$$\rho\langle G, D\Psi_f(P)[G]\rangle = \langle G, \mathcal{B}G\rangle - \langle X, G\rangle\langle G, \mathcal{B}X\rangle \tag{30}$$

$$= \langle H, TH\rangle + a\langle v, H\rangle. \tag{31}$$

The first term of (30) *is the response when the denominator is held at ρ; the second term is exactly the effect of differentiating that denominator. Suppose $T \succ 0$. If $v = 0$, the symmetric Jacobian is positive semidefinite with radial kernel. If $v \neq 0$, it has exactly one negative eigenvalue, and its negative-response set is*

$$\left\| T^{1/2}\left(H + \frac{a}{2}T^{-1}v\right)\right\|_F^2 < \frac{a^2}{4}\langle v, T^{-1}v\rangle. \tag{32}$$

In particular, a purely tangential or purely radial direction cannot have negative response. The condition $T \succ 0$ holds if $f' > 0$ throughout $[0, 1]$; it also holds for the classical cubic compositions at every full-rank input with at least two singular values.

Proof. Differentiating $X = P/\|P\|_F$ gives $D\Psi_f(P) = \rho^{-1}\mathcal{B}\Pi$. Self-adjointness of $\mathcal{B}$ proves (30). Substituting $G = aX + H$ and using $\Pi G = H$ proves (31). In the orthogonal decomposition $\mathrm{span}\{X\} \oplus X^\perp$, the dimensionless symmetric Jacobian is

$$\rho \operatorname{sym} D\Psi_f(P) = \begin{pmatrix} 0 & v^*/2 \\ v/2 & T \end{pmatrix}. \tag{33}$$

Completing the square gives (32). Congruence reduces the block matrix to $\mathrm{diag}(-\langle v, T^{-1}v\rangle/4, T)$, proving the inertia statements. The spectrum of $\mathcal{B}$ consists of $f'(s_i)$, the divided differences $(f(s_i) - f(s_j))/(s_i - s_j)$, $(f(s_i) + f(s_j))/(s_i + s_j)$, and $f(s_i)/s_i$ on rectangular complements, with their continuous limits. All are positive when $f' > 0$. For a classical cubic composition, $f' > 0$ on $[0, 1)$; a full-rank input of rank at least two has every $s_i < 1$, and its divided differences are positive as well. □

If $\mu I \preceq T \preceq MI$, the unique negative eigenvalue of $\rho \operatorname{sym} D\Psi_f(P)$ is $-\nu$, where

$$\nu = \frac{1}{4}\langle v, (T + \nu I)^{-1} v\rangle, \qquad \frac{\sqrt{M^2 + V_f(s)} - M}{2} \le \nu \le \frac{\sqrt{\mu^2 + V_f(s)} - \mu}{2}. \tag{34}$$

Indeed, the eigenvector equations give the scalar identity; its right side is decreasing and positive at zero when $v \neq 0$. The bounds follow by replacing the inverse with $(M+\nu)^{-1}I$ and $(\mu+\nu)^{-1}I$. This strengthens the existence margin in Proposition 12 to an inertia statement and an exact task-dependent cone. The tuned reference quintic has signed scalar derivatives, so T need not be positive: the decomposition still holds, but the one-negative-eigenvalue conclusion need not.

Overlap does not determine the sign. For any self-adjoint S, write $S = S_+ - S_-$ for its positive and negative spectral parts. Then $\langle G, SG\rangle = \langle G, S_+G\rangle - \langle G, S_-G\rangle$. In particular, $\|\Pi_- G\|_F^2/\|G\|_F^2$ must be accompanied by the eigenvalue-weighted energies. A large overlap with a weak negative subspace can be outweighed by a small overlap with strong positive eigenvalues. The diagnostic saves these two weighted energies separately.

D.3 A POSITIVE RESULT FOR RADIAL NOISE DENSITIES

The universal characterization quantifies over every bounded centrally symmetric noise law. Some broad subclasses guarantee alignment even for the nonmonotone reference map.

Proposition 18 (Alignment under radially decreasing noise)**.** *Let Ψ be a bounded measurable spectral map $\Psi(P) = U\,\mathrm{diag}(d_i(\sigma(P)))V^\top$, with $d_i \ge 0$ and the usual equivariance conditions making it well defined. If Z has a density $p(z) = \varphi(\|z\|_F^2)$ with φ nonincreasing, then*

$$\langle g, \mathbb{E}\Psi(g + Z)\rangle \ge 0 \qquad \textit{for every } g. \tag{35}$$

For isotropic Gaussian noise, the inequality is strict at $g \neq 0$ if every d_i is positive at positive full-rank spectra. The statement applies to uniform-ball noise and to the reference polynomial with positive spectral values, despite its negative Jacobian directions. It does not require monotonicity of the spectral response.

Proof. Integrate over the singular values and frames of $P = g + Z$. Outside a Lebesgue-null set, the positive singular values are distinct. The SVD volume factor depends on the singular values alone, and its frame measure is invariant under changing the sign of any one left singular vector. Write $a_i = \langle g, u_i v_i^\top\rangle$ and hold the other frame components fixed. Pairing a frame with the frame obtained by flipping u_i turns the ith contribution to the alignment integral into a positive volume factor times

$$\frac{d_i a_i}{2}\left[\varphi(C - 2\sigma_i a_i) - \varphi(C + 2\sigma_i a_i)\right], \qquad C = \|g\|_F^2 + \sum_j \sigma_j^2 - 2\sum_{j\neq i}\sigma_j a_j.$$

Both arguments are squared distances of actual matrices from g. The bracket has the same sign as a_i, since $\sigma_i > 0$ and φ is nonincreasing. Each component's integral is nonnegative. Boundedness permits the decomposition and pairing under the integral. For a Gaussian density the decrease is strict; when $g \neq 0$, $a_i \neq 0$ on a positive-measure set of frames, which makes the integral strictly positive. □

The frame-pairing argument is specific to the spectral update and the full radial noise density; central symmetry alone does not provide it. Isotropic Gaussian EMA noise remains in this class at every fixed momentum and batch size. The result is about a frozen field; conditioning on a realized momentum history introduces the additional term in Proposition 21.

D.4 Full spherical symmetry removes the zero-signal obstruction

Central symmetry and full rotational symmetry of the noise are different assumptions. The following comparison is in the entire $d = mn$ dimensional matrix space, not just in a selected noise subspace.

Proposition 19 (Positive recovery under full spherical noise)**.** *Let $d = mn > 1$ and let f be an odd polynomial with $f(s) > 0$ for $0 < s \le 1$. For $Z = \rho U$, where $\rho > 0$ is fixed and U is uniform on the unit Frobenius sphere, $h(g) = \mathbb{E}\Psi_f(g + Z)$ satisfies*

$$Dh(0) = \gamma_f I, \qquad \gamma_f = \frac{d-1}{d\rho}\mathbb{E}\langle U, \Psi_f(U)\rangle > 0. \tag{36}$$

For isotropic Gaussian matrix noise $Z \sim N(0, \tau^2 I_d)$, the corresponding identity is

$$Dh(0) = \frac{\mathbb{E}\langle Z, \Psi_f(Z)\rangle}{d\tau^2} I \succ 0. \tag{37}$$

Thus an EMA of independent isotropic Gaussian innovations has positive zero-signal recovery for every fixed $\beta < 1$ and batch size, since its stationary noise is still isotropic Gaussian. This does not assert positive response at every finite signal or conditional descent on an adaptive trajectory.

Proof. The noise laws are invariant under left and right orthogonal transformations, and the map is equivariant under them. Therefore the derivative at zero commutes with these actions. Row and column sign changes force its off-coordinate entries to vanish, and row and column permutations make all diagonal entries equal; it is a scalar multiple of the identity. For spherical noise, differentiating under the expectation is valid near $g = 0$ because Z stays a fixed distance from the origin. Zero homogeneity gives $D\Psi_f(rU) = D\Psi_f(U)/r$. Applying the divergence theorem on an annulus and integrating in polar coordinates gives

$$\mathbb{E}\,\mathrm{tr}\, D\Psi_f(U) = (d-1)\mathbb{E}\langle U, \Psi_f(U)\rangle.$$

Indeed, the common factor $1 - \varepsilon^{d-1}$ from an annulus of inner radius ε cancels between its volume integral and its two boundary fluxes. Taking the trace of the scalar derivative proves (36). For Gaussian noise, differentiation of its smooth density, justified by boundedness of Ψ_f, gives $Dh(0)[H] = \tau^{-2}\mathbb{E}[\Psi_f(Z)\langle Z, H\rangle]$. Taking its trace proves (37). Finally, $\langle P, \Psi_f(P)\rangle = \sum_i \sigma_i(P) f(\sigma_i(P)/\|P\|_F) > 0$ at every nonzero P, so both scalars are strictly positive. □

The spherical law is bounded, so it lies within the noise class of the structural characterization. The Gaussian comparison is unbounded but has finite required moments. For the two rank-three cubics at unit radius, the exact spherical coefficients are $64/99$ for $f_+(s) = (s + 3s^3)/4$ and $104/99$ for $f(s) = (3s - s^3)/2$, even though their specified single-input adverse responses are $-25/204$ and $-1/34$. To evaluate these coefficients, for U uniform on the matrix sphere use $\mathbb{E}\,\mathrm{tr}((UU^\top)^2) = (m + n + 1)/(mn + 2)$, obtained from the sphere's fourth moments $\mathbb{E}U_a^4 = 3/[d(d+2)]$ and $\mathbb{E}U_a^2 U_b^2 = 1/[d(d+2)]$ for $a \neq b$. The reference quintic also has $f_K(s) > 0$ for positive s, by its certified positive polynomial factor, so its zero-signal field is restoring under these full spherical laws at every depth. The constructed transverse and one-dimensional noise laws break this full-space symmetry. No such symmetry assumption is made about transformer gradient noise.

D.5 Unbiased estimation from independent gradients

The distinction matters statistically as well. Let P and the checkpoint be fixed, and let $G_1, \ldots, G_n$ be independent identically distributed estimates of the gradient g of one specified evaluation objective, with covariance operator Σ and finite second moment. The batches used for these estimates are independent of the batch forming P. Write $S = \mathrm{sym}\, D\Psi(P)$ and $\bar{G} = n^{-1}\sum_i G_i$.

Proposition 20 (Gradient covariance bias and its removal)**.** *For $n \geq 2$,*

$$\mathbb{E}\langle \bar{G}, S\bar{G}\rangle = \langle g, Sg\rangle + \frac{1}{n}\operatorname{tr}(S\Sigma), \tag{38}$$

$$\widehat{q}_U = \frac{1}{n(n-1)}\sum_{i\neq j}\langle G_i, SG_j\rangle, \qquad \mathbb{E}\widehat{q}_U = \langle g, Sg\rangle, \tag{39}$$

$$\operatorname{Var}(\widehat{q}_U) = \frac{4}{n}\langle Sg, \Sigma Sg\rangle + \frac{2}{n(n-1)}\operatorname{tr}(S\Sigma S\Sigma). \tag{40}$$

The exact sample identity

$$\langle \bar{G}, S\bar{G}\rangle - \widehat{q}_U = \frac{1}{n(n-1)}\sum_{i}\langle G_i - \bar{G}, S(G_i - \bar{G})\rangle \tag{41}$$

holds for indefinite S as well. Each quadratic component of (30)–(31) *and each spectral energy has the same unbiased cross-estimator. A quotient by an estimated gradient norm is not thereby unbiased; the signed numerator is the primary estimand.*

Proof. Put $\epsilon_i = G_i - g$. Expanding the quadratic form and using independence proves (38) and (39). The centered U-statistic equals $2n^{-1}\sum_i\langle Sg, \epsilon_i\rangle + [n(n-1)]^{-1}\sum_{i\neq j}\langle \epsilon_i, S\epsilon_j\rangle$. The two sums are uncorrelated. Distinct terms of the second sum are uncorrelated unless they correspond to the same unordered pair. Their variance is $\operatorname{tr}(S\Sigma S\Sigma) = \|\Sigma^{1/2}S\Sigma^{1/2}\|_F^2$ per pair, giving (40). Expanding $\sum_i\langle G_i - \bar{G}, S(G_i - \bar{G})\rangle$ proves (41). These arguments condition on the entire checkpoint, so randomness in its training history creates no bias in the conditional identities. □

This is the usual order-two U-statistic construction (Hoeffding, 1948), applied to a signed operator. It removes covariance bias, whose sign itself need not be positive. It does not certify the sign at every individual layer, remove a difference between training and evaluation objectives, or turn correlated checkpoints into independent samples. Our uncertainty intervals use complete training seeds. Negative point estimates of a positive spectral energy are possible under cross-estimation and are not truncated.

D.6 From a local response to an actual momentum step

An actual conditional update contains an additional term even under symmetric innovations. Fix the history before step t, so W_t, $g_t = \nabla F(W_t)$ and the previous momentum are fixed. Write its map input as $p + cg_t + Z$, where p is the history contribution, $c > 0$ is the current-signal weight, and Z is the current noise contribution. For plain EMA, $p = \beta M_{t-1}$, $c = 1 - \beta$, $Z = c\bar{\xi}_t$; for the Nesterov weighting used in the neural implementation, these are $p = \beta^2 M_{t-1}$, $c = 1 - \beta^2$, $Z = c\bar{\xi}_t$ in real arithmetic.

Proposition 21 (Conditional alignment and finite-step loss)**.** *Let $U = \Psi(p + cg_t + Z)$, and assume the following expectations exist. Then*

$$\begin{aligned}\mathcal{A}_t := \mathbb{E}[\langle g_t, U\rangle \mid \mathcal{F}_{t-1}] &= \mathcal{B}_t + \mathcal{D}_t, \\ \mathcal{B}_t &= \mathbb{E}[\langle g_t, \Psi(p+Z)\rangle \mid \mathcal{F}_{t-1}], \\ \mathcal{D}_t &= \mathbb{E}[\langle g_t, \Psi(p+cg_t+Z) - \Psi(p+Z)\rangle \mid \mathcal{F}_{t-1}].\end{aligned} \tag{42}$$

If the intervening segments stay in a differentiability region and the derivatives are integrably dominated, then

$$\mathcal{D}_t = c\mathbb{E}\left[\int_0^1 \langle g_t, D\Psi(p + Z + ucg_t)[g_t]\rangle\, du \,\middle|\, \mathcal{F}_{t-1}\right]. \tag{43}$$

If F has an L_F-Lipschitz gradient along the update segments and $V_t = \mathbb{E}[\|U\|_F^2 \mid \mathcal{F}_{t-1}]$, then

$$|\mathbb{E}[F(W_t - \eta U) - F(W_t) \mid \mathcal{F}_{t-1}] + \eta\mathcal{A}_t| \leq \frac{L_F\eta^2}{2}V_t. \tag{44}$$

Thus $\mathcal{A}_t > L_F\eta V_t/2$ suffices for expected descent, and $\mathcal{A}_t < -L_F\eta V_t/2$ suffices for expected ascent. For convex F, $\mathcal{A}_t < 0$ alone implies expected ascent for every $\eta > 0$.

Proof. Add and subtract $\Psi(p+Z)$ for (42); integrate its Jacobian along each segment for (43). The two-sided smoothness inequality proves (44). The convex lower supporting-plane inequality gives the final assertion. □

When $p = 0$, oddness and central symmetry imply $\mathcal{B}_t = 0$. In general, conditioning on the momentum history leaves $p \neq 0$, so this cancellation is unavailable. Neither a negative Jacobian direction nor a negative local task quadratic form by itself determines $\mathcal{A}_t$. Conversely, positive response in the neural audit is evidence against activation of that local mechanism on the measured task directions; it does not prove monotonicity of the map or all-time descent of training. The finite-difference identity (42) remains valid when a segment crosses the origin; only the derivative-integral representation then requires care.

D.7 A FINITE-RESPONSE GUARANTEE FOR APPROXIMATE CONVEX MAPS

Regularity has a concrete use even without a loss advantage: it gives a positive margin against perturbations of the input and against forward error in a computed secant.

Corollary 22 (Certified response with forward error)**.** *For the smoothed nuclear map, put* $\mu_\kappa = \kappa\sqrt{r}/\sqrt{1+\kappa}$ *and* $L_\kappa = \sqrt{r}(1+\kappa)/\sqrt{\kappa}$*. Let* $P \neq 0$*,* $\rho = \|P\|_F$*,* $\|E\|_F = 1$*,* $a = \langle E, P/\rho\rangle$ *and* $0 < h < \rho$*. If the two computed endpoints obey* $\|\widehat{\Psi}(P \pm hE) - \Psi_\kappa^{\mathrm{sn}}(P \pm hE)\|_F \le e_\pm$*, then*

$$\frac{\langle E, \widehat{\Psi}(P+hE) - \widehat{\Psi}(P-hE)\rangle}{2h} \ge \frac{\mu_\kappa \rho^2(1-a^2)}{(\rho+h)^3} - \frac{e_+ + e_-}{2h}. \tag{45}$$

Also, for $\|\Delta\|_F < \rho$*,*

$$\|\Psi_\kappa^{\mathrm{sn}}(P+\Delta) - \Psi_\kappa^{\mathrm{sn}}(P)\|_F \le \frac{L_\kappa \|\Delta\|_F}{\rho - \|\Delta\|_F}. \tag{46}$$

Both bounds include rank-deficient nonzero inputs.

Proof. The angular Hessian bound of Theorem 58 and the radial kernel give $\langle E, D\Psi_\kappa^{\mathrm{sn}}(Q)[E]\rangle \ge \mu_\kappa(\|Q\|_F^2 - \langle E, Q\rangle^2)/\|Q\|_F^3$. For $Q = P + tE$, the numerator is exactly $\rho^2(1-a^2)$ and the denominator is at most $(\rho+h)^3$. Integrating over $[-h, h]$ and bounding the endpoint errors proves (45). Integrating the Jacobian norm bound $L_\kappa/\|Q\|_F$ along $P + t\Delta$ proves (46). □

If $e_+ = e_- = e$, the sufficient error tolerance in (45) is maximized at $h = \rho/2$, where $e < (4\mu_\kappa/27)(1-a^2)$ certifies a positive response. For a radial direction $a^2 = 1$, zero homogeneity precludes a strictly positive margin. This is a finite-input response certificate, not a nonlinear training-loss bound or a backward-error theorem for a Gram eigensolver. The errors $e_\pm$ still require independent control. Schatten maps below exponent two lack a uniform input-Lipschitz bound at rank deficiency; their exact convex alignment remains valid nonetheless.

E THE FROZEN RESIDUAL ON THE SYMMETRIC CLASS

Here we hold the gradient fixed and ask how large it can be while the average update still has zero projection along it. The metric below measures the largest such squared gradient over the allowed noise distributions. We bound this quantity in terms of the noise weights. It describes the average field and does not track an optimizer trajectory.

Proposition 23 (Frozen residual on the symmetric class)**.** *Let* $\mathcal{P}$ *be all centrally symmetric laws in the unit Frobenius ball, discrete, degenerate and nondegenerate alike, and let* $\mathcal{E}_f(w)$ *be as in* (47) *below. If* $0 < a_f := \min_s f(s)/s \le b_f := \max_s f(s)/s < \infty$ *then* $\mathcal{E}_f(w) \le (b_f/a_f)^2 v$*, with no dimensional factor and no global derivative bound. If in addition* $r \ge 3$ *and* f *is a nonlinear odd polynomial, then along any weight family with* $v \to 0$ *and* $\ell \to 0$ *there is* $c_f > 0$ *with* $c_f v \le \mathcal{E}_f(w) \le (b_f/a_f)^2 v$ *for all large indices, the lower bound realized by a single fixed bounded full-dimensional centrally symmetric law and a single fixed matrix direction.*

E.1 What the metric is, and what it is not

For a nonnegative weight family w with $\sum_i w_i = 1$, let $\mathcal{P}$ be the set of *all* centrally symmetric laws supported in the unit Frobenius ball, discrete, degenerate and nondegenerate alike, and put

$$\mathcal{E}_f(w) = \sup_{\nu\in\mathcal{P},\ \|E\|_F=1} \sup\big\{x^2 :\ \mathbb{E}_\nu\langle E, \Psi_f(xE + Z_w)\rangle = 0\big\}. \tag{47}$$

This is the worst *frozen* projected stationary squared gradient of the convex ridge objective $F(W) = \phi(\langle E, W\rangle)$: the largest signal amplitude at which the averaged update has no component along the signal direction. It takes no infimum over step sizes and refers to no trajectory. Consequently (47) is not a minimax risk for the actual EMA at arbitrary schedules, and no bound on it can be rewritten as one. Its role in the paper is to show that the mechanism is not tied to one noise instance, which is why the hard law below is a single fixed full-dimensional one rather than a family that moves with the memory.

E.2 The derivative-free upper bound

Lemma 24 (Upper bound without derivatives)**.** *Suppose $0 < a_f := \min_{0\le s\le 1} f(s)/s \le b_f := \max_{0\le s\le 1} f(s)/s < \infty$, using $f'(0)$ at $s = 0$. Then $\langle P, \Psi_f(P)\rangle \ge a_f\|P\|_F$ and $\|\Psi_f(P)\|_F \le b_f$ for every P, and $\mathcal{E}_f(w) \le (b_f/a_f)^2 v$.*

Proof. With $\varsigma_i = \sigma_i(P)/\|P\|_F$, so $\sum_i \varsigma_i^2 = 1$,

$$\langle P, \Psi_f(P)\rangle = \|P\|_F \sum_i \varsigma_i f(\varsigma_i) = \|P\|_F \sum_i \varsigma_i^2 \frac{f(\varsigma_i)}{\varsigma_i} \ge a_f\|P\|_F,$$

and $\|\Psi_f(P)\|_F^2 = \sum_i f(\varsigma_i)^2 \le b_f^2 \sum_i \varsigma_i^2 = b_f^2$. If x realizes a zero in (47) then, taking the inner product with xE and using $\mathbb{E}\|Z_w\|_F \le (\mathbb{E}\|Z_w\|_F^2)^{1/2} \le \sqrt{v}$ together with $\mathbb{E}\|xE + Z_w\|_F \ge \|xE\|_F = |x|$ by Jensen,

$$0 = \mathbb{E}\langle xE, \Psi_f(xE + Z_w)\rangle \ge a_f\mathbb{E}\|xE + Z_w\|_F - b_f\mathbb{E}\|Z_w\|_F \ge a_f|x| - b_f\sqrt{v},$$

the middle step splitting $xE = (xE + Z_w) - Z_w$ and applying the two displayed bounds. The second moment obeys $\mathbb{E}\|Z_w\|_F^2 = v\,\mathbb{E}\|\xi\|_F^2 \le v$ rather than equality, since a law in $\mathcal{P}$ need only be supported in the unit ball and may have $\mathbb{E}\|\xi\|_F^2 < 1$; the inequality is the direction the bound needs. Hence $|x| \le (b_f/a_f)\sqrt{v}$. □

There is no dimensional factor and no global derivative bound, and central symmetry was not used, so the same conclusion holds for any zero-mean noise.

E.3 A truncation and Lindeberg transfer

The lower bound needs to move from a Gaussian calculation to the actual finite-memory law, and the difficulty is that Ψ_f is discontinuous at the origin, so no global derivative bound exists. The device is to differentiate an artificially smoothed test function instead.

Lemma 25 (Distributional transfer)**.** *Regard matrices as Euclidean vectors, let $\Sigma = \mathrm{Cov}(\xi)$ with $q = \operatorname{rank}\Sigma \ge 1$, let $G \sim N(0, \Sigma)$, and put $Y_w = Z_w/\sqrt{v}$. For any bounded zero-homogeneous Ψ that is C^3 away from the origin there are constants C_1, C_2 independent of w, B, β and T with*

$$\sup_g \big\|\mathbb{E}\Psi(g + Y_w) - \mathbb{E}\Psi(g + G)\big\| \le C_1\varrho^q + C_2\ell\varrho^{-3} \qquad (\varrho > 0), \tag{48}$$

and hence, at $\varrho = \ell^{1/(q+3)}$, $\sup_g \|h_w(\sqrt{v}\,g) - H_\Sigma(g)\| \le (C_1 + C_2)\ell^{q/(q+3)}$ where $H_\Sigma(g) = \mathbb{E}\Psi(g + G)$. No density is assumed for ξ, and q may be smaller than the vectorized dimension.

Proof. Fix a C^3 cutoff χ that vanishes on $[0, 1]$ and equals one on $[2, \infty)$, for instance the degree-seven smoothstep $\chi(\varsigma) = 35\theta^4 - 84\theta^5 + 70\theta^6 - 20\theta^7$ with $\theta = \varsigma - 1$ on $1 < \varsigma < 2$. Put

$$A_3 = \big\|D^3[\chi(\|\cdot\|)\Psi(\cdot)]\big\|_\infty, \quad B_3 = \big\|D^3[1 - \chi(\|\cdot\|/2)]\big\|_\infty, \quad M_3 = \mathbb{E}\|\xi\|^3 + \mathbb{E}\|G\|^3,$$

and $D_\Sigma = \mathrm{vol}(\mathbb{B}_q)/[(2\pi)^{q/2}\sqrt{\mathrm{pdet}\,\Sigma}]$, all finite; A_3 is determined by the fixed spectral polynomial and the fixed cutoff and is controlled by the coefficients alone. We take $C_1 = \|\Psi\|_\infty D_\Sigma(4^q + 2^q)$ and $C_2 = (M_3/6)(A_3 + \|\Psi\|_\infty B_3)$.

Let $\Psi_\varrho(P) = \chi(\|P\|/\varrho)\Psi(P)$. Zero-homogeneity of Ψ gives $\|D^3\Psi_\varrho\|_\infty = A_3\varrho^{-3}$, since scaling the argument by ϱ scales the third derivative by ϱ^{-3} and Ψ itself is scale invariant. Now replace the summands $w_i\xi_i/\sqrt{v}$ one at a time by independent Gaussians of matching covariance. In each replacement the first and second Taylor terms of Ψ_ϱ cancel because the means and covariances agree, so only the third-order remainder survives, and summing over i gives a total of at most $A_3M_3\ell/(6\varrho^3)$ where $\ell = \sum_i w_i^3/v^{3/2}$ is exactly the normalized third-moment weight.

It remains to charge the difference between Ψ and Ψ_ϱ, which is supported in the ball of radius 2ϱ and bounded by $\|\Psi\|_\infty$. Bound the corresponding probability for Y_w by applying the same Lindeberg replacement to a smooth upper bound for the indicator of that ball, supported in radius 4ϱ, whose third derivative is $B_3\varrho^{-3}$; the Gaussian's density on its q-dimensional support is at most $[(2\pi)^q\,\mathrm{pdet}\,\Sigma]^{-1/2}$, so the Gaussian probability of the radius-4ϱ ball is at most $D_\Sigma(4\varrho)^q$, and its own radius-2ϱ ball probability is at most $D_\Sigma(2\varrho)^q$. Adding the three contributions gives (48).

For an infinite weight family, carry out the finite replacement against the smooth truncated test functions and the smooth ball upper bound first, pass to the limit in L^2, and only then compare with the original map. Doing it in that order avoids exchanging a limit with a map that is discontinuous at the origin. □

What is differentiated is the smoothed test function, never Ψ itself, which is why no global derivative bound is needed. The constants may degrade as the fixed spectral map or the covariance conditioning worsens, but none of them varies with the memory.

E.4 The matching lower bound

Proposition 26 (Lower bound at a single fixed law)**.** *Let $r \geq 3$ and let f be a nonlinear odd polynomial with $0 < a_f \leq b_f < \infty$. For any nonnegative weight family with $v_n \to 0$ and $\ell_n \to 0$ there is $c_f > 0$ such that $c_f v_n \leq \mathcal{E}_f(w^{(n)})$ for all large n, and the lower bound is realized by one fixed bounded full-dimensional centrally symmetric law and one fixed matrix direction. Moreover the scalar averaged gradient flow has a forward-invariant interval away from the origin.*

Proof. By Theorem 2 and Proposition 12 there are unit matrices P, E with $\varkappa := \langle E, D\Psi_f(P)[E]\rangle < 0$, and E is not parallel to P, since $D\Psi_f(P)[P] = 0$ by Lemma 9. Let $S \sim N(0,1)$ and set $J(x) = \mathbb{E}\langle E, \Psi_f(xE + SP)\rangle$. Split the expectation over $|S| \leq Cx$, $Cx < |S| < 1$ and $|S| \geq 1$ for a large fixed C. On the middle range, $xE + SP$ is close to the unit matrix $\mathrm{sgn}(S)P$ after scaling, so zero-homogeneity lets us write $\Psi_f(xE + SP) = \Psi_f(\mathrm{sgn}(S)P + (x/|S|)E)$ and Taylor expand in $x/|S|$ about $\pm P$. Oddness makes the zeroth-order terms cancel between $\pm S$, the first-order term contributes $\varkappa x/|S|$ from each side, and the second-order remainder is $O(x^2/S^2)$. Since $\int_{Cx}^1 \varphi(u)/u\,du = \varphi(0)\log(1/x) + O(1)$ and $\int_{Cx}^1 x^2/u^2\,du = O(x)$,

$$J(x) = 2\varphi(0)\varkappa\, x\log(1/x) + O(x), \qquad x \downarrow 0,$$

with φ the standard normal density; the inner range contributes $O(x)$ because $\|\Psi_f\|_F \leq b_f$ there and the range has length $O(x)$, and the outer range is $O(x)$ by the same Taylor argument. As $\varkappa < 0$, this gives $J(x) < 0$ for small positive x. At the other end, zero-homogeneity gives $\lim_{x\to\infty} J(x) = \langle E, \Psi_f(E)\rangle \geq a_f > 0$. So there are $0 < x_- < x_+$ and $m > 0$ with $J(x_-) < -2m$ and $J(x_+) > 2m$.

Now perturb SP by a small independent full-dimensional centered Gaussian; at the two fixed nonzero test points both signs survive by dominated convergence, since Ψ_f is bounded. Realize the resulting covariance by a fixed bounded centrally symmetric law, for instance $\xi = (\epsilon P + \varsigma U)/(1 + \varsigma)$ with ϵ Rademacher, U independent and uniform on the unit ball of the whole matrix space, and $\varsigma > 0$ fixed and small, rescaling $x_\pm$ if needed. That law has norm at most one, positive definite covariance, and a full-dimensional density, which need not be smooth everywhere. For its covariance Σ choose the margin m so that $\langle E, H_\Sigma(x_-E)\rangle \leq -m$ and $\langle E, H_\Sigma(x_+E)\rangle \geq m$.

Lemma 25 now supplies a finite-memory certificate: whenever

$$(C_1 + C_2)\ell^{q/(q+3)} < m/2, \tag{49}$$

the actual frozen field is negative at $x_-\sqrt{v}$ and positive at $x_+\sqrt{v}$. Since the chosen law has a density, the scalar field $x \mapsto \mathbb{E}\langle E, \Psi_f(xE + Z_w)\rangle$ is continuous, so it has a zero in $(x_-\sqrt{v}, x_+\sqrt{v})$ and $\mathcal{E}_f(w) \geq x_-^2 v$. Because $\ell_n \to 0$, (49) holds for all large n, with $c_f = x_-^2$.

For the last claim, the scalar averaged gradient flow $\dot{x} = -\mathbb{E}\langle E, \Psi_f(xE + Z_w)\rangle$ is positive at the left endpoint and negative at the right endpoint of that interval, so the interval is forward invariant and its limit points form a nonzero stationary set. This is why the conclusion is not merely an algebraic root that would be repelling. For the convex Huber ridge the interval lies inside the quadratic region once v is small enough. □

E.5 Consequences for EMA, and the fence around them

For B independent samples per step and stationary EMA weights $w_{j,b} = (1-\beta)\beta^j/B$,

$$v = \frac{1-\beta}{B(1+\beta)}, \qquad u = \frac{(1-\beta)^2}{B^2(1+\beta+\beta^2)}, \qquad \ell = \sqrt{\frac{1-\beta}{B}}\,\frac{(1+\beta)^{3/2}}{1+\beta+\beta^2}, \tag{50}$$

by summing the geometric series $\sum_j \beta^{2j}$ and $\sum_j \beta^{3j}$. Both $B \to \infty$ and $\beta \uparrow 1$ send $v \to 0$ and $\ell \to 0$, so both land in the regime of Proposition 26. Writing $N_{\mathrm{var}} = 1/v$, the worst frozen residual squared gradient is $\Theta(N_{\mathrm{var}}^{-1})$ and *not* $\Theta(N_{\mathrm{var}}^{-2})$.

The correct reading is narrow. This does not say that every fixed task exhibits this error. It says that without separately controlling recovery geometry, the N^{-2} rule familiar from skewed scalar noise cannot be promoted to a general matrix mean-field rule, and the hard law used here has every odd tensor equal to zero, so no skewness explanation is available for it.

As a checkable illustration rather than a step of the proof, for two-dimensional circular noise the Gaussian limit gives $H_K(x) = \int_0^\infty q_K(x/\varrho)\, 2\varrho e^{-\varrho^2} d\varrho$, and at $K = 1$ one-dimensional quadrature puts the positive zero near 0.345, with square near 0.119, so that branch predicts a frozen residual near $0.12\,\sigma^2 v$. Those decimals are a numerical illustration and Proposition 26 does not depend on them.

Remark 27 (Relation to Section 5, and instance discipline)**.** This proposition is not a prerequisite for Theorem 31. That theorem handles finite memory directly through the uniform small-ball bound of Lemma 32 and needs no Gaussian limit, so the two are complementary evidence rather than one chain of implications. Three different noise instances appear in this paper: the circular or four-point law used in Section 3 for the pathwise cancellation, the fixed full-dimensional law used here, and the two-dimensional uniform law of Definition 7. They are distinct, so no dynamic comparison may be assembled across them: any statement comparing two algorithms must be made inside one instance, which is why every trajectory comparison in this paper uses Definition 7.

F From normalization coupling to an actual rank-three trajectory

This appendix isolates the effect of dividing by the input matrix's own norm. The scalar function is increasing everywhere, so a negative scalar slope cannot explain the failure. We first derive an exact scalar recursion for the chosen matrix task, then bound its expected update and random fluctuations. These bounds show that the actual EMA, starting from zero momentum, keeps a nonzero gradient over the stated million-update window.

F.1 An exact positive-slope witness and scalar recursion

Set

$$f_+(s) = \frac{s + 3s^3}{4}, \qquad S_3 = \frac{\mathrm{diag}(4,1,1)}{\sqrt{18}}, \qquad E_3 = \frac{\mathrm{diag}(3,2,2)}{\sqrt{17}}. \tag{51}$$

Both matrices have unit Frobenius norm and full rank. The scalar slope satisfies $f'_+(s) = (1 + 9s^2)/4 \geq 1/4$ on $[0, 1]$. In the common singular frame, the diagonal derivative matrix is $D = \mathrm{diag}(9/4, 3/8, 3/8)$. Consequently

$$\langle E_3, D\Psi_{f_+}(S_3)[E_3]\rangle = \frac{93}{68} - \frac{76}{51} = -\frac{25}{204} < 0. \tag{52}$$

The negative term is precisely the derivative of the endogenous normalization. All the off-diagonal blocks of Proposition 10 are positive because f_+ is strictly increasing. This is therefore an instance of the slope-heterogeneity obstruction in Theorem 2.

Take $F(W) = \frac{1}{2}\langle E_3, W\rangle^2$, $W_0 = x_0 E_3$, $x_0 > 0$, batch size one, and

$$\xi_t = \sigma\epsilon_t S_3 + R_t, \qquad \mathbb{P}(\epsilon_t = 1) = \mathbb{P}(\epsilon_t = -1) = \tfrac{1}{2}, \qquad \|R_t\|_F \leq \frac{\sigma}{8000}. \tag{53}$$

The signs are iid and independent of the preceding history. The perturbation R_t can be nonanticipating, including dependence on the current sign. In particular, independent centrally symmetric perturbations preserve a fixed symmetric additive noise law. A smooth compactly supported density in the full nine-dimensional matrix ball makes the total law smooth and full dimensional.

Use $\Psi = \Psi_{f_+}$ and $\beta = 1/10$ in (2). Put $k = \sqrt{17/18}$, $u_t = \langle E_3, W_t\rangle/(\sigma k)$ and $\alpha = \eta/(\sigma k)$. The exact momentum decomposition is

$$M_t = \sigma k\, b_t E_3 + \sigma z_t S_3 + \Delta_t, \quad b_t = \beta b_{t-1} + (1-\beta)u_t, \quad z_t = \beta z_{t-1} + (1-\beta)\epsilon_t,$$

where $b_{-1} = z_{-1} = 0$, $\Delta_t = \beta\Delta_{t-1} + (1-\beta)R_t$ and $\|\Delta_t\|_F \leq \sigma/8000$. For $R_t = 0$, diagonal evaluation gives

$$u_{t+1} = u_t - \alpha q(b_t, z_t), \qquad q(b, z) = \frac{628b^3 + 1932b^2 z + 2186bz^2 + 876z^3}{4\sqrt{17}(17b^2 + 32bz + 18z^2)^{3/2}}. \tag{54}$$

For $R_t \neq 0$ the same scalar recursion holds with $q(b_t, z_t)$ replaced by $Y_t = \langle E_3, \Psi_{f_+}(M_t)\rangle$.

F.2 Conditional barriers, including the unequal EMA radii

The current sign dominates the preceding noise EMA: $\mathrm{sign}(z_t) = \epsilon_t$ and $4/5 \leq |z_t| \leq 1$. The two conditional radii are generally different. Accordingly, define the conditional field, for $|v| \leq 1$, by

$$G(b, v) = \tfrac{1}{2}\left[q\big(b, \tfrac{9}{10} + \tfrac{v}{10}\big) + q\big(b, -\tfrac{9}{10} + \tfrac{v}{10}\big)\right]. \tag{55}$$

This is the conditional expectation given the actual preceding noise EMA $v = z_{t-1}$. It does not replace that EMA by an independent stationary draw.

Proposition 28 (Uniform conditional barriers)**.** *The following bounds hold simultaneously for every $|v| \leq 1$:*

$$G(b, v) \leq -\frac{21}{1000}, \qquad \frac{1}{5} \leq b \leq \frac{19}{40}, \tag{56}$$

$$G(b, v) > \frac{178}{1000}, \qquad b \geq \frac{9}{10}. \tag{57}$$

Moreover, $\|\Psi_{f_+}(P)\|_F \leq 1$ for every P, and the perturbation in (53) *changes a projected update by at most $1/1000$. Thus, if $|b_t - u_t| \leq 1/40$, the conditional mean of the actual update satisfies*

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] \leq -\gamma_- \quad \text{on } [a, 2a], \qquad \mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] \geq \gamma_+ \quad \text{on } [b, \infty), \tag{58}$$

where the intervals refer to u_t and

$$a = \frac{9}{40}, \qquad b = \frac{19}{20}, \qquad \gamma_- = \frac{1}{50}, \qquad \gamma_+ = \frac{17}{100}.$$

Proof. The inner bound is an exact rational interval certificate on $[1/5, 19/40] \times [-1, 1]$. We cover this rectangle by 43 rational boxes. On each box, we evaluate G at the rational centre with enclosed square roots, and use interval enclosures of both partial derivatives for the mean-value remainder. Every resulting upper endpoint is at most $-21/1000$. Square-root endpoints are rational numbers

with denominator 2^{96}; their enclosing property is checked by exact squaring. All subsequent interval operations use rational arithmetic. The largest certified upper endpoint is approximately -0.02137074; the claimed rational margin is weaker.

For completeness, the formulas used by this short certificate are

$$\partial_b q(b,z) = -\frac{25z(27b^3+95b^2z+101bz^2+27z^3)}{\sqrt{17}(17b^2+32bz+18z^2)^{5/2}}, \qquad \partial_z q(b,z) = -\frac{b}{z}\partial_b q(b,z).$$

The positive form $17b^2+32bz+18z^2 = 17(b+16z/17)^2+(50/17)z^2$ provides a strictly positive interval denominator throughout the rectangle.

The outer bound also has a short global verification. For $z>0$, $q(b,z)$ decreases in $b \geq 0$ to $157/289$. For $z<0$, zero homogeneity gives $q(b,z) = q(b/|z|,-1)$. The cubic $27w^3-95w^2+101w-27$ has no root in $[9/10,\infty)$ and is positive at $9/10$; Sturm's rule checks this exactly. Hence $q(w,-1)$ increases there. Since $|z| \leq 1$,

$$G(b,v) \geq \tfrac{1}{2}\left[\frac{157}{289} + q(9/10,-1)\right] = 0.1785858542\ldots > \frac{178}{1000}.$$

The same calculation gives $q(b,z) > 0.46$ for $b \geq 2$ and $4/5 \leq |z| \leq 1$.

For a normalized singular value $s \in [0,1]$, $0 \leq f_+(s) \leq s$, so the full update norm is at most one. The unnormalized spectral polynomial has derivative norm at most $5/2$ on the unit Frobenius sphere: its divided differences and its scalar derivatives are all between $1/4$ and $5/2$. Differentiating normalization therefore gives $\|D\Psi_{f_+}(P)\|_{\mathrm{op}} \leq 5/(2\|P\|_F)$. Since

$$\langle E_3, S_3\rangle = \frac{16}{\sqrt{306}}, \qquad \|\sigma k b E_3 + \sigma z S_3\|_F \geq \sigma\frac{4}{5}\frac{5}{\sqrt{153}},$$

every segment from this matrix to its perturbation by Δ_t has norm greater than $5\sigma/16$. Its projected response changes by at most $(8/\sigma)\|\Delta_t\|_F \leq 1/1000$. This estimate is uniform in b. Finally, $u_t \in [a,2a]$ and $|b_t-u_t| \leq 1/40$ put $b_t \in [1/5, 19/40]$, while $u_t \geq b$ gives $b_t \geq 37/40 > 9/10$. Subtracting the perturbation budget proves (58). □

F.3 An explicit finite-horizon theorem for the zero-initialized EMA

Theorem 29 (Normalization-only nonlinear residence)**.** *Use* (51)–(53)*,* $F(W) = \frac{1}{2}\langle E_3, W\rangle^2$*,* $\beta = 1/10$*, constant* $\eta > 0$ *and zero initial momentum. Let* $T > m \geq 1$ *be integers,* $\delta \in (0,1)$*,* $u_0 = x_0/(\sigma k)$*,* $\alpha = \eta/(\sigma k)$*, and put*

$$d = \beta^{m+1}u_0 + \frac{\alpha\beta}{1-\beta}, \qquad U = u_0 + m\alpha, \qquad L = \log\frac{4(T+1)}{\delta},$$
$$J = \alpha\left(1+\frac{L}{2\gamma_-}\right), \qquad t_{\mathrm{ent}} = m + \left\lceil \frac{2U}{\alpha\gamma_+} + \frac{2L}{\gamma_+^2}\right\rceil + 1.$$

Suppose

$$d \leq \frac{1}{40}, \qquad J \leq \frac{a}{2}, \qquad u_0 - m\alpha > 2a, \qquad U \geq 2+d+\alpha, \qquad T \geq t_{\mathrm{ent}}. \tag{59}$$

Then with probability at least $1-\delta$*, there is* $\tau \leq t_{\mathrm{ent}}$ *such that*

$$a\sigma k \leq \langle E_3, W_t\rangle \leq (b+a/2)\sigma k \quad \textit{for every } \tau \leq t \leq T. \tag{60}$$

In particular the squared population gradient is between $a^2k^2\sigma^2$ *and* $(b+a/2)^2k^2\sigma^2$ *throughout this interval.*

Proof. We first record bounds valid on every path, before introducing a probability event. The update norm is at most one, so $|u_{t+1}-u_t| \leq \alpha$. The exact signal EMA recursion gives

$$|b_t - u_t| \leq \beta^{t+1}u_0 + \frac{\alpha\beta}{1-\beta} \leq d \quad (t \geq m), \qquad u_0 - m\alpha \leq u_m \leq U.$$

If $|u_t| \geq 2 + d$, then $|b_t| \geq 2$ with the same sign. The last samplewise bound in Proposition 28, oddness, and the perturbation estimate ensure that the actual projected update has this sign. Therefore an induction gives $|u_t| \leq U$: above $2 + d$ updates point inward, and below that threshold the next magnitude is at most $2 + d + \alpha \leq U$. The objective is a convex, globally 1-smooth quadratic; in the numerical instance, this bound also keeps every population gradient below 0.251.

Put $H_t = \mathbb{E}[Y_t \mid \mathcal{F}_{t-1}]$ and $D_t = Y_t - H_t$. These are ordinary martingale differences for the preceding-step filtration, even though z_t is an EMA. Conditional on that filtration, $Y_t \in [-1, 1]$, so Hoeffding's lemma gives $\mathbb{E}[\exp(\lambda D_t) \mid \mathcal{F}_{t-1}] \leq \exp(\lambda^2/2)$. For each starting time $i \in \{m, \dots, T\}$, apply Ville's inequality to the three exponential supermartingales with $\lambda = 2\gamma_-$, $\lambda = -2\gamma_+$ and $\lambda = -\gamma_+$. A union bound costs at most $3(T + 1)e^{-L} \leq \delta$. On the resulting event, every $m \leq i < j \leq T$, $n = j - i$, satisfies

$$\sum_{t=i}^{j-1} D_t \leq \gamma_- n + \frac{L}{2\gamma_-}, \tag{61}$$

$$\sum_{t=i}^{j-1} D_t \geq -\gamma_+ n - \frac{L}{2\gamma_+}, \tag{62}$$

$$\sum_{t=i}^{j-1} D_t \geq -\frac{\gamma_+ n}{2} - \frac{L}{\gamma_+}. \tag{63}$$

The martingale differences are defined along the actual process at all states, so these inequalities continue to apply to intervals selected by a first-exit argument.

Because $u_m > 2a$, a first exit below a would have a last entry into $[a, 2a]$ from above, at a state at least $2a - \alpha$. On the subsequent run within this interval, $H_t \leq -\gamma_-$ by (58). Summing the recursion and using (61), the total additional decrease is at most $\alpha L/(2\gamma_-)$. Thus the alleged exit state is at least $2a - J \geq 3a/2 > a$, a contradiction. We have $u_t \geq a$ for every $m \leq t \leq T$.

If the process stays above b for n steps after m, its mean update is at least γ_+. Equation (63) gives

$$u_{m+n} \leq U - \frac{\alpha\gamma_+ n}{2} + \frac{\alpha L}{\gamma_+}.$$

At the stated entry budget this is nonpositive, contradicting the inner bound. Hence some $\tau \leq t_{\mathrm{ent}}$ satisfies $u_\tau \leq b$. After entry, any run above b starts at most α above it, and (62) bounds the further excursion by $\alpha L/(2\gamma_+)$. Since $\gamma_+ \geq \gamma_-$, its height is at most $b + J \leq b + a/2$. Rescaling proves the theorem. □

An evaluated instance. For $\sigma = 1/16$, $x_0 = 1/4$, $\eta = 10^{-5}$, $T = 10^6$, $\delta = 0.05$ and $m = 3$, the certificate evaluates

$$d = 0.000429890 < 0.025, \qquad J = 0.0750651 < 0.1125, \qquad t_{\mathrm{ent}} = 295417 < T.$$

The logarithm and square root in these budgets are enclosed by rational arithmetic: the logarithm uses an 80-term atanh series after binary argument reduction, and the square root uses the same checked enclosure as the drift certificate. Both endpoints give the integer entry bound 295417. All other inequalities in (59) hold, and

$$1.86767578125 \times 10^{-4} \;\leq\; \|\nabla F(W_t)\|_F^2 \;\leq\; 4.1648018 \times 10^{-3} \qquad (295417 \leq t \leq 10^6) \tag{64}$$

with probability at least 0.95. This is a sufficient finite-horizon guarantee for the same parameters executed in the experiment. It establishes a persistent nonzero signal; its band is broader than the observed stationary concentration. The pathwise upper bound in this instance is $|\langle E_3, W_t\rangle| \leq x_0 + m\eta = 0.25003$. Consequently replacing the scalar quadratic by the Huber function in Definition 7 leaves every iterate unchanged: until a hypothetical first exit from $|x| < 1$, the preceding proof applies and rules that exit out. The same example can therefore have a globally bounded population gradient and bounded sample gradients.

Strictly positive curvature is also admissible. Let $\Pi_{E_3}^{\perp}$ be the Frobenius projection orthogonal to E_3 and replace the objective by $F_\mu(W) = \langle E_3, W\rangle^2/2 + \mu\|\Pi_{E_3}^{\perp} W\|_F^2/2$. For $0 < \mu \leq 1$ this is

η	Updates	Endogenous (10^{-3})	Fixed scale (10^{-7})	Prediction error
5×10^{-6}	2,000,000	1.237121 ± 0.001382	0.655856 ± 0.026280	0.030%
1×10^{-5}	1,000,000	1.236738 ± 0.001972	1.292033 ± 0.045607	0.061%
2×10^{-5}	500,000	1.238535 ± 0.003096	2.619157 ± 0.108684	0.084%

Table 6: The same increasing cubic under two choices of scale. Entries are tail mean squared population gradients with 95% intervals across independent trajectories. The prediction column compares the endogenous-scale measurement with the positive stationary-field root, computed without fitting trajectory outputs.

strongly convex and 1-smooth. Since every update has norm at most one and the transverse parameter starts at zero, $\|\Pi_{E_3}^{\perp} W_t\|_F \le \eta t$. Its added gradient is therefore an admissible nonanticipating perturbation in (53) whenever $\mu\eta T \le \sigma/8000$. The same signal residence, and hence the same squared-gradient lower bound, follow. In the evaluated instance $0 < \mu \le 7.8125 \times 10^{-7}$ is allowed. This is an explicit open interval of positive curvatures on the stated horizon; the upper bound concerns the signal coordinate, since the full gradient has an additional transverse component. Curvature and added-noise perturbations share the same budget.

F.4 Removing the endogenous scale

The controlled ablation uses the identical scalar polynomial with an externally fixed scale:

$$\Psi_{\mathrm{fix}}(P) = \frac{1}{4}\left(\frac{P}{\sigma} + \frac{3PP^{\top}P}{\sigma^3}\right) = \nabla_P\left[\frac{\|P\|_F^2}{8\sigma} + \frac{3\,\mathrm{tr}((P^{\top}P)^2)}{16\sigma^3}\right].$$

It is monotone, and its directional derivative is at least $1/(4\sigma)$. In particular, the frozen field for the same stationary noise EMA has the exact projected expression

$$h_{\mathrm{fix}}(x) = \frac{x}{4\sigma}\left(1 + \frac{9(1-\beta)}{1+\beta}\frac{76}{153}\right) + \frac{3x^3}{4\sigma^3}\frac{113}{289}, \tag{65}$$

whose unique zero is $x = 0$ and whose recovery slope is positive. The normalized field has instead $h'_{\mathrm{norm}}(0) = -(25/(204\sigma))\mathbb{E}|Z_\infty|^{-1} < 0$. Thus central symmetry removes the origin mean in both methods, while changing only the scale dependence reverses the recovery sign. The two algorithms are run with the same noise sequence, objective, initial point, momentum and step size. Both execute every EMA and parameter update. The fixed-scale control has nonzero sample fluctuations at the optimum; its constant-step squared-gradient floor is expected to decrease with the step size.

F.5 Executed trajectories and the scale intervention

The experiment uses 128 independently seeded trajectories per step size, sharing each noise sequence between the endogenous-scale method and the fixed-scale control. Every run starts at $W_0 = E_3/4$ and $M_{-1} = 0$, with $\sigma = 1/16$, $\beta = 0.1$ and one sample per update. We keep $\eta T = 10$ across the step-size sweep and report the final 20% of each run. Confidence intervals use the 128 trajectory means, with half-width 1.96 times their standard error.

The stationary Bernoulli convolution is computed independently of the trajectories, using all 2^{12} sign histories. Its omitted geometric tail is at most 10^{-12}, which changes the projected field by less than 8×10^{-12}. The positive stationary-field root predicts $x_*/\sigma = 0.5628478660$ and $x_*^2 = 1.237491095 \times 10^{-3}$. The analytic prediction was evaluated after an exploratory pilot, before the final MT19937 runs; it has no parameters fitted to either set of trajectories. Across the three executed step sizes, the relative error of this predicted squared gradient is at most 0.084%. At the certified $\eta = 10^{-5}$ setting, the endogenous-scale tail mean is 9,572 times the fixed-scale control. All 128 trajectories stay in the certified band after at most 34,935 updates, within the theorem's 295,417-update entry bound.

The small-step control floor decreases in proportion to the step size, while the normalization-only floor stays at the independently predicted nonzero value. The stationary-field root is a numerical prediction; the high-probability residence claim is supplied separately by Theorem 29 and its rational certificate.

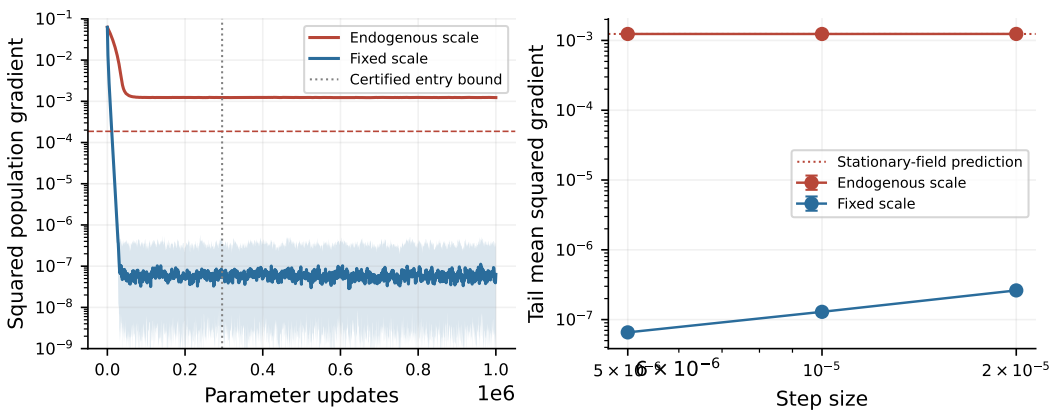


Figure 2: Normalization coupling creates a persistent rank-three optimization floor. Left: actual zero-initialized EMA trajectories at the certified million-step setting; curves show the median and 10–90% range of squared gradients across 128 trajectories. The dashed horizontal line is the proved lower bound. Right: the normalization-only floor remains fixed as the step size decreases; the fixed-scale control floor shrinks. The dotted prediction is computed from the stationary noise law without trajectory fitting.

G Exact cancellation and the polar endpoint

Two consequences on the transverse task. First, keep the objective of Definition 7, replace its noise by

$$\xi = \sigma(\cos\theta\, C_0 + \sin\theta\, C_1), \qquad \theta \sim \mathrm{Unif}[0, 2\pi), \tag{66}$$

and take $\beta = 0$, $B = 1$, $x_0 = \sigma t_\star$. Equation (15) makes the signal update zero on every sample: the squared gradient stays $\sigma^2 t_\star^2$ at every depth and for every positive step-size sequence. This is an algebraic invariant, with attraction depending on the depth and step size. The continuous-radius law in Definition 7 removes this exact cancellation; its actual EMA requires the separate argument of Section 5.

Proposition 30 (Exact polar has a per-input dead zone)**.** *For $Z \in \mathrm{span}\{C_0, C_1\}$ with $\|Z\|_F = \rho$ and $t = a/\rho > 0$, the reduced polar factor, sending zero singular values to zero, satisfies $\langle E, \mathrm{polar}(aE + Z)\rangle = 0$ for $0 < t < 1$, $= 1/\sqrt{2}$ at $t = 1$, and $= \sqrt{2}$ for $t > 1$.*

Monotonicity excludes reversed alignment but allows zero response. Under the continuous stationary radius, averaging gives $h_{\mathrm{polar}}(x) = \sqrt{2}\,\mathbb{P}(R < x) > 0$ for $x > 0$, with quadratic rather than linear decay near zero (Proposition 15). The per-input dead zone does not imply a dead zone of that mean field, and neither statement alone determines an EMA trajectory.

H Nonasymptotic residence certificate for transverse noise

The noise draws are generated independently of the parameters, but the current parameter already depends on earlier draws. Conditioning on the current signal x_t therefore changes the distribution of its accumulated noise. The signal momentum a_t also lags behind x_t. As a result, $\mathbb{E}[q_K(a_t/\rho_t) \mid x_t]$ need not equal the fixed-gradient field $h_K(x_t)$. The theorem below handles both effects for Definition 7, using the actual recursion (2) and the certified constants at $K \in \{1, 5\}$.

A noise scale that is uniform in the memory. Let Z_∞ be the stationary noise EMA and $R = \|Z_\infty\|_F$, with $s^2 = \mathbb{E}R^2 = (\sigma^2/3)v$ and $v = (1-\beta)/(B(1+\beta))$. Its independent, symmetric log-concave coordinates give $\mathbb{P}(R \le \varrho) \le \pi\varrho^2/s^2$ and $s^{-1} \le \mathbb{E}R^{-1} \le (1+2\pi)s^{-1}$, uniformly over B, β (Lemma 32). So $h_K'(0) = f_K'(1/\sqrt{2})\,\mathbb{E}R^{-1}$, fixing the sign of the recovery coefficient at every (B, β) with magnitude of order $1/s$: more memory shrinks the noise scale, it does not turn the negative coefficient of (16) positive. The same bound yields six memory-free barrier constants $(Q, D, a, b, \gamma_-, \gamma_+)$, a supremum and a *local* derivative bound on q_K with two radii and two drift margins, certified at $K = 1$ and $K = 5$ in Appendix P, the second by rational interval

arithmetic. For the bridge, fix an integer $m_{\mathrm{win}} \geq 1$ and a terminal time $T > m_{\mathrm{win}}$, and put $\mathcal{L} = \log[4m_{\mathrm{win}}(T+1)^2/\delta]$, $\gamma = \min\{\gamma_-, \gamma_+\}$, and

$$\begin{aligned} d &= \beta^{m_{\mathrm{win}}} x_0 + \frac{\eta Q \beta}{1-\beta}, \qquad J = \eta Q + \frac{4\eta Q^2 m_{\mathrm{win}} \mathcal{L}}{\gamma}, \\ e &= \frac{2D}{as}\Big[(x_0+\sigma)\beta^{m_{\mathrm{win}}} + \eta Q\big(m_{\mathrm{win}} + \tfrac{\beta}{1-\beta}\big)\Big], \end{aligned} \tag{67}$$

where d bounds the initialization transient and gradient lag, e the state drift and memory truncation, and J the correlated fluctuation. Paired with Theorem 51 for the separation claim, δ becomes $\delta/2$ throughout, changing $\mathcal{L}$ and hence J. Lemma 33 writes the recursion on $[as, 1/2]$, for $t \geq m_{\mathrm{win}}$ and under $d \leq as/2$, as $x_{t+1} = x_t - \eta[h_K(x_t) + D_t + E_t]$ with $|E_t| \leq e$ and D_t a bounded martingale difference *against the delayed filtration* $\mathcal{F}_{t-m_{\mathrm{win}}}$, not $\mathcal{F}_{t-1}$, where the claim would be false. Grouping times by residue modulo m_{win} gives uniform interval fluctuations. Evaluating the centred field at a truncated delayed state keeps that event defined after an exit, which is essential to the first-exit proof.

Theorem 31 (Finite-time residence of the actual nonzero EMA)**.** *Take Definition 7,* $0 < \beta < 1$, $0 < \delta < 1$*, and* $K \in \{1, 5\}$ *with the matching certified constants. Put* $U_0 = x_0 + \eta Q m_{\mathrm{win}}$*. Assume*

$$\eta Q m_{\mathrm{win}} \leq \tfrac{1}{16}, \qquad d \leq \tfrac{as}{2}, \qquad e \leq \tfrac{\gamma}{2}, \qquad J \leq \tfrac{as}{2}, \tag{68}$$

and put $t_{\mathrm{ent}} = m_{\mathrm{win}} + \lceil 4U_0/(\eta\gamma_+) + 32Q^2 m_{\mathrm{win}}\mathcal{L}/\gamma_+^2 \rceil + 1$*. If* $T \geq t_{\mathrm{ent}}$ *then with probability at least* $1-\delta$ *there is* $\tau \leq t_{\mathrm{ent}}$ *with* $as \leq x_t \leq (b + a/2)s < (b+1)s$ *for all* $\tau \leq t \leq T$*, hence for every* $t_{\mathrm{ent}} \leq t \leq T$*,*

$$a^2 \frac{\sigma^2}{3} \frac{1-\beta}{B(1+\beta)} \;\leq\; \|\nabla F(W_t)\|_F^2 \;\leq\; (b+1)^2 \frac{\sigma^2}{3} \frac{1-\beta}{B(1+\beta)}. \tag{69}$$

The proof first establishes pathwise boundedness and only then uses the quadratic objective region. Its inner barrier, entry and outer barrier follow by comparing restoring drift with uniform interval fluctuations (Appendix I). The conclusion is simultaneous high-probability residence on a specified finite horizon. Its conditions must be rechecked when η, T, B or β change. A separate pathwise linearization has Perron root above one when the recovery coefficient is negative (Proposition 35); it supplies local instability, not a substitute for the nonlinear residence proof.

What the theorem establishes quantitatively. This is an existence and scale bridge from the frozen field to a dependent EMA trajectory. It does not yet predict the simulation budget. Corollary 38 improves the proof using drift supermartingales, replacing the fluctuation cost by $\eta Q^2 m_{\mathrm{win}} \log[4m_{\mathrm{win}}(T+1)/\delta]/\gamma$. An analytic remainder bound and exact polynomial certificates also strengthen the depth-one barriers. At $\beta = 0.9$, $B = 1$, $\sigma = 1/16$, $\delta = 0.05$ and a formal $T = 10^6$ budget, the two improvements reduce the evaluated entry bound from 2.27×10^{13} to 1.11×10^{11}, which still exceeds the horizon. Recomputing all conditions at $T = 10^{12}$ gives a valid window with $\eta = 1.62 \times 10^{-11}$ and entry before 1.77×10^{11} (Table 20). Thus the guarantee is nonempty but conservative; the reported simulations remain outside it. These sufficient bounds do not establish a lower bound on the algorithm's entry time.

Robustness has two levels. At depth one, negative frozen recovery survives a rotation of the noise plane by any angle of magnitude below 0.3007 radians, uniformly over B, β, σ (Proposition 39). A separate trajectory result gives explicit budgets for bounded longitudinal noise, noise-plane perturbations and positive transverse curvature on a fixed horizon (Proposition 40). These budgets are conservative; the empirical perturbation study tests a broader regime rather than being asserted to satisfy them.

I Finite-time residence of the actual EMA

The proof has three parts. We first control how often the accumulated noise is small. We then compare the current gradient with its delayed momentum estimate. Finally, we use bounds on drift and random fluctuations to keep the signal inside a positive band. These steps prove Lemma 33 and Theorem 31. Throughout, the setting is Definition 7 with the algorithm exactly as in (2): plain EMA, $M_{-1} = 0$, a parameter update at every step, $\varepsilon = 0$, real arithmetic, and $K \in \{1, 5\}$ with the matching certified constants of Appendix P.

I.1 THE STATIONARY NOISE LAW, UNIFORMLY IN THE MEMORY

Lemma 32 (Uniform small-ball and inverse-moment bounds). *Let $Z_\infty = \sum_{j\ge 0}\sum_{b=1}^{B}\frac{(1-\beta)\beta^j}{B}\xi_{j,b}$ be the stationary noise EMA of Definition 7 and $R = \|Z_\infty\|_F$, with $s^2 = \mathbb{E}R^2 = (\sigma^2/3)v$. Then Z_∞ has a density ϱ_∞ on its two-dimensional support and, for every $B \ge 1$ and $0 < \beta < 1$,*

$$\|\varrho_\infty\|_\infty \le s^{-2}, \qquad \mathbb{P}(R \le \varrho) \le \frac{\pi\varrho^2}{s^2}, \qquad \frac{1}{s} \le \mathbb{E}\frac{1}{R} \le \frac{1+2\pi}{s}. \tag{70}$$

The same density bound also gives $\mathbb{E}[R^{-1}; R < c] \le 2\pi c/s^2$ for every $c > 0$.

Proof. In the orthonormal basis $\{C_0, C_1\}$ the two coordinates of Z_∞ are $\frac{\sigma}{\sqrt{2}}\sum_{j,b}\frac{(1-\beta)\beta^j}{B}U_{j,b}$ and the same with V, so they are independent, identically distributed, symmetric, and each is a weighted sum of independent uniforms. A uniform density is log-concave, log-concavity is preserved by convolution, and it is preserved under weak limits (Saumard & Wellner, 2014), so each coordinate has a symmetric log-concave density f. Its variance is $\frac{\sigma^2}{2}\cdot\frac{1}{3}\sum_{j,b}\big(\frac{(1-\beta)\beta^j}{B}\big)^2 = \frac{\sigma^2 v}{6} = s^2/2$.

Let $M = f(0)$ and $S(t) = \int_t^\infty f$. Log-concavity makes $t \mapsto f(t+u)/f(t)$ nonincreasing for each $u \ge 0$, hence $t \mapsto S(t)/f(t)$ is nonincreasing on $[0,\infty)$, so $S(t)/f(t) \le S(0)/f(0) = 1/(2M)$. Therefore $S'(t) = -f(t) \le -2MS(t)$, and Grönwall gives $2S(t) \le e^{-2Mt}$. Consequently

$$\mathrm{Var} = \int_{-\infty}^{\infty} t^2 f(t)\,dt = 2\int_0^\infty 2tS(t)\,dt \le \int_0^\infty 2te^{-2Mt}\,dt = \frac{1}{2M^2},$$

using integration by parts in the middle step. With $\mathrm{Var} = s^2/2$ this yields $M \le 1/s$, and since a symmetric log-concave density attains its maximum at the origin, $\|f\|_\infty \le 1/s$. The joint density is the product of the two coordinate densities, so $\|\varrho_\infty\|_\infty \le s^{-2}$, and integrating that bound over a disc of radius ϱ gives $\mathbb{P}(R \le \varrho) \le \pi\varrho^2/s^2$.

For the inverse moment, Jensen and $\mathbb{E}R \le (\mathbb{E}R^2)^{1/2} = s$ give $\mathbb{E}R^{-1} \ge 1/\mathbb{E}R \ge 1/s$. For the upper bound the small-ball estimate says the radial density satisfies $f_R(\varrho) \le 2\pi\varrho/s^2$, so

$$\mathbb{E}[R^{-1}; R < c] \le \int_0^c \frac{1}{\varrho}\cdot\frac{2\pi\varrho}{s^2}\,d\varrho = \frac{2\pi c}{s^2},$$

which is the last claim. Taking $c = s$ gives $\mathbb{E}[R^{-1}; R < s] \le 2\pi/s$, while $\mathbb{E}[R^{-1}; R \ge s] \le 1/s$ trivially, and adding the two gives the stated upper bound. □

Every constant in (70) is free of B and β, which is what makes the barriers below memory free. A first consequence concerns the frozen-field slope at the origin. Since q_K is odd with $q_K(0) = 0$ and is globally Lipschitz in its argument, with $C_K = \sup_{t\ge 0}|q_K'(t)|$ finite because q_K' is a continuous function of t vanishing at infinity, we have $|q_K(x/R)|/|x| \le C_K/R$ for every $x \ne 0$ and $R > 0$. That is the domination the exchange needs, and it is a bound on $|q_K|$ rather than on $t|q_K'|$: the weighted local bound (175) controls $t|q_K'(t)|$ and $t^2|q_K'(t)|$ and does not by itself give $|q_K(x/R)|/|x| \le C_K/R$, so we quote the global scalar Lipschitz constant here instead. Since $q_K(x/R)/x \to q_K'(0)/R$ pointwise as $x \downarrow 0$ and R^{-1} is integrable by (70), dominated convergence gives $h_K'(0) = f_K'(1/\sqrt{2})\,\mathbb{E}R^{-1}$. So the sign of the frozen-field slope equals the sign of $f_K'(1/\sqrt{2})$ at every (B,β), and more memory shrinks s without ever flipping it. The depth pattern (16) therefore transfers to the field over the range of depths where it holds, and in particular at $K \in \{1, 5\}$.

I.2 THE DYNAMIC BRIDGE

The exact reduction gives $x_{t+1} = x_t - \eta q_K(a_t/\rho_t)$ with a_t the EMA of $\phi'(x_t)$ and $\rho_t = \|Z_t\|_F$. Write $\mathcal{F}_t$ for the history through the tth innovation and $F_q(x,z) = q_K(x/\|z\|_F)$.

Lemma 33 (Lag, centring and uniform fluctuation). *Take Definition 7 with the algorithm (2), fix an integer $m_{\mathrm{win}} \ge 1$ and a terminal time $T > m_{\mathrm{win}}$, and let d, e and $\mathcal{L}$ be as in (67). Then:*

(i) the filtered gradient lags the current one by at most $\beta^{t+1}x_0 + \eta Q\beta/(1-\beta)$, for every t and every path, with no restriction on t and no condition on d;

(ii) for $t \geq m_{\mathrm{win}}$ and provided $d \leq as/2$, while $x_t \in [as, 1/2]$ the recursion can be written $x_{t+1} = x_t - \eta[h_K(x_t) + D_t + E_t]$ with $|E_t| \leq e$, where D_t is obtained by truncating the noise to its last m_{win} innovations and evaluating the field at the truncated delayed state, and satisfies $\mathbb{E}[D_t \mid \mathcal{F}_{t-m_{\mathrm{win}}}] = 0$ and $|D_t| \leq 2Q$;

(iii) with probability at least $1-\delta$, every interval $m_{\mathrm{win}} \leq i < j \leq T$ satisfies $\left|\sum_{t=i}^{j-1} D_t\right| \leq 2Q\sqrt{2m_{\mathrm{win}}(j-i)\mathcal{L}}$ simultaneously.

Both restrictions in (ii) are used: $t \geq m_{\mathrm{win}}$ is what makes the delayed state $x_{t-m_{\mathrm{win}}}$ and the truncated innovation window well defined, and $d \leq as/2$ is what keeps the segment joining a_t to $\phi'(x_t)$ inside the region where the local derivative bound holds.

Proof of Lemma 33. The lag bound. Unrolling $a_t = \beta a_{t-1} + (1-\beta)\phi'(x_t)$ from $a_{-1} = 0$ gives $a_t = (1-\beta)\sum_{j=0}^{t} \beta^j \phi'(x_{t-j})$. Pad the trajectory to negative indices by $x_u = x_0$ for $u < 0$ and put $\tilde{a}_t = (1-\beta)\sum_{j\geq 0} \beta^j \phi'(x_{t-j})$, a convex combination with weights summing to one. The padded terms carry total weight β^{t+1} and $\phi'(x_0) = x_0$, so

$$a_t = \tilde{a}_t - \beta^{t+1} x_0, \qquad |\tilde{a}_t - \phi'(x_t)| \leq (1-\beta)\sum_{j\geq 1} \beta^j \left|\phi'(x_{t-j}) - \phi'(x_t)\right|.$$

Since ϕ' is 1-Lipschitz and one step moves x by at most ηQ, the jth term is at most $\eta Q \min\{j, t\} \leq \eta Q j$, and $(1-\beta)\sum_{j\geq 1} j\beta^j = \beta/(1-\beta)$ gives $|a_t - \phi'(x_t)| \leq \beta^{t+1} x_0 + \eta Q \beta/(1-\beta)$. The padding is what produces the constant x_0 rather than a bound on $|\phi'(x_t)|$, so this step uses no probabilistic event and no bound on the trajectory, and the global bound below may invoke it without circularity.

The decomposition. Fix $m_{\mathrm{win}} \geq 1$ and truncate the noise to its last m_{win} innovations, $Z_t^{[m_{\mathrm{win}}]} = (1-\beta)\sum_{j=0}^{m_{\mathrm{win}}-1} \beta^j \bar{\xi}_{t-j}$, so $\|Z_t - Z_t^{[m_{\mathrm{win}}]}\|_F \leq \sigma\beta^{m_{\mathrm{win}}}$. Let y_t be the truncation of $x_{t-m_{\mathrm{win}}}$ to $[as, 1/2]$, put $h^{[m_{\mathrm{win}}]}(y) = \mathbb{E}F_q(y, Z_t^{[m_{\mathrm{win}}]})$, and define

$$D_t = F_q(y_t, Z_t^{[m_{\mathrm{win}}]}) - h^{[m_{\mathrm{win}}]}(y_t), \qquad E_t = q_K(a_t/\rho_t) - h_K(x_t) - D_t. \tag{71}$$

Because y_t is $\mathcal{F}_{t-m_{\mathrm{win}}}$-measurable while $Z_t^{[m_{\mathrm{win}}]}$ uses only innovations $t - m_{\mathrm{win}} + 1, \ldots, t$, which are independent of $\mathcal{F}_{t-m_{\mathrm{win}}}$,

$$\mathbb{E}[D_t \mid \mathcal{F}_{t-m_{\mathrm{win}}}] = \mathbb{E}\big[F_q(y, Z_t^{[m_{\mathrm{win}}]})\big]\big|_{y=y_t} - h^{[m_{\mathrm{win}}]}(y_t) = 0,$$

and $|D_t| \leq 2Q$. The conditioning index matters: $Z_t^{[m_{\mathrm{win}}]}$ is *not* independent of $\mathcal{F}_{t-1}$, so the analogous claim at lag one is false, and the grouped martingale below is what replaces it.

For E_t, telescope in five steps while $x_t \in [as, 1/2]$, so that $|y_t - x_t| \leq \eta Q m_{\mathrm{win}}$ and every argument stays in the barrier region where (175) gives $|\partial_x F_q| \leq D/x$ and $\|\nabla_z F_q\| \leq D/x$ with $x \geq as$. Replacing a_t by $\phi'(x_t) = x_t$ costs at most $(2D/(as))d$, because the segment joining them descends only to $x_t - d \geq as/2$, which is the one place $d \leq as/2$ is used. Replacing Z_t by $Z_t^{[m_{\mathrm{win}}]}$ costs at most $(D/(as))\sigma\beta^{m_{\mathrm{win}}}$, since $F_q(x, \cdot)$ is (D/x)-Lipschitz. Replacing x_t by y_t costs at most $(D/(as))\eta Q m_{\mathrm{win}}$. Replacing $h^{[m_{\mathrm{win}}]}$ by h_K at y_t costs at most $(D/(as))\sigma\beta^{m_{\mathrm{win}}}$, by the same Lipschitz bound under a coupling of the finite window with the stationary law that keeps their difference below $\sigma\beta^{m_{\mathrm{win}}}$ almost surely. Finally moving the stationary field from y_t back to x_t costs at most $(D/(as))\eta Q m_{\mathrm{win}}$, since h_K inherits the pointwise Lipschitz constant. Summing the five,

$$|E_t| \leq \frac{2D}{as}\big[d + \eta Q m_{\mathrm{win}} + \sigma\beta^{m_{\mathrm{win}}}\big] = e,$$

which is (67) after substituting d.

The uniform fluctuation bound. Partition $\{m_{\mathrm{win}}, \ldots, T-1\}$ by residue class modulo m_{win}. Within a class the indices differ by at least m_{win}, so by the martingale property just established each class is a bounded martingale difference sequence with increments at most $2Q$ in absolute value. Fix $m_{\mathrm{win}} \leq i < j \leq T$ and let n_ϱ be the number of indices of class ϱ in $[i, j)$, so $\sum_\varrho n_\varrho = j - i$. Azuma's inequality gives, for each class, $\mathbb{P}\big(|\sum_{t\in\varrho} D_t| > 2Q\sqrt{2n_\varrho\mathcal{L}}\big) \leq 2e^{-\mathcal{L}}$. There are at most

m_{win} classes and at most $(T+1)^2/2$ intervals, so a union bound over at most $m_{\mathrm{win}}(T+1)^2$ pairs at $\mathcal{L} = \log[4m_{\mathrm{win}}(T+1)^2/\delta]$ has total failure probability at most δ. On the complement, for every interval,

$$\Big|\sum_{t=i}^{j-1} D_t\Big| \le \sum_{\varrho} 2Q\sqrt{2n_\varrho \mathcal{L}} \le 2Q\sqrt{2\mathcal{L}}\sqrt{m_{\mathrm{win}}\sum_{\varrho} n_\varrho} = 2Q\sqrt{2m_{\mathrm{win}}(j-i)\mathcal{L}},$$

by Cauchy–Schwarz in the form $\sum_\varrho \sqrt{n_\varrho} \le \sqrt{m_{\mathrm{win}}\sum_\varrho n_\varrho}$. Since y_t is a truncation of the state, it is well defined whether or not the real trajectory remains in the barrier region, so the event just constructed does not presuppose the conclusion and may be used inside a first-exit argument. □

I.3 The residence theorem

Proof of Theorem 31. Work on the probability-$(1-\delta)$ event of Lemma 33. The four parts must be taken in this order, since the later ones use the objective's quadratic form, which the first establishes.

Cold start and global boundedness. For $t \le m_{\mathrm{win}}$ the bound $|x_t - x_0| \le \eta Q m_{\mathrm{win}}$ and $\eta Q m_{\mathrm{win}} \le 1/16$ give $x_{m_{\mathrm{win}}} \in [3/16, 5/16]$. Suppose $|x_t| > \sigma + d$ for some $t \ge m_{\mathrm{win}}$. The lag bound gives $|a_t - \phi'(x_t)| \le d$, so a_t has the sign of $\phi'(x_t)$, which is the sign of x_t, and $|a_t| \ge |\phi'(x_t)| - d > \sigma \ge \rho_t$. Hence $|a_t/\rho_t| > 1$, where $q_K > 0$ by (174) at $K = 1$ and by Appendix P at $K = 5$, so the update points toward the origin. Together with a step length of at most ηQ this gives $|x_{t+1}| \le \max\{|x_t|, \sigma + d + \eta Q\}$, and induction from $x_{m_{\mathrm{win}}}$ yields $|x_t| \le U_0 = x_0 + \eta Q m_{\mathrm{win}} \le 5/16$ pathwise for every $t \ge m_{\mathrm{win}}$. In particular $|x_t| < 1$ always, so $\phi'(x_t) = x_t$ throughout and the objective is in its quadratic region. The later use of the locally quadratic form is therefore not circular.

The inner barrier is not crossed. Suppose, for contradiction, that $x_t < as$ for some $t \in [m_{\mathrm{win}}, T]$, and let t_1 be the first such time. Since one step moves at most ηQ and $J \ge \eta Q$, the trajectory must have been in $[as, 2as]$ just before. Such an entry from above exists because $x_{m_{\mathrm{win}}} \ge 3/16$ while $2as < 1/2000$ at both certified rows. Let i be the last time before t_1 at which the trajectory entered $[as, 2as]$ from above $2as$, so $x_i \ge 2as - \eta Q$ and $x_t \in [as, 2as]$ for $i \le t < t_1$. On that interval the inner barrier (171) gives $h_K(x_t) \le -\gamma_- \le -\gamma$, and $|E_t| \le e \le \gamma/2$, so with $n = t_1 - i$,

$$x_{t_1} - x_i = -\eta \sum_{t=i}^{t_1-1} \big[h_K(x_t) + D_t + E_t\big] \ge \frac{\eta\gamma}{2} n - 2\eta Q\sqrt{2m_{\mathrm{win}} n\mathcal{L}} \ge -\frac{4\eta Q^2 m_{\mathrm{win}}\mathcal{L}}{\gamma},$$

the last step minimizing the quadratic $\frac{\eta\gamma}{2}\nu^2 - 2\eta Q\sqrt{2m_{\mathrm{win}}\mathcal{L}}\,\nu$ over $\nu = \sqrt{n} \ge 0$. Hence $x_{t_1} \ge 2as - \eta Q - 4\eta Q^2 m_{\mathrm{win}}\mathcal{L}/\gamma = 2as - J \ge \frac{3}{2}as > as$, contradicting $x_{t_1} < as$. So $x_t \ge as$ for all $m_{\mathrm{win}} \le t \le T$.

Entry by t_{ent}. Suppose $x_t > bs$ for all $m_{\mathrm{win}} \le t \le m_{\mathrm{win}} + n$. On that stretch the outer barrier gives $h_K(x_t) \ge \gamma_+$, so by the same summation

$$x_{m_{\mathrm{win}}+n} \le U_0 - \frac{\eta\gamma_+}{2} n + 2\eta Q\sqrt{2m_{\mathrm{win}} n\mathcal{L}} \le U_0 - \frac{\eta\gamma_+}{4} n + \frac{8\eta Q^2 m_{\mathrm{win}}\mathcal{L}}{\gamma_+},$$

using $2Q\sqrt{2m_{\mathrm{win}} n\mathcal{L}} \le \gamma_+ n/4 + 8Q^2 m_{\mathrm{win}}\mathcal{L}/\gamma_+$, which is the arithmetic-geometric mean inequality. The right side is negative once $n \ge 4U_0/(\eta\gamma_+) + 32Q^2 m_{\mathrm{win}}\mathcal{L}/\gamma_+^2$, contradicting the lower bound $x \ge as > 0$ just proved. So there is $\tau \le t_{\mathrm{ent}}$ with $x_\tau \le bs$; if $x_{m_{\mathrm{win}}} \le bs$ already then $\tau = m_{\mathrm{win}}$ works.

No exit after entry. Consider any maximal run above bs starting after τ. Its entry overshoot is at most ηQ, and repeating the drift-versus-fluctuation computation with γ_+ in place of γ bounds the additional upward excursion by $4\eta Q^2 m_{\mathrm{win}}\mathcal{L}/\gamma_+$. So the highest point reached is at most $bs + J \le (b + a/2)s < (b+1)s$. Combining with the inner bound gives (69), since $\nabla F(W_t) = x_t E$ and $\|E\|_F = 1$. □

Corollary 34 (Expectation form)**.** *Under the same hypotheses, for any fixed $t \in [t_{\mathrm{ent}}, T]$,*

$$(1-\delta)a^2 s^2 \le \mathbb{E}\|\nabla F(W_t)\|_F^2 \le (b+1)^2 s^2 + \delta U_0^2, \tag{72}$$

and for the whole-run average, $\frac{1}{T}\sum_{t<T} \mathbb{E}\|\nabla F(W_t)\|_F^2 \ge (1-\delta)a^2 s^2 (T - t_{\mathrm{ent}})_+/T$.

Proof. The lower bound discards the failure event, whose probability is at most δ. For the upper bound, on the good event $x_t^2 \le (b+1)^2 s^2$, and on the complement the pathwise bound $|x_t| \le U_0$ from the first part of the proof applies, contributing at most δU_0^2. The average retains the initial segment because no lower bound is claimed before t_{ent}. □

Choosing $\delta \le s^2$ makes both sides of (72) of order s^2, at which point every condition in (68) that involves δ through $\mathcal{L}$ must be rechecked with the new value, since $\mathcal{L}$ grows like $\log(1/\delta)$ and enters J.

I.4 Path linearization

Proposition 35 (Growth of the first variation)**.** *Along a fixed noise path with $\rho_t > 0$ for every t, which holds almost surely under Definition 7, the first variation of the reduced recursion about the invariant orbit $x = a = 0$ satisfies*

$$\begin{pmatrix} \delta x_{t+1} \\ \delta a_t \end{pmatrix} = \begin{pmatrix} 1 + \eta c_K (1-\beta)/\rho_t & \eta c_K \beta/\rho_t \\ 1-\beta & \beta \end{pmatrix} \begin{pmatrix} \delta x_t \\ \delta a_{t-1} \end{pmatrix}, \tag{73}$$

where $f_K'(1/\sqrt{2}) = -c_K$. If $c_K > 0$ then each variational matrix is dominated entrywise from below *by the fixed nonnegative matrix obtained by replacing ρ_t with σ, whose Perron root exceeds one, so a positive initial variation grows at least geometrically.*

Proof. Differentiating $x_{t+1} = x_t - \eta q_K(a_t/\rho_t)$ and $a_t = \beta a_{t-1} + (1-\beta)x_t$ in the quadratic region, and using $q_K'(0) = f_K'(1/\sqrt{2}) = -c_K$ at the orbit where $a_t = 0$, gives $\delta x_{t+1} = \delta x_t + \eta c_K \delta a_t/\rho_t$ and $\delta a_t = (1-\beta)\delta x_t + \beta \delta a_{t-1}$; substituting the second into the first yields (73). All four entries are nonnegative, and the two that involve ρ_t are *decreasing* in it, so $\rho_t \le \sigma$ makes each actual matrix dominate the fixed matrix at $\rho_t = \sigma$ entrywise, not the reverse. That is the direction the argument needs: the fixed matrix is the pessimistic one. It has characteristic polynomial $\lambda^2 - \big(1 + \beta + \eta c_K (1-\beta)/\sigma\big)\lambda + \beta$, whose value at $\lambda = 1$ is $-\eta c_K (1-\beta)/\sigma < 0$. A monic quadratic with a negative value at $\lambda = 1$ and positive leading coefficient has a root strictly greater than one, and for a nonnegative matrix that root is the Perron root. Since a product of nonnegative matrices that dominate a fixed nonnegative matrix entrywise dominates its product, a variation with both components positive grows at least at that geometric rate. The requirement $\rho_t > 0$ is what makes (73) defined, and it fails only on a null set. □

This says the negative recovery of (16) enters the first-order variation of the real nonzero EMA, not only the frozen field, which is why reducing the step size shrinks the displacement without changing the direction of the feedback. It is not an alternative proof of Theorem 31: a linearization gives no nonlinear finite-scale residence, and the two results are complementary.

I.5 Nonemptiness of the conditions

Proposition 36 (An explicit admissible choice)**.** *Take*

$$m_{\mathrm{win}} = \max\{1, \lceil \log(16D(x_0+\sigma)/(a\gamma s))/(-\log\beta) \rceil\}$$

and then any

$$0 < \eta \le \min\Big\{ \frac{1}{16 Q m_{\mathrm{win}}},\ \frac{a\gamma s}{16 D Q (m_{\mathrm{win}} + \beta/(1-\beta))},\ \frac{a\gamma s}{16 Q^2 m_{\mathrm{win}} \mathcal{L}} \Big\}.$$

Then (68) *holds and $H = \beta^{m_{\mathrm{win}}} x_0 + \eta Q/(1-\beta) \le a\gamma s/(8D) < as/4$, so the hypothesis of the separation statement following Theorem 51 is met by the same choice.*

Proof. The first term gives $\eta Q m_{\mathrm{win}} \le 1/16$. The choice of m_{win} makes $\beta^{m_{\mathrm{win}}} \le a\gamma s/(16D(x_0+\sigma))$, so $\beta^{m_{\mathrm{win}}}(x_0+\sigma) \le a\gamma s/(16D)$, and the second term bounds $\eta Q(m_{\mathrm{win}} + \beta/(1-\beta)) \le a\gamma s/(16D)$; adding the two and multiplying by $2D/(as)$ gives $e \le \gamma/4 \le \gamma/2$. The same two bounds give $d \le a\gamma s/(16D) + a\gamma s/(16D) \le as/2$ since $\gamma \le 1$ and $D \ge 1$. For J, the third term gives $4\eta Q^2 m_{\mathrm{win}} \mathcal{L}/\gamma \le as/4$ and also $\eta Q \le a\gamma s/(16 Q m_{\mathrm{win}} \mathcal{L}) \le as/4$, using $\gamma \le 1$, $Q \ge 1$, $m_{\mathrm{win}} \ge 1$ and $\mathcal{L} \ge \log 16$, so $J \le as/2$. The bound on H is the same two estimates without the $2D/(as)$ factor. □

For fixed B, β, σ, K, choose m_{win} as above and keep it fixed. Taking $\eta \downarrow 0$ with T of order η^{-2} makes $\mathcal{L} = O(\log(1/\eta))$ and hence $t_{\mathrm{ent}} \ll T$; all smallness conditions eventually hold. The next subsection strengthens the fluctuation argument. Appendix P supplies refined analytic constants and checks a concrete nonempty window, while recording that the evaluated million-step budget is still uncovered.

I.6 A SHARPER CERTIFICATE FROM DRIFT SUPERMARTINGALES

The square-root fluctuation estimate in Lemma 33 controls more than the barrier argument needs. For residence it suffices to control fluctuations against a fixed restoring drift. This gives a smaller, rigorously justified budget without an independence assumption on the residue classes.

Lemma 37 (Uniform fluctuation against drift)**.** *For the D_t in* (71)*, let $\mathcal{L}_* = \log[4m_{\mathrm{win}}(T+1)/\delta]$. With probability at least $1-\delta$, every $m_{\mathrm{win}} \le i < j \le T$, with $n = j - i$, satisfies both*

$$\left|\sum_{t=i}^{j-1} D_t\right| \le \frac{\gamma n}{2} + \frac{Q^2 m_{\mathrm{win}} \mathcal{L}_*}{\gamma}, \tag{74}$$

$$\left|\sum_{t=i}^{j-1} D_t\right| \le \frac{\gamma_+ n}{4} + \frac{2Q^2 m_{\mathrm{win}} \mathcal{L}_*}{\gamma_+}. \tag{75}$$

Here $\gamma = \min\{\gamma_-, \gamma_+\}$, and the constants can be any of the certified rows in Appendix P.

Proof. Conditional on $\mathcal{F}_{t-m_{\mathrm{win}}}$, the variable D_t has mean zero and lies in an interval of length $2Q$: centring a variable in $[-Q, Q]$ changes its endpoints but not its range. Hoeffding's conditional bound therefore gives $\mathbb{E}[e^{\lambda D_t} \mid \mathcal{F}_{t-m_{\mathrm{win}}}] \le e^{\lambda^2 Q^2/2}$ for every real λ. Fix a starting index i and a residue class modulo m_{win}. Along the indices of that class at or after i, the products

$$\exp\left\{\lambda \sum_{k=1}^{N} D_{t_k} - \frac{\lambda^2 Q^2 N}{2}\right\}$$

are nonnegative supermartingales: successive indices differ by m_{win}, exactly the delay in the conditional bound. Ville's inequality bounds the probability that any such product exceeds $e^{\mathcal{L}_*}$ by $e^{-\mathcal{L}_*}$. Apply this to both signs of λ and to $|\lambda| = \gamma/Q^2$ and $|\lambda| = \gamma_+/(2Q^2)$. There are at most $T+1$ starting indices, m_{win} classes, two signs and two choices of magnitude, so the union has probability at most $4m_{\mathrm{win}}(T+1)e^{-\mathcal{L}_*} = \delta$.

On the complement, for either positive value of λ, every class contributes at most $\lambda Q^2 n_r/2 + \mathcal{L}_*/\lambda$ in either signed direction on any interval $[i, j)$. Summing the classes, using $\sum_r n_r = n$, gives $|\sum D_t| \le \lambda Q^2 n/2 + m_{\mathrm{win}}\mathcal{L}_*/\lambda$. Substituting the two values proves the claim. Summation and the union bound do not require the classes to be independent. As in Lemma 33, the truncated delayed state keeps this event defined after any exit from the barrier region. □

Corollary 38 (Sharpened finite-time residence)**.** *Use either original row* (172) *or the refined depth-one row* (178)*, and define d, e, U_0 as in Theorem 31. Its conclusion continues to hold when J and t_{ent} are replaced by*

$$J_* = \eta Q + \frac{\eta Q^2 m_{\mathrm{win}} \mathcal{L}_*}{\gamma}, \qquad t_{\mathrm{ent},*} = m_{\mathrm{win}} + \left\lceil \frac{4U_0}{\eta\gamma_+} + \frac{8Q^2 m_{\mathrm{win}} \mathcal{L}_*}{\gamma_+^2} \right\rceil + 1, \tag{76}$$

provided $\eta Q m_{\mathrm{win}} \le 1/16$, $d \le as/2$, $e \le \gamma/2$, $J_* \le as/2$ and $T \ge t_{\mathrm{ent},*}$. In particular, the same m_{win} as in Proposition 36 and

$$0 < \eta \le \min\left\{\frac{1}{16 Q m_{\mathrm{win}}},\ \frac{a\gamma s}{16DQ(m_{\mathrm{win}} + \beta/(1-\beta))},\ \frac{a\gamma s}{4Q^2 m_{\mathrm{win}} \mathcal{L}_*}\right\} \tag{77}$$

satisfy all four smallness conditions. A joint statement with the repair theorem uses failure probability $\delta/2$ for each result, hence $\mathcal{L}_* = \log[8m_{\mathrm{win}}(T+1)/\delta]$ on this side; this does not change the original budget (67) *when the original theorem is used.*

Proof. The lag, decomposition and cold-start arguments are unchanged. For the refined row, the numerical facts needed in that argument also hold uniformly: $2as < 10^{-3} < 3/16$ and $\sigma + d + \eta Q \le \sigma + as < 3/16 \le U_0$. On a last run into the inner barrier, the outward deterministic drift is at least $\eta\gamma n/2$. Subtract (74) to obtain $x_{t_1} \ge 2as - \eta Q - \eta Q^2 m_{\mathrm{win}} \mathcal{L}_* / \gamma = 2as - J_*$. The same bound controls upward excursions above the outer barrier because $\gamma_+/2 \ge \gamma/2$. For entry, (75) gives

$$x_{m_{\mathrm{win}}+n} \le U_0 - \frac{\eta\gamma_+ n}{4} + \frac{2\eta Q^2 m_{\mathrm{win}} \mathcal{L}_*}{\gamma_+},$$

which contradicts the positive inner barrier by the stated entry time. These are precisely the three first-exit arguments in the original proof, with their respective budgets replaced.

The first two terms of (77) give the same cold-start, lag and approximation bounds as Proposition 36. The third gives $\eta Q^2 m_{\mathrm{win}} \mathcal{L}_* / \gamma \le as/4$ and $\eta Q \le a\gamma s/(4Q m_{\mathrm{win}} \mathcal{L}_*) \le as/4$, since $\gamma \le 1$, $Q \ge 1$ and $\mathcal{L}_* \ge 1$. Thus $J_* \le as/2$. The failure-probability allocation is a union bound between the two theorem events and does not assume that the two algorithms, which can share their noise, are independent. □

I.7 Recovery under a rotation of the noise plane

The noise being exactly orthogonal to the signal is useful for the unperturbed construction, but the negative recovery coefficient is not confined to that orientation. Define

$$\xi_\theta = \frac{\sigma}{\sqrt{2}} \big[U(\cos\theta\, C_0 + \sin\theta\, E) + V C_1 \big].$$

This rotates the noise plane while preserving its covariance eigenvalues, boundedness and central symmetry. Write the original stationary noise coordinates as A, B_0, so that under the rotation $L_\theta = \sin\theta\, A$ is the longitudinal coordinate and $R_\theta = (\cos^2\theta\, A^2 + B_0^2)^{1/2}$ is the transverse radius. The frozen signal field is $H_{K,\theta}(x) = \mathbb{E} q_K((x + L_\theta)/R_\theta)$.

Proposition 39 (An explicit open interval of adverse orientations)**.** *At depth one define*

$$u_c = \left(\frac{-95742 + \sqrt{95742^2 + 4 \cdot 23911 \cdot 9429}}{2 \cdot 23911} \right)^{1/2}, \qquad \theta_c = \arctan u_c = 0.300713738\ldots.$$

For every $|\theta| < \theta_c$ and every (B, β, σ) of Definition 7, $H_{1,\theta}(0) = 0$ and

$$H'_{1,\theta}(0) \le -\frac{c_\theta}{s_\theta} < 0, \qquad c_\theta = \frac{9429 - 95742\tan^2\theta - 23911\tan^4\theta}{8000(1+\tan^2\theta)^{7/2}}, \qquad s_\theta = s\sqrt{\frac{1+\cos^2\theta}{2}}. \tag{78}$$

At depth five, $|\tan\theta| \le 1/100$ likewise gives $H'_{5,\theta}(0) \le -0.9/s_\theta$.

Proof. The simultaneous sign reversal $(A, B_0) \mapsto (-A, -B_0)$ preserves the law, reverses L_θ and leaves R_θ fixed. Oddness of q_K gives zero field at the origin. Moreover $|L_\theta|/R_\theta \le |\tan\theta|$ and $R_\theta \ge |\cos\theta|(A^2 + B_0^2)^{1/2}$. The latter and Lemma 32 make R_θ^{-1} integrable; the bounded scalar derivative of q_K then justifies differentiating under the expectation. By (173), on $|t| \le |\tan\theta| < u_c$ its derivative is at most $-c_\theta$. Thus $H'_{1,\theta}(0) \le -c_\theta \mathbb{E} R_\theta^{-1} \le -c_\theta/s_\theta$ by Jensen, because $\mathbb{E} R_\theta^2 = s^2(1+\cos^2\theta)/2$. At depth five use the evenness of q_5' and Lemma 71 in the same argument. □

The depth-one interval is approximately ± 17.23 degrees. This is a uniform sign guarantee for a perturbed frozen field. It neither identifies an attracting nonzero equilibrium nor asserts that the finite-time residence conditions hold throughout that interval.

I.8 Finite-horizon residence under adapted perturbations

A separate result extends the actual trajectory conclusion. It records the smallness budget explicitly, so the field-level orientation result is not substituted for a trajectory proof. Let $\ell_t \in \mathbb{R}$ and $r_t \in \mathrm{span}\{C_0, C_1\}$ be nonanticipating perturbations with $|\ell_t| \le \nu_\parallel$ and $\|r_t\|_F \le \nu_\perp$ almost surely.

Innovations at future times remain independent of the enlarged history. Replace the gradient observation by

$$\big(\phi'(x_t) + \ell_t\big)E + \bar{\xi}_t + r_t,$$

and retain the same zero EMA initialization and parameter updates. Perturbations can depend on the current innovation and the previous trajectory.

Proposition 40 (A nonzero perturbation budget for actual residence)**.** *With either certificate row, set*

$$d_\nu = d + \nu_\parallel, \qquad e_\nu = e + \frac{2D\nu_\parallel + D\nu_\perp}{as}. \tag{79}$$

The signal residence conclusion of Corollary 38 holds for this perturbed algorithm if its conditions are imposed with d_ν, e_ν in place of d, e and additionally

$$\sigma + \nu_\perp + d_\nu + \eta Q \le U_0. \tag{80}$$

All probability statements use the original iid innovations, together with any auxiliary randomness in the nonanticipating perturbations. Strict inequalities in these sufficient conditions leave a nonzero open set of admissible $(\nu_\parallel, \nu_\perp)$.

Proof. Split the signal EMA into the EMA of $\phi'(x_t)$ and that of ℓ_t, the latter bounded by $\nu_\parallel$. The pathwise lag bound is therefore d_ν for $t \ge m_{\mathrm{win}}$. Likewise split the transverse EMA into the original exogenous Z_t^0 and the EMA of r_t; their difference is bounded by $\nu_\perp$, and the actual radius is at most $\sigma + \nu_\perp$. Under (80), the cold-start and inward-update argument gives $|x_t| \le U_0$ pathwise just as before. In particular the trajectory remains in the quadratic region.

Define D_t using the last m_{win} original innovations and the truncated delayed value of the actual perturbed trajectory. It remains centred conditional on the enlarged $\mathcal{F}_{t-m_{\mathrm{win}}}$; independence from future innovations is all this step uses. The five-step decomposition in Lemma 33 now pays an additional $2D\nu_\parallel/(as)$ for the longitudinal EMA and $D\nu_\perp/(as)$ to replace the transverse EMA by Z_t^0. The segment in the longitudinal replacement stays above $as/2$ by the assumed lag bound. Thus $|E_t| \le e_\nu$. Lemma 37 applies to these same centred variables, and all three first-exit arguments of Corollary 38 are unchanged. □

This proposition covers bounded longitudinal noise and the preceding rotation by taking $\nu_\parallel = \sigma|\sin\theta|/\sqrt{2}$ and $\nu_\perp = \sigma|1 - \cos\theta|/\sqrt{2}$. Its certified angular window can be much smaller than Proposition 39's sign window, since the two results answer different questions.

It also gives an explicit finite-horizon extension to small transverse curvature. Let $\Pi_\perp$ project onto $\mathrm{span}\{C_0, C_1\}$ and set

$$F_\mu(W) = \phi(\langle E, W\rangle) + \frac{\mu}{2}\|\Pi_\perp W\|_F^2, \qquad 0 < \mu \le 1.$$

This objective is convex and 1-smooth. The rotating-frame algebra of Lemma 8 preserves $\mathrm{span}\{E, C_0, C_1\}$ under the update. For $K = 1, 5$, the invariant scalar interval (176) bounds the full update by $\|\Psi_K(P)\|_F \le Q_{\mathrm{mat}} := 5\sqrt{2}/4$. Starting from zero transverse parameter, $\|\Pi_\perp W_t\|_F \le \eta Q_{\mathrm{mat}} t$ pathwise. Hence Proposition 40 applies up to T with $\nu_\parallel = 0$ and $\nu_\perp = \mu\eta Q_{\mathrm{mat}} T$ whenever the displayed budgets hold. Its signal lower bound implies $\|\nabla F_\mu(W_t)\|_F^2 \ge a^2 s^2$ throughout residence. This proves existence of strictly positive admissible curvature on every fixed certified horizon; it does not give a curvature bound uniform as $T \to \infty$, or certify the larger perturbations used in the empirical sensitivity study.

J An increasing and flattening response: what survives momentum

The rank-three trajectory of Appendix F uses the response $f_+(s) = (s + 3s^3)/4$, which is increasing but expands ratios of singular values, at $\beta = 1/10$. This appendix repeats the normalization-only mechanism with the classical Newton–Schulz cubic, which is increasing and flattens every ratio, and determines what part of the obstruction survives momentum. The supplementary dynamics study supplies the numerical results; the independent rational certificate supplies the constants used in Appendix J.1.

Lemma 41 (Monotone and flattening at every depth). *Let $p(u) = (3u - u^3)/2$ and $f_K = p^{\circ K}$. On $[0, 1]$ every f_K is strictly increasing and $f_K(u)/u$ is strictly decreasing.*

Proof. $p'(u) = 3(1 - u^2)/2 \geq 0$, $(p(u)/u)' = -u < 0$ and $p([0, 1]) = [0, 1]$. If g has both properties then $(p \circ g)' = p'(g)g' \geq 0$ and $p(g(u))/u = [p(g(u))/g(u)]\,[g(u)/u]$ is a product of nonnegative decreasing functions. Exact factorizations of f_K' and $(f_K/u)'$ for $K \leq 3$ are recorded in the certificate file. □

The witness. With $S_3 = \mathrm{diag}(4, 1, 1)/\sqrt{18}$, $D = \mathrm{diag}(f_1'(s_i)) = \mathrm{diag}(1/6, 17/12, 17/12)$ and the Jacobian $D(I - ss^\top)$ of Theorem 2, a direction $e \propto (a, b, b)$ has quadratic form $(a - 4b)(a - 34b)/(54(a^2 + 2b^2))$, negative exactly when $4 < a/b < 34$ and equal to $-1/34$ at $E = \mathrm{diag}(7, 1, 1)/\sqrt{51}$. Full 3×3 finite differences confirm the value to 10^{-8}. The adverse direction loads the dominant singular value more than the noise does; increasing it under normalization shrinks the minor normalized singular values, where the flattening slope $17/12$ is large. A numerical minimization over the positive octant of the unit sphere gives the most negative eigenvalue -0.0376 of the symmetric Jacobian at $K = 1$ (-0.0669 at $K = 3$): the cubic's normalization-only obstruction is small compared with its descent value $g_\infty = \sum_i E_i p(E_i) = 1.038$ along E at vanishing noise. This size difference decides what survives momentum.

Task and averaged field. Take $F(W) = \langle E, W\rangle^2/2$, $W_0 = x_0 E$, $M_{-1} = 0$, batch size one and $\xi_t = \sigma\epsilon_t S_3$ with iid Rademacher signs. Writing $x_t = \langle E, W_t\rangle$, $M_t = a_t E + \sigma\zeta_t S_3$ with $a_t = \beta a_{t-1} + (1 - \beta)x_t$ and $\zeta_t = \beta\zeta_{t-1} + (1 - \beta)\epsilon_t$, the recursion is $x_{t+1} = x_t - \eta q(a_t/\sigma, \zeta_t)$ with $q(b, z) = \langle E, \Psi_{f_K}(bE + zS_3)\rangle$, exactly. By zero-homogeneity $q(b, z) = g(b/z)$ for $z > 0$ and $q(b, z) = -g(-b/|z|)$ for $z < 0$, where $g(u) = q(u, 1)$, so for any symmetric law of ζ the averaged field is

$$H(b) = \mathbb{E}\, q(b, \zeta) = \mathbb{E}\, \varphi(b/|\zeta|), \qquad \varphi(u) = \tfrac{1}{2}[g(u) - g(-u)], \qquad \varphi'(0) = -\tfrac{1}{34}. \tag{81}$$

Here $b = x/\sigma$ and ζ has the stationary law of the noise EMA with unit innovations.

Lemma 42 (Small-ball threshold). *For $\zeta = (1 - \beta)\sum_{j\geq 0} \beta^j \epsilon_j$ with $\epsilon_j \in \{\pm 1\}$, $\min|\zeta| = \max(0, 1 - 2\beta)$; for $\beta < 1/2$ the minimum is attained at $\epsilon_0 = 1$, $\epsilon_j = -1$ for $j \geq 1$, and for $\beta \geq 1/2$ zero is a limit of atoms. At $\beta = 1/2$, ζ is uniform on $[-1, 1]$.*

Proof. $|\zeta| \geq (1 - \beta)(1 - \sum_{j\geq 1} \beta^j) = 1 - 2\beta$ with equality at the stated signs. For $\beta \geq 1/2$, the two affine images $\beta[-1, 1] \pm (1 - \beta)$ overlap and cover $[-1, 1]$. Recursively selecting an inverse branch for zero constructs an infinite sign address for it; its finite prefixes approach zero. At $\beta = 1/2$ the sum $\frac{1}{2}\sum_j 2^{-j}\epsilon_j$ is a signed binary expansion. □

Proposition 43 (What survives momentum with one-dimensional noise). *Let $K = 1$ and $t_0 = 0.780$. Interval arithmetic certifies $\varphi < 0$ on $(0, (1 - m)t_0]$ and $\varphi > 0$ on $[(1 + m)t_0, \infty)$ with $m = 0.02$. (a) For $\beta < 1/2$, $H(b) < 0$ for $0 < b < (1 - m)t_0(1 - 2\beta)$ and $H(b) > 0$ for $b > (1 + m)t_0$, so H has a positive root in $[(1 - m)t_0(1 - 2\beta), (1 + m)t_0]$. (b) For $\beta = 1/2$,*

$$H(b) = b\int_b^\infty \frac{\varphi(u)}{u^2}\, du > 0 \qquad \text{for every } b > b_0 = t_0 e^{-I_+/c}, \tag{82}$$

where $c = \sup_{0<u\leq t_0}(-\varphi'(u)) \leq 0.0448$ and $I_+ = \int_{t_0}^\infty \varphi(u)u^{-2}du \geq 1.032$ are interval-certified, so $b_0 \leq 10^{-10.1}$; and $H(b)/b \to -\infty$ as $b \downarrow 0$, so every positive root is below b_0.

Proof. (a) By Lemma 42, $b/|\zeta| \leq b/(1 - 2\beta) < (1 - m)t_0$ on the first range and $b/|\zeta| \geq b > (1 + m)t_0$ on the second, so the integrand in (81) has a fixed sign; H is continuous. (b) With ζ uniform, substitute $u = b/z$ in $H(b) = \int_0^1 \varphi(b/z)\, dz$. The mean value theorem gives $\varphi(u) \geq -cu$ on $(0, t_0]$, hence $\int_b^\infty \varphi u^{-2}du \geq I_+ - c\log(t_0/b)$, positive iff $b > b_0$. For $t_0 \leq b \leq (1 + m)t_0$, use $|\varphi| \leq 9/\sqrt{51}$ to get $\int_b^\infty \varphi(u)u^{-2}du \geq I_+ - (9/\sqrt{51})m/[(1 + m)t_0] > 0$; beyond $(1 + m)t_0$ the integrand is positive. As $b \downarrow 0$ the integral behaves like $\varphi'(0)\log(1/b) \to -\infty$. The constants are enclosed by mean-value interval evaluations on 2000 subintervals below t_0, 4000 log-spaced subintervals on $[t_0, 10^6]$ and the substitution $v = 1/u$ on the tail. □

Table 7: Momentum sweep for the increasing and flattening cubic. Root and RMS levels are in units of σ; the forecast is the noise-only level $\sqrt{b_*^2 + V_h}$; entry is the median first time $|x_t| \le 1.5\sigma b_*$ against the averaged-path forecast. "none" means the Monte Carlo field is positive on $[.005, 4]$.

K	β	Root b_*	Forecast RMS	Observed RMS	Fixed-scale RMS	Entry (median / bound)
1	0.1	0.656	0.656	0.657	0.0040	18,499 / 18,498
1	0.3	0.352	0.352	0.351	0.0040	23,500 / 23,499
1	0.45	0.097	0.097	0.092	0.0038	33,068 / 33,233
1	0.5	none	—	0.0110	0.0039	—
1	0.7	none	—	0.0146	0.0037	—
1	0.9	none	—	0.0189	0.0039	—
1	0.95	none	—	0.0223	0.0040	—
3	0.1	0.740	0.740	0.740	0.0058	16,726 / 16,726
3	0.3	0.384	0.384	0.384	0.0057	21,712 / 21,726
3	0.45	0.104	0.104	0.105	0.0057	31,370 / 31,528
3	0.5	none	—	0.0123	0.0056	—
3	0.7	none	—	0.0163	0.0057	—
3	0.9	none	—	0.0209	0.0057	—
3	0.95	none	—	0.0246	0.0057	—

The two regimes are the small-ball competition: for $|\zeta| \gg b$ the update is the obstruction $\varphi'(0)b/|\zeta|$, of size at most $cb\,\mathbb{E}|\zeta|^{-1}$ where that expectation is finite, while for $|\zeta| \ll b$ the normalized update is full descent along E, $\varphi \approx g_\infty$, an $O(1)$ term. When the radius is bounded away from zero the second term is absent and the obstruction always wins at small b; when the radius has small-ball mass the obstruction wins only below b_0, exponentially small in $1/c$. For $\beta \in \{.7, .9, .95\}$ and both depths the Monte Carlo field is positive on the whole grid $b \in [.005, 4]$ (Table 7), consistent with (b).

Trajectories across momentum. Table 7 and Figure 3 report 128 zero-initialized trajectories per cell with $\sigma = 1/16$, $x_0 = 1/4$, $\eta = 10^{-5}$ and $T = 10^6$, at $K \in \{1, 3\}$ and seven values of β, sharing each noise sequence with a fixed-scale control. The control replaces the endogenous norm by the scale $1/2 = 8\sigma$: the cubic is monotone only on the unit ball, and the scale must exceed the pathwise bound $x_0 + \sigma$ on the momentum entries (the largest normalized argument along all runs is 0.608). Appendix J.1 explains why the scale σ itself is inadmissible for this cubic. Below the threshold the endogenous method resides at a nonzero gradient whose tail RMS agrees with the noise-only forecast $\sigma\sqrt{b_*^2 + V_h}$, computed as in Proposition 48 from the root b_*, the slope $\kappa_* = H'(b_*)$ and the linear response to the centred noise functional $q(b_*, \zeta_t)$; entry times follow the averaged path $\dot{b} = -H(b)/\sigma$. At $\beta = .3$ the observed RMS is $0.351\,\sigma$ against the forecast $0.352\,\sigma$, with the control at $0.0040\,\sigma$, and every trajectory enters $|x| \le 1.5x_*$ by update 24,092 (median 23,500, forecast 23,499). At $\beta \ge 1/2$ the field has no root above the grid and the endogenous method's tail level drops by more than an order of magnitude, to $0.0110\,\sigma$ at $\beta = .5$ and $0.0189\,\sigma$ at $\beta = .9$, a small multiple of the control's $0.0039\,\sigma$. The two regimes differ in kind, not only in size: below the threshold the level σb_* does not depend on η, whereas at $\beta = .9$ a step-size sweep at fixed ηT ($\eta \in \{10^{-5}, 3 \times 10^{-6}, 1e-06\}$) gives log-log slopes of 0.50 for the endogenous method and 0.53 for the control, consistent with constant-step floors of order $\sqrt{\eta}$ ($0.0061\,\sigma$ against $0.0012\,\sigma$ at $\eta = 1e-06$). No step-size-independent level is resolved at these scales; this sweep does not exclude a root below the field grid. A dense 3×3 implementation of the polynomial iteration in a random orthogonal frame reproduces the scalar recursion to $2.8e-16$ over 4000 steps of four trajectories.

Multidimensional noise. Under $k \ge 2$ noise coordinates the stationary EMA radius has finite $\mathbb{E}R^{-1}$ at every β, so the frozen recovery slope along E is $\mathbb{E}[Q(\widehat{Z})/R]$ with $Q(\hat{z}) = \langle E, D\Psi_{f_K}(\hat{z})[E]\rangle$, and an obstruction that survives momentum requires a direction E on which this direction average is negative. We searched over ranks $r \in \{3, 4, 6, 8\}$, depths $K \le 3$, and noise subspaces spanned by $k \in \{2, 3\}$ diagonal directions with iid uniform coordinates at $\beta = 0.9$ (400 random and structured frames per cell, including anisotropic frames), minimizing the smallest eigenvalue of the symmetric part of $\mathbb{E}[(D(\hat{z}) - D(\hat{z})\hat{z}\hat{z}^\top)/R]$ over all unit E (Table 8). Every cell is positive; the smallest value is +4.6443 at $r = 3$, $k = 3$, $K = 1$. A targeted search around the one-dimensional witness, with a transverse coordinate of relative size $\delta \ge 0.01$, is also positive (smallest +4.712). For the rank-three witness the adverse directions occupy a restricted cone, and these searches suggest that the positive contributions dominate after averaging. This is an explanation of the observed small-rank results, not

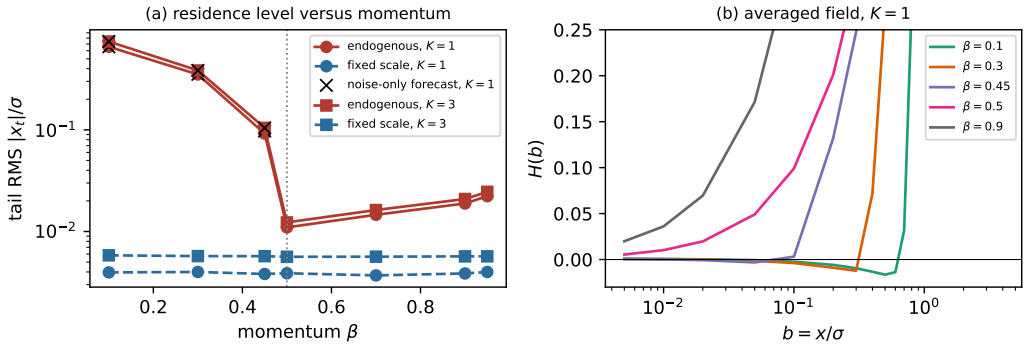


Figure 3: (a) Tail RMS signal of the endogenous-scale method and of the fixed-scale control against momentum, with the noise-only forecast where the field has a root; the dotted line is the threshold $\beta = 1/2$ of Lemma 42. (b) The averaged field $H(b)$ at $K = 1$: a negative region and a positive root below the threshold, positivity above it.

Table 8: Smallest eigenvalue of the direction-averaged symmetric Jacobian at $\beta = 0.9$ over the searched frames; positive values mean no adverse direction.

r	k	K	$\lambda_{\min}$ of averaged symmetric Jacobian	$\mathbb{E}[1/R]$
3	2	1	+5.2479	9.53
3	2	2	+5.3645	9.24
3	2	3	+5.3489	22.35
3	3	1	+4.6443	5.97
3	3	2	+5.2621	6.07
3	3	3	+5.5513	6.06
4	2	1	+5.0039	8.97
4	2	2	+5.5248	9.58
4	2	3	+5.3092	21.67
4	3	1	+4.6735	5.95
4	3	2	+5.1773	5.91
4	3	3	+5.3955	5.84
6	2	1	+5.0616	9.15
6	2	2	+5.6835	22.35
6	2	3	+5.6467	9.45
6	3	1	+4.7195	5.95
6	3	2	+5.3419	6.01
6	3	3	+5.5679	6.00
8	2	1	+5.2845	9.58
8	2	2	+5.5815	9.36
8	2	3	+5.3582	22.01
8	3	1	+4.7748	5.91
8	3	2	+5.3558	5.99
8	3	3	+5.5066	5.85

a general positivity argument: the higher-rank examples below have visible adverse fields at high momentum. The reference quintic's transverse obstruction (Definition 7) is different in kind: its slope $f'_K(1/\sqrt{2}) = -1.18$ at $K = 1$ is the same for every noise direction, so it survives every β with magnitude $\mathbb{E}R^{-1}$. These are numerical findings for the searched ranks $r \le 8$, depths $K \le 3$ and diagonal noise laws. They neither exclude higher ranks nor higher polynomial depth.

J.1 A RESIDENCE CERTIFICATE FOR AN INCREASING AND FLATTENING RESPONSE AT $\beta = \frac{3}{10}$

The witness of Appendix F uses $f_+(s) = (s + 3s^3)/4$, which is increasing but expands ratios of singular values, and the momentum $\beta = 1/10$. This subsection certifies the same phenomenon for

the classical Newton–Schulz cubic

$$f(u) = \frac{3u - u^3}{2}, \qquad S_3 = \frac{\operatorname{diag}(4,1,1)}{\sqrt{18}}, \qquad E = \frac{\operatorname{diag}(7,1,1)}{\sqrt{51}}, \tag{83}$$

at $\beta = \frac{3}{10}$. On $[0,1]$, $f'(u) = 3(1-u^2)/2 \geq 0$ and $(f(u)/u)' = -u < 0$: the response is strictly increasing and flattens every ratio of singular values. Both matrices have unit Frobenius norm. In the common singular frame $D = \operatorname{diag}(f'(s_i)) = \operatorname{diag}(1/6, 17/12, 17/12)$, and the Jacobian $D(I - ss^\top)$ of Theorem 2 gives

$$\langle E, D\Psi_f(S_3)[E]\rangle = \sum_i f'(s_i)E_i^2 - \langle S_3, E\rangle \sum_i f'(s_i)s_iE_i = \frac{11}{51} - \frac{25}{102} = -\frac{1}{34} < 0. \tag{84}$$

Every scalar slope is positive; the negative sign is the derivative of the normalization.

Task and scalar recursion. Take $F(W) = \frac{1}{2}\langle E, W\rangle^2$, $W_0 = x_0E$ with $x_0 > 0$, batch size one, $\beta = \frac{3}{10}$, and

$$\xi_t = \sigma\epsilon_t S_3 + R_t, \qquad \mathbb{P}(\epsilon_t = 1) = \mathbb{P}(\epsilon_t = -1) = \tfrac{1}{2}, \qquad \|R_t\|_F \leq \frac{\sigma}{60000}, \tag{85}$$

with iid signs independent of the preceding history and a nonanticipating perturbation R_t that may depend on the current sign. Put $k = \sqrt{51/18}$, $u_t = \langle E, W_t\rangle/(\sigma k)$ and $\alpha = \eta/(\sigma k)$. Exactly as in Appendix F,

$$M_t = \sigma k\, b_t E + \sigma z_t S_3 + \Delta_t, \qquad b_t = \beta b_{t-1} + (1-\beta)u_t, \qquad z_t = \beta z_{t-1} + (1-\beta)\epsilon_t,$$

with $b_{-1} = z_{-1} = 0$, $\Delta_t = \beta\Delta_{t-1} + (1-\beta)R_t$ and $\|\Delta_t\|_F \leq \sigma/60000$. The scaling k makes $(M_t - \Delta_t)/\sigma = \operatorname{diag}(7b_t + 4z_t,\, b_t + z_t,\, b_t + z_t)/\sqrt{18}$, whose squared Frobenius norm is $\rho^2/18$ with the positive form

$$\rho^2(b,z) = 51b^2 + 60bz + 18z^2 = 51\Big(b + \frac{10z}{17}\Big)^2 + \frac{6z^2}{17}. \tag{86}$$

For $R_t = 0$, diagonal evaluation of (1) with $m = (7b{+}4z,\, b{+}z,\, b{+}z)$ gives $u_{t+1} = u_t - \alpha q(b_t, z_t)$, where

$$q(b,z) = \sum_i E_i\, f\Big(\frac{m_i}{\rho}\Big) = \frac{9\,(300b^3 + 536b^2z + 322bz^2 + 65z^3)}{\sqrt{51}\,\rho(b,z)^3}. \tag{87}$$

This is odd and zero-homogeneous, $q(b,z) = q(b/|z|, \operatorname{sign} z)$ for $z \neq 0$, and its partial derivatives are

$$\partial_b q(b,z) = -\frac{27z\,P(b,z)}{\sqrt{51}\,\rho^5}, \qquad \partial_z q(b,z) = -\frac{b}{z}\,\partial_b q(b,z), \tag{88}$$

where $P(b,z) = 112b^3 + 188b^2z + 103bz^2 + 18z^3$. For $R_t \neq 0$ the same recursion holds with $q(b_t, z_t)$ replaced by $Y_t = \langle E, \Psi_f(M_t)\rangle$. Every identity in (84)–(88) is verified symbolically by the certificate script, and a separate full-matrix implementation (singular value decomposition, no scalar shortcut) reproduces (87) to 2.9×10^{-15} and replayed paths to 1.6×10^{-15}.

The noise EMA below $\beta = 1/2$. Since $\beta|z_{t-1}| \leq \beta < 1 - \beta$, induction from $z_{-1} = 0$ gives, for every $t \geq 0$,

$$\operatorname{sign} z_t = \epsilon_t, \qquad 1 - 2\beta = \tfrac{2}{5} \leq |z_t| \leq 1. \tag{89}$$

This covers the transient as well as the stationary law: $z_0 = \pm 7/10$, and the lower end is approached when every later sign opposes the earlier ones. The conditional field given the preceding EMA $v = z_{t-1} \in [-1,1]$ is

$$G(b,v) = \tfrac{1}{2}\Big[q\big(b, \tfrac{7}{10} + \tfrac{3v}{10}\big) + q\big(b, -\tfrac{7}{10} + \tfrac{3v}{10}\big)\Big]. \tag{90}$$

Both conditional radii lie in $[\frac{2}{5}, 1]$ and are generally unequal. Near $b = 0$ the drift is weak because $\partial_b q(0,z) = -k/(34|z|)$ is exactly the witness (84) rescaled by k; the margins below are therefore intrinsic to this response, not artefacts of the certificate.

Proposition 44 (Uniform conditional barriers at $\beta = \frac{3}{10}$)**.** *For every* $|v| \le 1$,

$$G(b,v) \le -\tfrac{11}{2000}, \qquad \tfrac{2}{25} \le b \le \tfrac{7}{40}, \tag{91}$$

$$G(b,v) \ge \tfrac{1501}{2000}, \qquad b \ge \tfrac{5}{8}. \tag{92}$$

Moreover $|q(b,z)| \le \frac{11}{10}$ *for all* $(b,z) \ne 0$*;* $q(b,z) \ge q(\frac{5}{8},-1) > 1/2$ *whenever* $b \ge \frac{5}{8}$ *and* $0 < |z| \le 1$*; and the perturbation in* (85) *changes a projected update by at most* $\frac{1}{2000}$*. Thus, if* $|b_t - u_t| \le \frac{1}{200}$*, the conditional mean of the actual update satisfies*

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] \le -\gamma_- \quad \textit{on } [a,2a], \qquad \mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] \ge \gamma_+ \quad \textit{on } [b_+,\infty), \tag{93}$$

where the intervals refer to u_t*,* $|Y_t| \le c$ *on every path, and*

$$a = \tfrac{17}{200}, \qquad b_+ = \tfrac{63}{100}, \qquad \gamma_- = \tfrac{1}{200}, \qquad \gamma_+ = \tfrac{3}{4}, \qquad c = \tfrac{11}{10} + \tfrac{1}{2000}.$$

Proof. The inner bound is an exact rational interval certificate on $[\frac{2}{25}, \frac{7}{40}] \times [-1,1]$, which we cover by 217 rational boxes. On each box, G is evaluated at the rational centre with enclosed square roots, and the mean-value remainder uses interval enclosures of both partial derivatives from (88); the completed square in (86) supplies a strictly positive interval denominator. Square-root endpoints are rationals with denominator 2^{96} whose enclosing property is checked by exact squaring; every other operation is exact rational arithmetic. Every upper endpoint is at most $-\frac{11}{2000}$; the largest is -0.005501807.

For the outer bound take first $z > 0$: $P(b,z) > 0$ for $b \ge 0$, so by (88) $q(b,z)$ decreases in $b \ge 0$ towards its limit $300/289$. For $z < 0$, zero homogeneity gives $q(b,z) = q(b/|z|,-1)$. The cubic $P(w,-1) = 112w^3 - 188w^2 + 103w - 18$ has no root in $[\frac{5}{8},\infty)$ and is positive at $\frac{5}{8}$, and $300w^3 - 536w^2 + 322w - 65$ is positive at $w = \frac{5}{8}$; Sturm's rule checks both exactly. Hence $q(w,-1)$ increases on $[\frac{5}{8},\infty)$ from $q(\frac{5}{8},-1) = 0.53897\ldots > 1/2$. Since $|z| \le 1$, $b \ge \frac{5}{8}$ implies $b/|z| \ge \frac{5}{8}$, and

$$G(b,v) \ge \tfrac{1}{2}\Big[\tfrac{300}{289} + q(\tfrac{5}{8},-1)\Big] = 0.7885152\ldots > \tfrac{1501}{2000}.$$

The samplewise bound $q^2 \le (\frac{11}{10})^2$ is equivalent to $6171\rho^6 \ge 8100N^2$, with N the cubic numerator in (87). In the variable $w = b/z$ this is a polynomial of degree six with no real root and positive value at $w = 0$ (Sturm's rule); oddness, $q(w,1) = -q(-w,-1)$, covers both noise signs, and $q(b,0) = \pm 300/289$. Numerically $\sup|q| = 1.098$, attained at the interior minimum of $q(\cdot,-1)$.

For the perturbation, write $\Psi_f(P) = \Phi(P/\|P\|_F)$ with $\Phi(Q) = \frac{1}{2}(3Q - QQ^\top Q) = \nabla_Q\big[\frac{3}{4}\|Q\|_F^2 - \frac{1}{8}\operatorname{tr}((Q^\top Q)^2)\big]$. In the singular frame of Q the self-adjoint derivative $D\Phi(Q)$ is block diagonal, with eigenvalues $f'(s_i) = 3(1-s_i^2)/2$ on the diagonal entries and $(3 - s_i^2 - s_j^2 \mp s_i s_j)/2$ on each pair of off-diagonal entries. On the unit Frobenius sphere all of them lie in $[0,3/2]$, so $\|D\Phi(Q)\|_{\mathrm{op}} \le 3/2$, and differentiating the normalization gives $\|D\Psi_f(P)\|_{\mathrm{op}} \le 3/(2\|P\|_F)$. By (86) and (89), $\|M_t - \Delta_t\|_F \ge \sigma|z_t|/\sqrt{51} \ge 2\sigma/(5\sqrt{51}) = 0.05601\ldots\sigma$, so every point of the segment from $M_t - \Delta_t$ to M_t has norm greater than $\sigma/20$, and the projected response changes by at most $\frac{3}{2}(\sigma/60000)/(\sigma/20) = \frac{1}{2000}$. This estimate is uniform in b; the radius is smaller than in (53) because E and S_3 are nearly parallel ($\langle E, S_3\rangle = 30/\sqrt{918}$), so the momentum can nearly cancel.

Finally, $u_t \in [a,2a]$ and $|b_t - u_t| \le \frac{1}{200}$ put $b_t \in [\frac{2}{25}, \frac{7}{40}]$, and $u_t \ge b_+$ gives $b_t \ge \frac{5}{8}$. Since b_t is $\mathcal{F}_{t-1}$-measurable and ϵ_t is independent of $\mathcal{F}_{t-1}$, $\mathbb{E}[q(b_t,z_t) \mid \mathcal{F}_{t-1}] = G(b_t,z_{t-1})$; subtracting the perturbation budget proves (93). □

Theorem 45 (Residence for the increasing and flattening witness)**.** *Use* (83)–(85)*,* $F(W) = \frac{1}{2}\langle E, W\rangle^2$*,* $\beta = \frac{3}{10}$*, constant* $\eta > 0$ *and zero initial momentum. Let* $T > m \ge 1$ *be integers,* $\delta \in (0,1)$*,* $u_0 = x_0/(\sigma k)$*,* $\alpha = \eta/(\sigma k)$*, and with the constants of Proposition 44 put*

$$d = \beta^{m+1}u_0 + \frac{\alpha c\beta}{1-\beta}, \qquad U = u_0 + m\alpha c, \qquad L = \log\frac{4(T+1)}{\delta},$$

$$J = \alpha\Big(c + \frac{Lc^2}{2\gamma_-}\Big), \qquad t_{\mathrm{ent}} = m + \Big\lceil \frac{2U}{\alpha\gamma_+} + \frac{2Lc^2}{\gamma_+^2} \Big\rceil + 1.$$

Suppose

$$d \le \tfrac{1}{200}, \qquad J \le \frac{a}{2}, \qquad u_0 - m\alpha c > 2a, \qquad U \ge b_+ + d + \alpha c, \qquad T \ge t_{\mathrm{ent}}. \tag{94}$$

Then with probability at least $1-\delta$ *there is* $\tau \le t_{\mathrm{ent}}$ *such that*

$$a\sigma k \le \langle E, W_t\rangle \le (b_+ + a/2)\sigma k \qquad \textit{for every } \tau \le t \le T, \tag{95}$$

so the squared population gradient lies between $a^2k^2\sigma^2$ *and* $(b_+ + a/2)^2k^2\sigma^2$ *on this interval. On every path,* $|\langle E, W_t\rangle| \le U\sigma k$ *for all* $t \le T$.

Proof. This is the argument of Theorem 29 with the samplewise bound c in place of 1. On every path $|u_{t+1} - u_t| = \alpha|Y_t| \le \alpha c$, and the exact signal recursion $b_t - u_t = \beta(b_{t-1} - u_{t-1}) + \beta(u_{t-1} - u_t)$ gives $|b_t - u_t| \le \beta^{t+1}u_0 + \alpha c\beta/(1-\beta) \le d$ for $t \ge m$, together with $u_0 - m\alpha c \le u_m \le U$. If $|u_t| \ge b_+ + d$ then $|b_t| \ge \frac{5}{8}$ with the sign of u_t, so by Proposition 44 and oddness $|q(b_t, z_t)| > 1/2$ with that sign, and Y_t keeps it because the perturbation moves the response by at most $\frac{1}{2000}$; hence $|u_{t+1}| \le |u_t|$. Otherwise $|u_{t+1}| \le b_+ + d + \alpha c \le U$. Induction gives $|u_t| \le U$ on every path.

Put $H_t = \mathbb{E}[Y_t \mid \mathcal{F}_{t-1}]$ and $D_t = Y_t - H_t$, ordinary martingale differences for the preceding-step filtration. Conditionally $Y_t \in [-c, c]$, so Hoeffding's lemma gives $\mathbb{E}[e^{\lambda D_t} \mid \mathcal{F}_{t-1}] \le e^{\lambda^2 c^2/2}$. For each starting time $i \in \{m, \dots, T\}$, Ville's inequality for the exponential supermartingales with $\lambda = 2\gamma_-/c^2$, $\lambda = -2\gamma_+/c^2$ and $\lambda = -\gamma_+/c^2$, and a union bound costing at most $3(T+1)e^{-L} \le \delta$, give an event on which every $m \le i < j \le T$, $n = j - i$, satisfies

$$\sum_{t=i}^{j-1} D_t \le \gamma_- n + \frac{Lc^2}{2\gamma_-}, \qquad \sum_{t=i}^{j-1} D_t \ge -\gamma_+ n - \frac{Lc^2}{2\gamma_+}, \qquad \sum_{t=i}^{j-1} D_t \ge -\frac{\gamma_+ n}{2} - \frac{Lc^2}{\gamma_+}. \tag{96}$$

The differences are defined along the actual process at all states, so these inequalities apply to intervals selected by first-exit arguments.

Since $u_m > 2a$, a first exit below a would be preceded by a last entry into $[a, 2a]$ from above, at a state at least $2a - \alpha c$. On the subsequent run inside the interval, $H_t \le -\gamma_-$ by (93), and the first bound in (96) limits the total additional decrease to $\alpha Lc^2/(2\gamma_-)$; the alleged exit state is at least $2a - J \ge 3a/2 > a$, a contradiction. Hence $u_t \ge a$ for $m \le t \le T$. If $u_t > b_+$ for all $m \le t < m + n$, the third bound gives $u_{m+n} \le U - \alpha\gamma_+ n/2 + \alpha Lc^2/\gamma_+$, which is nonpositive for $n = t_{\mathrm{ent}} - m - 1$ and contradicts $u \ge a$; so some $\tau \le t_{\mathrm{ent}}$ has $u_\tau \le b_+$. After entry, a run above b_+ starts at most αc above it, and the second bound limits its further excursion to $\alpha Lc^2/(2\gamma_+) \le \alpha Lc^2/(2\gamma_-)$, so its height is at most $b_+ + J \le b_+ + a/2$. Rescaling by σk proves the theorem. □

An evaluated instance. For $\sigma = 1/16$, $x_0 = 1/4$, $\eta = 10^{-6}$, $T = 10^7$, $\delta = 0.05$ and $m = 5$, so that $u_0 = 2.37635$ and $\alpha = 9.5054 \times 10^{-6}$, the certificate evaluates

$$d = 0.00173685 < 0.005, \qquad J = 0.0236102 < 0.0425, \qquad t_{\mathrm{ent}} = 666\,776 < T.$$

Every quantity is enclosed in rational arithmetic: the logarithm $L = 20.5001\dots$ by an 80-term atanh series after binary argument reduction, the square root in k by the checked enclosure of the drift certificate, and both endpoints of the entry expression give the same integer. The remaining inequalities in (94) hold, and

$$7.99642 \times 10^{-5} \;\le\; \|\nabla F(W_t)\|_F^2 \;\le\; 0.00500544 \qquad (666\,776 \le t \le 10^7) \tag{97}$$

with probability at least 0.95; the exact endpoints are $\frac{4913}{61440000}$ and $\frac{1230137}{245760000}$. The pathwise bound is $|\langle E, W_t\rangle| \le 0.2500055$, so the Huber remark of Appendix F applies verbatim. The step size is ten times smaller than in Corollary 4: the inner margin $\gamma_- = \frac{1}{200}$ is a quarter of the $1/50$ of Proposition 28, and J scales as η/γ_-. At this horizon, $\eta = 2 \times 10^{-6}$ gives $J = 0.04722 > a/2$ and $\eta = 10^{-5}$ gives $J = 0.236$, so both fail the jump budget; at $T = 10^6$ the pair $\eta = 2 \times 10^{-6}$ passes ($J = 0.04192$, entry bound 333 433) and $\eta = 10^{-6}$ has entry bound 666 766. Halving the step to 5×10^{-7} moves the entry bound to 1 333 443, beyond 10^6 but inside 10^7. The certified band is broad compared with the observed concentration: it establishes a persistent nonzero signal, not its location.

The fixed-scale control. Replacing the endogenous norm by a fixed scale s gives $\Phi_s(P) = \Phi(P/s)$, the gradient of a potential whose Hessian is positive semidefinite exactly on $\{\|P\|_{\mathrm{op}} \le s\}$, by

the eigenvalue list above; the Newton–Schulz cubic is not monotone beyond the unit spectral ball. For the stationary noise EMA, with $\mathbb{E}z_\infty^2 = (1-\beta)/(1+\beta) = 7/13$, $\sum_i E_i^2 S_i^2 = 131/153$ and $\sum_i E_i^4 = 267/289$, the frozen field is exactly

$$h_s(x) = \frac{3x}{2s}\Big(1 - \frac{7}{13}\cdot\frac{131}{153}\cdot\frac{\sigma^2}{s^2}\Big) - \frac{267}{289}\cdot\frac{x^3}{2s^3}. \tag{98}$$

Its recovery slope at the origin is positive for every $s \geq \sigma$, whereas the normalized field has $h'_{\mathrm{norm}}(0) = -(1/(34\sigma))\,\mathbb{E}|z_\infty|^{-1} < 0$. At $s = \sigma$, however, h_σ has a positive root at $x = 1.323\,\sigma < x_0 = 4\sigma$: the control started at x_0 lies outside the monotone region and moves away from the origin (in short runs every trajectory leaves $|\langle E, W\rangle| \leq 1$ within 4 254 updates). The admissible control therefore fixes the scale at the pathwise momentum bound $s = x_0 + \sigma = 5\sigma$, for which the root is $8.927\,\sigma > x_0$ and every normalized singular value along the executed runs stays below one (largest 0.973). It is the same cubic at a scale that keeps it monotone.

Executed trajectories. We ran 128 independently seeded trajectories at the certified setting, sharing each noise sequence between the endogenous-scale method and the control, from $W_0 = E/4$ and $M_{-1} = 0$, with 10^7 updates each. Every trajectory is inside the certified band (95) for good after at most 173 878 updates (median 173 845) and first crosses $u \leq b_+$ between updates 178 331 and 178 432, within the entry bound 666 776. The smallest noise EMA magnitude observed is 0.4, the lower end of (89). Over the final 20% of each run the endogenous-scale mean squared population gradient is $4.869 \times 10^{-4} \pm 5.5 \times 10^{-7}$ (95% interval across trajectory means). The stationary Bernoulli convolution with all 2^{14} sign histories, computed without trajectory fitting, has a positive field root at $u_* = 0.2098$, that is $x_* = 0.02208$ and $x_*^2 = 4.874 \times 10^{-4}$, inside the certified band and within 0.085% of the measured tail mean. The fixed-scale control's tail mean is 8.32×10^{-9}, a factor 58 523 smaller; as in Appendix F, its constant-step floor is expected to scale with the step size. Along perturbed replayed paths at the admissible radius, the full-matrix response deviates from $q(b_t, z_t)$ by at most 9.1×10^{-7}, well inside the $\frac{1}{2000}$ budget. The stationary root and the tail statistics are numerical observations; the residence statement is Theorem 45.

J.2 Visible high-momentum obstructions at larger rank

The negative high-momentum search above was restricted to ranks at most eight and classical Newton–Schulz depths at most three. Increasing rank changes the balance: one dominant singular coordinate can shrink many unsaturated minor coordinates under Frobenius normalization, while their combined contribution to the signal update grows with their multiplicity. We retain the strictly increasing, ratio-flattening response $f_K = p^{\circ K}$ of Lemma 41 and the same unbiased iid Rademacher innovation law. A wider exploratory search over $r \in \{3, 8, 32, 128, 512\}$ and $K \in \{1, 3, 5, 8, 12\}$ first used a Gaussian approximation to the standardized high-momentum EMA. All 25 search cells are archived, including all positive minima; those approximate searches are not the reported finite-β evidence.

Fixed simple spectra and the actual noise law. For $n = r - 1$, let

$$S = \mathrm{diag}(s_0, s_1/\sqrt{n}, \ldots, s_1/\sqrt{n}), \quad s_0^2 + s_1^2 = 1,$$
$$E = \mathrm{diag}(e_0, e_1/\sqrt{n}, \ldots, e_1/\sqrt{n}), \quad e_0^2 + e_1^2 = 1.$$

We fixed $(s_0, s_1) = (1,1)/\sqrt{2}$, $(e_0, e_1) = (3,1)/\sqrt{10}$ at $r = 128, 512$, $K = 5$, and $(s_0, s_1) = (4,3)/5$, $(e_0, e_1) = (4,1)/\sqrt{17}$ at $r = 512$, $K = 8$. The task is still $F(W) = \langle E, W\rangle^2/2$, with $\xi_t = \sigma\epsilon_t S$, $\sigma = 1/16$, $x_0 = 1/4$, zero signal and noise EMAs, and batch size one. Thus $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$ and $\|\xi_t\|_F = \sigma$ exactly. No temporally dependent innovations or imposed nonzero-radius EMA are used.

Write $s_\beta = \sigma\sqrt{(1-\beta)/(1+\beta)}$, $y = x/s_\beta$, $z = \sigma\zeta/s_\beta$ and $h = \eta/s_\beta$. Then z has unit stationary second moment. The scalar matrix reduction is exact:

$$u = ye_0 + zs_0, \qquad v = ye_1 + zs_1, \qquad R = \sqrt{u^2 + v^2},$$
$$q(y,z) = e_0 f_K(u/R) + \sqrt{n}\, e_1 f_K(v/(\sqrt{n}R)),$$
$$H(y) = \mathbb{E}q(y,z) = \tfrac{1}{2}\mathbb{E}[q(y,z) + q(y,-z)].$$

New stationary EMA banks, using the actual $\beta \in \{.9, .95\}$ and Rademacher innovations, give the fields and roots in Table 9. The $K = 5$, rank-128 roots are about .078 and .077 in units of s_β;

the rank-512, $K = 8$ roots are 0.3065 and 0.3020. Eight banks of 2^{19} noise samples are retained, including each bank's root and slope. These are numerical root estimates, not interval-certified roots. The earlier rank-three $K = 1$ witness remains positive on this grid. Dense matrix polynomial evaluations in independent left and right orthogonal frames agree with the reduction within 2.7e-15 at ranks 128 and 512.

Proposition 46 (Statistical bracket for the high-momentum averaged field)**.** *For the fixed rank-512, $K = 8$ spectra above, the independent noise-law evaluation bounds the true stationary fields at both $\beta = .9$ and $\beta = .95$ by*

$$H(.1) \le -0.0620 < 0, \qquad H(.5) \ge 0.1740 > 0$$

with simultaneous confidence at least $1 - 10^{-6}$. Consequently each field has a positive root in $(.1, .5)$ at that confidence level. The probability concerns the noise-law evaluation, not a trajectory residence event.

Proof. Generate $N = 524288$ independent innovation histories per β, with lengths $L = 256$ and 512, respectively, and let z_L be their zero-initialized normalized EMA. For $\varphi_y(z) = [q(y, z)+q(y, -z)]/2$, $|\varphi_y| \le B = e_0 + \sqrt{n}e_1$. Hoeffding's inequality and a union bound for the four stated signs give a simultaneous sampling radius $B\sqrt{2\log(8/\delta)/N}$, with $\delta = 10^{-6}$. Under stationary coupling, $|z - z_L| \le \beta^L/\sqrt{(1-\beta)/(1+\beta)}$. Because $\|yE+zS\|_F \ge y|s_0e_1 - s_1e_0|$ and $\sup|f'_K| \le (3/2)^K$, the scalar reduction has a global Lipschitz constant in z bounded by $(3/2)^K/[y|s_0e_1-s_1e_0|]$. Multiplying these bounds accounts deterministically for the stationary tail. The sampling radius is .05025; the largest tail allowance is below 1.7×10^{-8}. The measured negative values are $-.11347, -.11228$ and positive values .22429, .23538, giving the displayed bounds. Continuity in $y > 0$ gives the root bracket. As in the other computational studies, the stated sampling confidence models the separately seeded pseudorandom histories as iid draws; it is not a directed-rounding certificate. □

Paired trajectories and the five-step bridge. Before optimization runs, we sealed the root, slope, Green–Kubo variance and entry forecasts for the rank-512, $K = 8$ pair (sha256 prefix `ecf694650cb9`); the rank-128, $K = 5$, $\beta = .95$ bridge was sealed separately. Each cell uses 64 independently seeded trajectories, with identical innovations shared between endogenous normalization and a fixed scale of 8σ. The fixed-scale arguments remain below 0.5737, so every application of the classical cubic stays in its monotone range. The step sizes are $h \in \{3 \cdot 10^{-5}, 10^{-5}, 3 \cdot 10^{-6}\}$ for $K = 8$ and $h \in \{10^{-5}, 3 \cdot 10^{-6}, 10^{-6}\}$ for the $K = 5$ bridge. The RMS forecast is $\sqrt{y_*^2 + h\Gamma(y_*)/(2H'(y_*))}$; it omits the nonlinear mean shift and has no asserted finite-h remainder bound.

The first windows, of slow-time length at least the larger of four predicted entry times and $24/H'(y_*)$, already showed endogenous levels close to the predicted roots. They also left a visible fixed-scale-control transient: at the smallest step the control RMS was about .019, .032 and .092 in the three settings. Those complete results remain archived. After observing that transient, we fixed a uniform fourfold horizon extension, preserving every root, step, seed and prediction. Table 10 and Figure 4 report the extended windows; this extension is a disclosed horizon sensitivity check, not an additional prospective prediction claim. Tail statistics average the last half of each run. The endogenous levels remain near nonzero roots as h decreases, whereas the fixed-scale level decreases. All cases with tail sign changes, including the intermediate step of the five-step bridge, remain in the table; no stationary positive-branch law is inferred from their RMS. In particular, the bridge uses five classical cubic iterations and high momentum, while retaining the full matrix normalization mechanism and a fixed-scale control.

Scope. These examples close the earlier empirical absence of a visible high-momentum instance for an increasing, flattening response. They do not extend the rational-interval trajectory residence certificate of Theorem 45 to high momentum. The confidence bracket proves only signs of a noise-averaged field with statistical confidence; the actual high-momentum paths are finite numerical runs. Resolving the weaker five-step plateau uses $h = 10^{-6}$, corresponding to $\eta \approx 10^{-8}$, and the longest run has about 2.97×10^8 updates per trajectory. These runs do not establish a practical-step-size residence certificate. The construction has a deliberately anisotropic diagonal spectrum and a one-dimensional noise subspace, and is not a model of transformer gradient noise. The original rank-three negative searches and tiny-field-root limitations remain valid within their stated ranges.

Table 9: Actual Rademacher-EMA fields after the disclosed search. Signals and roots are in units of the stationary noise RMS s_β. "None on grid" refers to $y \in [.005, 2]$, not absence of every positive root.

Rank	K	β	Smallest grid H	Positive root y_*
3	1	0.90	+0.0035	none on grid
128	5	0.90	-0.0142	0.0784
512	5	0.90	-0.0221	0.1110
512	8	0.90	-0.1135	0.3065
3	1	0.95	+0.0035	none on grid
128	5	0.95	-0.0139	0.0766
512	5	0.95	-0.0218	0.1090
512	8	0.95	-0.1128	0.3020

Table 10: Extended high-momentum runs. RMS levels are divided by s_β; "Tail crossings" is the fraction of trajectories with a sign change in the last half. All cells are retained. The forecast adds leading linear-response variance to the root and does not include a nonlinear mean-shift correction.

r/K	β	h	y_*	RMS forecast	RMS observed	Fixed RMS	Tail crossings
512/8	0.90	3e-05	0.3065	0.3124	0.3077	0.0125	0.062
512/8	0.90	1e-05	0.3065	0.3085	0.3083	0.0072	0.000
512/8	0.90	3e-06	0.3065	0.3071	0.3067	0.0040	0.000
512/8	0.95	3e-05	0.3020	0.3132	0.3017	0.0153	0.781
512/8	0.95	1e-05	0.3020	0.3058	0.3022	0.0090	0.000
512/8	0.95	3e-06	0.3020	0.3031	0.3017	0.0050	0.000
128/5	0.95	1e-05	0.0766	0.0867	0.0834	0.0049	1.000
128/5	0.95	3e-06	0.0766	0.0798	0.0768	0.0027	0.891
128/5	0.95	1e-06	0.0766	0.0777	0.0770	0.0015	0.016

K Quantitative forecasts from the noise law

We use the transverse task to test how much the noise distribution can tell us about the optimization trajectory. First, the average update predicts where the signal stops moving on average. A linear filter then predicts fluctuations around that level, including the correlation introduced by momentum. Both calculations are performed before the optimization runs. We distinguish the limiting theorem, the approximation at a finite step size, and the measurements used to test that approximation.

K.1 Dimensionless dynamics and the Green–Kubo coefficient

Put $y_t = x_t/s$, $b_t = a_t/s$, $z_t = Z_t/s$ and $h = \eta/s$. While $|x_t| \le 1$, the exact recursion is

$$b_t = \beta b_{t-1} + (1-\beta) y_t, \qquad z_t = \beta z_{t-1} + (1-\beta)\bar{\xi}_t/s, \qquad y_{t+1} = y_t - h g_K(b_t, z_t), \tag{99}$$

where $g_K(y, z) = q_K(y/\|z\|_F)$, with its continuous extension in z at $z = 0$ for $y > 0$. Let π be the stationary law of z_t, let $r_t = \|z_t\|_F$, and define

$$H(y) = \int g_K(y, z)\, \pi(dz) = h_K(sy), \tag{100}$$

$$\Gamma(y) = \sum_{j \in \mathbb{Z}} \mathrm{Cov}_\pi\big(g_K(y, z_0), g_K(y, z_j)\big). \tag{101}$$

The two-sided covariance sum is defined using a stationary extension. It equals $\mathrm{Var}_\pi(g_K(y, z_0)) + 2\sum_{j \ge 1} \mathrm{Cov}_\pi(g_K(y, z_0), g_K(y, z_j))$. The normalized noise satisfies $\mathbb{E}_\pi r_t^2 = 1$. Its distribution still depends on both B and β; matching s alone does not identify H or Γ.

Theorem 47 (Averaging, fluctuations and entry prediction)**.** *Fix B, $0 \le \beta < 1$, $\sigma > 0$ and a finite polynomial depth K. Let I be a compact subinterval of $(0, 1/s)$, and let $\bar{y}'(u) = -H(\bar{y}(u))$, $\bar{y}(0) = y_0 \in I$. Consider any fixed slow-time horizon S on which $\bar{y}$ remains in the interior of I. For the exact recursion from zero EMA, stopped on leaving a slightly larger positive compact interval, as*

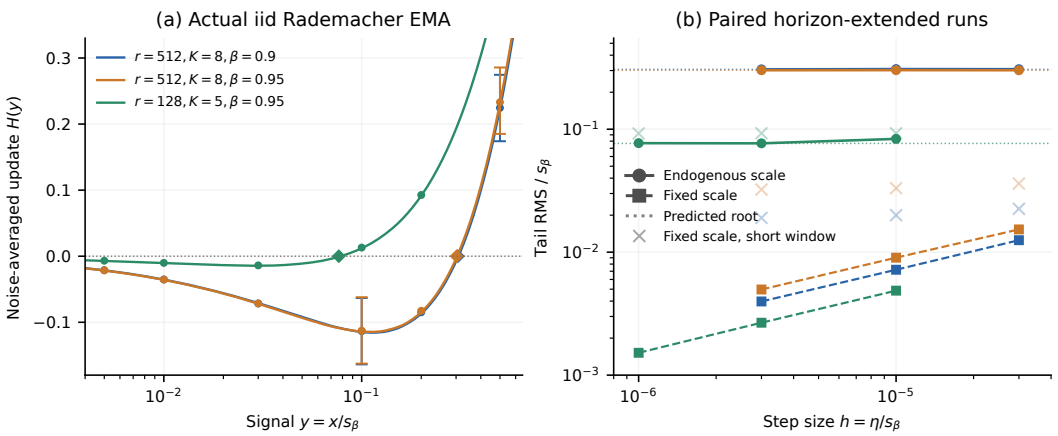


Figure 4: High-momentum increasing and flattening responses. Left: actual Rademacher-EMA fields from independent noise banks; dots at positive roots are numerical estimates, while the error bars at .1 and .5 give the simultaneous Hoeffding-plus-tail brackets for rank 512, $K = 8$. Right: observed tail RMS versus step size in the extended windows, with horizontal root predictions. Circles use endogenous normalization and squares the matched fixed scale. Faint crosses show the fixed-scale levels in the original shorter windows, exposing the control transient.

$h \downarrow 0$,

$$\sup_{0\le u\le S} |y_{\lfloor u/h \rfloor} - \bar{y}(u)| \longrightarrow 0 \quad \textit{in probability}, \qquad \frac{y_{\lfloor u/h \rfloor} - \bar{y}(u)}{\sqrt{h}} \Longrightarrow U(u), \tag{102}$$

where the second convergence is in $D([0,S])$ with the Skorokhod topology and

$$dU(u) = -H'(\bar{y}(u))U(u)\,du + \sqrt{\Gamma(\bar{y}(u))}\,d\mathcal{B}(u), \qquad U(0) = 0.$$

The stopping probability before S tends to zero. If $y_0 > c > 0$ and $H(y) > 0$ on $[c, y_0]$, let $\tau_h = \inf\{t : y_t \le c\}$. Then

$$\frac{h\tau_h - \displaystyle\int_c^{y_0} \frac{dy}{H(y)}}{\sqrt{h}} \Longrightarrow \mathcal{N}\left(0, \int_c^{y_0} \frac{\Gamma(y)}{H(y)^3}\,dy\right). \tag{103}$$

If $y_* > 0$ satisfies $H(y_*) = 0$ and $\kappa_* = H'(y_*) > 0$, then for $y_0 = y_*$ the limiting fluctuation process is an Ornstein–Uhlenbeck process. Its equilibrium variance is $\Gamma(y_*)/(2\kappa_*)$.

The last assertion takes $h \downarrow 0$ first and the slow time to infinity afterward. It does not assert an invariant probability law supported on the positive branch for a fixed nonzero h, nor does it exclude eventual sign changes. The theorem concerns fixed B, β; it is not uniform as $\beta \uparrow 1$.

Proof. The noise transition is a contraction under synchronous coupling: $\|z_t - z'_t\|_F = \beta^t \|z_0 - z'_0\|_F$. Its support is bounded, so it has a unique stationary law and is ergodic. On any positive compact interval of y, the function $g_K(y,z)$ and all derivatives needed below are bounded and Lipschitz in z. To check the apparently singular point $z = 0$, the explicit formula is

$$g_K(y,z) = \frac{1}{\sqrt{2}}\left[f_K\left(\frac{r+y}{\sqrt{2(y^2+r^2)}}\right) + f_K\left(\frac{y-r}{\sqrt{2(y^2+r^2)}}\right)\right], \qquad r = \|z\|_F.$$

The numerator is even in r because f_K is odd. For a polynomial f_K, this is a smooth function of y and r^2 whenever $y > 0$. In particular no inverse-radius moment is used in this argument.

Write P for the noise transition. The contraction and the preceding Lipschitz bounds imply uniform geometric convergence of $P^j g_K(y,\cdot)$ to $H(y)$, also after the required y derivatives. Consequently

$$G_y(z) = \sum_{j=0}^{\infty} \left[P^j g_K(y,z) - H(y)\right]$$

is bounded, differentiable in y, and solves $G_y - PG_y = g_K(y,\cdot) - H(y)$. The covariance series in (101) is absolutely summable. For fixed y, the centred noise sum has the decomposition into bounded martingale differences

$$D_t(y) = G_y(z_t) - PG_y(z_{t-1})$$

and a telescoping boundary term. Its stationary variance per step is

$$\mathbb{E}_\pi D_t(y)^2 = \lim_{n\to\infty} \frac{1}{n} \operatorname{Var}_\pi \left(\sum_{t=1}^{n} g_K(y, z_t) \right) = \Gamma(y) \geq 0.$$

With y_t varying by at most $h \sup |q_K|$ per step, summation by parts in this Poisson decomposition contributes a bounded boundary term and a remainder $O(hn)$ before stopping. After multiplying by the update step h, both contribute $O(h)$ on $n \leq S/h$; after division by $\sqrt{h}$ they vanish. The signal-EMA lag has the pathwise bound

$$|b_t - y_t| \leq \beta^{t+1}|b_{-1} - y_0| + \frac{h\beta \sup |q_K|}{1-\beta}.$$

The initialization term has a summable geometric tail. On positive compact intervals its accumulated effect is $O(h)$, as is the accumulated effect of the second term. A finite initial segment with b_t outside the interval is harmless because the update itself is bounded; after that segment b_t is in a positive compact interval. Starting the noise EMA at zero likewise changes the accumulated drift by $O(h)$ under synchronous coupling with stationarity. These bounds are at fixed β.

The martingale maximal inequality and the discrete Gronwall inequality now give the averaging statement. Apply the same Poisson decomposition to the conditional quadratic variation; ergodicity and boundedness give its limit $\int_0^u \Gamma(\bar{y}(v))\, dv$. The bounded martingale increments satisfy the Lindeberg condition. The martingale functional central limit theorem and the first-order Taylor expansion of H give the displayed linear stochastic differential equation. Compact containment follows from the averaging statement because the deterministic solution stays a positive distance from the stopping boundary.

For entry, the deterministic crossing speed is $-H(c) < 0$, so the hitting-time delta method gives variance $\operatorname{Var}(U(u_c))/H(c)^2$, where $u_c = \int_c^{y_0} H(y)^{-1} dy$. Solving the linear variance equation $V' = -2H'(\bar{y})V + \Gamma(\bar{y})$ yields

$$\frac{V(u_c)}{H(c)^2} = \int_c^{y_0} \frac{\Gamma(y)}{H(y)^3}\, dy.$$

This proves (103). At y_* the coefficients are constant, and $V(u) = \Gamma(y_*)[1 - e^{-2\kappa_* u}]/(2\kappa_*)$. □

K.2 Retaining the finite-step effect of momentum

The leading diffusion limit omits effects proportional to the step size divided by the memory relaxation rate. They are substantial in the reported experiments. Retain the exact signal EMA and drive its linearization by the stationary, *noise-only* forcing $\zeta_t = g_K(y_*, z_t) - H(y_*)$:

$$A_t = \beta A_{t-1} + (1-\beta)Y_t, \qquad Y_{t+1} = Y_t - h(\kappa_* A_t + \zeta_t). \tag{104}$$

The force is always evaluated at the quadrature-defined y_*, not at the evolving optimizer state. This is a linear response model, not a rerun of (99).

Proposition 48 (Exact covariance of the linear response)**.** *Suppose $\kappa_* > 0$. The homogeneous filter* (104) *is Schur stable exactly when*

$$0 < h\kappa_* < \frac{2(1+\beta)}{1-\beta}. \tag{105}$$

Let $S_\zeta(\omega) = \sum_{j\in\mathbb{Z}} \operatorname{Cov}(\zeta_0, \zeta_j) e^{-ij\omega}$, with the convention $\operatorname{Var}(\zeta_t) = (2\pi)^{-1} \int_{-\pi}^{\pi} S_\zeta(\omega) d\omega$. Under (105)*, the stationary filter variances are*

$$V_h := \operatorname{Var}(Y_t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{h^2 |1 - \beta e^{-i\omega}|^2 S_\zeta(\omega)}{|1 - [1 + \beta - h\kappa_*(1-\beta)]e^{-i\omega} + \beta e^{-2i\omega}|^2}\, d\omega, \tag{106}$$

$$\operatorname{Var}(A_t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{h^2 (1-\beta)^2 S_\zeta(\omega)}{|1 - [1 + \beta - h\kappa_*(1-\beta)]e^{-i\omega} + \beta e^{-2i\omega}|^2}\, d\omega. \tag{107}$$

At fixed β, both variances divided by h converge to $\Gamma(y_*)/(2\kappa_*)$ as $h \downarrow 0$.

Proof. For the state (Y_t, A_{t-1}), the homogeneous transition is

$$L_h = \begin{pmatrix} 1 - h\kappa_*(1-\beta) & -h\kappa_*\beta \\ 1-\beta & \beta \end{pmatrix}.$$

Its determinant is β and trace is $1 + \beta - h\kappa_*(1-\beta)$. The degree-two Schur criterion gives (105). Eliminating A_t gives

$$Y_{t+1} = [1 + \beta - h\kappa_*(1-\beta)]Y_t - \beta Y_{t-1} - h\zeta_t + h\beta\zeta_{t-1}.$$

The transfer function and the EMA relation give the two integrals. Absolute summability of the noise covariances makes S_ζ continuous. The pole near one has decay rate $h\kappa_* + O(h^2)$, while the other tends to β. Rescaling $\omega = hu$ in a neighborhood of zero gives the limit

$$\frac{1}{2\pi}\int_{\mathbb{R}} \frac{\Gamma(y_*)}{\kappa_*^2 + u^2}\, du = \frac{\Gamma(y_*)}{2\kappa_*}.$$

The integral away from that neighborhood is $o(h)$, by stability and boundedness of the spectrum, proving the variance limits. □

As a check, for hypothetical white forcing of variance v_ζ, the first integral simplifies to

$$V_h = \frac{v_\zeta h}{2\kappa_*} \frac{1 + 2\beta h\kappa_*/(1-\beta^2)}{1 - h\kappa_*(1-\beta)/(2(1+\beta))}. \tag{108}$$

The actual force is colored and the full integral is used in the forecasts. The formula isolates why replacing memory by the noise scale alone loses a finite-step amplification of fluctuations.

K.3 What is predicted, and what is not certified

The proposed level forecast is

$$\widehat{\mathbb{E}\|\nabla F\|_F^2}/s^2 = y_*^2 + V_h, \qquad \chi_{\mathrm{lin}} = \frac{\sqrt{\mathrm{Var}(A_t)}}{y_*}. \tag{109}$$

The equilibrium-only forecast is y_*^2. The ratio χ_{lin} compares the signal-EMA fluctuations to the distance from the expansion point to the singular branch at zero. We fixed $\chi_{\mathrm{lin}} \le 1/2$ as the local regime before the holdout runs. This is an approximation indicator, not a rigorous error bound or a sufficient condition for nonlinear stability; cases outside it are retained.

Theorem 47 justifies the leading local fluctuation law. Equation (109) is a finite-h Gaussian approximation. In particular, state–noise correlation can shift the local mean by $O(h)$; this is the same order as the leading variance, so the theorem does not prove that $y_*^2 + V_h$ is a complete first-order expansion of the nonlinear second moment. No finite-h remainder guarantee or interval-arithmetic certification is claimed for the numerical forecast. Its accuracy is assessed prospectively below.

For entry we use $c = 1$. The deterministic averaged-EMA recursion $\bar{b}_t = \beta\bar{b}_{t-1} + (1-\beta)\bar{y}_t$, $\bar{y}_{t+1} = \bar{y}_t - hH(\bar{b}_t)$, initialized at $(\bar{y}_0, \bar{b}_{-1}) = (x_0/s, 0)$, supplies a discrete correction to the leading integral in (103). Its first crossing of one predicts the median entry step. The asymptotic step-standard-deviation is $[h^{-1}\int_1^{y_0} \Gamma(y)H(y)^{-3}dy]^{1/2}$.

K.4 Locked forecasts and independent matrix validation

The forecasts are computed independently of the optimization trajectories. We evaluate the stationary-noise characteristic function using a truncated product of sinc functions, obtain the coordinate density by FFT, and integrate the product of coordinate densities in polar coordinates. An angle-adapted quadrature then resolves the rapidly varying depth-five response. Doubling the radial, angular and response resolutions must change both y_* and κ_* by less than 0.1%. These are floating-point convergence checks, not validated interval enclosures.

For each of the four noise laws, eight independent stationary noise banks each have 2^{20} samples after a $40/(1-\beta)$ burn-in. Filtering $g_K(y_*, z_t)$ through (104) estimates the exact linear-filter variance without estimating a high-dimensional transition kernel. Bank-to-bank standard errors quantify

Table 11: Prospective noise-only forecasts at held-out step sizes. The level is the tail mean of $\|\nabla F\|^2/s^2$. A star marks $\chi > 1/2$, outside the stated local approximation regime. Forecasts were locked before the independent matrix runs.

B	β	K	h	Predicted	Observed	Error	Entry pred./obs.
1	0.5	1	0.008	0.15761	0.15734	0.2%	961/961
1	0.5	1	0.03	0.15932	0.15823	0.7%	256/256
1	0.5	1	0.12	0.16744	0.16073	4.2%	63/63
1	0.5	5	0.008	0.00236	0.00279	15.5%	1048/1048
1	0.5	5	0.03	0.00244	0.00233	4.8%	279/279
1	0.5	5	0.12	0.00288	0.00257	12.2%	69/69
1	0.9	1	0.008	0.12545	0.12526	0.2%	2409/2409
1	0.9	1	0.03	0.13334	0.13318	0.1%	638/638
1	0.9	1	0.12*	0.18341	0.17700	3.6%	158/158
1	0.9	5	0.008	0.00191	0.00188	1.4%	2522/2522
1	0.9	5	0.03	0.00222	0.00207	6.8%	669/669
1	0.9	5	0.12*	0.00456	0.00463	1.5%	166/165
1	0.99	1	0.008	0.15685	0.15063	4.1%	7765/7762
1	0.99	1	0.03*	0.36829	0.40489	9.0%	2066/2064
1	0.99	1	0.12*	2.28798	2.05640	11.3%	517/516
1	0.99	5	0.008*	0.00306	0.00288	6.5%	7883/7880
1	0.99	5	0.03*	0.01124	0.01203	6.6%	2089/2087
1	0.99	5	0.12*	0.08600	0.52393	83.6%	520/518
4	0.9	1	0.008	0.12274	0.12231	0.4%	4822/4820
4	0.9	1	0.03	0.13116	0.13038	0.6%	1282/1281
4	0.9	1	0.12*	0.18419	0.17604	4.6%	319/319
4	0.9	5	0.008	0.00187	0.00185	1.1%	4939/4940
4	0.9	5	0.03	0.00219	0.00203	7.6%	1313/1313
4	0.9	5	0.12*	0.00461	0.00478	3.5%	327/326

numerical sampling uncertainty. For the entry CLT, the Green–Kubo integral is estimated by noise-only batch sums with block length the next power of two above $64/(1-\beta)$. All numerical inputs, sampling rules and predictions are fixed before validation.

The holdout grid crosses $(B,\beta) \in \{(1,.5),(1,.9),(1,.99),(4,.9)\}$, $K \in \{1,5\}$ and $h \in \{.008,.03,.12\}$. These step sizes differ from the exploratory grid $\{.001,.004,.015,.06,.19\}$; exploratory results are not counted as holdout evidence. All 24 cells are retained. Each uses 256 independent trajectories, zero momentum initialization and the original fixed task. The implementation updates the four entries of the actual 2×2 momentum matrix and applies the matrix polynomial by multiplication, rather than using the forecasting filter or the scalar projection formula. Validation seeds are disjoint from the noise-bank seeds. The horizon is the larger of 12,000 steps and four times the predicted entry; the tail statistic uses its last half. Confidence intervals use variation across independent trajectories, not correlated updates within one trajectory.

The primary holdout has median level error 1.38% over the 15 cases satisfying $\chi_{\mathrm{lin}} \le 1/2$, with maximum 15.53%. A uniform independent replication with 4,096 new trajectories in *every* cell, retaining the same locked forecasts and horizons, gives median 1.35% and maximum 13.85% in those 15 cases. Across all 24 cells, the primary median error is 4.15%, compared with 12.09% for the equilibrium-only forecast. In the largest failure, $K=5$, $\beta=.99$ and $h=.12$, the predicted level is 0.0860 while the primary observed level is 0.5239; $\chi_{\mathrm{lin}}=1.55$ flags departure from the local regime. The precision replication retains this failure. The locality indicator does not certify nonlinear error: even inside the local regime, a few late transients can affect the finite-run second moment, as the broad primary interval at $K=5$, $\beta=.5$, $h=.008$ shows.

The full results, including failures of the local approximation, are in Table 11. The forecast remains unchanged during validation. The comparison tests quantitative entry and gradient-level predictions at the tested horizons; it is separate from the simultaneous residence certificate of Theorem 31.

Table 12: Independent confirmation of the time-dependent tangent forecast. Entries are the 10%, 50% and 90% first-entry quantiles, in updates. All 12 new step-size settings are retained.

β	K	h	Predicted quantiles	Observed quantiles
0.9	1	0.01	1912/1925/1944	1912/1926/1945
0.9	1	0.04	474/477/484	474/477/484
0.9	1	0.1	189/189/191	189/189/191
0.9	5	0.01	2002/2015/2036	2002/2016/2037
0.9	5	0.04	496/500/507	496/500/506
0.9	5	0.1	197/199/202	197/199/202
0.99	1	0.01	6197/6205/6223	6197/6205/6222
0.99	1	0.04	1547/1548/1549	1547/1548/1549
0.99	1	0.1	619/619/619	619/619/619
0.99	5	0.01	6276/6299/6325	6276/6299/6326
0.99	5	0.04	1554/1561/1572	1554/1561/1572
0.99	5	0.1	620/621/623	620/621/624

K.5 Finite-step entry quantiles: independent confirmation

The leading CLT in (103) predicts the correct limiting dispersion, but it need not predict finite-step dispersion when memory is long. For example, at $\beta = .99$, $K = 1$ and $h = .12$ it predicts a roughly 33-update 10–90% span, whereas the primary matrix experiment has both quantiles at update 516. The deterministic averaged-EMA median remains accurate. At this point the signal EMA still exceeds the current signal substantially, making the actual forcing less sensitive to noise than the instantaneous-state CLT presumes.

We therefore tested the time-dependent version of the same noise-only filter. First compute the deterministic path $(\bar{y}_t, \bar{b}_t)$ defined above, and set

$$\begin{aligned} \mathcal{A}_t &= \beta \mathcal{A}_{t-1} + (1-\beta)\mathcal{Y}_t, \\ \mathcal{Y}_{t+1} &= \mathcal{Y}_t - h\big[H'(\bar{b}_t)\mathcal{A}_t + g_K(\bar{b}_t, z_t) - H(\bar{b}_t)\big], \qquad \mathcal{A}_{-1} = \mathcal{Y}_0 = 0. \end{aligned} \tag{110}$$

The forecast entry time is the first t with $\bar{y}_t + \mathcal{Y}_t \le 1$. The noise process starts at zero, retaining the actual cold-start forcing. Neither g_K nor its evaluation point uses the perturbed optimizer state: the evaluation point $\bar{b}_t$ is determined entirely by the averaged recursion. Thus this is a time-dependent linear response calculation, not an exact nonlinear simulation. As in (104), replacing the random derivative by its average leaves a finite-step approximation; we make no certified nonlinear remainder claim for it.

This refinement was formulated after the primary study, so it received a separate confirmation rather than being retroactively inserted into the original forecast. The new grid uses $B = 1$, $\beta \in \{.9, .99\}$, $K \in \{1, 5\}$ and previously unused $h \in \{.01, .04, .10\}$. For each of all 12 cells, 4,096 independent noise-only paths produce the 10%, 50% and 90% entry quantiles. The forecasts and validation seeds are fixed before the matrix holdout runs. The confirmation then uses 4,096 independent matrix trajectories per cell. All 36 predicted quantiles are within one update of their observed counterparts; the median absolute error is zero. No entry is censored in either the forecast or the matrix confirmation. Table 12 reports every cell. This confirms that retaining the signal lag improves a quantitative distributional prediction, beyond recovering the sign of local instability.

K.6 The stationary mean shift and the closure terms it requires

The approximation (109) accounts for fluctuations around y_*, but the mean itself can also move when the step size is nonzero. We estimate this shift using the correlation between the local slope and the momentum state, together with the average curvature. This gives a candidate first-order correction. To turn it into a theorem, one must also control the difference between the nonlinear process and the linear filter. The exact identity below lists the errors that still need such bounds.

Write $\zeta_t = g_K(y_*, z_t)$, $D_t = \partial_y g_K(y_*, z_t)$ and $J_t = \partial_y^2 g_K(y_*, z_t)$. Thus $\mathbb{E}_\pi \zeta_t = 0$, $\kappa_* = \mathbb{E}_\pi D_t > 0$ and $H''(y_*) = \mathbb{E}_\pi J_t$. Let (Y_t, A_t) be the stationary linear filter (104), with $V_h = \mathrm{Var}(Y_t)$ and $V_A = \mathrm{Var}(A_t)$, driven by the same noise path as the nonlinear recursion. Its EMA impulse response

satisfies

$$A_t = -h \sum_{j \geq 1} w_j \zeta_{t-j}, \qquad w_1 = 1 - \beta, \qquad w_j \longrightarrow 1 - \beta^j \quad (h \downarrow 0 \text{ at fixed } j). \tag{111}$$

The numerical candidate used in the archived forecasts is

$$\delta_h = -\frac{C_h + \frac{1}{2} H''(y_*) V_A}{\kappa_*}, \qquad C_h = \mathrm{Cov}(D_t, A_t). \tag{112}$$

Proposition 49 (Exact residual of the stationary mean-shift closure)**.** *Suppose the nonlinear recursion* (99) *and the Schur-stable linear filter admit a joint stationary realization driven by the same stationary noise chain. Put* $e_t = b_t - y_*$, $\mu = \mathbb{E} e_t = \mathbb{E} y_t - y_*$, $V_b = \mathrm{Var}(b_t)$, $V_y = \mathrm{Var}(y_t)$ *and* $C_b = \mathrm{Cov}(D_t, e_t)$. *Assume the displayed expectations exist, and define the exact Taylor residual*

$$R_3 = \mathbb{E}\left[g_K(y_* + e_t, z_t) - \zeta_t - D_t e_t - \tfrac{1}{2} J_t e_t^2 \right].$$

Then

$$\begin{aligned} \kappa_*(\mu - \delta_h) = & -(C_b - C_h) - \tfrac{1}{2} H''(y_*)(V_b - V_A) - \tfrac{1}{2} H''(y_*)\mu^2 \\ & - \tfrac{1}{2} \mathrm{Cov}(J_t, e_t^2) - R_3, \end{aligned} \tag{113}$$

$$\mathbb{E} y_t^2 - \left(y_*^2 + V_h + 2 y_* \delta_h \right) = (V_y - V_h) + 2 y_* (\mu - \delta_h) + \mu^2. \tag{114}$$

If the segment between y_* *and* b_t *lies in a fixed positive compact interval on which* $|\partial_y^3 g_K| \leq L_3$, *then* $|R_3| \leq L_3 \mathbb{E}|e_t|^3/6$.

Proof. Stationarity gives $\mathbb{E} g_K(b_t, z_t) = 0$ and $\mathbb{E} b_t = \mathbb{E} y_t$. The definition of R_3 gives

$$0 = \kappa_* \mu + C_b + \tfrac{1}{2} H''(y_*)(V_b + \mu^2) + \tfrac{1}{2} \mathrm{Cov}(J_t, e_t^2) + R_3.$$

Subtract $\kappa_* \delta_h = -C_h - H''(y_*) V_A/2$ to obtain (113). The identity $\mathbb{E} y_t^2 = (y_* + \mu)^2 + V_y$ gives (114), and Taylor's integral remainder gives the last bound. □

The sealed test contains 16 cells. Its corrected median error is 0.6% in the 10 cells flagged local before validation, 2.1% over all 16 cells, and its maximum error is 57.7%. The locality flag is not an error certificate. These summary populations are kept distinct throughout the interpretation.

What remains unproved. For the candidate $y_*^2 + V_h + 2 y_* \delta_h$ to be a complete first-order expansion, the covariance and variance closure errors in (113)–(114) must be $o(h)$, together with μ^2 and R_3. Moment scaling $\mathbb{E}|e_t|^k = O(h^{k/2})$ and a bounded third derivative make $R_3 = O(h^{3/2})$, but do not by themselves control $C_b - C_h$, $V_b - V_A$, $V_y - V_h$ or $\mathrm{Cov}(J_t, e_t^2)$ at that order. In particular, replacing the nonlinear state by the linear filter inside those covariances needs a separate argument. Synchronous coupling of the fast noise chain does not establish that stationary closure, nor an invariant law confined to one sign branch. We therefore do not claim an $O(h^2)$ remainder or a completed nonlinear first-order theorem. The finite-h level improvements below are numerical validation of the candidate correction.

The EMA lag is included. It matters that C_h uses A_t, the EMA state at the time the current noise functional is applied. If $C_Y = \mathrm{Cov}(D_t, Y_t)$ and $C_{\mathrm{lag}} = \mathrm{Cov}(D_t, A_t - Y_t)$, then, exactly,

$$C_h = C_Y + C_{\mathrm{lag}}, \qquad A_t - Y_t = h \sum_{j \geq 1} \beta^j (\kappa_* A_{t-j} + \zeta_{t-j}). \tag{115}$$

Thus replacing A_t by Y_t drops a contribution of the same order as C_h. Writing $c_j = \mathrm{Cov}(D_t, \zeta_{t-j})$ gives the linear-filter identity

$$C_h = -h \sum_{j \geq 1} w_j c_j = -h \sum_{j \geq 1} (1 - \beta^j) c_j + O(h^2) = h C_0 + O(h^2), \tag{116}$$

at fixed β. Here the cross covariances decay geometrically because both kernels are bounded Lipschitz functionals of the contracting noise chain; this makes the series absolutely summable. The fixed-lag perturbation of the stable filter and the geometrically weighted sum justify the last expansion for the

Table 13: EMA-lag audit using new noise banks. $C_A = C_Y + C_{\mathrm{lag}}$; each covariance is divided by h. The last column counts banks with a sign change after burn-in. All twelve cells are retained, including the three cells with frequent crossings.

β	K	h	C_Y/h	C_{lag}/h	C_A/h	Banks crossing
0.50	5	0.001	-0.053	0.009	-0.044	0/8
0.50	5	0.004	-0.038	0.009	-0.029	0/8
0.50	5	0.016	-0.029	0.009	-0.021	1/8
0.90	1	0.001	-0.941	0.648	-0.293	0/8
0.90	1	0.004	-0.950	0.645	-0.305	0/8
0.90	1	0.016	-0.941	0.636	-0.305	0/8
0.90	5	0.001	-0.342	0.265	-0.077	0/8
0.90	5	0.004	-0.364	0.265	-0.099	0/8
0.90	5	0.016	-0.379	0.258	-0.121	8/8
0.99	5	0.001	-4.465	3.200	-1.265	0/8
0.99	5	0.004	-4.240	3.016	-1.224	8/8
0.99	5	0.016	-3.438	2.433	-1.005	8/8

linear covariance. Its β^j factor is precisely the leading EMA-lag contribution. No nonlinear closure is inferred from this identity.

Fresh-bank closure audit. The supplementary mean-shift audit evaluates 12 cells at $h \in \{.001, .004, .016\}$ with 8 new noise banks of 524288 samples per cell, retaining all cells and all sign changes. The nonlinear signal and its EMA start at y_*; a discarded initial window is at least $40/(h\kappa_*)$ steps. This is a retrospective diagnostic, not a new prospective prediction test. The lag term cancels 16.7–77.4% of C_Y in these cells (Table 13), so omitting it would substantially change the correction. The exact lag split holds to 1.7e-15, and the finite-window version of (113), including the observed mean drift, noise-bank mean and slope-bank mean, holds to 2.4e-14. The artifact stores each covariance error, variance error, random-curvature term and Taylor residual separately. These are checks of the accounting; neither small closure residuals on these windows nor a small χ_{lin} supplies a remainder bound. Cells with sign changes cannot be interpreted as a stationary law confined to the positive branch.

Numerical evaluation. The kernel $g_y(y_*, z)$ is evaluated by the chain rule along the Newton–Schulz orbit, exactly as the quadrature slope is, and C_h is accumulated in the same filter pass that produces V_h and V_A, from the products $g_y(y_*, z_t)A_t$ with A_t the EMA state at the moment ζ_t is applied. The curvature $H''(y_*)$ is a central difference of the quadrature slope with steps 10^{-3} and $2 \cdot 10^{-3}$; the two agree to a relative 1.0×10^{-3}. The noise banks (8 banks of 2^{20} samples), seeds and burn-in are those of the locked forecast, so the recomputed y_*, κ_*, V_h and V_A reproduce the locked artifact with relative error 0. Three checks accompany the new term. The Monte Carlo mean of $g_y(y_*, z_t)$ deviates from the quadrature κ_* by at most 0.9% (bank standard error up to 0.7%). The representation (111) holds along the filter to relative precision 2.1×10^{-15}, and the lag-sum form of C_h in (116), summed to the burn-in length, agrees with the direct accumulator to 4.3×10^{-7}. For $\beta \le .9$, C_h/h varies by at most 19.4% over the fifteen-fold range of primary step sizes, and its ratio to the truncated series C_0 at $h = .008$ lies in $[0.83, 1.21]$, within the bank noise of C_0 (relative standard error up to 36.6%). At $\beta = .99$, where $h/(1-\beta)$ is not small at any primary step size, the ratio lies in $[0.55, 0.93]$: the $O(h^2)$ term in (116) is visible, as expected from the non-uniformity in β. In every evaluated cell, $C_h + H''(y_*)V_A/2 > 0$, hence $\delta_h < 0$ and the corrected forecast lies below $y_*^2 + V_h$. All quantities, with bank standard errors, are stored in the artifact.

Retrospective effect on the primary cells. Table 14 recomputes the 24 cells of Table 11 with the new term. The observed levels were already known when the term was derived, so this is a consistency check, not a locked test. In the 15 cells with $\chi_{\mathrm{lin}} \le 1/2$, the median relative level error falls from 1.4% to 0.3% against the primary runs and from 1.4% to 0.7% against the 4,096-trajectory replication; the error decreases in 13 of these cells. The maximum in this regime moves from 15.5% to 16.3% and belongs, as before, to $K = 5$, $\beta = .5$, $h = .008$, the cell whose broad primary interval and positive observed shift indicate late transients rather than a stationary effect. Across all 24 cells the median falls from 4.2% to 2.8%. In the 9 cells with $\chi_{\mathrm{lin}} > 1/2$ the median rises from 6.5% to 9.2%: there the tail mean of $|x_t|/s$ exceeds y_* by amounts unrelated to δ_h, because the chain leaves the neighbourhood in which the expansion is taken, and the small correction has the wrong sign. The last

column compares the observed mean shift with δ_h, both divided by h; a constant ratio across h is the signature of a first-order effect. In the local regime the sign agrees in 14 of 15 cells, the observed value lies within two approximate standard errors of δ_h in 9 (standard errors derived from the level by the delta method, since the primary artifact stores no direct one), and the median ratio of observed to predicted shift is 1.67.

Prospective test. The corrected forecast was then evaluated at $h \in \{.02, .06\}$: .02 was unused, whereas .06 had appeared in the earlier exploratory grid. These are new-seed validations after the forecast was sealed, not two unseen step sizes. The same four noise laws and $K \in \{1, 5\}$ give 16 cells, using the primary forecast's noise banks. The artifact was sealed by its sha256 (prefix `82ab05e3ae9c`, UTC 2026-09-08T20:06:00Z) before any trajectory was run and was not modified afterwards; the hash is re-checked after validation. Validation uses the independent 2×2 matrix implementation described above with 1,024 trajectories per cell, a new seed (4460913), horizons equal to the larger of 12,000 steps and four times the predicted entry, and the tail statistic over the last half. Both the original and the corrected forecast are scored; Table 15 lists every cell. In the 10 cells with $\chi_{\mathrm{lin}} \le 1/2$, the median relative level error is 0.6% for the corrected forecast against 2.9% for $y_*^2 + V_h$, with maxima 3.3% and 7.5%; the corrected forecast is closer in 10 of these cells. Over all 16 cells the medians are 2.1% and 3.6%. The 6 cells outside the local regime are retained; the correction is small there and does not repair them (median 7.1% to 5.0%, with the largest failure again at $\beta = .99$, $K = 5$), consistent with the unclosed terms in Proposition 49. Entry medians remain within 0.3% in every cell.

Separating the two step sizes retrospectively, the genuinely unused $h = 0.02$ has 6 local cells, with median error 1.60% to 0.23% (all 6 improve); both stress cells at this step worsen. The reused $h = 0.06$ has 4 local cells, with median 3.65% to 1.66%. This disclosed split leaves the original sealed aggregate unchanged.

The mean shift itself is predicted less sharply than the level. Its sign is correct in all 10 local cells, but the observed value lies within two standard errors of δ_h in only 3 of them, and at $\beta = .9$ the observed shift exceeds the prediction in magnitude. The local cells with the largest excess are those in which nearly every trajectory changes sign at least once during the run (crossing probabilities are recorded in the artifact); there the tail mean of $|x_t|/s$ includes passages through zero and is not the stationary mean on the positive branch, so the comparison in the last column is only indicative. Because the level depends on the shift through $2y_*\delta_h$, discrepancies in the shift can contribute to its residual error. Higher-order terms and sign changes are possible explanations, but this experiment does not separate them from the covariance and variance closure errors in (113). The candidate correction addresses the missing mean term empirically; the nonlinear first-order closure remains open, and no remainder bound is claimed.

L Quartic construction and its trajectory guarantee

This construction adds a shared correction to normalized descent. The correction changes singular-value ratios while allowing the whole update to remain the gradient of a convex potential. We state the map and its curvature bounds here; the following appendix gives the proofs and the comparison on the transverse task.

A two-product construction. For $P \neq 0$, write $R = \|P\|_F$, $X = P/R$, $\tau_4(X) = \mathrm{tr}((XX^\top)^2)$, and choose $0 < \lambda < 2/9$:

$$\Phi_\lambda(P) = R[1 - \lambda\tau_4(X)], \qquad \widetilde{\Psi}_\lambda(P) = \nabla\Phi_\lambda(P) = (1 + 3\lambda\tau_4(X))X - 4\lambda XX^\top X, \quad (117)$$

with both values zero at the origin. Computing $S = XX^\top$ and SX takes two matrix products; $\tau_4 = \|S\|_F^2$ adds only a scalar reduction.

Theorem 50 (Sharp angular curvature)**.** *Φ_λ is an orthogonally invariant norm. Put $\kappa_\lambda = 1 - \frac{9}{2}\lambda > 0$ and $\Lambda_\lambda = 1 + 3\lambda$. For every $P \neq 0$ and H,*

$$\frac{\kappa_\lambda}{\|P\|_F}\|\Pi_P^\perp H\|_F^2 \le D^2\Phi_\lambda(P)[H, H] \le \frac{\Lambda_\lambda}{\|P\|_F}\|\Pi_P^\perp H\|_F^2, \quad (118)$$

where $\Pi_P^\perp$ projects orthogonally to P. For $\min(m, n) \ge 2$, both constants are attained, including at repeated and zero singular values.

Table 14: Retrospective effect of the first-order mean shift on the 24 primary cells of Table 11. Observed is the primary 256-trajectory tail level of $\|\nabla F\|^2/s^2$; the last column compares the observed tail mean of $|x_t|/s$ minus y_* with the predicted δ_h, both divided by h. A star marks $\chi_{\mathrm{lin}} > 1/2$. These observations were available before the correction was derived, so this table is not a locked test.

B	β	K	h	$y_*^2 + V_h$	$+2y_*\delta_h$	Observed	Err. old	Err. new	$(\bar{y} - y_*)/h$ obs. / pred.
1	0.5	1	0.008	0.15761	0.15728	0.15734	0.2%	0.0%	-0.042 / -0.052
1	0.5	1	0.03	0.15932	0.15806	0.15823	0.7%	0.1%	-0.048 / -0.053
1	0.5	1	0.12	0.16744	0.16231	0.16073	4.2%	1.0%	-0.082 / -0.054
1	0.5	5	0.008	0.00236	0.00234	0.00279	15.5%	16.3%	$+0.095$ / -0.027
1	0.5	5	0.03	0.00244	0.00236	0.00233	4.8%	1.4%	-0.045 / -0.027
1	0.5	5	0.12	0.00288	0.00255	0.00257	12.2%	0.9%	-0.026 / -0.029
1	0.9	1	0.008	0.12545	0.12536	0.12526	0.2%	0.1%	-0.041 / -0.017
1	0.9	1	0.03	0.13334	0.13297	0.13318	0.1%	0.2%	-0.034 / -0.018
1	0.9	1	0.12*	0.18341	0.18133	0.17700	3.6%	2.4%	$+0.042$ / -0.025
1	0.9	5	0.008	0.00191	0.00189	0.00188	1.4%	0.3%	-0.043 / -0.028
1	0.9	5	0.03	0.00222	0.00215	0.00207	6.8%	3.7%	-0.057 / -0.025
1	0.9	5	0.12*	0.00456	0.00429	0.00463	1.5%	7.3%	$+0.107$ / -0.026
1	0.99	1	0.008	0.15685	0.15571	0.15063	4.1%	3.4%	-0.933 / -0.206
1	0.99	1	0.03*	0.36829	0.36164	0.40489	9.0%	10.7%	$+6.134$ / -0.321
1	0.99	1	0.12*	2.28798	2.24587	2.05640	11.3%	9.2%	$+6.857$ / -0.507
1	0.99	5	0.008*	0.00306	0.00297	0.00288	6.5%	3.2%	$+0.274$ / -0.142
1	0.99	5	0.03*	0.01124	0.01084	0.01203	6.6%	9.8%	$+1.540$ / -0.156
1	0.99	5	0.12*	0.08600	0.08430	0.52393	83.6%	83.9%	$+3.106$ / -0.168
4	0.9	1	0.008	0.12274	0.12260	0.12231	0.4%	0.2%	-0.074 / -0.026
4	0.9	1	0.03	0.13116	0.13060	0.13038	0.6%	0.2%	-0.062 / -0.027
4	0.9	1	0.12*	0.18419	0.18144	0.17604	4.6%	3.1%	$+0.066$ / -0.033
4	0.9	5	0.008	0.00187	0.00185	0.00185	1.1%	0.1%	-0.033 / -0.029
4	0.9	5	0.03	0.00219	0.00212	0.00203	7.6%	4.2%	-0.061 / -0.028
4	0.9	5	0.12*	0.00461	0.00431	0.00478	3.5%	9.9%	$+0.116$ / -0.030

Table 15: New-seed test of the corrected forecast at $h \in \{.02, .06\}$ (.06 reuses an exploratory step size), sealed (sha256 prefix `82ab05e3ae9c`) before 1,024 independent matrix trajectories per cell were run with a new seed. Columns as in Table 14.

B	β	K	h	$y_*^2 + V_h$	$+2y_*\delta_h$	Observed	Err. old	Err. new	$(\bar{y} - y_*)/h$ obs. / pred.
1	0.5	1	0.02	0.15853	0.15769	0.15763	0.6%	0.0%	-0.057 / -0.053
1	0.5	1	0.06	0.16182	0.15930	0.15879	1.9%	0.3%	-0.068 / -0.053
1	0.5	5	0.02	0.00240	0.00235	0.00234	2.6%	0.4%	-0.035 / -0.027
1	0.5	5	0.06	0.00257	0.00241	0.00239	7.5%	0.8%	-0.034 / -0.028
1	0.9	1	0.02	0.12953	0.12930	0.12936	0.1%	0.0%	-0.029 / -0.017
1	0.9	1	0.06	0.14697	0.14612	0.14247	3.2%	2.6%	-0.145 / -0.020
1	0.9	5	0.02	0.00206	0.00202	0.00198	3.9%	1.7%	-0.056 / -0.026
1	0.9	5	0.06*	0.00281	0.00268	0.00262	7.3%	2.5%	$+0.011$ / -0.025
1	0.99	1	0.02*	0.25488	0.25121	0.26251	2.9%	4.3%	$+4.041$ / -0.265
1	0.99	1	0.06*	0.84598	0.82864	0.90897	6.9%	8.8%	$+7.357$ / -0.418
1	0.99	5	0.02*	0.00682	0.00656	0.00696	1.9%	5.7%	$+1.314$ / -0.154
1	0.99	5	0.06*	0.03001	0.02917	0.06895	56.5%	57.7%	$+2.039$ / -0.167
4	0.9	1	0.02	0.12709	0.12673	0.12677	0.3%	0.0%	-0.035 / -0.026
4	0.9	1	0.06	0.14564	0.14445	0.13984	4.1%	3.3%	-0.165 / -0.028
4	0.9	5	0.02	0.00203	0.00198	0.00193	4.9%	2.5%	-0.065 / -0.027
4	0.9	5	0.06*	0.00280	0.00266	0.00261	7.2%	1.7%	$+0.016$ / -0.028

The proof exploits the joint constraint on the quartic moment and the angular second variation; bounding them separately gives only $1 - 6\lambda$. At $\lambda = 1/8$, the sharp bracket is $[7/16, 11/8]$. Symmetric pairing and integration of this Hessian bound give

$$\langle g, \mathbb{E}\widetilde{\Psi}_\lambda(g+Z)\rangle \geq \kappa_\lambda \mathbb{E}\frac{\|g\|_F^2\|Z\|_F^2 - \langle g, Z\rangle^2}{(\|g\|_F + \|Z\|_F)^3},$$

strictly positive for $g \neq 0$ whenever $\mathbb{P}(Z \notin \operatorname{span}\{g\}) > 0$ (Corollary 52). This bound needs neither a density nor a spectral gap.

The output singular values are $\widetilde{s}_i = s_i[1 + 3\lambda\tau_4 - 4\lambda s_i^2]$. For $s_i > s_j > 0$, they satisfy $1 < \widetilde{s}_i/\widetilde{s}_j < s_i/s_j$ by (127): the map preserves order and strictly compresses ratios. The shared coefficient evades pointwise rigidity. Its limitation is quantitative: $\|\widetilde{\Psi}_\lambda(P) - X\|_F \leq 7\lambda/\operatorname{sr}(P)$ by (128), so at high stable rank it is close to normalized descent. The Schatten family offers stronger flattening at SVD cost.

Theorem 51 (Repair on the same task and sample budget)**.** *Take Definition 7, changing only the map to $\widetilde{\Psi}_{1/8}$. Its projection is $\widetilde{q}(t) = t(7 + 34t^2 + 15t^4)/(16(1 + t^2)^{5/2})$, with sharp bounds $\kappa^{\mathrm{task}} t/\sqrt{1 + t^2} \leq \widetilde{q}(t) \leq \Lambda^{\mathrm{task}} t/\sqrt{1 + t^2}$, $\kappa^{\mathrm{task}} = 7/16$, $\Lambda^{\mathrm{task}} = 23/24$. Let Q bound the projected update, $m_{\mathrm{win}} \geq 1$, $\eta Q m_{\mathrm{win}} \leq 1/16$, $H = \beta^{m_{\mathrm{win}}} x_0 + \eta Q/(1 - \beta) \leq 1/8$, $\kappa_{\mathrm{c}} = \kappa^{\mathrm{task}}$, $U_0 = x_0 + \eta Q m_{\mathrm{win}}$ and $t_c = m_{\mathrm{win}} + \lceil \sqrt{2} U_0/(\eta\kappa_{\mathrm{c}}) \rceil + 1$. Every noise path with $\rho_t \leq \varrho$ for $t \leq T$ satisfies*

$$|x_t| \leq H + \varrho \exp[-\kappa_{\mathrm{c}}\eta(t - t_c)/(2(\varrho + H))], \qquad t_c \leq t \leq T.$$

Taking $\varrho = \sigma$ covers every path; taking $\varrho_T = \min\{\sigma, s\sqrt{6\log(8(T + 1)/\delta)}\}$ covers all such times with probability at least $1 - \delta/2$.

The curvature constants $\kappa_\lambda, \Lambda_\lambda$, task envelope $\kappa^{\mathrm{task}}, \Lambda^{\mathrm{task}}$, and contraction slope κ_{c} play different roles. In particular, $\Lambda^{\mathrm{task}} = 23/24$ differs from the global curvature ceiling $11/8$. The residual H is a lag and initialization bound, not a proved floor. For a simultaneous comparison with Theorem 31, impose $H \leq as/4$ and $T \geq \max\{t_{\mathrm{ent}}, t_{\mathrm{sep}}\}$, where t_{sep} is defined in (135), and allocate $\delta/2$ to each arm. The repaired squared gradient is then at most $a^2 s^2/4$ while the reference's is at least $a^2 s^2$ throughout the common window.

An ablation separating task success from universal alignment. Deleting $3\lambda\tau_4 X$ gives $\Psi_{\mathrm{unc}} = X - XX^\top X/2$ at $\lambda = 1/8$. Its task response $t(1 + 3t^2)/(4(1 + t^2)^{3/2})$ is positive, yet a three-dimensional witness satisfies $\langle H, D\Psi_{\mathrm{unc}}(P)[H]\rangle = -4/747$ (Proposition 56). Thus recovery on one construction does not establish universal monotonicity. The experiments include this ablation alongside the coupled potential.

M THE COUPLED CONVEX POTENTIAL AND ITS REPAIR GUARANTEE

We begin by proving that the quartic potential is a norm. This gives the alignment guarantee, after which we quantify how much the update flattens singular values. We then study the update on the same transverse task as the reference method. The remaining sections treat the Schatten family, show what changes when the shared term is deleted, and examine a positive denominator regularizer.

M.1 THE POTENTIAL IS A NORM

Recall $\Phi_\lambda(P) = \|P\|_F[1 - \lambda\tau_4(X)]$ with $X = P/\|P\|_F$ and $\tau_4(X) = \operatorname{tr}((XX^\top)^2) = \sum_i s_i^4$ for the singular values s_i of X, so $0 \leq \tau_4 \leq 1$ with $\tau_4 = 1$ exactly at rank one and $\tau_4 = 1/r$ at a flat spectrum. Since τ_4 depends only on the singular values of X, Φ_λ is orthogonally invariant, and one-homogeneity is immediate from the definition. Convexity is what needs proof.

Proof of Theorem 50. One-homogeneity makes the Hessian vanish in the radial direction, since $\Phi_\lambda(P + \theta P) = (1 + \theta)\Phi_\lambda(P)$ is affine in θ. So it suffices to bound $D^2\Phi_\lambda$ on directions H with $\langle X, H\rangle = 0$ at a point with $\|X\|_F = 1$, and then restore the factor $\|P\|_F^{-1}$ by homogeneity of degree -1 of the Hessian. Write

$$\mathcal{M}_X(H) = \langle H, HX^\top X + XH^\top X + XX^\top H\rangle. \tag{119}$$

Differentiating Φ_λ twice at $\|X\|_F = 1$ along such an H gives

$$D^2\Phi_\lambda(X)[H, H] = (1 + 3\lambda\tau_4(X))\|H\|_F^2 - 4\lambda\mathcal{M}_X(H), \tag{120}$$

which is a direct computation from $\nabla\Phi_\lambda$ in (117). It remains to show $0 \leq \mathcal{M}_X(H) \leq \frac{3}{2}\|H\|_F^2$ for unit H orthogonal to X.

Pass to the singular coordinates of X, so $X = \mathrm{diag}(x_1, \dots, x_r)$ padded with zeros and $\sum_i x_i^2 = 1$. Expanding (119) in those coordinates decouples it into blocks. The diagonal entries h_{ii} contribute $3\sum_i x_i^2 h_{ii}^2$. Each off-diagonal pair (h_{ij}, h_{ji}) with $i \neq j$ and both indices at most r contributes a 2×2 quadratic form in (h_{ij}, h_{ji}) whose eigenvalues are $x_i^2 + x_j^2 \pm x_i x_j$, so its largest coefficient is $x_i^2 + x_j^2 + x_i x_j \leq \frac{3}{2}(x_i^2 + x_j^2) \leq \frac{3}{2}$ and its smallest is $x_i^2 + x_j^2 - x_i x_j \geq 0$. The rectangular block, present when $m \neq n$, contributes $x_i^2 h_{ij}^2$ with coefficient at most 1 and at least 0. Hence $\mathcal{M}_X(H) \geq 0$ always, and the only coefficient that can exceed 1 is the diagonal one. For that, note $\|H\|_F = 1$ and $\langle X, H \rangle = 0$ mean the vectors $x = (x_i)$ and $h = (h_{ii})$ satisfy $\|x\|_2 = 1$, $\|h\|_2 \leq 1$ and $\langle x, h \rangle = 0$, so they are two orthonormal columns after normalizing h, and every row of that two-column matrix has norm at most one, giving $x_i^2 + h_{ii}^2 \leq 1$. Therefore

$$\sum_i x_i^2 h_{ii}^2 \leq \tfrac{1}{4} \sum_i (x_i^2 + h_{ii}^2)^2 \leq \tfrac{1}{4} \sum_i (x_i^2 + h_{ii}^2) \leq \tfrac{1}{2}, \tag{121}$$

using $xy \leq \frac{1}{4}(x+y)^2$, then $(x+y)^2 \leq x + y$ for $x + y \leq 1$, then $\sum_i x_i^2 + \sum_i h_{ii}^2 \leq 2$. So the diagonal contribution is at most $3/2$ as well, and $\mathcal{M}_X(H) \leq \frac{3}{2}\|H\|_F^2$ overall, giving $\Lambda_\lambda = 1 + 3\lambda$ at $\tau_4 = 1$ and the crude floor $1 - 6\lambda$ at $\tau_4 = 0$.

The floor is not sharp, because the two extremes are not jointly attainable: by the equality analysis above, $\mathcal{M}_X(H) = \frac{3}{2}$ needs $x_i^2 = h_{ii}^2 = \frac{1}{2}$ on two indices, which already forces $\tau_4 \geq \frac{1}{2}$. Both bounds must therefore be optimized together, that is over

$$\sup_{\|H\|_F = 1,\ \langle X, H \rangle = 0} \big[4\mathcal{M}_X(H) - 3\tau_4(X) \big]. \tag{122}$$

The block decomposition writes $\mathcal{M}_X(H)$ as a sum of contributions from the diagonal entries, from each off-diagonal pair, and from the rectangular part, and $\|H\|_F^2 = 1$ splits the mass among them. Since the bracket is linear in the block masses at fixed X, its supremum over unit H is a convex combination of the values obtained by putting the whole unit mass on one block, so it suffices to bound each normalized block separately. We do that in three cases, and each is where the argument replaces a step that does not survive: it is *not* true in general that moving mass outside a pair leaves that pair's first contribution unchanged while raising τ_4, because τ_4 is a function of X alone and redistributing H does not move it, while redistributing X changes both terms at once.

Off-diagonal pairs. Put the unit mass on the pair (i, j) and let $u = x_i^2 + x_j^2 \leq 1$. The largest eigenvalue of that 2×2 form is $x_i^2 + x_j^2 + x_i x_j \leq \frac{3}{2}u$, and the same pair contributes $\tau_4 \geq (x_i^4 + x_j^4) \geq u^2/2$. Hence

$$4\mathcal{M}_X - 3\tau_4 \ \leq\ 6u - \tfrac{3}{2}u^2 \ \leq\ \tfrac{9}{2}, \tag{123}$$

the last step because $6u - \frac{3}{2}u^2$ increases on $[0, 1]$ and equals $\frac{9}{2}$ at $u = 1$.

The rectangular block. For H supported there, that is on the columns beyond index r where X vanishes, two of the three terms of $\mathcal{M}_X$ drop and $\mathcal{M}_X(H) = \sum_k h_k^2 x_k^2 \leq \max_k x_k^2 \leq 1$, so the bracket is at most $4 < \frac{9}{2}$. The confinement is doing work and is supplied by the block decomposition, not by H being arbitrary: an unconfined unit H orthogonal to X can reach $\mathcal{M}_X = 1.27$.

The diagonal block. Normalize the diagonal vector h to unit length; with $\langle x, h \rangle = 0$ the vectors x, h are orthonormal, so writing $u_i = x_i^2$ and $y_i = h_i^2$ gives $\sum_i u_i = \sum_i y_i = 1$ and $u_i + y_i \leq 1$ for each i, the latter because every row of the two-column orthonormal matrix $[x\ h]$ has norm at most one. Let $u = u_1 \geq v = u_2$ be the two largest u_i. Then $\sum_i u_i y_i \leq u y_1 + v(1 - y_1)$, since the remaining y mass $1 - y_1$ meets coefficients at most v, and $y_1 \leq 1 - u$ gives $\sum_i u_i y_i \leq u(1 - u + v)$ as $u \geq v$. Also $\tau_4 = \sum_i u_i^2 \geq u^2 + v^2$. With $\mathcal{M}_X = 3\sum_i u_i y_i$ on this block,

$$4\mathcal{M}_X - 3\tau_4 \ \leq\ 12u(1 - u + v) - 3(u^2 + v^2) = 12u - 15u^2 + 12uv - 3v^2.$$

The right side has v-derivative $12u - 6v \geq 6u > 0$ on $0 \leq v \leq u$, so it increases in v, and $v \leq \min\{u, 1 - u\}$. For $u \leq \frac{1}{2}$ take $v = u$, giving $12u - 6u^2 \leq \frac{9}{2}$ on $[0, \frac{1}{2}]$. For $u \geq \frac{1}{2}$ take $v = 1 - u$, giving $30u(1 - u) - 3 \leq \frac{9}{2}$, with equality at $u = \frac{1}{2}$. Both branches meet $\frac{9}{2}$ only at $u = v = \frac{1}{2}$. So (122) equals $9/2$ and

$$\kappa_\lambda = 1 - \tfrac{9}{2}\lambda, \tag{124}$$

attained at two equal singular values with H the symmetric off-diagonal direction on that pair. Substituting into (120) gives (118) after restoring $\|P\|_F^{-1}$ and writing $\Pi_P^{\perp} H$ for the component of H

orthogonal to P. The ceiling $\Lambda_\lambda = 1 + 3\lambda$ is attained at $\tau_4 = 1$ with $\mathcal{M}_X(H) = 0$, for instance $P = \mathrm{diag}(1, 0)$ and $H = \mathrm{diag}(0, 1)$, where $D^2\Phi_\lambda(P)[H, H] = 1 + 3\lambda$ exactly; so neither constant can be improved, and no value below $1 + 3\lambda$ is an upper bound. Both equality cases need two available singular directions, so sharpness is asserted for $\min(m, n) \geq 2$.

The Hessian is therefore positive semidefinite at every $P \neq 0$ as soon as $\lambda < 2/9$, and on a line through the origin $\Phi_\lambda(\theta P) = |\theta|\Phi_\lambda(P)$ is convex, so Φ_λ is convex on all of $\mathbb{H}$. It is even and one-homogeneous, and $\Phi_\lambda(P) \geq (1 - \lambda)\|P\|_F$ since $\tau_4 \leq 1$, with $\Phi_\lambda(P) \leq \|P\|_F$ since $\tau_4 \geq 0$. A convex, even, one-homogeneous function that is bounded above and below by positive multiples of a norm is a norm. Rank deficiency and repeated singular values are permitted throughout, since the argument used only the singular values of X and never a gap. □

M.2 Alignment, without a density or a derivative bound

Corollary 52 (Alignment with no density or derivative hypothesis)**.** *For every bounded centrally symmetric matrix noise Z,*

$$\langle g, \mathbb{E}\widetilde{\Psi}_\lambda(g + Z)\rangle \;\geq\; \kappa_\lambda\, \mathbb{E}\frac{\|g\|_F^2\|Z\|_F^2 - \langle g, Z\rangle^2}{(\|g\|_F + \|Z\|_F)^3}, \tag{125}$$

the integrand being zero at $(0, 0)$, and its right side is strictly positive for $g \neq 0$ whenever $\mathbb{P}(Z \notin \mathrm{span}\{g\}) > 0$.

Proof of Corollary 52. By central symmetry and (18) applied to $\widetilde{\Psi}_\lambda$, which is odd, $\langle g, \mathbb{E}\widetilde{\Psi}_\lambda(g + Z)\rangle = \frac{1}{2}\mathbb{E}\langle g, \widetilde{\Psi}_\lambda(Z + g) - \widetilde{\Psi}_\lambda(Z - g)\rangle$. For a fixed realization of Z, the fundamental theorem of calculus along the segment $t \mapsto Z + tg$, $t \in [-1, 1]$, gives

$$\langle g, \widetilde{\Psi}_\lambda(Z + g) - \widetilde{\Psi}_\lambda(Z - g)\rangle = \int_{-1}^{1} D^2\Phi_\lambda(Z + tg)[g, g]\, dt \;\geq\; \kappa_\lambda \int_{-1}^{1} \frac{\|\Pi^\perp_{Z+tg} g\|_F^2}{\|Z + tg\|_F}\, dt.$$

The projection identity $\|\Pi^\perp_Q g\|_F^2 = (\|g\|_F^2\|Q\|_F^2 - \langle g, Q\rangle^2)/\|Q\|_F^2$ at $Q = Z + tg$ has numerator $\|g\|_F^2\|Z + tg\|_F^2 - \langle g, Z + tg\rangle^2 = \|g\|_F^2\|Z\|_F^2 - \langle g, Z\rangle^2$, independent of t because adding a multiple of g changes neither the transverse energy nor the Gram determinant of (g, Z). Hence the integrand is that fixed numerator divided by $\|Z + tg\|_F^3$, and $\|Z + tg\|_F \leq \|Z\|_F + \|g\|_F$ for $|t| \leq 1$ bounds the denominator, leaving a factor 2 from the length of the interval that cancels the $\frac{1}{2}$. That gives (125). If the segment passes through the origin, the numerator $\|g\|_F^2\|Z\|_F^2 - \langle g, Z\rangle^2$ vanishes, since Z is then a multiple of g, and the subgradient jump of a convex function contributes a nonnegative amount to the left side, so the inequality persists. Positivity of the right side for $g \neq 0$ under $\mathbb{P}(Z \notin \mathrm{span}\{g\}) > 0$ is immediate, the numerator being the Gram determinant of (g, Z).

For the derivative form, suppose $\mathbb{P}(Z = 0) = 0$ and $\mathbb{E}\|Z\|_F^{-1} < \infty$. Zero-homogeneity of $\widetilde{\Psi}_\lambda$ off the origin plus (118) give $\|D\widetilde{\Psi}_\lambda(Q)\|_{\mathrm{op}} \leq \Lambda_\lambda/\|Q\|_F$, so for a perturbation of size $\|\zeta\|_F \leq \|Z\|_F/2$ the difference quotient $\|\widetilde{\Psi}_\lambda(Z + \zeta) - \widetilde{\Psi}_\lambda(Z)\|_F/\|\zeta\|_F$ is bounded by $2\Lambda_\lambda/\|Z\|_F$, and for $\|\zeta\|_F > \|Z\|_F/2$ it is bounded by $4\sup_P\|\widetilde{\Psi}_\lambda(P)\|_F/\|Z\|_F$, since $\widetilde{\Psi}_\lambda$ is bounded. Both bounds are constant multiples of $1/\|Z\|_F$, which is integrable, so dominated convergence lets the derivative and the expectation be exchanged. Taking expectations in (118) and using $D^2\Phi_\lambda(Z) = D\widetilde{\Psi}_\lambda(Z)$ with $\Pi^\perp_Z = I - \widehat{Z} \otimes \widehat{Z}$, $\widehat{Z} = Z/\|Z\|_F$, gives the two-sided bound by

$$\mathcal{G}_Z = \mathbb{E}\Big[\frac{\Pi^\perp_Z}{\|Z\|_F}\Big], \tag{126}$$

the recovery matrix of the frozen field at the origin, in the sense that $\kappa_\lambda\, \mathcal{G}_Z \preceq A \preceq \Lambda_\lambda\, \mathcal{G}_Z$. □

The scale $1/\varrho$ in $\mathcal{G}_Z$ is attained and not an artifact: if Z lives on the sphere of radius ϱ and is almost surely orthogonal to a unit direction E, then $\langle E, \mathcal{G}_Z E\rangle = 1/\varrho$. If the inverse moment fails to exist one must use (125) and not extend the derivative form.

M.3 Ratio compression and the flattening bound

Proposition 53 (Strict ratio compression and its limit)**.** *Let s_i be the singular values of $X = P/\|P\|_F$ and $\widetilde{s}_i$ those of $\widetilde{\Psi}_\lambda(P)$, with $0 < \lambda < 2/9$. Then for $s_i > s_j > 0$,*

$$1 < \widetilde{s}_i/\widetilde{s}_j < s_i/s_j, \tag{127}$$

and with $\mathrm{sr}(P) = \|P\|_F^2/\|P\|_{\mathrm{op}}^2$ the stable rank,

$$\|\widetilde{\Psi}_\lambda(P) - X\|_F \le 3\lambda\tau_4 + 4\lambda\|XX^\top X\|_F \le \frac{7\lambda}{\mathrm{sr}(P)}. \tag{128}$$

Proof of (127) *and* (128). Write $\alpha = \alpha(X) = 1 + 3\lambda\tau_4(X)$, so the output singular values are $\widetilde{s}_i = s_i(\alpha - 4\lambda s_i^2)$. Since $s_i \le 1$ and $\lambda < 2/9$, the factor $\alpha - 4\lambda s_i^2 \ge 1 - 4\lambda > 0$, so signs and ordering are preserved and the singular subspaces are untouched, the map being a spectral function. For $s_i > s_j > 0$,

$$\frac{\widetilde{s}_i}{\widetilde{s}_j} = \frac{s_i}{s_j} \cdot \frac{\alpha - 4\lambda s_i^2}{\alpha - 4\lambda s_j^2} < \frac{s_i}{s_j},$$

the last step because $s_i > s_j$ makes the numerator smaller than the denominator, which is the statement that a larger singular value receives a smaller gain. For the other side, $\widetilde{s}_i - \widetilde{s}_j = (s_i - s_j)[\alpha - 4\lambda(s_i^2 + s_i s_j + s_j^2)]$. Here $s_i^2 + s_i s_j + s_j^2 \le \frac{3}{2}(s_i^2 + s_j^2)$, and the pair carries at most all of the unit Frobenius mass, so with $\varpi = s_i^2 + s_j^2 \le 1$ the bracket is at least $1 + 3\lambda\tau_4 - 6\lambda\varpi$. Since $\tau_4 \ge \varpi^2/2$ on that pair, the bracket is at least $1 + \frac{3}{2}\lambda\varpi^2 - 6\lambda\varpi$, which is decreasing in $\varpi \in [0, 1]$ and hence at least $1 - \frac{9}{2}\lambda > 0$ throughout $\lambda < 2/9$. So $\widetilde{s}_i > \widetilde{s}_j$ and the ratio exceeds one. For (128), $\tau_4 = \sum_i s_i^4 \le \|X\|_{\mathrm{op}}^2 \sum_i s_i^2 = \|X\|_{\mathrm{op}}^2$ and $\|XX^\top X\|_F^2 = \sum_i s_i^6 \le \|X\|_{\mathrm{op}}^4 \sum_i s_i^2$, so $\|XX^\top X\|_F \le \|X\|_{\mathrm{op}}^2$; the triangle inequality on $\widetilde{\Psi}_\lambda(P) - X = 3\lambda\tau_4 X - 4\lambda XX^\top X$ with $\|X\|_F = 1$ then gives $\|\widetilde{\Psi}_\lambda(P) - X\|_F \le 3\lambda\|X\|_{\mathrm{op}}^2 + 4\lambda\|X\|_{\mathrm{op}}^2 = 7\lambda\|X\|_{\mathrm{op}}^2$, and $\|X\|_{\mathrm{op}}^2 = \|P\|_{\mathrm{op}}^2/\|P\|_F^2 = 1/\mathrm{sr}(P)$. □

So at high stable rank the correction is close to Frobenius-normalized descent. That is the boundary a usefulness claim has to clear, and Section 7 measures the realized angle, the ratio compression and the relative difference rather than only the loss.

M.4 The repaired trajectory

On the task of Definition 7 at $\lambda = 1/8$, Lemma 8 applied to the spectral function of $\widetilde{\Psi}_{1/8}$ gives, after collecting terms,

$$\widetilde{q}(t) = \frac{t(7 + 34t^2 + 15t^4)}{16(1 + t^2)^{5/2}}, \tag{129}$$

which is strictly positive for $t > 0$ and satisfies $|\widetilde{q}| \le Q = 2$. Dividing by $t/\sqrt{1+t^2}$ gives the rational function $(15\vartheta^2 + 34\vartheta + 7)/(16(\vartheta+1)^2)$ in $\vartheta = t^2$, whose derivative vanishes only at $\vartheta = 5$; the value there is $23/24$, the value at $\vartheta = 0$ is $7/16$, and the limit at infinity is $15/16$. So

$$\tfrac{7}{16}\frac{t}{\sqrt{1+t^2}} \;\le\; \widetilde{q}(t) \;\le\; \tfrac{23}{24}\frac{t}{\sqrt{1+t^2}}, \qquad t \ge 0, \tag{130}$$

with both constants attained, so on this task $\kappa^{\mathrm{task}} = 7/16$ and $\Lambda^{\mathrm{task}} = 23/24$ are the sharp response constants. They must not be confused with the curvature constants of Theorem 50, which at $\lambda = 1/8$ are $\kappa_\lambda = 1 - \frac{9}{2}\lambda = 7/16$ and $\Lambda_\lambda = 1 + 3\lambda = 11/8$. The agreement of the two *lower* constants is not an accident of $\lambda = 1/8$. For general λ the exact profile is

$$\widetilde{q}_\lambda(t) = \frac{t}{\sqrt{1+t^2}}\Big[1 - \frac{\lambda}{2}\frac{t^4 - 2t^2 + 9}{(1+t^2)^2}\Big], \tag{131}$$

and writing $A_\lambda(t) = \widetilde{q}_\lambda(t)\sqrt{1+t^2}/t$, extended by its limit at $t = 0$, two exact identities hold for every $t \ge 0$:

$$A_\lambda(t) - \Big(1 - \frac{9}{2}\lambda\Big) = \frac{2\lambda t^2(2t^2+5)}{(1+t^2)^2}, \qquad \Big(1 - \frac{\lambda}{3}\Big) - A_\lambda(t) = \frac{\lambda(t^2-5)^2}{6(1+t^2)^2}. \tag{132}$$

Both right sides are nonnegative, and the first vanishes exactly at $t = 0$ while the second vanishes exactly at $t^2 = 5$. So the sharp task envelope is $(1 - \frac{9}{2}\lambda) \le A_\lambda \le (1 - \frac{\lambda}{3})$ throughout $0 \le \lambda < 2/9$, and $\kappa_\lambda^{\mathrm{task}} = \kappa_\lambda = 1 - \frac{9}{2}\lambda$ identically in λ, with $\Lambda_\lambda^{\mathrm{task}} = 1 - \lambda/3$ giving $23/24$ at $\lambda = 1/8$. The two constants remain *different objects*: the curvature pair brackets $D^2\Phi_\lambda$ over the whole matrix space at every rank, while the response pair brackets one scalar profile on a two-dimensional instance, and the upper constants genuinely differ ($1 + 3\lambda$ against $1 - \lambda/3$). What is false is the earlier claim that the lower two coincide only numerically at one λ. We quote κ^{task} where the argument is about this task and Λ_λ only where a bound must hold for every noise geometry.

Proof of Theorem 51. The unconditional lag bound of Lemma 33 does not depend on the map, so it holds here. Put $d = \beta^{m_{\mathrm{win}}} x_0 + \eta Q\beta/(1-\beta)$, so $H = d + \eta Q$. First, global boundedness. For $t \ge m_{\mathrm{win}}$, if $|x_t| > H \ge d$ then the lag bound makes a_t share the sign of x_t, and since $\widetilde{q}$ has the sign of its argument the update moves inward; as one step has length at most $\eta Q \le H$ it cannot cross zero. Hence $|x_{t+1}| \le \max\{|x_t|, H\}$, and induction from $|x_{m_{\mathrm{win}}}| \le x_0 + \eta Q m_{\mathrm{win}} \le 5/16$ gives $|x_t| \le 5/16$ for all $t \ge m_{\mathrm{win}}$, so $\phi'(x_t) = x_t$ throughout and the trajectory stays in the quadratic region. The same argument shows $[-H, H]$ is invariant: inside $|x_t| \le d$ the bounded step length keeps $|x_{t+1}| \le d + \eta Q = H$, and on $d < |x_t| \le H$ the update points inward, and no larger symmetric interval is exited in one step either.

Second, the contraction. For $|x_t| > H$, using the lower envelope in (130) at argument $|a_t|/\rho_t$ with $|a_t| \ge |x_t| - d$ and $\rho_t \le \varrho$,

$$|x_{t+1}| \le |x_t| - \eta\kappa_{\mathrm{c}} \frac{|x_t| - d}{\sqrt{(|x_t| + d)^2 + \varrho^2}}, \qquad \kappa_{\mathrm{c}} = \tfrac{7}{16}, \tag{133}$$

since $\widetilde{q}(t) \ge \kappa_{\mathrm{c}} t/\sqrt{1+t^2}$ translates into $\langle E, \widetilde{\Psi}(aE + Z)\rangle \ge \kappa_{\mathrm{c}} a/\sqrt{a^2 + \rho^2}$ and $|a_t| + d \ge |x_t|$ bounds the denominator. While $|x_t| > \varrho + H$ we have $|a_t| > \varrho \ge \rho_t$, so the argument exceeds one and each step moves inward by at least $\eta\kappa_{\mathrm{c}}/\sqrt{2}$; hence $|x_t| \le \varrho + H$ by $t_c = m_{\mathrm{win}} + \lceil \sqrt{2} U_0/(\eta\kappa_{\mathrm{c}}) \rceil + 1$ at the latest. After entry, set $y_t = (|x_t| - H)_+$. On $|x_t| \le \varrho + H$ the denominator in (133) is at most $\sqrt{(\varrho + 2H)^2 + \varrho^2} \le 2(\varrho + H)$ and the numerator is at least y_t, so $y_{t+1} \le (1 - \eta\kappa_{\mathrm{c}}/(2(\varrho + H)))y_t$, the factor being nonnegative because $H \ge \eta Q \ge \eta\kappa_{\mathrm{c}}$. Entering with $y \le \varrho$ and iterating gives the geometric decay

$$|x_t| \;\le\; H + \varrho \exp\Big[-\frac{\kappa_{\mathrm{c}}\eta(t - t_c)}{2(\varrho + H)}\Big], \qquad t_c \le t \le T. \tag{134}$$

Third, the noise bound. Each coordinate of Z_t is a weighted sum of independent variables bounded in $[-\sigma/\sqrt{2}, \sigma/\sqrt{2}]$ with total weight-squared at most v, so Hoeffding's sub-Gaussian proxy for one coordinate is at most $\sigma^2 v/2 = 3s^2/2$. Applying the two-sided bound to each of the two coordinates and adding, $\mathbb{P}(\rho_t > \varrho) \le 4\exp[-\varrho^2/(6s^2)]$. A union over the $T + 1$ times, combined with the deterministic $\rho_t \le \sigma$, gives the stated ϱ_T at level $\delta/2$, which is the allocation the separation statement uses; independence across times is not needed for a union bound. □

The separation statement follows by comparing the two conclusions on one noise stream, and it needs a nonempty common window rather than merely equal sampling. Assume $H \le as/4$ and put

$$t_{\mathrm{sep}} = t_c + \Big\lceil \frac{2(\varrho_T + H)}{\eta\kappa_{\mathrm{c}}} \log_+ \frac{4\varrho_T}{as} \Big\rceil, \tag{135}$$

which is exactly the time (134) needs to bring $|x_t|$ below $as/2$, so that the repaired squared gradient is at most $a^2 s^2/4$. Require

$$T \;\ge\; \max\{t_{\mathrm{ent}},\, t_{\mathrm{sep}}\}. \tag{136}$$

Allocate $\delta/2$ to the residence event and $\delta/2$ to the radius envelope, recomputing each logarithmic budget at its own level, so $\mathcal{L} = \log[8m_{\mathrm{win}}(T+1)^2/\delta]$ in (67) and $\varrho_T = \min\{\sigma, s\sqrt{6\log(8(T+1)/\delta)}\}$ here. Then on the intersection, of probability at least $1 - \delta$, for every $t \in [\max\{t_{\mathrm{ent}}, t_{\mathrm{sep}}\}, T]$ the reference squared gradient is at least $a^2 s^2$ while the repaired one is at most $a^2 s^2/4$, and both arms consume identical sample counts. Without (136) the two guarantees can hold on disjoint stretches and the comparison is empty; equal per-step sampling does not supply it. Nothing in H resists $\eta \downarrow 0$, so it is a transient and lag bound and not a proved floor. The explicit sufficient selection of Section 5 gives $H \le a\gamma s/(8D) < as/4$, so the hypothesis is met by the same choice that satisfies (68).

For decreasing steps at fixed B, β with $\eta_t \to 0$ and $\sum_t \eta_t = \infty$, the lag obeys $|a_t - \phi'(x_t)| \le \beta^{t+1} x_0 + Q \sum_{k<t} \beta^{t-k} \eta_k \to 0$ by dominated convergence on the convolution. Given $\epsilon > 0$, for all large t the lag is below $\epsilon/4$, so on $|x_t| \ge \epsilon/2$ the update has the inward sign with magnitude at least a fixed positive multiple of η_t, and $\sum_t \eta_t = \infty$ forces entry into $[-\epsilon, \epsilon]$; once there, the vanishing step length and lag prevent exit. Hence $x_t \to 0$ on every bounded noise path. This is specific to the transverse task and is not a general debiasing claim under state-dependent noise.

M.5 The Schatten family and its endpoint

Proof of Proposition 5. Write Ψ for the selection. The subgradient inequality at $x = g + z$ and $y = z$ gives

$$\langle g, \Psi(g+z) \rangle \;\ge\; \Phi(g+z) - \Phi(z),$$

and the same at $-z$ gives $\langle g, \Psi(g-z) \rangle \ge \Phi(g-z) - \Phi(-z) = \Phi(z-g) - \Phi(z)$ by evenness. Averaging over the symmetric law,

$$\langle g, \mathbb{E}\Psi(g+Z) \rangle \;\ge\; \tfrac{1}{2}\mathbb{E}\big[\Phi(g+Z) + \Phi(Z-g) - 2\Phi(Z)\big] \;\ge\; 0,$$

the last step being midpoint convexity at the midpoint Z of $g+Z$ and $Z-g$. Finiteness, convexity and one-homogeneity in finite dimension give a global Lipschitz bound on Φ, so every subgradient selection is bounded and its expectation exists. Neither oddness nor $\Psi(0) = 0$ is needed for the alignment conclusion, and neither holds for every selection: for $\Phi(x) = |x|$ on $\mathbb{R}$, the Borel selection $\Psi(x) = 1$ for $x \ge 0$ and -1 for $x < 0$ is a counterexample to both additional assertions.

For the existence claim, evenness and convexity imply $\Phi(y) \ge \Phi(0) = 0$, so $0 \in \partial\Phi(0)$. The unique minimum-norm selection $\Psi_{\min}(x) = \mathrm{proj}_{\partial\Phi(x)}(0)$ is consequently zero at the origin and odd, because $\partial\Phi(-x) = -\partial\Phi(x)$ and projection commutes with negation. It is Borel measurable. One explicit justification is to take the pointwise limit as $t \downarrow 0$ of the continuous maps $\Psi_t(x) = (x - \mathrm{prox}_{t\Phi}(x))/t$. Indeed, with $y_t = \mathrm{prox}_{t\Phi}(x)$ one has $\Psi_t(x) \in \partial\Phi(y_t)$, and monotonicity against $(x, \Psi_{\min}(x))$ gives $\|\Psi_t(x)\|_F^2 \le \langle \Psi_t(x), \Psi_{\min}(x) \rangle$, hence $\|\Psi_t(x)\|_F \le \|\Psi_{\min}(x)\|_F$. Thus $y_t \to x$ and every cluster point belongs to $\partial\Phi(x)$ by its closed graph and has minimum norm. Uniqueness forces $\Psi_t(x) \to \Psi_{\min}(x)$. This proves existence of a canonical odd choice, separately from the alignment guarantee for every choice. □

The alternative route is worth recording because it is the one that fails. Adding the two subgradient inequalities $\Phi(A) \ge \Phi(B) + \langle \Psi(B), A - B \rangle$ and $\Phi(B) \ge \Phi(A) + \langle \Psi(A), B - A \rangle$ gives $\langle A - B, \Psi(A) - \Psi(B) \rangle \ge 0$ at once, so Ψ is monotone, and Proposition 1 converts that into alignment. But Proposition 1 needs an *odd* Ψ, so this route proves less. It is also worth saying that monotonicity is a two-line consequence of convexity and not its definition: monotone maps need not be gradients at all, a rotation by a right angle plus $\frac{1}{2}I$ having symmetric part $\frac{1}{2}I$ yet failing cyclic monotonicity, and the argument above never uses convexity of the ambient function, only that Ψ selects from a subdifferential.

For completeness, Euler's relation $\langle \Psi(x), x \rangle = \Phi(x)$ follows from the subgradient inequality at tx, $t > 0$, by taking t on both sides of one. Evenness is necessary for the universal guarantee in this one-homogeneous class, and its failure is selection independent. If $\Phi(e) > \Phi(-e)$ for some e, then $g = -\gamma e$ and $Z = \pm\zeta e$ with $\zeta > \gamma > 0$ give, by Euler at the two points $(\zeta - \gamma)e$ and $-(\zeta + \gamma)e$,

$$\langle g, \mathbb{E}\Psi(g+Z) \rangle = -\tfrac{\gamma}{2}\big[\Phi(e) - \Phi(-e)\big] < 0 \tag{137}$$

for *every* selection, so no choice of subgradient can rescue a non-even potential. The concrete instance is $\Phi(u) = |u| + u/2$ on $\mathbb{R}$: finite, convex and positively one-homogeneous but not even, with $\partial\Phi$ a singleton off the origin, $\Psi(u) = \mathrm{sign}(u) + 1/2$, so no odd selection exists at all, and $g = -1/2$ with $Z = \pm 1$ equally likely gives $g\,\mathbb{E}\Psi(g+Z) = -1/4$. If an odd selection exists, Euler gives $\Phi(-x) = \langle -x, \Psi(-x) \rangle = \langle x, \Psi(x) \rangle = \Phi(x)$; conversely, evenness gives the odd minimum-norm selection constructed above.

The conclusion is nonnegative, not strict, and the slack is attained inside the class of genuine norms. For the spectral norm on diagonal 2×2 matrices, $g = \mathrm{diag}(1, 0)$ with $Z = \pm\,\mathrm{diag}(0, 3)$ has minimal-norm subgradients $\mathrm{diag}(0, \pm 1)$ and alignment exactly zero, although $g \neq 0$ and Z is almost surely transverse to it. So the strictness in Corollary 52 is a curvature property of the quartic and must not be read as a property of even convex potentials in general. For $\varsigma \ge 1$ the Schatten norm (10) is such a

Φ: it is a norm, hence finite, convex, even and one-homogeneous. Off the set where a singular value vanishes its gradient is the spectral map $\Psi_{(\varsigma)}(P) = U\,\mathrm{diag}[(\sigma_i/\|P\|_\varsigma)^{\varsigma-1}]V^\top$. Note this is *not* a counterexample to Theorem 2: the normalization is by $\|P\|_\varsigma$ rather than $\|P\|_F$, so the theorem's hypothesis fails, and it is precisely that change of normalizer that couples the directions.

Proposition 54 (Schatten recovery and its endpoint)**.** *On the transverse task, the unit-radius response has slope $\mathcal{R}'_{(\varsigma)}(0) = (\varsigma-1)2^{1/\varsigma-1/2}$. For $1 < \varsigma \le 2$, this is the sharp global lower-envelope constant: $\mathcal{R}_{(\varsigma)}(t) \ge \mathcal{R}'_{(\varsigma)}(0)t/\sqrt{1+t^2}$ for $t > 0$. At $\varsigma = 1$ the slope is zero, agreeing with Proposition 30.*

Proof of Proposition 54. By Lemma 8 the input $xE + Z$ with $\|Z\|_F = \rho$ has two signed spectral components proportional to $(1 \pm t)/\sqrt{2}$ at $t = x/\rho$, on the two eigenspaces of the symmetric orthogonal S. At unit radius the Schatten normalizer is $\|P\|_\varsigma = 2^{-1/2}[(1+t)^\varsigma + |1-t|^\varsigma]^{1/\varsigma}$, so for $0 \le t < 1$ the projection along E is

$$\mathcal{R}_{(\varsigma)}(t) = \frac{1}{\sqrt{2}}\left[\Big(\frac{1+t}{\sqrt{2}\|P\|_\varsigma}\Big)^{\varsigma-1} - \Big(\frac{1-t}{\sqrt{2}\|P\|_\varsigma}\Big)^{\varsigma-1}\right]. \tag{138}$$

At $t = 0$ both arguments equal $1/\sqrt{2}$ divided by the common normalizer, so the bracket vanishes and the profile passes through the origin. Differentiating (138) at $t = 0$, the normalizer contributes nothing, because $(1+t)^\varsigma + (1-t)^\varsigma$ has zero derivative at $t = 0$ and hence $\|P\|_\varsigma$ is stationary there; only the two arguments move, each with derivative $\pm 1/(\sqrt{2}\|P\|_\varsigma)$. Collecting the factors gives

$$\mathcal{R}'_{(\varsigma)}(0) = (\varsigma-1)\,2^{(2-\varsigma)/(2\varsigma)} = (\varsigma-1)\,2^{1/\varsigma-1/2}. \tag{139}$$

□

So the slope is strictly positive for $\varsigma > 1$ and vanishes identically at $\varsigma = 1$. The endpoint is the nuclear norm, whose subgradient at a *full-rank* point is the polar factor, uniquely; full rank rather than distinctness of the singular values is the right condition, since repeated positive singular values do not obstruct differentiability there. At rank-deficient points we fix the reduced-polar selection, which sends zero singular values to zero. So (139) vanishing at $\varsigma = 1$ is the same fact as the dead zone of Proposition 30 rather than a separate limitation: for $0 < t < 1$ the two components have equal modulus after normalization and the projection cancels exactly.

The local slope is not automatically a global envelope constant, and the distinction is what the contraction theorem consumes. Write $\mathcal{R}_{(\varsigma)}$ for the profile in (138), extended to $t \ge 1$ by $\mathrm{sign}(1-t)|1-t|^{\varsigma-1}$ in the second term, with the convention $\mathrm{sign}(0)|0|^{\varsigma-1} = 0$ at $t = 1$.

Proposition 55 (Global Schatten envelope)**.** *For $\varsigma > 1$ the profile is strictly increasing on $[0,\infty)$ with $|\mathcal{R}_{(\varsigma)}| \le \sqrt{2}$, and*

$$\mathcal{R}_{(\varsigma)}(t) \;\ge\; \kappa_{(\varsigma)}\frac{t}{\sqrt{1+t^2}}, \qquad \kappa_{(\varsigma)} = \begin{cases} (\varsigma-1)2^{1/\varsigma-1/2}, & 1 < \varsigma \le 2, \\ 1/\sqrt{2}, & \varsigma > 2. \end{cases} \tag{140}$$

The first constant is sharp, being the slope at the origin. The second is a valid conservative choice and is not claimed sharp.

Proof. For $t \ne 1$, differentiating gives

$$\mathcal{R}'_{(\varsigma)}(t) = \frac{2\sqrt{2}(\varsigma-1)\,|1-t^2|^{\varsigma-2}}{\big((1+t)^\varsigma + |1-t|^\varsigma\big)^{2-1/\varsigma}} \;\ge\; 0, \tag{141}$$

so the profile increases, the possible singularity at $t = 1$ having order $|t-1|^{\varsigma-2}$, which is integrable for $\varsigma > 1$. Each output singular value of a norm gradient is at most one in modulus, so the projection is at most $\sqrt{2}$. For $1 < \varsigma \le 2$, concavity of $u \mapsto u^{\varsigma/2}$ applied to $u = (1+t)^2$ and $u = (1-t)^2$ gives $(1+t)^\varsigma + |1-t|^\varsigma \le 2\big(\frac{(1+t)^2+(1-t)^2}{2}\big)^{\varsigma/2} = 2(1+t^2)^{\varsigma/2}$, while $\varsigma - 2 \le 0$ and $|1-t^2| \le 1+t^2$ give $|1-t^2|^{\varsigma-2} \ge (1+t^2)^{\varsigma-2}$. Substituting both into (141),

$$\mathcal{R}'_{(\varsigma)}(t) \;\ge\; \frac{2\sqrt{2}(\varsigma-1)(1+t^2)^{\varsigma-2}}{2^{2-1/\varsigma}(1+t^2)^{\varsigma-1}} = (\varsigma-1)2^{1/\varsigma-1/2}(1+t^2)^{-3/2},$$

and integrating from 0 to t against $\int_0^t (1+\theta^2)^{-3/2} d\theta = t/\sqrt{1+t^2}$ gives the envelope. Its value at the origin is the slope, so the constant is sharp. For $\varsigma > 2$ and $0 < t \le 1$ put $a = 1+t$ and $b = 1-t$, so $a \ge b \ge 0$. Convexity of $u \mapsto u^{\varsigma-1}$ gives $a^{\varsigma-1} - b^{\varsigma-1} \ge (a-b)b^{\varsigma-2}$, and the sharper comparison we need is $a^{\varsigma-1} - b^{\varsigma-1} \ge (a-b)a^{\varsigma-2}$, valid because $a^{\varsigma-1} - b^{\varsigma-1} - (a-b)a^{\varsigma-2} = b\,a^{\varsigma-2} - b^{\varsigma-1} = b\,(a^{\varsigma-2} - b^{\varsigma-2}) \ge 0$ for $\varsigma \ge 2$. Also $a^{\varsigma} + b^{\varsigma} \le a^{\varsigma-2}(a^2+b^2)$. Since $a - b = 2t$ and $a^2 + b^2 = 2(1+t^2)$,

$$\frac{\mathcal{R}_{(\varsigma)}(t)}{t/\sqrt{1+t^2}} \;\ge\; \frac{2t\,a^{\varsigma-2}\sqrt{1+t^2}}{\sqrt{2}\,t\,\big(a^{\varsigma-2}(a^2+b^2)\big)^{(\varsigma-1)/\varsigma}} = \Big(\frac{a^{\varsigma-2}}{(a^2+b^2)^{\varsigma/2-1}}\Big)^{1/\varsigma} = \Big(\frac{a}{\sqrt{a^2+b^2}}\Big)^{(\varsigma-2)/\varsigma},$$

and $a/\sqrt{a^2+b^2} \ge 1/\sqrt{2}$ because $a \ge b$, so the ratio is at least $2^{-(\varsigma-2)/(2\varsigma)} = 2^{1/\varsigma - 1/2} \ge 1/\sqrt{2}$. For $t \ge 1$ monotonicity gives $\mathcal{R}_{(\varsigma)}(t) \ge \mathcal{R}_{(\varsigma)}(1) = 1/\sqrt{2} \ge (1/\sqrt{2})t/\sqrt{1+t^2}$. □

Using the local slope globally is unsound above $\varsigma = 2$: at $\varsigma = 3$ it would demand $\mathcal{R}_{(3)}(1) \ge 2^{1/3} = 1.26$, whereas $\mathcal{R}_{(3)}(1) = 1/\sqrt{2}$ exactly. Since an input singular-value ratio $a/b > 1$ leaves the map as $(a/b)^{\varsigma-1}$, flattening relative to the input happens exactly on $1 < \varsigma < 2$, with $\varsigma = 2$ neutral and $\varsigma > 2$ expanding ratios, so the design range and the range where the two constants agree are the same. Consequently a claim of strict recovery must be stated on $\varsigma > 1$, and the contraction of Theorem 51 transfers to the family with κ_{c} replaced by $\kappa_{(\varsigma)}$ of (140) for any fixed $\varsigma > 1$. The entry time t_c scales as $1/\kappa_{(\varsigma)}$, but the invariant radius $H = \beta^{m_{\mathrm{win}}} x_0 + \eta Q/(1-\beta)$ does *not*: at fixed step size and window it does not involve $\kappa_{(\varsigma)}$ at all, so the interval cannot be rescaled by $1/\mathcal{R}'_{(\varsigma)}(0)$. The tradeoff this exposes is intrinsic. Flattening increases as $\varsigma \downarrow 1$ while the restoring force in (139) decreases to zero, so the two cannot both be maximized, and where the balance falls is an empirical question that Section 7 answers on one task.

M.6 Deleting the shared term, and a three-dimensional witness

Removing $3\lambda\tau_4 X$ from (117) leaves $\Psi_{\mathrm{unc}}(P) = X - 4\lambda X X^\top X$, a *fixed pointwise* spectral map with $f(s) = s - 4\lambda s^3$, so Theorem 2 applies and it cannot be universally monotone for $r \ge 3$. On the task, Lemma 8 gives at $\lambda = 1/8$ the profile $t(1+3t^2)/(4(1+t^2)^{3/2}) > 0$ with $\mathcal{R}'(0) = f'(1/\sqrt{2}) = 1/4 > 0$, so it repairs the sign here. A positive response on the two-dimensional task therefore does not identify shared spectral coupling as necessary.

Proposition 56 (Explicit witness)**.** *At $\lambda = 1/8$, with $P = \mathrm{diag}(\frac{2}{3}, \frac{2}{3}, \frac{1}{3})$ and $H = \mathrm{diag}(19, 19, 5)/(3\sqrt{83})$, both of unit Frobenius norm,*

$$\langle H, D\Psi_{\mathrm{unc}}(P)[H] \rangle = -\tfrac{4}{747} < 0. \tag{142}$$

Proof. Here $f(s) = s - s^3/2$ and $s = (2/3, 2/3, 1/3)$ has $\|s\|_2 = 1$, so $X = P$ and Lemma 9 gives $A = D(I - ss^\top)$ with $D = \mathrm{diag}(f'(s_i)) = \mathrm{diag}(1 - \frac{3}{2}s_i^2) = \mathrm{diag}(\frac{1}{3}, \frac{1}{3}, \frac{5}{6})$. Then $b = (I - ss^\top)Ds = (-1, -1, 4)/27$. Taking the unnormalized direction $s - b$ and using Proposition 12's computation, $(s-b)^\top A(s-b) = -\|b\|^2 + b^\top Db = -4/729$, while $\|s-b\|_F^2 = 1 + \|b\|^2 = 83/81$. Normalizing divides the quadratic form by $83/81$, giving $-\frac{4}{729} \cdot \frac{81}{83} = -\frac{4}{747}$. Finally $s - b = (19, 19, 5)/27$, which normalized is H as stated. □

So the two-dimensional task tests a specific sign repair while a three-dimensional witness tests the universal structural guarantee, and neither substitutes for the other. That is why Section 7 runs $\widetilde{\Psi}$ and Ψ_{unc} as separate arms and measures the Jacobian's minimum eigenvalue rather than only a directional response.

M.7 Positive regularizer: a local formula

The theorems are stated at $\varepsilon = 0$. For the implemented normalization $P/(\|P\|_F + \varepsilon)$ on the transverse plane at noise radius R, the two signed spectral arguments are $u_\pm(x) = (R \pm x)/[\sqrt{2}(\sqrt{R^2 + x^2} + \varepsilon)]$, whose derivatives at $x = 0$ are $\pm 1/[\sqrt{2}(R + \varepsilon)]$, so the small-signal recovery coefficient is

$$h'_{K,\varepsilon}(0) = \mathbb{E}\left[\frac{f'_K\big(R/[\sqrt{2}(R+\varepsilon)]\big)}{R+\varepsilon}\right]. \tag{143}$$

Table 16: Scale-explicit local recovery estimates. Both $\mathcal{C}_K(0)$ and δ_*^{num} are independent of σ. The last column uses $\sigma = 1/16$ and $\varepsilon_*^{\mathrm{num}} = s\delta_*^{\mathrm{num}}$. Here δ_*^{num} is the first zero detected by the refined scan, not a certified threshold and not a guarantee for the EMA trajectory.

B	β	K	$\mathcal{C}_K(0)$	δ_*^{num}	$\varepsilon_*^{\mathrm{num}}$
1	0.9	1	-2.063	0.11961	9.90×10^{-4}
1	0.9	2	-2.337	0.20302	1.68×10^{-3}
1	0.9	3	2.823	0.19734	1.63×10^{-3}
1	0.9	5	-3.292	0.01103	9.13×10^{-5}
16	0.99	1	-2.089	0.11762	7.52×10^{-5}
16	0.99	5	-3.333	0.01083	6.92×10^{-6}

For $\varepsilon > 0$ the exchange of derivative and expectation is supported by the bounded denominator; at $\varepsilon = 0$ it uses the inverse moment bound (70). This is closer to an implementation than a full regularized transition statement would be, and it makes the question sharp: the sign of the recovery coefficient depends on ε relative to the momentum noise scale, through the argument $R/[\sqrt{2}(R+\varepsilon)]$, which decreases from $1/\sqrt{2}$ toward 0 as ε/R grows.

Units and scale equivariance. The archived E1 configuration uses $\sigma = 1$; its radii and regularizers are in those units. E2 instead uses $\sigma = 1/16$, the largest scale allowed by Definition 7. A derivative with respect to the physical signal x cannot be transferred between these experiments unchanged. Put $U = R/s$, $\delta = \varepsilon/s$ and $H_{K,\delta}(y) = h_{K,s\delta}(sy)$. Then the exact, dimensionless relation is

$$\mathcal{C}_K(\delta; B, \beta) := H'_{K,\delta}(0) = s\, h'_{K,\varepsilon}(0) = \mathbb{E}\left[\frac{f'_K\big(U/[\sqrt{2}(U+\delta)]\big)}{U+\delta}\right]. \tag{144}$$

The law of U is independent of σ, so $\mathcal{C}_K$ depends on ε only through ε/s. Equivalently, $h^{(\sigma)}_{K,\varepsilon}(x) = h^{(1)}_{K,\varepsilon/\sigma}(x/\sigma)$. The previously quoted -15.58 is the archived estimate of the derivative with respect to x at $\sigma = 1$, $B = 1$, $\beta = 0.9$, $K = 1$, $\varepsilon = 0$. At $\sigma = 1/16$ the same derivative is approximately -249.24; in either convention $\mathcal{C}_1(0; 1, 0.9) \approx -2.0633$. This agrees with $\mathbb{E}[R^{-1}] \geq 1/s$ and the coefficient $f'_1(1/\sqrt{2}) = -1.178625$.

Numerical resolution and nonmonotone dependence on the regularizer. Table 16 uses adaptive integration of the smooth radial density ratio $f_U(u)/u$ multiplied by $u/(u+\delta)$. This resolves the boundary layer near $u = 0$ that an unmodified radial trapezoid misses when δ is smaller than one grid spacing. Doubling the radial, angular and Fourier grids from $(3000, 720, 131072)$ to $(6000, 1440, 262144)$ changes the dimensionless slopes by at most 5.1×10^{-4} on the reported audit grid. These are convergence checks, not interval error bounds. The comparison reproduces every E1 trapezoid value before applying the correction.

The sign need not vary monotonically with ε. At $B = 1, \beta = 0.9$, the refined scan detects zeros for $K = 2$ near $\delta = 0.2030, 0.6452, 1.9755$, and for $K = 3$ near $0.1973, 0.7129, 2.3525, 7.0248$. In particular, the old endpoint-only sweep missed two sign changes between adjacent sampled regularizers; its absence-of-crossing statement for these depths is withdrawn. Increasing ε can therefore remove a particular local negative slope, but this calculation supplies neither a monotone safety threshold nor a residence theorem. The scale that must accompany any implementation value is ε/s, together with (B, β, K).

M.8 A POSITIVE REGULARIZER THAT STAYS INSIDE THE CONVEX CLASS

Equation (143) prices ε for the reference iteration but makes no positive-ε map safe. The obstruction is specific and worth isolating first. Replacing $\|P\|_F$ by $\|P\|_F + \varepsilon$ in (117) destroys the potential: on the diagonal slice $P = \mathrm{diag}(x, y)$ with $r = \sqrt{x^2 + y^2}$, the resulting field G has

$$\partial_y G_{11} - \partial_x G_{22} = \frac{12\lambda\varepsilon xy(x^2 - y^2)}{(\varepsilon + r)^5 r}, \tag{145}$$

which at $\lambda = 1/8$ is $-7/2160$ at $(x, y, \varepsilon) = (3, 4, 1)$, so that field is not the gradient of anything and Proposition 5 does not reach it. The regularized radius is not by itself the culprit, and it is worth being

precise about what is: at $\lambda = 0$ the curl vanishes identically and $P/(\|P\|_F + \varepsilon)$ *is* a gradient, namely of $r - \varepsilon \log(\varepsilon + r)$, which is convex, since its radial second derivative $\varepsilon/(\varepsilon + r)^2$ is positive, and even. What obstructs is the interaction between that radius and the quartic term, and (145) is exactly proportional to λ. Adding $3\lambda(\varepsilon/R)^4 X$ does not help either: it moves the field by a radial multiple of P and leaves (145) unchanged. We found no monotonicity witness for the substituted map and claim none; the point is that it forfeits the proof. Lifting instead of substituting keeps everything.

Proposition 57 (Affine-lift regularization)**.** *For $0 \le \lambda < 2/9$ and $\varepsilon > 0$ put*

$$\Phi_{\lambda,\varepsilon}(P) = \Phi_\lambda \begin{pmatrix} P & 0 \\ 0 & \varepsilon \end{pmatrix} - (1 - \lambda)\varepsilon.$$

Then $\Phi_{\lambda,\varepsilon}$ is finite, convex, even and smooth on all of $\mathbb{H}$ including the origin, with $\Phi_{\lambda,\varepsilon}(0) = 0$. Writing $R_\varepsilon = \sqrt{\|P\|_F^2 + \varepsilon^2}$, $X = P/R_\varepsilon$ and $z = \varepsilon/R_\varepsilon$,

$$\widetilde{\Psi}_{\lambda,\varepsilon}(P) = \nabla\Phi_{\lambda,\varepsilon}(P) = \big(1 + 3\lambda[\tau_4(X) + z^4]\big)X - 4\lambda X X^\top X, \tag{146}$$

which is odd, bounded, monotone and still two matrix products. For every P and every $H \neq 0$,

$$D^2\Phi_{\lambda,\varepsilon}(P)[H, H] \;\ge\; \frac{\kappa_\lambda \varepsilon^2}{R_\varepsilon^3}\|H\|_F^2 > 0, \tag{147}$$

so for every $g \neq 0$ and every centrally symmetric Z with $\|Z\|_F \le \varrho$,

$$\langle g, \mathbb{E}\widetilde{\Psi}_{\lambda,\varepsilon}(g + Z)\rangle \;\ge\; \frac{\kappa_\lambda \varepsilon^2 \|g\|_F^2}{\big[(\|g\|_F + \varrho)^2 + \varepsilon^2\big]^{3/2}} > 0. \tag{148}$$

On the task of Definition 7 with $\|Z\|_F = \rho > 0$, $a > 0$ and $R_\varepsilon = \sqrt{a^2 + \rho^2 + \varepsilon^2}$,

$$\kappa_\lambda \frac{a}{R_\varepsilon} \;\le\; \langle E, \widetilde{\Psi}_{\lambda,\varepsilon}(aE + Z)\rangle \;\le\; \Lambda_\lambda \frac{a}{R_\varepsilon}, \tag{149}$$

by oddness the mirror statement holding for $a < 0$, and neither constant can be improved uniformly in ε.

Proof. By Theorem 50 Φ_λ is an orthogonally invariant norm, and $P \mapsto \mathrm{blockdiag}(P, \varepsilon)$ is affine, so the composition is convex and subtracting a constant preserves that. The lifted matrix has Frobenius norm $R_\varepsilon \ge \varepsilon > 0$, so it never meets the origin, the only point where Φ_λ fails to be smooth; hence smoothness in P everywhere, the origin included. Since $\mathrm{blockdiag}(-P, \varepsilon) = \mathrm{blockdiag}(-I_m, 1)\,\mathrm{blockdiag}(P, \varepsilon)$ with the left factor orthogonal, orthogonal invariance gives evenness. At $P = 0$ the lifted matrix is $\varepsilon\, e_{m+1} e_{n+1}^\top$, whose normalization has $\tau_4 = 1$, so $\Phi_\lambda = \varepsilon(1 - \lambda)$ and the constant cancels.

For the gradient, the chain rule through an affine map reads off the P-block of $\nabla\Phi_\lambda$ at the lifted matrix. There the normalized argument is $Y = \mathrm{blockdiag}(X, z)$, so $\tau_4(Y) = \tau_4(X) + z^4$ and $YY^\top Y = \mathrm{blockdiag}(XX^\top X, z^3)$, and (117) gives (146). Oddness follows from evenness, X being odd and $\tau_4(X)$ even. Boundedness holds because a gradient of a norm at a nonzero point lies in the dual unit ball, and monotonicity because it is the gradient of a convex function.

For (147), apply (118) at the lifted matrix Q in the direction $\mathrm{blockdiag}(H, 0)$, whose component orthogonal to Q has squared norm $\|H\|_F^2 - \langle P, H\rangle^2/R_\varepsilon^2$. Cauchy–Schwarz gives $\langle P, H\rangle^2 \le \|P\|_F^2\|H\|_F^2 = (R_\varepsilon^2 - \varepsilon^2)\|H\|_F^2$, so that squared norm is at least $\varepsilon^2\|H\|_F^2/R_\varepsilon^2$; dividing by $\|Q\|_F = R_\varepsilon$ and multiplying by κ_λ closes it. For (148), Proposition 1's identity turns the alignment into $\frac{1}{2}\mathbb{E}\langle g, \widetilde{\Psi}_{\lambda,\varepsilon}(Z + g) - \widetilde{\Psi}_{\lambda,\varepsilon}(Z - g)\rangle$, and for each fixed z the segment $s \mapsto z + sg$, $s \in [-1, 1]$, gives $\langle g, \widetilde{\Psi}_{\lambda,\varepsilon}(z + g) - \widetilde{\Psi}_{\lambda,\varepsilon}(z - g)\rangle = \int_{-1}^{1} D^2\Phi_{\lambda,\varepsilon}(z + sg)[g, g]\, ds$, since the integrand is now continuous everywhere and no subgradient jump arises. Bounding $\|z + sg\|_F \le \varrho + \|g\|_F$ inside (147) and halving gives the claim.

For (149), Lemma 8's algebra makes the lifted matrix's singular values $\rho|1 \pm t|/\sqrt{2}$ together with ε, where $t = a/\rho$; writing $u = \varepsilon/\rho$ and evaluating (146),

$$\frac{R_\varepsilon}{a}\langle E, \widetilde{\Psi}_{\lambda,\varepsilon}(aE + Z)\rangle - \kappa_\lambda = \frac{\lambda\big(8t^4 + 14t^2u^2 + 20t^2 + 15u^4 + 6u^2\big)}{2(t^2 + u^2 + 1)^2}, \tag{150}$$

$$\Lambda_\lambda - \frac{R_\varepsilon}{a}\langle E, \widetilde{\Psi}_{\lambda,\varepsilon}(aE+Z)\rangle = \frac{\lambda\big(7t^4+16t^2u^2+10t^2+24u^2+15\big)}{2(t^2+u^2+1)^2}, \tag{151}$$

both numerators having nonnegative coefficients. The first vanishes only at $t = u = 0$, so κ_λ is approached as a/ρ and ε/ρ tend to zero together and no larger floor holds for all ε; the second tends to zero as $u \to \infty$ at fixed t, where $\widetilde{\Psi}_{\lambda,\varepsilon}(P) \to (1+3\lambda)P/\varepsilon$ is normalized descent with gain Λ_λ, so the ceiling is approached too. □

Three readings, then the scope. First, the floor in (149) is the same κ_λ as at $\varepsilon = 0$, and the response is $\kappa_\lambda a/\sqrt{a^2+\rho^2+\varepsilon^2}$ rather than $\kappa_\lambda a/\sqrt{a^2+\rho^2}$, so Theorem 51 runs verbatim with the noise radius ϱ inflated to $\sqrt{\varrho^2+\varepsilon^2}$ and nothing else changed. The task ceiling does not survive: $\Lambda^{\mathrm{task}} = 23/24$ holds only at $\varepsilon = 0$, and at $\varepsilon = 2\rho$ the ratio already reaches 1.0975, which is why (149) carries the general Λ_λ. Second, (148) is strictly positive with no transverse-energy hypothesis at all, unlike (125), which vanishes when Z is almost surely collinear with g. The nonsmoothness at the origin is exactly what forced that hypothesis, and ε is what removes it, at the price of an $O(\varepsilon^2)$ floor that degrades as $\varepsilon \downarrow 0$. Third, ε shifts the small-signal slope nonmonotonically rather than only shrinking it: at $\lambda = 1/8$ the derivative of the response in t at $t = 0$ is $(22u^4+20u^2+7)/[16(1+u^2)^{5/2}]$, which is $7/16$ at $u = 0$, rises to 0.5472 at $u = (7/11+3\sqrt{34}/22)^{1/2} = 1.1965$, returns to $7/16$ at $u = 2.5029$ and decays as $11/(8u)$, so a moderate regularizer helps here and a large one does not.

This is a construction of the present revision, not a method the experiments measure. No arm of Section 7 implements (146), no reported number depends on it, and it does not extend Definition 7's algorithm, where a positive ε remains outside the theorems. What it settles is a narrower question: whether the convex route to safety is compatible with a positive regularizer at all. It is, provided the regularizer enters as an affine lift rather than as an added constant in the denominator, and (145) is the reason the distinction is not cosmetic.

N A CONVEX SPECTRAL NORM WITH RANK-INDEPENDENT FLATTENING

The quartic update is inexpensive, but it flattens less when we replicate the same spectrum in a larger matrix. We therefore study a smoothed version of the nuclear norm whose output ratios stay the same under this replication. The smoothing scale depends on the whole matrix, and its derivative must be included in the update. We prove the norm property first, then give bounds on how strongly the update responds to a change.

Theorem 58 (Homogeneous smoothed nuclear norm)**.** *Let $r = \min(m,n)$, $\kappa > 0$, $\tau = \kappa/r$ and*

$$\Phi^{\mathrm{sn}}_\kappa(P) = \sum_{i=1}^{r} \sqrt{\sigma_i(P)^2 + \tau\|P\|_F^2}, \qquad \Phi^{\mathrm{sn}}_\kappa(0) = 0. \tag{152}$$

This is an orthogonally invariant norm, smooth away from zero, whose gradient is odd, zero-homogeneous, and monotone. Transpose P if necessary so it has r rows. For $X = P/\|P\|_F$, $A = XX^\top + \tau I_r$, and $Y = A^{-1/2}$, its gradient is

$$\Psi^{\mathrm{sn}}_\kappa(P) = (Y + \tau\,\mathrm{tr}(Y) I_r)X, \qquad \Psi^{\mathrm{sn}}_\kappa(0) = 0. \tag{153}$$

For every $P \neq 0$ and H, it satisfies

$$\frac{\kappa\sqrt{r}}{\sqrt{1+\kappa}\|P\|_F}\|\Pi_P^\perp H\|_F^2 \;\le\; D^2\Phi^{\mathrm{sn}}_\kappa(P)[H,H] \;\le\; \frac{\sqrt{r}(1+\kappa)}{\sqrt{\kappa}\|P\|_F}\|\Pi_P^\perp H\|_F^2. \tag{154}$$

The bounds are dimension-explicit, not claimed sharp. Every positive singular-value ratio is strictly compressed, with order preserved.

Proof. For $s \in \mathbb{R}^r$, write $\varphi(s) = \sum_i \sqrt{s_i^2 + \tau\|s\|_2^2}$. Each summand is the Euclidean norm of a linear injective image of s; their sum is a permutation-invariant, unconditional norm. Its spectral lift is a matrix norm: one may use the symmetric-gauge characterization, or combine the dual representation of φ with von Neumann's trace inequality. Since the matrix under every square root in (152) is positive definite at $P \neq 0$, the trace function is smooth even at repeated and zero singular values.

Differentiating $\mathrm{tr}(PP^\top + \tau\|P\|_F^2 I)^{1/2}$ gives (153). Convexity gives monotonicity; evenness gives oddness of the unique gradient away from zero. Zero is a subgradient at the origin.

For the curvature bounds, normalize $\|s\|_2 = 1$, put $r_i = \sqrt{s_i^2 + \tau}$ and $c = \tau \sum_i r_i^{-1}$. The output singular values are $d_i = s_i(r_i^{-1} + c)$. Homogeneity makes the radial Hessian zero. For a diagonal angular variation h with $h^\top s = 0$, direct differentiation gives the exact identity

$$h^\top D^2\varphi(s)h = c\|h\|_2^2 + \tau \sum_i \frac{h_i^2}{r_i^3}. \tag{155}$$

Jensen's inequality and $r_i \geq \sqrt{\tau}$ imply

$$\frac{\kappa\sqrt{r}}{\sqrt{1+\kappa}} \leq c \leq \sqrt{\kappa r}, \qquad 0 \leq \frac{\tau}{r_i^3} \leq \frac{1}{\sqrt{\tau}}.$$

The other orthogonal blocks of the spectral Hessian have coefficients $(d_i - d_j)/(s_i - s_j)$, $(d_i + d_j)/(s_i + s_j)$, and, for a rectangular null block, d_i/s_i. At the fixed spectrum the common c adds c to each coefficient. The function $u \mapsto u/\sqrt{u^2 + \tau}$ is odd, increasing, and has derivative in $[0, 1/\sqrt{\tau}]$. All the remaining coefficients therefore lie in $[c, c + 1/\sqrt{\tau}]$. These blocks, together with (155), prove (154), first for distinct positive singular values and then by continuity everywhere away from $P = 0$.

For $s_i > s_j > 0$, the function $u \mapsto u/\sqrt{u^2 + \tau} + cu$ is strictly increasing, whereas its ratio to u is strictly decreasing. Consequently $1 < d_i/d_j < s_i/s_j$. □

Corollary 59 (Exact invariance under spectral replication)**.** *For an integer $k \geq 1$, put $P^{[k]} = k^{-1/2}\,\mathrm{blockdiag}(P, \ldots, P)$ with k blocks, using $r' = kr$ in* (152)*. Then*

$$\Psi_\kappa^{\mathrm{sn}}(P^{[k]}) = \mathrm{blockdiag}\big(\Psi_\kappa^{\mathrm{sn}}(P), \ldots, \Psi_\kappa^{\mathrm{sn}}(P)\big).$$

Thus every output singular-value ratio is exactly invariant under replication, while the input stable rank grows by a factor k. In contrast, the quartic map's output ratios tend to the input ratios as $k \to \infty$.

Proof. Replication preserves the Frobenius norm, divides each normalized singular value by $\sqrt{k}$, and replaces τ by τ/k. Hence r_i is divided by $\sqrt{k}$, each inverse root is multiplied by $\sqrt{k}$, and the trace has k identical blocks. Substitution in (153) proves the identity. For the quartic map, τ_4 and every s_i^2 are divided by k, so the shared and individual corrections both tend to zero. □

A quantitative polar approximation. Let $z_i = \sqrt{r}\,s_i$ and $\overline{u} = r^{-1}\sum_j (z_j^2 + \kappa)^{-1/2}$. The output value is exactly

$$d_i = \frac{z_i}{\sqrt{z_i^2 + \kappa}} + \kappa\overline{u}\,z_i. \tag{156}$$

Because $\overline{u} \leq \kappa^{-1/2}$, any pair of singular values with $0 < \ell \leq z_j \leq z_i \leq u$ satisfies the dimension-independent bound

$$1 \leq \frac{d_i}{d_j} \leq (1 + \sqrt{\kappa}\,u)\sqrt{1 + \kappa/\ell^2}.$$

For fixed positive normalized amplitudes, ratios tend to one as $\kappa \downarrow 0$. This bound and replication invariance address the magnitude of flattening, whereas strict compression alone does not. Stronger flattening also weakens the uniform curvature lower bound: the construction retains a quantitative tradeoff.

Alignment and computation. Proposition 5 applies directly. Replacing the quartic curvature constant by $\kappa\sqrt{r}/\sqrt{1+\kappa}$ in the symmetric-segment argument gives the corresponding strictly positive alignment bound whenever the noise has a component transverse to a nonzero signal. No spectral gap is needed. The reference implementation uses the eigendecomposition of the smaller Gram matrix in float64, with a cubic factorization cost. The Schatten update can use the same factorization route; Appendix O separates its cutoff and roundoff errors from exact-map regularity. Measured wall time is reported with the training comparison. The optional Newton–Schulz inverse-square-root implementation approximates this same map. For commuting positive iterates in exact arithmetic, an

operator residual $\|I - A\widehat{Y}^2\| \le \epsilon < 1$ bounds the relative scalar inverse-root error by $\epsilon/(1+\sqrt{1-\epsilon})$. A numerical residual is not by itself a proof that an approximate map is monotone as a function of its input. All reported training results use the eigensolver reference, and all exact monotonicity statements concern (153).

Proposition 60 (Recovery at a matched output scale)**.** *On the unit-radius transverse task, let q_{sn} and $q_{(\varsigma)}$ denote the projected responses of the smoothed nuclear and Schatten maps. At zero signal,*

$$\frac{q'_{\mathrm{sn}}(0)}{\|\Psi^{\mathrm{sn}}_{\kappa}(Z)\|_F} = \frac{\kappa(\kappa+2)}{(1+\kappa)^2} = 1-(1+\kappa)^{-2}, \qquad \frac{q'_{(\varsigma)}(0)}{\|\Psi_{(\varsigma)}(Z)\|_F} = \varsigma - 1. \tag{157}$$

The same normalized slopes hold for a traceless symmetric perturbation of a flat rank-r spectrum. They compare local recovery after rescaling each map to unit output norm at that noise input; no training-loss comparison is asserted.

Proof. At $s_i = 1/\sqrt{r}$, the shared coefficient in (155) is $c = \kappa\sqrt{r}/\sqrt{1+\kappa}$ and $\tau/(s_i^2+\tau)^{3/2} = \kappa\sqrt{r}/(1+\kappa)^{3/2}$. Thus every traceless symmetric unit perturbation has response $\kappa\sqrt{r}(\kappa+2)/(1+\kappa)^{3/2}$, whereas the output norm is $\sqrt{r(1+\kappa)}$. For Schatten the corresponding derivative and norm are $(\varsigma-1)r^{1/\varsigma-1/2}$ and $r^{1/\varsigma-1/2}$. The transverse reduction is the rank-two case in its rotated singular frame. □

O Gram-implementable convex spectral maps, gain at zero and Lipschitz constants

Schatten-1.05 is the natural baseline for the smoothed nuclear map because both can be computed through a Gram matrix. Sharing this computation does not make them equally sensitive to small singular values. We first derive the exact formulas, then study their response near zero singular values and check the errors made by finite-precision implementations. Training loss and measured runtime are reported separately. Throughout, $P \in \mathbb{R}^{r\times n}$ has been transposed so that $r = \min(m,n)$ is its number of rows, $P = U\,\mathrm{diag}(s)V^\top$ is a thin singular value decomposition with $U \in \mathbb{R}^{r\times r}$ orthogonal, $V \in \mathbb{R}^{n\times r}$ with $V^\top V = I_r$ and $s_1 \ge \cdots \ge s_r \ge 0$, and $PP^\top = U\,\mathrm{diag}(\mu)U^\top$ with $\mu_i = s_i^2$. We write $X = P/\|P\|_F$, $\hat{s} = s/\|P\|_F$ and $\hat{\mu}_i = \hat{s}_i^2$ for the normalized quantities, reserve λ for the quartic parameter and ς for the Schatten exponent, and put $\tau = \kappa/r$ as in (152). For a function c on the spectrum of $PP^\top$, $c(PP^\top) = U\,\mathrm{diag}(c(\mu_i))U^\top$ is the primary matrix function (Bhatia, 1997). A function $\varphi : \mathbb{R}^r \to \mathbb{R}$ is absolutely symmetric if it is invariant under permutations and sign changes of its coordinates; a symmetric gauge is an absolutely symmetric norm. The supplementary Gram-map verification suite supplies the implementations, numerical results and machine-readable records for this appendix.

O.1 Gram representations for differentiable spectral potentials

Lemma 61 (Odd spectral functions are continuous)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be continuous and odd. For $Q \in \mathbb{R}^{r\times n}$ with thin singular value decomposition $Q = U\,\mathrm{diag}(\sigma)V^\top$ put $M_g(Q) = U\,\mathrm{diag}(g(\sigma_i))V^\top$. Then M_g is well defined and continuous on $\mathbb{R}^{r\times n}$.*

Proof. Fix $M > 0$. By the Weierstrass theorem there are polynomials $q_k \to g$ uniformly on $[-M, M]$; their odd parts $q_k^{\mathrm{o}}(t) = (q_k(t) - q_k(-t))/2$ are odd polynomials and, since g is odd, $\sup_{[-M,M]} |q_k^{\mathrm{o}} - g| \le \sup_{[-M,M]} |q_k - g|$. For an odd polynomial $q(t) = \sum_j a_j t^{2j+1}$ one has $M_q(Q) = \sum_j a_j (QQ^\top)^j Q$ for every thin singular value decomposition, as in (1), so M_q is a polynomial in the entries of Q. On $\{\|Q\|_{\mathrm{op}} \le M\}$, for any fixed decomposition used to define $M_g(Q)$, $\|M_g(Q) - M_{q_k^{\mathrm{o}}}(Q)\|_F^2 = \sum_i (g(\sigma_i) - q_k^{\mathrm{o}}(\sigma_i))^2 \le r \sup_{[0,M]} |g - q_k^{\mathrm{o}}|^2 \to 0$ uniformly. Hence $M_g(Q)$ is the uniform limit of well-defined continuous maps, so it does not depend on the decomposition and is continuous on that ball; M was arbitrary. □

Proposition 62 (Gram implementability)**.** *Let $\varphi : \mathbb{R}^r \to \mathbb{R}$ be convex and absolutely symmetric, $\Phi = \varphi \circ \sigma$ on $\mathbb{R}^{r\times n}$, and $P \ne 0$ as above.*

(i) If φ is differentiable at s with gradient g, then $g_i \geq 0$, $g_i = 0$ whenever $s_i = 0$, $(g_i - g_j)(s_i - s_j) \geq 0$ and $g_i = g_j$ whenever $s_i = s_j$, and $G := U \operatorname{diag}(g) V^\top \in \partial\Phi(P)$. Any function c on the spectrum of $PP^\top$ with $c(\mu_i) = g_i/s_i$ for $s_i > 0$, with any finite value at $\mu = 0$, satisfies

$$G = c(PP^\top)\, P. \tag{158}$$

When Φ is differentiable at P, this G is its gradient. The coefficient function can depend on the entire spectrum of P, as the Schatten normalizer illustrates.

(ii) The smoothed nuclear potential (152) *and the quartic potential* (117) *are differentiable on $\mathbb{R}^{r\times n} \setminus \{0\}$ with gradient G (Theorems 58 and 50). For $1 < \varsigma < \infty$ the Schatten potential* (10) *is differentiable at every $P \neq 0$, including every rank-deficient P, with $\nabla\Phi_{(\varsigma)}(P) = \Psi_{(\varsigma)}(P)$, and $\Psi_{(\varsigma)}$ is continuous on $\mathbb{R}^{r\times n} \setminus \{0\}$.*

(iii) Relative to X, so that $\Psi(P) = \hat{c}(XX^\top)X$, the coefficients are

$$\begin{aligned}
&\hat{c}(\mu) = 1 && \textit{(normalized descent)},\\
&\hat{c}(\mu) = \textstyle\sum_j a_j^{(K)} \mu^j, \quad f_K(t) = \sum_j a_j^{(K)} t^{2j+1} && \textit{(reference } \Psi_K\textit{)},\\
&\hat{c}(\mu) = 1 + 3\lambda\tau_4(X) - 4\lambda\mu && \textit{(quartic)},\\
&\hat{c}(\mu) = (\mu+\tau)^{-1/2} + \tau \operatorname{tr}\big[(XX^\top + \tau I)^{-1/2}\big] && \textit{(smoothed nuclear)},\\
&\hat{c}(\mu) = \mu^{(\varsigma-2)/2} / \|\hat{s}\|_\varsigma^{\varsigma-1} \ \ (\mu > 0), \quad \|\hat{s}\|_\varsigma^\varsigma = \operatorname{tr}\big[(XX^\top)^{\varsigma/2}\big] && \textit{(Schatten-}\varsigma\textit{)}.
\end{aligned} \tag{159}$$

In particular, the smoothed nuclear and Schatten maps can both use one eigendecomposition of the $r \times r$ Gram matrix $XX^\top$, $O(r)$ scalar coefficient operations, and two products of an $r \times r$ matrix with an $r \times n$ matrix. Forming the Gram matrix also costs a matrix product. The quartic and reference maps need matrix products only. For a general φ, evaluating $g = \nabla\varphi(s)$ has its own cost; the representation alone does not bound that cost or establish a stable numerical algorithm.

Proof. (i) For each i the function $t \mapsto \varphi(s + (t - s_i)e_i)$ is even and convex, hence nondecreasing on $[0, \infty)$; so $g_i \geq 0$ when $s_i > 0$, and $g_i = 0$ when $s_i = 0$ because an even differentiable function has zero derivative at the origin. Let Π swap coordinates i and j. Convexity and $\varphi(\Pi s) = \varphi(s)$ give $0 \geq \langle g, \Pi s - s\rangle = (s_i - s_j)(g_j - g_i)$, which is the ordering claim; differentiating $\varphi \circ \Pi = \varphi$ at a point with $s_i = s_j$, where $\Pi s = s$, gives $\Pi g = g$, hence $g_i = g_j$. Consequently g is nonnegative and ordered like s, and the singular values of G are exactly g. For any $Q \in \mathbb{R}^{r\times n}$, convexity of φ gives $\Phi(Q) \geq \varphi(s) + \langle g, \sigma(Q) - s\rangle$; von Neumann's trace inequality (Bhatia, 1997, Ch. IV) gives $\langle G, Q\rangle \leq \sum_i \sigma_i(G)\sigma_i(Q) = \langle g, \sigma(Q)\rangle$; and $\langle G, P\rangle = \sum_i g_i s_i = \langle g, s\rangle$ because G and P share their singular frames. Hence $\Phi(Q) \geq \Phi(P) + \langle G, Q - P\rangle$, that is $G \in \partial\Phi(P)$. For (158), $U \operatorname{diag}(c(\mu_i)) U^\top U \operatorname{diag}(s) V^\top = U \operatorname{diag}(c(\mu_i)s_i) V^\top = U \operatorname{diag}(g) V^\top$, since $c(\mu_i)s_i = g_i$ when $s_i > 0$ and both sides vanish when $s_i = 0$. Because $g_i = g_j$ whenever $s_i = s_j$, c is constant on every eigenspace of $PP^\top$, so $c(PP^\top)$ does not depend on the eigenvectors chosen inside an eigenspace.

(ii) The first two statements are the cited theorems together with (158). For the Schatten potential with $\varsigma > 1$, the function $\varphi(s) = \|s\|_\varsigma$ is continuously differentiable away from the origin, where $\partial_i\varphi(s) = |s_i|^{\varsigma-1} \operatorname{sign}(s_i) / \|s\|_\varsigma^{\varsigma-1}$. By (i), at every $Q \neq 0$ the matrix $G_Q := \|\sigma(Q)\|_\varsigma^{1-\varsigma} M_h(Q)$, $h(t) = |t|^{\varsigma-1}\operatorname{sign}(t)$, belongs to $\partial\Phi_{(\varsigma)}(Q)$, and $Q \mapsto G_Q$ is continuous on $\mathbb{R}^{r\times n} \setminus \{0\}$ by Lemma 61 and the continuity of σ. For $P, Q \neq 0$ the two subgradient inequalities give

$$0 \ \leq \ \Phi_{(\varsigma)}(Q) - \Phi_{(\varsigma)}(P) - \langle G_P, Q - P\rangle \ \leq \ \langle G_Q - G_P, Q - P\rangle \ \leq \ \|G_Q - G_P\|_F \|Q - P\|_F,$$

and the right side is $o(\|Q - P\|_F)$ as $Q \to P$ by the continuity of $Q \mapsto G_Q$. So $\Phi_{(\varsigma)}$ is differentiable at P with gradient G_P, which is (10) with zero singular values sent to zero.

(iii) Substitute g_i/s_i. For $\varphi = \|\cdot\|_2$, $g = s/\|s\|_2$. For the reference, $f_K(\hat{s}_i)/\hat{s}_i = \sum_j a_j^{(K)} \hat{s}_i^{2j}$; the map is not a gradient, but (1) is the same identity. The quartic follows from (117) written as $[(1 + 3\lambda\tau_4)I - 4\lambda XX^\top]X$, the smoothed nuclear from (153), and the Schatten coefficient is $\hat{s}_i^{\varsigma-2} / \|\hat{s}\|_\varsigma^{\varsigma-1}$; all three maps are zero-homogeneous, so evaluating at X loses nothing. The cost count is read off from (158): $\hat{c}(XX^\top)X = U\big(\operatorname{diag}(\hat{c}(\hat{\mu}_i))(U^\top X)\big)$. □

The differentiability qualification matters. At $P = \operatorname{diag}(1, 0)$, for example, $\operatorname{diag}(1, a)$ is a nuclear-norm subgradient for every $|a| \le 1$, but for $a \neq 0$ it cannot equal $c(PP^\top)P$. Thus the proposition is not a statement that every subgradient selection of every orthogonally invariant convex potential has this form.

Remark 63 (The coefficient at zero, and finite precision)**.** For the reference, quartic and smoothed nuclear maps the coefficient in (159) extends continuously to $\mu = 0$, with values a_0^K, $1 + 3\lambda\tau_4(X)$ and $\tau^{-1/2} + \tau \operatorname{tr} Y$; these are the gains at zero of Proposition 66. For $\varsigma < 2$ the Schatten coefficient $\mu^{(\varsigma-2)/2}/\|\hat{s}\|_\varsigma^{\varsigma-1}$ diverges as $\mu \downarrow 0$ while the product $\hat{c}(\hat{\mu}_i)\hat{s}_i = \hat{s}_i^{\varsigma-1}/\|\hat{s}\|_\varsigma^{\varsigma-1}$ still tends to zero: this is why the map is continuous, by Proposition 62(ii), yet not Lipschitz, by Proposition 67. The historical Gram implementation assigns coefficient zero to estimated singular values at or below $10^{-8}s_{\max}$ and removes them from the normalizer; the SVD implementation instead uses $10^{-12}s_{\max}$. These are different truncated approximations when singular values lie between their cutoffs. Neither hard threshold inherits the exact map's convex-gradient guarantee. The initial numerical check compares it with the SVD formula on shapes up to 384×1536 and $\varsigma \in \{1.05, 1.25, 1.5\}$, and the maximum relative difference is 2.0×10^{-7} at condition number 10^5 and 1.5×10^{-11} at 10^3. The Gram route squares the condition number, since an absolute eigenvalue error $u\mu_{\max}$ is a relative error $u(s_{\max}/s_i)^2$ in μ_i (Proposition 69). In the matched-training study of Appendix W the Gram Schatten-1.05 arm costs $1.15\times$ the reference's wall time against $1.14\times$ for the smoothed nuclear arm. The exclusion of Schatten-1.05 from the earlier training grid, which rested on an SVD implementation, was therefore an implementation artefact and not a property of the map. This cost comparison does not establish equal fidelity near numerical rank deficiency.

Proposition 64 (A hard relative cutoff need not preserve monotonicity)**.** *Fix $p > 1$ and $0 < \theta < 1$. Replace the singular values at or below $\theta s_{\max}$ by zero and evaluate the Schatten gradient using the retained singular values, including its normalizer. This truncated map is not monotone, even on positive diagonal 2×2 matrices in exact arithmetic.*

Proof. Identify a positive diagonal matrix with its diagonal (x, y). Near $z = (1, \theta)$, below $y = \theta x$ the map equals $(1, 0)$; its limit from above is $(1, \theta^{p-1})/(1 + \theta^p)^{(p-1)/p}$. Write their difference as j, let $v = (1, \theta)$ be tangent to the cutoff boundary, and let $n = (-\theta, 1)$ be its normal pointing into the region retaining both values. Then

$$\langle j, v\rangle = (1 + \theta^p)^{1/p} - 1 > 0.$$

For sufficiently small $a > 0$, $w = -v + an$ satisfies $\langle n, w\rangle > 0$ and $\langle j, w\rangle < 0$. The points $P_h = z - hw/2$ and $Q_h = z + hw/2$ lie on opposite sides of the boundary, and $\langle T(Q_h) - T(P_h), Q_h - P_h\rangle = h\langle j, w\rangle + o(h) < 0$ for small $h > 0$. □

This implementation issue is separate from the continuous but non-Lipschitz exact Schatten gradient. It does not show that a training trajectory visits the cutoff boundary. It does require attributing convexity and monotonicity guarantees to the exact mathematical map, and measuring the effect of its numerical policy.

Fidelity audit. The additional fidelity audit uses axis-aligned and randomly rotated rectangular matrices, prescribed condition numbers from 1 to 10^{12}, prescribed rank deficiency, float64 and float32 arithmetic, and cutoffs $0, 10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}$. The requested zero cutoff still has the historical implementation's independent 10^{-12} lower bound; an untruncated SVD reference is recorded separately. For each arithmetic type, the SVD reference is evaluated in float64 on the same rounded input. Table 17 separates three comparisons: the default Gram map against the historical SVD map, Gram and SVD at the same 10^{-8} cutoff, and SVD at the two different cutoffs. These relative errors are separate comparisons, not additive terms. The smoothed nuclear control uses its SVD formula on the same rounded input. The data show the expected loss of Gram fidelity on poorly conditioned matrices, especially in float32; the analytic coefficient bound is not used as a substitute for these matrix-output comparisons.

Rank-deficient synthetic inputs require an additional distinction. Rounding the construction $U \operatorname{diag}(s)V^\top$ can replace intended zeros by tiny positive singular values; float32 storage introduces a larger perturbation. Raising these values to the power $.05$ can produce a visible output change even with an SVD. Accordingly the audit also compares both implementations with the prescribed spectrum before rounding, and retains the untruncated SVD results without calling them an exact

Arithmetic	Constructed condition	n	Total	Same cutoff	Cutoff only	Smooth
float64	10^3	7	7.4×10^{-12}	7.4×10^{-12}	0	3.6×10^{-13}
float64	10^5	7	3.2×10^{-8}	3.2×10^{-8}	0	8.3×10^{-13}
float64	10^7	7	4.7×10^{-3}	4.7×10^{-3}	0	9.0×10^{-13}
float64	10^9	7	2.1×10^{-1}	8.9×10^{-2}	1.9×10^{-1}	1.8×10^{-12}
float64	rank deficient	14	5.5×10^{-7}	5.5×10^{-7}	0	8.7×10^{-13}
float32	10^3	7	9.1×10^{-3}	9.1×10^{-3}	0	3.6×10^{-4}
float32	10^5	7	9.7×10^{-1}	9.7×10^{-1}	0	4.9×10^{-4}
float32	10^7	7	6.3×10^{-1}	6.3×10^{-1}	0	7.2×10^{-4}
float32	10^9	7	1.1×10^{0}	1.2×10^{0}	2.0×10^{-1}	9.8×10^{-4}
float32	rank deficient	14	4.6×10^{-1}	2.6×10^{-1}	4.1×10^{-1}	6.0×10^{-4}

Table 17: Synthetic Gram fidelity stress audit for Schatten-1.05, default cutoff 10^{-8}. Entries are maximum relative Frobenius errors over the n rotated matrices in each row. "Total" compares Gram with the historical SVD cutoff 10^{-12}; "same cutoff" compares both routes at 10^{-8}; "cutoff only" compares SVD at 10^{-8} and 10^{-12}. "Smooth" compares Gram and SVD for $\kappa = .001$. The supplementary audit records the source arrays, hardware and hashes. These deliberately adverse spectra do not estimate the frequency or size of errors during training.

oracle at numerical rank deficiency. Historical training spectra alone are not a directional check: similar output norms and quantile ratios do not imply similar matrix updates. The saved-input audit and cutoff training study address the relevant training regime separately from this synthetic stress test.

Saved training-state inputs. Table 18 repeats the comparisons on complete matrices saved from actual gradient and momentum states in the three cutoff training arms. These are the diagnostic pre-update expressions: their float32 evaluation order can differ by a few units in the last place from the in-place expression inside the optimizer. Each comparison nevertheless uses exactly the same saved tensor for both routes. The table demonstrates both close agreement on typical inputs and substantial differences for some weak-mode inputs. Larger cutoffs improve agreement with the equally truncated SVD while increasing the departure from the historical SVD cutoff; lowering the cutoff does not uniformly improve fidelity. The archived per-checkpoint and head/non-head strata retain this heterogeneity. Mode counts refer to the eigendecomposition used in the audit; the auxiliary training diagnostic uses an eigenvalues-only routine and is not a record of the exact mode decisions inside every training call.

To distinguish the diagnostic expression from the executed argument, an additional seed-200 run intercepts the exact float64 tensor passed to each map at all four checkpoints. This instrumented run is excluded from loss and timing comparisons. The nominal expression differs from the executed input by at most 3.27×10^{-8} in relative Frobenius norm. The 100 actual-call inputs reproduce the fidelity issue: with the default Gram cutoff, the median difference from the historical SVD route is 1.68×10^{-5} and the maximum is .2624; at matched 10^{-8} cutoffs the median is 3.32×10^{-10} but the maximum is .0984. The smoothed-map Gram/SVD difference on these same inputs is at most 2.84×10^{-11}. These are fixed-input comparisons on one captured trajectory, not bounds for all training updates or evidence that the errors worsen test loss.

The matched cutoff study in Appendix W finds small final-loss contrasts under the common grid, seeds and sample budget. This supports robustness of the observed training result to these numerical policies, not numerical equivalence to the exact Schatten gradient. We therefore refer to the measured arm as a cutoff-defined Gram Schatten baseline.

A bounded refinement attempt. To test whether the unresolved modes could be recovered without a full SVD, the subspace-refinement implementation fixes the split at $s/s_{\max} = 10^{-4}$, uses the Gram formula above it, performs an SVD of $U_{\text{low}}^{\top} X$ below it, and combines both parts in the full Schatten normalizer with the historical 10^{-12} SVD cutoff. The predetermined maximum normalized-direction-error target was 10^{-5}. The CPU check on the saved nominal inputs and synthetic matrices did not meet this target. A small residual high/low Gram cross block was not a sufficient fidelity certificate in the presence of weak singular modes. The supplementary refinement audit retains the implementation and failures; no training arm was launched for it and the threshold was not relaxed. A comparably

Archive	Arithmetic / cutoff	n	Total median	Total max	Same cutoff max	Cutoff max	Smooth max
1	float32 / 10^{-8}	100	3.6×10^{-2}	6.9×10^{-1}	6.9×10^{-1}	2.6×10^{-1}	8.5×10^{-3}
1	float64 / 10^{-10}	100	1.7×10^{-5}	2.2×10^{-1}	2.2×10^{-1}	2.5×10^{-2}	2.8×10^{-11}
1	float64 / 10^{-8}	100	1.7×10^{-5}	2.6×10^{-1}	9.7×10^{-2}	2.6×10^{-1}	2.8×10^{-11}
1	float64 / 10^{-6}	100	2.4×10^{-2}	3.0×10^{-1}	9.9×10^{-6}	3.0×10^{-1}	2.8×10^{-11}
2	float32 / 10^{-8}	100	3.5×10^{-2}	6.9×10^{-1}	6.9×10^{-1}	1.9×10^{-1}	6.8×10^{-3}
2	float64 / 10^{-10}	100	1.1×10^{-5}	7.6×10^{-2}	7.6×10^{-2}	2.4×10^{-2}	2.8×10^{-11}
2	float64 / 10^{-8}	100	1.2×10^{-5}	1.9×10^{-1}	7.6×10^{-2}	1.9×10^{-1}	2.8×10^{-11}
2	float64 / 10^{-6}	100	2.4×10^{-2}	3.3×10^{-1}	1.0×10^{-5}	3.3×10^{-1}	2.8×10^{-11}
3	float32 / 10^{-8}	100	3.5×10^{-2}	1.0×10^{0}	1.0×10^{0}	2.7×10^{-1}	8.6×10^{-3}
3	float64 / 10^{-10}	100	1.5×10^{-5}	1.9×10^{-1}	1.9×10^{-1}	2.5×10^{-2}	2.8×10^{-11}
3	float64 / 10^{-8}	100	1.5×10^{-5}	2.7×10^{-1}	8.8×10^{-2}	2.7×10^{-1}	2.8×10^{-11}
3	float64 / 10^{-6}	100	2.4×10^{-2}	3.0×10^{-1}	1.7×10^{-5}	3.0×10^{-1}	2.8×10^{-11}
4	float32 / 10^{-8}	100	3.6×10^{-2}	1.0×10^{0}	1.0×10^{0}	2.6×10^{-1}	8.6×10^{-3}
4	float64 / 10^{-10}	100	1.4×10^{-5}	1.8×10^{-1}	1.8×10^{-1}	2.5×10^{-2}	2.8×10^{-11}
4	float64 / 10^{-8}	100	1.7×10^{-5}	2.6×10^{-1}	9.8×10^{-2}	2.6×10^{-1}	2.8×10^{-11}
4	float64 / 10^{-6}	100	2.4×10^{-2}	3.0×10^{-1}	3.1×10^{-5}	3.0×10^{-1}	2.8×10^{-11}

Table 18: Saved-input fidelity. Archives 1, 2 and 3 are seed 200 of the training arms with cutoffs 10^{-8}, 10^{-6} and 10^{-10}, respectively, using diagnostic reconstruction. Archive 4 intercepts actual map-call arguments in a separate default-cutoff run. Each contains all 25 managed matrices at four checkpoints. The cutoff in the second column is the one used for this fixed-input audit, which need not equal that of the source trajectory. Error definitions match Table 17. These are descriptive matrix/checkpoint summaries, not independent-seed uncertainty intervals. The supplementary fidelity audit provides the complete strata and artifact hashes.

inexpensive implementation with verified fidelity to the historical SVD formula remains open in this study.

O.2 Gain at zero and Jacobian bounds

Definition 65 (Gain at zero)**.** Let $\Psi(P) = \hat{c}(XX^\top)X$ be a zero-homogeneous spectral map, so that $\Psi(U\,\mathrm{diag}(s)V^\top) = U\,\mathrm{diag}(d)V^\top$ with $d_i = \hat{c}(\hat{s}_i^2)\hat{s}_i$. Fix $r \geq 2$ and a unit vector $s^\circ \in \mathbb{R}^{r-1}$ with positive entries, the remaining spectrum. For $\varepsilon > 0$ let $P_\varepsilon = \mathrm{diag}(s^\circ, \varepsilon) \in \mathbb{R}^{r\times r}$, whose last normalized singular value is $\hat{s}_r(\varepsilon) = \varepsilon/\sqrt{1+\varepsilon^2}$ while the others tend to s°. Then $\Psi(P_\varepsilon)$ is diagonal; let $d_r(\varepsilon)$ be its last diagonal entry. The gain at zero is

$$g_0(\Psi; s^\circ) = \lim_{\varepsilon \downarrow 0} \frac{d_r(\varepsilon)}{\hat{s}_r(\varepsilon)} \ \in [0, \infty] \tag{160}$$

whenever the limit exists: the slope at zero of one output singular value as a function of its normalized input singular value, the rest of the spectrum being fixed in the limit. By zero-homogeneity it is the same at every input with this normalized spectrum, and zero padding to $m \times n$ changes nothing.

Proposition 66 (Gain at zero)**.** *For every $r \geq 2$ and every remaining unit spectrum s°:*

(a) normalized descent has $g_0 = 1$;

(b) the reference map Ψ_K has $g_0 = f_K'(0) = a_0^K$, independent of s° and r, equal to 3.4445 at $K = 1$ and 484.9 at $K = 5$;

(c) the quartic map has $g_0 = 1 + 3\lambda\tau_4(s^\circ) \in [1 + 3\lambda/(r-1), 1 + 3\lambda]$, at most 1.375 for $\lambda = 1/8$;

(d) the smoothed nuclear map has

$$g_0 = \sqrt{r/\kappa} + \kappa\sqrt{r}\,\overline{u}^\circ, \qquad \overline{u}^\circ = \frac{1}{r}\sum_{j=1}^{r} \frac{1}{\sqrt{r(s_j^\circ)^2 + \kappa}}, \quad s_r^\circ := 0, \tag{161}$$

with $0 < \overline{u}^\circ < \kappa^{-1/2}$, hence $\sqrt{r/\kappa} < g_0 < \sqrt{r/\kappa}\,(1+\kappa)$; at $r = 384$ and $\kappa = .001$ this is the interval $(619.68, 620.30)$, and the flat remaining spectrum gives 619.70;

(e) the Schatten map has $g_0 = +\infty$ for $1 < \varsigma < 2$, with $d_r(\varepsilon)/\hat{s}_r(\varepsilon) = \hat{s}_r^{\varsigma-2}/\|\hat{s}\|_\varsigma^{\varsigma-1}$, $g_0 = 1$ at $\varsigma = 2$ and $g_0 = 0$ for $\varsigma > 2$; at $\varsigma = 1$ the reduced polar selection has $d_r(\varepsilon) = 1$ for every $\varepsilon > 0$ and $d_r = 0$ at $\varepsilon = 0$, so it is discontinuous at P_0.

Proof. In each case $d_r(\varepsilon)/\hat{s}_r(\varepsilon) = \hat{c}(\hat{s}_r^2)$ evaluated at the normalized spectrum of P_ε, which tends to $(s^\circ, 0)$, and the claims follow from (159). (a) is immediate. (b): $f_K = p^{\circ K}$ with $p(0) = 0$, so the chain rule gives $f_K'(0) = p'(0)^K = a_0^K$. (c): $\hat{c}(0) = 1 + 3\lambda\tau_4$, and $\tau_4(s^\circ) = \sum_j (s_j^\circ)^4 \in [1/(r-1), 1]$ for a unit vector in $\mathbb{R}^{r-1}$, with the lower value attained only at the flat spectrum. (d): at unit norm $\hat{c}(0) = \tau^{-1/2} + \tau \sum_j (\hat{s}_j^2 + \tau)^{-1/2}$ evaluated at $(s^\circ, 0)$; writing $\tau = \kappa/r$ and $(\hat{s}_j^2 + \tau)^{-1/2} = \sqrt{r}\,(r\hat{s}_j^2 + \kappa)^{-1/2}$ gives (161), the term $j = r$ being the zero coordinate. Each summand in $\overline{u}^\circ$ is at most $\kappa^{-1/2}$, with equality only for $j = r$, so $0 < \overline{u}^\circ < \kappa^{-1/2}$ and the bracket follows. (e): the coefficient is $\hat{s}_r^{\varsigma-2}/\|\hat{s}\|_\varsigma^{\varsigma-1}$ with $\|\hat{s}\|_\varsigma \to \|s^\circ\|_\varsigma \in (0, \infty)$; its limit is $+\infty$, 1 or 0 according to the sign of $\varsigma - 2$. The polar statement is the definition of the reduced polar factor. □

Proposition 67 (Jacobian bounds)**.** *The following statements concern the exact maps. Suprema involving a derivative are taken over its differentiability points; the Schatten non-Lipschitz assertions require $r \geq 2$.*

(i) For $r \geq 2$ and $1 < \varsigma < 2$, $\sup_{P \neq 0} \|P\|_F \|D\Psi_{(\varsigma)}(P)\|_{\mathrm{op}} = +\infty$. Quantitatively, at $P = \mathrm{diag}(s) \in \mathbb{R}^{r \times n}$ with $\|P\|_F = 1$ and distinct positive s_i,

$$\|D\Psi_{(\varsigma)}(P)\|_{\mathrm{op}} \;\geq\; \frac{(\varsigma - 1)\, s_r^{\varsigma-2}}{\|s\|_\varsigma^{\varsigma-1}} \Big(1 - \frac{s_r^\varsigma}{\|s\|_\varsigma^\varsigma}\Big), \tag{162}$$

$$\|D\Psi_{(\varsigma)}(P)\|_{\mathrm{op}} \;\geq\; \frac{s_r^{\varsigma-2}}{\|s\|_\varsigma^{\varsigma-1}} \qquad \text{if } n > r, \tag{163}$$

both of which tend to infinity as $s_r \downarrow 0$ with the other entries fixed.

(ii) For every $P \neq 0$ and $\kappa > 0$,

$$\|D\Psi_\kappa^{\mathrm{sn}}(P)\|_{\mathrm{op}} \;\leq\; \frac{L_\kappa}{\|P\|_F}, \qquad L_\kappa = \frac{\sqrt{r}\,(1 + \kappa)}{\sqrt{\kappa}}. \tag{164}$$

(iii) Consequently $\Psi_\kappa^{\mathrm{sn}}$ is Lipschitz on the unit Frobenius sphere,

$$\|\Psi_\kappa^{\mathrm{sn}}(P) - \Psi_\kappa^{\mathrm{sn}}(Q)\|_F \;\leq\; \tfrac{\pi}{2} L_\kappa \|P - Q\|_F \qquad \text{for all unit } P, Q, \tag{165}$$

whereas, for $r \geq 2$, $\Psi_{(\varsigma)}$ is not Lipschitz on the sphere for any $1 \leq \varsigma < 2$.

Proof. (i) On the open set of matrices with distinct positive singular values the eigenvalues of $XX^\top$ are simple and positive, so $\hat{c}(XX^\top)$ and $\|\hat{s}\|_\varsigma$ are smooth functions of X and $\Psi_{(\varsigma)}$ is smooth there; its Fréchet derivative is the directional derivative along any direction. For $|t| < s_r$ the matrix $P + tE_{rr}$ is diagonal with positive entries, so by Proposition 62(ii), with the formula transported to an unsorted diagonal by permutation equivariance, $\Psi_{(\varsigma)}(P + tE_{rr}) = \mathrm{diag}\big(d(s + te_r)\big)$ with $d(\sigma) = (\sigma/\|\sigma\|_\varsigma)^{\varsigma-1}$ componentwise. Hence $D\Psi_{(\varsigma)}(P)[E_{rr}] = \mathrm{diag}(\partial d/\partial \sigma_r(s))$, whose (r, r) entry is

$$\frac{\partial d_r}{\partial \sigma_r}(s) = (\varsigma - 1)\frac{s_r^{\varsigma-2}}{\|s\|_\varsigma^{\varsigma-1}} - (\varsigma - 1)\frac{s_r^{\varsigma-1}\, s_r^{\varsigma-1}}{\|s\|_\varsigma^{2\varsigma-1}} = \frac{(\varsigma - 1)\, s_r^{\varsigma-2}}{\|s\|_\varsigma^{\varsigma-1}} \Big(1 - \frac{s_r^\varsigma}{\|s\|_\varsigma^\varsigma}\Big),$$

and $\|E_{rr}\|_F = 1$ gives the first bound. If $n > r$, take $H = E_{rn}$. The matrix $P + tH$ differs from P only in row r, which becomes $s_r e_r^\top + te_n^\top$ of norm $\rho_t = \sqrt{s_r^2 + t^2}$, so $P + tH = \mathrm{diag}(s_1, \ldots, s_{r-1}, \rho_t)\, W_t^\top$ for an $n \times r$ matrix W_t with orthonormal columns $e_1, \ldots, e_{r-1}, (s_r e_r + te_n)/\rho_t$, and

$$\Psi_{(\varsigma)}(P + tH) = \sum_{i<r} d_i(\sigma_t) E_{ii} + d_r(\sigma_t) \frac{s_r E_{rr} + tE_{rn}}{\rho_t}, \qquad \sigma_t = (s_1, \ldots, s_{r-1}, \rho_t) = s + O(t^2).$$

Its (r, n) entry $d_r(\sigma_t)t/\rho_t$ has derivative $d_r(s)/s_r = s_r^{\varsigma-2}/\|s\|_\varsigma^{\varsigma-1}$ at $t = 0$, which is (163). The right sides of (162) and (163) diverge as $s_r \downarrow 0$ because $\varsigma - 2 < 0$ and $\|s\|_\varsigma$ tends to the positive ς-norm

Table 19: Gain at zero at $r = 384$ for the flat remaining spectrum, with the range over all remaining spectra in Proposition 66; its finite-difference check $d_r/\hat{s}_r$ at $\hat{s}_r = 10^{-4}$ and 10^{-6}, computed with the matrix implementations; the uniform Jacobian bound on the unit Frobenius sphere (the global Lipschitz bound in Proposition 67 includes a factor $\pi/2$); and whether our implementation uses an eigendecomposition of the $r \times r$ Gram matrix. The other maps use matrix products only.

Map	g_0	$\hat{s}_r = 10^{-4}$	$\hat{s}_r = 10^{-6}$	Uniform Jacobian bound on sphere	Gram eigh
Normalized descent	1	1.0000	1.0000	1	no
Reference, $K = 1$	$a_0 = 3.4445$	3.4445	3.4445	finite	no
Reference, $K = 5$	$a_0^5 = 484.9$	484.7	484.9	finite	no
Quartic, $\lambda = 1/8$	1.001	1.001	1.001	$\leq \Lambda_\lambda$	no
Smoothed nuclear, $\kappa = .001$	619.70	618.5	619.7	≤ 620.30	yes
Smoothed nuclear, $\kappa = .01$	196.16	196.1	196.2	≤ 197.92	yes
Schatten-1.05	$+\infty$	5515	4.381×10^5	unbounded	yes
Schatten-1.25	$+\infty$	640.1	2.024×10^4	unbounded	yes

of the remaining entries. Finally $\|P\|_F \|D\Psi(P)\|_{\mathrm{op}}$ is scale invariant for every zero-homogeneous map (Lemma 9), which gives the supremum statement.

(ii) By Theorem 58, $\Phi_\kappa^{\mathrm{sn}}$ is smooth on $\mathbb{R}^{r\times n} \setminus \{0\}$ with gradient $\Psi_\kappa^{\mathrm{sn}}$, so $D\Psi_\kappa^{\mathrm{sn}}(P) = D^2\Phi_\kappa^{\mathrm{sn}}(P)$ is self-adjoint and, by convexity, positive semidefinite. The operator norm of such an operator is its largest eigenvalue, $\sup_{\|H\|_F=1} D^2\Phi_\kappa^{\mathrm{sn}}(P)[H,H]$, and the upper bound in (154) together with $\|\Pi_P^\perp H\|_F \leq \|H\|_F$ gives (164). The constant is the one in (154): in the notation of that proof every Hessian block coefficient is at most $c + \tau^{-1/2} \leq \sqrt{\kappa r} + \sqrt{r/\kappa} = L_\kappa$, which is also the supremum of the coefficient $\hat{c}$ in (159).

(iii) For unit $P \neq Q$ let $\theta = \arccos\langle P, Q\rangle \in (0, \pi]$ and let γ be a unit-speed great-circle arc of length θ from P to Q on the sphere (for $Q = -P$ any such arc). Then $\Psi(Q) - \Psi(P) = \int_0^\theta D\Psi(\gamma(\vartheta))[\gamma'(\vartheta)]\, d\vartheta$, so by (ii) $\|\Psi(Q) - \Psi(P)\|_F \leq L_\kappa\theta$, while $\|P - Q\|_F = 2\sin(\theta/2) \geq 2\theta/\pi$. For the Schatten map, suppose it were L-Lipschitz on the sphere. At a unit P with distinct positive singular values and any H, zero-homogeneity gives $D\Psi(P)[P] = 0$ (Lemma 9), hence $D\Psi(P)[H] = D\Psi(P)[H_\perp]$ with $H_\perp = H - \langle H, P\rangle P$, $\|H_\perp\|_F \leq \|H\|_F$, and $D\Psi(P)[H_\perp] = \lim_{t\to 0} t^{-1}\big[\Psi\big((P + tH_\perp)/\|P + tH_\perp\|_F\big) - \Psi(P)\big]$ with $\|(P + tH_\perp)/\|P + tH_\perp\|_F - P\|_F = t\|H_\perp\|_F + O(t^2)$; so $\|D\Psi_{(\varsigma)}(P)\|_{\mathrm{op}} \leq L$ at every such P, contradicting (i). At $\varsigma = 1$ the reduced polar selection is discontinuous at rank-deficient points (Proposition 66(e)). □

Numerical checks. Table 19 lists the closed forms next to the finite differences of Definition 65 at $r = 384$. At $\hat{s}_r = 10^{-6}$ every finite gain agrees with its closed form to a relative error below 2×10^{-3}, and the two Schatten values grow by the factor $100^{2-\varsigma}$ per two decades, as $\hat{s}_r^{\varsigma-2}$ predicts. Dense Jacobians by automatic differentiation on 6×6 and 4×6 diagonal inputs with smallest normalized singular value decreasing from 0.067 to 6.7×10^{-5} (square) and from 0.083 to 8.3×10^{-5} (rectangular) give the following observations. The Schatten-1.05 operator norm grows from 8.53 to 444 in the square case, where it coincides with the diagonal derivative in (162) once $s_{\min} \leq 7 \times 10^{-4}$, and from 10.4 to 7375 in the rectangular case, where it coincides with the null-block coefficient $s_r^{\varsigma-2}/\|s\|_\varsigma^{\varsigma-1}$, $1/(\varsigma - 1) = 20.0$ times the diagonal derivative. The smoothed nuclear operator norm at $\kappa = .001$ rises to 77.47 against the bound $L_\kappa = 77.54$ at $r = 6$ and to 63.26 against 63.31 at $r = 4$: the bound (164) is approached to within 0.1% at small $s_{\min}$, an observation and not a sharpness claim. The depth-five reference approaches a_0^5, at 485. At the rank-deficient input $\mathrm{diag}(1, .5, 0)$, the distance $\|\Psi_{(\varsigma)}(\mathrm{diag}(1, .5, \delta)) - \Psi_{(\varsigma)}(\mathrm{diag}(1, .5, 0))\|_F$ decays like $\delta^{\varsigma-1}$ (log-log slope 0.050 at $\varsigma = 1.05$), so its ratio to δ diverges, while the same ratio for the smoothed nuclear map is constant at 49.0, below $L_\kappa = 54.8$ for $r = 3$.

O.3 Consequences

Proposition 68 (Output energy in a noise block)**.** *Let $\|P\|_F = 1$ and split the indices of the normalized singular values into a signal block S with $|S| = r_s \geq 1$ and $s_i \geq s_+ > 0$ for $i \in S$, and a noise block N with $|N| = r_n \geq 1$ and $0 \leq s_i \leq \varepsilon$ for $i \in N$. Put $e_N = \sum_{i\in N} s_i^2 \leq r_n\varepsilon^2$, $e_S = 1 - e_N$, and for a spectral map with output singular values d_i let $E_N = \sum_{i\in N} d_i^2$, $E_S = \sum_{i\in S} d_i^2$*

and $F_N = E_N/(E_N + E_S)$. An empty noise block has $F_N = 0$ and needs no ratio bound involving r_n^{-1}.

(a) For the smoothed nuclear map, with $\varepsilon_\star = \sqrt{\kappa/r}$, put $\bar{e} = \min\{r_n\varepsilon^2, 1 - r_s s_+^2\} < 1$. Then

$$\frac{E_N}{E_S} \;\le\; (1+\tau)L_\kappa^2\,\frac{e_N}{e_S} \;\le\; (1+\tau)L_\kappa^2\,\frac{\bar{e}}{1-\bar{e}}. \tag{166}$$

If, additionally, $r_n\varepsilon^2 < 1$, this implies the simpler bound

$$\frac{E_N}{E_S} \;\le\; (1+\tau)(1+\kappa)^2\frac{r_n(\varepsilon/\varepsilon_\star)^2}{1-r_n\varepsilon^2}. \tag{167}$$

Writing x for any of these applicable nonnegative bounds, $F_N \le x/(1+x)$. When $r_n\varepsilon^2 \ge 1$, use (166)*; the denominator in* (167) *is then not a positive lower bound for e_S.*

(b) For the Schatten map with $1 < \varsigma < 2$, exactly $E_N/E_S = \sum_{i\in N} s_i^{2(\varsigma-1)} / \sum_{i\in S} s_i^{2(\varsigma-1)}$, so if every noise value lies in $[\varepsilon_-, \varepsilon]$ with $\varepsilon_- > 0$,

$$\frac{r_n\,\varepsilon_-^{2(\varsigma-1)}}{r_s} \;\le\; \frac{E_N}{E_S} \;\le\; \frac{r_n\,\varepsilon^{2(\varsigma-1)}}{r_s\,s_+^{2(\varsigma-1)}}. \tag{168}$$

Hence at fixed $\varepsilon_- > 0$, $F_N \to r_n/(r_n + r_s)$ as $\varsigma \downarrow 1$. At fixed ς and a fixed positive signal lower bound s_+, the upper bound gives $F_N = O(\varepsilon^{2(\varsigma-1)})$ as $\varepsilon \to 0$; this rate is attained when all noise values are proportional to ε. $F_N \le \frac{1}{2}$ requires $\varepsilon_- \le (r_s/r_n)^{1/(2(\varsigma-1))}$ and holds whenever $\varepsilon \le s_+(r_s/r_n)^{1/(2(\varsigma-1))}$.

(c) For the reduced polar factor, $F_N = r_n/(r_n + r_s)$ whenever every noise value is positive.

Proof. (a) By (156) in the variables of Theorem 58, $d_i = \gamma_i s_i$ with $\gamma_i = (s_i^2 + \tau)^{-1/2} + c$ and $c = \tau\sum_j (s_j^2+\tau)^{-1/2} \in [0, \sqrt{\kappa r}]$. Since $0 \le s_i \le 1$, $(1+\tau)^{-1/2} \le \gamma_i \le \tau^{-1/2} + \sqrt{\kappa r} = L_\kappa$. Therefore $E_N = \sum_N \gamma_i^2 s_i^2 \le L_\kappa^2 e_N$ and $E_S \ge e_S/(1+\tau)$, which is the first inequality. Since $e_S \ge r_s s_+^2 > 0$, we have $0 \le e_N \le \bar{e} < 1$; monotonicity of $e/(1-e)$ proves the second inequality in (166). Under the additional condition $r_n\varepsilon^2 < 1$, use $L_\kappa^2 = (1+\kappa)^2/\varepsilon_\star^2$ and $e_S \ge 1 - r_n\varepsilon^2 > 0$ to obtain (167). Finally, $F_N = (E_N/E_S)/(1 + E_N/E_S)$ is increasing in E_N/E_S. (b) $d_i = (s_i/\|s\|_\varsigma)^{\varsigma-1}$, and the common normalizer cancels in the ratio; the bounds use $s_+ \le s_i \le 1$ on S, $\varepsilon_- \le s_i \le \varepsilon$ on N and $2(\varsigma-1) > 0$. As $\varsigma \downarrow 1$ all three powers tend to one, so both sides of (168) tend to r_n/r_s. The two half-energy statements are $E_N/E_S \le 1$ read through the lower and upper bounds. (c) $d_i = 1$ whenever $s_i > 0$. □

For example, $s = (1/\sqrt{2}, 3/5, \sqrt{7}/10, \sqrt{7}/10)$ has unit squared norm. Taking its first entry as signal and $\varepsilon = 3/5$ gives $e_N = e_S = 1/2$ but $r_n\varepsilon^2 = 27/25$. Equation (166) remains valid, whereas (167) does not apply. This also explains why a positive-denominator condition cannot be omitted. The inequality (166) is the sense in which the noise energy of the smoothed nuclear map is controlled by its gain: the output noise-to-signal energy ratio exceeds the input ratio by at most the squared dynamic range $(1+\tau)L_\kappa^2$ of the gain γ_i, and below the threshold $\varepsilon_\star$ it decays quadratically in $\varepsilon/\varepsilon_\star$. For the Schatten map no such bound exists; the polar limit $r_n/(r_n + r_s)$ is approached as $\varsigma \downarrow 1$ at every fixed positive noise level.

Numbers. At $r = 384$ with a flat signal block of dimension $r_s = 8$ and $r_n = 376$ noise values all equal to ε, the computed fractions F_N are as follows. At $\varepsilon = 10^{-2}$, above the threshold $\varepsilon_\star = 0.00161$, every flattening map is near the polar value 0.979: 0.977 for the smoothed nuclear map at $\kappa = .001$, 0.951 for the depth-five reference, 0.971 for Schatten-1.05. At $\varepsilon = 10^{-3}$ they are 0.903, 0.901 and 0.963. At $\varepsilon = 10^{-4}$ they separate: 0.109 for the smoothed nuclear map (bound (166): 0.591), 0.088 for the reference, 0.006 at $\kappa = .01$, 3.8×10^{-6} for normalized descent, but 0.954 for Schatten-1.05 and 0.441 for Schatten-1.25. The noise level at which the noise block carries exactly half the output energy is 2.9×10^{-4} for the smoothed nuclear map, 3.2×10^{-4} for the reference, 1.6×10^{-4} for Schatten-1.25 and 6.7×10^{-18} for Schatten-1.05, the last being the closed form $s_+(r_s/r_n)^{1/(2(\varsigma-1))}$ of (168) and below normwise float64 roundoff at unit spectral scale in a generic dense matrix (the

scalar value itself is representable). In this computation the reference and the smoothed nuclear map behave alike, consistent with their comparable gains 484.9 and 619.70; these are computed values on one synthetic spectrum, not a statement about training gradients.

Proposition 69 (Eigenvalue conditioning of the coefficient)**.** *Let the eigenvalues $\hat{\mu}_i$ of $XX^\top$, $\hat{\mu}_{\max} \le 1$, be perturbed with absolute errors $|\delta_i| \le u\hat{\mu}_{\max}$. This is a coefficient-sensitivity calculation: the singular frames and the factors $\hat{s}_i$ multiplying the coefficients are held fixed. It does not account for eigenvector error or error in forming $XX^\top$.*

(a) For the Schatten coefficient with $1 \le \varsigma < 2$ and the normalizer held fixed, the relative change of $\hat{c}(\hat{\mu}_i)$ is $(1+\delta_i/\hat{\mu}_i)^{(\varsigma-2)/2} - 1$, whose first order term has modulus

$$\Big(1 - \frac{\varsigma}{2}\Big)\frac{|\delta_i|}{\hat{\mu}_i} \;\le\; \Big(1 - \frac{\varsigma}{2}\Big)u\Big(\frac{s_{\max}}{s_i}\Big)^2, \tag{169}$$

unbounded as $s_i \to 0$; for $\hat{\mu}_i \le u\hat{\mu}_{\max}$ the computed eigenvalue can be nonpositive and the coefficient is undefined.

(b) For the smoothed nuclear coefficient, if $u\hat{\mu}_{\max} \le \tau/2$ then every output singular value and every coefficient $\hat{c}(\hat{\mu}_i)$ changes relatively by at most $u\hat{\mu}_{\max}/\tau$, and to first order by at most

$$\frac{u\,\hat{\mu}_{\max}}{2\tau} \;\le\; \frac{u\,r}{2\kappa}, \tag{170}$$

which is 4.3×10^{-11} at $r = 384$, $\kappa = .001$ and $u = 2.2 \times 10^{-16}$.

Proof. (a) is the derivative of $\mu \mapsto \mu^{(\varsigma-2)/2}$. (b) Each term $(\hat{\mu}+\tau)^{-1/2}$ changes by the factor $(1+x)^{-1/2}$ with $x = \delta/(\hat{\mu}+\tau)$, $|x| \le u\hat{\mu}_{\max}/\tau \le \frac{1}{2}$, and $|(1+x)^{-1/2} - 1| \le |x|$ on $[-\frac{1}{2}, \frac{1}{2}]$, with first-order term $|x|/2$. The coefficient $\hat{c}(\hat{\mu}_i) = (\hat{\mu}_i+\tau)^{-1/2} + \tau\sum_j(\hat{\mu}_j+\tau)^{-1/2}$ is a positive combination of such terms, so it inherits the same relative bound, and so does $d_i = \hat{c}(\hat{\mu}_i)\hat{s}_i$. □

The script perturbs every eigenvalue by $u\hat{\mu}_{\max}$ on log-spaced spectra, using machine epsilon for the illustrative value of u; a full roundoff bound would require dimension- and backend-dependent factors as well. The Schatten-1.05 coefficient of the smallest eigenvalue changes relatively by 1.1×10^{-8} at $s_{\min}/s_{\max} = 10^{-4}$, 1.1×10^{-4} at 10^{-6}, 0.010 at 10^{-7} and 0.43 at 10^{-8}, matching (169) where it is small; the smoothed nuclear coefficients change by less than (170) on every spectrum. The threshold $10^{-8}s_{\max}$ in the historical Gram implementation is near the scale $\sqrt{u}\,s_{\max}$ at which the relative eigenvalue error can be of order one in float64. It is a historical numerical policy, not a rank certificate; spurious positive eigenvalues can survive it. The corresponding scale in float32 is much larger.

Which results consume an upper bound. The trajectory guarantee of Theorem 51 uses two properties of the task profile: the lower envelope $\kappa_c\, t/\sqrt{1+t^2} \le \widetilde{q}(t)$ of (130), in the contraction step (133), and a bound Q on $|\widetilde{q}|$, for the invariant interval. Its upper envelope Λ^{task} is sharp but is not used in that proof. Both properties hold for every Schatten exponent $\varsigma > 1$ (Proposition 55, with $|\mathcal{R}_{(\varsigma)}| \le \sqrt{2}$), so the contraction transfers to the Schatten family and does not separate the two convex maps. An upper bound on the Jacobian is consumed in three places. First, the derivative form of Corollary 52 exchanges derivative and expectation through $\|D\Psi(Q)\|_{\mathrm{op}} \le \Lambda_\lambda/\|Q\|_F$ and brackets the recovery matrix as $\kappa_\lambda\mathcal{G}_Z \preceq A \preceq \Lambda_\lambda\mathcal{G}_Z$ in (126); for the smoothed nuclear map the same argument runs with $(\kappa\sqrt{r}/\sqrt{1+\kappa},\ L_\kappa)$ in place of $(\kappa_\lambda, \Lambda_\lambda)$, by (154) and Proposition 67(ii), whereas for Schatten $\varsigma < 2$ no such constant exists by Proposition 67(i), and only the finite-scale response $\mathcal{R}_w$ of Section 2 is defined without further hypotheses on the noise law. Second, Proposition 76 assumes finite suprema of $\|D\Psi\|_{\mathrm{op}}$ and $\|D^2\Psi\|_{\mathrm{op}}$ on the sphere; the smoothed nuclear map satisfies this, with the first supremum at most L_κ and the second finite by smoothness on the compact sphere, and the Schatten map with $\varsigma < 2$ violates it. The finite-pulse measurements in Appendix V do not themselves prove this failure: two finite pulse sizes cannot establish a small-pulse limit, and the hard cutoff has its own discontinuities. The non-Lipschitz conclusion follows instead from Proposition 67. Third, Corollary 13 needs a Lipschitz Jacobian near the witness, a hypothesis that concerns nonmonotone maps. The upper bound therefore separates the two convex maps in the derivative-level statements and in the conditioning of Proposition 69, not in the contraction theorem.

O.4 What is and is not distinctive

Three facts should be stated plainly. First, invariance under spectral replication is not specific to the smoothed nuclear norm. The Schatten-ς output ratio is exactly $(s_i/s_j)^{\varsigma-1}$, independent of the rank, so every output singular-value ratio is invariant under the replication of Corollary 59 for every Schatten map as well (the block identity itself holds up to the common factor $k^{-(\varsigma-1)/\varsigma}$), and likewise for normalized descent and for the polar factor; Table 37 shows the constant Schatten ratios at every rank. What the corollary adds is that a map with a bounded gain has this property. Second, what distinguishes the smoothed nuclear map from Schatten $\varsigma < 2$ is structural: the bounded gain $g_0 < \sqrt{r/\kappa}(1+\kappa)$, the uniform Jacobian bound L_κ, the coefficient-conditioning bound $ur/(2\kappa)$, and the explicit threshold $\varepsilon_\star\|P\|_F = \sqrt{\kappa/r}\,\|P\|_F$ from (156): below it an output singular value is $z_i/\sqrt{\kappa}$ to leading order, normalized descent with gain $\sqrt{r/\kappa}$, and above it the value tends to one, the polar factor. The depth-five reference has the same qualitative structure with gain $a_0^5 = 484.9$ against 619.70 for $\kappa = .001$ at $r = 384$, and in Proposition 68's computation the two behave alike; what separates the smoothed nuclear map from the reference is monotonicity, not the gain. Third, the matched-training study of Appendix W does not separate the two convex maps by loss in favour of the smoothed nuclear norm. Gram Schatten-1.05 differs from the reference by -0.0013 $[-0.0040, +0.0014]$ on eight seeds, inside the declared margin, at $1.15\times$ the reference's time against $1.14\times$ for the smoothed nuclear map, whose difference $+0.0029$ $[+0.0004, +0.0053]$ crosses the margin; the direct paired difference, Schatten-1.05 minus smoothed nuclear on the same seeds, is -0.0041 $[-0.0069, -0.0014]$, which under the declared rule is inconclusive and nominally favours the Schatten map. The distinction between the two maps established in this appendix is therefore bounded gain, Lipschitz continuity and conditioning, and not an empirical loss advantage.

P Barrier certificates at depth one and depth five

Theorem 31 consumes six numbers that must not depend on B, β or T: a bound Q on $|q_K|$, a local derivative constant D, an inner radius a with barrier γ_-, and an outer radius b with barrier γ_+, the two drift margins meaning

$$h_K(x) \le -\gamma_- \ \ (as \le x \le 2as), \qquad h_K(x) \ge \gamma_+ \ \ (x \ge bs). \tag{171}$$

The certified values are

$$(Q, D, a, b, \gamma_-, \gamma_+) = \begin{cases} \big(2, 17, \frac{1}{200}, 4, \frac{1}{500}, \frac{1}{4}\big), & K = 1, \\ \big(2, 7776, \frac{1}{300000}, 8, 10^{-6}, \frac{1}{32}\big), & K = 5, \end{cases} \tag{172}$$

the second obtained by rational interval arithmetic. This appendix derives them and then reports how much each one gives away.

P.1 Depth one

From Lemma 8 at $K = 1$,

$$q_1(t) = \frac{t(-4.7145 + 9.671t^2 + 6.2595t^4)}{4(1+t^2)^{5/2}}, \qquad q_1'(t) = \frac{23911t^4 + 95742t^2 - 9429}{8000(1+t^2)^{7/2}}, \tag{173}$$

both exact rational identities in the reference coefficients. Four consequences are what the dynamics uses:

$$|q_1(t)| \le Q = 2, \qquad q_1(t) \le -\tfrac{4}{5}t \ \ (0 \le t \le \tfrac{1}{4}), \qquad q_1(t) \ge 0.49 \ \ (t \ge 1), \tag{174}$$

$$\sup_{t \ge 0} \max\{t|q_1'(t)|,\, t^2|q_1'(t)|\} \le D = 17. \tag{175}$$

Each is a statement about one explicit rational function of t and is verified by locating the finitely many critical points of the corresponding polynomial numerators. The last one gives the local bounds $|\partial_x q_1(x/\rho)| \le D/x$ and $\|\nabla_z q_1(x/\|z\|_F)\| \le D/x$ on the positive barrier region, which is what Lemma 33 needs; no global derivative bound on Ψ_1 is claimed, and none exists near the origin.

Lemma 70 (Inner and outer barriers, uniformly in the memory)**.** *With $a = 1/200$, $b = 4$, $\gamma_- = 1/500$ and $\gamma_+ = 1/4$, for every $0 < \beta < 1$ and $B \ge 1$, $h_1(x) \le -\gamma_-$ on $as \le x \le 2as$ and $h_1(x) \ge \gamma_+$ for $x \ge bs$.*

Proof. Write $y = x/s$. For the inner barrier, split the expectation $h_1(x) = \mathbb{E}q_1(x/R)$ at $R = 4x$. On $R \geq 4x$ the argument is at most $1/4$, so the negative linear bound in (174) applies. A positive contribution requires $x/R > t_\star$, that is $R < x/t_\star < (5/3)x$ since $t_\star > 3/5$. Hence

$$h_1(x) \leq -\tfrac{4}{5}x\mathbb{E}\tfrac{1}{R} + \tfrac{4}{5}x\mathbb{E}[R^{-1}; R < 4x] + 2\mathbb{P}\big(R < \tfrac{5}{3}x\big) \leq -\tfrac{4}{5}y + \pi\Big(\tfrac{32}{5} + \tfrac{50}{9}\Big)y^2 \leq -\tfrac{4}{5}y + 38y^2,$$

where the first term used $\mathbb{E}R^{-1} \geq 1/s$ from (70), the second used $\mathbb{E}[R^{-1}; R < c] \leq 2\pi c/s^2$ which follows by integrating the small-ball bound, and the third used $\mathbb{P}(R < c) \leq \pi c^2/s^2$. For $a \leq y \leq 2a = 1/100$ the right side is at most $-\frac{4}{5}y + 38 \cdot \frac{1}{100}y \leq -\frac{2}{5}y \leq -\frac{1}{500}$. For the outer barrier at $x \geq 4s$, use $q_1(x/R) \geq 0.49$ on $R \leq x$ and $q_1 \geq -2$ elsewhere, so with $\mathbb{P}(R > x) \leq s^2/x^2$ by Chebyshev, $h_1(x) \geq 0.49 - 2.49s^2/x^2 \geq 0.49 - 2.49/16 = 0.334375 > \frac{1}{4}$. □

Every constant here comes from (70), which holds at every (B, β), so the barriers are memory free as required.

P.2 Depth five

The composed map has no compact closed form, so the certificate is a chain of rational interval inclusions. Two preliminary facts about p itself:

$$p([0, \tfrac{5}{4}]) \subset [0, \tfrac{5}{4}], \qquad \sup_{[0,5/4]} |p'| < 6, \qquad \mu := a_0 - \frac{b_0^2}{4c_0} > 0.638 \text{ with } p(u) \geq \mu u \ (u \geq 0), \tag{176}$$

the last because $p(u) - \mu u = u(4063u^2 - 4775)^2/8{,}126{,}000 \geq 0$ identically, an exact factorization. From these, $|q_5| < 2$ so $Q = 2$ is admissible, and $\sup|f_5'| \leq 6^5 = 7776$. Differentiating (14) in t gives $|q_K'(t)| \leq \|f_K'\|_{\infty,[0,1]} \max\{1, t\}/(1+t^2)^{3/2}$, so $D = 7776$ is admissible. For $t \geq 1$ both signed spectral contributions are positive and $q_5(t) \geq \mu^5 t/\sqrt{1+t^2} \geq \mu^5/\sqrt{2} > 0.075$, so on $x \geq 8s$ the same Chebyshev split as above gives $h_5(x) \geq 0.075 - 2.075/64 > 1/32 = \gamma_+$.

Lemma 71 (Inner barrier at depth five)**.** *For $0 \leq t \leq 1/100$ one has $q_5'(t) \leq -0.98/(1+t^2)^{3/2} < -0.9$, hence $q_5(t) \leq -0.9t$; and with $a = 1/300{,}000$ and $\gamma_- = 10^{-6}$, $h_5(x) \leq -\gamma_-$ on $as \leq x \leq 2as$ for every (B, β).*

Proof. For $0 \leq t \leq 1/100$ both arguments of (14) lie in $I_0 = [0.700000, 0.714143]$, since at $t = 0$ both equal $1/\sqrt{2} = 0.70710678$ and at $t = 1/100$ they are 0.7141421 and 0.7000007. Treat every displayed decimal as an exact rational, with all enclosures rounded outward. The following five rows each satisfy $p(I_j) \subset I_{j+1}$ for $j \leq 3$, with $p'(I_j)$ contained in the stated interval:

j	I_j	$p(I_j) \subset$	$p'(I_j) \subset$
0	[0.700000, 0.714143]	[1.0980966, 1.1147593]	[−1.220, −1.135]
1	[1.098096, 1.114760]	[0.7033454, 0.7222050]	[0.940, 1.329]
2	[0.703345, 0.722205]	[1.0880857, 1.1109255]	[−1.264, −1.156]
3	[1.088085, 1.110927]	[0.6950338, 0.7172887]	[0.722, 1.237]
4	[0.695032, 0.717290]	(last row)	[−1.237, −1.105]

Inside each interval the signs of p' and p'' are determined by the corresponding lower-degree polynomials, so each inclusion is certified by evaluating at the endpoints together with the finitely many interior critical points, all of which are roots of explicit quartics. The first two image rows are rounded outward from the exact rational endpoints, which matters: $p(0.7)$ is exactly $222951841/200000000 = 1.114759205$, so an upper endpoint of 1.1147592 would exclude a point of the true image. The five derivative factors have a negative product, and the product of their minimum absolute values is

$$1.135 \cdot 0.940 \cdot 1.156 \cdot 0.722 \cdot 1.105 = 0.983968113284 > 0.98.$$

By the chain rule along the orbit, f_5' on the relevant range has absolute value at least 0.98 with a negative sign. Writing $u_\pm(t) = (1 \pm t)/\sqrt{2(1+t^2)}$, differentiating (14) gives

$$q_5'(t) = \frac{(1-t)f_5'(u_+(t)) + (1+t)f_5'(u_-(t))}{2(1+t^2)^{3/2}} \leq -\frac{0.98}{(1+t^2)^{3/2}} < -0.9 \tag{177}$$

for $0 \le t \le 1/100$, using $(1-t)+(1+t)=2$ and $(1+t^2)^{3/2} \le 1.00015$ on that range. The final step needs the numerator bound 0.98 and not 0.9: since $-0.9/(1+t^2)^{3/2} > -0.9$ for $t > 0$, the intermediate inequality $q_5'(t) < -0.9/(1+t^2)^{3/2} < -0.9$ is false as a chain, and 0.98 is what the certified derivative product supplies. With $q_5(0) = 0$ this integrates to $q_5(t) \le -0.9t$. For the barrier, the argument of Lemma 70 applies with a general (Q, c, t_0) whenever $|q_K| \le Q$ and $q_K(t) \le -ct$ on $[0, t_0]$: with $y = x/s$, $h(x) \le -cy + \pi(2c/t_0 + Q/t_0^2)y^2$. Taking $c = 0.9$, $t_0 = 1/100$, $Q = 2$ and $a = 1/300{,}000$, the quadratic coefficient times y is below 0.45 on $a \le y \le 2a$, so $h_5(x) \le -0.45y \le -1.5 \times 10^{-6} < -10^{-6}$. □

P.3 A STRONGER DEPTH-ONE BARRIER WITH A COMPLETE CERTIFICATE

The preceding values remain a reproducible baseline. We now derive a stronger row, using no approximation to the stationary density:

$$(Q, D, a, b, \gamma_-, \gamma_+) = \left(\frac{313}{200}, \frac{186}{125}, \frac{3}{250}, 4, \frac{9}{1000}, \frac{7}{20}\right) \quad (K = 1). \tag{178}$$

The improved inner barrier follows by integrating the complete nonlinear remainder, rather than separating the noise radius at a chosen cutoff.

Lemma 72 (An integrable remainder bound)**.** *For $y = x/s \ge 0$,*

$$h_1(x) \le -\frac{9429}{8000}y + \frac{25327\pi}{2400}y^2. \tag{179}$$

Consequently $h_1(x) \le -9/1000$ on $3s/250 \le x \le 6s/250$.

Proof. Write $c = 9429/8000$ and $r(t) = q_1(t) + ct$. Equation (173) shows $q_1'(t) \ge -c$ for $t \ge 0$: its numerator is at least -9429, and dividing a negative numerator by $(1+t^2)^{7/2} \ge 1$ can only raise it. Thus $r(0) = 0$ and $r(t) \ge 0$. Moreover $r(t) = O(t^3)$ at zero and $r(t) = O(t)$ at infinity. By the joint-density bound in Lemma 32, polar integration gives

$$\mathbb{E}r(x/R) \le \frac{2\pi}{s^2}\int_0^\infty r(x/\rho)\rho\, d\rho = 2\pi y^2 \int_0^\infty \frac{r(t)}{t^3}\, dt.$$

Two integrations by parts, whose boundary terms vanish by the stated endpoint orders, yield

$$\int_0^\infty \frac{r(t)}{t^3}\, dt = \frac{1}{2}\int_0^\infty \frac{q_1''(t)}{t}\, dt = \frac{1}{16000}\int_0^\infty \frac{257487 - 383066t^2 - 71733t^4}{(1+t^2)^{9/2}}\, dt = \frac{25327}{4800}.$$

For the last equality the integrals of $1, t^2, t^4$ against $(1+t^2)^{-9/2}$ are respectively $16/35, 8/105, 2/35$, obtainable by $t = \tan u$. Adding $-cx\mathbb{E}R^{-1} \le -cy$ proves (179). Finally use $\pi < 22/7$. The negative of the right-hand side is a concave quadratic, so its minimum on $[3/250, 6/250]$ is at an endpoint. Exact rational evaluation at both endpoints exceeds $9/1000$. □

For completeness, the remaining scalar constants also have a finite rational certificate. With $u = t^2/(1+t^2) \in [0,1]$, put

$$N(u) = 23911u^2 + 95742u(1-u) - 9429(1-u)^2,$$
$$M(u) = -4.7145(1-u)^2 + 9.671u(1-u) + 6.2595u^2.$$

The three required squared inequalities become nonnegativity on $[0,1]$ of

$$Q^2 - \frac{uM(u)^2}{16}, \qquad D^2 - \frac{u(1-u)^2N(u)^2}{8000^2}, \qquad D^2 - \frac{u^2(1-u)N(u)^2}{8000^2}.$$

The first is of degree five and the latter two of degree seven. We partition $[0,1]$ into 256 equal rational intervals and verify that every Bernstein coefficient of each polynomial on each interval is nonnegative. The convex-hull property of the Bernstein representation certifies each entire interval; this is not a pointwise grid test. Using the already certified $q_1(t) \ge 0.49$ for $t \ge 1$ and the smaller $Q = 1.565$, the outer barrier is $h_1(x) \ge 0.49 - (0.49 + 1.565)/16 = 0.3615625 > 7/20$ for $x \ge 4s$. Together with Lemma 72, this proves every entry of (178).

Certificate	T	m_{win}	sufficient η	entry bound	covered?
Original proof and row (172)	10^6	197	1.7605×10^{-13}	2.2721×10^{13}	no
Drift bound only, original constants	10^6	197	7.3871×10^{-13}	5.4148×10^{12}	no
Drift bound and row (178)	10^6	152	2.5855×10^{-11}	1.1051×10^{11}	no
Same refined certificate, longer horizon	10^{12}	152	1.6211×10^{-11}	1.7625×10^{11}	yes

Table 20: Certified constants and evaluated sufficient budgets. The first three rows diagnose the uncovered million-step budget. The fourth recomputes all conditions at its stated horizon and is a nonempty application of the sharpened theorem. Entries are rounded for presentation; all inequalities are checked using 60-digit arithmetic.

P.4 Quantitative coverage and its limits

Table 20 evaluates the explicit sufficient choices, first with the original proof and then with Corollary 38's drift supermartingales and refined scalar constants. All rows use $K = 1$, $\beta = 0.9$, $B = 1$, $\sigma = 1/16$, $x_0 = 1/4$ and $\delta = 0.05$; hence $s = 0.00827833\ldots$. The horizon enters the logarithmic budget and must be specified before selecting η. When the calculated entry exceeds that horizon, the row does not satisfy the theorem's complete hypotheses and conveys no trajectory guarantee.

The two rigorous refinements reduce the entry bound by a factor of about 206 at the same formal million-step budget, but the refined bound remains about 1.1×10^5 times that budget. This is a substantive tightening of the certificate, not a quantitative explanation of the simulations. The original repair-side calculation similarly gives $t_{\mathrm{sep}} \approx 1.9 \times 10^{13}$, so the original joint comparison is uncovered there as well. For a joint application of the refined result, both algorithms' budgets must be recomputed with failure probability $\delta/2$; Table 20 concerns residence alone.

For the refined row the last term in (77) binds. Substituting it into the leading entry term gives the certificate-level dependence

$$t_{\mathrm{ent},*} \approx \frac{16\, U_0 Q^2 m_{\mathrm{win}} \mathcal{L}_*}{a\, \gamma\, \gamma_+\, s}. \tag{180}$$

In this calculation the initial distance x_0 is fixed while the noise scale shrinks, and the certified step is proportional to s. This explains the inverse-scale dependence of this sufficient bound. It is neither an algorithmic lower bound nor a proof that no other choice of barriers, concentration argument or initialization can cover a practical budget. The unvalidated quadrature barriers reported in an earlier draft are not used in this table or in a theorem. The replacement above uses an analytic stationary-law bound, exact polynomial certificates and an explicit probability budget.

Q Related work in detail

Results about optimizer failure can look similar while using different losses, noise models or step-size schedules. The table below places these assumptions side by side. We then compare the spectral iterations and convex-map constructions used in the related literature with those studied here.

Q.1 Nonconvergence results for normalized and signed methods

Table 21: Selected negative results and their assumptions. Deterministic examples require no stochastic noise assumption. "Bounded grad" means sample gradients are globally bounded; bounded data alone does not imply this for a quadratic objective.

Result	Objective	Sym. noise	Bounded grad	Map	Conclusion
Karimireddy et al. (2019), Ex. 3	smooth least squares	yes, state-dep.	no	sign	subspace trap
Parshakova et al. (2026)	convex Lipschitz, nonsmooth	determ./n/a	yes	exact polar	nonconvergence
Do et al. (2026)	smooth quadratic	no	no	finite NS	nonconvergence a.e. B
Tao et al. (2026)	ℓ_1/ℓ_∞ geometry	n/a	yes	sign, polar	lower bound
This paper, Thm. 31	convex Huber ridge, smooth	**yes**	yes	finite NS, $K \in \{1, 5\}$	finite-window residence

The stochastic and deterministic mechanisms must be distinguished. Counterexample 1 of Karimireddy et al. (2019) uses asymmetric stochastic gradients; Counterexample 2 is deterministic and nonsmooth. Their smooth least-squares Counterexample 3 uses two equally likely gradient values, so its centered oracle noise is conditionally centrally symmetric. The coordinatewise sign updates remain in a subspace that excludes the optimizer. This is a failure under symmetric, state-dependent noise, and it already shows that symmetry alone does not imply convergence. In our task $\xi \overset{d}{=} -\xi$ and the signal gradient is observed exactly, so this origin shift vanishes. We distinguish negative restoring alignment from a zero-response trap, and impose a fixed additive noise law rather than the state-dependent oracle of that example. In contrast, Parshakova et al. (2026) do not require random noise at all: their objective is the deterministic convex Lipschitz function $f(W) = c|W_{11} + W_{22}| + |W_{11} - W_{22}|$, and the algorithm uses its subgradients. Their failure mechanism cannot be classified as asymmetric stochastic averaging.

The precise quantifiers in Parshakova et al. (2026). Counterexample 3.3 fixes $\beta \in [0, 1)$ and an offline positive nonincreasing sequence (λ_t), with $c = (1 - \beta)/2$. If $\lim_t \lambda_t > 0$, an open set of initializations has a positive objective gap at every step. If λ_t is strictly decreasing to zero, the result asserts existence of an initialization with that property. Counterexample 3.4 instead takes $\beta \in [0, 1/2)$, $c = 1/2 - \beta$, and step sizes given by a prescribed function of the subgradient history $G_0, \dots, G_t$. Outside a measure-zero set of initializations, the iterates of their Muon or regularized Muon cannot converge to a minimizer. These are different schedule and initialization quantifiers; neither should be abbreviated to "every schedule" at every momentum value. Their study treats exact polar and excludes Nesterov weighting.

Theorem 1.3 of Do et al. (2026) treats a quadratic stochastic objective with bounded data and a nonzero centered third moment, giving nonconvergence for almost every minibatch size under its step-size conditions. Its sample gradients are not globally bounded as a function of the parameter. Our Huber ridge is convex and 1-smooth with globally bounded sample gradients, while its hard noise is nondegenerate within a two-dimensional transverse plane. Our principal dynamic conclusion is a two-sided finite-window band from zero momentum; it is not an asymptotic nonconvergence theorem. These distinctions suffice to locate the contribution without placing all previous failures under one noise-bias explanation.

Q.2 The implemented iteration

Amsel et al. (2025) optimize the polynomial for approximating the matrix sign function, and Kim & Oh (2026) prove convergence to exact polar for a coefficient family whose monotonicity premise the reference quintic fails, a point their Appendix H.4 states explicitly. Since $p(1) = 0.701$, the reference iteration has no fixed point at one and does not approach exact polar, which is why the paper works with the response profile of Lemma 8 rather than with a distance to polar. Scalar non-monotonicity of the tuned quintic is already known and is not claimed here; Theorem 2 is a different statement, since it applies to spectral functions that *are* increasing on $[0, 1]$, including classical Newton–Schulz polynomials, and rules out every fixed nontrivial mixture with normalized descent.

The normalization in Li & Tsuchiya (2026) is an externally supplied constant: Algorithm 1 divides M_t by G_{op}, a uniform operator-norm bound on stochastic gradients. It does not divide by $\|M_t\|_{\mathrm{op}}$. Consequently the frame-preserving Jacobian is $\mathrm{diag}(f'(\sigma_i/G_{\mathrm{op}}))/G_{\mathrm{op}}$, without the derivative of an endogenous norm. The algorithm also uses the coefficients $(15/8, -5/4, 3/8)$ and a discounted online-to-nonconvex conversion with random query points on update segments. These are concrete algorithmic differences from our reference iteration. We do not infer that every normalization based on an operator norm is fixed, or that its derivative can be ignored.

Beneventano et al. (2026) study the spectral dynamics and noise geometry of Muon, asking which spectral directions the update emphasizes and how the noise interacts with the spectrum. Our question is the sign of the response along a fixed direction, which is a different functional of the same objects, and the two are compatible: a map can distribute attention across the spectrum in the way they describe while having the negative recovery direction of Proposition 12. Fan et al. (2026) report spectral failure modes outside pretraining and propose a high-pass remedy; that is a different intervention aimed at a different failure, and our repair is not a variant of it. Inexact-polar and finite-depth analyses (Shulgin et al., 2025; Do et al., 2026; Li & Tsuchiya, 2026) cover generic inexactness, and Choudhury et al.

(2026) give a nonconvex matrix framework with Nesterov momentum; none of these is contradicted by a finite-window residence at fixed (B, β) on a fixed instance.

Q.3 Repairs, and what is classical

That error feedback repairs biased compression and signSGD is classical (Karimireddy et al., 2019), and convex Lipschitz theory for an error-feedback Muon already exists (Parshakova et al., 2026). We therefore do not offer another error-feedback variant, and the repair of Section 6 adds no state, no extra sampling and no singular value decomposition. The abstract correspondence between zero-homogeneous monotone maps and gradients of convex one-homogeneous potentials is also classical (Kushnir & Krishnamoorthy, 2022); what Theorem 50 adds is a specific potential with a two-sided angular curvature bound valid at any rank and with repeated singular values, from which Corollary 52 gives a quantitative alignment inequality needing neither a density nor a derivative bound, and (127) gives strict ratio compression. The log-concavity facts we use, stability under convolution and weak limits, are from Saumard & Wellner (2014); the density-variance inequality and every barrier are proved here.

Smooth convergence analyses (Sato et al., 2025; Chang et al., 2025; Cutkosky & Mehta, 2020; Sun et al., 2023) hold at β fixed or approaching one on larger classes, or measure a different stationarity notion. None conflicts with Theorem 31, which concerns an explicit finite window at fixed (B, β) under explicit step-size conditions and makes no claim about the $T \to \infty$ limit at vanishing step size. Classical $\Omega(T^{-1/2})$ lower bounds (Arjevani et al., 2023; Drori & Shamir, 2021) constrain a larger class and therefore do not bound our smaller one from below.

Q.4 The claim we do make

The identifiable combination is: rigidity for Frobenius-normalized pointwise spectral responses including increasing ones; recovery failure under bounded centrally symmetric transverse matrix noise with a noiseless signal-direction observation; a finite-time two-sided variance-scale residence for the *actual* nonzero EMA from zero momentum on a fixed task; the finite-depth sign pattern (16); and a coupled convex-potential repair that keeps strict spectral flattening while removing the residence at the same sampling budget. The first and third ingredients concern different levels: map structure and an actual stochastic trajectory, respectively.

R Experimental protocol, and what each measurement is

This appendix specifies the experimental conditions for Section 7, including independent random streams, statistical units, and whether a reported fraction refers to a projected subspace or the full parameter space.

R.1 The synthetic experiments

All three synthetic experiments run on CPU in float64.

E1, the map and field identities. Profiles are evaluated at depths $K \in \{1, 2, 3, 5, 7\}$ plus exact polar, Frobenius normalization and both repairs, on a grid of 6001 amplitudes in $[0, 6]$, with $\sigma = 1$ and the frozen field taken over $B \in \{1, 4, 16, 64\}$ and $\beta \in \{0.5, 0.9, 0.99\}$ against a numerical approximation to the memory-truncated stationary law (radial quadrature with 3000 radii and 720 angles, and a 131072-point transform for the coordinate law). E2 uses a 65536-point transform and $\sigma = 1/16$. Four error sources are kept separate and none is called exact: the memory tail dropped by truncating the EMA at m_{win} innovations, bounded by $\sigma\beta^{m_{\mathrm{win}}}$; the transform-domain truncation of the characteristic-function product; the discretization of the radial and angular grids; and floating-point rounding. Every projection is computed twice, once from the reduction (14) and once from genuine matrix arithmetic on 2×2 matrices, and the recorded maximum discrepancy across all depths, amplitudes and noise angles is 2.1×10^{-14}. The regularizer ablation sweeps $\varepsilon \in \{0, 10^{-7}, 10^{-5}, 10^{-3}, 10^{-2}, 10^{-1}\}$ at $K \in \{1, 2, 3, 5\}$. Two additional controls establish that the signal-direction gradient is noiseless to 4.4×10^{-16}, and the matched-v cells agree in second moment to 2.8×10^{-17} while their frozen fields differ, which is the discrepancy we report as small rather than absent. The noise law is exactly centrally symmetric by construction; the finite-sample

third moment is only a numerical control, not the reason symmetry holds. Output length is likewise not the explanation: at $t = 0.3$ the reference and repaired maps have output norms 1.45 and 0.93, the same order, while their projections have opposite signs; and the frozen-field sign follows the local coefficient sign in all 24 (B, β, K) combinations tested.

E2, the trajectories. $T = 5 \times 10^4$ steps, 256 independent trajectories per cell, 62 cells comprising two 24-cell grids over $B \in \{1, 4, 16, 64\}$, $\beta \in \{0.5, 0.9, 0.99\}$ and $K \in \{1, 5\}$, plus 14 matched-v settings with additional momentum values. The three sweeps use fixed step size, step size scaled with memory, and matched v at different (B, β). The three stay separate because they ask different things: a fixed step size with varying memory probes the memory effect, a step size scaled with the memory probes the noise-scale limit, and fixed v at different (B, β) probes equal second moment against unequal history. Reporting them merged would confound the three. The unit of replication is the independent trajectory. Random streams are independent across cells, and the three arms within each cell consume the *same* noise stream; the realized radius sequences match in all 62 cells. We compare updates and samples, since a larger batch at the same update count is not the same sample budget. Step sizes come from a 2000-step pilot at the reference cell, never from the trajectories being reported. Metrics are taken at the pre-update momentum. The fixed tail window is the last fifth of the horizon. All 62 cells completed. Values below the reporting threshold are excluded from algorithmic comparisons. The earlier tail-separation interpretation is withdrawn: subtracting nearly equal spectral responses in float64 produces an arithmetic plateau, whose magnitude is not an optimization guarantee.

E2 threshold replay and the actual statistical unit. The revised threshold $|x_t| \le 10^{-8} x_0 = 2.5 \times 10^{-9}$ was selected after the original E2 runs and before the replay; it is not an originally preregistered outcome. The replay repeats the 62 original cell configurations with the same seeds and measures the first passage of each of the 256 individual paths after every completed update. It matches all original checkpoint and tail quantiles exactly and reproduces every complete radius sequence. All 15872 repaired paths and all 15872 normalized-descent paths reach the threshold and are below it at the 50000-update endpoint. Taking the median first passage within each cell, then the median of those 62 cell medians, gives 454.5 and 456 updates respectively. These two levels of aggregation are explicit; the different cells are not independent replications of one setting. For the displayed $B = 1, \beta = 0.9, K = 1$ cell, the path medians are 438 and 430 updates.

First passage alone does not distinguish recovery: 1606 reference paths cross transiently, but none is below threshold at the endpoint. Thus the corrected comparison concerns sustained small gradients over the observed run, with endpoint membership reported separately from first passage. Figure 5c ends each quantile curve at its first saved checkpoint below threshold and clips the visible quantile ribbon at that threshold. The downward markers indicate this censored checkpoint, rather than a per-path first-passage estimate. No sub-threshold value is displayed. The horizontal axis uses the recorded pre-update state's index t, which is the actual number of completed updates.

E3, the bridges. Two families: a μ-regularized quadratic with $\mu \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ in dimensions $\{2, 4, 8\}$, and multi-output regression at 16×16 and 64×64 with condition numbers $\{3, 10, 100\}$ and noise oriented at $\{0, \pi/8, \pi/4, \pi/2\}$ to the gradient. Seven arms including plain SGD and heavy-ball momentum. The equivariance control is the load-bearing one: these maps are bi-orthogonally equivariant to 5.6×10^{-15}, so rotating signal and noise by the same orthogonal pair changes nothing, and the only valid geometry ablation fixes the noise covariance spectrum and rotates it relative to the gradient. Label scale is recorded because it grows as μ shrinks, reaching 10^3 at $\mu = 10^{-3}$.

R.2 Archived transformer experiments

The archived runs use the 120 232 971-byte WikiText-103 corpus, with the first 90% for training and the last 10% for validation. The model is a pre-norm byte-level decoder: width 384, six layers, six heads, context 256, learned positions, GELU feed-forward blocks at four times the width, no biases and an untied output head. It has 10 921 728 parameters, including 25 matrix tensors with 10 715 136 parameters. Matrix maps are evaluated in float64 and cast back to float32. These losses are byte-level measurements, not standard tokenized WikiText benchmark scores.

The optimizer uses momentum .95 with Nesterov weighting, the reference depth $K = 5$, $\varepsilon = 10^{-7}$ and the same shape factor across arms. Non-matrix parameters use AdamW at 3×10^{-4} without

weight decay. Runs have no learning-rate schedule, warm-up or gradient clipping. These choices differ from the plain EMA and zero regularizer in the trajectory theorem. New paired experiments use the separate, fully recorded protocol of Appendix T.

Historical diagnostics. The earlier analysis reported a 58% detection rate with seed-level interval $[57, 60]\%$ across 1200 matrix-checkpoint measurements from eight trunk seeds, using Lanczos to seek a negative eigenvalue. This is a detection rate, not the population fraction of nonmonotone inputs: failure to find a negative direction certifies nothing. The old rank-24 diagnostic first truncates and renormalizes the momentum, so it measures a different map and cannot give full-space negative energy at the actual input. The revision replaces both as primary evidence with the exact full-spectrum decomposition. The historical fork analysis treated checkpoints as independent for standard errors and used a retrospective median split when one gradient-response sign group was missing. Its effect estimates are not used in the revised claims. Fresh runs verify complete state restoration and aggregate within seed first.

Step-size sweeps. At 2500 updates, the reference, quartic, uncoupled ablation and normalized descent each use ten step sizes at two seeds. Their best seed means are 1.261, 1.363, 1.365 and 1.384; the reference chooses .008 and the other three .064. The Schatten grids contain four step sizes with uneven replication; $\varsigma = 1.05$ gives 1.266 and $\varsigma = 1.15$ gives 1.274, while the $\varsigma = 1.5$ winner is at its grid edge. These results do not establish matched performance at equal tuning budgets. The uncoupled ablation is not universally monotone despite repairing the synthetic task.

R.3 Computational cost

Runs use one GPU each across six eight-GPU nodes. Every arm of the wall-clock comparison runs on the same node type, so the reported ratios ($18\times$ the reference and $23\times$ the quartic for the Schatten subgradient, at the median over runs) compare like with like; they are measurements on this hardware and not a claim about the maps in general. The median is taken over unique configurations, with duplicate records counted only once.

R.4 Map-level flattening and monotonicity

Table 22 uses a Jacobian assembled by central differences, independently of the differentiation used in the GPU experiments. The input spectrum decays geometrically to stable rank 2.5; flattening is measured at dimension 16 and the dense Jacobian at dimension 6. Bi-orthogonal equivariance of these maps holds to 3×10^{-15} in the same calculation, which is what licenses measuring at a diagonal representative.

Table 22: Update maps measured from their definitions on a decaying spectrum of stable rank 2.5, at dimension 16 for the flattening and 6 for the dense Jacobian. Output stable rank is the flattening achieved; $\lambda_{\min}/\lambda_{\max}$ of the symmetric Jacobian part is negative exactly when a non-monotonicity witness was found there. Values at 10^{-11} are the central-difference floor, that is numerical zero, and are not evidence of monotonicity beyond what the theorems already give.

	Ψ_5	Ψ_{unc}	$\widetilde{\Psi}_{1/8}$	$P/\|P\|_F$	$\Psi_{(\varsigma)}$, ς=1.05	polar
output stable rank	11.8	3.1	3.1	2.5	13.7	16.0
$\lambda_{\min}/\lambda_{\max}$	-0.47	-5×10^{-3}	$+1 \times 10^{-10}$	$+8 \times 10^{-11}$	$+9 \times 10^{-12}$	-4×10^{-12}
witness found	yes	yes	no	no	no	no

R.5 Numerical verification

Numerical checks cover the following identities and constants: the exact two-dimensional reduction against genuine matrix arithmetic, the shared zero $t_\star$ at every depth, the local recovery coefficients, the repaired profile and its sharp task envelope $[7/16, 23/24]$, the sharp curvature bracket (118) together with both of its equality cases and the fact that no transverse direction beats $1 - \frac{9}{2}\lambda$, convexity across $\lambda < 2/9$ and its failure above, the Schatten slope formula (139) against central differences and its vanishing at $\varsigma = 1$, the depth sign string including the break at $K = 32$ and the period-eight tail, the exact invariance of the fixed-radius stall at five depths with its multiplier, the three-dimensional

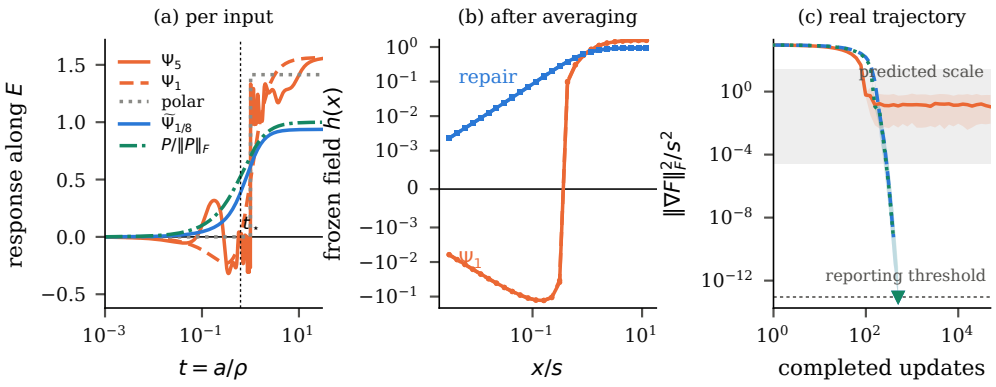


Figure 5: Recovery on the transverse construction. (a) Unit-radius response. (b) Response averaged over stationary momentum noise: solid $B = 1, \beta = 0.9$, dashed $B = 16, \beta = 0.99$. (c) Actual zero-initialized EMA at $B = 1, \beta = 0.9, K = 1$, showing the median and 10–90% quantiles across 256 paths. Values below $10^{-8}x_0$ are censored. The shaded residence scale is not a guarantee at the simulated step size. Downward markers denote the first saved checkpoint below the threshold.

witness (142), the singular-value ratio compression (127), the polar dead zone, and the stationary law's second and inverse moments against (70). Numerical agreement is not a proof of the theorems, and where a barrier is computed by quadrature rather than certified by interval arithmetic the text says which.

S Supplementary construction and checkpoint experiments

We separate three questions: whether the construction is numerically reproduced, whether controlled perturbations preserve its mechanism, and whether the diagnostic explains an intervention during training. Experimental settings, independent random streams and statistical aggregation are specified in Appendices R and T.

Synthetic residence and reliable recovery measurements. The original sweep contains 62 cells with 256 independent trajectories each: two 24-cell grids over $B \in \{1, 4, 16, 64\}$, $\beta \in \{0.5, 0.9, 0.99\}$ and $K \in \{1, 5\}$, plus 14 matched-variance settings. The memory-scaled sweep has reference tail residence 0.99–1.00 across its 24 cells; at fixed step size, the $\beta = 0.99$ cells lose residence as batch size grows. These simulated steps do not satisfy the sufficient theorem conditions. The shaded band in Figure 5 is a predicted scale, rather than a certified experimental guarantee.

For the coupled repair and normalized descent, tiny apparent plateaus reflect cancellation in the original float64 spectral subtraction. We therefore report the threshold $|x_t| \le 10^{-8}x_0$, chosen after the original runs and before an exact trajectory replay. The replay matches every original checkpoint quantile and the noise streams in all 62 cells. All 15872 paths of each repair reach the threshold and finish below it at update 50,000; the median of the 62 cell-wise median first-passage times is 454.5 updates for the quartic and 456 for normalized descent. For the reference, 1606 paths cross transiently, but none finish below the threshold. First passage alone would therefore confuse transient crossings with a low terminal gradient. Curves and uncertainty bands are censored at the reporting threshold; floating-point decades are not used as an effect size.

Changing noise orientation and transverse curvature. A new 40-cell experiment fixes the signal direction and rotates one noise basis toward it by $\theta \in \{0, .01, .05, .1, .2, .4, .8, 1.2\}$, preserving central symmetry, the covariance eigenvalues and noise bound. It adds transverse curvature $\mu \in \{0, .001, .01, .1, 1\}$ to the quadratic objective. Each cell has 64 independent paths, six map arms, zero momentum, $\sigma = 1/16$, $\beta = .9$, $B = 1$ and 30,000 steps. This is a local recovery test with $x_0 = .1s$, distinct from Definition 7's fixed initialization. Positive map scalings are fixed at one common input before the sweep. Frozen responses use independent noise draws; trajectories share the same noise within a path.

Table 23: Paired 200-update validation effect of changing the matrix map, relative to continuing the reference iteration. Positive values mean higher byte-level loss. Intervals use eight independent seed averages; learning rate and pilot gains are fixed.

Map replacement	Δ validation loss	95% seed CI
Quartic potential	$+0.1022$	$[+0.0993, +0.1050]$
Uncoupled quartic	$+0.1090$	$[+0.1073, +0.1108]$
Normalized descent	$+0.1025$	$[+0.0992, +0.1057]$

At $K = 1$, rotation from $\theta = 0$ to $.4$ changes the dimensionless finite-scale response ($h/s = .001$) from -2.04 to $-.35$, while the mean tail $|x|/s$ at $\mu = 0$ falls from .352 to .130. At $\theta = .8$ the response is $+1.41$ and the tail is .012. Depth five is more sensitive: $\theta = .1$ changes its response to $+4.07$ and the tail to .0013, while $\theta = .4$ restores a negative response ($-.65$) and residence (.210). The main pattern survives $\mu = 1$; the full grid and intervals are shown in Figure 6. Thus the mechanism persists beyond a flat transverse objective and noiseless signal observations, but is sensitive to relative noise geometry. When $\theta > 0$, the repaired arms also have residual noise.

At $\theta = .8$ their tail amplitude exceeds both reference arms. This is a controlled mechanism test, not independent validation of a general training predictor.

Paired transformer interventions. We rerun eight independent seeds of a 10.9M-parameter byte-level transformer on WikiText-103. At three checkpoints per seed, every map arm inherits the full training state and consumes the same 200 future minibatches. Layer scalings are fixed at a separate earlier pilot. Reference gradients, gradient replicates and validation occupy disjoint data partitions. We use the full spectral decomposition of Proposition 10, independently checked against dense Jacobians, to compute both the gradient-direction Rayleigh quotient and its negative-eigenspace energy. The finite-scale conditional response respects the implemented Nesterov input; this differs from the plain EMA in the trajectory theorem.

The reference has a negative direction at all 600 measured matrix inputs; the quartic has none. Yet all 600 reference-gradient Rayleigh quotients and conditional responses are positive. The reference gradients are held-out estimates, not known population gradients. Their mean energy in the negative eigenspace is 21.3%, while their net response remains positive. With zero negative primary checkpoints out of 24, the sign interaction is unidentifiable. Table 23 averages checkpoint differences within each seed before forming intervals across seeds. Under this fixed-gain intervention, all three replacements increase validation loss; the experiment does not support attributing training harm to the observed adverse eigenspaces.

Training loss and computation. The archived 2500-step sweep gives seed-mean validation losses 1.261 for the reference, 1.363 for the quartic, 1.365 for its uncoupled ablation, and 1.384 for normalized descent, each selected over ten step sizes at two seeds. Schatten $\varsigma = 1.05$ and 1.15 attain 1.266 and 1.274 under smaller, unevenly replicated grids. These are descriptive results with unequal tuning budgets, not evidence of matched optimizer performance. The quartic's weak flattening and the Schatten family's SVD cost are material tradeoffs. Structural alignment guarantees alone do not establish a training advantage.

T CONTROLLED ROBUSTNESS AND PAIRED CHECKPOINT EXPERIMENTS

These experiments test robustness to relative noise orientation and transverse curvature, and measure the effect of changing the update map at paired training checkpoints. All specified conditions are included in the analysis, including negative findings.

T.1 RELATIVE NOISE ROTATIONS AND TRANSVERSE CURVATURE

Write $W = xE + yC_0 + zC_1$ on the invariant three-dimensional subspace and use

$$F_\mu(W) = \tfrac{1}{2}[x^2 + \mu(y^2 + z^2)], \qquad \xi_\theta = \frac{\sigma}{\sqrt{2}}\{U(\cos\theta\, C_0 + \sin\theta\, E) + VC_1\},$$

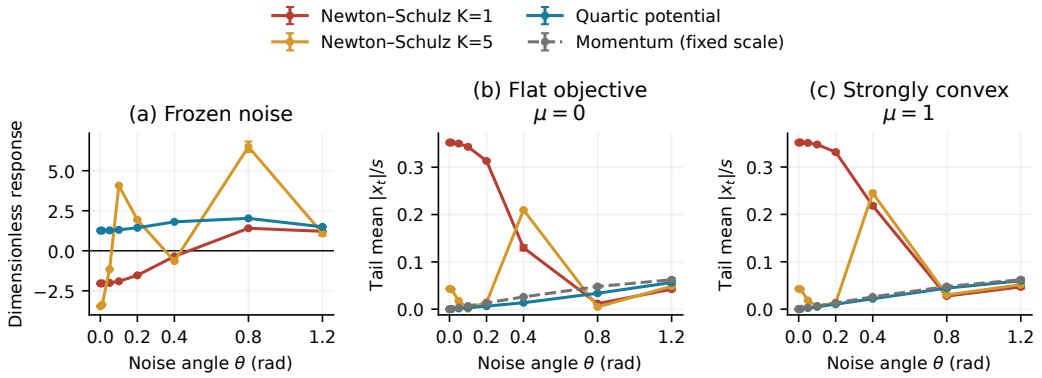


Figure 6: All prescribed relative-noise angles. (a) Independently estimated dimensionless local response. (b,c) Mean tail signal amplitude at transverse curvature $\mu = 0$ and $\mu = 1$. Intervals use independent noise draws in (a) and independent trajectories in (b,c). Negative recovery and residence track one another in these controlled interventions, with sensitivity to depth and angle; the repaired arms acquire residual signal noise when $\theta > 0$.

where U, V are independent uniforms on $[-1, 1]$. Extend the objective to the full matrix space with curvature μ on the remaining orthogonal coordinate. For $\mu > 0$ this is strongly convex; its sample losses can be chosen as shifted nonnegative quadratics. Their gradients are not assumed globally bounded. The two nonzero noise covariance eigenvalues, total variance, central symmetry and norm bound are unchanged by θ. The signal axis E and the objective axes are held fixed while the noise basis rotates; this is a relative rotation rather than a joint orthogonal change of coordinates. For $\theta > 0$ the signal observation becomes noisy.

We run the Cartesian product of $\theta \in \{0, .01, .05, .1, .2, .4, .8, 1.2\}$ and $\mu \in \{0, .001, .01, .1, 1\}$, with 64 independent complete paths per cell. The six arms are depth-one and depth-five reference iterations, the quartic potential, its uncoupled ablation, Frobenius normalization and ordinary momentum. All use zero initial momentum, $\sigma = 1/16$, $B = 1$, $\beta = .9$, $\eta = .001s$, and $T = 30{,}000$; the tail is the final 6000 updates. The initial signal $x_0 = .1s$ intentionally probes local recovery and differs from the fixed $x_0 = 1/4$ in Definition 7. These steps are not certified by the residence theorem. Each map receives one positive constant matching the depth-one output norm at the prespecified input $(x, y, z) = (.1s, s, 0)$, and that constant is frozen over the entire grid. No output normalization changes the map from step to step.

Paths use independent random streams derived from seed 918203, with each path's stream shared across maps and interventions. Frozen responses use a separate seed 179203, 131,072 independent draws and 256 EMA burn steps; the remaining stationary memory has norm at most $\sigma\beta^{256}$. Paired finite differences use the same draw at $h/s \in \{.001, .01, .1\}$, and the displayed local response uses .001. The response is multiplied by s and the fixed map gain. Confidence intervals for a tail statistic average time within a path first, then treat the 64 paths as independent.

The stable implementation propagates the half-sum and half-difference of the two signed spectral arguments analytically through the quintic. It therefore does not subtract nearly equal final spectral values to extract a small signal. The reduction agrees with independent full-matrix SVD evaluation to 2.6×10^{-14}. Raw tiny values are retained but are not interpreted as performance differences; figures use linear amplitude axes. At nonzero tilt, a residual noise floor is expected even for the repaired maps.

T.2 A FULL-STATE PAIRED TRANSFORMER PROTOCOL

Eight independent training seeds use the same 10.9M-parameter architecture as Appendix R.2, with 3000 trunk updates, $B = 32$, $K = 5$, momentum .95, learning rate .008, $\varepsilon = 0$ and float64 map evaluation. One H20 GPU runs each seed. The corpus is split by byte ranges: training $[0, .80N)$, reference $[.80N, .85N)$, gradient replicates $[.85N, .90N)$, and validation $[.90N, N)$. The reference

Table 24: All noise angles: finite-scale responses R_K at $h/s = .001$, and mean tail $|x_t|/s$ at $\mu = 1$. Values below 10^{-9} are censored for presentation.

θ	R_1	R_5	NS-1	NS-5	Quartic
0	-2.039	-3.466	0.3517	0.0430	$< 10^{-9}$
0.01	-2.038	-3.365	0.3517	0.0423	0.0005
0.05	-2.006	-1.165	0.3508	0.0186	0.0026
0.1	-1.907	4.069	0.3475	0.0058	0.0052
0.2	-1.534	1.936	0.3314	0.0139	0.0106
0.4	-0.349	-0.646	0.2182	0.2448	0.0221
0.8	1.405	6.496	0.0277	0.0303	0.0444
1.2	1.216	1.089	0.0470	0.0517	0.0601

gradient averages 512 sequences, each of eight gradient replicates uses 32, and the fixed validation set contains 768 sequences. Validation indices are identical across arms, checkpoints and seeds. These are disjoint data partitions, not merely distinct random generators sampling the same pool.

The pilot at update 199 fixes per-layer gains for all later forks. Immediately before updates 999, 1999 and 2999, each checkpoint is forked for 200 updates into reference, quartic, uncoupled and normalized-descent arms. The snapshot includes model parameters, matrix momentum, auxiliary AdamW moments and counters, all gradients, CPU/CUDA random states, data-generator state and training mode. Each arm starts from an identical complete state and consumes identical future examples; both equalities are verified. After all probes and forks, the trunk state must match its saved state bitwise. A separate audit compares a trunk without interventions, a normal arm order and a reversed arm order; the trunks and matched fork outcomes agree exactly.

At every actual optimizer input P, Proposition 10 supplies the complete symmetric-Jacobian spectrum using all singular values. Numerical negatives use a relative threshold 10^{-10} times the largest absolute eigenvalue. The diagnostic is checked on 72 independent small matrix cases against a dense automatic-differentiation Jacobian, including repeated, rank-deficient and rectangular inputs; the maximum normalized error is below 1.5×10^{-13}. The primary statistic is

$$\chi(P, G) = \frac{\|P\|_F}{\|\Psi(P)\|_F} \frac{\langle G, \mathcal{A}(P)[G] \rangle}{\|G\|_F^2}.$$

Negative-eigenspace energy and dimension fractions are secondary statistics. They are full-space quantities rather than properties of a truncated and renormalized momentum.

The conditional finite-scale probe uses the historical buffer M and independent gradient replicates G_j. The actual Nesterov input is $\beta^2 M + (1 - \beta^2) G_j$. We perturb this input by $\pm h G / \|G\|_F$ at three scales $h/\|P\|_F \in \{.001, .01, .1\}$, using paired draws. This measures a conditional response at the frozen checkpoint; it is not the frozen stationary law of the synthetic EMA theorem.

For the primary validation effect, subtract the reference arm within each checkpoint, average the three differences within each seed, and compute a Student-t interval over the eight seed means with seven degrees of freedom. The 24 checkpoints are repeated measurements, not independent training runs. A response-sign interaction is reported only when both signs are observed. Otherwise it is unidentifiable; no retrospective median split of another statistic replaces the missing group. The protocol estimates a map-replacement effect at fixed learning rate and pilot gains, not an equally tuned optimizer ranking.

U Fresh task-response audit with independent gradients

A response measured with a noisy gradient estimate can differ from the response to the underlying gradient. This study checks whether correcting that estimation effect changes the sign of the measured response. We also measure the contribution from the normalization denominator. The study uses eight new training seeds and 600 inputs; its results are kept separate from the earlier 600-input checkpoint-fork study.

Training setup and measured gradient. The source and protocol were frozen before the eight runs, using new training seeds 400–407. Each run trains the existing byte-level transformer with width 384, six layers, six heads, context 256, batch size 32, 3000 updates, learning rate .008, Nesterov momentum .95 and the five-step reference map. Matrix maps use float64; parameters use the existing float32 training implementation. All eight H20 runs use PyTorch 2.7.1. No learning rate, checkpoint or layer is selected according to response sign. The diagnostic differentiates the real-arithmetic spectral map, evaluated in float64 at the captured input. Final parameter casts belong to the executed-loss diagnostic and are not treated as a smooth operation covered by the Jacobian theorem.

The first 80% of the fixed WikiText corpus supplies training crops. The remaining 20% defines a finite evaluation objective: expected cross-entropy over a uniformly sampled context-length crop. At each checkpoint the 16 diagnostic batches, each of 64 crops, are independently sampled with replacement from this same pool, independently of the batch forming the actual optimizer input. This common distribution is needed for Proposition 20; different corpus ranges would not give two estimates of the same g. It is an evaluation gradient, not an exactly observed population training gradient. Conditional on the checkpoint, the sampling estimates are unbiased for that specified crop objective.

At updates 500, 1500 and 3000, the implementation captures the exact float64 argument passed into every reference matrix-map call. It first executes the ordinary training step, saves the complete post-step state, restores the pre-step state for diagnosis, and then restores the executed post-step state. Checkpoint hashes verify that both restorations preserve model parameters, gradients, matrix momentum, auxiliary AdamW state and RNG. The training step is neither repeated nor replaced by a diagnostic update. This avoids the operation-order discrepancy between a separately reconstructed input and the actual map argument noted in Appendix O.

Computing the response and its uncertainty. For each matrix block, the exact spectral decomposition of Proposition 10 yields all 16×16 bilinear products for $\rho \operatorname{sym} D\Psi(P)$, the fixed-denominator derivative, the normalization term, the tangential and cross terms, and the positive and negative spectral contributions. It uses the complete spectrum, including changes in singular vectors and rectangular complements. There is no singular-value truncation in this diagnostic. The saved matrix of each component rebuilds the plug-in estimate and the off-diagonal U-statistic.

The primary aggregate is $\sum_i a_i \langle g_i, D\Psi(P_i)[g_i] \rangle$, with the optimizer's shape factor $a_i = \max(1, m_i/n_i)^{1/2}$; the learning rate is not included in this response statistic. Its unbiased estimate sums the layer cross-estimates divided by their captured ρ_i and multiplied by a_i. Each seed first averages its three checkpoints; the primary 95% Student-t interval then uses the eight independent seed means. A complete-seed bootstrap with 50000 draws is a sensitivity analysis. The layer counts and component intervals are descriptive; no simultaneous layerwise sign test is claimed.

The primary cross-estimate is $8.778\,[8.247, 9.309]$, with seed-bootstrap interval $[8.366, 9.186]$. The plug-in value is $11.751\,[11.216, 12.287]$; their paired difference is $2.973\,[2.930, 3.016]$, or 25.3% of the plug-in mean. Despite this substantial estimation effect, all 600 layer point estimates and all 24 checkpoint aggregates remain positive after correction. Every one of the 600 inputs still has a negative eigendirection. The latter does not determine the task response.

The decomposition makes the distinction quantitative. The positive spectral contribution is $10.059\,[9.489, 10.630]$ and the negative magnitude is $1.281\,[1.219, 1.344]$. Their difference is the net response. Holding the denominator fixed gives $8.821\,[8.243, 9.399]$; the subtracted normalization term is $.043\,[-.078, .164]$. In radial and tangential coordinates, the tangential term is $8.777\,[8.247, 9.308]$ and the cross term is $.00091\,[-.03639, .03820]$. Thus the audit does not establish a negative normalization effect on the mean task response in this measured setting. It also does not infer that the term is identically zero.

Measuring the loss of the executed step. Sixteen independent outcome batches per checkpoint compare the executed update with its pre-step state on common crops, using a random stream independent of the gradient batches. The complete-seed mean loss change is -0.000381 with interval $[-0.000865, +0.000103]$. The interval crosses zero. The update includes the auxiliary AdamW parameters and ordinary finite-step curvature; its loss difference cannot be identified with the matrix Jacobian response. Proposition 21 states the additional conditional alignment and curvature requirements. No causal training-harm claim is extracted from this paired one-step diagnostic.

Checks and interpretation. The supplementary audit material contains the training, response-evaluation and analysis implementations, frozen execution copies, the protocol manifest, all eight completed seed records and the 16×16 component matrices. An input fixture records the first matrix block at the first checkpoint of the first prespecified seed. Independent forward-mode differentiation of the matrix polynomial agrees with its spectral bilinear response to relative error 4.3×10^{-15}. Small square, rectangular, repeated-spectrum and rank-deficient matrices are also checked against dense automatic differentiation. Exact rational enumeration checks the unbiased estimator's mean and variance, including an indefinite response operator.

Cross-estimation corrects one source of statistical bias. It does not prove positivity for every population layer, convert a checkpoint Jacobian into a zero-signal frozen field, average over the actual conditional EMA law, or remove a training/evaluation distribution difference. The eight-seed result is evidence against activation of the proposed adverse local mechanism on these measured task gradients. It strengthens the empirical boundary of the structural theorem; it does not establish practical harm that the theorem's worst-case quantifier alone cannot supply.

V GENERIC LOCAL LOSS DIAGNOSTICS (AUXILIARY)

From an optimizer response to a loss prediction. The sign of a map's Jacobian is a property of the optimizer; a training claim also requires the task gradient, the actual parameter displacement and loss curvature. The following identity makes that distinction operational. At parameters W, apply a single positive or negative task-gradient pulse to each matrix input,

$$P_i^{\pm} = P_i \pm h_i E_i, \qquad h_i = \rho \|P_i\|_F, \qquad \|E_i\|_F = 1.$$

Let $d_{\pm}$ denote the resulting complete parameter displacements, including the common auxiliary AdamW update, and write $c = (d_+ + d_-)/2$ and $b = (d_+ - d_-)/2$. For a twice differentiable population loss with gradient g and Hessian H at W,

$$F(W + d_+) - F(W + d_-) = 2\langle g, b\rangle + 2\langle c, Hb\rangle + \mathcal{R}. \tag{181}$$

If the Hessian is L_H-Lipschitz on the segments joining W to $W + c + tb$, $-1 \le t \le 1$, then

$$|\mathcal{R}| \le L_H \|b\| \left(\|c\|^2 + \frac{\|b\|^2}{3} \right). \tag{182}$$

Indeed, integrate $\langle \nabla F(W + c + tb), b\rangle$ over $t \in [-1, 1]$ and use the first-order Taylor remainder for the gradient. The odd cross term in $\int_{-1}^{1} \|c + tb\|^2 dt$ vanishes. This bound respects the common center of the paired updates and vanishes with the pulse magnitude.

For exact arithmetic, the matrix contribution to the first term is

$$2\langle g, b\rangle = -\sum_i \eta a_i \left\langle g_i, \Psi(P_i + h_i E_i) - \Psi(P_i - h_i E_i) \right\rangle,$$

where $a_i > 0$ includes the frozen pilot gain and shape factor. Its local limit is

$$-2 \sum_i \eta a_i h_i \langle g_i, D\Psi(P_i)[E_i]\rangle.$$

Thus the optimizer response enters a concrete local loss prediction through its projection on the task gradient; the curvature term quantifies how the ordinary update changes that projection. The identity alone does not identify an adverse recovery mechanism. With independent estimates $\widehat{g}$ and $\widehat{H}$, the additional error is bounded by $2\|b\|\|\widehat{g} - g\| + 2\|c\|\|b\|\|\widehat{H} - H\|_{\mathrm{op}}$. We do not claim certified population bounds for the neural Hessian constants; we test these frozen, quantitative predictions on a separate outcome pool.

Transported response with a sharper remainder. The common step can be large even when the pulse is small. Moving the task gradient to the known midpoint removes that common-step truncation error.

Proposition 73 (Midpoint task-response prediction)**.** *Let $d_\pm = c \pm b$ be two predetermined parameter updates, and suppose F has an L_H-Lipschitz Hessian on $\{W + c + tb : -1 \le t \le 1\}$. For any estimate $\widehat{g}_c$ of $\nabla F(W + c)$,*

$$|F(W + d_+) - F(W + d_-) - 2\langle \widehat{g}_c, b\rangle| \le 2|\langle \widehat{g}_c - \nabla F(W + c), b\rangle| + \frac{L_H}{3}\|b\|^3. \tag{183}$$

Consequently the prediction certifies the sign whenever its magnitude exceeds the right-hand side. Condition on the checkpoint and the complete predetermined intervention pair (c, b). If the independent scalar single-example derivative $\langle \nabla f(W + c; Z), b\rangle$ has sub-Gaussian variance proxy v_b and $\widehat{g}_c$ averages n such examples, its first error term is at most $2\sqrt{2v_b \log(2/\delta)/n}$ with probability at least $1 - \delta$.

Proof. Integrate $\langle \nabla F(W + c + tb), b\rangle$ over $[-1, 1]$. The constant term is $2\langle \nabla F(W + c), b\rangle$, and the linear Hessian term integrates to zero. The norm of the gradient Taylor remainder is at most $L_H t^2\|b\|^2/2$; integration gives $L_H\|b\|^3/3$. Adding and subtracting the estimated gradient proves the deterministic bound. The probability statement is the scalar sub-Gaussian concentration inequality applied to the independent directional-gradient estimates. □

The derivative factor remains the optimizer's finite task response: $2b_i = -\eta a_i[\Psi(P_i + h_i E_i) - \Psi(P_i - h_i E_i)]$ on matrix blocks. The transported predictor estimates its loss consequence at a single virtual midpoint; it does not evaluate the loss of either candidate update. Compared with (182), its deterministic remainder has no $L_H\|b\|\|c\|^2$ term. The theorem states assumptions under which the prediction is certified. The neural experiment independently measures accuracy; it does not assert an estimated sub-Gaussian constant or population Hessian bound.

Prospective experimental protocol. The study uses twelve new complete training seeds of the 10.9M-parameter transformer, with 3000 reference-optimizer updates, batch size 32, context 256, momentum .95, learning rate .008 and float64 matrix maps. Seeds 100–103 are calibration runs; seeds 104–111 are held-out runs. The corpus is split by byte ranges into training $[0, .80N)$, direction estimation $[.80N, .84N)$, prediction $[.84N, .90N)$ and outcome evaluation $[.90N, N)$. The direction and prediction gradients each average 512 sequences; outcomes use the same 768 validation sequences across all arms, checkpoints and seeds. The protocol and predictor definitions are fixed before the runs.

Immediately before trunk updates 499, 1499 and 2999, every state is forked for the Cartesian product of maps {reference, quartic, homogeneous smoothed nuclear $\kappa = .001$} and pulse magnitudes $\rho \in \{.025, .1, .4\}$. The value $\kappa = .001$ was fixed by a spectral flattening criterion before training outcomes. Positive layer gains are matched to the reference at pilot update 199 and are then fixed. The first update uses $P_i^{\pm}$; the next 49 updates use the ordinary map and identical future minibatches within the pair. The pulse changes the first map input, leaving the underlying momentum recursion intact. This is an intervention on the optimizer's response to an added task-gradient signal, not a claim that the untreated training process naturally realizes the negative-response construction.

The complete checkpoint includes parameters, gradients, matrix momentum, AdamW moments and counters, model mode, and CPU, CUDA and data-generator states. Every fork starts from the same training state and uses identical future examples; the original training state is restored after probing. The actual float32 parameter displacements are also checked against those used in the prediction. At every checkpoint all predictions, including curvature, are fixed before either outcome arm is evaluated.

Prespecified predictors and evaluation. The local-response predictor uses a float64 centered derivative at relative step 10^{-4}, with a $2\times$ step consistency check. The finite-response predictor evaluates the exact positive and negative map outputs. A second predictor adds the bilinear Hessian term in (181); Hessian-vector products use the prediction pool and the attention backend supporting second derivatives. This separates finite-pulse map nonlinearity from curvature of the neural objective.

The primary endpoint is the analytic one-update prediction on the eight held-out seeds. Secondary endpoints are the paired loss differences after 10 and 50 ordinary updates. For each horizon, a ridge model with fixed penalty 1 is fit only on the four calibration seeds; feature standardization is fit on

Table 25: Prospective prediction of actual transformer loss interventions. NRMSE is relative to a zero-effect predictor; sign is percent correct, and rank is within-checkpoint Spearman correlation. Analytic predictors use no outcome calibration; calibrated rows use four training seeds and eight held-out seeds.

Prediction	Horizon	NRMSE ↓	Sign	Rank ↑
Local response	1	2.038	77.3%	0.235
Finite response	1	1.956	77.3%	0.216
Finite response + curvature	1	0.417	94.9%	0.878
Calibrated cheap	1	0.641	76.4%	0.812
+ response	1	0.707	75.9%	0.778
+ response + curvature	1	0.443	92.1%	0.876
Calibrated cheap	10	0.978	39.4%	0.200
+ response	10	0.971	38.0%	0.308
+ response + curvature	10	0.970	38.4%	0.314
Calibrated cheap	50	0.995	38.0%	-0.053
+ response	50	0.984	40.7%	0.065
+ response + curvature	50	0.990	41.2%	0.156

those seeds as well. A cheap baseline uses amplitude interactions with ordinary update projection, update norm, gradient norm, momentum staleness, gradient signal-to-noise ratio, adverse gradient energy, checkpoint and map identity. Nested models add the local response, then the directional curvature term. No feature, penalty or checkpoint is selected using held-out outcomes.

Normalized RMSE divides prediction RMSE by the RMSE of predicting zero paired effect. Signs exclude only effects below the prespecified 10^{-7} threshold. Rank correlation is calculated within a checkpoint across its nine interventions, and across maps at fixed pulse amplitude. Confidence intervals resample complete held-out seeds, retaining their checkpoints, maps and amplitudes together. The study contains 324 paired interventions, of which 216 are held out, but only eight independent held-out training seeds. All magnitudes, maps and horizons are reported, including cases where the response or forecast does not help.

Independent confirmation of midpoint transport. The midpoint refinement was formulated after preliminary results of the first study, particularly the first checkpoint of calibration seed 100. We evaluate it in a separate registered study with eight new seeds 200–207, keeping the same architecture, trunk schedule, maps, pulse amplitudes, checkpoints and disjoint data pools. All 216 paired interventions are held out; there is no outcome calibration. At each pair, the predictor estimates one gradient at the virtual midpoint on the prediction pool, seals the result, and then evaluates the two actual one-update arms on the outcome pool. The original-state response and curvature predictors are evaluated on these same new checkpoints as baselines. This confirmation does not replace or change the original 10- and 50-update tests. A supplementary sign-choice analysis reports realized loss and regret when the sealed prediction selects the positive or negative pulse; because it was added after initial registration, it is explicitly exploratory.

All runs use H20 GPUs and Python 3.12.11. The original study uses PyTorch 2.7.1 with CUDA 12.8. In the confirmation study, seeds 200–203 use that stack and seeds 204–207 use PyTorch 2.7.1 with CUDA 12.6; all maps and pulse pairs within a seed share the same environment. A descriptive post-run environment audit gives midpoint NRMSE .135 and .092 in these two four-seed groups, with 98.1% sign accuracy in each. The prespecified eight-seed analysis is unchanged.

Results and practical scope. In the first study, local response alone has held-out NRMSE 2.038; finite response with curvature reduces it to .417 (seed-bootstrap 95% interval [.317, .535]) and improves sign accuracy from 77.3% to 94.9%. Against the calibrated cheap baseline, adding response alone worsens one-step RMSE by 10.2%, whereas response plus curvature improves it by 30.9% ([15.1%, 42.2%]). For horizons 10 and 50, however, the incremental improvement from response plus curvature is only .8% and .5%, with intervals $[-2.1\%, 4.0\%]$ and $[-1.5\%, 2.7\%]$. We retain this failed transfer test: a frozen local response does not establish prediction of extended neural training.

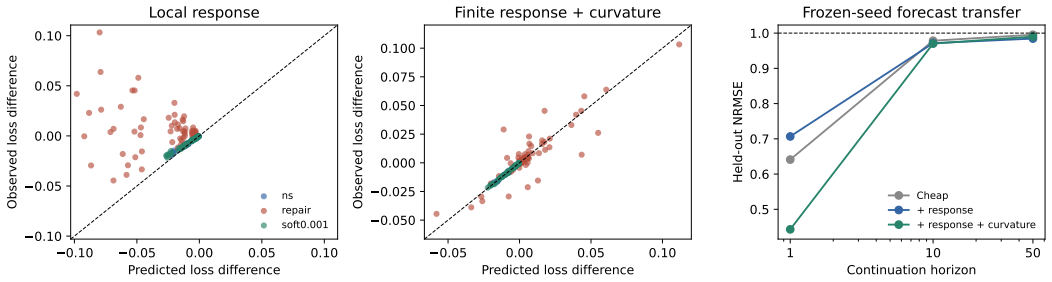


Figure 7: Held-out training-effect predictions for the prespecified task-gradient pulses. Each point is a paired loss effect; all inference treats training seeds as independent units. The right panel tests transfer of the frozen prediction features to natural continuation. These endpoints quantify controlled local training effects and do not establish prediction of final converged performance.

Table 26: Independent confirmation on eight entirely new seeds and 216 paired interventions. These analytic predictors use no outcome calibration.

Prediction	Horizon	NRMSE ↓	Sign	Rank ↑
Local response	1	1.789	76.4%	0.213
Finite response	1	1.731	76.4%	0.181
Finite response + curvature	1	0.353	97.7%	0.952
Midpoint task gradient	1	0.110	98.1%	0.974

On the eight fresh confirmation seeds, midpoint transport has NRMSE .110 ($[.089, .142]$), sign accuracy 98.1%, and within-checkpoint rank correlation .974. Its RMSE is 69.0% lower than the original-state curvature correction ($[51.7\%, 77.9\%]$). The NRMSEs at amplitudes .025, .1 and .4 are .040, .041 and .115, respectively; the larger-pulse error is retained. The fixed-amplitude map rank is .972, and the fixed-map amplitude rank is .958. Table 27 further separates map effects: all 72 reference and all 72 smoothed-nuclear outcomes are negative, so their 100% accuracy equals a constant-sign rule. The quartic stratum has $51/72$ harmful outcomes; its majority rule attains 70.8% and midpoint transport 94.4%. The NRMSE improvement over curvature is .374 to .115 on quartic, .035 to .035 on reference, and .037 to .036 on smoothed nuclear. The pooled sign score therefore does not by itself establish a mechanism across optimizers. In particular, the quartic map's positive task response coexists with both helpful and harmful loss interventions; the transported gradient predicts this distinction without asserting negative task response in natural training.

The exploratory decision rule chooses the positive pulse when the sealed prediction is negative, and the negative pulse otherwise. In the confirmation study it reduces mean outcome loss by .004405 compared with always adding the signal (95% interval $[.002515, .006154]$). Its mean excess loss above an oracle choosing the better realized arm is 2.50×10^{-5}; always adding the signal has excess loss .004430. This demonstrates an operational use of the local predictor, conditional on the prespecified intervention set. It requires 512 independent prediction sequences per candidate midpoint and is an offline diagnostic with extra gradient cost. It is not an evaluation of a continuously adapted training policy or a forecast of final performance.

Controls study: direct evaluation, shared gradient and response ablations. The midpoint identity holds for any predetermined pair of updates. Proposition 74 relates its sampling error to direct common-sample evaluation; it gives a small-pulse variance comparison, not exact equality of variances or universal mean-square-error dominance. A third registered study measures error and cost directly on 8 new seeds 300–307, with the architecture, trunk schedule, checkpoints 499, 1499, 2999, pulse amplitudes and disjoint pools of the confirmation study. Its four maps are reference, quartic, smoothed nuclear at $\kappa = .001$ and Gram-implemented Schatten-1.05, giving 288 pairs. At each checkpoint one gradient at each map's ordinary next iterate $W + d_0$ is estimated on 512 prediction sequences and reused for its three amplitudes through $2\langle \widehat{g}(W + d_0), b\rangle$ (Proposition 76). Each pair also has its own midpoint gradient, the finite response at W, the response with bilinear curvature, and direct paired evaluations on those same 512 sequences and a fixed subset of 128. Two ablations

Table 27: Fresh-seed results by map. "Harmful +" counts pairs where adding the task-gradient signal raises outcome loss relative to removing it. The three error columns are NRMSE, and sign is midpoint-prediction accuracy. Each row contains all three amplitudes and checkpoints across eight seeds.

Map	Harmful +	Local	+ curvature	Midpoint	Sign
Reference	0/72	0.272	0.035	0.035	100.0%
Quartic	51/72	1.892	0.374	0.115	94.4%
Smoothed nuclear	0/72	0.324	0.037	0.036	100.0%

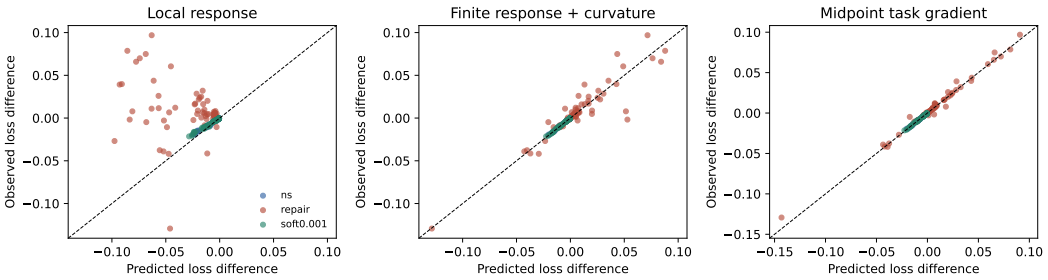


Figure 8: Fresh-seed, parameter-free prediction of actual one-update loss effects. The original response omits curvature of the task after the common update; the virtual-midpoint gradient accounts for that change and leaves a cubic pulse remainder. All maps, amplitudes and checkpoints are shown. Prediction and outcome gradients/losses use disjoint data pools.

retain the actual midpoint gradient and replace b_i by the pulse direction with the actual per-layer norm, $-\|b_i\|_F E_i$, or by the normalized-descent response with a pilot gain matching the ordinary update norm. The latter does not match the pulse response norm. All checkpoint predictions are sealed before either outcome arm is evaluated on the 768 validation sequences. Runs use PyTorch 2.7.1 with CUDA 12.6 on H20 GPUs; all pairs within a seed share the same device.

Paired uncertainty and strata. The following reanalysis uses the existing controls outcomes; it adds no training runs. All 10,000 bootstrap resamples draw eight complete seeds with replacement, retaining each seed's checkpoints, maps and amplitudes. The same draws are used for every predictor, stratum and paired difference. Intervals are pointwise percentile intervals from only eight independent training seeds, conditional on the corpus and common validation pool; they do not quantify corpus-sampling uncertainty or give simultaneous coverage across the table. The additional map-by-amplitude and amortization analyses were performed after the outcomes were available.

Pooled midpoint NRMSE is 0.110 against 0.042 for the direct 512-sequence control. Their paired difference is .0684 (95% interval $[.0413, .1123]$), and the midpoint has larger mean squared error on each of the eight seeds. Tables 30 and 31 show the concentration of this difference in quartic at $\rho = .4$: its midpoint NRMSE is .120 versus .043 for direct evaluation. On reference, smoothed nuclear and Schatten, the midpoint NRMSE is .040–.041 after pooling amplitudes. For $\rho \leq .1$, pooled midpoint and direct NRMSEs are .0378 and .0356; the small paired excess .0022 has interval $[.0008, .0048]$. These are small numerical gaps, not exact equality of the estimators. The median relative midpoint/direct discrepancy is 2.7×10^{-4} and its root mean square is 0.090 times the root mean square outcome. Same-sample discrepancies are consistent with the finite-pulse Taylor remainder, plus numerical evaluation error; total error against the disjoint outcome pool also contains sampling and pool-distribution differences. Response norms and direction cosines alone do not establish the size of the loss's third derivative. The 128-sequence direct control has pooled NRMSE 0.073, also below that of the per-pair midpoint gradient.

Sign baselines and what the ablations identify. All 216 reference, smoothed-nuclear and Schatten outcomes in the controls study are negative. Quartic has $52/72$ harmful outcomes; its majority rule attains 72.2% and midpoint transport 95.8%. A single pooled constant sign attains 81.9%, but a rule with one sign per optimizer attains 93.1%; these are different comparators. Leaving one seed out

Table 28: Controls study: one-update paired loss changes on 8 new seeds and 288 pairs. NRMSE is relative to predicting zero effect with seed-bootstrap intervals; sign is compared with the map-stratified majority rule's 93.1%; rank is within-checkpoint Spearman correlation; cost covers measured prediction blocks, with shared setup costs excluded. The direct rows evaluate both arms on the prediction pool.

Predictor	NRMSE [95% CI]	Sign	Rank	Timed cost (s)
Finite response at W	1.554 [1.485, 1.714]	81.9%	0.300	–
Response + curvature	0.290 [0.255, 0.349]	98.6%	0.966	–
Midpoint gradient (per pair)	0.110 [0.075, 0.172]	99.0%	0.990	0.56
Shared gradient (per map)	0.332 [0.290, 0.440]	98.3%	0.972	0.56/map
Direct paired evaluation, 512 seq.	0.042 [0.031, 0.063]	99.3%	0.999	0.36
Direct paired evaluation, 128 seq.	0.073 [0.051, 0.117]	99.3%	0.992	0.10
Midpoint gradient, pulse-direction response	0.621 [0.489, 0.926]	96.9%	0.862	0.56
Midpoint gradient, normalized-descent response	0.847 [0.606, 1.350]	98.6%	0.895	0.56

Table 29: Controls study by map. "Harmful $+$" counts pairs where adding the signal raises the next loss; majority is the retrospective stratum-majority sign accuracy. The three error columns are NRMSE of the midpoint gradient, direct common-sample evaluation, and the pulse-direction ablation.

Map	Harmful $+$	Majority	Midpoint sign	Midpoint	Direct	Pulse-dir. ablation
Reference	0/72	100.0%	100.0%	0.041	0.040	2.030
Quartic	52/72	72.2%	95.8%	0.114	0.042	0.366
Smoothed nuclear	0/72	100.0%	100.0%	0.040	0.039	1.592
Schatten-1.05 (Gram)	0/72	100.0%	100.0%	0.040	0.040	1.725

and learning the sign from the other seven gives the same optimizer-specific rules, as does further conditioning on amplitude. This is a retrospective baseline audit, not a newly sealed prediction study. Table 32 reports the paired sign gain. The degenerate bootstrap intervals on one-signed strata describe these eight seeds; they do not certify that unseen interventions always have the same sign.

The pulse-direction ablation has NRMSE 0.621 pooled and exceeds 1.5 on reference, smoothed nuclear and Schatten; their actual response cosines with the pulse direction are $.4$–$.5$. The normalized-descent ablation has pooled NRMSE 0.847 and changes both direction and response magnitude. Quartic's proxies are closer, with NRMSEs .37 and .16. These comparisons establish that the actual displacement direction is useful for the Taylor prediction. They do not isolate information unique to recovery geometry: any directional loss approximation can degrade when the actual displacement is replaced by a different vector. Given the checkpoint and specified pulse directions, the finite response is computable without additional samples; the directions themselves were estimated from 512 direction-pool sequences. Neither these ablations nor the midpoint sign score demonstrates that adverse recovery geometry adds predictive value beyond a matched generic local-loss model. That narrower mechanism question remains open.

Measured cost and sharing across pairs. The archived synchronized timers give 0.56 s per midpoint gradient, 0.36 s per direct 512-sequence pair and 0.10 s per 128-sequence pair. The shared gradient costs 0.56 s *per map*, not per pair. These are timed prediction blocks. Direction estimation, construction of the candidate updates and outcome evaluation are excluded. Direct-evaluation timers include parameter/state restoration, whereas gradient timers start after restoration; thus the timings are not strictly matched end-to-end compute budgets. Finite-response and curvature costs in the main table are left unreported because their recorded timers omit the gradient at W; the stored $.002$ s projection and 1.73 s Hessian-vector product are not complete predictor costs.

Table 33 reconstructs costs for the existing three-pair families and estimates the two-small-amplitude cost by scaling the recorded typical per-pair call. For $\rho \in \{.025, .1\}$, shared-gradient NRMSE is 0.038 versus 0.036 for direct evaluation, with paired difference $.0025$ ($[-.0052, .0122]$). The estimated shared cost is 0.78 times two direct 512-sequence evaluations; this is evidence for possible amortization at small pulses, without a prespecified equivalence margin. For all three amplitudes the measured cost ratio is 0.52, but shared NRMSE rises to 0.332, including $.365$ on quartic at $\rho = .4$.

Table 30: Retrospective controls reanalysis by optimizer and amplitude. Each row contains 24 pairs from eight seed clusters. Entries are NRMSE [pointwise 95% cluster-bootstrap interval]; direct and midpoint use the same 512 prediction sequences. All registered amplitudes and maps are retained.

Map	ρ	Harmful +	Midpoint	Direct (512)	Shared
Reference	0.025	0/24	0.044 [0.037, 0.051]	0.044 [0.037, 0.051]	0.044 [0.037, 0.051]
Reference	0.1	0/24	0.044 [0.037, 0.050]	0.044 [0.037, 0.050]	0.044 [0.037, 0.049]
Reference	0.4	0/24	0.041 [0.035, 0.046]	0.040 [0.034, 0.046]	0.083 [0.065, 0.097]
Quartic	0.025	18/24	0.035 [0.025, 0.059]	0.035 [0.025, 0.058]	0.035 [0.025, 0.058]
Quartic	0.1	18/24	0.038 [0.027, 0.064]	0.035 [0.026, 0.059]	0.038 [0.023, 0.069]
Quartic	0.4	16/24	0.120 [0.081, 0.201]	0.043 [0.031, 0.068]	0.365 [0.313, 0.519]
Smoothed nuclear	0.025	0/24	0.042 [0.036, 0.048]	0.042 [0.036, 0.048]	0.042 [0.036, 0.047]
Smoothed nuclear	0.1	0/24	0.042 [0.036, 0.047]	0.042 [0.036, 0.047]	0.042 [0.036, 0.046]
Smoothed nuclear	0.4	0/24	0.040 [0.034, 0.045]	0.039 [0.034, 0.045]	0.086 [0.070, 0.100]
Schatten-1.05 (Gram)	0.025	0/24	0.042 [0.037, 0.048]	0.042 [0.037, 0.048]	0.042 [0.037, 0.048]
Schatten-1.05 (Gram)	0.1	0/24	0.042 [0.036, 0.048]	0.042 [0.036, 0.048]	0.041 [0.036, 0.046]
Schatten-1.05 (Gram)	0.4	0/24	0.040 [0.035, 0.045]	0.040 [0.034, 0.046]	0.075 [0.061, 0.087]

Table 31: Paired NRMSE differences against direct common-sample evaluation on 512 sequences. Positive values favor direct evaluation. Intervals use identical seed resamples for both predictors. The last column compares mean squared errors within each seed.

Interventions	Predictor	ΔNRMSE vs. direct (512) [95% CI]	Seeds with larger MSE
All pairs	Midpoint	0.0684 [0.0413, 0.1123]	8/8
All pairs	Shared	0.2907 [0.2578, 0.3768]	8/8
All pairs	Direct (128)	0.0313 [0.0127, 0.0587]	6/8
$\rho \leq .1$	Midpoint	0.0022 [0.0008, 0.0048]	7/8
$\rho \leq .1$	Shared	0.0025 [-0.0052, 0.0122]	6/8
$\rho \leq .1$	Direct (128)	0.0282 [0.0128, 0.0510]	6/8
Quartic, $\rho = .4$	Midpoint	0.0774 [0.0462, 0.1359]	8/8
Quartic, $\rho = .4$	Shared	0.3225 [0.2806, 0.4506]	8/8
Quartic, $\rho = .4$	Direct (128)	0.0323 [0.0133, 0.0640]	6/8

Direct evaluation on 128 sequences remains cheaper than sharing the 512-sequence gradient in both families. The experiment therefore exhibits an error/cost tradeoff, not a demonstrated advantage at matched total compute. "Multiple pairs" means predicting their individual loss differences; it does not evaluate ranking all candidate loss levels or an adaptive many-action training policy.

Finite-pulse center geometry. Table 34 reports the center shift relative to the response norm. The small-pulse theorem gives $\|c_\rho - c_0\| = O(\rho^2)$ and $\|b_\rho\| = O(\rho)$ for smooth maps; the ratio is $O(\rho)$ if the first-order response is nonzero. Reference and smoothed nuclear show approximately linear decrease over the sampled amplitudes. Schatten-1.05 gives ratios .34, .19, .18 at $\rho = .4, .1, .025$. This finite-range plateau is consistent with sensitivity near small singular values, but three amplitudes do not establish a nonzero limit as $\rho \to 0$. At a fixed full-rank matrix the Schatten map is smooth locally, so a sufficiently small pulse can also enter the quadratic-center regime; its constants need not be uniform as rank deficiency is approached. These observations do not by themselves diagnose a training instability or explain the training-loss differences.

V.1 What the midpoint predictor is and is not

Proposition 73 concerns two predetermined updates $d_\pm = c \pm b$ of any optimizer; the spectral map enters only through (c, b). Three consequences fix its scope: common-sample evaluation and the midpoint projection have the same leading variance when the bound below is small, but neither has universal mean-square-error dominance; one gradient can transfer to a family of pairs with an explicit remainder; and sign accuracy must be compared with a stratum-specific majority rule. Here $\|\cdot\|$ is the Euclidean norm on the full parameter vector.

Table 32: Controls sign accuracy with optimizer-specific majority baselines. Gains are percentage points (pp); pointwise intervals resample complete seeds and recompute the majority rule. The pooled row uses one majority sign per map.

Map	Majority (%)	Midpoint sign (%) [95% CI]	Gain (pp) [95% CI]
Reference	100.0	100.0 [100.0, 100.0]	0.0 [0.0, 0.0]
Quartic	72.2	95.8 [90.3, 100.0]	23.6 [9.7, 38.9]
Smoothed nuclear	100.0	100.0 [100.0, 100.0]	0.0 [0.0, 0.0]
Schatten-1.05 (Gram)	100.0	100.0 [100.0, 100.0]	0.0 [0.0, 0.0]
Pooled, map-specific rule	93.1	99.0 [97.6, 100.0]	5.9 [2.4, 9.7]

Table 33: Reusing one gradient within each map. Two pairs means amplitudes .025 and .1; three includes .4. Three-pair costs reconstruct recorded totals; two-pair direct and per-pair-midpoint costs scale the aggregate recorded call cost, since individual amplitude timings were not saved. All costs are averaged across maps and checkpoints, with the recording limitations stated in the text.

Pairs per map	Predictor	NRMSE [95% CI]	Seconds/family	Seconds/pair
2	Per-pair midpoint	0.038 [0.027, 0.061]	1.12	0.56
2	Shared	0.038 [0.024, 0.066]	0.56	0.28
2	Direct (512)	0.036 [0.026, 0.057]	0.72	0.36
2	Direct (128)	0.064 [0.046, 0.103]	0.19	0.10
3	Per-pair midpoint	0.110 [0.075, 0.172]	1.68	0.56
3	Shared	0.332 [0.290, 0.440]	0.56	0.19
3	Direct (512)	0.042 [0.031, 0.063]	1.08	0.36
3	Direct (128)	0.073 [0.051, 0.117]	0.29	0.10

Proposition 74 (Common-sample paired evaluation equals the midpoint projection to cubic order)**.** *Let $f(\cdot; Z)$ have an L_Z-Lipschitz Hessian on $\{W + c + tb : |t| \le 1\}$. Then*

$$f(W+c+b;Z) - f(W+c-b;Z) = 2\langle \nabla f(W+c;Z), b\rangle + R_Z, \qquad |R_Z| \le \tfrac{1}{3} L_Z \|b\|^3, \tag{184}$$

and the constant $1/3$ is attained. Condition on (W, c, b) independently of the sampled data. For samples $Z_1, \dots, Z_n$ with $L_{Z_i} \le L$, let $\widehat{D}_n = \frac{1}{n}\sum_i [f(W+c+b;Z_i) - f(W+c-b;Z_i)]$ and $\widehat{M}_n = 2\langle \frac{1}{n}\sum_i \nabla f(W+c;Z_i), b\rangle$. (i) For every realization, $|\widehat{D}_n - \widehat{M}_n| \le L\|b\|^3/3$. (ii) If the Z_i are independent with a common law, $L_Z \le L$ almost surely, and $b \ne 0$ with finite $v_1 = \mathrm{Var}\langle \nabla f(W+c;Z), b/\|b\|\rangle$, then $\mathrm{sd}(\widehat{M}_n) = 2\|b\|\sqrt{v_1/n}$,

$$\big|\mathrm{sd}(\widehat{D}_n) - \mathrm{sd}(\widehat{M}_n)\big| \le \frac{L\|b\|^3}{3\sqrt{n}}, \qquad \Big|\frac{\mathrm{sd}(\widehat{D}_n)}{\mathrm{sd}(\widehat{M}_n)} - 1\Big| \le \varepsilon := \frac{L\|b\|^2}{6\sqrt{v_1}}, \tag{185}$$

where the relative bound requires $v_1 > 0$. The absolute bound also holds when $v_1 = 0$, and $|\operatorname{Var}\widehat{D}_n - \operatorname{Var}\widehat{M}_n| \le [2\,\mathrm{sd}(\widehat{M}_n) + L\|b\|^3/(3\sqrt{n})]\, L\|b\|^3/(3\sqrt{n}) = O(\|b\|^4/n)$. (iii) With $F = \mathbb{E} f(\cdot; Z)$, integrable losses, and differentiation permitted under the expectation, let $D = F(W+c+b) - F(W+c-b)$. Then $\widehat{D}_n$ is unbiased for D, whereas $\mathbb{E}\widehat{M}_n - D = -\mathbb{E}R_Z$, at most $L\|b\|^3/3$ in absolute value. (iv) If the two arms instead use independent samples of size n and their losses have finite variances, the direct estimator $\widehat{D}_n^{\mathrm{ind}}$ satisfies

$$\operatorname{Var}\widehat{D}_n^{\mathrm{ind}} - \operatorname{Var}\widehat{D}_n = \frac{2}{n}\operatorname{Cov}\big(f(W+c+b;Z), f(W+c-b;Z)\big).$$

For $v_1 > 0$, its variance divided by $\operatorname{Var}\widehat{M}_n$ is $[\operatorname{Var} f(W+c+b;Z) + \operatorname{Var} f(W+c-b;Z)]/(4\|b\|^2 v_1)$; this ratio need not exceed one.

Proof. With $\varphi(t) = \langle \nabla f(W+c+tb;Z), b\rangle$, the left side of (184) is $\int_{-1}^{1}\varphi$. Subtracting $\varphi(0)$ and the linear Hessian term, which integrates to zero, leaves the gradient Taylor remainder, of size at most $L_Z t^2\|b\|^2/2$; integrating gives $1/3$. For attainment take $f(w) = L\langle u, w\rangle^3/6$ with $\|u\| = 1$, whose Hessian $L\langle u, w\rangle\, uu^\top$ is exactly L-Lipschitz, and $b \parallel u$. Averaging gives (i) with $\widehat{D}_n = \widehat{M}_n + \bar{R}_n$.

Table 34: Response geometry in the controls study: medians over pairs of $\|b\|$, $\|c_\rho - c_0\|/\|b\|$, and the cosine between response and pulse direction. For a smooth map with nonzero first-order response, the ratio is $O(\rho)$ locally.

Map	ρ	Median $\|b\|$	Median $\|c_\rho - c_0\|/\|b\|$	Median cosine
Reference	0.025	0.021	0.031	0.39
Reference	0.1	0.081	0.121	0.39
Reference	0.4	0.291	0.399	0.42
Quartic	0.025	0.025	0.013	0.99
Quartic	0.1	0.098	0.051	0.99
Quartic	0.4	0.365	0.197	0.99
Smoothed nuclear	0.025	0.021	0.023	0.45
Smoothed nuclear	0.1	0.082	0.090	0.45
Smoothed nuclear	0.4	0.300	0.309	0.46
Schatten-1.05 (Gram)	0.025	0.021	0.181	0.41
Schatten-1.05 (Gram)	0.1	0.080	0.185	0.45
Schatten-1.05 (Gram)	0.4	0.286	0.340	0.48

In (ii), $\bar{R}_n$ averages n independent variables bounded by $L\|b\|^3/3$, so $\mathrm{sd}(\bar{R}_n) \le L\|b\|^3/(3\sqrt{n})$; the standard deviation is a seminorm, and $\mathrm{Var}(X+Y) - \mathrm{Var}\,X = \mathrm{Var}\,Y + 2\,\mathrm{Cov}(X,Y)$ with $|\,\mathrm{Cov}(X,Y)| \le \mathrm{sd}\,X\,\mathrm{sd}\,Y$. When $v_1 > 0$, dividing by $\mathrm{sd}(\widehat{M}_n)$ gives ε. Expectations give (iii); independence of the arms gives (iv). □

Thus direct common-sample evaluation has no Taylor bias, whereas the midpoint projection may have a cubic bias; their standard errors agree to relative order ε only when $v_1 > 0$ and ε is small. The sign of $\mathrm{Cov}(\widehat{M}_n, \bar{R}_n)$ is unrestricted, so the proposition does not establish a universal variance or mean-square-error ordering. Common samples reduce variance relative to independent arms exactly when their loss covariance is positive. For example, $f(w;Z) = Zw$, $Z \in \{-1,1\}$ equiprobable, at center zero gives $\widehat{D}_n = \widehat{M}_n$ but independent-arm variance only one half as large. At a nondegenerate center with positive loss variance and fixed positive v_1, the usual independent-arm variance ratio instead grows as $\|b\|^{-2}$ for small b. These distinctions are checked explicitly in the accompanying numerical script.

In this appendix's protocol the outcome is a common-sample paired evaluation on 768 validation sequences, while prediction uses 512 sequences from a disjoint corpus range. The unbiasedness statement applies to the distribution sampled by an estimator: direct prediction-pool evaluation is unbiased for that pool's mean contrast, not automatically for a different validation-pool population. Error against the archived outcome includes sampling variation and differences between the corpus ranges. The difference between direct and midpoint estimates on the same prediction sequences isolates the sample-mean Taylor remainder in exact arithmetic; actual float32 evaluations also introduce numerical error. It does not isolate sampling variance against the validation target. Which method has lower validation error is therefore measured, rather than inferred from this identity. Numerically, on the cubic per-sample loss $f(w;Z) = \frac{1}{2}w^\top Aw + \langle \beta_Z, w\rangle + \gamma_Z\langle u_Z, w\rangle^3/6$ in $\mathbb{R}^{50}$ with $\|u_Z\| = 1$ and $|\gamma_Z| \le L = 2$, whose per-sample Hessian $A + \gamma_Z\langle u_Z, w\rangle u_Z u_Z^\top$ is exactly $|\gamma_Z|$-Lipschitz, (184) holds to 2×10^{-15} with loss values up to 3.6, the bound is attained to ratio 1.000000 for b along the cubic direction, and at $\|b\| = .02$ and .32 with $\|c\| = 1$ the observed $|\,\mathrm{sd}(\widehat{D}_n)/\mathrm{sd}(\widehat{M}_n) - 1|$ is 8.6×10^{-9} and 2.3×10^{-6} against $\varepsilon = 2.7 \times 10^{-4}$ and 0.068, with independent-arm inflation 2240 and 9.2. On a random two-layer tanh network in $\mathbb{R}^{50}$, with no certified constant, the root mean square remainder has log-log slope 2.976 to 3.000 over $\|b\| \in [.01, .32]$ and the inflation factor is 77 at $\|b\| = .32$ and 77273 at .01. These are observations on synthetic losses, not neural constants.

Proposition 75 (One gradient, many interventions)**.** *Let (c_j, b_j), $1 \le j \le m$, be predetermined pairs, c_0 a reference center with $\|c_j - c_0\| \le \Delta$, and $D_j = F(W + c_j + b_j) - F(W + c_j - b_j)$. If F has an L_H-Lipschitz Hessian on the convex hull of $W + c_0$ and the points $W + c_j \pm b_j$, then with $g_0 = \nabla F(W + c_0)$ and $H_0 = \nabla^2 F(W + c_0)$, for every j,*

$$|D_j - 2\langle g_0, b_j\rangle| \le 2\|H_0\|_{\mathrm{op}}\Delta\|b_j\| + L_H\Delta^2\|b_j\| + \tfrac{1}{3}L_H\|b_j\|^3, \tag{186}$$

$$|D_j - 2\langle g_0 + H_0(c_j - c_0), b_j\rangle| \le L_H\Delta^2\|b_j\| + \tfrac{1}{3}L_H\|b_j\|^3. \tag{187}$$

Proof. Proposition 73 at the center c_j leaves the cubic term; $\nabla F(W + c_j) = g_0 + H_0(c_j - c_0) + r_j$ with $\|r_j\| \le L_H\|c_j - c_0\|^2/2$, and Cauchy–Schwarz bounds the two inner products. □

The correction in (187) needs one Hessian-vector product per pair, which costs about as much as a gradient, so it is not a saving. Numerically, let F be the mean of the cubic per-sample loss above over a fixed mixture of 4096 samples. Its Hessian $A + T[w]$ is affine in w, with $T[\delta] = \frac{1}{n}\sum_i \gamma_i\langle u_i, \delta\rangle u_i u_i^\top$, so $\|H_0\|_{\mathrm{op}} = 2.000$ is exact and the Hessian-Lipschitz constant is $L_H = \sup_{\|\delta\|=1}\|T[\delta]\|_{\mathrm{op}}$. This is not the per-sample constant $\frac{1}{n}\sum_i|\gamma_i| = 1.004$, which is valid but overstates L_H by a factor 280 because the γ_i are symmetric about zero and average out in F. We use the certified value $L_H = 3.59\times10^{-3}$, the spectral norm of the $d^2\times d$ unfolding of the tensor $\frac{1}{n}\sum_i \gamma_i u_i^{\otimes 3}$, which bounds $\|T[\delta]\|_F \ge \|T[\delta]\|_{\mathrm{op}}$ for unit δ; an ascent over unit δ reaches $\|T[\delta]\|_{\mathrm{op}} = 1.84\times10^{-3}$, so the certified value is within a factor 1.95 of the true constant. Over 360 random cells ($\Delta \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, $\|b_j\| \in \{.02, .08, .32\}$, forty random directions each) the error never exceeds 0.218 of the right side of (186) and 0.170 of that of (187). With $c_j - c_0$ and b_j along the top eigenvector of H_0 the first ratio reaches 1.000, so the first-order term is attained; with both along the ascent direction δ the second ratio is 0.513 in every cell, the ratio of the reached to the certified constant, since the corrected error is then exactly $(\Delta^2\|b_j\| + \|b_j\|^3/3)\langle\delta, T[\delta]\delta\rangle$ and δ is the top eigenvector of $T[\delta]$ at the ascent's fixed point. The random-direction ratios are small because Cauchy–Schwarz is loose for generic directions in $\mathbb{R}^{50}$, not because the constants are.

Proposition 76 (Shared center of the pulse family)**.** *Fix the checkpoint, the map Ψ and unit directions E_i, and let ρ vary. Let c_0 be the ordinary update ($\rho = 0$), so $W + c_0$ is the ordinary next iterate. With $h_i = \rho\|P_i\|_F$,*

$$c_\rho - c_0 = -\frac{\eta}{2}\sum_i a_i\big[\Psi(P_i + h_iE_i) + \Psi(P_i - h_iE_i) - 2\Psi(P_i)\big] \tag{188}$$

on the matrix blocks and zero elsewhere, since the auxiliary update is common to all arms. If Ψ is zero-homogeneous and twice continuously differentiable away from zero, let κ_1 and κ_2 be the suprema over the unit Frobenius sphere of the operator norms of $D\Psi$ and $D^2\Psi$. Then for $0 < \rho < 1$, uniformly in the E_i,

$$\|c_\rho - c_0\| \le \frac{\rho^2\eta\kappa_2}{2(1-\rho)^2}\Big(\sum_i a_i^2\Big)^{1/2}, \qquad \|b_\rho\| \le \frac{\rho\,\eta\kappa_1}{1-\rho}\Big(\sum_i a_i^2\Big)^{1/2}. \tag{189}$$

Hence the right side of (186) *with $(c_j, b_j) = (c_\rho, b_\rho)$ and $\Delta = \|c_\rho - c_0\|$ is $O(\rho^3)$: one gradient at the ordinary next iterate predicts every amplitude and direction to the same order in ρ as the per-pair midpoint gradient.*

Proof. The identity is the definition of $c_\rho = (d_+ + d_-)/2$. Zero-homogeneity gives $D^k\Psi(Q) = D^k\Psi(Q/\|Q\|_F)/\|Q\|_F^k$, and $\|P_i + th_iE_i\|_F \ge (1-\rho)\|P_i\|_F$ for $|t| \le 1$. The second-order Taylor remainder bounds each bracket in (188) by $\rho^2\kappa_2/(1-\rho)^2$, the mean value inequality bounds $\|\Psi(P_i + h_iE_i) - \Psi(P_i - h_iE_i)\|_F$ by $2\rho\kappa_1/(1-\rho)$, and the blocks add in quadrature. The κ_k are finite by continuity on the compact sphere. □

The hypothesis holds for the polynomial maps (1), which are smooth away from the origin, and for the smoothed nuclear map by Theorem 58; it fails for the Schatten-ς map, $1 < \varsigma < 2$, at rank-deficient inputs. Two limits matter in practice. Across maps the ordinary updates differ at order η, so a gradient at one map's next iterate predicts another map's pairs only with a first-order error $2\|H_0\|_{\mathrm{op}}\|c_0^{\mathrm{map}} - c_0\|\|b\| = O(\rho)$; one gradient per map is needed. And (189) is asymptotic with a constant κ_2 that can be large: the depth-five quintic has $f_5'(0) = a_0^5 \approx 485$ and varies rapidly on singular values below the smallest normalized one. On a synthetic 8×6 input with smallest normalized singular value 0.0092, the reference map's centre shift $\|c_\rho - c_0\|$ is 0.78 times $\|b_\rho\|$ at $\rho = .4$ and 0.167 times at $\rho = .025$, and its log-log slope stays within .05 of 2 only below $\rho = 3.1\times10^{-3}$; with smallest normalized singular value 0.166 the ratios are 0.635 and 0.0773 and the onset is $\rho = 0.10$. The smoothed nuclear map at $\kappa = .001$ behaves like the reference (0.70 and 0.220), the quartic map, a cubic polynomial in X, gives 0.199 and 0.0132. Whether the first-order

term of (186) is small at the neural amplitudes is therefore empirical; the controls study records $\|c_\rho - c_0\|$ for every pair and the shared predictor's error directly.

Cost. Counting a backward pass as two forwards, a gradient over the n prediction sequences costs three forward units and direct common-sample evaluation of one pair costs two, one forward per arm. The map evaluations producing (c_j, b_j) are common to every row of Table 35. Per-pair midpoint gradients are never cheaper than direct evaluation; a shared gradient is cheaper from $m = 2$ candidates on, since $3 \le 2m$. For the map being trained, the run itself computes a gradient at the ordinary next iterate on the next minibatch; reusing it requires no additional gradient evaluation but averages 32 training sequences instead of 512 prediction sequences. Only under the same sampling law, independence assumptions and per-sequence directional variance would this multiply the variance by 16; training and prediction pools need not satisfy that comparison. The archived study uses the separate 512-sequence gradient, not this proposed reuse.

Table 35: Cost of predicting m predetermined pairs with a common center, in forward passes over the n prediction sequences, a gradient counting as three. The bias column is the deterministic remainder of Propositions 73 and 75; for a common center, the midpoint and direct rows have standard error $2\|b_j\|\sqrt{v_{1,j}/n}$ up to the factor in (185); a shared gradient at a different center has its own directional variance. The last column is this protocol's three amplitudes per map.

Predictor	Deterministic bias for pair j	Cost	$m = 3$
Per-pair midpoint gradient	$\le L_H\|b_j\|^3/3$	$3m$	9
Shared gradient at $W + c_0$	right side of (186)	3	3
Direct common-sample paired evaluation	none	$2m$	6

Sign accuracy on one-signed strata. Where every outcome in a stratum has the same sign, the majority-sign rule is 100% accurate and so is any predictor of the right sign. In the confirmation study (seeds 200–207), all 144 reference and smoothed-nuclear pairs have negative outcomes, so their 100% sign accuracy says nothing about the predictor. On the quartic stratum 51/72 pairs are harmful; the majority rule attains 70.8% and the midpoint predictor 94.4%. This demonstrates within-family discrimination, without identifying a recovery-specific mechanism. Pooled over 216 pairs a single majority sign attains 76.4% against 98.1% for the midpoint predictor and 76.4% for the frozen local response, which predicts a negative effect for every pair and therefore carries exactly the majority rule's sign information. On one-signed strata the informative quantities are the NRMSE and the within-stratum rank correlation. Table 36 recomputes the fresh-seed strata from the archived rows with the majority baseline alongside; its rank column is the Spearman correlation across the 24 pairs (eight seeds, three checkpoints) at fixed map and amplitude, averaged over amplitudes, so that amplitude ordering cannot inflate it. The revised tables report the majority-sign baseline next to every sign accuracy.

Table 36: Fresh-seed strata with the majority-sign baseline. "Harmful +" counts pairs where adding the signal raises the next loss; rank is the within-stratum Spearman correlation at fixed amplitude, averaged over amplitudes. Recomputed from the archived rows of the confirmation study; no new outcomes.

Map	Harmful +	Majority sign	Midpoint sign	Midpoint NRMSE	Rank at fixed ρ
Reference	0/72	100.0%	100.0%	0.035	0.990
Quartic	51/72	70.8%	94.4%	0.115	0.992
Smoothed nuclear	0/72	100.0%	100.0%	0.036	0.987

What recovery geometry contributes. The predictor $2\langle \widehat{g}(W + c), b\rangle$ is a product of two factors. The response $b = (d_+ - d_-)/2$ is computed from the map, the momentum state and the specified pulse directions, with no additional samples once these are fixed: $2b_i = -\eta a_i[\Psi(P_i + h_i E_i) - \Psi(P_i - h_i E_i)]$ on matrix blocks and zero on the auxiliary parameters. The gradient at the common point $W + c$ needs data and carries the loss geometry after the ordinary step. The original-state local-response predictor uses the gradient at W and the infinitesimal approximation to b; it omits transport over the common step and has NRMSE 1.79 on the fresh seeds; the bilinear curvature term of (181) recovers part of it (0.353) and the midpoint gradient recovers it to 0.110 (Table 27). The common step has norm 1.00

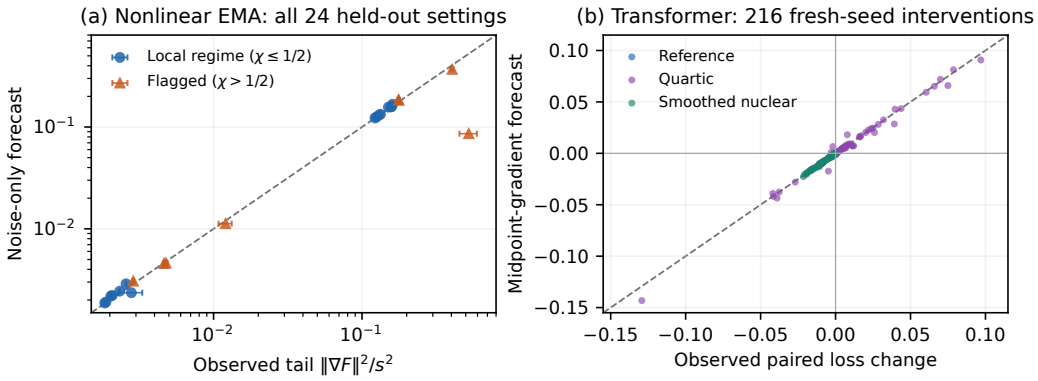


Figure 9: Auxiliary prospective quantitative predictions. (a) Noise-only forecasts versus independent nonlinear EMA measurements at all 24 held-out settings; orange markers retain cases flagged outside the local regime. Bars are trajectory-based 95% intervals. (b) Midpoint-gradient forecasts of all 216 paired loss changes on eight fresh transformer seeds. Positive changes mean the added signal harms the next loss. Dashed lines indicate perfect prediction.

to 1.05 against 0.022 to 0.32 for the response, so what the frozen predictor misses is the gradient change over the common step, not the pulse. This does not identify a recovery-specific source of predictability: the Taylor formula for any paired updates requires their actual displacement direction. The controls study therefore keeps the actual midpoint gradient and replaces b by two proxies, the pulse direction $-\|b_i\|_F E_i$ with matched per-layer norm and the normalized-descent response with matched pilot gain. If these proxies fall short, the actual displacement direction matters for this Taylor prediction. That is a generic requirement of directional loss prediction and does not show that adverse recovery geometry adds information beyond a matched general local-loss model. Moreover, only the pulse-direction ablation matches the actual per-layer response norm: the normalized-descent ablation matches an ordinary-update pilot gain, so it changes response magnitudes as well as directions. The controls and their limitations are reported on eight new seeds, with retrospective stratified uncertainty and timing analysis in Tables 30–33.

W Matched training and the magnitude of spectral flattening

Synthetic spectra and independent mathematical checks. We check the analytic gradient of the smoothed nuclear norm against differentiation of the potential, compute full dense Jacobians for square and rectangular matrices, and check both angular-curvature bounds, spectral replication and finite directional second differences at repeated and zero singular values. Twenty nondegenerate matrix/parameter combinations and fifteen degenerate-spectrum directions are checked. These numerical checks supplement the proof of Theorem 58; they are not the proof of convexity. The Gram implementation of the Schatten maps (Proposition 62) is checked against the SVD formula on spectra with condition numbers up to 10^5, where the maximum relative difference is 2.0×10^{-7}; at condition 10^3 the two agree to 10^{-11}. The Gram route squares the condition number in the small eigenvalues, which is the source of this difference (Appendix O).

Table 37 evaluates every map on the same diagonal matrix whose singular values are logarithmically spaced between 10^{-1} and 10, then normalized to unit Frobenius norm. The reported ratio is the 90th percentile divided by the 10th percentile of the output singular values. The reference has depth five, the quartic has $\lambda = 1/8$, and the smoothed nuclear map has $\kappa = .001$. Each ratio of two positive singular values under Schatten-ς is exactly the input ratio to the power $\varsigma - 1$. The table's interpolated percentile ratios are computed numerically; interpolation does not in general commute with this power. At $r = 4$ the reference flattens more than the smoothed nuclear map, whereas at $r \ge 16$ the ordering reverses. The construction guarantees rank-independent retention of flattening, not dominance on every spectrum.

Table 37: Magnitude of flattening on common synthetic inputs. Smaller spectral ratios mean flatter updates. No learning-rate or output-norm adjustment changes these ratios.

r	Input	Reference	Quartic	Smoothed nuclear	Schatten-1.05	Schatten-1.25
4	36.54	1.12	24.62	1.56	1.20	2.51
16	39.81	1.49	36.23	1.26	1.20	2.51
64	39.81	1.63	38.76	1.22	1.20	2.51
256	39.81	1.62	39.53	1.21	1.20	2.51
512	39.81	1.65	39.67	1.21	1.20	2.51

Training protocol and model selection. Every arm trains the same 10.9M-parameter byte-level transformer: width 384, six layers, six heads, context 256, batch size 32, 2500 updates, Nesterov momentum .95, and zero additive normalization epsilon. The model uses float32 and every matrix map is evaluated in float64. Auxiliary parameters use AdamW with learning rate .0003 and no weight decay. TF32 is disabled. WikiText-103 bytes are partitioned into training $[0, .8N)$, tuning $[.8N, .9N)$, and test $[.9N, N)$. Tuning uses seeds 100 and 101; the final comparison uses eight new seeds 200–207. Every method has the same learning-rate candidates $\{.002, .004, .008, .016, .032\}$; the selected rate minimizes the mean final tuning loss, breaking ties toward the smaller value, and the test uses this selection with no further search. Tuning runs never evaluate the test loss. The same seed fixes identical initial parameters, training positions and evaluation positions across all arms; each evaluation averages 32 minibatches. At the first common gradient, each layer receives the positive fixed multiplier $\|\Psi_{\mathrm{ref}}(P_i)\|_F / \|\Psi(P_i)\|_F$, constant for all later updates, so the comparison does not re-normalize the output at each step, which would change the convex map.

The first study compares the reference depth-five map, the quartic map with $\lambda = 1/8$ and the smoothed nuclear map with $\kappa = .001$; the last value was fixed by the synthetic spectrum criterion before any training losses. Its initial four-value grid started at .004; the quartic's tuning optimum was at that boundary, so .002 was added to all three methods before any test outcomes were inspected. A three-step smoke run of an SVD-based Schatten-1.05 map took 20.48 seconds against 1.62 for the reference, and the SVD arm was excluded from that grid. Proposition 62 shows the exclusion was an implementation artefact: the same map is the Loewner function of the smaller Gram matrix applied to the input, exactly as the smoothed nuclear map is computed. The second study therefore adds, on the same seeds, batches and grid, Gram-implemented Schatten-1.05 (selected rate .008) and Schatten-1.25 (.004), the smoothed nuclear map at $\kappa = .01$ (.004), a rerun of the reference at its selected rate .008 on different devices of the same type, and an equal-wall-clock reference trained for 2850 updates, the first study's time ratio of 1.14 applied to 2500. All tuning and test runs use PyTorch 2.7.1 with CUDA 12.8 on H20 GPUs.

Margin and the direction of the difference. The archived margin protocol uses $\pm.005$ nats, about twice the original reference seed standard deviation (.0027). It was recorded before the added v4 outcomes, after the original smoothed-map outcome was known. It is consequently not an independently justified equivalence margin for that original comparison. Tables report where the paired 95% interval lies relative to this margin; the historical verdict labels remain in the archived protocol. An interval inside the margin can still lie strictly above zero, as for the wider smoothed-map arm. An interval crossing the negative margin can still lie strictly below zero, as for the wider Schatten arm. Neither label overrides the direction or uncertainty of the measured loss difference. Intervals are nominal rather than simultaneous.

Outcomes, spectral measurements and uncertainty. The primary endpoint is final test loss at equal training-sample budget. Differences are paired within the eight complete seeds against the first study's reference losses, and their 95% intervals use a Student-t distribution with seven degrees of freedom. Layers are not independent replicates. At the last update of the rerun reference, every map's spectrum is evaluated at the same pre-update momentum matrices; we take the geometric mean of the layer-wise 90th/10th-percentile ratios within each seed, then average the eight statistics, so the spectral comparison isolates the maps rather than their training histories. Training time includes data access, forward and backward passes, map computation and optimizer updates, with synchronization at common timing boundaries, and excludes held-out evaluations and spectrum measurements. The

Table 38: Equal-sample training on eight held-out seeds, plus the explicitly longer reference. Loss differences are paired against the first study's reference with 95% intervals. Spectral ratio is the 90th/10th-percentile singular-value ratio on common reference inputs; smaller means flatter. Time is training wall time relative to the reference, with float64 maps.

Map	Test loss	Difference vs. reference	Ratio	Time	Margin location
Reference	1.2559	—	2.70	1.00×	—
Quartic	1.3619	$+0.1059$ $[+0.0980, +0.1138]$	21.38	0.38×	above $+.005$
Smoothed nuclear ($\kappa = .001$)	1.2588	$+0.0029$ $[+0.0004, +0.0053]$	1.85	1.14×	crosses margin
Reference (rerun)	1.2568	$+0.0009$ $[-0.0007, +0.0025]$	2.70	1.00×	within $\pm.005$
Schatten-1.05 (Gram)	1.2547	-0.0013 $[-0.0040, +0.0014]$	1.17	1.15×	within $\pm.005$
Schatten-1.25 (Gram)	1.2736	$+0.0176$ $[+0.0127, +0.0226]$	2.15	1.15×	above $+.005$
Smoothed nuclear ($\kappa = .01$)	1.2675	$+0.0116$ $[+0.0042, +0.0190]$	3.01	1.14×	crosses margin
Reference, equal wall clock (2850 updates)	1.2423	-0.0136 $[-0.0150, -0.0121]$	—	1.14×	below $-.005$

archived quartic implementation forms the larger Gram matrix for tall layers, so its timing is an implementation measurement rather than an optimized quartic cost bound.

Table 38 gives the complete comparison from both studies. The rerun reference reproduces the original to $+0.0009$ $[-0.0007, +0.0025]$ in the mean, with a largest per-seed absolute difference of .0027 under identical configuration; this is a measured reproducibility variation, not a threshold below which every effect is meaningless. The paired mean retains its own uncertainty. Gram Schatten-1.05 trains to 1.2547, a paired difference of -0.0013 $[-0.0040, +0.0014]$, inside the margin, at time ratio $1.15\times$ and output spectral ratio 1.17, the flattest of all arms. Its paired difference to the smoothed nuclear map at $\kappa = .001$ is $-.0041$ $[-.0069, -.0014]$: at equal cost the Schatten map is nominally better, and the study does not separate the two maps in favour of the smoothed nuclear norm. Schatten-1.25 ($+0.0176$ $[+0.0127, +0.0226]$, ratio 2.15) is worse, and $\kappa = .01$ ($+0.0116$ $[+0.0042, +0.0190]$, ratio 3.01) is inconclusive because its lower limit falls just inside the margin; its point estimate is the second largest. The equal-wall-clock reference reaches 1.2423, a difference of -0.0136 $[-0.0150, -0.0121]$: the 14% of time that float64 Gram arithmetic costs buys the reference more loss reduction than any flattening arm's difference. Ordered by spectral ratio (1.17, 1.85, 2.15, 2.70, 3.01) the paired differences are $-.0013$, $+.0029$, $+.0176$, 0 and $+.0116$, so loss is not monotone in flattening on this task, and the direct loss contrast favours Schatten-1.05. The smoothed map instead has the bounded gain, Lipschitz constant and eigenvalue conditioning of Appendix O. These values quantify one model and corpus at equal sample budget with float64 maps; they are not evidence of universal optimizer superiority, of equivalence at equal wall clock, or of low-precision implementation speed. **Second model scale.** The same protocol at width 512, eight layers and eight heads, with the learning-rate grid, tuning seeds, eight held-out seeds, sample budget, selection rule and margin unchanged (selected rates .004, .004 and .008 for the reference, the smoothed nuclear map and Schatten-1.05), gives Table 39. The reference reaches 1.2255. The smoothed nuclear map at $\kappa = .001$ differs by $+0.0011$ $[+0.0004, +0.0018]$ is above zero and within the declared margin, at output spectral ratio 1.88. Gram Schatten-1.05 differs by -0.0038 $[-0.0055, -0.0022]$ at ratio 1.17: the interval lies below zero, but its lower limit crosses -0.005. The historical margin rule returned inconclusive, but the nominal interval still supports a lower loss than the reference. At this width both flattening arms ran at $0.73\times$ the reference's wall clock, the reverse of the ordering at width 384, an observed dependence on matrix shape and implementation; no scaling law is established by these two measurements; the time column is implementation- and hardware-specific, and no equal-wall-clock claim is made at this scale. The second scale supports feasibility at another scale of the same model and corpus. In both scales, the direct paired loss contrast favours Schatten-1.05 over smoothed nuclear; neither regularity of the exact map nor a chosen tolerance reverses this measured ordering.

Direct paired contrasts and multiplicity sensitivity. Table 40 makes the relevant comparisons explicit rather than inferring them from separate comparisons to the reference. This is a retrospective reanalysis of the same eight seeds. The two Schatten-minus-smoothed contrasts have Holm-adjusted two-sided p values .0093 and .0003 within that two-comparison family. Both differences are negative on all eight seeds; exact two-sided sign tests give .0078125 at each scale, or .015625 after Holm adjustment. This robustness check does not assume Gaussian seed differences. Other rows are descriptive nominal paired intervals. The scales share corpus and seeds and are not independent datasets. At the base scale the 2850-step reference uses approximately the smoothed map's measured

Table 39: Second model scale (width 512, eight layers): paired final held-out byte loss against the reference on the eight held-out seeds, output spectral ratio, wall-clock ratio and interval location relative to the declared margin ± 0.005.

Map	Test loss	Difference vs. reference	Ratio	Time	Margin location
Reference	1.2255	—	2.82	1.00×	—
Smoothed nuclear ($\kappa = .001$)	1.2266	$+0.0011$ $[+0.0004, +0.0018]$	1.88	0.73×	within $\pm.005$
Schatten-1.05 (Gram)	1.2216	-0.0038 $[-0.0055, -0.0022]$	1.17	0.73×	crosses margin

time: its paired time ratio to the smoothed arm is 1.001. There is no corresponding wider-model equal-time experiment.

Table 40: Retrospective direct paired contrasts. Negative values favour the first named method. The interval for each row is a nominal paired 95% interval across eight seeds; Holm adjustment applies only to the two Schatten-minus-smoothed tests; p_t uses the paired t test and p_{sign} the exact sign test.

Scale	Paired contrast	Loss difference [95% CI]	Holm p_t	Holm p_{sign}
10.9M	Schatten minus smoothed	-0.0041 $[-0.0069, -0.0014]$	0.0093	0.0156
25.6M	Schatten minus smoothed	-0.0050 $[-0.0066, -0.0034]$	0.0003	0.0156
10.9M	Smoothed minus reference	$+0.0029$ $[+0.0004, +0.0053]$	—	—
25.6M	Smoothed minus reference	$+0.0011$ $[+0.0004, +0.0018]$	—	—
10.9M	Schatten minus reference rerun	-0.0022 $[-0.0046, +0.0002]$	—	—
10.9M	2850-step reference minus Schatten	-0.0123 $[-0.0146, -0.0100]$	—	—
10.9M	2850-step reference minus smoothed	-0.0165 $[-0.0179, -0.0150]$	—	—

Prospective cutoff sensitivity under matched tuning. The archived Gram implementation uses a relative singular cutoff 10^{-8} and recomputes its Schatten normalizer after removing modes; the SVD implementation uses 10^{-12}. Thus these are distinct numerical maps on sufficiently ill-conditioned inputs. Appendix O separates cutoff error from same-cutoff arithmetic error, and proves that a hard relative cutoff need not preserve exact-map monotonicity.

We therefore fixed two additional cutoffs, 10^{-6} and 10^{-10}, before their training outcomes. Both receive the same five learning rates and two tuning seeds as before, followed by seeds 200–207 at 2500 updates; a contemporaneous 10^{-8} arm repeats its previously selected rate .008. This adds 20 tuning and 24 test runs. Both new grids select .008. The lower cutoff is a roundoff stress test below $\sqrt{u}$, not a claim of greater accuracy. Table 41 shows small point differences with intervals spanning zero; there is no prespecified cutoff-equivalence margin. These results describe final-loss sensitivity in this model, not fidelity of every update to the exact Schatten gradient.

The nominal-cutoff rerun differs from its archived counterpart by $-.000052$ $[-.000891, +.000788]$. All 44 training records pass configuration, source-hash, seed, stream and completion checks. A transient dispatcher failure label for one completed file is retained in the completion audit: its shared-filesystem listing became visible after the completed log. No failed training record is dropped. The four extra per-run cutoff diagnostics use an eigenvalues-only routine on reconstructed map inputs. Their retention counts are diagnostic estimates rather than recorded masks from the update's eigendecomposition. Their timing exclusion can absorb pending GPU work; consequently the cutoff table makes no runtime comparison. The independent map microbenchmark and original training timing remain the timing evidence.

Table 41: New cutoff sensitivity study: eight paired test seeds after identical learning-rate searches for the two added cutoffs. Differences use the concurrent 10^{-8} rerun. Holm correction covers the two cutoff contrasts.

Cutoff	Selected LR	Test loss	Difference vs. 10^{-8} [95% CI]	Holm p
10^{-6}	0.008	1.2545	-0.0001 $[-0.0010, +0.0008]$	1.000
10^{-8}	0.008	1.2546	—	—
10^{-10}	0.008	1.2545	-0.0001 $[-0.0011, +0.0008]$	1.000

X EXECUTABLE PROOF ARTIFACTS

The trajectory proofs require inequalities to hold throughout ranges of possible states. The supplementary proof archive divides these ranges into small boxes and bounds the response on every box using rational arithmetic. An independent checker verifies both the bounds and that the boxes cover the whole required region. The archive also includes the reference-polynomial checks, their outputs and file hashes. These checks need only the Python standard library; no training run or GPU is required.

Certificate	Accepted boxes	Certified entry bound	Horizon
Normalization only, $\beta = .1$	43	295417	10^6
Increasing and flattening, $\beta = .3$	217	666776	10^7

The archive's usage guide supplies the execution instructions and expected outputs. The trajectory checker verifies all binary subdivision paths, so a missing region or an overlapping leaf cannot pass merely by having a correct bound on the remaining boxes. It independently recomputes the rational drift enclosures, checks the scalar matrix reduction and its derivative polynomials, verifies outer signs using Sturm sequences, and recomputes the perturbation, lag, excursion and entry budgets. Square-root endpoints are verified by squaring rationals; logarithms use an atanh-series tail enclosure. It recovers the exact squared-gradient bands $[153/819200, 4913/1179648]$ and $[4913/61440000, 1230137/245760000]$.

The ordered leaf digests coincide with the archived searches used in Appendices F and J.1. All leaf coordinates and upper bounds are stored as exact rational strings. The repository additionally exercises rejection of altered bounds, duplicate or missing leaves, and inconsistent subdivision coordinates. The reference-polynomial checker verifies the selected algebra of the tuned quintic, including the depth-one and depth-five local barriers. The accompanying guide maps each checker to its scope; digests establish file identity, not mathematical validity.

These programs check the computational premises of the written probabilistic arguments. They are not a formalization of the entire paper. In particular, no finite-window certificate is claimed for the high-momentum numerical constructions, and the finite-step nonlinear forecast has no certified remainder. Neural protocols, raw per-seed sufficient statistics, exact map-input fixtures and state-restoration hashes are supplied separately with the experiment code.