

Sketch Placement Matters: Equal Stability and Different Bias in Gaussian-Target Langevin Dynamics

Jay Paek

Abstract

Unbiased estimates of the same Langevin drift can produce different stationary sampling errors. For a Gaussian target, we compare PQx with $Q^{1/2}PQ^{1/2}x$, where P is a fresh Gaussian sketch Gram matrix at each step. The methods have identical mean-square stability regions and expected quadratic energies, but their stationary variance profiles cross at an explicit spectral threshold. For $Q_\mu = \text{diag}(\mu, 1, \dots, 1)$ at fixed dimension, sketch size, and admissible positive step size, their non-Gaussian invariant laws have Euclidean W_2 errors $\Theta(\mu^{-1/2})$ and $\Theta(1)$ as $\mu \downarrow 0$. A covariance lower bound and a stationary coupling to ordinary Langevin sampling establish this separation. Simulations confirm the predicted variance profiles.

1 Introduction

Langevin samplers often replace an expensive gradient with an unbiased random estimate. The resulting state-dependent noise changes the stationary distribution. We study how the placement of a random sketch allocates this error across the target’s eigendirections.

We examine this question on a Gaussian target with precision $Q \succ 0$. At each step, draw $S \in \mathbb{R}^{m \times d}$ with independent $\mathcal{N}(0, 1/m)$ entries and set $P = S^\top S$, so $\mathbb{E}P = I$. The two drift estimates are

$$g_A(x) = PQx, \quad g_B(x) = Q^{1/2}PQ^{1/2}x. \quad (1)$$

Model A compresses the completed gradient, while B sketches a quadratic factor before reassembling the gradient. Their means are both Qx . Their conditional noise covariances, however, are

$$\begin{aligned} \text{Cov}(g_A(x)) &= \frac{1}{m} [Qxx^\top Q + \|Qx\|^2 I], \\ \text{Cov}(g_B(x)) &= \frac{1}{m} [Qxx^\top Q + (x^\top Qx)Q]. \end{aligned} \quad (2)$$

In the first case, the squared force excites every coordinate through an isotropic term. In the second, the additional excitation is shaped by Q , reducing its effect in flat directions. Thermal noise keeps the state moving at stationarity, so this difference persists.

The two placements share stability and expected energy but redistribute stationary variance across eigendirections. For $Q_\mu = \text{diag}(\mu, 1, \dots, 1)$ at a fixed admissible step, this yields stationary Euclidean W_2 errors of orders $\mu^{-1/2}$ and one. The separation concerns absolute error in a target whose width grows as $\mu \downarrow 0$; it depends on the metric and does not imply that B is universally more accurate.

Relation to existing results. Exact finite-step quadratic covariance identities appear in Liu et al. [5]. The matrix-learning-rate analysis of Ziyin et al. [11] recovers Model A’s covariance after a change of variables and matching second moments; Wang et al. [10, Theorem B.2] gives the Gaussian-SGD stability condition used below. Gaussian-target SGLD invariant covariances are studied by Brosse et al. [1].

Compressed-gradient Langevin methods and their variance reduction are studied in QLSD [9]; randomized-coordinate Langevin and variance reduction are analyzed by Ding and Li [2]. Random-direction zeroth-order sampling has related

estimators [8]. Subspace Langevin Monte Carlo [6] projects both the drift and thermal noise; its update is therefore different from either model here. Here we compare two placements of the same sketch and relate their directional variance profiles to transport errors of the invariant laws.

Paulin and Whalley [7, Theorem 1] give step-size bias bounds under moment assumptions. Their sufficient step-size condition (5) requires $h \leq \min\{1/\beta, \alpha/(\alpha^2 + 2C_G)\}$, where α, β bound the target Hessian and C_G is a mean-square Lipschitz constant for the gradient error. On our family, $\alpha = \mu$, $\beta = 1$, and (2) gives $C_{G,A} = (d+1)/m$ and $C_{G,B} = (d+\mu)/m$. That condition therefore shrinks with μ and does not directly cover our fixed-step separation. Our comparison resolves this fixed-step conditioning dependence.

2 Models and the common stability region

Let Q be a real symmetric positive definite $d \times d$ matrix, with eigenvalues $0 < \lambda_1 \leq \dots \leq \lambda_d = L$. The target is

$$\pi_Q = \mathcal{N}(0, Q^{-1}), \quad V(x) = \frac{1}{2}x^\top Qx.$$

Fix an integer $m \geq 1$ and $\eta > 0$. Draw independent sketches $S_k \in \mathbb{R}^{m \times d}$ with independent $\mathcal{N}(0, 1/m)$ entries and independent thermal noises $\xi_k \sim \mathcal{N}(0, I_d)$. The two sequences are mutually independent and independent of the initial state. Set $P_k = S_k^\top S_k$. We compare

$$\begin{aligned} X_{k+1}^A &= (I - \eta P_k Q) X_k^A + \sqrt{2\eta} \xi_k, \\ X_{k+1}^B &= (I - \eta Q^{1/2} P_k Q^{1/2}) X_k^B + \sqrt{2\eta} \xi_k. \end{aligned} \quad (3)$$

Both use full isotropic thermal noise. The Gram matrix P_k is not an orthogonal projection; every finite $m \geq 1$ is allowed. For a random update matrix M_k , *mean-square stability* means $\mathbb{E} \|M_{k-1} \cdots M_0 x\|^2 \rightarrow 0$ for every deterministic x . Write $\mathcal{P}_2$ for probability laws with finite second moment. Define

$$\begin{aligned} a &= 1 + 1/m, & t &= \eta/(2m), & b_i &= 1 - \eta a \lambda_i / 2, \\ D &= \sum_i \frac{\lambda_i}{b_i}, & H &= \sum_i \frac{1}{b_i}, & z &= 1 - tD. \end{aligned} \quad (4)$$

The quantities involving b_i^{-1} will be used only when all $b_i > 0$.

The covariance and stability formulas specialize the Gaussian-SGD results cited above.

Theorem 2.1 (Covariance and stability). *For either model in (3), mean-square stability is equivalent to existence of an invariant law in $\mathcal{P}_2$, and holds exactly when*

$$\eta < \frac{2}{aL}, \quad tD = \sum_i \frac{\eta \lambda_i}{2m - (m+1)\eta \lambda_i} < 1. \quad (5)$$

In this region, the invariant law in $\mathcal{P}_2$ is unique and centered. Denote the laws by ν_A, ν_B . Their covariances commute with Q and have eigenvalues

$$c_i^A = \frac{1}{z \lambda_i b_i}, \quad c_i^B = \frac{1 + t \lambda_i H / z}{\lambda_i b_i}. \quad (6)$$

Outside (5), there is no invariant law in $\mathcal{P}_2$. This last statement does not rule out invariant laws with infinite second moment.

Proof. Gaussian fourth moments give, for symmetric C ,

$$\mathbb{E}[PCP] = aC + m^{-1} \text{tr}(C)I. \quad (7)$$

Indeed, write $P = m^{-1} \sum_{r=1}^m u_r u_r^\top$ with independent standard Gaussian u_r . The terms with unequal row indices contribute $(1 - 1/m)C$ and the equal-index terms contribute $m^{-1}(2C + \text{tr}(C)I)$. This also proves (2).

Let $B_0 = I - \eta Q$. The homogeneous second-moment operators are

$$\begin{aligned}\mathcal{T}_A(C) &= B_0 C B_0 + \frac{\eta^2}{m} [Q C Q + \text{tr}(Q C Q) I], \\ \mathcal{T}_B(C) &= B_0 C B_0 + \frac{\eta^2}{m} [Q C Q + \text{tr}(Q C) Q].\end{aligned}\tag{8}$$

The noisy recursion adds $2\eta I$. Work in an eigenbasis of Q , and put $g_i = (1 - \eta\lambda_i)^2 + \eta^2\lambda_i^2/m$. The diagonal second moments in Model B evolve under

$$K_B = \text{diag}(g_i) + \frac{\eta^2}{m} \lambda \lambda^\top, \quad \lambda = (\lambda_1, \dots, \lambda_d)^\top.\tag{9}$$

This diagonal recursion closes for every initial second moment, not just diagonal ones. Consequently, $\rho(K_B) < 1$ is equivalent to full mean-square stability: the diagonal controls the trace of every positive semidefinite second moment. Conversely, testing the coordinate-vector initial states gives decay of every column of the nonnegative matrix K_B^n .

The matrix K_B is symmetric positive definite. Thus $\rho(K_B) < 1$ is equivalent to $I - K_B \succ 0$. Since

$$I - K_B = G - \frac{\eta^2}{m} \lambda \lambda^\top, \quad G = \text{diag}(2\eta\lambda_i b_i),$$

positive definiteness first requires every $b_i > 0$. Factoring by $G^{1/2}$ then gives the equivalent rank-one condition

$$\frac{\eta^2}{m} \lambda^\top G^{-1} \lambda = \frac{\eta}{2m} \sum_i \frac{\lambda_i}{b_i} < 1.$$

The two update matrices satisfy $M_B = Q^{1/2} M_A Q^{-1/2}$ pointwise, so their products are conjugate and their mean-square stability regions coincide.

Strict stability gives geometric decay of product second moments in finite dimension: there are $C < \infty$ and $r < 1$ such that $\mathbb{E} \|M_{n-1} \cdots M_0\|_F^2 \leq C r^n$. This follows from (9), conjugacy, and $\mathbb{E} \|M_{n-1} \cdots M_0\|_F^2 = \text{tr}(\mathcal{T}^n(I))$. On two-sided independent innovations, the series

$$X_0 = \sqrt{2\eta} \sum_{j=0}^{\infty} M_{-1} \cdots M_{-j} \xi_{-j-1}\tag{10}$$

converges in L^2 ; the product for $j = 0$ is I . Conditional on all sketches, its summands are orthogonal centered Gaussian vectors. Their squared L^2 norms sum to $2\eta \sum_{j \geq 0} \text{tr}(\mathcal{T}^j(I)) < \infty$ by the geometric bound. Its time shifts construct an invariant law in $\mathcal{P}_2$. A shared-innovation coupling contracts differences in L^2 , establishing uniqueness within $\mathcal{P}_2$.

Conversely, an invariant second moment has positive diagonal c satisfying $K_B c = c - 2\eta \mathbf{1} < c$ in Model B. The weighted maximum norm with weights c_i implies $\rho(K_B) \leq \max_i (K_B c)_i / c_i < 1$. For A the diagonal matrix is $K_A = \text{diag}(g_i) + (\eta^2/m) \mathbf{1}(\lambda_i^2)_i^\top$; it is similar to K_B by $\text{diag}(\lambda_i)$, so the same conclusion holds. Stationarity of the mean gives $\eta Q \mathbb{E} X = 0$.

Coordinate-sign symmetries preserve Q , the Gaussian sketch, and the thermal noise in its eigenbasis. Uniqueness therefore makes the stationary covariance diagonal in that basis. Let $T = \sum_j \lambda_j^2 c_j^A$. The stationary A equations give $\lambda_i b_i c_i^A = 1 + tT$, whence $T = (1 + tT)D$ and $1 + tT = z^{-1}$. For B let $R = \sum_j \lambda_j c_j^B$. Its equations give $\lambda_i b_i c_i^B = 1 + t\lambda_i R$ and $R = H + tDR = H/z$. These identities yield (6). $\square$

Exact relationship to Gaussian SGD. For A, set $w = QX$. Then $w_{k+1} = (I - \eta Q P_k) w_k + \sqrt{2\eta} Q \xi_k$. Its second-moment recursion matches Gaussian least-squares SGD with feature covariance I , learning-rate matrix ηQ , batch size m , zero momentum, and label-noise variance $2m/\eta$. Substitution in Ziyin et al. [11, Theorem 2] recovers $Q C_A Q$ and hence (6) for A. The innovation distributions differ, so this is second-moment equivalence, not equality of Markov kernels. The reduction and the existing stability result are the reason Theorem 2.1 is explicitly treated as background.

3 An exact comparison of stationary uncertainty

The common stability region does not determine how the invariant covariances allocate variance.

Theorem 3.1 (Spectral crossing and equal energy). *Assume (5) and define $\lambda_* = D/H$ and $J = \sum_i (\lambda_i b_i)^{-1}$. Then*

$$c_i^A - c_i^B = \frac{t(D - \lambda_i H)}{z \lambda_i b_i}, \quad (11)$$

$$\text{tr}(QC_A) = \text{tr}(QC_B) = H/z, \quad (12)$$

$$\text{tr}(C_A - C_B) = \frac{t}{z}(DJ - H^2) \geq 0. \quad (13)$$

Thus A has greater variance when $\lambda_i < \lambda_*$ and B when $\lambda_i > \lambda_*$. For finite m and positive η , equality in (13) holds exactly when the spectrum is constant.

Proof. Subtract (6) and use $1 - z = tD$ to obtain (11). Summing after multiplication by λ_i gives (12); summing without that multiplication gives (13). Finally,

$$H = \sum_i \sqrt{\lambda_i/b_i} \sqrt{1/(\lambda_i b_i)} \leq \sqrt{DJ}.$$

Equality in Cauchy–Schwarz holds precisely when all λ_i are equal. The crossing value is a weighted average of the eigenvalues with positive weights b_i^{-1} . $\square$

The common expected potential energy need not equal the target value $d/2$. For $Q = \lambda I$, the two updates coincide pathwise.

4 A separation between the actual sampling errors

We use $W_2(\nu, \pi)^2 = \inf \mathbb{E} \|X - Y\|^2$, where the infimum is over couplings with marginals ν, π and $\|\cdot\|$ is the Euclidean norm. The invariant laws in Theorem 2.1 are not Gaussian at positive step size and finite m ; Appendix A gives a short argument. Their transport errors therefore require bounds beyond the covariance formulas.

For any centered law $\nu \in \mathcal{P}_2$ with covariance $\text{diag}(c_i)$ in an eigenbasis of Q ,

$$W_2(\nu, \pi_Q) \geq \left[\sum_i \left(\sqrt{c_i} - \lambda_i^{-1/2} \right)^2 \right]^{1/2}. \quad (14)$$

In any coupling, $\mathbb{E}[X_i Y_i] \leq \sqrt{c_i/\lambda_i}$ by Cauchy–Schwarz, which proves the bound after summing. This is the commuting-covariance case of the general covariance transport bound [3]; it is an equality for the corresponding Gaussian laws, but only a lower bound here. Denote its right side for Model A by $\mathcal{L}_A$.

Let ν_U be the ordinary unadjusted Langevin invariant law for

$$Y_{k+1} = (I - \eta Q)Y_k + \sqrt{2\eta} \xi_k.$$

The stability assumption ensures $\eta < 2/L$, so $\nu_U = \mathcal{N}(0, C_U)$ with $u_i = [\lambda_i(1 - \eta\lambda_i/2)]^{-1}$. Set

$$W_U = \left[\sum_i \left(\sqrt{u_i} - \lambda_i^{-1/2} \right)^2 \right]^{1/2} = W_2(\nu_U, \pi_Q). \quad (15)$$

The lower bound is attained here by coupling both diagonal Gaussian laws through shared standard Gaussian coordinates.

Proposition 4.1 (Stationary shared-noise bound). *Under (5), there is a stationary coupling of $X^B \sim \nu_B$ and $Y \sim \nu_U$ satisfying*

$$\mathbb{E}[(X_i^B - Y_i)^2] = \frac{\eta}{m(2 - \eta\lambda_i)} \left(\lambda_i c_i^B + H/z \right). \quad (16)$$

Consequently,

$$W_2(\nu_B, \pi_Q) \leq \mathcal{U}_B := \left[\sum_i \frac{\eta(\lambda_i c_i^B + H/z)}{m(2 - \eta\lambda_i)} \right]^{1/2} + W_U. \quad (17)$$

Proof. Use the same two-sided thermal innovations in (10) and the stable ULA series, with the sketches independent of those innovations. Their joint L^2 limits yield a stationary coupling. Write $E_k = X_k^B - Y_k$ and $\Delta H_k = Q^{1/2}(P_k - I)Q^{1/2}$. Then

$$E_{k+1} = (I - \eta Q)E_k - \eta \Delta H_k X_k^B.$$

The fresh centered matrix ΔH_k is independent of both current states, so the covariance cross terms vanish. Equation (7) gives

$$\mathbb{E}[\Delta H_k C_B \Delta H_k] = m^{-1} [Q C_B Q + \text{tr}(Q C_B) Q].$$

In coordinate i , stationarity therefore gives

$$\mathbb{E}E_i^2 = (1 - \eta\lambda_i)^2 \mathbb{E}E_i^2 + \frac{\eta^2}{m} (\lambda_i^2 c_i^B + \lambda_i H/z).$$

This proves (16). Its sum is a valid coupling cost. Apply the triangle inequality through ν_U and use (15). $\square$

Theorem 4.2 (Different conditioning dependence). *Fix $d \geq 2$, finite $m \geq 1$, and*

$$0 < \eta < \frac{2m}{m + d + 1}. \quad (18)$$

For $Q_\mu = \text{diag}(\mu, 1, \dots, 1)$, $0 < \mu \leq 1$, both chains have the finite-second-moment invariant laws above. As $\mu \downarrow 0$,

$$W_2(\nu_{A,\mu}, \pi_{Q_\mu}) = \Theta(\mu^{-1/2}), \quad W_2(\nu_{B,\mu}, \pi_{Q_\mu}) = \Theta(1). \quad (19)$$

All constants may depend on (d, m, η) , but not on μ .

Proof. Put $b_* = 1 - \eta a/2$, $z_* = 1 - td/b_*$, and

$$\delta = b_* z_* = 1 - \frac{\eta(m + d + 1)}{2m} > 0.$$

For all $\mu \in (0, 1]$, $b_i \geq b_* > 0$, $D \leq d/b_*$, and $z \geq z_* > 0$, which proves uniform admissibility of (5). Also $H/z \leq d/\delta$. Hence both relative variance profiles satisfy

$$\lambda_i c_i^A \leq \delta^{-1}, \quad \lambda_i c_i^B \leq \delta^{-1}. \quad (20)$$

For the B inequality, use $\lambda_i \leq 1$ and $(1 + td/\delta)/b_* = (1/z_*)/b_* = \delta^{-1}$.

Define $z_0 = 1 - t(d - 1)/b_* \in (0, 1)$. An exact identity is

$$z_\mu b_{1,\mu} = z_0 - \mu(\eta a z_0/2 + t) \leq z_0.$$

The first coordinate of (14) consequently gives, for every $0 < \mu \leq 1$,

$$W_2(\nu_{A,\mu}, \pi_{Q_\mu}) \geq \frac{z_0^{-1/2} - 1}{\sqrt{\mu}}. \quad (21)$$

An independent coupling of the two centered laws and (20) give the matching bound

$$W_2^2(\nu_{A,\mu}, \pi_{Q_\mu}) \leq \text{tr } C_A + \text{tr } Q_\mu^{-1} \leq (1 + \delta^{-1})(d - 1 + \mu^{-1}) \leq \frac{d(1 + \delta^{-1})}{\mu}.$$

For B, (20), $H/z \leq d/\delta$, and (17) bound the coupling term by $[\eta d(d + 1)/(m(2 - \eta)\delta)]^{1/2}$. The function $f(s) = (1 - s)^{-1/2} - 1$ is convex with $f(0) = 0$, so $f(\eta\lambda_i/2) \leq \lambda_i f(\eta/2)$. Therefore

$$W_2(\nu_{B,\mu}, \pi_{Q_\mu}) \leq \sqrt{\frac{\eta d(d + 1)}{m(2 - \eta)\delta}} + \sqrt{d}[(1 - \eta/2)^{-1/2} - 1] =: K. \quad (22)$$

Finally, any stiff coordinate has $c_i^B = (1 + tH/z)/b_* \geq b_*^{-1}$. Its one-coordinate transport lower bound is at least $b_*^{-1/2} - 1 > 0$. This proves both orders in (19). $\square$

At fixed positive step size, A retains a relative variance error in the broad coordinate, whereas B’s relative variance tends to one. B’s stiff-coordinate error remains positive. The result concerns stationary error; the broad-coordinate mean contracts by $1 - \eta\mu$, so stability does not imply uniform mixing as $\mu \downarrow 0$.

5 Numerical study

We test the covariance formulas, coupling identity, and directional effect across 41 configurations, each with 20 independent runs of coupled A, B, and ULA chains.

Design. The principal family uses $d = 4$, $m = 2$, $\eta = 0.25$, and nine μ values from 1 to 10^{-4} . Additional configurations vary step size, sketch size, spectrum, and dimension from 2 to 256; include two flat directions; double burn-in under cold and hot starts; and stress the method at 85% of its second-moment stability threshold. The protocol is summarized in Appendix B; complete configurations accompany the code. Burn-in uses at least 12 decay e-folds at the slower mean/second-moment spectral rate. Measurement spans at least 640 slow drift time units when $\lambda_{\min} \leq 0.03$, 320 for the dimension sweep and otherwise at least 80, and always at least 30,000 steps. These burn-in choices are heuristic, with hot-start and longer-run controls.

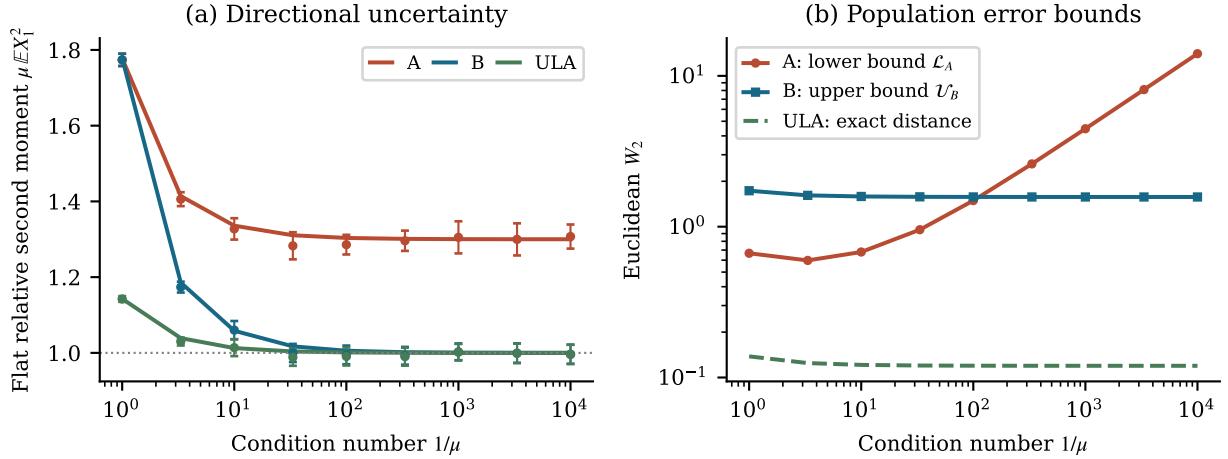


Figure 1: One-flat-direction family, $d = 4$, $m = 2$, $\eta = 0.25$. (a) Exact relative variances (lines) and post-burn-in second-moment estimates (markers, pointwise 95% intervals across 20 independent runs). The dotted line is the target value one. (b) Deterministic population bounds from (14) and Proposition 4.1, and the exact ULA distance. These curves are not empirical Wasserstein estimates.

Directional effect and transport bounds. Figure 1 shows the persistent flat-mode distortion in A and the different conditioning dependence of the bounds. At $\mu = 10^{-4}$, the measured relative second moments are 1.307 ± 0.032 for A and 0.996 ± 0.025 for B. The exact predictions are 1.300035 and 1.000057. B’s predicted excess is below the sampling uncertainty. The paired A-minus-B relative difference is 0.3108 ± 0.0194 . At these parameters, $\mathcal{L}_A = 14.027$ and $\mathcal{U}_B = 1.574$, so the analytic bounds already distinguish the actual invariant-law errors. Table 1 gives four conditioning levels.

Validation. Full covariance-operator solves on rotated precision matrices in dimensions 2, 3, 4, and 8 agree with (6) to maximum absolute discrepancy 6.1×10^{-14} . Direct tensor Gaussian quadrature independently averages the random update matrices, with maximum operator discrepancy 7.8×10^{-16} . Across all configurations except the designated near-boundary stress case, the simulated total coupling cost $\mathbb{E}\|X^B - Y\|^2$ agrees with the sum of the coordinatewise predictions in (16) to within 3.77% relative error. Hot-start and doubled-burn-in checks at $\mu = 10^{-3}$ support the same

$1/\mu$	A: measured $\mu\mathbb{E}X_1^2$	B: measured $\mu\mathbb{E}X_1^2$	$\mathcal{L}_A$	$\mathcal{U}_B$
10	1.327 ± 0.028	1.060 ± 0.024	0.680	1.586
100	1.286 ± 0.026	0.995 ± 0.025	1.490	1.575
1000	1.305 ± 0.042	1.003 ± 0.022	4.461	1.574
10000	1.307 ± 0.032	0.996 ± 0.025	14.027	1.574

Table 1: Post-burn-in estimates with pointwise 95% confidence half-widths; the target relative second moment is one. $d = 4, m = 2, \eta = 0.25$. The last two columns are population lower/upper bounds, not measured distances.

moment predictions. 3 of 41 pointwise paired energy-difference intervals exclude zero; these intervals are pointwise. Near the stability boundary, exploratory fourth-moment estimates are large, illustrating that finite variance does not ensure easy Monte Carlo estimation.

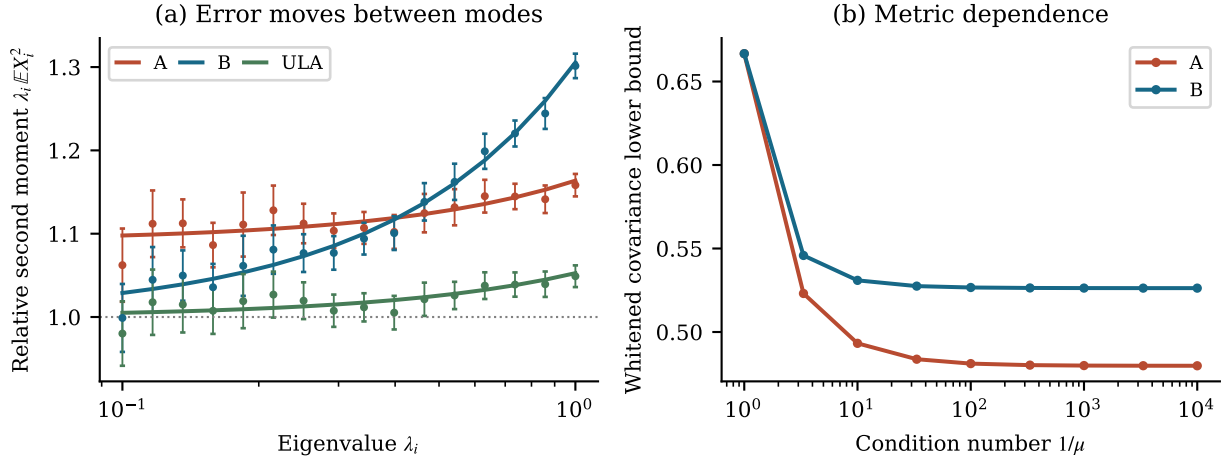


Figure 2: (a) Relative covariance profiles for a geometric spectrum with $d = 16, m = 4, \eta = 0.1$: exact curves and independent-run estimates illustrate the crossing. (b) Covariance transport lower bounds after whitening by $Q_\mu^{1/2}$ for the principal family. B’s larger lower bound in this metric does not establish an ordering of the actual whitened Wasserstein errors.

Dimension sweep. We sweep $d \in \{2, 4, 8, 16, 32, 64, 128, 256\}$ for two spectra: geometric eigenvalues between 0.1 and 1, and $(0.1, 1, \dots, 1)$. Both use $m = 2$ and $\eta = 0.4\eta_{\text{crit}}(Q, m)$, where η_{crit} is the exact second-moment stability boundary. Condition number and sketch size are fixed; step size adapts to dimension. Each configuration measures at least 320 slow drift time units and 30,000 steps. Figure 3 shows the flattest and stiffest modes. Across the dimension sweep, the largest relative second-moment discrepancy over all methods and coordinates is 5.01%, and 95.8% of pointwise intervals cover their exact predictions. The intervals are pointwise. This exploratory sweep used a uniform extension from an initial 80 to 320 drift time units.

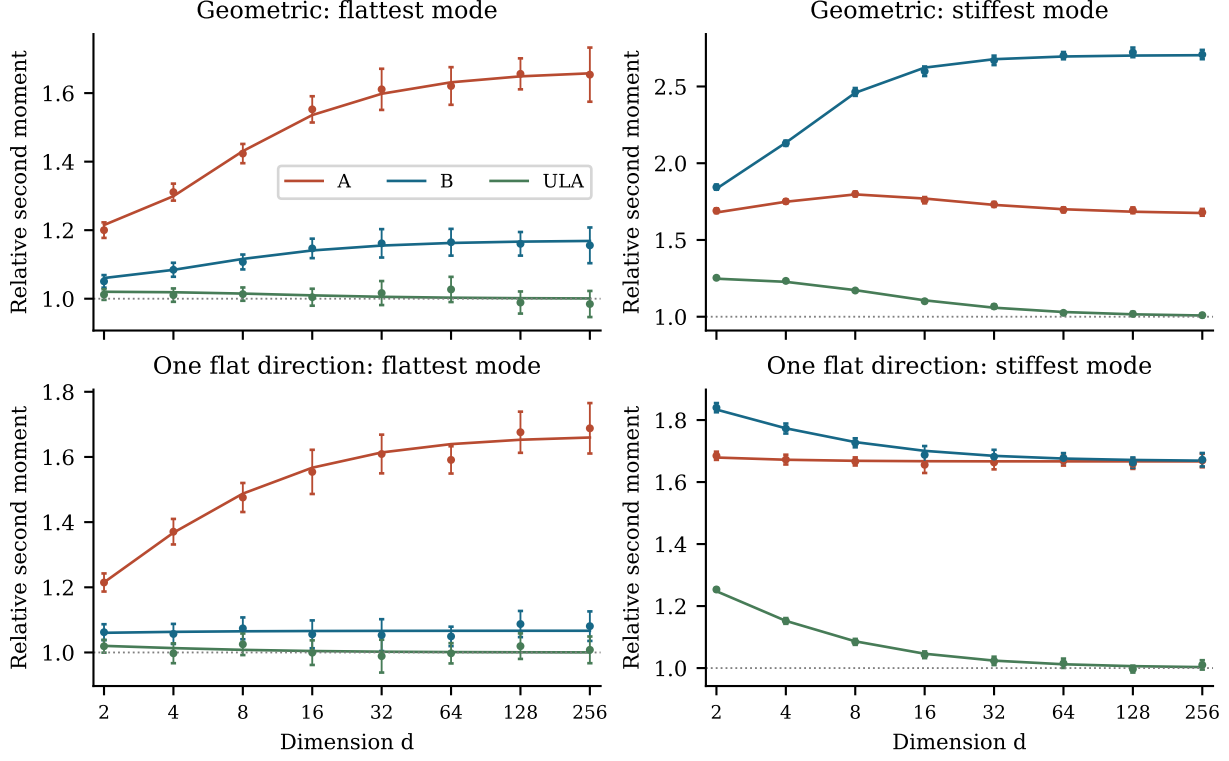


Figure 3: Dimension controls at fixed condition number 10 and sketch size $m = 2$. Lines are exact relative variances; markers and pointwise 95% intervals use 20 independent runs. Rows distinguish geometric and one-flat-direction spectra; columns show the smallest and largest eigenvalues. Step sizes equal 40% of each configuration’s stability threshold.

Two-dimensional distributions. For $Q = \text{diag}(0.01, 1)$ and $m = 2$, we use $\eta \in \{0.25, 0.5\}$, with 20 runs and 256,000 measured states per run and method. Figure 4 uses $\eta = 0.5$; Appendix C gives $\eta = 0.25$. The heatmaps and marginal curves use paired coordinate snapshots every ten steps, pooled across runs. Histograms share a grid and color scale, with Gaussian smoothing of one bin. Second moments use all measured states. At $\eta = 0.5$, the exact relative variance profiles are $(1.2567, 2.0031)$ for A and $(1.0079, 2.2520)$ for B. Thus B preserves the broad direction more closely while increasing narrow-direction error.

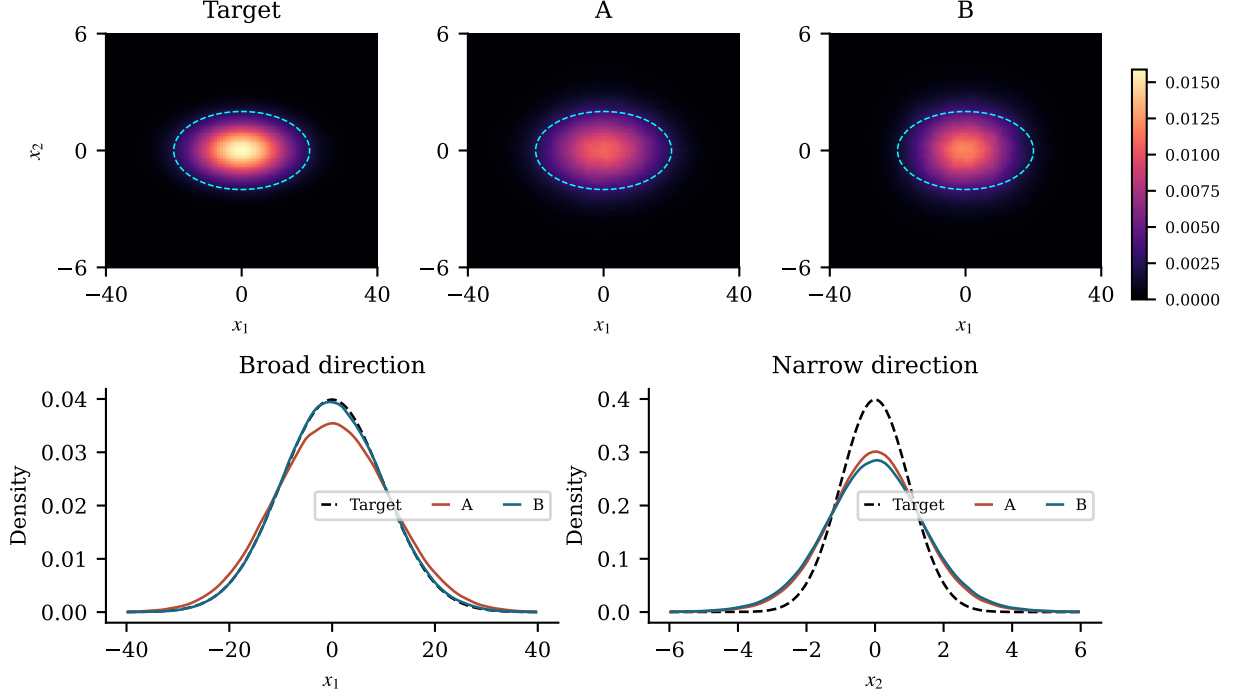


Figure 4: Two-dimensional illustration, $Q = \text{diag}(0.01, 1)$, $m = 2$, $\eta = 0.5$. Top: analytic target density and empirical heatmaps for A and B, using the same axes and color scale. Dashed cyan curves mark the target Mahalanobis-radius-two ellipse. Bottom: empirical marginal densities and exact target marginals. The displayed samples remain correlated; the figure does not estimate population Wasserstein distance.

An explicit variance profile. For $d = 4$, $m = 2$, and $\eta = 1/4$, the relative variance profiles simplify to

$$\begin{aligned} \mu c_1^A &= \frac{208}{160 - 43\mu}, & c_i^A &= \frac{256 - 48\mu}{160 - 43\mu} \quad (i > 1), \\ \mu c_1^B &= \frac{160 + 48\mu}{160 - 43\mu}, & c_i^B &= \frac{272 - 64\mu}{160 - 43\mu} \quad (i > 1). \end{aligned} \quad (23)$$

As $\mu \downarrow 0$, the relative variance profiles tend to $(1.3, 1.6, 1.6, 1.6)$ for A and $(1, 1.7, 1.7, 1.7)$ for B. The same expected energy therefore conceals different directional distortions.

Step size and conditioning. Figure 5 compares A's predicted broad-coordinate lower bound, $\sqrt{c_1^A} - \mu^{-1/2}$, with the empirical broad-coordinate Wasserstein distance to $\mathcal{N}(0, \mu^{-1})$. Panel (a) varies conditioning at $\eta = 0.25$; panel (b) varies $\eta \in \{0.0625, 0.125, 0.25, 0.375, 0.5\}$ at $\mu = 0.01$. Both use $d = 4$ and $m = 2$. The empirical distance integrates the quantile function of the pooled snapshots against the exact Gaussian quantile function. Shading gives pointwise percentile intervals from 200 resamples of the 20 entire runs. These intervals describe resampling uncertainty, not finite-sample bias. Empirical and exact ULA marginal distances provide a reference for this bias. The comparison is one-dimensional and does not estimate full-dimensional W_2 .

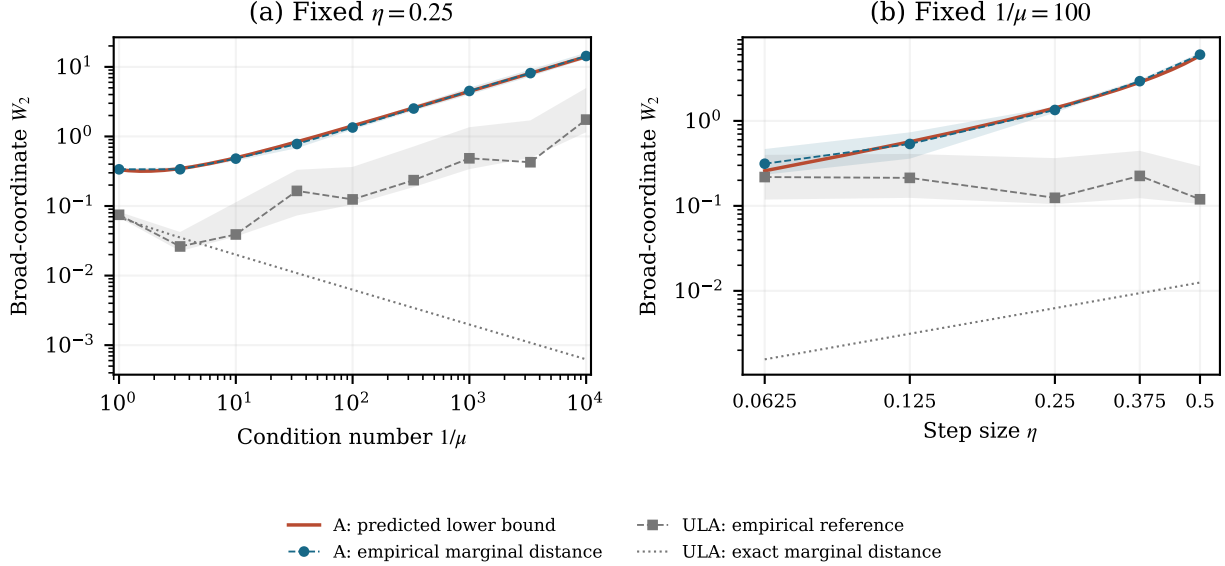


Figure 5: Predicted lower bound and empirical broad-coordinate sampling error for A, with $d = 4$, $m = 2$. (a) Conditioning varies at fixed $\eta = 0.25$. (b) Step size varies at fixed $1/\mu = 100$. The solid red curve is the population marginal covariance lower bound; blue markers estimate the marginal W_2 from pooled snapshots. Shading shows whole-run bootstrap intervals. Gray curves compare empirical and exact ULA distances, exposing the finite-sample error in the transport diagnostic.

6 Discussion

Sketch placement changes stationary directional uncertainty despite equal mean drift, stability, and expected energy. The covariance crossing identifies this redistribution; the stationary coupling establishes the resulting separation in actual Euclidean W_2 error.

The metric matters. After whitening, both covariance lower bounds remain bounded, and B’s larger lower bound does not order the actual transport errors. In (23), B protects the broad direction while increasing stiff-direction error; both methods have limiting expected energy 3.05, compared with the target’s 2. Non-Gaussian stationary laws and related heavy-tail effects are already established in SGD [4].

No computational speedup is established: PQx may require the full gradient, and $Q^{1/2}$ may be costly. Sharp step-size bounds, target-normalized error, and inexpensive factorizations remain natural directions for further study.

References

- [1] N. Brosse, A. Durmus, and E. Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL <https://arxiv.org/abs/1811.10072>.
- [2] Z. Ding and Q. Li. Langevin Monte Carlo: Random coordinate descent and variance reduction. *Journal of Machine Learning Research*, 22(205):1–51, 2021. URL <https://www.jmlr.org/papers/v22/20-1205.html>.
- [3] M. Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990. doi: 10.1002/mana.19901470121. URL <https://onlinelibrary.wiley.com/doi/10.1002/mana.19901470121>.
- [4] M. Gurbuzbalaban, U. Simsekli, and L. 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*, pages 3964–3975. PMLR, 2021. URL <https://proceedings.mlr.press/v139/gurbuzbalaban21a.html>.
- [5] K. Liu, L. Ziyin, and M. Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7045–7056. PMLR, 2021. URL <https://proceedings.mlr.press/v139/liu21ad.html>.
- [6] T. Maunu and J. Yao. Subspace Langevin Monte Carlo. *SIAM Journal on Mathematics of Data Science*, 8(3): 850–886, 2026. doi: 10.1137/25M1739327. URL <https://epubs.siam.org/doi/10.1137/25M1739327>. Preprint arXiv:2412.13928, first posted 2024.
- [7] D. Paulin and P. A. Whalley. Theoretical guarantees for stochastic gradient Langevin dynamics. arXiv preprint arXiv:2610.01651v1, 2026. URL <https://arxiv.org/abs/2610.01651>. Version dated October 1, 2026.
- [8] A. Roy, L. Shen, K. Balasubramanian, and S. Ghadimi. Stochastic zeroth-order discretizations of Langevin diffusions for Bayesian inference. *Bernoulli*, 28(3):1810–1834, 2022. doi: 10.3150/21-BEJ1400. URL <https://arxiv.org/abs/1902.01373>.
- [9] M. Vono, V. Plassier, A. Durmus, A. Dieuleveut, and E. Moulines. QLSG: Quantised Langevin stochastic dynamics for Bayesian federated learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6459–6500. PMLR, 2022. URL <https://proceedings.mlr.press/v151/vono22a.html>.
- [10] J.-N. Wang, Z. Huang, K. Li, Z. Wang, and L. Wu. Momentum in large-batch training: Polyak enlarges the critical batch size, Nesterov improves data efficiency. arXiv preprint arXiv:2609.02728v3, 2026. URL <https://arxiv.org/abs/2609.02728>. Version dated October 27, 2026.
- [11] L. Ziyin, K. Liu, T. Mori, and M. Ueda. Strength of minibatch noise in SGD. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=uorVGbWV5sw>. Full text also available as arXiv:2102.05375v3.

A Why the invariant laws are not Gaussian

The finite-second-moment invariant covariance is positive definite because a step adds $2\eta I$. Suppose the invariant law were $\mathcal{N}(0, C)$ with $C \succ 0$. Conditional on S , a stationary step would be Gaussian with covariance $M(S)CM(S)^\top + 2\eta I$. For any nonzero direction v , its conditional variance $q(S) = v^\top M(S)CM(S)^\top v + 2\eta \|v\|^2$ is a nonconstant polynomial in the Gaussian entries of S . To see this, set the first row of S to $sw^\top$ and all other rows to zero. Choose $w = v$ for A and $w = Q^{1/2}v$ for B. In each case the coefficient of s^4 in $q(S)$ is strictly positive because $C, Q \succ 0$. The Gaussian sketch has full support, so $q(S)$ has positive variance and finite moments. The unconditional fourth moment of this projection would be $3\mathbb{E}[q(S)^2] > 3(\mathbb{E}q(S))^2$, contradicting the Gaussian fourth moment. This argument does not assert a finite stationary fourth moment for the actual invariant law.

B Experimental protocol

Each run starts at zero and draws fresh sketches and thermal noise. A and B share each sketch, and all three chains share thermal noise; runs use independent seeds. The hot-start control uses $\mathcal{N}(0, 25Q^{-1})$. Second moments average X_i^2 about the known zero mean. Pointwise 95% Student- t intervals use independent-run summaries, accounting for dependence within each chain. Geometric spectra have $\lambda_i = 0.1 \cdot 10^{(i-1)/(d-1)}$; the two-flat control uses (0.01, 0.01, 1, 1) and the stress case uses (0.1, 1, 1, 1). Heatmaps retain every tenth paired state. Empirical transport diagnostics are excluded from the error claims because finite-sample error grows with target width. Code, seeds, complete configurations, and retained summaries accompany the paper.

C Two-dimensional smaller-step control

Figure 6 repeats the two-dimensional diagnostic at the principal study's step size $\eta = 0.25$. The exact relative variances are (1.0861, 1.3342) for A and (1.0034, 1.4169) for B. Each method uses 20 runs, 4,794 burn-in steps and 256,000 measured states per run. At $\eta = 0.5$, burn-in is 3,000 steps; the measured count and snapshot stride are unchanged. The 2D estimates are separate from the four-dimensional study.

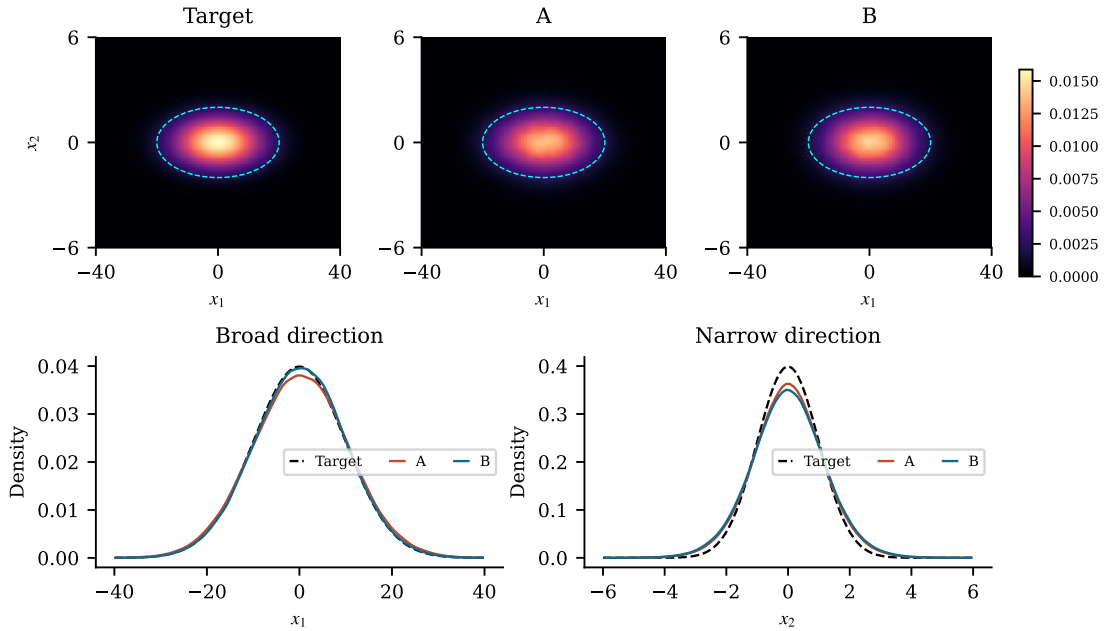


Figure 6: Smaller-step control for Figure 4, with $\eta = 0.25$. The same histogram grid, smoothing, axes and color scale are used.