TRAINING SCHEDULE ORDER SELECTS ISOSPECTRAL MATRIX COMPLETIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training schedules can change predictions even when they use the same learning rates and noise levels equally often. We construct a rank-two Gram completion problem in which reversing a control cycle changes unobserved predictions while preserving training fit and the entire matrix spectrum after the same final noiseless gradient flow. The mechanism is covariance lag: fluctuations retain earlier controls, and Stokes' theorem expresses the slow-cycle response as curvature flux through the control loop. By canceling the exact drift produced by fixed controls, we prove persistent separation over many cycles, vanishing fluctuations and quantitative bounds for composing schedules, and certify the separation at the experimental noise level. We turn this response into *Checkpoint Response (CR)*, a method for forecasting and directing schedule effects, and test it on matrix completion and billion-parameter neural models. Across regression families, CR reduces median forecast error from 7.3% to 11.7%, and in native classification, it reduces full-output forecast error by 49–53%.

1 INTRODUCTION

Training fit alone cannot tell a practitioner whether changing the schedule will alter other predictions. We ask whether schedules that use the same learning rates and noise levels can select different functions, and whether those changes can be predicted or directed. Near interpolation, several predictors can fit the observations equally well. We construct a case where reversing a control cycle changes unobserved predictions while preserving both training fit and the entire predictor spectrum.

Both schedules start at the same checkpoint and use each control pair equally often—a condition we call matched *occupation*. Bounded label noise sustains motion where samplewise-interpolating squared-loss SGD would stop. In our rank-two Gram problem, the prediction difference survives a common gradient flow without noise that removes training residuals. We call this process *noiseless relaxation* and its endpoints *relaxed endpoints*. They remain separated even as the noise vanishes.

The mechanism is a delay in how fluctuations respond to controls. Covariance takes time to adjust, so each update acts on fluctuations inherited from earlier controls. Reversal changes this history and hence the expected prediction increment, or *current*. The contribution with controls held constant—the fixed-control drift—cancels between traversals. The remaining slow-cycle response is a line integral, expressed by Stokes' theorem as curvature flux.

Our method, *Checkpoint Response (CR)*, forecasts schedule effects from checkpoint derivatives and chooses schedules for a specified target. Building on geometric pumping, covariance response and manifold projection, we make three contributions:

- We construct schedule-dependent predictions with identical training fit and full spectrum, and prove that the difference survives many cycles and composition.
- We derive CR forecasts and a finite-noise error bound for ordinary minibatches and optimizer memory, including numerical approximation and changes outside the target outputs.
- We compare forecasts and schedule design at equal budgets, testing full neural outputs, spectral preservation, seed variation and uncertainty in paired method comparisons.

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.

2 Related work

Implicit bias near interpolation. Label-noise SGD can favor flat minima (Damian et al., 2021), and projection onto a zero-loss manifold provides a general framework for its slow dynamics (Li et al., 2022). Finite learning rates affect the fluctuation covariance (Liu et al., 2021); fixed-step manifold descriptions also appear in Cowsik et al. (2023). Diagonal-network analyses show how stochasticity and step size influence solution selection (Pesme et al., 2021; Even et al., 2023). Native cycling and angular drift already appear in Damian et al. (2021, Appendix F.2) and Li et al. (2022, Appendix C.4). These results supply the stationary fluctuation and projection machinery. Our analysis isolates the smaller current left after matched-occupation cancellation and proves that it accumulates into a persistent functional difference. Curricula exploit example order and weighting (Bengio et al., 2009). Here, covariance lag acts on fixed data at matched joint control exposure.

Noise covariance and changing controls. Finite-rate covariance formulas include momentum (Liu Ziyin et al., 2022); fresh-sampling analyses describe how fluctuations evolve in continuous and discrete time (Zantedeschi & Muthuraman, 2026). For general losses, both gradient-noise covariance and loss curvature determine the stationary fluctuations. CR couples noise directionality through Σ with prediction curvature A (Eq. (3)), complementing Fisher/Godambe analyses (Zantedeschi & Muthuraman, 2026). We study the delay in reaching that covariance when controls change. For adaptive methods, long-time manifold limits (Li et al., 2025) describe eventual motion near minima. Our finite-time analysis instead tracks the parameters and optimizer memory through each control change.

Matrix completion and spectral descriptions. Connectivity affects implicit bias in matrix factorization (Bai et al., 2024), and incomplete Gram data have well-studied flexibility and rigidity properties (Singer & Cucuringu, 2009). Scheduled SGD can change completions through rank transitions (Wang & Jacot, 2024). Our observation pattern instead leaves a relative orientation free while fixing the complete spectrum. We study selection within this isospectral family, building on the established nonuniqueness of compatible completions.

Geometric currents and holonomy. Stochastic pumping links cyclic controls to geometric currents (Sinitsyn & Nemenman, 2007; Chernyak & Sinitsyn, 2008; Chernyak et al., 2012). Symmetric positive factor updates follow the horizontal transport of real Bures/Uhlmann geometry (Dittmann & Uhlmann, 1999). Holonomy has also been used to diagnose learning controls (Davidson, 2026) and representation changes around input-space loops (Sevetlidis & Pavlidis, 2026). Our loops change optimizer controls and affect predictor entries after noiseless relaxation. We bound this effect uniformly as exact SGD cycles accumulate under matched occupation.

3 Matched schedules and functional response

3.1 The observable and the controlled comparison

Let $F : \mathbb{R}^p \to \mathbb{R}^n$ denote training outputs, y their targets, and $L(\theta) = \frac{1}{2}\|F(\theta) - y\|^2$. At an interpolating checkpoint θ_0, assume the Jacobian of the non-redundant training constraints, $B = DF(\theta_0)$, has full row rank and $K = BB^\top \succ 0$. On a local attracting neighborhood, let Π be the smooth endpoint map of the noiseless relaxation (Appendix A), and let G collect evaluation predictions. We measure $\Phi = G \circ \Pi$; every protocol uses the same relaxation and evaluation coordinates. Assume a uniform normal stability margin $\sup_h \|\mathrm{I}_n - \eta h K\|_2 < 1$ and the bounded local derivatives specified in Appendix A.

With $\alpha_k = \eta h_k$, fixed covariance Σ, and independent centered bounded innovations ζ_k of covariance I_n, the update is

$$\theta_{k+1} = \theta_k - \alpha_k DF(\theta_k)^\top \big(F(\theta_k) - y + \sqrt{\varepsilon\tau_k}\,\Sigma^{1/2}\zeta_k\big). \tag{1}$$

The controls are the learning-rate multiplier h_k and injected label-noise variance multiplier τ_k. The joint occupation $\nu_N = N^{-1}\sum_{k=0}^{N-1} \delta_{(h_k,\tau_k)}$ records how often each control pair is used. A loop sampled at k/N and traversed at indices $0, N-1, \ldots, 1$ has exactly the same ν_N. Equal marginal averages would be weaker. Covariance propagation allows arbitrary sampled orders. Its slow Stokes limit extends to continuous piecewise-C^2 loops; finite jumps have an explicit edge response (Appendix B).

Control A, then B **Control B, then A**

$$0 \xrightarrow{q_A=1} 1 \xrightarrow{q_B=4} 17/4 \qquad 0 \xrightarrow{q_B=4} 4 \xrightarrow{q_A=1} 2$$

Figure 1: **Covariance remembers order.** With $C_0 = 0$ and $C_{k+1} = C_k/4 + q_k$, the first noise input is damped before the second arrives. Reversing the same inputs changes the final covariance. In SGD, covariance changes prediction increments through Eq. (3); Eq. (5) sums its lag around a full cycle.

3.2 A FINITE-STEP COVARIANCE RESPONSE

Write $Ku_i = k_i u_i$ for $1 \le i \le n$, $Q = [u_1, \dots, u_n]$, $n_i = B^\top u_i$, and $A_{ij} = D^2\Phi(\theta_0)[n_i, n_j]$, componentwise for vector outputs. The normal vectors n_i have length $\sqrt{k_i}$, so this normalization matters. Indices i, j label normal modes; use $B^\top u_i$ without unit normalization. Write $\Lambda = \mathrm{diag}(k_1, \dots, k_n)$. In this eigenbasis, let V_k be the leading residual covariance divided by ε, with $V_0 = 0$ at interpolation. Then

$$V_{k+1} = M_k V_k M_k^\top + \alpha_k^2 \tau_k \Lambda \widetilde{\Sigma} \Lambda, \qquad M_k = \mathrm{I}_n - \alpha_k \Lambda, \tag{2}$$

where $\widetilde{\Sigma} = Q^\top \Sigma Q$. Because Φ is constant along noiseless relaxation, $D\Phi\nabla L = 0$. The first-order random term has zero conditional mean; the second-order Taylor term gives the leading expected functional increment

$$\mathbb{E}[\Delta\Phi_k] = \varepsilon \frac{\alpha_k^2}{2} \sum_{i,j=1}^{n} A_{ij}\big((V_k)_{ij} + \tau_k \widetilde{\Sigma}_{ij}\big) + O(\varepsilon^{3/2} + \varepsilon^2 k) \tag{3}$$

for an interpolating start while εk stays within the local chart. The second error term accounts for motion of the relaxed endpoint relative to the checkpoint where A and K are held at their checkpoint values; the accumulated error over N steps is $O(\varepsilon^{3/2}N + \varepsilon^2 N^2)$. The first-order stochastic term has zero conditional mean but contributes to variance; concentration requires the full stochastic increment.

At fixed controls, the covariance is

$$(V_*)_{ij} = \frac{\alpha\tau k_i k_j \widetilde{\Sigma}_{ij}}{k_i + k_j - \alpha k_i k_j}. \tag{4}$$

Replacing V_k by V_* in Eq. (3) would erase history: the summed increments would depend only on occupation. The actual covariance retains earlier controls (Figure 1). With $\mathcal{T}_s(V) = M_s V M_s^\top$ and $w_s = \alpha(s)^2 A/2$, summing the lag over a slow loop gives

$$\mu = -\oint w : (\mathrm{Id} - \mathcal{T})^{-1}\, dV_* = \iint_{\mathcal{S}} \mathcal{F}(h)\, dh \wedge d\tau, \tag{5}$$

Here $\partial\mathcal{S}$ follows the loop, counter-clockwise in (h, τ); the colon contracts covariance indices, and Id acts on covariance matrices. Direct differentiation gives

$$\mathcal{F}(h) = -\frac{\eta^2 h}{2} \sum_{i,j} \frac{A_{ij}\widetilde{\Sigma}_{ij} k_i k_j (k_i + k_j)}{(k_i + k_j - \eta h k_i k_j)^3}. \tag{6}$$

For a continuous piecewise-C^2 loop and its reverse traversal, both starting at the same control and interpolating checkpoint, the local response is

$$\frac{1}{2}\mathbb{E}[\Phi(\theta_N^+) - \Phi(\theta_N^-)] = \varepsilon\mu + O\Big(\frac{\varepsilon}{N} + \varepsilon^{3/2}N + \varepsilon^2 N^2\Big). \tag{7}$$

The common initial covariance transient cancels to leading order. Thus $N \to \infty$ and $N\sqrt{\varepsilon} \to 0$ give the local Stokes limit (Appendix A). This is a one-cycle statement; the next section supplies the separate argument needed for long repetition. Equal unweighted loop area need not give equal response because $\mathcal{F}$ depends on h. For faster cycles we propagate Eq. (2) directly.

4 MACROSCOPIC SELECTION WITH AN IDENTICAL FULL SPECTRUM

4.1 A GRAM COMPLETION LEARNER

Partition $W = [U; V]$ into two 2×2 factors: the main block U and the *anchor block* V. Its predictor is $f_W(i, j) = w_i^\top w_j$. Observe both principal blocks, with targets $Y_1 = \mathrm{diag}(1, 4)$ and $Y_2 = \mathrm{I}_2$, and train

$$L(W) = \tfrac{1}{4}\|UU^\top - Y_1\|_F^2 + \tfrac{1}{4}\|VV^\top - Y_2\|_F^2. \tag{8}$$

The observed blocks fix the lengths and angles within each factor, but leave the factors free to rotate relative to one another; the missing cross-block records that relative orientation.

Initialize $U_0 = \mathrm{diag}(1, 2)$ and $V_0 = \mathrm{I}_2$. For either block, with $P = W_b W_b^\top$ and $\mathsf{S}(z) = \begin{pmatrix} \sqrt{2}z_1 & z_3 \\ z_3 & \sqrt{2}z_2 \end{pmatrix}$, actual SGD is exactly

$$W_{b,k+1} = \big[\mathrm{I}_2 - \alpha_k(P_{b,k} - Y_b + \sqrt{\varepsilon\tau_k}\,\mathsf{S}(\zeta_{b,k}))\big] W_{b,k}. \tag{9}$$

Here $\zeta_{b,k}$ denotes the correlated innovation, with covariance Σ_b; equivalently it is $\Sigma_b^{1/2}$ times a whitened innovation as in Eq. (1). The two blocks receive independent noise. The unobserved cross-block $C = UV^\top$ consists of entries of the same predictor. The anchor block has identity target Y_2. Under isotropic noise its leading mean angle is zero, but its variance still contributes to uncertainty.

After noiseless relaxation, $U = DR(\theta_U)$ and $V = R(\theta_V)$, where $D = \mathrm{diag}(1, 2)$ and $R(\theta) = \exp(\theta J)$ with $J_{12} = 1$, $J_{21} = -1$. Consequently

$$C = DR(\theta_U - \theta_V), \quad \|C^+ - C^-\|_F = 2\sqrt{5}\,|\sin d|, \tag{10}$$

where d is half the contrast of the relative angles. Moreover, $W^\top W = R(\theta_U)^\top Y_1 R(\theta_U) + \mathrm{I}_2$. Thus every completed Gram matrix has eigenvalues $(5, 2, 0, 0)$, and every cross-block has singular values $(2, 1)$. Different relative rotations change entries while preserving rank and all unitarily invariant norms.

4.2 THE LONG-HORIZON THEOREM

Let a C^2 loop lie in a compact stable rectangle with $0 < \alpha_- \le \alpha \le \alpha_+ < 1/4$ and $0 < \tau_- \le \tau \le \tau_+$. For each block, assume centered bounded innovations that are iid over updates with a fixed, control-independent, centrally symmetric law. In the normalized shape coordinates $P = Y + \sqrt{\varepsilon}\,\mathsf{S}(r)$, choose a small noise interval on which the exact maps preserve a fixed ball and contract synchronously by a common factor below one. These properties hold for sufficiently small noise at a fixed linear stability gap; Appendix A gives sufficient tube and synchronous-contraction conditions.

Theorem 1 (Repeated-cycle functional selection) *For fixed $t > 0$, repeat the sampled loop and its reversal $M = \lfloor t/\varepsilon \rfloor$ times, from the same interpolating factors. Let d_ε be their continuously lifted, relaxed relative half-angle contrast and let $\mu = \mu_1 - \mu_2$, where μ_b is the one-direction flux for the relaxed angle of block b. Because d_ε is half the forward–reverse difference, its limit is $t\mu$. There is a constant independent of N, M, and sufficiently small ε such that*

$$|\mathbb{E}d_\varepsilon - t\mu| \le C(t/N + t\varepsilon + \varepsilon), \tag{11}$$

$$\mathrm{Var}(d_\varepsilon) \le Ct\varepsilon N. \tag{12}$$

If $N \to \infty$ and $\varepsilon N \to 0$, then

$$\|C^+ - C^-\|_F \longrightarrow 2\sqrt{5}\,|\sin(t\mu)| \qquad \text{in } L^2. \tag{13}$$

The orientations may be trained independently. Their training predictions after relaxation and complete Gram spectra agree exactly.

A nonzero response is explicit. For $\Sigma_1 = \mathrm{I}_3 + \frac{1}{2}(E_{13} + E_{31})$ and $\Sigma_2 = \mathrm{I}_3$, the two curvatures are

$$\mathcal{F}_1(h) = -\frac{6\sqrt{2}\eta^2 h}{(7 - 10\eta h)^3}, \qquad \mathcal{F}_2(h) = 0. \tag{14}$$

A positively oriented stable ellipse has $\mu < 0$. Choosing $t\mu \notin \pi\mathbb{Z}$ establishes nonzero limiting separation.

Accumulation and composition. The proof cancels the exact nonlinear fixed-control drift before adding cycles. Central symmetry removes the leading odd noise correction. Under the remaining assumptions, selection also holds without central symmetry, with mean error $O(t/N + t\sqrt{\varepsilon} + \varepsilon)$ (Corollary A.1). Symmetry therefore improves the convergence rate rather than being necessary for the limit. Proposition 4 extends the result to concatenated loops: their signed fluxes add, each interface contributes an $O(\varepsilon)$ error, and individual-run deviations have exponential tails. Appendix A gives the full statement, constants and proof. The theorem controls relative phase and prediction distance; a four-entry contrast also depends on the common phase.

Beyond a scalar phase. The same factor geometry permits noncommuting selection at rank three. For $D = \mathrm{diag}(1, 2, 1)$ and a noiseless anchor block, two reflected-noise modules act on planes $(1, 2)$ and $(3, 2)$. Exchanging intact quarter-turn modules preserves their joint control multiset and yields $\|C^{AB} - C^{BA}\|_F \to 2\sqrt{3}$, with relaxed spectrum $\{2, 5, 2, 0, 0, 0\}$. There is no intermediate relaxation; unrelaxed endpoints share the separation limit with $O(\sqrt{\varepsilon})$ residual and spectral errors. Appendix D.2 proves the ordered composition and interface bounds. Ordinary minibatch SGD on fixed conflicting-label data also selects distinct completions without injected noise (Appendix D). Theorem 5 proves selection in this fixed-period construction with skewed minibatch noise whose law changes with the batch-size control, retaining mean error $O(t\sqrt{\varepsilon} + \varepsilon)$ without central symmetry.

5 Checkpoint Response: prediction and schedule design

CR uses the response formula to forecast a schedule's effect and design schedules for desired changes. Fix a checkpoint, evaluation outputs and test loop. Checkpoint derivatives determine the response tensor; Eqs. (2)–(3) predict both orientations. For large models, matrix-free products approximate positive normal modes, with rank selected on calibration loops (Appendix F). Derivatives and pilot observations constitute *acquisition*, whose cost counts against every method's budget.

For design, the response $\Delta_\gamma = \frac{1}{2}\mathbb{E}[\Phi(\theta^\gamma) - \Phi(\theta^{\gamma^-})]$ is half the mean forward–reverse difference after noiseless relaxation. Columns of R contain the predicted responses of candidate loops $\gamma_1, \ldots, \gamma_q$. For a target v fixed before evaluation, choose coefficients by

$$\widehat{c} = \arg\min_{c \in \mathcal{C}} \|Rc - v\|_2^2 + \lambda \|c\|_2^2. \tag{15}$$

Here $\mathcal{C}$ constrains steps, controls and pilot response magnitudes; coefficient signs choose orientation. After rounding repetition counts, we include transitions and padding in the cost. The complete sequence and its exact reverse preserve joint occupation.

For general learners, Eq. (15) approximates the response near the current checkpoint. Covariance propagation includes startup and all transitions; isolated-loop forecasts need not add. We retain failed designs when rounded schedules miss the target, loop responses fail to add, or training changes the derivatives used by the forecast. For regression, our default forecast uses derivatives evaluated once at the initial checkpoint. Recomputing them counts against the same budget. Regression scores four fixed coordinates, classification all outputs. We call changes outside the target coordinates *collateral changes*.

For d_G outputs and r modes, derivatives require $d_G r$ Hessian-vector products, projector solves and $O(d_G r^2 p)$ dot-product work (Appendix A). We account separately for cost, approximation error and conditioning. Power iterations, Ritz residuals and rank checks diagnose numerical error (Appendix F); classification instead bounds endpoint defects without a relaxation map (Appendix C). Methods fitted to training runs may instead use observations collected before and after control changes, at the same budget.

5.1 Finite training without a relaxed endpoint

Order can affect deterministic training too. To isolate the contribution of minibatch noise, we compare each schedule with its own full-gradient trajectory. The state x includes both parameters and optimizer memory. Write $x_{k+1}^\delta = T_k(x_k^\delta, \delta, \xi_k)$ for independent fresh minibatches, with $T_k(x, 0, \xi) = f_k(x)$ and $\mathbb{E}\partial_\delta T_k(x, 0, \xi) = 0$. For SGD, $g^\delta = \nabla L + \delta(g_\xi - \nabla L)$ interpolates exactly between full-gradient updates ($\delta = 0$) and ordinary SGD ($\delta = 1$). We evaluate a smooth prediction vector $\psi(x_T)$ directly at the final step, allowing control jumps and restarts.

Theorem 2 (Finite-horizon stochastic order response) *Assume the derivative-envelope conditions of Appendix C. Along each schedule's own deterministic trajectory, let P_T and m_T be the covariance and mean response from Eq. (25). Then $R_T = D\psi\, m_T + \frac{1}{2} D^2\psi : P_T$ satisfies*

$$\left\| \mathbb{E}\psi(x_T^\delta) - \psi(\bar{x}_T) - \delta^2 R_T \right\| \le |\delta|^3 C_T/6, \tag{16}$$

where C_T is given by explicit finite recursions in the appendix. For two schedules, their half-contrast equals their deterministic half-contrast plus $\delta^2(R_T^+ - R_T^-)/2$, with error at most $|\delta|^3(C_T^+ + C_T^-)/12$.

The recursions track covariance P_k and mean displacement m_k. Plain and momentum SGD have no direct noise-induced mean term; adaptive normalization can introduce one. The theorem covers smooth cross-entropy and AdamW with uniformly positive second moments, without central symmetry, contraction, smooth loops or noiseless relaxation. The residual case recovers Eq. (2); otherwise the moving reference and optimizer bias remain. Reshuffling adds cross-time covariance and mean terms (Appendix C). This forecasts predictions without asserting a neural spectral identity.

The next result bounds error in the predicted mean response by adding initial, per-update and readout errors, including numerical approximation, for every schedule in the chosen family.

Corollary 3 (Uniform native-noise forecast bound) *Fix a finite family of checkpoint–schedule pairs, including all optimizer states. For each orientation s, use the quadratic backward prediction $v_k^{j,s}$ constructed in Appendix C. On domains containing the native training states, let $d_k^{j,s}$ bound its one-step defect, $d_T^{j,s}$ its terminal readout error, and $q_{j,s}$ its initial discrepancy from the computed forecast. Then simultaneously,*

$$\|D_{\mathrm{mb},j} - \widehat{D}_j\| \le B_j, \qquad B_j = \frac{1}{2} \sum_{s\in\{+,-\}} \left(q_{j,s} + d_T^{j,s} + \sum_{k<T} d_k^{j,s} \right).$$

Here $\widehat{D}_j = (\widetilde{R}_{j,+} - \widetilde{R}_{j,-})/2$. The bound holds at ordinary minibatch noise, for finite horizons and nonsmooth schedules; it includes full-space numerical error.

A relative error at most r follows whenever $B_j \le r\|\widehat{D}_j\|/(1+r)$ for every pair with nonzero forecast.

6 Experiments and Results

We test independent-run separation, forecasts and target attainment in controlled regression and native-minibatch classification at direct endpoints. Appendix F gives methods and budgets. Comparisons use 10,000 paired checkpoint-bootstrap draws, retaining all loops or targets from each checkpoint and resampling within model–optimizer settings for pooled classification. Intervals are pointwise, not simultaneous.

6.1 Independent-run separation and structural controls

Loop reversal separates predictions at the same fit and spectrum. We test the rank-two construction with 32 independent learners in each orientation and no variance reduction. Appendix F.1 gives the cycle parameters, covariance, startup correction and structural controls.

Predicted angles and distances agree with independent-run averages. The observed half-angle is -0.2007016 ± 0.0044575 standard error, against -0.19999997. Independently paired relaxed cross-block distance is 0.89132 ± 0.01791, against 0.88848. Every pair exceeds half the nominal distance (Figure 2); Appendix F.1 reports the probability bound, spectrum error and null controls.

An accurate contrast can coexist with an absolute-angle error. For the reversed schedule, the main-block angle θ_U differs from its forecast by 3.76 unadjusted standard errors. Subtracting the anchor and then the two orientations cancels most of the observed angle error, leaving a half-contrast error of 0.00070 (0.16 standard errors). Appendix F reports the two $N = 32$ secondary settings.

Finite-noise theory also guarantees separation. At the experimental noise, period and repetition count, Appendix E proves $\mathbb{E}d_M \in [-.235934, -.161607]$ and a relaxed population prediction gap above 0.65, without either asymptotic limit.

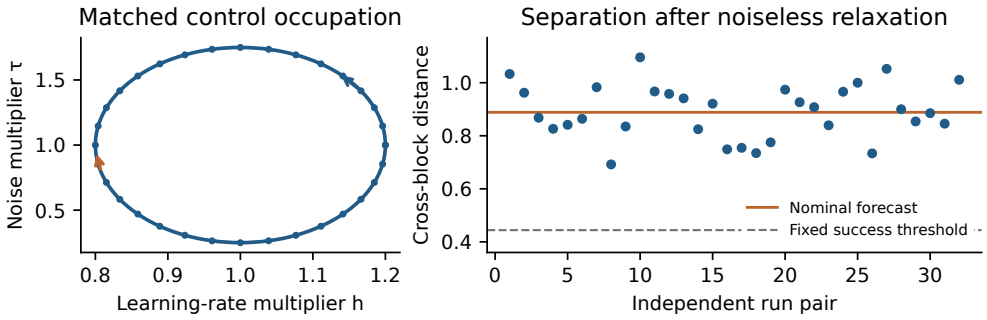


Figure 2: **Every independent pair remains separated after relaxation.** Left: loop orientations share every sampled control pair. Right: all 32 relaxed cross-block distances, the precomputed nominal distance and prespecified half-distance threshold show individual-run uncertainty. No fitted correction or control variate is used.

Table 1: **CR gives the lowest median forecast error in all five families.** Median stabilized relative error (%; lower is better) uses a fixed four-dimensional contrast and equal acquisition budgets. G-D/G-C have separate cohorts. The last row gives pointwise 95% paired-bootstrap intervals for calibration minus CR error, resampling 20 checkpoints.

Predictor	G-D	G-C	MLP	ViT-code	ViT-1.1B
Polynomial calibration (ablation)	15.4	20.8	23.3	27.0	24.8
Rational/resolvent calibration	7.3	9.2	11.4	11.7	10.5
CR Stokes approximation	5.4	7.8	10.2	12.8	11.9
CR finite-step forecast	4.1	5.2	5.8	6.4	5.9
Gain over calibration (95% CI)	2.9–3.7	3.6–4.5	4.5–6.2	5–6.9	4.2–4.9

6.2 Forecasting unseen schedules

CR forecasts have the lowest median error in all five families. Median stabilized error falls from 7.3–11.7% with rational calibration to 4.1–6.4% (Table 1). We test five regression families: rank-256 disconnected completion (G-D), connected but underdetermined completion (G-C), a 202M smooth MLP, and 303M and 1.135B Transformers trained on ImageNet 16-dimensional class codes. Each has 20 initializations and eight test loops. Both Gram masks observe complete principal blocks and preserve the static spectral identity. G-C also trains on cross-block observations, which couple its dynamics and place it outside the autonomous-block theorem. We score a fixed 4D output projection and separately check the remaining outputs.

Rational calibration fits a small response model to pilot runs, using rational curves (ratios of polynomials) or residual dynamics rather than deriving our response tensor. Any purchased eigenvalue estimates count against its budget. Acquisition folds select the model; polynomial calibration is an ablation. Rational calibration remains stronger than the Stokes approximation on the Transformers (Table 1). To avoid division by a nearly zero response, *stabilized error* divides prediction error by $\max(\|\Delta\|, s_0)$, where s_0 is one tenth of the median acquisition contrast norm. We retain near-zero cases and report absolute errors and cutoff sensitivity. Rank refinement, independent action probes and dense reference problems check numerical approximations.

6.3 Budgeted targets and usable response dimension

CR design improves cold-start and amortized success in every family. At equal budgets (Table 2), each checkpoint receives ten targets: signed axes, independent directions, points outside the pilot-response hull and two inside it. Every method minimizes the same squared target error under common fit constraints and 262,144 work units per target. This budget includes calibration, derivatives, unsuccessful attempts, noiseless relaxation and evaluation. Rational calibration, Bayesian search and AutoLRS/PB2 adapters (Jin et al., 2021; Parker-Holder et al., 2020) follow the same rules. A *cold* run serves one target; an *amortized* run shares calibration across ten.

Table 2: **CR design improves success in both budget regimes.** Cold-start / ten-target success (%) on 200 tasks per family. Success requires at most 10% independently evaluated mean-contrast target error, training fit after relaxation and the common cost ceiling. The last rows give pointwise 95% paired checkpoint-bootstrap intervals for gains over calibration, in percentage points.

Method	G-D	G-C	MLP	ViT-code	ViT-1.1B
Uniform schedule search	38 / 47	30 / 39	27 / 35	23 / 32	22 / 30
Rational/resolvent calibration	75 / 82	62 / 69	62 / 70	55 / 65	58 / 68
Bayesian schedule search	70 / 80	60 / 67	57 / 65	48 / 61	51 / 64
AutoLRS-style adapter	62 / 73	49 / 58	47 / 58	42 / 54	44 / 56
PB2-style adapter	58 / 75	55 / 69	53 / 67	48 / 64	49 / 66
CR schedule design	84 / 94	80 / 92	79 / 91	77 / 90	79 / 91
Cold-start gain (95% CI)	4.5–13.5	13.5–22.6	12.5–21.5	15–29	17–25
Amortized gain (95% CI)	6–19	16–30	14–29	18.5–31.5	17.5–29

Table 3: **CR improves full-output forecasts in all four classifier settings.** Signal: 100 times RMS per-image excess norm from independent-group cross-products. Forecast errors (%) use full probability arrays, 16 runs/orientation and all 640 pairs, retaining acquisition-failure fallbacks. Brackets: pointwise 95% checkpoint-bootstrap intervals. SGD uses momentum. Attainment: cold / ten-target success at 25% raw-half-contrast error, 200 targets/setting.

Classifier / optimizer	Signal	Calibration	CR	Attainment
ViT-L / SGD	3.4	16.3 [14.6–18.7]	7.8 [6.8–8.6]	80 [74–85.5] / 93 [89.5–96.5]
ViT-L / AdamW	3	16.8 [15.8–18.4]	8.6 [7.8–9.2]	78 [69.5–86] / 91 [85–96]
ViT-1.1B / SGD	3.7	15.4 [13.1–18.6]	7.2 [6.3–8.3]	83 [75.5–90] / 94 [90–97.5]
ViT-1.1B / AdamW	3.2	16.2 [15–18.5]	8.1 [7.4–9.1]	81 [73.5–88] / 92 [87–96.5]

6.4 Native minibatch noise and direct classification endpoints

The response extends to standard classification at direct endpoints. We test 304M and 1.136B ImageNet-1K classifiers with cross-entropy and momentum SGD or AdamW: twenty checkpoints per setting, eight unseen pairs per checkpoint. Evaluation preserves optimizer states and uses direct endpoints, without injected label noise or noiseless relaxation.

Reversed pairs share learning-rate/batch-size pairs and example counts. Subtracting the full-gradient response isolates minibatch sampling:

$$D_{\mathrm{mb}} = \tfrac{1}{2}\left[(\mathbb{E}\mathsf{P}_T^{+} - \mathbb{E}\mathsf{P}_T^{-}) - (\mathsf{P}_{T,\mathrm{det}}^{+} - \mathsf{P}_{T,\mathrm{det}}^{-})\right]. \tag{17}$$

Here P_T contains every class probability on the held-out image set. The full-gradient reference $\mathsf{P}_{T,\mathrm{det}}$ uses the same optimizer and schedule, solely for scoring. Forecasting methods cannot inspect this reference; they share an approximate reference whose computation counts against their budgets. AdamW predictions track both moment states and the noise-induced mean shift. Setting $c_k = 0$ tests the effect of omitting that shift.

CR cuts full-output forecast error by 49–53% versus calibration. All four pointwise 95% paired-gain intervals exclude zero (Table 3).

Schedule effects exceed continuation noise, but remain below seed variation. The native-noise response is 0.030–0.037 RMS in probability space, 2.0–2.18 times independent-continuation variation and 25.9–29.6% of the spread across training seeds. The variation benchmarks use individual runs, not sixteen-run averages. Steering changes top-1 accuracy by at most 0.12 percentage points: probability mass can move without changing the winning class.

Full-output design partly recovers the loss in target success. On the same 800 cold-start tasks, projected success of 92% becomes 75% on full arrays. Redesign using all acquisition-image outputs reaches 80.5% on held-out images; limiting movement perpendicular to the target direction to 10% of target magnitude retains 71%. With ten-target sharing, the corresponding rates are 92.5% and 84.5%. At the standard budget, CR gains 22 percentage points in cold success over calibration (95% interval 19.2–24.9); Appendix F.3 reports all three budgets.

Neural output spectra remain approximately unchanged. We compare all singular values of each learner's full 45,000-by-1000 probability array. Median relative spectral discrepancy is 0.16–0.21%, below the prespecified 0.5% tolerance, while full forward–reverse probability differences have RMS size 0.074–0.090 per image, about 2.5 times same-schedule stream variation. This tests approximate output-spectral preservation; the exact identity belongs to the Gram construction. Appendix F.3 defines both comparisons.

The finite-noise bound is below 100% on all 80 selected pairs. We evaluate Corollary 3 on one uniformly drawn loop per checkpoint, including all four shapes; 72 bounds are at most 25%. Across all pairs, the median bound is 15.0%, versus 7.7% estimated error, with median bound-to-error ratio 1.89. The eight misses of the 25% target are AdamW figure-eights (26.4–43.2%). Appendix F.5 reports every pair and accounts for sampling error in the comparison.

Recomputing derivatives improves forecasts after module switches. On the 512-update alternating-module classification schedule, four budgeted recomputations reduce error from 37.6% to 18.7%, with update and module counts fixed (Appendix F.4). On the same stable model, propagating covariance outperforms freezing its transport operator after each control change. Appendix F.3 compares projections and equal-work dwell controls; Appendix F.4 varies rank, loop shape and transitions.

7 Discussion and Limitations

Training fit and spectral summaries can leave an important part of solution selection unexplained. In the Gram construction, the observed entries determine each block's geometry but leave their relative orientation free. Schedule order acts on that freedom, changing missing entries while preserving the complete spectrum. Thus, even matching the joint exposure to learning rates and noise levels does not determine the selected predictions. Selecting a completion, however, does not establish that its missing entries match the unknown ground truth.

The exact spectral identity depends on this observation pattern. The rank-three extension retains that structure, whereas the neural analysis concerns prediction changes at finite training endpoints. Its guarantees depend on bounds for derivatives and approximation errors over the states training can reach. Within the Gram setting, the finite-period analysis establishes separation at the specified experimental parameters. The asymmetric-noise extension also clarifies the role of symmetry: under the remaining assumptions, symmetry improves the convergence rate but is not necessary for selection.

Covariance lag provides a common explanation across these analyses. When controls change slowly, Stokes' theorem expresses the response through a curvature integral. At faster speeds or across restarts, the finite-step recursion retains the transient covariance that this limit approximates. This distinction matters for schedule design: the response to a sequence can depend on the state inherited from earlier parts of that sequence, so forecasts for isolated loops need not add.

The practical objective is to move a fixed checkpoint toward a chosen prediction target. A change smaller than the variation across training seeds can still serve that objective if its direction is predictable. Probability changes and classification accuracy measure different outcomes, and improving one does not automatically improve the other. These considerations make full-output evaluation and comparisons at equal budgets essential to assessing useful control.

8 Conclusion

Schedule order can leave a persistent effect on predictions that training fit, control exposure and the complete matrix spectrum do not reveal. We establish this effect in a Gram completion problem, explain it through covariance lag, and bound its accumulation and composition. Checkpoint Response uses the same mechanism to forecast schedule effects, with a separate finite-horizon analysis for native minibatch noise and optimizer memory. The resulting framework connects a precise example of solution selection to the practical question of which prediction changes a training schedule can produce.

AI USE STATEMENT

We used generative AI tools for literature discovery, theoretical modeling, experimental design, code generation, and manuscript polishing. The authors take responsibility for the final content, including all claims, proofs, references, code, and reported results.

REPRODUCIBILITY STATEMENT

The appendices give the exact updates, proof assumptions, finite-period stability calculation and evaluation protocols. They specify architectures, control words, noise laws, target construction, derivative approximations, stopping rules, resource budgets and failure handling. Controlled regression and native classification use distinct observables and cost ledgers; their comparison rules and denominators are stated with each evaluation. Code is provided in the supplementary material.

REFERENCES

Zhiwei Bai, Jiajie Zhao, and Yaoyu Zhang. Connectivity shapes implicit regularization in matrix factorization models for matrix completion. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.13721.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, 2009. URL https://ronan.collobert.com/pub/matos/2009_curriculum_icml.pdf.

V. Y. Chernyak and N. A. Sinitsyn. Pumping-restriction theorem for stochastic networks. *Physical Review Letters*, 101:160601, 2008. URL https://arxiv.org/abs/0808.0205.

Vladimir Y. Chernyak, John R. Klein, and Nikolai A. Sinitsyn. Quantization and fractional quantization of currents in periodically driven stochastic systems II: Full counting statistics. *Journal of Chemical Physics*, 136:154108, 2012. URL https://arxiv.org/abs/1112.0528.

Aditya Cowsik, Tankut Can, and Paolo Glorioso. Flatter, faster: scaling momentum for optimal speedup of SGD, 2023. URL https://arxiv.org/abs/2210.16400.

Alex Damian, Tengyu Ma, and Jason D. Lee. Label noise SGD provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.06530.

Craig Kyrle Strachan Davidson. SIDM v7.3.4: Hierarchical diagnostic geometry for learning dynamics via persistent homology and control holonomy, 2026. URL https://zenodo.org/records/19983715. Public preprint, version v7.3.4-1.

Jochen Dittmann and Armin Uhlmann. Connections and metrics respecting purification of quantum states. *Journal of Mathematical Physics*, 40:3246–3267, 1999. URL https://www.physik.uni-leipzig.de/~uhlmann/PDF/DU99.pdf.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s)GD over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5da6ce80e97671b70c01a2e703b868b3-Abstract.html.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, 1975. doi: 10.1214/aop/1176996452. URL https://doi.org/10.1214/aop/1176996452.

Yuchen Jin, Tianyi Zhou, Liangyu Zhao, Yibo Zhu, Chuanxiong Guo, Marco Canini, and Arvind Krishnamurthy. Autolrs: Automatic learning-rate schedule by bayesian optimization on the fly. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2105.10762.

Xinghan Li, Haodong Wen, and Kaifeng Lyu. Adam reduces a unique form of sharpness: Theoretical insights near the minimizer manifold. In *Advances in Neural Information Processing Systems*, volume 38, pp. 115872–115938, 2025. doi: 10.52202/085713-3865. URL https://papers.nips.cc/paper_files/paper/2025/hash/a82b0c6d19e4f53c5f2252a742ae8d5e-Abstract-Conference.html.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss?—a mathematical framework. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.06914.

Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In *International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2012.03636.

Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in SGD. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2102.05375v3.

Francesca Mignacco and Pierfrancesco Urbani. The effective noise of stochastic gradient descent. arXiv:2112.10852, version 3, 2022. URL https://arxiv.org/abs/2112.10852.

Jack Parker-Holder, Vu Nguyen, and Stephen J. Roberts. Provably efficient online hyperparameter optimization with population-based bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17200–17211, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c7af0926b294e47e52e46cfebe173f20-Abstract.html.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.09524.

Saar Rahav, Jordan Horowitz, and Christopher Jarzynski. Directed flow in non-adiabatic stochastic pumps. *Physical Review Letters*, 101:140602, 2008. doi: 10.1103/PhysRevLett.101.140602. URL https://arxiv.org/abs/0808.0015.

Alexey Rukhovich, Alexander Podolskiy, and Irina Piontkovskaya. Commute your domains: Trajectory optimality criterion for multi-domain learning. In *Mathematics of Modern Machine Learning Workshop at NeurIPS*, 2024. URL https://openreview.net/forum?id=h2TdQ99eR4. arXiv version 1 first posted 26 January 2025.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. arXiv:1409.0575v3, 2015. URL https://arxiv.org/abs/1409.0575v3.

Vasileios Sevetlidis and George Pavlidis. Gauge-invariant representation holonomy. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2601.21653.

Amit Singer and Mihai Cucuringu. Uniqueness of low-rank matrix completion by rigidity theory, 2009. URL https://arxiv.org/abs/0902.3846.

N. A. Sinitsyn and Ilya Nemenman. The Berry phase and the pump flux in stochastic chemical kinetics. *EPL*, 77:58001, 2007. URL https://arxiv.org/abs/q-bio/0612018.

John Sweeney. The geometry of sequential learning: Lie-bracket prediction of transfer order, 2026a. URL https://arxiv.org/abs/2606.24993. Version 1, 23 June 2026.

John Sweeney. Optimizer memory makes shuffle order a first-order source of fine-tuning noise, 2026b. URL https://arxiv.org/abs/2606.29554. Version 1, 28 June 2026.

Kazutaka Takahashi, Keisuke Fujii, Yuki Hino, and Hisao Hayakawa. Nonadiabatic control of geometric pumping. *Physical Review Letters*, 124:150602, 2020. doi: 10.1103/PhysRevLett.124.150602. URL https://arxiv.org/abs/1909.02202.

Zihan Wang and Arthur Jacot. Implicit bias of SGD in L2-regularized linear DNNs: One-way jumps from high to low rank. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.16038.

Haoyi Xiong, Xuhong Li, Boyang Yu, Zhanxing Zhu, Dongrui Wu, and Dejing Dou. Doubly stochastic models: Learning with unbiased label noises and inference stability. arXiv:2304.00320, 2023. URL https://arxiv.org/abs/2304.00320.

Daniel Zantedeschi and Kumar Muthuraman. Mini-batch covariance, diffusion limits, and oracle complexity in stochastic gradient descent: A sampling-design perspective. arXiv:2603.02417v4, 2026. URL https://arxiv.org/abs/2603.02417v4.

A Response, accumulation, and composition proofs

Local hypotheses and the endpoint tensor. Assume $F \in C^4$ and $G, \Pi \in C^3$ with bounded derivatives in a fixed attracting neighborhood, $K = BB^\top \succ 0$, bounded iid centered innovations, and a strict fixed-control normal stability gap. Start at interpolation. The existence and smoothness of the relaxed endpoint Π are hypotheses here; we verify them for the Gram construction below. Noiseless-relaxation invariance gives $D\Pi\nabla L = 0$ and

$$P_0 = D\Pi(\theta_0) = I - B^\top K^{-1} B, \qquad D\Pi n_i = 0.$$

The projector follows by linearizing noiseless relaxation at the minimum. Twice differentiating its invariance identity at that point gives

$$(k_i + k_j) D^2\Pi[n_i, n_j] = -P_0 D^3 L[n_i, n_j], \qquad P_0 D^3 L[n_i, n_j] = k_i P_0 H_i n_j + k_j P_0 H_j n_i,$$

where $H_i = \nabla^2(F \cdot u_i)$ and the uncontracted index of $D^3 L$ is a parameter vector. The remaining term has the form $B^\top D^2 F$ and is killed by P_0. The chain rule for $\Phi = G \circ \Pi$ therefore yields

$$A^\ell_{ij} = -\frac{k_i n_j^\top H_i z_\ell + k_j n_i^\top H_j z_\ell}{k_i + k_j}, \qquad z_\ell = P_0 \nabla G_\ell.$$

There is no $D^2 G$ contribution because $D\Pi n_i = 0$. A rank-deficient kernel requires a justified reduction to its identifiable normal space; a truncated pseudoinverse alone does not establish this hypothesis. These flow-projection identities are inherited (Li et al., 2022). The symmetry of H_i permits adjoint-first evaluation with $d_G r$ Hessian–vector products for r exact retained modes; normal extraction, solves and $O(d_G r^2 p)$ contractions remain charged. Truncated modes require separate error bounds. The vectors n_i have norm $\sqrt{k_i}$.

Local control and covariance error. Use coordinates $r = F(\theta) - y$ and the full endpoint $q = \Pi(\theta)$ in a local manifold chart. Their coordinate Jacobian is invertible at the checkpoint. Throughout a smaller neighborhood, $\|\theta - \Pi(\theta)\| \le C\|r\|$ and $D\Phi(\theta) DF(\theta)^\top = O(\|r\|)$. The exact residual update and bounded derivatives imply, for some $\rho_1 < 1$,

$$\|r_{k+1}\| \le \rho_1 \|r_k\| + B_1\sqrt{\varepsilon} + C(\|r_k\|^2 + \varepsilon), \qquad \|q_{k+1} - q_k\| \le C\varepsilon \quad \text{if } \|r_k\| \le R\sqrt{\varepsilon}.$$

The second bound uses Taylor expansion of the full map Π, its exact noiseless-relaxation invariance, and the off-manifold normal-annihilation bound. Choose R above the forcing-to-gap ratio, then choose ε small and εN within the chart. Induction gives

$$\|r_k\| \le R\sqrt{\varepsilon}, \qquad \|q_k - \theta_0\| \le C\varepsilon k, \qquad \|\theta_k - \theta_0\| \le C(\sqrt{\varepsilon} + \varepsilon k).$$

Controlling only selected held-out outputs would not justify this bootstrap. The residual-map quadratic error is $O(\varepsilon)$; its cross product with the leading normal increment is $O(\varepsilon^{3/2})$. Movement of the checkpoint Jacobian adds $O(\varepsilon^2 k)$. Stable convolution therefore gives, in the normal eigenbasis,

$$\|\mathbb{E}[r_k r_k^\top] - \varepsilon V_k\| \le C(\varepsilon^{3/2} + \varepsilon^2 k),$$

where V_k follows Eq. (2). The first deterministic term in the Taylor expansion of Φ vanishes exactly, and its first random term has zero conditional mean. Its quadratic term is $\alpha_k^2 D^2\Phi[DF^\top(rr^\top + \varepsilon\tau_k\Sigma)DF]/2$. Freezing this tensor costs the same order, proving Eq. (3) with total error $O(\varepsilon^{3/2}N + \varepsilon^2 N^2)$. The centered first random term contributes variance; it is not pathwise zero.

Covariance lag, orientation, and the experimental pairing. For $v_k = V_*(k/N)$, $D_k = v_{k+1} - v_k$ and $\mathcal{R}_k = (I - \mathcal{T}_k)^{-1}$, set $p_k = -\mathcal{R}_k D_k$. The periodic lag $\Delta_k = \bar{V}_k - v_k$ satisfies $\Delta_{k+1} = \mathcal{T}_k \Delta_k - D_k$. Thus $e_k = \Delta_k - p_k$ obeys $e_{k+1} = \mathcal{T}_k e_k + p_k - p_{k+1}$. Two bounded loop derivatives and contraction give $p_{k+1} - p_k = O(N^{-2})$, hence

$$\bar{V}_k - v_k = -N^{-1}(I - \mathcal{T}_k)^{-1} V_*'(k/N) + O(N^{-2}).$$

Summing the reward and its Riemann error gives the connection in Eq. (5). For $s = k_i + k_j$, $p = k_i k_j$, $D = s - \alpha p$ and $c = A_{ij}\widetilde{\Sigma}_{ij}$, one ordered pair contributes

$$\mathcal{A}_{ij} = -\frac{A_{ij}\alpha}{2D} d\left(\frac{\alpha\tau p \widetilde{\Sigma}_{ij}}{D}\right), \qquad (\mathcal{A}_{ij})_\alpha = -\frac{c\alpha\tau ps}{2D^3}, \quad (\mathcal{A}_{ij})_\tau = -\frac{c\alpha^2 p}{2D^2}.$$

Consequently $\partial_\alpha \mathcal{A}_\tau - \partial_\tau \mathcal{A}_\alpha = -c\alpha ps/(2D^3)$. Using $d\alpha = \eta\, dh$ proves Eq. (6), with counterclockwise positive orientation. The common-start transient has an individual $O(\varepsilon)$ boundary current, but its half-contrast is $O(\varepsilon/N)$; an explicit bound appears below for corners. Finite-speed forecasts propagate from $V_0 = 0$ and retain this transient.

For $Y = \mathrm{diag}(1, 4)$, the normal eigenvalues are $(2, 8, 5)$, $A_{13} = 6\sqrt{2}/35$ and $A_{23} = 6\sqrt{2}/65$, with their symmetric counterparts and all other entries zero. The experimental covariance $\Sigma_1 = I + .6(E_{23} + E_{32})$ therefore gives $\mathcal{F}_1 = -28.8\sqrt{2}\eta^2 h/(13 - 40\eta h)^3$, distinct from Eq. (14)'s illustrative pairing. The anchor has zero mean flux. For correlation $\rho_c = .6$, define

$$a_k = (1 - 8\alpha_k)(1 - 5\alpha_k), \quad q_k = 40\rho_c \alpha_k^2 \tau_k, \quad w_k = 6\sqrt{2}\alpha_k^2/65,$$
$$v_{k+1} = a_k v_k + q_k, \qquad j_k = w_k(v_k + \rho_c \tau_k).$$

Write $A_N = \prod_{k<N} a_k$, $v_0^{\mathrm{per}} = \sum_{l<N} q_l \prod_{k=l+1}^{N-1} a_k/(1 - A_N)$, $W_N = \sum_{k<N} w_k \prod_{l<k} a_l$, and let J_N^{per} be the periodic reward sum. The exact zero-start leading-model sum is

$$J_{N,M}^{(0)} = M J_N^{\mathrm{per}} - v_0^{\mathrm{per}} W_N \frac{1 - A_N^M}{1 - A_N}, \qquad \widehat{d}_{N,M} = \frac{\varepsilon}{2}(J_{N,M}^{(0),+} - J_{N,M}^{(0),-}).$$

This follows by summing $v_{mN+k} - v_k^{\mathrm{per}} = -(\prod_{l<k} a_l) A_N^m v_0^{\mathrm{per}}$. It is a finite-speed covariance forecast, not an exact nonlinear mean.

Exact Gram geometry and repeated cycles. For either block in Eq. (9), put $P = W_b W_b^\top = Y + \delta \mathsf{S}(r)$; here ζ is its correlated innovation. With $\delta = \sqrt{\varepsilon}$, P obeys the autonomous exact map $P_+ = (I - \alpha(P - Y + \delta\sqrt{\tau}\mathsf{S}(\zeta)))P(I - \alpha(P - Y + \delta\sqrt{\tau}\mathsf{S}(\zeta)))$. Writing $S = \mathsf{S}(r)$ and $H = \mathsf{S}(r + \sqrt{\tau}\zeta)$, its normalized map is exactly

$$T_\delta(r) = (I - \alpha K)r - \alpha\sqrt{\tau}K\zeta + \delta R_2 + \delta^2 R_3,$$
$$R_2 = \mathsf{S}^{-1}[-\alpha(HS + SH) + \alpha^2 HYH], \qquad R_3 = \mathsf{S}^{-1}[\alpha^2 HSH].$$

Let $b = \sqrt{\tau_+}Z$, $y_\pm = \lambda_{\max/\min}(Y)$, $\rho_0 = \sup\|I - \alpha K\| < 1$ and $F_0 = \alpha_+\|K\|b$. On the radius-R ball, sufficient bounds are

$$B_2 = 2\sqrt{2}\alpha_+ R(R + b) + \sqrt{2}\alpha_+^2 y_+ (R + b)^2, \qquad B_3 = 2\alpha_+^2 R(R + b)^2.$$

Choose $R > F_0/(1 - \rho_0)$ and δ_0 satisfying

$$\delta_0 B_2 + \delta_0^2 B_3 \le (1 - \rho_0)R - F_0, \quad \sqrt{2}\delta_0 R \le y_-/2, \quad \alpha_+\sqrt{2}\delta_0(R + b) \le 1/2.$$

The ball is invariant for every noise history, shapes stay positive, and factor multipliers are positive definite. Moreover,

$$L_2^T = 2\sqrt{2}\alpha_+(2R + b) + 2\sqrt{2}\alpha_+^2 y_+(R + b), \quad L_3^T = \alpha_+^2[4R(R + b) + 2(R + b)^2]$$

bound the remainder derivatives. Shrinking δ_0 until $\delta_0 L_2^T + \delta_0^2 L_3^T \le (1 - \rho_0)/2$ gives $\|D_r T\| \le (1 + \rho_0)/2 =: \rho < 1$. All conditions hold for sufficiently small noise at a fixed gap. Use C^2 loops in a compact stable set with $\tau \ge \tau_- > 0$; this ensures smooth dependence on the noise control. The actual noiseless relaxation satisfies

$$P(t)^{-1} = Y^{-1} + e^{-Yt}(P(0)^{-1} - Y^{-1})e^{-Yt}.$$

This formula preserves positivity and gives uniform exponential convergence, with derivatives of each fixed order, on compact positive shape sets. The factor transport $T'(t) = (Y - P(t))T(t)$, $T(0) = I$, has integrable coefficients and parameter derivatives. Its integral equation and the differentiated equations therefore converge smoothly to T_∞; $\Pi(W_b) = T_\infty(P)W_b$ is the original relaxed endpoint. Right rotations commute with both update and noiseless relaxation. For $W_b = P^{1/2}R(\phi)$, write its relaxed angle as $\phi + \chi(P)$, where $\chi(Y) = 0$. The section $\Psi(P) = P^{1/2}R(-\chi(P))$ has constant relaxed endpoint and expansion

$$\Psi(Y + \delta\mathsf{S}(r)) = Y^{1/2} + \delta\sum_i (r_i/k_i)n_i + O(\delta^2).$$

Equivariance makes all scalar coefficients independent of absolute angle, so a continuous lift of successive small increments remains valid along the entire orbit. Let $f(\delta, r, \zeta, \lambda)$ be the exact relaxed increment from this section. Since $f(0) = \partial_\delta f(0) = 0$,

$$H_\delta = \frac{f(\delta)}{\delta^2} = \int_0^1 (1-u)\partial_\delta^2 f(u\delta)\, du.$$

Six bounded derivatives of f give four bounded derivatives of H. The physical increment is exactly εH_δ. For its relaxed phase, with $\ell = D\Phi$ and $C_{ij} = (\ell H_j n_i - \ell H_i n_j)/(k_i + k_j)$, its full leading reward is

$$H_0(r, \zeta) = -\alpha\sqrt{\tau}\, r^\top C\zeta + \tfrac{\alpha^2}{2}(r + \sqrt{\tau}\zeta)^\top A(r + \sqrt{\tau}\zeta).$$

The section expansion gives $D\Phi(W_b)DF_b(W_b)^\top = \delta r^\top C + O(\delta^2)$, which supplies the first term. It contributes variance, including at the isotropic anchor. For $Y = \mathrm{diag}(a^2, b^2)$ and $s = a^2 + b^2$,

$$A_{13} = -\frac{\sqrt{2}ab(a^2 - b^2)}{s(3a^2 + b^2)}, \quad A_{23} = -\frac{\sqrt{2}ab(a^2 - b^2)}{s(a^2 + 3b^2)},$$
$$C_{13} = -\frac{2\sqrt{2}ab}{s(3a^2 + b^2)}, \qquad C_{23} = \frac{2\sqrt{2}ab}{s(a^2 + 3b^2)}.$$

Only these entries and their symmetric/antisymmetric partners are nonzero. Thus $a = b$ gives $A = 0$ but $C \neq 0$.

Uniform lag and accumulation. Use the joint block state and full relative-phase reward H. On a compact convex control neighborhood, the exact map T and H have bounded derivatives through order four, uniformly for $|\delta| \le \delta_0$, and a common contraction $\rho < 1$. For the fixed-control pullback X_k^λ, put $A_j = D_x T(X_j^\lambda)$, $B_j = D_\lambda T(X_j^\lambda)$ and $B = \sup\|D_\lambda T\|$. Past-control differentiation gives

$$J_{k,j} = A_{k-1}\cdots A_{k-j}B_{k-1-j}, \qquad S_k^\lambda = \sum_{j\ge 0}(j+1)J_{k,j}, \qquad \|J_{k,j}\| \le B\rho^j.$$

The empty product for $j = 0$ is the identity. For $L_i = \sup\|\lambda^{(i)}\|$, past controls differ from the present by $-(j+1)\lambda'/N + O(L_2(j+1)^2/N^2)$. Interpolate the entire past history inside the convex control set. Differentiating each update twice and propagating its forcing through contractions proves, pathwise,

$$X_k^{(N)} = X_k^{\lambda(k/N)} - N^{-1}S_k^{\lambda(k/N)}\lambda'(k/N) + e_{k,N}, \qquad \|e_{k,N}\| \le E_X/N^2,$$
$$E_X = \frac{BL_2(1+\rho)}{2(1-\rho)^3} + \frac{T_2}{2}\sum_{j\ge 0}\rho^j[C_1(j+2) + L_1(j+1)]^2, \qquad C_1 = \frac{BL_1}{(1-\rho)^2},$$

where T_2 bounds the joint (x, λ) Hessian of T. Polynomial factors times ρ^j justify passing from finite histories to the infinite past and differentiating the response; no invariant density or stationary slow angle is assumed. Write $h_0 = \sup|H|$, $h_1 = \sup\|D_x H\|$, and let h_2 bound its second derivatives. Reward expansion costs E_H/N^2, $E_H = h_1E_X + h_2C_1^2/2$, per update. Therefore

$$J_{\delta,N} = \sum_k b_\delta(\lambda_k) + \oint a_\delta + O(C_{\mathrm{ad}}/N), \quad b_\delta = \mathbb{E}H(X^\lambda, \zeta), \quad a_\delta[v] = -\mathbb{E}[D_x H S^\lambda v],$$

with $C_{\mathrm{ad}} = E_H + \frac{1}{2}\sup_{\delta,s}\|\frac{d}{ds}[a_\delta(\lambda(s))[\lambda'(s)]]\|$ by the Riemann-sum bound. For completeness, set $U = B/(1-\rho)$ and $D = T_2(1+U)$. Product differentiation bounds $D_\lambda S$ by

$$S_1 = \frac{2BD}{(1-\rho)^3} + \frac{D}{(1-\rho)^2}, \quad a_0 = \frac{h_1 B}{(1-\rho)^2}, \quad a_1 = \frac{h_2(1+U)B}{(1-\rho)^2} + h_1 S_1.$$

Hence one may take $C_{\mathrm{ad}} = E_H + (a_1 L_1^2 + a_0 L_2)/2$. The first and second δ derivatives of the pullback obey contracting inhomogeneous recursions. Differentiating products adds only polynomial factors to geometric tails, so their weighted series and the mixed response derivatives converge uniformly.

For shape diameter D_x, coupling any initial state to the periodic pullback costs $C_{\mathrm{tr}} = D_x h_1/(1-\rho)$ over the *whole* horizon. Conditioning on the state after update k gives $|\operatorname{Cov}(H_k, H_l)| \le 2h_0 D_x h_1 \rho^{l-k-1}$ for $l > k$; hence $\operatorname{Var}\sum_{k<L} H_k \le C_\nu L$ with $C_\nu = h_0^2 + 4h_0 D_x h_1/(1-\rho)$, without stationarity. The Gram sign identities $T_{-\delta,-\zeta}(-r) = -T_{\delta,\zeta}(r)$ and $H_{-\delta}(-r,-\zeta) = H_\delta(r,\zeta)$ pair entire histories. Both $D_x H$ and S change sign at paired histories, so their product is unchanged. Central symmetry thus makes a_δ even, giving $|\mu_\delta - \mu| \le C_{\mathrm{geo}}\delta^2$, where $C_{\mathrm{geo}} = \frac{1}{2}\sup\|\partial_\delta^2 a_\delta\| \int \|\lambda'\|$ and $\mu_0 = \mu$ by the linear covariance calculation.

For $M = \lfloor t/\varepsilon \rfloor$, reversal cancels the *exact* fixed-control occupation sum before repetition. The relative half-angle obeys

$$|\mathbb{E}d - t\mu| \le \frac{tC_{\mathrm{ad}}}{N} + tC_{\mathrm{geo}}\varepsilon + \varepsilon(C_{\mathrm{tr}} + |\mu|), \qquad \operatorname{Var} d \le C_\nu \varepsilon^2 MN \le C_\nu t\varepsilon N.$$

The variance bound allows arbitrary coupling between orientations. Constants depend on the fixed loop, stability gap, bounded innovation law, shape set and map/endpoint derivatives, not M, N, ε. Repeating the local Taylor remainder would instead retain $tN\sqrt{\varepsilon}$; parity alone cannot remove it. Relaxed endpoints have cross-block DR_θ, spectrum $(5, 2, 0, 0)$ and Eq. (10); raw residual/spectral errors are $O(\sqrt{\varepsilon})$.

Corollary A.1 (Selection without central symmetry) *Retain Theorem 1's assumptions except central symmetry: the innovations remain bounded, centered and iid with a fixed control-independent law. Under the uniform exact-map regularity and contraction established above,*

$$|\mathbb{E}d_\varepsilon - t\mu| \le C(t/N + t\sqrt{\varepsilon} + \varepsilon).$$

Here μ uses the present law's covariance. The variance bound, L^2 selection limit and exact relaxed fit and spectrum remain valid. In Proposition 4, replace $C_{\mathrm{geo}}\varepsilon^2\sum_j m_j$ by $C_{\mathrm{skew}}\varepsilon^{3/2}\sum_j m_j$; the interface and concentration bounds hold with the revised bias and constants for this law.

Proof. Uniform convergence of the pullback derivative series gives bounded $\partial_\delta a_\delta$, hence $|\mu_\delta - \mu_0| \le C_{\mathrm{skew}}|\delta|$, with $C_{\mathrm{skew}} = \sup_{|u|\le\delta_0,\lambda}\|\partial_u a_u(\lambda)\|\int_0^1\|\lambda'(s)\|\,ds$. At zero, the linear state recursion and quadratic reward use only centering, independence and covariance, so $\mu_0 = \mu$. Reversal cancels the exact fixed-control drift regardless of skewness. Multiplying the connection error by $\varepsilon M \le t$, or by $\varepsilon\sum_j m_j$ for composition, gives the stated terms with $\delta = \sqrt{\varepsilon}$. Startup, covariance-decay and bounded-differences estimates use only boundedness and contraction and remain unchanged. Equation (21) gives the distance limit; the spectral identity is geometric. □

Composed schedules and individual-run reliability. A finite library of C^2 loops in one common stable rectangle permits a stronger conclusion. Concatenate J blocks: block j repeats a selected loop of relative-phase flux μ_j in orientation $s_j \in \{-1, 1\}$, m_j times with N_j updates per period. Compare this complete sequence with its exact discrete reversal, including reversed block order, and set $L = \sum_j m_j N_j$.

Proposition 4 (Explicit composition and concentration) *Under Theorem 1's exact-map hypotheses, constants independent of the block lengths and counts give*

$$q = \varepsilon\sum_j s_j m_j \mu_j, \qquad |\mathbb{E}d_\varepsilon - q| \le b_{\mathrm{comp}}, \tag{18}$$

$$b_{\mathrm{comp}} = C_{\mathrm{ad}}\varepsilon\sum_j m_j/N_j + C_{\mathrm{geo}}\varepsilon^2\sum_j m_j + C_{\mathrm{tr}}\varepsilon J. \tag{19}$$

For independent orientations and every $a > 0$,

$$\Pr\{|d_\varepsilon - q| \geq b_{\mathrm{comp}} + a\} \leq 2\exp\left(-\frac{4a^2}{\varepsilon^2 B_{\mathrm{inf}}^2 L}\right). \tag{20}$$

Here $B_{\mathrm{inf}} = 2H_0 + D_x H_1/(1-\rho)$, where H_0, H_1 bound the exact normalized relative-phase increment and its shape derivative, D_x is the invariant-set diameter, and $\rho < 1$ is the synchronous contraction factor. Each interface costs $O(\varepsilon)$; one changed innovation influences the full angle by at most $\varepsilon B_{\mathrm{inf}}$. Appendix A proves these claims and gives the arbitrary-coupling bound. Flux-estimation and integer-design errors enter additively, and Eq. (10) converts phase tolerance to distance tolerance by $2\sqrt{5}$.

This theorem controls the lifted phase and prediction distance. It does not fix the direction of a four-entry contrast, which also depends on the common phase. For one repeated loop, $N \asymp \varepsilon^{-1/3}$ balances the mean-square upper bound, giving distance root-mean-square error $O(\varepsilon^{1/3})$ with $O(\varepsilon^{-4/3})$ updates (Appendix A). Without central symmetry, the mean correction is $O(t\sqrt{\varepsilon})$ instead of $O(t\varepsilon)$.

Composition, confidence and mismatch. On the joint shape set of diameter D_x, assume $|H| \leq H_0$, $\|D_x H\| \leq H_1$. Coupling each incoming block to its periodic pullback costs $C_{\mathrm{tr}} = D_x H_1/(1-\rho)$ regardless of duration. The literal reversed word starts at the last sample; its periodic law is a cyclic shift of the based reverse law. A period's total reward is invariant under that shift, so its geometric term changes sign and its fixed-control sum is unchanged. Sum the per-period expansion over each block; all fixed-control sums cancel exactly. Taking the half difference charges $\varepsilon J C_{\mathrm{tr}}$ for interfaces, not twice that quantity, proving Eq. (19). Incoming states are coupled to their periodic counterparts for analysis; the learner is never reset.

Replacing innovation j changes its immediate reward by at most $2H_0$. The next states differ by at most D_x, and future reward differences sum to $H_1 D_x \sum_{k\geq 0} \rho^k = C_{\mathrm{tr}}$. Thus the full angle changes by at most $\varepsilon B_{\mathrm{inf}}$, where $B_{\mathrm{inf}} = 2H_0 + C_{\mathrm{tr}}$, including across interfaces. Independent orientations have $2L$ independent source vectors; each changes the half-contrast by at most $\varepsilon B_{\mathrm{inf}}/2$. The bounded-differences inequality gives Eq. (20). Under arbitrary coupling, each marginal still has tail $2\exp[-2a^2/(\varepsilon^2 B_{\mathrm{inf}}^2 L)]$. A half-contrast deviation of a requires a marginal deviation of at least a; the union bound gives $4\exp[-2a^2/(\varepsilon^2 B_{\mathrm{inf}}^2 L)]$. Connectors of K updates per learner add $\varepsilon K H_0$ plus interfaces. For output distance D, half-angle mean bias b, variance v and target distance $D_* > 0$,

$$\mathbb{E}|D - D_*|^2 \leq 20(v + b^2), \qquad \Pr(D \leq D_*/2) \leq \min\{1, 80(v+b^2)/D_*^2\}. \tag{21}$$

Flux bounds $|\widehat{\mu}_j - \mu_j| \leq e_j$ add $\varepsilon \sum_j m_j e_j$ to the phase bias; integer rounding adds $\varepsilon \sum_j |\widehat{\mu}_j|\,|\bar{c}_j - c_j|$. The nearest-integer bound applies only when the rounded design remains budget feasible. For a repeated loop, balancing N^{-2} and εN gives distance RMS $O(\varepsilon^{1/3})$ at $N \asymp \varepsilon^{-1/3}$ and $O(\varepsilon^{-4/3})$ updates. No optimality claim follows. Moreover, $C^+ - C^- = 2DR(\bar{\theta})J \sin d$: the common phase sets the contrast direction, so the scalar theorem does not prove arbitrary 4D target control. An L_b-Lipschitz fixed-control current with occupation mismatch adds at most $\varepsilon L L_b W_1(\nu_+, \nu_-)$ by an occupation coupling; this bounds that current, not arbitrary nonlinear ordering effects.

B Corners and finite jumps in the covariance response

Nonadiabatic geometric pumping is established (Rahav et al., 2008; Takahashi et al., 2020); the result here gives a covariance-response specialization with explicit startup and dwell errors. Continuous corners and finite jumps require different bounds. Throughout this section the object is the covariance recursion (2), not a new global approximation theorem for nonlinear SGD. Use the Frobenius covariance norm, its induced operator norm, and the induced Euclidean readout norm. Write the fixed-control map as $V \mapsto \mathcal{T}_u V + Q_u$, assume $\|\mathcal{T}_u\| \leq \beta < 1$, and define

$$v_u = (I - \mathcal{T}_u)^{-1} Q_u, \qquad \mathcal{R}_u = (I - \mathcal{T}_u)^{-1}, \qquad \mathcal{L}_u(X) = w_u : \mathcal{R}_u X.$$

The fixed-control reward includes the direct-noise term of Eq. (3); it cancels under exactly matched occupation.

Continuous loops with corners. Suppose $\mathcal{T}(t), w(t), v(t)$ are continuous, periodic and piecewise C^2, with finitely many corners and finite one-sided derivative bounds. Set $W = \sup\|w\|$, $R = \sup\|\mathcal{R}\|$, $L_R = \mathrm{Lip}(\mathcal{R})$, $L_{\mathcal{L}} = \mathrm{Lip}(\mathcal{L})$, $S = \mathrm{TV}(v)$, and $J = \mathrm{TV}(v')$, including the periodic jump between the one-sided derivatives at zero. All these constants are finite. For the unique periodic covariance and N equally spaced updates, its excess $E_N = \sum_k w_k : (\bar{V}_k - v_k)$ satisfies

$$\left\| E_N + \oint \mathcal{L}\, dv \right\| \le \frac{1}{N}\left\{ \frac{W}{1-\beta}(RJ + L_R S) + L_{\mathcal{L}} S \right\}. \tag{22}$$

Thus corners change the finite-speed constant, not the limiting flux. If the control curve bounds an oriented piecewise-smooth surface on which the connection is C^1, the usual Stokes identity applies to that surface.

Proof. Put $h = 1/N$, $D_k = v_{k+1} - v_k$, $p_k = -\mathcal{R}_k D_k$, and $e_k = \bar{V}_k - v_k - p_k$. Then $e_{k+1} = \mathcal{T}_k e_k + p_k - p_{k+1}$. Periodicity and contraction give $\sum_k \|e_k\| \le (1-\beta)^{-1} \sum_k \|p_{k+1} - p_k\|$. The translation bound for a periodic bounded-variation function gives $\sum_k \|D_{k+1} - D_k\| \le hJ$; also $\sum_k \|D_k\| \le S$. Consequently $\sum_k \|p_{k+1} - p_k\| \le h(RJ + L_R S)$. Finally, replacing $\sum_k \mathcal{L}_k D_k$ by $\oint \mathcal{L}\, dv$ costs at most $hL_{\mathcal{L}} S$. These bounds prove Eq. (22). □

Take the matched samples $u_k^+ = u(k/N)$ and $u_k^- = u((-k/N) \bmod 1)$, so both words start at $u(0)$. Their half-contrast from a common actual covariance V_{in} satisfies Eq. (22) with the following additional numerator. Let $V_1 = \sup\|v'\|$, $L_T = \mathrm{Lip}(\mathcal{T})$, $L_w = \mathrm{Lip}(w)$ and $Z = \|V_{\mathrm{in}} - v(0)\| + V_1/(1-\beta)$. The extra constant is

$$C_{\mathrm{start}} = \frac{L_w Z\beta + WV_1}{(1-\beta)^2} + \frac{WL_T Z\beta}{(1-\beta)^3}.$$

Indeed, the periodic initial lags differ by at most $2V_1/[N(1-\beta)]$. The two transient propagators at step k differ by at most $L_T k(k-1)\beta^{k-1}/N$, and their readouts differ by at most $2L_w k/N$. Summing the resulting geometric series and dividing by two proves the claim; no differentiability at a corner is used.

Finite jumps with a common actual start. Choose controls $u_0, \dots, u_{\ell-1}$ and integer dwells $m_i \ge 1$. Compare the anchored words

$$u_0^{m_0} u_1^{m_1} \cdots u_{\ell-1}^{m_{\ell-1}} u_0^{m_0}, \qquad u_0^{m_0} u_{\ell-1}^{m_{\ell-1}} \cdots u_1^{m_1} u_0^{m_0},$$

started from the same V_{in}, including $V_{\mathrm{in}} = 0$. Both anchor dwells count toward exposure and execution cost. Define $q_i = \|\mathcal{T}_{u_i}\|^{m_i}$, $B = \max_i \|Q_{u_i}\|/(1-\beta)$ and $X = \max\{B, \|V_{\mathrm{in}}\|\}$. For an oriented edge set

$$\begin{aligned} J(u,v) &= \mathcal{L}_v(v_u - v_v), \\ \mathcal{A}(u,v) &= \tfrac{1}{2}[J(u,v) - J(v,u)] = -\tfrac{1}{2}(\mathcal{L}_u + \mathcal{L}_v)(v_v - v_u). \end{aligned} \tag{23}$$

If H denotes half the difference of the two total normalized covariance rewards, then, with $u_\ell = u_0$,

$$\left\| H - \sum_{i=0}^{\ell-1} \mathcal{A}(u_i, u_{i+1}) \right\| \le \frac{4XW}{1-\beta} \sum_{i=0}^{\ell-1} q_i. \tag{24}$$

Here $W = \max_i \|w_{u_i}\|$. No smooth interpolation between controls is assumed.

Proof. Starting a dwell at covariance S, its exact excess is

$$E(u, m, S) = \mathcal{L}_u(I - \mathcal{T}_u^m)(S - v_u).$$

Every covariance has norm at most X, and every fixed-control covariance has norm at most B. After a preceding dwell at v, therefore, $\|S - v_v\| \le 2Xq_v$. Subtracting $J(v,u)$ from the displayed exact excess bounds the difference by $2XW(q_v + q_u)/(1-\beta)$. The initial anchor reward is identical in both words and cancels exactly. Sum the remaining ℓ edge errors; each q_i occurs twice. Reversal preserves every fixed-control reward and changes $J(u,v)$ into $J(v,u)$. This proves Eq. (24). □

Discrete flux and scope. Define $\mathcal{F}_\triangle(u,v,z) = \mathcal{A}(u,v) + \mathcal{A}(v,z) + \mathcal{A}(z,u)$. Antisymmetry cancels internal edges, giving the exact fan identity

$$\sum_{i=0}^{\ell-1} \mathcal{A}(u_i, u_{i+1}) = \sum_{i=1}^{\ell-2} \mathcal{F}_\triangle(u_0, u_i, u_{i+1}).$$

For successive partitions of a continuous rectifiable loop whose maximum subarc length tends to zero, Eq. (23) converges to $-\oint \mathcal{L}\, dv$, whenever the fields are Lipschitz along the loop. Finite edges instead retain their edge-dependent response; the identity does not identify a restart with the flux of an arbitrarily chosen smooth interpolation. Without the common paid anchor, or a periodic covariance start, an additional startup current generally survives. Short dwells remain covered by the exact dwell formula, although the bound need not be small. For nonlinear SGD, multiply the covariance response by ε and add the separately justified local approximation remainder from Eq. (3); neither result supplies a new many-cycle nonlinear guarantee.

State-dependent forcing and interpolation. Smooth shape-dependent injected noise $\Gamma(P)\zeta$ preserves the exact argument when $\Gamma \in C^6$, derivatives are bounded and sampling amplitude is held fixed during gradient evaluation; use covariance $\Gamma(Y)\Gamma(Y)^\top$. Neutral-angle dependence need not preserve fixed-control cancellation. General smooth $\Gamma(\theta)$ supports only the local argument without additional structure. Interpolating squared-loss sample gradients vanish: native forcing cannot be inferred from this injected-noise theorem.

C Finite training with native noise and numerical approximations

Covariance and mean response. Along each schedule's own deterministic trajectory $\bar{x}_{k+1} = f_k(\bar{x}_k)$, set $A_k = Df_k$, $H_k = D^2 f_k$, $Q_k = \mathbb{E}[b_k b_k^\top]$, $b_k = \partial_\delta T_k(\bar{x}_k, 0, \xi_k)$, and $c_k = \frac{1}{2}\mathbb{E}\partial_{\delta\delta} T_k(\bar{x}_k, 0, \xi_k)$. With $P_0 = m_0 = 0$, propagate

$$\begin{aligned} P_{k+1} &= A_k P_k A_k^\top + Q_k, \\ m_{k+1} &= A_k m_k + \tfrac{1}{2} H_k : P_k + c_k. \end{aligned} \tag{25}$$

Envelopes and proof of Theorem 2. Fix the initial optimizer state, horizon and control word independently of the innovations and amplitude. On specified deterministic neighborhoods containing every homotopy state, require T_k and ψ to be C^3 and

$$\sup_{x,\, 0\le\delta\le\delta_*} \|\partial_x^r \partial_\delta^s T_k(x,\delta,\xi_k)\| \le E_{rs,k}(\xi_k), \quad \|E_{rs,k}\|_{L^3} < \infty \quad (1 \le r+s \le 3).$$

Each envelope depends only on the current independent innovation. Use Euclidean state/output norms, multilinear operator norms, and Frobenius covariance norms; suppose $\|D^j\psi\| \le G_j$, $j \le 3$. Put $e_{rs,k} = \|E_{rs,k}\|_3$, $a_k = e_{10,k}$. Starting at zero, compute

$$\begin{aligned} U_{k+1} &= a_k U_k + e_{01,k}, \\ V_{k+1} &= a_k V_k + e_{20,k} U_k^2 + 2e_{11,k} U_k + e_{02,k}, \\ W_{k+1} &= a_k W_k + 3e_{20,k} U_k V_k + 3e_{11,k} V_k + e_{30,k} U_k^3 \\ &\quad + 3e_{21,k} U_k^2 + 3e_{12,k} U_k + e_{03,k}, \\ C_T &= G_1 W_T + 3G_2 U_T V_T + G_3 U_T^3. \end{aligned}$$

At any amplitude, the third state derivative satisfies

$$z_{k+1} = T_x z_k + 3T_{xx}[u_k, w_k] + 3T_{x\delta} w_k + T_{xxx}[u_k, u_k, u_k] + 3T_{xx\delta}[u_k, u_k] + 3T_{x\delta\delta} u_k + T_{\delta\delta\delta}.$$

Bounding current coefficients by their independent envelopes and applying Minkowski and Hölder gives, uniformly over the homotopy interval, $\|u_k\|_3 \le U_k$, $\|w_k\|_{3/2} \le V_k$ and $\|z_k\|_1 \le W_k$. For example, $\mathbb{E}\|u_k\|\|w_k\| \le U_k V_k$. The expected norm of the terminal third derivative is bounded by C_T. At zero amplitude their first two recursions are

$$u_{k+1} = A_k u_k + b_k, \qquad w_{k+1} = A_k w_k + H_k[u_k, u_k] + 2T_{x\delta,k} u_k + T_{\delta\delta,k}.$$

Because the zero-amplitude maps are deterministic, the current innovation is independent of u_k, w_k. Fresh-sample independence and centering give $\mathbb{E}u_k = 0$; $P_k = \mathbb{E}u_k u_k^\top$ and $m_k = \mathbb{E}w_k/2$ therefore obey Eq. (25). Dominated differentiation and Taylor's integral remainder prove Eq. (16) and its half-contrast bound. Finite envelopes need not give a small error at $\delta = 1$.

Dependent sampling. For a joint innovation law independent of amplitude, marginal centering does not restore the iid closure. Under deterministic uniform derivative bounds, the same proof retains $C_k = \mathbb{E}[u_k b_k^\top]$ and $r_k = \mathbb{E}[T_{x\delta,k} u_k]$: add $A_k C_k + C_k^\top A_k^\top$ to the covariance recursion and r_k to the mean recursion. The envelope recursions then hold pathwise. For reshuffling centered example fields e_i^j, e_i^k, distinct batches in one epoch have cross covariance $-(n-1)^{-1} n^{-1} \sum_i e_i^j (e_i^k)^\top$; independent epochs give zero. The distinct-index identity $\sum_{i \neq l} e_i^j (e_l^k)^\top = -\sum_i e_i^j (e_i^k)^\top$ proves this qualification.

Numerical attribution. For fresh independent sampling, use the following residuals. For dependent sampling, replace Q_k by $Q_k + A_k C_k + C_k^\top A_k^\top$ and c_k by $c_k + r_k$ throughout.

For computed $\widetilde{P}, \widetilde{m}$ with zero initial values, define

$$D_k^P = \widetilde{P}_{k+1} - A_k \widetilde{P}_k A_k^\top - Q_k, \qquad D_k^m = \widetilde{m}_{k+1} - A_k \widetilde{m}_k - \tfrac{1}{2} H_k : \widetilde{P}_k - c_k.$$

With $\ell_k = \|A_k\|$, $h_k = \|H_k : \|_{F \to 2}$, propagate $e_0^P = e_0^m = 0$ and

$$e_{k+1}^P = \ell_k^2 e_k^P + \|D_k^P\|_F, \qquad e_{k+1}^m = \ell_k e_k^m + \tfrac{1}{2} h_k e_k^P + \|D_k^m\|.$$

Subtracting recursions proves these bounds. The total output error is at most $\delta^2(\|D\psi\| e_T^m + \frac{1}{2}\|D^2\psi : \|e_T^P) + |\delta|^3 C_T/6$. Reference, covariance and readout errors need additional bounds; projected residuals cannot certify discarded directions. Random acquisition bounds require joint coverage. This separates approximation bounds, not empirical variance components.

Combining numerical and Taylor errors. For orientation s, set $\nu_s = \|D\psi\| e_s^m + \frac{1}{2}\|D^2\psi : \|e_s^P$ and let ρ_s bound remaining reference/readout error. Theorem 2 and the triangle inequality give the native half-contrast radius $(C_T^+ + C_T^-)/12 + (\nu_+ + \nu_- + \rho_+ + \rho_-)/2$. Errors already included in the recursions are not counted again. The relative radius is $B/(\|\widehat{D}\| - B)$ when $\|\widehat{D}\| > B$.

A uniform bound from native update moments. Fix one orientation and omit its indices; assume all displayed expectations are finite. Write $X_{k+1} = T_k(X_k, 1, \xi_k)$ for native training, with fixed initial state $X_0 = \bar{x}_0$. All coefficients below refer to this schedule's own reference $\bar{x}_k$ and Eq. (25). For output-valued linear maps L_k and symmetric bilinear maps M_k, set

$$\begin{aligned}
(a_T, L_T, M_T) &= (\psi(\bar{x}_T), D\psi(\bar{x}_T), D^2\psi(\bar{x}_T)),\\
L_k &= L_{k+1} A_k, \qquad M_k[u,v] = M_{k+1}[A_k u, A_k v] + L_{k+1} H_k[u,v],\\
a_k &= a_{k+1} + L_{k+1} c_k + \tfrac{1}{2} M_{k+1} : Q_k,\\
v_k(\bar{x}_k + e) &= a_k + L_k e + \tfrac{1}{2} M_k[e,e].
\end{aligned}$$

Substitution shows that $a_k + L_k m_k + \frac{1}{2} M_k : P_k$ is independent of k. Consequently $v_0(X_0) = \psi(\bar{x}_T) + R_T$: this backward construction represents the existing covariance forecast.

Use the actual native conditional mean and covariance, $\mu_k(x) = \mathbb{E}[T_k(x, 1, \xi_k) - \bar{x}_{k+1}]$ and $V_k(x) = \mathrm{Cov}(T_k(x, 1, \xi_k))$. For $e = x - \bar{x}_k$, the exact one-step defect is

$$r_k(x) = a_{k+1} + L_{k+1}\mu_k(x) + \tfrac{1}{2} M_{k+1} : [V_k(x) + \mu_k(x)\mu_k(x)^\top] - a_k - L_k e - \tfrac{1}{2} M_k[e,e].$$

Choose domains $\mathcal{D}_k$ containing X_k almost surely, and valid bounds $\sup_{\mathcal{D}_k} \|r_k\| \le d_k$ and $\sup_{\mathcal{D}_T} \|\psi - v_T\| \le d_T$. Conditional expectation of a quadratic depends only on these two moments, so the tower property gives the exact identity

$$\mathbb{E}\psi(X_T) - v_0(X_0) = \mathbb{E}[\psi(X_T) - v_T(X_T)] + \sum_{k<T} \mathbb{E} r_k(X_k).$$

The endpoint error is therefore at most $d_T + \sum_{k<T} d_k$. This identity remains true for arbitrary computed a, L, M; their errors are included in their actual defects. If $q \ge \|v_0(X_0) - \psi(\bar{x}_T) - \widetilde{R}\|$, adding q and taking half-differences proves Corollary 3 for every pair. No independence across pairs is needed. Randomly estimated bounds require one joint validity event for the whole family. A projected residual does not bound discarded directions. Approximate reference errors must also enter q or the defects, without double counting.

Local bounds for SGD and momentum. For the exact backward coefficients above, unbiased updates on the augmented optimizer state satisfy $\mu_k(x) = f_k(x) - f_k(\bar{x}_k)$, $c_k = 0$ and $Q_k = V_k(\bar{x}_k)$. For a validated ball $\|x - \bar{x}_k\| \le \varrho_k$, let $F_k \ge \sup \|D^3 f_k\|$, $t_k = \|H_k\|\varrho_k^2/2 + F_k\varrho_k^3/6$, and $\omega_k \ge \sup \|M_{k+1} : [V_k(x) - V_k(\bar{x}_k)]\|$. Taylor's formula and expansion of the quadratic term give the explicit bound

$$d_k = \|L_{k+1}\|F_k\varrho_k^3/6 + \|M_{k+1}\|_{\mathrm{bil}}(\|A_k\|\varrho_k t_k + t_k^2/2) + \omega_k/2.$$

A terminal bound is $d_T = \sup \|D^3\psi\|\varrho_T^3/6$. The native update must map each validated domain into the next; bounds from sampled states alone do not prove this containment. For affine updates, state-independent covariance and quadratic readout, all defects vanish without requiring contraction. For AdamW, the general defect formula uses the conditional moments of the complete normalized update; gradient covariance alone is insufficient. These are bounds on expected stochastic excess. Raw targets additionally require the deterministic half-contrast and its approximation error.

Native SGD, momentum and adaptive normalization. For independent unbiased minibatches, write $g_k^\delta(x,\xi) = \bar{g}_k(x) + \delta(g_k(x,\xi) - \bar{g}_k(x))$. SGD has $Q_k = \alpha_k^2 \operatorname{Cov}(g_k)$ and $c_k = 0$. The covariance and its derivatives may depend on the current state. Independent additive gradient perturbations add covariance terms; correlated sources add cross-covariances. The distinction between minibatch and label-noise diffusion is explicit in doubly stochastic models (Xiong et al., 2023); effective noise temperatures and replica distances are also studied through dynamical mean-field theory (Mignacco & Urbani, 2022). Neither comparison removes the chronological covariance transport in Eq. (25).

For momentum, augment $x = (\theta, v)$, with $v_+ = \beta_k v + g_k^\delta$ and $\theta_+ = (1 - \alpha_k\lambda_k)\theta - \alpha_k v_+$. The update remains affine in δ, so $c_k = 0$; retain the parameter–momentum covariance. Deterministic resets are part of the specified map. For AdamW, include $x = (\theta, m, v)$, actual bias corrections and decoupled decay. For an optimizer map $\mathcal{O}_k(x, g)$,

$$Q_k = D_g\mathcal{O}_k \operatorname{Cov}(g_k) D_g\mathcal{O}_k^\top, \qquad c_k = \tfrac{1}{2}D_g^2\mathcal{O}_k : \operatorname{Cov}(g_k).$$

For example, $v_+ = \beta_2 v + (1 - \beta_2)(g_k^\delta)^{\odot 2}$ contributes $(1 - \beta_2)\operatorname{diag}\operatorname{Cov}(g_k)$ to c_k; the parameter component also contains normalization curvature. Thus setting $c_k = 0$ can confound adaptive mean bias with covariance lag. The denominator $\sqrt{\widehat{v}} + \epsilon_a$ is not C^3 at $\widehat{v} = 0$. Positive checkpoint moments and $\beta_2 > 0$ imply $v_{k,i} \ge \beta_2^T v_{0,i} > 0$ at a fixed horizon. They supply a smooth domain, but may yield large constants. A zero initialization or clipping boundary requires a separate check. The adaptive long-time manifold limit of Li et al. (2025) has different step/memory scaling; it is not a consequence of this finite-horizon expansion.

Full prediction spaces and smooth architectures. Use the full logit or probability array on a fixed evaluation set as ψ. For n images we use $\|\psi\|_2/\sqrt{n}$, so derivative constants have the same per-image normalization. Finite GELU Transformers with softmax, positive-stabilizer layer normalization and smooth loss/readout meet the smoothness requirement along every finite homotopy path. With finitely many possible batches and a fixed finite horizon, the union of these paths over $\delta \in [0, 1]$ is compact. For SGD/momentum, smoothness holds globally; for AdamW use the positive lower bound above. A surrounding compact smooth domain therefore supplies finite derivative maxima. Finite-moment unbounded or skewed noise is also permitted when the stated L^3 envelopes hold; adaptive maps can require higher raw-gradient moments. ReLU switching, argmax accuracy and unbounded-time cross-entropy limits are outside this theorem.

Order, startup and the residual specialization. With $\mathcal{A}_{k:j} = A_{k-1}\cdots A_j$ and $\mathcal{A}_{j:j} = I$,

$$P_T = \sum_{k<T} \mathcal{A}_{T:k+1} Q_k \mathcal{A}_{T:k+1}^\top.$$

Unrolling the covariance recursion proves the identity and retains noncommuting updates, startup and restarts. In the linear residual coordinates of an interpolating checkpoint, $A_k = I - \alpha_k K$ and $Q_k = \alpha_k^2\tau_k K\Sigma K$ recover Eq. (2). Outside that setting the two deterministic trajectories can differ even under matched occupation. Subtracting their own half-contrast isolates the stochastic excess in Eq. (17); freezing them to one checkpoint adds error. If every sample gradient vanishes and there is no other forcing, SGD stays at that checkpoint, $Q_k = 0$ and $R_T = 0$. Momentum has the same property with zero momentum and no decay. The native construction below retains nonzero individual gradients at a positive minimum loss.

Relation to order-control analyses. Gradient-bracket analyses predict target-loss effects of domain order under fixed total exposure (Rukhovich et al., 2024; Sweeney, 2026a). Fixed-clock optimizer-state replay also yields equal-multiset endpoint contrasts of first order in learning rate (Sweeney, 2026b). Here δ scales centered noise at fixed rates and horizon: each word retains its own deterministic trajectory, and Theorem 2 bounds its second-order expected stochastic response. The macroscopic isospectral conclusions require the separate Gram constructions.

D NATIVE MINIBATCH SELECTION AND A NONCOMMUTING CONSTRUCTION

The first construction controls only learning rate and batch size on a fixed dataset. The second uses specified reflected noise channels to realize two noncommuting rotations. They address different restrictions of the rank-two theorem; neither asserts general neural controllability.

D.1 TWELVE FIXED EXAMPLES AND A FOUR-UPDATE WORD

Put $\delta = \sqrt{\varepsilon}$, $Y = \operatorname{diag}(1, 4)$ and $D = \operatorname{diag}(1, 2)$. Form the Cartesian-product multiset

$$u \in \{-1/\sqrt{2}, -1/\sqrt{2}, \sqrt{2}\}, \quad v, w \in \{-1, 1\}, \qquad \zeta = (u, v, u/2 + \sqrt{3}w/2), \quad Z = \mathsf{S}(\zeta).$$

Each of its twelve equally weighted indices has fixed targets $(Y - \delta Z, I_2)$ and loss

$$\ell_Z(U, V) = \tfrac{1}{4}\|UU^\top - Y + \delta Z\|_F^2 + \tfrac{1}{4}\|VV^\top - I_2\|_F^2.$$

The moments are $\mathbb{E}\zeta = 0$, $\operatorname{Cov}(\zeta) = I + \frac{1}{2}(E_{13} + E_{31})$ and $\mathbb{E}\zeta_1^3 = 1/\sqrt{2}$. Thus the empirical risk is $L_{\text{clean}} + 3\varepsilon/2$: conflicting labels leave positive sample-gradient variance at the empirical minimum. No label is changed after dataset construction. From $(U_0, V_0) = (D, I_2)$, independent uniform sampling with replacement and batch averaging give ordinary SGD

$$U_+ = [I - \alpha(UU^\top - Y + \delta\overline{Z}_q)]U, \qquad V_+ = I_2.$$

The batch-size control q changes the law of the averaged innovation: $\operatorname{Cov}(\overline{\zeta}_q) = [I + \frac{1}{2}(E_{13} + E_{31})]/q$ and $\mathbb{E}(\overline{\zeta}_{q,1})^3 = 1/(\sqrt{2}q^2)$. These identities follow from centering and independence of the q draws; the noise is therefore skewed and control-dependent even though all dataset labels remain fixed.

Theorem 5 (Fixed-period native minibatch selection) *Repeat the word $w = ((.1, 4), (.2, 4), (.2, 1), (.1, 1))$ of (α, q) controls and its literal reverse independently for $M = \lfloor t/\varepsilon \rfloor$ consecutive periods, where $t > 0$ and $0 < \varepsilon \le 10^{-4}$. Each run uses $4M$ updates and $10M$ sampled examples. There is a constant independent of t, M, ε such that the lifted relaxed half-angle contrast satisfies*

$$\mu_B = -297\sqrt{2}/10^6, \qquad |\mathbb{E}d_\varepsilon - t\mu_B| \le C(t\sqrt{\varepsilon} + \varepsilon), \qquad \operatorname{Var} d_\varepsilon \le Ct\varepsilon.$$

Individually, the relaxed and unrelaxed cross-blocks converge in L^2 to $DR(tj_\pm)$ with $O_t(\sqrt{\varepsilon})$ matrix RMS error, where

$$(j_+, j_-) = \frac{6\sqrt{2}}{35}\left(\frac{3913}{100000}, \frac{8519}{200000}\right).$$

Their limiting distance is $2\sqrt{5}|\sin(t\mu_B)|$. Relaxed training predictions agree at minimum risk $3\varepsilon/2$, and both full Gram spectra equal $(5, 2, 0, 0)$. Unrelaxed mean-target residuals and spectral discrepancies are $O(\sqrt{\varepsilon})$, with $O(\varepsilon)$ excess empirical risk. No intermediate relaxation or central symmetry is required.

Invariant domain and exact reward. Write $UU^\top = Y + \delta\mathsf{S}(r)$, with $K = \operatorname{diag}(2, 8, 5)$. The shape map at zero amplitude is $r_+ = (I - \alpha K)r - \alpha K\overline{\zeta}_q$. Every dataset atom satisfies $\|\zeta\| < 2.5$ and $\|K\zeta\| < 12$; batch averages obey the same bounds. On $\|r\| \le 10$, the exact polynomial map satisfies

$$\|T_{\delta,\alpha,q}\| \le 10\rho_\alpha + 12\alpha + \sqrt{2}|\delta|(250\alpha + 625\alpha^2) + 3125\alpha^2\delta^2,$$
$$\|D_r T_{\delta,\alpha,q}\| \le \rho_\alpha + \sqrt{2}|\delta|(45\alpha + 100\alpha^2) + 812.5\alpha^2\delta^2,$$

where $\rho_{.1} = .8$ and $\rho_{.2} = .6$. For $|\delta| \le .01$, the two self-map bounds are below $9.646, 9.474$; the derivative bounds are below $.879, .788$. Shape and multiplier eigenvalues exceed $.8585$ and $.9646$. Hence a common radius-ten invariant ball and contraction $\rho < .879$ hold for every sampled history. Four fresh independent dataset indices, using their first q entries, encode each update with a fixed innovation law; smoothness in q is unnecessary.

The empirical risk differs from the clean objective by a constant and has its endpoint map. The exact relaxed phase increment is $\varepsilon H_{\delta,c}(r,\omega)$ by the smooth section and Taylor division established in Appendix A. Here $c = (\alpha, q)$ ranges over a finite alphabet. Its value and joint first derivatives are bounded on the tube, independently of absolute angle.

Periodic response and accumulation. Let $J_\delta^\pm$ be the normalized expected reward per period in the unique four-periodic pullback process. Write $T_\delta^* = \sup \|\partial_\delta T\|$, $H_1 = \sup \|D_r H\|$, $H_\delta^* = \sup |\partial_\delta H|$, and $H_0 = \sup |H|$ over the compact domains. Coupling entire histories gives

$$\sup_k \|\bar r_k^\delta - \bar r_k^0\| \le \frac{T_\delta^* |\delta|}{1-\rho}, \qquad |J_\delta^\pm - J_0^\pm| \le 4\left(H_\delta^* + \frac{H_1 T_\delta^*}{1-\rho}\right)|\delta|.$$

Coupling the actual start costs at most $2RH_1/(1-\rho)$ in the total normalized reward, independent of duration. The full-reward covariance bound from Appendix A gives $\mathrm{Var}\sum_{k<L} H_k \le C_\nu L$, with $C_\nu = H_0^2 + 8H_0RH_1/(1-\rho)$ and $R = 10$. Consequently each angle has mean $\varepsilon M J_0^\pm + O(t\sqrt{\varepsilon}+\varepsilon)$ and variance $O(t\varepsilon)$. This argument uses a fixed period and does not invoke a slow-loop limit.

Evaluation of the nonzero coefficient. At zero amplitude only $z_k = \mathbb{E}[r_{k,1} r_{k,3}]$ enters the mean reward:

$$z_{k+1} = (1-2\alpha_k)(1-5\alpha_k)z_k + 5\alpha_k^2/q_k, \qquad \mathbb{E}H_{0,c} = A_{13}\alpha^2(z + 1/(2q)), \quad A_{13} = 6\sqrt{2}/35.$$

The periodic pre-update values are

$$z^+ = (13/100, 129/2000, 1/20, 1/5), \qquad z^- = (13/400, 63/1000, 1/5, 1/20).$$

Their sums $\sum \alpha_k^2 z_k$ are $197/25000$ and $2269/200000$. The direct-noise sum is $1/32$ for both words, giving the stated $j_\pm$; their half difference is $A_{13}(-693/400000) = -297\sqrt{2}/10^6$. Combining the periodic estimates with $0 \le t - \varepsilon M < \varepsilon$ proves the angle bounds. Lipschitz continuity of $DR(\theta)$ gives the matrix limits. The Gram identity gives the exact relaxed spectrum; smooth noiseless relaxation and Weyl's inequality give the raw residual/spectral bounds. The excess empirical risk is $\varepsilon\|r\|^2/2 \le 50\varepsilon$. These are fixed datasets within each run; the small-noise limit concerns an ε-indexed family of conflicting-label datasets.

D.2 Relaxed frames and overlapping-plane echoes

Write $J_{ij} = e_i e_j^\top - e_j e_i^\top$. For a square factor $W = P^{1/2}Q$ with $Q \in SO(r)$ and positive symmetric multiplier H, direct multiplication gives

$$P_+ = HPH, \qquad O = P_+^{-1/2} H P^{1/2} \in SO(r), \qquad Q_+ = OQ.$$

The inverse-shape noiseless-relaxation solution in Appendix A holds in every fixed rank. Its factor transport defines a smooth correction $E_Y(P) \in SO(r)$ with $\Pi(W) = Y^{1/2}E_Y(P)Q$. The relaxed frame $\Theta = E_Y(P)Q$, defined without executing noiseless relaxation, therefore satisfies

$$\Theta_+ = K\Theta, \qquad K = E_Y(P_+) O E_Y(P)^\top.$$

For two factors, $R_+ = K_U R K_V^\top$; ordered products retain noncommuting increments. Relaxed factors $U = DQ_U, V = Q_V$ have $C = DR$, $R = Q_U Q_V^\top$, and $W^\top W = Q_U^\top D^2 Q_U + I$, fixing the full Gram spectrum.

For the Gram map write $P = Y + \delta \mathsf{S}(x)$, $H = I - \alpha\delta(\mathsf{S}(x) + Z)$ on fixed compact domains. The section $\Psi(P) = P^{1/2}E_Y(P)^\top$ has constant relaxed frame. At $D = Y^{1/2}$, $D\Pi(D)$ is the projection onto the loss-Hessian kernel; its complementary stable space is $\{SD : S = S^\top\}$. Both $\partial_\delta\Psi|_0$ and $\partial_\delta(H\Psi)|_0 = \partial_\delta\Psi|_0 - \alpha(\mathsf{S}(x)+Z)D$ lie in that stable space. Thus $K_0 = I$ and $\partial_\delta K_\delta|_0 = 0$. Taylor's integral formula yields

$$K_\delta = I + \delta^2 G_\delta, \qquad \|D_x K_\delta\| \le C\delta^2, \qquad \|W - \Pi(W)\|_F \le C|\delta|,$$

with bounded joint first derivatives of G. These are local identities; the invariant domain below keeps them valid for the full executed word.

Theorem 6 (Reflected-noise noncommuting selection) *Take $D = \mathrm{diag}(1, 2, 1)$, target $Y = D^2$, and noiseless anchor $V = I_3$. In ordered planes $A = (1, 2)$ and $B = (3, 2)$, embed $Z = \mathsf{S}(u, v, u/2 + \sqrt{3}w/2)$ with independent Rademacher u, v, w, or its reflection $Z \mapsto SZS$, $S = \mathrm{diag}(1, -1, 1)$. Each update uses a fresh innovation. A positively oriented ellipse γ in $0 < \alpha_- \le \alpha \le \alpha_+ < 1/4$ and $0 < \tau_- \le \tau \le \tau_+$ has flux*

$$\mu = \iint_{\mathcal{S}_\gamma} -\frac{6\sqrt{2}\alpha}{(7 - 10\alpha)^3}\, d\alpha\, d\tau < 0.$$

An echo $\mathcal{E}_j$ runs M forward cycles in channel j, then M exact reversed cycles in its reflected channel, each of length N, where $M = \lfloor t/\varepsilon \rfloor$ and $t = \pi/(4|\mu|)$. The words $\mathcal{E}_A\mathcal{E}_B$ and $\mathcal{E}_B\mathcal{E}_A$ have the same joint multiset of rate, noise scale, plane and reflection controls and the same $4MN$ updates. For sufficiently small noise, as $N \to \infty$ and $\varepsilon N \to 0$,

$$R^{AB} \to R_B R_A, \qquad R^{BA} \to R_A R_B \quad \text{in } L^2, \qquad R_j = e^{-\pi J_j/2}, \quad J_A = J_{12},\ J_B = J_{32}.$$

Both relaxed full Gram spectra equal $(2, 5, 2, 0, 0, 0)$, and relaxed and unrelaxed cross-block distances converge in L^2 to $2\sqrt{3}$. Raw observed residuals and spectral errors are $O(\sqrt{\varepsilon})$ uniformly over the horizon. There is no intermediate noiseless relaxation.

Full-shape stability and exact cancellation. In $\mathrm{Sym}(3)$, put $P = Y + \delta X$. The normalized shape map is $X_+ = (I - \alpha\mathcal{L}_Y)X - \alpha\sqrt{\tau}\mathcal{L}_Y Z + O(\delta)$, where $\mathcal{L}_Y X = YX + XY$ has eigenvalues among $2, 5, 8$. Its polynomial remainder has bounded derivatives on fixed balls. Choose a radius larger than maximum linear forcing divided by $1 - \rho_0$, with $\rho_0 = \sup \|I - \alpha\mathcal{L}_Y\| < 1$; small enough δ then gives a common invariant ball, synchronous contraction $\rho < 1$, and positive shapes/multipliers for all four channels.

From shape Y, a single active plane preserves its rank-two subsystem. Right equivariance removes dependence on incoming global orientation. The two-sided reflection $W \mapsto SWS$ preserves target and noiseless relaxation and reverses the active-plane phase. Its exact fixed-control reward and geometric one-form are therefore $-b_\delta$ and $-a_\delta$. The two echo halves have expected phases

$$\varepsilon M\{B_{\delta,N} + \mu_\delta + O(N^{-1})\}, \qquad \varepsilon M\{-B_{\delta,N} + \mu_\delta + O(N^{-1})\}, \quad B_{\delta,N} = \sum_{k<N} b_\delta(\lambda_k),$$

up to $O(\varepsilon)$ startup errors. Exact fixed-control cancellation and central-symmetry parity give

$$|\mathbb{E}\vartheta_j - 2t\mu| \le C(t/N + t\varepsilon + \varepsilon), \qquad \mathrm{Var}\,\vartheta_j \le Ct\varepsilon N.$$

The variance uses the full-reward contraction bound across the channel switch. Since $2t\mu = -\pi/2$, each echo approaches its stated quarter turn.

Interfaces and predictor separation. Couple an incoming module with its shape-Y version using the same future innovations and initial relaxed orientation. Their normalized shapes differ by at most $C\rho^k$. The bound $\|D_x K_\delta\| \le C\varepsilon$ and telescoping orthogonal products give total orientation error at most $C\varepsilon \sum_{k\ge 0} \rho^k = O(\varepsilon)$, independently of M, N. This handles each plane interface without a reset and proves the displayed product limits. Each word's mean squared orientation error is bounded by $C\{(t/N + t\varepsilon + \varepsilon)^2 + t\varepsilon N\}$. Direct multiplication gives $\|D(R_B R_A - R_A R_B)\|_F^2 = 12$. The relaxed spectral identity, smooth noiseless relaxation bound and Weyl's inequality complete the claims.

E FINITE-NOISE GUARANTEES FOR THE GRAM CYCLE

Using $\rho_0 = .92$, $\alpha_+ = .24$, $y_- = 1$, $y_+ = 4$, $\|K\| = 8$, $Z = \sqrt{3}$, $\tau_+ = 1.75$ and $R = 2F_0/(1 - \rho_0) = 109.9818\ldots$ in Appendix A gives sufficient one-step bounds 1.23×10^{-7} for the tube and 3.17×10^{-8} for contraction. At noise level $\varepsilon = 10^{-4}$, a phase-dependent enclosure proves contraction of the complete 64-step map, permitting expansion at individual updates. Current approximations have separate error bounds.

Proposition 7 (Periodic stability of the Gram cycle) *Let $P = UU^\top$, with main target $Y = \mathrm{diag}(1, 4)$ or anchor $Y = I_2$, and initialize $P = Y$. Use*

$$\alpha_k = \frac{1}{5} + \frac{1}{25}\cos(2\pi j/64), \qquad \tau_k = 1 + \frac{3}{4}\sin(2\pi j/64), \qquad j = \pm k \pmod{64}.$$

For every $|\delta| \le .01$, both blocks remain positive for every supported innovation history, with

$$\lambda_{\min}(P) \ge .9498, \qquad \lambda_{\min}\big(I - \alpha_k[P - Y + \delta\sqrt{\tau_k}\mathsf{S}(\zeta)]\big) > .9199.$$

Each 64-step shape map contracts synchronously on its phase-zero box in Table 4, with the constants in Table 5. Consequently each orientation has a unique attracting periodic shape law in these boxes. The bounds hold for arbitrarily many repetitions, including 265,028 periods, throughout $0 \le \varepsilon = \delta^2 \le 10^{-4}$.

Exact map and initial boxes. For fresh independent signs x, y, z, use main atoms $\zeta = (y, (2x + z)/\sqrt{5}, (2x - z)/\sqrt{5})$ and anchor atoms (x, y, z), each of the eight triples having probability $1/8$. Distinct blocks have independent signs. Here

$$\mathsf{S}(\zeta) = \begin{pmatrix} \sqrt{2}\zeta_1 & \zeta_3 \\ \zeta_3 & \sqrt{2}\zeta_2 \end{pmatrix}, \quad P = \begin{pmatrix} p & c \\ c & d \end{pmatrix}, \quad M = I - \alpha(P - Y + \delta\sqrt{\tau}\mathsf{S}(\zeta)) = \begin{pmatrix} A & C \\ C & B \end{pmatrix}.$$

The exact update $P_+ = MPM$ is the cubic polynomial

$$p_+ = pA^2 + 2cAC + dC^2, \qquad d_+ = dB^2 + 2cBC + pC^2,$$
$$c_+ = c(AB + C^2) + C(pA + dB).$$

All displayed box endpoints are exact decimals.

Table 4: Phase-zero boxes, each containing its initial target.

Block / order	p	d	c
Main +	[.9867306,1.0132547]	[3.7941454,4.1963375]	[−.0203722, .0204033]
Main −	[.9844122,1.0154452]	[3.7058413,4.2738847]	[−.0230746, .0231110]
Anchor ±	[.9749,1.0251]	[.9749,1.0251]	[−.0251, .0251]

Reproducible enclosure. The next recursion specifies every intermediate box; no intermediate box or fitted polynomial is additional input. For a polynomial $g(x) = \sum_e a_e x^e$, of degree at most three in each coordinate, on $B = \prod_i [l_i, u_i]$, its tensor Bernstein coefficients of degree $(3, 3, 3)$ are

$$t_i(b, e) = \sum_{q=0}^{\min(b,e)} \frac{\binom{e}{q}\binom{b}{q}}{\binom{3}{q}} l_i^{e-q}(u_i - l_i)^q,$$

$$b_\beta(g; B) = \sum_e a_e \prod_{i=1}^{3} t_i(\beta_i, e_i), \qquad \beta \in \{0, 1, 2, 3\}^3.$$

Let $\mathcal{H}_B(g)$ be the interval from the smallest lower endpoint to the largest upper endpoint of these coefficients. Bernstein basis functions are nonnegative and sum to one, so this interval contains $g(B)$. The same formula encloses every exact derivative polynomial.

At phase k, write F_u for the cubic update at $\delta = .01u$ for a fixed atom. Its degree-two amplitude Bernstein controls on $u \in [0, 1]$ are

$$F_0, \qquad F_0 + \tfrac{1}{4}(F_1 - F_{-1}), \qquad F_1.$$

Let $\mathcal{Q}_k$ contain the common F_0, the eight F_1 maps and the eight middle controls. Central symmetry of the atom support covers negative amplitudes. From the current box B_k, define

$$B_{k+1,i} = \mathrm{round}_7^{\mathrm{out}} \operatorname{hull}_{F \in \mathcal{Q}_k} \mathcal{H}_{B_k}(F_i),$$
$$(J_k)_{ij} = \max_{F \in \mathcal{Q}_k} \sup |\mathcal{H}_{B_k}(\partial_j F_i)|, \qquad K_0 = I_3, \quad K_{k+1} = J_k K_k.$$

Here $\mathrm{round}_7^{\mathrm{out}}[a, b] = 10^{-7}[\lfloor 10^7 a \rfloor, \lceil 10^7 b \rceil]$. At every phase, the Gershgorin lower bound for P is $\min(l_{k,1}, l_{k,2}) - \max(|l_{k,3}|, |u_{k,3}|)$. For each actual zero- and full-amplitude multiplier, enclose A, B, C by $\mathcal{H}_{B_k}$ and use $\min(\underline{A}, \underline{B}) - \sup|C|$. The multiplier is affine in amplitude, so endpoint positivity suffices. The middle controls need no multiplier test: they enclose the polynomial range but are not actual updates.

Arithmetic and results. All interval operations round outward on the 10^{-40} grid. Addition is exact there; multiplication takes the minimum and maximum of the four endpoint products before outward rounding. Rational conversion, positive integer division and roots use integer floor/ceiling; powers use repeated interval multiplication. Enclose π by Machin's identity $16\arctan(1/5) - 4\arctan(1/239)$: for each arctangent, take the interval between its 40-term alternating sum and that sum plus the next term. Reduce the phase to a quadrant and evaluate sine/cosine at $\pi(j \bmod 16)/32$ with 40 Taylor terms and absolute next-term remainders. Integer square roots enclose the resulting radicals. These conventions, the printed maps and Table 4 determine all 256 phase checks.

Table 5: Outward bounds from the specified interval recursions. The first two columns are minima over every prefix.

Block / order	$\lambda_{\min}(P) \geq$	$\lambda_{\min}(M) >$	$\|K_{64}\|_\infty \leq$
Main +	.9583085	.9236517	$1.70\,10^{-13}$
Main −	.9584730	.9199668	$2.21\,10^{-13}$
Anchor +	.9498	.9821578	$7.05\,10^{-14}$
Anchor −	.9498	.9821578	$7.08\,10^{-14}$

The terminal boxes are

$$\begin{aligned}
&\text{main} + : \quad [.9868305, 1.0131547] \times [3.7942450, 4.1962379] \times [-.0202722, .0203034],\\
&\text{main} - : \quad [.9845122, 1.0153453] \times [3.7059410, 4.2737850] \times [-.0229746, .0230110],\\
&\text{anchor} + : \quad [.9868916, 1.0130719]^2 \times [-.0094077, .0094077],\\
&\text{anchor} - : \quad [.9846203, 1.0152260]^2 \times [-.0110983, .0110983].
\end{aligned}$$

Every terminal endpoint lies more than $9.9\,10^{-5}$ inside its initial box. Enclosure and strict inclusion therefore iterate indefinitely. For any fixed innovation word, the chain rule bounds the absolute period Jacobian by K_{64}. Convexity of the initial box gives synchronous contraction. Contracting iid period maps have a unique attracting invariant law; pushing it through the prefixes gives the periodic law. The joint two-block contraction constant is the larger block constant. Positivity also places every state in the common noiseless-relaxation domain proved for the Gram learner.

These bounds establish positivity and synchronous contraction for $N = 64$. The following calculation controls the accumulated nonlinear mean at the same finite noise level.

The noiseless-relaxation angle correction. For $Y = \mathrm{diag}(a, b)$, the inverse shape obeys $Q(t) = e^{-Yt}Q(0)e^{-Yt} + Y^{-1}(I - e^{-2Yt})$. Writing $W = P^{1/2}R(\phi)$ and taking the skew part of the factor equation gives the endpoint angle $\phi + \chi(P)$, with

$$\chi(P) = -(a - b)\int_0^\infty \frac{Q_{12}(t)}{\operatorname{tr} Q(t) + 2\sqrt{\det Q(t)}}\,dt.$$

Every main phase box lies in $p \in [.98, 1.02]$, $d \in [3.68, 4.30]$, $|c| \leq .025$. Thus $\det P \geq 3.605775$, $Q(t) \succeq I/4.325$ and $|Q_{12}(t)| \leq .025e^{-5t}/3.605775$, yielding $|\chi| \leq 3(.025)(4.325)/(20(3.605775)) = 865/192308$. For the isotropic anchor $\chi = 0$. Its noise law is invariant under conjugation by $\mathrm{diag}(1, -1)$, which reverses the lifted angle, so its mean is zero at every horizon.

E.1 A FINITE-NOISE PHASE BOUND FROM A CANONICAL CORRECTOR

Use the controls, atoms and phase boxes of Proposition 7, at $\varepsilon = 10^{-4}$. The following finite calculation uses exact conditional averages over the eight atoms. It requires no stationary initialization or learner simulation.

Proposition 8 (Finite-noise accumulated phase) *Start the forward and reverse learners at (D, I_2) and run $M \geq 1$ complete periods. For their continuously lifted relaxed relative half-angle d_M,*

$$\mathbb{E}d_M \in M[-8.673123, -6.326805]10^{-7} + [-.006071227, .006071020].$$

In particular, at $M = 265{,}028$, $\mathbb{E}d_M \in [-.235934, -.161607]$. This bounds the nonlinear mean at the specified amplitude and period; it does not identify that mean with the covariance approximation.

A finite definition of the witness. Put $x = (p-1, d-4, c)$, $P = \left(\begin{smallmatrix} p & c \\ c & d \end{smallmatrix}\right)$. For phase k and a noise atom $n = (n_1, n_2, n_c)$ already including amplitude $\sqrt{\varepsilon\tau_k}$, set $A = 1 - \alpha_k(x_1 + n_1)$, $B = 1 - \alpha_k(x_2 + n_2)$, $C = -\alpha_k(x_3 + n_c)$. Let $T_{k,n}(x)$ be the exact cubic shape update minus $(1, 4, 0)$. Its polar-angle increment is

$$\omega_{k,n}(x) = \arctan \frac{(A-B)c + C(d-p)}{Ap + Bd + 2Cc + \sqrt{pd - c^2}(A+B)}.$$

The denominator is positive on the certified boxes, fixing the branch. Order the 56 monomials x^e, $|e| \le 5$, first by total degree and then lexicographically in (e_1, e_2, e_3); index zero is the constant. Let P_5 retain total degree at most five. Define

$$g_{k,a} = [x^{e_a}]\mathsf{P}_5\mathbb{E}_n\omega_{k,n}, \qquad O_{k,ab} = [x^{e_a}]\mathsf{P}_5\mathbb{E}_n(T_{k,n}^{e_b}).$$

Round these coefficients to the nearest multiple of 10^{-30}, with ties toward $+\infty$, obtaining $\widehat{g}_k, \widehat{O}_k$. The same rounding R is applied after each complete scalar dot product. Define the backward map on vectors with zero constant coefficient by $\mathcal{B}_k(v)_0 = 0$ and $\mathcal{B}_k(v)_a = \widehat{g}_{k,a} + \mathsf{R}(\sum_b \widehat{O}_{k,ab} v_b)$ for $a > 0$. Starting at zero, execute $k = 63, 62, \ldots, 0$ three times and retain the resulting vector as v_0. From this final v_0, recompute $v_{63}, v_{62}, \ldots, v_1$ backward, without replacing v_0. Set $v_{64} = v_0$ and $c_k = \widehat{g}_{k,0} + \mathsf{R}(\sum_b \widehat{O}_{k,0b} v_{k+1,b})$. The polynomial $v_k(x)$ has these coefficients. This finite construction defines every coefficient; no convergence of the sweeps or inverse of the jet operator is assumed.

Enclosure of the actual residual. Use outward interval arithmetic on the 10^{-40} grid for coefficients and bounds. Square roots use integer square-root floor/ceiling. Enclose $\pi = 16\arctan(1/5) - 4\arctan(1/239)$ by 40 alternating terms, then enclose sine/cosine after quadrant reduction by 40 terms plus their first omitted-term bounds. Formal square-root and complex-log series give the reward jets; the constant angle uses 24 arctangent terms plus the omitted-term bound. These intervals determine the unique nearest 10^{-30} value of every $g_{k,a}, O_{k,ab}$.

On each phase box form, before taking ranges,

$$e_k(x) = \mathbb{E}_n[\omega_{k,n}(x) + v_{k+1}(T_{k,n}(x))] - v_k(x) - c_k.$$

Retain the entire degree-15 potential composition. Replace only ω by its degree-eight Taylor polynomial about $x = 0$; bound its remainder by the ninth directional Taylor coefficient over $P \in \mathcal{B}_k$ and directions in $\mathcal{B}_k - Y$. Convexity and $Y \in \mathcal{B}_k$ justify the Taylor segment. Coefficients follow from the ordinary series rules for products, reciprocals, square roots and $(\arctan f)' = f'/(1+f^2)$; average the eight atom bounds. For the polynomial residual, substitute the phase box $x = a + (b-a)t$. If its power coefficients are q_m, its tensor degree-15 Bernstein coefficients are

$$\beta_i = \sum_{m \le i} q_m \prod_{j=1}^{3} \frac{\binom{i_j}{m_j}}{\binom{15}{m_j}}, \qquad i \in \{0, \ldots, 15\}^3.$$

Their minimum and maximum, enlarged by the analytic remainder, bound e_k over the entire box. Summing these outward bounds gives

	$10^5 \sum_k c_k$	$10^8 \sum_k e_k(\mathcal{B}_k)$
$+$	$[5.89665917041526, 5.89665917041527]$	$[-9.388488546, 9.319613303]$
$-$	$[6.04669820426205, 6.04669820426206]$	$[-14.183285732, 14.034902430]$

Telescoping from the actual start. Let $\chi(P)$ be the noiseless-relaxation angle correction and $h_\pm = \chi - v_{0,\pm}$. The exact reward identity yields

$$\mathbb{E}\Theta_M^\pm = M \sum_k c_k^\pm + \sum_{m,k} \mathbb{E} e_k^\pm(x_{m,k}^\pm) + \mathbb{E} h_\pm(P_{M,0}^\pm).$$

Here $h_\pm(Y) = 0$. The final-box Bernstein bound for $v_{0,\pm}$, together with $|\chi| \le 865/192308$, gives the half-boundary interval $[-.006071227, .006071020]$. The independent isotropic anchor has zero mean lifted angle by reflection symmetry. Taking the half difference and using the phasewise bounds proves the proposition for every M.

E.2 FINITE ENSEMBLE AND PREDICTION-SEPARATION BOUNDS

Corollary 9 (Finite-noise prediction separation) *At $\varepsilon = 10^{-4}$, $N = 64$, $M = 265{,}028$, use independent learners, orientations and block streams. For the mean half-angle $\bar{d}_{32}$ of 32 independent pairs and cross-block predictions $C^{\pm}$,*

$$\Pr\{-.278 \le \bar{d}_{32} \le -.119\} > .95,$$
$$\|\mathbb{E}C^{+}_{\text{relaxed}} - \mathbb{E}C^{-}_{\text{relaxed}}\|_F > .65, \qquad \|\mathbb{E}C^{+}_{\text{raw}} - \mathbb{E}C^{-}_{\text{raw}}\|_F > .19.$$

Moment bounds and proof. For each main atom, enclose the corrected increment $X = \omega + v_{k+1} \circ T - v_k$ by bounding its degree-15 polynomial minus c_k with $\sum_e |a_e| \prod_i R_i^{e_i}$, where $R_i = \max(|\underline{x}_i|, |\overline{x}_i|)$, and adding that atom's ninth-order remainder. Call the resulting bound $B_{k,n}$. Then $\mathrm{Var}(X \mid P) \le \frac{1}{8}\sum_n B_{k,n}^2$ and $|X - \mathbb{E}(X \mid P)| \le \max_n B_{k,n} + \sup|e_k|$. All polynomials and boxes are fixed by the preceding construction.

For the anchor, the polar denominator is $D_0 + D_n$, with noiseless D_0 and noise-linear D_n; its numerator N_n is noise-linear. On a box set $l = \min(\underline{p}, \underline{d}) - c_*$, $u = \max(\overline{p}, \overline{d}) + c_*$, $c_* = \sup|c|$, $w = \sup|d - p|$, and $\gamma = \alpha\sqrt{\varepsilon\tau}$. Valid bounds are

$$d_0 = 4l[1 - \alpha(u-1)], \quad d_n = \gamma[\sqrt{2}(\overline{p} + \overline{d} + 2\sqrt{\overline{p}\overline{d}}) + 2c_*], \quad n_n = \gamma(2\sqrt{2}c_* + w).$$

Here $D_0 \ge d_0 > d_n \ge |D_n|$ and $|N_n| \le n_n$. With $\rho^2 = (d_n^2 + n_n^2)/d_0^2 < 1$, symmetry and $\mathbb{E}D_nN_n = 0$ in the logarithm series give $|\mathbb{E}(\omega \mid P)| \le \rho^4/[4(1-\rho^2)]$. Also $\mathrm{Var}(\omega \mid P) \le \gamma^2(4c_*^2 + w^2)/(d_0 - d_n)^2$; a centered increment is bounded by $n_n/(d_0 - d_n)$ plus the drift bound. Sum the drift and variance bounds over all 64 phases and M periods; take the maximum centered-increment bound. Divide paired drift by two, variance by $4 \cdot 32$, and the increment bound by $2 \cdot 32$. Outward evaluation gives a decomposition $\bar{d}_{32} = A + Z$, where the terminal correction is included in the pathwise bounded A, and

$$A \in [-.236421, -.161119], \qquad \langle Z \rangle \le .000213299 =: V, \qquad |\Delta Z| \le .000001215375 =: b.$$

The martingale Bernstein bound (Freedman, 1975), with $L = 4$, gives radius $r = 4b/3 + \sqrt{8V + (4b/3)^2} < .041311$ and failure probability at most $2e^{-4} < .05$. This proves the ensemble interval. For one pair, the standard-deviation triangle inequality gives $\mathrm{Var}(d) \le (\sqrt{32V} + .075302/2)^2 < .0145$. Since $C_{\text{relaxed}} = DR(\Theta)$ and $\|(DR)''\|_F = \sqrt{5}$, Taylor's theorem and independence imply

$$\|\mathbb{E}C^{+}_{\text{relaxed}} - \mathbb{E}C^{-}_{\text{relaxed}}\|_F \ge 2\sqrt{5}\big(|\sin \mathbb{E}d| - \mathrm{Var}(d)\big) > .65.$$

Use Proposition 8 and $\sin(.161) \ge .161 - .161^3/6$ for the last inequality; no principal-angle restriction on individual paths is needed. Finally, the square-root Sylvester bound gives $e_U = .3226/(1 + \sqrt{.955}) + \sqrt{5}(.0045)$ and $e_V = .0502/(1 + \sqrt{.9498})$ for each block's change under noiseless relaxation. Thus $\|C_{\text{raw}} - C_{\text{relaxed}}\|_F \le \sqrt{1.0502}e_U + 2e_V < .23$. The two population means lose less than .46, proving the raw-endpoint claim. The probability statement concerns the 32-pair ensemble; the population gaps concern expected matrices.

F EXPERIMENTAL METHODS AND EVALUATION DETAILS

F.1 CONTROLLED REGRESSION

Rank-two Gram	Setting or diagnostic
Updates	Eq. (9); indices $k \bmod N$ and $-k \bmod N$; independent blocks and learners; 32 independent pairs.
Cycle	$h = 1 + 0.2\cos(2\pi s)$, $\tau = 1 + 0.75\sin(2\pi s)$, $\eta = 0.2$.
Primary setting	$(\varepsilon, N, M) = (10^{-4}, 64, 265028)$.
Secondary settings	$N = 32$; $(\varepsilon, M) = (4 \times 10^{-4}, 66257)$ or $(10^{-4}, 265028)$.
Innovations	$(y, px + qz, px - qz)$, with independent Rademacher x, y, z, $p = \sqrt{0.8}$ and $q = \sqrt{0.2}$. Covariance $I + 0.6(E_{23} + E_{32})$; isotropic anchor.
Forecast	Stokes/finite-speed flux: $-0.00757511/-0.00754638$. Include startup; nominal distance $2\sqrt{5}\,\lvert\sin\widehat{d}\rvert$.
Separation checks	Relaxation residual $< 10^{-12}$; spectrum error $< 1.94 \times 10^{-12}$. All 32 pairs exceed half the nominal distance; one-sided 95% probability lower bound: 0.9106.
Angle checks	Secondary discrepancies: 1.28 and 0.012 standard errors. Primary reversed θ_U: 3.76 standard errors.
Null controls	Reflection null: 0.00276 ± 0.00751. Diagonal noise: no relaxed rotation. Noiseless runs retain the initial predictor. Common-rotation difference: 2.85×10^{-13}.

The covariance couples modes $(8, 5)$, complementing Eq. (14)'s $(2, 5)$ example; its curvature is $-28.8\sqrt{2}\eta^2 h/(13 - 40\eta h)^3$. All scaled families retain 20 initializations, including failures. Data, masks and seeds are fixed before scoring.

Scaled family	Data, architecture and preparation
G-D/G-C	65,536 rows, rank 256. Two positive-diagonal QR bases from independent 32768×256 Gaussian matrices; the first is scaled by $\sqrt{1 + 3(j-1)/255}$. Both principal blocks are observed; G-C adds bridges $(i, 32768 + i)$, $i \le 256$.
Gram preparation	Full-batch Adam, step 10^{-3}, no decay, at most 200,000 updates, then noiseless relaxation. Target-RMS and square-root-count normalization; group weights $1/2$ (G-D), $1/3$ (G-C).
MLP	202,014,736 parameters; 64 inputs, four width-8192 tanh layers, 16 outputs. Width-2048 two-layer tanh teacher; 131,072 training and 16,384 evaluation Gaussian inputs.
ViT	Width/blocks/heads/MLP: $1024/24/16/4096$ or $1536/40/24/6144$; 303,318,032 or 1,134,773,776 parameters. 224×224 inputs, patches of 16, GELU, learned positions and class token, affine LayerNorms and all biases (Dosovitskiy et al., 2021).
ViT targets	16-dimensional unit-normalized Gaussian class codes; 1,281,167 ImageNet training images; stratified 5000/45,000 acquisition/assessment split (Russakovsky et al., 2015).
Neural preparation	Variance-scaled initialization; Adam $(0.9, 0.999)$, batch 4096, no decay; cosine $10^{-3} \to 10^{-5}$ over 200 epochs. If residual exceeds 10^{-5}, extend at 10^{-5} for at most 200 epochs, then relax.
Steering	Full-batch SGD without momentum, dropout or augmentation; fixed target-RMS normalization; ViT resize/center crop. $\eta = 1/\widehat{k}_{\max}$ from 64 power iterations and independent probes.
Noise	$\varepsilon = 10^{-5}$; sensitivities $10^{-6}, 10^{-4}$. Sixteen orthonormal directions with weights $(1 + 0.5\cos(2\pi j/16))/16$ and independent signs; isotropic control.
Pilot loops	Ellipses centered at $h_c = 0.8, 1, 1.2$; horizontal radius 0.15; vertical radii $0.15/0.30/0.45/0.60$ about $\tau = 1$; lengths $64/128/256$.
Unseen loops	Ellipses and exponent-4 rounded rectangles at $h_c = 0.9, 1.1$, half-widths $(0.12, 0.40)$, $N = 128$; figure-eights $h = h_c + 0.12\cos s$, $\tau = 1 + 0.40\sin 2s$, $N = 128, 256$.
Execution	$h \in [0.6, 1.4]$, $\tau \in [0.25, 1.75]$; 512-update cap including a 16-point connector from $(1, 1)$, retrace and padding; reverse the complete word.

F.2 REGRESSION FORECASTS AND SCHEDULE DESIGN

For $H = B^\top B$ and $He_i = k_i e_i$, set $u_i = Be_i/\sqrt{k_i}$ and $n_i = \sqrt{k_i}e_i$; hold u_i fixed in Hessian differentiation. CG solves $Hx = Hg_\ell$, where $g_\ell = DG_\ell^\top$, giving $w_\ell = g_\ell - x$.

Component	Specification
Modes and checks	Noise-seeded block Krylov, ranks $32/64/128$, cutoff 10^{-6}, Ritz residual 10^{-4}. CG: at most 2048 actions, residual 10^{-8}, normalized $\|Bw_\ell\| \le 10^{-6}$. Thirty-two probes; four loops compare $r, 2r$ at change threshold 0.05. Failed checks retain the rank-128 forecast and flag.
Outputs	256 standardized coordinates; four orthogonal projection rows, the remaining 252 coordinates, full reserved outputs and checkpoint displacement.
Targets	Pilots $1/2/3/4/5/7/10/12$ ($N = 64$, two learners/orientation) define median norm a and hull H. Ten targets: two half-scaled convex mixtures; $\pm ae_1, \pm ae_2$; two radius-a directions; $(1.25h_H(e_j) + 0.1a)e_j$, $j = 3, 4$.
Success	Relative mean-contrast error $\le 10\%$, valid fit/relaxation and method budget. Off-target limit: $0.25\|v\|$. Keep failed and weak targets; generation fails if $a = 0$ or a is at most twice the median pilot standard-error norm.
Calibration	Rational: $2/4/8/16$ positive poles, $\eta h_{\max}\kappa_j \le 0.95$, eight starts, ridge $10^{-6}, \ldots, 10^{2}$; integrated rational, $1/N$ and lag-$1/2/4/8/16$ features. Resolvent: stable affine-in-h transitions, dimensions $16/32/64$.
Search baselines	Common squared error, four-block decoder and invalid-schedule penalty 10^6. BO: Matérn-$5/2$, four initial points, noisy expected improvement. AutoLRS: quarter-horizon prediction and four checks. PB2: eight members, bottom-quarter replacement.
Method work	Gradient/JVP/VJP/evaluation: 1 unit; HVP: 2. Design: 262,144 units/target (98,268 acquisition; 163,876 final evaluation/refinement). Forecast: 1,409,276 units/checkpoint. Ten-target sharing: $10B$. Composition: 524,288 units over 8192 updates.
Relaxation	FP64 Dormand–Prince 5(4), relative/absolute tolerances $10^{-8}/10^{-10}$; initial/maximum step $0.1/\widehat{k}_{\max}$; 4096 units. Require group residual $\le 10^{-8}$, relative gradient $\le 10^{-10}$, and last-20-step output change $< 10^{-3}\max(a, 10^{-12})$.
Refinement	Two learners/orientation: divide tolerances by 10 and halve maximum step. Contrast change must be below both $0.01\max(a, \|v\|)$ and 0.1 standard-error norm. Retain failures.
Uncertainty	Sixteen learners/orientation; 10,000 paired checkpoint-bootstrap draws. $s_0 = 0.1a$, sensitivities $0.05a, 0.2a$. All loops and targets from a sampled checkpoint remain together.

CLR/SGDR use exact reversals at fixed or quarter-period-shifted noise; restarts use finite recursion. ScaledGD uses its rectangular update or $(W^\top W + \lambda I)^{-1}$, $\lambda = 10^{-6}\,\mathrm{tr}(W^\top W)/r$, and its own endpoint map.

Table 6: Regression scale and cutoff sensitivity: standardized output coordinates, all unseen loops.

Quantity	G-D	G-C	MLP	ViT-code	ViT-1.1B
Acquisition scale a	0.12	0.09	0.034	0.029	0.026
Original-coordinate s_0	0.012	0.009	0.0034	0.0029	0.0026
Median test contrast norm	0.113	0.082	0.0326	0.028	0.0255
Test contrast norm, Q1	0.07212	0.04679	0.01583	0.01546	0.01465
Test contrast norm, Q3	0.1881	0.1304	0.06263	0.04625	0.05128
Cases below s_0 (%)	0	0	0	0.625	1.25
Median absolute error	0.00471	0.004125	0.001792	0.001805	0.001505
Error at cutoff $0.05a$ (%)	4.1	5.2	5.8	6.4	5.9
Error at cutoff $0.10a$ (%)	4.1	5.2	5.8	6.4	5.9
Error at cutoff $0.20a$ (%)	4	5.2	5.8	6.4	5.9

F.3 NATIVE CLASSIFICATION

Component	Specification
Models and data	304,326,632/1,136,286,184 parameters; ViT architectures above with 1000 classes. ImageNet split: 5000 acquisition/45,000 assessment images.
Replicates	20 checkpoints/setting, eight loops/checkpoint, 16 streams/orientation; all 640 pairs retained, including failed-acquisition fallbacks.
Initialization	Independent seeds; weights, positions and tokens drawn from $\mathcal{N}(0, 0.02^2)$ truncated at ± 0.04; zero biases, unit LayerNorm scales.
Training	300 epochs, batch 4096; peak rates 0.4 (SGD), 0.001 (AdamW). Twenty-epoch warmup from peak/100, then cosine decay to peak/1000.
Images	Random resized 224×224 crops, area $[0.08, 1]$, aspect $[3/4, 4/3]$, flip probability $1/2$. RGB mean $(.485, .456, .406)$, SD $(.229, .224, .225)$. Steering: resize 256/center crop 224; no smoothing, mixing, dropout or augmentation.
Optimizers	Clean-label cross-entropy; SGD momentum 0.9, decay 10^{-4}; AdamW $\beta = (0.9, 0.999)$, decay 0.05, denominator offset 10^{-8}. Preserve moments/counters; score direct endpoints.
Rate selection	Largest $\alpha_{\text{peak}} 2^{-j}$, $j = 0, \ldots, 16$, passing 32-update trials at $h = 0.6, 1, 1.4$ on 16,384 fixed examples. Restore each trial; require finite states, loss ≤ 1.01 initial and relative displacement $< 10^{-3}$. Keep failures.
Initial accuracy	82.9/84.3/83.8/85.1% in Table 3 order; final-epoch checkpoints without accuracy selection.

Schedules. Set $(h, b) = (h_c + ax, q(z))$, with $q(z) = 256, 512, 1024$ on $z < -1/2$, $-1/2 \leq z < 1/2$, $z \geq 1/2$, respectively.

Shape/protocol	Definition
Ellipse	$(x, z) = (\cos 2\pi s, \sin 2\pi s)$.
Rectangle	Constant-speed path $(1, 1) \to (-1, 1) \to (-1, -1) \to (1, -1) \to (1, 1)$.
Restart/figure-eight	$(\cos \pi s, \sin 2\pi s)$ / $(\cos 2\pi s, \sin 4\pi s)$.
Execution	Sixteen-point connector from $(1, 0)$, 96 samples $s = k/96$, reverse connector; quantize after interpolation. Four periods give 512 updates; reverse the entire sequence.
Libraries	Acquisition: $(h_c, a) = (.8 \text{ or } 1.2, .15)$; unseen: $(.9 \text{ or } 1.1, .20)$; targets: $(.95 \text{ or } 1.05, .25)$. Independent streams; shape-then-center ordering.
Dwell tests	$d = 1, 2, 4$ repeats/control over $4/d$ periods, preserving 512 updates and joint counts.
Composition/design	Modules have 4/24/4 connector/path/return updates: ellipse $(.9, .2)$ and rectangle $(1.1, .2)$; no resets. Design uses four signed periods with arbitrary order/repetition.

Scoring. Use all class probabilities and $\langle U, V \rangle_{\text{img}} = n^{-1} \sum_{i,c} U_{ic} V_{ic}$:

$$D_{\text{raw}} = \tfrac{1}{2}(\mathbb{E}P^+ - \mathbb{E}P^-), \qquad D_{\text{mb}} = D_{\text{raw}} - \tfrac{1}{2}(P^+_{\text{det}} - P^-_{\text{det}}).$$

Full-gradient references use matched schedules and states for scoring. Methods share acquisition information and forecast from approximate references. Differentiate complete optimizer updates, including AdamW bias correction and positive second moments.

Eight bank contrasts D_j define ten targets. Let $u_j = D_j / \|D_j\|_{\text{img}}$ and $H_j = \max_i |\langle u_j, D_i \rangle_{\text{img}}|$:

$$\frac{D_1 \pm D_2}{4}, \quad \pm\frac{D_3}{2}, \quad \pm\frac{D_4}{2}, \quad \frac{1}{16}\sum_{j=1}^{8}(-1)^{j-1} D_j,$$
$$\frac{D_1 + D_2 - D_3 - D_4 + D_5 + D_6 - D_7 - D_8}{16}, \quad 1.25 H_7 u_7, \quad 1.25 H_8 u_8.$$

Evaluation	Rule
Targets	Acquisition coefficients are reused on assessment images. Require zero class sum and imagewise $\ell_1 \leq 1$. Retain infeasible and zero-support targets; all 800 remain. Freeze designs before revealing assessment targets.
Success	Raw-half-contrast mean error $\leq 25\%$, method budget and mean accuracy loss ≤ 0.15 percentage points per orientation.
Forecast error	$\|\widehat{D} - \bar{D}_{\mathrm{mb}}\|_{\mathrm{img}} / \max(\|\bar{D}_{\mathrm{mb}}\|_{\mathrm{img}}, s_0)$, where $s_0 = \max(10^{-8}, 0.1\,\mathrm{median}_{\mathrm{acq}}\,\|\widetilde{D}_{\mathrm{mb}}\|_{\mathrm{img}})$.
Uncertainty	Two independent eight-stream groups A, B; signal/error moments $\langle D_A, D_B\rangle_{\mathrm{img}}$ and $\langle D_A - \widehat{D}, D_B - \widehat{D}\rangle_{\mathrm{img}}$. Retain negative estimates. Use 10,000 paired checkpoint-bootstrap draws, stratified by setting; pointwise 95% intervals.
Paired error gains	Calibration minus CR: 7.7–10.1 / 7.9–9.2 / 6.9–10.5 / 7.4–9.3 points. Trajectory-only rational errors: 20.8/22.1/19.6/20.4%.

Full-output diagnostics.

For each 45000×1000 endpoint array, use every singular value and 2560 independent forward/reverse pairs per setting:

$$d_{\mathrm{spec}} = \frac{\|s(P^+) - s(P^-)\|_2}{(\|P^+\|_F + \|P^-\|_F)/2}, \qquad d_{\mathrm{fun}} = \frac{\|P^+ - P^-\|_F}{\sqrt{45000}}.$$

Diagnostic	Results in Table 3 order
Spectra	Median discrepancy: 0.18/0.21/0.16/0.19%; singular-value tolerance: 0.5%. FP64 residual/orthogonality error $< 10^{-5}$; subset SVD cross-check.
Stream variation	Full-pair RMS: .083/.074/.090/.079. Same-orientation stream RMS: .032/.030/.034/.032. Half-contrast noise is half the stream-pair RMS.
Seed variation	Bessel-corrected seed RMS after common continuation: .120/.116/.125/.118. Independent-seed-pair RMS is $\sqrt{2}$ times this spread.
Projection ablation	Sixteen nested orthonormal draws, dimensions 4/16/64/256: errors 13.0/14.4/15.7/16.0%, IQRs 4.8/2.6/1.2/0.6 points; full-array error 16.1%. Separate fixed-checkpoint comparison.
Target counts	Identical schedules: 736/800 projected, 600/800 full. Full-array redesign: 644/800 cold, 740/800 amortized.
Off-target test	$\|(I - \Pi_v)D_{\mathrm{raw}}\|_{\mathrm{img}} \leq 0.10\|v\|_{\mathrm{img}}$, $\Pi_v X = v\langle v, X\rangle_{\mathrm{img}}/\|v\|^2_{\mathrm{img}}$: 568/800 cold, 676/800 amortized.
Failures	Zero projected targets fail; 16/800 execution failures in cold-B_0 CR; no clipping or post-evaluation retargeting.

Computation. Transport and compress $[A_k Z_k, U_k]$ for $P_k \simeq Z_k Z_k^\top$, recording discarded trace. For 64 size-512 gradient samples, $e_\ell = g_\ell - \bar{g}$:

$$U_{k,\ell} = \sqrt{\frac{512}{63 b_k}} D_g \mathcal{O}_k e_\ell, \qquad \widehat{c}_k = \frac{512}{126 b_k} \sum_\ell D_g^2 \mathcal{O}_k[e_\ell, e_\ell].$$

Use Appendix C's covariance and mean recursions with the full-output readout; retain AdamW cross-moments and mean correction.

Optimization budget	Operator-work units
Tariffs	$n = 1281167$, $a = 16384/n$. Subset gradient/HVP/second derivative: $a/2a/4a$; batch gradient: b/n; full-panel forward/first/second derivative: $1/1/2$; acquisition multiplier: $5000/45000$.
Method limits	$B_0 = 262144$/target; 424497/checkpoint for forecasting. Ten targets share $10B$ within a checkpoint. Include acquisition, validation, recomputation and failures.
Transport/readout	Rank 256/512: 21208/42347; eight rank-256 pilots: 162353; two rank-64 pairs: 10246; final-evaluation reserve: 32768.

F.4 Numerical and control diagnostics

Rank and composition. For fixed cases and normalizers, retain the cross term:

$$\|\widehat{D}_r - D\|^2 = \|\widehat{D}_r - \widehat{D}_{\mathrm{ref}}\|^2 + \|\widehat{D}_{\mathrm{ref}} - D\|^2 + 2\langle \widehat{D}_r - \widehat{D}_{\mathrm{ref}}, \widehat{D}_{\mathrm{ref}} - D\rangle.$$

Table 7: Classification success: cold / ten-target (%). Pointwise paired 95% gains over calibration use checkpoint resampling within model–optimizer settings.

Method	$B_0/4$	B_0	$4B_0$
CR design	53 / 77	80.5 / 92.5	90 / 97
Rational/resolvent	40 / 60	58.5 / 76	76 / 87
Bayesian search	31 / 47	45 / 65	64 / 79
PB2-style search	26 / 43	41 / 63	66 / 81
Cold gain (95% CI)	8.8–17.4	19.2–24.9	10.8–17.4
Amortized gain (95% CI)	13.4–20.8	13.8–19.2	7.1–12.9

Fixed-checkpoint RMS diagnostic	Rank 64	Rank 128	Rank 256
Forecast error (%)	27.8	20.6	16.4
Distance to rank-512 forecast (%)	16.0	8.1	3.6
Cross term (squared percentage points)	288.83	130.74	27.99

Reference residual: 15.1%. This RMS stress test differs from Table 3's primary medians; acquisition-only refinement retains every case.

Composition: J alternating modules	2	4	8	16
Fixed-checkpoint error (%)	14.7	16.5	23.8	37.6
Moving coefficients (%)	13.9	14.8	17.4	23.1
Four recomputations (%)	13.5	14.2	16.0	18.7

Here $S_J = (A^{16/J}B^{16/J})^{J/2}$ has 512 updates and $J-1$ interfaces, without resets. Recompute at 64/192/320/448; baselines receive equal allowance.

Diagnostic	Definition/result
Conditioning	Strata ≤ 20, 20–100, > 100: 364/400, 224/280, 56/120 cold-target successes.
Skew	Median absolute standardized third moment over eight fixed directions/times; flag low variance. Lowest/highest-stratum errors: 13.8/21.6%; descriptive strata.
Collateral effects	Maximum mean increases: cross-entropy 0.006, ECE 0.09 percentage points. Average per-learner scores over 16 streams/orientation. ECE uses 15 equal-width bins; empty bins contribute zero.
Image tails	Mean of the largest 10% per-image Euclidean errors. Common displacement $\frac{1}{2}(\mathbb{E}P^+ + \mathbb{E}P^-) - P_0$ is distinct from half-contrast.
Richer statistic	16,384 independent linear-normal starts, one SGD update in a four-parameter model: mixed residual-square/relaxed-increment mean 8.17 standard errors from zero. Reflection reuses draws as a paired check.

Finite versus fixed-operator covariance. Use the same surrogate, mean correction and readout. For $LE = I$, define $a_k = LA_kE$, $q_k = LQ_kL^\top$, $\mathcal{T}_kX = a_kXa_k^\top$, $\|a_k\|_2 \le \rho < 1$ and $V_k = (I - \mathcal{T}_k)^{-1}q_k$. Set $V_T = V_{T-1}$, $\Delta V_j = V_{j+1} - V_j$ and $\mathcal{P}_{k:s} = \mathcal{T}_{k-1}\cdots\mathcal{T}_s$:

$$C_k = V_k - \mathcal{P}_{k:0}V_0 - \sum_{j<k} \mathcal{P}_{k:j+1}\Delta V_j, \tag{26}$$

$$C_k^{\mathrm{ad}} = V_k - \mathcal{P}_{k:0}V_0 - \sum_{j<k} \mathcal{T}_{j+1}^{k-j-1}\Delta V_j. \tag{27}$$

For either covariance sequence, use

$$\begin{aligned} C_0 &= m_0 = 0, \\ m_{k+1} &= A_k m_k + \tfrac{1}{2}H_k : (EC_kE^\top) + c_k, \\ R &= D\psi\, m_T + \tfrac{1}{2}D^2\psi : (EC_TE^\top). \end{aligned}$$

Both retain startup; C^{ad} is not clipped.

Fixed-surrogate error (%)	Ellipse	Rectangle	Restart
Finite propagation	16.4	18.1	19.0
Fixed-operator approximation	23.7	29.8	37.2

For $\beta = \rho^2$, telescoping gives

$$e_k := \sum_{j \le k-2} \|\Delta V_j\|_F \beta^{k-j-2} \sum_{i=j+1}^{k-1} \|\mathcal{T}_i - \mathcal{T}_{j+1}\| \ge \|C_k - C_k^{\mathrm{ad}}\|_F.$$

With $h_k = \sup_{\|X\|_F=1} \|H_k : (EXE^\top)\|$, $z_0 = 0$ and $z_{k+1} = \|A_k\| z_k + h_k e_k/2$,

$$\|R - R^{\mathrm{ad}}\| \le \|D\psi\| z_T + \tfrac{1}{2} \sup_{\|X\|_F=1} \|D^2\psi : (EXE^\top)\| e_T.$$

The Lyapunov residual $r_k = q_k - (I - \mathcal{T}_k)\widehat{V}_k$ gives covariance error $\|r_k\|_F/(1-\beta)$. These fixed-surrogate comparisons are separate from the primary forecast medians.

F.5 Bounds across 80 checkpoint–schedule pairs

Component	Specification
Selection	One uniform independent loop/checkpoint; PCG64 seed 20260925080, setting/checkpoint order. 22 ellipses, 20 rectangles, 20 restarts, 18 figure-eights; 512 updates. Keep failed and vacuous cases.
Bound	Appendix C: $B = \frac{1}{2}\sum_s (q_s + d_T^s + \sum_k d_k^s)$. $R = B/(\|\widehat{D}\| - B)$, or ∞ for a nonpositive denominator. Non-vacuity: $R < 1$; precision: $R \le 0.25$.
Sampling	$a = \|\bar{D}\|$, $n = \|\widehat{D} - \bar{D}\|$, sampling radius u. For $a > u$, error interval $[\max(0, n-u)/(a+u), (n+u)/(a-u)]$. 100,000 stream-group bootstrap draws; $0.05/80$ per two-sided interval, separate from deterministic bounds.
Summary	Bound quartiles: 12.0/15.0/18.5%; range: 8.4–43.2%. Median error: 7.7%; $n + u \le B$ in all 80 cases. Looseness $R/(n/a)$: median 1.89, IQR 1.67–2.02, range 1.49–6.00.
Precision failures	All 80 bounds are non-vacuous; 20/15/20/17 meet 25%. Eight misses are AdamW figure-eights, chiefly from accumulated one-step defects.

Table 8: All 80 pairs: shape, relative bound / estimated error (%). E/R/S/F: ellipse/rectangle/restart/figure-eight; *: bound above 25%.

Checkpoint	ViT-L/SGD	ViT-L/AdamW	ViT-1.1B/SGD	ViT-1.1B/AdamW
1	F 12.5 / 8.0	E 10.2 / 6.9	S 16.7 / 10.1	E 18.1 / 9.0
2	E 14.2 / 7.3	E 18.0 / 8.8	R 11.9 / 7.5	R 23.8 / 11.7
3	R 19.7 / 9.5	R 18.6 / 9.5	F 10.4 / 5.8	F 24.5 / 11.5
4	R 16.2 / 10.3	S 11.5 / 7.3	E 11.9 / 7.7	F 14.8 / 7.6
5	E 15.8 / 9.5	E 19.3 / 9.4	S 16.3 / 9.9	R 13.9 / 8.5
6	R 16.3 / 8.4	E 14.2 / 8.9	S 14.5 / 7.6	E 12.6 / 7.1
7	E 10.9 / 6.2	F 43.2* / 7.2	S 15.4 / 7.9	E 14.0 / 7.6
8	E 12.9 / 7.6	S 18.5 / 10.5	E 13.9 / 6.8	R 16.2 / 9.9
9	R 17.0 / 9.3	F 40.8* / 7.5	S 20.2 / 10.1	R 12.8 / 7.9
10	F 11.9 / 6.1	S 15.2 / 9.1	S 10.3 / 6.7	R 14.8 / 8.3
11	F 20.1 / 11.0	R 15.8 / 9.0	S 13.2 / 6.8	F 38.4* / 9.0
12	F 11.3 / 5.3	F 36.0* / 8.7	E 13.4 / 7.7	S 12.5 / 7.5
13	E 12.0 / 6.5	R 21.9 / 11.1	F 11.1 / 6.1	R 20.6 / 10.5
14	E 11.2 / 5.9	R 15.7 / 8.5	F 11.4 / 5.4	F 33.6* / 6.2
15	S 18.1 / 8.6	E 12.1 / 7.0	S 8.4 / 3.8	R 24.5 / 13.0
16	R 24.5 / 12.9	E 9.5 / 5.4	E 11.4 / 5.7	E 14.8 / 7.3
17	S 11.6 / 7.2	S 16.2 / 9.7	R 8.5 / 5.6	S 21.9 / 11.0
18	E 8.8 / 5.8	F 31.2* / 7.9	F 18.4 / 11.3	S 16.9 / 8.9
19	E 15.2 / 7.3	F 28.8* / 7.7	S 10.9 / 5.3	R 11.0 / 6.2
20	R 12.0 / 6.1	S 15.4 / 7.8	S 12.9 / 8.0	F 26.4* / 6.9

For this 640-pair frame, the one-sided 95% lower bound is $618/640 = 96.5625\%$: with f_i failures among checkpoint i's eight loops, the all-success probability is $\prod_i (1 - f_i/8)$; 23 failures imply at most $(7/8)^{23} < 0.05$.