AFTER THE FACT: COUPLED UPDATES SHAPE THE IMPLICIT BIAS OF MUON

Anonymous authors
Paper under double-blind review

ABSTRACT

When Muon updates one matrix layer, its motion changes gradients seen by surrounding parameters, so learning depends on how those signals are transformed downstream. We study this interaction through covariance, starting from Qwen3-0.6B Muon trajectories where similar NS5 update magnitudes can hide very different local responses, including negative elasticities. Persistent motion then turns these responses into curvature gradients. For stable cycles, we characterize when coupled spectral updates admit scalar curvature objectives, identify power compatibility, and show why cube-root updates remain aligned under decay. Under noisy training, Gaussian averaging can restore stability even when instantaneous responses are negative, while finite covariance fluctuations can change whether a scalar objective exists. Together, these results connect spectral response, persistent matrix motion, and optimizer coupling within one theory of curvature dynamics.

1 INTRODUCTION

When we choose an optimizer for a matrix layer, we also choose how that layer will change the gradients received by the rest of the model. In a Transformer block, moving an attention projection changes the features passed forward and the errors passed backward, so the resulting training behavior depends on how several parameter groups respond to one another. With Muon, these interactions pass through a spectral transformation of momentum (Jordan et al., 2024), while the surrounding parameters can follow a different update rule (Liu et al., 2025).

We are interested in whether this coupled evolution follows a scalar curvature objective, which would describe the relative movement of the remaining parameters up to a positive change of speed. Persistent motion provides a good connection as excursions around a minimum produce gradients on parameters that control its curvature, as developed in edge of stability and central flow analyses (Arora et al., 2022; Cohen et al., 2024). For a matrix layer, these excursions carry a spatial covariance, and its distribution determines the signals presented to the receiving optimizer.

Our description follows the fast update into a statistical state before applying the slow update to the gradients produced within that state (Figure 1). This intermediate distribution matters even when the fast motion is stable, since a nonlinear spectral function can respond differently to a fluctuating covariance and to its average. We think it provides a useful way to connect optimizer choices across a model, with the first group determining which signals persist and the second determining how those signals become parameter changes.

We begin with actual Muon inputs from a Qwen3-0.6B training trajectory, where NS5 produces fairly similar update magnitudes even though its local responses vary sharply across singular directions and often become negative. This raises a natural question about whether such negative responses imply unstable motion. Gaussian averaging identity shows that they need not, because changes in singular orientation contribute alongside amplitude changes and together can produce a positive restoring response.In the corresponding small step limit, the fast matrix has an attracting Gaussian equilibrium, giving a second explicit setting alongside the stable oscillations of our deterministic model.

For centered oscillations, we classify the fast and slow spectral functions that admit a scalar objective, with pure powers preserving compatibility across an entire family of receiving updates and cube root updates retaining a special compatibility with decay. Under Gaussian equilibrium, we show that covariance fluctuations can destroy the same scalar representation through a normalized square

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

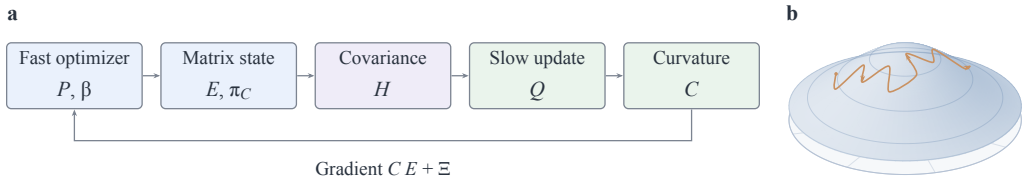


Figure 1: Through matrix motion and its covariance, the fast and slow updates form the feedback loop in (a), while (b) illustrates fluctuations of the fast matrix state.

update, while a normalized linear update admits an explicit potential. Both conclusions arise from the way a receiving optimizer transforms the covariance distribution, connecting microscopic spectral response to the geometry of coupled learning.

2 Preliminaries

2.1 Matrix Layers

For a linear map $z = Wh$, backpropagation gives an outer product $rh^\top$, where h is the input activation and r is the output loss derivative. Across a minibatch, the gradient is a sum of these products, so its singular directions depend on representations learned elsewhere in the network, including the projections used by attention and feedforward layers (Vaswani et al., 2017; Dao & Gu, 2024). We retain this structure by writing a matrix displacement $E \in \mathbb{R}^{k\times n}$ and a trainable curvature $C = C^\top \succ 0$, with loss and gradients

$$\ell_t(C, E) = \frac{1}{2}\operatorname{tr}(E^\top C E) + \langle \Xi_t, E\rangle_F, \qquad G_t = CE_t + \Xi_t, \qquad H_t = \frac{1}{2}E_t E_t^\top. \tag{1}$$

With whitened layer inputs, the quadratic term is the expected output error measured by C, retaining an identity covariance across columns and a curvature matrix across rows (Martens & Grosse, 2015; Gupta et al., 2018). Every $(C, 0)$ minimizes the expected loss, while motion in E supplies the gradient H that changes the surrounding curvature. We use $\Xi_t = 0$ for periodic motion and independent Gaussian entries for statistical equilibrium, so the same coupling can be followed through both settings. For model coordinates y that generate $C(y)$, the corresponding gradient is $\frac{1}{2}\operatorname{tr}[(\partial C/\partial y_a)EE^\top]$. Thus a coordinate that sharpens one direction while flattening another receives the difference between their covariance weighted contributions, which connects our independent curvature chart to the way an architecture couples its parameter groups.

In this description, we hold the current matrix displacement fixed when computing the gradient received by the surrounding parameters, then average that gradient over the motion produced by training. Differentiating the loss after substituting a curvature dependent cycle or equilibrium would also differentiate the state itself, introducing a different contribution. By following the gradient that the receiving group actually uses, we can ask whether its accumulated direction has a scalar description even as the distribution of matrix motion changes with curvature.

2.2 Spectral Updates

Given a momentum input $N = U\operatorname{diag}(\sigma_i)V^\top$, we apply a positive scalar function P after Frobenius normalization,

$$\Phi_P(N) = U\operatorname{diag}(P(x_i))V^\top, \qquad x_i = \frac{\sigma_i}{\|N\|_F}, \qquad \sum_i x_i^2 = 1. \tag{2}$$

Exact polar updates use $P = 1$, while powers $P(x) = x^\alpha$ retain a graded dependence on input magnitude (Qi et al., 2026). For NS5 we retain the five original quintic compositions of Jordan et al. (2024), whose coefficients and full matrix derivatives appear in Appendix F. With fixed momentum coefficient $0 \le \beta < 1$, the fast update is

$$M_t = \beta M_{t-1} + (1-\beta)G_t, \quad N_t = \beta M_t + (1-\beta)G_t, \quad E_{t+1} = \rho E_t - \eta\Phi_P(N_t). \tag{3}$$

For $H = U\operatorname{diag}(h_i)U^\top$, the receiving matrix uses $\Psi_Q(H) = U\operatorname{diag}(Q(h_i/\|H\|_F))U^\top$, with its own spectral function and a smaller learning rate. Through the shared norm, a change in one gradient

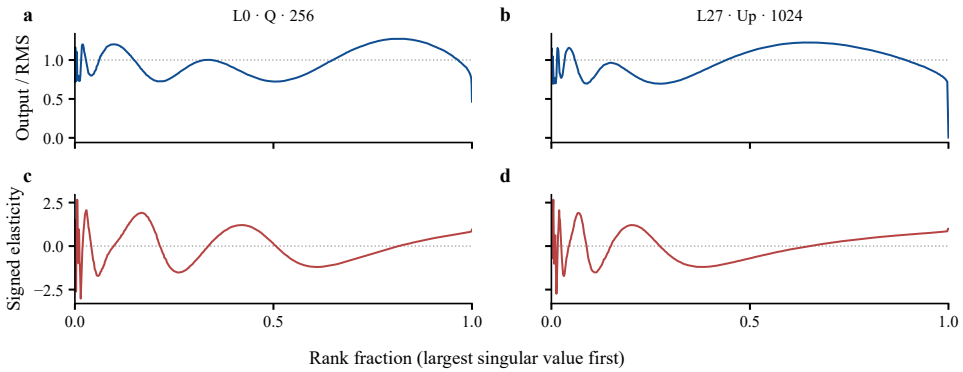


Figure 2: We compare output values and signed elasticity at two prescribed Transformer inputs, with singular directions ordered by input magnitude and outputs divided by their root mean square.

direction can change the relative response of another, while multiplying the whole input by a positive scalar leaves the update unchanged. Fixed momentum on this group is included in the complete averaging arguments in Appendices B and J.

To describe what the receiving optimizer sees, we collect the fast training state at frozen C into a distribution π_C^P. Passing that state through $H = EE^\top/2$ gives its covariance distribution, and the averaged curvature force is

$$F_{P,Q}(C) = \mathbb{E}_{\pi_C^P}[\Psi_Q(H)], \qquad \dot{C} = -F_{P,Q}(C). \tag{4}$$

For a stable cycle this average is over its phases, while for a mixing equilibrium it is over the stationary distribution. We derive both reductions from the complete updates, then ask when $F_{P,Q} = b\nabla_C V$ with $b > 0$ in the fixed Frobenius geometry, so one objective describes all relative curvature changes.

3 Spectral Response

3.1 Observed Response

To follow how an update changes with its input, we use the elasticity $e(x) = xP'(x)/P(x)$, which expresses the fractional output change per fractional input change. Because the raw singular values share a normalization, a perturbation $r_i = \delta \log \sigma_i$ gives $\delta \log P(x_i) = e(x_i)(r_i - \sum_j x_j^2 r_j)$, coupling the response of one direction to changes in the others.

We measure this response at two checkpoints of one Qwen3 0.6B training trajectory, reconstructing the effective momentum input from the next background minibatch. Across attention and feedforward projections, the NS5 output values are moderately concentrated while their elasticities cross zero and vary widely within each matrix (Figure 2). Directional perturbations of the full polynomial update agree with the analytic response, giving a direct way to observe how the same update values can respond differently to changes in their input.

3.2 Gaussian Response

For a fixed matrix perturbation, the input singular frame changes along with the noise, so the statistical response includes directions that rotate this frame. We can compute the combined response exactly at a centered isotropic Gaussian input ζ, by differentiating $\overline{\Phi}_P(G) = \mathbb{E}\Phi_P(G + \zeta)$. For every fixed matrix direction A, orthogonal symmetry and Gaussian integration by parts give the identity

$$\mathrm{D}\overline{\Phi}_P(0)[A] = a_\beta A, \qquad a_\beta = \frac{\mathbb{E}\langle \Phi_P(\zeta), \zeta\rangle_F}{kn\sigma_\beta^2} > 0. \tag{5}$$

Here σ_β^2 is the entry variance of the noise after the fixed Nesterov filter, and positivity follows because every nonzero singular direction contributes $\sigma_i P(x_i) > 0$ to the inner product. Through this identity,

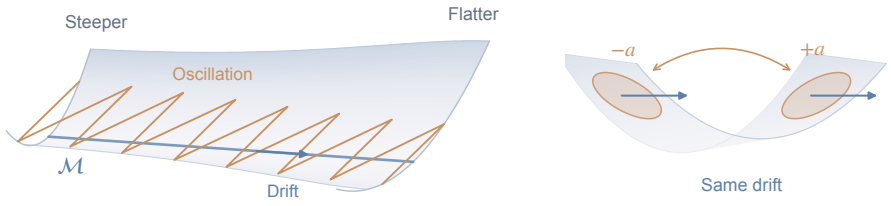


Figure 3: Persistent excursions generate motion along the manifold of minima, with both signs of the displacement contributing the same covariance gradient.

we can express the averaged restoring force directly through the alignment between the input and its update.

For the original NS5 with four rows, a wide Gaussian input has all four normalized singular values close to $1/2$, where its elasticity is less than -3. Its full matrix response still includes positive changes of singular orientation, and their isotropic average gives the positive coefficient in (5). We therefore obtain a concrete mechanism through which strongly negative amplitude responses coexist with statistical restoration, which will determine the equilibrium covariance in Section 5.

We can understand this distinction by viewing a matrix perturbation in the input's own singular frame. Changes of singular amplitude alter the lengths associated with that frame, while changes of orientation redistribute the output among directions held fixed by the rest of the model. Across an isotropic ensemble, the same fixed perturbation encounters many frames, so its mean response combines both effects. The positive inner product in (5) expresses their combined alignment, giving a statistical restoring coefficient even when amplitude changes within an individual input act in the opposite direction.

4 Curvature Objectives

4.1 Persistent Covariance

We first consider a square matrix block with $0 < \rho < 1$ and a centered oscillation $E_\pm = \pm aDR$, where $a = \eta/(1+\rho)$, $D = \mathrm{diag}(d_i) \succ 0$, and R is orthogonal. Opposite displacements produce the same $EE^\top$, so their gradients on curvature accumulate even as the matrix alternates between phases (Figure 3). In the eigenbasis of C, the cycle equations are

$$d_i = P(x_i), \qquad \lambda_i d_i = \tau x_i, \qquad \sum_i x_i^2 = 1, \qquad \overline{EE^\top} = a^2 D^2. \tag{6}$$

Under Euclidean updates on curvature, the averaged force is proportional to $w_i = d_i^2$, whose relative values change as the cycle adjusts to C. With a normalized slow spectral function, the corresponding force becomes $v_i = Q(d_i^2/T)$, where $T = (\sum_j d_j^4)^{1/2}$. Normal attraction of the full matrix cycle connects this force to the actual parameter and momentum trajectory, with an $O(\gamma)$ tracking error over fixed intervals of slow time as the curvature learning rate γ decreases (Appendix B).

As a curvature increases, its effective gradient input changes and the spectral map adjusts the amplitude of that direction, which feeds back into the next gradient. The shared normalization redistributes this adjustment across the matrix, so the branch equations express a coupled response even though C is diagonal in its eigenbasis. Through the elasticity, we can follow this feedback into the changing covariance, which explains why the derivative of the spectral function enters the curvature objective question alongside its current values.

4.2 Reciprocity

To understand when this force has a potential, we compare how changing curvature j alters the force on direction i with the reverse response. On a homogeneous branch, scaling all curvatures preserves

v, so a positive change of speed can produce a scalar gradient exactly when

$$\partial_j \left(\frac{v_i}{L} \right) = \partial_i \left(\frac{v_j}{L} \right), \qquad L = \sum_i \lambda_i v_i > 0. \tag{7}$$

We think this reciprocal response is the useful geometric content of a curvature objective, since it specifies how all directions fit together when a model parameter changes several curvatures at once.

Allowing a positive multiplier leaves the relative force between any two curvature directions unchanged, so the objective describes a preference among directions while the multiplier sets the training speed. In the periodic model, a common normalization of the deterministic covariance can be absorbed into this multiplier. Once the covariance itself fluctuates, a different scalar can be applied to each sample before averaging, which changes how those samples contribute to the mean direction. This distinction connects the same objective question to the statistical transformations in Section 6.

For Euclidean slow updates, differentiating (6) turns reciprocity into collinearity of the occupied points (λ_i, e_i). With at least four independent curvatures, positive analytic P, and an attracting branch with $0 < e_i < 1$, this yields exactly

$$P(x) = cx^\alpha + \ell x \quad (\alpha \neq 1), \qquad \text{or} \qquad P(x) = x(c + \ell \log x). \tag{8}$$

Pure powers have constant elasticity, while the linear correction permits elasticity to vary affinely with curvature along the branch. One classical cubic Newton–Schulz step belongs to this family, whereas two compositions leave it while preserving a unique normally attracting centered spectrum near uniform curvature. Thus the response within one stable branch already distinguishes finite orthogonalization choices, with the classification and quantitative NS5 example developed in Appendices C and F.

4.3 Joint Compatibility

When the slow function is $Q(s) = s^\kappa$ with $\kappa > 0$, we obtain the same power plus linear and logarithmic families for $P^{2\kappa-1}$, while at $\kappa = 1/2$ the family becomes $P(x) = cx^A e^{Bx}$. The exponent $2\kappa-1$ combines the squared fast amplitude, the receiving power, and the division by P in $\lambda_i = \tau x_i / P(x_i)$, giving a direct interpretation of how both updates enter the classification.

Conversely, for a fast power $P(x) = cx^\alpha$ with $0 < \alpha < 1$, let $p = 2\alpha/(1-\alpha)$ and $h = 2 - 1/p$. The positive analytic slow functions that admit a scalar objective are exactly

$$\begin{aligned} &Q(s) = c_1 s^r + c_2 s^h \quad (r \neq h), \\ \text{or} \qquad &Q(s) = s^h (c_1 + c_2 \log s). \end{aligned} \tag{9}$$

Requiring the fast function to remain compatible with both $\kappa = 1$ and $\kappa = 1/2$ already forces it to be a pure power, while every such power remains compatible with all $\kappa > 0$. For power pairs, the potential exponent is $1 + \kappa p$, so changing the receiving optimizer changes the curvature preference while preserving its scalar description.

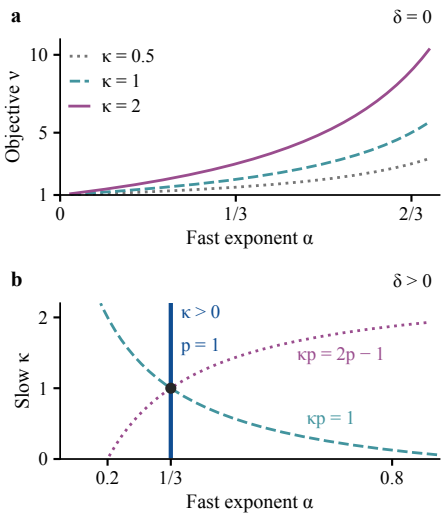


Figure 4: Power pairs determine a common curvature preference, while positive slow decay preserves scalar representation along the compatibility curves.

For these pairs, the branch relation gives $d_i^2 \propto \lambda_i^p$, and the receiving power changes the relative force to $\lambda_i^{\kappa p}$. Increasing either exponent therefore concentrates curvature reduction on the stiffer directions, with the resulting preference represented by $V = (1+\kappa p)^{-1} \log \sum_i \lambda_i^{1+\kappa p}$. Powers remain compatible because a common rescaling of their input produces one common output multiplier, which is absorbed into speed, allowing the relative curvature weights to retain this form as the other update changes.

With positive decay on the slow matrix, the cube root fast power is uniquely compatible with every positive slow power, because its covariance satisfies $D^2 \propto C$. Its slow normalization is then radial in the same coordinates as decay, allowing both terms to enter one potential (Figure 4). We derive the complete decay classification in Appendix E, while this covariance alignment gives the geometric reason for the distinguished exponent.

We can see why the alignment matters by comparing $(\alpha, \kappa) = (1/2, 1)$ with $(1/3, 2)$, which both produce a force proportional to λ_i^2 before decay. Adding the same positive decay preserves a scalar objective for the cube root pair and removes it for the other pair, because their different normalization magnitudes now change the balance between the spectral force and radial contraction. Thus even two choices that agree on their original curvature preference can respond differently when the receiving update acquires another component.

At the aligned exponent, we can express this combination through a power sum and a radial term in the same potential, because the slow spectral force has the common scale $\theta^{-\kappa}$ with $\theta = \|\lambda\|_2$. After this factor is extracted, its directional contribution integrates to $\sum_i \lambda_i^{\kappa+1}/(\kappa+1)$, while decay contributes $\delta\theta^{\kappa+2}/(\kappa+2)$. The resulting potential therefore combines relative curvature preference with contraction of the entire curvature vector, which gives a direct geometric interpretation of the cube root alignment.

5 Gaussian Equilibrium

5.1 Statistical Restoration

We now follow the positive response in (5) into matrix motion by taking four rows and independent gradient noise with entry variance σ^2. To compare how the receiving rule uses covariance, we choose $Q(s) = s$ or $Q(s) = s^2$ on the slow matrix. With $E = \sqrt{\eta}X$, fast decay factor $\rho = 1$, and curvature step size $\eta\varepsilon$, we first let the fast step size decrease and then separate the remaining time scales through $\varepsilon \to 0$. On time $u = \eta t$, the complete noisy NS5 update approaches

$$dX_u = -a_\beta C_u X_u \, du + \sqrt{b_\beta} \, dB_u, \qquad a_\beta > 0, \quad b_\beta > 0. \tag{10}$$

Fixed momentum changes the temporally correlated noise and its accumulated variance, while its lag on the slowly changing signal disappears in this scaling. The vector martingale limit and weighted response averaging are proved in Appendix I, using the full matrix state throughout.

At frozen C, the resulting Ornstein–Uhlenbeck process has independent columns with equilibrium law $X_{\cdot a} \sim N(0, T_\beta C^{-1})$, where $T_\beta = b_\beta/(2a_\beta)$. Coupling two trajectories through the same Brownian motion contracts their difference exponentially, so we obtain a stable statistical state even when the instantaneous NS5 inputs have strongly negative elasticities. The inverse curvature covariance also explains where the persistent signal is largest, since flatter directions sustain larger fluctuations under the same noise.

The frozen equilibrium has density proportional to $\exp[-\operatorname{tr}(X^\top C X)/(2T_\beta)]$ and satisfies detailed balance, so its statistical structure is especially simple. Through the same quadratic coupling as in the periodic model, these equilibrium fluctuations generate gradients on C, while the slow update converts their distribution into a deterministic drift. On slow time $s = \eta\varepsilon t$, the complete trajectory converges to (4), with the mixing and averaging argument developed in Appendix J.

5.2 Spectral Universality

For the smooth homogeneous spectral maps in this diffusion model, the fast function and momentum change the restoring rate and the scalar temperature, while the shape of the equilibrium covariance remains C^{-1}. A normalized slow update removes the temperature, so its average depends on the distribution of covariance shapes produced by that equilibrium. We regard this as a useful form of universality, because it identifies which details of the fast optimizer survive the statistical state and which are absorbed into its scale.

Indeed, we can couple two such equilibrium matrices by writing $X = \sqrt{T_\beta}Y$ with the same Gaussian Y, so their sample covariances differ by one positive scalar before the slow update. Since $\Psi_Q(cH) = \Psi_Q(H)$, the normalized outputs agree sample by sample under this coupling. The

universality therefore follows from an explicit relation between the distributions and holds for the transformation of each covariance sample.

For $z_a \sim N(0, C^{-1})$, define $H_n = n^{-1} \sum_{a=1}^{n} z_a z_a^\top$. The slow force can then be written as $F_{n,Q}(C) = \mathbb{E}\Psi_Q(H_n)$, where the sample covariance has a Wishart distribution and retains finite fluctuations across matrix directions. In the centered cycle, P determines a specific relative covariance through (6), while in Gaussian equilibrium the normalized update depends on this Wishart law. Both states therefore fit (4), with their different covariance distributions determining the role of the receiving update.

When we compare the allocations of persistent motion in these states, covariance grows as a positive power of curvature for a deterministic power cycle, whereas Gaussian forcing sustains larger variance along softer directions through C^{-1}. Following the same matrix statistic through these states therefore reveals how the fast dynamics determines which curvature directions send stronger signals to the remaining parameters. Applying the receiving rule to those signals then determines which combinations preserve a scalar description and how the remaining combinations change its geometric structure.

6 Covariance Fluctuations

6.1 Normalization Order

For a normalized square update, the averaged curvature force is

$$F_{n,2}(C) = \mathbb{E}\left[\frac{H_n^2}{\operatorname{tr}(H_n^2)}\right]. \tag{11}$$

If the covariance were fixed at its mean, this would reduce to $C^{-2}/\operatorname{tr}(C^{-2})$, which follows a scalar objective. At finite n, the matrix numerator and the shared normalization fluctuate together, so their correlation changes the relative force across directions.

We can interpret the two orders by considering a sample that places unusually large energy in one covariance direction. Normalizing that sample first reduces every contribution from it through the shared denominator, while normalizing after averaging leaves its larger energy available to influence the common numerator. In fact, the second order is the average of $\Psi_2(H_n)$ weighted by $\operatorname{tr}(H_n^2)$, whereas the first gives equal weight to each sample. Since curvature changes which directions carry these energy fluctuations, the two averages can respond differently to the same curvature perturbation.

We prove that for every sufficiently large finite n, on a neighborhood of $C = \operatorname{diag}(1, 2, 3, 4)$, the force in (11) lies outside every representation $b(C)\nabla_C V(C)$ with $b > 0$. The leading Wishart correction introduces an asymmetric normalized curvature response of order $1/n$, whose nonzero coefficient is established in Appendix K. Through this construction, a unique Gaussian equilibrium produces a slow force whose relative directions resist compression into one scalar objective.

The same reciprocal response test used for the cycle applies here, because rescaling C inversely rescales the covariance samples and leaves their normalized transformation unchanged. We can therefore compare the curvature derivatives of $F_i/\sum_j \lambda_j F_j$ across directions, while the Wishart expansion computes how their symmetry changes with finite covariance fluctuations. For a symmetric curvature perturbation A, differentiating the Gaussian density also gives

$$\mathrm{D}F(C)[A] = -\frac{n}{2}\operatorname{Cov}_C\left(\Psi_Q(H_n), \operatorname{tr}(AH_n)\right). \tag{12}$$

Through this relation, we can read the response of the mean force from correlations between its output and the covariance direction perturbed by C. For a cycle, the corresponding response follows by differentiating its branch equations, while here the distribution supplies the change through its statistical weights.

The distinction becomes more informative if we move normalization after averaging. Writing $S_1 = \operatorname{tr}(C^{-1})$, the exact Wishart second moment satisfies

$$\mathbb{E}H_n^2 = \frac{S_1}{n}\nabla_C\left[\log\det C - (n+1)\log S_1\right]. \tag{13}$$

Dividing this expectation by $\mathbb{E}\,\mathrm{tr}(H_n^2)$ preserves the same potential and changes only speed. Thus the order in which we normalize and average determines whether a scalar description survives, even though both constructions use the same Gaussian state and the same covariance samples.

6.2 Scalar Potentials

We can also preserve a scalar objective while transforming individual covariance samples, by using the normalized linear map $\Psi_1(H) = H/\|H\|_F$. For the Gaussian partition function $Z(C)$ and its weighted counterpart

$$J(C) = \int \frac{\exp[-\,\mathrm{tr}(Y^\top C Y)/2]}{\|YY^\top\|_F}\, dY, \qquad F_{n,1}(C) = -\frac{2}{Z(C)}\nabla_C J(C), \tag{14}$$

we obtain a positive speed multiplying the gradient of $-J$. The same argument applies when each sample is multiplied by a scalar weight before averaging, since the weight can be incorporated into the partition function (Appendix L).

Here $Z(C)$ normalizes the Gaussian density, while $J(C)$ collects the same covariance samples with their linear normalization weights. Through that scalar integral, the weighting of the distribution and the matrix force remain tied to one another, yielding the gradient identity in (14). This gives a constructive way to understand compatibility between a statistical state and its spectral transformation, with a potential obtained from the distribution itself.

In particular, $H/\,\mathrm{tr}\,H$ gives a compatible map with unit trace at every sample, matching the trace of the normalized square. Both averaged flows therefore reduce total curvature at the same speed, while their relative curvature changes have different scalar representations. We think this comparison captures the architectural meaning of compatibility, because a parameter group receives a distribution of signals and its update rule determines how those signals are combined across directions.

For a coordinate that changes two curvatures in opposite directions, this difference is visible in the component of the averaged update along that coordinate, even when the sum of the curvature forces agrees. A common measure of total progress can therefore accompany different preferences among the changes available to that coordinate. By retaining the covariance distribution between the two updates, we can follow how those preferences arise from the statistical state and the receiving rule together.

7 Discussion

We think the main implication concerns where an implicit curvature objective comes from, since even an equilibrium distribution of fast matrix states can produce a slow force whose directional responses resist representation by one potential. When we eliminate the fast variables, we retain their interaction with the receiving update through an expectation, within which nonlinear spectral transformations can preserve information about covariance fluctuations. From this perspective, we obtain scalar curvature preferences through compatibility between how motion is distributed and how its gradients are combined, which brings the construction of optimizer statistics into the same geometric question as the choice of spectral function.

We can develop this connection by changing how much covariance information enters each slow update, while keeping both spectral functions fixed. If we average m independent copies of H_n before applying the normalized square, we obtain exactly the Wishart covariance law with mn samples, so the leading obstruction to scalar representation decreases from order $1/n$ to order $1/(mn)$. By increasing this averaging, we approach the curvature direction obtained from the mean covariance, whose scalar representation follows from our analysis. We therefore see independent feature aggregation and independent temporal aggregation as interchangeable resources within this model, through which width and optimizer memory can influence the geometry of the mean update as well as the variability of individual steps.

For successive states along an actual trajectory, temporal correlations give this connection further structure, because different curvature directions relax over different time scales. If we retain the covariance over a window of duration w, we feed $\bar{H}_w = w^{-1}\int_0^w X_u X_u^\top\, du$ into the slow transformation, whose input distribution now depends on the evolution of the fast state throughout that window.

In our Gaussian diffusion, we can absorb temperature into an overall covariance scale, while temporal dependence remains through $e^{-a_\beta Ct}$ and hence through the dimensionless window length $a_\beta w$. We can consequently recover sensitivity to the fast spectral rule through covariance averaging, even when the instantaneous normalized forces agree, since changing the restoring rate changes how much statistically distinct motion enters each update. For sufficiently long windows at frozen curvature, we recover the mean covariance, whereas shorter windows retain the fluctuations responsible for the additional directional interaction.

We think this dependence gives parameter grouping an architectural meaning that extends beyond deciding which tensors share an optimizer. When we split a projection into blocks, we choose which directions contribute to each shared normalization, while the statistics retained within those blocks determine how their fluctuations influence one another. Through these choices, we alter how a change in one feature direction modifies the updates assigned to other directions, even when the underlying loss and the scalar spectral function remain fixed. We can therefore understand grouping and covariance accumulation together as choices about which interactions survive into the evolution of surrounding parameters.

We are especially interested in using this freedom to design coupled updates around a desired curvature geometry, with the receiving transformation and its accumulated statistics treated as explicit design variables. Within the periodic model, power compatibility identifies fast rules whose scalar representation survives changes in the slow spectral exponent, while covariance averaging in the Gaussian model controls how strongly fluctuations modify an otherwise integrable force.

8 Related Work

Matrix optimization connects the curvature factors of K-FAC and Shampoo (Martens & Grosse, 2015; Gupta et al., 2018) to Adam in a preconditioner's eigenbasis in SOAP (Vyas et al., 2025) and norm constrained directions in Scion (Pethick et al., 2025). For Muon, orthogonalized momentum (Jordan et al., 2024), update scaling and decay (Liu et al., 2025), and layerwise norm aggregation (Crawshaw et al., 2025) provide concrete choices whose interaction motivates our study of coupled spectral updates.

Finite spectral transformations have been analyzed through minimax approximation in Polar Express (Amsel et al., 2026), stationarity guarantees for Newton–Schulz (Kim & Oh, 2026), and distributed block orthogonalization (Khaled et al., 2026). Spectral powers broaden the update family (Qi et al., 2026), while margin and associative memory analyses connect these methods to learning preferences (Fan et al., 2026; Gronich & Vardi, 2026; Wang et al., 2026). We follow their microscopic response into the distribution of gradients received by another optimizer.

Discrete oscillations and their averaged curvature effects are developed through edge of stability and central flows (Cohen et al., 2021; Arora et al., 2022; Cohen et al., 2024), while stochastic analyses relate gradient noise to motion among minima (Damian et al., 2021; Li et al., 2021). The broader question of how implicit bias can be represented also appears in matrix factorization (Razin & Cohen, 2020). Our classification and Gaussian construction identify how stable covariance statistics combine with a receiving spectral update, with further connections to optimization geometry and structured models in Appendix O.

9 Conclusion

We have described curvature preference through the covariance distribution produced by one matrix optimizer and transformed by another, connecting actual spectral responses to periodic motion and Gaussian equilibrium. For a centered cycle, the spectral functions determine a covariance whose compatibility can be classified, while isotropic averaging produces a positive restoring force and a universal covariance shape. Finite fluctuations in that shape then expose a further choice, since transforming and normalizing each sample can destroy a scalar objective that survives when averaging comes first. We think this distributional description provides a direct way to reason about coupled parameter groups, with their microscopic response, statistical motion, and receiving update contributing distinct parts of the learning dynamics.

10 AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing. We have not used generative AI tools for generating ideas or assisting with math or coding, as well as all other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

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. In *International Conference on Learning Representations*, volume 2026, pp. 138323–138360, 2026.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pp. 948–1024. PMLR, 2022.

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

Jeremy M Cohen, Alex Damian, Ameet Talwalkar, J Zico Kolter, and Jason D Lee. Understanding optimization in deep learning with central flows. *arXiv preprint arXiv:2410.24206*, 2024.

Michael Crawshaw, Chirag Modi, Mingrui Liu, and Robert M Gower. An exploration of non-euclidean gradient descent: Muon and its many variants. *arXiv preprint arXiv:2510.09827*, 2025.

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359, 2022.

Jared Q Davis, Albert Gu, Krzysztof Choromanski, Tri Dao, Christopher Re, Chelsea Finn, and Percy Liang. Catformer: Designing stable transformers via sensitivity analysis. In *International Conference on Machine Learning*, pp. 2489–2499. PMLR, 2021.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. *Advances in Neural Information Processing Systems*, 38:39622–39669, 2026.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Eitan Gronich and Gal Vardi. The implicit bias of adam and muon on smooth homogeneous neural networks. *arXiv preprint arXiv:2602.16340*, 2026.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850. PMLR, 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

Rustem Islamov, Michael Crawshaw, Jeremy Cohen, and Robert Gower. Non-euclidean gradient descent operates at the edge of stability. *arXiv preprint arXiv:2603.05002*, 2026.

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

Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization. In *International Conference on Learning Representations*, volume 2026, pp. 17389–17416, 2026.

Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with newton-schulz. In *International Conference on Learning Representations*, volume 2026, pp. 45672–45718, 2026.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

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

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

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

Grigorios A Pavliotis and Andrew M Stuart. *Multiscale methods: averaging and homogenization*, volume 53. Springer New York, 2008.

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

Xianbiao Qi, Marco Chen, Jiaquan Ye, Yelin He, and Rong Xiao. Delving into muon and beyond: Deep analysis and extensions. *arXiv preprint arXiv:2602.04669*, 2026.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. pmlr, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 93423–93444, 2025.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Tan. Muon outperforms adam in tail-end associative memory learning. In *International Conference on Learning Representations*, volume 2026, pp. 37382–37419, 2026.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR, 2021.

APPENDIX

A Matrix Cycles **15**
A.1 Cycle Equation . . . 15
A.2 Matrix Derivative . . . 15
A.3 Attraction and Continuation . . . 16

B Matrix Averaging **17**
B.1 Coupled Updates . . . 17
B.2 Augmented Derivative . . . 18
B.3 Averaging Proof . . . 19
B.4 Curvature Covector . . . 20

C Curvature Objectives **20**
C.1 Integrating Factors . . . 20
C.2 Slow Powers . . . 21
C.3 Spectral Composition . . . 22

D Joint Classification **22**
D.1 Normalized Pairs . . . 22
D.2 Fast Powers . . . 23

E Decay and Covariance **24**
E.1 Cube-Root Compatibility . . . 24
E.2 Power Pairs . . . 25
E.3 Polar Endpoint . . . 26

F Finite Iterations **26**
F.1 Explicit Branch . . . 27
F.2 Force Derivative . . . 27
F.3 Force and Decay . . . 28

G Statistical Model **29**

H Gaussian Response **30**
H.1 Noise Filter . . . 30
H.2 Average Response . . . 30
H.3 Negative Elasticities . . . 31

I Diffusion Limit **31**
I.1 Matrix Invariance . . . 31
I.2 Remainder Bounds . . . 32
I.3 Derivative Averaging . . . 32

J Covariance Averaging **33**
J.1 Gaussian Equilibrium . . . 33
J.2 Radial Bounds . . . 33
J.3 Poisson Equation . . . 34
J.4 Momentum and Decay . . . 35

K Wishart Fields **35**
K.1 Distribution Response . . . 35
K.2 First Correction . . . 36
K.3 Curvature Obstruction . . . 37

L Statistical Potentials **37**
L.1 Linear Updates . . . 37
L.2 Averaged Squares . . . 38
L.3 Equal Trace Updates . . . 39

M Spectral Measurements **39**

N Additional Comparisons **39**

O More Related Work **42**
O.1 Optimization Geometry 42
O.2 Implicit Selection 42
O.3 Curvature Dynamics 42
O.4 Statistical Averaging 43
O.5 Spectral Updates 43
O.6 Structured Matrices 43

A Matrix Cycles

All matrix spaces carry the Frobenius inner product. Let $C = C^\top \succ 0$ and

$$f(C, E) = \frac{1}{2}\operatorname{tr}(E^\top C E), \qquad \nabla_E f = CE, \qquad \nabla_C f = \frac{1}{2} E E^\top. \tag{15}$$

For a full rank input $N = U \operatorname{diag}(\sigma_i) V^\top$, define

$$\Phi_P(N) = U \operatorname{diag}\left(P\left(\frac{\sigma_i}{\|N\|_F}\right)\right) V^\top. \tag{16}$$

The scalar function P is positive and C^3 on a neighborhood of the occupied normalized singular values. The associated matrix map is C^3 on the input neighborhoods used below. It is odd, orthogonally equivariant, and homogeneous of degree zero. The fast update is

$$\begin{aligned} M_t &= \beta M_{t-1} + (1-\beta) C E_t, \\ N_t &= \beta M_t + (1-\beta) C E_t, \qquad \eta > 0, \quad 0 \le \beta < 1, \quad 0 < \rho < 1. \\ E_{t+1} &= \rho E_t - \eta \Phi_P(N_t), \end{aligned} \tag{17}$$

The state before gradient evaluation is (E_t, M_{t-1}).

A.1 Cycle Equation

Proposition A.1 (Centered cycles)**.** *Work in an eigenbasis $C = \operatorname{diag}(\lambda_i)$. Suppose*

$$x_i > 0, \qquad \sum_i x_i^2 = 1, \qquad d_i = P(x_i), \qquad \lambda_i d_i = \tau x_i, \qquad \tau > 0. \tag{18}$$

Put $D = \operatorname{diag}(d_i)$ and

$$a = \frac{\eta}{1+\rho}, \qquad b = \frac{(1-\beta)a}{1+\beta}, \qquad c = \frac{(1-\beta)(1+2\beta)a}{1+\beta}. \tag{19}$$

For each $R \in O(k)$, the matrices

$$E_t = (-1)^t a D R, \qquad M_{t-1} = -(-1)^t b C D R, \qquad N_t = (-1)^t c C D R \tag{20}$$

form a centered period two trajectory. Every full rank centered period two trajectory has this form after choosing a left eigenbasis of C.

Proof. On an alternating trajectory the momentum equation gives $(1+\beta)M_t = (1-\beta)CE_t$. The Nesterov input then has the coefficient c in (19). Since $CD = \tau \operatorname{diag}(x_i)$, normalization gives $\Phi_P(CDR) = DR$. The parameter recurrence is the scalar identity $-a = \rho a - \eta$.

For the converse, put $D_0 = E_0/a$, $\tau = \|CD_0\|_F$, and $X = CD_0/\tau$. The cycle equation is $\Phi_P(X) = D_0$. With $S = XX^\top$ and $A(q) = P(\sqrt{q})/\sqrt{q}$, it becomes $CA(S)X = \tau X$. Invertibility of X gives $CA(S) = \tau I$. Hence C and S commute. Their common eigenbasis and the singular value decomposition of X give (18) and a right orthogonal factor. □

A.2 Matrix Derivative

At $X = \operatorname{diag}(x_i)$ with $\|X\|_F = 1$, a diagonal perturbation $H = \operatorname{diag}(u)$ satisfies

$$\operatorname{diag}\big(\mathrm{D}\Phi_P(X)[H]\big) = \operatorname{diag}(P'(x_i))(I - xx^\top)u. \tag{21}$$

For an off-diagonal pair, write

$$A_{ij} = \frac{d_i - d_j}{x_i - x_j}, \qquad B_{ij} = \frac{d_i + d_j}{x_i + x_j}. \tag{22}$$

At a repeated input, their values are $P'(x_i)$ and $P(x_i)/x_i$. The pair derivative is

$$\begin{pmatrix} (\mathrm{D}\Phi_P(X)[H])_{ij} \\ (\mathrm{D}\Phi_P(X)[H])_{ji} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} A_{ij} + B_{ij} & A_{ij} - B_{ij} \\ A_{ij} - B_{ij} & A_{ij} + B_{ij} \end{pmatrix} \begin{pmatrix} H_{ij} \\ H_{ji} \end{pmatrix}. \tag{23}$$

To obtain (21), differentiate $N/\|N\|_F$ at X. Its derivative is $H - X\langle X, H\rangle$. For (23), infinitesimal left and right frame rotations with entries u, v give

$$H_{ij} = ux_j - x_i v, \qquad H_{ji} = -ux_i + x_j v.$$

The output has the same expressions with x replaced by d. Their sum and difference give (22). Continuity supplies the repeated-input values. Homogeneity gives $\mathrm{D}\Phi_P(sX) = s^{-1}\mathrm{D}\Phi_P(X)$ for $s > 0$.

Lemma A.2 (Spatial responses)**.** *At a branch satisfying (18), define*

$$\mathcal{B}[H] = \tau^{-1}\mathrm{D}\Phi_P(X)[CH], \qquad e_i = \frac{x_i P'(x_i)}{d_i}. \tag{24}$$

Let $V^\top V = I$ and $V^\top x = 0$. The diagonal responses are zero and the eigenvalues of

$$K_x = V^\top \operatorname{diag}(e_i) V. \tag{25}$$

Each off-diagonal pair has a response one from a right rotation. Its other response is

$$\mu_{ij} = \frac{x_i x_j (d_i^2 - d_j^2)}{d_i d_j (x_i^2 - x_j^2)}. \tag{26}$$

At $x_i = x_j$, this value is $e_i = e_j$. All responses are real. The conditions

$$0 < e_i < 1, \qquad 1 < d_i/d_j < x_i/x_j \quad \textit{whenever } x_i > x_j \tag{27}$$

place every response outside the radial and right-rotation directions in $(0, 1)$.

Proof. The diagonal response matrix is

$$\operatorname{diag}(P'(x_i))(I - xx^\top)\operatorname{diag}(x_i/d_i).$$

Conjugation by $\operatorname{diag}(d_i/x_i)$ gives $\operatorname{diag}(e_i)(I - xx^\top)$. In the basis $[x, V]$, this matrix has diagonal blocks zero and K_x. For a pair, multiply the matrix in (23) by $\operatorname{diag}(\lambda_i, \lambda_j)/\tau$. The vector $(d_i, -d_j)$ is an eigenvector with eigenvalue one. The determinant of the pair block is $A_{ij}B_{ij}\lambda_i\lambda_j/\tau^2$, which gives (26). Repeated inputs have equal curvatures by (18), and their remaining response is e_i. Finally, with $r = x_i/x_j > 1$ and $s = d_i/d_j$, the pair response is $(s - s^{-1})/(r - r^{-1})$. It lies in $(0, 1)$ exactly when $1 < s < r$. The diagonal conclusion follows by compression of $0 < \operatorname{diag}(e_i) < I$. □

A.3 Attraction and Continuation

For a perturbation of the before-gradient state, use coordinates $u = \delta E$ and $r = (1-\beta)^{-1}C^{-1}\delta M_{t-1}$. Each spatial response μ gives the matrix

$$J_\mu = \begin{pmatrix} \rho - (1+\beta)\varkappa\mu & -\beta^2\varkappa\mu \\ 1 & \beta \end{pmatrix}, \qquad \varkappa = \frac{(1+\beta)(1+\rho)}{1+2\beta}. \tag{28}$$

Indeed $N_t = \beta^2 M_{t-1} + (1-\beta^2)CE_t$, and substitution of the cycle input into its derivative gives (28). Its characteristic polynomial is

$$z^2 - [\rho + \beta - (1+\beta)\varkappa\mu]z + \beta(\rho - \varkappa\mu). \tag{29}$$

The three quadratic stability inequalities are

$$1 - \beta\rho + \beta\varkappa\mu > 0, \qquad (1-\rho)(1-\beta) + \varkappa\mu > 0, \qquad (1+\rho)(1+\beta)(1-\mu) > 0. \tag{30}$$

For $\beta > 0$, the second lower bound is stronger than the first because

$$\frac{1-\beta\rho}{\beta} - (1-\rho)(1-\beta) = \frac{1-\beta}{\beta} + \beta(1-\rho) > 0.$$

For $\beta = 0$, the roots are zero and $\rho - (1+\rho)\mu$. Thus strict stability holds exactly when

$$-\frac{(1-\rho)(1-\beta)}{\varkappa} < \mu < 1. \tag{31}$$

At $\mu = 0$, the roots are ρ, β. At $\mu = 1$, they are -1 and $\beta(1+\beta-\beta\rho)/(1+2\beta)$. Oddness of Φ_P makes the derivative identical at opposite phases, so the return multipliers are the squares of these roots.

Theorem A.3 (Normal attraction)**.** *Suppose every eigenvalue of K_x and every physical pair response in (26) belongs to $(0,1)$. Then*

$$\mathcal{M}_C = \{(aDR, -bCDR) \mid R \in O(k)\} \tag{32}$$

is a compact normally attracting manifold of fixed points of the two-step return. Its unit eigenspace is semisimple and equals its tangent space. Every sufficiently close state approaches a member of this family under the return map.

Proof. Invertibility of D makes the displayed parameterization a smooth embedding. The physical modes and the radial mode contract by (31). Every off-diagonal pair supplies one response one, with simple temporal root -1. These are precisely the $k(k-1)/2$ right-rotation tangents. All complementary multipliers have modulus below one, so the return unit eigenspace is semisimple.

The spectral projection onto the stable bundle varies smoothly along the compact family. On this bundle a convergent sum of iterated quadratic norms gives an adapted norm with a uniform contraction factor. In the associated tubular coordinates, the transverse return is bounded by this contraction plus a quadratic remainder. The tangent increment has a quadratic remainder because the derivative preserves the stable bundle. Transverse errors decay geometrically and tangent increments are summable. The resulting trajectory converges to a point of the family. Orthogonal equivariance makes these estimates uniform in R. □

Lemma A.4 (Branch continuation)**.** *If (18) holds with $0 < e_i < 1$, its labelled solution continues smoothly to nearby independent curvatures. It is analytic when P is analytic. If $A(q) = P(\sqrt{q})/\sqrt{q}$ has $A' < 0$ throughout $(0,1)$, each curvature vector has at most one interior allocation. If $P' > 0$ also holds, every existing allocation satisfies Theorem A.3.*

Proof. Writing $q_i = x_i^2$, the equations become $\lambda_i A(q_i) = \tau$ and $\sum_i q_i = 1$. Their derivative satisfies

$$\lambda_i A'(q_i) = \frac{\tau(e_i - 1)}{2q_i} < 0. \tag{33}$$

A kernel vector obeys $\lambda_i A'(q_i)\delta q_i = \delta\tau$ and $\sum_i \delta q_i = 0$. The sum of the reciprocals of the negative coefficients is negative, so the kernel is zero. The implicit function theorem gives continuation. For fixed τ, strict monotonicity of A gives at most one q_i, and each q_i decreases with τ. Their sum can equal one for at most one multiplier. The branch equations order q_i with λ_i. Positivity of P' and negativity of A' give $0 < e_i < 1$ and the pair inequalities in (27). □

B Matrix Averaging

B.1 Coupled Updates

Let $C_t \in \mathrm{Sym}_{++}(k)$, $E_t, M_{t-1} \in \mathbb{R}^{k\times k}$, and $S_{t-1} \in \mathrm{Sym}(k)$. All entries use the Frobenius geometry. For an SPD input $Z = U\,\mathrm{diag}(z_i)U^\top$, put

$$\Psi_Q(Z) = U\,\mathrm{diag}\left(Q\left(\frac{z_i}{\|Z\|_F}\right)\right)U^\top. \tag{34}$$

Let $\mathcal{R}$ denote either Ψ_Q or the identity map. The complete update is

$$\begin{aligned} M_t &= \beta M_{t-1} + (1-\beta)C_tE_t, & N_t &= \beta M_t + (1-\beta)C_tE_t,\\ E_{t+1} &= \rho E_t - \eta\Phi_P(N_t), & H_t &= \frac{1}{2}E_tE_t^\top,\\ S_t &= \beta_s S_{t-1} + (1-\beta_s)H_t, & Z_t &= \beta_s S_t + (1-\beta_s)H_t,\\ C_{t+1} &= C_t - \gamma[\mathcal{R}(Z_t) + \delta C_t], \end{aligned} \tag{35}$$

where $0 \le \beta_s < 1$ and $\delta \ge 0$. Both gradients use (C_t, E_t). The coefficients $\eta, \beta, \rho, \beta_s, \delta$ are fixed as γ tends to zero.

Let $\mathcal{U}$ be an open set of SPD matrices with simple spectrum. Order the eigenvalues consistently, write their projectors as $\Pi_i(C)$, and suppose a C^2 labelled branch defines

$$D(C) = \sum_i d_i(\lambda(C))\Pi_i(C) \succ 0, \qquad \Phi_P(CD(C)) = D(C). \tag{36}$$

Assume its full frozen return satisfies Theorem A.3. Define

$$h_*(C) = \frac{a^2}{2} D(C)^2, \qquad F_\delta(C) = \mathcal{R}(h_*(C)) + \delta C. \tag{37}$$

Choose compact sets $K \subset \operatorname{int} K_1 \subset \mathcal{U}$, with positive separation between K and the complement of $\operatorname{int} K_1$. All branch, map, rank, and spectral margins hold on a neighborhood of K_1, and the maps are C^3 on the corresponding input neighborhoods. Normalization uses the exact norm.

Theorem B.1 (Matrix averaging)**.** *Under these assumptions, let $\bar{C}$ solve*

$$\frac{d\bar{C}}{ds} = -F_\delta(\bar{C}), \qquad \bar{C}(0) = C_0, \tag{38}$$

and suppose $\bar{C}(s) \in K$ for $0 \le s \le H$. There are a fixed small full-state tube around

$$\mathcal{N}_C = \{(aD(C)R, -bCD(C)R, h_*(C)) \mid R \in O(k)\}, \tag{39}$$

constants $q < 1$, $A_0, A_H < \infty$, and $\gamma_0 > 0$ with the following property. If the initial state lies in this tube at quotient distance r_0, the iteration with $0 < \gamma < \gamma_0$ remains in the branch neighborhood for $0 \le \gamma t \le H$. Its distances at $t = 2j$ satisfy

$$r_j \le q^j r_0 + \frac{A_0 \gamma}{1-q}. \tag{40}$$

For the linear interpolation $C^\gamma(s)$ through C_t at $s = \gamma t$,

$$\sup_{0 \le s \le H} \|C^\gamma(s) - \bar{C}(s)\|_F \le A_H \gamma \left(1 + \frac{r_0}{1-q}\right). \tag{41}$$

The state tube includes all entries of E, M and all symmetric entries of C, S. The constants may depend on the fixed coefficients and compact margins.

B.2 Augmented Derivative

We first establish the stable frozen quotient used in the proof. At $\gamma = 0$, the curvature is fixed and both phases satisfy $H_t = h_*$. Thus the slow buffer has fixed value h_*. Let J be the complete fast one-step derivative on $(\delta E, \delta M_{t-1})$, and let Π_E extract its first component. Formula (28) describes its spatial modes, including every off-diagonal pair. Define

$$\mathcal{H}_E[u] = \frac{1}{2}(uE^\top + Eu^\top). \tag{42}$$

The augmented one-step derivatives at the two phases are

$$\mathcal{J}_\pm = \begin{pmatrix} J & 0 \\ (1-\beta_s)\mathcal{H}_{E_\pm}\Pi_E & \beta_s I \end{pmatrix}. \tag{43}$$

Since $\mathcal{H}_{E_-} = -\mathcal{H}_{E_+}$, their product is

$$\mathcal{J}_-\mathcal{J}_+ = \begin{pmatrix} J^2 & 0 \\ B_C & \beta_s^2 I \end{pmatrix}, \qquad B_C = (1-\beta_s)\mathcal{H}_{E_+}(\beta_s \Pi_E - \Pi_E J). \tag{44}$$

This adds $k(k+1)/2$ stable buffer multipliers β_s^2. Stable resonances can produce Jordan blocks, which the adapted norm below accommodates.

For a skew matrix Ω, the right-rotation tangent is

$$w_\Omega = (E_+\Omega, M_-\Omega), \qquad Jw_\Omega = -w_\Omega.$$

Equation (42) gives

$$\mathcal{H}_{E_+}[E_+\Omega] = \frac{1}{2}E_+(\Omega + \Omega^\top)E_+^\top = 0,$$

so $B_C w_\Omega = 0$. The tangent unit eigenspace of the augmented return is therefore the tangent of (39), with zero slow-buffer component. Splitting the fast space into tangent and stable invariant subspaces makes the remaining augmented block strictly stable. This proves semisimplicity of the unit eigenspace and includes every slow-buffer orientation.

B.3 AVERAGING PROOF

Proof of Theorem B.1. The complete update descends under the right action

$$(E, M, S, C) \longmapsto (ER, MR, S, C).$$

The fast map is equivariant and $EE^\top$ is invariant. A local quotient chart uses

$$R_C(E) = \mathrm{polar}(D(C)E), \qquad \widehat{E} = ER_C(E)^\top, \qquad \widehat{M} = MR_C(E)^\top. \tag{45}$$

Near the family, $D(C)\widehat{E}$ is SPD. Its independent symmetric entries, all entries of $\widehat{M}$, and all symmetric entries of S form coordinates. The frozen fixed point is

$$z_*(C) = (aD(C)^2, -bCD(C), h_*(C)).$$

The chart and inverse representative are uniformly smooth on a sufficiently small tube over K_1. Compactness of $O(k)$ makes the bounds uniform along the orbit.

Let A_C be the frozen two-step derivative in this quotient. Equation (44) makes it strictly stable. The discrete Lyapunov equation

$$A_C^\top W_C A_C - W_C = -I \tag{46}$$

has the positive solution $W_C = \sum_{n\geq 0}(A_C^n)^\top A_C^n$. Equivalently, its entries solve an invertible finite linear system. Hence W_C, W_C^{-1} and $\mathrm{D}W_C$ are uniformly bounded on K_1. Since $W_C \succeq I$, there is $q_0 < 1$ such that $\|A_C v\|_{W_C} \leq q_0 \|v\|_{W_C}$ uniformly. This controls both stable eigenvalues and their eigenspace conditioning.

Write $\Lambda_j = C_{2j}$ and use the chart centered at $z_*(\Lambda_j)$. Up to a first exit from the chosen neighborhoods, Taylor expansion of the frozen map gives $q_0 r_j + K r_j^2$. Each actual curvature update has size $O(\gamma)$. Replacing the two changing curvatures by Λ_j, recentering at $z_*(\Lambda_{j+1})$, and changing the metric therefore give

$$r_{j+1} \leq q_0 r_j + K r_j^2 + K\gamma r_j + A_0\gamma. \tag{47}$$

Choose the tube radius and then γ_0 so that the first three terms are at most $q r_j$ for some $q_0 < q < 1$. Taking r_0 and $A_0\gamma/(1-q)$ each below half the tube radius gives (40). A slightly larger admissible tube covers intermediate odd iterates by one-step continuity.

At an exact frozen family point, the first fast step uses Λ_j and reaches the opposite phase. The two covariance gradients used in the slow pair are therefore both $h_*(\Lambda_j)$, and the buffer stays at that value. The exact pair increment from this point is

$$-2\gamma F_\delta(\Lambda_j) + \gamma^2 \delta F_\delta(\Lambda_j).$$

For a state at distance r_j, local Lipschitz bounds on the first fast step, covariance, buffer, and slow readout give

$$\Lambda_{j+1} = \Lambda_j - 2\gamma F_\delta(\Lambda_j) + \mathcal{E}_j, \qquad \|\mathcal{E}_j\|_F \leq K_1\gamma r_j + K_2\gamma^2. \tag{48}$$

The second fast step uses the changed curvature. Its contribution is already included in (47). Equality of the two leading covariance gradients follows from the simultaneous update and the opposite parameter phases.

Set $h = 2\gamma$ and $e_j = \|\Lambda_j - \bar{C}(jh)\|_F$. Smoothness gives a Lipschitz constant L_F for F_δ and an $O(h^2)$ local Euler error. Thus

$$e_{j+1} \leq (1 + hL_F)e_j + K_1\gamma r_j + K_3\gamma^2.$$

For $J \leq H/(2\gamma) + 1$, summation and (40) yield

$$e_J \leq e^{L_F(H+2\gamma)} \left[\frac{K_1\gamma r_0}{1-q} + J\gamma^2 \left(\frac{K_1 A_0}{1-q} + K_3 \right) \right]. \tag{49}$$

One actual slow step and one interpolated ODE step each have size $O(\gamma)$, extending this bound to (41). Finally decrease γ_0 so that (49) and one step are smaller than a fixed fraction of the separation between K and the boundary of K_1. The limiting curve remains in K, so a first curvature exit is impossible. Equation (47) and the odd-iterate tube exclude a first state exit. This closes all provisional bounds. □

B.4 Curvature Covector

For $\mathcal{R} = \Psi_Q$, write

$$T = \left(\sum_j d_j^4 \right)^{1/2}, \qquad s_i = d_i^2/T, \qquad v_i = Q(s_i). \tag{50}$$

Then $F_\delta(C) = \sum_i (v_i + \delta\lambda_i)\Pi_i(C)$. For arbitrary symmetric H, eigenvalue differentiation gives

$$\operatorname{tr}(F_\delta(C)H) = \sum_i (v_i + \delta\lambda_i)\, d\lambda_i(C)[H]. \tag{51}$$

The orientation contribution vanishes because the trace pairing of a diagonal matrix with an eigenframe commutator vanishes.

Consequently the full matrix field has a positive scalar representation $F_\delta = b\nabla_C \mathcal{V}$ locally if and only if its eigenvalue covector has such a representation. Restriction to a fixed diagonal slice proves necessity. For sufficiency, compose an eigenvalue potential V and multiplier b with the ordered eigenvalues and use

$$\nabla_C V(\lambda(C)) = \sum_i \partial_i V(\lambda(C))\Pi_i(C).$$

This is a local statement on the specified simple-spectrum neighborhood. For the identity slow readout and $\delta = 0$, the coefficients are $a^2 d_i^2/2$.

C Curvature Objectives

C.1 Integrating Factors

A positive scalar representation of a smooth nonzero one-form ω means $\omega = b\, dV$ on a neighborhood, with smooth $b > 0$. The Frobenius condition for this representation is $\omega \wedge d\omega = 0$ throughout that neighborhood. Indeed, this identity makes $d\omega$ vanish on pairs of vectors in $\ker\omega$. The identity $d\omega(X, Y) = -\omega([X, Y])$ then makes that distribution involutive. The local Frobenius theorem supplies a function V with $\ker dV = \ker\omega$. Their nonzero proportionality factor has a constant sign after shrinking the neighborhood, and reversing V makes it positive. Conversely, $d\omega = db \wedge dV$ directly gives the condition.

Lemma C.1 (Homogeneous normalization)**.** *Suppose $\omega = \sum_i w_i(\lambda)d\lambda_i$ has smooth positive coefficients of degree zero on an open positive curvature region. Set $L = \sum_i \lambda_i w_i$. Then the following are locally equivalent,*

$$\omega = b\, dV \quad (b > 0), \qquad \omega \wedge d\omega = 0, \qquad d(\omega/L) = 0. \tag{52}$$

When they hold, a potential can be chosen with $dV = \omega/L$.

Proof. Let $\mathcal{E} = \sum_i \lambda_i \partial_{\lambda_i}$. Euler's identity gives

$$\iota_{\mathcal{E}} d\omega = \omega - dL, \qquad \iota_{\mathcal{E}}(\omega \wedge d\omega) = L\, d\omega - dL \wedge \omega.$$

Thus the middle condition implies $d(\omega/L) = 0$, and wedging the latter identity with ω gives the converse. A scalar representation implies the middle condition by direct differentiation. Finally a closed form on a sufficiently small ball has a primitive. For $\zeta = \omega/L$, one such primitive is

$$V(\lambda) = \int_0^1 \zeta_{\lambda_0 + t(\lambda - \lambda_0)}[\lambda - \lambda_0]\, dt.$$

Differentiation and equality of cross derivatives give $dV = \zeta$. Since $L > 0$, this supplies a positive factor. □

The full-curvature hypothesis in the results below means that every λ_i varies independently on an open set. A positive scalar rescaling multiplies $\omega \wedge d\omega$ by the square of that scalar.

C.2 Slow Powers

Assume P is positive and real analytic on a connected interval $I \subset (0,1)$ containing all occupied fast input levels. Work on a full open branch satisfying (18), $0 < e_i < 1$, and the attraction assumptions above. A normalized slow power $Q(s) = s^\kappa$ gives, up to a common positive factor,

$$v_i = P(x_i)^{2\kappa}, \qquad \omega_\kappa = \sum_i v_i \, d\lambda_i. \tag{53}$$

There is no separate slow decay in this section.

Differentiating the branch equations with $q_i = x_i^2$ gives

$$\begin{aligned} dq_i &= \frac{2q_i}{1-e_i}(d\log\lambda_i - d\log\tau), \\ d\log\tau &= \sum_i \frac{\zeta_i}{Z}\, d\log\lambda_i, \qquad \zeta_i = \frac{q_i}{1-e_i}, \quad Z = \sum_i \zeta_i. \end{aligned} \tag{54}$$

For $i \neq j$, it follows that

$$\partial_j v_i = -v_i c_i \frac{\zeta_j}{Z\lambda_j}, \qquad c_i = \frac{2\kappa e_i}{1-e_i}. \tag{55}$$

Writing $a_i = \zeta_i/(Z\lambda_i v_i)$, the triple coefficients of $\omega_\kappa \wedge d\omega_\kappa$, divided by $v_i v_j v_l$, vanish exactly when every triple (a_i, c_i) is collinear. For $\kappa > 0$, multiply the rows of this determinant by $1 - e_i$ and use $\lambda_i = \tau x_i / P(x_i)$. Invertible column operations show that the equivalent points are

$$(h(x_i), e(x_i)), \qquad h(x) = xP(x)^{1-2\kappa}. \tag{56}$$

Theorem C.2 (Power slow updates). *Let $k \geq 4$ and $\kappa > 0$. Under the analytic branch assumptions above, ω_κ admits a positive local scalar representation if and only if the following holds on I. For $m = 2\kappa - 1 \neq 0$, the positive function P^m has one of the forms*

$$P(x)^m = cx^r + \ell x \quad (r \neq 1), \qquad \text{or} \qquad P(x)^m = x(c + \ell \log x). \tag{57}$$

For $\kappa = 1/2$, the class is

$$P(x) = cx^A e^{Bx}, \qquad c > 0. \tag{58}$$

All coefficients retain the stated positivity and branch conditions.

Proof. Lemma A.4 also shows that τ and q in the positive simplex form local curvature coordinates. Three q coordinates vary independently while a fourth maintains their sum. Hence (56) applies to independently chosen points in three occupied scalar intervals.

If h is nonconstant, analyticity makes it nonconstant on every occupied interval. Choose points in the first two intervals with distinct h values. They determine a nonvertical line. Varying the third point forces

$$e(x) = A + BxP(x)^{1-2\kappa} \tag{59}$$

on an open scalar interval, and analyticity extends the identity over I. If h is constant, then $xh'/h = 1 - me$ gives $e = 1/m$ and $P = cx^{1/m}$, which is already included in (57). Its admissibility requires $1/m \in (0,1)$.

For $m \neq 0$, set $Y = P^m$. Equation (59) becomes $xY' = mAY + mBx$. The nonresonant and resonant solutions are (57). When $m = 0$, (59) is $xP'/P = A + Bx$, yielding (58).

Conversely each displayed family satisfies (59), including its constant-h case. Thus every determinant in (56) vanishes. The coefficients in (53) are positive and degree zero, so Lemma C.1 supplies a local primitive after division by $\sum_i \lambda_i v_i$. The common normalization factor of the slow power preserves this representation. □

Theorem C.3 (Euclidean slow updates). *Under the same assumptions with $k \geq 4$, the averaged Euclidean slow force admits a local scalar objective up to positive speed exactly for*

$$P(x) = cx^r + \ell x \quad (r \neq 1), \qquad \text{or} \qquad P(x) = x(c + \ell \log x). \tag{60}$$

For a pure power $P(x) = cx^\alpha$, $0 < \alpha < 1$, a canonical potential is

$$V(\lambda) = \frac{1}{\nu} \log \sum_i \lambda_i^\nu, \qquad \nu = \frac{1+\alpha}{1-\alpha}. \tag{61}$$

Proof. The averaged Euclidean force is $(a^2/2)(d_i^2)_i$ by Theorem B.1. Its positive constant factor leaves scalar representability unchanged. Apply Theorem C.2 at $\kappa = 1$. For the pure power, the branch relation gives x_i proportional to $\lambda_i^{1/(1-\alpha)}$, so d_i^2 is a common positive factor times $\lambda_i^{\nu-1}$. Dividing by $\sum_i \lambda_i d_i^2$ gives the differential of (61). □

At $\kappa = 0$, the slow power sends every SPD gradient to the identity. Its eigenvalue force is $(1, \dots, 1)$ for every positive fast readout, with potential $\sum_i \lambda_i$. This endpoint has a separate algebraic condition because the factor κ in (55) vanishes.

C.3 Spectral Composition

Corollary C.4 (Two slow exponents)**.** *Under the preceding hypotheses, a fast readout admits scalar curvature objectives for both slow exponents $\kappa = 1$ and $\kappa = 1/2$ if and only if it is a pure power $P(x) = cx^\alpha$. These are exactly the admissible fast readouts that admit such objectives for every $\kappa > 0$.*

Proof. The half-power condition gives $e = A + Bx$, and the identity-power condition gives $e = C + Dx/P(x)$. If $B \neq 0$, then $D \neq 0$ and $x/P(x) = u + vx$ with $v \neq 0$. Differentiating $P = x/(u + vx)$ gives $e = u/(u + vx)$. Equality to $A + Bx$ forces the quadratic coefficient Bv in $(A + Bx)(u + vx) = u$ to vanish, a contradiction. Hence $B = 0$ and P is a pure power. The converse follows because a pure fast power gives

$$v_i \propto \lambda_i^{2\alpha\kappa/(1-\alpha)}, \qquad V = \frac{1}{\nu} \log \sum_i \lambda_i^\nu, \qquad \nu = 1 + \frac{2\alpha\kappa}{1-\alpha}. \tag{62}$$

□

The equivalent fast power under Euclidean slow descent is

$$\phi_\kappa(\alpha) = \frac{\kappa\alpha}{1 + (\kappa - 1)\alpha}, \qquad \phi_\kappa \circ \phi_\mu = \phi_{\kappa\mu}. \tag{63}$$

Pairs sharing this effective exponent follow the same oriented averaged curvature paths.

A polynomial in the Euclidean class has at most one nonzero monomial whose degree differs from one. This follows directly from the necessary identity $xP' = rP + bx$. For the classical cubic $p_c(x) = 3x/2 - x^3/2$, each finite iterate is increasing on $(0, 1)$ and has elasticity in $(0, 1)$, since elasticities multiply under composition. The uniform shape $x_i = k^{-1/2}$ gives an allocation at equal curvature. Its gain decreases, so Lemma A.4 gives a unique stable allocation nearby. One step belongs to (60), whereas

$$p_c(p_c(x)) = \frac{36x - 39x^3 + 27x^5 - 9x^7 + x^9}{16} \tag{64}$$

lies outside it. The cubic and quintic coefficients retain their negative and positive signs under further cubic compositions. Thus the same obstruction holds for every finite number of classical steps at least two, even though the stable allocation remains unique.

D Joint Classification

D.1 Normalized Pairs

Use the joint force in (50) and set

$$a_P(x) = \frac{x}{P(x)}, \qquad \mathcal{S} = \sum_i a_P(x_i) Q(s_i), \qquad \xi = \sum_i a_P'(x_i) Q(s_i)\, dx_i. \tag{65}$$

For $\omega = \sum_i Q(s_i) d\lambda_i$ and $L = \sum_i \lambda_i Q(s_i)$, substitution of $\lambda_i = \tau a_P(x_i)$ gives

$$\omega = \mathcal{S}\, d\tau + \tau\xi, \qquad L = \tau\mathcal{S}, \qquad \omega/L = d\log\tau + \xi/\mathcal{S}. \tag{66}$$

Define $\zeta = \sum_i a'_P(x_i)s_iQ'(s_i)dx_i$. Since

$$ds_i = \frac{(P(x_i)^2)'}{T}dx_i - s_i\, d\log T,$$

the diagonal wedge terms vanish and $d\xi = -d\log T \wedge \zeta$. Lemma C.1 therefore gives the exact criterion

$$d\mathcal{S} \wedge \xi + \mathcal{S}\, d\log T \wedge \zeta = 0 \tag{67}$$

on the tangent space of $\sum_i x_i^2 = 1$. The second term retains the coupling introduced by slow input normalization. For $Q(s) = s^\kappa$, both $\mathcal{S}$ and ξ contain $T^{-\kappa}$, which cancels in their ratio and recovers Theorem C.2.

D.2 Fast Powers

Lemma D.1 (Separable sums)**.** *Let A, B be real analytic on a connected scalar interval. On a full open positive simplex neighborhood with $k \geq 4$, the form*

$$d\sum_i A(q_i) \wedge d\sum_i B(q_i), \qquad \sum_i q_i = 1,$$

vanishes throughout the neighborhood exactly when constants u, v, c, with $(u, v) \neq (0, 0)$, satisfy $uA'(q) + vB'(q) = c$ on the interval.

Proof. Eliminating one simplex coordinate shows that the two-form vanishes exactly when every triple of points $(A'(q_i), B'(q_i))$ is collinear. Three coordinates vary independently with a fourth balancing their sum. If the analytic curve is constant, an affine relation follows immediately. Otherwise choose distinct points from two occupied intervals. Their line contains every point in a third occupied interval. The analytic affine relation then extends over the connected interval. Conversely the displayed relation gives a linear dependence between the two differentials on the simplex, so their wedge vanishes. □

Theorem D.2 (General slow readouts)**.** *Let $k \geq 4$ and $P(x) = c_0x^\alpha$ with $c_0 > 0$ and $0 < \alpha < 1$. Set*

$$p = \frac{2\alpha}{1-\alpha}, \qquad h = 2 - \frac{1}{p}. \tag{68}$$

Let Q be positive and real analytic on a connected interval containing the occupied slow input levels. The normalized joint field admits a local positive scalar representation exactly when

$$Q(s) = c_1s^r + c_2s^h \quad (r \neq h), \qquad \text{or} \qquad Q(s) = s^h(c_1 + c_2\log s). \tag{69}$$

The coefficients retain positivity on the stated domain.

Proof. The fast branch equations give

$$s_i = \frac{\lambda_i^p}{(\sum_j \lambda_j^{2p})^{1/2}}, \qquad \lambda_i = \theta s_i^{1/p}, \qquad \theta = (\sum_j \lambda_j^{2p})^{1/(2p)}. \tag{70}$$

Thus θ and the positive unit sphere coordinates s_i parameterize the full curvature region. Put $a(s) = s^{1/p}$ and choose $b'(s) = a'(s)Q(s)$. With

$$\mathcal{S} = \sum_i a(s_i)Q(s_i), \qquad \mathcal{B} = \sum_i b(s_i),$$

direct differentiation gives

$$\omega/L = d\log\theta + d\mathcal{B}/\mathcal{S}, \qquad d(\omega/L) = -d\mathcal{S} \wedge d\mathcal{B}/\mathcal{S}^2. \tag{71}$$

Apply Lemma D.1 with $q = s^2$. Its affine relation is equivalent to

$$u\, a'(s)Q(s) + v\, a(s)Q'(s) = 2cs, \qquad (u, v) \neq (0, 0). \tag{72}$$

If $v \neq 0$, this is $sQ' + AQ = Bs^{2-1/p}$, whose ordinary and resonant solutions are (69). If $v = 0$, it gives $Q \propto s^{2-1/p}$, already included. Conversely (72) yields $v\, d\mathcal{S} + (u - v)d\mathcal{B} = 0$ on the sphere. Equation (71) is therefore closed, and Lemma C.1 supplies the positive scalar representation. □

The potentials can also be written explicitly. In the first family of (69), let $\mathcal{S} = \sum_i s_i^{1/p} Q(s_i) > 0$. For $r \neq -1/p$,

$$V = \log\theta + \frac{1}{pr+1}\log\mathcal{S}. \tag{73}$$

For $r = -1/p$, $\mathcal{S} = kc_1 + c_2$ is constant and

$$V = \log\theta + \frac{c_1}{p(kc_1+c_2)}\sum_i \log s_i. \tag{74}$$

For the logarithmic family,

$$V = \log\theta + \frac{1}{2p}\log\mathcal{S}. \tag{75}$$

To verify these formulas, integrate $b' = p^{-1}s^{1/p-1}Q(s)$ in (71). In the ordinary case $d\mathcal{B} = (pr+1)^{-1}d\mathcal{S}$. At $r = -1/p$, the nonconstant term in $\mathcal{B}$ is $(c_1/p)\sum_i \log s_i$. For the logarithmic family, $d\mathcal{B} = (2p)^{-1}d\mathcal{S}$. Each displayed potential therefore has $dV = \omega/L$.

Corollary D.3 (Finite slow iterations)**.** *For a pure fast power, one classical cubic slow step admits a scalar objective exactly at $\alpha = 1/3$. Two classical cubic slow steps, the original quintic slow step, and every positive finite composition of that quintic lie outside the class for every $0 < \alpha < 1$.*

Proof. The special exponent $h = 2 - 1/p$ is below two. The powers one and three in the classical cubic can fit (69) only when $h = 1$, which gives $p = 1$ and $\alpha = 1/3$. Two cubic steps have the five nonzero powers in (64). For a quintic $ax + bx^3 + cx^5$ with $a, c > 0$ and $b < 0$, if an iterate begins with coefficients A_1, A_3, A_5, the next coefficients are

$$aA_1, \qquad aA_3 + bA_1^3, \qquad aA_5 + 3bA_1^2A_3 + cA_1^5.$$

Their signs remain positive, negative, and positive. Hence every such iterate contains at least three distinct nonzero powers. A polynomial in (69) has at most two. Equivalently, the necessary Euler equation in (72) forces every monomial except the special degree h to have the same degree. This excludes the stated compositions. □

E Decay and Covariance

E.1 Cube-Root Compatibility

Theorem E.1 (Cube-root compatibility)**.** *Let $k \geq 4$. For a pure fast power $P(x) = c_0x^\alpha$, $0 < \alpha < 1$, the covariance D^2 is a positive scalar multiple of C throughout a full open curvature region exactly when $\alpha = 1/3$.*

At this exponent, let Q be positive and real analytic on a connected interval containing the occupied slow inputs. Adding any fixed decoupled slow decay $\delta \geq 0$ preserves the local scalar-representation class. It consists exactly of

$$Q(s) = c_1s^r + \ell s \quad (r \neq 1), \qquad \text{or} \qquad Q(s) = s(c_1 + \ell\log s). \tag{76}$$

The corresponding full matrix fields have scalar potentials in the Frobenius geometry.

Proof. For a pure fast power, the branch gives d_i^2 proportional to λ_i^p, with $p = 2\alpha/(1-\alpha)$. Proportionality to C on a full open region requires $p = 1$, equivalently $\alpha = 1/3$. At this exponent, set

$$\theta = \left(\sum_i \lambda_i^2\right)^{1/2}, \qquad s_i = \lambda_i/\theta, \qquad \mathcal{S}_Q = \sum_i s_iQ(s_i), \qquad \beta_Q = \sum_i Q(s_i)\,ds_i.$$

The form β_Q is closed because each coefficient depends on its own scalar coordinate. The force with decay is

$$\omega_\delta = (\mathcal{S}_Q + \delta\theta)d\theta + \theta\beta_Q. \tag{77}$$

Consequently

$$d\omega_0 = d\theta \wedge (\beta_Q - d\mathcal{S}_Q), \qquad \omega_\delta \wedge d\omega_\delta = \omega_0 \wedge d\omega_0. \tag{78}$$

The forms have positive coefficients, so the local Frobenius criterion applies. Theorem D.2 at $p = 1$ gives exactly (76).

Explicit integrating factors also give the potentials. For $Q = c_1 s^r + \ell s$, multiplication by $\theta^r > 0$ yields $\theta^r \omega_\delta = dW$, where for $r \neq -1, -2$,

$$W = \frac{c_1}{r+1} \sum_i \lambda_i^{r+1} + \frac{\ell}{r+1} \theta^{r+1} + \frac{\delta}{r+2} \theta^{r+2}. \tag{79}$$

At $r = -1$ and $r = -2$, the respective primitives are

$$\begin{aligned} W_{-1} &= c_1 \sum_i \log \lambda_i + \ell \log \theta + \delta\theta, \\ W_{-2} &= -c_1 \sum_i \lambda_i^{-1} - \ell\theta^{-1} + \delta \log \theta. \end{aligned} \tag{80}$$

For $Q = s(c_1 + \ell \log s)$, multiplication by θ gives the primitive

$$W = \frac{c_1}{2}\theta^2 + \frac{\ell}{2} \left(\sum_i \lambda_i^2 \log \lambda_i - \theta^2 \log \theta \right) + \frac{\delta}{3}\theta^3. \tag{81}$$

Differentiation verifies these identities. The potentials are symmetric spectral functions and lift through (51) to the full SPD matrix. □

The identity (78) applies to each smooth positive Q. It preserves both the vanishing and the nonzero value of its triple form. The relation $D^2 \propto C$ supplies the mechanism by which radial decay preserves this admissibility boundary.

E.2 Power Pairs

For general pure fast and slow powers, set $p = 2\alpha/(1-\alpha) > 0$, $r = \kappa p$, and

$$\mathcal{S} = \sum_i \lambda_i^{2p}, \qquad c(\lambda) = \mathcal{S}^{-\kappa/2}, \qquad F = \frac{1}{r+1} \sum_i \lambda_i^{r+1}, \qquad R_0 = \frac{1}{2} \sum_i \lambda_i^2. \tag{82}$$

The actual force form retains the normalization magnitude,

$$\omega_\delta = c\, dF + \delta\, dR_0. \tag{83}$$

Theorem E.2 (Power pairs with decay)**.** *Let $k \geq 3$, $p > 0$, $\kappa > 0$, and $\delta > 0$. On a full open positive curvature neighborhood, (83) admits a positive local scalar representation exactly when*

$$p = 1, \qquad \text{or} \qquad r = 1, \qquad \text{or} \qquad r = 2p - 1. \tag{84}$$

Proof. Differentiating (83) gives

$$\omega_\delta \wedge d\omega_\delta = \delta\, dR_0 \wedge dc \wedge dF. \tag{85}$$

The coefficient for distinct indices i, j, l is

$$-\delta\kappa p\, \mathcal{S}^{-\kappa/2-1} \det \begin{pmatrix} \lambda_i & \lambda_i^{2p-1} & \lambda_i^r \\ \lambda_j & \lambda_j^{2p-1} & \lambda_j^r \\ \lambda_l & \lambda_l^{2p-1} & \lambda_l^r \end{pmatrix}. \tag{86}$$

For three distinct positive arguments and three distinct real exponents, this determinant is nonzero. To prove the assertion, order the exponents as $a < b < c$, factor the smallest power from each row, and set $t = \lambda^{b-a}$. The remaining columns are $1, t, t^q$ with $q = (c-a)/(b-a) > 1$. Strict convexity of t^q makes three distinct points noncollinear. Every full open curvature neighborhood contains a triple of distinct curvatures, so vanishing throughout the neighborhood requires two of $1, 2p-1, r$ to coincide. These coincidences give (84).

For sufficiency, explicit primitives cover the three cases. At $p = 1$, $\mathcal{S} = 2R_0$ and

$$V = F + \frac{\delta}{\kappa+2} (2R_0)^{\kappa/2+1}, \qquad \omega_\delta = c\, dV. \tag{87}$$

At $r = 1$, $F = R_0$ and $\omega_\delta = (c + \delta)dR_0$, with positive factor. At $r = 2p - 1$, one has $c = \mathcal{S}^{1/(2p)-1}$ and $F = \mathcal{S}/(2p)$, so

$$\omega_\delta = d\left(\mathcal{S}^{1/(2p)} + \delta R_0\right). \tag{88}$$

□

Corollary E.3 (Unique power compatibility)**.** *Among pure fast powers $0 < \alpha < 1$, the exponent $\alpha = 1/3$ is the unique one that admits a scalar objective for every normalized slow power $\kappa > 0$ in the presence of positive slow decay.*

Proof. At $p = 1$, the first condition in (84) holds for every κ. For $p \neq 1$, the remaining conditions permit at most the two values $\kappa = 1/p$ and $\kappa = 2 - 1/p$. The identity $p = 1 \iff \alpha = 1/3$ completes the claim. □

The normalization in (82) enters the relative strength of the two terms in (83). The integrating factors in (87)–(88) apply directly to this mixed-degree field. The homogeneous normalization lemma applies to the degree-zero fields of the preceding sections.

E.3 Polar Endpoint

For $Q \equiv 1$, every SPD slow input has readout I. If E_t has full rank and the initial slow buffer is positive semidefinite, both its EMA and Nesterov input are SPD after the gradient is included. Hence the complete slow update is exactly

$$C_{t+1} = (1 - \gamma\delta)C_t - \gamma I. \tag{89}$$

It is the gradient step for

$$\mathcal{V}(C) = \operatorname{tr} C + \frac{\delta}{2}\|C\|_F^2. \tag{90}$$

This identity holds for every fast readout and follows from positivity of the slow gradient $E_t E_t^\top/2$.

F Finite Iterations

Let the original quintic and its five-fold composition be

$$p_0(x) = \frac{6889x - 9550x^3 + 4063x^5}{2000}, \qquad P = p_0^{\circ 5}. \tag{91}$$

The polynomial $p_0(x)/x$ is a quadratic in x^2 with positive leading and constant coefficients and negative discriminant. Thus every positive input has positive output.

Proposition F.1 (Joint NS5 obstruction)**.** *For fast and slow readouts both given by (91), there is a positive diagonal 3×3 curvature with a normally attracting fast centered branch at $\beta = 19/20$ and $\rho = 99/100$. Its fast update singular values satisfy*

$$\frac{\max_i d_i}{\min_i d_i} < \frac{253}{250}. \tag{92}$$

At zero slow decay, let $u_i = \lambda_i v_i / \sum_j \lambda_j v_j$ and $B_{ij} = \partial u_i / \partial \log \lambda_j$. Then

$$|B_{ij} - B_{ji}| > \frac{4}{5} \quad (i < j), \qquad \frac{\|\operatorname{skew} B\|_F}{\|\operatorname{sym} B\|_F} > \frac{19}{100}. \tag{93}$$

At the same curvature, the coefficient of $d\lambda_1 \wedge d\lambda_2 \wedge d\lambda_3$ in

$$\omega_\delta \wedge d\omega_\delta, \qquad \omega_\delta = \sum_i (v_i + \delta\lambda_i)d\lambda_i,$$

exceeds five for every fixed $\delta \geq 0$.

F.1 Explicit Branch

Proof. Choose the rational squared inputs

$$(q_1, q_2, q_3) = \left(\frac{1179}{12500}, \frac{225361}{500000}, \frac{227479}{500000} \right). \tag{94}$$

Their sum is one. Define $G(q) = P(\sqrt{q})/\sqrt{q}$, $J(q) = P'(\sqrt{q})$, and

$$x_i = \sqrt{q_i}, \qquad \tau = G(q_2), \qquad \lambda_i = \frac{\tau}{G(q_i)}, \qquad w_i = d_i^2 = q_i G(q_i)^2. \tag{95}$$

These identities give (18) exactly.

All values needed below follow from a rational recursion. Put $G_0 = J_0 = 1$ and write $a_0 = 6889/2000$, $b_0 = -191/40$, $c_0 = 4063/2000$. For five iterations use

$$\begin{aligned} z_j &= q G_j(q)^2, \\ G_{j+1}(q) &= G_j(q)(a_0 + b_0 z_j + c_0 z_j^2), \\ J_{j+1}(q) &= J_j(q)(a_0 + 3b_0 z_j + 5c_0 z_j^2). \end{aligned} \tag{96}$$

Thus $G = G_5$, $J = J_5$, and $e_i = J(q_i)/G(q_i)$. The physical pair responses are rational as well,

$$\mu_{ij} = \frac{w_i - w_j}{G(q_i)G(q_j)(q_i - q_j)}. \tag{97}$$

The following open intervals have rational endpoints.

Quantity	Lower endpoint	Upper endpoint
λ_1	0.4611502878	0.4611502879
λ_3	1.0008487807	1.0008487808
e_1	0.9649049895	0.9649049896
e_2	0.9651825278	0.9651825279
e_3	0.6761589706	0.6761589707
μ_{12}	0.0093112827	0.0093112828
μ_{13}	0.0136437729	0.0136437730
μ_{23}	0.8186033344	0.8186033345

Here $\lambda_2 = 1$ exactly. These bounds give every elasticity and every physical pair response in $(0, 1)$. Lemma A.2 and Theorem A.3 prove full matrix normal attraction. The same rational evaluation gives $w_1 < w_2 < w_3$ and $w_3/w_1 < (253/250)^2$, proving (92).

F.2 Force Derivative

For the slow map, put

$$T^2 = \sum_i w_i^2, \qquad q_i^s = \frac{w_i^2}{T^2}, \qquad V_i = w_i G(q_i^s), \qquad \varepsilon_i = \frac{J(q_i^s)}{G(q_i^s)}. \tag{98}$$

The actual slow output is $v_i = V_i/T$, so the common factor cancels from u_i. Define

$$\begin{aligned} \pi_j &= \frac{q_j/(1-e_j)}{\sum_l q_l/(1-e_l)}, \\ K_{ij}^f &= \frac{2e_i}{1-e_i}(\delta_{ij} - \pi_j), \\ K_{ij}^s &= \varepsilon_i \left(K_{ij}^f - \sum_l q_l^s K_{lj}^f \right), \\ u_i &= \frac{\lambda_i V_i}{\sum_l \lambda_l V_l}, \\ B_{ij} &= u_i \left[\delta_{ij} + K_{ij}^s - \sum_l u_l (\delta_{lj} + K_{lj}^s) \right]. \end{aligned} \tag{99}$$

Equation (54) gives $K^f_{ij} = \partial \log w_i / \partial \log \lambda_j$. Differentiating $s_i = w_i/T$ gives the second response $K^s_{ij} = \partial \log v_i / \partial \log \lambda_j$. The final row follows by differentiating the normalized mass u_i.

Applying (96) at the rational inputs q^s_i and then (99) gives

$$\begin{aligned} -0.811121489010424 < B_{12} - B_{21} < -0.811121489010422, \\ 0.811121489010422 < B_{13} - B_{31} < 0.811121489010424, \\ -0.811121489010424 < B_{23} - B_{32} < -0.811121489010422, \end{aligned} \tag{100}$$

and

$$0.03860765916359 < \frac{\|\operatorname{skew} B\|_F^2}{\|\operatorname{sym} B\|_F^2} < 0.03860765916360. \tag{101}$$

The lower endpoint exceeds $(19/100)^2$, proving (93).

F.3 Force and Decay

Define

$$\widetilde{D}_{ij} = \frac{V_j K^s_{ji}}{\lambda_i} - \frac{V_i K^s_{ij}}{\lambda_j} = T(\partial_i v_j - \partial_j v_i), \tag{102}$$

and the rational quantities

$$\begin{aligned} A &= V_1 \widetilde{D}_{23} - V_2 \widetilde{D}_{13} + V_3 \widetilde{D}_{12}, \\ H &= \lambda_1 \widetilde{D}_{23} - \lambda_2 \widetilde{D}_{13} + \lambda_3 \widetilde{D}_{12}. \end{aligned} \tag{103}$$

The decay form is closed, so the desired triple coefficient is

$$\mathcal{F}_\delta = \frac{A}{T^2} + \delta \frac{H}{T}. \tag{104}$$

Rational evaluation gives

$$\begin{aligned} 5.03390636301575 < A/T^2 < 5.03390636301576, \\ 10.38334820176642 < H < 10.38334820176643. \end{aligned} \tag{105}$$

Since $T > 0$, (104) exceeds five for every $\delta \geq 0$.

The displayed decimal endpoints are rational numbers. Evaluating (96) and (99)–(103) with outward rational bounds gives the stated intervals, and comparison of rational squares proves (92), (101), and (105). □

Corollary F.2 (Euclidean curvature response)**.** *At the same fast NS5 branch, define the normalized Euclidean curvature force by*

$$u^0_i = \frac{\lambda_i w_i}{\sum_j \lambda_j w_j}, \qquad B^0_{ij} = \frac{\partial u^0_i}{\partial \log \lambda_j}.$$

Then

$$\frac{\|\operatorname{skew} B^0\|_F}{\|\operatorname{sym} B^0\|_F} > \frac{13}{100}, \tag{106}$$

while the fast update singular-value ratio obeys (92)*.*

Proof. The fast logarithmic response K^f in (99) gives

$$B^0_{ij} = u^0_i \left[\delta_{ij} + K^f_{ij} - \sum_l u^0_l (\delta_{lj} + K^f_{lj}) \right].$$

Evaluation with the same rational bounds yields

$$-1.571840403058 < B^0_{12} - B^0_{21} < -1.571840403057$$

and

$$0.01830095772374 < \frac{\|\operatorname{skew} B^0\|_F^2}{\|\operatorname{sym} B^0\|_F^2} < 0.01830095772375.$$

The lower endpoint exceeds $(13/100)^2$, giving (106). □

G STATISTICAL MODEL

Let $E \in \mathbb{R}^{4\times n}$ and $C \in \mathrm{Sym}_{++}(4)$, with $n \geq 4$. All matrix spaces carry the Frobenius inner product. The random loss and its two gradients are

$$\ell_t(C,E) = \frac{1}{2}\operatorname{tr}(E^\top CE) + \langle \Xi_t, E\rangle, \qquad G_t = CE_t + \Xi_t, \qquad H_t = \frac{1}{2}E_tE_t^\top, \tag{107}$$

where all entries of the matrices Ξ_t are independent $N(0,\sigma^2)$ variables and $\sigma > 0$.

Write $\phi = \Phi_P$ for the normalized spectral map on $4 \times n$ matrices. We assume that ϕ is odd, orthogonally equivariant, homogeneous of degree zero, bounded, and C^2 on the nonzero matrices, with a positive scalar readout on positive occupied singular values. The original NS5 satisfies these assumptions. More generally, an odd polynomial $P(x) = xR(x^2)$ gives

$$\phi(N) = R\left(\frac{NN^\top}{\|N\|_F^2}\right)\frac{N}{\|N\|_F}, \qquad N \neq 0, \tag{108}$$

which also defines the map at deficient rank. Set $\phi(0) = 0$. Compactness of the unit sphere and homogeneity give constants K_j such that

$$\|\phi(N)\|_F \leq K_0, \qquad \|D^j\phi(N)\| \leq K_j\|N\|_F^{-j}, \qquad j = 1,2. \tag{109}$$

The two principal slow readouts on nonzero positive semidefinite matrices are

$$\Psi_1(H) = \frac{H}{\|H\|_F}, \qquad \Psi_2(H) = \frac{H^2}{\operatorname{tr}(H^2)}. \tag{110}$$

We set both maps to zero at zero. Each is positive semidefinite and has operator norm at most one. In this section Ψ denotes either map, and $q(X) = \Psi(XX^\top)$. The complete iteration is

$$\begin{aligned} M_t &= \beta M_{t-1} + (1-\beta)(C_tE_t + \Xi_t),\\ N_t &= \beta M_t + (1-\beta)(C_tE_t + \Xi_t),\\ E_{t+1} &= E_t - \eta\phi(N_t),\\ C_{t+1} &= C_t - \eta\varepsilon\Psi(H_t), \end{aligned} \tag{111}$$

with fixed $0 \leq \beta < 1$, $M_{-1} = 0$, and $E_0 = \sqrt{\eta}X_0$ for a fixed matrix X_0. The scaling of the slow readout removes the common factor in H_t, so $\Psi(H_t) = q(E_t/\sqrt{\eta})$.

For $S < \lambda_{\min}(C_0)/2$, the discrete curvature remains in a fixed compact positive region on $0 \leq \eta\varepsilon t \leq S$. Indeed,

$$C_0 - \eta\varepsilon t I \preceq C_t \preceq C_0. \tag{112}$$

Let $C^{\eta,\varepsilon}(s)$ interpolate these iterates at $s = \eta\varepsilon t$. For independent columns $z_a \sim N(0, C^{-1})$, put

$$H_n = \frac{1}{n}\sum_{a=1}^{n} z_a z_a^\top, \qquad F_{\Psi,n}(C) = \mathbb{E}\Psi(H_n). \tag{113}$$

Theorem G.1 (Statistical curvature limit)**.** *Fix a finite $n \geq 4$ and the coefficients specified above. Let $\bar{C}$ solve*

$$\frac{d\bar{C}}{ds} = -F_{\Psi,n}(\bar{C}), \qquad \bar{C}(0) = C_0. \tag{114}$$

For every $S < \lambda_{\min}(C_0)/2$ and every $r > 0$,

$$\lim_{\varepsilon\downarrow 0}\limsup_{\eta\downarrow 0}\Pr\left[\sup_{s\leq S}\|C^{\eta,\varepsilon}(s) - \bar{C}(s)\|_F > r\right] = 0. \tag{115}$$

The force in (113) is the same for every fast readout satisfying the assumptions above and for every fixed β and $\sigma > 0$.

The following sections establish the diffusion underlying this limit and the averaging of its covariance. The final universality statement follows because the fast readout and momentum enter the invariant covariance through a common scalar, which both slow readouts remove.

H GAUSSIAN RESPONSE

H.1 NOISE FILTER

Introduce a stationary noise buffer on a two-sided Gaussian innovation sequence,

$$U_t = \beta U_{t-1} + (1-\beta)\Xi_t, \qquad \zeta_t = \beta U_t + (1-\beta)\Xi_t. \tag{116}$$

The explicit filter is

$$\zeta_t = (1-\beta^2)\Xi_t + (1-\beta)\sum_{j\geq 1}\beta^{j+1}\Xi_{t-j}. \tag{117}$$

Its coefficients are nonnegative and sum to one. If v_β denotes the variance of each entry, direct summation gives

$$v_\beta = \sigma^2\frac{1-\beta}{1+\beta}(1+2\beta-2\beta^3). \tag{118}$$

For $\ell \geq 1$, the covariance of the same entry at lag ℓ is

$$\sigma^2\frac{1-\beta}{1+\beta}\beta^{\ell+1}(1+\beta-\beta^2). \tag{119}$$

Different entries have independent Gaussian sequences. Writing their normalized correlation as ρ_ℓ, we have $\rho_0 = 1$, $\rho_\ell \geq 0$, and $\sum_\ell \rho_\ell < \infty$. The filter sum also gives

$$v_\beta \sum_{\ell\in\mathbb{Z}} \rho_\ell = \sigma^2. \tag{120}$$

H.2 AVERAGE RESPONSE

Proposition H.1 (Gaussian response identity)**.** *For $\bar{\phi}(G) = \mathbb{E}\phi(G+\zeta_0)$, there is a scalar $a_\beta > 0$ such that*

$$D\bar{\phi}(0)[A] = a_\beta A, \qquad a_\beta = \frac{\mathbb{E}\langle\phi(\zeta_0),\zeta_0\rangle}{4nv_\beta}. \tag{121}$$

The long-run covariance of the centered matrix sequence $\phi(\zeta_t)$ is $b_\beta I_{4n}$, where

$$b_\beta = \frac{1}{4n}\sum_{\ell\in\mathbb{Z}}\mathbb{E}\langle\phi(\zeta_0),\phi(\zeta_\ell)\rangle, \qquad b_\beta \geq a_\beta^2\sigma^2 > 0. \tag{122}$$

The series is absolutely convergent.

Proof. Gaussian convolution is smooth. Differentiating its density gives

$$D\bar{\phi}(0)[A] = v_\beta^{-1}\mathbb{E}[\phi(\zeta_0)\langle\zeta_0, A\rangle].$$

Row and column sign changes make the off-diagonal entries of this linear operator vanish in the matrix-entry basis. Row and column permutations make all its diagonal entries equal. Taking the trace gives (121). For a nonzero input with singular values σ_i,

$$\langle\phi(N), N\rangle = \sum_i \sigma_i P(\sigma_i/\|N\|_F) > 0.$$

The Gaussian input is nonzero almost surely, so $a_\beta > 0$. Equation (109) also gives $\mathbb{E}D\phi(\zeta_0) = a_\beta I$ by Gaussian integration by parts.

The joint distribution of (ζ_0, ζ_ℓ) has the same row and column symmetries. They make each lag covariance of the output a scalar matrix. Expand one output coordinate into Gaussian Hermite chaoses. Its covariance at lag ℓ has the form

$$\sum_{m\geq 1} V_m\rho_\ell^m, \qquad V_m \geq 0, \qquad \sum_m V_m < \infty.$$

The last inequality follows from boundedness of ϕ. Since $0 \leq \rho_\ell \leq 1$ and the correlations are summable, this covariance series is absolutely summable. The first-chaos projection of the full matrix output is $a_\beta\zeta_0$, by (121). Its contribution to the long-run variance of one coordinate is $a_\beta^2 v_\beta\sum_\ell\rho_\ell = a_\beta^2\sigma^2$. All remaining contributions are nonnegative, proving (122). □

H.3 Negative Elasticities

For the original NS5, rational composition and differentiation give

$$-3.5519290 < \frac{P'(1/2)}{2P(1/2)} < -3.5519289. \tag{123}$$

For each fixed index t, the law of large numbers across columns yields

$$\frac{\zeta_t \zeta_t^\top}{\|\zeta_t\|_F^2} \longrightarrow \frac{1}{4} I_4 \quad \text{in probability as } n \to \infty. \tag{124}$$

Continuity of the singular values and of $e(x) = xP'(x)/P(x)$ therefore gives

$$\Pr\left[\max_{1 \le i \le 4} e_i(\zeta_t) < -3\right] \longrightarrow 1. \tag{125}$$

The full matrix derivative explains how this event coexists with (121). At $N = sR$ with $RR^\top = I_4$, write $H = AR + H_\perp$, where $H_\perp R^\top = 0$, and set $d_0 = P(1/2)$ and $e_0 = e(1/2)$. Differentiation gives

$$D\phi(sR)[H] = \frac{d_0}{s}\left[e_0\left(\operatorname{sym} A - \frac{\operatorname{tr} A}{4} I\right) R + \operatorname{skew} A\, R + H_\perp\right]. \tag{126}$$

The radial response is zero, the nine symmetric traceless directions have relative coefficient e_0, and the six skew directions and $4(n-4)$ external directions have coefficient one. Thus

$$\operatorname{div} \phi(N) = \frac{2d_0}{\|N\|_F}(9e_0 + 4n - 10). \tag{127}$$

Gaussian averaging combines all these matrix channels in the scalar restoring response. The positive identity in (121) holds for every fixed width in the model.

I Diffusion Limit

I.1 Matrix Invariance

Let $f_t = \phi(\zeta_t)$ and $\mathcal{F}_t = \sigma(\Xi_j, j \le t)$. The sequence is centered by oddness. Conditional on $\mathcal{F}_t$, the mean of ζ_{t+k} is $\beta^{k+1} U_t$ for $k \ge 1$, and its covariance is bounded below by

$$s_*^2 I_{4n}, \qquad s_* = (1-\beta^2)\sigma > 0. \tag{128}$$

The convolution of a bounded function with such a Gaussian has a uniform Lipschitz constant. Its value at zero mean vanishes by oddness. Hence

$$\|\mathbb{E}[f_{t+k} \mid \mathcal{F}_t]\|_F \le K \beta^{k+1} \|U_t\|_F. \tag{129}$$

The series

$$h_t = \sum_{k \ge 1} \mathbb{E}[f_{t+k} \mid \mathcal{F}_t], \qquad D_t = f_t + h_t - h_{t-1} \tag{130}$$

therefore converges in every finite L^p, with $\|h_t\|_F \le K \|U_t\|_F$, and $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] = 0$. Its partial sums satisfy

$$\sum_{t=1}^{N} f_t = \sum_{t=1}^{N} D_t + h_0 - h_N. \tag{131}$$

This is a martingale approximation with a directly bounded remainder.

For $p > 2$, a union bound gives $\sqrt{\eta} \max_{t \le U/\eta} \|h_t\|_F \to 0$ in probability. Stationarity and ergodicity make the normalized conditional quadratic variation of $\sum_t D_t$ converge to its mean. Its finite p moment gives the conditional Lindeberg condition. Applying the martingale invariance principle to every matrix projection yields

$$\sqrt{\eta} \sum_{t < u/\eta} \phi(\zeta_t) \Rightarrow \sqrt{b_\beta}\, B_u \tag{132}$$

on each fixed interval, where the entries of B are independent standard Brownian motions. The covariance in (122) follows either from (131) or from summing the absolutely convergent lag covariances.

I.2 Remainder Bounds

Set $X_t = E_t/\sqrt{\eta}$ and $A_t = C_t E_t$. Decompose the actual buffer as $M_t = U_t + V_t$, using the stationary auxiliary U from (116). Then

$$V_t = \beta V_{t-1} + (1-\beta)A_t, \qquad N_t = \zeta_t + A_t + \beta(V_t - A_t). \tag{133}$$

With $R_t = V_t - A_t$, the recurrence is $R_t = \beta R_{t-1} - \beta(A_t - A_{t-1})$. Stop on a fixed curvature region and $\|X_t\|_F \le R$. Boundedness of ϕ gives $E_t - E_{t-1} = O(\eta)$, and (111) gives $C_t - C_{t-1} = O(\eta\varepsilon)$. Consequently

$$\|A_t - A_{t-1}\|_F \le K_R\eta, \qquad \mathbb{E}\|R_t\|_F \le K_R\eta + K\beta^t. \tag{134}$$

The geometric term includes $V_{-1} = -U_{-1}$ from the zero actual momentum initialization.

To control the derivatives in (109), a shifted Gaussian $Z \sim N(m, s_*^2 I_d)$, with $d = 4n$, satisfies

$$\sup_m \mathbb{E}\|Z\|^{-r} \le s_*^{-r} 2^{-r/2} \frac{\Gamma((d-r)/2)}{\Gamma(d/2)}, \qquad 0 < r < d. \tag{135}$$

For a direct proof, represent the inverse power as a Laplace integral and use

$$\mathbb{E}e^{-t\|Z\|^2} = (1 + 2s_*^2 t)^{-d/2} \exp\left(-\frac{t\|m\|^2}{1+2s_*^2 t}\right).$$

The resulting integral is maximized at $m = 0$ and evaluates to (135).

In (133), every signal shift is measurable before the fresh innovation Ξ_t. Thus (135) applies uniformly along each Taylor segment. The first derivative controls the signal lag, and the second derivative controls the quadratic remainder. We obtain

$$X_{t+1} - X_t = -\sqrt{\eta} f_t - \eta D\phi(\zeta_t)[C_t X_t] + r_t^\eta, \tag{136}$$

with

$$\mathbb{E}\sum_{t<U/\eta} \|r_t^\eta\|_F \le K_{R,U}\sqrt{\eta} + o(1). \tag{137}$$

The geometric initialization term contributes at most $K\sqrt{\eta}\sum_t \beta^t$. All constants here are chosen after the finite width and the fixed momentum coefficient.

I.3 Derivative Averaging

Let $Z_t^\circ = D\phi(\zeta_t) - a_\beta I$. Its entries lie in L^2 by (135). For any two centered functions of the Gaussian inputs, the maximal-correlation bound is controlled by ρ_ℓ. One way to obtain it is to expand both functions in Hermite polynomials, whose cross-lag factors are ρ_ℓ^m, and then apply Cauchy–Schwarz. Summability therefore gives

$$\mathbb{E}\left\|\sum_{t=j}^{j+L-1} Z_t^\circ\right\|^2 \le KL. \tag{138}$$

Choose $L_\eta = \lfloor \eta^{-1/4} \rfloor$ and freeze $C_t X_t$ at the beginning of each block. On the stopped region, its block oscillation is bounded by $K_R\sqrt{\eta}L_\eta$. The frozen block value is bounded, so (138) controls its product with the block sum even though both depend on the same innovations. Summing block norms yields

$$\mathbb{E}\sup_{u\le U}\left\|\eta\sum_{t<u/\eta} Z_t^\circ[C_t X_t]\right\|_F \le K_{R,U}\left(L_\eta^{-1/2} + \sqrt{\eta}L_\eta + \sqrt{\eta L_\eta}\right) + o(1) \longrightarrow 0. \tag{139}$$

The final term follows from Cauchy–Schwarz on the largest incomplete block and also covers a block containing the stopping index. This argument averages the coefficient while retaining its dependence on the current state.

Proposition I.1 (Coupled diffusion). *For fixed $\varepsilon > 0$, the interpolated $(E_t/\sqrt{\eta}, C_t)$ at $u = \eta t$ converges on the interval supplied by (112) to the solution of*

$$dX_u = -a_\beta C_u X_u\, du + \sqrt{b_\beta}\, dB_u, \qquad dC_u = -\varepsilon q(X_u)\, du, \tag{140}$$

with initial state (X_0, C_0).

Proof. Equations (132), (137), and (139) identify the stopped martingale problem for the first equation. The bounded curvature increments give tightness of C in the space of continuous curves. The leading martingale has constant limiting quadratic variation $b_\beta I_{4n}$, so it supplies the Brownian motion in the joint filtration. Linear-growth drift and the martingale maximal bound give compact containment for X, allowing the stopping radius to increase.

The radial estimate proved in Lemma J.1 gives zero occupation of the origin for each possible limit. The bounded function q is continuous elsewhere, so the interpolated curvature integral converges to the second equation by dominated convergence in a pathwise representation. The same lemma supplies the inverse-radius integrability used for pathwise uniqueness. Indeed,

$$\|q(x) - q(y)\|_F \le K \frac{\|x - y\|_F}{\|x\|_F}, \qquad x, y \neq 0,$$

because q is homogeneous of degree zero and smooth on the unit sphere. Under a common Brownian motion, the difference of two solutions satisfies Grönwall's inequality with integrable coefficient

$$K \left(1 + \|X^{(2)}\|_F + \|X^{(1)}\|_F^{-1}\right).$$

Thus the same initial state gives the same path. Smooth bounded regularizations of q provide existence by compactness and the same zero-occupation argument, completing identification of the limit. □

The drift in (140) is the positive Gaussian response from (121). The leading kick f_t and its corrector in (130) are independent of X and C. Their derivatives with respect to those states are zero, while (139) accounts for the random linear response.

At each fixed $u > 0$, (133) and (134) give $N_{\lfloor u/\eta \rfloor} - \zeta_{\lfloor u/\eta \rfloor} = O_{\mathbb{P}}(\sqrt{\eta})$. The Gaussian matrix is nonzero almost surely, and the boundary of the NS5 event in (125) has Gaussian measure zero. Thus its probability for the actual input converges to the stationary Gaussian probability, connecting the negative elasticities to the same iteration whose limit is (140).

J Covariance Averaging

J.1 Gaussian Equilibrium

For fixed C, the generator of the fast diffusion is

$$\mathcal{L}_C = -a_\beta C X \cdot \nabla_X + \frac{b_\beta}{2} \Delta_X. \tag{141}$$

Its unique invariant law has independent columns

$$X_{\cdot a} \sim N(0, T_\beta C^{-1}), \qquad T_\beta = \frac{b_\beta}{2a_\beta} > 0. \tag{142}$$

The covariance follows by solving $a_\beta(C\Sigma + \Sigma C) = b_\beta I$. Two solutions driven by the same Brownian motion have difference $e^{-a_\beta C u}(X_0 - \widetilde{X}_0)$, so their Frobenius distance contracts by $e^{-a_\beta \lambda_{\min}(C) u}$.

The invariant density is proportional to $\exp[-\operatorname{tr}(X^\top C X)/(2T_\beta)]$, and the generator has the divergence form

$$\mathcal{L}_C f = \frac{b_\beta}{2} \pi_C^{-1} \operatorname{div}(\pi_C \nabla_X f). \tag{143}$$

Thus it is reversible in its Gaussian measure. Since Ψ is homogeneous of degree zero, scaling X by $\sqrt{T_\beta}$ removes T_β from the averaged readout and gives (113).

J.2 Radial Bounds

Lemma J.1 (Radial bounds)**.** *Suppose the adapted curvature obeys $mI \preceq C_u \preceq MI$ on the interval considered in (140). With $d = 4n > 2$, there is a constant K such that, for $0 < h \le 1$,*

$$\varepsilon \int_0^{S/\varepsilon} \Pr(\|X_u\|_F \le h)\, du \le K(Sh^d + \varepsilon h^2). \tag{144}$$

The process has zero occupation of the origin and $\int_0^U \|X_u\|_F^{-1} du < \infty$ *almost surely on every finite interval. Its moments of every fixed positive order are uniformly bounded, and the maximum over* $[0, U]$ *has the moment bounds of a confining radial OU process.*

Proof. For $R_u = \|X_u\|_F^2$, Itô's formula gives

$$dR_u = (b_\beta d - 2a_\beta c_u R_u)\, du + 2\sqrt{b_\beta R_u}\, dW_u, \qquad c_u \in [m, M]. \tag{145}$$

The coefficient c_u is the Rayleigh quotient of C_u in the matrix direction X_u. Comparison for this square-root diffusion bounds R_u below by the radial OU process with constant coefficient M and above by the one with coefficient m, using the same radial Brownian motion and initial radius.

The lower process is the squared norm of a d-dimensional Gaussian OU process. Its covariance is bounded below by a positive multiple of $(1 \wedge u) I_d$, so

$$\Pr(\|X_u\|_F \le h) \le K \min\{1, h^d (1 \wedge u)^{-d/2}\}.$$

Integrating first over $[0, 1]$ and then over $[1, S/\varepsilon]$ gives (144). The same comparison and Gaussian inverse moments yield $\mathbb{E}\|X_u\|_F^{-1} \le K(1 \wedge u)^{-1/2}$, which is integrable at zero. The upper comparison gives the positive moment and maximum bounds. These arguments include zero initial radius. □

J.3 Poisson Equation

Choose a smooth cutoff q_h that vanishes for $\|X\|_F \le h$ and equals q for $\|X\|_F \ge 2h$, with $\|q_h\|_F \le 1$. Write $F_h(C) = \int q_h \, d\mu_C$, where μ_C is (142). The centered Poisson equation has the semigroup solution

$$\mathcal{L}_C \chi_{C,h} = q_h - F_h(C), \qquad \chi_{C,h}(x) = -\int_0^\infty [P_u^C q_h(x) - F_h(C)]\, du. \tag{146}$$

This is the averaging construction of Pavliotis & Stuart (2008). Here the Gaussian transition formula gives the bounds needed to apply it uniformly on the compact curvature region.

In particular, differentiation with respect to the initial matrix gives

$$\|D_X P_u^C q_h\|_\infty \le K \frac{e^{-cu}}{\sqrt{1 - e^{-2cu}}}, \tag{147}$$

which is integrable at zero and infinity. Differentiating the Gaussian mean and covariance with respect to C gives an integrable bound for $D_C(P_u^C q_h - F_h)$. On $u \le 1$, the mean contribution is at most $K\sqrt{u}\|X\|_F$, and the covariance contribution is at most Ku, together with the bounded derivative of F_h. On $u \ge 1$, centering gives a bound $K(1+u)e^{-cu}(1 + \|X\|_F^2)$. Consequently, uniformly in h,

$$\begin{aligned} \|D_X \chi_{C,h}\| &\le K, \\ \|D_C \chi_{C,h}(X)\| &\le K(1 + \|X\|_F^2), \\ \|\chi_{C,h}(X)\|_F &\le K[1 + \log(1 + \|X\|_F)]. \end{aligned} \tag{148}$$

For the last inequality, integrate the Gaussian mixing bound $K \min\{1, e^{-cu}(1 + \|X\|_F)\}$ after a fixed positive initial interval. All estimates use the uniform bound on q_h, so their constants are independent of the cutoff radius.

Proof of Theorem G.1. Proposition I.1 supplies the first limit for every fixed ε. In the variable $s = \varepsilon u$, write the limiting process again as (X_s, C_s). It satisfies

$$dX_s = -\frac{a_\beta}{\varepsilon} C_s X_s \, ds + \sqrt{\frac{b_\beta}{\varepsilon}}\, dB_s, \qquad dC_s = -q(X_s)\, ds.$$

Apply Itô's formula to the smooth corrector in (146). It gives

$$\begin{aligned} \int_0^s [q_h(X_r) - F_h(C_r)]\, dr &= \varepsilon[\chi_{C_s,h}(X_s) - \chi_{C_0,h}(X_0)] \\ &\quad + \varepsilon \int_0^s D_C \chi_{C_r,h}[q(X_r)]\, dr \\ &\quad - \sqrt{\varepsilon b_\beta} \int_0^s D_X \chi_{C_r,h}\, dB_r. \end{aligned} \tag{149}$$

The bounded gradient in (148) controls the martingale maximum by $O_{L^2}(\sqrt{\varepsilon})$. The curvature derivative and the moment bound in Lemma J.1 make the finite-variation term $O_{L^1}(\varepsilon)$. The logarithmic growth of χ and the radial maximum bound make the boundary term $O_{\mathbb{P}}(\varepsilon \log(2+\varepsilon^{-1}))$.

The difference $q - q_h$ is bounded and supported in the ball of radius $2h$. Equation (144) bounds its integrated contribution by $K(Sh^d + \varepsilon h^2)$. The frozen Gaussian density also gives $\sup_C \|F_{\Psi,n}(C) - F_h(C)\|_F \le Kh^d$. Letting h decrease establishes

$$\sup_{s\le S} \left\| \int_0^s [q(X_r) - F_{\Psi,n}(C_r)]\, dr \right\|_F \longrightarrow 0 \quad \text{in probability as } \varepsilon \downarrow 0. \tag{150}$$

Differentiation of the Gaussian density shows that $F_{\Psi,n}$ is smooth on the SPD cone, hence Lipschitz on the compact region. Grönwall's inequality applied to (150) gives $\sup_{s\le S}\|C_s - \bar{C}(s)\|_F \to 0$ in probability. The deterministic bound (112) applies to the discrete, diffusion, and averaged curves, so all estimates hold on the full stated interval. Combining this with the first weak limit and using that the final limit is deterministic gives (115). □

J.4 Momentum and Decay

A fixed slow momentum coefficient $0 \le \beta_s < 1$ gives the same limit. If its buffer starts at zero, divide that buffer by η. It is then an EMA of $X_t X_t^\top/2$. On compact path sets, its difference from the current covariance is bounded by the modulus of continuity of X over the last L steps and a tail $K\beta_s^L$. Choose L proportional to $|\log \eta|$. The first term tends to zero because $\eta L \to 0$, and the tail also vanishes. The initial L updates contribute at most $K\eta\varepsilon L$ to curvature. Apply the same argument to the slow Nesterov input and use the cutoff estimates above to pass through Ψ.

For fast decay $E_{t+1} = (1 - \eta d_f)E_t - \eta\phi(N_t)$ with fixed $d_f \ge 0$, the diffusion drift is $-(a_\beta C + d_f I)X$, and its invariant column covariance is

$$\frac{b_\beta}{2a_\beta}\left(C + \frac{d_f}{a_\beta} I\right)^{-1}. \tag{151}$$

The averaged normalized force is therefore $F_{\Psi,n}(C + d_f I/a_\beta)$. A slow decay factor $1 - \eta\varepsilon d_s$ adds $d_s C$ to the averaged force. These coefficients enter the same compact-region argument, with the interval chosen inside the resulting positive curvature region.

K Wishart Fields

The squared readout in (113) gives

$$F_{2,n}(C) = \mathbb{E}\frac{H_n^2}{\operatorname{tr}(H_n^2)}. \tag{152}$$

Let $R = C^{-1}$ and $S_j = \operatorname{tr}(R^j)$. The sample covariance has mean R, while its fluctuations enter both the numerator and the shared denominator in (152).

K.1 Distribution Response

For any bounded matrix readout Ψ independent of C, differentiation of its Gaussian expectation gives the full matrix identity

$$DF_{\Psi,n}(C)[A] = -\frac{n}{2}\operatorname{Cov}_C\left(\Psi(H_n), \operatorname{tr}(AH_n)\right), \qquad A = A^\top. \tag{153}$$

Indeed, the density of the $4 \times n$ matrix of Gaussian columns has logarithmic derivative

$$\frac{n}{2}\operatorname{tr}(C^{-1}A) - \frac{n}{2}\operatorname{tr}(AH_n).$$

Its expectation is zero, yielding (153). Gaussian moments justify differentiation and show that $F_{\Psi,n}$ is smooth on the SPD cone. Thus cross-curvature response is determined by correlations between the slow readout and the covariance fluctuation.

K.2 First Correction

Lemma K.1 (Differentiable Wishart expansion). *Uniformly on every compact subset of* $\mathrm{Sym}_{++}(4)$,

$$F_{2,n}(C) = \frac{R^2}{S_2} + \frac{1}{n}\mathcal{A}(C) + O_{C^1}(n^{-2}), \tag{154}$$

where

$$\mathcal{A}(C) = \frac{S_1 R}{S_2} - \frac{S_1^2 R^2}{S_2^2} - \frac{8R^4}{S_2^2} + \frac{8S_4 R^2}{S_2^3}. \tag{155}$$

The C^1 estimate covers every symmetric perturbation of C.

Proof. Write $H_n = R + \Delta$, and define

$$\begin{aligned} N_0 &= R^2, \quad N_1 = R\Delta + \Delta R, \quad N_2 = \Delta^2, \\ D_0 &= S_2, \quad D_1 = 2\,\mathrm{tr}(R\Delta), \quad D_2 = \mathrm{tr}(\Delta^2). \end{aligned} \tag{156}$$

The Gaussian fourth moment gives

$$\mathbb{E}\Delta_{ab}\Delta_{cd} = \frac{1}{n}(R_{ac}R_{bd} + R_{ad}R_{bc}). \tag{157}$$

Consequently

$$\begin{aligned} n\mathbb{E}N_2 &= S_1 R + R^2, \quad n\mathbb{E}D_2 = S_1^2 + S_2, \\ n\mathbb{E}(N_1 D_1) &= 8R^4, \qquad\quad n\mathbb{E}D_1^2 = 8S_4. \end{aligned} \tag{158}$$

For the cross term, (157) first gives $\mathbb{E}[\Delta\,\mathrm{tr}(R\Delta)] = 2R^3/n$ and multiplication on the two sides supplies the third identity. Expanding $(N_0 + N_1 + N_2)/(D_0 + D_1 + D_2)$ through quadratic order and substituting (158) yields (155). The contributions R^2/S_2 from the numerator and denominator corrections cancel.

For the differentiable remainder, use a standard Gaussian matrix $G \in \mathbb{R}^{4\times n}$ and write

$$W_n = \frac{1}{n}GG^\top, \qquad H_n = C^{-1/2}W_n C^{-1/2}, \qquad \Gamma_n = W_n - I. \tag{159}$$

The random variable W_n is independent of C. On the fixed event $\|\Gamma_n\|_F < 1/2$, the map

$$\mathcal{F}(C, W) = \frac{(C^{-1/2}WC^{-1/2})^2}{\mathrm{tr}[(C^{-1/2}WC^{-1/2})^2]}$$

has uniformly bounded fourth derivatives in W and their first derivatives in C on the chosen compact curvature set. Gaussian column independence gives

$$\mathbb{E}\Gamma_n = 0, \qquad \mathbb{E}[(\Gamma_n)_a(\Gamma_n)_b(\Gamma_n)_c] = O(n^{-2}), \qquad \mathbb{E}\|\Gamma_n\|_F^4 = O(n^{-2}). \tag{160}$$

The cubic coordinate moment has this order because only sample-index triples with all indices equal survive centering. Retain the cubic Taylor term and bound the fourth-order remainder by (160). Their expectations are both $O(n^{-2})$.

The same expansion applies to $D_C\mathcal{F}$, using the mixed derivative bounds. To control the complementary event, observe that, for nonzero PSD W,

$$\|C^{-1/2}WC^{-1/2}\|_F \asymp \|W\|_F, \qquad \|D_C(C^{-1/2}WC^{-1/2})[A]\|_F \le K\|W\|_F\|A\|_F.$$

The degree-zero output map has derivative of order $\|C^{-1/2}WC^{-1/2}\|_F^{-1}$, so $\mathcal{F}$ and $D_C\mathcal{F}$ are uniformly bounded on all such W. Gaussian quadratic-form exponential moments give $\Pr(\|\Gamma_n\|_F \ge 1/2) \le Ke^{-cn}$ by a union bound over its finitely many entries. The Taylor polynomials on this event are also exponentially small in expectation, by Cauchy–Schwarz and uniform higher moments. The splitting event is independent of C, so the same estimates hold after differentiation. This proves the uniform C^1 expansion. □

K.3 Curvature Obstruction

Theorem K.2 (Finite-width curvature obstruction). *There are a finite integer n_0 and an open neighborhood of $\mathrm{diag}(1,2,3,4)$ on which $F_{2,n}$ has a nonzero Frobenius obstruction for every $n \geq n_0$. In diagonal curvature coordinates, write $r_i = \lambda_i^{-1}$ and $\omega_n = \sum_i (F_{2,n})_i \, d\lambda_i$. Then*

$$\omega_n \wedge d\omega_n = \frac{8}{nS_2^4} dS_1 \wedge dS_2 \wedge d(S_3/3) + O(n^{-2}), \tag{161}$$

uniformly on a compact neighborhood. At $C = \mathrm{diag}(1,2,3,4)$, the coefficient of $d\lambda_1 \wedge d\lambda_2 \wedge d\lambda_3$ has expansion

$$\frac{10616832}{1766100625} \frac{1}{n} + O(n^{-2}). \tag{162}$$

Hence the force is locally nonintegrable up to every positive scalar multiplier in the full SPD Frobenius geometry.

Proof. Orthogonal invariance makes the force diagonal when C is diagonal. Lemma K.1 gives $\omega_n = \omega_0 + n^{-1}\omega_1 + O_{C^1}(n^{-2})$, where

$$\omega_0 = -\frac{dS_1}{S_2}, \qquad \omega_1 = \frac{S_1}{S_2} d\mathcal{L} + \left(\frac{S_1^2}{S_2^2} - \frac{8S_4}{S_2^3} \right) dS_1 + \frac{8}{S_2^2} d(S_3/3), \tag{163}$$

and $\mathcal{L} = \log\det C$. These identities also hold on the full SPD space, since $dS_j[A] = -j\,\mathrm{tr}(R^{j+1}A)$ and $d\mathcal{L}[A] = \mathrm{tr}(RA)$. The leading form has $\omega_0 \wedge d\omega_0 = 0$. In the first correction $\omega_0 \wedge d\omega_1 + \omega_1 \wedge d\omega_0$, the two terms containing $d\mathcal{L}$ cancel. The contribution parallel to dS_1 vanishes, leaving exactly (161).

Since $dS_j = -j \sum_i r_i^{j+1} d\lambda_i$, the coefficient for $i < j < l$ in the leading term is

$$\frac{16 r_i^2 r_j^2 r_l^2 (r_i - r_j)(r_i - r_l)(r_j - r_l)}{nS_2^4}. \tag{164}$$

At the specified diagonal curvature, $S_2 = 205/144$, and substitution for the first three directions gives (162). Its coefficient is positive on a sufficiently small fixed neighborhood. The uniform C^1 remainder then gives one finite n_0 for which the triple form stays nonzero throughout that neighborhood for every $n \geq n_0$.

If a full matrix force had the form $b\nabla_C V$ with $b > 0$, restriction to the diagonal slice would give $\omega_n = b\,dV$. Such a form has zero wedge with its exterior derivative. The strictly positive coefficient therefore proves the stated nonintegrability. □

Together with (125), this yields a finite-width construction for every prescribed probability level. Choose n large enough for both the negative-elasticity probability and Theorem K.2, then apply Theorem G.1 at that fixed n. Positive Gaussian restoring response, attracting Gaussian equilibrium, and the nonzero averaged curvature form all belong to the same original NS5 model.

L Statistical Potentials

L.1 Linear Updates

For $Y \in \mathbb{R}^{4\times n}$, define

$$\begin{aligned} Z(C) &= \int e^{-\mathrm{tr}(Y^\top C Y)/2} \, dY = (2\pi)^{2n} (\det C)^{-n/2}, \\ J(C) &= \int \frac{e^{-\mathrm{tr}(Y^\top C Y)/2}}{\|YY^\top\|_F} \, dY. \end{aligned} \tag{165}$$

Theorem L.1 (Linear statistical force). *For every positive integer n, the normalized linear force satisfies*

$$F_{1,n}(C) = -\frac{2}{Z(C)} \nabla_C J(C). \tag{166}$$

It therefore has the canonical scalar representation

$$F_{1,n} = L_{1,n}\nabla_C V_{1,n}, \qquad V_{1,n} = -\frac{\log J}{2n-1}, \qquad L_{1,n} = \frac{(4n-2)J}{Z} > 0. \tag{167}$$

Proof. For a PSD matrix with at most four nonzero eigenvalues,

$$\frac{1}{2}\|Y\|_F^2 \le \|YY^\top\|_F \le \|Y\|_F^2.$$

Thus the singularity in J has radial order two in dimension $4n > 2$, and Gaussian decay makes its integral finite at infinity. Differentiation with respect to C inserts $-YY^\top/2$, which cancels the singularity and gives

$$\nabla_C J(C) = -\frac{1}{2}\int \frac{YY^\top}{\|YY^\top\|_F} e^{-\operatorname{tr}(Y^\top CY)/2}\, dY.$$

Gaussian moments justify the derivative uniformly on compact SPD sets. Dividing by $Z(C)$ proves (166). The change of variable $Y = c^{-1/2}\widetilde{Y}$ gives $J(cC) = c^{1-2n}J(C)$. Euler's identity then yields

$$\operatorname{tr}(CF_{1,n}) = \frac{(4n-2)J}{Z},$$

and differentiation of $V_{1,n}$ proves (167) on the full matrix space. □

The same calculation applies to a nonnegative scalar weight $w(YY^\top)$ independent of C, with its weighted Gaussian integral and matrix moment locally uniformly finite. Define $J_w(C) = \int w(YY^\top)e^{-\operatorname{tr}(Y^\top CY)/2}\, dY$. Its corresponding force is

$$\frac{1}{Z}\int w(YY^\top)YY^\top e^{-\operatorname{tr}(Y^\top CY)/2}\, dY = -\frac{2}{Z}\nabla_C J_w(C), \tag{168}$$

so the scalar weight enters a potential through the Gaussian partition function.

L.2 Averaged Squares

The unnormalized square has the exact Wishart expectation

$$F_n^{\mathrm{raw}}(C) = \mathbb{E}H_n^2 = \frac{n+1}{n}R^2 + \frac{S_1}{n}R. \tag{169}$$

It admits the potential

$$V_n^{\mathrm{raw}} = \log\det C - (n+1)\log S_1, \qquad F_n^{\mathrm{raw}} = \frac{S_1}{n}\nabla_C V_n^{\mathrm{raw}}. \tag{170}$$

Indeed $\nabla_C \log\det C = R$ and $\nabla_C S_1 = -R^2$. Normalizing this expectation also preserves the same potential,

$$\frac{\mathbb{E}H_n^2}{\mathbb{E}\operatorname{tr}(H_n^2)} = \frac{S_1}{(n+1)S_2 + S_1^2}\nabla_C V_n^{\mathrm{raw}}. \tag{171}$$

The first correction of (171) contains the first two terms of (155). Subtracting it from the samplewise normalized expectation gives

$$F_{2,n} - \frac{\mathbb{E}H_n^2}{\mathbb{E}\operatorname{tr}(H_n^2)} = -\frac{8}{n}\left(\frac{R^4}{S_2^2} - \frac{S_4R^2}{S_2^3}\right) + O_{C^1}(n^{-2}). \tag{172}$$

The second term is parallel to the leading force, while the R^4 term supplies the new dS_3 direction in (161). Thus the obstruction is produced by correlation between the fluctuating numerator and its shared normalization.

L.3 Equal Trace Updates

The linear trace-normalized map $\Psi_{\mathrm{tr}}(H) = H/\operatorname{tr} H$ gives another exact scalar representation. Set

$$J_{\mathrm{tr}}(C) = \int \frac{e^{-\operatorname{tr}(Y^\top CY)/2}}{\operatorname{tr}(YY^\top)}\, dY = \int_0^\infty Z(C + 2tI)\, dt. \tag{173}$$

The second equality follows from $a^{-1} = \int_0^\infty e^{-ta} dt$. The same integrability and differentiation argument gives

$$F_{\mathrm{tr},n}(C) = -\frac{2}{Z}\nabla_C J_{\mathrm{tr}}(C). \tag{174}$$

Both Ψ_{tr} and Ψ_2 have trace one at every nonzero sample. Their averaged flows therefore have the same total curvature derivative $d\operatorname{tr} C/ds = -1$, while their relative curvature allocation has the distinct scalar-representation properties proved above.

At the rank-one sample endpoint $n = 1$, $H^2 = \operatorname{tr}(H)H$ and $\|H\|_F = \operatorname{tr} H$, so the two maps in (110) agree samplewise. At the large-width endpoint, the squared force tends to $C^{-2}/\operatorname{tr}(C^{-2})$, represented by $-dS_1/S_2$. The finite-width correction in (161) identifies how covariance fluctuations change compatibility between these endpoints.

M Spectral Measurements

We measure the NS5 readout at two continued pretraining states of Qwen3-0.6B-Base,[1] following 256 and 1,024 updates along one trajectory, where we evaluate the next background text gradient with the original global clipping and form the Nesterov input

$$N = \beta^2 M_{\mathrm{prev}} + (1-\beta^2)G, \qquad \beta = 0.95.$$

Across layers 0, 14, and 27, we select the attention query and output projections together with the feedforward expansion and contraction projections, yielding 24 matrix and state pairs. For the main text we use the first layer query matrix at state 256 and the final layer expansion matrix at state 1,024, which were both included in the original selection for directional checks.

For each input, we obtain $x_i = \sigma_i/\|N\|_F$ through an FP64 singular value decomposition and evaluate both the polynomial and its derivative through the five original quintic compositions, retaining the sign of $e_i = x_i P'(x_i)/P(x_i)$. We divide the output values by their full spectrum root mean square and display the modes satisfying $\sigma_i/\sigma_{\max} \geq 10^{-5}$, which retains 24,564 of 24,576 modes at their original rank positions. Across the measured matrices, the twelve modes below this threshold carry at most 7.3×10^{-11} of any input's squared Frobenius norm, while each plotted curve retains every resolved mode of its specified matrix and state.

For directional checks, we use the first layer query and final layer expansion matrices at both states, with a rank trend and a fixed sign vector defining two relative amplitude directions. After centering with weights x_i^2, each direction v satisfies $x^\top v = 0$ and $|v_i| \leq x_i$, allowing us to preserve the input norm and singular frames through

$$x(\epsilon) = \frac{x + \epsilon v}{\sqrt{1 + \epsilon^2\|v\|_2^2}}, \qquad N(\epsilon) = \|N\|_F U \operatorname{diag}(x(\epsilon))V^\top.$$

We compare the analytic derivative $U\operatorname{diag}(P'(x_i)v_i)V^\top$ with the centered difference quotient for $\epsilon \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ in FP64 and FP32, giving 80 measurements across the four matrices and two directions. At $\epsilon = 10^{-5}$, we obtain FP64 relative discrepancies of order 10^{-9}, while the full curves show how decreasing perturbations approach the derivative and eventually encounter rounding effects.

N Additional Comparisons

We can inspect the deterministic compatibility results through the update values, elasticities, and curvature response on one stable NS5 branch, where the first two quantities describe its current

[1] https://huggingface.co/Qwen/Qwen3-0.6B-Base

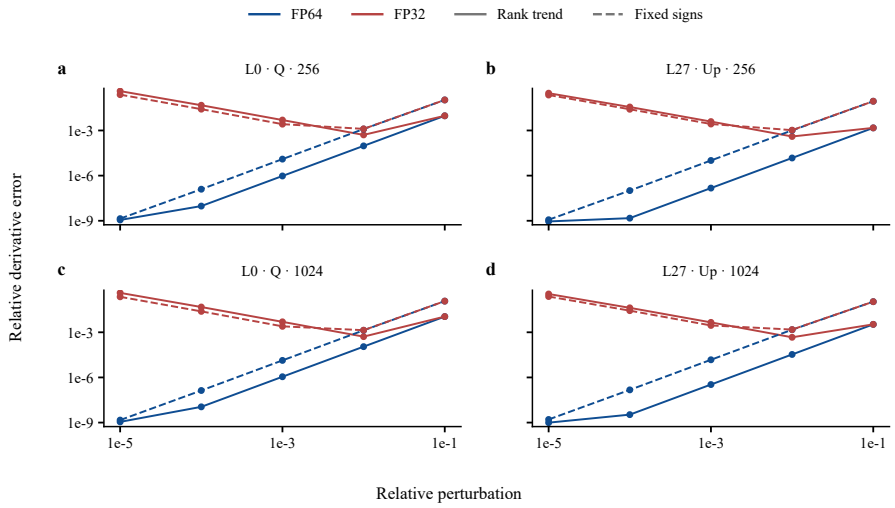


Figure 5: Across four selected inputs, we compare the centered matrix difference quotient with the analytic response as the perturbation decreases, using blue and red for FP64 and FP32 and solid and dashed lines for the rank trend and fixed sign directions. FP64 approaches the derivative prediction across the displayed amplitudes, while the smallest FP32 displacements expose rounding effects.

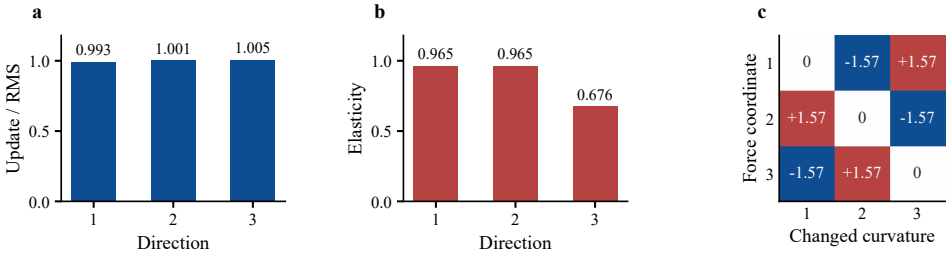


Figure 6: On the stable NS5 branch of Appendix F, nearly equal update values coexist with different elasticities and an asymmetric normalized curvature response under Euclidean slow updates. The final panel gives $B - B^\top$ in log curvature coordinates, with direction indices shared across the panels.

motion and the last describes how that motion changes with curvature. For power updates, varying the two exponents then connects the resulting scalar preference to the covariance alignment that preserves compatibility under decay.

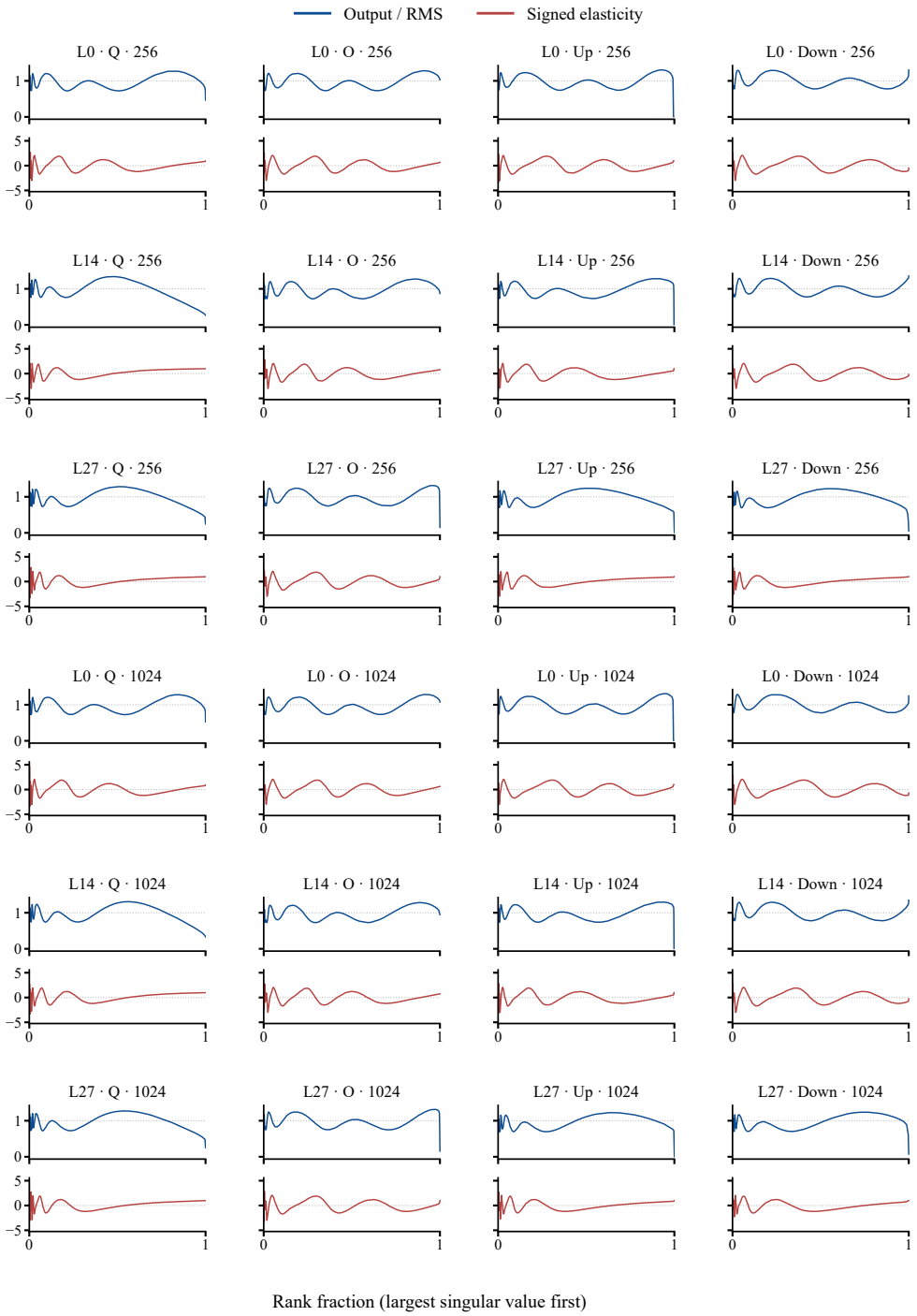


Figure 7: Across all 24 selected matrices and states, we pair the normalized NS5 output in blue with its signed elasticity in red at the same singular value ranks, retaining 24,564 resolved modes. Titles identify the state, layer, projection, and retained count n, with Q and O denoting attention query and output projections and Up and Down denoting feedforward expansion and contraction.

O More Related Work

O.1 Optimization Geometry

When we interpret an optimizer through the information it retains and the displacement it produces, momentum, preconditioning, and regularization become distinct parts of the dynamics. For Adam, first and second moment estimates retain different statistics of past gradients and jointly determine the coordinatewise update (Kingma & Ba, 2014). Initialization and momentum jointly influence the training trajectory (Sutskever et al., 2013), and decoupled weight decay introduces a separate contraction whose combination with adaptive scaling changes regularization (Loshchilov & Hutter, 2019). Keeping these operations explicit leads naturally to our question about the composition of two matrix optimizers, because the covariance produced by one group supplies the gradient processed by the other.

Once the parameters are organized as matrices or tensors, the preconditioner can retain structure across coordinates through Kronecker curvature factors or separate tensor dimensions (Martens & Grosse, 2015; Gupta et al., 2018). In SOAP, Adam operates within the evolving eigenbasis of a Shampoo preconditioner, connecting coordinatewise moment estimates to a matrix geometry (Vyas et al., 2025). Norm-constrained linear minimization oracles and choices of layerwise norm aggregation extend this perspective across spectral optimizers (Pethick et al., 2025; Crawshaw et al., 2025). Our classification follows how the resulting joint geometry transforms a covariance generated by one parameter group into curvature motion in the other.

O.2 Implicit Selection

Because fitting the observations can leave many solutions available, the optimization rule can influence the predictor through its choice among those solutions. Adaptive and nonadaptive methods exhibit this distinction in both explicit examples and neural-network comparisons (Wilson et al., 2017), while geometry-based analyses relate update directions to associated implicit-bias and margin problems (Gunasekar et al., 2018). For homogeneous networks, constrained margin formulations describe the asymptotic selection induced by gradient descent or flow (Lyu & Li, 2019), while analyses of smooth homogeneous models characterize the hybrid margin geometry induced by Muon–Adam parameter groups (Gronich & Vardi, 2026). Spectral descent and Muon also admit a margin analysis on multiclass separable data (Fan et al., 2026). Our curvature-flow representation develops the corresponding selection question for persistent matrix motion followed by a second spectral transformation.

Within matrix factorization, the evolution of singular values offers a concrete route from the training dynamics to preferences for low-complexity solutions (Gunasekar et al., 2017; Arora et al., 2019). Rank-based descriptions become especially useful in examples where the trajectory escapes every bounded norm description (Razin & Cohen, 2020). By classifying the spectral functions for which a curvature force has a positive scalar integrating factor, we identify when one objective describes its relative directions in the fixed Frobenius geometry.

O.3 Curvature Dynamics

When discrete training persists near a stability threshold, its rapid motion can influence a much slower evolution of curvature. Edge-of-stability observations and subsequent analyses connect this behavior to learning-rate-dependent sharpness and the interaction between oscillation and descent (Cohen et al., 2021; Arora et al., 2022). Self-stabilization and central-flow analyses develop mechanisms and averaged descriptions for that interaction (Damian et al., 2023; Cohen et al., 2024). Our full matrix averaging result extends this line of reasoning to a coupled state containing both matrix blocks and both momentum buffers, so the limiting covariance retains its role as the input to another optimizer.

Curvature-directed motion also emerges from stochastic gradients, through the regularization associated with label noise and the limiting evolution along low-loss parameter sets (Damian et al., 2021; Li et al., 2021). Explicit neighborhood objectives, as used by sharpness-aware minimization, provide a complementary route to changing curvature preferences (Foret et al., 2021). Since parameter rescalings can change sharpness while preserving the predictor (Dinh et al., 2017), the geometry used to interpret that preference belongs in the mathematical description. Non-Euclidean stability

analysis further connects update geometry with the relevant sharpness measure (Islamov et al., 2026), reinforcing our treatment of the two spectral maps as joint ingredients of the induced force.

O.4 Statistical Averaging

For temporally correlated gradient inputs, martingale approximation connects accumulated microscopic noise to a diffusion process through control of the maximal approximation error. We use this connection together with a weighted average of the spectral Jacobian, so the restoring drift and accumulated noise arise from the same complete matrix iteration. Once the limiting fast process is identified, the invariant measure and its Poisson equation determine the averaged curvature force through stochastic averaging (Pavliotis & Stuart, 2008). In the Gaussian construction, this force reads a Wishart covariance distribution, allowing its finite fluctuations to enter the reciprocal curvature response explicitly.

O.5 Spectral Updates

Beginning with the orthogonalized momentum and tuned quintic of the original Muon design (Jordan et al., 2024), practical developments have connected spectral geometry to iteration count and parameter grouping. Scaling Muon to large language models brings weight decay and update-magnitude scaling into this design (Liu et al., 2025), while block-periodic orthogonalization addresses its distributed execution (Khaled et al., 2026). For finite iterations, stationarity guarantees quantify the role of approximation depth (Kim & Oh, 2026), and Polar Express adapts successive polynomials through a minimax approximation criterion (Amsel et al., 2026).

When we vary the scalar transformation, spectral powers applied to momentum and RMS-normalized updates already provide a family of Muon-related optimizers (Qi et al., 2026). Associative-memory analysis further relates the update spectrum to balanced learning across frequent and rare classes (Wang et al., 2026). Our measurements place spectral outputs and signed elasticities beside one another at the same Transformer inputs, and our classification follows the response through a second matrix update to determine which pairs admit a scalar curvature objective.

O.6 Structured Matrices

Across sequence models, identifying an operative matrix representation has repeatedly enabled a connection between mathematical structure and practical computation. Attention organizes interactions among positions through a data-dependent matrix (Vaswani et al., 2017), while S4 and Mamba develop structured state-space computations with different forms of temporal dependence and selection (Gu et al., 2021; Gu & Dao, 2023). Structured state space duality relates these viewpoints through semiseparable transformations and leads to Mamba-2 (Dao & Gu, 2024). For our coupled dynamics, the covariance $EE^\top$ connects spectral response to the gradients received by surrounding parameters, while its distribution distinguishes persistent cycles from Gaussian equilibrium.

Memory-aware exact kernels and architectural sensitivity analyses further connect matrix computations to implementation and model design (Dao et al., 2022; Davis et al., 2021). Width limits connect parameterization to feature learning (Yang & Hu, 2021), while the column count in our Gaussian matrix model determines the fluctuations of a covariance sample. Following that distribution through the slow update reveals how matrix dimensions enter its curvature preference through the statistics available to the receiving parameter group.