MOMENT-ACCURATE GAUSSIAN MIXTURES FOR CONSTANT-STEP STOCHASTIC APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Local Gaussian models of constant-step learning predict output variability and expected losses, but weak convergence alone does not justify these moment predictions. We establish moment-accurate Gaussian mixtures by matching stationary energy with local Ornstein–Uhlenbeck limits, ruling out quadratic tail mass invisible to weak convergence. For step size a, the second-order Wasserstein error is $o(\sqrt{a})$, uniformly over invariant laws, using each law's actual root weights. The assumptions combine confinement, descent, finitely many hyperbolic equilibria and root continuity with finite-variance innovations. The result yields observable covariances, expected objective gaps and first-order mean shifts, while allowing singular covariances, compatible saddles and weights without a limit. For additive noise given by a fixed invertible transform of independent standardized Student t_3 coordinates, symmetry gives an order-sharp $\sqrt{a}$ smooth-test bound. Numerical transport calculations demonstrate the value of root-specific covariances; controlled SGD studies assess observable predictions across step sizes, batch sizes and model geometries.

1 INTRODUCTION

With a fixed learning rate a, stochastic gradient descent (SGD) keeps fluctuating near a trained solution. Local Gaussian models turn curvature and gradient-noise covariance into predictions of output variability and expected objective gaps (Mandt et al., 2017; Yaida, 2019; Liu et al., 2021). These predictions concern moments, so their validity requires more than a Gaussian description of typical fluctuations. When does local Gaussian theory accurately predict these quantities in a landscape with several equilibria?

The first obstacle is that weak convergence can hide the very quantities these predictions require. As $R \to \infty$, $(1 - R^{-2})\mathsf{N}(0,1) + R^{-2}\delta_R$ converges weakly to $\mathsf{N}(0,1)$, while its second moment tends to 2. A vanishing fraction of observations can therefore carry a nonvanishing share of quadratic energy. Single-center moment results address this difficulty under their respective stability and regularity conditions (Dieuleveut et al., 2020; Wei et al., 2025). Multiple equilibria introduce a second obstacle: their local covariances differ, while their occupation weights depend on global dynamics and the selected invariant law.

We study stochastic approximation (SA) with constant step size a,

$$X_{n+1}^{(a)} = X_n^{(a)} - a g(X_n^{(a)}) + a\xi(X_n^{(a)}, \zeta_{n+1}), \qquad X_n^{(a)} \in \mathbb{R}^d, \tag{1}$$

where the observations are i.i.d. and the innovations are conditionally centered. For an invariant draw $X \sim \pi_a$, let J_a label its nearest root m_j and write $w_{a,j} = \mathbb{P}_{\pi_a}(J_a = j)$. At a stable root, set $A_j = Dg(m_j)$ and $C_j = \mathbb{E}[\xi(m_j,\zeta)\xi(m_j,\zeta)^T]$; the local covariance solves $A_j S_j + S_j A_j^T = C_j$. For components with $\liminf_{a\downarrow 0} w_{a,j} > 0$, the moment predictions take the familiar form

$$\begin{aligned} \mathrm{Cov}(X \mid J_a = j) &= aS_j + o(a), \\ \mathrm{Cov}(f(X) \mid J_a = j) &= aDf(m_j)S_j Df(m_j)^T + o(a). \end{aligned} \tag{2}$$

The second formula applies to globally Lipschitz f differentiable at m_j. For gradient drift $g = \nabla F$, the mean objective gap is $\mathbb{E}[F(X) - F(m_j) \mid J_a = j] = a\,\mathrm{tr}(C_j)/4 + o(a)$. Our question is when these moment statements hold jointly across roots under second-moment noise assumptions.

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.

We resolve this question by matching two descriptions of stationary energy. Descent controls squared distance to the root set, while stationarity determines its leading energy balance. Local Ornstein–Uhlenbeck (OU) limits satisfy the same balance, excluding quadratic mass hidden in rare displacements. This proves uniform integrability of $\|Z_a\|^2$, where $Z_a = (X - m_{J_a})/\sqrt{a}$, and upgrades weak Gaussian limits to moment accuracy. Under confinement, descent, finitely many hyperbolic roots and the noise regularity in Assumption 2.1, the resulting mixture has $o(\sqrt{a})$ Wasserstein error, uniformly over invariant laws. Each mixture uses its law's actual finite-a weights: local derivatives determine fluctuations, while global communication and invariant-law selection determine occupation.

Contributions. **Moment-accurate mixtures.** Theorems 3.2–3.3 cover all roots, including compatible saddles, singular covariances and weights without a limit. **Observable predictions.** Section 4 derives covariance, loss and first-order mean formulas; Section 5 quantifies smooth-test accuracy, including a sharp $\sqrt{a}$ endpoint for transformed independent t_3 noise. **Numerical evidence.** Root-specific transport separates the mixture from pooled covariance models; SGD and heavy-tail studies assess the moment predictions using matched OU measurement references.

Relation to existing theory. Classical constant-gain analyses establish stationary asymptotics (Kushner & Huang, 1981; Pflug, 1986; Kushner & Shwartz, 1984; Kushner & Yin, 2003). The finite-variance single-center results of Chen et al. (2022) identify scaled limits through an integral equation and Gaussian uniqueness conditions. Further single-center analyses establish moment expansions, Gaussian approximation and stationary concentration (Dieuleveut et al., 2020; Wei et al., 2025; Wang et al., 2026; Merad & Gaïffas, 2025); Table 1 compares their assumptions and conclusions. Our additional step is uniform quadratic-tail control across all roots using second innovation moments and root continuity. Finite-time W_1 bounds with recursively defined Gaussian covariances address a complementary time scale (Haque et al., 2026).

Nonconvex SA theory characterizes invariant support, saddle exclusion and nonunique invariant laws (Fort & Pagès, 1999; Benaïm, 1996; 1999; Bianchi et al., 2022; Shirokoff & Zaleski, 2025). We resolve moments within the root-labeled components; temporal-average normality addresses a different random object (Yu et al., 2021). The covariance, quarter-trace loss and bias formulas have single-center antecedents (Yaida, 2019; Liu et al., 2021; Dieuleveut et al., 2020; Allmeier & Gast, 2024); energy matching justifies their rootwise use with low-moment remainders. Occupation weights instead use global communication information, as in light- and heavy-tail metastability theory (Freidlin & Wentzell, 2012; Azizian et al., 2024; Wang & Rhee, 2023).

Diffusion and modified equations describe finite-step dynamics (Li et al., 2017; 2021); stationary approximation requires control through the infinite-time limit, as provided by Poisson methods under their assumptions (Mattingly et al., 2010). Our energy argument works directly with discrete invariant laws. Multiplicative noise links innovation and iterate tails (Gürbüzbalaban et al., 2021; Hodgkinson & Mahoney, 2021), while equilibrium manifolds add tangential dynamics (Li et al., 2022). Appendix V develops these comparisons.

Table 1: Gaussian and moment results for SA in their stated model, moment and time settings.

Work	Setting and moment control	Relevant conclusion
Chen et al. (2022)	Single-center stability; additive iid finite-variance noise	Stationary scaling limit; Gaussian identification under uniqueness conditions
Dieuleveut et al. (2020)	Strong convexity; higher regularity and moments	Stationary moment and bias expansions
Wei et al. (2025), Thms. 3.4–3.5	Strong convexity; L^p regularity; joint long-time/small-step limit	CLT at $p = 2$; convex-set bounds at $p \geq 4$
Wang et al. (2026), Thm. 3.8	Contractive setting; additive noise with finite third moment	Stationary W_1 rate; derived scaled third-moment bound
This paper	Finite hyperbolic roots; descent; second innovation moments and root continuity	Law-uniform all-root moments and actual-weight W_2 approximation

2 MODEL AND ASSUMPTIONS

The observations (ζ_n) are i.i.d. and the measurable innovation map ξ does not depend on a. Write $P_a f(x) = \mathbb{E} f(x - ag(x) + a\xi(x, \zeta_1))$ and $C(x) = \mathbb{E}[\xi(x, \zeta_1)\xi(x, \zeta_1)^{\mathsf{T}}]$. The deterministic flow on physical time $t = na$ is $\dot{x} = -g(x)$. We use $W_2(\mu, \nu)^2 = \inf_{\gamma \in \Pi(\mu,\nu)} \int \|x - y\|^2 \, \gamma(\mathrm{d}x, \mathrm{d}y)$, where $\Pi(\mu, \nu)$ denotes all couplings. Throughout, constants are independent of sufficiently small a and of the selected invariant law.

The assumptions combine confinement, locally linearizable roots and a descent energy with second-moment noise control and root continuity.

Assumption 2.1 (Confinement, descent and local regularity)**.** The following conditions hold.

1. For every x, $\mathbb{E}\xi(x, \zeta_1) = 0$ and
$$\operatorname{tr} C(x) \le c_0 + c_1 \|g(x)\|^2. \tag{3}$$
The kernels P_a are Feller for sufficiently small a. At every root m_j, $C(x) \to C_j := C(m_j)$ as $x \to m_j$.

2. The drift g is C^2 and globally Lipschitz, and
$$\langle x, g(x) \rangle \ge \kappa \|x\|^2 - b, \qquad \kappa > 0. \tag{4}$$
Its root set $\mathcal{M} = \{m_1, \dots, m_\ell\}$ is finite and nonempty. Each $A_j = Dg(m_j)$ is hyperbolic, meaning $\operatorname{spec}(A_j) \cap i\mathbb{R} = \varnothing$.

3. There is $U \in C^2(\mathbb{R}^d)$ with
$$\begin{gathered} |U(x)| \le C_U(1 + \|x\|^2), \qquad \sup_x \|D^2 U(x)\|_{\mathrm{op}} \le L_U, \\ \langle \nabla U(x), g(x) \rangle \ge c_U \|g(x)\|^2, \end{gathered} \tag{5}$$
where $c_U > 0$.

Concentration and root continuity supply stationary innovation-tail control (Lemma A.3), even without uniform integrability on fixed root neighborhoods (Appendix T). For additive noise, a finite second moment suffices for the noise conditions. For a gradient drift $g = \nabla F$, one may take $U = F$ when the displayed growth and smoothness conditions hold.

Let $\mathcal{S} = \{j : \min_{\lambda \in \operatorname{spec}(A_j)} \Re\lambda > 0\}$ and $\mathcal{U} = \{1, \dots, \ell\} \setminus \mathcal{S}$. Thus $\mathcal{S}$ indexes the attracting equilibria of $\dot{x} = -g(x)$. Break nearest-root ties by the smallest index and denote the resulting Voronoi cells by $\mathcal{C}_j$. For $X \sim \pi_a$, define $J_a = j$ on $\mathcal{C}_j$, $X_{\mathcal{M}} = m_{J_a}$, $Z_a = (X - m_{J_a})/\sqrt{a}$, and $w_{a,j} = \pi_a(\mathcal{C}_j)$. All roots, including saddles, enter this labeling convention.

3 GAUSSIAN MIXTURES AND MOMENT ACCURACY

We first locate stationary mass and identify its fluctuation scale around the roots. This concentration estimate supplies the starting point for local Gaussian limits.

Theorem 3.1 (Existence and concentration)**.** *Under Assumption 2.1, an invariant probability exists for every sufficiently small a. Every invariant probability has a finite second moment and satisfies*
$$\mathbb{E}\|X\|^2 \le K, \qquad \mathbb{E}\|g(X)\|^2 \le Ka, \qquad \mathbb{E}\operatorname{dist}(X, \mathcal{M})^2 \le Ka. \tag{6}$$
The constants are uniform over invariant-law selections. In particular, (J_a, Z_a) is tight.

From concentration to moments. Concentration fixes the scale, but rare excursions could still carry a nonvanishing share of the rescaled second moment. The descent dissipation $D(x) = \langle \nabla U(x), g(x) \rangle$ controls this contribution: the global error bound $\|g(x)\| \ge c_g \operatorname{dist}(x, \mathcal{M})$ gives $D(X)/a \ge c_U c_g^2 \|Z_a\|^2$. Writing $E_j = D^2 U(m_j)$, stationarity and Lemma A.3 yield
$$\frac{\mathbb{E}D(X)}{a} = \frac{1}{2} \sum_j w_{a,j} \operatorname{tr}(E_j C_j) + o(1). \tag{7}$$

Along any convergent subsequence of root weights and localized laws, the OU moment identity gives exactly the same limiting energy. Agreement of the distributional limit and its mean excludes hidden energy in the tails, yielding uniform integrability of $\|Z_a\|^2$. Appendices B–D prove this matching and its uniformity over invariant laws.

Local Gaussian components. At a stable root, A_j has eigenvalues with positive real parts, and the covariance is the unique solution

$$A_j S_j + S_j A_j^T = C_j, \qquad \Gamma_j = \mathsf{N}(0, S_j). \tag{8}$$

To include saddle components, we retain fluctuations on their stable subspaces and identify when noise excites an unstable direction. Let P_j^s and $P_j^u = I - P_j^s$ be the real spectral projectors onto the positive- and negative-real-part generalized eigenspaces of A_j. They need not be orthogonal. Define

$$\bar{S}_j = \int_0^\infty e^{-A_j t} P_j^s C_j (P_j^s)^T e^{-A_j^T t}\, \mathrm{d}t, \qquad \bar{\Gamma}_j = \mathsf{N}(0, \bar{S}_j). \tag{9}$$

The integral converges on the stable subspace; if that subspace is empty, $\bar{\Gamma}_j = \delta_0$. At a stable root, $P_j^s = I$, so $\bar{S}_j = S_j$ and $\bar{\Gamma}_j = \Gamma_j$.

The next theorem supplies the tail control needed to turn these local Gaussian descriptions into covariance predictions.

Theorem 3.2 (Rootwise moment accuracy)**.** *Under Assumption 2.1,*

$$\lim_{R\to\infty} \limsup_{a\downarrow 0} \sup_{\pi\in\mathcal{I}_a} \mathbb{E}_\pi[\|Z_a\|^2 \mathbf{1}\{\|Z_a\| > R\}] = 0, \tag{10}$$

where $\mathcal{I}_a$ is the set of invariant probabilities. At every root, uniformly over $\pi \in \mathcal{I}_a$,

$$\mathbb{E}[\mathbf{1}_{\mathcal{C}_j}(X) Z_a Z_a^T] - w_{a,j}\bar{S}_j \to 0, \qquad \mathbb{E}[\mathbf{1}_{\mathcal{C}_j}(X) Z_a] \to 0. \tag{11}$$

For any selection with $\liminf_{a\downarrow 0} w_{a,j} > 0$,

$$W_2\big(\mathcal{L}(Z_a \mid J_a = j), \bar{\Gamma}_j\big) \to 0, \qquad a^{-1}\,\mathrm{Cov}(X \mid J_a = j) \to \bar{S}_j. \tag{12}$$

Such a component necessarily satisfies $P_j^u C_j (P_j^u)^T = 0$.

Thus a saddle can retain mass only if its root noise acts within its stable subspace; Appendix J gives an exact example.

We now assemble the local predictions using each invariant law's actual root weights. Matching observations and Gaussian samples within the same root makes the transport guarantee explicit.

Theorem 3.3 (All-root Gaussian approximation)**.** *Under Assumption 2.1, let $\bar{G}_{a,\pi} = \sum_j w_{a,j}\mathsf{N}(m_j, a\bar{S}_j)$. Define the same-label scaled transport cost by*

$$D_a(\pi)^2 = \sum_{j: w_{a,j}>0} w_{a,j} W_2^2\big(\mathcal{L}(Z_a \mid J_a = j), \bar{\Gamma}_j\big). \tag{13}$$

Then

$$\sup_{\pi\in\mathcal{I}_a} D_a(\pi) \longrightarrow 0, \qquad W_2(\pi, \bar{G}_{a,\pi}) \le \sqrt{a}\, D_a(\pi). \tag{14}$$

In particular,

$$\sup_{\pi\in\mathcal{I}_a} a^{-1/2} W_2(\pi, \bar{G}_{a,\pi}) \longrightarrow 0. \tag{15}$$

If $P_j^u C_j (P_j^u)^T \neq 0$, then $w_{a,j} \to 0$ uniformly.

The same-label cost preserves each root's fluctuation model; rescaling its couplings gives the original-scale bound. The weights may vary without converging, covering both communicating wells and disjoint invariant regions (Appendices K and S.2).

Noise excitation provides a useful specialization in which only stable components retain asymptotic mass.

Assumption 3.4 (Excitation of unstable roots)**.** For every $j \in \mathcal{U}$,

$$P_j^u C_j (P_j^u)^T \neq 0. \tag{16}$$

A sufficient condition is excitation of a left eigenvector:

$$A_j^T v = \lambda v, \qquad \Re\lambda < 0, \qquad v^* C_j v > 0. \tag{17}$$

The projection condition also covers defective blocks missed by this sufficient condition (Lemma B.2). For the following marked W_2, use squared cost $\mathbf{1}\{j \neq k\} + \|z - y\|^2$ between (j, z) and (k, y).

Corollary 3.5 (Stable components under excitation)**.** *Under Assumptions 2.1 and 3.4, let* $s_a = \sum_{j\in\mathcal{S}} w_{a,j}$. *Then, uniformly,*

$$s_a \to 1, \quad W_2(\mathcal{L}(J_a, Z_a), Q_a) \to 0, \quad Q_a = \sum_{j\in\mathcal{S}} \frac{w_{a,j}}{s_a}(\delta_j \otimes \Gamma_j), \tag{18}$$

and $\mathbb{E}[\mathbf{1}\{J_a \in \mathcal{U}\}\|Z_a\|^2] \to 0$.

4 Observable covariances, losses and mean shifts

The moment theorem turns local parameter fluctuations into predictions for outputs, losses and systematic mean shifts. We first propagate covariance through a differentiable observable.

Corollary 4.1 (Observable fluctuations and covariance)**.** *Assume Theorem 3.2 and fix a root with* $\liminf_a w_{a,j} > 0$. *For measurable* $f : \mathbb{R}^d \to \mathbb{R}^k$ *differentiable at* m_j,

$$\mathcal{L}\left(\frac{f(X) - f(m_j)}{\sqrt{a}} \,\middle|\, \mathcal{C}_j\right) \Rightarrow \mathsf{N}(0, Df(m_j)\bar{S}_j Df(m_j)^{\mathsf{T}}). \tag{19}$$

If f *is also globally Lipschitz, convergence holds in* W_2 *and* $a^{-1}\operatorname{Cov}(f(X) \mid \mathcal{C}_j) \to Df(m_j)\bar{S}_j Df(m_j)^{\mathsf{T}}$.

At a minimum, the objective has no linear term, so its leading fluctuation is quadratic. Second-moment accuracy makes the resulting prediction valid for the expected loss as well as its distribution.

Corollary 4.2 (Loss law and mean)**.** *Assume Theorem 3.2, let* $g = \nabla F$ *and fix a stable root with* $\liminf_a w_{a,j} > 0$. *For* $H_j = D^2F(m_j) \succ 0$ *and* $G_j \sim \Gamma_j$,

$$\frac{F(X) - F(m_j)}{a} \mid \mathcal{C}_j \Rightarrow \tfrac{1}{2} G_j^{\mathsf{T}} H_j G_j, \tag{20}$$

$$\frac{\mathbb{E}[F(X) - F(m_j) \mid \mathcal{C}_j]}{a} \to \tfrac{1}{2}\operatorname{tr}(H_j S_j) = \tfrac{1}{4}\operatorname{tr} C_j. \tag{21}$$

The limit law is $\frac{1}{2}\sum_r \mu_r \chi^2_{1,r}$ *with independent chi-squares, where* μ_r *are the eigenvalues of* $S_j^{1/2} H_j S_j^{1/2}$, *including zeros. Without individual weight lower bounds, centering at the nearest root gives*

$$\frac{\mathbb{E}[F(X) - F(m_{J_a})]}{a} - \frac{1}{4}\sum_{j=1}^{\ell} w_{a,j} \operatorname{tr} C_j \to 0. \tag{22}$$

The global Lipschitz property controls observable growth; the bounded Hessian of F controls quadratic loss by squared displacement. Appendix I proves the resulting expectation limits. The quarter-trace coefficient agrees with the local harmonic prediction of Yaida (2019); here its remainder and conditional interpretation follow from the moment theorem.

A centered Gaussian describes fluctuations of order $\sqrt{a}$, while nonlinear drift can shift the mean at order a. The next expansion identifies this smaller systematic effect from the same rootwise covariance.

Theorem 4.3 (Rootwise mean expansion). *Under Assumption 2.1, fix any root j and define*

$$(t_j)_r = \operatorname{tr}[D^2 g_r(m_j)\bar{S}_j], \qquad b_j = -\tfrac{1}{2}A_j^{-1}t_j.$$

Uniformly over invariant-law selections,

$$\mathbb{E}[\mathbf{1}_{\mathcal{C}_j}(X - m_j)] = a w_{a,j} b_j + o(a). \tag{23}$$

For $f : \mathbb{R}^d \to \mathbb{R}^k$ of class C^2 with bounded coordinatewise Hessians, let $k_{f,j} = Df(m_j)b_j + \frac{1}{2}(D^2 f(m_j) : \bar{S}_j)$. Then

$$\mathbb{E}[\mathbf{1}_{\mathcal{C}_j}\{f(X) - f(m_j)\}] = a w_{a,j} k_{f,j} + o(a). \tag{24}$$

If $\liminf_a w_{a,j} > 0$, the conditional means are $m_j + ab_j + o(a)$ and $f(m_j) + ak_{f,j} + o(a)$, respectively.

Here the rth coordinate of $D^2 f : \bar{S}_j$ is $\operatorname{tr}[D^2 f_r \bar{S}_j]$. The proof refines the cutoff flux to $o(a^2)$ and expands the rootwise drift balance (Appendix E), extending single-center bias expansions to rootwise low-moment asymptotics (Dieuleveut et al., 2020; Allmeier & Gast, 2024). A nonzero mean shift at a root whose weight stays bounded away from zero contributes order $\sqrt{a}$ to the same-label scaled error (Appendix E.1). For a fixed test loss, the coefficient is

$$k_{\mathrm{test},j} = \nabla \ell_{\mathrm{test}}(m_j)^T b_j + \tfrac{1}{2}\operatorname{tr}[D^2 \ell_{\mathrm{test}}(m_j)\bar{S}_j]. \tag{25}$$

The first term vanishes for the full objective generating a gradient drift, yielding (21). For a scalar globally Lipschitz observable with bounded Hessian, these coefficients also separate global variance into between-root variation, within-root Gaussian variance and a mean-shift correction (Corollary E.3).

For each fixed independent with-replacement batch size B, the matrices $C_j, \bar{S}_j$ and coefficients $b_j, k_{f,j}$ scale by $1/B$. Thus fluctuations scale as $\sqrt{a/B}$ and leading mean shifts as a/B.

Computing predictions. Evaluate A_j, C_j at each relevant root, solve the Lyapunov equation on its stable subspace, and propagate covariance through observable derivatives; second derivatives give $b_j, k_{f,j}$. These coefficients give leading conditional predictions; mixture aggregation also uses invariant-law weights from a global model or ergodically justified time occupations (Appendix L). Output intervals quantify algorithmic variability about fixed-root predictions.

5 Noise tails and smooth accuracy

The preceding results justify moments and qualitative W_2 accuracy. Additional noise structure gives explicit rates for bounded smooth tests of the root-scaled law. For $0 < \delta \le 1$, let $\mathcal{H}_{2,\delta}$ contain real functions satisfying $\|h\|_\infty + \|\nabla h\|_\infty + \|D^2 h\|_\infty + [D^2 h]_\delta \le 1$. The last term is the global Hölder seminorm, and $\Gamma_j h = \int h(z)\,\Gamma_j(\mathrm{d}z)$. At a stable root, measure the unnormalized local error by

$$\mathcal{E}_{a,j} = \sup_{h \in \mathcal{H}_{2,\delta}} \left| \mathbb{E}[\mathbf{1}_{\mathcal{C}_j}(X) h((X - m_j)/\sqrt{a})] - w_{a,j}\Gamma_j h \right|. \tag{26}$$

Corollary 5.1 (Additive-noise accuracy). *Under Assumption 2.1, centered additive innovations with a finite $(2+\delta)$ moment give $\mathcal{E}_{a,j} = O(a^{\delta/2})$ at every stable root. For a fixed invertible linear transform of d independent standardized t_3 coordinates, $\mathcal{E}_{a,j} = O(\sqrt{a})$ on $\mathcal{H}_{2,1}$.*

Despite infinite third absolute moments, the t_3 bound follows from symmetry and OU smoothing; invertibility ensures $C_j \succ 0$ (Corollary H.3). In the scalar linear example, the same cosine test has error $O(a)$ for Gaussian or Rademacher innovations and order $\sqrt{a}$ for t_3 noise (Proposition H.4). For state-dependent noise, a uniformly bounded local $(2+\delta)$ moment and a locally β-Hölder covariance give $\mathcal{E}_{a,j} = O(a^{\delta/2} + a^{\beta/2} + \sqrt{a})$ at stable roots, for $0 < \delta, \beta \le 1$ (Theorem F.1). Appendices F–H derive these smooth-test rates, tail moduli and quantitative exclusion of noise-excited saddles.

6 Numerical studies

We first test root-specific moments and transport, then examine covariance, loss and mean predictions in controlled SGD. Heavy-tail and communication studies connect these measurements to the

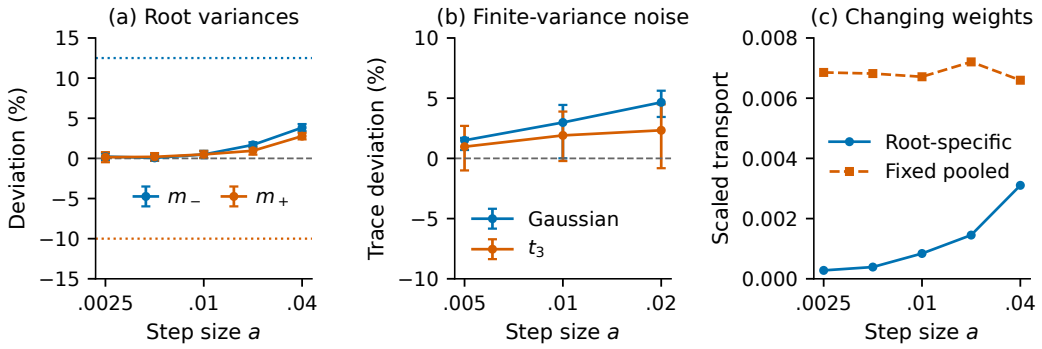

Figure 1: **Root-specific moments and transport.** (a) Scaled variance deviations from each root's OU target; dotted lines show the fixed equal-weight pooled reference's limiting errors. Bars are one standard error across 16 runs. (b) Second-moment trace deviations in the nongradient model; medians and interquartile ranges retain t_3 dispersion. (c) Deterministic same-label scaled transport with alternating weights, using root-specific covariances or the same fixed pooled target.

noise assumptions and occupation weights. Predictions use roots, derivatives and noise covariances evaluated independently of the measured trajectories. Covariance and loss settings use 16 independent paths; mean confirmation uses 32 per step. Error bars denote one standard error across paths unless stated otherwise. Percentage deviations are $100(\widehat{\theta}/\theta_{\mathrm{pred}} - 1)$ for an estimate $\widehat{\theta}$ and its prediction θ_{pred}.

6.1 Root-specific moments and transport

The two-dimensional example of Proposition K.1 has invariant rectangles surrounding roots with $S_- = \mathrm{diag}(1/75, 1/50)$ and $S_+ = \mathrm{diag}(1/60, 1/50)$. At $a = 0.0025$, measured scaled first-coordinate variances are 0.013363 ± 0.000054 and 0.016691 ± 0.000104. A fixed equal-weight pooled reference assigns 0.015 to both roots, giving limiting relative errors of $+12.5\%$ and -10%. Root labels therefore matter even when an aggregate appears accurate (Figure 1a). Proposition K.2 bounds the scaled initialization error by 10^{-6}. We also compute the invariant product laws by a refined one-dimensional transfer discretization. For the same-label cost D_a in (13), select invariant mixtures whose root weights alternate between $1/4$ and $3/4$. As a decreases from 0.04 to 0.0025, D_a falls from 0.00310 to 0.00028, while the fixed pooled cost remains near 0.0069 (Figure 1c). Even an optimally chosen shared covariance has a positive limiting same-label cost of 0.005902 along this sequence (Proposition U.1). Appendix U gives the quantile calculation and grid refinement.

6.2 Neural SGD: prediction accuracy and measurement variability

We train 64 input weights in a four-unit $\tanh$ network with fixed output weights on the UCI optical digits 3/8 task. The writer-disjoint split has 769 training and 357 test examples, represented by 16 training-fitted principal components (Alpaydin & Kaynak, 1998). Ridge 0.02 confines the model; a fixed random tilt $10^{-4} r^T x$ makes roots generically nondegenerate, and bounded dither with covariance $0.015^2 I$ supplies noise excitation (Proposition N.1). Twenty optimization starts locate one numerical root m, with Hessian $H = D^2 F(m)$ having spectrum $[0.02435, 0.24009]$. We freeze this root and 12 test-logit probes before measuring fluctuations.

Per-run mean absolute relative error (MARE) averages absolute variance errors relative to the 12 probe predictions. Aggregate MARE averages each probe's variance across paths before taking absolute relative errors. At the smallest tested normalized step, $a\lambda_{\max}(H) = 0.000625$, per-run MARE is 6.1–7.0% and aggregate MARE is 0.9–2.0%; the $B = 1$ mean variance and objective-gap ratios are 0.986 and 1.029. Stationary OU paths with the same save schedule give per-run MARE $5.89 \pm 0.19\%$, showing the contribution of finite-trajectory variability (Figure 2b). Fourfold longer

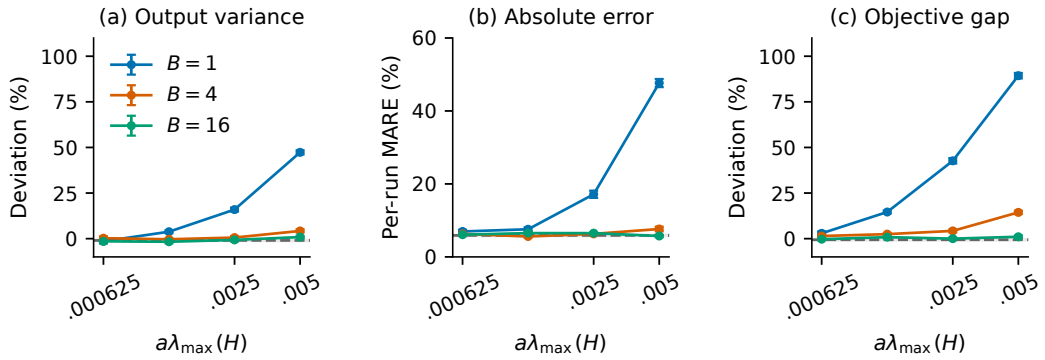

Figure 2: **Neural predictions across step sizes.** Output-variance and actual objective-gap deviations share the $[-10, 110]\%$ range; MARE starts at zero. All four steps, three batches and 16 paths per setting are included. Dashed lines and bands show the mean and one standard error of 64 matched stationary OU paths, whose loss is quadratic.

Table 2: Small-step prediction ratios across models and batches at $a\lambda_{\max}(H) = 0.000625$. Entries are means $\pm$ one standard error across 16 independent paths; all use physical exposure 8000. Output ratios average the 12 fixed probes within each path.

Model	Output variance $B = 1$	Objective gap $B = 1$	Output variance $B = 16$	Objective gap $B = 16$
Digits-256	1.012 ± 0.005	1.030 ± 0.005	0.998 ± 0.003	0.997 ± 0.003
Digits-1024	1.024 ± 0.003	1.027 ± 0.003	0.996 ± 0.005	0.999 ± 0.003
WDBC-256	1.003 ± 0.003	1.012 ± 0.005	0.997 ± 0.003	1.003 ± 0.004

exposure reduces SGD MARE to 3.0–3.2% for $B = 1, 16$, alongside OU $3.34 \pm 0.10\%$. The $B = 1$ objective-gap ratio remains 1.028 ± 0.004 (Figure 16).

The full covariance is important: small-step logit MARE is 6.1–7.0%, compared with 228–233% for a diagonal-noise reference and 185–190% for an isotropic reference (Table 4). At the largest normalized step, $a\lambda_{\max}(H) = 0.005$, with $B = 1$, output variances and objective gaps exceed their predictions by 47.3% and 89.3%. Replacing the OU covariance by the exact discrete-linear reference changes logit MARE from 47.65% to 47.41% at this setting. The frozen quadratic model's relative loss correction is at most 0.251% (Proposition O.1), far below the observed objective-gap deviation. Paired-state diagnostics, augmentation ablations and longer exposures characterize these finite-step departures (Appendix R).

Model dimension and data. An independent 192-path study covers 256/1024-parameter digits networks and a 256-parameter Wisconsin Diagnostic Breast Cancer (WDBC) classifier. Small-step MARE is 2.74–3.69%, alongside OU 2.91–3.15%, with residual mean-ratio biases of roughly 2–3% for digits-1024 (Table 2). Appendix S.1 reports all steps and batches.

6.3 Nonlinear observables, objective gaps and mean shifts

Consider uniform sampling from four component losses,

$$F(x) = \tfrac{1}{2}\|x\|^2 - 2\log\cosh x_1 - 1.5\log\cosh x_2,$$
$$\ell_{s,t}(x) = F(x) + sb^T x + 0.6t\log\cosh(v^T x), \quad b = (0.5, 0.2)^T, \quad v = (-0.3, 0.7)^T, \tag{27}$$

where $s, t \in \{-1, 1\}$ are independent. The single-sample covariance is $C_1(x) = bb^T + 0.36\tanh^2(v^T x)vv^T$. At the positive root m, write $H = D^2F(m)$ and let S_1 solve $HS_1 + S_1H = C_1(m)$. For $f_2(x) = \tanh(x_1 - x_2/2)$ and $a = 0.005$, variance ratios are 1.009 ± 0.007, 1.003 ± 0.005 and 0.999 ± 0.005 for $B = 1, 4, 16$. The actual mean objective-gap ratios are

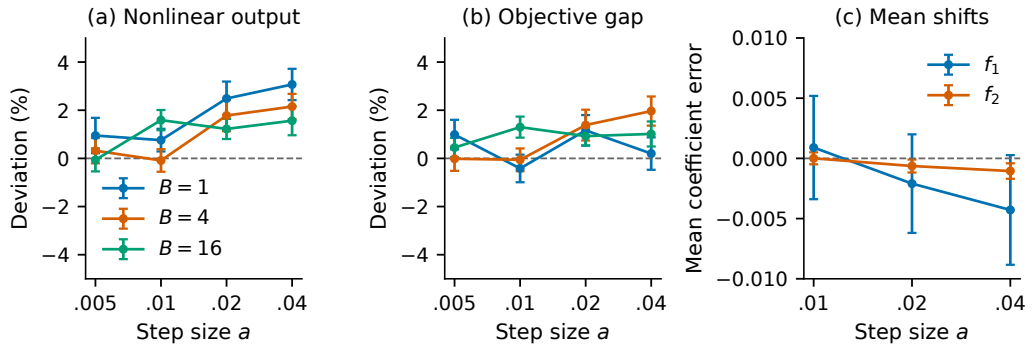

Figure 3: **Predictions beyond parameter covariance.** Percentage deviations of (a) nonlinear output variance from $(a/B)Df_2(m)S_1Df_2(m)^T$ and (b) actual objective gap from $(a/4B)\,\mathrm{tr}\,C_1(m)$ share the range $[-5,5]\%$. (c) Signed mean-coefficient errors, with $T = 1600/a$. Bars are one standard error across 16 paths in (a,b) and 32 in (c); the two outputs share paths. Zero denotes the prediction.

1.010 ± 0.006, 1.000 ± 0.005 and 1.005 ± 0.004 (Figure 3a,b). The generalized chi-square prediction also describes the loss distribution (Figure 9).

For $f = (x_1 + x_2, \tanh(x_1 - x_2/2))$, Theorem 4.3 predicts $k_f = (-0.055598, -0.027255)^T$ at the positive root for $B = 1$. Writing $T = aN$ for the physical exposure of N retained updates, a 96-path measurement fixes $aT = 1600$, with 32 independent paths at each of $a = 0.04, 0.02, 0.01$. At the smallest step, the signed estimates are -0.05470 ± 0.00429 and -0.027249 ± 0.000498 (Figure 3c). Appendix R.3 also reports an independent 48-path measurement with larger discrepancies, paired burn-in windows and an empirical drift decomposition.

6.4 Heavy-tailed fluctuations and global occupation

The nonlinear nongradient model (72) compares matched Gaussian and t_3 innovations across three steps and 96 paths, labeling all nine roots. At $a = 0.005$, median stable root-centered trace ratios are 1.015 and 1.010, respectively (Figure 1b). Figure 11 reports the broader t_3 dispersion and all-root truncated moments; Figure 4 gives the separate analytic smooth-test comparison.

Global communication supplies occupation information in the separate scalar model $g(x) = x + 5\cos x$. Long paths give 388 and 413 transitions at $a = 0.04$ and 0.02, with occupation and directional rates agreeing with the heavy-tail proxy at the measurement resolution (Appendix S.2).

7 Discussion

Stationary energy matching turns local Gaussian limits into reliable moment asymptotics across multiple equilibria. It gives covariance, loss and mean-shift formulas a common foundation under second innovation moments, separating local coefficients from global occupation weights.

The heterogeneous-root comparison links this distinction directly to transport: even the best common covariance retains a positive limiting same-label cost. In neural SGD, matched OU paths calibrate finite-trajectory variability, while step-size and batch-size comparisons identify departures from the local moment predictions.

The present theory concerns stationary laws near finite hyperbolic roots. Extending energy matching to equilibrium manifolds would connect normal moment control with tangential dynamics (Li et al., 2022). Quantitative W_2 bounds and equilibration estimates would connect these stationary predictions to prescribed training horizons.

REPRODUCIBILITY STATEMENT

The appendices contain complete proofs, model-specific certificates and experimental protocols. Source code, frozen configurations, seeds, per-run results and deterministic reference calculations accompany the paper.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript writing and editing, literature search and identification, drafting portions of the manuscript, experimental design and code development, exploration of new research directions based on existing results, and the development and verification of mathematical arguments and proofs.

We have reviewed all AI-assisted work. Suggested references were checked against original sources, experimental code and numerical results were executed and verified by the authors, and mathematical claims, derivations, and proofs were independently checked for correctness.

We take responsibility for the final content of this work, including all text, claims, proofs, code, experiments, and other artifacts produced with the aid of generative AI.

REFERENCES

Sebastian Allmeier and Nicolas Gast. Computing the bias of constant-step stochastic approximation with Markovian noise. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4379. URL https://papers.nips.cc/paper_files/paper/2024/hash/f949c1f490beb42124a267b7476cd353-Abstract-Conference.html.

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. URL https://doi.org/10.24432/C50P49.

Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. What is the long-run distribution of stochastic gradient descent? A large deviations analysis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2168–2229, 2024.

Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The global convergence time of stochastic gradient descent in non-convex landscapes: Sharp estimates via large deviations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 1982–2044, 2025.

Michel Benaïm. A dynamical system approach to stochastic approximations. *SIAM Journal on Control and Optimization*, 34(2):437–472, 1996.

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pp. 1–68. Springer, Berlin, 1999.

Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, 30: 1117–1147, 2022.

Zaiwei Chen, Shancong Mou, and Siva Theja Maguluri. Stationary behavior of constant stepsize SGD-type algorithms: An asymptotic characterization. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(1):1–24, 2022. doi: 10.1145/3508039.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348–1382, 2020.

Jean-Claude Fort and Gilles Pagès. Asymptotic behavior of a Markovian stochastic algorithm with constant step. *SIAM Journal on Control and Optimization*, 37(5):1456–1482, 1999.

Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer, Heidelberg, 3 edition, 2012.

Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3964–3975, 2021. URL https://proceedings.mlr.press/v139/gurbuzbalaban21a.html.

Shaan Ul Haque, Zedong Wang, Zixuan Zhang, and Siva Theja Maguluri. How accurately can a Gaussian approximate stochastic approximation iterates?, 2026. URL https://arxiv.org/abs/2602.13906v2. Version 2, revised May 30, 2026.

Liam Hodgkinson and Michael W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4262–4274, 2021. URL https://proceedings.mlr.press/v139/hodgkinson21a.html.

Michael Högele and Ilya Pavlyukevich. Metastability in a class of hyperbolic dynamical systems perturbed by heavy-tailed Lévy type noise. *Stochastics and Dynamics*, 15(3):1550019, 2015. doi: 10.1142/S0219493715500197.

Peter Imkeller and Ilya Pavlyukevich. Metastable behaviour of small noise Lévy-driven diffusions. *ESAIM: Probability and Statistics*, 12:412–437, 2008.

Yuri Kifer. A discrete-time version of the Wentzell–Freidlin theory. *The Annals of Probability*, 18 (4):1676–1692, 1990.

Harold J. Kushner and Hai Huang. Asymptotic properties of stochastic approximations with constant coefficients. *SIAM Journal on Control and Optimization*, 19(1):87–105, 1981.

Harold J. Kushner and Adam Shwartz. Weak convergence and asymptotic properties of adaptive filters with constant gains. *IEEE Transactions on Information Theory*, 30(2):177–182, 1984. doi: 10.1109/TIT.1984.1056897.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, New York, 2 edition, 2003.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, 2 edition, 2013. doi: 10.1007/978-1-4419-9982-5.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2101–2110, 2017. URL https://proceedings.mlr.press/v70/li17f.html.

Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/69f62956429865909921fa916d61c1f8-Abstract.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://openreview.net/forum?id=siCt4xZn5Ve.

Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7045–7056, 2021. URL https://proceedings.mlr.press/v139/liu21ad.html.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017. URL https://jmlr.org/papers/v18/17-214.html.

Jonathan C. Mattingly, Andrew M. Stuart, and Michael V. Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM Journal on Numerical Analysis*, 48(2):552–577, 2010. doi: 10.1137/090770527.

Ibrahim Merad and Stéphane Gaïffas. Convergence and concentration properties of constant step-size SGD through Markov chains. *Electronic Journal of Statistics*, 19(2):5843–5894, 2025.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Georg Ch. Pflug. Stochastic minimization with constant step-size: Asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

David Shirokoff and Philip Zaleski. Convergence of Markov chains for constant step-size stochastic gradient descent with separable functions. *SIAM Journal on Applied Dynamical Systems*, 24(3):2005–2043, 2025. doi: 10.1137/24M1694598.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

Xingyu Wang and Chang-Han Rhee. Large deviations and metastability analysis for heavy-tailed dynamical systems. arXiv preprint arXiv:2307.03479, 2023.

Xingyu Wang, Sewoong Oh, and Chang-Han Rhee. Eliminating sharp minima from SGD with truncated heavy-tailed noise. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2102.04297.

Zedong Wang, Yuyang Wang, Ijay Narang, Felix Wang, Yuzhou Wang, and Siva Theja Maguluri. Constant-stepsize stochastic approximation: Finite-time convergence, Gaussian approximation, and tail bounds. arXiv preprint arXiv:2602.13960v2, 2026.

Ziyang Wei, Jiaqi Li, Zhipeng Lou, and Wei Biao Wu. Gaussian approximation and concentration of constant learning-rate stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast cancer wisconsin (diagnostic). UCI Machine Learning Repository, 1993. URL https://doi.org/10.24432/C5DW2B.

Sho Yaida. Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkNksoRctQ.

Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A. Erdogdu. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2006.07904.

GUIDE TO THE APPENDICES

The proofs follow the dependence of the results: concentration and stationary innovation tails precede local OU limits, energy matching, transport and mean expansions. The experimental appendices give the model certificates and measurement protocols behind the main figures.

Mathematical question	Location
Existence, concentration and stationary innovation tails	Appendix A
Finite OU measures and generalized unstable directions	Appendix B
Root localization, energy matching and Gaussian transport	Appendices C–D
Rootwise mean shifts and global observable variance	Appendix E
Smooth-test rates, unstable mass and the symmetric t_3 endpoint	Appendices F–H
Observable and loss consequences	Appendix I
Exact saddle components, selected weights and local equilibration	Appendices J–K
Root continuity without fixed-neighborhood uniform integrability	Appendix T
Experimental question	**Location**
Same-label invariant-law transport and numerical resolution	Appendix U
Structural assumptions for the augmented neural model	Appendix N
Recursions, deterministic references and independent replication	Appendix Q
OU measurement references, neural augmentation and observation windows	Appendix R
Independent mean measurements and drift decomposition	Appendix R.3
Neural dimension and data comparisons	Appendix S.1
Occupation and directional transition rates at long horizons	Appendix S.2

A EXISTENCE, ENERGY AND LOCALIZATION

We prove Theorem 3.1. Constants K may change between lines; all bounds are uniform over invariant laws and sufficiently small a. Let $V(x) = \|x\|^2$. Conditional centering gives the exact identity

$$P_aV(x) - V(x) = -2a\langle x, g(x)\rangle + a^2\{\|g(x)\|^2 + \operatorname{tr} C(x)\}. \tag{28}$$

Global Lipschitz continuity implies $\|g(x)\|^2 \le K(1 + \|x\|^2)$. Together with (3) and (4), this gives, after reducing the step-size upper bound,

$$P_aV(x) \le (1 - \kappa a)V(x) + Ka. \tag{29}$$

Starting at any fixed point, iteration yields $\sup_n \mathbb{E}V(X_n) \le V(X_0) + K/\kappa$. The Cesàro averages of the laws are tight. Any weak limit is invariant: for bounded continuous f, the telescoping difference between the average law and its image under P_a tends to zero, and the Feller property permits passage to the weak limit. Lower semicontinuity of V gives a finite second moment.

To cover every invariant probability, we use a bounded truncation before integrating the drift inequality.

Lemma A.1 (Truncated Foster bound)**.** *Let V be finite and nonnegative. If $PV \le \rho V + b$ with $0 < \rho < 1$ and $0 \le b < \infty$, every invariant probability satisfies $\pi V \le b/(1 - \rho)$.*

Proof. Set $V_N = V \wedge N$. Concavity and monotonicity of truncation give $PV_N \le (PV) \wedge N \le (\rho V + b) \wedge N$. Invariance for the bounded V_N implies

$$\int \{V \wedge N - (\rho V + b) \wedge N\}\, d\pi \le 0.$$

The integrand plus b is nonnegative and converges pointwise to $(1 - \rho)V$. Fatou's lemma yields $(1 - \rho)\pi V \le b$, without assuming integrability in advance. □

Apply Lemma A.1 to (29); every invariant probability has $\mathbb{E}V(X) \le K/\kappa$.

Write $\Delta = X^+ - X = a\{\xi(X, \zeta_1) - g(X)\}$. Taylor's upper bound for U and conditional centering imply

$$P_aU(x) - U(x) \le -ac_U\|g(x)\|^2 + \tfrac{1}{2}L_Ua^2\{c_0 + (1 + c_1)\|g(x)\|^2\}.$$

All terms are integrable: U has quadratic growth, ∇U has at most linear growth, and the covariance and drift bounds control the second increment moment. Stationarity therefore gives, for $2c_U > L_U(1+c_1)a$,

$$\mathbb{E}\|g(X)\|^2 \le \frac{L_U c_0 a}{2c_U - L_U(1+c_1)a}. \tag{30}$$

It also proves

$$\mathbb{E}\|\Delta\|^2 = a^2\mathbb{E}\{\operatorname{tr} C(X) + \|g(X)\|^2\} \le Ka^2. \tag{31}$$

We next justify the global error bound used in the main text. Since A_j is invertible, Taylor expansion and its smallest singular value give $\|g(x)\| \ge c_j\|x - m_j\|$ in a sufficiently small ball about m_j. Outside a large ball, Cauchy–Schwarz and dissipativity give $\|g(x)\| \ge \kappa\|x\| - b/\|x\| \ge (\kappa/2)\|x\|$. There $\operatorname{dist}(x,\mathcal{M}) \le \|x\| + \max_j\|m_j\| \le K\|x\|$. On the remaining compact set separated from every root, the continuous positive ratio $\|g(x)\|/\operatorname{dist}(x,\mathcal{M})$ has a positive minimum. Combining these three regions gives

$$\|g(x)\| \ge c_g \operatorname{dist}(x,\mathcal{M}) \qquad (x \in \mathbb{R}^d). \tag{32}$$

The last estimate in (6) follows. In particular, every measurable set at distance at least $r > 0$ from $\mathcal{M}$ has stationary mass at most Ka/r^2.

A.1 Root continuity and stationary innovation tails

The concentration estimates above control the mass away from the root set. We next combine them with continuity at the roots to control large innovations under stationary sampling. The argument precedes, and does not use, uniform integrability of the rescaled displacement.

Lemma A.2 (Innovation laws near the roots)**.** *Under Assumption 2.1, as $x \to m_j$, the law of $\xi(x,\zeta_1)$ converges weakly to that of $\xi(m_j,\zeta_1)$ and its second moment converges. In particular,*

$$\lim_{R\to\infty}\lim_{r\downarrow 0}\sup_{\operatorname{dist}(x,\mathcal{M})\le r}\mathbb{E}[\|\xi(x,\zeta_1)\|^2\mathbf{1}\{\|\xi(x,\zeta_1)\| > R\}] = 0. \tag{33}$$

Proof. Fix a sufficiently small $a_* > 0$ for which P_{a_*} is Feller. For any sequence $x_n \to m_j$, the next-state laws $P_{a_*}(x_n,\cdot)$ converge weakly to $P_{a_*}(m_j,\cdot)$. If Feller is defined on C_0, it first gives vague convergence; the probability limit then gives weak convergence and tightness. The maps

$$F_n(y) = (y - x_n)/a_* + g(x_n)$$

converge uniformly on compact sets to $F(y) = (y - m_j)/a_* + g(m_j)$. Tightness and this local uniform convergence imply convergence of the image laws under F_n to the image law under F. These are precisely the innovation laws, because the mechanism ξ is independent of the step size. Continuity of C at m_j gives convergence of their second moments to $\operatorname{tr} C_j$.

For completeness, set $h_R(u) = \min\{\|u\|^2, R^2\}$. This is bounded and continuous, and

$$\mathbb{E}[\|\xi\|^2\mathbf{1}\{\|\xi\| > \sqrt{2}R\}] \le 2\{\operatorname{tr} C(x) - \mathbb{E}h_R(\xi(x,\zeta_1))\}.$$

Both terms in braces are continuous at m_j. Their difference there is $\mathbb{E}(\|\xi(m_j,\zeta_1)\|^2 - R^2)_+$, which tends to zero as $R \to \infty$. Taking the finite maximum over roots proves (33). The order of the two limits is part of the statement. □

Lemma A.3 (Stationary innovation and increment tails)**.** *Under Assumption 2.1,*

$$\lim_{R\to\infty}\limsup_{a\downarrow 0}\sup_{\pi\in\mathcal{I}_a}\mathbb{E}_\pi[\|\xi(X,\zeta_1)\|^2\mathbf{1}\{\|\xi(X,\zeta_1)\| > R\}] = 0. \tag{34}$$

The same assertion holds with $Y_a = \xi(X,\zeta_1) - g(X)$ in place of $\xi(X,\zeta_1)$. Consequently, for each fixed $\rho > 0$ and $\Delta = aY_a$,

$$\mathbb{E}[\|\Delta\|^2\mathbf{1}\{\|\Delta\| > \rho\}] = o(a^2) \tag{35}$$

uniformly over invariant-law selections.

Proof. Let $K_r = \{x : \mathrm{dist}(x, \mathcal{M}) \le r\}$. The covariance bound and concentration imply

$$\begin{aligned}\mathbb{E}_\pi[\|\xi\|^2 \mathbf{1}_{K_r^c}(X)] &= \mathbb{E}_\pi[\mathrm{tr}\, C(X)\mathbf{1}_{K_r^c}(X)] \\ &\le c_0\pi(K_r^c) + c_1\mathbb{E}_\pi\|g(X)\|^2 \le Ka/r^2 + Ka.\end{aligned}$$

Inside K_r, the tail expectation is bounded by the supremum in (33). For fixed R, r, take $\limsup_{a\downarrow 0}\sup_{\pi\in\mathcal{I}_a}$, then send $r \downarrow 0$, and finally $R \to \infty$. This proves (34) in its stated order.

For the increment, use the pointwise inequality

$$\begin{aligned}\|u - v\|^2\mathbf{1}\{\|u - v\| > R\} \le 4\|u\|^2\mathbf{1}\{\|u\| > R/2\} \\ + 4\|v\|^2\mathbf{1}\{\|v\| > R/2\}.\end{aligned}$$

The innovation term has just been controlled, and the drift contribution is bounded by $4\mathbb{E}\|g(X)\|^2 = O(a)$. Thus the same asymptotic tail assertion holds for Y_a. At the diverging threshold ρ/a, it gives $\mathbb{E}[\|Y_a\|^2\mathbf{1}\{\|Y_a\| > \rho/a\}] = o(1)$. Multiplication by a^2 proves (35). □

These are tail statements under stationary sampling as $a \downarrow 0$. Appendix T shows that this stationary control is strictly weaker than uniform integrability on a fixed root neighborhood.

B IDENTIFICATION OF FINITE STATIONARY OU MEASURES

Lemma B.1. *Let C be symmetric nonnegative and let ν be a finite nonnegative measure with $\int \|z\|^2\nu(\mathrm{d}z) < \infty$. Suppose*

$$\int \mathcal{L}f\,\mathrm{d}\nu = 0 \quad (f \in C_c^2(\mathbb{R}^d)), \qquad \mathcal{L}f = -\langle Az, \nabla f\rangle + \tfrac{1}{2}\,\mathrm{tr}(CD^2 f). \tag{36}$$

Write $q = \nu(\mathbb{R}^d)$ and $M = \int zz^{\mathsf{T}}\nu(\mathrm{d}z)$. Then $AM + MA^{\mathsf{T}} = qC$. If A is positive stable, $\nu = q\mathsf{N}(0, S)$ with $AS + SA^{\mathsf{T}} = C$. If $A^{\mathsf{T}}v = \lambda v$, $\Re\lambda < 0$ and $v^*Cv > 0$, then $q = 0$.

Proof. Choose a smooth cutoff ψ_R equal to one on $B(0, R)$ and zero outside $B(0, 2R)$, with derivative bounds $\|D^k\psi_R\|_\infty \le KR^{-k}$ for $k = 1, 2$. Apply the stationary equation to $z_r z_s \psi_R(z)$. The additional derivative terms are supported on $R \le \|z\| \le 2R$ and bounded by $K(1 + \|z\|^2)$ there, uniformly for $R \ge 1$. Their integrals vanish as $R \to \infty$ by the finite second moment. The remaining terms converge by dominated convergence and give $AM + MA^{\mathsf{T}} = qC$. Hence

$$2\Re\lambda\, v^*Mv = qv^*Cv.$$

The left side is nonpositive and the right side is strictly positive if $q > 0$, proving the last assertion, including complex eigendirections.

For the stable case, define $\phi(t) = \int e^{it^{\mathsf{T}}z}\nu(\mathrm{d}z)$. Apply the stationary equation to the real and imaginary parts of $e^{it^{\mathsf{T}}z}\psi_R(z)$. Cutoff terms are bounded by $K_t(1 + \|z\|)$ on the same annulus; passage to the limit is again justified. Differentiation of ϕ is valid by its finite first moment, and the equation becomes

$$(A^{\mathsf{T}}t) \cdot \nabla\phi(t) = -\tfrac{1}{2}t^{\mathsf{T}}Ct\,\phi(t).$$

For $t_s = e^{-A^{\mathsf{T}}s}t$, it follows that $\frac{\mathrm{d}}{\mathrm{d}s}\phi(t_s) = \frac{1}{2}t_s^{\mathsf{T}}Ct_s\,\phi(t_s)$. Solving this scalar linear equation, without dividing by ϕ, gives

$$\phi(t) = \phi(t_s)\exp\left\{-\tfrac{1}{2}\int_0^s t^{\mathsf{T}}e^{-Au}Ce^{-A^{\mathsf{T}}u}t\,\mathrm{d}u\right\}.$$

Stability implies $t_s \to 0$ and $\phi(t_s) \to q$. Thus $\phi(t) = q\exp(-t^{\mathsf{T}}St/2)$. Uniqueness of Fourier transforms of finite measures completes the proof; no nondegeneracy of C is required. □

Lemma B.2 (Hyperbolic OU classification)**.** *Let A be hyperbolic and let ν satisfy Lemma B.1, with mass q. Define P^s, P^u and $\bar{S}$ as in (9). If $q > 0$, then $P^uC(P^u)^T = 0$ and $\nu = q\mathsf{N}(0, \bar{S})$. Conversely, this Gaussian is stationary whenever $P^uC(P^u)^T = 0$. In particular, $P^uC(P^u)^T \ne 0$ forces $q = 0$.*

Proof. Choose a real invertible matrix T such that $T^{-1}AT = \mathrm{diag}(A_s, A_u)$, with spectra in the positive and negative half-planes, respectively. This is a splitting into generalized invariant subspaces and does not presume diagonalizability. Transform the measure by T^{-1} and set $\widetilde{C} = T^{-1}CT^{-T}$ and $\widetilde{M} = T^{-1}MT^{-T}$. These are congruences, so both matrices remain nonnegative. The moment identity in Lemma B.1 gives

$$A_u\widetilde{M}_{uu} + \widetilde{M}_{uu}A_u^T = q\widetilde{C}_{uu}.$$

Since A_u has negative-real-part spectrum, $H = \int_0^\infty e^{A_u^T t}e^{A_u t}\,\mathrm{d}t$ is positive definite and satisfies $A_u^T H + HA_u = -I$. Taking a trace pairing yields

$$-\operatorname{tr}\widetilde{M}_{uu} = q\operatorname{tr}(H\widetilde{C}_{uu}).$$

For $q > 0$, the opposite signs and $\operatorname{tr}(H\widetilde{C}_{uu}) \geq \lambda_{\min}(H)\operatorname{tr}\widetilde{C}_{uu}$ imply $\widetilde{M}_{uu} = \widetilde{C}_{uu} = 0$. Nonnegativity of $\widetilde{C}$ forces $\widetilde{C}_{su} = 0$: evaluate its quadratic form at (x, ty) for every real t. Also $\int \|z_u\|^2\,\mathrm{d}\widetilde{\nu} = 0$, so $\widetilde{\nu}$ is supported on $z_u = 0$.

To obtain the stationary equation for its stable marginal, multiply a compactly supported stable test by a compact unstable cutoff equal to one near zero. The unstable derivative terms vanish on the support of the measure. The stable case of Lemma B.1 identifies this marginal as the Gaussian with covariance $\int_0^\infty e^{-A_s t}\widetilde{C}_{ss}e^{-A_s^T t}\,\mathrm{d}t$. Transforming back gives $q\mathsf{N}(0, \bar{S})$. If the stable subspace is empty, the measure is $q\delta_0$; if the unstable subspace is empty, the conclusion is precisely the stable OU lemma. If $q = 0$, nonnegativity gives $\nu = 0$.

Finally, positivity and $P^u C(P^u)^T = 0$ imply $P^u C = C(P^u)^T = 0$. Thus $P^s C(P^s)^T = C$, and the Gaussian supported on the stable subspace indeed solves the full stationary equation. Without this compatibility condition, $A\bar{S} + \bar{S}A^T = P^s C(P^s)^T$ still holds, but its right side need not equal C. □

The projection condition can exclude roots not covered by eigenvector excitation. For $A = \left(\begin{smallmatrix} -1 & 1 \\ 0 & -1 \end{smallmatrix}\right)$ and $C = \mathrm{diag}(1, 0)$, every left eigenvector is proportional to $(0, 1)^T$, whereas $P^u C(P^u)^T = C \neq 0$. The classification therefore excludes positive local mass even though (17) fails. This matrix example isolates the local excitation criterion; global confinement is supplied separately by Assumption 2.1.

C Localization at all roots

Lemma C.1 (All-root local limits)**.** *Under Assumption 2.1, every sequence $a \downarrow 0$ and every invariant-law selection admit a subsequence such that*

$$(J_a, Z_a) \Rightarrow \sum_{j=1}^{\ell} \delta_j \otimes \nu_j, \quad q_j := \nu_j(\mathbb{R}^d) = \lim w_{a,j}, \quad \sum_j q_j = 1. \tag{37}$$

The measures have finite second moments $M_j = \int zz^T\nu_j(dz)$ and satisfy $A_jM_j + M_jA_j^T = q_jC_j$. At every root, $\nu_j = q_j\bar{\Gamma}_j$, with $q_j = 0$ whenever $P_j^u C_j(P_j^u)^T \neq 0$.

Lemma C.2 (Marked weak approximation)**.** *Under Assumptions 2.1 and 3.4,*

$$\sum_{j\in\mathcal{U}} w_{a,j} \to 0, \qquad d_{\mathrm{BL}}\left(\mathcal{L}(X_{\mathcal{M}}, Z_a), \sum_{j\in\mathcal{S}} w_{a,j}(\delta_{m_j} \otimes \Gamma_j)\right) \to 0. \tag{38}$$

Along every subsequence with convergent weights,

$$(J_a, Z_a) \Rightarrow \sum_{j\in\mathcal{S}} w_j(\delta_j \otimes \Gamma_j). \tag{39}$$

We first prove Lemma C.1, and then apply excitation to obtain Lemma C.2. For finite measures define $d_{\mathrm{BL}}(\mu, \nu) = \sup_{\|h\|_\infty + \mathrm{Lip}(h)\leq 1} |\mu h - \nu h|$. Choose $r > 0$ sufficiently small that the closed balls $\overline{B}(m_j, 4r)$ are disjoint and have their nearest root equal to m_j. Choose $\chi_j \in C_c^\infty$ with $0 \leq$

$\chi_j \le 1$, equal to one on $B(m_j, r)$ and zero outside $B(m_j, 2r)$. Define the finite localized measure and its mass by

$$\nu_{a,j}(D) = \mathbb{E}[\chi_j(X)\mathbf{1}\{(X - m_j)/\sqrt{a} \in D\}], \qquad q_{a,j} = \mathbb{E}\chi_j(X). \tag{40}$$

Theorem 3.1 gives

$$\int \|z\|^2 \nu_{a,j}(\,\mathrm{d}z) \le K, \qquad 0 \le w_{a,j} - q_{a,j} \le Ka. \tag{41}$$

Indeed, on the cutoff support $\|X - m_j\| = \operatorname{dist}(X, \mathcal{M})$; the difference between the cell indicator and cutoff is supported where $\operatorname{dist}(X, \mathcal{M}) \ge r$. These finite measures are uniformly tight.

Cutoff change. Let $D_j = \operatorname{supp} \nabla\chi_j$ and choose $\rho > 0$ so the closed ρ-enlargement D_j^ρ is compact and separated from all roots. On $\{\|\Delta\| \le \rho\}$ a nonzero cutoff change implies $X \in D_j^\rho$, by the fundamental theorem of calculus along the increment segment. On this compact set, $\mathbb{E}[\|\Delta\| \mid X = x] \le a(\|g(x)\| + \sqrt{\operatorname{tr} C(x)}) \le Ka$. The small-increment contribution to $\mathbb{E}|\chi_j(X^+) - \chi_j(X)|$ is at most $Ka\pi_a(D_j^\rho) = O(a^2)$. The large-increment contribution is at most $\mathbb{P}(\|\Delta\| > \rho) \le \rho^{-2}\mathbb{E}\|\Delta\|^2 = O(a^2)$. Consequently,

$$\mathbb{E}|\chi_j(X^+) - \chi_j(X)| \le Ka^2. \tag{42}$$

Localized generator equation. Fix $f \in C_c^2$ and abbreviate $Z = (X - m_j)/\sqrt{a}$, $H = \Delta/\sqrt{a}$. Stationarity of $\chi_j(x)f((x - m_j)/\sqrt{a})$ and (42) give

$$a^{-1}\mathbb{E}\{\chi_j(X)[f(Z + H) - f(Z)]\} = O(a). \tag{43}$$

Let ω_f be a bounded modulus of continuity of $D^2 f$, tending to zero at zero. Taylor's formula gives

$$f(Z + H) - f(Z) = \nabla f(Z)^\mathsf{T} H + \tfrac{1}{2} H^\mathsf{T} D^2 f(Z) H + R, \qquad |R| \le K\|H\|^2 \omega_f(\|H\|).$$

Write $Y_a = \xi(X, \zeta_1) - g(X)$, so $H = \sqrt{a}Y_a$. The divided Taylor remainder is bounded by

$$a^{-1}\mathbb{E}[\chi_j|R|] \le K\mathbb{E}[\chi_j(X)\|Y_a\|^2 \omega_f(\sqrt{a}\|Y_a\|)].$$

On $\|Y_a\| \le R_0$, it tends to zero uniformly as $a \downarrow 0$, using bounded second moments. On the complement, boundedness of ω_f and Lemma A.3 control the expectation uniformly. Let $a \downarrow 0$ and then $R_0 \to \infty$ to obtain $a^{-1}\mathbb{E}[\chi_j|R|] \to 0$.

Conditional centering gives $\mathbb{E}[H \mid X] = -\sqrt{a}g(X)$ and $\mathbb{E}[HH^\mathsf{T} \mid X] = a\{C(X) + g(X)g(X)^\mathsf{T}\}$. The drift-squared contribution is $O(a)$ by (6). Local C^2 regularity and (41) give

$$\mathbb{E}\left[\chi_j(X)\|g(X)/\sqrt{a} - A_j Z\|\right] \le K\sqrt{a}\mathbb{E}[\chi_j(X)\|Z\|^2] \le K\sqrt{a}. \tag{44}$$

Finally, C is bounded on the cutoff support by (3). Its continuity at m_j, together with concentration outside smaller root balls, implies $\mathbb{E}[\chi_j(X)\|C(X) - C_j\|] \to 0$. Substitution in (43) yields

$$\int \left[-\langle A_j z, \nabla f(z)\rangle + \tfrac{1}{2}\operatorname{tr}(C_j D^2 f(z))\right] \nu_{a,j}(\,\mathrm{d}z) \longrightarrow 0.$$

Identification and removal of subsequences. Along any sequence $a_n \downarrow 0$, extract a common subsequence on which all finite measures $\nu_{a_n,j}$ converge weakly. Their limits ν_j have finite second moment by lower semicontinuity, and their masses are the limits of $q_{a_n,j}$. The integrand in the last display is bounded continuous because f has compact support, so each limit satisfies (36). Lemma B.1 gives the matrix identity for every root; Lemma B.2 identifies $\nu_j = q_j\bar{\Gamma}_j$. The total lost cutoff mass is $O(a)$, so $\sum_j q_j = 1$ and $q_j = \lim w_{a,j}$. Replacing cutoffs by cells costs $O(a)$ for bounded tests, proving (37) and Lemma C.1.

Under excitation, Lemma B.2 gives $q_j = 0$ for every unstable root. The marked law and the stable comparison measure then have the same subsequential weak limit. The bounded-Lipschitz metric metrizes weak convergence for these tight finite measures. If (38) failed uniformly, a violating sequence of step sizes and invariant laws would have such a subsequence, a contradiction. This proves Lemma C.2.

D EXACT ENERGY, UNIFORM INTEGRABILITY AND W_2

Write $D = \langle \nabla U, g \rangle$ and $E_j = D^2U(m_j)$. The preceding concentration theorem also gives $\mathbb{E}\|Z_a\|^2 \le K$.

Lemma D.1 (Quadratic behavior of the descent dissipation)**.** *At every root, $\nabla U(m_j) = 0$, and, as $y \to 0$,*

$$D(m_j + y) = y^T E_j A_j y + o(\|y\|^2). \tag{45}$$

Moreover $B_j = (E_j A_j + A_j^T E_j)/2 \succeq c_U A_j^T A_j \succ 0$.

Proof. Since $D \ge 0$ and $D(m_j) = 0$, differentiability gives $\nabla D(m_j) = 0$. The product rule gives $\nabla D(m_j) = A_j^T \nabla U(m_j)$, because $g(m_j) = 0$. Hyperbolicity implies invertibility of A_j, hence $\nabla U(m_j) = 0$. Now $\nabla U(m_j + y) = E_j y + o(\|y\|)$ and $g(m_j + y) = A_j y + O(\|y\|^2)$. Multiplication proves (45). Applying (5) at $m_j + ty$ and dividing by t^2 proves the matrix inequality. In particular this argument does not presume that E_j is positive definite at unstable roots. □

Lemma D.2 (Exact stationary energy asymptotics)**.** *Under Assumption 2.1, without excitation,*

$$\frac{\mathbb{E}D(X)}{a} = \frac{1}{2}\sum_{j=1}^{\ell} w_{a,j} \operatorname{tr}(E_j C(m_j)) + o(1). \tag{46}$$

The remainder is uniform over selected invariant laws.

Proof. Put $Y = \xi(X, \zeta) - g(X)$, so $\Delta = aY$. Taylor expansion with an integral remainder gives

$$U(X + aY) - U(X) = a\nabla U(X)^T Y + \frac{a^2}{2} Y^T D^2U(X) Y + a^2 r_a(X, Y),$$

where, with $L_U = \sup_x \|D^2U(x)\|$,

$$|r_a(x, y)| \le \|y\|^2 \int_0^1 (1 - t)\|D^2U(x + tay) - D^2U(x)\|\, dt,$$
$$|r_a(x, y)| \le L_U \|y\|^2.$$

We show $\mathbb{E}|r_a(X, Y)| = o(1)$; this is the step for which a bare second-moment bound would not suffice.

Choose a fixed compact union K_0 of small closed root balls. For any fixed R, continuity of D^2U gives uniform continuity on a compact enlargement of K_0, and hence

$$\sup_{x \in K_0,\, \|y\| \le R} |r_a(x, y)| \longrightarrow 0.$$

The expectation on $\|Y\| > R$ is bounded by $L_U \mathbb{E}[\|Y\|^2 \mathbf{1}\{\|Y\| > R\}]$. Lemma A.3 makes this contribution vanish after $a \downarrow 0$ and then $R \to \infty$. Thus the required remainder limit holds inside K_0.

Outside K_0, concentration gives $\pi_a(K_0^c) = O(a)$. Conditional centering and the covariance growth assumption imply

$$\begin{aligned}\mathbb{E}[\|Y\|^2 \mathbf{1}_{\{X \notin K_0\}}] &= \mathbb{E}[\{\operatorname{tr} C(X) + \|g(X)\|^2\}\mathbf{1}_{\{X \notin K_0\}}] \\ &\le c_0 \pi_a(K_0^c) + (1 + c_1)\mathbb{E}\|g(X)\|^2 = O(a).\end{aligned}$$

The global bounded-Hessian estimate therefore makes the outside remainder vanish as well.

All terms in the Taylor identity are integrable: U, $\nabla U^T g$, and g^2 have at most quadratic growth and the covariance is controlled. Stationarity and conditional centering consequently give

$$\frac{\mathbb{E}D(X)}{a} = \frac{1}{2}\mathbb{E}\operatorname{tr}\big[D^2U(X)\{C(X) + g(X)g(X)^T\}\big] + o(1).$$

The drift-squared term is $O(a)$ by (30). To replace the covariance term by root values, first restrict to small root balls. There, continuity of D^2U and of C at each root controls the error as the radii decrease. Outside these balls its absolute expectation is bounded by a constant times $c_0\pi_a(\text{outside}) + c_1\mathbb{E}\|g(X)\|^2 = O(a)$ for each fixed radius. The difference between ball mass and Voronoi mass is also $O(a)$ for fixed radii. Taking $a \to 0$ and then shrinking the radii proves (46). All estimates are law-uniform. □

Proof of Theorem 3.2. Take any sequence of steps and invariant laws, and extract the common all-root subsequence of Lemma C.1. Write $R_a = D(X)/a$. Lemma D.1 identifies $R_a \Rightarrow R$, where R is the image of the marked limit under $(j, z) \mapsto z^T E_j A_j z$. Indeed, on $\|Z_a\| \le R_0$ the divided Taylor error is uniformly $o(1)$, and tightness controls the complement. Both variables are nonnegative. Symmetry of E_j, M_j and the matrix moment equation imply

$$\mathbb{E}R = \sum_j \operatorname{tr}(E_j A_j M_j) = \tfrac{1}{2} \sum_j \operatorname{tr}\{E_j (A_j M_j + M_j A_j^T)\} = \tfrac{1}{2} \sum_j q_j \operatorname{tr}(E_j C_j).$$

The two traces in the middle sum agree by transposition and cyclicity. This identity applies at unstable roots without identifying their measures as Gaussians. Lemma D.2 and $w_{a,j} \to q_j$ give $\mathbb{E}R_a \to \mathbb{E}R$.

For fixed $T > 0$, bounded continuous truncation gives

$$\mathbb{E}(R_a - T)_+ = \mathbb{E}R_a - \mathbb{E}(R_a \wedge T) \longrightarrow \mathbb{E}(R - T)_+.$$

Since $r\mathbf{1}\{r > 2T\} \le 2(r - T)_+$, the variables R_a are asymptotically uniformly integrable along this subsequence. The domination $R_a \ge c_U c_g^2 \|Z_a\|^2$ transfers the property to $\|Z_a\|^2$. If (10) failed, there would be $a_n \downarrow 0$, invariant laws π_n and $R_n \to \infty$ with tail expectations bounded away from zero. The preceding subsequence argument contradicts this, proving the law-uniform assertion.

At every root, Lemma B.2 gives $\nu_j = q_j \bar{\Gamma}_j$; incompatible roots have $q_j = 0$. Uniform integrability permits linear and quadratic tests multiplied by the discrete root label, proving (11) on every extracted subsequence and hence on the full family. For weights bounded below, division gives conditional weak convergence and square uniform integrability. The weak-plus-second-moment criterion for W_2 proves (12); subtracting conditional means gives its covariance assertion. □

Coupling at a common root. We prove the same-label assertion of Theorem 3.3:

$$\sup_{\pi \in \mathcal{I}_a} \sum_{j : w_{a,j} > 0} w_{a,j} W_2^2(\mathcal{L}(Z_a \mid J_a = j), \bar{\Gamma}_j) \longrightarrow 0. \tag{47}$$

For the marked cost $d((j, z), (k, y))^2 = \mathbf{1}\{j \ne k\} + \|z - y\|^2$, it implies

$$W_2\left(\mathcal{L}(J_a, Z_a), \sum_j w_{a,j} (\delta_j \otimes \bar{\Gamma}_j)\right) \to 0 \tag{48}$$

uniformly over invariant-law selections.

Proof of Theorem 3.3. Take any sequence $a_n \downarrow 0$ and invariant laws π_n. Extract a subsequence on which all weights converge to q_j and the marked laws have the limits in Lemma C.1. Lemma B.2 identifies each limit as $q_j \bar{\Gamma}_j$. If $q_j > 0$, conditional weak convergence and the square uniform integrability in Theorem 3.2 yield conditional W_2 convergence (Villani, 2009). If $q_j = 0$, an independent coupling gives, whenever $w_{a,j} > 0$,

$$w_{a,j} W_2^2(\mathcal{L}(Z_a \mid J_a = j), \bar{\Gamma}_j) \le 2\mathbb{E}[\mathbf{1}\{J_a = j\} \|Z_a\|^2] + 2 w_{a,j} \operatorname{tr} \bar{S}_j.$$

For any fixed R, the first expectation is bounded by $R^2 w_{a,j} + \mathbb{E}[\|Z_a\|^2 \mathbf{1}\{\|Z_a\| > R\}]$. First let $n \to \infty$, then $R \to \infty$. The bound tends to zero by (10). Terms with $w_{a,j} = 0$ contribute zero without defining a conditional law. Summing over the finite root set proves (47) along the extracted subsequence. Failure of uniform convergence would give a violating sequence with such a subsequence, a contradiction.

Choose optimal conditional couplings of Z_a and $G_j \sim \bar{\Gamma}_j$, and mix them with the exact weights $w_{a,j}$. Matching labels gives zero label cost and proves (48). Mapping each pair to $(m_j + \sqrt{a} Z_a, m_j + \sqrt{a} G_j)$ produces marginals π_a and $\bar{G}_{a,\pi}$. Its squared cost is a times the sum in (47), proving (15). For an incompatible root, positive limiting weight would contradict Lemma B.2; the same subsequence argument makes the exclusion uniform. □

Proof of Corollary 3.5. Excitation and Lemma B.2 exclude every unstable component, so $s_a \to 1$ uniformly. Along any weight-convergent subsequence, the marked law and Q_a have the same stable Gaussian weak limit. Equation (10) upgrades the first convergence to W_2; the finite Gaussian mixtures Q_a also converge in W_2. The triangle inequality and the subsequence argument prove (18). For each unstable root and fixed R, its second-moment contribution is at most $R^2 w_{a,j} + \mathbb{E}[\|Z_a\|^2 \mathbf{1}\{\|Z_a\| > R\}]$. Letting $a \to 0$ and then $R \to \infty$ proves the last assertion. □

E ROOTWISE STATIONARY MEANS

We prove Theorem 4.3. All estimates in this section use Assumption 2.1 alone and are uniform over invariant-law selections.

Lemma E.1 (Fixed-neighborhood tails)**.** *For every fixed $r > 0$,*

$$\mathbb{E}[\operatorname{dist}(X, \mathcal{M})^2 \mathbf{1}\{\operatorname{dist}(X, \mathcal{M}) \geq r\}] = o(a), \qquad \mathbb{P}\{\operatorname{dist}(X, \mathcal{M}) \geq r\} = o(a).$$

Proof. The first expectation divided by a is $\mathbb{E}[\|Z_a\|^2 \mathbf{1}\{\|Z_a\| \geq r/\sqrt{a}\}]$, which tends to zero by (10). The probability is at most the first expectation divided by r^2. □

Lemma E.2 (Original-scale cutoff flux)**.** *For the cutoffs in Appendix C,*

$$\mathbb{E}[|\chi_j(X^+) - \chi_j(X)|(1 + \|X^+ - m_j\|)] = o(a^2). \tag{49}$$

Moreover, $\mathbb{E}[\mathbf{1}_{\mathcal{C}_j}(X) g(X)] = o(a)$ for every root.

Proof. Choose $\rho > 0$ so the ρ-enlargement of $\operatorname{supp} \nabla \chi_j$ is compact and separated from every root. If $\|\Delta\| \leq \rho$, a nonzero cutoff change places both endpoints in a fixed compact set and X in this enlargement. The integrand in (49) is then bounded by $K\|\Delta\|$. On the enlargement, $\mathbb{E}[\|\Delta\| \mid X] \leq Ka$ by boundedness of the drift and conditional second moments. Its stationary mass is $o(a)$ by Lemma E.1. This contribution is $o(a^2)$.

If $\|\Delta\| > \rho$ and the cutoff changes, at least one endpoint lies in the compact cutoff support. Hence $\|X^+ - m_j\| \leq K + \|\Delta\|$, and the remaining contribution is bounded by

$$K\mathbb{E}[(1 + \|\Delta\|)\mathbf{1}\{\|\Delta\| > \rho\}] \leq K_\rho \mathbb{E}[\|\Delta\|^2 \mathbf{1}\{\|\Delta\| > \rho\}] = o(a^2).$$

This proves the flux estimate using Lemma A.3.

Apply stationarity componentwise to the bounded function $\chi_j(x)(x - m_j)$. Expanding its increment and using conditional centering yields

$$a\mathbb{E}[\chi_j(X) g(X)] = \mathbb{E}[(\chi_j(X^+) - \chi_j(X))(X^+ - m_j)] = o(a^2).$$

On $\mathcal{C}_j$, $\|x - m_j\| = \operatorname{dist}(x, \mathcal{M})$ and $\|g(x)\| \leq L\|x - m_j\|$. Where $\mathbf{1}_{\mathcal{C}_j} - \chi_j$ is nonzero, this distance is at least a fixed $r > 0$. Thus

$$\mathbb{E}[(\mathbf{1}_{\mathcal{C}_j} - \chi_j)\|g(X)\|] \leq (L/r)\mathbb{E}[\operatorname{dist}(X, \mathcal{M})^2 \mathbf{1}\{\operatorname{dist}(X, \mathcal{M}) \geq r\}] = o(a).$$

Replacing the cutoff proves the drift balance. The argument uses stationarity of the original chain, without assigning stationarity to a cell-conditioned law. □

Proof of Theorem 4.3. Fix any root and write $y = x - m_j$. Local C^2 regularity gives

$$g(m_j + y) = A_j y + \tfrac{1}{2}(D^2 g(m_j) : yy^T) + r_j(y), \qquad \|r_j(y)\| = o(\|y\|^2).$$

For $\epsilon > 0$, choose $r > 0$ so $\|r_j(y)\| \leq \epsilon \|y\|^2$ when $\|y\| \leq r$. Outside this ball, global Lipschitz continuity and the fixed Taylor coefficients give $\|r_j(y)\| \leq K_r \|y\|^2$. Consequently,

$$\frac{1}{a}\mathbb{E}[\mathbf{1}_{\mathcal{C}_j}\|r_j(X - m_j)\|] \leq \epsilon \mathbb{E}\|Z_a\|^2 + K_r \mathbb{E}[\|Z_a\|^2 \mathbf{1}\{\|Z_a\| > r/\sqrt{a}\}].$$

Let $a \downarrow 0$ and then $\epsilon \downarrow 0$. Theorem 3.2 makes the remainder $o(a)$ and also gives $\mathbb{E}[\mathbf{1}_{\mathcal{C}_j}(X - m_j)(X - m_j)^T] = a w_{a,j} \tilde{S}_j + o(a)$. Combine these facts with Lemma E.2 to obtain

$$A_j \mathbb{E}[\mathbf{1}_{\mathcal{C}_j}(X - m_j)] + \tfrac{a}{2} w_{a,j} t_j = o(a).$$

Invertibility of A_j proves (23).

For a C^2 observable with coordinatewise bounded Hessian, the same truncation argument controls its second-order Taylor remainder by $o(a)$. Its linear term uses (23), and its quadratic term uses (11), yielding (24). Bounded Hessians ensure at most quadratic growth and integrability. Dividing by weights bounded below proves both conditional expansions. □

For gradient drift and $f = F$, the linear term vanishes and $k_{F,j} = \frac{1}{2}\operatorname{tr}(H_j \bar{S}_j)$. At a compatible root, $H_j \bar{S}_j + \bar{S}_j H_j = C_j$, giving $k_{F,j} = \frac{1}{4}\operatorname{tr} C_j$. Incompatible roots have vanishing weights, so their coefficient difference does not affect the weighted limit (22). For a fixed evaluation loss, the gradient term in (25) generally remains. The expansion concerns algorithmic variation of this fixed loss, rather than uncertainty in a population risk.

E.1 Mean offsets and root-centered Gaussian accuracy

For an invariant-law family π_a with $\underline{w}_j := \liminf_{a\downarrow 0} w_{a,j} > 0$ and $b_j \neq 0$, the mean-difference bound for W_2 gives

$$D_a(\pi_a) \geq \sqrt{w_{a,j}}\, \|\mathbb{E}[Z_a \mid J_a = j]\|, \qquad \liminf_{a\downarrow 0} \frac{D_a(\pi_a)}{\sqrt{a}} \geq \sqrt{\underline{w}_j}\, \|b_j\|.$$

Indeed, Theorem 4.3 gives $\mathbb{E}[Z_a \mid J_a = j] = \sqrt{a}\, b_j + o(\sqrt{a})$, proving the second inequality. This mean-offset contribution to the same-label error is consistent with the scalar nonlinear lower bound of Wang et al. (2026, Proposition 3.9).

E.2 Between-root and within-root observable variance

The weighted expansions also describe global variability, including the contribution of different root predictions.

Corollary E.3 (Global observable variance)**.** *Under Assumption 2.1, let $f : \mathbb{R}^d \to \mathbb{R}$ be C^2, globally Lipschitz, and have bounded Hessian. Set $f_j = f(m_j)$, $v_{f,j} = Df(m_j)\bar{S}_j Df(m_j)^T$, and $\bar{f}_a = \sum_j w_{a,j} f_j$. With $k_{f,j}$ from Theorem 4.3, uniformly over invariant-law selections,*

$$\operatorname{Var}(f(X)) = \sum_j w_{a,j}(f_j - \bar{f}_a)^2 + a \sum_j w_{a,j}\{v_{f,j} + 2(f_j - \bar{f}_a)k_{f,j}\} + o(a). \tag{50}$$

The weights are the actual finite-a weights; no convergence or positive lower bound is required.

Proof. Write $\Delta_j f = f(X) - f_j$ on $\mathcal{C}_j$. Theorem 4.3 gives $\mathbb{E}[\mathbf{1}_{\mathcal{C}_j}\Delta_j f] = a w_{a,j} k_{f,j} + o(a)$. We first show

$$\mathbb{E}[\mathbf{1}_{\mathcal{C}_j}(\Delta_j f)^2] = a w_{a,j} v_{f,j} + o(a). \tag{51}$$

For $y = x - m_j$, differentiability implies $(f(m_j + y) - f_j)^2 - (Df(m_j)y)^2 = o(\|y\|^2)$ as $y \to 0$. Global Lipschitz continuity bounds the absolute difference by $K\|y\|^2$ everywhere. For any $\eta > 0$, choose $r > 0$ so the local bound is $\eta\|y\|^2$. After division by a, the expected absolute difference on $\mathcal{C}_j$ is at most

$$\eta \mathbb{E}\|Z_a\|^2 + K\mathbb{E}[\|Z_a\|^2 \mathbf{1}\{\|Z_a\| > r/\sqrt{a}\}].$$

Theorem 3.2, followed by $\eta \downarrow 0$, makes this remainder $o(1)$ uniformly. The rootwise second-moment formula (11) then proves (51).

Summing over the finite root set yields

$$\mathbb{E}f(X) = \bar{f}_a + a \sum_j w_{a,j} k_{f,j} + o(a),$$

$$\mathbb{E}f(X)^2 = \sum_j w_{a,j} f_j^2 + a \sum_j w_{a,j}(v_{f,j} + 2 f_j k_{f,j}) + o(a).$$

The coefficients and $\bar{f}_a$ are uniformly bounded because the root set is finite. Squaring the first expansion and subtracting it from the second proves (50). Every estimate is weighted before summation, so vanishing component weights cause no division by small probabilities. □

The leading term is between-root variability and can remain of order one. The order-a terms combine within-root Gaussian variance with a signed correction from the shifted component means. The argument uses only second moments: it controls squared observable increments directly, without applying the mean expansion to f^2.

F WEIGHTED LOCALIZATION AND LOCAL SMOOTH-TEST RATES

Theorem F.1 (Local smooth-test rate). *Assume Assumption 2.1 and fix $j \in \mathcal{S}$. For centered additive innovations with $\mathbb{E}\|\varepsilon\|^{2+\delta} < \infty$,*

$$\mathcal{E}_{a,j} \le K_j a^{\delta/2}. \tag{52}$$

For state-dependent noise, suppose in a fixed neighborhood of m_j that

$$\sup_x \mathbb{E}\|\xi(x,\zeta_1)\|^{2+\delta} < \infty, \qquad \|C(x) - C_j\|_{\mathrm{op}} \le K\|x - m_j\|^\beta, \qquad 0 < \beta \le 1. \tag{53}$$

Then $\mathcal{E}_{a,j} \le K_j(a^{\delta/2} + a^{\beta/2} + \sqrt{a})$. Whenever $w_{a,j} > 0$, the corresponding conditional smooth-test error is bounded by $\mathcal{E}_{a,j}/w_{a,j}$. No higher stationary moment, positive-definite covariance, or unstable-root excitation is needed for this stable-root statement.

Lemma F.2 (Weighted cutoff flux). *For the cutoffs in Appendix C, let $Z_j^+ = (X^+ - m_j)/\sqrt{a}$. Under Assumption 2.1, uniformly over invariant-law selections,*

$$\frac{1}{a}\mathbb{E}[|\chi_j(X^+) - \chi_j(X)|(1 + \|Z_j^+\|)] \le K\sqrt{a}. \tag{54}$$

Proof. Use D_j^ρ and Δ from that appendix. On $\{\|\Delta\| \le \rho\}$, a cutoff change forces both endpoints into a fixed compact set and $X \in D_j^\rho$. For $0 < a \le 1$,

$$|\chi_j(X^+) - \chi_j(X)|(1 + \|Z_j^+\|) \le Ka^{-1/2}\|\Delta\|\mathbf{1}_{D_j^\rho}(X).$$

Its expectation is at most $Ka^{-1/2}\, a\pi_a(D_j^\rho) \le Ka^{3/2}$. On $\{\|\Delta\| > \rho\}$, a nonzero cutoff change means at least one endpoint lies in the compact support of χ_j. Therefore $\|X^+ - m_j\| \le K + \|\Delta\|$ and the weighted change is bounded by

$$Ka^{-1/2}(1 + \|\Delta\|)\mathbf{1}\{\|\Delta\| > \rho\}.$$

Both $\mathbb{P}(\|\Delta\| > \rho)$ and $\mathbb{E}[\|\Delta\|\mathbf{1}\{\|\Delta\| > \rho\}]$ are bounded by constants times $\mathbb{E}\|\Delta\|^2 \le Ka^2$. The large-increment contribution is thus also $O(a^{3/2})$. Adding the two bounds and dividing by a proves the lemma. □

OU Poisson equation. Fix a stable root and write $A = A_j$, $C = C_j$, $\Gamma = \Gamma_j$. Choose $M_A \ge 1$ and $\rho_A > 0$ with $\|e^{-At}\| \le M_A e^{-\rho_A t}$. Let P_t^{OU} be the OU semigroup and, for $h \in \mathcal{H}_{2,\delta}$, define

$$u_h(z) = -\int_0^\infty (P_t^{\mathrm{OU}}h(z) - \Gamma h)\,\mathrm{d}t. \tag{55}$$

Coupling a process starting at z to a stationary process starting at $G \sim \Gamma$ with the same noise bounds their displacement by $M_A e^{-\rho_A t}\|z - G\|$. Hence the integral converges and

$$|u_h(z)| \le \frac{M_A}{\rho_A}(\|z\| + \mathbb{E}\|G\|), \qquad \|\nabla u_h\|_\infty \le \frac{M_A}{\rho_A}, \tag{56}$$
$$\|D^2 u_h\|_\infty \le \frac{M_A^2}{2\rho_A}, \qquad [D^2 u_h]_\delta \le \frac{M_A^{2+\delta}}{(2+\delta)\rho_A}.$$

The derivative bounds follow by differentiating the affine OU transition map under the integral, with the displayed exponential bounds providing domination. They use the regularity of h, not elliptic smoothing, and remain valid when C is singular. Integrating $\partial_t P_t^{\mathrm{OU}}h = \mathcal{L}_j P_t^{\mathrm{OU}}h$ gives $\mathcal{L}_j u_h = h - \Gamma h$; the same bounds justify interchange at each fixed z.

Proof of Theorem F.1. Put $Z = (X - m_j)/\sqrt{a}$ and $H = (X^+ - X)/\sqrt{a}$. Shrink the cutoff support, if necessary, to the neighborhood of (53). For each fixed a, $x \mapsto \chi_j(x) u_h((x-m_j)/\sqrt{a})$ is bounded, so its stationary equation is valid without extra integrability assumptions. Lemma F.2 and (56) imply

$$\left|a^{-1}\mathbb{E}\{\chi_j(X)[u_h(Z+H) - u_h(Z)]\}\right| \le K\sqrt{a}. \tag{57}$$

Taylor expansion with the Hölder Hessian has remainder $|R_h| \le K\|H\|^{2+\delta}$. Only X inside the compact cutoff support contributes. Therefore

$$a^{-1}\mathbb{E}[\chi_j|R_h|] \le Ka^{\delta/2}\mathbb{E}[\chi_j\|\xi(X,\zeta_1) - g(X)\|^{2+\delta}] \le Ka^{\delta/2}.$$

For additive noise this uses $\mathbb{E}\|\varepsilon\|^{2+\delta} < \infty$; for state-dependent noise it uses the local moment in (53). It never requires a $(2+\delta)$ moment of Z. The nonlinear drift error is $O(\sqrt{a})$ by (44), and the drift-squared second-order term is $O(a)$. For state-dependent noise the additional covariance error is bounded by

$$K\mathbb{E}[\chi_j\|C(X) - C_j\|] \le Ka^{\beta/2}\mathbb{E}[\chi_j\|Z\|^{\beta}] \le Ka^{\beta/2},$$

where the last inequality follows from (41) and Hölder's inequality for a measure of mass at most one. This error is zero for additive noise. Substituting in (57) yields, uniformly in h,

$$|\nu_{a,j}h - q_{a,j}\Gamma_j h| \le K(a^{\delta/2} + \sqrt{a} + a)$$

in the additive case, with $Ka^{\beta/2}$ added in the state-dependent case. Replacing the cutoff by its Voronoi cell and $q_{a,j}$ by $w_{a,j}$ costs at most $2Ka$, since $\|h\|_\infty \le 1$. For $0 < \delta \le 1$ and $a \le 1$, $a + \sqrt{a} \le 2a^{\delta/2}$. This proves the two rates; division by a positive $w_{a,j}$ proves the conditional bound. An arbitrarily small weight is allowed in the absolute bound, but the resulting relative bound need not be informative. □

G Tail moduli and quantitative unstable mass

This section uses the weighted cutoff estimate. Choose a fixed root cutoff $\chi_j \in C_c^\infty$ supported within the local regularity neighborhood, and write

$$\nu_{a,j}h = \mathbb{E}[\chi_j(X)h((X - m_j)/\sqrt{a})], \quad q_{a,j} = \mathbb{E}\chi_j(X).$$

The localization proof gives $|w_{a,j} - q_{a,j}| \le Ka$, $\int\|z\|^2\nu_{a,j}(dz) \le K$, and (54). With a compact neighborhood N_j containing the cutoff support, define the stationary averaged modulus and covariance variation

$$\overline{T}_j(a;\pi_a) = \mathbb{E}[\chi_j(X)\|\xi(X,\zeta)\|^2 \min\{\sqrt{a}\|\xi(X,\zeta)\|, 1\}], \tag{58}$$

$$R_{C,j}(a;\pi_a) = \mathbb{E}[\chi_j(X)\|C(X) - C_j\|]. \tag{59}$$

An optional upper bound is the fixed-neighborhood supremum

$$T_j^{\sup}(a) = \sup_{x\in N_j} \mathbb{E}[\|\xi(x,\zeta)\|^2 \min\{\sqrt{a}\|\xi(x,\zeta)\|, 1\}].$$

Since $0 \le \chi_j \le 1$, $\overline{T}_j \le T_j^{\sup}$. The averaged modulus tends to zero under the standing assumptions; the supremum need not do so.

Lemma G.1 (Localized generator residual)**.** *Suppose f is C^2, grows at most linearly, and has bounded first and second derivatives with globally Lipschitz Hessian. With constants allowed to depend on these bounds and the fixed root,*

$$\left|\int \mathcal{L}_j f \, d\nu_{a,j}\right| \le K\{\sqrt{a} + \overline{T}_j(a;\pi_a) + R_{C,j}(a;\pi_a)\},$$
$$\mathcal{L}_j f(z) = -\langle A_j z, \nabla f(z)\rangle + \tfrac{1}{2}\operatorname{tr}(C(m_j)D^2 f(z)). \tag{60}$$

Only Assumption 2.1 is needed; the root need not be stable.

Proof. For fixed a, stationarity applies to the bounded function $x \mapsto \chi_j(x)f((x - m_j)/\sqrt{a})$. Put $Z = (X - m_j)/\sqrt{a}$ and $H = \sqrt{a}(\xi - g)$. Moving the cutoff change to the other side and using (54) bounds the localized one-step generator by $K\sqrt{a}$.

Boundedness and Lipschitz continuity of the Hessian give the Taylor remainder estimate

$$|f(z+h) - f(z) - \nabla f(z)^T h - \tfrac{1}{2} h^T D^2 f(z) h| \le K\|h\|^2 \min\{\|h\|, 1\}.$$

On the compact cutoff support, g is bounded. Splitting $\|\xi\|$ above and below a bound for $\|g\|$, and using $u^2 \min\{su, 1\}$ at $2u$ being at most eight times its value at u, bounds the remainder after averaging with χ_j and dividing by a by $K(\overline{T}_j(a;\pi_a) + \sqrt{a})$.

Conditional centering gives the drift $-g/\sqrt{a}$ and conditional uncentered second moment $C(X) + g(X)g(X)^T$ in the generator. The local C^2 expansion of g and the existing localized second moment give nonlinear drift error $K\sqrt{a}$. The drift-squared term is $O(a)$ by (30); the covariance variation is bounded by a constant times (59). This proves (60) for $a \le 1$. □

Proposition G.2 (Stable-root modulus bound)**.** *Let $\mathcal{H}_{2,1}$ be the bounded smooth test class with Lipschitz Hessian. At any stable root,*

$$\sup_{h \in \mathcal{H}_{2,1}} \left| \mathbb{E}[\mathbf{1}_{\mathcal{C}_j}(X) h((X - m_j)/\sqrt{a})] - w_{a,j} \Gamma_j h \right| \le K\{\sqrt{a} + \overline{T}_j(a;\pi_a) + R_{C,j}(a;\pi_a)\}.$$

Both $\overline{T}_j(a;\pi_a)$ and $R_{C,j}(a;\pi_a)$ tend to zero under Assumption 2.1. If C is locally β-Hölder, $0 < \beta \le 1$, then $R_{C,j}(a;\pi_a) \le Ka^{\beta/2}$; for additive noise it is zero. If local $(2+\delta)$ moments are uniformly bounded, $0 < \delta \le 1$, then $\overline{T}_j(a;\pi_a) \le Ka^{\delta/2}$.

Proof. The OU Poisson construction in Appendix F supplies u_h with $\mathcal{L}_j u_h = h - \Gamma_j h$, linear growth, and uniformly bounded derivatives up through a Lipschitz Hessian, even for singular $C(m_j)$. Apply Lemma G.1 to u_h and replace cutoff mass by cell mass at cost $O(a)$.

For the averaged tail modulus, split $\|\xi\| \le R$ and its complement. The first contribution is at most $\sqrt{a} R \mathbb{E}\|\xi\|^2$, and the second is at most $\mathbb{E}[\|\xi\|^2 \mathbf{1}\{\|\xi\| > R\}]$. Lemma A.3 proves convergence uniformly over invariant laws by sending $a \downarrow 0$ and then $R \to \infty$. For covariance continuity, with $\omega_j(r) = \sup_{\|x - m_j\| \le r} \|C(x) - C(m_j)\|$, concentration gives the useful uniform estimate

$$R_{C,j}(a;\pi_a) \le K \inf_{0 < r \le r_j} \{\omega_j(r) + a/r^2\}.$$

Under Hölder regularity a sharper direct estimate is $R_{C,j}(a;\pi_a) \le Ka^{\beta/2}\mathbb{E}[\chi_j(X)\|Z\|^\beta] \le Ka^{\beta/2}$. Finally $\min(u,1) \le u^\delta$ proves the moment implication for the stated class $\mathcal{H}_{2,1}$. □

Proposition G.3 (An explicit endpoint for standardized t_3 noise)**.** *For scalar $\varepsilon = t_3/\sqrt{3}$, whose density is $2/[\pi(1+x^2)^2]$, and $s > 0$,*

$$\mathbb{E}[\varepsilon^2 \min\{s|\varepsilon|, 1\}] = \frac{2s}{\pi} \log(1 + s^{-2}) + \frac{2}{\pi} \arctan s. \tag{61}$$

Consequently additive standardized t_3 innovations give a smooth-test upper bound $O(\sqrt{a}\log(1/a))$ for $\mathcal{H}_{2,1}$. The same order holds for a fixed finite-dimensional linear transform of independent such variables. This bound uses only the general modulus; Appendix H removes the logarithm when the root covariance is positive definite by using conditional symmetry.

Proof. Split the symmetric density integral at $1/s$:

$$\frac{4}{\pi}\left(s \int_0^{1/s} \frac{x^3}{(1+x^2)^2}\, dx + \int_{1/s}^{\infty} \frac{x^2}{(1+x^2)^2}\, dx \right).$$

Antiderivatives are $\frac{1}{2}[\log(1+x^2) + (1+x^2)^{-1}]$ and $\frac{1}{2}[\arctan x - x/(1+x^2)]$, respectively. The rational boundary terms cancel, giving (61). Substitute $s = \sqrt{a}$. For a fixed linear transform L of d such variables, set $c = \|L\|_{\mathrm{op}}$ and $u_i = |\varepsilon_i|$. Since $\|L\varepsilon\| \le c\sum_i u_i \le cd \max_i u_i$ and $q_s(u) = u^2 \min(su, 1)$ is increasing,

$$q_s(\|L\varepsilon\|) \le c^2 d^2 \sum_{i=1}^{d} u_i^2 \min(scdu_i, 1).$$

Taking expectations and applying the scalar formula at scd proves the stated order for fixed c, d; $c = 0$ is immediate. □

G.1 NOISE ON THE UNSTABLE SUBSPACE

Theorem G.4 (Excited unstable mass)**.** *Under Assumption 2.1, fix an unstable root with $P_j^u C_j (P_j^u)^T \neq 0$. Then, uniformly over invariant-law selections,*

$$w_{a,j} \le K_j\{\sqrt{a} + \overline{T}_j(a;\pi_a) + R_{C,j}(a;\pi_a)\}. \tag{62}$$

Under the local moment and covariance conditions (53), the bound is $K_j(a^{\delta/2} + a^{\beta/2} + \sqrt{a})$. For additive noise the covariance-variation term vanishes. No diagonalizability or positive definiteness of the projected covariance is required.

Proof. Choose real coordinates $T^{-1}A_jT = \operatorname{diag}(A_s, A_u)$ on the generalized stable and unstable subspaces. The spectrum of A_u has negative real part. Let L_u be the unstable rows of T^{-1}, so $L_uA_j = A_uL_u$, and set

$$B = -A_u, \qquad C_u = L_uC_jL_u^T, \qquad H = \int_0^\infty e^{-B^Tt}e^{-Bt}\,\mathrm{d}t.$$

The integral converges, $H \succ 0$, and $B^TH + HB = I$. The projection condition is equivalent to $C_u \neq 0$: $P_j^uC_j(P_j^u)^T$ is the congruence by T of the block matrix with only its uu block C_u nonzero. Define

$$\kappa_u = \frac{1}{2\lambda_{\max}(H)}, \qquad \tau = \operatorname{tr}(HC_u) > 0, \qquad s = \tau/\kappa_u.$$

For $y = L_uz$ and $q = y^THy$, use the fixed test

$$f(z) = \log(1 + q/s).$$

Writing $M = H^{1/2}L_u$, the derivatives of $\log(1 + \|v\|^2)$ give, for every fixed $k \ge 1$,

$$\sup_z \|D^kf(z)\| \le c_k\|M\|^k s^{-k/2}.$$

Indeed, each derivative is a rational function with a denominator power of $1 + \|v\|^2$ and a numerator of lower degree. Composition with $Mz/\sqrt{s}$ gives the bound. The function also grows at most linearly, so Lemma G.1 applies.

On functions of y, the root generator is $\mathcal{L}_u = (By)\cdot\nabla_y + \frac{1}{2}C_u : D_y^2$. Consequently,

$$\mathcal{L}_uq = \|y\|^2 + \tau \ge 2\kappa_uq + \tau.$$

The nonnegative matrix $H^{1/2}C_uH^{1/2}$ has trace τ, hence $y^THC_uHy \le \tau q$. The diffusion chain rule now yields

$$\begin{aligned}\mathcal{L}_jf(z) &= \frac{\|y\|^2 + \tau}{s+q} - \frac{2y^THC_uHy}{(s+q)^2} \\ &\ge \frac{2\kappa_uq + \tau}{s+q} - \frac{2\tau q}{(s+q)^2} = \kappa_u\frac{2t^2 + t + 1}{(1+t)^2} \ge \frac{\kappa_u}{2}, \qquad t = q/s.\end{aligned}$$

For the last inequality, subtracting half the denominator from the numerator leaves $3t^2/2 + 1/2 \ge 0$. With the cutoff mass $q_{a,j}$, Lemma G.1 gives

$$(\kappa_u/2)q_{a,j} \le \int \mathcal{L}_jf\,\mathrm{d}\nu_{a,j} \le K_j\{\sqrt{a} + \overline{T}_j(a;\pi_a) + R_{C,j}(a;\pi_a)\}.$$

Since $0 \le w_{a,j} - q_{a,j} \le Ka$, this proves (62). Proposition G.2 supplies the stated moment and Hölder orders. The constants depend on the fixed root and noise; they can deteriorate as hyperbolicity or excitation is lost. □

Corollary G.5 (Symmetric unstable-root bound)**.** *Under the conditions of Theorem G.4, assume also conditional symmetry of the innovation on the cutoff neighborhood. Then*

$$w_{a,j} \le K_j\{\sqrt{a} + \overline{T}_j^{\mathrm{sym}}(a;\pi_a) + R_{C,j}(a;\pi_a)\},$$

where $\overline{T}_j^{\mathrm{sym}}$ is defined in (63). This conclusion does not require ellipticity.

Proof. The same f has a bounded fourth derivative. Apply the symmetric residual of Lemma H.1 and the positive generator bound above. The smoothing assumption in Lemma H.2 is unnecessary for this explicit smooth test. □

The Jordan example after Lemma B.2 illustrates the strict gain over left-eigenvector excitation. For its unstable block, $H = \left(\begin{smallmatrix} 1/2 & 1/4 \\ 1/4 & 3/4 \end{smallmatrix}\right)$ solves the displayed Lyapunov equation. Thus the qualitative and quantitative exclusion criteria cover the same projected-noise condition.

Corollary G.6 (Quantitative marked smooth approximation)**.** *Under Assumptions 2.1 and 3.4, suppose the stable-root bound and Theorem G.4 give the same error order $r(a)$ at every root, where, for example, $r(a) = a^{\delta/2} + a^{\beta/2} + \sqrt{a}$. Then, for tests $h(j, z)$ whose slice at each label belongs to $\mathcal{H}_{2,1}$,*

$$\left| \mathbb{E}h(J_a, Z_a) - \sum_{j \in \mathcal{S}} \frac{w_{a,j}}{s_a} \Gamma_j h(j, \cdot) \right| \le K r(a).$$

For the $\mathcal{H}_{2,\delta}$ class the same moment-based conclusion follows by combining Theorem F.1 with Theorem G.4.

Proof. Sum the absolute stable-cell errors; bound every unstable cell by its mass because the test is bounded by one; then normalize the stable weights at cost at most $1 - s_a$. Finiteness of the root set completes the proof. □

H Symmetric innovations and a sharp Student endpoint

Fix any root and a compact neighborhood N_j containing the cutoff support. Assume $\xi(x, \zeta) \overset{d}{=} -\xi(x, \zeta)$ for every $x \in N_j$ and define

$$\overline{T}_j^{\mathrm{sym}}(a; \pi_a) = \mathbb{E}[\chi_j(X) \|\xi(X, \zeta)\|^2 \min\{a\|\xi(X, \zeta)\|^2, 1\}]. \tag{63}$$

This modulus tends to zero uniformly by Lemma A.3. It is bounded above by the corresponding fixed-neighborhood supremum when an explicit uniform noise bound is available. Symmetry here concerns the innovation conditional on the state, not the full increment $\xi - g$.

Lemma H.1 (Symmetric generator residual)**.** *Let $f \in C^4$ have at most linear growth and bounded derivatives of orders one through four. For the cutoff measure $\nu_{a,j}$ in (40), under Assumption 2.1 and the preceding symmetry condition,*

$$\left| \int \mathcal{L}_j f \, \mathrm{d}\nu_{a,j} \right| \le K\{\sqrt{a} + \overline{T}_j^{\mathrm{sym}}(a; \pi_a) + R_{C,j}(a; \pi_a)\}.$$

The constant may depend on the growth and derivative bounds of f, but not on a or the invariant-law selection. No stability, ellipticity or fourth innovation moment is needed for this residual.

Proof. Bounded second and fourth derivatives imply, for all v, h,

$$\left| \frac{f(v + h) + f(v - h) - 2f(v)}{2} - \frac{1}{2} h^T D^2 f(v) h \right| \le K\|h\|^2 \min\{\|h\|^2, 1\}. \tag{64}$$

For $\|h\| \le 1$, use fourth-order Taylor expansion with its integral remainder and cancellation of the odd terms. For $\|h\| > 1$, use the integral second-order formula and bounded Hessian. Put $z = (x - m_j)/\sqrt{a}$ and $v = z - \sqrt{a} g(x)$. Conditional symmetry and (64) give

$$\mathbb{E}[f(v + \sqrt{a}\xi) - f(v) \mid X = x] = \frac{a}{2} \operatorname{tr}(C(x) D^2 f(v)) + r_a(x),$$
$$|r_a(x)| \le Ka\mathbb{E}[\|\xi\|^2 \min\{a\|\xi\|^2, 1\} \mid X = x].$$

Expand the deterministic difference $f(v) - f(z)$ separately to first order. Its remainder is bounded by $Ka\|g(x)\|^2$, whose localized expectation divided by a is $O(a)$. Replacing $D^2 f(v)$ by $D^2 f(z)$ costs at most $K\sqrt{a}\mathbb{E}[\chi_j \|g(X)\|] = O(a)$ after division by a; the covariance is bounded on the cutoff support. Drift linearization costs $O(\sqrt{a})$ by (44), and covariance replacement costs $KR_{C,j}$. Finally apply stationarity to $\chi_j(x) f((x - m_j)/\sqrt{a})$. The weighted cutoff estimate (54) contributes $O(\sqrt{a})$, proving the claim. □

Lemma H.2 (An integrable fourth derivative)**.** *Suppose A is positive stable and $C \succ 0$. For smooth mollifications h_ϵ of $h \in \mathcal{H}_{2,1}$, the OU Poisson solutions $u_\epsilon = -\int_0^\infty (P_t h_\epsilon - \Gamma h_\epsilon)\,\mathrm{d}t$ have at most linear growth and bounded derivatives of orders one through four, uniformly over h and ϵ. The constants depend on A, C and the dimension, and need not remain bounded as C becomes singular.*

Proof. Mollification preserves the bounded first and second derivatives and Lipschitz Hessian. In particular, $\|D^3 h_\epsilon\|_\infty \le K$, independently of ϵ. Write $E_t = e^{-At}$ and $Q_t = \int_0^t E_s C E_s^T\,\mathrm{d}s$, with $G_t \sim \mathsf{N}(0, Q_t)$. Differentiate $P_t h_\epsilon(z) = \mathbb{E} h_\epsilon(E_t z + G_t)$ three times on the test function. For the fourth derivative, Gaussian integration by parts yields, in unit directions $v_1, \dots, v_4$,

$$D^4 P_t h_\epsilon(z)[v_1, \dots, v_4] = \mathbb{E}\big[D^3 h_\epsilon(E_t z + G_t)[E_t v_1, E_t v_2, E_t v_3]\langle Q_t^{-1} G_t, E_t v_4\rangle\big]\,.$$

Cauchy–Schwarz bounds the Gaussian score by $\|Q_t^{-1/2} E_t v_4\|$, so

$$\|D^4 P_t h_\epsilon\|_\infty \le K \|E_t\|^4 \|Q_t^{-1/2}\|.$$

Since $C \succ 0$, $Q_t \succeq ctI$ for $0 < t \le 1$; the bound is $O(t^{-1/2})$ near zero. For $t \ge 1$, $Q_t \succeq Q_1 \succ 0$ and exponential decay of E_t makes it integrable. The same affine-semigroup differentiation gives integrable bounds for the first three derivatives. Synchronous coupling gives linear growth as in (56). These majorants justify differentiation of the time integral and are uniform in the mollifier. □

Corollary H.3 (Symmetric innovations)**.** *Under Assumption 2.1, fix a stable root with $C_j \succ 0$ and $\xi(x,\zeta) \overset{d}{=} -\xi(x,\zeta)$ on N_j. For $\mathcal{H}_{2,1}$ tests,*

$$\mathcal{E}_{a,j} \le K_j\{\sqrt{a} + \overline{T}_j^{\mathrm{sym}}(a;\pi_a) + R_{C,j}(a;\pi_a)\}. \tag{65}$$

For additive noise obtained by a fixed invertible linear transform of d independent standardized t_3 coordinates, $\mathcal{E}_{a,j} = O(\sqrt{a})$.

Proof of Corollary H.3. Apply Lemma H.1 to the mollified Poisson solutions from Lemma H.2. Their generators equal $h_\epsilon - \Gamma_j h_\epsilon$. Replacing cutoff mass by cell mass costs $O(a)$ for bounded tests, as in the proof of Theorem F.1. The constants are uniform in h and ϵ. Since h is globally Lipschitz and bounded, $h_\epsilon \to h$ uniformly, so both finite-measure integrals converge. This proves the displayed bound on exactly $\mathcal{H}_{2,1}$, without assuming classical third derivatives of every unmollified test.

For a locally Lipschitz covariance, $R_{C,j} \le K\mathbb{E}[\chi_j \|X - m_j\|] = O(\sqrt{a})$. To control the other term, a sufficient uniform tail condition is

$$\sup_{x \in N_j} \mathbb{E}[\|\xi(x,\zeta)\|^2 \mathbf{1}\{\|\xi(x,\zeta)\| > R\}] \le K/R, \qquad R \ge 1. \tag{66}$$

Indeed, for $Y = \|\xi(x,\zeta)\|$,

$$\mathbb{E}[Y^2 \min\{aY^2, 1\}] = 2a \int_0^{a^{-1/2}} t\mathbb{E}[Y^2 \mathbf{1}\{Y > t\}]\,\mathrm{d}t = O(\sqrt{a})$$

uniformly over x. Fixed linear transforms of independent standardized t_3 coordinates satisfy (66). To see this in fixed dimension, bound $\|L\varepsilon\|$ by $K \max_r |\varepsilon_r|$ and its squared tail by a sum of scalar squared tails. The same argument applies to uniformly bounded state-dependent transforms when the covariance regularity and root ellipticity also hold. Local Lipschitz covariance alone does not give an explicit rate for $\overline{T}_j^{\mathrm{sym}}$. □

H.1 An exact finite-variance endpoint

The density of $\varepsilon = t_3/\sqrt{3}$ is $2/[\pi(1+x^2)^2]$. With $s = \sqrt{a}$, splitting the integral at $1/s$ gives

$$\begin{aligned} T^{\mathrm{sym}}(s^2) &= \frac{4}{\pi}\left\{ s^2 \int_0^{1/s} \frac{x^4}{(1+x^2)^2}\,\mathrm{d}x + \int_{1/s}^\infty \frac{x^2}{(1+x^2)^2}\,\mathrm{d}x \right\} \\ &= \frac{6s}{\pi} - 3s^2 + \frac{2}{\pi}(1+3s^2)\arctan s = \frac{8}{\pi}s + O(s^2). \end{aligned} \tag{67}$$

Thus symmetry improves the general $\sqrt{a}\log(1/a)$ modulus to $\sqrt{a}$ for this noise.

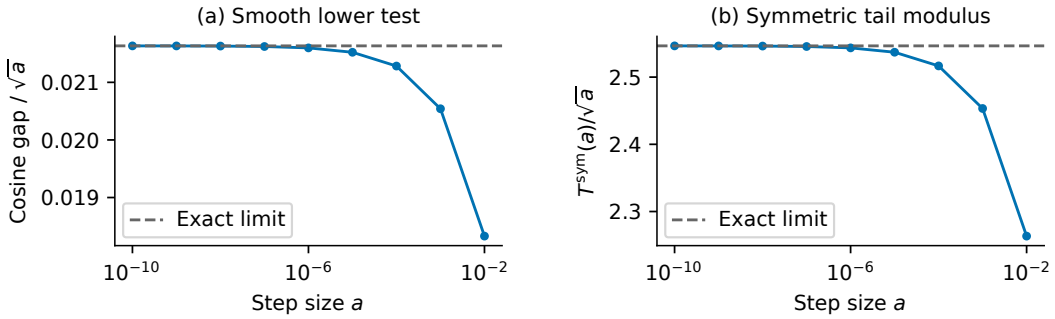

Figure 4: **A sharp symmetric finite-variance endpoint.** Analytic computations for standardized t_3 innovations. (a) The cosine gap divided by $\sqrt{a}$ approaches $e^{-1/4}/36$. (b) The symmetric tail modulus divided by $\sqrt{a}$ approaches $8/\pi$. Dashed lines are the exact limits. The characteristic function uses 100 terms of a convergent logarithmic series at 80-digit arithmetic precision; the modulus uses (67). No trajectories or sampling error bars enter these calculations.

Proposition H.4 (Matching smooth lower test)**.** *For $X^+ = (1-a)X + a\varepsilon$, $\varepsilon = t_3/\sqrt{3}$ and $0 < a < 1$, let $Z_a = X/\sqrt{a}$ have its invariant law and let $G \sim \mathsf{N}(0, 1/2)$. Then*

$$\mathbb{E}[\cos(Z_a)/4] - \mathbb{E}[\cos(G)/4] = \frac{e^{-1/4}}{36}\sqrt{a} + O(a). \tag{68}$$

Consequently the $O(\sqrt{a})$ bound of Corollary H.3 is order-sharp in this subclass.

Proof. The innovation characteristic function is $(1+|t|)e^{-|t|}$, and $Z_a = \sqrt{a}\sum_{k\geq 0}(1-a)^k \varepsilon_{-k}$ in L^2. For fixed t, use $\log \phi_\varepsilon(u) = -u^2/2 + |u|^3/3 + O(u^4)$ uniformly near zero. Geometric summation yields

$$\log \phi_{Z_a}(t) = -\frac{t^2}{2(2-a)} + \frac{|t|^3\sqrt{a}}{3(3-3a+a^2)} + O(a),$$

because the remainder sum is bounded by $Ka^2\sum_{k\geq 0}(1-a)^{4k} = O(a)$. Exponentiation gives $\phi_{Z_a}(t) - e^{-t^2/4} = e^{-t^2/4}|t|^3\sqrt{a}/9 + O(a)$. Set $t = 1$ and divide by four. Each of the four summands defining the $\mathcal{H}_{2,1}$ norm of $\cos(z)/4$ equals $1/4$, so this is an admissible lower test. □

Comparison with Gaussian and Rademacher innovations. In the same linear recursion, take $\varepsilon \sim \mathsf{N}(0,1)$ or $\Pr(\varepsilon = \pm 1) = 1/2$. Near zero, their log characteristic functions are $-u^2/2$ and $\log\cos u = -u^2/2 - u^4/12 + O(u^6)$, respectively. For fixed t, the stationary series gives

$$\log \phi_{Z_a}(t) = -\frac{t^2}{2(2-a)} + O(a) = -\frac{t^2}{4} + O(a),$$

since $a^2\sum_{k\geq 0}(1-a)^{4k} = O(a)$ controls the quartic remainder. Exponentiating and taking $t = 1$ shows that the same cosine-test gap is $O(a)$. For standardized t_3 noise, the even but nonanalytic $|u|^3$ term instead sums to order $\sqrt{a}$. Thus the difference in this example reflects tail regularity, even though all three innovations are symmetric.

The ellipticity condition enters only the extra smoothing step for $\mathcal{H}_{2,1}$. The general singular-covariance bounds remain valid, and the specialized result concerns a smooth test metric rather than a Wasserstein rate.

I Observable and loss consequences

Proof of Corollary 4.1. Write $f(m_j + y) - f(m_j) = Df(m_j)y + r(y)$, where $\|r(y)\|/\|y\| \to 0$ as $y \to 0$. Within $\mathcal{C}_j$, $y = \sqrt{a}Z_a$. For each fixed R, the scaled remainder on $\{\|Z_a\| \leq R\}$ is bounded by $R\sup_{\|y\|\leq\sqrt{a}R}\|r(y)\|/\|y\| \to 0$ (with ratio zero at $y = 0$). Conditional tightness

makes the complementary probability arbitrarily small as $R \to \infty$. Thus the remainder tends to zero in conditional probability, and Slutsky's theorem proves (19).

If f is globally Lipschitz with constant L_f, its squared scaled displacement is at most $L_f^2\|Z_a\|^2$ within the cell. More explicitly, its conditional tail above R^2 is bounded by $L_f^2 w_{a,j}^{-1}\mathbb{E}[\|Z_a\|^2\mathbf{1}\{\|Z_a\| > R/L_f\}]$. Theorem 3.2 and the weight lower bound make this tend to zero uniformly in small a; the constant-observable case is immediate. Weak convergence and this uniform integrability give W_2 convergence and convergence of both first and second moments. Subtracting the limiting zero mean proves covariance convergence. □

Proof of Corollary 4.2. Since $\nabla F(m_j) = 0$, the second-order expansion at m_j and conditional tightness give (20). Global Lipschitz continuity of ∇F gives a globally bounded Hessian, and hence $|F(x) - F(m_j)| \le \frac{L}{2}\|x - m_j\|^2$ for all x. The conditional uniform integrability established above thus permits expectation convergence in (20). Taking the trace in $H_jS_j + S_jH_j = C(m_j)$ gives the factor $1/4$. Diagonalizing the symmetric nonnegative matrix gives the chi-square representation, including zero eigenvalues. The batch claim applies the same theorem separately for each fixed B, using $C_B = C_1/B$ and $S_{j,B} = S_{j,1}/B$. It is not a joint limit theorem for $B \to \infty$. □

For the weighted assertion (22), put $L_a = [F(X) - F(m_{J_a})]/a$. The global Hessian bound gives $|L_a| \le (L/2)\|Z_a\|^2$. Along a weight-convergent subsequence, Taylor expansion and marked weak convergence identify the all-root quadratic limit of L_a; uniform integrability permits expectation convergence. Let M_j be the second moment of the all-root limit measure from Lemma C.1. The limiting mean is $\frac{1}{2}\sum_j \operatorname{tr}(H_jM_j)$. At every root, $H_jM_j + M_jH_j = q_jC_j$, so this mean equals $\frac{1}{4}\sum_j q_j \operatorname{tr} C_j$. The subsequence argument proves (22) without excitation or convergent weights. The argument also implies W_1 convergence of each nonnegligible conditional loss law to its quadratic limit, but does not control the loss variance.

Batch and discretization formulas. For independent sampling with replacement, covariance of a batch average is exactly $C_1(x)/B$. The local Lyapunov equation is linear in C, giving $S_{j,B} = S_{j,1}/B$ for each fixed B. For a batch sampled uniformly without replacement from N summands, the corresponding factor is $(N - B)/(B(N - 1))$, when C_1 is the population covariance with denominator N. For a stable linear recursion the scaled stationary covariance solves (74) whenever the spectral radius of $I - aA$ is below one. Expansion gives $S_a = S + aS^{(1)} + O(a^2)$, where $AS^{(1)} + S^{(1)}A^\mathsf{T} = ASA^\mathsf{T}$. This follows from invertibility of the continuous Lyapunov operator and a finite-dimensional perturbation expansion, and quantifies the linear discretization correction.

J AN EXACT UNEXCITED SADDLE COMPONENT

Consider $g(x, y) = (x - 2\tanh x, y)$ with additive noise $(0, \varepsilon)$, where $\mathbb{E}\varepsilon = 0$ and $\mathbb{E}\varepsilon^2 = 1$. The potential $U(x, y) = x^2/2 - 2\log\cosh x + y^2/2$ satisfies $\nabla U = g$, $\|D^2U\| \le 1$ and quadratic growth. Moreover,

$$\langle (x, y), g(x, y)\rangle \ge \tfrac{1}{2}x^2 + y^2 - 2.$$

The drift is globally Lipschitz. The scalar equation $x = 2\tanh x$ has roots $0, \pm m$. On $x > 0$, $2\tanh x - x$ is strictly concave, initially increasing, and eventually negative; hence it has exactly one positive root, with negative derivative there. Thus the nonzero roots are stable, and the origin has $A = \operatorname{diag}(-1, 1)$. All roots are hyperbolic. Translation by additive finite-variance noise gives a Feller kernel and constant covariance, satisfying the noise conditions. These facts verify Assumption 2.1.

For $0 < a < 2$, each line $x \in \{-m, 0, m\}$ is invariant and supports

$$\pi_{a,x} = \delta_x \otimes \mathcal{L}\left(a\sum_{k\ge 0}(1 - a)^k\varepsilon_{-k}\right).$$

The series converges in L^2. On the central line every point has the origin as its nearest root, so the saddle weight is one. Its scaled covariance is $\operatorname{diag}(0, 1/(2 - a))$, converging to $\bar{S}_0 = \operatorname{diag}(0, 1/2)$.

For Gaussian innovations the finite-step law is exactly Gaussian, and

$$W_2\big(\pi_{a,0}, \mathsf{N}(0, a\bar{S}_0)\big) = \sqrt{a}\left|(2-a)^{-1/2} - 2^{-1/2}\right| = O(a^{3/2}).$$

This is an exact example of Theorem 3.3, including a singular saddle component; its faster rate is specific to this linear coordinate. Arbitrary convex combinations of the three invariant laws allow fixed or step-dependent component weights. The all-root mixture remains defined even when the total stable weight is zero.

K AN EXPLICIT FAMILY OF SELECTED INVARIANT LAWS

Absorbing decompositions are established in separable finite-sum SGD (Shirokoff & Zaleski, 2025). This example supplies explicit root covariances and a local equilibration certificate within our assumptions. For $d \geq 2$, let

$$g(x) = (h(x_1), x_2, \ldots, x_d)^{\mathsf{T}}, \qquad h(u) = \frac{(u+1)u(u-2)}{1+u^2}, \tag{69}$$

and use independent Rademacher innovations of amplitude $\delta = 1/5$ in every coordinate. The roots are $m_- = (-1, 0, \ldots, 0)$, $m_0 = 0$ and $m_+ = (2, 0, \ldots, 0)$. The first-coordinate derivatives are $3/2, -2, 6/5$, respectively.

Proposition K.1 (Arbitrary selected weights)**.** *For every $0 < a \leq 1/6$, the recursion with (69) has invariant laws supported in the two disjoint rectangles*

$$R_- = [-6/5, -4/5] \times [-1/5, 1/5]^{d-1}, \quad R_+ = [9/5, 11/5] \times [-1/5, 1/5]^{d-1}. \tag{70}$$

Denote any such laws by π_a^- and π_a^+. For every function $c(a) \in [0,1]$, $c(a)\pi_a^- + (1-c(a))\pi_a^+$ is invariant. Its local Gaussian covariances are

$$S_- = \mathrm{diag}(\delta^2/3, \delta^2/2, \ldots), \qquad S_+ = \mathrm{diag}(5\delta^2/12, \delta^2/2, \ldots). \tag{71}$$

An oscillating selection $c(a)$ can prevent even the marginal law of Z_a from converging.

Proof of Proposition K.1. Polynomial division gives

$$h(u) = u - 1 + \frac{-3u+1}{1+u^2}, \qquad h'(u) = 1 + \frac{3u^2 - 2u - 3}{(1+u^2)^2}.$$

Thus h is smooth, $h(u) - u$ is bounded and $|h'(u)| \leq 6$. For example, bound the three numerator terms separately using $u^2/(1+u^2)^2 \leq 1/4$, $2|u|/(1+u^2)^2 \leq 1$ and $(1+u^2)^{-2} \leq 1$. Consequently $u \mapsto u - ah(u) + a\eta$ is nondecreasing whenever $a \leq 1/6$, for every $\eta \in [-1/5, 1/5]$. The exact endpoint values are

$$h(-6/5) = -96/305 < -1/5, \quad h(-4/5) = 56/205 > 1/5,$$
$$h(9/5) = -63/265 < -1/5, \qquad h(11/5) = 88/365 > 1/5.$$

At each left endpoint the update with the smallest innovation points inward; at each right endpoint the update with the largest innovation points inward. Monotonicity then maps the entire interval into itself for every allowed innovation. For the remaining coordinates, $x_r^+ = (1-a)x_r + a\eta_r$ preserves $[-1/5, 1/5]$ when $a \leq 1$. Both rectangles in (70) are therefore invariant. The finite-noise transition kernel is Feller, so Cesàro averaging on either compact rectangle supplies an invariant probability measure supported there.

The drift is globally Lipschitz and dissipative because $g(x) = x + O(1)$. It is the gradient of $U(x) = \int_0^{x_1} h(u)\,\mathrm{d}u + \frac{1}{2}\sum_{r=2}^d x_r^2$. This potential has at most quadratic growth, bounded Hessian and $\langle \nabla U, g\rangle = \|g\|^2$. The three roots are m_-, m_0, m_+, with Jacobians $\mathrm{diag}(3/2, 1, \ldots)$, $\mathrm{diag}(-2, 1, \ldots)$ and $\mathrm{diag}(6/5, 1, \ldots)$. The innovation covariance is $(1/25)I$, which excites the unstable first coordinate at the middle root. All qualitative assumptions hold.

Each rectangle lies in its corresponding stable root's Voronoi cell. Linearity of the stationary equation makes every stated convex combination invariant, and its cell weights are exactly $c(a)$ and $1 - c(a)$. Solving the diagonal Lyapunov equations gives (71). For example, choosing $c(a) = (1 + \sin(\log(1/a)))/2$ gives sequences tending to zero on which $c(a)$ is respectively zero and one. Theorem 3.3 identifies distinct marginal Gaussian limits on these sequences, because their first-coordinate variances differ. This proves the final assertion. □

Proposition K.2 (Certified local equilibration). *For Proposition K.1 with $h(u) = (u+1)u(u-2)/(1+u^2)$ and coordinatewise Rademacher noise of amplitude $1/5$, each rectangle R_- and R_+ contains a unique invariant law $\pi_a^{\pm}$ for $0 < a \le 1/6$. For $x \in R_\pm$, after n updates,*

$$W_2\left(\mathcal{L}\left(\frac{X_n - m_\pm}{\sqrt{a}}\right), \mathcal{L}_{\pi_a^\pm}\left(\frac{X - m_\pm}{\sqrt{a}}\right)\right) \le \frac{\operatorname{diam}(R_\pm)}{\sqrt{a}}(1-a/2)^n \le \frac{\operatorname{diam}(R_\pm)}{\sqrt{a}} e^{-an/2}.$$

Proof. The preceding endpoint argument proves invariance. The derivative satisfies

$$h'(u) = \frac{u^4 + 5u^2 - 2u - 2}{(1+u^2)^2}.$$

On $[-6/5, -4/5]$, separate bounds on the numerator and denominator give $h'(u) \ge [(4/5)^4 + 5(4/5)^2 + 8/5 - 2]/[1+(6/5)^2]^2 = 2006/3721 > 1/2$; on $[9/5, 11/5]$ they give $h'(u) \ge [(9/5)^4 + 5(9/5)^2 - 22/5 - 2]/[1+(11/5)^2]^2 = 12686/21316 > 1/2$. Together with the global bound $h'(u) \le 6$, the first-coordinate update derivative lies in $[0, 1-a/2]$. Every other coordinate has derivative $1-a$. Thus, under a common noise realization, the update contracts Euclidean distances within each convex rectangle by at most $1 - a/2$. Existence follows from the compact Feller argument above; synchronous coupling implies uniqueness and $W_2(P_a^n(x,\cdot), \pi_a^\pm) \le (1-a/2)^n \operatorname{diam}(R_\pm)$. Scaling by $1/\sqrt{a}$ proves the statement. □

The heterogeneous-root experiment uses physical burn-in $an \ge 2\log[\operatorname{diam}(R_\pm)/(\epsilon\sqrt{a})]$ to make its scaled initialization error at most ϵ. Retained observations are temporally dependent; uncertainty is measured across independent trajectories.

L Global communication and component weights

Local moment accuracy determines the fluctuations of each component. Selecting its occupation weight requires additional communication and stationary-limit hypotheses, supplied in different settings by metastability theory.

Neighborhood and cell weights. For a stable root and sufficiently small fixed r, write $p_{a,j}(r) = \pi_a(B(m_j, r))$. The concentration theorem gives $0 \le w_{a,j} - p_{a,j}(r) \le K_r a$. This transfers order-one limits between neighborhood and Voronoi weights. It cannot transfer exponential asymptotics: $p_a = e^{-1/a}$ and $w_a = e^{-1/a} + a$ differ by $O(a)$ but have different limits of $a\log(\cdot)$. Likewise, an empirical time occupation from a chosen initialization is not either stationary weight unless an appropriate ergodic approximation is justified.

Light tails. For additive innovations, let $\Lambda(\theta) = \log \mathbb{E} e^{\theta^\top \varepsilon}$ and let Λ^* be its Legendre transform. When an applicable path large-deviation theorem holds on physical time $t = na$, the candidate action is

$$I_{0T}(\phi) = \int_0^T \Lambda^*(\dot\phi(t) + g(\phi(t)))\,\mathrm{d}t$$

for absolutely continuous paths, at speed $1/a$. Under the required uniform path, communication and tightness hypotheses, directed minimum-cost trees built from restricted inter-attractor path costs describe logarithmic neighborhood occupations; see Kifer (1990) and the SGD analyses of Azizian et al. (2024; 2025). If W_j is the resulting tree cost toward root j, the neighborhood exponent takes the form $-(W_j - \min_k W_k)$, with the neighborhood limits prescribed by the applicable theorem. Equal tree costs do not identify the relative order-one weights: prefactors or another selection argument are needed.

For isotropic Gaussian innovations $\mathsf{N}(0, \sigma^2 I)$ and $g = \nabla F$, the action satisfies

$$I_{0T}(\phi) = \frac{1}{2\sigma^2}\int_0^T \|\dot\phi - \nabla F(\phi)\|^2\,\mathrm{d}t + \frac{2}{\sigma^2}\{F(\phi(T)) - F(\phi(0))\}.$$

The familiar barrier cost follows when the relevant reverse-gradient connection exists, interpreted through arbitrarily long approximating paths. A reverse-gradient solution started exactly at an equilibrium remains there, so finite-time attainment from that exact initial point is not implied. The

diffusion $\mathrm{d}X = -\nabla F(X)\,\mathrm{d}t + \sqrt{a}\sigma\,\mathrm{d}W$ has a Gibbs stationary density proportional to $e^{-2F/(a\sigma^2)}$ under the usual integrability conditions. Its Laplace prefactors do not automatically transfer to the discrete recursion studied here.

Regularly varying tails. For an additive innovation with regularly varying tail index $\alpha > 2$, write $H(R) = \mathbb{P}(\|\varepsilon\| > R)$ and suppose the rescaled tail measures converge to ν_{tail} away from zero. A displacement of order one in an update $a\varepsilon$ occurs with probability on the scale $H(1/a)$ per iteration. The corresponding waiting scales are $1/H(1/a)$ iterations and $a/H(1/a)$ units of physical time. If the deterministic attraction basins are $\mathcal{B}_j$, the one-jump candidate rates are

$$q_{ij} = \nu_{\mathrm{tail}}(\mathcal{B}_j - m_i), \qquad i \neq j, \qquad q_{ii} = -\sum_{j\neq i} q_{ij}.$$

Turning this formula into a well-chain limit requires stability and return control, negligible boundary landings such as $\nu_{\mathrm{tail}}(\partial(\mathcal{B}_j - m_i)) = 0$, and the hypotheses of an appropriate discrete-time metastability theorem. Convergence of stationary weights to the stationary vector of this chain additionally requires control of the stationary limit and, for a unique vector, irreducibility of the limiting chain. The diffusion results of Imkeller & Pavlyukevich (2008); Högele & Pavlyukevich (2015) and discrete heavy-tail analyses of Wang et al. (2022); Wang & Rhee (2023) establish versions of these mechanisms in their respective settings. The clipped dynamics in the former discrete reference need not have the same one-jump rates as unclipped updates. Equation (75) is the symmetric scalar $\alpha = 3$ benchmark for the unclipped experiment, not a consequence of finite variance alone.

M A SHARP-ORDER EXAMPLE FOR SMOOTH TESTS

Proposition M.1 (A bounded-noise $\sqrt{a}$ lower example)**.** *Assumption 2.1, even with bounded additive innovations and a globally linear drift, does not imply an $O(a)$ Gaussian smooth-test error. There is a one-root example whose error on $\mathcal{H}_{2,1}$ is at least a positive constant times $\sqrt{a}$ for sufficiently small a.*

Proof. Take $g(x) = x$, $0 < a < 1$, and innovations

$$\Pr(\varepsilon = \sqrt{2}) = 1/3, \qquad \Pr(\varepsilon = -1/\sqrt{2}) = 2/3.$$

They have mean zero, variance one, and third moment $m_3 = 1/\sqrt{2}$. The stationary scaled variable is

$$Z_a = \sqrt{a}\sum_{k\geq 0}(1-a)^k\varepsilon_k.$$

The Gaussian limit has variance $1/2$. For fixed real t, choose the continuous logarithm near zero with $\log\phi_\varepsilon(0) = 0$. Boundedness of the innovations permits a uniform fourth-order remainder in the logarithm of their characteristic function near zero. Summing that expansion over $\sqrt{a}(1-a)^k t$ gives

$$\log \mathbb{E}e^{itZ_a} = -\frac{t^2}{2(2-a)} - \frac{im_3t^3\sqrt{a}}{6(3-3a+a^2)} + O(a),$$

since $\sum_{k\geq 0} a^2(1-a)^{4k} = O(a)$. Hence

$$\mathbb{E}\sin(tZ_a) = -e^{-t^2/4}\frac{m_3t^3}{18}\sqrt{a} + O(a).$$

At $t = 1$, the bounded test $h(z) = \sin z/4$ belongs to $\mathcal{H}_{2,1}$: each of its sup-norm, gradient norm, Hessian norm, and Hessian Lipschitz seminorm is at most $1/4$. Its Gaussian expectation is zero, so it yields the claimed lower bound. This example also explains why symmetric-noise or covariance-only experiments may fail to display the worst-case $\sqrt{a}$ order. □

N A CONTROLLED NEURAL FINITE-SUM WITNESS

A fixed linear tilt makes nondegeneracy generic, while bounded coordinate dither excites every direction. These devices define the augmented finite-sum model used in the neural experiment.

Proposition N.1 (Generic tilt and bounded full-rank finite-sum noise)**.** *Fix a finite dataset* $(x_i, y_i)_{i=1}^N$ *with* $y_i \in \{-1, 1\}$*, fixed output weights* c*, and a network* $f_W(x) = c^T \tanh(Wx)$ *in which only* W *is trained. Write* $\theta = \mathrm{vec}(W) \in \mathbb{R}^p$ *and*

$$\ell_i(\theta) = \log(1 + e^{-y_i f_W(x_i)}), \quad F_0(\theta) = \frac{1}{N}\sum_i \ell_i(\theta) + \frac{\lambda}{2}\|\theta\|^2, \qquad \lambda > 0.$$

Draw $r \in \mathbb{R}^p$ *from an absolutely continuous distribution* once*, and hold it fixed. For fixed* $\tau, \gamma > 0$*, define*

$$F_r(\theta) = F_0(\theta) + \tau r^T \theta,$$

and the augmented finite sum, indexed uniformly by $i \in \{1, \dots, N\}$*,* $k \in \{1, \dots, p\}$*,* $\sigma \in \{-1, 1\}$*,*

$$\widetilde{\ell}_{i,k,\sigma}(\theta) = \ell_i(\theta) + \frac{\lambda}{2}\|\theta\|^2 + \tau r^T \theta + \gamma\sqrt{p}\,\sigma\theta_k.$$

For almost every fixed r*, iid finite-sum SGD on these* $2Np$ *equally weighted summands satisfies Assumption 2.1 and excitation at every unstable root, in every fixed finite dimension. Its mean objective is* F_r*, and its one-sample noise covariance is*

$$C(\theta) = \mathrm{Cov}_i(\nabla \ell_i(\theta)) + \gamma^2 I_p.$$

All local innovation moments exist, and the covariance is locally Lipschitz. The theorem constants may depend on the fixed tilt and root conditioning; no uniform spectral gap over random tilts is asserted.

Proof. Because the dataset and c are fixed and finite, $f_W(x_i)$ and its derivatives of any fixed positive order in θ are globally bounded: derivatives of $\tanh$ of any fixed order are bounded, and each differentiation contributes fixed input factors. The logistic loss is smooth; on the bounded logit range its derivatives are bounded. Thus every ℓ_i has globally bounded derivatives of all fixed positive orders.

Consequently $g_r = \nabla F_r$ is smooth, globally Lipschitz, and of the form $\lambda\theta + O(1)$ for fixed r. It is dissipative, F_r is coercive with at most quadratic growth, its Hessian is globally bounded, and $U = F_r$ satisfies $\langle \nabla U, g_r \rangle = \|g_r\|^2$.

Apply Sard's regular-value theorem to the smooth map $G = \nabla F_0 : \mathbb{R}^p \to \mathbb{R}^p$. Its critical values have Lebesgue measure zero. Since $-\tau r$ has an absolutely continuous law, it is a regular value with probability one. Every root of g_r therefore has nonsingular Hessian. The Hessian is real symmetric, so these roots are hyperbolic for the gradient flow. The root set is closed, and it is bounded because $G(\theta) = \lambda\theta + O(1)$. All its points are isolated by the inverse function theorem; a compact set of such regular roots is finite (otherwise an accumulation point would contradict local invertibility). Coercivity supplies a minimizer, so the root set is nonempty.

The gradient of each augmented summand is a continuous function. Finite averaging therefore gives a Feller transition. The ridge and fixed tilt cancel when forming centered gradient noise. The remaining data gradient fluctuation and the coordinate dither are uniformly bounded, centered, and independent when their indices are sampled independently. The dither covariance is $\gamma^2 p\,\mathbb{E}(e_k e_k^T) = \gamma^2 I_p$, so the displayed covariance formula follows. In particular, for every nonzero complex vector v, $v^* C(\theta) v \geq \gamma^2 \|v\|^2 > 0$. This implies the required unstable-direction excitation at every unstable root. Bounded noise supplies the covariance growth bound with $c_1 = 0$ and every local innovation moment. Smooth finite-sum gradients give covariance continuity and local Lipschitz regularity. □

The regular-value result is Sard's theorem (Lee, 2013). It verifies the augmented model's structural assumptions almost surely in the fixed tilt. The unaugmented arm requires separate verification, and finite-time equilibration and positive component weights remain separate questions in either model.

For a batch of size B, sample B independent triples (i, k, σ) and average their complete gradients, including the dither. Then $C_B = (C_{\mathrm{data}} + \gamma^2 I)/B$.

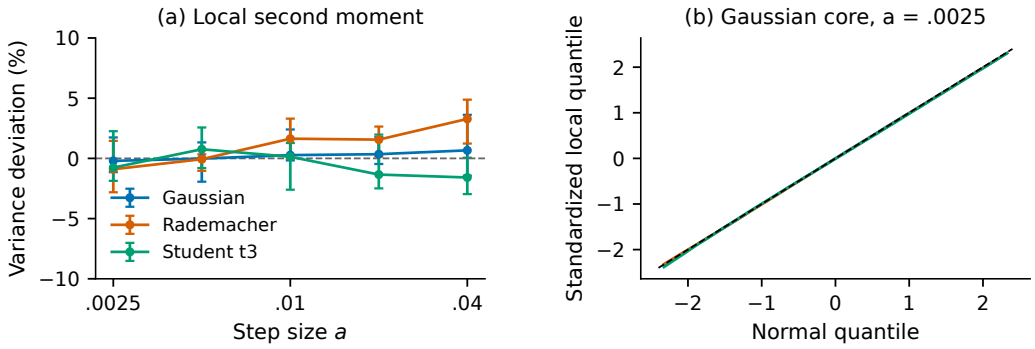

Figure 5: Local fluctuations with matched innovation covariance. Variance deviations in (a) are percentages relative to the OU target on a zero-centered axis. (a) Median variance deviation and interquartile range across 16 runs; the continuous OU prediction is zero. (b) Pooled first-coordinate Q–Q curves at $a = 0.0025$. The intervals describe between-run dispersion.

O LOCAL FLUCTUATIONS AND GLOBAL COMMUNICATION

O.1 MATCHED-COVARIANCE LOCAL FLUCTUATIONS

Consider

$$F(x) = \sum_{r=1}^{2} \{x_r^2/2 - c_r \log\cosh x_r\}, \qquad (c_1, c_2) = (2, 1.5),$$
$$g = (I + 0.8J)\nabla F, \qquad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{72}$$

There are four attracting roots $(\pm m_1, \pm m_2)$, with $m_1 = 1.91500805$ and $m_2 = 1.28783945$. All four have the same local Jacobian and innovation covariance,

$$A = \begin{pmatrix} 0.83362791 & -0.48454958 \\ 0.66690233 & 0.60568697 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 2 \end{pmatrix}, \quad S = \begin{pmatrix} 0.79678347 & 0.33891463 \\ 0.33891463 & 1.27784990 \end{pmatrix}. \tag{73}$$

We match C across Gaussian, Rademacher and standardized Student t_3 innovations. Each of five step sizes uses 16 independent runs of physical duration 5,000, with duration 50 discarded. We assign samples to the nearest of all nine roots and pool stable marked displacements. Pooling is valid here because their theoretical Gaussian components coincide.

Figure 5 shows central Q–Q curves close to the common Gaussian reference at the smallest step, alongside appreciable run-to-run variance dispersion.

Finite-step reference. An exactly linear recursion has scaled stationary covariance

$$S_a = (I - aA)S_a(I - aA)^\mathsf{T} + aC, \quad AS_a + S_aA^\mathsf{T} - aAS_aA^\mathsf{T} = C. \tag{74}$$

This reference isolates linear discretization effects; it is not the nonlinear chain's covariance. For gradient drift, the corresponding quadratic loss admits a simple relative bound.

Proposition O.1 (Discrete correction for a frozen quadratic model)**.** *Let $H = H^T \succ 0$, let $C \succeq 0$ be the covariance of centered iid additive innovations, and suppose $\epsilon = a\lambda_{\max}(H) < 2$. For the stationary linear recursion $Y^+ = (I - aH)Y + a\xi$, define $L_a = \frac{1}{2}\mathbb{E}(Y^T HY)$ and $L_{\mathrm{OU}} = a\,\mathrm{tr}\,C/4$. Then*

$$0 \le L_a - L_{\mathrm{OU}} \le \frac{\epsilon}{2 - \epsilon} L_{\mathrm{OU}}.$$

If $\mathrm{tr}\,C > 0$, division by L_{OU} gives the same bound on the relative correction.

Proof. All eigenvalues of $I - aH$ have modulus less than one, so the stationary covariance V_a exists and satisfies $HV_a + V_aH - aHV_aH = aC$. In an orthonormal eigenbasis of H, write its

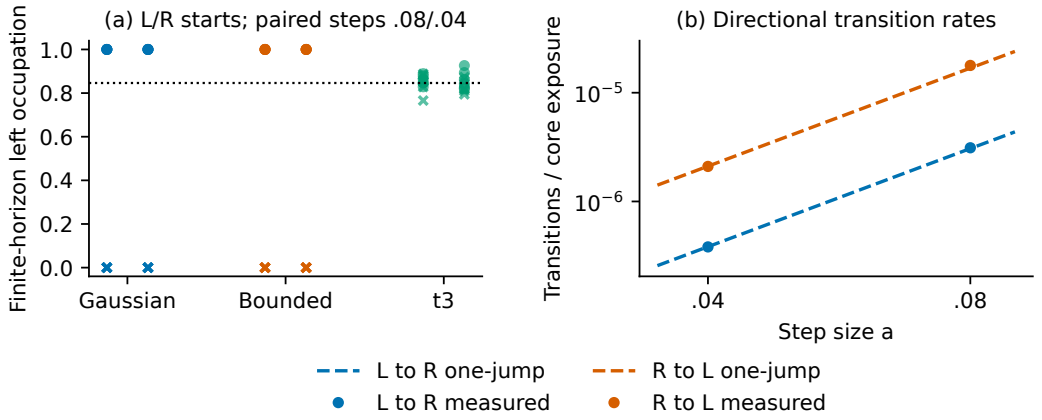

Figure 6: Global occupation and transition rates. (a) Per-run left occupation, with paired positions for $a = 0.08$ and 0.04; circles start left and crosses start right. The dotted line is the heavy-tail limit (75). (b) Heavy-tail completed transitions divided by exposure to the last visited well core, compared with one-jump probabilities per iteration. Run lengths are reported in Table 3.

eigenvalues as λ_i and the diagonal entries of C as $c_i \geq 0$. The diagonal covariance equations give $(V_a)_{ii} = ac_i/[\lambda_i(2 - a\lambda_i)]$, hence

$$L_a = \frac{a}{2}\sum_i \frac{c_i}{2 - a\lambda_i}, \qquad L_a - L_{\mathrm{OU}} = \frac{a}{4}\sum_i c_i \frac{a\lambda_i}{2 - a\lambda_i} \leq \frac{\epsilon}{2 - \epsilon}\frac{a}{4}\sum_i c_i.$$

This calculation requires no commutation between H and C. □

At the largest neural normalized step, $\epsilon = 0.005$, the bound is 0.2507% after rounding upward. It quantifies the loss correction within the frozen quadratic reference; the nonlinear objective is evaluated separately in the SGD experiment.

Student t_3 noise has an infinite fourth moment. The Q–Q comparison and bounded smooth probes in Appendix P assess distributional shape without requiring fourth moments or fitting a convergence-rate exponent.

O.2 Communication and finite-horizon occupation

For $g(x) = x + 5\cos x$, the stable roots are $m_L = -1.30644001$ and $m_R = 3.83746711$, separated by $s = 1.97738303$. We again compare three innovation laws of variance one, using 12 independent runs per step and noise law. Half start at each minimum. For standardized t_3 noise, the one-large-jump prediction is

$$q_{LR} \propto (s - m_L)^{-3}, \quad q_{RL} \propto (m_R - s)^{-3}, \quad w_L^{\mathrm{jump}} = \frac{q_{RL}}{q_{LR} + q_{RL}} = 0.84620744. \tag{75}$$

We choose heavy-tail run lengths before simulation using the jump proxy, targeting 400 completed transitions in total at each step size. Gaussian and bounded-noise comparators use two million iterations per run. The horizons are chosen to resolve transitions and occupations under each noise law.

We distinguish separatrix crossings from completed transitions between radius-0.25 well cores. Figure 6 shows that local covariance matching does not imply similar finite-horizon basin exploration. For runs without transitions, averaging equally many initializations in both wells returns $1/2$ by design. That average does not identify a stationary weight. Conversely, a Gaussian run without observed transitions does not establish noncommunication. Detailed counts and horizons appear in Table 3.

P NUMERICAL METHODS AND ADDITIONAL MEASUREMENTS

The supplementary local-shape, occupation and output-coverage studies comprise 240, 72 and 192 seeded trajectories, respectively. Appendix Q specifies the root-specific moment and objective-gap measurements. The files `code/reproduce.py` and `code/simulation.py` generate these measurements directly from the recursions. Running the first script without arguments regenerates figures and tables from the shipped data; adding `-run` also regenerates the simulations. Per-run CSV files retain the seeds and estimators; compressed arrays retain the local Q–Q samples and running occupation curves. Version and duration information is recorded in `data/metadata.json`.

P.1 LOCAL-NOISE EXPERIMENT

The roots of the four-well system are the nine combinations of $\{-m_1, 0, m_1\}$ and $\{-m_2, 0, m_2\}$. The four combinations with nonzero coordinates are stable; the other roots are unstable. The scalar curvatures at m_1, m_2 are $1 - c_r \operatorname{sech}^2(m_r)$, giving (73). Here $U = F$ is a valid descent function because $\langle \nabla F, (I + 0.8J)\nabla F\rangle = \|\nabla F\|^2$ and $\|(I + 0.8J)\nabla F\|^2 = (1 + 0.8^2)\|\nabla F\|^2$. Moreover, $\nabla F = x + O(1)$, so $\langle x, g(x)\rangle = \|x\|^2 + O(\|x\|)$. These identities verify descent and dissipativity for this nonreversible example.

Each step in $(0.04, 0.02, 0.01, 0.005, 0.0025)$ uses 16 independent trajectories starting at the positive minimum. The total iteration count is $\lfloor 5000/a \rfloor$, the burn-in count is $\lfloor 50/a \rfloor$, and retained samples are spaced by $\lfloor 0.5/a \rfloor$ iterations. Thus the target retained physical spacing is 0.5 (it is 0.48 at $a = 0.04$ owing to integer rounding). Uncertainty and dispersion are summarized at the independent-run level.

Draw two independent scalar innovations of variance one, either standard normal, Rademacher, or $t_3/\sqrt{3}$, and multiply them by $L = \left(\begin{smallmatrix} 1 & 0 \\ 0.4 & \sqrt{1.84} \end{smallmatrix}\right)$. This gives $LL^\mathsf{T} = C$ for all three laws. The t_3 law has finite $(2 + \delta)$ moments only for $\delta < 1$ and has infinite third absolute and fourth moments. The qualitative theorem applies; the moment-based additive rate theorem applies for any $0 < \delta < 1$, while Corollary H.3 gives the sharper $O(\sqrt{a})$ smooth-test bound because this noise is symmetric and $C \succ 0$.

All labeling uses the nearest of all nine roots. Within a run, retain stable-labeled displacements $Z = (X - X_{\mathcal{M}})/\sqrt{a}$ and use the usual centered sample covariance with denominator $n - 1$. The retained stable fraction is at least 0.9795 over these runs; other labels are excluded and their fractions are recorded. The four stable components have the same S, so pooling them has a common local target without estimating global weights. Figure 5(a) reports the median and interquartile range of the run-level variance estimates. Panel (b) pools retained first-coordinate samples at the smallest step and divides by $\sqrt{S_{11}}$; the reference quantiles range from 0.01 to 0.99. These are finite-horizon measurements of local fluctuations.

We also use bounded smooth distributional probes at six fixed frequencies:

$$\Theta = \{(1,0), (0,1), (1,1), (1,-1), (2,0), (0,2)\}.$$

$$\widehat{e} = \max_{\theta \in \Theta} \left| \frac{1}{n} \sum_{k=1}^{n} e^{i\theta^\mathsf{T} Z_k} - e^{-\theta^\mathsf{T} S\theta/2} \right|. \tag{76}$$

Real and imaginary parts are bounded smooth tests; fixed-frequency normalization places them within a constant multiple of the theorem's test class. This fixed-frequency diagnostic contains both Monte Carlo variation and finite-step error. Figure 7(a) reports run-level medians and interquartile ranges, without fitting a rate exponent.

P.2 COMMUNICATION EXPERIMENT AND TRANSITION BUDGET

For $g(x) = x + 5\cos x$, use independent scalar innovations from the same three variance-one laws. At each $a \in \{0.08, 0.04\}$, use 12 runs, alternating their initial states between m_L and m_R. The first 100,000 iterations are discarded. For the $t_3/\sqrt{3}$ innovation, define the exact one-step tail probabilities at the minima

$$p_{LR}(a) = \mathbb{P}(t_3 > \sqrt{3}(s - m_L)/a), \qquad p_{RL}(a) = \mathbb{P}(t_3 > \sqrt{3}(m_R - s)/a).$$

Table 3: Occupation experiment: 12 runs per row. N includes burn-in; counts sum post-burn-in events over runs. Heavy-tail occupation is the mean with one across-run standard error in parentheses. *The value 0.500 is fixed by six starts in each well and no observed crossings; it is not a stationary-weight estimate.

Noise	a	N per run	Left occupation	Raw	Completed
Gaussian	0.08	2,000,000	0.500^*	0	0
Gaussian	0.04	2,000,000	0.500^*	0	0
Rademacher	0.08	2,000,000	0.500^*	0	0
Rademacher	0.04	2,000,000	0.500^*	0	0
Student t3	0.08	6,600,000	0.852 (0.010)	477	413
Student t3	0.04	51,500,000	0.845 (0.011)	522	400

Under the two-state proxy the stationary mean number of transitions per iteration is $r(a) = 2p_{LR}(a)p_{RL}(a)/(p_{LR}(a) + p_{RL}(a))$. Before running any trajectory, set the post-burn-in length to

$$N_{\mathrm{post}}(a) = 10^5 \left\lceil \frac{400}{12r(a)10^5} \right\rceil .$$

This gives 6,500,000 and 51,400,000 post-burn-in iterations per heavy-tail run at the two steps. The corresponding total lengths are 6,600,000 and 51,500,000. The proxy fixes each run's horizon before observing any realized transition counts. Gaussian and Rademacher comparisons use 2,000,000 total iterations per run, including the same burn-in.

Left occupation is the fraction of post-burn-in iterates with $X_n < s$. Raw crossings count changes in that indicator. Completed transitions instead track the last visited radius-0.25 core about a minimum: the well label changes only when the opposite core is reached. Its state is tracked throughout burn-in; reported counts and exposure use only the retained time interval. Directional rate estimates divide completed $i \to j$ transitions by the number of iterations whose pre-update last-core label is i. They are per-iteration rates and are compared with $p_{ij}(a)$, not directly with rates in a normalized limiting clock.

The heavy-tail runs produce 413 completed transitions at $a = 0.08$ and 400 at $a = 0.04$ (raw crossings: 477 and 522). Their mean left occupations are 0.8517 and 0.8448, with across-run standard errors 0.0099 and 0.0113. The smaller-step estimate is close to the asymptotic benchmark 0.8462 at this precision. Two step sizes and these horizons do not identify an asymptotic bias exponent. Zero observed crossings in the other noise laws describe the sampled horizon, while Proposition K.1 separately proves exact noncommunication in a different explicit bounded-noise model.

P.3 Finite-sum SGD and prediction functions

For the four losses in (27), each iteration samples B independent sign pairs with replacement and averages their gradients. The simulator uses the equivalent symmetric centered-noise sign convention in (1). For each $B \in \{1, 4, 16\}$ and $a \in \{0.04, 0.02, 0.01, 0.005\}$, run 16 trajectories with the same total physical duration, burn-in and thinning as the local experiment. All start at the positive minimum and all retained samples in these runs remain in its nearest-root cell. This is a local finite-horizon study and does not infer their global stationary weights. For this gradient drift, the local covariance solves $HS_1 + S_1H = C_1(m)$.

Fix $f_1(x) = x_1 + x_2$ and $f_2(x) = \tanh(x_1 - x_2/2)$ before simulation. Let $u = (1, -1/2)^{\mathsf{T}}$ and let G have rows $(1, 1)$ and $\operatorname{sech}^2(u^{\mathsf{T}}m)u^{\mathsf{T}}$. The predicted covariance of $\sqrt{B/a}\{f(X) - f(m)\}$ is GS_1G^{T}. For each run, measure the centered sample variance ratio against the corresponding diagonal entry and the fraction satisfying

$$\left|\sqrt{B/a}\{f_k(X) - f_k(m)\}\right| \le 1.9599639845\sqrt{(GS_1G^{\mathsf{T}})_{kk}}.$$

This last quantity is nominal Gaussian fluctuation coverage around a deterministic local value; it is not coverage of a confidence interval for a population parameter. Neither S_1 nor G is estimated from trajectories. Figure 7(b) gives the nonlinear observable's run-averaged coverage; individual-

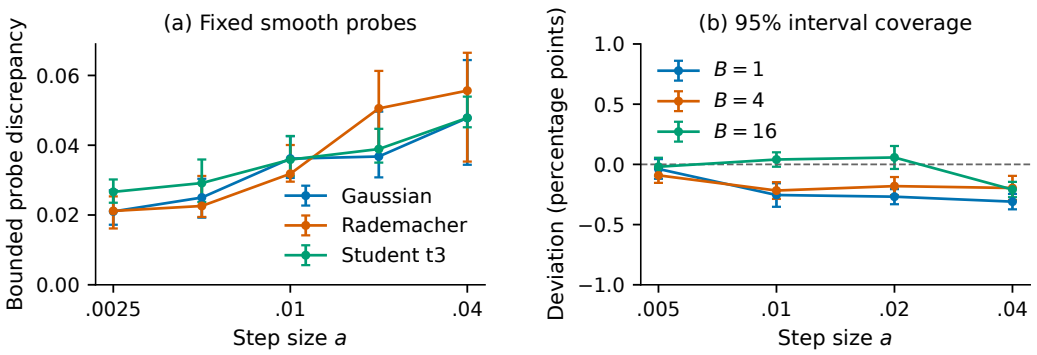

Figure 7: Additional local measurements. Coverage in (b) is shown as percentage-point deviation from 95 percent on a zero-centered axis. (a) Bounded characteristic-function discrepancy (76), shown as median and interquartile range across 16 runs. (b) Nominal 95% local fluctuation coverage for the nonlinear observable, shown as mean and one standard error across runs. These finite-horizon measurements describe distributional agreement and coverage; the asymptotic rates are established by the analytic results.

run values range from 0.9379 to 0.9557. The parameter covariance, both observable variance ratios and both coverages are retained in `data/sgd.csv`.

Seed allocation. Indices start at zero in the listed parameter order. Local seeds are $270900 + 10000k + 100i + r$, where k indexes (Gaussian, Rademacher, Student t_3), i indexes the five descending steps and $r = 0, \dots, 15$. Occupation seeds are $370900 + 10000k + 100i + r$, with steps $(0.08, 0.04)$ and $r = 0, \dots, 11$. SGD seeds are $470900 + 10000b + 100i + r$, where b indexes batch sizes $(1, 4, 16)$, i indexes the four descending steps and $r = 0, \dots, 15$. Random draws use NumPy's random functions within Numba-compiled loops; the shipped dependency versions are recorded to support reruns. Every standardized t_3 draw is constructed as $N_0/\sqrt{N_1^2 + N_2^2 + N_3^2}$ from four independent standard Gaussian variables. This has law $t_3/\sqrt{3}$ and variance one; the denominator is not truncated at a fixed positive threshold.

Q Experimental protocols

Estimands and independent replication. Covariance and loss studies use 16 independently seeded trajectories per setting, mean confirmation uses 32, and the supplementary occupation study uses 12. For R independent run-level statistics T_r, the reported standard error is $s(T_r)/\sqrt{R}$, with sample variance denominator $R - 1$. CDF bands use the 16 run-level ECDFs pointwise and are not simultaneous bands. For infinite-fourth-moment innovations, raw quadratic estimates are displayed as independent-run distributions or medians and interquartile ranges. The fixed neural probes are averaged within each run. Mean absolute relative error (MARE) for path r is $K^{-1} \sum_k |\widehat{v}_{r,k}/v_k - 1|$; its reported mean averages this statistic across paths. Aggregate MARE is $K^{-1} \sum_k |R^{-1} \sum_r \widehat{v}_{r,k}/v_k - 1|$. The signed error averages the latter ratios without absolute values. Root-centered second moments use the deterministic-root output instead of the trajectory mean.

Heterogeneous invariant components. The two-dimensional recursion (69) uses independent noise coordinates ± 0.2, initialized at m_- or m_+. The step grid is $0.04, 0.02, 0.01, 0.005, 0.0025$. The burn-in count is

$$n_{\mathrm{burn}} = \left\lceil \frac{2}{a} \log \frac{0.4\sqrt{2}}{10^{-6}\sqrt{a}} \right\rceil .$$

Proposition K.2 bounds the scaled W_2 initialization error by 10^{-6}. The retained physical exposure is $a\lceil 5000/a \rceil$; states are saved every $\max\{1, \mathrm{round}(0.5/a)\}$ updates. The seed is $27092700 +$

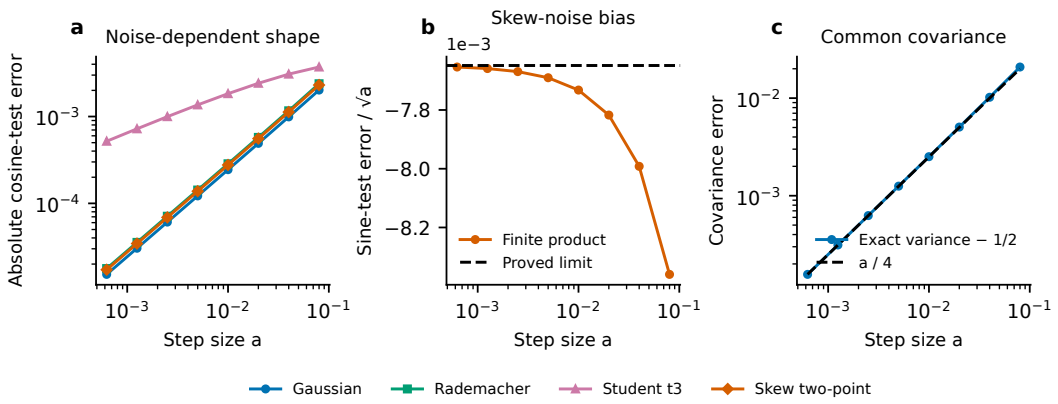

Figure 8: **Distributional shape and covariance at $t = 1$.** Normalized cosine discrepancies (a), skew sine bias divided by $\sqrt{a}$ (b), and covariance error (c). Curves are deterministic characteristic-function calculations with analytic truncation error below 10^{-12}. The nonzero skew limit distinguishes shape error from covariance discretization.

$100000i + 1000k + r$, for zero-based root, step and replicate indices. The full objective is

$$F(x) = \tfrac{1}{2}x_1^2 - x_1 - \tfrac{3}{2}\log(1 + x_1^2) + \arctan x_1 + \tfrac{1}{2}x_2^2.$$

Both roots satisfy $H_j S_j = 0.02 I_2$, so both loss limits are $0.01\chi_2^2$ with mean 0.02. At $a = 0.0025$, scaled mean losses are 0.020089 ± 0.000074 and 0.019990 ± 0.000072. For selected left-component weights $0.25, 0.5, 0.75$, weighted within-component variances are recorded separately from temporal occupation estimates.

Analytical shape error. For $X_{n+1} = (1 - a)X_n + a\varepsilon_{n+1}$, the scaled stationary characteristic function is

$$\phi_a(t) = \prod_{k \geq 0} \phi_\varepsilon\big(t\sqrt{a}(1 - a)^k\big).$$

Gaussian, Rademacher, standardized t_3 and the skew law of Proposition M.1 all have variance one. Standardized t_3 has density $2/[\pi(1 + x^2)^2]$ and characteristic function $(1 + |t|)e^{-|t|}$. Eight steps halve from 0.08 to 0.000625, and $t \in \{0.5, 1, 2\}$, giving 96 deterministic evaluations. Tests are divided by $1 + |t| + t^2 + |t|^3$ to belong to $\mathcal{H}_{2,1}$. Stopping before index K omits an independent centered sum of variance $(1 - a)^{2K}/(2 - a)$. The characteristic-function truncation error is at most $t^2(1 - a)^{2K}/[2(2 - a)]$, which is set below 10^{-12}. The implementation sums complex logarithms and exponentiates; the bound concerns analytic truncation. These individual test errors differ from the supremum over the test class. The common exact scaled covariance is $1/(2 - a)$.

State-dependent finite-sum learning. At the positive root $m = (1.9150080482, 1.2878394550)$ of (27),

$$H = \operatorname{diag}(0.8336279122, 0.6056869745), \qquad S_1 = \begin{pmatrix} 0.1518851672 & 0.0642388947 \\ 0.0642388947 & 0.0475438154 \end{pmatrix}.$$

The 192 output/loss trajectories use $B \in \{1, 4, 16\}$ and $a \in \{0.04, 0.02, 0.01, 0.005\}$. Each discards $\lceil 50/a \rceil$ updates and retains physical exposure 5000 with saving interval approximately 0.5. Seeds are $28092700 + 100000i + 1000k + r$ in the listed batch and step orders. Each update averages B independent complete component gradients with replacement. All nine roots enter the nearest-root label; every saved observation in this experiment remained in the positive-root cell. For $f = (x_1 + x_2, \tanh(x_1 - x_2/2))$,

$$Df(m)S_1 Df(m)^T = \begin{pmatrix} 0.3279067719 & 0.0433471556 \\ 0.0433471556 & 0.0072842122 \end{pmatrix}.$$

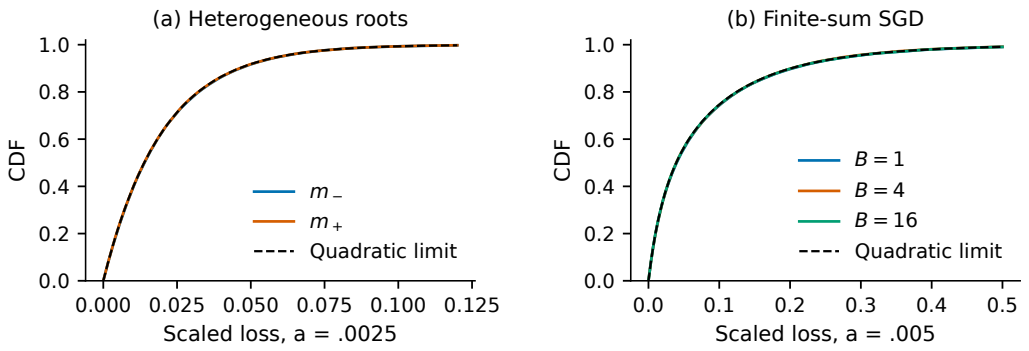

Figure 9: **Loss distributions beyond their means.** (a) Actual nonlinear losses divided by a in the heterogeneous rectangles at $a = 0.0025$, with the shared $0.01\chi^2_2$ limit. (b) Actual finite-sum losses multiplied by B/a at $a = 0.005$, with the generalized chi-square limit computed by deterministic quadrature. Bands are pointwise one standard error across 16 independent runs. The corresponding moment curves appear in Figures 1 and 3.

The predicted scaled mean loss is 0.0777061923; the quadratic-form eigenvalues are 0.0108051295 and 0.1446072551. Gaussian polar coordinates give, for $q \geq 0$,

$$\mathbb{P}\{\tfrac{1}{2}(\mu_1 G_1^2 + \mu_2 G_2^2) \leq q\} = 1 - \frac{2}{\pi}\int_0^{\pi/2} \exp\left\{-\frac{q}{\mu_1 \cos^2\theta + \mu_2 \sin^2\theta}\right\} d\theta.$$

We evaluate this expression by 256-node Gauss–Legendre quadrature.

Neural data and root selection. The UCI optical digits training and test files have disjoint writers. Pixel centering/scaling, the 16-dimensional PCA basis and whitening scales are fitted on the training subset only. The network trains a 4×16 matrix without bias parameters and fixes output weights $(0.6, -0.8, 1.1, -1.3)$. The baseline uses $\lambda = 0.02$, $\tau = 10^{-4}$, $\gamma = 0.015$ and $r \sim \mathsf{N}(0, I_{64}/64)$ with seed 27010001. Root search uses 20 Gaussian initializations of scale 0.2 (seed 27020001), BFGS and exact-Hessian Newton refinement. Candidates require gradient norm below 10^{-8} and minimum Hessian eigenvalue above 10^{-6}; the qualified candidate with smallest objective is selected. All qualified candidates found coincide within 10^{-4}. At that root, training accuracy is 99.74%, test accuracy 97.20% and test logistic loss 0.13946. These metrics describe the deterministic root.

The first six test examples of each class define 12 fixed probes. The logit Jacobians are evaluated at the root, and the one-sample covariance uses the full training sum with denominator N, plus $\gamma^2 I$. Its trace is 0.34814885, giving scaled mean loss 0.08703721. Gradient and Hessian formulas are analytic; a directional Hessian check at the root gives maximum discrepancy 7.1×10^{-12}. An independent eight-trajectory pilot fixed the evaluation range before the evaluation seeds were used; its settings and records accompany the code.

Neural sampling. The normalized steps are $0.000625, 0.00125, 0.0025, 0.005$ and batches are $1, 4, 16$. Physical burn-in is 600 and retained exposure is approximately 2000. States are saved every $\max\{1, \mathrm{round}(1/a)\}$ updates, giving 2001 saved states with exact spacing 0.9996112563352947 at every setting. Full objective loss is evaluated every tenth save, giving 201 values per run. Seeds are $29092700 + 100000i + 1000k + r$ in ascending batch/step order. All saved states lie within Euclidean distance 0.75 of the fixed root, documenting the local region explored during measurement. The initialization certificate of Proposition K.2 applies only to the invariant-rectangle experiment.

Covariance references. Full covariance solves $HS + SH = C_1$. Diagonal-noise and isotropic-noise references replace C_1 by $\mathrm{diag}(C_1)$ and $(\mathrm{tr}\, C_1/d)I$, keeping the same Hessian and output Jacobian. The discrete-linear reference solves $(I - aH)S_a(I - aH)^T + aC_1 = S_a$. Table 4 compares these references at each problem's smallest step. At the largest neural step with $B = 1$, full and discrete-linear logit MARE are 47.65% and 47.41%.

Table 4: Covariance-reference comparison. Output-variance MARE (percent), mean ± one standard error across 16 runs, at $a = 0.005$ for the two-dimensional finite sum and $a\lambda_{\max}(H) = 0.000625$ for the neural model. Each problem uses its fixed root Hessian and output Jacobian. The discrete-linear reference is evaluated for the neural model.

Problem	B	Full	Diagonal noise	Isotropic noise	Discrete linear
Finite sum	1	2.32 ± 0.37	52.40 ± 0.37	34.70 ± 0.30	–
Finite sum	4	1.74 ± 0.23	51.58 ± 0.36	34.03 ± 0.30	–
Finite sum	16	1.34 ± 0.26	52.36 ± 0.27	34.79 ± 0.23	–
Neural	1	6.97 ± 0.47	228.29 ± 4.15	184.92 ± 3.49	6.98 ± 0.47
Neural	4	6.31 ± 0.36	233.11 ± 3.50	189.61 ± 2.99	6.31 ± 0.37
Neural	16	6.09 ± 0.36	227.63 ± 2.49	185.48 ± 2.16	6.09 ± 0.36

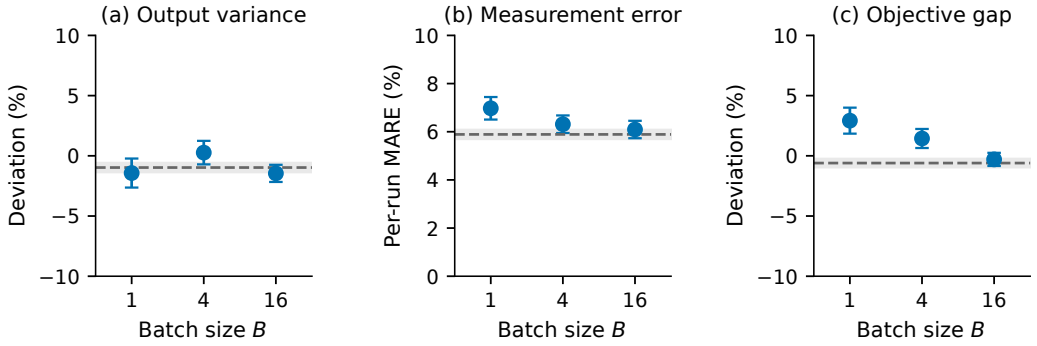

Figure 10: **Small-step neural predictions.** At $a\lambda_{\max}(H) = 0.000625$, (a) logit-variance and (c) actual objective-gap deviations use a common percentage scale. (b) Per-run MARE measures absolute relative error across the 12 probes. Bars summarize 16 paths; dashed lines and bands show the mean and one standard error of 64 matched stationary OU paths. The OU loss is quadratic. The main-text Figure 2 reports all four steps.

The supplement provides complete configurations, per-run CSV/JSON records, selected arrays and figure-generation code. Its README documents arrays omitted for the upload-size limit and the steps to regenerate them from the original seeds and budgets. The experiments use Python 3.12.14, NumPy 2.5.3, SciPy 1.18.1 and Numba 0.67.0 with one BLAS thread. Dataset files, attribution and download hashes are included.

R Moment measurements and neural diagnostics

R.1 Exact OU measurement references

For the scaled variable $Z = \sqrt{B/a}(X - m)$, the neural reference solves $HS + SH = C_1$. At save interval Δt, the exact stationary OU skeleton is

$$Z_{k+1} = DZ_k + \eta_k, \qquad D = e^{-H\Delta t}, \qquad \eta_k \sim \mathsf{N}(0, S - DSD^T), \qquad Z_0 \sim \mathsf{N}(0, S).$$

Each reference uses 64 independent paths, 2001 saves, $\Delta t = 0.9996112563352947$, and loss evaluation every tenth save. No initialization transient is present in this stationary reference. The matrices and save schedules coincide across all baseline step/batch settings after scaling, so the same reference is shared across those comparisons. The four augmentation arms each have their own reference covariance. Seeds are $30092700 + 10000i$ for arm indices $i = 0, 1, 2, 3$; Gaussian streams within each arm generate 64 independent paths. Four groups of 16 paths also provide replicates of the aggregated-variance statistic.

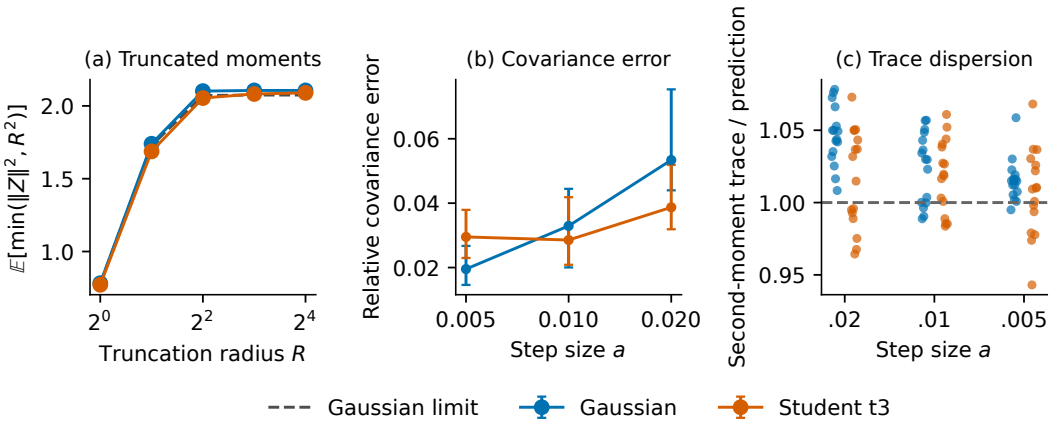

Figure 11: **Finite-second-moment evidence in the nonlinear nongradient system.** (a) All-root truncated moments at $a = 0.005$, with one standard error across 16 runs and a deterministic Gaussian reference. (b) Median relative covariance error and interquartile range. (c) All individual stable root-centered trace ratios; blue is Gaussian noise and orange is t_3. Raw quadratic measurements use dispersion summaries because t_3 has no fourth moment.

For a stationary scalar process with variance v and lag covariance γ_k, the centered sample variance satisfies

$$\mathbb{E}\widehat{v} = v - \frac{2}{n(n-1)} \sum_{k=1}^{n-1} (n-k)\gamma_k. \tag{77}$$

To see this, use $\sum_i \mathbb{E}(Y_i - \bar{Y})^2 = nv - n\,\mathrm{Var}(\bar{Y})$ and expand the variance of the sample mean. For Gaussian samples with covariance matrix K_n and $P = I - \mathbf{1}\mathbf{1}^T/n$, $\mathrm{Var}(\widehat{v}) = 2\,\mathrm{tr}(PK_nPK_n)/(n-1)^2$. Using the frozen neural matrices, the 12 linearized probes have relative centering biases of 0.544–0.793% and relative single-path variance-estimator standard deviations of 7.59–8.79%. The simulated OU per-run MARE is $5.89 \pm 0.19\%$, and its full covariance relative Frobenius error is $23.54 \pm 0.19\%$. These quantify finite-trajectory resolution under the Gaussian reference, rather than the nonlinear chain's approximation error.

R.2 Nonlinear finite-second-moment experiment

The recursion (72) uses Gaussian or $t_3/\sqrt{3}$ coordinates transformed by $L = \begin{pmatrix} 1 & 0 \\ 0.4 & \sqrt{1.84} \end{pmatrix}$. For each noise and $a = 0.02, 0.01, 0.005$, 16 paths start at the positive stable root, discard physical time 50 and retain physical time 5000. The save interval is 0.5 and seeds are $31092700 + 100000i + 1000k + r$, with noise index i and descending step index k. All observations and their nine-root labels are saved. Stable conditional statistics pool the four stable labels, whose root covariances coincide. The stable root-centered second moment divides by the number of stable observations; cellwise weighted second moments instead divide by the full sample count. All-root truncated moments include unstable labels and use radii $R = 1, 2, 4, 8, 16$.

For a two-dimensional Gaussian with covariance eigenvalues λ_1, λ_2, set $v(\theta) = \lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta$. Its truncated moment is

$$\mathbb{E}\min(\|G\|^2, R^2) = \frac{4}{\pi} \int_0^{\pi/2} v(\theta)\{1 - e^{-R^2/(2v(\theta))}\}\, d\theta,$$

evaluated with 256-node Gauss–Legendre quadrature. Bounded truncations admit independent-run standard errors; raw quadratic statistics under t_3 are reported with median/IQR and individual-run points. These finite-horizon measurements assess moment and tail behavior without claiming an empirical proof of uniform integrability.

Table 5: Signed mean coefficients from 32 independent confirmation paths per step, with one standard error. Both outputs use the same paths.

a	$(\overline{f}_1 - f_1(m))/a$	$(\overline{f}_2 - f_2(m))/a$
0.04	-0.05988 ± 0.00455	-0.028306 ± 0.000644
0.02	-0.05769 ± 0.00409	-0.027889 ± 0.000539
0.01	-0.05470 ± 0.00429	-0.027249 ± 0.000498
Theory	-0.05560	-0.027255

R.3 ROOTWISE MEAN MEASUREMENTS

At the positive root of (27), Theorem 4.3 gives

$$b = (-0.0290244628, -0.0265740280)^T, \qquad k_f = (-0.0555984908, -0.0272547532)^T$$

for the two preselected outputs $f = (x_1 + x_2, \tanh(x_1 - x_2/2))$ and $B = 1$. An independent set of 48 paths uses steps $0.04, 0.02, 0.01$, 16 paths per step, retained physical exposure $T = 400/a$, burn-in 50, and seeds $32092700 + 1000k + r$. Its smallest-step estimates are -0.08186 ± 0.00926 and -0.03026 ± 0.00173. The confirmation uses 32 new independent paths per step and fixes $T = 1600/a$ before execution, with seeds $34092700 + 1000k + r$. Here $k = 0, 1, 2$ indexes descending steps and r starts at zero. Every path starts at the positive root; all nine labels are recorded, and all retained states in these runs belong to that root's cell. No stopping or retention rule depends on agreement with the coefficient.

Let $b_0 = \lceil 50/a \rceil$, $b_1 = \lceil 200/a \rceil$, $N = \lceil 1600/a^2 \rceil$ and $q = \mathrm{round}(0.5/a)$. The confirmation records the post-update state X_{i+1} at indices $i = b_0 + kq < b_1 + N$. The primary window uses $b_0 \le i < b_0 + N$; the paired later-burn window uses $b_1 \le i < b_1 + N$ on the same grid. Actual save intervals are $0.48, 0.5, 0.5$, respectively. The record of each path includes exact first and last save times and the root-conditioned count. Quarter, half and full primary windows form nested measurements, not independent replicates. All 96 paths were completed, totaling 673,120,000 updates.

For R independent paths of physical exposure T, the OU approximation gives

$$\mathrm{Var}\left(\frac{B}{a}\{\overline{f} - f(m)\}\right) \approx \frac{B}{RTa}\,\mathrm{diag}(JA^{-1}C_1A^{-T}J^T), \qquad J = Df(m).$$

The diagonal constants are 0.887625 and 0.016139; with $R = 32, aT = 1600$ the predicted standard errors are 0.00416 and 0.000561. Keeping aT fixed controls this leading sampling scale as a decreases. The later-minus-primary coefficient differences at $a = 0.01$ are -0.000130 and -0.0000191; their uncertainty is computed from paired differences, not independent-window error bars. The windows overlap by approximately 99.625%, 99.813% and 99.906% at the three steps, so this is a sensitivity comparison rather than independent evidence of equilibration.

An exact empirical decomposition. Write $\delta = X - m$, $M = \overline{\delta\delta^T}/a$, and $q_g = (g_1''(m_1), g_2''(m_2))^T$. For this separable drift define

$$r_g(X) = g(X) - A\delta - \tfrac{1}{2}q_g \odot \delta^{\odot 2}, \qquad r_f(X) = f(X) - f(m) - J\delta - \tfrac{1}{2}(D^2 f(m) : \delta\delta^T).$$

All bars in the following identity average the same saved states:

$$\begin{aligned}\widehat{k}_f - k_f = {} & JA^{-1}\overline{g}/a \\ & + \tfrac{1}{2}D^2 f(m) : (M - S) - \tfrac{1}{2}JA^{-1}\{q_g \odot \mathrm{diag}(M - S)\} \\ & + (\overline{r}_f - JA^{-1}\overline{r}_g)/a. \end{aligned} \tag{78}$$

This is an algebraic decomposition, without replacing the empirical gradient by zero. For f_1 at $a = 0.01$, the three terms averaged over the confirmation paths are 0.001277, -0.000305 and -0.0000724. For f_2, they are 0.000194, -0.000158 and -0.0000297. The maximum algebraic residual across all recorded windows is 2.8×10^{-15}. The 48-path measurement's larger f_1 error at this step is likewise predominantly empirical drift: its total error is -0.02626, with drift contribution -0.02600. For this measurement, states are recovered from the saved invertible observables: $u =$

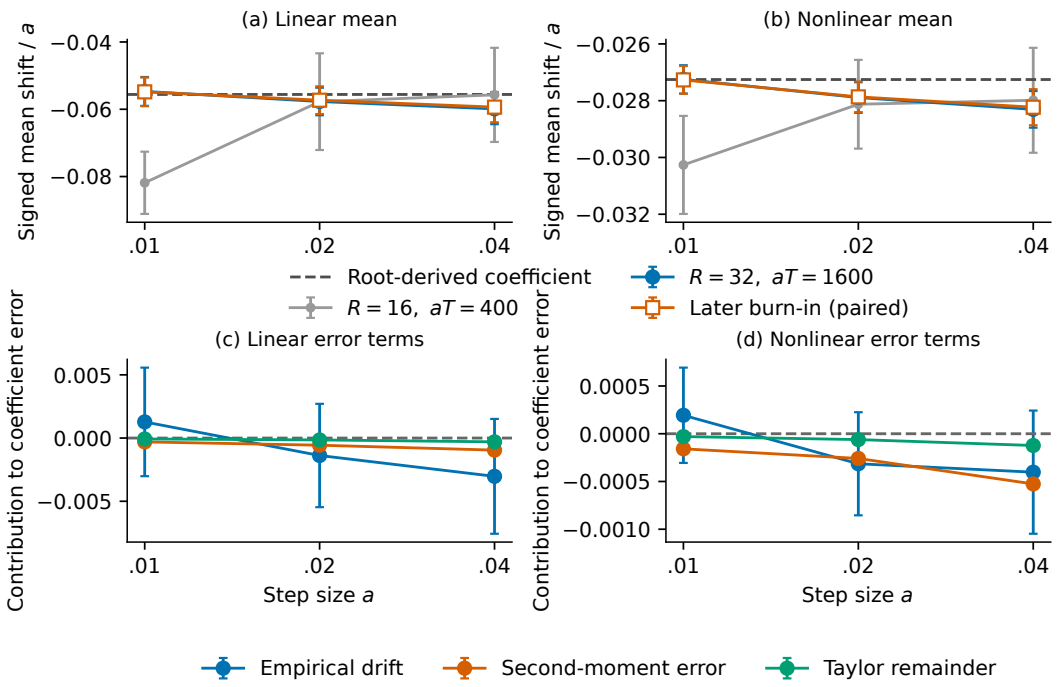

Figure 12: **Independent mean measurements and error sources.** (a,b) Signed coefficients for both preselected observables. Gray paths use $R = 16, aT = 400$; the independent confirmation uses $R = 32, aT = 1600$, with paired burn-in windows. Dashed lines are root-derived coefficients. (c,d) The three terms in (78) for the confirmation's primary window. Bars are one standard error across independent paths; outputs and windows within a path are dependent.

Table 6: Neural augmentation ablation. Mean logit-variance ratios (one standard error across 16 trajectories) use each arm's root-derived reference. The common steps are $a_s = 0.002603$ and $a_l = 0.020825$.

Augmentation	$B = 1, a_s$	$B = 1, a_l$	$B = 16, a_s$	$B = 16, a_l$
Tilt + dither	0.986 ± 0.012	1.473 ± 0.011	0.985 ± 0.007	1.009 ± 0.006
Tilt only	0.986 ± 0.012	1.507 ± 0.015	0.986 ± 0.007	1.008 ± 0.006
Dither only	0.986 ± 0.012	1.469 ± 0.011	0.985 ± 0.007	1.009 ± 0.006
Neither	0.986 ± 0.012	1.502 ± 0.014	0.986 ± 0.007	1.008 ± 0.006

$\operatorname{arctanh}(f_2)$, $x_2 = (f_1 - u)/1.5$ and $x_1 = f_1 - x_2$. These observations identify the source of measured error, without furnishing a numerical bound on the theorem's $o(a)$ remainder.

To distinguish boundary effects from innovation averaging, the confirmation also accumulates gradients and innovations at *every pre-update state*. For each complete update window, the recursion gives the exact identity

$$\frac{1}{N}\sum_{i=b}^{b+N-1} g(X_i) = \frac{1}{N}\sum_{i=b}^{b+N-1} \xi_i - \frac{X_{b+N} - X_b}{aN}.$$

The recorded balance residual is at most 7.4×10^{-17}. These full-update diagnostics use a different schedule from the thinned, post-update averages in (78) and are stored separately.

R.4 Tilt, dither and local neural geometry

The four arms are $(\tau, \gamma) = (10^{-4}, 0.015), (10^{-4}, 0), (0, 0.015), (0, 0)$. They share the data preprocessing, ridge, output weights and test probes. Removing dither leaves the objective and its root

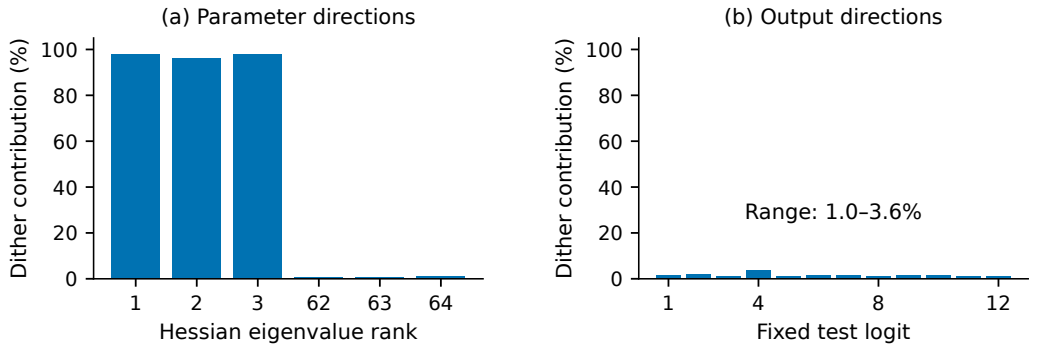

Figure 13: **Directional effect of dither.** Its fraction of theoretical variance in six Hessian eigenvectors (a) and 12 fixed logit directions (b). Both panels use the same 0–100 percent scale; the output range is annotated. These are deterministic contributions to the local Lyapunov covariance.

unchanged; only C_1 and S_1 are recomputed. For $\tau = 0$, deterministic continuation and Newton refinement give a root at distance 0.001988 from the baseline, with minimum Hessian eigenvalue 0.024356. The four deterministic roots/arms have the same test accuracy to the reported precision. Each arm defines a controlled comparison around its separately computed root.

Each arm uses common actual steps $a_s = 0.0026031543133731632$ and $a_l = 0.020825234506985306$, batches 1 and 16, and 16 paths per setting. The baseline endpoint paths are also used in Figure 10; the other arms supply 192 paths. Endpoint seed indices are paired across arms: $29092700 + 100000i + 1000k + r$, where $i = 0, 2$ and $k = 0, 3$. Burn-in, exposure and saving match the baseline. All retained states are included in the moment statistics.

Because $C_{\text{dither}} = \gamma^2 I$ and H is symmetric, its contribution is $S_{\text{dither}} = \gamma^2 H^{-1}/2$. Directional fractions are $u^T S_{\text{dither}} u/(u^T S_1 u)$; output fractions use rows of the frozen logit Jacobian for u. The resulting decomposition in Figure 13 distinguishes weak-curvature parameter directions from prediction directions.

R.5 Same-state finite-step diagnostics

The same 32 baseline $B = 1$ paths at the two step endpoints supply these diagnostics and the corresponding variance and loss statistics. Parameters are recorded on the 201-state loss schedule. On each of those states we compare $F(X) - F(m)$ with $\frac{1}{2}(X - m)^T H(X - m)$ and record output linearization residuals. On every hundredth saved state, we evaluate $g(X) - H(X - m)$ and the complete training-sum covariance $C(X)$. All diagnostics use common state schedules within a comparison.

The actual and quadratic mean-gap ratios are 1.029 ± 0.011 and 1.023 ± 0.010 at a_s, versus 1.893 ± 0.015 and 3.000 ± 0.041 at a_l. The scaled mean drift-residual norm rises from 0.0185 to 0.1471. The mean relative covariance variation rises from 0.261 to 2.605. These measurements distinguish the objective's Taylor remainder from changes in the fluctuation covariance; they do not identify a universal step threshold.

R.6 Observation-window sensitivity

To separate trajectory measurement error from persistent finite-step departures, 64 independent baseline paths use $B = 1, 16$, steps a_s, a_l, burn-in 600 and retained physical exposure 8000. Seeds are $33092700 + 100000i + 1000k + r$, with ascending batch and step indices. The actual save interval is 0.9996112563352947, giving 8004 states per path. Each path's first 2001 states provide a paired short-window estimate; the two window estimates are correlated. A further 64 stationary OU paths use the 8004-state schedule, seed stream 30132700 and the baseline matrices.

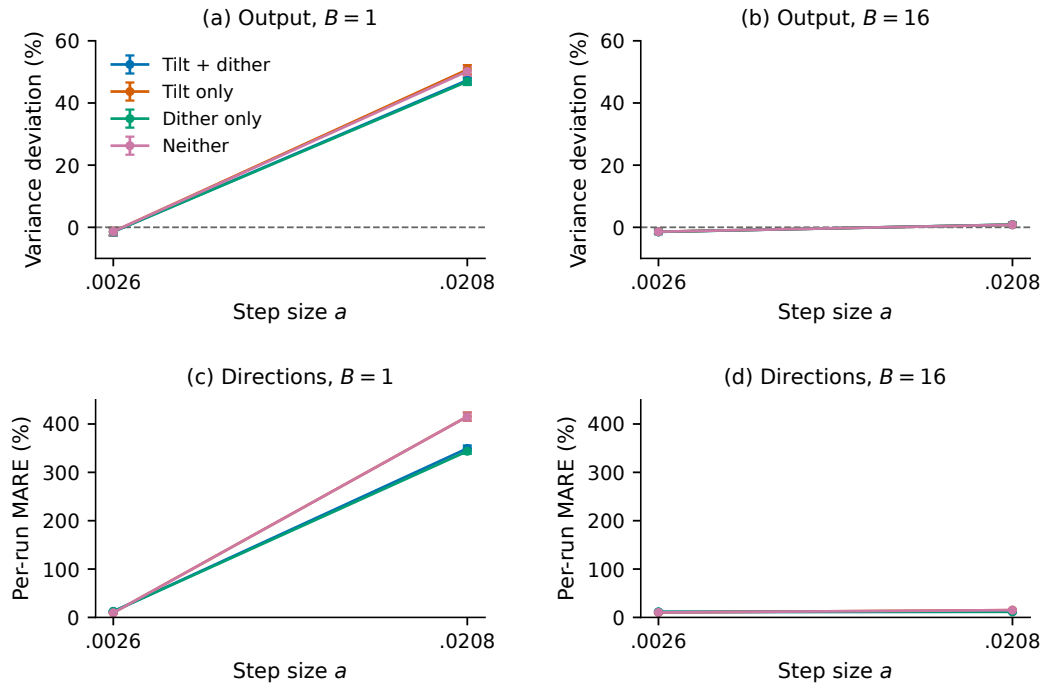

Figure 14: **Augmentation ablation across step and batch.** Logit-variance deviations from prediction (top) and MARE in six Hessian directions (bottom). Each row shares its vertical range across batches. Each arm uses its own root-derived reference; bars are one standard error across 16 paths. Comparable small-step output accuracy coexists with substantial directional errors at the largest step and $B = 1$.

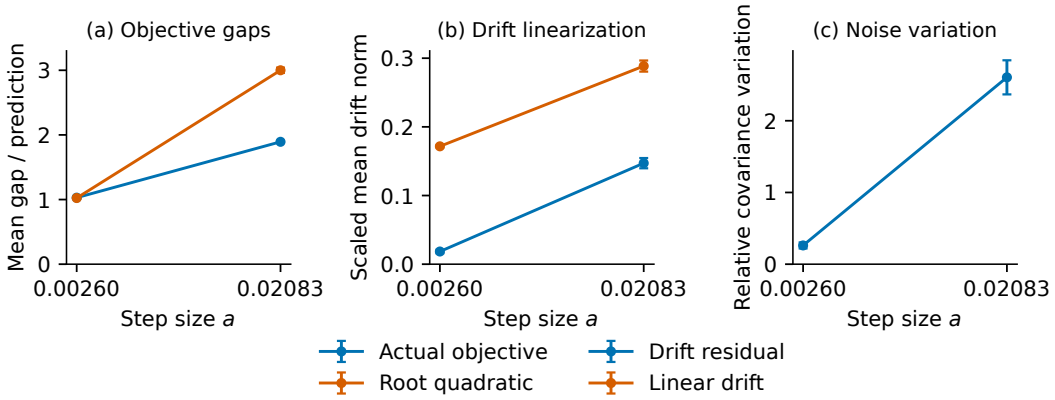

Figure 15: **Neural finite-step decomposition on paired states.** Actual and quadratic objective gaps (a), scaled drift residual and linear drift norm (b), and relative state-dependent covariance variation (c). Bars are one standard error across 16 paths at each step. The covariance norm is Frobenius; drift norms are Euclidean.

At a_s, the per-run logit MARE falls from $6.70 \pm 0.41\%$ to $3.20 \pm 0.17\%$ for $B = 1$, and from $6.81 \pm 0.29\%$ to $2.99 \pm 0.15\%$ for $B = 16$. The corresponding OU references are $5.89 \pm 0.19\%$ and $3.34 \pm 0.10\%$. At a_l, $B = 1$, the long-window variance ratio remains 1.478 ± 0.008 and the actual loss ratio is 1.916 ± 0.010. At a_s, $B = 1$, the long-window loss ratio is 1.028 ± 0.004, resolving a smaller finite-step objective bias even where output variance is close to its local reference. Thus

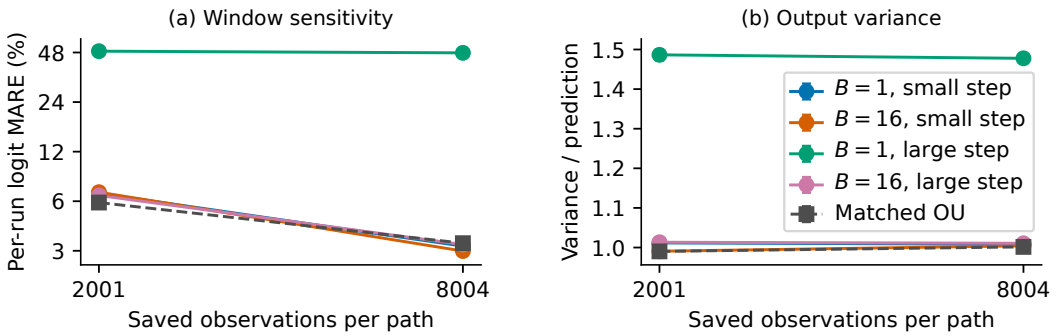

Figure 16: **Observation-window sensitivity.** Per-run logit MARE (a) and mean variance ratio (b) for paired short and full windows of 16 independent neural paths per setting. The stationary OU references use 64 paths at each window length. Bars are one standard error across paths; shared paths make the two neural estimates dependent.

increasing the observation window reduces the variance-estimation floor while preserving the large-step departures.

S Model dimension, data geometry and communication horizons

S.1 Controlled increases in neural dimension

We extend the fixed-output model of Proposition N.1 to 16 and 64 hidden units, training 256 and 1024 input weights, respectively. The output vector repeats $(0.6, -0.8, 1.1, -1.3)$ and is multiplied by $\sqrt{4/h}$ at width h. The optical-digits task retains the same writer-disjoint split and 16 training-fitted whitened principal components. Ridge, tilt and dither remain $\lambda = 0.02$, $\tau = 10^{-4}$ and $\gamma = 0.015$. Each model has its own reference computed from its full empirical gradient covariance. This comparison tests moment prediction in higher fixed dimension; both the architecture and its covariance change with width.

To assess a different data geometry, a 16-unit model uses the UCI Wisconsin Diagnostic Breast Cancer dataset (Wolberg et al., 1993). The 569 examples have 30 real-valued features; identifier columns are excluded. A fixed within-class permutation (seed 35092717) assigns $\lfloor 0.7 n_c \rfloor$ examples of each class to training, giving 397 training and 172 test examples. Feature centering, scaling, PCA and whitening are fitted on training data only, retaining 16 components. The labels are -1 for benign and $+1$ for malignant. This is a fixed classification benchmark for fluctuation measurements.

For each of the three cases, four Gaussian initializations of scale 0.2 undergo L-BFGS optimization and exact-Hessian Newton refinement. The root-search stream is $35092700 + 1000c$, where $c = 0, 1, 2$ indexes digits at widths 16 and 64, followed by WDBC at width 16; it also generates the fixed tilt direction $r \sim \mathsf{N}(0, I_p/p)$. Every candidate satisfies gradient norm below 10^{-8} and minimum Hessian eigenvalue above 10^{-6}; the smallest qualified objective selects the reference. All candidates and their objective values are retained. The three Hessian spectral intervals are $[0.02088, 0.24696]$, $[0.02018, 0.25429]$ and $[0.02044, 0.34261]$. The respective test accuracies of the deterministic roots are 97.20%, 97.20% and 95.93%. No fluctuation outcome or test accuracy enters root selection.

Four pilot paths per case and step use $B = 1$, retained physical exposure 1000 and seeds $35192700 + 100000c + 1000i + r$. The independent study uses 16 paths at every combination of $a\lambda_{\max}(H) \in \{0.000625, 0.005\}$ and $B \in \{1, 16\}$, with physical exposure 8000 and burn-in 600. Its seeds are $37092700 + 100000c + 10000i + 1000b + r$, with ascending step and batch indices. All 192 evaluation paths are retained separately from the 24 pilots. The first six test examples of each class define the 12 probes before any trajectories are sampled.

With $b_0 = \lceil 600/a \rceil$, $N = \lceil 8000/a \rceil$ and $q = \max\{1, \mathrm{round}(1/a)\}$, the saved states are X_{b_0+1+kq} for $kq < N$. Full objective values and parameters are retained every tenth save; all probe series, six parameter projections and root distances are retained. No observations are excluded according to

Table 7: Independent dimension/data study. Each row uses 16 paths with retained physical exposure 8000. Values are mean ± one standard error. MARE is a percentage.

Model	$a\lambda_{\max}, B$	Variance ratio	MARE (%)	Loss ratio
Digits 256	$0.000625, 1$	1.012 ± 0.005	3.018 ± 0.225	1.030 ± 0.005
Digits 256	$0.000625, 16$	0.998 ± 0.003	3.245 ± 0.203	0.997 ± 0.003
Digits 256	$0.005, 1$	1.520 ± 0.006	52.034 ± 0.571	1.927 ± 0.010
Digits 256	$0.005, 16$	1.001 ± 0.004	3.004 ± 0.207	1.012 ± 0.003
Digits 1024	$0.000625, 1$	1.024 ± 0.003	3.689 ± 0.273	1.027 ± 0.003
Digits 1024	$0.000625, 16$	0.996 ± 0.005	3.288 ± 0.185	0.999 ± 0.003
Digits 1024	$0.005, 1$	1.525 ± 0.008	52.514 ± 0.764	1.675 ± 0.007
Digits 1024	$0.005, 16$	1.006 ± 0.004	2.622 ± 0.140	1.011 ± 0.002
WDBC 256	$0.000625, 1$	1.003 ± 0.003	2.742 ± 0.174	1.012 ± 0.005
WDBC 256	$0.000625, 16$	0.997 ± 0.003	2.769 ± 0.144	1.003 ± 0.004
WDBC 256	$0.005, 1$	1.108 ± 0.004	10.979 ± 0.396	1.170 ± 0.006
WDBC 256	$0.005, 16$	1.005 ± 0.004	3.099 ± 0.164	1.009 ± 0.005

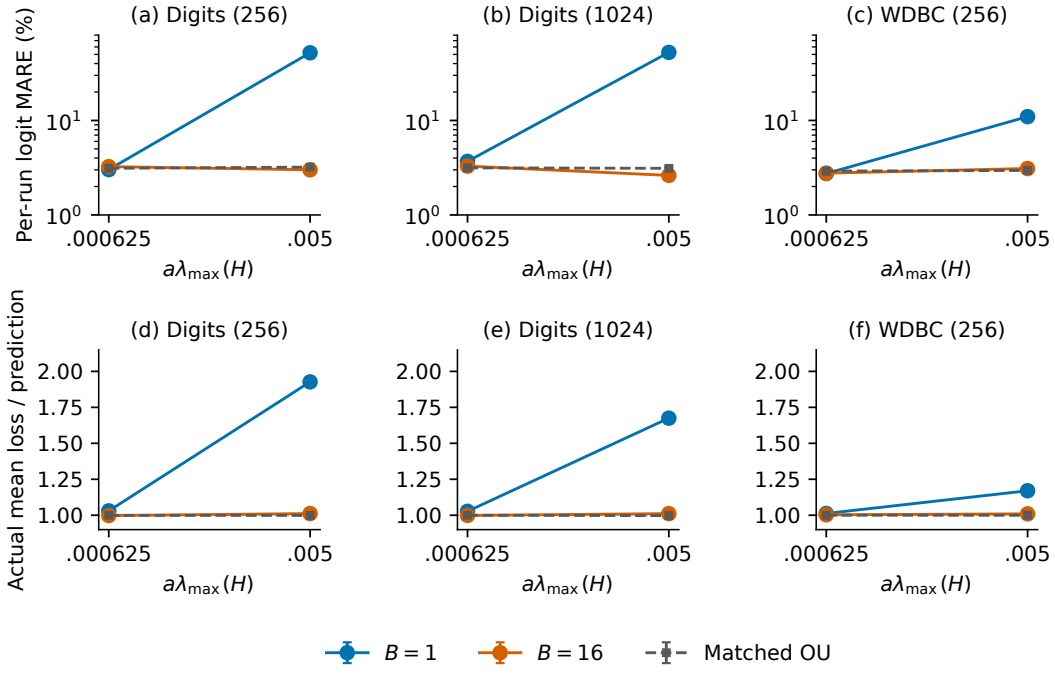

Figure 17: **Moment predictions across dimension and data.** Per-run logit MARE (top) and actual objective-gap ratio (bottom) for 16 independent paths per setting. Exact stationary OU references use 64 paths per case and step, matching the save and loss schedules and sharing the scaled reference across batches. All bars are one standard error across paths. The OU loss is quadratic.

their displacement. These local finite-horizon measurements do not enumerate the neural root set or estimate its invariant component weights.

Matched OU measurements. Every case and step has 64 independent stationary OU paths with the exact same number and spacing of saves as its nonlinear paths. The reference is shared across batches after $\sqrt{B/a}$ scaling. If $H = V\Lambda V^T$ and $\widetilde{S} = V^T S V$, the eigenbasis transition has diagonal mean map $D = e^{-\Lambda\Delta t}$ and innovation covariance $\widetilde{S} - D\widetilde{S}D$. We initialize with $\mathsf{N}(0, \widetilde{S})$, evaluate the linearized probes at every save, and evaluate the quadratic objective on the same loss schedule. The reference seed streams are $38092700 + 100000c + 1000i$. This yields 384 OU paths; within each path all probes share the same multivariate process. Table 7 and Figure 17 report the resulting measurements.

At the smaller normalized step, per-run logit MARE ranges from 2.74% to 3.69% across the six model/batch settings, compared with matched OU means of 2.91–3.15%. The corresponding aggregate MARE ranges from 0.64% to 2.35%. Actual objective-gap ratios range from 0.997 to 1.030. These statistics support small-step moment prediction across the tested dimensions while resolving modest objective bias in some settings.

At the larger step with $B = 1$, the two digits models have variance ratios 1.520 ± 0.006 and 1.525 ± 0.008, and actual loss ratios 1.927 ± 0.010 and 1.675 ± 0.007. WDBC has smaller departures, with variance ratio 1.108 ± 0.004 and loss ratio 1.170 ± 0.006. At $B = 16$, all three larger-step loss ratios lie between 1.009 and 1.012, and per-run MARE between 2.62% and 3.10%. Thus increasing batch size improves the approximation at this fixed step; dimension alone does not explain the observed departures.

Dither accounts for 13.86%, 38.39% and 7.00% of the full noise trace in the three cases, respectively, but only 0.88–2.82%, 0.84–2.31% and 0.49–1.22% of the predicted probe variances. The growing parameter dimension therefore changes the role of dither in parameter-space measurements even when its output-space contribution stays small. These are controlled fixed-dimensional networks with trained input weights and fixed output weights; they do not test the theorem for unrestricted deep architectures or nonisolated stationary sets.

S.2 Independent long-horizon communication measurements

The scalar model $g(x) = x + 5\cos x$ separates local fluctuation accuracy from basin exploration. A fixed-budget pilot extends the Gaussian and standardized t_3 comparators to 20 million retained updates at each of $a = 0.08, 0.04$, with four paths per noise and step, half initialized at each minimum. All use 100,000 burn-in updates and the same radius-0.25 core definition as Appendix P. The seed is $36092700 + 100000k + 1000i + r$, with $k = 0$ for Gaussian and $k = 2$ for t_3. The Gaussian paths have no completed transitions; the t_3 paths have 450 at 0.08 and 46 at 0.04. The Gaussian result describes initialization dependence over this horizon, rather than noncommunication of the chain.

An independent t_3 study uses $a = 0.04, 0.02$ and 12 paths per step, half from each minimum. The one-jump proxy fixes retained budgets of 51,400,000 and 410,900,000 updates per path before execution, targeting 400 transitions per step in expectation under that proxy. Each path nevertheless runs to its fixed horizon regardless of the realized count. The seed is $39292700 + 1000i + r$, and the burn-in is 100,000 updates. The total executed budget is $5{,}550{,}000{,}000$ updates. The nearly eightfold budget increase for a halved step follows the a^{-3} iteration scale of the scalar tail model.

We retain every completed transition time and measurements at one tenth, one quarter, one half and the full post-burn horizon. These windows are nested within paths. Full-window left occupations are 0.83646 ± 0.01579 and 0.83964 ± 0.01012, compared with the limiting proxy 0.84621. The completed-transition totals are 388 and 413, respectively. Completion means reaching the opposite radius-0.25 root neighborhood; raw separator crossings total 486 and 587 and include returns before completion. The left-to-right rate divided by its root-based one-jump probability is 0.970 and 1.028; the corresponding right-to-left ratios are 0.921 and 1.002. Both occupation and directional-rate measurements are compatible with the proxy at the resolution of this fixed-budget study. Rate intervals resample the 12 independent paths 10,000 times and take percentile endpoints at 2.5% and 97.5%; complete paths, including their counts and exposures, are resampled together. The bootstrap seeds are $40192700 + \mathrm{int}(100a)$.

At the full horizon, all four path-bootstrap rate intervals include their corresponding one-jump probability. At $a = 0.02$, the left-to-right and right-to-left ratios have 95% intervals approximately $[0.901, 1.165]$ and $[0.893, 1.117]$. The smaller-step budget yields a comparable event count, rather than the eightfold loss of information that an unchanged update budget would incur.

To examine the one-jump mechanism directly, each update also records whether its innovation exceeds the root-to-separatrix threshold in either direction. At $a = 0.02$, the full streams contain 233 positive and 1322 negative threshold exceedances; the exact noise law predicts 236.4 and 1300.5 in expectation. Restricting these offers to times when the last visited core is the departure well gives 203 left-to-right and 209 right-to-left offers, compared with 204 and 209 completed transitions. An offer is defined at the deterministic root threshold and need not coincide with a completed transition

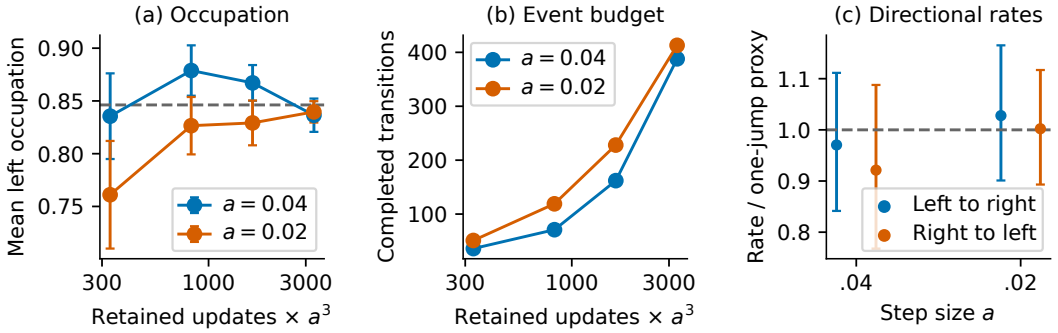

Figure 18: **Long-horizon communication at a smaller step.** (a) Mean occupation and one standard error over 12 paths at nested horizons; the dashed line is the heavy-tail limit. (b) Total completed transitions against retained updates scaled by a^3. (c) Directional rates relative to the finite-step one-jump proxy, with 95% path-bootstrap intervals. Nested windows are dependent.

from the actual state. Their close counts support the one-jump approximation at this step, without establishing a transition-law limit or a mixing-time bound.

T Root continuity without neighborhood uniform integrability

The stationary tail control in Lemma A.3 does not require innovation-square uniform integrability on a fixed root neighborhood. The following example separates these two properties.

Proposition T.1. *There is a scalar recursion satisfying Assumption 2.1 with uniformly bounded innovation variance, but, for every $r > 0$,*

$$\lim_{R\to\infty} \sup_{|x|\le r} \mathbb{E}[\xi(x,\zeta)^2 \mathbf{1}\{|\xi(x,\zeta)| > R\}] > 0.$$

Proof. Set $g(x) = x$ and $U(x) = x^2/2$, with the single hyperbolic root 0. Let η_0, η_1 be independent Rademacher variables and let V be independent and uniform on $[0,1]$. For integers $k, n \ge 1$, define

$$x_k = 2^{-k}, \quad x_{kn} = x_k + 2^{-k-n-2}, \quad r_{kn} = 2^{-k-n-5}, \quad v_k = 2^{-k}, \quad M_{kn} = 2^{k+n}.$$

The intervals $I_{kn} = (x_{kn} - r_{kn}, x_{kn} + r_{kn})$ are pairwise disjoint. For adjacent intervals at fixed k, the center separation is 2^{-k-n-3} and the sum of radii is $3 \cdot 2^{-k-n-6}$. Intervals farther apart in that cluster are therefore also disjoint. The whole kth cluster lies between x_k and $(73/64)x_k$, so distinct clusters are separated as well. Define

$$\psi_{kn}(x) = \left(1 - \frac{|x - x_{kn}|}{r_{kn}}\right)_+, \qquad p_{kn}(x) = v_k \psi_{kn}(x)/M_{kn}^2.$$

For $x \in I_{kn}$ let

$$\xi(x,\zeta) = \eta_0 + M_{kn}\eta_1 \mathbf{1}\{V \le p_{kn}(x)\},$$

and set $\xi(x,\zeta) = \eta_0$ elsewhere. This mechanism is measurable, centered and independent of a. Its variance equals $1 + v_k\psi_{kn}(x)$ on an active interval and equals 1 elsewhere. Thus $1 \le C(x) \le 3/2$, and $C(x) \to 1 = C(0)$ as $x \to 0$.

The innovation laws are weakly continuous. Inside each interval and at its endpoints, this follows from continuity of its mixing probability. At a cluster point x_k, the active probabilities tend to zero as $n \to \infty$. At 0, they tend to zero as $k \to \infty$; these are the only other accumulation points. For a bounded continuous test, its expectation differs from the Rademacher expectation by at most twice its supremum norm times the active probability. This verifies continuity also at the accumulation points, even though the rare jump sizes diverge. The affine next-state map consequently gives a Feller kernel for every sufficiently small $0 < a < a_0 < 1$. For the C_0 convention, vanishing at infinity follows because outside the bounded interval union the update is $(1-a)x + a\eta_0$. All remaining conditions follow from $g(x) = x$ and $U(x) = x^2/2$.

Fix $r > 0$ and choose k so the whole kth cluster lies in $(0, r)$. For any $R > 1$, choose n with $M_{kn} - 1 > R$. At the interval center, only the rare-jump event contributes above this threshold, and

$$\mathbb{E}[\xi(x_{kn}, \zeta)^2 \mathbf{1}\{|\xi(x_{kn}, \zeta)| > R\}] = \frac{v_k}{M_{kn}^2}(M_{kn}^2 + 1) \geq v_k > 0. \qquad \square$$

U Numerical invariant-law transport with alternating weights

Model and purpose. Use the invariant rectangles of Proposition K.1, with $g(x, y) = ((x+1)x(x-2)/(1+x^2), y)$ and independent coordinate noises taking values ± 0.2 with equal probabilities. The roots are $m_- = (-1, 0)$ and $m_+ = (2, 0)$, and the covariance targets are $S_- = \mathrm{diag}(1/75, 1/50)$ and $S_+ = \mathrm{diag}(1/60, 1/50)$. The rectangle laws are unique by contraction. Since the coordinate updates and noises are independent, the product of their marginal invariant laws is invariant, and uniqueness identifies it with the rectangle law. Let $\mu_{a,j}$ be this law centered at m_j and divided by $\sqrt{a}$. We choose actual invariant mixtures with weights

$$w_a = \tfrac{1}{2} + \tfrac{1}{4}\cos(\pi \log_2(0.04/a)), \qquad \pi_a = w_a \pi_{a,-} + (1 - w_a)\pi_{a,+}.$$

Along the tested dyadic steps, the weights alternate between $3/4$ and $1/4$. This selection extends to a nonconvergent sequence of weights as $a \downarrow 0$. The experiment directly examines the same-label assertion of Theorem 3.3.

Transport calculation. For product laws $\mu = \mu_1 \otimes \mu_2$ and $\nu = \nu_1 \otimes \nu_2$ with finite second moments,

$$W_2^2(\mu, \nu) = W_2^2(\mu_1, \nu_1) + W_2^2(\mu_2, \nu_2).$$

Indeed, each marginal of any coupling is an admissible one-dimensional coupling, which proves the lower bound after summing coordinate costs. The product of the optimal marginal couplings attains the bound. Consequently, one-dimensional quantile integrals give the full two-dimensional component costs, and

$$D_a^2 = w_a W_2^2(\mu_{a,-}, \mathsf{N}(0, S_-)) + (1 - w_a) W_2^2(\mu_{a,+}, \mathsf{N}(0, S_+)). \tag{79}$$

Choosing a common label and rescaling its optimal coupling proves $W_2(\pi_a, w_a \mathsf{N}(m_-, aS_-) + (1 - w_a)\mathsf{N}(m_+, aS_+)) \leq \sqrt{a} D_a$. D_a is the optimal scaled cost among label-preserving couplings; unrestricted couplings may have smaller cost. As a structural comparison, replace the first-coordinate variance in both targets by 0.015, leaving their second-coordinate variances unchanged. This is a fixed equal-weight pooled reference: its covariance stays fixed as the selected mixture weights alternate. Proposition U.1 below gives the corresponding comparison with an optimally chosen common covariance.

Discretization and quadrature. For each marginal, discretize its invariant interval by a grid with spacing h. Apply each of the two deterministic update maps to each grid point and split its mass linearly between adjacent destination points. This gives a positive, mass-preserving finite-state transition operator $P_{a,h}$. Starting at the root, iterate until the one-step probability-vector residual is below $a \times 10^{-11}$ in ℓ^1. For the resulting sorted scaled grid values z_i with masses p_i, define $c_i = \sum_{k \leq i} p_k$, $c_0 = 0$, and $q_i = \Phi^{-1}(c_i)$. The one-dimensional quantile representation of W_2 (Peyré & Cuturi, 2019) gives, for target variance s^2,

$$W_2^2\left(\sum_i p_i \delta_{z_i}, \mathsf{N}(0, s^2)\right) = \sum_i p_i z_i^2 + s^2 - 2s \sum_i z_i \{\phi(q_{i-1}) - \phi(q_i)\},$$

where Φ and ϕ are the standard Gaussian distribution and density and $\phi(\pm\infty) = 0$. This follows by expanding the squared quantile difference and using $\int_{c_{i-1}}^{c_i} \Phi^{-1}(u)\, du = \phi(q_{i-1}) - \phi(q_i)$. Thus the remaining numerical approximation is the finite-grid invariant law.

Resolution and results. All five steps $a \in \{0.04, 0.02, 0.01, 0.005, 0.0025\}$ use grids with 16 385, 32 769, 65 537, 131 073 and 262 145 points per coordinate interval. Figure 1c uses the finest grid; the full sequence, probability vectors, residuals and iteration counts are supplied. The two finest resolutions change the reported root-specific costs by at most 1.05×10^{-6} in absolute value, or

0.38% relative to the finest value; both maxima occur at $a = 0.0025$. At the endpoint steps, the costs are 0.00310464 and 0.00027922, a reduction by a factor of 11.1. The pooled endpoint costs are 0.00659665 and 0.00685505. These deterministic calculations support decreasing same-label scaled transport and the importance of root-specific covariances. Grid refinement measures numerical sensitivity; the theorem supplies the asymptotic conclusion independently of the discretization.

Proposition U.1 (Optimal shared-covariance comparison)**.** *Set* $\sigma_- = \sqrt{1/75}$ *and* $\sigma_+ = \sqrt{1/60}$. *For* $w \in (0,1)$, *define*

$$J_w(\Sigma)^2 = wW_2^2(\mathsf{N}(0,S_-),\mathsf{N}(0,\Sigma)) + (1-w)W_2^2(\mathsf{N}(0,S_+),\mathsf{N}(0,\Sigma)), \qquad \Sigma \succeq 0.$$

The minimum is

$$\inf_{\Sigma\succeq 0} J_w(\Sigma) = \sqrt{w(1-w)}\,|\sigma_- - \sigma_+|, \tag{80}$$

attained at $\Sigma_w^* = \mathrm{diag}((w\sigma_- + (1-w)\sigma_+)^2, 1/50)$. *For the actual scaled invariant components, let*

$$\widetilde{J}_a(\Sigma)^2 = w_a W_2^2(\mu_{a,-},\mathsf{N}(0,\Sigma)) + (1-w_a)W_2^2(\mu_{a,+},\mathsf{N}(0,\Sigma)).$$

Along $a_k = 0.04\,2^{-k}$,

$$\inf_{\Sigma\succeq 0} \widetilde{J}_{a_k}(\Sigma) \longrightarrow c_* := \frac{\sqrt{3}}{4}|\sigma_- - \sigma_+| = 0.00590169943749\ldots, \qquad D_{a_k} \longrightarrow 0.$$

Proof. Projection onto the first coordinate contracts W_2. Writing $s = \sqrt{\Sigma_{11}}$, the one-dimensional Gaussian formula gives

$$\begin{aligned} J_w(\Sigma)^2 &\geq w(\sigma_- - s)^2 + (1-w)(\sigma_+ - s)^2 \\ &= \{s - w\sigma_- - (1-w)\sigma_+\}^2 + w(1-w)(\sigma_- - \sigma_+)^2. \end{aligned}$$

This lower bound covers every positive-semidefinite covariance, including nondiagonal matrices. At Σ_w^*, the second marginals coincide and the product of the optimal first- and second-coordinate couplings attains the bound.

For the actual components, the reverse triangle inequality for each Wasserstein distance, followed by the weighted Euclidean triangle inequality, gives

$$|\widetilde{J}_a(\Sigma) - J_{w_a}(\Sigma)| \leq D_a \qquad (\Sigma \succeq 0).$$

Taking infima in both resulting inequalities yields

$$\left|\inf_{\Sigma\succeq 0}\widetilde{J}_a(\Sigma) - \inf_{\Sigma\succeq 0} J_{w_a}(\Sigma)\right| \leq D_a.$$

Theorem 3.3 gives $D_a \to 0$ for these invariant-law selections. On the dyadic sequence $w_a \in \{1/4, 3/4\}$, both Gaussian optima equal c_*, proving the limit. □

Thus root-specific covariances permit vanishing same-label scaled transport even when every shared covariance retains a positive limiting cost. The associated original-scale label-preserving cost is multiplied by $\sqrt{a}$. The comparison fixes component centers at the roots and concerns label-preserving couplings; unrestricted mixture couplings may exchange mass between labels.

V Relation to diffusion, heavy-tail and manifold theories

The random object and order of limits. The object here is an arbitrary invariant law of the discrete recursion at step a, followed by $a \downarrow 0$. The additive finite-variance analysis of Chen et al. (2022) addresses this order of limits around a single center. Its strongly convex, linear and contractive settings lead to stationary scaling limits and Gaussian identification under the stated uniqueness conditions. Our assumptions instead permit multiple hyperbolic roots and state-dependent innovations. Energy matching establishes the quadratic uniform integrability needed for moment accuracy across all components and invariant-law selections. With a uniform scaled third moment, a Gaussian weak limit already implies qualitative W_2 convergence; energy matching supplies the required tail control using second innovation moments instead. Classical stochastic approximation distinguishes

this constant-gain problem from diminishing-gain convergence (Kushner & Yin, 2003). Stochastic modified equations approximate the weak dynamics of SGD over specified time intervals and support algorithmic interpretations (Li et al., 2017). An approximation on a fixed interval does not by itself identify a stationary limit: passing to infinite time requires additional uniform control. Mattingly et al. (2010) develop Poisson-equation methods for time averages and invariant-measure errors of numerical SDE approximations under smoothness and ergodicity conditions. Our proof instead uses the discrete stationary energy balance to obtain quadratic uniform integrability under second innovation moments, with potentially several invariant probabilities. The Poisson equation enters later, for the separate smooth-test rate. The finite-step study of Li et al. (2021) further motivates checking the accuracy of diffusion descriptions at the actual learning rate. Our matched OU trajectories calibrate the variability of local moment measurements; the nonlinear trajectories independently measure their departures from the limiting prediction.

Heavy-tailed innovations and heavy-tailed iterates. Gürbüzbalaban et al. (2021) show how multiplicative SGD noise can produce heavy-tailed stationary iterates even with light-tailed data; Hodgkinson & Mahoney (2021) develop related mechanisms beyond linear models. Innovation moments and stationary-iterate moments are therefore distinct assumptions. We assume finite innovation variance with covariance growth controlled by the drift, together with dissipativity and descent. For sufficiently small steps these conditions imply finite stationary second moments, and energy matching controls the second moments after rootwise rescaling. They allow the standardized t_3 examples, whose third absolute moment is infinite. The symmetric smooth-test argument exploits this particular innovation structure; it does not assert a universal $\sqrt{a}$ Wasserstein rate or cover stationary infinite-variance regimes.

Local geometry and global communication. At isolated hyperbolic roots, the local covariance follows a Lyapunov equation and the mean shift follows a second-order drift balance. The new guarantee is that these local formulas remain moment-accurate across root-labeled components, with actual weights, singular covariances and second innovation moments. Single-center stationary expansions and harmonic loss identities provide the corresponding antecedents (Dieuleveut et al., 2020; Yaida, 2019; Allmeier & Gast, 2024). Equilibrium manifolds have additional tangential dynamics: Li et al. (2022) study how SGD moves along a zero-loss manifold under general noise covariance. Such motion is absent from a finite set of hyperbolic roots, so a manifold extension would require an additional limit along the equilibrium set. Separately, Freidlin–Wentzell theory concerns global rare transitions in small-noise diffusions (Freidlin & Wentzell, 2012), while discrete learning dynamics and heavy-tail communication require their own transition analysis (Azizian et al., 2024; Wang & Rhee, 2023). The invariant-rectangle example retains arbitrary weights, and the scalar communicating example uses a heavy-tail transition proxy. Both use the same principle: local moment coefficients and global occupation probabilities carry different information.