MAKING GRADIENT HISTORY COUNT: SPECTRALLY CALIBRATED MONTE CARLO

Anonymous authors
Paper under double-blind review

ABSTRACT

Bayesian inference and uncertainty quantification rely on Monte Carlo estimates, whose accuracy can remain limited by correlated samples under a finite gradient budget. We ask how gradients already computed during sampling can reduce this error. Spectrally calibrated score repulsion Monte Carlo (SC-SRMC) uses the running average of target scores, the gradients of the log density, to adjust acceptance in the Metropolis adjusted Langevin algorithm (MALA). An exact Gaussian recursion identifies the largest score covariance eigenvalue as the scale for increasing tilt strength. A radial cap bounds the resulting tilt. Under uniform stability and regularity conditions, we bound the score history over finite horizons, with explicit terms for transition correlations. An exact projection transfers the bound to the score component of observable error. The method preserves MALA's rule for generating candidate states and requires no extra online gradients or second derivatives. At matched online gradient budgets, one configuration selected on five development targets gives lower observed mean squared error of coordinate means across 11 Gaussian and quartic validation targets, including six new geometries. The geometric mean reduction relative to diagonally preconditioned MALA is 5.65 fold at the same step size and 6.59 fold with separately tuned step sizes. The error decomposition ties parameter choice to both the observable and target geometry.

1 INTRODUCTION

Bayesian inference and uncertainty quantification require expectations under complex distributions. Markov chain Monte Carlo (MCMC) estimates these expectations from correlated samples. Initialization and sample dependence affect accuracy over finite runs, and gradient evaluations often account for much of the computational cost in gradient based methods. The question is how to use gradients from earlier steps to reduce this error without adding gradient evaluations.

Gradient information supports several approaches to this problem. Hamiltonian Monte Carlo (HMC) uses auxiliary momentum to generate trajectories (Neal, 2011). The adaptive No U Turn Sampler (NUTS) selects their lengths (Hoffman & Gelman, 2014). Gradient based control variates (Mira et al., 2013) and Stein control functionals (Oates et al., 2017) reuse gradients in estimator corrections with zero target expectation. Score Repellent Monte Carlo (SRMC) uses a running average of target scores, the gradients of the log density, to guide subsequent transitions (Hu et al., 2026). That work introduces an exponential tilt and rational warmup, and decomposes asymptotic estimation error using the score.

With exact draws from tilted Gaussian targets, stronger tilts can amplify initial fluctuations even when they reduce asymptotic variance. Amplification occurs in directions with large score variance. Tilt strength therefore needs a scale that accounts for the early response. Reducing the score history also need not reduce every estimation error. We must identify the component of each observable's error controlled by that history.

Spectrally calibrated score repulsion Monte Carlo (SC-SRMC) sets the warmup scale from the largest score covariance eigenvalue, estimated in a preliminary run, and caps the norm of the target tilt. The exact Gaussian recursion gives a spectral condition preventing early amplification. Under uniform stability and regularity conditions, a Poisson equation accounts for correlated transitions in a finite horizon bound on the score history. An exact projection separates each observable's error into the controlled score component and a residual.

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

We implement the method with the Metropolis adjusted Langevin algorithm (MALA). MALA draws a candidate state from a gradient guided probability distribution, called the proposal distribution. It then accepts the candidate or remains at the current state. Preconditioning rescales candidate displacements using a diagonal or full matrix. Past gradients adjust acceptance, preserving the candidate generation rule and online gradient cost.

Our contributions are:

1. Under uniform stability and regularity conditions, we bound score history over finite horizons with explicit transition mixing terms. An exact projection transfers the bound to the score component of observable error, including after burnin.
2. We use score covariance to calibrate spectral warmup (SW) and a radial cap. The MALA implementation reuses cached scores without extra online gradients or second derivatives.
3. Across 11 validation targets, including six new geometries, one SC-SRMC configuration reduces the observed mean squared error (MSE) of coordinate means by a geometric mean factor of 5.65 relative to diagonally preconditioned MALA, at equal online gradient budgets and the same step size.

2 Spectral calibration with cached scores

In the exact Gaussian model, strong tilts can amplify the early score history. We calibrate tilt strength to the largest score covariance eigenvalue and cap the tilt. The acceptance correction favors candidate states with lower score projections along the history. Both the correction and history update reuse scores computed by MALA (Figure 1).

2.1 Score history and target tilting

Let π be the target probability distribution on $\mathbb{R}^d$. Assume its density $\pi(x)$ is positive and continuously differentiable. Its score is $s(x) = \nabla \log \pi(x)$, and we assume $\pi(s) = 0$. For an observable f, $\pi(f) = \mathbb{E}_\pi f(X)$ is its target expectation. Score repulsion guides transitions through an exponential target tilt (Hu et al., 2026). Choose an open tilt domain $\mathcal{U} \subset \mathbb{R}^d$ containing zero, where the log normalizer is finite and twice differentiable. For a tilt vector $u \in \mathcal{U}$, define

$$A(u) = \log \mathbb{E}_\pi e^{-u^\intercal s(X)}, \qquad \pi_u(\,\mathrm{d}x) = e^{-u^\intercal s(x) - A(u)} \pi(\,\mathrm{d}x), \qquad S(u) = \mathbb{E}_{\pi_u} s(X). \tag{1}$$

Let P_u be a transition kernel preserving π_u. Write $P_u g(x)$ for the expected value of g after one transition from x. The state and score history evolve as

$$X_{n+1} \sim P_{u_n}(X_n, \cdot), \qquad \theta_{n+1} = \theta_n + a_n\{s(X_{n+1}) - \theta_n\}. \tag{2}$$

The averaging weight a_n controls each new score's contribution. With $a_n = \frac{1}{n+1}$ and $\theta_0 = 0$, as in the experiments, θ_n averages $s(X_1), \ldots, s(X_n)$, including repeated states after rejection. The tilt vector u_n is fixed before each transition, and the retained state updates the history after acceptance or rejection. We estimate $\pi(f)$ by averaging $f(X_k)$ after burnin, an initial segment discarded from estimation. The retained states in this segment remain in the score history.

2.2 Acceptance correction with cached scores

For a fixed positive definite preconditioner (metric) M (Girolami & Calderhead, 2011) and step size h, we draw a candidate state y from the standard MALA proposal distribution at the current state x (Roberts & Rosenthal, 1998):

$$q_0(y \mid x) = \mathcal{N}\left(y \mid x + \frac{h}{2} M s(x), hM\right). \tag{3}$$

We accept y with probability given by the correction of Metropolis et al. (1953) and Hastings (1970), applied to the current tilted target:

$$\mathrm{acc}_u(x, y) = 1 \wedge \frac{\pi(y) e^{-u^\intercal s(y)} q_0(x \mid y)}{\pi(x) e^{-u^\intercal s(x)} q_0(y \mid x)}. \tag{4}$$

On rejection, the chain remains at x. The correction follows the score repulsion acceptance rule of Hu et al. (2026, Algorithm 2), here applied to the standard MALA proposal. Their Algorithm 4 uses the tilted score in the proposal itself. The resulting transition kernel is reversible under π_u for fixed u. Relative to base MALA, its log acceptance ratio adds $-u^{\mathsf{T}}\{s(y) - s(x)\}$. Both scores are already available from the proposal and reverse proposal calculations. The acceptance correction and history update therefore add vector arithmetic with no additional gradient evaluations.

2.3 Spectral warmup and radial cap

For a Gaussian target, score covariance equals the precision matrix. Its largest eigenvalue determines how strongly the score history affects the exact recursion of Section 3. We use this scale to calibrate the tilt strength schedule (Hu et al., 2026). A preliminary run of the base transition produces score samples. Let $\widehat{C}_s$ denote their empirical covariance, and set $\widehat{\Lambda} = \lambda_{\max}(\widehat{C}_s)$. These quantities estimate $\mathrm{Cov}_\pi\{s(X)\}$ and its largest eigenvalue. For $\widehat{\Lambda} > 0$ and positive parameters $\bar{\alpha}, c, \delta$, set

$$\alpha_n^{\mathrm{SW}} = \frac{n+1}{c\widehat{\Lambda} + \frac{n+1}{\bar{\alpha}}}, \qquad u_n = \Pi_{B(0,r)}\left(\alpha_n^{\mathrm{SW}}\theta_n\right), \qquad r = \frac{\delta\sqrt{d}}{\sqrt{\widehat{\Lambda}}}. \tag{5}$$

Here $\Pi_{B(0,r)}$ denotes Euclidean projection onto the ball of radius r centered at zero. The parameter $\bar{\alpha}$ sets the eventual tilt strength, and c controls the warmup duration. The parameter δ sets the cap radius. The projection preserves the history direction and shortens the tilt when its norm exceeds r. Appendix D gives the diagonal and dense metric constructions and setup costs.

For $\theta_n \neq 0$, the projection is exactly equivalent to $u_n = \alpha_n\theta_n$ with

$$\alpha_n = \min\left\{\alpha_n^{\mathrm{SW}}, \frac{\delta\sqrt{d}}{\sqrt{\widehat{\Lambda}}\|\theta_n\|}\right\}. \tag{6}$$

At $\theta_n = 0$ we set $\alpha_n = \alpha_n^{\mathrm{SW}}$. The implementation uses a 10^{-12} denominator guard. Under Theorem 1's conditions with $a_n = \frac{1}{n+1}$, the history tends to zero in mean square. Since $\|u_n\| \leq \bar{\alpha}\|\theta_n\|$, the tilt also vanishes in mean square. At $u = 0$, the tilted target equals π.

The Kullback Leibler (KL) divergence is the expected log density ratio. If $\nabla^2 A(u) = \mathrm{Cov}_{\pi_u}(s) \preceq \Lambda I$ and $\widehat{\Lambda} \geq \Lambda$ along the segment from zero to u, then

$$D_{\mathrm{KL}}(\pi_u\|\pi) = \int_0^1 t\, u^{\mathsf{T}}\nabla^2 A(tu)u \,\mathrm{d}t \leq \frac{1}{2}\Lambda\|u\|^2 \leq \frac{1}{2}\delta^2 d. \tag{7}$$

A finite normalizer throughout the tilt ball ensures that each tilted density is well defined. Appendix E verifies normalizability for the benchmark targets and gives the corresponding condition for Poisson regression.

3 Theory over finite horizons

Bounding the score history requires control of both its mean drift and correlations between successive transitions. Score covariance determines the drift averaged under the tilted target, called the mean field. A Poisson equation accounts for the transition correlations.

Lemma 1 (Drift from the covariance of scores).

If differentiation under the integral is valid, the Jacobian $DS(u)$ of the mean score satisfies

$$DS(u) = -\mathrm{Cov}_{\pi_u}(s). \tag{8}$$

For fixed $\alpha \geq 0$ and covariance bounds $0 < \kappa_s \leq K < \infty$, if $\kappa_s I \preceq \mathrm{Cov}_{\pi_u}(s) \preceq KI$ on a convex tilt region containing zero, $\alpha\theta$, and $\alpha\theta'$, then $h_\alpha(\theta) = S(\alpha\theta) - \theta$ satisfies

$$\langle\theta, h_\alpha(\theta)\rangle \leq -(1+\alpha\kappa_s)\|\theta\|^2, \qquad \|h_\alpha(\theta) - h_\alpha(\theta')\| \leq (1+\alpha K)\|\theta - \theta'\|. \tag{9}$$

For the radial projection (5), the first inequality holds with the effective scalar $\alpha_n(\theta) = \frac{\|u_n(\theta)\|}{\|\theta\|}$ for $\theta \neq 0$. At zero, set $\alpha_n(0) = \alpha_n^{\mathrm{SW}}$.

Because the transition kernel changes with the history, we solve a Poisson equation for each tilt. This separates a martingale difference, a noise term with zero conditional mean, from the remaining error:

$$g_u - P_u g_u = s - S(u), \qquad G_u = P_u g_u. \tag{10}$$

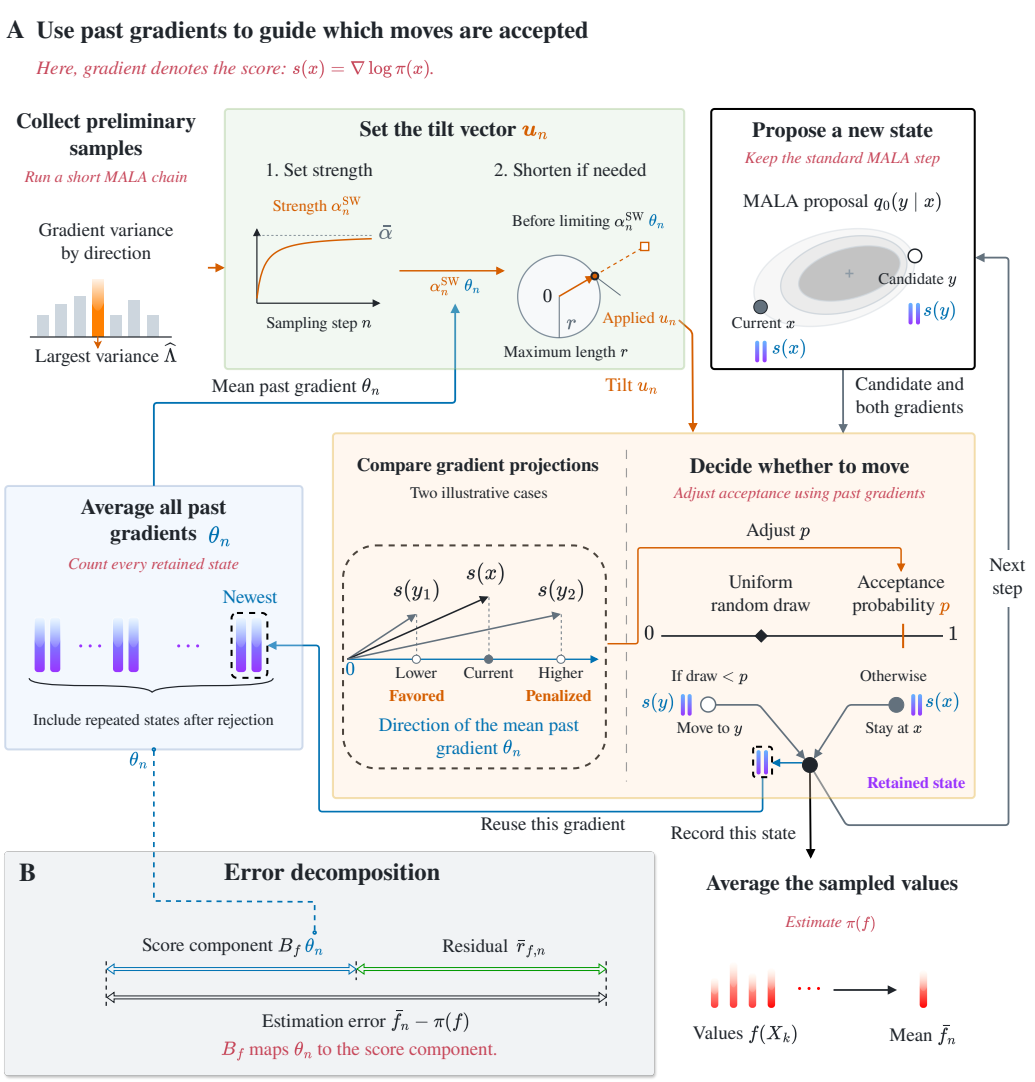


Figure 1: Cached gradients connect the acceptance decision to the score component of estimation error. A: The average gradient θ_n sets the tilt direction, preliminary score covariance calibrates warmup, and radial projection bounds u_n. The proposal ellipses remain unchanged, with their mean at the cross. Examples y_1, y_2 show how lower score projections receive larger acceptance multipliers. Only one candidate is proposed per transition. The acceptance probability is $p = \mathrm{acc}_{u_n}(x, y)$ from (4). The retained state's score updates history, including repetitions after rejection. B: The full history identity $\bar{f}_n - \pi(f) = B_f\theta_n + \bar{r}_{f,n}$ separates the component controlled through history from the mean residual. Section 4 defines the projection coefficient B_f, sample mean $\bar{f}_n$, and mean residual $\bar{r}_{f,n}$. Residual fluctuations also enter the total error. Segment lengths are schematic.

We assume uniform conditional second moment bounds on the martingale increments, root mean square (L^2) bounds on G_u, and Lipschitz dependence of G_u on u. The steps satisfy $|a_{n+1} - a_n| = \mathcal{O}(a_n^2)$, and the schedule satisfies $\|u_{n+1} - u_n\|_{L^2} = \mathcal{O}(a_n)$. The analysis uses stochastic approximation for controlled Markov noise (Andrieu et al., 2005). For a fixed kernel, Poisson equations yield bounds over finite horizons (Haque & Maguluri, 2025), and Wasserstein contraction provides Poisson solutions (Hofstadler, 2026). Appendix A states the moment and parameter regularity conditions needed as the kernel changes with the history.

Theorem 1 (Contraction of the score history over finite horizons).

Condition on the preliminary sampling run. Assume the covariance and quantitative Poisson conditions in Appendix A. The radial schedule must be uniformly Lipschitz. A convex domain must contain zero, θ_n, and the corrected iterate $z_n = \theta_n + a_n G_{u_n}(X_n)$. Let $m_n = 1 + \kappa_s \underline{\alpha}_n$. Here $\underline{\alpha}_n$ bounds the effective tilt strength from below on this domain. For sufficiently small a_n, finite constants C_0, C_1, C_2, C_3 give

$$\mathbb{E}\|\theta_n\|^2 \le C_0 \exp\left\{-\frac{3}{2}\sum_{j=0}^{n-1} m_j a_j\right\} + C_1 \sum_{k=0}^{n-1} a_k^2 \exp\left\{-\frac{3}{2}\sum_{j=k+1}^{n-1} m_j a_j\right\} + C_2 \sum_{k=0}^{n-1} a_k^3 \exp\left\{-\frac{3}{2}\sum_{j=k+1}^{n-1} m_j a_j\right\} + C_3 a_n^2. \tag{11}$$

For fixed α, take $a_n = \frac{\gamma}{(n+n_0+1)^\rho}$ with scale $\gamma > 0$, exponent $\rho \in (\frac{1}{2}, 1)$, and initial offset $n_0 \ge 0$ chosen to meet the step conditions. Then

$$\mathbb{E}\|\theta_n\|^2 \le C_\star e^{-c_\star(1+\alpha\kappa_s)\{(n+n_0)^{1-\rho} - n_0^{1-\rho}\}} + \mathcal{O}\left(\frac{\sigma_\alpha^2 a_n}{1+\alpha\kappa_s} + \frac{C_{R,\alpha}^2 a_n^2}{1+\alpha\kappa_s}\right). \tag{12}$$

The positive constants $C_\star, c_\star$ do not depend on n. Here σ_α^2 bounds the conditional martingale variance and $C_{R,\alpha}$ controls the remainder of the corrected iterate for the kernel family at fixed α. Appendix A defines their uniform counterparts. For $\rho = 1$ and $\frac{3}{2}\gamma(1+\alpha\kappa_s) > 1$, the leading stochastic term is

$$\mathcal{O}\left\{\frac{\gamma^2\sigma_\alpha^2}{\left[\frac{3}{2}\gamma(1+\alpha\kappa_s) - 1\right]n}\right\}.$$

For the experimental choice $\gamma = 1$, this simplifies to $\mathcal{O}\{\frac{\sigma_\alpha^2}{(1+\alpha\kappa_s)n}\}$.

For fixed α, the factor $1 + \alpha\kappa_s$ in (12) strengthens contraction. The constants σ_α and $C_{R,\alpha}$ depend on mixing under the tilted kernel, so the bound does not imply that increasing α always reduces the history error. For the capped schedule on an unbounded history domain, $\inf_\theta \min\{\alpha_n^{\mathrm{SW}}, r/\|\theta\|\} = 0$. The uniform bound (11) then uses $m_n = 1$. The drift gain in (9) depends on the current history. On a deterministic domain $\|\theta\| \le H$ containing both iterates, one may instead take $\underline{\alpha}_n = \min\{\alpha_n^{\mathrm{SW}}, r/H\} > 0$.

Proposition 1 (Verification on a compact state space).

Let $\mathsf{X} = \mathbb{T}^d$ be the torus of d circular coordinates. For concentration $\eta > 0$, use the von Mises density $\pi(x) \propto \exp\{\eta \sum_{j=1}^d \cos x_j\}$. Use the projected schedule in (5). Let P_u target π_u through a Metropolis kernel with a wrapped Gaussian random walk, adding Gaussian noise modulo 2π in each coordinate. Fix the proposal scale and cap radius. The covariance and quantitative Poisson assumptions of Theorem 1 then hold uniformly over every generated u. After absorbing finitely many initial steps into C_0, its bound applies to the compact experiment with $d = 6$ and $\eta = 2$.

The radial cap confines u to a compact ball. The bounded score makes the tilted density uniformly bounded above and below on this ball. The continuous wrapped proposal also has a positive minimum. Thus every P_u has a common Doeblin minorization, a uniform positive lower bound by one shared reference measure. The score covariance eigenvalues vary continuously with u and remain strictly positive on the compact domain. The Poisson series converges uniformly in sup norm, the maximum magnitude over states. It gives uniform bounds on the solution and its variation with the tilt. Appendix A gives the details.

3.1 Exact Gaussian transient and spectral warmup

Let $\pi = \mathcal{N}(\mu, Q^{-1})$ have mean μ and positive definite precision matrix Q. Write $\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ for the sample mean. At each step, draw exactly from the current tilted density. The score noises ξ_n below are independent. With $a_{n-1} = \frac{1}{n}$, $\theta_0 = 0$, and fixed $\alpha \ge 0$,

$$\theta_n = \left(I - \frac{I + \alpha Q}{n}\right)\theta_{n-1} + \frac{\xi_n}{n}, \qquad \xi_n \sim \mathcal{N}(0, Q), \qquad \bar{X}_n - \mu = -Q^{-1}\theta_n. \tag{13}$$

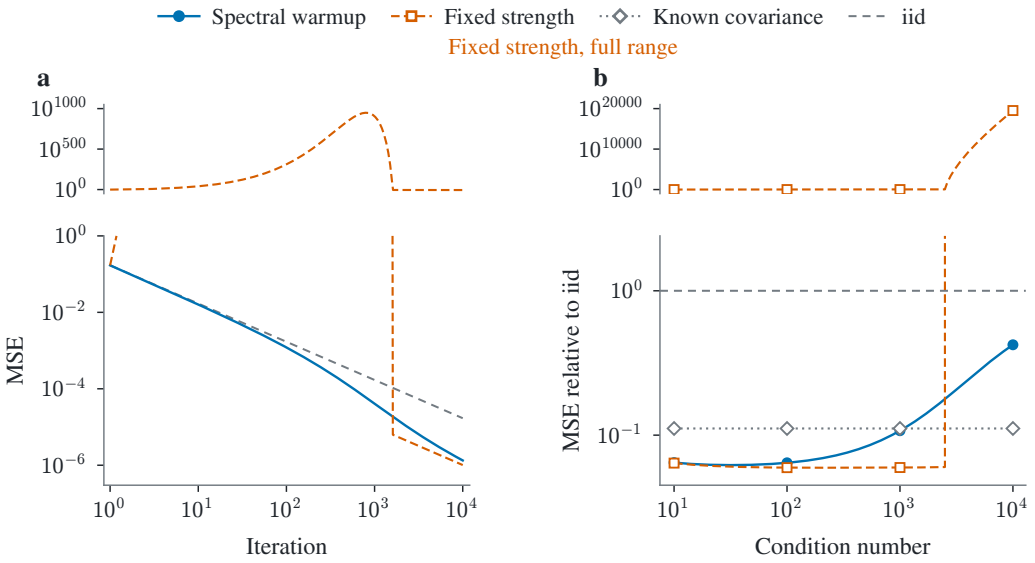


Figure 2: Spectral warmup controls early amplification before the asymptotic variance reduction becomes useful. Exact Gaussian draws isolate this effect from transition correlations. Panel (a) follows the transient, and (b) tests sensitivity to anisotropy relative to independent and identically distributed (iid) target draws. Upper strips retain the full fixed strength curves that exceed the lower plots' range. Lower plots resolve the small errors. The known covariance comparison rescales the target to unit covariance. These calculations use no radial cap. Appendix C gives the settings and exact second moment evaluation.

The subscript q denotes a coordinate along an eigenvector of Q with eigenvalue q:

$$n\mathbb{E}(\bar{X}_{n,q} - \mu_q)^2 \longrightarrow \frac{1}{q(1+2\alpha q)}. \tag{14}$$

Increasing α reduces the asymptotic variance. Early in the run, the multiplier $1 - \frac{1+\alpha q}{n}$ can have magnitude much larger than one and amplify the history. Writing $C = c\widehat{\Lambda}$, the warmup $\alpha_{n-1}^{\mathrm{SW}} = \frac{n}{C+\frac{n}{\alpha}}$, consistent with (5), gives the multiplier

$$1 - \frac{1}{n} - \frac{q}{C + \frac{n}{\alpha}}. \tag{15}$$

If $C \geq q_{\max} := \lambda_{\max}(Q)$, the multiplier lies in $[-1, 1]$ for every n and every eigendirection. In the exact Gaussian model, warmup therefore prevents amplification by individual steps. Figure 2 compares the resulting MSE curves.

4 Score Alignment and Observable Error

Let $f \in L^2(\pi)$ be a scalar observable with finite second moment. Its centered value separates into a linear score component and a residual. For invertible score covariance, define

$$B_f = \mathrm{Cov}_\pi(f, s)\mathrm{Cov}_\pi(s)^{-1}, \qquad r_f(x) = f(x) - \pi(f) - B_f s(x). \tag{16}$$

Write $\bar{s}_n = \frac{1}{n}\sum_{k=1}^n s(X_k)$ for the sample mean score, $\bar{f}_n = \frac{1}{n}\sum_{k=1}^n f(X_k)$ for the observable mean, and $\bar{r}_{f,n} = \frac{1}{n}\sum_{k=1}^n r_f(X_k)$ for the mean residual. The population least squares coefficient B_f minimizes marginal residual variance, giving the control variate identity $\bar{f}_n - B_f\bar{s}_n - \pi(f) = \bar{r}_{f,n}$ (Mira et al., 2013). Minimizing long run MCMC variance also involves lag covariances. SC-SRMC reduces $\bar{s}_n$ through acceptance and uses the ordinary sample mean. With $a_{k-1} = \frac{1}{k}$ and $\theta_0 = 0$, the full history satisfies, on every trajectory,

$$\bar{f}_n - \pi(f) = B_f\theta_n + \bar{r}_{f,n}. \tag{17}$$

Discarding the first b samples, $0 \le b < n$, gives the exact score contribution

$$\frac{B_f}{n-b}\{n\theta_n - b\theta_b\}. \tag{18}$$

The history bounds on θ_n and θ_b control this contribution.

Assume a residual Poisson equation, smooth dependence on u, and $\mathbb{E}\|\theta_k\|^4 \le C_4/(k+1)^2$ for a finite constant C_4. This additional fourth moment condition, beyond Theorem 1, controls the quadratic change in residual mean under tilting. Appendix B gives

$$\|\bar{r}_{f,n}\|_{L^2} \le \frac{\tau_f}{\sqrt{n}} + \mathcal{O}\left(\frac{1+\log n}{n}\right). \tag{19}$$

Here τ_f bounds the conditional standard deviation of the residual martingale increments. Constants are uniform in n for a fixed kernel family and tilt schedule.

For $X \sim \mathcal{N}(0,1)$, $s(X) = -X$, and $f(X) = X^2 - 1$, we have $B_f = 0$, so the residual carries all estimation error. With exact tilted draws, its leading fluctuation is $\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\varepsilon_k^2 - 1)$. The innovations ε_k are iid standard Gaussian variables, giving asymptotic variance 2 for every fixed α (Appendix B).

For $\mathrm{Var}_\pi(f) > 0$, define the population score alignment as

$$R_f^2 = \frac{\mathrm{Var}_\pi(B_f s(X))}{\mathrm{Var}_\pi(f(X))} \in [0,1]. \tag{20}$$

The value R_f^2 is the fraction of observable variance explained by its score projection. Estimation gains also depend on the tilted kernel's mixing constants. The extension estimates R_f^2 from the second half of a shared preliminary run before evaluation (Appendix E). The public data study separately screens chain movement (Section 5).

5 Experiments

With the proposal and online gradient budget fixed, we test mean accuracy across Gaussian and quartic targets. Matched component comparisons examine warmup and clipping, and nonlinear observables test the scope of the gain. Posterior benchmarks compare dense pMALA, HMC, and adaptive NUTS.

5.1 One configuration across target geometries

We select $(\bar{\alpha}, c, \delta) = (16, 1, 6.4)$ on four rotated Gaussians ($d = 60$, $\kappa \in \{1, 10, 100, 400\}$) and a quartic target ($d = 16$). Here κ is the condition number of the Gaussian precision matrix. Selection minimizes the geometric mean MSE ratio to the reference SC-SRMC configuration $(8, 4, 1.6)$ across these development targets, subject to acceptance and error constraints. The parameters remain fixed during validation on all 11 configurations. Five repeat the development geometries with fresh preliminary runs and evaluation seeds. Six change the dimension, spectrum, rotation, or quartic coefficient. Appendix G gives the search and selection rule.

We compare against preconditioned MALA (pMALA) with a diagonal metric and the same proposal step h, calibrated by preliminary sampling. A second comparison uses multipliers selected separately for each method on the development targets, giving $1.25h$ for SC-SRMC and h for pMALA. Each method uses 24 chains with 6,000 transitions and 6,001 online score calls per chain. The first 1,500 states are excluded from estimation and kept in history. Methods share the metric, initial states, proposal noise, and acceptance uniforms. For chain r, $\widehat{\mu}_r$ is its mean estimate after burnin and μ is the target mean. Coordinate mean MSE averages $d^{-1}\|\widehat{\mu}_r - \mu\|^2$ over chains. Setup and development costs appear in Appendix G.5, including the larger search cost for SC-SRMC.

5.2 Mean estimation with a shared proposal

With the shared proposal, SC-SRMC gives lower observed mean MSE on every validation target (Figure 3). The common metric and step keep the proposal rule fixed, so this comparison isolates the

Table 1: The mean estimation gain persists when the selected configuration is applied to new geometries. Geometric means of SC-SRMC MSE relative to pMALA weight each target equally. The shared step comparison isolates the acceptance change, while joint tuning also allows the proposal step to change. Intervals are descriptive 95% paired bootstrap intervals at matched online budgets.

Validation configurations	Shared step	Joint step tuning
Original geometries (5)	0.163 [0.148, 0.177]	0.136 [0.123, 0.151]
New geometries (6)	0.190 [0.171, 0.212]	0.166 [0.150, 0.184]
All targets (11)	0.177 [0.165, 0.191]	0.152 [0.141, 0.163]

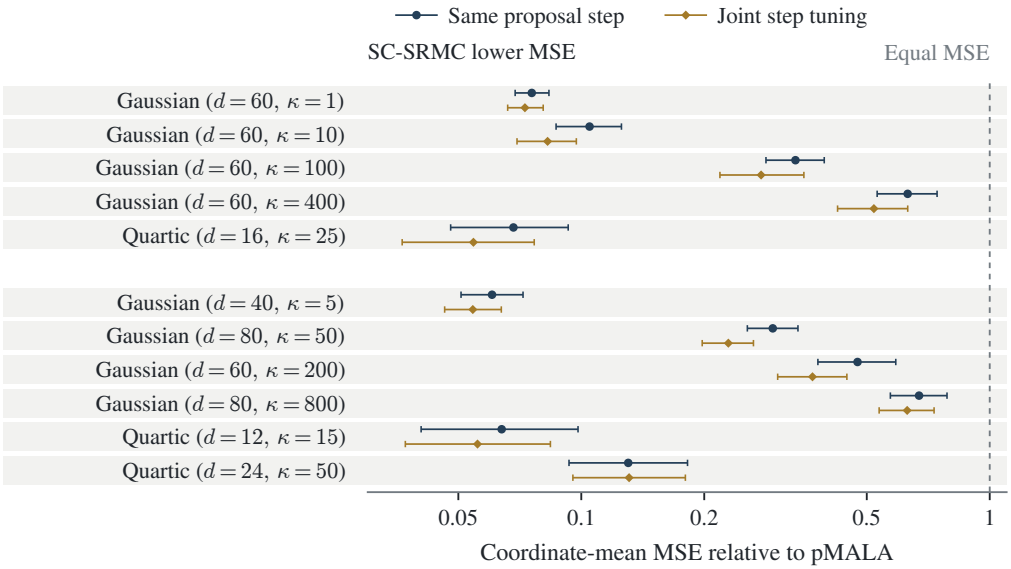


Figure 3: Lower mean error with a shared proposal isolates the benefit of using gradient history in acceptance. The gain extends to geometries excluded from parameter selection. Circles share the pMALA proposal step, and diamonds use separately selected steps. The upper and lower blocks show repeated and new geometries. Points give SC-SRMC coordinate mean MSE relative to pMALA, with descriptive 95% paired bootstrap intervals. The dashed line marks equality. Both comparisons use the fixed configuration and matched online budgets described in Section 5.1.

use of gradient history in acceptance. The geometric mean reduction relative to pMALA is 5.65 fold, rising to 6.59 fold with separately selected steps. The gain is similar on repeated and new geometries (Table 1), supporting reuse of the selected configuration beyond the targets that determined it.

Calibration also matters within SC-SRMC. At the same proposal step, the selected configuration improves on the reference in seven of the 11 prespecified comparisons after Holm correction (Holm, 1979), including on new geometries. These tests assess parameter selection, while the pMALA comparisons above describe the gain from the complete method. Appendix G.2 gives both analyses. Their intervals condition on the target set and realized preliminary runs.

5.3 How calibration and the observable affect the gain

Warmup and the radial cap jointly set the realized strength in Equation (6). For the Gaussian with $\kappa = 400$, the selected configuration permits larger tilts and gives lower mean MSE than the reference, although both are clipped at almost every step. Clipping frequency alone therefore does not measure the acceptance adjustment.

Matched component comparisons hold the eventual strength, metric, and proposal step fixed (Appendix G.3). Spectral warmup has lower geometric mean MSE than warmup selected without the

spectral scale. Fixed strength is slightly more accurate here, and the interval comparing capped and uncapped tilts includes equality. At high anisotropy, the cap can make fixed and gradual schedules yield identical estimates. The exact Gaussian calculation explains how warmup prevents early amplification. For MALA transitions, the realized capped strength and transition correlations also govern the history error.

The benefit also depends on the observable. Gaussian coordinate means have zero projection residual, so the retained score average determines their error. For nonlinear observables, residual fluctuations and their cross term with the score component can also contribute. Tuning for coordinate means increases second moment error on the isotropic Gaussian, illustrating the distinction in Section 4. Appendix G.4 reports these errors and Stein residuals, which measure departures from identities with zero target expectation.

5.4 Posterior targets and stronger transition kernels

We keep the selected configuration and each target's pMALA step fixed across 15 benchmarks, including synthetic and public posteriors with dense preconditioning. Benchmark definitions and random streams are unchanged, with no further parameter selection. All 27 baseline comparisons form one Holm family (Appendix H).

SC-SRMC improves over diagonal pMALA on all targets in the first group. With dense metrics, it improves on both pMALA and HMC on five of eight targets, including logistic and probit posteriors (Figure 8). Gradient history can therefore complement a proposal that already accounts for correlations. The more anisotropic Gaussian improves over pMALA, with an inconclusive HMC comparison.

For the Diabetes, Wisconsin Diagnostic Breast Cancer (WDBC), Wine, and Digits posteriors built from scikit-learn data (Pedregosa et al., 2011), Diabetes and Digits pass the dense pMALA comparison and transition quality checks. Wine also passes the statistical comparison, but the preliminary screen flags low local MALA acceptance on Wine and WDBC. At the same nominal online gradient budget, adaptive BlackJAX NUTS (Cabezas et al., 2024) starts from adaptation endpoints and MALA from displaced states. Under these initializations, NUTS has lower observed MSE on WDBC and Digits, while its comparisons on Diabetes and Wine remain inconclusive.

The selected parameters improve five targets over the reference after Holm correction. On dense logistic and probit, however, smaller score averages accompany larger coordinate error with little change in jump distance. The projection weights and residual cross term are not estimated here, so score norm alone does not explain the change in total error. Both configurations also retain nearly immobile quartic and Poisson chains. Appendix H relates these cases to the mixing conditions in the history bound.

6 Conclusion

SC-SRMC uses score covariance to calibrate tilt strength and reuses past gradients to adjust MALA acceptance. Under uniform stability and regularity conditions, the finite horizon bound controls the score component of estimation error, with explicit dependence on transition mixing. One configuration gives lower coordinate mean error across the 11 validation targets, including six new geometries, at the same online gradient budget. The choice of SC-SRMC parameters also depends on the observable, because residual fluctuations contribute to errors beyond the score component.

Reproducibility statement

Appendices A, B, and C give the assumptions and proofs. Appendices D and E specify the targets, preprocessing, and sampling protocols. Calibration checks, parameter selection, and transfer comparisons appear in Appendices F, G, and H. The supplement contains code, seeds, frozen protocols, source hashes, and chain outputs for these studies. A single command builds the manuscript, and the LaTeX source is compatible with Overleaf.

Artificial intelligence use statement

AI tools assisted with wording and proofreading of the manuscript.

Ethics statement

The tuning study, original confirmation, and eight-target extension use synthetic targets or data. The external replication and earlier exploratory studies use public, deidentified benchmark datasets from scikit-learn. These include medical measurements from the Wisconsin Diagnostic Breast Cancer (WDBC) and Diabetes datasets. The analyses concern sampling performance, with no use of direct identifiers or clinical interpretation. Dataset provenance, hashes of processed data, and loading code are included.

References

Christophe Andrieu, Éric Moulines, and Pierre Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM Journal on Control and Optimization*, 44(1):283–312, 2005. DOI.

Nicolas Brosse, Alain Durmus, Sean Meyn, Eric Moulines, and Sergey Samsonov. Diffusion approximations and control variates for MCMC. *Computational Mathematics and Mathematical Physics*, 64(4):693–738, 2024. DOI.

Alberto Cabezas, Adrien Corenflos, Junpeng Lao, Rémi Louf, Antoine Carnec, Kaustubh Chaudhari, Reuben Cohn-Gordon, Jeremie Coullon, Wei Deng, Sam Duffield, Gerardo Durán-Martín, Marcin Elantkowski, Dan Foreman-Mackey, Michele Gregori, Carlos Iguaran, Ravin Kumar, Martin Lysy, Kevin Murphy, Juan Camilo Orduz, Karm Patel, Xi Wang, and Rob Zinkov. BlackJAX: Composable Bayesian inference in JAX. *arXiv preprint arXiv:2402.10797*, 2024. DOI.

Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214, 2011. DOI.

Shaan Ul Haque and Siva Theja Maguluri. Stochastic approximation with unbounded Markovian noise: A general-purpose theorem. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3718–3726. PMLR, 2025.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. DOI.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

Julian Hofstadler. Solving Poisson's equation for Wasserstein contractive Markov chains. *arXiv preprint arXiv:2602.19119*, 2026. DOI.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Jie Hu, Lingyun Chen, Geeho Kim, Jinyoung Choi, Bohyung Han, and Do Young Eun. Score-repellent Monte Carlo: Toward efficient non-Markovian sampler with constant memory in general state spaces. *arXiv preprint arXiv:2604.22948*, 2026. DOI. ICML 2026 Spotlight.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. DOI.

Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, second edition, 2009. DOI.

Antonietta Mira, Reza Solgi, and Daniele Imparato. Zero variance Markov chain Monte Carlo for Bayesian estimators. *Statistics and Computing*, 23(5):653–662, 2013. DOI.

Radford M. Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng (eds.), *Handbook of Markov Chain Monte Carlo*, chapter 5, pp. 113–162. CRC Press, 2011. DOI.

Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):695–718, 2017. DOI.

Barak A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994. DOI.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. DOI.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(1): 255–268, 1998. DOI.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718, 2021. DOI.

A Proof of the bound over finite horizons

A.1 Assumptions and score covariance

The history update uses correlated scores from a kernel that changes with the history. We analyze this stochastic approximation recursion (Robbins & Monro, 1951) by separating its noise into a martingale difference and a remainder. The Poisson equation gives this decomposition for Markov chain sums (Meyn & Tweedie, 2009).

Condition on the preliminary sampling output. Let Θ be a deterministic convex region containing zero. Almost surely, it must contain both the histories θ_n and the corrected iterates $z_n = \theta_n + a_n G_{u_n}(X_n)$ defined below. Assume $\mathbb{E}\|\theta_0\|^2 < \infty$. Write the radial schedule as $u_n(\theta) = c_n(\theta)\theta$, with $c_n(\theta) \geq 0$. Let $\mathcal{U}$ be a convex tilt domain containing zero and every $u_n(\Theta)$. Assume $u_n : \Theta \to \mathcal{U}$ has Lipschitz constant ℓ_n and $\sup_n \ell_n < \infty$. Define $\underline{\alpha}_n = \inf_{\theta\in\Theta} c_n(\theta)$. For radial clipping on unbounded Θ, this infimum is zero, so $m_n = 1$. An enhanced uniform factor requires a deterministic bound on all of Θ, including the corrected iterates. This domain bound must hold independently of the realized history maximum.

Assumption 1 (Covariance of scores).

For constants $0 < \kappa_s \leq K < \infty$ and every $u \in \mathcal{U}$,

$$\kappa_s I \preceq C(u) := \mathrm{Cov}_{\pi_u}(s) \preceq KI. \tag{21}$$

The log normalizer is twice differentiable and differentiation may pass under the integral.

For a fixed target, MALA has quantitative mixing bounds under strong log concavity and smoothness (Dwivedi et al., 2019). The following assumption states the moment and parameter regularity conditions used for the evolving kernel family.

Assumption 2 (Poisson regularity for controlled kernels).

For each u, let $g_u - P_u g_u = s - S(u)$ and $G_u = P_u g_u$. Define

$$M_{n+1} = g_{u_n}(X_{n+1}) - G_{u_n}(X_n). \tag{22}$$

Here $\mathcal{F}_n$ contains the randomness observed through step n. Assume $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = 0$. For finite constants σ, G, L_G, B_u, c_a, also assume

$$\mathbb{E}[\|M_{n+1}\|^2 \mid \mathcal{F}_n] \le \sigma^2,$$
$$\|G_{u_n}(X_n)\|_{L^2} \vee \|G_{u_n}(X_{n+1})\|_{L^2} \le G, \tag{23}$$
$$\|G_{u_{n+1}}(X_{n+1}) - G_{u_n}(X_{n+1})\|_{L^2} \le L_G \|u_{n+1} - u_n\|_{L^2}, \tag{24}$$
$$\|u_{n+1} - u_n\|_{L^2} \le B_u a_n, \qquad |a_{n+1} - a_n| \le c_a a_n^2. \tag{25}$$

Assume the steps are deterministic, $0 < a_n \le 1$, $m_n a_n \le \frac{2}{3}$, and $a_n L_n^2 \le \frac{m_n}{4}$. Here $m_n = 1 + \kappa_s \underline{\alpha}_n$ and $L_n = 1 + K\ell_n$. Fixed and projected schedules satisfy the map regularity with $\ell_n = \alpha$ and $\ell_n = \alpha_n^{\mathrm{SW}}$, respectively.

Differentiating the normalized exponential family gives Lemma 1. For coordinates i, j,

$$\partial_{u_j} S_i(u) = -\{\pi_u(s_i s_j) - \pi_u(s_i)\pi_u(s_j)\}. \tag{26}$$

This proves (8). If $u = \alpha\theta$, then $S(0) = 0$ and

$$\langle \theta, S(\alpha\theta)\rangle = \int_0^\alpha \theta^\mathsf{T} DS(t\theta)\theta \, \mathrm{d}t = -\int_0^\alpha \theta^\mathsf{T} C(t\theta)\theta \, \mathrm{d}t \le -\alpha\kappa_s \|\theta\|^2. \tag{27}$$

Adding $-\|\theta\|^2$ gives strong contraction. The covariance upper bound and the fundamental theorem of calculus give the Lipschitz bound. For a radial map $u = c(\theta)\theta$, the same ray integral proves the pointwise contraction with $c(\theta)$. The projected map $\theta \mapsto \Pi_{B(0,r)}(\beta\theta)$ has Lipschitz constant β. The composite mean field therefore has Lipschitz constant $1 + K\beta$.

A.2 Recursion for the corrected iterate

The Poisson equation implies

$$s(X_{n+1}) - S(u_n) = M_{n+1} + G_{u_n}(X_n) - G_{u_n}(X_{n+1}) \tag{28}$$
$$= M_{n+1} + G_n - G_{n+1} + \Delta_{n+1}. \tag{29}$$

Here $G_n = G_{u_n}(X_n)$ and $\Delta_{n+1} = G_{u_{n+1}}(X_{n+1}) - G_{u_n}(X_{n+1})$. Define $z_n = \theta_n + a_n G_n$. Direct substitution yields

$$z_{n+1} = z_n + a_n\{h_n(z_n) + M_{n+1} + R_{n+1}\}. \tag{30}$$

Here $h_n(\theta) = S(u_n(\theta)) - \theta$. The remainder is

$$R_{n+1} = \Delta_{n+1} + \frac{a_{n+1} - a_n}{a_n} G_{n+1} + h_n(\theta_n) - h_n(z_n). \tag{31}$$

By Assumption 2 and $\|z_n - \theta_n\|_{L^2} \le a_n G$, with $L_* := \sup_n L_n < \infty$,

$$\|R_{n+1}\|_{L^2} \le \{L_G B_u + c_a G + L_* G\} a_n =: C_R a_n. \tag{32}$$

Set $Y_{n+1} = z_n + a_n\{h_n(z_n) + M_{n+1}\}$ and $e_n = \mathbb{E}\|z_n\|^2$. Martingale orthogonality, strong contraction, and the Lipschitz bound give

$$\mathbb{E}\|Y_{n+1}\|^2 \le (1 - 2m_n a_n + L_n^2 a_n^2) e_n + \sigma^2 a_n^2. \tag{33}$$

Young's inequality with parameter $\frac{m_n}{4}$ gives

$$e_{n+1} \le (1 + \frac{m_n a_n}{4})\mathbb{E}\|Y_{n+1}\|^2 + (a_n^2 + \frac{4a_n}{m_n})\mathbb{E}\|R_{n+1}\|^2 \tag{34}$$
$$\le (1 - \tfrac{3}{2} m_n a_n) e_n + 2\sigma^2 a_n^2 + 5C_R^2 a_n^3. \tag{35}$$

The last line uses $a_n L_n^2 \le \frac{m_n}{4}$, $m_n a_n \le \frac{2}{3}$, $a_n \le 1$, and $m_n \ge 1$. Iterating (35) and using $1 - x \le e^{-x}$ yields the bound for z_n. Finally,

$$\mathbb{E}\|\theta_n\|^2 \le 2e_n + 2a_n^2 G^2. \tag{36}$$

This proves Theorem 1.

A.3 Polynomial step sizes

For fixed α, set $m = 1 + \alpha\kappa_s$ and use $a_n = \frac{\gamma}{(n+n_0+1)^\rho}$ with $\gamma > 0$, $n_0 \geq 0$, and $\rho \in (\frac{1}{2}, 1)$, subject to the step conditions. Comparison over exponential windows gives the following bounds with positive constants $c_\star, C_\star$ independent of n:

$$\sum_{k<n} a_k^2 e^{-c_\star m \sum_{j=k+1}^{n-1} a_j} \leq C_\star \frac{a_n}{m}, \tag{37}$$

$$\sum_{k<n} a_k^3 e^{-c_\star m \sum_{j=k+1}^{n-1} a_j} \leq C_\star \frac{a_n^2}{m}. \tag{38}$$

To bound these sums, let $p \in \{2, 3\}$ and split at $k = \lfloor \frac{n}{2} \rfloor$. For $k \geq \lfloor \frac{n}{2} \rfloor$, we have $a_k \asymp a_n$ and $\sum_{j=k+1}^{n-1} a_j \geq (n-k-1)a_n$. Geometric summation and the smallness condition on the steps bound this part by $\frac{C_\star a_n^p}{m a_n}$. For $k < \lfloor \frac{n}{2} \rfloor$, the accumulated step mass over the second half of the run is at least $c' n^{1-\rho}$ for a constant $c' > 0$. Since $\sum_{k\geq 0} a_k^p < \infty$, this part is bounded by $C_\star \exp\{-c'' m n^{1-\rho}\}$ for a constant $c'' > 0$ and can be absorbed into $\frac{C_\star a_n^{p-1}}{m}$ for large n. Finitely many initial indices enter the constants. This proves (12). For $\rho = 1$, set $b = \frac{3}{2}\gamma(1 + \alpha\kappa_s)$ and compare with the following recurrence, where c_2, c_3 are nonnegative constants:

$$e_{n+1} \leq \left(1 - \frac{b}{n + n_0 + 1}\right) e_n + \frac{c_2}{(n + n_0 + 1)^2} + \frac{c_3}{(n + n_0 + 1)^3}. \tag{39}$$

For $b > 1$, the leading stochastic term is $\mathcal{O}\{\frac{\gamma^2 \sigma_\alpha^2}{(b-1)n}\}$. The transient is $\mathcal{O}(n^{-b})$ when $1 < b < 2$. The term of second order is $\mathcal{O}(n^{-2} \log n)$ at $b = 2$ and $\mathcal{O}(n^{-2})$ thereafter.

A.4 Variation of the calibrated schedule

The projection $\Pi_{B(0,r)}$ is nonexpansive. Let $\beta_n = \alpha_n^{\mathrm{SW}}$. Then

$$\|u_{n+1} - u_n\| \leq \|\beta_{n+1}\theta_{n+1} - \beta_n\theta_n\| \tag{40}$$

$$\leq \bar{\alpha}\|\theta_{n+1} - \theta_n\| + |\beta_{n+1} - \beta_n|\|\theta_n\|. \tag{41}$$

On the stability region $\|\theta\| \leq B$, the covariance bound and Poisson moment bounds in (23) bound the score second moment along the chain. The history recursion then gives $\|\theta_{n+1} - \theta_n\|_{L^2} = \mathcal{O}(a_n)$. With $C = c\widehat{\Lambda}$, direct differentiation of $\frac{n+1}{C + \frac{n+1}{\bar{\alpha}}}$ gives $|\beta_{n+1} - \beta_n| = \mathcal{O}(n^{-2}) = \mathcal{O}(a_n)$ for $a_n = \frac{1}{n+1}$. Hence $\|u_{n+1} - u_n\|_{L^2} = \mathcal{O}(a_n)$.

A.5 Proof of Proposition 1

Write ν for normalized Haar measure, the uniform probability measure on $\mathbb{T}^d$. Let r be the fixed cap radius, and set $\mathcal{U} = \{u : \|u\| \leq r\}$. For the von Mises target,

$$s_j(x) = -\eta \sin x_j, \qquad \frac{\mathrm{d}\pi_u}{\mathrm{d}\nu}(x) \propto \exp\left\{\eta \sum_{j=1}^{d} (\cos x_j + u_j \sin x_j)\right\}.$$

The exponent is uniformly bounded over $\mathbb{T}^d \times \mathcal{U}$. Hence finite constants $0 < m_\pi < M_\pi < \infty$ satisfy

$$m_\pi \leq \frac{\mathrm{d}\pi_u}{\mathrm{d}\nu}(x) \leq M_\pi \quad \text{for all } (x, u) \in \mathbb{T}^d \times \mathcal{U}. \tag{42}$$

All normalizer derivatives are therefore justified by bounded convergence.

The tilted density factorizes by coordinate, so $C(u) = \mathrm{Cov}_{\pi_u}(s)$ is diagonal. Each diagonal entry is $\eta^2 \mathrm{Var}_{\pi_{u_j}}(\sin X_j)$. Each marginal density has full support, and $\sin$ is nonconstant. The variances are strictly positive and continuous in u_j. Compactness of $[-r, r]$ supplies a common positive lower bound κ_0, while boundedness of the score supplies a finite upper bound. This verifies Assumption 1 throughout the convex tilt ball $\mathcal{U}$.

Let $q_{\mathrm{rw}}(y-x)$ be the wrapped Gaussian proposal density at the fixed random walk scale. It is continuous and strictly positive on the compact torus, so $q_{\min} := \min_z q_{\mathrm{rw}}(z) > 0$. The accepted part of the Metropolis transition with a random walk has density

$$q_{\mathrm{rw}}(y-x)\min\left\{1, \frac{\frac{\mathrm{d}\pi_u}{\mathrm{d}\nu}(y)}{\frac{\mathrm{d}\pi_u}{\mathrm{d}\nu}(x)}\right\} \geq q_{\min}\min\{1, \frac{m_\pi}{M_\pi}\} =: \epsilon.$$

Consequently $P_u(x,\cdot) \geq \epsilon\nu(\cdot)$ for every x, u. The common Doeblin minorization contracts centered bounded functions in sup norm at a uniform geometric rate.

The score, $S(u)$, and the log Metropolis acceptance ratio are Lipschitz in u on $\mathcal{U}$. The map $z \mapsto \min(1, e^z)$ is Lipschitz on the bounded range in (42). Thus P_u is Lipschitz in u in the operator norm on bounded functions. The uniformly convergent series

$$g_u = \sum_{k=0}^{\infty} P_u^k\{s - S(u)\}$$

solves the Poisson equation and is uniformly bounded. The resolvent identity relates changes in the Poisson solution operator to changes in the transition kernel, giving Lipschitz dependence on u. The same holds for $G_u = P_u g_u$. The martingale moments in Assumption 2 are then uniformly bounded.

Let $G_\infty = \sup_{u\in\mathcal{U}, x\in\mathbb{T}^d}\|G_u(x)\| < \infty$. Since $a_n \leq 1$ and $\|\theta_n\| \leq \eta\sqrt{d}$, the deterministic ball $\Theta = B(0, \eta\sqrt{d} + G_\infty)$ contains both θ_n and z_n. The projected schedule maps this region into the same tilt ball $\mathcal{U}$. The covariance and Poisson bounds hold throughout this ball. The bounded score also gives the required schedule variation. The steps meet the smallness conditions after a deterministic finite index. The finitely many earlier terms enter C_0. This verifies all conditions of Theorem 1 for the implemented Metropolis kernels with a wrapped random walk.

A.6 Poisson conditions from Wasserstein contraction

Proposition 2.

Let d_V be a metric on X, possibly incorporating a tail weight V. Fix a reference state x_* and define

$$\|f\|_{\mathrm{Lip},V} = \sup_{x\neq y}\frac{\|f(x) - f(y)\|}{d_V(x,y)}.$$

Use this norm on the Banach space $\mathcal{B}_*$ of Lipschitz functions with $f(x_*) = 0$. Assume $\sup_u \pi_u\{d_V(\cdot, x_*)\} < \infty$ and $b = s - s(x_*) \in \mathcal{B}_*$. Assume $P_u f$ is finite at every state for $f \in \mathcal{B}_*$. Define the anchored operator $T_u f = P_u f - P_u f(x_*)$. Suppose these operators satisfy

$$\|T_u\|_{\mathcal{B}_*\to\mathcal{B}_*} \leq \rho < 1, \qquad \|T_u - T_v\|_{\mathcal{B}_*\to\mathcal{B}_*} \leq L\|u - v\|. \tag{43}$$

Then the anchored Poisson solution is

$$g_u = (I - T_u)^{-1}b = \sum_{k=0}^{\infty} T_u^k b, \tag{44}$$

with $g_u(x_*) = 0$ and

$$\|g_u\|_{\mathrm{Lip},V} \leq \frac{\|b\|_{\mathrm{Lip},V}}{1-\rho}, \qquad \|g_u - g_v\|_{\mathrm{Lip},V} \leq \frac{L\|b\|_{\mathrm{Lip},V}}{(1-\rho)^2}\|u - v\|.$$

The contraction gives convergence of the series in $\mathcal{B}_*$. The identity $(I - T_u)g_u = b$ makes $g_u - P_u g_u - s$ constant in the state. The moment assumption ensures π_u integrability. Integrating against the invariant distribution identifies this constant as $-S(u)$, proving (10). The parameter bound follows from the resolvent identity:

$$g_u - g_v = (I - T_u)^{-1}(T_u - T_v)(I - T_v)^{-1}b.$$

For each fixed kernel, Hofstadler (2026) develops the centered Poisson construction under Wasserstein contraction. The operator bounds in (43) give Lipschitz dependence on the tilt for the anchored family. To apply Theorem 1 with these solutions, require the conditional martingale bound, endpoint moments, and L^2 parameter bound in Assumption 2. These additional moment conditions control tail growth along the adaptive chain.

B OBSERVABLE DECOMPOSITION AND SHARPNESS

The coefficient B_f minimizes the marginal residual variance and identifies the component controlled by the score history. For correlated samples, minimizing long run variance also involves temporal dependence. Diffusion based control variates use a Poisson equation to approximate long run variance minimization (Brosse et al., 2024).

From $\theta_n = n^{-1}\sum_{k=1}^n s(X_k)$ and (16),

$$\frac{1}{n-b}\sum_{k=b+1}^{n}\{f(X_k)-\pi(f)\} = \frac{B_f(n\theta_n - b\theta_b)}{n-b} + \frac{1}{n-b}\sum_{k=b+1}^{n} r_f(X_k).$$

This proves both (17) and (18).

For the residual, let $R_f(u) = \pi_u(r_f)$ and suppose

$$v_u - P_u v_u = r_f - R_f(u), \qquad H_u = P_u v_u. \tag{45}$$

The constants below may depend on the fixed kernel family and tilt schedule and are uniform over iteration indices. Assume the associated martingale difference has conditional variance at most τ_f^2. Assume the endpoint H_u terms have L^2 norm at most H_f, and require

$$\|H_{u_k}(X_k) - H_{u_{k-1}}(X_k)\|_{L^2} \le \frac{L_f}{k}.$$

Orthogonality gives $R_f(0) = 0$ and $DR_f(0) = -\mathrm{Cov}_\pi(r_f, s) = 0$. If $\sup_u \|D^2 R_f(u)\| \le C_f$, $\|u_k\| \le \bar{\alpha}\|\theta_k\|$, and $\mathbb{E}\|\theta_k\|^4 \le \frac{C_4}{(k+1)^2}$, then

$$\|R_f(u_{k-1})\|_{L^2} \le \frac{C_f \bar{\alpha}^2 \sqrt{C_4}}{2k}.$$

Summing the residual Poisson decomposition gives

$$\begin{aligned}\|\bar{r}_{f,n}\|_{L^2} \le{}& \frac{\tau_f}{\sqrt{n}} + \frac{2H_f + L_f(1+\log n)}{n} \\ &+ \frac{C_f\bar{\alpha}^2\sqrt{C_4}(1+\log n)}{2n}.\end{aligned} \tag{46}$$

The residual martingale variance bound τ_f^2 depends on the tilted transition family. Orthogonality removes the linear change in the residual mean. The additional fourth moment assumption bounds the quadratic term in L^2, beyond the second moment control in Theorem 1. In the Gaussian example below, centered Gaussian history and $\mathbb{E}\theta_k^2 = \mathcal{O}(k^{-1})$ do give $\mathbb{E}\theta_k^4 = 3(\mathbb{E}\theta_k^2)^2 = \mathcal{O}(k^{-2})$.

The quadratic Gaussian observable has zero score alignment. Take $\pi = \mathcal{N}(0,1)$, $s(x) = -x$, and $f(x) = x^2 - 1$. Under exact draws from the current tilted target, write $X_k = \alpha\theta_{k-1} + \varepsilon_k$ with iid $\varepsilon_k \sim \mathcal{N}(0,1)$. The exact second moment recursion in Appendix C gives $\mathbb{E}\theta_k^2 = \mathcal{O}(k^{-1})$ for $k \ge 1$. Expanding,

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} f(X_k) = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\varepsilon_k^2 - 1) + \frac{2\alpha}{\sqrt{n}}\sum_{k=1}^{n}\theta_{k-1}\varepsilon_k + \frac{\alpha^2}{\sqrt{n}}\sum_{k=1}^{n}\theta_{k-1}^2.$$

The second term has variance $\mathcal{O}(\frac{\log n}{n})$, and the third is $\mathcal{O}_p(\frac{\log n}{\sqrt{n}})$. Both vanish. The innovation term converges to $\mathcal{N}(0,2)$, so the asymptotic variance is 2 for every fixed α.

C EXACT GAUSSIAN DERIVATIONS

Completing the square in $\pi(x)e^{-\alpha\theta^\top s(x)}$ gives $\pi_{\alpha\theta} = \mathcal{N}(\mu + \alpha\theta, Q^{-1})$. Write $X_n - \mu = \alpha\theta_{n-1} + \varepsilon_n$ with iid innovations $\varepsilon_n \sim \mathcal{N}(0, Q^{-1})$. Then $s(X_n) = -\alpha Q\theta_{n-1} + \xi_n$. Here $\xi_n = -Q\varepsilon_n$ and $\mathrm{Cov}(\xi_n) = Q$. This proves (13).

In an eigendirection of Q with eigenvalue q, let $c_n = \mathbb{E}\theta_{n,q}^2$. With $\theta_0 = 0$,

$$c_n = \left(1 - \frac{1+\alpha q}{n}\right)^2 c_{n-1} + \frac{q}{n^2} = q\sum_{k=1}^{n}\frac{1}{k^2}\prod_{j=k+1}^{n}\left(1 - \frac{1+\alpha q}{j}\right)^2. \tag{47}$$

Setting $d_n = nc_n$ gives

$$d_n = \left(1 - \frac{1 + 2\alpha q}{n} + \mathcal{O}(n^{-2})\right) d_{n-1} + \frac{q}{n}, \tag{48}$$

so $d_n \to \frac{q}{1+2\alpha q}$. Since $\bar{X}_n - \mu = -Q^{-1}\theta_n$, (14) follows. For the warmup, the nth draw uses $\alpha_{n-1}^{\mathrm{SW}} = \frac{n}{C + \frac{n}{\bar{\alpha}}}$. Thus (15) is at most 1 and at least $1 - 1 - \frac{q}{C} \geq -1$ whenever $C \geq q_{\max}$.

Figure 2 evaluates these exact second moments in log space with $d = 100$ and $\bar{\alpha} = 4$, without a radial cap. Panel (a) fixes the precision condition number at 400. Precision is inverse covariance, and its condition number is the largest eigenvalue relative to the smallest. Panel (b) evaluates 251 condition numbers at $n = 10{,}000$, with markers at the four original settings. Warmup uses $C = 2q_{\max}$, where $q_{\max}$ is the largest precision eigenvalue. The known covariance comparison rescales the target to unit covariance, giving $q_{\max} = 1$. Fixed strength curves exceed 10^{948} in the transient and 10^{19522} in the condition sweep. The upper strips preserve those magnitudes, while the lower plots resolve the variance reduction regime.

D Experimental details and reproducibility

The studies below evaluate the reference configuration $(\bar{\alpha}, c, \delta) = (8, 4, 1.6)$. Appendix H evaluates the selected configuration $(16, 1, 6.4)$ used in Section 5. Appendix G describes its selection and validation with fresh preliminary runs and evaluation seeds.

D.1 Original confirmatory results

The shared diagonal metric and proposal step isolate the change in acceptance. Lower coordinate mean errors on the quartic and logistic targets extend the empirical result beyond the exact Gaussian model. This paired study uses the reference configuration and forms a separate inferential family from the parameter validation in Appendix G.

Table 2: Changing acceptance lowers mean error with the diagonal proposal held fixed. At the reference configuration, all target comparisons pass joint Holm correction. Entries give SC-SRMC coordinate mean MSE relative to pMALA, with 95% paired bootstrap intervals.

Target	pMALA acceptance	SC-SRMC acceptance	MSE ratio [95% interval]
Rotated Gaussian	0.569	0.617	0.398 [0.334, 0.474]
Quartic	0.563	0.604	0.116 [0.079, 0.157]
Synthetic logistic	0.563	0.618	0.045 [0.032, 0.063]

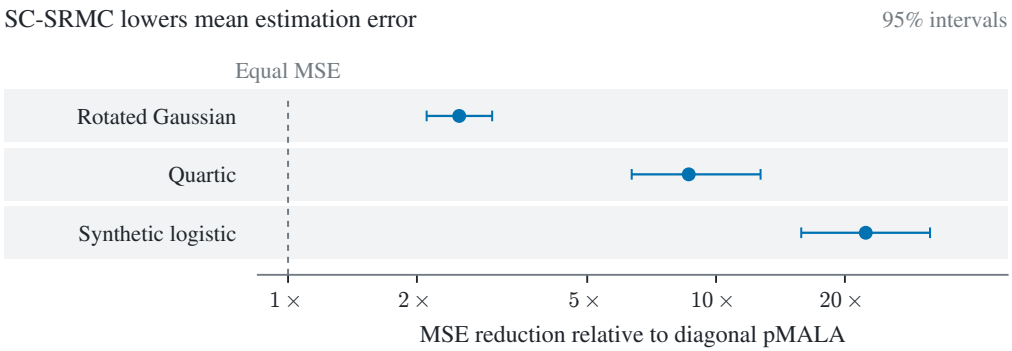


Figure 4: Gradient history yields greater mean accuracy without increasing the online score budget. The reference configuration retains the diagonal pMALA proposal. Points show pMALA MSE relative to SC-SRMC, so values above one indicate improvement. Bars are 95% paired bootstrap intervals, shaded bands group targets, and the dashed line marks equality. Each improvement passes the joint Holm correction reported in Table 2.

D.2 PRESPECIFIED TARGET DEFINITIONS

Table 3: The targets separate anisotropy, nonlinear curvature, and posterior estimation. Analytic or independently estimated means provide a common basis for measuring coordinate error. Here d is parameter dimension and N is observation count.

Target	d	Mean reference
Rotated Gaussian, condition number 100	60	analytic zero
Quartic, coefficient 0.05	16	symmetry (zero)
Synthetic Bayesian logistic, $N = 600$	12	importance sampling

The Gaussian precision eigenvalues span $[1, 100]$. The quartic density is proportional to $\exp\{-x^{\mathsf{T}}Qx/2 - 0.05\sum_j x_j^4/4\}$, and the eigenvalues of its quadratic precision matrix Q span $[1, 25]$. Both spectra are geometrically spaced. Orthogonal rotations come from QR factorizations with fixed seeds. A QR factorization separates a matrix into orthogonal and upper triangular factors. The logistic design has 11 standardized Gaussian features with first order autoregressive correlation, decaying geometrically from 0.6 between adjacent features, an intercept, Bernoulli labels from fixed seeds, and a $\mathcal{N}(0, 4I)$ prior. Importance sampling reweights draws from a proposal with heavy tails, centered at a numerical posterior mode. The final gradient norm at this mode is 1.3×10^{-9}. The reference combines two independent runs of 80,000 draws, with importance effective sample size (ESS) 104,118. This diagnostic expresses weight concentration as an equivalent count of equally weighted draws. The maximum coordinate standard error is 4.24×10^{-4}, and the maximum difference between run means is 0.00163.

D.3 TUNING, INITIALIZATION, AND COSTS

After tuning with an identity metric, a separate preliminary run of 12 chains and 1,500 transitions estimates score covariance. The diagonal metric uses inverse score variances, normalizes by their median, and clips to $[0.05, 20]$. It remains fixed during a second tuning stage and evaluation. Both stages use 12 chains and 500 transitions per candidate. A declared step size is multiplied by 2^k, $k = -6, \ldots, 10$, and the interval bracketing acceptance 0.57 receives six deterministic bisection rounds on the log scale. The baseline evaluation acceptances are $0.569, 0.563, 0.563$.

Evaluation uses 24 paired chains from displaced states with fixed seeds and 6,000 transitions. The first 1,500 transitions are excluded from the mean estimator and remain in the score history. The retained states define checkpoints at 500, 1,500, 3,000, and 4,500 samples. The SC-SRMC parameters $(\bar{\alpha}, c, \delta) = (8, 4, 1.6)$ are fixed across these original confirmatory, strong baseline, and public data studies.

Each method uses 6,001 evaluation score calls per chain. Tuning and calibration use 294,564 score calls per target. Dividing by 24 chains gives 12,273.5 setup calls and 18,274.5 total calls per chain. The diagonal baseline and the spectrally calibrated method share this setup.

D.4 INFERENCE AND SAVED OUTPUTS

For each target, 5,000 paired bootstrap resamples form the ratio interval. Paired sign randomization uses 20,000 draws for one sided tests, and Holm adjustment (Holm, 1979) covers the three targets. The intervals and tests resample evaluation chains with the reference mean and calibration fixed. The strong baseline and public data analyses also condition on these quantities. Section F.2 quantifies sensitivity to the saved importance references. The `confirmatory_*` files in the `results` subdirectory of `project` contain chain errors, acceptance and nonfinite event rates, the four horizon checkpoints, tuning candidates, protocol hashes, and environment versions. No confirmatory chain produces a nonfinite proposal, score, or log acceptance calculation.

D.5 CONFIRMATORY TUNING AND REFERENCE CHECKS

The coarse logistic tuning grid gave preliminary acceptance 0.403, outside the declared interval. Protocol v2 therefore added six rounds of bracket refinement. The two independent reference means

differed by 0.00163. Protocol v3 combined them and saved each chain's coordinate means for reference sensitivity analysis. All reported confirmatory results use v3, with targets, methods, seeds, initialization, chain counts, and horizons unchanged. The `protocol` subdirectory of `project` records these amendments and the source and test hashes.

D.6 Startup and budget accounting

Differentiating the score tilt in a Langevin proposal would require a Hessian vector product, which can be evaluated without forming the full Hessian (Pearlmutter, 1994). SC-SRMC instead applies the tilt in acceptance and reuses cached scores. Its online cost therefore contains no second derivative operations.

Table 4: Cached scores give SC-SRMC and pMALA the same online gradient cost. Matching that cost still leaves sampler setup and initialization different. Setup is reported per target before amortization over R chains, and "Per chain" gives online score calls. Screening and reference construction are accounted for separately in the text.

Study	Method	Setup	R	Per chain	Start
Confirmatory	MALA pair	294,564	24	6,001	Displaced
Extension	MALA pair	350,832	20	5,001	Displaced
Extension	HMC	570,432	20	5,001	Displaced
Public (all)	MALA pair	350,832	12	6,001	Displaced
Diabetes	NUTS	87,688	12	6,001–6,007	Adaptation end
WDBC	NUTS	73,693	12	6,001–6,007	Adaptation end
Wine	NUTS	87,033	12	6,001–6,007	Adaptation end
Digits	NUTS	197,784	12	6,002–6,016	Adaptation end

pMALA and SC-SRMC share tuning and metric calibration. Running them as a pair incurs this setup cost once. Running either method alone divides the full setup cost across its R chains. Including setup, the per chain costs are 18,274.5 calls for the confirmation, 22,542.6 for each extension MALA method, and 33,522.6 for extension HMC. The public MALA screen costs 3,606 calls per chain pair, or 1,803 per method per chain. Including this allocation gives 37,040 calls per public MALA chain. Public NUTS amortized setup is 7,307.33, 6,141.08, 7,252.75, and 16,482 calls per chain for Diabetes, WDBC, Wine, and Digits. Its online cost is then added as shown in Table 4. The separately generated references serve all compared methods and are excluded from algorithm setup and online cost.

NUTS starts evaluation from its adaptation endpoint, while both MALA methods return to the specified displaced initialization after setup. The NUTS comparison therefore matches online gradient budgets, with different startup procedures and total costs. Table 4 reports these costs alongside the online budget.

D.7 Strong baseline design and tuning checks

The extension fixes eight targets, two baselines, 20 chains, 64 linear observables, and a Holm rule over 16 tests. The shared preliminary run uses 12 chains and 1,200 transitions. The dense metric starts from the inverse score covariance, divides it by its median diagonal entry, and clips its eigenvalues to $[0.02, 50]$. This construction requires $\mathcal{O}(d^2)$ storage and $\mathcal{O}(d^3)$ factorization during setup. Tuning MALA with each of the identity and dense metrics uses 10 chains and 800 transitions per candidate, targeting acceptance 0.57. HMC tuning uses 10 chains and 250 transitions per candidate. Evaluation starts from displaced states with fixed seeds and discards the first quarter of the online budget. MALA comparisons use paired chains. HMC comparisons use independent ensembles. Bootstrap and randomization use 5,000 and 20,000 replicates, respectively.

The HMC protocol selects the largest tested step with acceptance at least 0.65 and nonfinite event rate at most 0.001. Tuning starts from the shared preliminary run's terminal states and refines the numerical stability boundary. Evaluation ramps the step from half to full size during the initial discarded transitions. Boundary refinement and the ramp were introduced in protocol v3 after the

incomplete v2 run: Poisson tuning gave acceptance 0.975 at step 0.02, but acceptance 0.431 and nonfinite event rate 0.50 at step 0.04. Version 1 used shortened tuning for a smoke test. Version 3 also widened the HMC evaluation acceptance gate from $[0.55, 0.95]$ to $[0.55, 0.99]$. Poisson's mean acceptance 0.97025 passes the revised gate, although its MSE comparison contributes no joint win. All eight targets were rerun under v3. The protocol archive contains these amendments, quality control records, seven partial v2 target outputs, and source and test hashes.

D.8 Public data targets and reference checks

Diabetes defines Gaussian linear regression with a standardized response and $\mathcal{N}(0, 10I)$ prior. WDBC, Wine classes 0 and 1, and Digits classes 0 and 1 define logistic models. We standardize their features and retain 8, 8, and 12 whitened principal component analysis (PCA) components, respectively. PCA selects orthogonal directions of greatest feature variance, and whitening scales their variances to one. Each model adds an intercept and uses a $\mathcal{N}(0, 4I)$ prior. The resulting dimensions are 11, 9, 9, and 13. Hashes of processed data identify each target. These datasets were also used in exploratory work. The replication fixes methods, seeds, budgets, screening criteria, and inferential rules before evaluation.

Each NUTS chain adapts separately and starts evaluation from its adaptation endpoint. Independent NUTS runs supply reference means. Protocol v3 introduced these choices after v2 failed the WDBC divergence gate and logistic reference precision checks. Version 2 had shared NUTS parameters across dispersed starting states.

The v3 Digits reference had maximum coordinate Monte Carlo standard error (MCSE) 0.0111 against the 0.01 threshold. MCSE estimates uncertainty in a sample mean due to finite sampling. Version 4 uniformly doubled all three logistic reference budgets from 120,000 to 240,000 integration steps per chain, preserving the threshold and evaluation design. All reported public results use v4. The v2 and v3 outputs remain archived with reference files and SHA-256 file integrity hash manifests. The protocol directory records the earlier adaptation quality amendment.

D.9 Exploratory mechanism study

The exploratory study measures alignment for 240 observables across ten targets from independent reference samples. Its analysis ranks and centers quantities within groups defined by target and observable type, bootstraps whole targets, permutes within groups, and repeats the calculation while omitting each target in turn. The adjusted rank correlation is 0.393, the 95% interval from resampled targets is $[0.288, 0.503]$, and the permutation p value is 0.00010. Omitting each target in turn gives a range of $[0.359, 0.420]$. The positive association persists when each target is omitted, providing empirical evidence that greater score alignment accompanies larger MSE gains. Saved exploratory results and the original SHA-256 manifest are maintained separately from the confirmatory results.

E Additional experimental results

The results below use SC-SRMC parameters $(\bar{\alpha}, c, \delta) = (8, 4, 1.6)$ across sampling budgets, baseline samplers, and observables. Appendix G evaluates the configuration $(16, 1, 6.4)$ selected by parameter tuning, and Appendix H applies it to all posterior and strong baseline benchmarks reported here.

E.1 Horizon behavior

Prefixes of the same chains vary the sampling budget while holding tuning and initialization fixed. The declining ratios in Figure 5 show a growing relative gain as the retained averages lengthen. The four checkpoints share trajectories, so their intervals describe each horizon rather than independent replications. Target labels use dimension d, observation count N, precision condition number κ, and quartic coefficient λ.

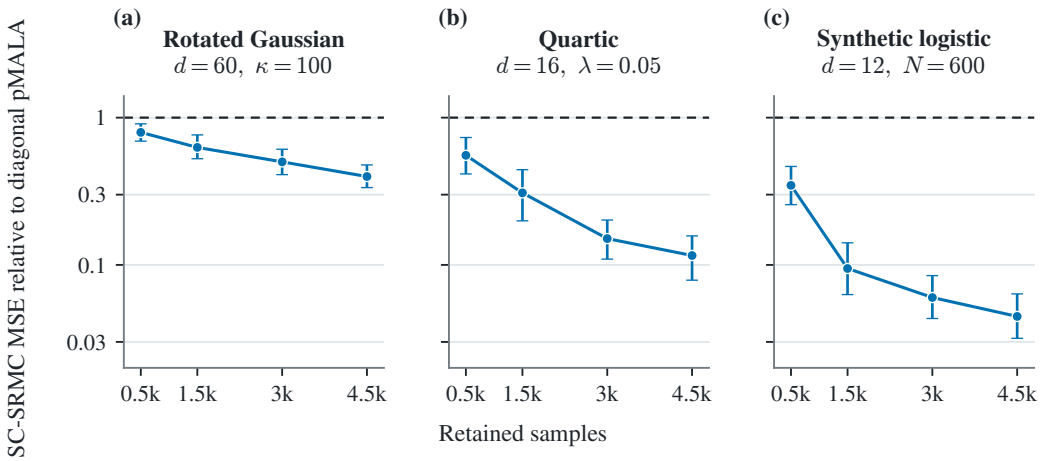


Figure 5: The relative MSE advantage grows as the same trajectories accumulate samples. Holding the reference configuration fixed separates this pattern from repeated tuning. Points give SC-SRMC coordinate mean MSE relative to pMALA, with pointwise 95% paired bootstrap intervals at prespecified checkpoints. Panels share a logarithmic ratio scale, and the dashed line marks equality. Target labels follow Table 3.

The absolute errors in Table 5 show how this relative gain arises: both methods become more accurate, and the SC-SRMC errors fall faster over these checkpoints. Values are rounded to three significant figures.

Table 5: Both samplers become more accurate with more samples, while SC-SRMC gains more in relative terms. Absolute coordinate mean MSEs complement the ratios in Figure 5. Columns give retained sample counts from the same runs, with tuning held fixed.

Target and method	500	1,500	3,000	4,500
Rotated Gaussian, SC-SRMC	0.0539	0.0181	0.00781	0.00401
Rotated Gaussian, pMALA	0.0679	0.0288	0.0156	0.0101
Quartic, SC-SRMC	0.0110	0.00214	0.000562	0.000300
Quartic, pMALA	0.0198	0.00694	0.00372	0.00259
Logistic, SC-SRMC	2.37×10^{-4}	1.99×10^{-5}	6.78×10^{-6}	3.08×10^{-6}
Logistic, pMALA	6.82×10^{-4}	2.10×10^{-4}	1.13×10^{-4}	6.89×10^{-5}

E.2 Strong baseline extension

A dense metric already rescales the proposal to account for correlations. The extension tests whether using score history adds value after that geometric correction. HMC provides a second comparison based on longer trajectories at the same online gradient budget.

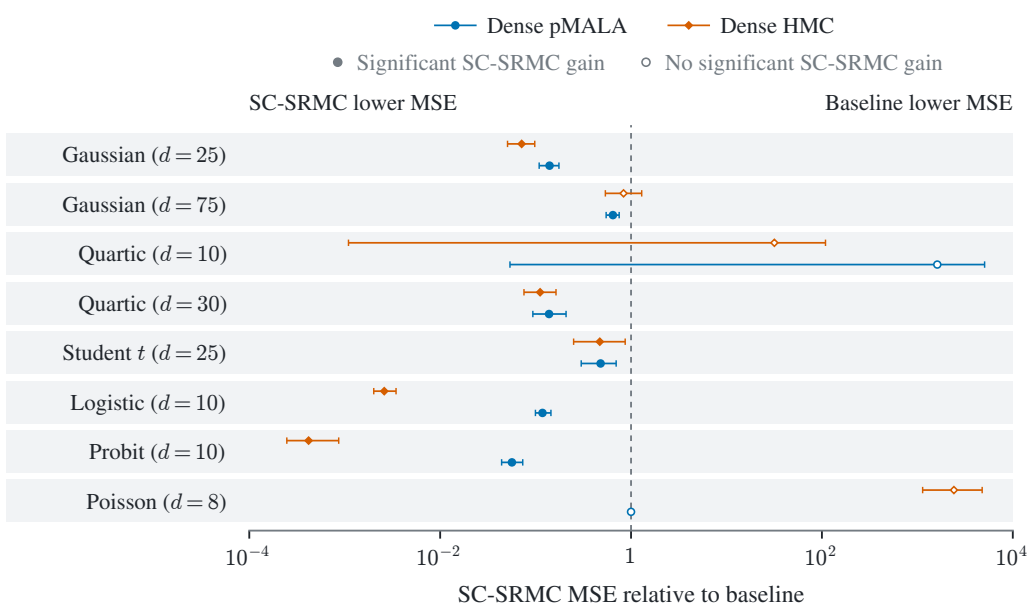


Figure 6: Using gradient history remains useful after dense preconditioning accounts for target correlations. At the reference configuration, circles compare SC-SRMC coordinate mean MSE with dense pMALA, and diamonds compare it with HMC. Bars are 95% bootstrap intervals, paired for pMALA and independent for HMC. Filled markers identify gains after joint Holm correction, and open markers identify the remaining comparisons. Bands group targets, and the dashed line marks equal MSE. Table 6 reports the corresponding tests.

Here d is dimension, N is the number of observations, κ is the Gaussian precision condition number, λ is the quartic coefficient, and ν gives the Student t degrees of freedom.

Every method uses 5,001 online score evaluations per chain. HMC runs 1,000 kernel transitions with $L = 5$ leapfrog steps and five score calls per trajectory. MALA runs 5,000 transitions. Both budgets include one initial score evaluation. Setup costs are separate. Two MALA tuning stages and the shared preliminary run cost 350,832 score calls per target. MALA tuning with an identity metric, the shared preliminary run, and HMC stability tuning cost 570,432.

Each generalized linear model (GLM) reference combines two independent importance runs of 50,000 draws from a proposal with heavy tails. Combined ESS is 69,021 for logistic, 67,459 for probit, and 68,967 for Poisson. The largest coordinate standard error is 5.30×10^{-4}, and the largest difference between run means is 0.00116. The logistic and Poisson terminal gradients at the mode have norms 6.2×10^{-7} and 4.3×10^{-9}, respectively.

Table 6: Dense proposals and gradient history provide complementary gains on several targets. The large quartic and Poisson errors instead reflect nearly immobile chains, examined below. Entries give reference SC-SRMC coordinate mean MSE relative to each baseline. Parentheses contain one sided randomization p values with joint Holm correction across the table.

Target	vs. dense pMALA	vs. dense HMC ($L = 5$)
Gaussian $d = 25, \kappa = 30$	0.140 (0.0008)	0.071 (0.0008)
Gaussian $d = 75, \kappa = 500$	0.645 (0.0008)	0.835 (1.00)
Quartic $d = 10, \lambda = 0.15$	1,618 (1.00)	31.8 (1.00)
Quartic $d = 30, \lambda = 0.03$	0.138 (0.0008)	0.112 (0.0008)
Student t, $d = 25, \nu = 7$	0.480 (0.0008)	0.471 (0.0177)
Logistic $d = 10, N = 400$	0.118 (0.0008)	0.00260 (0.0008)
Probit $d = 10, N = 400$	0.0565 (0.0008)	0.000418 (0.0008)
Poisson $d = 8, N = 500$	1.000 (1.00)	2,415 (1.00)

After joint Holm correction, SC-SRMC improves on both baselines for five targets (Table 6). These gains include logistic and probit posteriors, where the dense metric and history based acceptance change address different parts of the sampling procedure.

Nearly immobile chains account for the large errors on the $d = 10$ quartic and Poisson targets. On the quartic, one of the 20 SC-SRMC chains has acceptance 0.0004 and MSE 8.81, while the other 19 have MSE below 3.4×10^{-5}. On Poisson, ten chains from each MALA method remain at a displaced initialization. HMC moves across all chains, with acceptance from 0.958 to 0.979 and MSE of coordinate means 1.18×10^{-4}. All prespecified aggregate checks pass, with zero nonfinite proposals, scores, or log acceptance calculations. These checks do not detect the nearly immobile chains. The public study screens acceptance for each chain.

E.3 Alignment from preliminary sampling

Score alignment measures how much observable variance lies in the score projection. It does not by itself measure how accurately correlated samples estimate that component. To test its predictive value, the prespecified analysis uses eight directions per target and estimates alignment from the second half of the shared preliminary run. Within each target, it ranks and centers alignment and log MSE gain, then bootstraps whole targets, permutes within targets, and repeats the analysis while omitting each target. The adjusted correlation is 0.230, the 95% interval from resampled targets is $[-0.131, 0.561]$, and the permutation p value is 0.047. Omitting each target in turn gives a range of $[0.126, 0.342]$. The bootstrap interval spans both signs and leaves the association's direction unresolved. Estimates near $R_f^2 = 1$ leave little separation among observables in Figure 7. The difference between marginal alignment and trajectory error is explicit in Equation (17): temporal dependence and residual fluctuations also enter the estimation error. Appendix D reports alignment estimated from independent reference samples.

E.4 Verification on a compact state space

For the $d = 6, \eta = 2$ von Mises target on $\mathbb{T}^6$, both methods use random walk Metropolis (RWM) with a wrapped Gaussian proposal of scale 0.8, 20 paired chains, and 5,000 transitions. Proposition 1 verifies the theorem assumptions for every controlled kernel generated by the schedule with a radial cap. The first 1,250 transitions are discarded from the mean estimator. The retained wrapped coordinates in $[-\pi, \pi)^6$ are averaged arithmetically. The deterministic bound on score covariance $\widehat{\Lambda} = \eta^2 = 4$ sets the spectral scale. The MSE ratio for coordinate means is 0.188, with paired bootstrap 95% interval $[0.134, 0.271]$ and zero nonfinite numerical events. The MSE reduction is empirical, since the arithmetic mean also contains an orthogonal residual that the history bound does not control.

E.5 Adaptive NUTS replication on public data

Diabetes and Digits pass the transition screen and have lower mean MSE than dense pMALA after Holm correction (Table 7). None of the four comparisons with adaptive NUTS passes the correction.

Table 7: Lower mean error supports a gain only when the transition screen also passes. Diabetes and Digits meet both criteria against dense pMALA at the reference configuration. Wine's statistical comparison passes, but its screen fails, as does WDBC's. Entries give SC-SRMC coordinate mean MSE relative to each baseline, with jointly Holm adjusted one sided p values in parentheses. "Screen" reports acceptance and numerical checks for each chain. WDBC denotes Wisconsin Diagnostic Breast Cancer.

Target	vs. dense pMALA	vs. dense NUTS	Screen	Reference
Diabetes $d = 11$	0.380 (0.0227)	0.777 (1.00)	Yes	Analytic
WDBC $d = 9$	0.972 (0.392)	2,569 (1.00)	No	Pass
Wine $d = 9$	0.000527 (0.0020)	0.757 (0.724)	No	Pass
Digits $d = 13$	0.833 (0.0315)	27.9 (1.00)	Yes	Pass

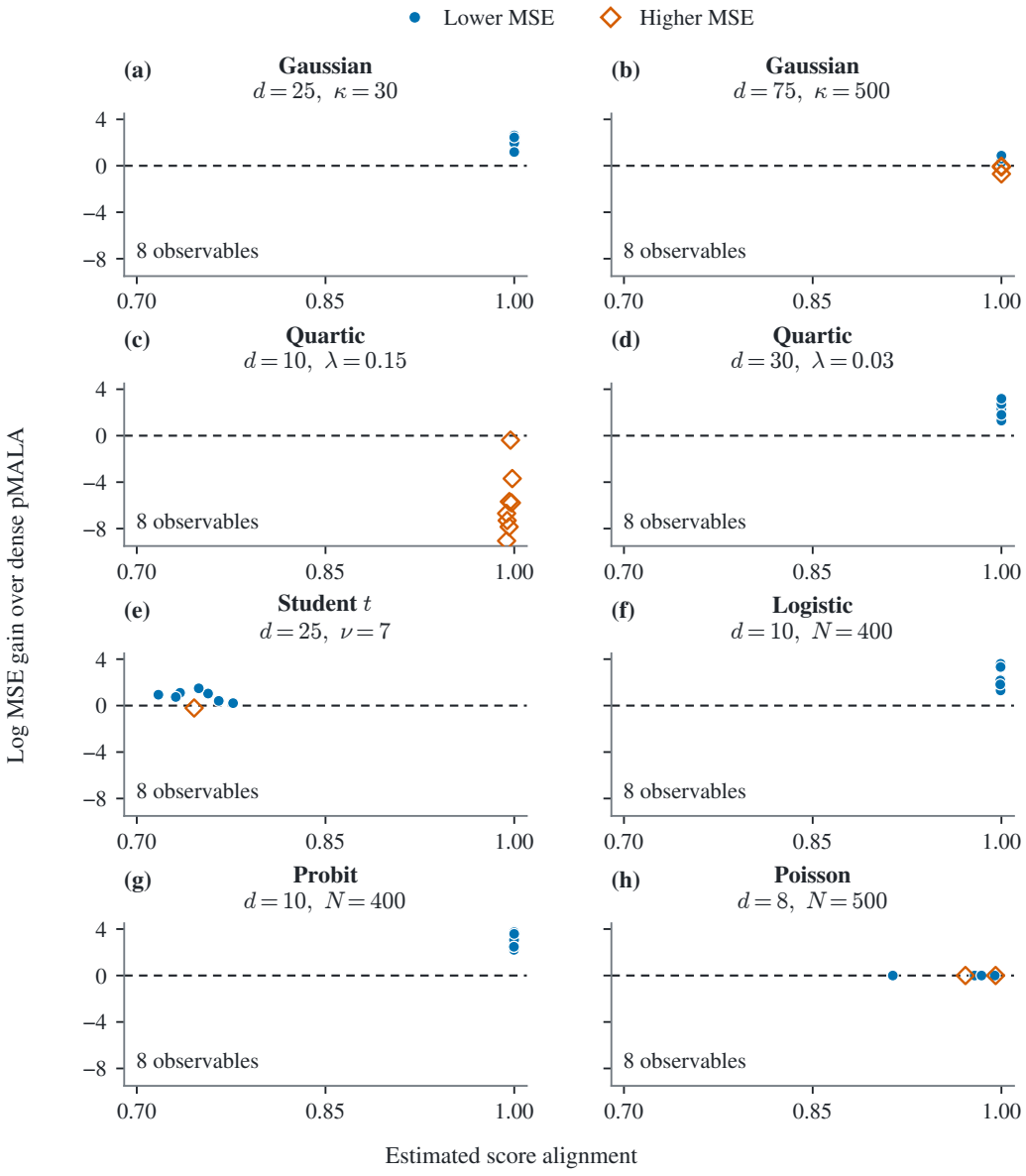


Figure 7: High score alignment alone does not determine the estimation gain. Linear observables with similar alignment can have different MSE gains, consistent with the additional role of transition correlations and residual error in Equation (17). The preliminary run estimates R_f^2, the variance fraction explained by scores. Gain is the natural logarithm of dense pMALA MSE relative to reference SC-SRMC. Filled circles indicate lower SC-SRMC MSE, open diamonds indicate higher MSE, and the dashed line marks equality. Points can overlap near full alignment.

Each method evaluates 12 chains. Dense pMALA and our method use 6,001 online score calls per chain, one initial evaluation and 6,000 transitions. They discard the first 1,500 transitions, with fixed seeds for displaced starts. Each NUTS chain adapts its step size and dense mass matrix for 800 steps from the corresponding displaced state, then evaluates from its adaptation endpoint. Target acceptance is 0.80, and maximum tree depth 8 bounds trajectory doubling. NUTS retains endpoints once cumulative integration cost exceeds 1,500 steps. Completing the final trajectory gives 6,001 to 6,016 score calls. Divergences flag excessive numerical integration error. Mean NUTS acceptance across four targets spans 0.848 to 0.906. Per chain, maximum adaptation divergence rate is 0.02, with zero evaluation divergences. Separate setup costs per target are 350,832 score calls for MALA tuning and the preliminary run. NUTS adaptation uses 73,693 to 197,784 calls per target, or 6,141 to 16,482 per chain. Screening adds 43,272 calls per target.

Before evaluation, the screen runs three disjoint batches of 12 chain pairs for 600 transitions. Every chain must have an acceptance rate of at least 0.05. Each SC-SRMC chain must attain at least 0.2 times its paired pMALA acceptance, with nonfinite event rate at most 0.001. WDBC fails with pMALA and SC-SRMC acceptance minima 0 and 0.0033. Wine's corresponding minima are 0 and 0.628. Table 7 includes the complete evaluations of both flagged targets. On Wine, the large MSE reduction accompanies poor movement of the baseline chains, which the acceptance screen detects.

The Diabetes mean is analytic. Each logistic reference uses four independent dense NUTS chains with seeds disjoint from evaluation, 1,200 adaptation steps at target acceptance 0.90 and maximum tree depth 9, and 240,000 integration steps per chain. Classical split $\widehat{R}$ compares variation within and between chain halves. Chain ESS is the equivalent independent draw count. Rounded diagnostics give maximum $\widehat{R}$ 1.00031, minimum ESS from split chains of raw draws 16,792, maximum coordinate MCSE 0.00783, and zero divergences. Vehtari et al. (2021) describe rank normalization and folding alternatives. Appendix D details protocol amendments.

E.6 Normalizer conditions for the benchmark targets

The tilted density must be integrable to define an invariant probability law. Gaussian tilts preserve integrability and shift the mean. For quartic targets, the fourth degree potential dominates the cubic score term. Compact and Student t targets have bounded scores. Logistic and probit scores grow at most linearly, so Gaussian priors control the tilted tails.

For Poisson regression with coefficients β, design matrix X with rows x_i, counts y, and prior $\mathcal{N}(0, 4I)$, the tilted log density is, up to a constant independent of β,

$$\log\Big\{\pi(\beta)e^{-u^{\mathsf{T}}s(\beta)}\Big\} = -\frac{1}{8}\|\beta\|^2 + \beta^{\mathsf{T}}\Big(X^{\mathsf{T}}y + \frac{u}{4}\Big) - \sum_i (1 - x_i^{\mathsf{T}}u)e^{x_i^{\mathsf{T}}\beta} + \text{const.} \tag{49}$$

When $x_i^{\mathsf{T}}u \le 1$ for every i, the exponential terms are nonpositive and Gaussian tails ensure integrability. A pure intercept tilt above one reverses the exponential terms and causes divergence. In the Poisson experiment, $\widehat{\Lambda} \approx 1.49957 \times 10^7$ and $r \approx 0.00116864$, with $\max_i\|x_i\| \approx 6.14039$. Thus $r\max_i\|x_i\| \approx 0.00718 < 1$ verifies integrability throughout the tilt ball.

For the funnel $z \mid y \sim \mathcal{N}(0, e^y I)$ with Gaussian y, the score component is $s_z(y, z) = -e^{-y}z$. At $u_y = 0$, integrating out z contributes $\exp\{\frac{1}{2}e^{-y}\|u_z\|^2\}$. This factor is nonintegrable for $u_z \neq 0$, so every tilt ball of positive radius around zero contains a nonnormalizable tilt. Score clipping changes the target gradient.

F Calibration ablation and reference sensitivity

F.1 Calibration with MALA transitions

To separate the effects of warmup and clipping, we vary each component while holding the MALA proposal fixed. Targets, seeds, the parameter selection rule, estimands, and analysis were fixed before evaluation in a study separate from the original confirmation. Four rotated Gaussian targets have $d = 60$ and geometrically spaced precision eigenvalues from 1 to $\kappa \in \{1, 10, 100, 400\}$, sharing the original rotation. The fifth target is the existing quartic density with $d = 16$, quadratic condition

number 25, and quartic coefficient 0.05. All five means are exactly zero, and all their score tilts are normalizable for every finite u.

Each target uses the original two stage baseline tuning and a preliminary sample of 12 chains with 1,500 transitions each. All variants share the resulting diagonal metric, baseline MALA step, displaced initial states, proposal noise, and acceptance uniforms. There are 24 independent chain pairs per comparison, 6,000 transitions and one initial score call per chain, with the first 1,500 states discarded only from the primary mean estimator. History always includes retained scores after both acceptance and rejection. Across targets, seeds occupy disjoint blocks rooted at 20260917.

Writing n for the transition index starting at one, SC-SRMC uses $\alpha_{n-1}^{\mathrm{SW}} = n/(4\widehat{\Lambda} + n/8)$ and the common cap $r = 1.6\sqrt{d/\widehat{\Lambda}}$. Fixed + cap uses strength 8 with this same cap. Nonspectral + cap uses strength $n/(C + n/8)$ with the same cap. Its cap retains $\widehat{\Lambda}$, so this comparison isolates the warmup scale. No cap keeps the spectral strength with $u_{n-1} = \alpha_{n-1}^{\mathrm{SW}}\theta_{n-1}$. pMALA uses its usual acceptance rule.

For nonspectral warmup, $C \in \{4, 40, 400, 4000\}$ is selected by a separate sample of 12 chains and 1,000 transitions per candidate, excluding the first 250 from the selection statistic. The statistic is expected squared jumping distance (ESJD). It averages $d^{-1}\sum_j M_{jj}^{-1}(X_{n,j} - X_{n-1,j})^2$ across transitions, including rejections, to standardize movement by the metric. Table 10 reports this statistic alongside acceptance and clipping. A candidate qualifies when every chain has acceptance at least 0.05 and nonfinite rate at most 0.001. Among qualifying candidates, we choose the largest ESJD and break ties using smaller C. If none qualifies, the same maximization over all candidates is recorded as a selection failure. Selection uses the preliminary samples, and evaluation includes every target and chain. All selections qualify. The selected C values are 40, 400, 4, 4000, and 4 in the target order above. Preliminary spectral estimates are 1.383, 10.478, 99.881, 409.286, and 25.245.

Shared setup costs 294,564 score calls per target, or 12,273.5 per evaluation chain. Nonspectral selection costs an additional 48,048 per target, or 2,002 per chain. Total amortized cost is therefore 18,274.5 calls per chain for the other methods and 20,276.5 for this tuned alternative. Online costs are matched across methods, with the additional selection cost assigned to nonspectral warmup.

Table 8: An active radial cap can make different strength schedules produce the same mean estimates. Entries compare reference SC-SRMC MSE with each matched component variant, with pointwise 95% paired bootstrap intervals. * marks significance after joint Holm correction of the two sided paired tests. Transition diagnostics in Table 10 explain where clipping dominates the comparison.

Target	Fixed + cap	Nonspectral + cap	No cap
Gaussian $\kappa = 1$	0.988 [0.947, 1.032]	0.927 [0.886, 0.970]	1.003 [1.000, 1.008]
Gaussian $\kappa = 10$	0.963 [0.871, 1.052]	0.740 [0.687, 0.802]*	0.993 [0.898, 1.104]
Gaussian $\kappa = 100$	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.086 [0.969, 1.219]
Gaussian $\kappa = 400$	1.000 [1.000, 1.000]	1.001 [0.949, 1.054]	1.291 [1.081, 1.549]
Quartic	1.200 [1.038, 1.440]	1.105 [0.961, 1.297]	0.884 [0.812, 0.959]

Conditional on the common preliminary run and selected C, intervals use 5,000 paired bootstrap resamples of chain errors. Two sided paired sign randomization tests use 20,000 draws. Holm correction covers the 15 SC-SRMC comparisons with the three ablated variants jointly. Baseline and early horizon comparisons are descriptive, and the unadjusted intervals are pointwise.

Spectral warmup improves on nonspectral warmup for the Gaussian with condition number 10 after joint correction ($p_{\mathrm{Holm}} = 0.000750$). The other component comparisons remain statistically inconclusive under this testing rule. Fixed strength has lower point MSE on the quartic target, and the uncapped variant has lower point MSE at condition 400. The complete SC-SRMC configuration has lower observed MSE than pMALA on every target in this study.

Early estimates and transition diagnostics locate the effects of the schedule and cap. Table 9 examines the initial history, while Table 10 records how often clipping changes the requested tilt.

Table 9: Early estimates reveal schedule differences that terminal MSE can obscure. Spectral warmup has lower observed early MSE than nonspectral warmup on the less anisotropic Gaussians. Ratios give SC-SRMC MSE relative to each variant using all states through transition n.

Target	Variant	$n = 100$	$n = 500$	$n = 1{,}500$
Gaussian $\kappa = 1$	Fixed + cap	0.998	1.020	1.007
Gaussian $\kappa = 1$	Nonspectral + cap	0.911	0.843	0.901
Gaussian $\kappa = 1$	No cap	0.998	1.023	1.006
Gaussian $\kappa = 10$	Fixed + cap	1.000	1.000	0.946
Gaussian $\kappa = 10$	Nonspectral + cap	0.874	0.845	0.643
Gaussian $\kappa = 10$	No cap	0.983	1.078	0.942
Gaussian $\kappa = 100$	Fixed + cap	1.000	1.000	1.000
Gaussian $\kappa = 100$	Nonspectral + cap	1.000	1.000	1.000
Gaussian $\kappa = 100$	No cap	1.005	0.915	1.034
Gaussian $\kappa = 400$	Fixed + cap	1.000	1.000	1.000
Gaussian $\kappa = 400$	Nonspectral + cap	0.998	0.969	0.926
Gaussian $\kappa = 400$	No cap	1.008	1.005	0.999
Quartic	Fixed + cap	0.990	0.975	1.115
Quartic	Nonspectral + cap	0.990	0.975	1.198
Quartic	No cap	1.037	1.183	0.932

Table 10: Frequent clipping explains why distinct nominal schedules can behave similarly. Acceptance and jump distance remain close among capped variants in this regime. Accept. and Clip give averages across chains and transitions, Min. gives the lowest chain acceptance, and ESJD is the metric standardized expected squared jumping distance defined in the text.

Target	Method	Accept.	Min.	Clip (%)	ESJD
Gaussian $\kappa = 1$	pMALA	0.573	0.560	0.00	0.4653
Gaussian $\kappa = 1$	SC-SRMC	0.611	0.590	0.39	0.5017
Gaussian $\kappa = 1$	Fixed + cap	0.611	0.592	1.98	0.501
Gaussian $\kappa = 1$	Nonspectral + cap	0.611	0.590	0.00	0.5008
Gaussian $\kappa = 1$	No cap	0.612	0.590	0.00	0.502
Gaussian $\kappa = 10$	pMALA	0.567	0.558	0.00	0.08636
Gaussian $\kappa = 10$	SC-SRMC	0.610	0.595	13.27	0.09476
Gaussian $\kappa = 10$	Fixed + cap	0.610	0.595	18.59	0.09487
Gaussian $\kappa = 10$	Nonspectral + cap	0.610	0.597	0.00	0.09408
Gaussian $\kappa = 10$	No cap	0.611	0.597	0.00	0.09505
Gaussian $\kappa = 100$	pMALA	0.560	0.539	0.00	0.01075
Gaussian $\kappa = 100$	SC-SRMC	0.613	0.600	99.90	0.01201
Gaussian $\kappa = 100$	Fixed + cap	0.613	0.600	99.98	0.01201
Gaussian $\kappa = 100$	Nonspectral + cap	0.613	0.600	99.98	0.01201
Gaussian $\kappa = 100$	No cap	0.599	0.583	0.00	0.01184
Gaussian $\kappa = 400$	pMALA	0.552	0.532	0.00	0.002927
Gaussian $\kappa = 400$	SC-SRMC	0.602	0.590	99.95	0.003237
Gaussian $\kappa = 400$	Fixed + cap	0.602	0.590	99.98	0.003237
Gaussian $\kappa = 400$	Nonspectral + cap	0.601	0.578	99.13	0.003231
Gaussian $\kappa = 400$	No cap	0.593	0.581	0.00	0.003248
Quartic	pMALA	0.569	0.557	0.00	0.06088
Quartic	SC-SRMC	0.609	0.597	11.59	0.06771
Quartic	Fixed + cap	0.612	0.602	25.76	0.06834
Quartic	Nonspectral + cap	0.612	0.600	25.76	0.06843
Quartic	No cap	0.609	0.594	0.00	0.06773

At Gaussian conditions 100 and 400, the cap is active for 99.902% and 99.950% of SC-SRMC transitions, respectively. Fixed + cap gives identical chain means and errors at both conditions. ESJD and acceptance are also similar among the capped variants. Clipping sets the effective tilt strength in

this regime, making the choice of warmup less influential. All five variants remain numerically finite, including the uncapped variant. The Gaussian recursion predicts the transient under exact draws from each tilted density. MALA uses correlated transitions, and the ablation shows that clipping can make distinct warmup schedules yield the same estimates.

The reproducibility package contains every chain mean and error, all seven checkpoints (full history $n = 100, 500, 1500$ and retained $n = 500, 1500, 3000, 4500$), pointwise paired intervals, and diagnostics for transitions 1–500, 501–1500, and 1501–6000. It also records every preliminary candidate, the selected step and metric, effective strength, runtime, protocol hash, and implementation and test hashes. The baseline and full SC-SRMC implementations reproduce the original kernel outputs in numerical equivalence tests.

F.2 Sensitivity to estimated reference means

An estimated reference mean contributes uncertainty to the reported MSE. The chain bootstrap intervals condition on that reference. To assess sensitivity, we recompute every comparison based on importance sampling using each independent reference run in place of their average. Evaluation chains and sampler settings stay fixed.

Table 11: The direction of each comparison persists under independent reference estimates. Ratios compare reference SC-SRMC coordinate mean MSE with each baseline. Run 1 and Run 2 replace the combined importance sampling reference while keeping every evaluation trajectory fixed.

Target	Baseline	Combined	Run 1	Run 2
Confirmatory logistic	pMALA	0.044691	0.049984	0.046479
Extension logistic	HMC	0.002595	0.0028872	0.0026966
Extension logistic	pMALA	0.11804	0.13182	0.12011
Extension probit	HMC	0.00041832	0.00048286	0.00044112
Extension probit	pMALA	0.056472	0.064899	0.059119
Extension Poisson	HMC	2415.2	2417	2411.9
Extension Poisson	pMALA	0.99999	0.99999	0.99999

For independent importance runs $\ell = 1, 2$ with separately normalized weights $w_{\ell i}$, draws $x_{\ell i}$, and means $\widehat{\mu}_\ell = \sum_i w_{\ell i} x_{\ell i}$, a first order self normalized importance sampling variance diagnostic for coordinate j of the combined reference is

$$\widehat{V}_j = \frac{1}{4} \sum_{\ell=1}^{2} \sum_i w_{\ell i}^2 (x_{\ell i, j} - \widehat{\mu}_{\ell, j})^2.$$

The delta method gives this first order variance diagnostic. It excludes importance sampling bias. The chain intervals remain conditional on the reference used.

Table 12: Reference precision matters most for the probit comparison. The variance fraction measures estimated mean coordinate reference variance relative to reference SC-SRMC MSE. These first order diagnostics describe the combined importance sampling reference and exclude importance sampling bias. Table 11 checks the resulting sensitivity.

Target	Mean variance	Max. standard error	Variance fraction
Confirmatory logistic	1.38e-07	0.000405	4.47%
Extension logistic	2.13e-07	0.000503	11.7%
Extension probit	8.64e-08	0.000313	26.1%
Extension Poisson	5.4e-08	0.000253	1.9e-05%

Estimated reference variance is largest relative to sampler error on extension probit, reaching 26.1% of SC-SRMC MSE. Even in this case, substituting either independent reference preserves the comparison direction. The substitutions also preserve the near tie with pMALA on Poisson and

HMC's substantially lower error there. Reference variability changes the estimated magnitude of these differences without changing their direction in the saved runs.

For the public logistic datasets, independent NUTS runs with larger budgets provide the reference means. Appendix E reports their MCSE, split chain, and convergence diagnostics. Their comparison intervals also condition on the reference, while the sensitivity calculations above use the saved importance samples.

G Parameter selection and independent validation

We select one SC-SRMC configuration for coordinate mean estimation on five targets. Independent chains test its MSE on both repeated and new geometries at a fixed sampling budget. This study follows a separate protocol from the original confirmation and Appendix F.1. The search grid, selection rule, targets, seed blocks, and analysis are prespecified. Validation fixes the selected parameters and includes every prespecified target.

G.1 Targets and development rule

Development uses the five targets of Appendix F.1: rotated Gaussians with dimension 60 and precision condition numbers 1, 10, 100, and 400, and the quartic density with dimension 16, quadratic condition number 25, and coefficient 0.05. Their exact means are zero. The saved diagonal metrics, baseline steps, and spectral estimates of score covariance are reused, while development uses new random seeds. Each candidate runs 12 chains for 6,000 transitions, discarding the first 1,500 states from estimation while keeping them in score history.

The search first evaluates $\delta \in \{0.8, 1.6, 3.2, 6.4\}$ and $c \in \{1, 2, 4, 8\}$ with $\bar{\alpha} = 8$ and the baseline step. It retains the two best qualifying pairs and evaluates $\bar{\alpha} \in \{4, 8, 16\}$ on each, keeping the reference $(8, 4, 1.6)$ configuration as a fallback. The objective is the geometric mean, over the five targets, of the MSE ratio to reference SC-SRMC. Each target's MSE is averaged across chains. A candidate qualifies if every chain has acceptance at least 0.05, invalid rate at most 0.001, finite diagnostics, and no target has mean MSE above 1.25 times its reference. Ties use canonical configuration order. The saved outputs include every candidate.

The selected SC-SRMC parameters are $(\bar{\alpha}, c, \delta) = (16, 1, 6.4)$. A final development stage separately tests step multipliers $\{0.75, 1, 1.25, 1.5\}$ for the selected SC-SRMC configuration and for pMALA, using each method's reference configuration. The selected global multipliers are 1.25 and 1, respectively. The development objectives are 0.679 for SC-SRMC parameter selection and 0.551 for joint selection, relative to reference SC-SRMC. Candidate reuse yields 27 distinct configurations on each of five targets, or 135 evaluations. The selected parameter values lie on the boundaries of the searched grids.

G.2 Independent evaluation and inference

Validation includes the five original geometries and six new configurations: Gaussians with $(d, \kappa) = (40, 5), (80, 50), (60, 200), (80, 800)$, and quartics with $(d, \kappa, \lambda) = (12, 15, 0.1), (24, 50, 0.02)$. All new configurations use new orthogonal rotations. Gaussian precision eigenvalues are geometrically spaced from 1 to κ. Quartics have density proportional to $\exp\{-x^\mathsf{T} Q x/2 - \lambda \sum_j x_j^4/4\}$ with the same quadratic spectral convention. Symmetry gives exact zero means throughout. New preliminary runs calibrate the baseline step in two stages and estimate score covariance, independently of the evaluation streams.

There are 24 independent chains per method and target, 6,000 transitions, and 4,500 retained states per mean estimate. The score history includes the gradients at all accepted or repeated states from the first transition. Within a target, methods use the same initial states, Gaussian innovations, and acceptance uniforms. The resulting trajectories form paired comparisons. SC-SRMC at the baseline step and pMALA have identical proposal rules. The joint method uses $1.25h$ with the same metric. Seeds for development, preliminary calibration, nonspectral screening, and evaluation occupy separate prespecified blocks.

For every comparison, 5,000 paired chain bootstrap resamples give pointwise 95% intervals. Geometric means across targets use the ratios of MSEs averaged across chains, giving each target equal weight. Their intervals independently resample chain pairs within each fixed target, then recompute the geometric mean. These intervals quantify chain uncertainty for the fixed target set and realized preliminary runs.

The primary family comprises 11 prespecified comparisons of selected and reference SC-SRMC parameters at the same step. Two sided paired sign randomization tests use 20,000 draws with joint Holm correction. Seven comparisons show lower MSE at the adjusted 0.05 threshold. Comparisons with pMALA, joint step tuning, component variants, and earlier horizons are descriptive. On the new Gaussian with $d = 80$ and $\kappa = 50$, the MSE ratio to reference SC-SRMC is 0.973 with interval $[0.874, 1.094]$, spanning equality.

Table 13: The shared proposal gain extends beyond the geometries used for parameter selection. Ratios compare selected SC-SRMC coordinate mean MSE with pMALA, with descriptive 95% paired bootstrap intervals. The upper and lower blocks use repeated and new geometries with independent validation chains. G denotes Gaussian (d, κ) and Q quartic (d, κ, λ). Both step choices follow Section 5.1.

Target	Shared step	Joint tuning
G $(60, 1)$	0.076 [0.069, 0.083]	0.073 [0.066, 0.081]
G $(60, 10)$	0.105 [0.087, 0.125]	0.083 [0.070, 0.097]
G $(60, 100)$	0.334 [0.283, 0.393]	0.276 [0.219, 0.351]
G $(60, 400)$	0.630 [0.530, 0.743]	0.520 [0.424, 0.630]
Q $(16, 25, .05)$	0.068 [0.048, 0.093]	0.054 [0.036, 0.077]
G $(40, 5)$	0.060 [0.051, 0.072]	0.054 [0.046, 0.064]
G $(80, 50)$	0.294 [0.255, 0.339]	0.229 [0.198, 0.264]
G $(60, 200)$	0.475 [0.380, 0.589]	0.368 [0.303, 0.447]
G $(80, 800)$	0.671 [0.571, 0.786]	0.628 [0.536, 0.731]
Q $(12, 15, .1)$	0.064 [0.041, 0.098]	0.056 [0.037, 0.084]
Q $(24, 50, .02)$	0.130 [0.093, 0.182]	0.131 [0.095, 0.180]

Table 14: Parameter selection improves mean estimation beyond the reference calibration while retaining the proposal. Ratios compare selected and reference SC-SRMC, with pointwise 95% paired bootstrap intervals. The shared step comparisons form the primary family of two sided tests reported by p_{Holm}. Joint step tuning is descriptive.

Target	Parameter tuning	p_{Holm}	Joint tuning
G $(60, 1)$	0.604 [0.572, 0.639]	0.00055	0.581 [0.542, 0.622]
G $(60, 10)$	0.791 [0.697, 0.891]	0.01120	0.624 [0.563, 0.697]
G $(60, 100)$	0.819 [0.690, 0.995]	0.11539	0.675 [0.552, 0.843]
G $(60, 400)$	0.754 [0.619, 0.923]	0.05060	0.623 [0.518, 0.760]
Q $(16, 25, .05)$	0.585 [0.449, 0.809]	0.04830	0.467 [0.365, 0.623]
G $(40, 5)$	0.683 [0.617, 0.753]	0.00055	0.613 [0.550, 0.682]
G $(80, 50)$	0.973 [0.874, 1.094]	0.63577	0.757 [0.669, 0.868]
G $(60, 200)$	0.790 [0.680, 0.909]	0.03395	0.613 [0.522, 0.716]
G $(80, 800)$	0.818 [0.710, 0.942]	0.04830	0.765 [0.665, 0.877]
Q $(12, 15, .1)$	0.605 [0.491, 0.739]	0.00495	0.527 [0.427, 0.681]
Q $(24, 50, .02)$	0.711 [0.569, 0.900]	0.05060	0.714 [0.586, 0.901]

G.3 Matched warmup and cap comparisons

The component comparisons keep the eventual tilt strength 16, cap scale 6.4, preliminary metric, and baseline step h. Fixed + cap replaces the warmup by constant strength 16. Nonspectral + cap uses $n/(C + n/16)$, where $C \in \{4, 40, 400, 4000\}$ is selected before evaluation by an independent sample of 12 chains with 1,000 transitions each. It maximizes ESJD standardized by the metric after

excluding the first 250 transitions, with the same acceptance and invalid rate requirements as the original ablation. Ties choose smaller C. Its cap retains the spectral estimate, isolating the warmup scale. The uncapped variant keeps the spectral schedule and removes radial clipping.

Table 15: The realized tilt determines whether schedule changes remain visible in the estimates. An active cap can make fixed and gradual schedules coincide, even though their nominal strengths differ. Ratios compare selected SC-SRMC mean MSE with matched component variants at the shared proposal step, with descriptive pointwise 95% paired bootstrap intervals.

Target	Fixed + cap	Nonspectral + cap	No cap
G $(60, 1)$	0.963 [0.923, 1.001]	0.906 [0.869, 0.946]	1.000 [1.000, 1.000]
G $(60, 10)$	1.041 [0.928, 1.170]	0.708 [0.614, 0.808]	1.072 [0.983, 1.169]
G $(60, 100)$	0.983 [0.936, 1.016]	0.652 [0.562, 0.767]	1.220 [1.024, 1.480]
G $(60, 400)$	1.000 [1.000, 1.000]	0.878 [0.776, 1.004]	0.960 [0.842, 1.085]
Q $(16, 25, .05)$	1.228 [1.004, 1.522]	1.018 [0.832, 1.276]	0.938 [0.807, 1.076]
G $(40, 5)$	1.062 [0.951, 1.168]	1.033 [0.952, 1.117]	1.007 [1.000, 1.023]
G $(80, 50)$	1.123 [1.045, 1.214]	0.676 [0.589, 0.787]	1.080 [0.981, 1.194]
G $(60, 200)$	1.000 [1.000, 1.000]	0.911 [0.801, 1.035]	0.927 [0.817, 1.063]
G $(80, 800)$	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.967 [0.835, 1.123]
Q $(12, 15, .1)$	1.128 [0.919, 1.410]	1.125 [0.954, 1.366]	1.008 [0.946, 1.081]
Q $(24, 50, .02)$	1.084 [0.932, 1.282]	0.939 [0.825, 1.100]	0.818 [0.664, 1.019]

Across all 11 targets, the selected spectral schedule has MSE ratio 1.053 $[1.016, 1.091]$ to fixed strength, 0.882 $[0.848, 0.919]$ to nonspectral warmup selected in preliminary sampling, and 0.995 $[0.959, 1.034]$ to the uncapped variant. Fixed strength has the lower geometric mean MSE in this matched setting, and the interval for the uncapped variant includes equality. At condition numbers 400 and 800, fixed and spectral capped variants give identical chain means. Radial projection couples the nominal schedule to the current history, so different requested strengths can lead to the same realized tilt. The gain over pMALA in this experiment therefore concerns the complete selected configuration. Appendix F.1 compares components under the reference parameters with a separate family of 15 tests.

G.4 Observable and chain diagnostics

Selection targets MSE of the coordinate means. Second moments and diagonal Stein identities provide separate diagnostics of nonlinear observables. Write $\widehat{\mathbb{E}}$ for an empirical average over one chain's retained samples after burnin. For Gaussian target marginal variance v_j, the reported relative second moment error is the average of $(\widehat{\mathbb{E}}[X_j^2]/v_j - 1)^2$ across chains and coordinates. All means are zero. The Stein residual is $1 + \widehat{\mathbb{E}}[X_j s_j(X)]$, with target value zero. Its squared error is averaged in the same way. The outputs also include quartic second moment estimates, evaluated without exact reference values.

Table 16: Parameters selected for coordinate means need not improve nonlinear observables. Ratios compare selected and reference SC-SRMC for relative second moment MSE and diagonal Stein residual MSE. Values above one indicate larger error. F shares the proposal step, and J uses joint step tuning. Exact quartic second moments are unavailable (n.a.).

Target	Second (F)	Second (J)	Stein (F)	Stein (J)
G $(60, 1)$	1.67	2.62	1.67	2.62
G $(60, 10)$	1.23	1.48	1.54	2.37
G $(60, 100)$	1.04	0.83	1.44	1.82
G $(60, 400)$	0.80	0.68	1.01	0.99
Q $(16, 25, .05)$	n.a.	n.a.	2.02	3.50
G $(40, 5)$	1.62	2.49	1.88	3.10
G $(80, 50)$	1.16	0.99	1.49	1.89
G $(60, 200)$	1.01	0.86	1.20	1.33
G $(80, 800)$	0.81	0.80	0.89	0.83
Q $(12, 15, .1)$	n.a.	n.a.	2.28	4.09
Q $(24, 50, .02)$	n.a.	n.a.	1.93	3.20

The isotropic Gaussian's relative second moment root mean squared error (RMSE) is 5.7%, 7.4%, and 9.2% for reference SC-SRMC, the selected configuration, and joint tuning. This quadratic observable has zero linear projection onto the score, so its error lies entirely in the residual of Equation (17). Reducing the score history does not directly control that error. Several other configurations also have increased second moment or Stein residual errors.

All validation outputs are finite, with zero invalid proposals and minimum acceptance 0.52 per chain across the displayed methods.

G.5 Search costs and reproducibility

Table 17: A reusable configuration has a development cost beyond the matched online sampling budget. Separating parameter search, target calibration, component screening, and evaluation makes that distinction explicit. Entries count new score calls at each stage. Reused preliminary runs are accounted for in the text.

Stage	New score calls
Parameter development	9,721,620
Validation baseline calibration and preliminary runs	3,240,204
Nonspectral component screening	528,528
Validation trajectories	11,089,848
Total	24,580,200

Development includes 8,281,380 calls for SC-SRMC candidates and 1,440,240 for pMALA candidates. Development reuses earlier preliminary runs that cost 1,472,820 score calls. Each new validation preliminary run costs 294,564 calls per target, or 12,273.5 per chain when amortized over 24 chains. Setup plus online cost is 18,274.5 per chain before allocating global development cost. Nonspectral screening adds 48,048 calls per target. The globally selected pMALA step equals its original step, so those two labels share one trajectory record and are counted once. Eight method labels correspond to seven actual configurations per target, giving 77 validation evaluations and 1,848 executed chain trajectories.

The `parameter_tuning_20260918` result bundle stores the frozen protocol, selected-configuration hash, all development candidates, all chain and coordinate outputs, seven checkpoints, three diagnostic windows, preliminary choices, and complete cost accounting. The parameterized kernel passes 12 tests against the frozen implementation. A replay of the reference configuration over 6,000 transitions is bitwise identical and adds 72,012 audit calls. Independent recomputation recovers every chain MSE exactly and all primary intervals, tests, and Holm adjustments to numerical

precision. Source, protocol, selection, and result hashes connect the manuscript tables to these saved data.

H SELECTED PARAMETERS ON POSTERIOR TARGETS AND STRONG BASELINES

We test whether the SC-SRMC parameters selected in Appendix G improve coordinate mean estimation on posterior targets and with stronger proposals. The configuration $(\bar{\alpha}, c, \delta) = (16, 1, 6.4)$ is fixed across all 15 benchmarks from Appendices D and E. Each target retains its original pMALA step, metric, preliminary sampling seed, initial states, reference mean, and evaluation streams. The reused geometries and streams measure transfer within this benchmark collection, with no selection from the new outcomes.

H.1 SHARED PROPOSALS, BUDGETS, AND STATISTICAL COMPARISONS

The three MALA methods share initial states, proposal noise, and acceptance uniforms within each target. Replays of pMALA and SC-SRMC with the reference parameters $(8, 4, 1.6)$ reproduce archived acceptance and invalid rates exactly, with coordinate estimates agreeing to numerical precision. HMC and NUTS use their archived trajectories with independently seeded evaluation streams, unchanged data, reference means, and settings.

The diagonal group uses 24 chains with 6,000 transitions and excludes the first 1,500 states from estimation. The dense group uses 20 chains and a budget of 5,001 score calls per chain. This gives 5,000 MALA transitions or 1,000 HMC transitions with five leapfrog steps, including one initial score call. Both discard the first quarter of their transitions. The public group uses 12 chains. MALA runs 6,000 transitions and discards 1,500 states, while NUTS excludes states through 1,500 integration steps and completes the trajectory that crosses its 6,000 step budget. Its realized cost is 6,001 to 6,016 score calls per chain, including the initial score. NUTS starts at its adaptation endpoint, whereas MALA starts from the original displaced states. Every SC-SRMC run retains discarded states in its score history.

The outcome is the mean over chains of $d^{-1}\|\widehat{\mu}_r - \mu\|^2$. Here r indexes chains, $\widehat{\mu}_r$ is the mean estimate after burnin, and μ is the target mean, analytic or independently estimated. The primary family contains all 27 comparisons of selected SC-SRMC with pMALA, HMC, or NUTS. We use 20,000 randomizations and one sided tests for lower MSE, with one joint Holm correction. MALA comparisons use paired sign randomization, and HMC and NUTS comparisons use independent label permutations. A separate family of 15 two sided paired tests compares selected and reference SC-SRMC configurations, again with Holm correction. Pointwise 95% intervals use 5,000 bootstrap resamples of whole chains, paired for the MALA methods and independent for HMC and NUTS. The intervals condition on these targets, realized preliminary runs, and reference means.

A qualified gain must pass both the adjusted statistical test and the prespecified quality checks. All targets and chains enter the tables. These checks retain the original reference and baseline criteria and require finite selected estimates with at most 0.1% invalid proposals. For each public target, we repeat three preliminary batches of 12 chains and 600 transitions for pMALA and selected SC-SRMC. The screen requires every chain to have acceptance at least 0.05, a selected to pMALA acceptance ratio of at least 0.2, and an invalid rate at most 0.001. A failed screen does not remove a target or stop evaluation. The dense benchmark criteria use aggregate baseline acceptance, so individual immobile chains remain visible in the diagnostics.

The Poisson posterior also satisfies a uniform integrability check over the selected tilt ball. With radius $r = \delta\sqrt{\frac{d}{\bar{\Lambda}}}$, the reconstructed calibration gives $r \max_i \|x_i\| = 0.028704 < 1$ for the design rows x_i. The exponential likelihood terms in every tilted log density therefore retain negative coefficients, and the Gaussian prior controls the remaining linear terms.

H.2 WHAT CARRIES OVER TO THE POSTERIOR BENCHMARKS

With the proposal steps fixed, the selected SC-SRMC configuration improves over diagonal pMALA on all three targets and over both dense baselines on five of eight targets (Figure 8). On the dense Gaussian, Student t, logistic, and probit targets, score history reduces coordinate mean error even

after the proposal accounts for correlations. The highly anisotropic Gaussian improves over pMALA, with an inconclusive HMC comparison. Tables 18 and 19 report the tests for the diagonal and dense groups within the joint comparison family.

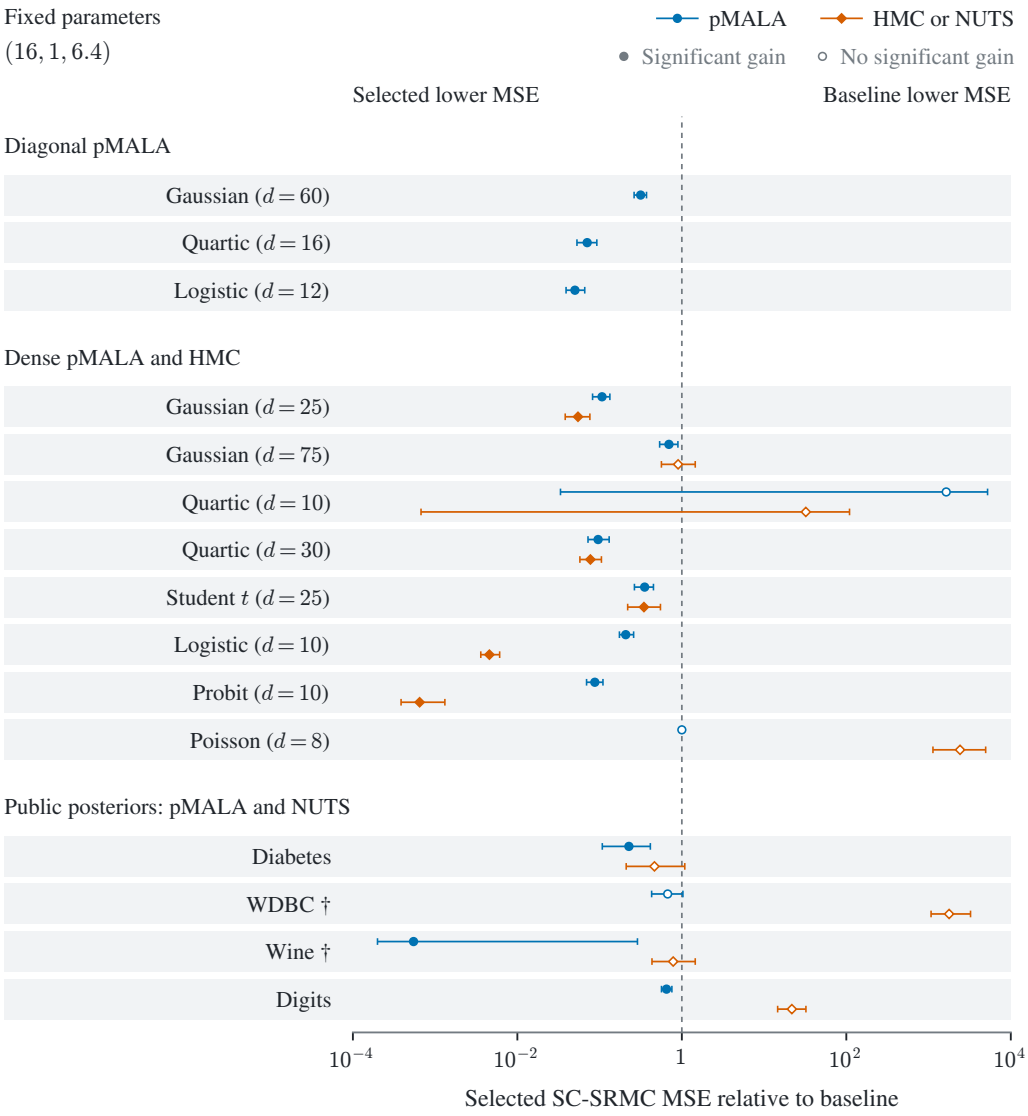


Figure 8: A fixed configuration retains mean estimation gains on several posterior targets and with dense proposals. Points give selected SC-SRMC MSE relative to each baseline. Circles compare pMALA, diagonal in the first group and dense thereafter. Diamonds compare HMC in the middle group and NUTS in the public group. Bars are pointwise 95% bootstrap intervals, paired for pMALA and independent otherwise. Filled markers pass the jointly Holm adjusted gain tests. Open markers do not. Daggers mark failed quality checks, which preclude qualified gains. The dashed line marks equality. The text explains startup differences and chain movement.

Table 18: The selected configuration preserves the shared proposal advantage across the diagonal benchmarks. Ratios give SC-SRMC coordinate mean MSE relative to pMALA, with pointwise 95% paired bootstrap intervals. One sided p values belong to the joint Holm family across all selected baseline comparisons.

Target	Baseline	MSE ratio [95% interval]	Holm p
Gaussian ($d = 60$)	pMALA	0.315 [0.263, 0.372]	0.00135
Quartic ($d = 16$)	pMALA	0.0706 [0.053, 0.0925]	0.00135
Logistic ($d = 12$)	pMALA	0.0501 [0.0391, 0.066]	0.00135

Table 19: Dense preconditioning leaves room for gradient history to improve mean estimation. Ratios compare selected SC-SRMC coordinate mean MSE with each baseline at unchanged proposal steps. Intervals are pointwise 95% bootstrap intervals, paired for pMALA and independent for HMC. One sided p values use the joint Holm correction across all selected baseline comparisons.

Target	Baseline	MSE ratio [95% interval]	Holm p
Gaussian ($d = 25$)	pMALA	0.107 [0.0821, 0.133]	0.00135
	HMC	0.0546 [0.0382, 0.0762]	0.00135
Gaussian ($d = 75$)	pMALA	0.697 [0.537, 0.896]	0.0451
	HMC	0.901 [0.564, 1.46]	1
Quartic ($d = 10$)	pMALA	1.64×10^{3} [0.0334, 5.22×10^{3}]	1
	HMC	32.2 [6.74×10^{-4}, 109]	1
Quartic ($d = 30$)	pMALA	0.0959 [0.0723, 0.131]	0.00135
	HMC	0.0773 [0.0576, 0.105]	0.00135
Student t ($d = 25$)	pMALA	0.353 [0.265, 0.452]	0.00135
	HMC	0.346 [0.219, 0.549]	0.00135
Logistic ($d = 10$)	pMALA	0.208 [0.174, 0.259]	0.00135
	HMC	0.00457 [0.00359, 0.00609]	0.00135
Probit ($d = 10$)	pMALA	0.0875 [0.0694, 0.11]	0.00135
	HMC	6.48×10^{-4} [3.85×10^{-4}, 0.00131]	0.00135
Poisson ($d = 8$)	pMALA	0.999957 [0.999846, 1.000013]	0.782
	HMC	2.42×10^{3} [1.13×10^{3}, 4.95×10^{3}]	1

Diabetes and Digits pass both the comparison with dense pMALA and the public transition screen. Wine passes the statistical comparison, but pMALA has an immobile chain in preliminary sampling. WDBC also fails the local acceptance screen. With NUTS starting from its adapted endpoint and MALA from displaced states, NUTS has lower observed MSE on WDBC and Digits. The NUTS comparisons on Diabetes and Wine are inconclusive under the joint testing rule (Table 20).

Table 20: Diabetes and Digits retain qualified gains over dense pMALA after transferring the selected configuration. NUTS gives lower observed MSE on WDBC and Digits under the different startup procedures described in the text. Ratios give SC-SRMC coordinate mean MSE relative to each baseline, with pointwise 95% bootstrap intervals, paired for pMALA and independent for NUTS. One sided p values use the joint Holm correction across all selected baseline comparisons. Daggers mark failed quality checks.

Target	Baseline	MSE ratio [95% interval]	Holm p
Diabetes	pMALA	0.227 [0.108, 0.414]	0.0078
	NUTS	0.464 [0.211, 1.09]	0.684
WDBC†	pMALA	0.673 [0.429, 1.03]	0.778
	NUTS	1.78×10^{3} [1.07×10^{3}, 3.24×10^{3}]	1
Wine†	pMALA	5.46×10^{-4} [1.99×10^{-4}, 0.289]	0.00135
	NUTS	0.785 [0.433, 1.46]	1
Digits	pMALA	0.651 [0.565, 0.757]	0.0078
	NUTS	21.8 [14.6, 32.3]	1

Immobile chains dominate the large errors on two synthetic targets. Both SC-SRMC configurations retain one nearly immobile chain on the $d = 10$ quartic target and ten immobile chains on Poisson. The quartic error is dominated by that chain, producing a wide bootstrap interval. The Poisson chains leave selected and reference MALA errors almost unchanged, while HMC moves throughout its ensemble. A bounded tilt does not ensure that the chain leaves these states. The finite horizon bound assumes mixing within the tilted family. Minimum chain acceptance in Table 21 makes the poor movement visible even when aggregate checks pass.

Table 21: Average acceptance can hide nearly immobile chains. Minimum acceptance exposes poor movement on quartic and Poisson targets that aggregate checks miss. For selected SC-SRMC, acceptance gives the chain mean and minimum, and invalid rate gives the maximum. Quality combines reference, baseline, numerical, and preliminary public transition checks.

Target	Mean acceptance	Minimum	Invalid rate	Quality
Diagonal preconditioning				
Gaussian ($d = 60$)	0.583	0.564	0	Pass
Quartic ($d = 16$)	0.612	0.6	0	Pass
Logistic ($d = 12$)	0.617	0.601	0	Pass
Strong transition baselines				
Gaussian ($d = 25$)	0.605	0.594	0	Pass
Gaussian ($d = 75$)	0.54	0.531	0	Pass
Quartic ($d = 10$)	0.583	4.00×10^{-4}	0	Pass
Quartic ($d = 30$)	0.588	0.573	0	Pass
Student t ($d = 25$)	0.653	0.629	0	Pass
Logistic ($d = 10$)	0.607	0.593	0	Pass
Probit ($d = 10$)	0.601	0.586	0	Pass
Poisson ($d = 8$)	0.437	0	0	Pass
Public posterior targets				
Diabetes	0.6	0.589	0	Pass
WDBC	0.862	0.755	0	Fail
Wine	0.76	0.739	0	Fail
Digits	0.847	0.821	0	Pass

H.3 Tilt strength and observable error

Relative to $(8, 4, 1.6)$, the selected parameters give lower observed MSE on nine targets. Five improvements pass the separate Holm correction, and the reference has significantly lower error on the dense logistic and probit targets (Table 22). The other eight comparisons remain inconclusive. On these two GLMs, the selected parameters reduce both the final history norm and the root mean square of the retained score averages, while increasing coordinate mean error. After burnin, ESJD changes by less than 1%. Clipping is also less frequent with the selected cap. The observable decomposition leaves the residual outside the score span uncontrolled by the score average. Here the projection weights and residual cross term are not estimated, so the smaller score norm establishes neither a smaller projected error nor the cause of the total error change.

Table 22: The preferred configuration varies across targets even with proposals held fixed. Dense logistic and probit favor the reference configuration despite the selected configuration's gains elsewhere. Ratios compare selected and reference SC-SRMC mean MSE, with pointwise 95% paired bootstrap intervals. Two sided p values use a separate joint Holm correction across these configuration comparisons.

Target	Selected MSE relative to reference	Holm p
Diagonal preconditioning		
Gaussian ($d = 60$)	0.792 [0.674, 0.915]	0.0436
Quartic ($d = 16$)	0.609 [0.51, 0.763]	0.0176
Logistic ($d = 12$)	1.12 [0.906, 1.4]	1
Strong transition baselines		
Gaussian ($d = 25$)	0.764 [0.655, 0.885]	0.0375
Gaussian ($d = 75$)	1.08 [0.888, 1.31]	1
Quartic ($d = 10$)	1.01 [0.537, 1.01]	1
Quartic ($d = 30$)	0.693 [0.599, 0.82]	0.00715
Student t ($d = 25$)	0.736 [0.607, 0.994]	0.625
Logistic ($d = 10$)	1.76 [1.65, 1.89]	7.50×10^{-4}
Probit ($d = 10$)	1.55 [1.4, 1.71]	7.50×10^{-4}
Poisson ($d = 8$)	0.999964 [0.999863, 1.000015]	1
Public posterior targets		
Diabetes	0.597 [0.339, 1.85]	1
WDBC	0.692 [0.443, 1.03]	1
Wine	1.04 [0.734, 1.43]	1
Digits	0.781 [0.683, 0.871]	0.0138

Reference sensitivity checks substitute each independent importance run for the combined reference on the four synthetic posteriors. For WDBC, Wine, and Digits, they substitute each of the four saved NUTS reference chain means. The evaluation trajectories are unchanged. The logistic and probit increases relative to the reference configuration persist under every substitution, as do their gains over pMALA and HMC. All seven posterior cases retain the directions of their MALA comparisons. The first order reference variance is about 6.6% and 16.9% of selected MSE on dense logistic and probit, respectively. The substitutions assess reference variability but do not bound importance sampling bias. NUTS evaluation coordinate means were not archived, so this sensitivity analysis does not recompute its public comparisons.

H.4 Execution and provenance

Setup costs are reported in Appendices D and E. The new runs reuse saved references, HMC runs, and NUTS runs. Selected sampling uses 1,520,280 score calls, the two MALA controls use 3,040,560, metric reconstruction uses 226,980, and public screens use 173,088. Technical replay checks before freezing add 2,792,232 calls. These costs are separate from the per chain online comparison budgets. The supplement records the frozen protocol, source and data hashes, chain estimates, preliminary metrics, screens, and execution ledger. Selected and reference configurations have separate result files.