GAUSSIAN LIMITS FOR SGD WITHOUT STATIONARY MOMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal dependence can separate the Gaussian approximation of stochastic gradient descent from its stationary moments. For unmodified least-squares SGD, we construct a design with standard Gaussian marginals whose stationary error has every positive moment infinite. Independent observations with the same marginals instead give finite stationary variance. Both regimes retain a Gaussian small-step limit. Our general theory establishes pathwise contraction from a finite second design moment, then uses score cancellation and localization to obtain stationary Gaussian and Ornstein–Uhlenbeck limits. Independent Gaussian regression errors yield an exact conditional Gaussian law and total-variation convergence under the same design integrability. Stronger design conditions identify a positive first-order total-variation constant and a deterministic covariance correction with $o(a)$ error. A scalar coverage expansion translates this correction into its inference consequence. Experiments examine distributional error, coverage, and calibration with dependent scores. Together, these results establish precise probability-law approximation beyond moment-based stationary analysis.

1 INTRODUCTION

Constant-step stochastic gradient descent (SGD) keeps adapting to a stream of observations. Its stationary last iterate describes the fluctuations that remain after initialization is forgotten. Gaussian and diffusion approximations connect these fluctuations to the learning rate and support uncertainty quantification (Kushner & Huang, 1981; Mandt et al., 2017; Li et al., 2019; Dieuleveut et al., 2020; Chen et al., 2022). We ask how far this description extends when the stationary iterate has no positive moments.

We study unmodified least-squares SGD,

$$Y_i = X_i^\top \beta^* + Z_i, \qquad \beta_i = \beta_{i-1} + aX_i(Y_i - X_i^\top \beta_{i-1}), \tag{1}$$

with a fixed step $a > 0$ during each run and stationary dependent observations. An extreme covariate can persist across updates and repeatedly amplify an earlier regression error.

Our clearest example holds the design marginal fixed at $N(0,1)$ and refreshes its value with probability $1-\rho$ at each step. Take iid centered errors of variance $\sigma^2 > 0$, independent of the design. Writing $e_0^{(a)}$ for the stationary error, for $0 < a < 2/3$ we obtain

$$\rho = 0: \quad \mathbb{E}(e_0^{(a)})^2 = \frac{a\sigma^2}{2-3a} < \infty, \qquad 0 < \rho < 1: \quad \mathbb{E}|e_0^{(a)}|^r = \infty \quad (r > 0).$$

With independent Gaussian regression errors, both regimes share a Gaussian small-step limit and the same first-order correction. Temporal persistence changes moment stability while preserving a precise probability-law approximation.

Total variation (TV) controls absolute errors in event probabilities, whereas moments integrate unbounded functions. Rare design environments can dominate every stationary moment while contributing little to a bounded probability distance. With independent Gaussian errors, conditioning on the entire design gives an exact Gaussian law with a vanishing random covariance perturbation. This identity explains the coexistence formalized in Corollary 5.3.

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.

Gaussian limits and quantitative approximation. Constant-gain limits (Kushner & Huang, 1981) and diffusion approximations (Li et al., 2019) establish the Gaussian/Lyapunov framework. Stationary moment expansions (Dieuleveut et al., 2020) and distributional characterizations (Chen et al., 2022) connect it to constant-step SGD. For the finite-history last iterate, Khusainov et al. (2025) obtain convex-set Gaussian bounds under iid sampling, bounded design and a finite third score moment. Our Gaussian-error branch permits dependent unbounded designs and gives a sharp TV expansion beyond every positive-order stationary moment class. The error-law assumptions and approximation metrics differ.

Convergence to an invariant law at a fixed step (Merad & Gaïffas, 2023) and Gaussian approximation as $a \downarrow 0$ are distinct limits; our finite-history bound connects them. Wasserstein and tail bounds (Wang et al., 2026) address moment-controlled settings. Averaged-iterate CLTs (Yu et al., 2021; Mou et al., 2020) concern the covariance of accumulated iterates, while our correction concerns the stationary last iterate.

Product stability under temporal dependence. Least-squares SGD is the least-mean-squares (LMS) recursion. Its L^p stability theory (Guo et al., 1997), Markov-operator perturbations (Delyon & Juditsky, 1999), and tracking-error expansions (Aguech et al., 2000) provide routes to moment control. Markovian analyses cover bounded coefficients (Srikant & Ying, 2019; Huo et al., 2026) and unbounded coefficients under super-Lyapunov drift and growth conditions (Durmus et al., 2021a). Our positive-semidefinite (PSD) product bound establishes pathwise contraction from design integrability. Appendix M compares the temporal conditions, including the complementary martingale-design setting of Shen et al. (2026). Our causal framework uses functional dependence (Wu, 2005) and projective martingale approximation (Wu & Woodroofe, 2004; Wu, 2007).

Heavy tails and central probabilities. Affine-recursion theory establishes stationary existence and power-law tails (Brandt, 1986; Kesten, 1973; Goldie, 1991), including Markov-dependent coefficients (Roitershtein, 2007). Multiplicative SGD noise can generate heavy stationary tails from light-tailed inputs (Gurbuzbalaban et al., 2021; Hodgkinson & Mahoney, 2021); Damek & Mentemeier (2026) refine the iid matrix-recursion justification using irreducible-proximal theory. Even bounded iid coefficients can obstruct exponential concentration (Durmus et al., 2021b). Here all positive stationary moments fail at every small fixed step, while the same laws admit a sharp Gaussian approximation in TV. Table 1 summarizes the distinctions.

Route through the results. The design assumption gives pathwise contraction; a logarithmic score moment adds stationary existence (Theorem 4.1). An L^2 projective score condition yields Gaussian and Ornstein–Uhlenbeck (OU) limits (Theorem 4.2). For iid centered errors independent of the design, finite variance gives conditional weak convergence, strengthened to TV for Gaussian errors (Theorems 4.3 and 4.4). With Gaussian errors, a fourth design moment identifies the conditional first-order correction; quantitative Gram dependence permits its unconditional sharp expansion (Theorems 5.1 and 5.2). Section 7 tests the expansion and its coverage consequence, then studies long-run covariance calibration for dependent scores.

2 Model and Assumptions

Let $(\varepsilon_i)_{i\in\mathbb{Z}}$ be iid random elements, and let $(X_i, Z_i) = G(\ldots, \varepsilon_{i-1}, \varepsilon_i)$ be measurable. Set $\mathcal{F}_i = \sigma(\varepsilon_j : j \le i)$ and $\mathcal{P}_i U = \mathbb{E}(U \mid \mathcal{F}_i) - \mathbb{E}(U \mid \mathcal{F}_{i-1})$. The dimension p is fixed. For vectors and matrices, $\|\cdot\|$ denotes Euclidean and operator norm, respectively; $\|A\|_{\mathrm{F}} = (\operatorname{tr} A^{\mathsf{T}} A)^{1/2}$ is Frobenius norm. Write $\|U\|_q = (\mathbb{E}\,\|U\|^q)^{1/q}$ and $U \in L^q$ when this is finite. We use $d_{\mathrm{TV}}(\mu,\nu) = \sup_{A\in\mathcal{B}(\mathbb{R}^p)} |\mu(A)-\nu(A)|$, where $\mathcal{B}(\mathbb{R}^p)$ denotes the Borel sets; this equals $\frac{1}{2}\int |f-g|$ for densities f, g. The standard Gaussian density and CDF are ϕ and Φ, with quantile $z_u = \Phi^{-1}(u)$; ϕ_p is the $N_p(0, I)$ density. Write

$$e_i = \beta_i - \beta^*, \qquad H_i = X_i X_i^{\mathsf{T}}, \qquad H = \mathbb{E}H_0, \qquad \xi_i = X_i Z_i, \qquad e_i = (I - aH_i)e_{i-1} + a\xi_i. \tag{2}$$

Assumption 1 (Design). $\mathbb{E}\,\|X_0\|^2 < \infty$ and $H \succ 0$.

Assumption 2 (Score). $\mathbb{E}\xi_0 = 0$, $\xi_0 \in L^2$, and

$$\Theta_2 = \sum_{k=0}^{\infty} \|\mathcal{P}_0 \xi_k\|_2 < \infty. \tag{3}$$

Assumption 2 is a projective short-range dependence condition. For any causal process U_k, let $U_k^{\{0\}}$ replace ε_0 by an independent copy, and define

$$\delta_{U,q}(k) = \left\| U_k - U_k^{\{0\}} \right\|_q, \qquad k \geq 0. \tag{4}$$

This notation applies to $U = X, H, \xi$. The sufficient condition $\sum_{k\geq 0} \delta_{\xi,2}(k) < \infty$ implies (3). Square-integrable martingale-difference scores also satisfy (3), since $\mathcal{P}_0 \xi_k = 0$ for $k \geq 1$. In particular, independent centered errors with finite variance satisfy the score condition for every causal design obeying Assumption 1. The design itself may have arbitrarily slow ergodic averaging. Neither a bounded conditional design covariance nor a moment of the iterate is assumed.

Define backward products in their actual multiplication order:

$$P_0(a) = I, \qquad P_j(a) = (I - aH_0)(I - aH_{-1}) \cdots (I - aH_{-j+1}). \tag{5}$$

Write $(\beta_i^{(a)}, e_i^{(a)})$ for the jointly stationary solution at step a, when it exists, and π_a for its marginal iterate law. Unadorned iterates with a specified β_0 denote an initialized recursion. Stationarity and uniqueness concern the iterate coupled to the data process; the iterate alone need not be Markov.

3 Gaussian marginals with no stationary moments

Let $B_i \sim \text{Bernoulli}(\rho)$ and $G_i \sim N(0,1)$ be independent iid sequences. The stationary design

$$X_i = B_i X_{i-1} + (1 - B_i) G_i \tag{6}$$

equals the most recent refreshed Gaussian value. Its marginal law is $N(0,1)$ for every ρ, and its transition kernel satisfies $P^n(x,\cdot) = \rho^n \delta_x + (1-\rho^n)N(0,1)$. Hence its uniform total-variation mixing distance is exactly ρ^n.

This construction isolates whether temporal persistence alone can destroy stationary moments while preserving a Gaussian limit.

Theorem 3.1 (Same Gaussian marginals, different stationary moments). *For the design (6) with $0 < \rho < 1$ and iid nondegenerate centered errors of variance $\sigma^2 < \infty$, independent of the entire design, the stationary error exists throughout the sufficient range $0 < a < 2/3$ and*

$$\mathbb{E}|e_0^{(a)}|^r = \infty \quad (r > 0), \qquad e_0^{(a)}/\sqrt{a} \Rightarrow N(0, \sigma^2/2) \quad (a \downarrow 0). \tag{7}$$

For Gaussian errors, the convergence is in total variation, also conditional on almost every design path. In contrast, at $\rho = 0$,

$$\mathbb{E}(e_0^{(a)})^2 = a\sigma^2/(2-3a) < \infty \qquad (0 < a < 2/3). \tag{8}$$

For each finite $q \geq 1$, $\delta_{X,q}(k) \leq 2\|G_0\|_q \rho^{k/q}$ and $\delta_{H,q}(k) \leq 2\|G_0^2\|_q \rho^{k/q}$.

A single extreme design value can survive across many updates, repeatedly amplifying an earlier regression error. For $|X| \geq u$ with $au^2 > 2$, each retained value multiplies its coefficient by at least $au^2 - 1$. A run of length m has probability proportional to ρ^m. For any $r > 0$, choosing u so that $\rho(au^2-1)^r > 1$ makes the rth-moment amplification dominate this probability decay. Appendix G formalizes the argument by isolating one error coefficient.

With Gaussian errors, each observation (X_i, Z_i, Y_i) is jointly Gaussian, but the *time series* is not jointly Gaussian. With Rademacher errors, observations are marginally sub-Gaussian and the score has an exactly Gaussian marginal; every stationary moment still fails. Appendix G proves these assertions and a subpolynomial tail lower bound. A second construction uses the positive AR(1) design $X_i = \rho X_{i-1} + \eta_i$, $0 < \rho < 1$. Pareto-3 innovations give a finite second design moment, while $\eta_i = \exp(cE_i^{2/3})$, $E_i \sim \text{Exp}(1)$, gives moments of every order. Both yield failure of all stationary iterate moments. These examples, proved in Appendix E, show how the qualitative and quantitative results apply to ordinary linear-filter designs as well as refresh processes.

4 STATIONARY LIMITS WITHOUT ITERATE MOMENTS

The examples call for stability at the level of individual data paths. We establish contraction of the update products first, so stationary existence does not require an iterate L^r bound.

Theorem 4.1 (Pathwise stability and stationary existence)**.** *Under Assumption 1, there exist deterministic $a_0, c > 0$ and a random $K < \infty$ almost surely such that, on a common probability-one event,*

$$\|P_j(a)\| \le K e^{-caj} \quad \text{for all } j \ge 0,\ 0 < a \le a_0. \tag{9}$$

If $\mathbb{E}\log(1 + \|\xi_0\|) < \infty$, the recursion has a unique jointly stationary solution, which is causal and is represented by

$$e_0^{(a)} = a \sum_{j=0}^{\infty} P_j(a)\xi_{-j}. \tag{10}$$

The series converges absolutely almost surely. For a fixed $a \le a_0$, any finite initialization is forgotten almost surely when coupled to this solution on the same data.

The estimate (9) is derived from the design distribution. It gives pathwise stability even when $\mathbb{E}K^r = \infty$ for every $r > 0$, as the examples in Appendix E illustrate.

Once the products contract, score cancellation determines the central fluctuation. Under Assumption 2, the long-run covariance

$$\Omega = \sum_{k\in\mathbb{Z}} \mathbb{E}(\xi_0 \xi_k^{\mathsf{T}}) \tag{11}$$

exists with absolute entrywise convergence. Let V be the unique positive semidefinite solution of

$$HV + VH = \Omega, \qquad V = \int_0^{\infty} e^{-tH}\Omega e^{-tH}\, dt. \tag{12}$$

Theorem 4.2 (Stationary central limit theorem)**.** *Under Assumptions 1–2, for $B_a \sim \pi_a$,*

$$a^{-1/2}(B_a - \beta^*) \Rightarrow N_p(0, V), \qquad a \downarrow 0. \tag{13}$$

For every finite T, the stationary process $U_a(t) = a^{-1/2}(\beta^{(a)}_{\lfloor t/a \rfloor} - \beta^*)$ converges in $D([0,T], \mathbb{R}^p)$, the space of right-continuous paths with left limits equipped with the Skorokhod J_1 topology, to the stationary solution of

$$dU_t = -HU_t\, dt + \Omega^{1/2}\, dW_t. \tag{14}$$

Here W is standard p-dimensional Brownian motion. Both limits hold without any positive moment of B_a.

The centering in (13) is β^*, even if $\int \beta\, \pi_a(d\beta)$ is undefined. Likewise, V is the covariance of the limiting Gaussian distribution and need not be a limit of stationary covariance matrices.

4.1 AN EXACT IDENTITY FOR GAUSSIAN REGRESSION ERRORS

Independent Gaussian errors turn the stationary series into an exact Gaussian law conditional on the design. Thus distributional approximation reduces to controlling a random covariance perturbation. Put $\mathcal{G}_X = \sigma(X_i : i \in \mathbb{Z})$.

Theorem 4.3 (Conditional total-variation limit)**.** *Suppose Assumption 1 holds and (Z_i) is iid $N(0, \sigma^2)$, $\sigma > 0$, independent of $\mathcal{G}_X$. For all sufficiently small a, define*

$$\mathcal{R}_a = a^2 \sum_{j=0}^{\infty} P_j(a) H_{-j}^2 P_j(a)^{\mathsf{T}}. \tag{15}$$

This matrix is finite and positive semidefinite almost surely, and

$$\mathcal{L}\big(a^{-1/2} e_0^{(a)} \mid \mathcal{G}_X\big) = N_p\left(0, \frac{\sigma^2}{2}(I + \mathcal{R}_a)\right), \qquad \|\mathcal{R}_a\| \longrightarrow 0 \quad a.s. \tag{16}$$

Consequently, conditional and unconditional convergence to $N_p(0, \sigma^2 I/2)$ hold in total variation. Quantitatively, the conditional distance is bounded by

$$d_{\mathrm{TV}} \le \min\{1, \|\mathcal{R}_a\|_{\mathrm{F}} / (2\sqrt{2})\}. \tag{17}$$

The covariance identity follows by telescoping $P_j P_j^{\mathsf{T}} - P_{j+1} P_{j+1}^{\mathsf{T}}$ and requires only $\mathbb{E}\,\|H_0\| < \infty$. The perturbation vanishes along almost every design path. Bounded convergence then applies to (17) even when $\mathbb{E}\,\|\mathcal{R}_a\|^r = \infty$. For the population noise variance, Gaussian central spheres have stationary coverage at most their nominal level and converge to that level; Appendix D gives the argument.

4.2 Conditional limits for non-Gaussian errors

For independent errors with finite variance, the same conditional weights still yield a Gaussian weak limit. This extends the qualitative conclusion to continuous and discrete error laws.

Theorem 4.4 (Quenched CLT under finite error variance)**.** *Under Assumption 1 and arbitrary iid centered errors with $\sigma^2 = \mathbb{E}Z_0^2 < \infty$, independent of the entire design,*

$$\mathcal{L}(e_0^{(a)}/\sqrt{a} \mid \mathcal{G}_X) \Rightarrow N_p(0, \sigma^2 I/2) \quad \textit{for almost every design realization.}$$

The conclusion holds under the qualitative design assumption alone.

The proof combines the Gaussian identity's sum of squared weights with a vanishing maximal weight and the independent-array Lindeberg theorem.

5 A sharp first-order Gaussian approximation

The exact conditional law also identifies the leading approximation error. A fourth design moment determines its coefficient: when $\mathbb{E}\,\|H_0\|^2 < \infty$, define

$$J = \int_0^\infty e^{-tH}\mathbb{E}(H_0^2)e^{-tH}\,dt, \qquad HJ + JH = \mathbb{E}(H_0^2). \tag{18}$$

The matrix J corrects the covariance of the approximating Gaussian law. The stationary covariance itself may be infinite or undefined.

Theorem 5.1 (Conditional first-order expansion)**.** *Retain Assumption 1 and additionally assume $\mathbb{E}\,\|H_0\|^2 < \infty$. With J from (18),*

$$\mathcal{R}_a/a \longrightarrow J \quad \textit{almost surely.} \tag{19}$$

Under the Gaussian-error assumptions of Theorem 4.3, this implies conditional $O(a)$ total-variation error relative to $N_p(0, \sigma^2 I/2)$, and conditional $o(a)$ error relative to $N_p(0, \sigma^2(I + aJ)/2)$, almost surely.

Averaging the conditional expansion requires control of rare design environments at the a scale. For this unconditional rate, impose

$$H_0 \in L^q, \qquad \sum_{k\geq 0} \delta_{H,q}(k) < \infty \quad \text{for some } q > 2. \tag{20}$$

It follows, for example, from $X_0 \in L^{2q}$ and $\sum_k \delta_{X,2q}(k) < \infty$. The qualitative limits and conditional expansion retain their weaker assumptions.

Theorem 5.2 (Unconditional first-order Gaussian approximation)**.** *Assume Assumption 1, (20), and the Gaussian-error assumptions of Theorem 4.3. Let*

$$\mu_a = \mathcal{L}(e_0^{(a)}/\sqrt{a}), \quad \gamma_0 = N_p(0, \sigma^2 I/2), \quad \gamma_a = N_p(0, \sigma^2(I + aJ)/2),$$

where J is defined in (18). There is a deterministic finite C, depending only on the design law and p, such that, for all sufficiently small positive a,

$$d_{\mathrm{TV}}(\mu_a, \gamma_0) \leq Ca, \tag{21}$$

$$d_{\mathrm{TV}}(\mu_a, \gamma_a) = o(a), \tag{22}$$

$$\lim_{a\downarrow 0} \frac{d_{\mathrm{TV}}(\mu_a, \gamma_0)}{a} = \frac{1}{4}\mathbb{E}\left|G^\top JG - \operatorname{tr} J\right|, \qquad G \sim N_p(0, I). \tag{23}$$

The constant in (23) is positive. None of these conclusions assumes any stationary iterate moment. The deterministic correction removes the leading order-a error.

The refresh design now brings the moment obstruction and the sharp expansion together.

Corollary 5.3 (Sharp Gaussian approximation with no stationary moments)**.** *For the refresh design with fixed $0 < \rho < 1$ and independent $N(0, \sigma^2)$ errors, let $\mu_a = \mathcal{L}(e_0^{(a)}/\sqrt{a})$. Then*

$$\begin{aligned} \mathbb{E}|e_0^{(a)}|^r &= \infty && (r > 0,\ 0 < a < 2/3), \\ d_{\mathrm{TV}}(\mu_a, N(0, \sigma^2/2)) &= \tfrac{3}{2}\phi(1)a + o(a), \\ d_{\mathrm{TV}}(\mu_a, N(0, \sigma^2(1 + 3a/2)/2)) &= o(a) && (a \downarrow 0). \end{aligned}$$

Indeed, Gaussian marginal moments and the dependence bounds in Theorem 3.1 verify (20), and $J = 3/2$. The log-Weibull AR example also satisfies the quantitative assumptions; the Pareto-3 example illustrates the weaker qualitative theory. Each conclusion holds for a fixed dependence law.

The expansion also predicts the leading coverage error of a Gaussian interval.

Corollary 5.4 (First-order scalar coverage)**.** *Under Theorem 5.2 with $p = 1$, let $\alpha \in (0, 1)$ be fixed and put $z = z_{1-\alpha/2}$ and $h_a = \sigma\sqrt{a/2}\, z$. Then*

$$\mathbb{P}(|e_0^{(a)}| \le h_a) = 1 - \alpha - aJz\phi(z) + o(a), \tag{24}$$

$$\mathbb{P}(|e_0^{(a)}| \le h_a\sqrt{1 + aJ}) = 1 - \alpha + o(a). \tag{25}$$

The deterministic correction removes the leading undercoverage. These intervals use the population noise variance; Appendix I.1 proves the expansion directly from the TV bound.

5.1 Initialization and central confidence regions

To use the stationary theory after a finite run, initialization must be negligible on the approximation's scale. Weak convergence uses the $\sqrt{a}$ scale; preserving the first-order expansion requires a sharper bound.

Corollary 5.5 (Last iterate after initialization)**.** *Under Assumptions 1–2, initialize at a fixed finite β_0. If*

$$a^{-1/2}e^{-can(a)} \longrightarrow 0, \tag{26}$$

with c from Theorem 4.1, then $a^{-1/2}(\beta_{n(a)} - \beta^*) \Rightarrow N_p(0, V)$. One sufficient condition is $an(a)/\log(1/a) \to \infty$.

Corollary 5.6 (Finite-history first-order approximation)**.** *Under the assumptions of Theorem 5.2, initialize with deterministic error $b = \beta_0 - \beta^*$. There are $a_0, c, C, t_0 > 0$ and a nonnegative L_* with $\mathbb{E}L_* < \infty$ such that, for $0 < a \le a_0$ and $an \ge t_0$,*

$$d_{\mathrm{TV}}\big(\mathcal{L}(e_n/\sqrt{a}), \mu_a\big) \le C\big(\|b\|\ a^{-1/2}e^{-can} + e^{-2can}\big) + \mathbb{P}(L_* > 1/a). \tag{27}$$

If $e^{-can(a)} = o(a^{3/2})$, the initialized law $\mu_{a,n} = \mathcal{L}(e_{n(a)}/\sqrt{a})$ satisfies

$$d_{\mathrm{TV}}(\mu_{a,n}, \gamma_a) = o(a), \qquad \lim_{a\downarrow 0} \frac{d_{\mathrm{TV}}(\mu_{a,n}, \gamma_0)}{a} = \frac{1}{4}\mathbb{E}|G^{\mathsf{T}}JG - \operatorname{tr} J|,$$

with γ_a, γ_0, G as in Theorem 5.2. For $b = 0$, it suffices that $e^{-2can(a)} = o(a)$. The schedule $n(a) = \lceil \max\{24, 2\log^2(1/a)\}/a \rceil$ satisfies both conditions for each fixed admissible law.

The transfer compares finite and stationary conditional Gaussian laws at the same endpoint (Appendix I.2). Under its stated history conditions, Corollary 5.4 also holds with $e_{n(a)}$ in place of $e_0^{(a)}$.

Under independent centered errors with variance $\sigma^2 \in (0, \infty)$, the score is a martingale difference and $\Omega = \sigma^2 H$, so $V = \sigma^2 I/2$ even for non-Gaussian errors. Let $\widehat{\sigma}_m^2$ be the ordinary least-squares residual variance from $m \to \infty$ observations. Ergodicity and second moments imply $\widehat{\sigma}_m^2 \to \sigma^2$ in probability. Therefore, for a stationary iterate or an iterate satisfying (26),

$$\mathcal{C}_a = \left\{ b \in \mathbb{R}^p : \frac{2\left\|\beta_{n(a)} - b\right\|^2}{a\widehat{\sigma}_m^2} \le \chi^2_{p,1-\alpha} \right\} \quad \text{satisfies} \quad \mathbb{P}(\beta^* \in \mathcal{C}_a) \longrightarrow 1 - \alpha. \tag{28}$$

Here $\chi^2_{p,u}$ is the u-quantile of the chi-square law with p degrees of freedom. Calibration data may overlap the SGD data. Variance consistency gives asymptotic marginal coverage; a first-order expansion would additionally require control of calibration error at the a scale. Appendix F proves this result and extends it to fixed positive definite preconditioning and fixed nonoverlapping mini-batches.

Calibration before burn-in. At a fixed earlier endpoint, Gaussian conditioning gives a complementary exact calibration. Propagating the initialization and noise coefficients yields a Gaussian pivot; a full-rank residual-variance block with disjoint error indices gives an exact F pivot. Appendix K states the rank conditions and formulas, allowing dependent designs across blocks. The experiment reports both coverage and interval width to show the precision available before burn-in.

6 Proof mechanism: pathwise contraction and cancellation

Positive-semidefinite update products contract pathwise, while score partial sums satisfy an invariance principle. Over $O(1/a)$ updates, or $O(1)$ learning time, the products approach e^{-tH}. Localization then removes the remote stationary past. Appendices A–C supply the proofs.

A truncated block-product inequality. For positive semidefinite $B_1, \dots, B_m$, let $s = \sum_\ell \|B_\ell\|$ and $g = \lambda_{\min}(\sum_\ell B_\ell)$. Direct expansion gives

$$\log \|(I - aB_m) \cdots (I - aB_1)\| \le -ag + 2as \min(as, 1). \tag{29}$$

For $as \le 1$, the remainder about $I - a\sum B_\ell$ is at most $2a^2s^2$; for $as > 1$, bound the log norm by as. The right side has finite expectation after division by a under only $\mathbb{E}s < \infty$. Ergodicity and $H \succ 0$ provide a fixed block length with $\mu = \mathbb{E}g > 0$. Choose a_0 so that $\mathbb{E}[2s \min(a_0 s, 1)] < \mu/2$. Block log norms are then bounded by a times a stationary integrable sequence of negative mean, simultaneously for $a \le a_0$. The ergodic theorem proves (9).

Finite learning-time windows. Put $S_n = \sum_{j=0}^{n-1} \xi_{-j}$. Assumption 2 gives

$$\sqrt{a}\, S_{\lfloor t/a \rfloor} \Rightarrow \Omega^{1/2} W_t, \qquad \mathbb{E} \max_{j \le n} \|S_j\|^2 \le Cn. \tag{30}$$

The ergodic theorem gives uniform small-step Cesaro convergence of the centered Gram sums. Abel summation in the matrix Euler equation then yields $P_{\lfloor t/a \rfloor}(a) \to e^{-tH}$ uniformly for $t \le T$, directly for the original unbounded coefficients. The products also have bounded total variation in probability. Summation by parts with (30) yields

$$\sqrt{a} \sum_{j < \lceil T/a \rceil} P_j(a) \xi_{-j} \Rightarrow \int_0^T e^{-tH} \Omega^{1/2}\, dW_t. \tag{31}$$

This argument accommodates dependence between the products and scores.

The infinite stationary past. The essential interchange of limits is established separately. Set $N = \lceil T/a \rceil$ and $R_{a,T} = \sqrt{a} \sum_{j \ge N} P_j(a) \xi_{-j}$. Abel summation gives

$$R_{a,T} = -\sqrt{a} P_N(a) S_N + a^{3/2} \sum_{j \ge N} P_j(a) H_{-j} S_{j+1}. \tag{32}$$

The random variable $D = \sup_{n \ge 1} n^{-1} \sum_{j<n} \|H_{-j}\|$ is finite almost surely. On $\{K \le M, D \le M\}$, dyadic decomposition and (30) imply

$$\mathbb{P}(\|R_{a,T}\| > \epsilon) \le \mathbb{P}(K \vee D > M) + \frac{CM^2}{\epsilon} (1 + T)^{3/2} e^{-c'T}. \tag{33}$$

Only score maxima are integrated. Taking $T \to \infty$ and then $M \to \infty$ proves that the remote past is negligible uniformly as $a \downarrow 0$. Combining this with (31) gives Theorem 4.2. Proposition C.1 expresses this localization as an explicit bound for bounded test functions.

Averaging the first-order correction. For the sharp TV expansion, the product bound becomes $\|P_j(a)\| \le e^{aL_*}e^{-caj}$ with $\mathbb{E}L_* < \infty$. On $\{aL_* \le 1\}$, weighted Gram squares provide an integrable bound; the complementary event contributes $\mathbb{P}(L_* > 1/a) = o(a)$. This localization averages the conditional expansion without requiring $\mathbb{E}\mathcal{R}_a < \infty$ (Appendix I).

7 Experiments

We first test the sharp distributional expansion and its coverage consequence in the refresh design. A dependent-score experiment then examines long-run covariance calibration under the general Gaussian limit. Every configuration uses independent replicated histories; methods share paths within a configuration. Gaussian scalar experiments integrate regression errors conditional on the design. Actual SGD is used for non-Gaussian errors and finite-history inference. Appendix N provides all protocols and Monte Carlo uncertainty.

Direct distributional error and the first-order correction. The refresh experiment holds the $N(0,1)$ design marginal fixed and varies $\rho \in \{0, .7, .9\}$. Five steps range from .04 to .0025, with 24,000 histories per configuration and $n = \lceil \max\{24, 2\log^2(1/a)\}/a \rceil$ updates. Corollary 5.6 shows that this schedule preserves the stationary first-order expansion asymptotically. We estimate TV by integrating the absolute density difference between the empirical conditional Gaussian mixture and its reference Gaussian. Figure 1(a) compares both approximations on a common zero-based scale: the baseline approaches the predicted constant $3\phi(1)/2$, and correction reduces the normalized error. At $\rho = .7, a = .0025$, estimated TV changes from 8.87×10^{-4} to 1.85×10^{-5}. Panel (b) shows concentration of the finite-history correction $\mathcal{R}_{a,n}/a$ toward $J = 1.5$. Panel (c) connects this coefficient to coverage. At the same step, Corollary 5.4 predicts approximately 0.043 percentage points of baseline undercoverage; the observed deficit is 0.0421 points. Correction raises coverage from 94.9579% to 95.0009%, a 0.0430-point increase. The leading term thus predicts the observed scale of the coverage change. Table 2 reports absolute errors and uncertainty.

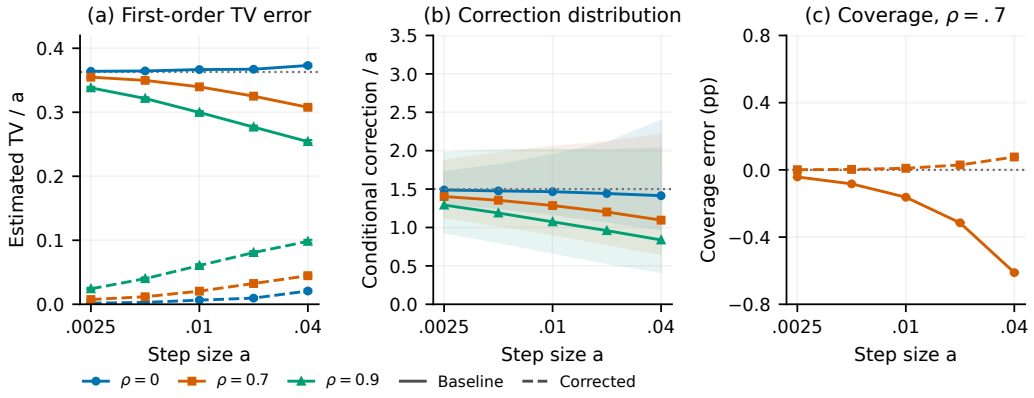


Figure 1: The first-order approximation in the Gaussian-refresh design. (a) Estimated TV divided by a for baseline (solid) and corrected (dashed) laws; the dotted line is $3\phi(1)/2$. Bars are 1.96 linearized Monte Carlo standard errors. (b) Median and 10th–90th percentiles of the finite-history correction $\mathcal{R}_{a,n}/a$; the dotted line is $J = 1.5$, and shading represents variation across histories. (c) Coverage minus 95%, in percentage points (pp), with 1.96 standard-error bars; Corollary 5.4 predicts baseline error $-100aJz_{.975}\phi(z_{.975}) + o(a)$ pp. Each configuration uses 24,000 histories.

Feasible inference with a dependent score. Theorem 4.2 also covers serially dependent scores, whose long-run covariance determines the Gaussian limit. We use a noncommuting two-dimensional Gaussian VAR(1) design. Gaussian AR(1) errors, independent of the design, have coefficient .7, giving $\Omega \neq H$. With 6,000 replications, we compare oracle Lyapunov calibration, lag-zero calibration, and residual-score Bartlett estimation using an earlier block on the same stream. This provides heteroskedasticity and autocorrelation consistent (HAC) calibration. At $a = .0025$, oracle coverage

is 95.23%, lag-zero coverage is 76.17%, and HAC coverage reaches 94.38% with 8,192 calibration observations and bandwidth 19. The lag-zero Gaussian limiting coverage is 76.01%. A population Bartlett benchmark isolates the effect of bandwidth: at bandwidth 19 its limiting coverage is 94.37%, close to the observed HAC result (Monte Carlo standard error 0.30 pp). Figure 2(b) compares this benchmark with feasible estimates. The agreement supports finite-bandwidth bias as an explanation of the remaining gap; larger bandwidths capture more temporal covariance in this experiment.

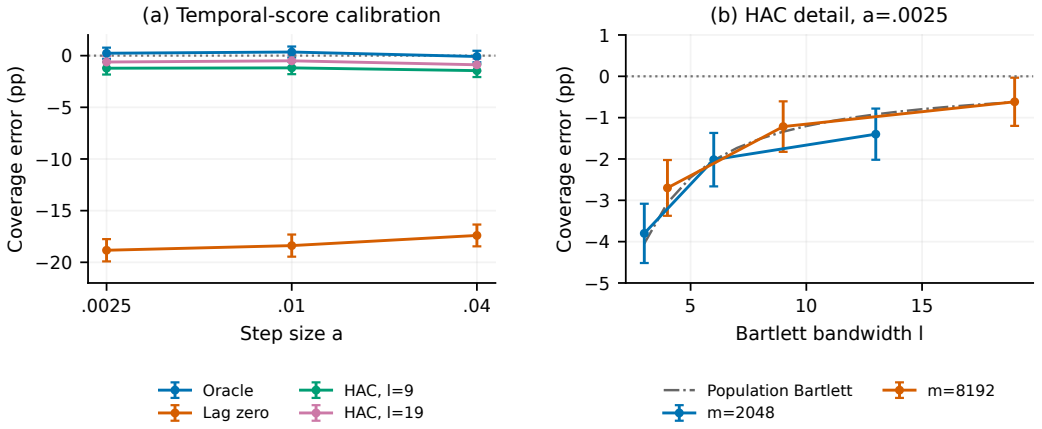


Figure 2: Temporal covariance matters for last-iterate inference. Coverage errors are in percentage points relative to 95%, with 1.96 Monte Carlo standard-error bars from 6,000 replications. (a) Oracle, lag-zero and feasible HAC calibration; HAC uses $m = 8192$. (b) HAC bandwidth sensitivity at $a = .0025$, with fixed axis limits $[-5, 1]$ pp. The gray dash-dotted curve is the population Bartlett Gaussian-limit benchmark; feasible estimates use observed calibration data.

Scope and initialization. Pareto AR designs, finite-variance non-Gaussian errors and bounded two-time characteristic functions support the qualitative stationary and OU predictions (Figure 3). The log-Weibull comparison shows that entry into the first-order regime depends on the design law. Correction is accurate at innovation scale $c = .5$; at $c = 1$, the baseline remains closer to nominal over the tested steps (Tables 4 and 7). Actual Gaussian SGD with nonzero initialization and a separate residual-variance calibration block attains near-nominal history-corrected coverage from early iterations; Figure 4 reports the accompanying interval widths. At $a = .01$, the median estimated-variance interval width decreases from 2.186 at $an = .05$ to .283 at $an = 24$; the corresponding 90th percentiles are 6.242 and .316. These widths quantify the precision cost of early inference alongside its coverage.

8 Discussion

Temporal persistence can destroy every stationary moment while preserving a sharp Gaussian approximation. Pathwise contraction and score cancellation separate stability along typical paths from integrability across environments. Conditional Gaussian structure turns this distinction into a quantitative law approximation and a leading coverage correction.

For fixed-dimensional least-squares SGD, these results support central inference even when stationary means and covariances are unavailable. Finite-history transfer connects the expansion to initialized runs; exact Gaussian calibration also permits earlier inference, with interval widths recording precision. With dependent scores, calibration instead uses their long-run covariance. The log-Weibull results show that entry into the first-order regime depends on the design law. Extending quantitative approximation to general dependent scores remains an open direction.

REPRODUCIBILITY STATEMENT

Appendices provide complete proofs and the numerical protocol. The accompanying source package contains independently implemented simulation and analysis scripts, declared configurations and seeds, saved replication outputs, generated tables and figures, dependency versions, and algebraic checks. The experiments use synthetic data and retain the endpoint and path-summary outputs needed for analysis from every replication. The supplied seeds and code regenerate the synthetic streams; full per-update trajectories were processed online rather than stored. The README gives reproduction commands.

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

Rafik Aguech, Eric Moulines, and Pierre Priouret. On a perturbation approach for the analysis of stochastic tracking algorithms. *SIAM Journal on Control and Optimization*, 39(3):872–899, 2000. doi: 10.1137/S0363012998333852.

Andreas Brandt. The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Advances in Applied Probability*, 18(1):211–220, 1986. URL https://doi.org/10.1017/S0001867800015639.

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), Article 19:1–24, 2022. URL https://doi.org/10.1145/3508039.

Ewa Damek and Sebastian Mentemeier. Analysing heavy-tail properties of stochastic gradient descent by means of stochastic recurrence equations. *Journal of Applied Probability*, 63(2):459–483, 2026. URL https://doi.org/10.1017/jpr.2025.10036.

B. Delyon and A. Juditsky. On small perturbations of stable Markov operators: Unbounded case. *Theory of Probability and Its Applications*, 43(4):577–587, 1999. doi: 10.1137/S0040585X97977173.

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. URL https://arxiv.org/abs/1707.06386.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. On the stability of random matrix product with Markovian noise: Application to linear stochastic approximation and TD learning. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1711–1752, 2021a. URL https://proceedings.mlr.press/v134/durmus21a.html.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight high probability bounds for linear stochastic approximation with fixed stepsize. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30063–30074, 2021b. URL https://arxiv.org/abs/2106.01257.

Charles M. Goldie. Implicit renewal theory and tails of solutions of random equations. *The Annals of Applied Probability*, 1(1):126–166, 1991. URL https://doi.org/10.1214/aoap/1177005985.

Lei Guo, Lennart Ljung, and Guan-Jun Wang. Necessary and sufficient conditions for stability of LMS. *IEEE Transactions on Automatic Control*, 42(6):761–770, 1997. URL https://doi.org/10.1109/9.587328.

Mert Gurbuzbalaban, Umut Simsekli, 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.

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.

Dongyan Huo, Yudong Chen, and Qiaomin Xie. Bias and extrapolation in Markovian linear stochastic approximation with constant stepsizes. *Mathematics of Operations Research*, published online June 1, 2026. URL https://doi.org/10.1287/moor.2024.0471.

Harry Kesten. Random difference equations and renewal theory for products of random matrices. *Acta Mathematica*, 131:207–248, 1973. URL https://doi.org/10.1007/BF02392040.

Marat Khusainov, Marina Sheshukova, Alain Durmus, and Sergey Samsonov. On the rate of Gaussian approximation for linear regression problems. arXiv:2509.14039v1, 2025. URL https://arxiv.org/abs/2509.14039v1.

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. URL https://doi.org/10.1137/0319007.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019. URL https://jmlr.org/papers/v20/17-526.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.

Ibrahim Merad and Stéphane Gaïffas. Convergence and concentration properties of constant step-size SGD through Markov chains. arXiv:2306.11497, 2023. Version 3, revised November 2025. URL https://arxiv.org/abs/2306.11497v3.

Wenlong Mou, Chris Junchi Li, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. On linear stochastic approximation: Fine-grained Polyak–Ruppert and non-asymptotic concentration. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2947–2997, 2020. URL https://proceedings.mlr.press/v125/mou20a.html.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. URL https://www.nber.org/papers/t0055.

Alexander Roitershtein. One-dimensional linear recursions with Markov-dependent coefficients. *The Annals of Applied Probability*, 17(2):572–608, 2007. doi: 10.1214/105051606000000844. URL https://arxiv.org/abs/math/0409335.

Yinan Shen, Yichen Zhang, and Wen-Xin Zhou. SGD with dependent data: Optimal estimation, regret, and inference. arXiv:2601.01371, 2026. URL https://arxiv.org/abs/2601.01371.

R. Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2803–2830, 2019. URL https://proceedings.mlr.press/v99/srikant19a.html.

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:2602.13960, 2026. URL https://arxiv.org/abs/2602.13960.

Wei Biao Wu and Michael Woodroofe. Martingale approximations for sums of stationary processes. *The Annals of Probability*, 32(2):1674–1690, 2004. doi: 10.1214/009117904000000351. URL https://arxiv.org/abs/math/0410160.

Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005. URL https://doi.org/10.1073/pnas.0506715102.

Wei Biao Wu. Strong invariance principles for dependent random variables. *The Annals of Probability*, 35(6):2294–2320, 2007. doi: 10.1214/009117907000000060. URL https://arxiv.org/abs/0711.3674.

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://proceedings.neurips.cc/paper/2021/hash/21ce689121e39821d07d04faab328370-Abstract.html.

A PATHWISE STABILITY AND STATIONARY EXISTENCE

All almost sure statements below can be imposed on a common probability-one event. The underlying shift of an iid sequence is mixing in the ergodic-theoretic sense, and its nonzero integer powers are ergodic. Every measurable causal factor inherits these properties. This fact is used only to apply the ergodic theorem to fixed nonoverlapping blocks; it is not a quantitative mixing assumption.

Lemma A.1 (Integrable sequences and exponential weights)**.** *Let $(h_j)_{j\ge 0}$ be a stationary nonnegative sequence with $\mathbb{E}h_0 < \infty$. Almost surely,*

$$D_h := \sup_{n\ge 1} \frac{1}{n} \sum_{j=0}^{n-1} h_j < \infty, \qquad \frac{h_j}{j+1} \longrightarrow 0. \tag{34}$$

For every $c > 0$, there is a deterministic $C_c < \infty$ such that, for $0 < a \le 1$,

$$a \sum_{j\ge 0} e^{-caj} h_j \le C_c D_h, \tag{35}$$

$$a^2 \sum_{j\ge 0} (j+1) e^{-caj} h_j \le C_c D_h. \tag{36}$$

Furthermore,

$$a^2 \sum_{j\ge 0} e^{-caj} h_j^2 \longrightarrow 0 \quad \text{a.s.} \qquad (a \downarrow 0). \tag{37}$$

Proof. The first assertion of (34) follows from the ergodic theorem, with the possibly random invariant conditional expectation as limit. For every $\epsilon > 0$, $\sum_{j\ge 1} \mathbb{P}(h_j > \epsilon j) = \sum_{j\ge 1} \mathbb{P}(h_0 > \epsilon j) < \infty$. Borel–Cantelli proves the second assertion, first for rational $\epsilon > 0$ and then for all $\epsilon > 0$. If $q = e^{-ca}$ and $C_n = \sum_{j<n} h_j$, Abel summation gives

$$\sum_{j\ge 0} q^j h_j = (1-q) \sum_{j\ge 0} q^j C_{j+1} \le \frac{D_h}{1-q}.$$

The ratio $a/(1-e^{-ca})$ is bounded on $(0, 1]$, proving (35). Since $(j+1)e^{-caj/2} \le C_c/a$, equation (36) follows from (35) with $c/2$. Finally, given $\epsilon > 0$, choose a random finite J such that $h_j \le \epsilon(j+1)$ for $j \ge J$. The finite sum over $j < J$ in (37) tends to zero, while the remainder is at most $\epsilon C_c D_h$ by (36). Let $\epsilon \downarrow 0$. □

Lemma A.2 (Deterministic block inequality)**.** *For positive semidefinite matrices $B_1, \dots, B_m$, write $s = \sum_{\ell=1}^m \|B_\ell\|$ and $g = \lambda_{\min}(\sum_{\ell=1}^m B_\ell)$. For all $a > 0$,*

$$\log \|(I - aB_m) \cdots (I - aB_1)\| \le -ag + 2as \min(as, 1),$$

where $\log 0 = -\infty$.

Proof. Let $x = as$ and $Q = \prod_{\ell=m}^1 (I - aB_\ell)$. Expansion in powers of a does not require commutativity. The sum of norms of terms of order r is bounded by $x^r/r!$. Consequently,

$$\left\| Q - \left(I - a \sum_\ell B_\ell \right) \right\| \le e^x - 1 - x.$$

For $x \le 1$, every eigenvalue of $a\sum_\ell B_\ell$ belongs to $[0, 1]$, so the norm of $I - a\sum_\ell B_\ell$ is $1 - ag$. Since $e^x - 1 - x \le (e/2)x^2 < 2x^2$, the inequality $\log u \le u - 1$ gives the result. For $x > 1$,

$$\log \|Q\| \le \sum_\ell \log(1 + a\|B_\ell\|) \le as \le -ag + 2as,$$

using $g \le s$. □

Proof of Theorem 4.1. By the ergodic theorem, $m^{-1}\sum_{j=0}^{m-1} H_{-j} \to H \succ 0$ almost surely as $m \to \infty$. Hence there exists a deterministic finite m such that

$$\mu := \mathbb{E}\lambda_{\min}\left(\sum_{j=0}^{m-1} H_{-j}\right) > 0.$$

This expectation is finite because the eigenvalue is bounded by the sum of matrix norms. Partition the past into blocks $\{-(r-1)m, \dots, -rm+1\}$, $r \geq 1$, and let g_r and s_r be the corresponding smallest Gram eigenvalue and sum of norms. Dominated convergence gives

$$\mathbb{E}[2s_1 \min(a_0 s_1, 1)] \longrightarrow 0 \quad (a_0 \downarrow 0),$$

since $\mathbb{E}s_1 < \infty$. Fix $a_0 \leq 1$ such that this expectation is less than $\mu/2$, and set

$$z_r = -g_r + 2s_r \min(a_0 s_r, 1), \qquad \eta = -\mathbb{E}z_1 > 0.$$

Lemma A.2 bounds each block log norm by az_r, simultaneously for $0 < a \leq a_0$. Ergodicity under the block shift implies $N^{-1}\sum_{r=1}^{N} z_r \to -\eta$ almost surely. Also $s_{N+1}/N \to 0$ almost surely, by the integrability and stationarity of block sums. Therefore

$$L = \sup_{N \geq 0}\left\{\sum_{r=1}^{N} z_r + s_{N+1} + \frac{\eta}{2}N\right\} < \infty \quad \text{a.s.}$$

For $j = Nm + r$, $0 \leq r < m$, the incomplete block costs at most as_{N+1} in log norm. Submultiplicativity gives

$$\log \|P_j(a)\| \leq aL - \frac{a\eta}{2}N \leq a(L + \eta/2) - \frac{a\eta}{2m}j.$$

Thus (9) holds with $c = \eta/(2m)$ and $K = \exp\{a_0(L + \eta/2)\}$. The same argument, applied to forward blocks, proves

$$\|(I - aH_n)\cdots(I - aH_1)\| \leq K^{+}e^{-can}, \qquad 0 < a \leq a_0, \tag{38}$$

for another almost surely finite random K^{+}. No integrability of L, K, or K^{+} is used.

If $\mathbb{E}\log(1 + \|\xi_0\|) < \infty$, stationarity and Borel–Cantelli imply $j^{-1}\log(1 + \|\xi_{-j}\|) \to 0$ almost surely. Exponential contraction then implies absolute convergence of (10) for every $a \in (0, a_0]$. Shifting the same series constructs a causal stationary process, and factoring out $I - aH_i$ verifies the recursion.

If e_i and $\widetilde{e}_i$ are stationary solutions coupled to the same data, backward iteration gives $e_0 - \widetilde{e}_0 = P_n(a)(e_{-n} - \widetilde{e}_{-n})$. The differences on the right form a tight family, since the two marginal laws are stationary. The product tends to zero almost surely, hence the right side tends to zero in probability. The left side is therefore zero almost surely. This also proves uniqueness without a moment condition on a competing stationary solution. Finally, the difference from a recursion initialized at a finite value equals the forward product in (38) times its finite initial discrepancy. It tends to zero almost surely. In particular, the marginal law from that initialization converges weakly to π_a for fixed a. □

B SCORE INVARIANCE PRINCIPLE AND MAXIMAL INEQUALITY

For completeness, we record the precise consequences of the projective condition. They are standard martingale-approximation facts; compare Wu & Woodroofe (2004); Wu (2007).

Lemma B.1 (Projective score bounds)**.** *Under Assumption 2, let $\theta_k = \|\mathcal{P}_0\xi_k\|_2$ and $D_i = \sum_{k\geq 0}\mathcal{P}_i\xi_{i+k}$. Then D_i is a stationary ergodic L^2 martingale-difference sequence, $\Omega = \mathbb{E}D_0D_0^{\mathsf{T}}$, and*

$$\left\|\max_{1\leq j\leq n}\left\|\sum_{i=1}^{j}\xi_i\right\|\right\|_2 \leq 2\sqrt{n}\sum_{k\geq 0}\theta_k. \tag{39}$$

The backward maximum satisfies the same bound with constant 4 instead of 2. The partial-sum processes satisfy a functional CLT on every compact time interval, including jointly over finite intervals extending to negative times.

Proof. The iid past tail sigma-field is trivial. Causality and centering therefore give the L^2 expansion $\xi_i = \sum_{k\geq 0} \mathcal{P}_{i-k}\xi_i$. For a fixed k, the sequence $(\mathcal{P}_{i-k}\xi_i)_i$ is a stationary martingale difference relative to $(\mathcal{F}_{i-k})_i$. Doob's inequality gives a maximal L^2 norm at most $2\sqrt{n}\theta_k$ for its partial sums. Minkowski's inequality and $\sum_k \theta_k < \infty$ prove (39). By stationarity, a backward partial sum over a window of length n has the distribution of a difference between the endpoint and a forward partial sum on that window. Its maximum is bounded by twice the forward maximum.

Orthogonality of innovation projections yields, for $k \geq 0$,

$$\left\|\mathbb{E}\xi_0\xi_k^{\mathsf{T}}\right\| \leq \sum_{j\geq 0} \theta_j\theta_{j+k}.$$

Summing over positive and negative lags proves absolute covariance summability. The series defining D_i converges in L^2 by Minkowski's inequality; its terms are $\mathcal{F}_i$-measurable with zero conditional mean given $\mathcal{F}_{i-1}$. Hence D_i is a stationary ergodic martingale difference.

Here are details of the functional approximation. For fixed M, define

$$\xi_i^{[M]} = \sum_{k=0}^{M} \mathcal{P}_{i-k}\xi_i, \qquad D_i^{[M]} = \sum_{k=0}^{M} \mathcal{P}_i\xi_{i+k}.$$

Their partial sums differ by $R_0^{[M]} - R_n^{[M]}$, where

$$R_i^{[M]} = \sum_{k=1}^{M}\sum_{r=0}^{k-1} \mathcal{P}_{i-r}\xi_{i-r+k}$$

is stationary and square integrable. Since $t^2\mathbb{P}(\left\|R_0^{[M]}\right\| > t) \to 0$,

$$\mathbb{P}\left(\max_{j\leq n}\left\|R_j^{[M]}\right\| > \epsilon\sqrt{n}\right) \leq (n+1)\mathbb{P}(\left\|R_0^{[M]}\right\| > \epsilon\sqrt{n}) \longrightarrow 0.$$

The stationary ergodic martingale $\sum_i D_i^{[M]}$ satisfies the functional CLT: its averaged conditional quadratic variation converges by ergodicity, and its conditional Lindeberg condition follows from the L^2 tail and Markov's inequality. Thus the same CLT holds for $\xi_i^{[M]}$. The maximal approximation remainder satisfies

$$n^{-1/2}\left\|\max_{j\leq n}\left\|\sum_{i=1}^{j}(\xi_i - \xi_i^{[M]})\right\|\right\|_2 \leq 2\sum_{k>M}\theta_k \longrightarrow 0.$$

Letting $M \to \infty$ proves the score functional CLT with covariance $\mathbb{E}D_0D_0^{\mathsf{T}}$. The analogous nonmaximal L^2 approximation shows that this covariance equals the limit of $n^{-1}\operatorname{Cov}(\sum_{i=1}^n \xi_i)$, which is Ω by absolute covariance summability.

Time reversal on a fixed window maps a forward partial-sum process to its endpoint minus its reversed path. Applied to its Brownian limit this gives a Brownian process with the same covariance. More generally, by stationarity one can begin a window at $-\lfloor L/a\rfloor$ and apply the forward functional CLT on a window of length of order $(L+T)/a$. Subtracting its value at time zero gives the asserted two-sided formulation and independent positive- and negative-time Brownian increments in the limit. □

To check (4), observe that $\mathcal{P}_0\xi_k = \mathbb{E}(\xi_k - \xi_k^{\{0\}} \mid \mathcal{F}_0)$. Indeed, conditioning the coupled variable on $\mathcal{F}_0$ integrates out the replaced innovation and equals $\mathbb{E}(\xi_k \mid \mathcal{F}_{-1})$. Conditional Jensen's inequality implies $\theta_k \leq \delta_{\xi,2}(k)$.

C Finite windows, remote past, and process convergence

For a quantitative comparison that remains meaningful without moments, let d_{BL} be the supremum of expectation differences over functions bounded by one with Lipschitz constant at most one. Put $U_{a,T} = \sqrt{a}\sum_{j<\lceil T/a\rceil} P_j(a)\xi_{-j}$ and $V_T = \int_0^T e^{-tH}\Omega e^{-tH}\,dt$.

Proposition C.1 (A bounded-test-function approximation bound)**.** *Under Assumptions 1–2, for $a \le a_0$, $T \ge 1$, and $M \ge 1$,*

$$d_{\mathrm{BL}}\big(\mathcal{L}(a^{-1/2}e_0^{(a)}), N_p(0,V)\big) \le d_{\mathrm{BL}}\big(\mathcal{L}(U_{a,T}), N_p(0,V_T)\big) + 2\mathbb{P}(K \vee D > M) + CM^2(1+T)^{3/2}e^{-c'T} + Ce^{-\lambda_{\min}(H)T}. \tag{40}$$

The first term concerns a finite learning-time window; the remaining terms explicitly separate localization of the random environment from removal of the remote past. A rate in a requires quantitative information about that finite-window approximation and the environment. The present assumptions ensure convergence, without assigning such a rate. The bound also explains why an approximation metric based on bounded events is appropriate when every Wasserstein distance of positive order is infinite.

C.1 Averaging the matrix products

Lemma C.2 (Uniform finite-window product limit)**.** *Under Assumption 1, for every finite T,*

$$\sup_{0\le t\le T} \big\|P_{\lfloor t/a\rfloor}(a) - e^{-tH}\big\| \longrightarrow 0 \quad a.s. \tag{41}$$

The total variation of $j \mapsto P_j(a)$ over $0 \le j \le \lceil T/a\rceil$ is bounded almost surely, uniformly for sufficiently small a.

Proof. Write $h_j = \|H_{-j}\|$ and $N = \lceil T/a\rceil$. By the ergodic theorem,

$$M_a = a\sum_{j<N} h_j = O(1) \quad \text{a.s.}, \qquad \delta_a := \max_{k\le N}\left\| a\sum_{j<k}(H_{-j}-H)\right\| \longrightarrow 0 \quad \text{a.s.}.$$

Uniformity of the second statement follows directly from Cesaro convergence: after a fixed index the unscaled partial sums are at most ϵk, and the finite initial segment vanishes after multiplication by a. For each $j \le N$,

$$\|P_j(a)\| \le \prod_{\ell<j}(1+ah_\ell) \le e^{M_a}, \quad \sum_{j<N}\|P_{j+1}(a) - P_j(a)\| \le e^{M_a}M_a.$$

Let $C_k = a\sum_{j<k}(H_{-j}-H)$. Discrete integration by parts yields

$$a\sum_{j<n} P_j(a)(H_{-j}-H) = P_{n-1}(a)C_n + \sum_{j=1}^{n-1}(P_{j-1}(a)-P_j(a))C_j.$$

Its norm, uniformly for $n \le N$, is at most $\delta_a e^{M_a}(1+M_a) = o(1)$ almost surely. The product recursion becomes

$$P_n(a) = I - a\sum_{j<n} P_j(a)H + o(1)$$

uniformly over the window. Comparing with $(I-aH)^n$ and applying discrete Gronwall proves (41). No coefficient truncation is needed for this finite-window argument. □

We use the following elementary continuity principle for the random weights. Suppose matrix-valued f_n converge uniformly to a continuously differentiable f, their total variations are bounded, and vector-valued g_n converge uniformly to a continuous g. Define $\int f_n\,dg_n$ by discrete summation and integration by parts. These integrals converge to $f(T)g(T) - f(0)g(0) - \int_0^T df(t)g(t)$. To see this, approximate g uniformly by a piecewise linear function. The approximation error in integration by parts is bounded by the approximation's uniform error times the total variation of f_n and its endpoint norms. For the piecewise linear approximation, uniform convergence of f_n suffices. Left- versus right-endpoint conventions contribute at most $\mathrm{TV}(f_n)\max_j\|\Delta g_n(j)\|$, which vanishes when the limit of g_n is continuous. The same argument applies when total variations are only bounded in probability, by localization and the usual subsequence representation of convergence in distribution to continuous paths.

Applying this principle to Lemmas B.1 and C.2 proves (31). The limiting integral is the ordinary stochastic integral with deterministic integrand e^{-tH}. No stochastic predictability of the finite-a backward weights is used.

C.2 Uniform control of the remote past

Lemma C.3 (Remote-past localization). *Under Assumptions 1–2, let* $N = \lceil T/a \rceil$ *and* $R_{a,T} = \sqrt{a}\sum_{j\geq N} P_j(a)\xi_{-j}$. *There are deterministic* $C, c' > 0$ *such that for* $0 < a \leq a_0 \wedge 1$, $T \geq 1$, $M \geq 1$, *and* $\epsilon > 0$,

$$\mathbb{P}(\|R_{a,T}\| > \epsilon) \leq \mathbb{P}(K \vee D > M) + \epsilon^{-1} C M^2 (1+T)^{3/2} e^{-c'T},$$

where $D = \sup_{n\geq 1} n^{-1}\sum_{j<n}\|H_{-j}\| < \infty$ *almost surely. In particular,*

$$\lim_{T\to\infty} \limsup_{a\downarrow 0} \mathbb{P}(\|R_{a,T}\| > \epsilon) = 0. \tag{42}$$

Proof. Since $\xi_0 \in L^1$, ergodicity gives $S_n/n \to 0$ almost surely. The boundary term at infinity therefore vanishes when summing $P_j(S_{j+1} - S_j)$ by parts. The identity $P_j - P_{j+1} = aP_jH_{-j}$ proves (32). Its series converges pathwise: the products decay exponentially, whereas H_{-j} and S_j grow at most polynomially almost surely by Lemma A.1 and the ergodic theorem.

On $E_M = \{K \leq M, D \leq M\}$, put $b_\ell = 2^\ell N$. For $b_\ell \leq j < 2b_\ell$, sum the matrix norms before taking any expectation:

$$\sum_{j=b_\ell}^{2b_\ell-1} e^{-caj}\,\|H_{-j}\|\,\|S_{j+1}\| \leq 2Db_\ell e^{-cab_\ell} \max_{k\leq 2b_\ell}\|S_k\|.$$

Consequently,

$$\begin{aligned}\mathbb{E}\big[\mathbf{1}_{E_M}\|R_{a,T}\|\big] &\leq M\sqrt{a}e^{-caN}\mathbb{E}\,\|S_N\| + 2M^2a^{3/2}\sum_{\ell\geq 0} b_\ell e^{-cab_\ell}\mathbb{E}\max_{k\leq 2b_\ell}\|S_k\| \\ &\leq CM^2\left\{\sqrt{aN}e^{-caN} + \sum_{\ell\geq 0}(ab_\ell)^{3/2}e^{-cab_\ell}\right\},\end{aligned}$$

by Lemma B.1. Since $T \leq aN \leq T+1$, the last expression is at most $CM^2(1+T)^{3/2}e^{-c'T}$ after reducing $c' > 0$ if necessary. Markov's inequality on E_M, and the bound $\mathbb{P}(E_M^c)$ on its complement, prove the claimed inequality. Independence between K, D, and the score is never needed. For fixed M, let $T \to \infty$, and then let $M \to \infty$. This proves (42). □

Proof of the marginal assertion of Theorem 4.2. The series representation gives

$$a^{-1/2}e_0^{(a)} = \sqrt{a}\sum_{j<\lceil T/a\rceil} P_j(a)\xi_{-j} + R_{a,T}.$$

The first term converges by (31) to $G_T = \int_0^T e^{-tH}\Omega^{1/2}\,dW_t$. Since $H \succ 0$, G_T converges in L^2 as $T \to \infty$ to a centered Gaussian vector with covariance V in (12). Lemma C.3 and the converging-together argument prove (13). Differentiating $e^{-tH}\Omega e^{-tH}$ and integrating over $t \geq 0$ gives $HV + VH = \Omega$. If $HV + VH = 0$, diagonalization of H shows each entry in its eigenbasis is multiplied by a strictly positive sum of eigenvalues, proving uniqueness. No expectation of the stationary iterate has been passed to a limit. □

Proof of Proposition C.1. Let f be bounded by one and have Lipschitz constant at most one. On $E_M = \{K \leq M, D \leq M\}$, the absolute difference $|f(U_{a,T} + R_{a,T}) - f(U_{a,T})|$ is at most $\|R_{a,T}\|$; on E_M^c it is at most two. The localized expectation estimate in the proof of Lemma C.3 bounds the first contribution by $CM^2(1+T)^{3/2}e^{-c'T}$, and the second by $2\mathbb{P}(E_M^c)$. Couple G_T and G_∞ using the same Brownian motion in their integral representations. Then $\mathbb{E}\,\|G_\infty - G_T\| \leq Ce^{-\lambda_{\min}(H)T}$ by the Gaussian covariance integral. The triangle inequality, followed by the supremum over f, proves (40). □

C.3 Stationary process convergence and initialization

Proof of the process assertion of Theorem 4.2. Fix $L, T > 0$ and initialize an auxiliary recursion at zero at index $-\lceil L/a \rceil$. On the interval $[-L, T]$, the cumulative centered Gram matrices converge uniformly to zero after multiplication by a, and the cumulative sums of Gram norms are bounded after that scaling. The proof of Lemma C.2 applies uniformly to every subinterval: its bounds involve only the total Gram norm over the whole interval and twice the maximum cumulative centered Gram sum. Thus the two-parameter transition matrices converge uniformly to $e^{-H(t-s)}$, $-L \le s \le t \le T$. Their total variations in s are uniformly bounded.

Lemma B.1 gives a joint score functional CLT on $[-L, T]$. The integration-by-parts continuity principle used above, now uniformly in the endpoint t, shows that the rescaled auxiliary recursion converges in $D([0, T], \mathbb{R}^p)$ to

$$U^{[L]}(t) = \int_{-L}^{t} e^{-H(t-s)} \Omega^{1/2} \, dW_s.$$

To make the functional step explicit, take a Skorokhod representation of the joint score and transition-matrix convergence, localizing their variation bounds. Continuity of the score limit gives uniform score convergence on this representation. Approximate that limit by a single piecewise linear function on $[-L, T]$; the variation and endpoint bounds control the approximation error uniformly in t. The resulting deterministic integrals converge uniformly by the uniform transition-matrix limit. This proves weak convergence of the auxiliary processes in the J_1 topology, and hence their tightness, on the original probability space.

At time zero, the discrepancy between the stationary and auxiliary recursions is the remote-past term $R_{a,L}$, up to the harmless rounding of L/a. Over the future interval this discrepancy is multiplied by forward transitions. Their largest norm is bounded by

$$\exp\left\{ a \sum_{i=1}^{\lceil T/a \rceil} \|H_i\| \right\} = O_P(1),$$

independently of L. Denote this bound by $M_{a,T}$, the rescaled stationary process by U_a, and the rescaled auxiliary process by $U_a^{[L]}$. For every $B \ge 1$,

$$\mathbb{P}\left(\sup_{t \le T} \left\| U_a(t) - U_a^{[L]}(t) \right\| > \epsilon \right) \le \mathbb{P}(M_{a,T} > B) + \mathbb{P}(\|R_{a,L}\| > \epsilon / B).$$

Lemma C.3, followed by $B \to \infty$, gives $\lim_{L \to \infty} \limsup_{a \downarrow 0} \mathbb{P}(\sup_{t \le T} \left\| U_a(t) - U_a^{[L]}(t) \right\| > \epsilon) = 0$. On a common two-sided Brownian motion, $U^{[L]}$ converges uniformly in probability on $[0, T]$ as $L \to \infty$ to $U(t) = \int_{-\infty}^{t} e^{-H(t-s)} \Omega^{1/2} \, dW_s$. The difference equals e^{-Ht} times a centered Gaussian initial discrepancy whose variance tends to zero exponentially in L. The converging-together argument in $D([0, T], \mathbb{R}^p)$ gives the stationary OU limit, whose initial value is independent of future Brownian increments. □

Proof of Corollary 5.5. Couple the initialized and stationary recursions on the same observations. By (38),

$$\frac{\left\| \beta_{n(a)} - \beta_{n(a)}^{(a)} \right\|}{\sqrt{a}} \le K^{+} e^{-can(a)} \left\{ \frac{\|\beta_0 - \beta^*\|}{\sqrt{a}} + \frac{\left\| e_0^{(a)} \right\|}{\sqrt{a}} \right\}.$$

The second term in braces is tight by Theorem 4.2, and K^{+} is finite almost surely and independent of a as a random variable on the common environment. Equation (26) makes the right side converge to zero in probability. Stationarity and Slutsky's theorem finish the proof. □

D Exact conditional covariance and total variation

Proof of Theorem 4.3. The assumptions imply the score condition, because $X_i Z_i$ is an L^2 martingale difference for the joint design-innovation and error filtration. Conditional on $\mathcal{G}_X$, the series (10)

is a linear series of independent centered Gaussian variables. Its covariance is

$$C_a = \sigma^2 a^2 \sum_{j\geq 0} P_j H_{-j} P_j^\mathsf{T}.$$

This covariance is finite almost surely: its norm is bounded by $\sigma^2 a^2 K^2 \sum_{j\geq 0} e^{-2caj} \|H_{-j}\|$, which is finite by Lemma A.1. Thus the conditional distribution is a well-defined Gaussian law, regardless of unconditional moments.

Symmetry of H_{-j} gives the exact matrix identity

$$P_j P_j^\mathsf{T} - P_{j+1} P_{j+1}^\mathsf{T} = 2a P_j H_{-j} P_j^\mathsf{T} - a^2 P_j H_{-j}^2 P_j^\mathsf{T}.$$

The boundary $P_n P_n^\mathsf{T}$ vanishes by (9). Summing and dividing by 2 proves

$$a \sum_{j\geq 0} P_j H_{-j} P_j^\mathsf{T} = \tfrac{1}{2}(I + \mathcal{R}_a), \qquad C_a/a = \sigma^2 (I + \mathcal{R}_a)/2. \tag{43}$$

Let $h_j = \|H_{-j}\|$. For all small a,

$$\|\mathcal{R}_a\| \leq K^2 a^2 \sum_{j\geq 0} e^{-2caj} h_j^2 \longrightarrow 0 \quad \text{a.s.}$$

by (37). The common, step-uniform random bound in Theorem 4.1 is what permits this conclusion without a second moment of H_0.

Let $r_1, \ldots, r_p \geq 0$ be the eigenvalues of $\mathcal{R}_a$. The Kullback–Leibler divergence from the conditional Gaussian law in (16) to $N_p(0, \sigma^2 I/2)$ is

$$\frac{1}{2} \sum_{j=1}^{p} \{r_j - \log(1 + r_j)\} \leq \frac{1}{4} \sum_{j=1}^{p} r_j^2.$$

Pinsker's inequality, for $d_{\mathrm{TV}}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$, gives (17). The bound tends to zero almost surely and is at most one. Integrating the conditional bound and applying bounded convergence proves unconditional total-variation convergence. □

Corollary D.1 (Finite-step central coverage)**.** *Under Theorem 4.3, for any* $0 < \alpha < 1$ *and every sufficiently small fixed* a*,*

$$\mathbb{P}\left(\frac{2\,\|B_a - \beta^*\|^2}{a\sigma^2} \leq \chi^2_{p,1-\alpha} \right) \leq 1 - \alpha.$$

The left side converges to $1 - \alpha$ *as* $a \downarrow 0$*.*

Proof. Conditional on $\mathcal{G}_X$, diagonalize $\mathcal{R}_a$ and represent the statistic as $\sum_{j=1}^{p} (1 + r_j) G_j^2$ with iid standard normal G_j. It is at least $\sum_j G_j^2$ on this coupling. The bound follows conditionally and hence unconditionally. Total-variation convergence proves the limit. This finite-step inequality uses the population noise variance and stationarity; it is not asserted for a finite-history recursion or for an estimated variance. □

E AR(1) LAWS WITH NO POSITIVE STATIONARY MOMENTS

Consider $p = 1$, independent standard Gaussian errors, and

$$X_i = \rho X_{i-1} + \eta_i = \sum_{k\geq 0} \rho^k \eta_{i-k}, \qquad 0 < \rho < 1, \tag{44}$$

with iid positive innovations satisfying $\eta_i \geq 1$. Then $X_i \geq (1 - \rho)^{-1}$. Whenever $\eta_0 \in L^q$, $q \geq 1$,

$$\delta_{X,q}(k) = \rho^k \left\| \eta_0 - \eta_0' \right\|_q. \tag{45}$$

Theorem E.1 (Persistent extremes destroy every stationary moment)**.** *For either innovation law below,* (44) *satisfies Assumptions 1–2. For every sufficiently small fixed* $a > 0$*,*

$$\int |b-\beta^*|^r \, \pi_a(db) = \infty \quad \text{for every } r > 0, \qquad d_{\mathrm{TV}}\left(\mathcal{L}\left(\frac{B_a-\beta^*}{\sqrt{a}}\right), N(0,1/2)\right) \to 0. \tag{46}$$

(i) Pareto innovations: $\mathbb{P}(\eta_0 > u) = u^{-\kappa}$ *for* $u \geq 1$*,* $\kappa > 2$*. For each* $r > 0$*,* $\mathbb{E}|P_m(a)|^r = \infty$ *whenever* $2mr \geq \kappa$*.*

(ii) All-moment innovations: $\mathbb{P}(\eta_0 > u) = \exp\{-(\log u)^\nu\}$ *for* $u \geq 1$*,* $1 < \nu < 2$*. Every finite product and every observation* (X_i, Y_i, Z_i) *has all positive moments, and the observations have geometric functional dependence at every finite moment order. The stationary iterate nevertheless satisfies* (46)*.*

For sufficiently large x*, the stationary tail has a lower bound*

$$\mathbb{P}(|B_a-\beta^*| > x) \geq c_a \exp\{-C_a(\log x)^\gamma\}, \quad \gamma = \tfrac{1}{2} \text{ in (i)}, \quad \gamma = \tfrac{\nu}{2} \text{ in (ii)}. \tag{47}$$

The mechanism can be seen from one innovation $U = \eta_{-m+1}$. It enters all m design values in $P_m(a)$, and, for $U \geq \sqrt{2/a}\,\rho^{-(m-1)}$,

$$|P_m(a)| \geq (a/2)^m \rho^{m(m-1)} U^{2m}. \tag{48}$$

Conditional on the design, the stationary error is Gaussian with variance at least $a^2(1-\rho)^{-2}|P_m(a)|^2$. In the Pareto case, (48) immediately makes every prescribed stationary moment infinite by choosing m sufficiently large. For the second law, choose $U > e^{dm}$ with $d > \log(1/\rho)$. The log amplification is of order m^2, whereas the negative log probability is of order $m^\nu = o(m^2)$. Thus observation moments of every finite order coexist with infinite stationary iterate moments. The product and moment arguments are in Appendices E.2–E.4; Appendix E.5 proves (47).

The Pareto example tests the qualitative theory under a finite second design moment. The log-Weibull example additionally has all finite product moments and satisfies the first-order approximation assumptions, so it tests the distinction between finite-product integrability and stationary moment stability.

E.1 Proofs for the persistent-innovation examples

Let $\lambda_\rho = \log(1/\rho) > 0$ and $\ell = (1-\rho)^{-1}$. Positive innovations bounded below by one give $X_i \geq \ell$. For $q \geq 1$ with $\eta_0 \in L^q$, Minkowski's inequality gives $\|X_0\|_q \leq \|\eta_0\|_q/(1-\rho)$, and replacing η_0 gives the exact dependence coefficient (45). The independent Gaussian errors may be included in the iid innovations defining the causal process. The score has zero conditional mean given the past and finite second moment whenever $\mathbb{E}\eta_0^2 < \infty$.

E.2 A lower bound from a single persistent innovation

Fix $m \geq 1$ and put $U = \eta_{-m+1}$. For $0 \leq j < m$, $X_{-j} \geq \rho^{m-1-j}U$. On

$$U \geq u_m := \sqrt{2/a}\,\rho^{-(m-1)},$$

each $aX_{-j}^2 \geq 2$, so $|1 - aX_{-j}^2| \geq (a/2)X_{-j}^2$. Multiplying proves (48). Let Q_a be the unscaled conditional variance of $e_0^{(a)}$. The $j = m$ summand in its series gives

$$Q_a \geq a^2 P_m(a)^2 X_{-m}^2 \geq a^2\ell^2 P_m(a)^2. \tag{49}$$

For $r > 0$, Tonelli's theorem and the conditional Gaussian moment formula imply

$$\mathbb{E}|e_0^{(a)}|^r = c_r \mathbb{E} Q_a^{r/2} \geq c_r (a\ell)^r \mathbb{E}|P_m(a)|^r, \qquad c_r = \mathbb{E}|N(0,1)|^r > 0. \tag{50}$$

These identities allow infinite expectations.

E.3 Pareto innovations

For $\mathbb{P}(\eta_0 > u) = u^{-\kappa}$, $u \geq 1$, with $\kappa > 2$, the design and score satisfy the main assumptions. For each $r > 0$ and $2mr \geq \kappa$,

$$\mathbb{E}|P_m(a)|^r \geq (a/2)^{mr} \rho^{rm(m-1)} \mathbb{E}[U^{2mr} \mathbf{1}_{\{U \geq u_m\}}] = \infty.$$

Equation (50) proves the failure of every positive stationary moment. Notice the quantifiers: a finite product can retain sufficiently small fractional moments, but for each fixed positive order a sufficiently long product loses that moment.

E.4 All moments of observations, but none of the stationary iterate

For $1 < \nu < 2$, let $\mathbb{P}(\eta_0 > u) = \exp\{-(\log u)^\nu\}$ for $u \geq 1$. One explicit representation is $\eta_0 = \exp(E_0^{1/\nu})$ with E_0 standard exponential. For any $q > 0$,

$$\mathbb{E}\eta_0^q = 1 + q \int_0^\infty \exp\{qt - t^\nu\}\, dt < \infty.$$

Thus the AR(1) design, errors, and responses have all moments, as do all finite products by Hölder's inequality. Equation (45) gives geometric functional dependence at every finite order $q \geq 1$. For $k \geq 1$, the score coefficient also obeys

$$\delta_{\xi,q}(k) = \rho^k \left\|\eta_0 - \eta_0'\right\|_q \left\|Z_0\right\|_q,$$

and its lag-zero coefficient is finite. The response coefficients are bounded by the corresponding design and error coefficients.

Fix $r > 0$ and select $d > \lambda_\rho$. For all sufficiently large m, the event $\{U > e^{dm}\}$ is contained in $\{U \geq u_m\}$. The product lower bound yields

$$\begin{aligned}\log \mathbb{E}|P_m(a)|^r &\geq rm \log(a/2) - r\lambda_\rho m(m-1) + 2rdm^2 - d^\nu m^\nu \\ &= r(2d - \lambda_\rho)m^2 - d^\nu m^\nu + O_{a,r,\rho}(m) \longrightarrow \infty.\end{aligned}$$

For each fixed m this expectation is finite; its divergence as $m \to \infty$, together with (50), forces $\mathbb{E}|e_0^{(a)}|^r = \infty$. Since r was arbitrary, all positive stationary moments fail. Theorem 4.3 applies to both innovation laws and proves the total-variation limit in Theorem E.1.

E.5 Subpolynomial stationary lower tails

On $\{U \geq u_m\}$, equation (48) simplifies to

$$|P_m(a)| \geq e^{\lambda_\rho m(m-1)}.$$

For large x, choose the smallest integer $m \geq 2$ satisfying $a\ell e^{\lambda_\rho m(m-1)} \geq x$. Then $m = O_{a,\rho}(\sqrt{\log x})$ and, by (49), $Q_a^{1/2} \geq x$ on that event. Conditional Gaussianity gives

$$\mathbb{P}(|e_0^{(a)}| > x) \geq \mathbb{P}(|N(0,1)| > 1)\mathbb{P}(\eta_0 \geq u_m).$$

For Pareto innovations, the last probability equals $(a/2)^{\kappa/2} \exp\{-\kappa\lambda_\rho(m-1)\}$. For the all-moment innovations it equals $\exp\{-[\frac{1}{2}\log(2/a) + \lambda_\rho(m-1)]^\nu\}$. These expressions give (47) with constants depending on a and the innovation law. Since $\gamma < 1$, $x^r\mathbb{P}(|e_0^{(a)}| > x) \to \infty$ for every $r > 0$, which also directly proves the moment assertion. No matching upper tail estimate is claimed.

E.6 Why persistence is essential to the all-moment failure

For comparison, suppose $p = 1$ and the observations are iid over time, with $\mathbb{E}X_0^2 \in (0, \infty)$ and $\mathbb{E}|\xi_0|^r < \infty$ for some $0 < r \leq 1$. The function $t \mapsto \left||1-t|^r - 1\right|$ is bounded by $C_r t$ on $[0, \infty)$. Dominated convergence therefore gives

$$\mathbb{E}|1 - aX_0^2|^r = 1 - ra\mathbb{E}X_0^2 + o(a) < 1$$

for sufficiently small a. Subadditivity for order $r \leq 1$, and independence of the score from subsequent update factors, imply

$$\mathbb{E}|e_0^{(a)}|^r \leq a^r \mathbb{E}|\xi_0|^r \sum_{j \geq 0} \{\mathbb{E}|1 - aX_0^2|^r\}^j < \infty.$$

Thus iid designs with these moments retain some positive stationary moments at small steps. In the examples, temporal persistence of the same large innovation is responsible for losing all of them.

F INFERENCE, PRECONDITIONING, AND FIXED MINI-BATCHES

F.1 FEASIBLE CENTRAL INFERENCE WITH INDEPENDENT ERRORS

Suppose (Z_i) is iid, centered, with variance $\sigma^2 \in (0, \infty)$, independent of the design process. Gaussianity is not required here. The score is a martingale difference, and $\Omega = \mathbb{E}X_0X_0^\mathsf{T}Z_0^2 = \sigma^2 H$, giving $V = \sigma^2 I/2$. Define

$$G_m = \frac{1}{m}\sum_{i=1}^{m} X_iX_i^\mathsf{T}, \qquad \widetilde{\beta}_m = G_m^{-1}\frac{1}{m}\sum_{i=1}^{m} X_iY_i, \qquad \widehat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - X_i^\mathsf{T}\widetilde{\beta}_m)^2.$$

On the event that G_m is singular, assign any finite value to $\widetilde{\beta}_m$; this event eventually does not occur almost surely because $G_m \to H \succ 0$. The ergodic theorem gives $m^{-1}\sum_{i=1}^{m}\xi_i \to 0$ almost surely, so $\Delta_m := \widetilde{\beta}_m - \beta^* = G_m^{-1}m^{-1}\sum_i \xi_i \to 0$ almost surely. Expanding the residual variance,

$$\widehat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m} Z_i^2 - 2\Delta_m^\mathsf{T}\frac{1}{m}\sum_{i=1}^{m}\xi_i + \Delta_m^\mathsf{T}G_m\Delta_m \longrightarrow \sigma^2 \quad \text{a.s..}$$

Theorem 4.2 or Corollary 5.5, combined with Slutsky's theorem, gives the chi-square limit in (28). The calibration size can be any deterministic $m(a) \to \infty$; choosing $m(a) = n(a)$ uses only observations already processed. Independence between the variance estimator and the final SGD iterate is unnecessary. The computation can be implemented using the cumulative Gram matrix, cross-product vector, and sum of Y_i^2; it does not estimate the variance from stationary SGD iterates.

F.2 FIXED POSITIVE DEFINITE PRECONDITIONING

Let $M \succ 0$ be deterministic and consider

$$\beta_i = \beta_{i-1} + aMX_i(Y_i - X_i^\mathsf{T}\beta_{i-1}).$$

With $u_i = M^{-1/2}(\beta_i - \beta^*)$, the transformed recursion has positive semidefinite coefficient $\widetilde{H}_i = M^{1/2}H_iM^{1/2}$ and score $\widetilde{\xi}_i = M^{1/2}\xi_i$. All assumptions of the main theorems are preserved. Transforming back, the stationary limiting covariance V_M solves

$$MHV_M + V_MHM = M\Omega M.$$

Under independent homoscedastic errors, $V_M = \sigma^2 M/2$. Under independent Gaussian errors, the conditional total-variation statement also transfers through the invertible linear transformation. Thus the central region in (28) uses the quadratic form $\left\|M^{-1/2}(\beta - b)\right\|^2$ for this algorithm.

F.3 FIXED NONOVERLAPPING MINI-BATCHES

Fix an integer $b \geq 1$. Updating once per block with the average of the b sample gradients gives

$$e_k = (I - a\overline{H}_k)e_{k-1} + a\overline{\xi}_k, \qquad \overline{H}_k = \frac{1}{b}\sum_{i=(k-1)b+1}^{kb} H_i, \quad \overline{\xi}_k = \frac{1}{b}\sum_{i=(k-1)b+1}^{kb}\xi_i.$$

The coefficient is positive semidefinite, integrable, and has expectation H. Grouping iid innovations into blocks produces a causal representation. The proofs of Theorems 4.1–4.2 apply to arbitrary positive semidefinite coefficients, not only rank-one matrices. For the score, the block projection is the sum of the b individual innovation projections in that block. Applying Minkowski's inequality and summing over future blocks bounds its projective coefficient sum by Θ_2. Hence the score condition is inherited. Also,

$$\sum_{k=1}^{n}\overline{\xi}_k = \frac{1}{b}\sum_{i=1}^{nb}\xi_i, \qquad \overline{\Omega} = \Omega/b.$$

The covariance of the limiting Gaussian law, on the block-update learning-time scale, is V/b. For independent Gaussian errors, conditional block scores are independent centered Gaussian vectors with covariance $\sigma^2\overline{H}_k/b$. The telescoping calculation gives conditional covariance $\sigma^2(I+\overline{\mathcal{R}}_a)/(2b)$ for the normalized stationary iterate, with $\overline{\mathcal{R}}_a \to 0$ almost surely. These assertions concern fixed b; no joint growth regime in b, p, and a^{-1} is asserted.

G PROOFS FOR THE GAUSSIAN-REFRESH EXAMPLE

Lemma G.1 (One-coefficient anti-concentration)**.** *For any nondegenerate real random variable Z, there are $d, q_0 > 0$ such that*

$$\inf_{v\in\mathbb{R}} \mathbb{P}(|v + cZ| > d|c|) \geq q_0 \qquad (c \neq 0).$$

Consequently, if $e = \sum_j c_j Z_j$ is a convergent series with iid copies of Z and the coefficients are fixed after conditioning on an independent environment, then, for any one index m and $r > 0$,

$$\mathbb{E}(|e|^r \mid \textit{environment}) \geq q_0 d^r |c_m|^r, \tag{51}$$

$$\mathbb{P}(|e| > x) \geq q_0 \mathbb{P}(d|c_m| \geq x), \tag{52}$$

with an immaterial adjustment of strict inequalities.

Proof. Choose $z_1 < z_2$ for which both $\mathbb{P}(Z \leq z_1)$ and $\mathbb{P}(Z \geq z_2)$ are positive. Take $d = (z_2 - z_1)/4$ and let q_0 be the smaller probability. An interval of length $2d$ cannot meet both separated sets. Hence at least one of these sets is disjoint from $\{z : |v + cz| \leq d|c|\}$, proving the first claim. For the series, condition on the environment and all errors except Z_m; the remainder is a fixed shift. Integrating the conditional bound proves both assertions, also for infinite expectations. Almost sure convergence of the series makes this conditioning legitimate on a set of environments of full measure. □

Proof of Theorem 3.1. Almost surely there is a most recent refresh at or before any fixed time; its location is a measurable function of the iid past. Setting X_i equal to its Gaussian mark constructs the stationary causal solution. In n future steps, no refresh has probability ρ^n; otherwise the latest Gaussian mark is independent of the starting state. This proves the kernel formula and, since the Gaussian law is nonatomic, the total-variation equality.

Replacing the innovation at time zero can affect X_k only if there is no refresh at times $1, \ldots, k$. This event has probability ρ^k and is independent of the time-zero coupled difference. Both coupled marginals are standard Gaussian. The L^q triangle inequality gives the displayed dependence bounds. In particular, $H = 1$ and all design moments exist. The score is a square-integrable martingale difference for the joint design/error innovation filtration, so all base theorems apply.

For every fixed $0 < a < 2/3$, Jensen's inequality gives

$$\mathbb{E} \log |1 - aG_0^2| \leq \tfrac{1}{2} \log \mathbb{E}(1 - aG_0^2)^2 = \tfrac{1}{2} \log(1 - 2a + 3a^2) < 0.$$

The logarithm is integrable (its finitely many real zeros have integrable logarithmic singularities under a Gaussian density). Ergodicity makes the scalar product decay exponentially almost surely, and the score has a finite logarithmic moment. The stationary-series construction and coupling argument therefore apply throughout this interval, for any fixed $\rho < 1$. This is a sufficient interval, not the maximal stability threshold.

Fix such an a and any $r > 0$. Isolate the coefficient of Z_{-m}:

$$c_{a,m} = aP_m(a)X_{-m}.$$

On the event

$$A_{m,u} = \{|X_{-m}| \geq u,\ B_{-m+1} = \cdots = B_0 = 1\},$$

all of the m subsequent designs equal X_{-m}. For $au^2 > 2$,

$$|c_{a,m}| \geq au(au^2 - 1)^m, \qquad \mathbb{P}(A_{m,u}) = \mathbb{P}(|G_0| \geq u)\rho^m.$$

Lemma G.1 therefore implies

$$\mathbb{E}|e_0^{(a)}|^r \geq q_0 d^r (au)^r \mathbb{P}(|G_0| \geq u)\{\rho(au^2 - 1)^r\}^m.$$

Choose a finite u large enough that the quantity in braces exceeds one, then let $m \to \infty$. The left side is independent of m, so it is infinite. This proof works for each $r > 0$ without assuming any stationary moments. The Gaussian-limit claims are the preceding theorems.

When $\rho = 0$, the current design/error pair is independent of the previous iterate. Now

$$\mathbb{E}(1 - aX_i^2)^2 = 1 - 2a + 3a^2 < 1 \quad (0 < a < 2/3).$$

The independent affine recursion has a square-integrable stationary solution. Its second-moment recursion is $M = (1 - 2a + 3a^2)M + a^2\sigma^2$, which gives (8). □

Remark G.2 (Two particularly transparent error choices)**.** With independent Gaussian errors, each time point (X_i, Z_i, Y_i) is jointly Gaussian (possibly singular as a triple), although the full time series is *not* a jointly Gaussian process. With independent Rademacher errors, every observation is sub-Gaussian, the errors are bounded, and $X_i Z_i$ has an exactly Gaussian marginal distribution. The all-moment failure persists in both cases. Only the Gaussian-error case is assigned total-variation convergence here; bounded errors by themselves do not justify that metric.

Corollary G.3 (A subpolynomial lower tail in the Gaussian-refresh example)**.** *For the model in Theorem 3.1, there are positive constants depending on a, ρ and the error distribution such that, for all sufficiently large x,*

$$\mathbb{P}(|e_0^{(a)}| > x) \geq c \exp\left\{-C \frac{\log x}{\log \log x}\right\}.$$

This is a lower bound, not a matching asymptotic.

Proof. Write $y = \log x$ and take $u = \sqrt{y}/\log y$. For large y, $b = au^2 - 1 > 1$ and $\log b \geq \frac{1}{2} \log y$. Choose

$$m = \left\lceil \frac{\log(x/(adu))}{\log b} \right\rceil.$$

For all sufficiently large y, $m \leq 3y/\log y$ and $daub^m \geq x$. The event in the preceding proof and (52) give $\mathbb{P}(|e| > x) \geq q_0 \rho^m \mathbb{P}(|G| \geq u)$. The elementary Gaussian lower tail bound $\mathbb{P}(|G| \geq u) \geq c\, u(1+u^2)^{-1} e^{-u^2/2}$ shows that its negative logarithm is at most $Cy/\log y$. This proves the claim. □

H First-order conditional covariance under a fourth design moment

Proof of Theorem 5.1. Let $h_j = \|H_{-j}\|$. On a probability-one event, the preceding finite-window lemma yields uniform convergence of $P_j(a)$ to e^{-ajH} for $j \leq T/a$, and $a \sum_{j<T/a} h_j^2 = O(1)$ for fixed T. Hence

$$a \sum_{j<T/a} \{P_j H_{-j}^2 P_j^\top - e^{-ajH} H_{-j}^2 e^{-ajH}\} \longrightarrow 0.$$

For each matrix entry, the ergodic theorem and summation by parts against the continuously differentiable deterministic matrix weights give

$$a \sum_{j<T/a} e^{-ajH} H_{-j}^2 e^{-ajH} \longrightarrow \int_0^T e^{-tH} \mathbb{E}(H_0^2) e^{-tH}\, dt.$$

To justify the weighted ergodic step explicitly, centered partial sums of H_{-j}^2 are $o(n)$; multiplying by a makes their maximum up to T/a vanish. Abel summation then bounds the weighted error by this maximum times the finite total variation and endpoint norms of the deterministic weights.

Let $D_2 = \sup_{n \geq 1} n^{-1} \sum_{j<n} h_j^2 < \infty$ almost surely. The tail after $N = \lceil T/a \rceil$ is bounded using (9) and dyadic intervals:

$$a \sum_{j \geq N} \left\| P_j H_{-j}^2 P_j^\top \right\| \leq C K^2 D_2 (1+T) e^{-c'T}.$$

Indeed, on $2^\ell N \leq j < 2^{\ell+1} N$, sum h_j^2 first to get a bound $2^{\ell+1} N D_2$, and use the exponential weight at the left endpoint. This is uniform over $0 < a \leq a_0 \wedge 1$. Letting $T \to \infty$ proves (19). Differentiation under the deterministic integral gives the Lyapunov identity.

The conditional Gaussian covariance formula and local Lipschitz continuity in total variation of nonsingular Gaussian laws with respect to their covariance matrices prove the final statements. The Lipschitz assertion follows directly from the Gaussian relative-entropy formula and Pinsker's inequality in a neighborhood of the identity. □

I INTEGRABLE LOGARITHMIC OVERSHOOTS AND AN UNCONDITIONAL SHARP RATE

Define $\delta_{H,q}(k) = \left\| H_k - H_k^{\{0\}} \right\|_q$, using operator norm inside L^q. We use condition (20). A directly checkable sufficient condition is $X_0 \in L^{2q}$ and $\sum_k \delta_{X,2q}(k) < \infty$, because

$$\delta_{H,q}(k) \le 2 \|X_0\|_{2q} \delta_{X,2q}(k).$$

The qualitative results retain their weaker conditions.

Lemma I.1 (A step-sensitive prefactor with integrable logarithm)**.** *Under (20) and Assumption 1, there exist deterministic $a_0, c > 0$ and $L_* \ge 0$ with $\mathbb{E}L_* < \infty$ such that*

$$\|P_j(a)\| \le e^{aL_*} e^{-caj} \quad \text{for every } j \ge 0,\ 0 < a \le a_0. \tag{53}$$

Only integrability of L_* is needed in the probability-law bounds below.

Proof. Follow the preceding block proof, choosing a fixed block length m and a small $a_0 \le 1$. For backward block r, let s_r be the sum of the Gram norms and g_r the smallest eigenvalue of the sum of the Grams. Put

$$z_r = -g_r + 2s_r \min(a_0 s_r, 1), \qquad \mathbb{E}z_r = -\eta < 0.$$

The existing deterministic product inequality bounds each block log norm by az_r simultaneously in $a \le a_0$. With $W_n = \sum_{r=1}^n (z_r + \eta)$ and $W_0 = 0$, define

$$L = \sup_{n \ge 0} \{W_n - \eta n/2 + s_{n+1}\}.$$

The block bound gives

$$\log \|P_j(a)\| \le a(L + \eta/2) - \frac{a\eta}{2m} j.$$

It remains to prove $\mathbb{E}L < \infty$, rather than freezing e^{aL} at $a = a_0$.

First, $s_r, z_r \in L^q$. The function $s \mapsto 2s \min(a_0 s, 1)$ is globally Lipschitz with constant at most four. The minimum eigenvalue is Lipschitz in operator norm. Thus the block functional defining z_r is Lipschitz, with constant at most five relative to the sum of the changes in its m Gram matrices. Introduce grouped iid innovations and their projections,

$$E_k = (\varepsilon_{(k-1)m+1}, \dots, \varepsilon_{km}), \quad \mathcal{F}_k^{(m)} = \sigma(E_i : i \le k), \quad \mathcal{P}_k^{(m)} U = \mathbb{E}(U \mid \mathcal{F}_k^{(m)}) - \mathbb{E}(U \mid \mathcal{F}_{k-1}^{(m)}).$$

Let $s_k^+ = \sum_{i=(k-1)m+1}^{km} \|H_i\|$ and define $z_k^+ = -\lambda_{\min}(\sum_{i=(k-1)m+1}^{km} H_i) + 2s_k^+ \min(a_0 s_k^+, 1)$. This forward-indexed functional is causal for (E_k), and the backward block satisfies $z_r = z_{1-r}^+$. Define $\delta_{z^+,q}^{\mathrm{block}}(k)$ by replacing E_0 with an independent copy. Replacing its m innovations one at a time gives

$$\sum_{k \ge 0} \delta_{z^+,q}^{\mathrm{block}}(k) \le 5m \sum_{u \ge 0} \delta_{H,q}(u) < \infty.$$

Here a coefficient at a negative lag is interpreted as zero. For a fixed replaced time $\ell \in \{1-m, \dots, 0\}$, the intervals $\{(k-1)m+1, \dots, km\}$, $k \ge 0$, partition $\{1-m, 2-m, \dots\}$. Thus each lag $u = i - \ell \ge 0$ occurs exactly once in the sum over future blocks and their time points, giving the factor m.

For the maximal bound, let $\theta_k = \left\| \mathcal{P}_0^{(m)} z_k^+ \right\|_q$. Conditional Jensen's inequality for the coupled block gives $\theta_k \le \delta_{z^+,q}^{\mathrm{block}}(k)$. The trivial remote-past tail of the iid block innovations and reverse martingale convergence yield the L^q expansion

$$z_i^+ - \mathbb{E}z_0^+ = \sum_{k \ge 0} \mathcal{P}_{i-k}^{(m)} z_i^+.$$

For fixed k, the summands indexed by i form a martingale-difference sequence with L^q norm θ_k. The martingale inequality $\|\sum_i D_i\|_q \le C_q (\sum_i \|D_i\|_q^2)^{1/2}$ and Doob's inequality, followed by Minkowski's inequality over k, give

$$\left\| \max_{j \le n} \left| \sum_{i=1}^j (z_i^+ - \mathbb{E}z_0^+) \right| \right\|_q \le C_q \sqrt{n} \sum_{k \ge 0} \theta_k \le C\sqrt{n}.$$

The summable projection norms justify passage from finite projection sums to the full expansion. Reversing a stationary window bounds the backward maximum by twice the corresponding forward maximum. Therefore

$$\Big\| \max_{j \le n} |W_j| \Big\|_q \le C\sqrt{n}. \tag{54}$$

Take $d = \eta/4$. We have

$$L \le \sup_{n \ge 0} (W_n - dn)_+ + \sup_{n \ge 0} (s_{n+1} - dn)_+.$$

For $N = 2^k$, (54) and the tail integral inequality yield

$$\mathbb{E}\big(\max_{j \le 2N} |W_j| - dN \big)_+ \le \frac{\mathbb{E} \max_{j \le 2N} |W_j|^q}{(q-1)(dN)^{q-1}} \le C N^{1-q/2}.$$

Summing over $k \ge 0$ proves integrability of the first supremum since $q > 2$. For the second,

$$\mathbb{E} \sup_{n \ge 0} (s_{n+1} - dn)_+ \le \sum_{n \ge 0} \mathbb{E}(s_1 - dn)_+ \le C_d(\mathbb{E}s_1 + \mathbb{E}s_1^2) < \infty.$$

Thus $\mathbb{E}L < \infty$. Set $L_* = L + \eta/2$ and $c = \eta/(2m)$. Since the $n = 0$ term in L is $s_1 \ge 0$, $L_* \ge 0$. □

The condition $q > 2$ makes $\sum_k 2^{k(1-q/2)}$ finite and is sufficient for this unconditional integrability step. The fourth design moment in Theorem 5.1 instead controls the pathwise covariance expansion; at $q = 2$, the displayed maximal bound alone does not sum over dyadic blocks.

Proof of Theorem 5.2. Let $E_a = \{aL_* \le 1\}$ and $h_j = \|H_{-j}\|$. By Lemma I.1, on E_a,

$$\|\mathcal{R}_a\| \le e^2 a^2 \sum_{j \ge 0} e^{-2caj} h_j^2.$$

The exact conditional Gaussian law, Pinsker's inequality, and convexity of total variation under mixtures give

$$\begin{aligned} d_{\mathrm{TV}}(\mu_a, \gamma_0) &\le \mathbb{P}(L_* > 1/a) + C_p a^2 \sum_{j \ge 0} e^{-2caj} \mathbb{E} h_0^2 \\ &\le a\mathbb{E}L_* + C_p \frac{a^2 \mathbb{E} h_0^2}{1 - e^{-2ca}} \le Ca. \end{aligned}$$

This step integrates neither e^{aL_*} nor the stationary covariance.

For the refined assertion, define $U_a = a \sum_{j \ge 0} e^{-2caj} h_j^2$. This family is uniformly integrable. One direct proof chooses an increasing convex Ψ with $\Psi(t)/t \to \infty$ and $\mathbb{E}\Psi(h_0^2) < \infty$. The sums of the deterministic weights are uniformly bounded. Normalize them and apply Jensen's inequality to bound $\mathbb{E}\Psi(U_a/C)$ uniformly. Such a Ψ exists for any integrable nonnegative variable, by a piecewise-linear construction from its integrable tails.

Write $\gamma(R) = N_p(0, \sigma^2(I + R)/2)$. By Theorem 5.1,

$$a^{-1} d_{\mathrm{TV}}(\gamma(\mathcal{R}_a), \gamma(aJ)) \longrightarrow 0 \quad \text{almost surely.}$$

On E_a, the triangle inequality via $\gamma(0)$ and the Gaussian Pinsker bound dominate this quantity by

$$C_p\big(\|\mathcal{R}_a\|/a + \|J\| \big) \mathbf{1}_{E_a} \le C_p(e^2 U_a + \|J\|),$$

which is uniformly integrable. Its expectation on E_a therefore tends to zero. On E_a^c use total variation at most one and

$$a^{-1}\mathbb{P}(L_* > 1/a) \longrightarrow 0,$$

which follows from $\mathbb{E}L_* < \infty$. Mixture convexity proves (22).

It remains to evaluate a deterministic Gaussian comparison. After the common rescaling by $\sqrt{2}/\sigma$, differentiating the density of $N_p(0, I + aJ)$ at zero gives

$$\frac{d}{da} f_a(x)\Big|_{a=0} = \tfrac{1}{2}(x^\top Jx - \operatorname{tr} J)\phi_p(x).$$

This differentiation holds in L^1: for a in a small interval, the derivatives are dominated by a polynomial times a Gaussian density with a fixed enlarged covariance. Therefore $d_{\mathrm{TV}}(\gamma_a, \gamma_0)/a$ tends to the right side of (23). The reverse triangle inequality and (22) transfer this limit to μ_a. Finally, $\mathbb{E}H_0^2 \succ 0$: if $v^\top \mathbb{E}H_0^2 v = 0$ then $H_0 v = 0$ almost surely and hence $Hv = 0$, contradicting $H \succ 0$. Thus $J \succ 0$ and the random quadratic form is not constant. □

Corollary I.2 (Quantitative coexistence in both all-moment examples)**.** *The Gaussian-refresh construction of Theorem 3.1, with Gaussian errors, satisfies (21)–(23) and has no positive stationary moments. In this case $H = 1$, $\mathbb{E}H_0^2 = 3$, and $J = 3/2$, independently of ρ. In particular,*

$$d_{\mathrm{TV}}\Big(\mathcal{L}(e_0^{(a)}/\sqrt{a}), N(0, \sigma^2(1 + 3a/2)/2)\Big) = o(a),$$

while the uncorrected distance is asymptotic to $\frac{3}{2}\phi(1)a$. The positive AR(1) construction with $\mathbb{P}(\eta > u) = \exp\{-(\log u)^\nu\}$, $1 < \nu < 2$, also satisfies the assumptions of Theorem 5.2 for any finite $q > 2$. Its qualitative no-moment conclusion therefore coexists with the same order-a unconditional total-variation rate.

Proof. For the refresh model use its displayed L^q dependence bound and Gaussian moments. For the positive AR(1) model, all innovation/design moments are finite and $\delta_{X,2q}(k) = \rho^k \|\eta_0 - \eta_0'\|_{2q}$; apply the sufficient condition preceding Lemma I.1. The scalar constant follows from $\mathbb{E}|G^2 - 1| = 4\phi(1)$. □

Remark I.3 (Localization of the probability-law correction)**.** Even in these examples, $\mathbb{E}\mathcal{R}_a$ can be infinite for every fixed $a > 0$. Localization on $\{aL_* \le 1\}$ and a bounded probability distance make the expansion possible. The deterministic matrix J corrects the Gaussian approximation to the law, while the conditional covariance retains the rare-environment fluctuations responsible for infinite moments.

I.1 First-order coverage from the probability-law expansion

Proof of Corollary 5.4. Fix the scalar design law and $\alpha \in (0, 1)$. Theorem 5.2 gives $d_{\mathrm{TV}}(\mu_a, \gamma_a) = o(a)$, where $\gamma_a = N(0, \sigma^2(1 + aJ)/2)$. For $A = [-\sigma z/\sqrt{2}, \sigma z/\sqrt{2}]$,

$$\mu_a(A) = \gamma_a(A) + o(a) = 2\Phi\left(\frac{z}{\sqrt{1 + aJ}}\right) - 1 + o(a) = 1 - \alpha - aJz\phi(z) + o(a).$$

The last step is a Taylor expansion at $a = 0$, with $z = z_{1-\alpha/2}$ and J fixed. It proves (24). For $A_a = \sqrt{1 + aJ}\, A$, the Gaussian probability is exactly $\gamma_a(A_a) = 1 - \alpha$. The same TV bound applies uniformly to measurable sets, including A_a, proving (25). Both arguments integrate bounded indicators and require no stationary iterate moment.

Under Corollary 5.6, $d_{\mathrm{TV}}(\mu_{a,n}, \gamma_a) = o(a)$ under its stated history conditions. Replacing μ_a by $\mu_{a,n}$ therefore proves the finite-history versions. Mere consistency of an estimated noise variance gives an $o(1)$ coverage conclusion; retaining the first-order expansion requires a separate $o(a)$ control of its effect on coverage. □

I.2 Transfer of the first-order expansion to finite histories

Proof of Corollary 5.6. Put $A_i = I - aH_i$ and $P_{n,j} = A_n \cdots A_{n-j+1}$, with $P_{n,0} = I$ and $T_n = P_{n,n}$. Shifting Lemma I.1 to endpoint n gives

$$\|P_{n,j}\| \le e^{aL_{*,n}} e^{-caj}, \qquad L_{*,n} \overset{d}{=} L_*, \qquad \mathbb{E}L_* < \infty.$$

The constants are independent of n; intersecting the shifted probability-one events makes the bound simultaneous in integer endpoints and $0 < a \le a_0$.

Conditional on $\mathcal{G}_X$, the initialized normalized error is Gaussian with mean $m_n = a^{-1/2}T_n b$ and covariance

$$S_n = \sigma^2 a \sum_{j=0}^{n-1} P_{n,j} H_{n-j} P_{n,j}^{\mathsf{T}}.$$

The stationary conditional covariance S_∞ at the same endpoint extends this sum over all $j \geq 0$. Both are finite almost surely. Telescoping the identity $A_i A_i^{\mathsf{T}} = I - 2aH_i + a^2 H_i^2$ gives

$$S_n = \frac{\sigma^2}{2}(I - T_n T_n^{\mathsf{T}} + \mathcal{R}_{a,n}), \qquad \mathcal{R}_{a,n} = a^2 \sum_{j=0}^{n-1} P_{n,j} H_{n-j}^2 P_{n,j}^{\mathsf{T}} \succeq 0. \tag{55}$$

On $E_{a,n} = \{aL_{*,n} \leq 1\}$, choose t_0 so that $ee^{-ct_0} \leq 1/\sqrt{2}$. For $an \geq t_0$,

$$S_\infty \succeq S_n \succeq \sigma^2 I/4, \qquad \|m_n\| \leq e \|b\| \; a^{-1/2} e^{-can},$$

and

$$\mathbf{1}_{E_{a,n}} \|S_\infty - S_n\| \leq \sigma^2 e^2 a \sum_{j \geq n} e^{-2caj} \|H_{n-j}\|.$$

Taking expectations of the last bound uses only stationarity and $\mathbb{E}\|H_0\| < \infty$, and gives Ce^{-2can}. The event indicator has already been bounded pathwise; no independence of $L_{*,n}$ and the design is used.

For covariance matrices $S_1, S_2 \succeq \nu I$, Gaussian comparison gives

$$d_{\mathrm{TV}}\{N_p(m, S_1), N_p(0, S_2)\} \leq C_{p,\nu}(\|m\| + \|S_1 - S_2\|).$$

Indeed, Pinsker's inequality controls the mean shift. Along the covariance segment $S_t = S_1 + t(S_2 - S_1)$, density differentiation and integration give an L^1 derivative bounded by $p\|S_2 - S_1\|/\nu$, since $\mathbb{E}|G^{\mathsf{T}} DG - \operatorname{tr} D| \leq 2p\|D\|$ for $G \sim N_p(0, I)$ and symmetric D. The covariance lower bound is the only spectral bound needed.

Apply this comparison on $E_{a,n}$, and use TV at most one on its complement. Convexity under mixtures gives (27). Finally,

$$\frac{\mathbb{P}(L_* > 1/a)}{a} \leq \mathbb{E}[L_* \mathbf{1}_{\{L_* > 1/a\}}] \longrightarrow 0.$$

The stated burn-in conditions make every term in (27) $o(a)$. The triangle inequality transfers the corrected $o(a)$ result, and the reverse triangle inequality transfers the uncorrected first-order constant. The asserted history schedule has $an(a) \geq 2\log^2(1/a)$ and therefore meets both sufficient conditions for every fixed $c > 0$. □

J Quenched weak limits and non-Gaussian extensions

Proof of Theorem 4.4. For a fixed design path put $h_j = \|H_{-j}\|$ and $w_{a,j} = \sqrt{a}\, P_j(a) X_{-j}$. Identity (43) gives

$$\sum_j w_{a,j} w_{a,j}^\top \to I/2.$$

We also have $\max_j \|w_{a,j}\| \to 0$. Indeed, by stationarity and integrability, Borel–Cantelli gives $h_j/(j+1) \to 0$ almost surely. For any $\varepsilon > 0$, split the bound

$$\sup_j \|w_{a,j}\|^2 \leq K^2 a \sup_j e^{-2caj} h_j$$

at a finite random J_ε after which $h_j \leq \varepsilon(j+1)$. The initial part tends to zero; the remaining part has limsup at most $CK^2\varepsilon$. Let $\varepsilon \to 0$.

Fix $v \in \mathbb{R}^p$. Conditional on the design, the summands $v^\top w_{a,j} Z_{-j}$ are independent and centered. Let $b_a = \sup_j |v^\top w_{a,j}| \to 0$. For any $\varepsilon > 0$, their Lindeberg sum is at most

$$\left(\sum_j |v^\top w_{a,j}|^2 \right) \mathbb{E}\{Z_0^2 \mathbf{1}_{\{|Z_0| > \varepsilon / b_a\}}\} \longrightarrow 0.$$

Their total variance converges to $\sigma^2 \|v\|^2 /2$. The independent triangular-array CLT applies. For the countably infinite row, truncate each row so that its remaining conditional variance is arbitrarily small, then use the finite-row theorem and Chebyshev's inequality. Cramer–Wold proves the result. Applying this argument to an arbitrary sequence $a \downarrow 0$ on the same probability-one design event gives the stated limit. □

Corollary J.1 (Non-Gaussian errors in the AR(1) examples)**.** *Replace the Gaussian errors in the preceding two positive AR(1) examples by any nondegenerate centered iid errors of finite variance, independent of the design. Every positive stationary moment still fails, the same lower-tail exponents $1/2$ and $\nu/2$ still hold (with different constants), and the normalized stationary law has a quenched weak Gaussian limit. In particular one may take bounded Rademacher errors.*

Proof. The deterministic inequality gives, with $\ell = (1-\rho)^{-1}$, $|aP_m X_{-m}| \geq a\ell |P_m|$. Substitute this coefficient into Lemma G.1 instead of using a conditional Gaussian variance. The product-moment bounds in Appendices E.2–E.4 and the tail events in Appendix E.5 give the same moment and tail conclusions. The score remains a square-integrable martingale difference. The quenched limit follows from Theorem 4.4. □

K Exact finite-history Gaussian inference

The following algebraic consequence provides central inference before stationarity. It uses a conditional Gaussian linear-model calculation. Let $A_i = I - aH_i$, $T_n = A_n \cdots A_1$, and

$$Q_0 = 0, \qquad Q_n = A_n Q_{n-1} A_n^\top + a^2 H_n.$$

For a deterministic initialization and Gaussian errors independent of the design,

$$\beta_n - T_n \beta_0 = (I - T_n)\beta^* + \varepsilon_n, \qquad \varepsilon_n \mid X_{1:n} \sim N_p(0, \sigma^2 Q_n).$$

If $Q_n \succ 0$, then for known σ^2 the set

$$\left\{ b : [\beta_n - T_n\beta_0 - (I-T_n)b]^\top Q_n^{-1} [\beta_n - T_n\beta_0 - (I-T_n)b] \leq \sigma^2 \chi^2_{p,1-\alpha} \right\}$$

has exact conditional coverage $1-\alpha$, without a burn-in assumption. It is a bounded ellipsoid when $I - T_n$ is nonsingular. The center then is

$$\widetilde{\beta}_n = (I - T_n)^{-1}(\beta_n - T_n \beta_0).$$

These facts follow by iterating the linear recursion; Q_n is the Gram matrix of the noise coefficients. Rank-one updates require $O(p^2)$ storage and work per update for T_n, Q_n, with a linear solve when reporting. Overlapping residual-variance calibration gives the asymptotic coverage established in Appendix F; a disjoint error block yields the exact finite-sample pivot below.

K.1 Exact calibration from a disjoint error block

Suppose the Gaussian error assumptions of Theorem 4.3 hold. Take $m > p$ calibration observations whose indices are disjoint from the SGD history $1{:}n$, and whose design matrix has full column rank. Let s_m^2 be their OLS residual sum of squares divided by $m - p$. Conditional on the entire design, $(m-p)s_m^2/\sigma^2 \sim \chi^2_{m-p}$ and this variable is independent of ε_n in Appendix K. Therefore, writing $w_n(b) = \beta_n - T_n\beta_0 - (I - T_n)b$, the exact pivot is

$$\frac{w_n(\beta^*)^\mathsf{T} Q_n^{-1} w_n(\beta^*)}{p s_m^2} \sim F_{p,m-p} \quad \text{conditionally on the design.} \tag{56}$$

Replacing the chi-square threshold in the known-variance set by $p s_m^2 F_{p,m-p;1-\alpha}$ gives exact conditional coverage, where $F_{d_1,d_2;u}$ is the u-quantile of the F law with degrees of freedom (d_1, d_2). Designs in the two blocks may be dependent. The proof uses the orthogonal projection decomposition of Gaussian calibration errors and the ratio of independent chi-square variables. Full-rank calibration and disjoint independent regression errors ensure the exact pivot.

L FEASIBLE CALIBRATION FOR A DEPENDENT SCORE

The general Lyapunov limit in Theorem 4.2 also permits feasible central inference. Suppose $V \succ 0$ and estimators satisfy $\widehat{H}_m \to H$ and $\widehat{\Omega}_m \to \Omega$ in probability. Define $\widehat{V}_m$ by

$$\widehat{H}_m \widehat{V}_m + \widehat{V}_m \widehat{H}_m = \widehat{\Omega}_m.$$

On the event of positive definite matrices, continuity of the Lyapunov inverse gives $\widehat{V}_m \to V$. For a stationary iterate or one satisfying Corollary 5.5, Slutsky's theorem yields

$$\frac{(\beta_{n(a)} - \beta^*)^\mathsf{T} \widehat{V}_m^{-1} (\beta_{n(a)} - \beta^*)}{a} \Rightarrow \chi^2_p, \qquad m = m(a) \to \infty.$$

The estimator and iterate may be dependent. Feasibility rests on consistent estimation of the input score's long-run covariance, which can be established directly from the data process.

For the experiment we use the residual-score Bartlett estimator (Newey & West, 1987). With OLS fit $\widetilde{\beta}_m$, put $\widehat{\xi}_i = X_i(Y_i - X_i^\mathsf{T}\widetilde{\beta}_m)$ and center these scores by their sample average. Set

$$\widehat{\Gamma}_k = \frac{1}{m} \sum_{i=k+1}^{m} (\widehat{\xi}_i - \overline{\widehat{\xi}})(\widehat{\xi}_{i-k} - \overline{\widehat{\xi}})^\mathsf{T}, \quad \widehat{\Omega}_{m,\ell} = \widehat{\Gamma}_0 + \sum_{k=1}^{\ell} \left(1 - \frac{k}{\ell+1}\right) (\widehat{\Gamma}_k + \widehat{\Gamma}_k^\mathsf{T}).$$

The fixed-parameter stable Gaussian VAR/AR input model in Appendix N has geometric dependence and all input moments. For completeness, consistency with $\ell = O(m^{1/4}) \to \infty$ follows directly. For the true score, each entry of $\xi_i \xi_{i-k}^\mathsf{T}$ has uniformly bounded variance. Its autocovariances are bounded on lags $|h| \le k$ and decay geometrically in $|h| - k$ beyond that range: the underlying Gaussian Markov process is geometrically mixing and the score products have moments of every order. Consequently each lag-k sample covariance has L^2 fluctuation $O(\sqrt{(k+1)/m})$. Summing over $k \le \ell$ gives $O(\ell^{3/2}/\sqrt{m}) = o(1)$. Absolute summability of the population score covariances removes Bartlett truncation bias by dominated convergence. Boundary bias is $O(\ell^2/m)$, also negligible.

OLS error is $O_P(m^{-1/2})$ in this model, because the sample Gram converges to $H \succ 0$ and score averages have variance $O(m^{-1})$. Cauchy–Schwarz, uniform bounded input moments, and Markov's inequality bound the sum of residual-replacement errors over all ℓ lags by $O_P(\ell/\sqrt{m} + \ell/m) = o_P(1)$. Centering the scores contributes $O_P(\ell/\sqrt{m})$ by the same bounds. This proves consistency for the stated Gaussian VAR/AR model and bandwidth growth.

M COMPARISON WITH TEMPORAL ASSUMPTIONS IN PRIOR WORK

Table 1 summarizes the objects and assumptions compared below.

Stationary Gaussian characterization. Chen et al. (2022) study a deterministic drift with additive iid centered noise of finite covariance. Under their smooth strongly convex or Hurwitz linear drift assumptions, the stationary scaled iterate has a finite second moment. Our population-drift representation has noise $\xi_i - (H_i - H)e_{i-1}$, which depends on the iterate and the temporally dependent design. Their characterization therefore supplies the Gaussian/Lyapunov framework but does not cover the infinite-moment stationary laws constructed here.

Temporal energy. Theorem 2 of Guo et al. (1997) considers a deterministic uniformly stable linear filter of a ϕ-mixing source, together with a bounded deterministic term, as specified in their equation (14). Write ζ_i for that source. Their equation (15) requires positive constants α, M, K such that

$$\mathbb{E} \exp\left\{ \alpha \sum_{j=1}^{n} \|\zeta_{i_j}\|^2 \right\} \le M e^{Kn}$$

for every $n \ge 1$ and increasing nonnegative integer sequence $i_1 < \cdots < i_n$. Within this class, their excitation condition (17) characterizes the L^p exponentially stable family of update products. The

Table 1: Different objects and assumptions in closely related theories. The surrounding text examines the temporal conditions for the refresh example.

Representative work	Setting or requirement	Result and relation to this paper
Chen et al. (2022); Dieuleveut et al. (2020)	Stationary small-step analysis	Law characterization and moment expansions, respectively; our laws may have all moments infinite.
Khusainov et al. (2025)	Iid, bounded designs; finite third score moment	Finite-history convex-distance bound; our Gaussian-error TV expansion permits dependent unbounded designs.
Guo et al. (1997); Durmus et al. (2021a)	Temporal energy or drift/growth control	Product-moment stability; our PSD argument gives pathwise contraction.
Roitershtein (2007); Damek & Mentemeier (2026)	Markov-modulated or iid affine recursions	Power-law tails; our examples combine all-moment failure with a Gaussian limit.
Mou et al. (2020)	Fixed-step averaged linear SA	CLT and correction for averages; our correction is for the stationary last iterate.
Merad & Gaïffas (2023); Wang et al. (2026)	Invariant-law convergence or quantitative SA with moment control	TV to stationarity or Wasserstein Gaussian bounds; our expansion is from stationarity to Gaussian.

direct refresh source isolates the role of joint temporal energy: retaining a single Gaussian value for n observations gives

$$\mathbb{E}e^{\alpha\sum_{i=1}^{n}X_i^2} \geq \rho^{n-1}\mathbb{E}e^{\alpha n G^2} = \infty \quad \text{when } \alpha n \geq 1/2.$$

Thus the direct refresh representation lies outside this joint-energy condition despite its Gaussian marginal. This checks the applicability of that sufficient representation, rather than treating its source condition as necessary for every representation.

Likewise, the refresh design has $\mathbb{E}(X_i \mid \mathcal{F}_{i-1}) = \rho X_{i-1}$, so it fails the conditional design martingale assumption in Shen et al. (2026). Its conditional second moment is $\rho X_{i-1}^2 + (1-\rho)$, which is unbounded, despite standard Gaussian marginals and uniform geometric ergodicity. The frameworks are therefore not nested. Wang et al. (2026) separate an iid component from a finite-state Markov component in their quantitative setting; the unbounded refresh Gram process is not covered merely by its uniform ergodicity. These distinctions locate the coexistence result within complementary theories of dependent-data SGD.

Unbounded-state product stability. Durmus et al. (2021a) allow unbounded coefficients through a super-Lyapunov drift and a matching growth bound. In their Assumption UE1, a Lyapunov function $\mathcal{V}$ obeys $P\mathcal{V} \leq e^{-c(\log \mathcal{V})^{\delta}}\mathcal{V}$ outside a sublevel set; Assumption A1 controls the matrix entries by a power of $\log \mathcal{V}$. For the refresh state x,

$$P\mathcal{V}(x) = \rho\mathcal{V}(x) + (1-\rho)\mathbb{E}\mathcal{V}(G) \geq \rho\mathcal{V}(x).$$

An unbounded $\mathcal{V}$ therefore cannot satisfy this drift: dividing by $\mathcal{V}(x)$ along levels tending to infinity would give $\rho \leq e^{-c(\log \mathcal{V}(x))^{\delta}} \to 0$. A bounded $\mathcal{V}$ cannot meet their growth condition for the unbounded Gram function x^2. Uniform geometric ergodicity thus coexists here with failure of this route to product-moment stability.

Tail results and the object being approximated. Roitershtein (2007), Assumption 1.2(A4)–(A5), use a bounded additive coefficient and a multiplicative coefficient bounded above and away from zero, together with Markov-renewal conditions, to obtain power-law tails. The refresh recursion has unbounded coefficients and a multiplier arbitrarily close to zero; its all-moment failure arises from a different regime. Damek & Mentemeier (2026) provide an irreducible-proximal treatment of iid symmetric-matrix SGD recursions, refining the justification in Gurbuzbalaban et al. (2021). Our scalar persistent constructions use direct anti-concentration and amplification bounds. Meanwhile,

Durmus et al. (2021b), Assumption A1 and Example 1, show that bounded iid matrices with sub-Gaussian additive noise can already obstruct exponential concentration. Our TV approximation controls absolute probabilities of all measurable events despite failure of every positive stationary moment.

N Experimental protocol and complete diagnostics

N.1 Replications, history lengths and numerical representation

The declared main grid contains 68 configurations. Scalar refresh designs use $\rho \in \{0, .7, .9\}$ and $a \in \{.04, .02, .01, .005, .0025\}$, with two independent seeds and 12,000 histories per seed. Refresh designs start exactly in their stationary Gaussian marginal. Their history length is

$$n(a) = \left\lceil \frac{\max\{24, 2\log^2(1/a)\}}{a} \right\rceil .$$

For the refresh design, this schedule satisfies both the weak-limit condition $an(a)/\log(1/a) \to \infty$ and the sharper finite-history condition in Corollary 5.6. A zero-initialized paired recursion starts halfway through each history and uses the same final observations. The paired histories measure initialization sensitivity of the reported central probabilities. Rates and seeds use independent histories, except for the explicit paired histories. Pareto and baseline vector comparisons use learning time 24 as a finite approximation, with paired histories of learning time 12; refresh, log-Weibull and feasible-vector experiments use growing learning times.

The largest paired grid discrepancy among individual scalar configurations was 2.11×10^{-5} for the Pareto design at $\rho = .7, a = .04$; among individual refresh seeds it was 7.35×10^{-7}. All per-configuration paired diagnostics are saved.

For independent Gaussian errors with variance one, propagate the exact finite-history variance and the coefficient of initialization:

$$Q_i = (1 - aX_i^2)^2 Q_{i-1} + a^2 X_i^2, \quad Q_0 = 0, \qquad T_i = (1 - aX_i^2) T_{i-1}, \quad T_0 = 1. \tag{57}$$

The simulation stores $\log Q_i$ via log-add-exp, the sign and log magnitude of T_i, and the analogous finite sum of $a^2 X_i^4$. It never clips a coefficient or deletes a replication. For $v_r = Q_r/a$, conditional integration gives

$$\widehat{F}_a(x) = R^{-1} \sum_{r=1}^{R} \Phi(x/\sqrt{v_r}), \qquad \widehat{C}_{1-\alpha} = R^{-1} \sum_{r} \left[2\Phi \left\{ z_{1-\alpha/2} \sqrt{v_{\mathrm{ref}}/v_r} \right\} - 1 \right],$$

where $v_{\mathrm{ref}} = 1/2$ or $(1 + aJ)/2$. Monte Carlo standard errors are sample standard deviations of these bounded conditional probabilities divided by $\sqrt{R}$. Saved summaries include 90%, 95%, and 99% coverage, widths, expansion counts, startup coefficients, and quantiles of the conditional correction. The grid CDF uses float64 evaluation. Bootstrap resampling operates on centered float32 CDF values for memory efficiency, over the full 1,601-point grid; the reported 95% band is the 95th percentile of the resampled suprema. It quantifies Monte Carlo error on the finite grid only. Histories are resampled as units, so the Gaussian errors are not spuriously treated as extra replications.

The scalar Kolmogorov first-order constant follows from Theorem 5.2, which implies that the corrected CDF is uniformly $o(a)$ away from the true CDF. Differentiating $\Phi(x/\sqrt{(1 + aJ)/2})$ gives a maximal absolute derivative $J\phi(1)/2$, hence $3\phi(1)/4$ in the refresh example. Saved finite-grid CDF diagnostics supplement the direct TV estimates plotted in Figure 1.

N.2 Direct TV integration and conditional correction

Let $v_r = Q_n^{(r)}/a$ for the $R = 24000$ refresh histories, and define

$$\widehat{f}_a(x) = R^{-1} \sum_{r=1}^{R} \frac{1}{\sqrt{v_r}} \phi(x/\sqrt{v_r}), \qquad \widehat{d}_{\mathrm{TV}} = \frac{1}{2} \int_{\mathbb{R}} |\widehat{f}_a(x) - g_a(x)| \, dx,$$

Table 2: Refresh design, $\rho = .7$, 24,000 histories. Coverage is in percent. TV estimates and the linearized Monte Carlo standard error of corrected TV are in units of 10^{-4}. The empirical conditional Gaussian mixture defines the estimated finite-history marginal law. Full outputs also include coverage standard errors, seed-wise estimates and CDF diagnostics.

a	Base coverage	Corrected coverage	TV	Corrected TV	MCSE (corr.)
0.04	94.3885	95.0766	123.130	17.859	0.469
0.02	94.6849	95.0293	65.035	6.488	0.199
0.01	94.8373	95.0095	33.976	2.050	0.084
0.005	94.9167	95.0027	17.492	0.588	0.035
0.0025	94.9579	95.0009	8.872	0.185	0.013

where g_a is the baseline or corrected Gaussian density. This is a Monte Carlo estimate of the finite-history marginal-law TV. For a reference variance v_*, the density ratio as a function of $z = x^2$ is

$$\widehat{f}_a(\sqrt{z})/g_a(\sqrt{z}) = R^{-1}\sum_r \sqrt{v_*/v_r}\exp\{z(1/v_* - 1/v_r)/2\}.$$

Its logarithm is convex, so the densities have at most two positive crossings unless they coincide identically. We locate the minimum through the mixture's posterior mean precision and bracket crossings in $\log(x/\sqrt{v_*})$, using log variances throughout. Gaussian CDF differences integrate each sign interval. The omitted positive-tail integral is bounded by the sum of the mixture and reference tail probabilities; all reported bounds are below 10^{-13}. Analytic Gaussian comparisons, independent adaptive quadrature, boundary-crossing cases and extreme log-variance checks verify the implementation. Every positive-weight history is retained.

For Monte Carlo uncertainty, integrate each component-minus-reference density over the fitted sign intervals. The sample standard deviation of these bounded per-history contributions divided by $\sqrt{R}$ is the linearized standard error. In the $\rho = .7$ endpoint configurations, 199 whole-history bootstrap resamples re-estimate crossings and TV; their standard deviations agree with the linearized errors. Separate-seed estimates and paired half-history checks are supplied. Finite-history and Monte Carlo approximations are documented separately from numerical integration error.

Convexity implies $\mathbb{E}d_{\mathrm{TV}}(\widehat{f}_a, g_a) \geq d_{\mathrm{TV}}(f_{a,n}, g_a)$, where $f_{a,n} = \mathbb{E}\widehat{f}_a$ is the finite-history marginal density. To assess sampling sensitivity, we also integrate the distance between the two independent 12,000-history seed mixtures. At $\rho = .7$, these split distances are 4.41×10^{-5} at $a = .01$ and 2.40×10^{-6} at $a = .0025$, compared with pooled corrected TV estimates 2.05×10^{-4} and 1.85×10^{-5}. Across all 15 refresh configurations, the split distance is .009–.215 times the corrected TV estimate. Adaptive quadrature and a separate partitioned integration agree within 10^{-10}; all results are saved. These are descriptive sampling diagnostics, distinct from the paired checks of history length; a single split does not bound the plug-in bias.

The finite-history conditional correction obeys

$$\mathcal{R}_{a,i} = (1 - aX_i^2)^2\mathcal{R}_{a,i-1} + a^2X_i^4, \qquad \mathcal{R}_{a,0} = 0.$$

Its median and quantiles track concentration toward $J = 1.5$ without averaging a stationary iterate or its square. Table 3 reports the fraction farther than .5 from J. Histories are independently generated across step sizes, so these are distributional diagnostics of the conditional expansion. Together with the propagated T_i and paired histories, this separates loss of initialization from concentration of $\mathcal{R}_{a,n}/a$ near J. Increasing history length addresses the former; the latter is a small-step property of the design law. Neither empirical diagnostic replaces the asymptotic conditions of Corollary 5.6.

N.3 Pareto and log-Weibull AR designs

For the raw design $\widetilde{X}_i = \rho\widetilde{X}_{i-1} + \eta_i$, normalize by its population second moment. The Pareto innovation satisfies $\mathbb{P}(\eta > u) = u^{-3}$, $u \geq 1$, giving

$$h_{\mathrm{raw}} = \frac{3/4}{1-\rho^2} + \left(\frac{3/2}{1-\rho}\right)^2, \qquad X_i = \widetilde{X}_i/\sqrt{h_{\mathrm{raw}}}.$$

Table 3: Finite-history conditional correction $\mathcal{R}_{a,n}/a$ in the refresh design, 24,000 histories per row. Quantiles describe between-history variation; $J = 1.5$ for every persistence.

ρ	a	10th percentile	Median	90th percentile	$\widehat{P}(\lvert\mathcal{R}_{a,n}/a - 1.5\rvert > .5)$
0	0.04	0.975	1.414	2.389	0.3043
0	0.02	1.078	1.442	2.102	0.1746
0	0.01	1.176	1.466	1.941	0.0897
0	0.005	1.254	1.476	1.811	0.0381
0	0.0025	1.318	1.488	1.719	0.0091
0.7	0.04	0.653	1.095	2.207	0.5507
0.7	0.02	0.786	1.202	2.110	0.4276
0.7	0.01	0.913	1.285	2.049	0.2943
0.7	0.005	1.030	1.354	1.967	0.1669
0.7	0.0025	1.135	1.403	1.867	0.0777
0.9	0.04	0.421	0.838	2.033	0.7159
0.9	0.02	0.538	0.959	2.004	0.6342
0.9	0.01	0.674	1.074	2.010	0.5298
0.9	0.005	0.811	1.189	2.009	0.3962
0.9	0.0025	0.937	1.292	1.969	0.2536

Table 4: Selected Pareto and log-Weibull configurations, 12,000 histories each. The parameter is ρ for Pareto and c for log-Weibull. Coverage is in percent. An expansion path contains at least one factor with $|1 - aX_i^2| > 1$ anywhere in its history. Pareto-3 does not meet the correction theorem's fourth design moment, so no corrected result is assigned. Every configuration is retained in the supplied results.

Model	Parameter	a	Coverage	Corrected coverage	Expansion paths
pareto	0	0.04	93.1154	–	3432
pareto	0	0.01	94.0687	–	1829
pareto	0	0.0025	94.5519	–	941
pareto	0.7	0.04	94.6931	–	226
pareto	0.7	0.01	94.9047	–	87
pareto	0.7	0.0025	94.9703	–	36
pareto	0.9	0.04	94.7574	–	10
pareto	0.9	0.01	94.9384	–	3
pareto	0.9	0.0025	94.9841	–	2
logweibull	0.5	0.04	94.7422	95.0008	0
logweibull	0.5	0.01	94.9358	95.0003	0
logweibull	0.5	0.0025	94.9839	95.0000	0
logweibull	1	0.04	94.5327	97.8303	967
logweibull	1	0.01	94.8405	95.9553	631
logweibull	1	0.0025	94.9314	95.2344	350

We use $\rho \in \{0, .7, .9\}$ and the same five steps, with 12,000 independent replications per configuration. For log-Weibull inputs, take $\rho = .7$, $\eta = \exp(cE^{1/\nu})$, $\nu = 1.5$, $c \in \{.5, 1\}$ and $E \sim \mathrm{Exp}(1)$, at steps $.04, .01, .0025$, also with 12,000 replications. The tail is $\exp[-(\log u/c)^{\nu}]$, so the no-moment argument holds at both scales. Population innovation moments are computed by scaled adaptive quadrature of

$$\mathbb{E}\eta^k = \int_0^\infty \exp(kct^{1/\nu} - t)\, dt.$$

Raw AR moments M_k obey $(1 - \rho^k)M_k = \sum_{j=0}^{k-1} \binom{k}{j} \rho^j M_j \mathbb{E}\eta^{k-j}$, $M_0 = 1$. Thus the normalized correction is $J = M_4/(2M_2^2)$. Quadrature tolerances and errors are saved. AR designs start at their population mean and warm up for 512 steps; their omitted initial contribution is attenuated by ρ^{512}. No normalization is fitted to simulated paths.

Table 5: Dependent-score ellipse coverage in percent, with Monte Carlo standard errors in parentheses, from 6,000 replications. HAC column headings give calibration size and bandwidth. All runs and all bandwidth choices are in the saved summaries.

a	Oracle	Lag zero	HAC $(2048, 13)$	HAC $(8192, 9)$	HAC $(8192, 19)$
0.04	94.92 (0.28)	77.60 (0.54)	93.60 (0.32)	93.55 (0.32)	94.12 (0.30)
0.01	95.35 (0.27)	76.62 (0.55)	93.92 (0.31)	93.82 (0.31)	94.50 (0.29)
0.0025	95.23 (0.28)	76.17 (0.55)	93.60 (0.32)	93.78 (0.31)	94.38 (0.30)

N.4 Noncommuting design and serial score covariance

The vector experiment uses

$$\widetilde{X}_i = B\widetilde{X}_{i-1} + \eta_i, \quad B = \begin{pmatrix} .6 & .15 \\ 0 & .4 \end{pmatrix}, \quad \mathrm{Cov}(\eta_i) = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}, \quad Z_i = .7Z_{i-1} + \sqrt{1 - .7^2}\,\zeta_i, \tag{58}$$

where Gaussian innovations are independent between design and errors, and ζ_i is standard normal. Solve $H_{\mathrm{raw}} = BH_{\mathrm{raw}}B^{\mathsf{T}} + \mathrm{Cov}(\eta_i)$ and rescale $X_i = s\widetilde{X}_i$, $s = \sqrt{2/\operatorname{tr} H_{\mathrm{raw}}}$. Initialize both input states exactly in their stationary Gaussian laws. With $A = (I - .7B)^{-1}$, independence gives

$$\Omega = AH + HA^{\mathsf{T}} - H, \qquad HV + VH = \Omega. \tag{59}$$

The population matrices are

$$H \simeq \begin{pmatrix} 1.18802063 & .33333890 \\ .33333890 & .81197937 \end{pmatrix}, \qquad \Omega \simeq \begin{pmatrix} 3.07622956 & .90851551 \\ .90851551 & 1.44351887 \end{pmatrix},$$
$$V \simeq \begin{pmatrix} 1.26604476 & .10208074 \\ .10208074 & .84698206 \end{pmatrix}.$$

The baseline vector runs use all five steps and 12,000 independent replications, with zero SGD initialization and $\lceil 24/(a\lambda_{\min}(H)) \rceil$ updates. The feasible-HAC runs use steps $.04, .01, .0025$ and 6,000 replications. Each processes 8,192 calibration observations, followed by $\lceil \max\{24, 2\log^2(1/a)\}/(a\lambda_{\min}(H)) \rceil$ observations before recording the endpoint. SGD processes both blocks; the calibration score estimates use only the earlier block. The last 2,048 or all 8,192 calibration observations are used with $\ell = \lfloor c_m m^{1/4} \rfloor$, $c_m \in \{.5, 1, 2\}$, yielding $(3, 6, 13)$ or $(4, 9, 19)$. Matrices are symmetrized only through their defining algebra; no eigenvalue clipping or oracle replacement is used. A non-positive estimate is counted as a failed, empty confidence set.

For the normalized endpoint error $u = (\beta_n - \beta^*)/\sqrt{a}$, the oracle statistic is $u^{\mathsf{T}}V^{-1}u$ and the lag-zero statistic is $2\|u\|^2$, compared with $\chi^2_{2,.95}$. For a Gaussian limit the latter has coverage

$$\frac{1}{2\pi}\int_0^{2\pi} \left[1 - \exp\{-\chi^2_{2,.95}/(4v(\theta))\}\right] d\theta, \quad v(\theta) = (\cos\theta, \sin\theta)V(\cos\theta, \sin\theta)^{\mathsf{T}}.$$

A deterministic 100,000-angle quadrature gives .760097279998, quantifying the central miscalibration caused by ignoring serial score covariance.

To isolate finite-bandwidth tapering, define the population Bartlett benchmark

$$\Gamma_k = (.7B)^k H, \quad \Omega_\ell = H + \sum_{k=1}^{\ell} \left(1 - \frac{k}{\ell + 1}\right)(\Gamma_k + \Gamma_k^{\mathsf{T}}), \qquad HV_\ell + V_\ell H = \Omega_\ell. \tag{60}$$

Let λ_1, λ_2 be the generalized eigenvalues of (V, V_ℓ). For $U \sim N_2(0, V)$, the coverage of $U^{\mathsf{T}}V_\ell^{-1}U \le \chi^2_{2,.95}$ is

$$C_\ell = \frac{1}{2\pi}\int_0^{2\pi} \left[1 - \exp\left\{-\frac{\chi^2_{2,.95}}{2(\lambda_1\cos^2\theta + \lambda_2\sin^2\theta)}\right\}\right] d\theta. \tag{61}$$

Adaptive quadrature gives $C_9 = .936597378164$ and $C_{19} = .943743069940$. Figure 2(b) plots this population Gaussian reference alongside the feasible estimates. It holds the bandwidth fixed and uses the known model to quantify tapering; finite-sample covariance estimation and finite-step approximation remain separate sources of error. The exact matrices, eigenvalues and quadrature errors for $\ell = 3, \ldots, 19$ are supplied with the outputs.

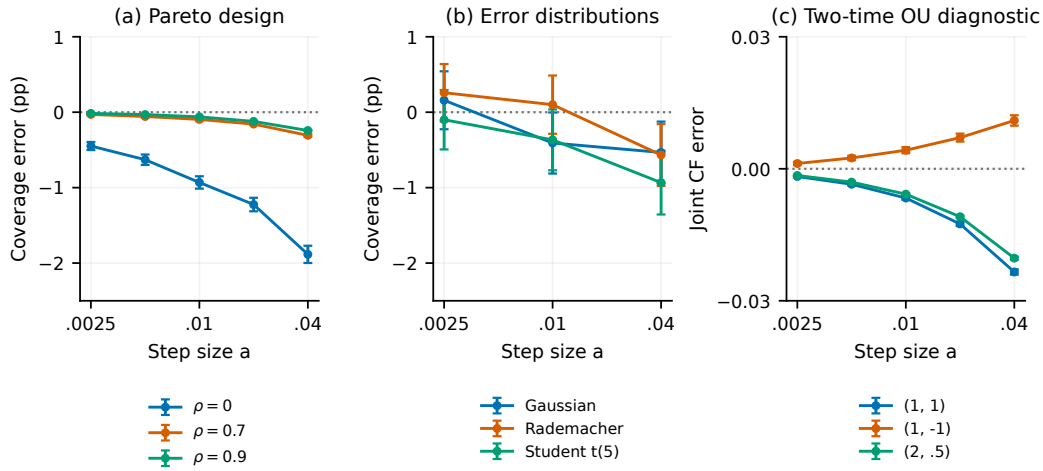


Figure 3: Qualitative stationary and process predictions. (a,b) Coverage error relative to 95%, in percentage points, on the same fixed axis range: Pareto AR designs and actual Gaussian/Rademacher/Student-t_5 errors, respectively. (c) Bounded two-time characteristic-function errors relative to the OU prediction, with a zero reference. Bars show 1.96 Monte Carlo standard errors. Log-Weibull results appear in Tables 4 and 7.

N.5 Non-Gaussian errors and bounded process diagnostics

At $\rho = .7$ and $a \in \{.04, .01, .0025\}$, actual scalar SGD is run with Gaussian, Rademacher, and variance-one Student-t_5 errors, with 12,000 independent replications each. Signed-log addition represents the state without clipping, including cancellation between opposite signs. These distributions test the finite-variance conditional CLT through marginal coverage. Actual-error coverage estimates have Bernoulli Monte Carlo standard errors.

For an OU diagnostic, save two endpoints separated by $k = \lceil .5/a \rceil$ updates, with learning-time lag $\tau = ak$. Conditional on a Gaussian-error design history, write $U_{\text{new}} = T_k U_{\text{old}} + \varepsilon$ with conditionally independent new noise. For $(s, t) \in \{(1, 1), (1, -1), (2, .5)\}$, evaluate

$$\mathbb{E}[e^{i(sU_{\text{old}} + tU_{\text{new}})} \mid X] = \exp\{-\tfrac{1}{2}[(s + tT_k)^2 v_{\text{old}} + t^2 v_{\text{noise}}]\}.$$

The sum of nonnegative variance terms is evaluated in log space, so rare large variances contribute their correct bounded limiting value zero. The OU target is $\exp[-(s^2 + t^2 + 2ste^{-\tau})/4]$. This bounded diagnostic targets two-time distributions of the OU limit without requiring stationary covariance moments.

N.6 Observed inference before burn-in

Set $\beta^* = 0$, $\beta_0 = 1$ and use the refresh design with $\rho = .7$. For each of $a = .04, .01$, simulate 48,000 independent design/error trajectories and record endpoints near learning times $an = .05, .25, 1, 4, 24$. Before SGD starts, generate 65 calibration observations on the same design stream, fit scalar OLS and compute its residual variance s^2 with 64 degrees of freedom. Regression errors are actual independent standard Gaussian draws in both blocks. The calibration block and SGD history have disjoint errors and potentially dependent designs.

All methods use the same endpoint and design history. Alongside SGD, propagate T_n, Q_n from (57), and put $w_n = \beta_n - T_n\beta_0$. The stationary interval uses center β_n and radius $z_{.975}\sqrt{a/2}$. The history-corrected center is $w_n/(1 - T_n)$, with radius $z_{.975}\sqrt{Q_n}/|1 - T_n|$ for known variance, or $t_{64,.975}s\sqrt{Q_n}/|1 - T_n|$ for estimated variance, where $t_{d,u}$ is the u-quantile of Student's t law with d degrees of freedom. We record actual coverage indicators and interval widths. Figure 4 shows $a = .01$, and Table 6 reports both step sizes. Endpoint estimates share trajectories across times; error bars use independent trajectories as the replication unit.

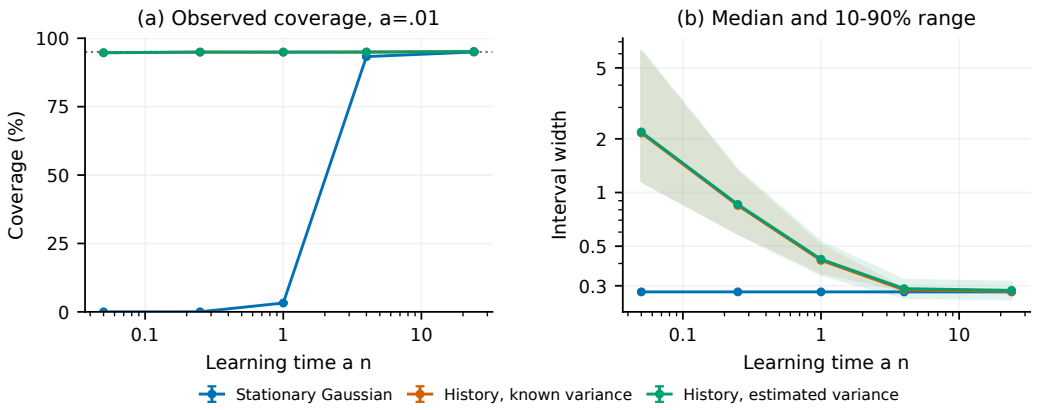


Figure 4: Observed calibration and precision before burn-in, $a = .01$, 48,000 trajectories. (a) Actual 95% coverage with 1.96 Monte Carlo standard-error bars; the dotted line is nominal. Both history-corrected curves are measured from simulated outcomes. (b) Median interval widths and 10th–90th percentile bands across the same trajectories, on a logarithmic scale. Estimated-variance intervals use an earlier block of 65 observations.

Table 6: Actual finite-history inference, 48,000 trajectories per step size. Coverage is in percent with Monte Carlo standard errors in parentheses. Width columns describe the estimated-variance history-corrected interval.

a	an	Naive	Known σ^2	Estimated σ^2	Median width	90th width
0.04	0.08	0.12 (0.02)	95.03 (0.10)	95.06 (0.10)	3.811	16.863
0.04	0.28	2.86 (0.08)	94.86 (0.10)	94.97 (0.10)	1.781	4.308
0.04	1	29.89 (0.21)	95.04 (0.10)	95.02 (0.10)	0.884	1.344
0.04	4	93.05 (0.12)	95.08 (0.10)	95.01 (0.10)	0.594	0.669
0.04	24	94.40 (0.10)	95.01 (0.10)	95.08 (0.10)	0.577	0.645
0.01	0.05	0.00 (0.00)	94.69 (0.10)	94.72 (0.10)	2.186	6.242
0.01	0.25	0.02 (0.01)	94.90 (0.10)	94.97 (0.10)	0.857	1.342
0.01	1	3.23 (0.08)	94.90 (0.10)	94.94 (0.10)	0.423	0.529
0.01	4	93.30 (0.11)	94.98 (0.10)	94.93 (0.10)	0.289	0.323
0.01	24	94.96 (0.10)	95.13 (0.10)	95.13 (0.10)	0.283	0.316

N.7 An imposed shock and its amplification

For the normalized Pareto AR design with $\rho = .7$, $a = .01$, start at the population mean, impose one innovation $u \in [10^2, 10^7]$, and use the innovation mean for the next 599 updates. Record the maximum partial log product, including the empty product. With $s = h_{\text{raw}}^{-1/2}$ and

$$m = 1 + \left\lfloor \frac{\log(\sqrt{a/2}\, su)}{\log(1/\rho)} \right\rfloor,$$

the analytic lower bound is $m \log(a/2) + m(m-1) \log \rho + 2m \log(su)$. The positive background increases factor magnitudes over the selected expansion interval. This deterministic construction isolates the amplification mechanism used in the tail proof.

N.8 Additional log-Weibull stress checks

After the declared main grid revealed overcorrection at $c = 1$, we fixed two additional configurations: 48,000 independent histories at $a = .0025$ and 24,000 at $a = .000625$, with new seeds 940001 and 940002. They retain the same normalization, $\rho = .7$, $\nu = 1.5$, and growing histories. These checks assess sampling precision and the effect of a smaller learning rate. The results are summarized in the generated stress table.

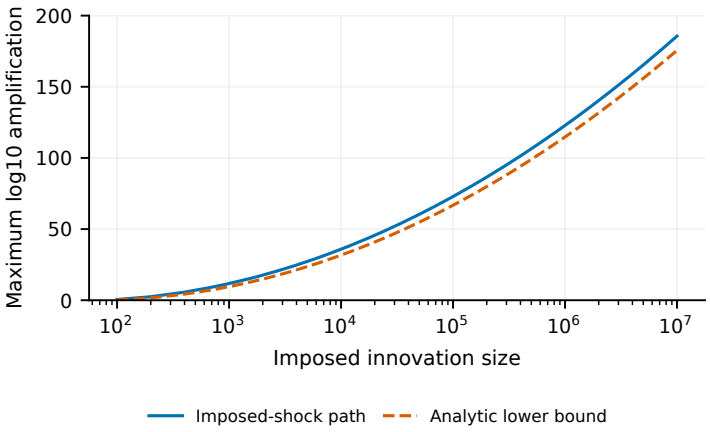


Figure 5: Persistence amplifies one extreme innovation. The deterministic path and analytic lower bound describe amplification conditional on an imposed shock; the vertical axis is its base-10 logarithm.

Table 7: Log-Weibull stress checks at $c = 1$. Coverage is reported as a percentage, with Monte Carlo standard errors in percentage points.

a	Histories	Baseline coverage	Corrected coverage	MCSE (baseline, corrected)
0.0025	48000	94.9426	95.2455	(0.0026, 0.0026)
0.000625	24000	94.9772	95.0545	(0.0031, 0.0031)

The smaller step reduces both coverage errors. At $a = .000625$, baseline and corrected errors are approximately -0.0228 and $+0.0545$ percentage points, respectively. The population correction is $J \simeq 10.8553$, whereas the sampled median of $\mathcal{R}_{a,n}/a$ is about 1.44 and its 99th percentile is about 25.53. The sampled conditional correction remains dispersed relative to its limiting coefficient, consistent with delayed entry into the first-order regime. Over the tested steps, the baseline interval remains closer to nominal coverage in the $c = 1$ setting.

Reproducibility artifacts. Configuration files and the README specify all seeds and software versions. Saved arrays, execution metadata and algebraic checks accompany the scripts used to generate every table and figure.