Geometry-Aware Score-Repellent Monte Carlo

Jie Hu[1] **Lingyun Chen**[2] **Do Young Eun**[2]
[1]Oakland University [2]NC State University
jiehu@oakland.edu, lchen47@ncsu.edu, dyeun@ncsu.edu

Abstract

The recently proposed Score-Repellent Monte Carlo (SRMC) method improves MCMC sampling by tilting the target density π in $\mathbb{R}^d$ with the factor $\exp(-\alpha\,\theta_n^\top s(x))$, where the score function $s(x) = \nabla_x \log \pi(x)$, and θ_n is the score average from past samples. However, a single scalar repellence strength $\alpha \geq 0$ cannot handle both the steep and the flat directions in anisotropic energy landscapes at once. We introduce *Geometry-Aware SRMC* (GA-SRMC), which generalizes the Euclidean alignment to $\theta^\top M s(x)$ for any symmetric positive-definite matrix M. A stochastic-approximation central limit theorem under a trace constraint on M identifies the trace-normalized inverse Fisher $M_{\mathrm{opt}} \propto S^{-1}$ as the unique minimizer of the worst-direction covariance bound, elevating the Fisher information $S = \mathrm{Cov}_\pi(s, s)$ from a proposal-side preconditioner (e.g., FisherMALA, natural gradient) to the bound-optimal repulsion-side matrix, with a single online estimate $\widehat{S}_n$ serving both roles at no additional computational cost. The full-Fisher online metric M_n converges almost surely to M_{opt} and satisfies a CLT, providing a theoretical guarantee for the online implementation of the GA-SRMC framework. On an 8-target simulation spanning the condition number $\kappa \in [1.3, 100]$ and dimension $d \in [2, 785]$, GA-SRMC dominates SRMC in ESS by $\mathbf{4.3}\times$ on anisotropic Gaussian ($\kappa = 100$) and $\mathbf{6.8}\times$ on MNIST under the MALA baseline, by $\mathbf{16}$–$\mathbf{32}\%$ on three well-conditioned logistic-regression posteriors, and by up to $+\mathbf{48}\%$ under FisherMALA; gains peak on the most anisotropic targets, and posterior-mean MSE is lower than with SRMC in all 16 full-Fisher implementations.

1 Introduction

Efficiently sampling from complex high-dimensional continuous distributions remains one of the central algorithmic challenges for modern Markov chain Monte Carlo (MCMC) in Bayesian inference and statistical machine learning. The challenge goes beyond designing an unbiased Markov chain with the correct stationary distribution π: the chain must also explore efficiently under anisotropy, curved geometry, and metastable basins within a practical computational budget.

Methods such as the Metropolis-Adjusted Langevin Algorithm (MALA) [51] and Hamiltonian Monte Carlo (HMC) [21, 44] marked a major advance by using first-order information to steer proposals toward informative regions of the target. Subsequent developments, including Riemannian-manifold variants [26], adaptive covariance schemes [28, 4], and Fisher-based proposal metrics [54, 30, 53], further sharpened the role of geometry in accelerating exploration. These methods act on the proposal mechanism of a sampler whose target distribution remains fixed.

A complementary philosophy is to use trajectory history to reshape the effective landscape itself. Metadynamics [38, 6, 29], adaptive biasing potential/force methods [18, 7–9], the Wang-Landau algorithm [57], contour SGLD [19], and Stein self-repulsive dynamics [59] all accumulate 'summaries' of past exploration, e.g., bias histograms, collective-variable grids, or past samples used to evaluate repulsive corrections. While these summaries are effective, they either lead to increasing

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

memory usage over time or demand explicit reweighting to recover unbiased Monte Carlo estimates. Combining self-repellence, asymptotic unbiasedness, and fixed memory usage remains difficult.

On finite state spaces, Self-Repellent Random Walk (SRRW) [20] and History-Driven Target (HDT)-MCMC [33] employ the empirical occupation measure to build a history-driven approach that still converges to π while penalizing states that have been visited too frequently. This mechanism can drive the asymptotic variance close to zero, while the full occupation measure does not scale to continuous state spaces $\mathbb{R}^d$. Score-Repellent Monte Carlo (SRMC) [34] resolves this scalability bottleneck by compressing history into a fixed-dimensional running average of the score function $s(x) \triangleq \nabla_x \log \pi(x)$ in $\mathbb{R}^d$, yielding an auxiliary iterate $\theta_n \in \mathbb{R}^d$ such that

$$X_{n+1} \sim K_{\theta_n}(X_n, \cdot), \qquad \theta_{n+1} = \theta_n + \gamma_{n+1}\big(s(X_{n+1}) - \theta_n\big), \tag{1}$$

where γ_n is the step size and K_θ is any MCMC kernel with its invariant distribution

$$\pi_\theta(x) \propto \pi(x) \exp\big(-\alpha\, \theta^\top s(x)\big). \tag{2}$$

Here, $\alpha \geq 0$ quantifies the strength of self-repellency, and π_θ is a history-dependent surrogate target, parameterized by the score average θ. Choosing either $\alpha = 0$ or $\theta = 0$ in (2) recovers the original target π. The SRMC framework can be wrapped around essentially any base MCMC scheme. Hu et al. [34] proved that the stochastic approximation (SA) recursion (1) converges almost surely to $\theta^\star = 0$, where the surrogate target $\pi_{\theta^\star} = \pi$ makes SRMC asymptotically unbiased and improves the sampling efficiency over the base MCMC algorithm. SRMC therefore occupies a distinctive position in this landscape: it is neither a purely kinetic adaptive sampler nor a biased history-flattening method; instead, it is a target-adaptive framework that discourages revisitation while converging back to π, with $O(d)$ memory and normalization-free implementation.

Despite its attractive properties, however, SRMC remains *geometry-blind* in a way that complicates tuning. A practitioner must choose the repellence strength α, but in (2), it multiplies the raw Euclidean alignment $\theta^\top s(x)$, whose magnitude depends on the units, curvature, and anisotropy of the score field. The original SRMC work treats α as a target-dependent tuning parameter, using heuristic diagnostics such as keeping the exponent scale $\alpha|\theta_n^\top s(X_n)|$ moderate or increasing α adaptively, rather than providing a systematic treatment [34]. For an anisotropic Gaussian target $\pi = \mathcal{N}(0, \Sigma)$ with $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, the coordinate scores satisfy $s_i(x) = -x_i/\sigma_i^2$, so $\theta^\top s(x) = -\sum_{i=1}^d \theta_i x_i/\sigma_i^2$. Because θ is a running average of past scores, the Euclidean inner product $\theta^\top s(x)$ amplifies curvature scale in the surrogate target: low-variance/high-curvature directions dominate the repulsion exponent simply because their scores are numerically large. A single scalar α must therefore handle directions with different variances: if it is large enough to move the chain along broad, flat directions, it overshoots in steep directions where scores are already large; if it is made small enough to avoid this overreaction, the repulsion is too weak to help in broad directions where mixing is slow. As a result, geometry-blind SRMC applies repulsion where the score is numerically large, rather than where the chain actually has difficulty mixing.

This mismatch suggests a more refined question: *instead of tuning α around raw score scales $\theta^\top s(x)$, should the repulsion term compare the history so far with the current score using a different metric?* Common scale-free normalizations such as cosine similarity and $\theta^\top s(x)/\|s(x)\|$, while seemingly effective at first sight, can in fact prevent the surrogate target from recovering π at equilibrium, thereby inducing bias; see Appendix D.1 for a detailed account. On the other hand, in the aforementioned Gaussian example, taking the weighted inner product with the weight matrix $M = \Sigma$ gives $\theta^\top M s(x) = \theta^\top \Sigma s(x) = -\sum_{i=1}^d \theta_i x_i$, which removes the variance-dependent denominators from the repulsion term. The issue is therefore not the score-based history itself, but the metric used to measure the alignment between the history variable θ and the current score $s(x)$. This naturally leads to the weighted inner product $\theta^\top M s(x)$, where M should encode the geometry of the target.

Thus motivated, in this paper, we introduce **Geometry-Aware** (GA)-SRMC, defined by the surrogate target family with a weight matrix $M \in \mathbb{R}^{d\times d}$:

$$\pi_\theta^{(M)}(x) \propto \pi(x) \exp\big(-\alpha\, \theta^\top M s(x)\big), \qquad M = M^\top \succ 0, \tag{3}$$

where $M = I_d$ recovers the original SRMC surrogate target. The SA update for θ_n is unchanged; only the metric through which the score average acts back on the surrogate target is modified. This identifies a new degree of freedom for the *repulsion-side* geometry within the surrogate target $\pi_\theta^{(M)}$, which is distinct from proposal-side preconditioning. In preconditioned MALA [54, 30], Riemannian or preconditioned HMC [26, 53], and adaptive Metropolis [28], a symmetric matrix A

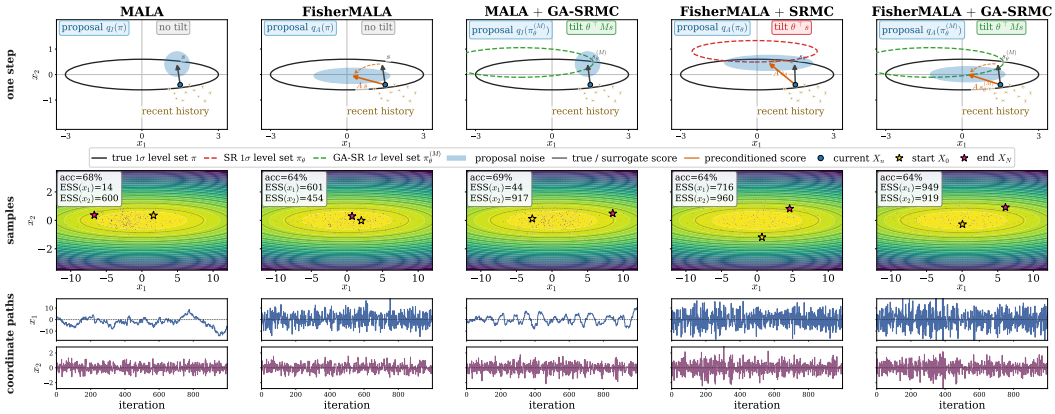

Figure 1: **Proposal-side and repulsion-side geometry are complementary.** Top: one-step mechanism. The preconditioner A shapes proposal movements, while the GA-SRMC matrix M shapes the history-dependent surrogate target. Middle/bottom: samples and coordinate paths on an anisotropic Gaussian target. Combining FisherMALA with GA-SRMC yields the most balanced effective sample sizes (ESS) across coordinates, computed from $N = 1000$ post-burn-in samples with ESS estimation in Geyer [25]. Full setup and details are deferred to Appendix A.1.

determines the proposal distribution for a chosen target. For example, if a preconditioned Langevin proposal is applied to a generic target ν with score $s_\nu(x) = \nabla_x \log \nu(x)$, then $q_A(y \mid x) = \mathcal{N}\big(y \mid x + \frac{1}{2} A s_\nu(x), A\big)$. Classically, one takes $\nu = \pi$. In GA-SRMC, by contrast, the matrix M specifies the history-dependent surrogate target $\pi_\theta^{(M)}$ in (3). Consequently, M deforms the energy landscape passed to the base kernel, whereas A shapes how the base kernel proposes moves on that landscape by enforcing this surrogate target. For a Metropolis–Hastings kernel with target ν and proposal q_A, imposing $\nu = \pi_\theta^{(M)}$ in (3) yields the following normalization-free acceptance ratio:

$$\frac{\pi_\theta^{(M)}(y)\, q_A(x \mid y)}{\pi_\theta^{(M)}(x)\, q_A(y \mid x)} = \underbrace{\frac{\pi(y)\, q_A(x \mid y)}{\pi(x)\, q_A(y \mid x)}}_{\text{proposal-side correction}} \underbrace{\exp\{-\alpha \theta^\top M(s(y) - s(x))\}}_{\text{Ours: geometry-aware score repellence}}, \tag{4}$$

which highlights that the two mechanisms are complementary, with M acting as a repulsion-side matrix and A as a proposal-side preconditioner. In the smooth continuous setting, the matrix $S \triangleq \mathrm{Cov}_\pi(s, s) = \mathbb{E}_\pi[s(X)s(X)^\top]$ is the Fisher information matrix by Stein's identity $\mathbb{E}_\pi[s(X)] = 0$.[1] Our Theorem 3.1 identifies the inverse-Fisher choice $M_{\text{opt}} \propto S^{-1}$ as the minimizer of a worst-direction bound derived from the central limit theorem (CLT). This finding strengthens the role of S^{-1} as a preconditioner for gradient-based proposals [54, 30, 53] and natural-gradient optimization [2, 42, 58]. We show that the same Fisher geometry governs not only proposal-side preconditioning but also the repulsion term of a score-repellent surrogate target.

Figure 1 illustrates the geometric issue motivating GA-SRMC. Standard MALA [51] uses isotropic proposal noise and mixes poorly along the elongated coordinate x_1, whose trace is highly autocorrelated, resulting in the imbalanced $(\mathrm{ESS}(x_1), \mathrm{ESS}(x_2)) = (14, 600)$. FisherMALA [54] preconditions the local move on the proposal side, so its x_1 trace is less 'sticky' than under MALA, and $\mathrm{ESS}(x_1)$ rises from 14 to 601. MALA with GA-SRMC adds history-dependent repulsion and improves over MALA along the slow direction, but without preconditioning its x_1 trajectory remains clearly correlated and still underperforms FisherMALA, showing that the base sampler's quality is crucial. Conversely, FisherMALA with SRMC [34] shows that score repellence can further improve a preconditioned sampler, $(\mathrm{ESS}(x_1), \mathrm{ESS}(x_2)) = (716, 960)$, yet its Euclidean tilt is not aligned with the target geometry. FisherMALA with GA-SRMC aligns both the preconditioner A and the Fisher-weighted matrix M with this geometry and attains the most balanced pair, $(\mathrm{ESS}(x_1), \mathrm{ESS}(x_2)) = (949, 919)$, at essentially the same acceptance rate ('acc' in the legend).

[1] As in Titsias [54], Fisher information matrix S is the score covariance under π, not to be confused with the classical position-dependent Fisher information $\mathbb{E}_{Y \sim p(\cdot|x)}[\nabla_x \log p(Y|x) \nabla_x \log p(Y|x)^\top]$ of a model $p(y|x)$ with parameter x.

Our contributions are as follows:

1. We generalize SRMC by replacing the Euclidean alignment $\theta^\top s(x)$ in the surrogate target π_θ in (2) with a geometry-aware alignment $\theta^\top M s(x)$. For any fixed symmetric positive-definite M, we show that the surrogate family is well-defined, GA-SRMC remains asymptotically unbiased, and its limiting covariance is determined by SM, with the choice of M playing a role.

2. We show that the inverse-Fisher matrix $M_{\mathrm{opt}} \propto S^{-1}$ equalizes the spectrum of SM under the trace constraint $\mathrm{tr}(SM) = \tau$, uniquely minimizing the universal worst-direction bound on the score-history covariance. We then estimate S online from the score that GA-SRMC already computes, with diagonal and full-Fisher versions that trade computational costs against geometric fidelity. Under additional stability assumption, the learned full-Fisher metric M_n converges almost surely to M_{opt}, and the score history and the estimator obey the same CLT as under the fixed metric M_{opt} in hindsight.

3. We evaluate GA-SRMC on 8 targets with two sampler families (MALA and FisherMALA) and three variants {base, + SRMC, + GA-SRMC}. In the worst-coordinate ESS under MALA, GA-SRMC dominates SRMC by $\mathbf{4.3}\times$ on anisotropic Gaussian ($\kappa = 100$) and $\mathbf{6.8}\times$ on MNIST at each method's best stability-capped α; under FisherMALA, it adds $+48\%$ on the strongly anisotropic Gaussian and ties with SRMC on near-isotropic targets. Its posterior-mean MSE is lower than SRMC's in all 16 full-Fisher comparisons.

The message is straightforward: *historical trajectory should be used in combination with the appropriate geometry.* SRMC introduces a constant-memory score-repellent mechanism for target adaptation [34]; GA-SRMC identifies the metric in this mechanism as an object with its own asymptotic geometry. For smooth continuous targets, this geometry is governed by the Fisher information matrix $S = \mathrm{Cov}_\pi(s, s)$, which shapes not only proposal-side preconditioning in gradient-based MCMC, but also the repellent surrogate target itself.

2 GA-SRMC with a Fixed Weight Matrix M

Notation. All vectors are column vectors. For symmetric A, B, Loewner ordering $A \succeq B$ (*resp.* $A \succ B$) means $A - B$ is positive semidefinite (*resp.* definite). We write I_d for the $d \times d$ identity matrix, $\mathrm{tr}(A)$ for the trace, and $A^{1/2}$ for the principal square root when $A \succeq 0$. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and largest eigenvalues of a square A. The operator $\mathrm{diag}(A)$ extracts the diagonal of A, and $\mathrm{diag}(a)$ is the diagonal matrix with entries from the vector a. For vector-valued functions g, h and a probability distribution π, denote their marginal covariance by $\mathrm{Cov}_\pi(g, h) \triangleq \mathbb{E}_\pi[(g(X) - \mathbb{E}_\pi[g(X)])(h(X) - \mathbb{E}_\pi[h(X)])^\top]$. We use $\xrightarrow{a.s.}$ and $\xrightarrow{dist.}$ for almost sure and weak convergence, respectively. Let $\mathcal{N}(\mu, V)$ be a multivariate Gaussian with mean μ and covariance V.

Basic structure of GA-SRMC. We analyze the GA-SRMC surrogate target $\pi_\theta^{(M)}$ in (3) with a fixed symmetric matrix $M \succ 0$. For each $\theta \in \mathbb{R}^d$, define

$$Z_\theta^{(M)} \triangleq \int_{\mathbb{R}^d} \pi(x) \exp\{-\alpha\, \theta^\top M s(x)\}\, dx, \;\; S_M(\theta) \triangleq \mathbb{E}_{\pi_\theta^{(M)}}[s(X)], \;\; F_M(\theta) \triangleq \mathbb{E}_{\pi_\theta^{(M)}}[f(X)], \quad (5)$$

where the score function $s = \nabla_x \log \pi$, and $f : \mathbb{R}^d \to \mathbb{R}^m$ is a test function of interest, with $\mu \triangleq \mathbb{E}_\pi[f(X)]$ for the quantity to be estimated. We write the joint iterate and its update function as:

$$\vartheta_n \triangleq (\theta_n, \mu_n) \in \mathbb{R}^{d+m}, \quad H(\vartheta_n, x) \triangleq (s(x) - \theta_n,\; f(x) - \mu_n). \quad (6)$$

Let K_ν denote an arbitrary base MCMC kernel with invariant distribution ν, e.g., random walk Metropolis, MALA, or HMC. At each iteration, GA-SRMC runs a base kernel $K_{\theta,M}$ targeting the current surrogate $\pi_\theta^{(M)}$ in (3) and updates the joint iterate online via the SA recursion

$$X_{n+1} \sim K_{\theta_n, M}(X_n, \cdot),\; \vartheta_{n+1} = \vartheta_n + \gamma_{n+1} H(\vartheta_n, X_{n+1}),\; \text{where } \gamma_n = (n+1)^{-\rho},\; \rho \in (0.5, 1]. \quad (7)$$

Note that the weight matrix M enters (7) only through the surrogate target used by the transition kernel; the SA recursion itself is unchanged from SRMC.

Assumptions. We adopt the regularity assumptions of the SA framework for SRMC [34], specialized to the surrogate family $\{\pi_\theta^{(M)}\}_{\theta \in \mathbb{R}^d}$ with fixed M: Lipschitz continuity and appropriate tail behavior of the score s [22, 16] (see also Hu et al. [34, Appendix C.1]), along with a drift condition and Lipschitz continuity of the Markov kernel [3, 50, 24]. They make the surrogates $\pi_\theta^{(M)}$ well-defined, and the base MCMC algorithm uniformly geometrically ergodic while θ stays bounded. Formal statements and proofs are in Appendix B.

Proposition 2.1 (Well-posedness & equilibrium)**.** *Under Assumption 1, for every symmetric $M \succ 0$:*

1. *(**Well-posedness**) For every $\theta \in \mathbb{R}^d$ and every $\alpha \geq 0$, the normalizing constant $Z_\theta^{(M)} < \infty$, so the surrogate $\pi_\theta^{(M)}$ in (3) is a well-defined probability distribution on $\mathbb{R}^d$.*
2. *(**Equilibrium**) The point $\vartheta^\star \triangleq (0, \mu) \in \mathbb{R}^{d+m}$ is the unique globally asymptotically stable equilibrium of the mean-field ODE $\dot{\vartheta}_t = h_M(\vartheta_t)$, where $h_M(\vartheta_t) \triangleq (S_M(\theta_t) - \theta_t,\ F_M(\theta_t) - \mu_t)$.*

Proposition 2.1 shows that the geometry-aware tilt is a valid target-adaptation mechanism. While M changes how the accumulated score history contributes to the surrogate exponent, it does not change the desired equilibrium: for every θ and M, the surrogate remains a proper probability distribution and coincides with $\pi_0^{(M)} = \pi$ at the unique equilibrium $\theta^\star = 0$. Moreover, the joint SA recursion (7) retains the correct equilibrium with the mean field $h_M(\vartheta^\star) = 0$ and is globally stable. The proof verifies that the weighted score tilt does not introduce any alternative stable equilibria, so M can reshape the transient dynamics and asymptotic covariance without altering the limiting target π.

The SA viewpoint shows how this 'reshaping' occurs. The recursion (7) is a noisy, decreasing-step Euler discretization of the mean-field ODE $\dot{\vartheta}_t = h_M(\vartheta_t)$ with controlled Markovian noise from the current surrogate kernel (see also [37, 13, 12] for ODE analysis of general SA). Around the stable equilibrium $\vartheta^\star = (0, \mu)$, the deterministic part is governed by the Jacobian

$$J_\star^{(M)} \triangleq \nabla_\vartheta h_M(\vartheta^\star) = \begin{bmatrix} -I_d - \alpha SM & 0 \\ -\alpha CM & -I_m \end{bmatrix}, \quad S \triangleq \mathbb{E}_\pi[s(X)s(X)^\top], \quad C \triangleq \mathrm{Cov}_\pi(f, s), \tag{8}$$

which captures the local deterministic damping of the mean-field ODE: the CLT covariance below balances this damping against the long-run covariance of the Markovian noise.

Asymptotic analysis for GA-SRMC with fixed M. The following proposition extends the SA theory of SRMC [34]: for fixed M, the surrogate $\pi_\theta^{(M)}$ still converges to π, and a CLT holds that Section 3 uses to choose optimal M_{opt}. Let $\{Z_k : k \in \mathbb{Z}\}$ be a trajectory of the base MCMC kernel K_π targeting π. For vector-valued functions a, b on $\mathbb{R}^d$, define the long-run covariance

$$U_{ab} \triangleq \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\left(\sum_{k=0}^{T-1} (a(X_k) - \mathbb{E}_\pi[a(X)]) \right) \left(\sum_{k=0}^{T-1} (b(X_k) - \mathbb{E}_\pi[b(X)]) \right)^\top \right], \tag{9}$$

and we write the long-run covariance associated with the function $H(\vartheta^\star, \cdot)$, as introduced in (6) and evaluated at the equilibrium point $\vartheta^\star$, as $U_{HH} = \left[\begin{smallmatrix} U_{ss} & U_{sf} \\ U_{fs} & U_{ff} \end{smallmatrix}\right]$. Let $\mathbf{1}_{(\cdot)}$ be the indicator function.

Proposition 2.2 (Almost sure convergence and CLT for fixed M)**.** *The iterate in (7) satisfies $\vartheta_n \xrightarrow[n\to\infty]{a.s.} \vartheta^\star$ and $\gamma_n^{-1/2}(\vartheta_n - \vartheta^\star) \xrightarrow[n\to\infty]{dist.} \mathcal{N}(0, \Sigma_\vartheta^{(M)})$, where $\Sigma_\vartheta^{(M)}$ is the unique solution of*

$$\left(\tfrac{\mathbf{1}_{\{\rho=1\}}}{2} I + J_\star^{(M)}\right)\Sigma_\vartheta^{(M)} + \Sigma_\vartheta^{(M)}\left(\tfrac{\mathbf{1}_{\{\rho=1\}}}{2} I + J_\star^{(M)}\right)^\top + U_{HH} = 0. \tag{10}$$

Setting $M = I_d$ in Proposition 2.2 recovers Hu et al. [34, Theorem 3.3]. Proposition 2.2 has two consequences that will be used in Section 3. First, GA-SRMC remains asymptotically unbiased for every fixed symmetric $M \succ 0$: the score-history variable satisfies $\theta_n \to 0$ almost surely, the surrogate target returns to $\pi_0^{(M)} = \pi$, and the Monte Carlo estimator converges to $\mu = \mathbb{E}_\pi[f(X)]$. Second, the proposition turns the mean-field picture into a covariance equation. The matrix $J_\star^{(M)}$ in (8) represents the deterministic local linearization of the ODE that accounts for the influence of M, whereas U_{HH} denotes the long-run covariance associated with the baseline MCMC sampler, in which the preconditioning matrix A (e.g., as in preconditioning techniques) plays a role. The Lyapunov equation (10) balances these two effects. Let $\Sigma_{\theta\theta}^{(M)}$ and $\Sigma_{\mu\mu}^{(M)}$ denote the limiting covariance of θ- and μ-iterates, respectively, within $\Sigma_\vartheta^{(M)}$. Focusing on the sequence $\{\theta_n\}$ alone leads to the Lyapunov equation $D_M \Sigma_{\theta\theta}^{(M)} + \Sigma_{\theta\theta}^{(M)} D_M^\top = U_{ss}$, whose solution $\Sigma_{\theta\theta}^{(M)}$ admits the closed form:

$$\Sigma_{\theta\theta}^{(M)} = \int_0^\infty e^{-tD_M} U_{ss} e^{-tD_M^\top} dt, \quad \text{where} \quad D_M \triangleq \beta I_d + \alpha SM, \quad \beta \triangleq 1 - \tfrac{\mathbf{1}_{\{\rho=1\}}}{2}. \tag{11}$$

We call $D_M \succ 0$ the *CLT damping operator*. The Jacobian has the negative drift block $-I_d - \alpha SM$, while the Lyapunov equation uses its positive damping counterpart $D_M = \beta I_d + \alpha SM$. Directions

associated with small eigenvalues of D_M are weakly damped and therefore dominate the worst-direction covariance bounds. This is the analytical counterpart of the top row for Figure 1: proposal-side preconditioning changes how the chain moves via U_{ss}, whereas repulsion-side preconditioning changes the geometry of the score-history damping through SM and in turn D_M. Under the SRMC ($M = I_d$), anisotropy of S produces uneven damping across score directions; under the inverse-Fisher choice $M \propto S^{-1}$, the damping becomes balanced across directions. This is the bridge to Section 3.

3 Choosing and Learning the Optimal Weight Matrix M_{opt}

Optimal choice of M_{opt}. By Proposition 2.2, any fixed symmetric $M \succ 0$ retains the limiting target π and the equilibrium $\theta^\star = 0$ unchanged, while only redistributing the damping in D_M across score directions. The worst-direction bound on $\Sigma_{\theta\theta}^{(M)}$ is governed by the smallest eigenvalue of D_M, i.e., $\lambda_{\min}(D_M) = \beta + \alpha\lambda_{\min}(SM)$, so a poor choice of M produces a slow eigendirection of SM that no single scalar α can rescue uniformly. Multiplying M by scalar c has the same effect on the surrogate exponent (3) as multiplying α by c, so an unconstrained optimization over M would only re-parametrize the repellence strength. We therefore impose the trace constraint

$$\mathrm{tr}(SM) \;=\; \mathbb{E}_\pi\big[s(X)^\top M s(X)\big] \;=\; \tau \;>\; 0, \tag{12}$$

which fixes the mean squared score magnitude under π and the trace of αSM in D_M. Our GA-SRMC framework embeds an optimization question: *which weight matrix M should be selected to best enhance the worst-case performance?* The relationship between $\Sigma_{\theta\theta}^{(M)}$ and $\lambda_{\min}(SM)$, along with the resulting optimal choice M_{opt}, is given below.

Theorem 3.1 (Trace-normalized inverse-Fisher matrix)**.** *Assume the fixed-M CLT in Proposition 2.2 and suppose $S \succ 0$. Then, the limiting covariance $\Sigma_{\theta\theta}^{(M)}$ of the θ_n-iterate satisfies*

$$\Sigma_{\theta\theta}^{(M)} \preceq \frac{\lambda_{\max}(U_{ss}S^{-1})}{2\{\beta + \alpha\lambda_{\min}(SM)\}}\, S. \tag{13}$$

Among all $M{=}M^\top{\succ}0$ with (12)*, the upper bound in* (13) *is uniquely minimized in Loewner order by*

$$M_{\mathrm{opt}} = \frac{\tau}{d} S^{-1}. \tag{14}$$

At this choice, limiting covariances associated with score history and Monte Carlo estimator become

$$\Sigma_{\theta\theta}^{(M_{\mathrm{opt}})} = \frac{U_{ss}}{2(\beta + \alpha\tau/d)}, \quad \Sigma_{\mu\mu}^{(M_{\mathrm{opt}})} = \frac{U_{gg}}{2\beta} + O(\alpha^{-1}), \tag{15}$$

where U_{gg} is defined by (9) *with the function $g(x) \triangleq f(x) - CS^{-1}s(x)$.*

Proofs are in Appendices C.1 and C.2. Intuitively, SM is similar to $S^{1/2}MS^{1/2} \succ 0$, so its eigenvalues are positive, and the trace constraint fixes their average at τ/d. A larger $\lambda_{\min}(SM)$ leads to a tighter bound in (13). Because $\lambda_{\min}(SM) \le \tau/d$, and equality occurs only when $S^{1/2}MS^{1/2} = (\tau/d)I_d$ (equivalently, $M = (\tau/d)I_d$), the inverse Fisher choice minimizes the bound: it maximizes the weakest guaranteed score-history contraction under the trace constraint. Moreover, $D_{M_{\mathrm{opt}}} = (\beta + \alpha\tau/d)I_d$, which turns (11) into the closed form (15). This limiting covariance is strictly Loewner-monotone in α, recovers the base sampler's long-run covariance $U_{ss}/(2\beta)$ at $\alpha = 0$, and vanishes as $\alpha \to \infty$. Hence, it gives a strictly stronger result than the Frobenius-norm ordering obtained for the suboptimal $M = I_d$ in [34, Proposition 3.4].

The limiting covariance $\Sigma_{\mu\mu}^{(M_{\mathrm{opt}})}$ in (15) reflects a variance-reduction principle: the score history damps only the score-aligned component $CS^{-1}s$ of the test function f, while the score-orthogonal residual $g \triangleq f - CS^{-1}s$, with $\mathrm{Cov}_\pi(g, s) = 0$, sets the floor $U_{gg}/(2\beta)$, which is the long-run covariance of the residual noise $g(X) - \mu$ under the base MCMC sampler and is independent of M for a fixed kernel. As $\alpha \to \infty$, GA-SRMC reduces the score-aligned covariance at rate $O(\alpha^{-1})$, but leaves a residual $U_{gg}/2\beta$ that score-based repulsion alone cannot remove. This decomposition aligns with established Monte Carlo variance-reduction methods, namely zero-variance MCMC [5, 43] and the control-functional approach [45], in which the score-aligned term $CS^{-1}s$ corresponds exactly to the portion of f that is eliminated by score-based control variates, with a detailed discussion in Appendix C.4. In Appendix C.3, we discuss two special cases of the independent surrogate sampler and Gaussian target to specify the role of M_{opt}.

Algorithm 1 GA-SRMC with online trace-normalized Fisher estimation

Require: score s; kernel family $K_{\theta,M}$; repellence α; small δ; stepsizes γ_n, ω_n; type $\in \{\text{diag}, \text{full}\}$.

1: Initialize X_0, $\theta_0 = 0$, μ_0, and $\hat{S}_n \succ 0$.
2: **for** $n = 0, \ldots, N-1$ **do**
3: Compute M_n via (16); draw $X_{n+1} \sim K_{\theta_n, M_n}(X_n, \cdot)$ targeting (3).
4: $s_{n+1} \leftarrow s(X_{n+1})$; update $\vartheta_{n+1} \leftarrow \vartheta_n + \gamma_{n+1} H(\vartheta_n, X_{n+1})$ as in (7).
5: FISHERUPDATE: $(\hat{S}_{n+1} \leftarrow \hat{S}_n + \gamma_{n+1}(s_{n+1}s_{n+1}^\top - \hat{S}_n)$ (diag: diagonal only).
6: **return** samples $\{X_n\}$, Monte Carlo estimator μ_N.

Remark 1 (why optimize the score-history covariance). The score history θ_n is the memory of GA-SRMC and the control signal that reshapes the next surrogate target. Proposition 2.2 implies that, at large time, $\mathbb{E}[\theta_n \theta_n^\top] = \gamma_n \Sigma_{\theta\theta}^{(M)} + o(\gamma_n)$, so $\Sigma_{\theta\theta}^{(M)}$ is the leading covariance constant of the running score imbalance. Reducing this block means that the trajectory cancels score directions more efficiently and that the surrogate feedback learns the local target geometry with less residual fluctuation. The inverse-Fisher choice M_{opt} therefore optimizes the asymptotic covariance of the very statistic that drives score repellence. The residual decomposition above explains when this improvement transfers directly to Monte Carlo efficiency.

The inverse Fisher S^{-1} has an established *proposal-side* role in FisherMALA [54, 30], FisherHMC [53], and in natural-gradient methods [2, 26, 42]; Theorem 3.1 identifies it as the bound-optimal *repulsion-side* choice. Although these two criteria act on different objects, i.e., expected squared jumped distance or the condition number of the target distribution versus CLT damping operator D_M in the score-history feedback, they both select S^{-1} because both whiten a score-driven object against the same anisotropy of π. Practically, an online estimate of S^{-1} can simultaneously serve both purposes without any additional overhead, placing GA-SRMC at the unique intersection of two closely related design choices: preconditioned-only samplers leverage Fisher geometry on the proposal side but lack score repellence, whereas SRMC modifies the target on the repulsion side while using the geometry-blind metric $M = I_d$. GA-SRMC uniquely employs Fisher geometry on *both* sides whenever a preconditioning sampler is used, which can further improve sampling efficiency.

Online Fisher implementation. We replace the unknown Fisher S by an online estimate $\widehat{S}_n$ from the score evaluations $\{s(X_k)\}_{k \le n}$ already computed along the chain and then normalize the trace via

$$M_n \;=\; \frac{\operatorname{tr}(\widehat{S}_n)}{d}\, \widehat{S}_n^{-1}, \tag{16}$$

which decouples α from the geometry: $\operatorname{tr}(\widehat{S}_n M_n) = \operatorname{tr}(\widehat{S}_n)$ matches the Fisher-weighted trace of the SRMC choice $M = I_d$, so α solely sets the repellence strength and M_n only its geometry. The *diagonal* version stores coordinate-wise score second moments and inverts a vector at $O(d)$ cost per iteration; it captures axis-aligned anisotropy but ignores cross-coordinate score correlations, as in TensorFlow/Stan [1, 15], and can be counterproductive on highly correlated targets [30, Section 3.5]. The *full* version, in the spirit of FisherMALA [54], maintains $\widehat{S}_{n+1} = (1-\gamma_{n+1})\widehat{S}_n + \gamma_{n+1} s_n s_n^\top$, making it suitable when score correlations are dense and the dimension is moderate. This recursion costs $O(d^2)$ from the Sherman-Morrison update (Appendix D.2.2). In Algorithm 1, FISHERUPDATE updates the inverse Fisher estimate on the fly. Complete implementations of the two Fisher estimates and Algorithm 1 are provided in Appendix D.2.

For Langevin-type base kernels, GA-SRMC adds *one Hessian-vector product (HVP) per surrogate-score evaluation*, two per MALA step, over its base sampler. By (3), the surrogate score is

$$\nabla_x \log \pi_\theta^{(M)}(x) = s(x) - \alpha \nabla_x^2 \log \pi(x) M\theta, \tag{17}$$

so that a preconditioned Langevin proposal with proposal matrix A_n and step size ϵ_n takes the form $Y = X_n + \epsilon_n A_n \nabla_x \log \pi_\theta^{(M_n)}(x) + \sqrt{2\epsilon_n} A_n^{1/2} \xi_n$, followed by the Metropolis accept–reject step in (4). The HVP $\nabla_x^2 \log \pi(X_n)\, M_n \theta_n$ can be computed via automatic differentiation [46, 27] or finite differences of the gradient [17], without forming the Hessian; for samplers that already maintain an online inverse-Fisher preconditioner [54, 53], one trace-normalized inverse-Fisher estimate serves both purposes. As in [54], the step size ϵ_n is adapted on the log scale, i.e., $\log \epsilon_{n+1} = \log \epsilon_n + c_\epsilon(a_n - 0.574)$, where $c_\epsilon \in (0, 1)$ is a positive adaptation gain, a_n is the realized Metropolis acceptance probability, and 0.574 is the standard MALA optimal-acceptance target [48, 49]. This is the adaptive scheme used in Figure 1.

Propositions 2.1–2.2 and Theorem 3.1 hold M fixed. In Algorithm 1, M_n is updated online at every step, so the augmented iterate $z_n \triangleq (\theta_n, \mu_n, \widehat{S}_n)$ follows one SA recursion of the form (7), with the extra coordinate $\widehat{S}_n$ and kernel K_{θ_n, M_n}.

Theorem 3.2 (Online inverse-Fisher metric)**.** *Suppose additionally that $\widehat{S}_n$ remains bounded in the positive definite cone (Assumption 3). Then, with $\tau = \operatorname{tr} S$ in M_{opt}, (i) $\theta_n \to 0$, $\mu_n \to \mu$, $\widehat{S}_n \to S$, and $M_n \to M_{\mathrm{opt}}$ almost surely; and (ii) z_n satisfies a joint CLT whose (θ, μ) marginal is the CLT of Proposition 2.2 at $M = M_{\mathrm{opt}}$: $\gamma_n^{-1/2}(\vartheta_n - \vartheta^\star) \xrightarrow{dist.} \mathcal{N}(0, \Sigma_\vartheta^{(M_{\mathrm{opt}})})$.*

The proof in Appendix C.5 follows the steps of Proposition 2.2 for the augmented iterate z_n with two modifications: the moving-metric Lyapunov function $\theta^\top \widehat{S}^{-1}\theta$ and the identity $\pi_0^{(M)} = \pi$ for every $M \succ 0$. Because of the latter, the Fisher state does not contribute to the linearization of the (θ, μ) mean field at equilibrium, which implies that, in the asymptotic regime, the online estimate M_n does not compromise the optimality of the hindsight selection M_{opt} in Theorem 3.1.

Remark 2 (Scope of the fixed-M and online results). Proposition 2.2 and Theorem 3.1 focus on a fixed weight matrix M, whereas Theorem 3.2 analyzes Algorithm 1 in a fully online regime using the full Fisher information estimator. In this setting, an inaccurate Fisher estimate during the early iterations may induce a mis-scaling of the proposal through A_n, and through M_n, it can reduce $\lambda_{\min}(SM_n)$ and thereby weaken the bound in (13). Theorem 3.2 establishes only that this degradation is asymptotically negligible. Practical stabilization is provided by a regularization parameter δ, trace normalization, and initialization based on the MAP Hessian, as stated in Section 4.

4 Experiments

We evaluate GA-SRMC on an anisotropic Gaussian target with condition number $\kappa = 100$ (GAUSS-2D), chosen to isolate geometric effects, and on seven Bayesian logistic-regression (BLR) posteriors (18) with a Gaussian prior of precision $\tau_0 = 0.1$ on standard real-world benchmarks: RIPLEY [47], the UCI [41] datasets PIMA, HEART, AUSTRALIAN, and GERMAN, the CoIL Challenge CARAVAN dataset [55], and MNIST [39] (full target specification in Appendix A.2). Collectively, these problems span dimensionalities $d \in [2, 785]$ and posterior condition numbers $\kappa \in [1.3, 100]$. Following the implementation similar to Titsias [54], we run 40 independent Markov chains for each algorithm, collecting $20{,}000$ post-burn-in samples after a burn-in of $2{,}000$ iterations, and evaluate each chain on two complementary metrics (mean $\pm$ standard deviation across the 40 chains): (i) the *worst-coordinate ESS* (Table 1), defined as the minimum across coordinates of the single-chain bulk effective sample size computed via ArviZ [36], which quantifies mixing along the slowest coordinate; and (ii) the *per-dimension posterior-mean MSE* $\frac{1}{d}\|\bar{X} - \mu_{\mathrm{ref}}\|_2^2$ (Table 2 with details in Appendix A.5), which quantifies the estimation error of the posterior mean across all d coordinates. All preconditioned variants in Tables 1 and 2 use the full online Fisher estimator from Algorithm 1 with $\widehat{S}_{n,\delta} \triangleq \widehat{S}_n + \delta I_d$ and $\delta = 0.01$ to ensure invertibility; the sensitivity to δ is analyzed in Appendix A.3.

Stability cap on α. For a fixed value of θ, the GA-SRMC chain samples the surrogate $\pi_\theta^{(M)}$, which coincides with the true target π only in the asymptotic regime $\theta_n \to \theta^\star = 0$ established in Proposition 2.2 and Theorem 3.2. For finite n and sufficiently large α, the chain can temporarily explore a distorted distribution, so relying on ESS comparisons alone can be misleading. To guard against this, we impose a *stability cap* defined separately for each target and each method: it is the largest α in our search grid for which the post-burn-in marginal standard deviations remain within 30% of those under a Gaussian Laplace approximation at the maximum a posteriori (MAP), following the validation diagnostic of Titsias [54] (complete details in Appendix A.4). Because the unscaled tilt in SRMC destabilizes the chain at smaller α than the geometry-aware tilt in GA-SRMC, the two methods typically admit different stability caps. Accordingly, all results in Tables 1 and 2 are reported at each method's best-ESS α under its stability cap.

Two sampler families. Table 1 reports the worst-coordinate ESS for the MALA family ($A = I_d$) and the FISHERMALA family ($A \propto S^{-1}$), each evaluated in three versions: the base sampler, + SRMC, and + GA-SRMC. The MALA family isolates the repulsion-side geometry most clearly, since $M\theta$ is the only channel through which Fisher geometry enters the sampler. On GAUSS-2D, MALA+SRMC reaches 104 at $\alpha = 1$ compared with 452 for MALA+GA-SRMC at $\alpha = 3$, a $4.3\times$ gain; on MNIST, 74 at $\alpha = 1$ versus 501 at $\alpha = 20$, a $6.8\times$ gain; on CARAVAN, 922 at $\alpha = 5$ versus 1353 at $\alpha = 10$, a 47% increase; and on AUSTRALIAN, HEART, and PIMA, GA-SRMC yields 16–32% higher ESS

Table 1: Worst-coordinate ESS (mean±std over 40 chains, 20,000 post-burn-in samples) at each method's best stability-capped α; $\delta = 0.01$. **Bold**: better of {SRMC, GA-SRMC} per family; **underlined**: row best; κ: condition number of the posterior precision at the MAP.

	MALA family			FISHERMALA family		
Target (n, d, κ)	base	+ SRMC	+ GA-SRMC	base	+ SRMC	+ GA-SRMC
Gauss 2D (2,2,10^2)	94±21	104±21	**452±17**	9894±371	12459±224	**18425±204**
Australian (690,15,2.8)	2245±170	4885±122	**5937±149**	4087±215	11170±228	**11638±260**
Heart (270,14,2.7)	2821±183	6310±150	**7324±163**	4330±133	9451±150	**9549±159**
Pima (768,9,2.7)	3448±119	7638±137	**10084±298**	5422±177	10975±206	**13358±279**
Ripley (1250,3,1.3)	8405±250	13372±180	**13591±180**	9514±277	15353±207	**15477±167**
German (1000,25,4.6)	1437±83	**4663±152**	4173±110	3259±114	9338±205	**9449±238**
Caravan (5822,86,16)	401±35	922±126	**1353±32**	1694±91	**6381±202**	6022±172
MNIST (4000,785,55)	49±22	74±31	**501±7**	612±46	1389±38	**1399±27**

Table 2: Posterior-mean MSE $\frac{1}{d}\|\bar{X} - \mu_{\text{ref}}\|_2^2$ (mean±std over 40 runs, units of 10^{-4}) for the samplers of Table 1. **Bold**: best per family; **underlined**: row best. Details in Appendix A.5.

	MALA family			FISHERMALA family		
Target	base	+ SRMC	+ GA-SRMC	base	+ SRMC	+ GA-SRMC
Gauss 2D	1880.317±3070.938	1827.031±2775.423	**1.236±1.775**	10.647±12.809	6.563±7.206	**0.227±0.252**
Australian	17.702±9.018	1.273±0.832	**0.908±0.474**	13.722±5.160	0.977±0.436	**0.830±0.333**
Heart	18.157±8.370	1.176±0.527	**0.871±0.328**	12.745±5.779	0.847±0.334	**0.719±0.255**
Pima	14.418±10.815	0.627±0.422	**0.370±0.167**	7.884±4.325	0.346±0.143	**0.289±0.105**
Ripley	5.786±5.515	0.205±0.188	**0.198±0.178**	6.512±5.968	0.232±0.195	**0.223±0.185**
German	24.062±10.052	1.624±0.953	**0.904±0.394**	13.791±4.586	0.937±0.437	**0.635±0.233**
Caravan	126.083±22.058	45.631±8.525	**10.217±1.671**	34.951±6.629	5.949±1.317	**2.938±0.496**
MNIST	685.244±33.782	541.485±26.944	**55.500±2.556**	136.093±5.676	56.842±2.694	**55.000±2.740**

at $\alpha \in \{25, 30\}$. In the FisherMALA family, where the proposal is already preconditioned by the inverse Fisher matrix, the geometry-aware correction is partly additive: the GA-SRMC/SRMC ratio is 1.48 on GAUSS-2D ($\alpha = 20$), 1.01–1.22 on AUSTRALIAN, PIMA, HEART, RIPLEY, GERMAN, and MNIST, and 0.94 on CARAVAN ($\alpha = 10$). Across both families, GA-SRMC matches or improves upon SRMC in 14 of 16 pairs, with the only two ESS exceptions being GERMAN under MALA (4173 vs. 4663) and CARAVAN under FisherMALA (6022 vs. 6381).

Stability cap of GA-SRMC vs. SRMC. Figure 2 traces the worst-coordinate ESS as α is swept on four representative targets. Solid lines with filled markers denote α values within the stability cap; dotted lines with hollow markers denote unstable α values beyond the cap, where the chain mixes rapidly on the distorted surrogate $\pi_\theta^{(M)}$ but exhibits inflated marginal standard deviations relative to π; red and blue stars mark the Table 1 entries for SRMC and GA-SRMC, respectively. Within the MALA family, GA-SRMC remains stable for one (AUSTRALIAN, HEART) or two (PIMA) additional α-grid steps beyond SRMC, matches SRMC's cap on RIPLEY, GERMAN, and CARAVAN (where MALA+SRMC's ESS peaks earlier at $\alpha = 5$), and on high-dimensional MNIST ($d = 785$) remains stable through $\alpha = 20$ whereas MALA+SRMC exits between $\alpha = 3$ and $\alpha = 10$ (Appendix A.4). Under FisherMALA on MNIST, both methods lose stability at $\alpha = 10$ and perform comparably within noise for $\alpha \leq 3$, where the cap is imposed.

Posterior-mean accuracy. Table 2 complements worst-coordinate ESS with the per-dimension MSE of the posterior mean $\bar{X}$ (details in Appendix A.5). Both repulsive samplers are substantially more accurate than their base samplers: on RIPLEY, PIMA, HEART, AUSTRALIAN, and GERMAN, SRMC already reduces the MSE by 14–28× in both families, and GA-SRMC lowers it further in all 16 full-Fisher comparisons, so that GA-SRMC $<$ SRMC $<$ base holds in every row. The reduction over SRMC is largest on the most anisotropic targets (1.236 vs. 1827.031 on GAUSS-2D, 55.500 vs. 541.485 on MNIST, and 10.217 vs. 45.631 on CARAVAN under MALA) and smallest on the near-isotropic RIPLEY ($\kappa = 1.3$; 0.198 vs. 0.205), where M_{opt} is close to a multiple of the identity and GA-SRMC reduces to SRMC. Notably, even in the two pairs where worst-coordinate ESS favored SRMC (GERMAN under MALA and CARAVAN under FisherMALA), GA-SRMC yields a more accurate posterior mean on every one of the 40 chains (0.904 vs. 1.624 on GERMAN and 2.938 vs. 5.949 on CARAVAN): worst-coordinate ESS captures only the slowest-mixing coordinate,

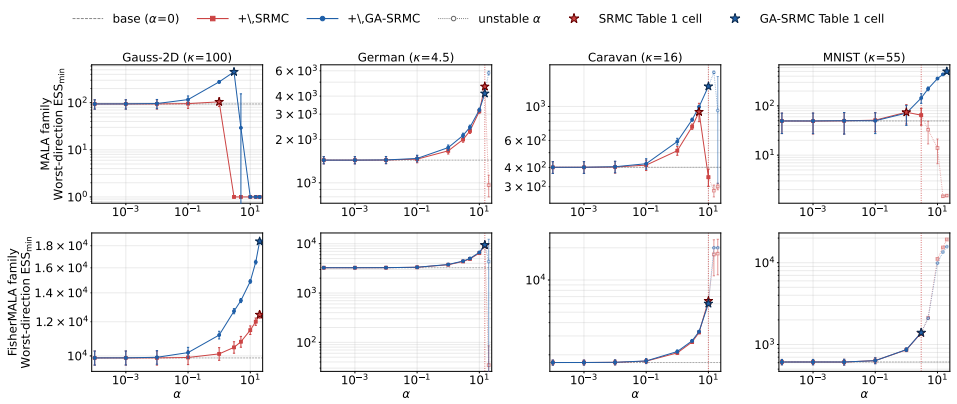

Figure 2: Worst-coordinate ESS versus α on four representative targets (log-log axes) for the MALA (top) and FisherMALA (bottom) proposals. Markers and bars are mean $\pm$ std across 40 chains.

whereas the MSE aggregates estimation error across all d coordinates. Moreover, adding GA-SRMC to MALA lowers the MSE more than replacing MALA with FisherMALA on all eight targets (by $2.45\times$ on MNIST to $32.9\times$ on RIPLEY).

When SRMC is the safer choice. GA-SRMC's advantage on coupled posteriors relies on a faithful Fisher estimate; Appendix A.6 compares the full $O(d^2)$ estimator against an $O(d)$ diagonal estimator that retains only per-coordinate squared scores. In worst-coordinate ESS (Table 5), the diagonal estimator preserves GA-SRMC's advantage on axis-aligned targets such as GAUSS-2D ($2.4\times$ under MALA) and in five of eight rows under diagonal FisherMALA, but flips to SRMC under MALA on the coupled BLR posteriors. In posterior-mean MSE (Table 6), using the diagonal estimate in the history feedback alone (MALA+diagonal GA-SRMC) still achieves a lower mean MSE than MALA+SRMC on all eight targets at $O(d)$ cost. Reusing the same imperfect diagonal estimate in both the proposal and the repulsion (diagonal FisherMALA+diagonal GA-SRMC), in contrast, yields a higher MSE than SRMC on six of the seven BLR targets, consistent with a double penalty when one coarse approximation enters both channels; in that dual-use configuration, SRMC is the safer choice.

5 Conclusion and Future directions

We introduced *Geometry-Aware* SRMC, an improvement over SRMC [34], which substitutes the Euclidean alignment $\theta^\top s(x)$ in the surrogate target with a matrix-weighted expression $\theta^\top M s(x)$, where M is optimally set to the trace-normalized inverse Fisher information. Because the same inverse Fisher already serves as the proposal-side preconditioner in modern Fisher-adaptive samplers [54, 30], GA-SRMC adds essentially no bookkeeping over a Fisher-preconditioned base. Like SRMC, GA-SRMC is a *wrapper*. In this paper, we apply MALA and FisherMALA to a geometry-aware surrogate target, so the main contribution is essentially a recipe that elevates the Fisher information from a mere proposal-side preconditioner to a geometric quantity that simultaneously shapes both the proposal and the score-history feedback. The same recipe should be applicable to any gradient-based sampler equipped with a Fisher-aware preconditioner.

Future directions. The wrapper view of GA-SRMC opens a broader design space than this paper can explore. A natural next step is the Hamiltonian family, e.g., HMC, Riemannian HMC, and Fisher-adaptive HMC [26, 53, 31], which already use a proposal-side metric. We expect the same dual-role Fisher recipe to extend to them with similar ease, making an empirical study on these Hamiltonian samplers the most immediate follow-up. In addition, the closed-form expression for the limiting covariance in our CLT suggests a self-adjusting repellence schedule that may eliminate the need for a stability sweep over individual targets. Discrete-state spaces and strongly multimodal settings pose deeper, more fundamental issues regarding the appropriate Fisher-like quantity and will require longer-term investigation. We interpret GA-SRMC as an initial move toward a more general geometry-aware framework for history-based MCMC.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their constructive comments and suggestions. Jie Hu's contribution to this work was supported in part by startup funding and Oakland University's 2026 Research Committee (URC) Faculty Research Fellowship. Lingyun Chen and Do Young Eun's contributions were supported in part by National Science Foundation grant IIS-2421484.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, 2016.

[2] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998.

[3] Christophe Andrieu and Éric Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, pages 1462–1505, 2006.

[4] Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, 2008.

[5] Roland Assaraf and Michel Caffarel. Zero-variance principle for Monte Carlo algorithms. *Physical Review Letters*, 83(23):4682, 1999.

[6] Andrea Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical Review Letters*, 100:020603, 2008.

[7] Michel Benaim and Charles-Edouard Brehier. Convergence of adaptive biasing potential methods for diffusions. *Comptes Rendus. Mathematique*, 354(8):842–846, 2016.

[8] Michel Benaïm and Charles-Edouard Bréhier. Convergence analysis of adaptive biasing potential methods for diffusion processes. *Communications in Mathematical Sciences*, 17(1):81–130, 2019.

[9] Michel Benaim, Charles-Edouard Brehier, and Pierre Monmarche. Analysis of an adaptive biasing force method based on self-interacting dynamics. *Electronic Journal of Probability*, 25, 2020.

[10] Rajendra Bhatia. *Matrix analysis*. Springer Science & Business Media, 2013.

[11] Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2nd edition, 1999.

[12] Vivek Borkar, Shuhang Chen, Adithya Devraj, Ioannis Kontoyiannis, and Sean Meyn. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *The Annals of Applied Probability*, 35(2):936–982, 2025.

[13] Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2009.

[14] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[15] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1–32, 2017.

[16] Martin Chak, Nikolas Kantas, Tony Lelièvre, and Grigorios A Pavliotis. Optimal friction matrix for underdamped Langevin sampling. *ESAIM: Mathematical Modelling and Numerical Analysis*, 57(6):3335–3371, 2023.

[17] Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. How to compute Hessian-vector products? In *ICLR Blogposts 2024*, 2024. URL `https://iclr-blogposts.github.io/2024/blog/bench-hvp/`.

[18] Eric Darve and Andrew Pohorille. Calculating free energies using average force. *Journal of Chemical Physics*, 115(20):9169–9183, 2001.

[19] Wei Deng, Guang Lin, and Faming Liang. A contour stochastic gradient Langevin dynamics algorithm for simulations of multi-modal distributions. In *Advances in Neural Information Processing Systems*, 2020.

[20] Vishwaraj Doshi, Jie Hu, and Do Young Eun. Self-repellent random walks on general graphs — Achieving minimal sampling variance via nonlinear Markov chains. In *International Conference on Machine Learning*, 2023.

[21] Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

[22] Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

[23] Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2nd edition, 1999.

[24] G Fort, E Moulines, and P Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *The Annals of Statistics*, pages 3262–3289, 2011.

[25] Charles J Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

[26] 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.

[27] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

[28] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[29] Jerome Henin, Tony Lelievre, Michael R Shirts, Omar Valsson, and Lucie Delemotte. Enhanced sampling methods for molecular dynamics simulations [article v1. 0]. *Living Journal of Computational Molecular Science*, 4(1), 2022.

[30] Max Hird and Samuel Livingstone. Quantifying the Effectiveness of Linear Preconditioning in Markov Chain Monte Carlo. *Journal of Machine Learning Research*, 26(119):1–51, 2025.

[31] 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.

[32] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[33] Jie Hu, Yi-Ting Ma, and Do Young Eun. Beyond self-repellent kernels: History-driven target towards efficient nonlinear MCMC on general graphs. In *International Conference on Machine Learning*, 2025.

[34] 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 (To appear in ICML 2026)*, 2026.

[35] Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 3rd edition, 2002.

[36] Ravin Kumar, Colin Carroll, Ari Hartikainen, and Osvaldo Martin. ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, 4(33): 1143, 2019.

[37] Harold J Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.

[38] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[39] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[40] John M. Lee. *Introduction to Smooth Manifolds*. Springer, 2nd edition, 2013.

[41] Kolby Nottingham Markelle Kelly, Rachel Longjohn. The uci machine learning repository. `https://archive.ics.uci.edu`, 2007.

[42] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

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

[44] Radford M Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman and Hall/CRC, 2011.

[45] 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.

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

[47] Brian D Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007.

[48] 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.

[49] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various Metropolis–Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[50] Gareth O Roberts and Jeffrey S Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458–475, 2007.

[51] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[52] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition, 1976.

[53] Adrian Seyboldt, Eliot L Carlson, and Bob Carpenter. Preconditioning Hamiltonian Monte Carlo by minimizing Fisher divergence. *arXiv preprint arXiv:2603.18845*, 2026.

[54] Michalis K. Titsias. Optimal preconditioning and Fisher adaptive Langevin sampling. In *Advances in Neural Information Processing Systems*, 2023.

[55] Peter van der Putten and Maarten van Someren. CoIL Challenge 2000: The insurance company case. In *Technical Report 2000–09, Sentient Machine Research, Amsterdam*, 2000.

[56] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[57] Fugao Wang and David P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters*, 86(10):2050–2053, 2001.

[58] Xiaodong Wu, Wenyi Yu, Chao Zhang, and Phil Woodland. An improved empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, pages 134151–134194, 2024.

[59] Mao Ye, Tongzheng Ren, and Qiang Liu. Stein self-repulsive dynamics: Benefits from past samples. In *Advances in Neural Information Processing Systems*, 2020.

A Additional Experimental Notes

Summary of Appendix A. This appendix collects the experimental details that did not fit into Section 4, organized into six subsections. Appendix A.1 documents the complete setup for Figure 1 on the 2D anisotropic Gaussian: it defines the target, explains the five panels in the first row, and provides unified pseudocode that specializes Algorithm 1 to each of the five samplers. Appendix A.2 writes out the explicit Bayesian logistic-regression (BLR) posterior (18) sampled from in all seven real-world experiments, together with the data preprocessing and the score and Hessian (19) that drive the proposal, the surrogate target, and the stability diagnostic. Appendix A.3 reports the sensitivity sweep over the regularizer δ in $\widehat{S}_{n,\delta} = \widehat{S}_n + \delta I_d$, and shows that the choice $\delta = 0.01$ used in all main-text experiments outperforms the conventional $\delta = 10$ of Titsias [54] on every tested target, with a gain that generally grows with the posterior condition number. Appendix A.4 details the per-target stability cap on the repellence strength α used in Table 1, which is computed via the standardized-marginal diagnostic $\max_j(\widehat{\sigma}_j/\sigma_{\mathrm{Laplace},j}) \le 1.30$ and is determined separately for each (target, method) pair. Appendix A.5 specifies the posterior-mean MSE evaluation used in Tables 2 and 6 and explains how worst-coordinate ESS and posterior-mean MSE complement each other. Finally, Appendix A.6 compares diagonal and full Fisher estimates across all 8 targets, two sampler families, and three repellence variants in both worst-coordinate ESS and posterior-mean MSE, identifying when a full Fisher estimate is required and when SRMC is the safer choice.

A.1 Setup for Figure 1 (anisotropic Gaussian)

Target. We use a 2D axis-aligned anisotropic Gaussian

$$\pi(x_1, x_2) = \mathcal{N}(0, \Sigma), \qquad \Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2), \qquad (\sigma_1, \sigma_2) = (5,\, 0.5),$$

so the per-coordinate condition number is $\kappa = (\sigma_1/\sigma_2)^2 = 100$. This clearly illustrates the geometry-blind effect: the x_1 direction has a small score, $|s_1(x)| = |x_1|/25$, whereas the x_2 direction has a large score, $|s_2(x)| = 4|x_2|$. Consequently, using a single repellence strength α on the raw inner product $\theta^\top s(x)$ pushes too hard along x_2 before it has any meaningful influence along x_1. All chains are initialized at $X_0 = (6,\, 3)^\top$, which is off-center with respect to both axes, so the burn-in phase must traverse both the wide and the narrow directions.

Choosing $\kappa = 100$ makes the geometry-blind effect clearly observable: a single-α SR-MALA cannot properly handle both axes at once. For a milder setting $\kappa = 10$, the geometry-blind tilt from SRMC is still inferior to GA-SRMC, but the worst-coordinate ESS gap is not that significant. For a more extreme $\kappa = 10^4$, the standard MALA chain essentially never explores the slow axis, so the study reduces to comparing the underlying samplers alone. Thus, $\kappa = 100$ provides a sweet spot where it is simultaneously evident both that preconditioning helps and that GA-SRMC strictly outperforms SRMC for a hard target density.

Interpretation of row 1 (one-step schematic). The five panels in the top row of Figure 1 are constructed around a common reference configuration consisting of an interior state X_n, the recent-history samples $\{X_k\}_{k=1}^{n-1}$, and the local score field $s(X) = -\Sigma^{-1}X$. The panels differ exclusively with respect to: (a) the surrogate target ν implicitly targeted by the proposal kernel, (b) the presence or absence of Langevin preconditioning, and (c) whether the surrogate exponent uses the alignment $\theta^\top s$ or $\theta^\top M s$. For each panel, define the empirical average of past scores

$$\theta \triangleq \frac{1}{n-1}\sum_{k=1}^{n-1} s(X_k).$$

The individual cases are as follows:

- **(i) MALA.** The surrogate coincides with the target, $\nu = \pi$, and no tilting term is introduced. The proposal covariance is isotropic, with kernel
$$q_I(\cdot \mid X) = \mathcal{N}\big(X + \epsilon\, s(X),\, 2\epsilon I_d\big).$$
Accordingly, the schematic displays the unmodified score vector s together with a circular noise contour, emphasizing that the drift direction is governed by s and the stochastic scale by ϵ.
- **(ii) FisherMALA [54].** Again $\nu = \pi$ and no tilt is applied, but the proposal is preconditioned:
$$q_A(\cdot \mid X) = \mathcal{N}\big(X + \epsilon\, A\, s(X),\, 2\epsilon\, A\big),$$

where the preconditioner $A \propto S^{-1}$ is the trace-normalized inverse Fisher information. The schematic depicts both s and the preconditioned direction As (with a dashed arc between them), as well as an anisotropic noise ellipse oriented according to A, indicating that preconditioning reorients the drift toward the slow principal axis.

- **(iii) MALA + GA-SRMC.** The proposal noise remains isotropic, but the kernel targets the surrogate
$$\nu = \pi_\theta^{(M)} \propto \pi(x) \exp\big(-\alpha\, \theta^\top M s(x)\big), \qquad M = A,$$
so that the drift is determined by the surrogate score
$$s_\theta^{(M)}(x) \triangleq \nabla_x \log \pi_\theta^{(M)}(x) = s(x) - \alpha\, \nabla_x^2 \log \pi(x)\, M\, \theta.$$
The schematic highlights the modified drift vector $s_\theta^{(M)}$ pointing away from the recent-history cloud, and a shifted (green dashed) level set of $\pi_\theta^{(M)}$, illustrating that score-history induces a repulsive displacement of the surrogate mass away from previously visited regions.
- **(iv) FisherMALA + SRMC.** Here the surrogate uses a geometry-blind (Euclidean) alignment,
$$\nu = \pi_\theta \propto \pi(x) \exp\big(-\alpha\, \theta^\top s(x)\big), \qquad M = I_d,$$
with corresponding surrogate score
$$s_\theta(x) = s(x) - \alpha\, \nabla_x^2 \log \pi(x)\, \theta,$$
while the proposal remains preconditioned by A. The schematic shows both s_θ and As_θ (dashed arc), the shifted (red dashed) level set of π_θ displaced along the dominant-score direction (here x_2, associated with higher curvature), and an anisotropic noise ellipse. Because the tilt direction implied by $M = I_d$ need not coincide with the geometry imposed by A, part of the Metropolis correction is expended compensating for a misaligned surrogate-induced drift.
- **(v) FisherMALA + GA-SRMC.** The surrogate is $\nu = \pi_\theta^{(M)}$ with matched geometry $M = A$ (the same trace-normalized inverse Fisher used for preconditioning). The schematic indicates that the surrogate displacement (green dashed level set) and the preconditioned drift align with the slow x_1 axis: the repulsion induced by score history and the proposal anisotropy act coherently, simultaneously biasing motion toward the slow direction and allocating proposal noise along that direction. This is the only configuration in which both the surrogate shift and the proposal covariance consistently promote exploration along the slow axis.

In particular, panels (iv) and (v) share the proposal-side preconditioner A (the trace-normalized inverse Fisher) but differ in the repulsion-side matrix M: panel (iv) uses $M = I_d$ (SRMC), while panel (v) uses $M = A$ (GA-SRMC). The diagram makes the consequence visible: in panel (iv) the red-dashed surrogate level set of π_θ shifts along the dominant-score axis (the narrow x_2-direction), *against* the proposal preconditioner that stretches noise along the slow x_1-direction; the chain then spends acceptance budget undoing the wrong-axis tilt. In panel (v), the green-dashed surrogate level set of $\pi_\theta^{(M)}$ shifts along the slow x_1-direction, so the score-history correction reinforces rather than opposes the proposal anisotropy. This is exactly the alignment that Theorem 3.1 promotes, from Figure 1 observation to a closed-form universal upper bound on $\Sigma_{\theta\theta}^{(M)}$.

Unified pseudocode for the five samplers. The five panels of Figure 1 differ only in two parts: whether the proposal is preconditioned (precond $\in \{\mathsf{none}, \mathsf{Fisher}\}$, controlling A_n) and which surrogate target is sampled (tilt $\in \{\mathsf{none}, \mathsf{SRMC}, \mathsf{GA\text{-}SRMC}\}$, controlling M_n). Algorithm 2 below is a five-way specialization of the main-body Algorithm 1 (Section 3): it inherits the same online Fisher state, the same trace normalization, and the same SA recursion, but covers all five panels by branching on (precond, tilt) via if-else. Table 3 lists the corresponding configurations; in particular, panel (v) (Fisher,GA-SRMC) is exactly Algorithm 1 with $A_n = M_n$, while the other four panels are obtained by switching off either the proposal-side preconditioning or the repulsion-side metric. All variants share the same constant-step-size Robbins–Monro update for $\log \epsilon$ and (where used) the same online Fisher state $\widehat{S}_n$. As in Algorithm 1, the metric used at iteration n is computed from $\hat{S}_n$ and the Fisher state is updated with the score at the new sate X_{n+1}, using the same step size γ_{n+1} as the score history. We write $\xi_n \sim \mathcal{N}(0, I_d)$, a_n for the realized Metropolis acceptance probability, and $\widehat{M}_n \triangleq (\mathrm{tr}(\widehat{S}_{n,\delta})/d)\widehat{S}_{n,\delta}^{-1}$ for the trace-normalized inverse Fisher. In this paper, the Hessian-vector product $\nabla_x^2 \log \pi(X_n)\, M_n \theta_n$ never materializes the $d \times d$ Hessian: it is an elementwise multiplication for the diagonal Gaussian target here and is approximated via automatic differentiation [46, 27] or via finite differences of gradients $\nabla_x^2 \log \pi(x)\, v \approx (\nabla_x \log \pi(x + \zeta v) - \nabla_x \log \pi(x))/\zeta$ as in [17].

Algorithm 2 Unified per-iteration update covering all five panels of Figure 1.

Require: state X_n, score history θ_n (used only when tilt$\neq$none), Fisher state $\widehat{S}_n$, switches (precond, tilt) from Table 3.

Compute $\widehat{M}_n$ from $\widehat{S}_n$ (needed only if precond $=$ Fisher or tilt $=$ GA-SRMC).
Proposal preconditioner: $A_n \leftarrow \widehat{M}_n$ if precond $=$ Fisher, else $A_n \leftarrow I_d$.
Repulsion metric: $M_n \leftarrow \widehat{M}_n$ if tilt $=$ GA-SRMC, else $M_n \leftarrow I_d$.
if tilt $=$ none **then** ▷ *Langevin drift*
  $\widetilde{s}_n \leftarrow s(X_n)$.
else
  $\widetilde{s}_n \leftarrow s(X_n) - \alpha\, \nabla_x^2 \log \pi(X_n)\, M_n \theta_n$ ▷ *Surrogate score computation*
Propose $Y \leftarrow X_n + \epsilon\, A_n \widetilde{s}_n + \sqrt{2\epsilon}\, A_n^{1/2} \xi_n$.
if tilt $=$ none **then** ▷ *Metropolis acceptance*
  $a_n \leftarrow \min\{1,\, \pi(Y) q_{A_n}(X_n|Y) \,/\, [\pi(X_n) q_{A_n}(Y|X_n)]\}$.
else
  a_n via Eq. (4) with $A = A_n$ and $M = M_n$.
Update $X_{n+1} \leftarrow Y$ with prob. a_n else $X_{n+1} \leftarrow X_n$; $\log \epsilon \leftarrow \log \epsilon + c_\epsilon(a_n - 0.574)$.
if tilt $\neq$ none **then** ▷ *Score-history SA update*
  $\theta_{n+1} \leftarrow \theta_n + \gamma_{n+1}(s(X_{n+1}) - \theta_n)$.
if precond $=$ Fisher **or** tilt $=$ GA-SRMC **then** ▷ *Online Fisher update*
  $\widehat{S}_{n+1} \leftarrow (1-\gamma_{n+1})\widehat{S}_n + \gamma_{n+1}\, \mathrm{diag}(s(X_{n+1}) s(X_{n+1})^\top)$.

The bottom-row panels of Figure 1 are the empirical sample clouds and per-coordinate trace plots produced by Algorithm 2 under the five configurations of Table 3. Effective sample size is computed coordinate-wise using Geyer's initial positive-sequence estimator [25]; the reported pairs $(\mathrm{ESS}(x_1), \mathrm{ESS}(x_2))$ correspond to the post-burn-in portion of $N = 1000$ samples.

Table 3: Configuration switches for Algorithm 2 corresponding to the five panels of Figure 1; $\widehat{M}_n \propto \widehat{S}_{n,\delta}^{-1}$ is the trace-normalized inverse Fisher.

Panel	Method	precond	tilt	A_n	M_n
(i)	MALA	none	none	I_d	—
(ii)	FisherMALA	Fisher	none	$\widehat{M}_n$	—
(iii)	MALA + GA-SRMC	none	GA-SRMC	I_d	$\widehat{M}_n$
(iv)	FisherMALA + SRMC	Fisher	SRMC	$\widehat{M}_n$	I_d
(v)	FisherMALA + GA-SRMC	Fisher	GA-SRMC	$\widehat{M}_n$	$\widehat{M}_n$ (shared)

A.2 Bayesian logistic regression: target distribution

The seven real-world targets in Section 4 (Ripley, Pima, Heart, Australian, German, Caravan, MNIST) are all Bayesian logistic-regression (BLR) posteriors over the parameter $\beta \in \mathbb{R}^d$, defined as follows. Given a dataset $\{(x_i, y_i)\}_{i=1}^N$ with feature vectors $x_i \in \mathbb{R}^d$ and binary labels $y_i \in \{0, 1\}$, we collect the features into the design matrix $\Phi \in \mathbb{R}^{N \times d}$ whose i-th row is $x_i^\top$. Each x_i consists of an intercept entry followed by the covariates; for all datasets except MNIST, every covariate is z-scored (centered and scaled to unit variance), and the whole design matrix is then divided by $\sqrt{N}$, which keeps the score magnitude comparable across datasets of very different sizes. For MNIST, which is binarized to a two-class problem in the spirit of Titsias [54], we use the digits 3 versus 8, draw a random subsample of $N = 4000$ images, and keep the raw pixel intensities scaled to $[0, 1]$ without z-scoring; with the intercept this gives $d = 785$, and the design is again divided by $\sqrt{N}$. Combining the standard logistic likelihood with the isotropic Gaussian prior $\pi_0(\beta) = \mathcal{N}(\beta; 0, \tau_0^{-1} I_d)$ of precision $\tau_0 > 0$ gives the (unnormalized) BLR posterior

$$\pi(\beta) \;\propto\; \exp\left\{ -\frac{\tau_0}{2}\|\beta\|^2 + \sum_{i=1}^{N} \Big[y_i\, x_i^\top \beta - \log\Big(1 + e^{x_i^\top \beta}\Big) \Big] \right\}, \tag{18}$$

Table 4: Worst-direction ESS for FisherMALA+GA-SRMC across δ values (mean$\pm$std over 40 replicate chains, 20000 post-burn-in samples) at each target's stability-capped best α. The smallest δ tested is preferred on every row; bold marks the row maximum.

Target	κ	$\alpha^\star$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 1$	$\delta = 10$
Ripley	1.3	20	**13958**$\pm$138	13574$\pm$197	12500$\pm$262	12027$\pm$262
Heart	2.7	20	**9549**$\pm$159	8716$\pm$162	6982$\pm$168	6402$\pm$148
Pima	2.7	20	**11164**$\pm$193	10319$\pm$217	8409$\pm$143	7719$\pm$128
Australian	2.8	20	**9148**$\pm$124	7699$\pm$150	5480$\pm$127	4950$\pm$118
German	4.5	15	**9449**$\pm$238	7798$\pm$184	5354$\pm$155	4763$\pm$210
Caravan	16.3	10	**6022**$\pm$172	4693$\pm$116	1096$\pm$1177	317$\pm$89
Gauss-2D	100	3	**12695**$\pm$173	4995$\pm$106	671$\pm$37	10$\pm$40

which is the target sampled from in all BLR experiments of Section 4, with the prior precision fixed to $\tau_0 = 0.1$ (prior standard deviation $\sqrt{10}$) throughout. The corresponding score and Hessian are

$$s(\beta) \;=\; -\tau_0\beta + \Phi^\top\big(y - \hat{p}(\beta)\big), \qquad -\nabla^2 \log \pi(\beta) \;=\; \tau_0 I_d + \Phi^\top D(\beta)\Phi, \tag{19}$$

where $\hat{p}_i(\beta) = \sigma(x_i^\top \beta) = 1/(1 + e^{-x_i^\top \beta})$ and $D(\beta) = \mathrm{diag}(\hat{p}_i(\beta)(1 - \hat{p}_i(\beta)))$. Evaluating (19) at the MAP $x^\star = \arg\max_\beta \pi(\beta)$ yields the observed Fisher $Z = -\nabla^2 \log \pi(x^\star)$, which initializes the Fisher state in all BLR experiments (Section 4) and is used in the standardized-marginal stability diagnostic of Appendix A.4 (Eq. (20)); the off-diagonal coupling through $\Phi^\top D(x^\star)\Phi$ is what governs the diagonal-vs-full-Fisher gap discussed in Appendices A.6 and D.2.1.

A.3 Choice of the regularizer δ

The online inverse-Fisher state employed by GA-SRMC and FisherMALA is based on the regularized matrix $\widehat{S}_{n,\delta} = \widehat{S}_n + \delta I_d$. The scalar δ therefore bounds the eigenvalues of $\widehat{S}_{n,\delta}^{-1}$ above by $1/\delta$ and, equivalently, governs the effective proposal step size in directions where $\widehat{S}_n$ has small eigenvalues. The original FisherMALA work [54] fixes $\delta = 10$. We performed a sweep over $\delta \in \{0.01, 0.1, 1.0, 10.0\}$ on seven targets, running each target at the value $\alpha^\star$ listed in Table 4 (taken from an earlier tuning round restricted to $\alpha \leq 20$) and holding all other hyperparameters constant; Table 4 shows the resulting worst-coordinate ESS for FisherMALA+GA-SRMC. Its $\delta = 0.01$ column therefore coincides with Table 1 on Heart, German, and Caravan, where $\alpha^\star$ equals the Table 1 value, but not on Ripley, Pima, Australian, and Gauss-2D, where Table 1 uses $\alpha = 30$, 25, 25, and 20, respectively. Across all tested targets, the smallest δ in this grid yields the highest ESS, and the improvement relative to the conventional $\delta = 10$ generally grows with the posterior's condition number. On Ripley, whose posterior is nearly isotropic, the gain is modest (16%); on Gauss-2D, it exceeds three orders of magnitude. Mechanistically, this is because a smaller δ allows the preconditioner $A_n = (d/\operatorname{tr} \widehat{S}_{n,\delta}^{-1})\, \widehat{S}_{n,\delta}^{-1}$ to more faithfully approximate the true inverse Fisher, so the sampler can take larger steps along low-curvature directions and mix much faster in those components. We did not explore $\delta < 0.01$, the smallest value in the sweep, and we adopt $\delta = 0.01$ in all main-text experiments.

A.4 Per-target stability caps from $\widehat{\sigma}/\sigma_{\mathrm{Laplace}}$

A score-repellent sampler at fixed θ in fact samples from the surrogate distribution $\pi_\theta^{(M)}(x) \propto \pi(x)\exp(-\alpha\theta^\top M s(x))$, not from π itself. These two coincide only in the limiting regime $\theta_n \to \theta^\star = 0$ guaranteed by Proposition 2.1. For a finite number of iterations n and sufficiently large α, the chain may temporarily explore a distorted target, in which case comparing effective sample sizes across samplers would not be meaningful. To guard against this, we use the standardized-marginal diagnostic of Titsias [54] originally proposed for FisherMALA validation. After burn-in, we compute the per-coordinate empirical standard deviation $\widehat{\sigma}_j$ of the j-th component of the chain (based on the pooled sample $\{X_n^{(j)}\}_{n=N_{\mathrm{burn}}+1}^{N_{\mathrm{burn}}+N_{\mathrm{collect}}}$ across the 40 replicate chains) and compare it to the Laplace prediction

$$\sigma_{\mathrm{Laplace},j} \;=\; \sqrt{(Z^{-1})_{jj}}, \qquad Z \;\triangleq\; -\nabla^2 \log \pi(x^\star) \;\succ\; 0, \tag{20}$$

where $x^\star = \arg\max_x \pi(x)$ is the MAP, Z is the Hessian of the negative log-density at $x^\star$ (also known as the observed Fisher information at $x^\star$, and equal to $\mathrm{Cov}_\pi(s,s) = S$ when π is Gaussian), and Z^{-1} denotes its inverse. The j-th diagonal entry $(Z^{-1})_{jj}$ approximates the j-th marginal posterior variance under the Gaussian Laplace approximation $\pi(x) \approx \mathcal{N}(x; x^\star, Z^{-1})$. We define the stability cap on α as the largest value in our grid for which $\max_j(\widehat{\sigma}_j/\sigma_{\mathrm{Laplace},j}) \leq 1.30$, meaning that it flags the chain as unstable as soon as any marginal standard deviation inflates by more than 30% relative to the Laplace benchmark. The ratio is rounded to two decimals before this comparison; three cells of Table 1 pass only after this rounding (GA-SRMC on Australian and Pima under FisherMALA at $\alpha = 25$, and on Pima under MALA at $\alpha = 30$).

The caps in Table 1 are determined separately for each method. We swept over $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 3, 5, 10, 15, 20\}$ on all targets, extended by $\alpha \in \{25, 30, 50\}$ on the six BLR targets other than MNIST, and recorded the std-ratio for every configuration. For each (target, method) pair, the cap is the largest α in this set that still satisfies the std-ratio condition, and Table 1 reports each method at its best-ESS α at or below its cap; for example, MALA+SRMC is reported at $\alpha = 5$ on Caravan and at $\alpha = 1$ on MNIST, below its caps of 10 and 3. For SRMC, this procedure yields caps of $\alpha = 30$ on Ripley, $\alpha = 25$ on Australian for FisherMALA alone, $\alpha = 20$ for the remaining Australian chains as well as Heart and Pima, $\alpha = 15$ on German, $\alpha = 10$ on Caravan, and $\alpha = 3$ on MNIST. For GA-SRMC, the caps are $\alpha = 30$ on Ripley and on Pima under MALA, $\alpha = 25$ on Australian, on Heart under MALA, and on Pima under FisherMALA, $\alpha = 20$ on Heart under FisherMALA, $\alpha = 15$ on German, and $\alpha = 10$ on Caravan; within the MALA family, they are thus higher than SRMC's on Australian, Heart, and Pima and equal on Ripley, German, and Caravan. On MNIST MALA the GA-SRMC cap is $\alpha = 20$, with std-ratios at $\alpha \in \{3, 5, 10, 15, 20\}$ equal to $\{0.96, 0.93, 0.90, 0.88, 0.87\}$, decreasing monotonically as α grows. In contrast, SRMC on MNIST leaves the stability regime between $\alpha = 3$ (std-ratio 1.14) and $\alpha = 10$ (std-ratio 1.83). This discrepancy in caps matches the mechanism described in Section 4: using MALA as the baseline, GA-SRMC's weighted tilt distributes repulsion according to the Fisher geometry, allowing the chain to remain well-behaved as α increases, whereas SRMC's unscaled tilt overemphasizes the coordinate with the largest score.

A.5 Posterior-mean MSE metric

Tables 2 and 6 measure Monte Carlo accuracy by the error of the posterior-mean estimate. Each configuration is run as 40 independent chains with 2,000 burn-in iterations and 20,000 retained samples, matching the ESS metric. For chain b, let $\bar{X}^{(b)}$ be the plain, equally weighted average of its 20,000 retained samples, i.e., the estimator of $\mathbb{E}_\pi[X]$ obtained with $f(x) = x$; this is the estimator we evaluate, and it differs from the SA estimate μ_N returned by Algorithm 1. Each table entry reports the mean $\pm$ standard deviation over the 40 chains of $\frac{1}{d}\|\bar{X}^{(b)} - \mu_{\mathrm{ref}}\|_2^2$, in units of 10^{-4}. For the BLR targets, the reference mean μ_{ref} is computed from four NUTS chains [31], each with 2,000 warm-up iterations and 200,000 retained draws (800,000 draws in total); for Gauss-2D, μ_{ref} is the exact mean 0. Because the MSE is non-negative and right-skewed (most visibly for MALA and MALA+SRMC on Gauss-2D), $\pm$ one standard deviation describes the spread across runs rather than a symmetric confidence interval. The MALA and MALA+SRMC columns of Table 6, which do not use a Fisher estimate, are carried over from Table 2; its four diagonal-Fisher configurations come from a separate set of 40 runs. Comparisons between these two sets of runs, including those between Tables 2 and 6, are therefore descriptive. On Caravan, where the two MALA-family samplers are evaluated at their respective best-ESS values ($\alpha = 5$ for SRMC and $\alpha = 10$ for GA-SRMC) in Tables 1 and 2, GA-SRMC also achieves a lower MSE at the common value $\alpha = 5$ (24.468 vs. 45.631, lower on all 40 chains).

ESS and MSE are complementary. The per-chain worst-coordinate ESS of Tables 1 and 5 quantifies the autocorrelation of the slowest-mixing coordinate of a chain, whereas the MSE quantifies the finite-run error of the full d-dimensional posterior-mean estimate, aggregating all coordinates. The two metrics can therefore rank methods differently, as for German under MALA and Caravan under FisherMALA (Section 4) and for the diagonal estimator under MALA (Appendix A.6); we report both.

Table 5: Diagonal vs full Fisher estimators across all 8 targets, MALA and FisherMALA families, and three repellence variants, at the same per-cell α as Table 1 (the stability caps were not re-evaluated for the diagonal estimator). Mean worst-coordinate ESS across 40 replicate chains.

		MALA family						FISHERMALA family					
		diagonal			full			diagonal			full		
Target	κ	base	+ SR	+ GA	base	+ SR	+ GA	base	+ SR	+ GA	base	+ SR	+ GA
Gauss-2D	100	94	104	**250**	94	104	**452**	9884	10717	**13220**	9894	12459	**18425**
Australian	2.8	2245	**4885**	3077	2245	4885	**5937**	2912	3983	**4030**	4087	11170	**11638**
Heart	2.7	2821	**6310**	3870	2821	6310	**7324**	2963	3930	**3933**	4330	9451	**9549**
Pima	2.7	3448	**7638**	4921	3448	7638	**10084**	3313	4365	**4539**	5422	10975	**13358**
Ripley	1.3	8405	**13372**	9917	8405	13372	**13591**	9478	11266	**11305**	9514	15353	**15477**
German	4.6	1437	**4663**	2034	1437	**4663**	4173	1394	**1992**	1984	3259	9338	**9449**
Caravan	16	401	**922**	695	401	922	**1353**	375	**632**	624	1694	**6381**	6022
MNIST	55	49	**74**	52	49	74	**501**	67	**154**	143	612	1389	**1399**

A.6 Geometric fidelity: diagonal vs full Fisher state

The Fisher state $\widehat{S}_n$ in Algorithm 1 can be instantiated at two levels of geometric fidelity. Writing $s_{n+1} \triangleq s(X_{n+1})$, the *diagonal* estimator stores only the per-coordinate squared-score recursion $\widehat{r}_{n+1} = (1-\gamma_{n+1})\widehat{r}_n + \gamma_{n+1}\, s_{n+1} \odot s_{n+1}$ at memory and per-step cost $O(d)$; it captures axis-aligned anisotropy but discards cross-coordinate score correlations. The *full* estimator maintains the rank-one recursion $\widehat{S}_{n+1} = (1-\gamma_{n+1})\widehat{S}_n + \gamma_{n+1} s_{n+1} s_{n+1}^\top$ at $O(d^2)$ memory and per-step cost (plus an amortized $O(d^3/K)$ Cholesky-of-inverse refresh). The same regularized matrix $\widehat{S}_{n,\delta} = \widehat{S}_n + \delta I_d$ feeds the proposal preconditioner A_n and (for GA-SRMC) the repulsion-side metric M_n in both cases. As in Section 4, the ESS ablation of Table 5 initializes the Fisher state at the MAP Hessian (at its diagonal for the diagonal estimator) and freezes it after one update (Remark 2); it therefore compares the metrics induced by the MAP Hessian and by its diagonal, not two online estimation trajectories. Detailed implementations are in Appendix D.2.

Table 5 reports the worst-coordinate ESS for both estimators on the full 8-target suite, MALA and FisherMALA families, three repellence variants {base, + SRMC, + GA-SRMC}, all at the same per-target α used in Table 1. The diagonal estimator costs $O(d)$ memory and per-step time, while the full estimator costs $O(d^2)$ memory plus an amortized $O(d^3/K)$ Cholesky refresh. By construction, the MALA-family base and + SRMC cells do not exercise the Fisher state ($A_n = I_d$ for MALA; $M_n = I_d$ for SRMC), so the diagonal and full columns coincide there. The informative diagonal-vs-full comparison is therefore the four cells where the Fisher state enters: MALA + GA-SRMC (through M_n) and the entire FisherMALA family (through A_n, and additionally through M_n for + GA-SRMC).

We observe three patterns in ESS. First, on Gauss-2D, whose covariance is exactly diagonal in the standard basis, GA-SRMC keeps its advantage under the diagonal estimator: the diagonal MALA GA-SRMC reaches 250 versus SRMC's 104 (2.4×) and the diagonal FisherMALA GA-SRMC reaches 13,220 versus SRMC's 10,717 (1.23×), smaller than, but in the same direction as, the corresponding full-Fisher ratios (4.3× and 1.48×). Second, under diagonal FisherMALA, GA-SRMC remains the within-family winner in five of the eight rows (Gauss-2D, Pima, Australian, and, by margins below 0.5%, Ripley and Heart), while German, Caravan, and MNIST flip to SRMC by margins of 0.4–7.7%. Third, on the BLR posteriors with a MALA proposal and a diagonal M_n, SRMC attains a higher worst-coordinate ESS than GA-SRMC: by 1.3–1.6× on Australian, Heart, Pima, Ripley, and Caravan, by 2.3× on German, and by 1.4× on MNIST. This third pattern is not specific to GA-SRMC: the broader preconditioning literature has long documented that diagonal preconditioners are a coarse approximation on posteriors whose dominant directions are not coordinate-aligned [26, 54], and Hird and Livingstone [30, Theorem 3.7] formalizes the failure quantitatively as a function of off-diagonal correlation. The mechanism in our setting is the analog on the repulsion side: when M_n is constrained to a coordinate-wise rescaling of θ_n but the dominant posterior directions are not coordinate-aligned, GA-SRMC's metric-weighted tilt $\theta^\top M_n s(x)$ can amplify the wrong directions. In worst-coordinate ESS alone, SRMC would thus appear to be the safer choice with a diagonal M_n under MALA; the posterior-mean MSE below qualifies this conclusion.

Table 6: Posterior-mean MSE $\frac{1}{d}\|\bar{X} - \mu_{\mathrm{ref}}\|_2^2$ with diagonal Fisher estimates (mean±std over 40 runs, in units of 10^{-4}; lower is better; details in Appendix A.5). The MALA and MALA+SRMC columns do not use a Fisher estimate and are carried over from Table 2; the four diagonal-Fisher configurations come from a separate set of 40 runs, so comparisons across the two sets of runs are descriptive. **Bold**: best within each sampler family; **underlined**: best in the row.

	MALA family			Diagonal-FISHERMALA family		
Target	base	+ SRMC	+ diag. GA-SRMC	base	+ SRMC	+ diag. GA-SRMC
Gauss 2D	1880.317±3070.938	1827.031±2775.423	**1.234±1.777**	10.244±11.280	6.999±6.965	**0.224±0.260**
Australian	17.702±9.018	1.273±0.832	**0.869±0.384**	17.008±7.469	**0.947±0.484**	0.957±0.456
Heart	18.157±8.370	1.176±0.527	**0.986±0.422**	18.692±9.034	**0.963±0.452**	1.020±0.518
Pima	14.418±10.815	0.627±0.422	**0.624±0.434**	14.059±10.421	**0.573±0.370**	0.585±0.364
Ripley	5.786±5.515	0.205±0.188	**0.177±0.154**	4.447±3.954	0.175±0.187	**0.173±0.178**
German	24.062±10.052	1.624±0.953	**1.595±0.765**	25.876±10.092	**1.796±0.963**	1.933±1.123
Caravan	126.083±22.058	45.631±8.525	**24.693±5.422**	144.471±32.160	**28.896±8.045**	30.695±9.182
MNIST	685.244±33.782	541.485±26.944	**356.518±18.107**	896.389±68.161	**519.022±39.772**	550.291±50.981

Posterior-mean MSE with diagonal Fisher estimates. Table 6 repeats the MSE comparison of Table 2 with diagonal Fisher estimates. Four observations follow.

- **(i) Geometric fidelity depends on the cross-coordinate structure of the Fisher matrix, not on κ alone.** Full and diagonal MALA+GA-SRMC are similar on Ripley (0.198 vs. 0.177) but differ on Caravan (10.217 vs. 24.693) and MNIST (55.500 vs. 356.518). On Gauss-2D, the MSE changes little despite $\kappa = 100$ (MALA+GA-SRMC: 1.236 vs. 1.234; FisherMALA+GA-SRMC: 0.227 vs. 0.224) because its Fisher matrix is diagonal, although its ESS does change (Table 5).
- **(ii) Under MALA, a diagonal metric still lowers the MSE relative to SRMC.** MALA+diagonal GA-SRMC has a lower mean MSE than MALA+SRMC on all eight targets, with differences within noise on Pima (0.624 vs. 0.627) and German (1.595 vs. 1.624); at $O(d)$ cost, it reduces the MALA MSE from 14.418 to 0.624 on Pima and from 126.083 to 24.693 on Caravan. In MSE, the ESS-based preference for SRMC with a diagonal M_n under MALA is therefore not borne out (a descriptive comparison, since the MALA+SRMC entries come from the full-Fisher runs).
- **(iii) Double penalty.** Diagonal-FisherMALA is worse than MALA on Heart (18.692 vs. 18.157), German (25.876 vs. 24.062), Caravan (144.471 vs. 126.083), and MNIST (896.389 vs. 685.244). Reusing this diagonal estimate in GA-SRMC gives a higher mean MSE than SRMC on six of the seven BLR targets (all except Ripley), and this dual-use configuration is also worse than MALA+diagonal GA-SRMC on five of the seven BLR targets, including Caravan (30.695 vs. 24.693) and MNIST (550.291 vs. 356.518). These results are consistent with a double penalty when the same imperfect estimate enters both the proposal and the repulsion.
- **(iv) Where a diagonal estimate is best used.** MALA+diagonal GA-SRMC outperforms diagonal-FisherMALA+SRMC on five of the eight targets (Gauss-2D, Australian, German, Caravan, MNIST), e.g., 356.518 vs. 519.022 on MNIST. The former keeps the proposal unpreconditioned and uses the marginal scales only in the history feedback; the latter changes the proposal scales without representing rotated directions (i.e., off-diagonal terms), while its SRMC term has no geometry-aware weight matrix M.

Practical takeaway. GA-SRMC's MALA-family ESS advantages in Table 1 (4.3× on Gauss-2D, 6.8× on MNIST, 1.47× on Caravan, and 16–32% on Australian, Heart, and Pima) require a full Fisher estimate on coupled posteriors, whereas a diagonal one suffices on axis-aligned targets such as Gauss-2D. When only a diagonal estimate is affordable on a coupled posterior, the two metrics disagree on whether SRMC or MALA+diagonal GA-SRMC is preferable (worst-coordinate ESS favors the former, MSE the latter), but reusing the same diagonal estimate in both the proposal and the repulsion is the configuration to avoid: there, SRMC is the safer choice.

B Proofs of Section 2

Summary of Appendix B. This appendix includes the assumptions on the target and on the fixed-M iterates (Appendix B.1), the derivative identity for the score-tilted log-partition function

(Appendix B.2, Lemma B.1), the self-contained proof of the well-posedness and equilibrium claim (Proposition 2.1; Appendix B.3), an explicit derivation of the Jacobian $J_\star^{(M)}$ (Appendix B.4), and the proof of the almost-sure convergence and CLT for fixed M (Proposition 2.2; Appendix B.5), which is also the template for the online-metric proof in Appendix C.5. Throughout, we use $\beta \triangleq 1 - \mathbf{1}_{\{\rho=1\}}/2 \in \{1/2, 1\}$.

B.1 Assumptions

We adopt verbatim the regularity assumptions of Hu et al. [34, Section 3] on the target π and the family of base kernels, specialized to our surrogate family $\{\pi_\theta^{(M)}\}_{\theta\in\mathbb{R}^d,\, M=M^\top\succ 0}$. Throughout, the test function $f : \mathbb{R}^d \to \mathbb{R}^m$ is π-square-integrable.

Assumption 1 (Properties of the target distribution)**.** Following Hu et al. [34, Assumption 1]: on $\mathbb{R}^d$, the score function $s(x) = \nabla_x \log \pi(x) = -\nabla_x U(x)$ is L-Lipschitz,[2] i.e. $\|s(x)-s(x')\| \le L\|x-x'\|$ for all $x, x' \in \mathbb{R}^d$, and the potential energy U has super-linear tail growth: there exist constants $c, R > 0$ and $p \in (1, 2]$ such that $U(x) \ge c\|x\|^p$ for all $\|x\| \ge R$, together with the asymptotic regularity $r^{-(p-2)}\nabla_x^2 U(r\hat{x}) \to W_\infty(\hat{x})$ as $r \to \infty$ uniformly in $\hat{x} \in \mathbb{S}^{d-1}$, for some measurable positive-definite $W_\infty : \mathbb{S}^{d-1} \to \mathbb{R}^{d\times d}$.

Assumption 2 (Uniform drift and kernel Lipschitzness)**.** Adopting Hu et al. [34, Assumption 2] based on Borkar et al. [12], the family $\{K_{\theta,M}\}$ targeting the surrogates $\{\pi_\theta^{(M)}\}$ is geometrically ergodic with the surrogate as its invariant distribution, satisfies the uniform (DV3) drift condition, and is jointly Lipschitz in (θ, M) on compact sets; full statements are reproduced from Hu et al. [34, Appendix C.3].

Assumption 3 (Online stability)**.** For the online iterates of Theorem 3.2, with $\theta_0 = 0$ and $\widehat{S}_0 \succ 0$: (a) almost surely, $\sup_{n\ge 0} \lambda_{\max}(\widehat{S}_n) < \infty$ and $\inf_{n\ge 0} \lambda_{\min}(\widehat{S}_n) > 0$, where the bounds may depend on the sample path; (b) on every compact set of pairs (θ, M) with $M = M^\top \succ 0$, Assumption 2 holds in the (DV3) form of Borkar et al. [12, A2], including its conditions on the sublevel sets of W, with functions and constants that may depend on the compact set, and $\|s(x)\|^2 + \|f(x)\| \le c\, W(x)$ for all x.

The Lipschitz score in Assumption 1 is the standard regularity used throughout MCMC theory [22, 16]; without it, the surrogate-score correction $-\alpha\, \nabla_x^2 \log \pi(x)\, M\theta$ in the GA-SRMC drift would be unbounded. The super-linear tail growth ($p > 1$) is what guarantees the normalizing constant $Z_\theta^{(M)} < \infty$ for every θ and α (Lemma B.2 below): without it, the score-tilted exponent $-\alpha\theta^\top M s(x)$ could grow linearly enough to overpower the polynomial decay of $\pi(x)$ in the tails. Assumption 2 is the controlled-Markov-noise condition needed by the SA CLT [12]; verified for SR-MH and SR-MALA at $M = I_d$ in Hu et al. [34, Appendix C.3], the same proof extends verbatim to general fixed $M \succ 0$ once θ is replaced by $M\theta$ in the surrogate density, since $\theta \mapsto M\theta$ is a fixed linear isomorphism.

Assumption 3 is used only for Theorem 3.2. Part (a) takes the place of the ODE@∞ condition Borkar et al. [12, A3], which is not available for the augmented iterate (Step 3 of Appendix C.5). It concerns $\widehat{S}_n$ alone, because θ_n and $\widehat{S}_n$ share their gain, so that $\|\theta_n\|^2 < \lambda_{\max}(\widehat{S}_n)$ (Lemma C.2). The lower bound is needed because (16) inverts $\widehat{S}_n$. Part (a) is also necessary for Theorem 3.2(i), since $\widehat{S}_n \succ 0$ for every n and $\widehat{S}_n \to S \succ 0$ force it. Part (b) reads Assumption 2 on compact sets of (θ, M), where the iterates stay under part (a), and extends the moment envelope to the new noise coordinate $s(x)s(x)^\top$.

B.2 A derivative identity for the score-tilted surrogate

Lemma B.1 (Score-tilted derivative identity)**.** *In this lemma only, let $h : \mathbb{R}^d \to \mathbb{R}^k$ be any $\pi_\theta^{(M)}$-integrable function and define*

$$\Phi_h(\theta) \;\triangleq\; \mathbb{E}_{\pi_\theta^{(M)}}[h(X)] \;=\; \frac{\int h(x)\,\pi(x)\, e^{-\alpha\,\theta^\top M s(x)}\, dx}{Z_\theta^{(M)}}.$$

[2] We slightly abuse mathematical notation here and in Appendix B.5 by using $U(x) \triangleq -\log \pi(x)$ for the potential. This should never be confused with the long-run covariance U_{ab} defined in (9).

Then

$$\nabla_\theta \Phi_h(\theta) \;=\; -\alpha \operatorname{Cov}_{\pi_\theta^{(M)}}(h, Ms) \;=\; -\alpha \operatorname{Cov}_{\pi_\theta^{(M)}}(h, s)\, M. \tag{21}$$

In particular, evaluated at $\theta = 0$ (so that $\pi_0^{(M)} = \pi$), $\nabla_\theta \mathbb{E}_{\pi_\theta^{(M)}}[h(X)]\big|_{\theta=0} = -\alpha \operatorname{Cov}_\pi(h, s)\, M$.

Proof. Differentiating the ratio under the integral sign in θ (justified by Assumption 1, which makes the integrand and its θ-derivative dominated for θ in a neighborhood of any fixed point) yields the standard log-partition identity for an exponential family with sufficient statistic $-\alpha M s(X)$:

$$\nabla_\theta \Phi_h(\theta) = \mathbb{E}_{\pi_\theta^{(M)}}[h(X) \cdot (-\alpha M s(X))^\top] - \Phi_h(\theta) \cdot \mathbb{E}_{\pi_\theta^{(M)}}[-\alpha M s(X)]^\top = -\alpha \operatorname{Cov}_{\pi_\theta^{(M)}}(h, Ms).$$

The second equality in (21) uses the linearity of covariance: $\operatorname{Cov}(h, Ms) = \operatorname{Cov}(h, s)\, M^\top = \operatorname{Cov}(h, s)\, M$ since M is symmetric. □

We invoke Lemma B.1 twice with $h = s$ and $h = f$ at $\theta = 0$, giving the two Jacobian blocks

$$\nabla_\theta S_M(\theta)\big|_{\theta=0} = -\alpha\, S\, M, \qquad \nabla_\theta F_M(\theta)\big|_{\theta=0} = -\alpha\, C\, M,$$

which combine with the explicit $-\theta$ and $-\mu$ blocks of h_M in (6) to yield the Jacobian $J_\star^{(M)}$ of (8). Because the identity factors as $\operatorname{Cov}_\pi(h, s) \cdot M$, the matrix M enters the Jacobian only as a right-multiplier on S and C, which is the technical reason the optimal-M analysis in Section 3 depends only on SM (rather than on a more elaborate function of M).

B.3 Proof of Proposition 2.1 (well-posedness and equilibrium)

We prove the two parts in turn. Lemma B.2 below is part **(i)**; the equilibrium and global-stability claim is derived afterward.

Lemma B.2 (Well-posedness of $\pi_\theta^{(M)}$)**.** *For every symmetric $M \succ 0$, every $\theta \in \mathbb{R}^d$, and every $\alpha \geq 0$, $Z_\theta^{(M)} = \int_{\mathbb{R}^d} \pi(x) \exp(-\alpha\, \theta^\top M s(x))\, dx < \infty$, so $\pi_\theta^{(M)}$ is a valid probability density.*

Proof. This proof largely follows Hu et al. [34, Appendix C.2].

Step 1: Linear bound on the score-tilted exponent. Pick any $x_0 \in \mathbb{R}^d$. By the score-Lipschitz hypothesis in Assumption 1, $\|s(x) - s(x_0)\| \leq L\|x\| + L\|x_0\|$, so $\|s(x)\| \leq L\|x\| + c_0$ with $c_0 \triangleq L\|x_0\| + \|s(x_0)\|$. Setting $C_M \triangleq \alpha\, \|M\|_{\mathrm{op}}\, \|\theta\| < \infty$ (with $\|M\|_{\mathrm{op}}$ as the operator norm of M), Cauchy-Schwarz gives the pointwise upper bound

$$-\alpha\, \theta^\top M s(x) \;\leq\; C_M\, \|s(x)\| \;\leq\; C_M\, (L\|x\| + c_0). \tag{22}$$

Step 2: Integrability of the score-tilted density. Split $Z_\theta^{(M)} = \int_{\|x\|<R} + \int_{\|x\|\geq R}$ and use $\pi(x) = Z_0^{-1} e^{-U(x)}$. The first integral is bounded by $e^{C_M(LR+c_0)} < \infty$ because both π and the exponential of a continuous function are bounded on the compact set $\{\|x\| \leq R\}$. For the tail,

$$\int_{\|x\|\geq R} \pi(x)\, e^{-\alpha\, \theta^\top M s(x)}\, dx \;\leq\; Z_0^{-1} \int_{\|x\|\geq R} e^{-c\|x\|^p + C_M L\|x\| + C_M c_0}\, dx,$$

which is finite for R large enough that $-c\|x\|^p + C_M L\|x\| \leq -\frac{c}{2}\|x\|^p$, since $p > 1$ ensures the polynomial decay dominates the linear growth. Hence $Z_\theta^{(M)} < \infty$. □

The proof of Lemma B.2 depends critically on $p > 1$ in Assumption 1: a target with merely linear tail decay (e.g., $U(x) = c\|x\|$ for $\|x\|$ large) would yield $-c\|x\| + C_M L\|x\| > 0$ once $C_M L > c$, i.e., once $\alpha\|M\|_{\mathrm{op}}\|\theta\|$ exceeds c/L. In that regime, the surrogate $\pi_\theta^{(M)}$ would have an infinite normalizing constant and be undefined. Super-linear growth ensures the surrogate is well-defined for *every* θ and *every* $\alpha \geq 0$, and is what enables the global stability argument below.

Proof of Proposition 2.1. Lemma B.2 establishes part **(i)** (well-posedness). We now prove part **(ii)** (uniqueness and global stability of the equilibrium $\vartheta^\star = (0, \mu)$) in three steps.

Step 1: $\vartheta^\star = (0, \mu)$ is a fixed point. At $\theta = 0$, the surrogate reduces to $\pi_0^{(M)} = \pi$ and Stein's identity gives $S_M(0) = \mathbb{E}_\pi[s(X)] = 0$ (this is the integration-by-parts identity, which holds under the tail

growth in Assumption 1). Hence the first block of h_M in (6) vanishes at $\theta=0$, and the second block vanishes at $\mu_0=\mu\triangleq\mathbb{E}_\pi[f(X)]$.

Step 2: Local exponential stability at $\vartheta^\star$ via the Hurwitz Jacobian. The Jacobian $J_\star^{(M)}$ in (8) is derived block-by-block in Lemma B.3 (Appendix B.4); its $(1,1)$ block $-I_d-\alpha SM$ is similar to the symmetric matrix $-I_d-\alpha\, S^{1/2}MS^{1/2}$ via the conjugation $A\mapsto S^{-1/2}AS^{1/2}$ (direct calculation: $S^{-1/2}(SM)S^{1/2}=S^{1/2}MS^{1/2}$), and $-I_d-\alpha\, S^{1/2}MS^{1/2}\prec 0$ since $S\succ 0$ and $M\succ 0$ imply $S^{1/2}MS^{1/2}\succ 0$; the $(2,2)$ block is $-I_m\prec 0$. Thus all eigenvalues of $J_\star^{(M)}$ have negative real part (i.e., $J_\star^{(M)}$ is Hurwitz), and $\vartheta^\star$ is locally exponentially stable.

Step 3: Global stability via a metric-modified strict Lyapunov function. Define

$$V(\theta)\ \triangleq\ \tfrac{1}{2}\,\theta^\top M\,\theta, \tag{23}$$

which is strictly convex and coercive (since $M\succ 0$), with $V(0)=0$ and $V(\theta)>0$ for $\theta\neq 0$. Its gradient is $\nabla V(\theta)=M\theta$, so along the deterministic θ-iterate ODE

$$\dot\theta_t=h_M^\theta(\theta_t)\triangleq S_M(\theta_t)-\theta_t, \tag{24}$$

the chain rule gives

$$\dot V(\theta)\ =\ \nabla V(\theta)^\top\dot\theta\ =\ (M\theta)^\top\big(S_M(\theta)-\theta\big)\ =\ \theta^\top M\,S_M(\theta)\ -\ \theta^\top M\,\theta. \tag{25}$$

Step 3a: Gradient and Hessian of the log-partition. Let $\phi_M(\theta)\triangleq\log Z_\theta^{(M)}$. Differentiating $\log Z_\theta^{(M)}=\log\!\int\pi(x)e^{-\alpha\theta^\top Ms(x)}dx$ in θ (justified by Assumption 1, identical to the dominated-convergence argument used for Lemma B.1) gives

$$\nabla\phi_M(\theta)\ =\ \frac{\nabla_\theta Z_\theta^{(M)}}{Z_\theta^{(M)}}\ =\ -\alpha M\,\mathbb{E}_{\pi_\theta^{(M)}}[s(X)]\ =\ -\alpha M\,S_M(\theta). \tag{26}$$

The Hessian is then obtained by differentiating (26) once more and applying Lemma B.1 with $h=s$ at general θ (so $\nabla_\theta S_M(\theta)=-\alpha\,\mathrm{Cov}_{\pi_\theta^{(M)}}(s,s)\,M$):

$$\nabla_\theta^2\,\phi_M(\theta)\ =\ -\alpha M\,\nabla_\theta S_M(\theta)\ =\ \alpha^2\,M\,\mathrm{Cov}_{\pi_\theta^{(M)}}(s,s)\,M\ \succ\ 0, \tag{27}$$

where the strict positivity uses $M\succ 0$ and $\mathrm{Cov}_{\pi_\theta^{(M)}}(s,s)\succ 0$ (the score is non-degenerate under the smooth surrogate density). Hence ϕ_M is strictly convex on $\mathbb{R}^d$.

Step 3b: A convexity inequality for the log-partition. At $\theta=0$, the surrogate equals π, so $\phi_M(0)=\log\!\int\pi=0$ and Stein's identity gives $\nabla\phi_M(0)=-\alpha M\,\mathbb{E}_\pi[s]=0$ via (26). A direct (and equivalent) argument that $\log Z_\theta^{(M)}\geq 0$ comes from Jensen's inequality: for any random variable Y, convexity of the exponential gives $\mathbb{E}[e^Y]\geq e^{\mathbb{E}[Y]}$, i.e., $\log\mathbb{E}[e^Y]\geq\mathbb{E}[Y]$. Applied with $Y=-\alpha\theta^\top Ms(X)$ under π,

$$\log Z_\theta^{(M)}\ =\ \log\mathbb{E}_\pi\big[e^{-\alpha\theta^\top Ms(X)}\big]\ \geq\ \mathbb{E}_\pi\big[-\alpha\theta^\top Ms(X)\big]\ =\ -\alpha\theta^\top M\,\mathbb{E}_\pi[s(X)]\ =\ 0, \tag{28}$$

i.e., $Z_\theta^{(M)}\geq 1$ for every θ, with equality if an only if $\theta=0$ (strict by strict convexity). Now apply the gradient inequality (i.e., $f(x)\geq f(y)+\nabla f(y)^\top(x-y)$ for a convex function f, see also Boyd and Vandenberghe [14, Section 3.1.3 (First-order conditions)]) of the strictly convex function ϕ_M at the points 0 and θ:

$$\phi_M(0)\ \geq\ \phi_M(\theta)+\nabla\phi_M(\theta)^\top(0-\theta)\ =\ \phi_M(\theta)+\alpha\,\theta^\top M\,S_M(\theta),$$

where the second equality substitutes (26). Rearranging, and combining with (28),

$$\theta^\top M\,S_M(\theta)\ \leq\ -\alpha^{-1}\,\phi_M(\theta)\ =\ -\alpha^{-1}\log Z_\theta^{(M)}\ \leq\ 0, \tag{29}$$

with both inequalities tight if and only if $\theta=0$.

Step 3c: V is a strict Lyapunov function. Substituting (29) into (25),

$$\dot V(\theta)\ =\ \theta^\top MS_M(\theta)-\theta^\top M\,\theta\ \leq\ -\theta^\top M\,\theta\ <\ 0\quad\text{for all }\theta\neq 0. \tag{30}$$

Step 3d: Global asymptotic stability via LaSalle's invariance principle. Since V is coercive, every sublevel set $\{\theta:V(\theta)\leq c\}$ is compact and forward-invariant under the ODE; in particular, all trajectories are bounded. By Step 3c, $\dot V(\theta)<0$ on $\mathbb{R}^d\setminus\{0\}$, so the only invariant subset of $\{\theta:\dot V(\theta)=0\}$ is $\{0\}$. By LaSalle's invariance principle, $\theta_t\to 0$ as $t\to\infty$ from any initial condition

$\theta_0 \in \mathbb{R}^d$. The μ-iterate ODE $\dot{\mu}_t = F_M(\theta_t) - \mu_t$ is a linear contraction driven by $F_M(\theta_t) \to \mu$, so $\mu_t \to \mu$ along the same trajectory. Hence $\vartheta^\star = (0, \mu)$ is the unique globally asymptotically stable equilibrium. This generalizes Hu et al. [34, Appendix C.5] from the SRMC case $M = I_d$ to arbitrary symmetric $M \succ 0$. □

For SRMC ($M = I_d$), Hu et al. [34] use the augmented potential $F(\theta) = \alpha^{-1} \log Z_\theta + \frac{1}{2}\|\theta\|^2$ rather than the bare quadratic $\frac{1}{2}\|\theta\|^2$. Both functions *are* Lyapunov functions for the deterministic flow: applying our Step 3 argument with $M = I_d$ shows that $\frac{1}{2}\|\theta\|^2$ alone is strictly decreasing along trajectories, so it certifies global asymptotic stability without the $\log Z_\theta$ correction. SRMC's choice of F adds the log-partition term not because it is needed for stability, but because it exhibits an additional structural property: the deterministic dynamics is a literal gradient flow $\dot{\theta}_t = -\nabla F(\theta_t)$, since $\nabla F = -S(\theta) + \theta = -(S(\theta) - \theta) = -h^\theta$.[3] This gradient-flow representation gives $\dot{F} = -\|\dot{\theta}\|^2$, which is convenient for finer rate analysis. Both arguments yield the same global-stability conclusion at $M = I_d$; we work with the simpler $\frac{1}{2}\|\theta\|^2$ because it is the one that generalizes.

For GA-SRMC with general $M \succ 0$ the gradient-flow representation *fails*: with $F_M(\theta) = \alpha^{-1} \log Z_\theta^{(M)} + \frac{1}{2}\theta^\top M\theta$, the gradient identity (26) gives $\nabla F_M(\theta) = -MS_M(\theta) + M\theta = M(\theta - S_M(\theta))$, which equals $-h^\theta(\theta) = -(S_M(\theta) - \theta)$ only when $M = I_d$. The bare quadratic $V(\theta) = \frac{1}{2}\theta^\top M\theta$ in (23) sidesteps this obstruction: even though the dynamics is no longer a Euclidean gradient flow, V is still strictly decreasing along trajectories thanks to the convexity bound (29). At $M = I_d$ this reduces to the bare $\frac{1}{2}\|\theta\|^2$ stand-alone Lyapunov argument; for $M \neq I_d$ the metric-weighted form is essential. (Equivalently, V is a Lyapunov function for the Riemannian gradient flow under the metric M, since $M^{-1}\nabla F_M(\theta) = \theta - S_M(\theta) = -h^\theta$.)

B.4 Derivation of the Jacobian $J_\star^{(M)}$

Lemma B.3 (Jacobian of the mean field h_M at $\vartheta^\star$)**.** *Under Assumptions 1–2, the Jacobian of the mean-field map $h_M : \mathbb{R}^{d+m} \to \mathbb{R}^{d+m}$ defined in* (6) *at the equilibrium $\vartheta^\star = (0, \mu)$ is*

$$J_\star^{(M)} \;=\; \begin{bmatrix} -I_d - \alpha SM & 0 \\ -\alpha CM & -I_m \end{bmatrix}, \tag{31}$$

where $S \triangleq \mathrm{Cov}_\pi(s, s)$ is the Fisher information matrix of π, $C \triangleq \mathrm{Cov}_\pi(f, s)$ is the cross-covariance of the test function with the score, and 0 denotes the $d \times m$ zero block.

Proof. Recall $h_M(\theta, \mu) = (S_M(\theta) - \theta,\, F_M(\theta) - \mu)$ with $S_M(\theta) \triangleq \mathbb{E}_{\pi_\theta^{(M)}}[s(X)]$ and $F_M(\theta) \triangleq \mathbb{E}_{\pi_\theta^{(M)}}[f(X)]$. We compute the four partial-derivative blocks separately.

Step 1: θ-derivative of the θ-iterate. By Lemma B.1 applied at $h = s$ and $\theta = 0$ (so that $\pi_0^{(M)} = \pi$ and $\mathrm{Cov}_{\pi_0^{(M)}} = \mathrm{Cov}_\pi$),

$$\left.\frac{\partial S_M}{\partial \theta}\right|_{\theta=0} \;=\; -\alpha\, \mathrm{Cov}_\pi(s, s)\, M \;=\; -\alpha\, SM. \tag{32}$$

Since the deterministic $-\theta$ term contributes the constant $-I_d$, we obtain $\partial(S_M(\theta) - \theta)/\partial\theta|_{\vartheta^\star} = -I_d - \alpha\, SM$.

Step 2: θ-derivative of the μ-iterate. Analogously, Lemma B.1 at $h = f$ and $\theta = 0$ gives

$$\left.\frac{\partial F_M}{\partial \theta}\right|_{\theta=0} \;=\; -\alpha\, \mathrm{Cov}_\pi(f, s)\, M \;=\; -\alpha\, CM, \tag{33}$$

so $\partial(F_M(\theta) - \mu)/\partial\theta|_{\vartheta^\star} = -\alpha\, CM$.

Step 3: μ-derivative of both blocks. Neither $S_M(\theta)$ nor $F_M(\theta)$ depends on μ, so $\partial S_M/\partial\mu = 0$ and $\partial F_M/\partial\mu = 0$. The deterministic $-\theta$ and $-\mu$ terms contribute 0 and $-I_m$ respectively, giving the second column of (31).

Stacking the blocks from Steps 1–3 into the $(d+m) \times (d+m)$ Jacobian gives (31). □

[3] For notation simplicity, we let $h^\theta \equiv h_I^\theta$ when $M = I_d$ in (24).

The block-triangular structure of $J_\star^{(M)}$ in (31) (zero upper-right block) is the technical reason the (θ,θ)-Lyapunov equation $D_M\Sigma_{\theta\theta}^{(M)} + \Sigma_{\theta\theta}^{(M)} D_M^\top = U_{ss}$ decouples from the μ-iterate analysis. This decoupling is what allows the optimal-M argument in Appendix C.1 to be carried out at the level of the θ-iterate alone, rather than requiring a joint optimization over both blocks of $\Sigma_\vartheta^{(M)}$. The same block-triangular structure is exploited by the residual-decomposition argument in Appendix C.2, which solves the θ-iterate first and substitutes the result into the μ-iterate.

B.5 Proof of Proposition 2.2 (a.s. convergence and CLT for fixed M)

The proof verifies the five SA conditions of Borkar et al. [12, Assumptions A1–A5] and quotes Borkar et al. [12, Theorems 1 and 4] to obtain the conclusion. The structure mirrors Hu et al. [34, Appendix C.8], with the additional matrix M inserted at the appropriate places.

Proof of Proposition 2.2. **Step 1: Step-size conditions (A1, A5b).** Our schedule $\gamma_n = (n+n_0)^{-\rho}$ with $\rho \in (\frac{1}{2}, 1]$ satisfies $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$, matching the Robbins–Monro conditions [12, A1, A5b].

Step 2: Lipschitz/linear growth of H (A2, A4). The mean-field map $H(\vartheta, x) = (s(x) - \theta,\ f(x) - \mu)$ inherits Lipschitz and linear-growth bounds in x directly from the Lipschitz score in Assumption 1 and the integrability of f. These bounds are uniform in ϑ and so satisfy [12, A2, A4]; the M-dependence enters only through the surrogate $\pi_\theta^{(M)}$ used by the kernel, not through H itself.

Step 3: ODE@∞ stability (A3). Symmetry of M gives $\theta^\top M s(x) = (M\theta)^\top s(x)$, so the surrogate satisfies $\pi_\theta^{(M)} = \pi_{M\theta}^{(I_d)}$ and $S_M(\theta) = S_{I_d}(M\theta)$. Plugging $\eta = M\theta_0$ into the SRMC A3 verification of Hu et al. [34, Appendix C.8] yields the Hurwitz limit $r^{-1} h_M(r\vartheta_0) \to h_M^\infty(\vartheta_0)$ for general $M \succ 0$ under Assumption 1, and thus Borkar et al. [12, A3] holds.

Step 4: Differentiable mean field and Hurwitz Jacobian (A5a). By Appendix B.3, $\vartheta^\star = (0, \mu)$ is the unique globally asymptotically stable equilibrium and the Jacobian $\frac{\mathbf{1}_{\{\rho=1\}}}{2} I + J_\star^{(M)}$ in (8) is Hurwitz. Uniform Lipschitzness of $\theta \mapsto S_M(\theta)$ on compact sets follows from Lemma B.1 together with continuity of $\mathrm{Cov}_{\pi_\theta^{(M)}}(s, s)$ in θ, mirroring Hu et al. [34, Appendix C.7].

Step 5: Base kernel regularity (DV3 and Lipschitzness). Assumption 2 provides exactly the uniform drift and kernel Lipschitzness needed by the controlled-Markov-noise version of the SA CLT [12].

Step 6: Conclusion via the SA CLT. By Borkar et al. [12, Theorems 1 and 4] applied to the joint $(d+m)$-dimensional iterate ϑ_n, $\vartheta_n \xrightarrow{a.s.} \vartheta^\star$ and

$$\gamma_n^{-1/2}(\vartheta_n - \vartheta^\star) \xrightarrow{dist.} \mathcal{N}(0, \Sigma_\vartheta^{(M)}),$$

where $\Sigma_\vartheta^{(M)}$ is the unique solution of

$$\big(\tfrac{\mathbf{1}_{\{\rho=1\}}}{2} I + J_\star^{(M)}\big)\Sigma_\vartheta + \Sigma_\vartheta \big(\tfrac{\mathbf{1}_{\{\rho=1\}}}{2} I + J_\star^{(M)}\big)^\top + U_{HH} = 0,$$

which is exactly (10).

Step 7: Decoupling the θ-iterate. The block-triangular structure of $J_\star^{(M)}$ in (8) (with zero upper-right block) decouples the (θ,θ) block, yielding the Lyapunov equation in (11) with damping operator $D_M = \beta I_d + \alpha S M$.

Step 8: Asymptotic unbiasedness. Almost-sure convergence $\theta_n \to 0$ implies $\pi_{\theta_n}^{(M)} \to \pi$ pointwise (continuity of the surrogate density in θ); the μ-iterate then satisfies $\mu_n \to \mu = \mathbb{E}_\pi[f(X)]$, recovering exact estimation of the target mean. □

C Proofs of Section 3

Summary of Appendix C. This appendix is organized as follows. Appendix C.1 proves Theorem 3.1: the trace-normalized inverse Fisher uniquely minimizes the worst-direction score-history covariance bound in the Loewner order. Appendix C.2 gives the residual decomposition for the μ-iterate that justifies the second equality in (15). Appendix C.3 derives the explicit limiting covariance

for the two special cases referenced in Section 3 of the main body: the independent surrogate sampler and the Gaussian target. Appendix C.4 positions the optimization relative to the original SRMC analysis [34] and the classical zero-variance/control-variate literature [5, 43, 45]. Appendix C.5 proves Theorem 3.2 on the online metric under the assumptions of Appendix B.1.

C.1 Proof of Theorem 3.1

The proof works directly on the damping matrix $D_M = \beta I_d + \alpha SM$ from (11). Although D_M is generically non-symmetric (since $SM \neq MS$ unless S and M commute), it has real positive eigenvalues, and a one-line trace argument suffices to characterize M_{opt}:

Lemma C.1 (Smallest-eigenvalue maximizer under a trace constraint)**.** *For any matrix $A \in \mathbb{R}^{d\times d}$ whose eigenvalues are real and positive with $\sum_{i=1}^{d} \lambda_i(A) = \tau > 0$,*

$$\lambda_{\min}(A) \;\le\; \tau/d,$$

with equality if and only if all eigenvalues coincide. In particular, if A is also diagonalizable, equality holds if and only if $A = (\tau/d) I_d$.

Proof. **(i)** The eigenvalues are positive reals with an average τ/d, and the minimum of a finite tuple is at most its average, with equality if and only if all entries are equal. A diagonalizable $A = P\Lambda P^{-1}$ with $\Lambda = (\tau/d) I_d$ equals $(\tau/d) I_d$, and conversely.

(ii) Since $S \succ 0$, its symmetric square root $S^{1/2}$ is invertible, and $SM = S^{1/2}\,(S^{1/2} M S^{1/2})\,S^{-1/2}$. Thus SM is similar to $B \triangleq S^{1/2} M S^{1/2}$, which is symmetric and positive definite ($x^\top B x = (S^{1/2}x)^\top M (S^{1/2}x) > 0$ for $x \neq 0$). Hence, the eigenvalues of SM are those of B: they are real and positive, and they sum to $\mathrm{tr}(B) = \mathrm{tr}(SM) = \tau$ by the cyclic invariance of the trace. Part (i) gives

$$\lambda_{\min}(SM) \;=\; \lambda_{\min}(B) \;\le\; \frac{1}{d}\sum_{i=1}^{d} \lambda_i(SM) \;=\; \frac{\tau}{d},$$

with equality if and only if every eigenvalue of B equals τ/d. Because B is symmetric and thus orthogonally diagonalizable, this holds if and only if $B = (\tau/d) I_d$, i.e., $M = S^{-1/2} B S^{-1/2} = (\tau/d) S^{-1}$. Conversely, $M = (\tau/d) S^{-1}$ is symmetric positive definite, satisfies $\mathrm{tr}(SM) = \tau$, and gives $SM = (\tau/d) I_d$, so equality holds; the maximizer of $\lambda_{\min}(SM)$ under the trace constraint is therefore unique. □

Proof of Theorem 3.1. The proof has four steps: characterize the spectrum of D_M via a similarity transformation to a symmetric form L_M (Step 1), derive the universal upper bound (13) on $\Sigma_{\theta\theta}^{(M)}$ via the operator-norm contraction of e^{-tL_M} (Step 2), maximize $\lambda_{\min}(D_M)$ under the trace constraint via Lemma C.1 to identify M_{opt} (Step 3), and compute the closed form $\Sigma_{\theta\theta}^{(M_{\mathrm{opt}})} = U_{ss}/(2(\beta + \alpha\tau/d))$ at the optimum (Step 4).

Step 1: Spectrum of D_M. Although D_M is non-symmetric, the similarity transformation $D_M = S^{1/2} L_M S^{-1/2}$ with the symmetric matrix $L_M \triangleq \beta I_d + \alpha\, S^{1/2} M S^{1/2} \succ 0$ shows that D_M is diagonalizable with real positive eigenvalues. In particular, since $\alpha \ge 0$, $\lambda_{\min}(D_M) = \lambda_{\min}(L_M) = \beta + \alpha\lambda_{\min}(S^{1/2} M S^{1/2}) = \beta + \alpha\lambda_{\min}(SM) > 0$, the last identity by Lemma C.1(ii). Linearity of trace and cyclic invariance ($\mathrm{tr}(SM) = \mathrm{tr}(S^{1/2} M S^{1/2}) = \tau$) then give

$$\mathrm{tr}(D_M) \;=\; \mathrm{tr}(\beta I_d) + \alpha\ \mathrm{tr}(SM) \;=\; d\beta + \alpha\ \mathrm{tr}(SM) \;=\; d\beta + \alpha\tau$$

under the trace constraint.

Step 2: Universal upper bound on $\Sigma_{\theta\theta}^{(M)}$ via the symmetric form. Starting from the Lyapunov solution $\Sigma_{\theta\theta}^{(M)} = \int_0^\infty e^{-tD_M} U_{ss} e^{-tD_M^\top} dt$ and substituting the similarity transformation $D_M = S^{1/2} L_M S^{-1/2}$ from Step 1 (so $e^{-tD_M} = S^{1/2} e^{-tL_M} S^{-1/2}$ and $e^{-tD_M^\top} = S^{-1/2} e^{-tL_M} S^{1/2}$ since L_M is symmetric), the integrand factors as $S^{1/2}\big(e^{-tL_M}\widetilde{U}\, e^{-tL_M}\big) S^{1/2}$ with $\widetilde{U} \triangleq S^{-1/2} U_{ss} S^{-1/2}$. Pulling the $S^{\pm 1/2}$ factors outside the integral yields

$$\Sigma_{\theta\theta}^{(M)} = S^{1/2}\,\Gamma_M\, S^{1/2}, \qquad \Gamma_M \triangleq \int_0^\infty e^{-tL_M}\,\widetilde{U}\, e^{-tL_M}\, dt.$$

We bound Γ_M via two ingredients. First, symmetry of L_M gives the operator-norm contraction $\|e^{-tL_M}\|_2 = e^{-t\lambda_{\min}(L_M)} = e^{-t\lambda_{\min}(D_M)}$ (the second equality holds because D_M and L_M are similar). Second, $\widetilde{U}$ is symmetric positive semi-definite, so $\widetilde{U} \preceq \lambda_{\max}(\widetilde{U})\, I_d$ in the Loewner order. We then chain three standard matrix-analysis facts (each cited at the point of use) to bound the integrand pointwise in $t \geq 0$:

(a) *Congruence preserves the Loewner order* [32, Observation 7.1.8]: if $A \preceq B$ and X are any matrices of compatible size, then $X^\top A X \preceq X^\top B X$. Applying this with $A = \widetilde{U}$, $B = \lambda_{\max}(\widetilde{U})\, I_d$, and $X = e^{-tL_M}$ (which is symmetric since L_M is symmetric, so $X^\top = X$) gives

$$e^{-tL_M}\widetilde{U}\, e^{-tL_M} \;\preceq\; \lambda_{\max}(\widetilde{U})\, e^{-tL_M}\, I_d\, e^{-tL_M} \;=\; \lambda_{\max}(\widetilde{U})\, e^{-2tL_M},$$

where the final equality uses $e^{-tL_M} e^{-tL_M} = e^{-2tL_M}$ (a single matrix commutes with itself, and the matrix exponential satisfies $e^A e^A = e^{2A}$).

(b) *Spectral lower bound on L_M as a Loewner inequality*: since L_M is symmetric with $\lambda_{\min}(L_M) = \lambda_{\min}(D_M)$ (Step 1 above), it admits an orthogonal diagonalization $L_M = Q\Lambda Q^\top$ with $\Lambda \succeq \lambda_{\min}(D_M)\, I_d$, hence $L_M \succeq \lambda_{\min}(D_M)\, I_d$ in the Loewner order.

(c) *Operator antitonicity of $A \mapsto e^{-2tA}$ on symmetric matrices*: by the Loewner-Heinz theorem and the spectral mapping theorem [10, Theorem V.1.9 and Corollary V.2.6], the matrix function $A \mapsto e^{-2tA}$ reverses the Loewner order on symmetric matrices for every $t \geq 0$, i.e., $A \succeq B \Rightarrow e^{-2tA} \preceq e^{-2tB}$. Combined with (b), this gives

$$e^{-2tL_M} \;\preceq\; e^{-2t\lambda_{\min}(D_M)\, I_d} \;=\; e^{-2t\lambda_{\min}(D_M)}\, I_d.$$

Equivalently, this can be verified directly by orthogonally diagonalizing $L_M = Q\Lambda Q^\top$ from (b), which gives $e^{-2tL_M} = Q\, \mathrm{diag}(e^{-2t\lambda_i(L_M)})\, Q^\top$ with each diagonal entry bounded above by $e^{-2t\lambda_{\min}(D_M)}$.

Combining (a) and (c) yields the pointwise bound $e^{-tL_M}\widetilde{U}\, e^{-tL_M} \preceq \lambda_{\max}(\widetilde{U})\, e^{-2t\lambda_{\min}(D_M)}\, I_d$. Integrating in t,

$$\Gamma_M \;\preceq\; \lambda_{\max}(\widetilde{U}) \int_0^\infty e^{-2t\lambda_{\min}(D_M)}\, dt \;\cdot\; I_d \;=\; \frac{\lambda_{\max}(\widetilde{U})}{2\lambda_{\min}(D_M)}\, I_d.$$

Conjugating by $S^{1/2}$ and using $\lambda_{\max}(\widetilde{U}) = \lambda_{\max}(S^{-1/2} U_{ss} S^{-1/2}) = \lambda_{\max}(U_{ss} S^{-1})$ (similarity transform shares the spectrum), we recover the universal bound

$$\Sigma_{\theta\theta}^{(M)} \;=\; S^{1/2}\Gamma_M S^{1/2} \;\preceq\; \frac{\lambda_{\max}(U_{ss}S^{-1})}{2\lambda_{\min}(D_M)}\, S^{1/2} I_d S^{1/2} \;=\; \frac{\lambda_{\max}(U_{ss}S^{-1})}{2\lambda_{\min}(D_M)}\, S,$$

which is (13) rewritten with $\lambda_{\min}(D_M)$ in the denominator.

Step 3: Maximizing $\lambda_{\min}(D_M)$ under the trace constraint. The bound (13) has the form $c(M)\, S$ with the scalar $c(M) \triangleq \lambda_{\max}(U_{ss}S^{-1})/(2\lambda_{\min}(D_M))$. Since $S \succ 0$, $c(M)S \preceq c(M')S$ if and only if $c(M) \leq c(M')$, so minimizing the bound in the Loewner order is equivalent to minimizing $c(M)$. This comparison across M is made at a common equilibrium kernel: $K_{0,M}$, and hence U_{ss} is held fixed as M varies. This is automatic when the proposal does not depend on M (then $K_{0,M} = K_\pi$); for a proposal tied to M, such as FisherMALA with $A = dM/\operatorname{tr}(M)$, it means freezing the proposal at a common choice, since otherwise $U_{ss} = U_{ss}(M)$ and the numerator of $c(M)$ vary with M. The optimality statement of Theorem 3.1 is made in this fixed-kernel sense, as is its use at the learned metric in Theorem 3.2. Under it, the numerator $\lambda_{\max}(U_{ss}S^{-1}) = \lambda_{\max}(\widetilde{U}) \geq 0$ does not depend on M. It vanishes only if $U_{ss} = 0$, in which case $\Sigma_{\theta\theta}^{(M)} = 0$ for every M by (11), every admissible M attains the (zero) bound, and uniqueness fails for $d \geq 2$; we therefore assume $U_{ss} \neq 0$. For $U_{ss} \neq 0$, minimizing the bound is equivalent to maximizing $\lambda_{\min}(D_M) = \beta + \alpha\lambda_{\min}(SM)$ (Step 1) over $\{M = M^\top \succ 0 : \operatorname{tr}(SM) = \tau\}$. By Lemma C.1(ii),

$$\lambda_{\min}(D_M) \;=\; \beta + \alpha\lambda_{\min}(SM) \;\leq\; \beta + \alpha\tau/d.$$

If $\alpha > 0$, equality holds if and only if $\lambda_{\min}(SM) = \tau/d$, i.e., if and only if $M = (\tau/d)S^{-1} = M_{\mathrm{opt}}$ (Lemma C.1(ii)); equivalently, all eigenvalues of D_M coincide and $D_M = (\beta + \alpha\tau/d)I_d$. Thus, for $\alpha > 0$ and $U_{ss} \neq 0$, at a common equilibrium kernel, M_{opt} uniquely minimizes the universal upper bound (13) in the Loewner order, proving (14). The condition $\alpha > 0$ is needed for uniqueness: if $\alpha = 0$, then $D_M = \beta I_d$ and the bound is the same for every M.

Step 4: Closed form for $\Sigma_{\theta\theta}^{(M_{\mathrm{opt}})}$. At $M = M_{\mathrm{opt}}$, the equality condition of Step 3 gives $D_{M_{\mathrm{opt}}} = (\beta + \alpha\tau/d) I_d$ directly, so the Lyapunov equation collapses to the scalar identity $2(\beta+\alpha\tau/d)\Sigma_{\theta\theta}^{(M_{\mathrm{opt}})} = U_{ss}$, yielding

$$\Sigma_{\theta\theta}^{(M_{\mathrm{opt}})} \;=\; \frac{U_{ss}}{2(\beta + \alpha\tau/d)}.$$

The second identity in (15) (the μ-iterate) is established below in Appendix C.2. □

The argument in Step 3 reduces the optimal-M problem to a single criterion: among matrices D_M with a fixed trace $d\beta + \alpha\tau$ and a real positive spectrum, the smallest eigenvalue is maximized at the scalar matrix $D_M = (\beta + \alpha\tau/d) I_d$. Geometrically, this is *isotropic damping*: every direction in $\mathbb{R}^d$ is damped at the same rate $\beta + \alpha\tau/d$ in the Lyapunov equation (equivalently, the linearized mean-field ODE contracts every θ-direction at the same rate $1 + \alpha\tau/d$), so no direction is left as a slow-decaying mode that would dominate the worst-direction CLT covariance bound. Because $D_M = \beta I_d + \alpha S M$ depends affinely on M, for $\alpha > 0$ the unique M satisfying (12) that makes D_M scalar is $M_{\mathrm{opt}} = (\tau/d) S^{-1}$. The symmetric form L_M enters Step 2 only as a technical device to obtain the exact decay $\|S^{-1/2} e^{-tD_M} S^{1/2}\|_2 = \|e^{-tL_M}\|_2 = e^{-t\lambda_{\min}(D_M)}$ in the S-weighted norm. For the non-normal D_M, the Euclidean norm $\|e^{-tD_M}\|_2$ itself can exceed $e^{-t\lambda_{\min}(D_M)}$ and is in general only bounded by $\sqrt{\kappa(S)}\, e^{-t\lambda_{\min}(D_M)}$;[4] the congruence argument of Step 2 avoids this factor. The optimization itself is intrinsic to D_M.

In addition, the optimization in Step 3 minimizes only the *prefactor* of the universal upper bound (13), which is itself a conservative Loewner relaxation: replacing e^{-2tL_M} with $e^{-2t\lambda_{\min}(D_M)} I_d$ collapses the full directional decay of the matrix exponential into its slowest mode, and replacing $\widetilde{U}$ with $\lambda_{\max}(\widetilde{U}) I_d$ similarly drops the directional structure of the noise covariance. Maximizing $\lambda_{\min}(D_M)$ therefore tightens only the slowest-mode rate; it does not control the directional dependence of the actual $\Sigma_{\theta\theta}^{(M)}$ in the original θ-basis. Consequently, $\Sigma_{\theta\theta}^{(M_{\mathrm{opt}})}$ may be larger than $\Sigma_{\theta\theta}^{(I_d)}$ along directions where the latter happens to be small, and even its largest eigenvalue can be larger: for $U_{ss} = S = \mathrm{diag}(1, 100)$ (independent sampling, Appendix C.3), $\alpha = \beta = 1$ and $\tau = \mathrm{tr}\, S$, $\lambda_{\max}(\Sigma_{\theta\theta}^{(M_{\mathrm{opt}})}) = 100/103 \approx 0.97$ while $\lambda_{\max}(\Sigma_{\theta\theta}^{(I_d)}) = 100/202 \approx 0.50$, although the bound (13) equals $S/103$ at M_{opt} versus $S/4$ at I_d. Since (13) is equivalent to $\lambda_{\max}(S^{-1/2}\Sigma_{\theta\theta}^{(M)} S^{-1/2}) \le c(M)$, the bound controls the S-normalized worst direction; even that quantity is not minimized by M_{opt} in general (for $U_{ss} = \mathrm{diag}(0.01, 100)$ and the same S, α, β, τ, it equals $1/103 \approx 0.0097$ at M_{opt} and ≈ 0.0050 at I_d). Theorem 3.1 is therefore a statement about the guaranteed worst-direction contraction under the trace normalization, not about Loewner dominance, or even worst-direction dominance, of the covariance itself.

C.2 Residual decomposition for the μ-block

We isolate the part of the μ-iterate CLT covariance $\Sigma_{\mu\mu}^{(M_{\mathrm{opt}})}$ that the score-history mechanism can shrink (the α-dependent terms, which decay at a rate $O(\alpha^{-1})$) from its large-repellence limit, the residual floor $U_{gg}/(2\beta)$.

Step 1: Score-aligned residual decomposition of the test function. Decompose the test function $f : \mathbb{R}^d \to \mathbb{R}^m$ as

$$f(x) \;=\; g(x) + CS^{-1}s(x), \qquad g(x) \;\triangleq\; f(x) - CS^{-1}s(x),$$

where $C = \mathrm{Cov}_\pi(f, s)$ and $S = \mathrm{Cov}_\pi(s, s)$ as in (8). The residual g satisfies

$$\mathrm{Cov}_\pi(g, s) \;=\; \mathrm{Cov}_\pi(f, s) - CS^{-1}\mathrm{Cov}_\pi(s, s) \;=\; C - CS^{-1}S \;=\; 0,$$

i.e., g is the score-orthogonal component of f. Since $\mathbb{E}_\pi[s(X)] = 0$ from Stein's identity, $\mathbb{E}_\pi[g(X)] = \mathbb{E}_\pi[f(X)] = \mu$.

[4] The condition number is defined as $\kappa(S) \triangleq \frac{\lambda_{\max}(S)}{\lambda_{\min}(S)} \ge 1$.

Step 2: Block-Lyapunov system at $M = M_{\mathrm{opt}}$. The joint Lyapunov equation (10) yields the following equations:

$$D_M \Sigma_{\theta\theta}^{(M)} + \Sigma_{\theta\theta}^{(M)} D_M^\top = U_{ss}, \tag{34}$$

$$D_M \Sigma_{\theta\mu}^{(M)} + \beta \Sigma_{\theta\mu}^{(M)} + \alpha \Sigma_{\theta\theta}^{(M)} (CM)^\top = U_{sf}, \tag{35}$$

$$2\beta \Sigma_{\mu\mu}^{(M)} + \alpha \big[CM \Sigma_{\theta\mu}^{(M)} + \Sigma_{\mu\theta}^{(M)} (CM)^\top \big] = U_{ff}. \tag{36}$$

Setting $M = M_{\mathrm{opt}} = (\tau/d) S^{-1}$ scalarizes the operators: $D_M = (\beta + \alpha\tau/d) I_d$, $CM_{\mathrm{opt}} = (\tau/d) CS^{-1}$, and $(CM_{\mathrm{opt}})^\top = (\tau/d) S^{-1} C^\top$.

Step 3: Solving for $\Sigma_{\theta\mu}^{(M_{\mathrm{opt}})}$. Write $\eta \triangleq \alpha\tau/d$ for brevity throughout, so that $D_M = (\beta + \eta) I_d$ and $\alpha\, CM_{\mathrm{opt}} = \eta\, CS^{-1}$ at $M = M_{\mathrm{opt}}$. Equation (34) then yields $\Sigma_{\theta\theta}^{(M_{\mathrm{opt}})} = U_{ss}/(2(\beta + \eta))$, as in Appendix C.1. Substituting into (35) gives the linear equation $(2\beta + \eta)\Sigma_{\theta\mu}^{(M_{\mathrm{opt}})} = U_{sf} - \alpha \Sigma_{\theta\theta}^{(M_{\mathrm{opt}})} (CM_{\mathrm{opt}})^\top$, hence the closed form

$$\Sigma_{\theta\mu}^{(M_{\mathrm{opt}})} \;=\; \frac{U_{sf}}{2\beta + \eta} \;-\; \frac{\eta\, U_{ss}\, S^{-1} C^\top}{2(\beta + \eta)(2\beta + \eta)}.$$

Step 4: Substitution into the μ-iterate. Substituting the closed form for $\Sigma_{\theta\mu}^{(M_{\mathrm{opt}})}$ from Step 3 into the μ-Lyapunov equation (36), the bracketed term in (36) expands to

$$2\beta \Sigma_{\mu\mu}^{(M_{\mathrm{opt}})} \;=\; U_{ff} - \frac{\eta}{2\beta + \eta} \big(CS^{-1} U_{sf} + U_{fs} S^{-1} C^\top \big) + \frac{\eta^2}{(\beta + \eta)(2\beta + \eta)}\, CS^{-1} U_{ss} S^{-1} C^\top. \tag{37}$$

Sending $\eta \to \infty$ in (37) gives the leading-order limit

$$\lim_{\eta \to \infty} 2\beta \Sigma_{\mu\mu}^{(M_{\mathrm{opt}})} \;=\; U_{ff} - \big(CS^{-1} U_{sf} + U_{fs} S^{-1} C^\top \big) + CS^{-1} U_{ss} S^{-1} C^\top. \tag{38}$$

Step 5: Identification of the residual floor. The leading-order limit (38) equals U_{gg}. Substituting (39) below into (38) produces U_{gg} plus eight matrix products that fall into two groups: four involving the cross-covariances U_{sg}, U_{gs} that cancel pairwise and four involving U_{ss} that telescope to zero, leaving U_{gg} as the only surviving term. Throughout this step, we use U_{sg}, U_{gs} for the long-run cross-covariances of s and g under the equilibrium kernel, defined by extending (9) of the main body to the pair (s, g); these are well-defined because $g = f - CS^{-1} s$ inherits the polynomial growth of f. Decompose the long-run noise covariances under the residual identity $f = g + CS^{-1} s$ (Step 1):

$$\begin{aligned} U_{ff} \;&=\; U_{gg} + CS^{-1} U_{sg} + U_{gs} S^{-1} C^\top + CS^{-1} U_{ss} S^{-1} C^\top, \\ U_{sf} \;&=\; U_{sg} + U_{ss} S^{-1} C^\top, \qquad U_{fs} \;=\; U_{gs} + CS^{-1} U_{ss}. \end{aligned} \tag{39}$$

Substituting (39) into the leading-order limit (38) expands $U_{ff} - (CS^{-1} U_{sf} + U_{fs} S^{-1} C^\top) + CS^{-1} U_{ss} S^{-1} C^\top$ into U_{gg} plus eight matrix products in $CS^{-1}, U_{sg}, U_{gs}, U_{ss}, S^{-1} C^\top$. Of these, the four involving U_{sg} or U_{gs} cancel pairwise (each $+CS^{-1} U_{sg}$ from U_{ff} matches a $-CS^{-1} U_{sg}$ from $-CS^{-1} U_{sf}$, and analogously for U_{gs}), and the remaining four involving U_{ss} telescope to zero ($+CS^{-1} U_{ss} S^{-1} C^\top$ from U_{ff} minus $CS^{-1} U_{ss} S^{-1} C^\top$ from $-CS^{-1} U_{sf}$ minus $CS^{-1} U_{ss} S^{-1} C^\top$ from $-U_{fs} S^{-1} C^\top$ plus $CS^{-1} U_{ss} S^{-1} C^\top$ from the third term, summing to zero). What remains is U_{gg}:

$$\lim_{\eta \to \infty} 2\beta \Sigma_{\mu\mu}^{(M_{\mathrm{opt}})} \;=\; U_{gg}.$$

For finite η, substituting (39) into (37) instead and collecting terms with $1 - \frac{\eta}{2\beta + \eta} = \frac{2\beta}{2\beta + \eta}$ and $1 - \frac{2\eta}{2\beta + \eta} + \frac{\eta^2}{(\beta + \eta)(2\beta + \eta)} = \frac{\beta}{\beta + \eta}$ gives the exact identity

$$\Sigma_{\mu\mu}^{(M_{\mathrm{opt}})} \;=\; \frac{U_{gg}}{2\beta} + \frac{CS^{-1} U_{sg} + U_{gs} S^{-1} C^\top}{2\beta + \eta} + \frac{CS^{-1} U_{ss} S^{-1} C^\top}{2(\beta + \eta)}, \tag{40}$$

and hence

$$2\beta \Sigma_{\mu\mu}^{(M_{\mathrm{opt}})} \;=\; U_{gg} + O(\eta^{-1}), \tag{41}$$

i.e., $U_{gg}/(2\beta)$ is the large-repellence limit of $\Sigma_{\mu\mu}^{(M_{\mathrm{opt}})}$ and the remaining α-dependent part is $O(\alpha^{-1})$ as $\alpha \to \infty$. Dividing (41) by 2β produces the second identity in (15). The last term of (40) is positive semidefinite and Loewner nonincreasing in α, but the middle (cross) term need not be positive

semidefinite: $\mathrm{Cov}_\pi(g,s)=0$ does not imply $U_{gs}=0$ because the long-run covariance also collects lagged cross-covariances along the chain. Hence $U_{gg}/(2\beta)$ is a lower bound for $\Sigma_{\mu\mu}^{(M_{\mathrm{opt}})}$, and $\Sigma_{\mu\mu}^{(M_{\mathrm{opt}})}$ is Loewner nonincreasing in α whenever $U_{gs}=0$, e.g., under independent sampling (Appendix C.3); for a general kernel, we claim only the limit and the $O(\alpha^{-1})$ rate. This is the GA-SRMC analog of the SRMC residual term in Hu et al. [34, Section 3, Eq. (15), and Appendix C.10].

The leading-order term $U_{gg}/(2\beta)$ in (41) does not contain α because the score-history mechanism only acts on the score-aligned component $CS^{-1}s$ of the test function, not on the residual g. Algebraically, this manifests as the cancelation in Step 5: substituting the expressions in (39) into (38) produces U_{gg} plus eight cross terms; the four involving U_{sg} or U_{gs} cancel pairwise, and the four involving U_{ss} sum to zero, leaving precisely U_{gg}. Because g is constructed to have $\mathrm{Cov}_\pi(g,s)=0$, the surrogate target $\pi_\theta^{(M)}$ at $\theta\neq 0$ does not bias $\mathbb{E}[g(X)]$ at first order, and the residual term $U_{gg}/(2\beta)$ therefore remains however large α becomes (for finite α and $U_{gs}\neq 0$, the cross term in (40) can move $\Sigma_{\mu\mu}^{(M_{\mathrm{opt}})}$ in either direction). Reducing this limit would require a different mechanism, e.g., a non-score-based history average, which lies outside the class analyzed here (Section 2 retains $s(x)$ as the inner statistic); see also Appendix D.1.

The decomposition $f=g+CS^{-1}s$ is the orthogonal projection of f onto the linear span of the score s (in the inner product $\mathrm{Cov}_\pi(\cdot,\cdot)$), with $CS^{-1}s$ as the score-aligned component and g as the score-orthogonal residual. The repulsion mechanism in GA-SRMC controls θ_n (a running average of s), so it can directly dampen the score-aligned variance of f, but leaves the residual g unchanged; this is why $\Sigma_{\mu\mu}^{(M_{\mathrm{opt}})}$ shrinks at a rate of $O(\alpha^{-1})$ on its score-aligned component and saturates at the floor $U_{gg}/(2\beta)$ on its score-orthogonal component (exactly so under independent sampling, where $U_{gs}=0$; in general, up to the cross term in (40), which also vanishes as $\alpha\to\infty$).

C.3 Special cases: independent surrogate sampler and Gaussian target

Theorem 3.1 is most informative on two special cases that the original SRMC analysis already isolates [34]: **(a)** the *independent surrogate sampler*, where the base kernel $K_{\theta,M}$ draws X_{n+1} independently from $\pi_\theta^{(M)}$ at each step, and **(b)** the *Gaussian target*, where $\pi=\mathcal{N}(\mu_\pi,V)$ and the test function are the mean estimator $f(x)=x$. Each isolates a complementary aspect of the variance-reduction story: **(a)** cleanly separates the M-independent residual floor $U_{gg}/(2\beta)$ from the M-dependent term that GA-SRMC shrinks at rate $O(\alpha^{-1})$, while **(b)** is the boundary case in which the score-aligned subspace contains the entire test function ($g\equiv\mu_\pi$, $U_{gg}=0$) and the limiting covariance shrinks to zero with no floor. In both cases, we give the explicit limiting covariances for both GA-SRMC ($M=M_{\mathrm{opt}}$) and SRMC ($M=I_d$) and compare them side by side; the comparison exposes the algebraic mechanism behind GA-SRMC's improvement along the slow eigendirections of S: scalar (uniform) shrinkage at rate $\beta+\alpha\tau/d$ for GA-SRMC versus matrix (direction-dependent) shrinkage at rate $\beta+\alpha\lambda_i(S)$ for SRMC, with a gap along the slowest eigendirection that tends to the mean-to-minimum eigenvalue ratio $\bar\lambda(S)/\lambda_{\min}(S)\geq 1$ as $\alpha\to\infty$, where $\bar\lambda(S)\triangleq\mathrm{tr}(S)/d$; along the fast eigendirections ($\lambda_i(S)>\bar\lambda(S)$), SRMC damps more strongly. In these comparisons, we take $\tau=\mathrm{tr}(S)$, the value of $\mathrm{tr}(SM)$ at the SRMC choice $M=I_d$, so that both metrics satisfy (12) with the same τ.

C.3.1 Independent surrogate sampling

When the base kernel draws independent samples from $\pi_\theta^{(M)}$, the long-run covariance of the noise $H(\vartheta^\star,\cdot)$ at equilibrium reduces to the per-sample covariance under π:

$$U_{ss}=\mathrm{Cov}_\pi(s,s)=S,\qquad U_{sf}=\mathrm{Cov}_\pi(s,f)=C^\top,\qquad U_{ff}=\mathrm{Cov}_\pi(f,f). \tag{42}$$

We derive the closed forms for $\Sigma_{\theta\theta}^{(M_{\mathrm{opt}})}$ and $\Sigma_{\mu\mu}^{(M_{\mathrm{opt}})}$ in turn.

Step 1: θ - iterate covariance and compare with SRMC. Substituting $U_{ss}=S$ into the closed form from Appendix C.1 (writing $\eta\triangleq\alpha\tau/d$),

$$\Sigma_{\theta\theta}^{(M_{\mathrm{opt}})} \;=\; \frac{S}{2(\beta+\eta)}\qquad(\text{GA-SRMC}). \tag{43}$$

For comparison, under SRMC's $M = I_d$ the symmetric Lyapunov equation $(\beta I_d + \alpha S)\Sigma_{\theta\theta}^{(I_d)} + \Sigma_{\theta\theta}^{(I_d)}(\beta I_d + \alpha S) = S$ has the closed form

$$\Sigma_{\theta\theta}^{(I_d)} \;=\; S\left(2(\beta I_d + \alpha S)\right)^{-1} \qquad \text{(SRMC)}, \tag{44}$$

with the inverse computed in the eigenbasis of S. The two forms differ structurally in their denominators: GA-SRMC has the *scalar* $2(\beta + \eta)$ that shrinks every direction at the same rate, while SRMC has the *matrix* $2(\beta I_d + \alpha S)$ whose eigenvalues range from $2(\beta + \alpha\lambda_{\min}(S))$ in the slow direction to $2(\beta + \alpha\lambda_{\max}(S))$ in the fast direction. On the worst-case (slow) direction, the prefactor ratio at high repellence is

$$\frac{\beta + \alpha\tau/d}{\beta + \alpha\lambda_{\min}(S)} \;\xrightarrow[\alpha\to\infty]{}\; \frac{\bar{\lambda}(S)}{\lambda_{\min}(S)} \;\geq\; 1,$$

the mean-to-minimum eigenvalue ratio of S; the same ratio compares the largest eigenvalues of the S-normalized covariances $S^{-1/2}\Sigma_{\theta\theta}^{(M)}S^{-1/2}$, the quantity bounded in (13). In the Euclidean θ-basis, GA-SRMC's θ-variance is smaller along every eigendirection of S with $\lambda_i(S) < \bar{\lambda}(S)$ and larger along those with $\lambda_i(S) > \bar{\lambda}(S)$, so its largest eigenvalue can be the larger one (Appendix C.1). For example, if $d = 10$ and the eigenvalues of S are log-spaced between $\lambda_{\min}(S)$ and $100\,\lambda_{\min}(S)$ (so $\kappa(S) = 100$), the ratio is ≈ 25.

Step 2: μ-iterate covariance. Substituting (42) into the substituted Lyapunov equation (37) from Step 4 of Appendix C.2 gives

$$\begin{aligned} 2\beta\,\Sigma_{\mu\mu}^{(M_{\text{opt}})} &= U_{ff} - \frac{\eta}{2\beta+\eta}\left(CS^{-1}C^\top + CS^{-1}C^\top\right) + \frac{\eta^2}{(\beta+\eta)(2\beta+\eta)}\,CS^{-1}C^\top \\ &= U_{ff} - CS^{-1}C^\top \cdot \frac{2\eta(\beta+\eta) - \eta^2}{(\beta+\eta)(2\beta+\eta)} \\ &= U_{ff} - CS^{-1}C^\top \cdot \frac{\eta(2\beta+\eta)}{(\beta+\eta)(2\beta+\eta)} \;=\; U_{ff} - \frac{\eta\,CS^{-1}C^\top}{\beta+\eta}, \end{aligned} \tag{45}$$

where the second equality uses $CS^{-1}U_{ss}S^{-1}C^\top = CS^{-1}\cdot S\cdot S^{-1}C^\top = CS^{-1}C^\top$ at $U_{ss} = S$.

Step 3: Decomposition into residual floor plus M-dependent term. Adding and subtracting $CS^{-1}C^\top/(2\beta)$ on the right-hand side of (45) divided by 2β produces

$$\Sigma_{\mu\mu}^{(M_{\text{opt}})} \;=\; \underbrace{\frac{U_{ff} - CS^{-1}C^\top}{2\beta}}_{\text{residual floor} = U_{gg}/(2\beta)} \;+\; \underbrace{\frac{CS^{-1}C^\top}{2(\beta + \alpha\tau/d)}}_{M\text{-dependent},\ O(\alpha^{-1})} \qquad \text{(GA-SRMC)}. \tag{46}$$

The first term is the residual floor $U_{gg}/(2\beta)$ promised by Theorem 3.1 (using the residual identity $U_{ff} - CS^{-1}C^\top = \mathrm{Cov}_\pi(g,g) = U_{gg}$ under independent sampling), and the second term decays at a rate $O(\alpha^{-1})$ as $\alpha\to\infty$; since $U_{gs} = \mathrm{Cov}_\pi(g,s) = 0$ here, (46) is (40) with a vanishing cross term, so the floor is a genuine lower bound and $\Sigma_{\mu\mu}^{(M_{\text{opt}})}$ is Loewner nonincreasing in α. The first term does not depend on M. For a general $M = M^\top \succ 0$ satisfying (12), the block equations (34)–(36) with $U_{ss} = S$ and $U_{sf} = C^\top$ are solved by $\Sigma_{\theta\theta}^{(M)} = \frac{1}{2}S^{1/2}L_M^{-1}S^{1/2}$ and $\Sigma_{\theta\mu}^{(M)} = \frac{1}{2}S^{1/2}L_M^{-1}S^{-1/2}C^\top$, with L_M from Step 1 of Appendix C.1, so that

$$\Sigma_{\mu\mu}^{(M)} = \frac{U_{gg}}{2\beta} + \frac{1}{2}\,CS^{-1/2}L_M^{-1}S^{-1/2}C^\top, \quad \frac{1}{2}\,CS^{-1/2}L_M^{-1}S^{-1/2}C^\top \preceq \frac{CS^{-1}C^\top}{2(\beta + \alpha\lambda_{\min}(SM))}.$$

For a scalar test function ($m = 1$), equality holds when $S^{-1/2}C^\top$ is an eigenvector of L_M for its smallest eigenvalue. Hence, for $\alpha > 0$, by Lemma C.1(ii), M_{opt} uniquely minimizes the *worst case* of the M-dependent term over scalar test functions with a given nonzero score-aligned variance $\mathrm{Var}_\pi(CS^{-1}s) = CS^{-1}C^\top$, in the same worst-direction sense as Theorem 3.1. It need not minimize this term for a particular f: for $S = I_2$, $f = s_1$, $\alpha = \beta = 1$ and $\tau = 2$, the term equals $1/(2(1+a))$ at $M = \mathrm{diag}(a, 2-a)$, which is smaller at $a = 1.9$ than at $M_{\text{opt}} = I_2$ ($a = 1$). The SRMC counterpart at $M = I_d$ is (47) below (cf. the residual term in Hu et al. [34, Section 3, Eq. (15), and Appendix C.10]).

Step 4: SRMC's μ-iterate counterpart and worst-direction comparison. Carrying the same algebra at $M = I_d$ (using the identity $\alpha(\beta I_d + \alpha S)^{-1} = S^{-1} - \beta S^{-1}(\beta I_d + \alpha S)^{-1}$ to extract the

residual floor) gives

$$\Sigma_{\mu\mu}^{(I_d)} \;=\; \underbrace{\frac{U_{ff} - CS^{-1}C^\top}{2\beta}}_{\text{residual floor} = U_{gg}/(2\beta)} \;+\; \underbrace{\frac{1}{2}\, CS^{-1}(\beta I_d + \alpha S)^{-1}C^\top}_{M\text{-dependent, } O(\alpha^{-1})} \qquad \text{(SRMC)}. \tag{47}$$

Side-by-side with the GA-SRMC closed form (46), three structural points stand out.

- *Same residual floor.* Both methods leave the identical α-independent floor $U_{gg}/(2\beta) = (U_{ff} - CS^{-1}C^\top)/(2\beta)$. The score-history mechanism, regardless of M, damps only the score-aligned component $CS^{-1}s$ of f and leaves the score-orthogonal residual g untouched.
- *Different M-dependent term.* GA-SRMC's $CS^{-1}C^\top/(2(\beta+\alpha\tau/d))$ has a *scalar* denominator that shrinks every direction at the same rate $\beta+\alpha\tau/d$; SRMC's $\frac{1}{2}\, CS^{-1}(\beta I_d+\alpha S)^{-1}C^\top$ has a *matrix* denominator whose worst-direction shrinkage rate is only $\beta + \alpha\lambda_{\min}(S)$.
- *Worst-direction gain.* When the score-aligned component of f lies along the slowest eigendirection of S (i.e., $S^{-1/2}C^\top$ is an eigenvector of S for $\lambda_{\min}(S)$), GA-SRMC's M-dependent term is smaller than SRMC's by the same factor $(\beta + \alpha\tau/d)/(\beta + \alpha\lambda_{\min}(S)) \to \bar{\lambda}(S)/\lambda_{\min}(S)$ as in the θ-iterate (Step 1); along the fast eigendirections ($\lambda_i(S) > \bar{\lambda}(S)$), SRMC's term is the smaller one. This idealized comparison (independent sampling, the exact M_{opt}, and the SA estimate μ_n) is qualitative intuition for why a uniform damping rate protects the slowest score direction; it is not a prediction for the experiments of Section 4, whose kernels and estimators differ.

The independent surrogate sampler is a conceptual idealization that neatly isolates the *variance reduction* arising from GA-SRMC's surrogate target, separating it from the asymptotic-variance effects of the underlying base kernel. At $\alpha = 0$ ($\eta = 0$), (46) reduces to $\Sigma_{\mu\mu}^{(M)} = U_{ff}/(2\beta)$, which coincides with the asymptotic covariance of the empirical mean for independent samples from the original target π (i.e., without any GA-SRMC tilt), as expected, because the score-history mechanism is deactivated in this case. At $\alpha \to \infty$ ($\eta \to \infty$), the M-dependent term vanishes and $\Sigma_{\mu\mu}^{(M_{\text{opt}})} \to U_{gg}/(2\beta)$: the residual floor that no M can shrink because g is score-orthogonal. Both limits are consistent with the general kernel result (41) of Appendix C.2.

C.3.2 Gaussian target with mean estimation

For $\pi = \mathcal{N}(\mu_\pi, V)$ the score is $s(x) = -V^{-1}(x - \mu_\pi)$ and the Fisher information is $S = \mathrm{Cov}_\pi(s, s) = V^{-1}$, so the optimal metric becomes $M_{\text{opt}} = (\tau/d)V$. Direct calculation gives

$$M_{\text{opt}}\, s(x) \;=\; \frac{\tau}{d} V\big[-V^{-1}(x - \mu_\pi)\big] \;=\; -\frac{\tau}{d}\,(x - \mu_\pi), \tag{48}$$

which is the *whitened* state, free of the curvature scale V^{-1}. This is exactly why the surrogate exponent $-\alpha\theta^\top M_{\text{opt}} s(x) = (\alpha\tau/d)\theta^\top(x - \mu_\pi)$ is balanced across coordinates, while the SRMC choice $M = I_d$ yields $-\alpha\theta^\top s(x) = \alpha\theta^\top V^{-1}(x - \mu_\pi)$, which overweighs small-variance directions.

Step 1: Linear relation between θ- and μ-iterates. For the test function $f(x) = x$, the algebraic identity $x - \mu_\pi = -Vs(x)$ implies that the empirical mean iterate μ_n and the score-history iterate θ_n differ only by the fixed linear map V, up to an asymptotically negligible initialization term:

$$\mu_n - \mu_\pi \;=\; -V\,\theta_n + o_{a.s.}(1),$$

since $-V(s(X_n) - \theta_n) = (X_n - \mu_\pi) + V\theta_n$ has mean zero under the surrogate at the limit $\theta = 0$.

Remark 3 (Why the Gaussian case collapses the joint (θ, μ) CLT to a single block). On a Gaussian target with $f(x) = x$, the test function f and the score s are linearly related by $x - \mu_\pi = -Vs(x)$, so the residual $g(x) = \mu_\pi$ in (55) is constant. This is a degenerate case of the score-aligned decomposition: f has *no* score-orthogonal component, so the residual floor $U_{gg}/(2\beta)$ in (46) is identically zero and the μ-block CLT reduces to a deterministic linear map of the θ-block CLT. This is what makes the Gaussian case in Section 4 so clean: every term in $\Sigma_x^{(M)}$ is captured by the θ-block analysis, with no asymptotic floor to obscure the α-dependence.

Step 2: μ-block CLT for $f(x) = x$. Hence $\Sigma_x^{(M)} = V\,\Sigma_{\theta\theta}^{(M)}\,V$, and substituting the closed form from (15) gives

$$\Sigma_x^{(M_{\text{opt}})} \;=\; \frac{V\,U_{ss}\,V}{2(\beta + \alpha\tau/d)}. \tag{49}$$

Step 3: Comparison with SRMC, with and without independent-sampling specialization.

General Gaussian target (any base kernel). Without specializing the base kernel, GA-SRMC's closed form $\Sigma_x^{(M_{\mathrm{opt}})} = VU_{ss}V/(2(\beta+\alpha\tau/d))$ from (49) compares to the SRMC analog

$$\Sigma_x^{(I_d)} \;=\; V\,\Sigma_{\theta\theta}^{(I_d)}\,V, \quad \text{where} \quad (\beta I_d + \alpha V^{-1})\Sigma_{\theta\theta}^{(I_d)} + \Sigma_{\theta\theta}^{(I_d)}(\beta I_d + \alpha V^{-1}) = U_{ss}. \tag{50}$$

The structural difference mirrors part (a) Step 1: GA-SRMC's denominator $2(\beta+\alpha\tau/d)$ is a *scalar* that contracts the entire matrix $VU_{ss}V$ uniformly, while SRMC's $\Sigma_{\theta\theta}^{(I_d)}$ inherits a direction-dependent damping matrix $\beta I_d + \alpha V^{-1}$ whose worst-direction eigenvalue (along the broadest axis of V, i.e., $\sigma_{\max}^2$) is only $\beta + \alpha/\sigma_{\max}^2$. Equivalently, in the eigenbasis Q of V ($V = Q\,\mathrm{diag}(\sigma_i^2)Q^\top$, $\widetilde{U}_{ss} \triangleq Q^\top U_{ss} Q$):

$$\big(Q^\top \Sigma_x^{(M_{\mathrm{opt}})} Q\big)_{ij} \;=\; \frac{\sigma_i^2\sigma_j^2\,(\widetilde{U}_{ss})_{ij}}{2(\beta+\alpha\tau/d)}, \qquad \big(Q^\top \Sigma_x^{(I_d)} Q\big)_{ij} \;=\; \frac{\sigma_i^2\sigma_j^2\,(\widetilde{U}_{ss})_{ij}}{2\beta + \alpha/\sigma_i^2 + \alpha/\sigma_j^2}.$$

So, in the worst-case direction ($i=j=\arg\max_k \sigma_k$), the GA-SRMC-to-SRMC variance ratio tends to $(2\beta + 2\alpha/\sigma_{\max}^2)/(2(\beta+\alpha\tau/d)) \to \lambda_{\min}(S)/\bar{\lambda}(S) = (\bar{\lambda}(S)/\lambda_{\min}(S))^{-1} \le 1$ as $\alpha\to\infty$ (with $S = V^{-1}$): *regardless of the base kernel*, GA-SRMC's worst-direction Σ_x-variance is tighter than SRMC's by exactly the mean-to-minimum eigenvalue ratio of S. The independent-sampling case below is one concrete instance of this general phenomenon.

Independent-sampling specialization. Under independent surrogate sampling, $U_{ss} = S = V^{-1}$ and (49) simplify to

$$\Sigma_x^{(M_{\mathrm{opt}})} \;=\; \frac{V\,V^{-1}\,V}{2(\beta+\alpha\tau/d)} \;=\; \frac{V}{2(\beta+\alpha\tau/d)} \qquad \text{(GA-SRMC)}, \tag{51}$$

which shrinks every coordinate at a *uniform* rate $\beta + \alpha\tau/d$. At $\rho = 1$ ($\beta = 1/2$) this becomes $V/(1+2\alpha\tau/d)$, recovering the formula stated in Section 3. Plugging $\widetilde{U}_{ss} = \mathrm{diag}(1/\sigma_i^2)$ into the general SRMC form above gives the coordinate-wise closed form

$$\Sigma_x^{(I_d)} \;=\; \mathrm{diag}\!\left(\frac{\sigma_i^2}{2\beta + 2\alpha/\sigma_i^2}\right) \qquad \text{(SRMC)}, \tag{52}$$

whose shrinkage rate $\beta + \alpha/\sigma_i^2$ is direction-dependent: narrow (σ_i small) coordinates over-contract while broad (σ_i large) coordinates under-contract. The worst-direction prefactor ratio is the same $\bar{\lambda}(S)/\lambda_{\min}(S) = (1/d)\sum_j (\sigma_{\max}/\sigma_j)^2 \approx 25$ on the $d=10$ Gaussian as in Appendix C.3.1, confirming the general bound above with an explicit constant.

Step 4: Vanishing residual floor. We compute $g(x) = f(x) - CS^{-1}s(x)$ for $f(x) = x$ using three substitutions. First, using $s(x) = -V^{-1}(x-\mu_\pi)$ and Stein's identity for the centered test function $f(x) - \mu_\pi$,

$$\begin{aligned} C \;&=\; \mathrm{Cov}_\pi(x,s) \;=\; \mathbb{E}_\pi\big[(x-\mu_\pi)\,s(x)^\top\big] \\ &=\; -\mathbb{E}_\pi\big[(x-\mu_\pi)(x-\mu_\pi)^\top\big]\,V^{-1} \;=\; -V\,V^{-1} \;=\; -I_d. \end{aligned} \tag{53}$$

Second, with $S = V^{-1}$, we have $CS^{-1} = (-I_d)V = -V$, so

$$CS^{-1}s(x) \;=\; (-V)\big[-V^{-1}(x-\mu_\pi)\big] \;=\; x - \mu_\pi. \tag{54}$$

Third, the residual is a constant function

$$g(x) \;=\; f(x) - CS^{-1}s(x) \;=\; x - (x-\mu_\pi) \;=\; \mu_\pi. \tag{55}$$

Hence, $\mathrm{Cov}_\pi(g,g) = 0$, $U_{gg} = 0$, and the residual floor in (46) vanish. The entire limiting covariance shrinks at the $O(\alpha^{-1})$ rate (49), with no α-independent floor.

For the SRMC comparison ($M = I_d$) under the same Gaussian target, $D_M = \beta I_d + \alpha V^{-1}$ is diagonal in the eigenbasis of V, and the coordinatewise covariance is $\sigma_i^2/(2\beta + 2\alpha/\sigma_i^2)$. This expression spreads across orders of magnitude with σ_i and is dominated by the smallest variance directions, formalizing the *geometry-blind failure mode* described in Section 1. The empirical consequence is consistent with the per-coordinate behavior reported in Section 4: GA-SRMC under M_{opt} contracts every coordinate uniformly, while SRMC under $M = I_d$ over-contracts the small-σ_i direction and under-contracts the large-σ_i direction. Equalizing contraction across coordinates, not coordinatewise dominance, is the optimality criterion induced by the trace constraint.

C.4 Position relative to SRMC and to classical control-variate ideas

The optimization in Theorem 3.1 sits at the intersection of two existing lines of work: the *SRMC line* that introduces the score-history surrogate target [34] and the *zero-variance/control-variate line* that formalizes the residual decomposition [5, 43, 45]. Each line, taken alone, stops short of identifying an optimal metric on the score space; combining them is what yields the closed forms in (15), as the five paragraphs below make precise.

Relationship to the original SRMC analysis. Theorem 3.1 delivers a strictly stronger conclusion than the original SRMC analysis: the closed form $\Sigma_{\theta\theta}^{(M_{\mathrm{opt}})} = U_{ss}/(2(\beta + \alpha\tau/d))$ implies that the *entire matrix* is nonincreasing in α in the Loewner order (strictly decreasing, with a positive definite difference, if $U_{ss} \succ 0$; for a singular $U_{ss} \succeq 0$, the decrease is strict exactly along the directions v with $v^\top U_{ss} v > 0$), while Hu et al. [34, Proposition 3.4] can only establish the monotone decrease of its Frobenius norm. Concretely, SRMC corresponds to the special case $M = I_d$ of the GA-SRMC family analyzed here, and Hu et al. [34, Theorem 3.3] establishes a fixed-M CLT for $M = I_d$, whose conclusion our Proposition 2.2 yields at $M = I_d$ (under the hypotheses of Proposition 2.2, which differ from those of Hu et al. [34]), with $D_{I_d} = \beta I_d + \alpha S$. At $M = I_d$, the operator $\beta I_d + \alpha S$ is non-scalar whenever S is non-isotropic, and a Loewner ordering in α over the entire matrix need not hold: for $S = \mathrm{diag}(1, 4)$, $U_{ss} = \left[\begin{smallmatrix} 1 & 0.9 \\ 0.9 & 1 \end{smallmatrix}\right] \succ 0$ and $\beta = 1$, the difference $\Sigma_{\theta\theta}^{(I_d)}|_{\alpha=0} - \Sigma_{\theta\theta}^{(I_d)}|_{\alpha=1}$ is indefinite, so a scalar summary such as the Frobenius norm is used instead. The reason GA-SRMC strengthens this is structural: under M_{opt}, the damping operator becomes scalar, $D_{M_{\mathrm{opt}}} = (\beta + \alpha\tau/d) I_d$, so the Lyapunov equation (11) collapses to a scalar identity in $\Sigma_{\theta\theta}$.

The μ-iterate: closed form vs. coupled Sylvester equations. The same scalarization phenomenon explains why $\Sigma_{\mu\mu}^{(M_{\mathrm{opt}})}$ admits an explicit closed form in (15). In contrast, Hu et al. [34, Appendix C.9] reduces the joint (θ, μ) CLT under $M = I_d$ to coupled Lyapunov/Sylvester equations for $(\Sigma_{\theta\theta}, \Sigma_{\theta\mu}, \Sigma_{\mu\mu})$. For any fixed M, the system (34)–(36) is block-triangular and can be solved in sequence: after the Lyapunov equation (34) for $\Sigma_{\theta\theta}^{(M)}$, (35) gives

$$\Sigma_{\theta\mu}^{(M)} \;=\; (D_M + \beta I_d)^{-1} \big[U_{sf} - \alpha\, \Sigma_{\theta\theta}^{(M)} (CM)^\top \big]$$

($D_M + \beta I_d$ has eigenvalues $2\beta + \alpha\lambda_i(SM) > 0$), and (36) then gives $\Sigma_{\mu\mu}^{(M)}$ explicitly. For general M, $\Sigma_{\theta\theta}^{(M)}$ solves a Lyapunov equation, e.g., $\mathrm{vec}\, \Sigma_{\theta\theta}^{(M)} = (I_d \otimes D_M + D_M \otimes I_d)^{-1} \mathrm{vec}\, U_{ss}$ (invertible because every eigenvalue sum of D_M is positive), whose solution couples all directions. At M_{opt}, the substitution $D_{M_{\mathrm{opt}}} = (\beta + \alpha\tau/d) I_d$ scalarizes the entire system, so all three blocks become explicit rational functions of α in the long-run covariances, as in (40); this isolates $U_{gg}/(2\beta)$ as the only α-independent term (the large-repellence limit) and identifies the $O(\alpha^{-1})$ rate at which the remaining terms vanish. The two special cases that the original SRMC analysis isolates, i.e., the independent surrogate sampler and the Gaussian target with mean estimation [34, Section 3], reappear here as concrete instances under the *optimal* metric (full derivations in Appendix C.3). In the independent-sampler case, $\Sigma_{\mu\mu}^{(M_{\mathrm{opt}})}$ separates exactly into a residual floor $U_{gg}/(2\beta)$ plus an M-dependent term that contracts at the uniform rate $\beta + \alpha\tau/d$ and whose worst case over scalar test functions with a given score-aligned variance is uniquely minimized by M_{opt} (Step 3 of the independent-sampler case in Appendix C.3); it is smaller than its SRMC counterpart along the slow eigendirections of S, not in every direction. In the Gaussian-mean case, $g \equiv \mu_\pi$ so the residual floor vanishes and the entire covariance shrinks uniformly at rate $O(\alpha^{-1})$; along the broadest axis of π and at a common equilibrium kernel, its variance is below SRMC's on the same target by a factor tending to $\bar{\lambda}(S)/\lambda_{\min}(S)$ as $\alpha \to \infty$.

The residual floor as a shared phenomenon. GA-SRMC at M_{opt} and SRMC at $M = I_d$ share a basic structure: the score-history mechanism damps only the score-aligned component of f, at rate $O(\alpha^{-1})$, and as $\alpha \to \infty$ the covariance tends to the score-orthogonal residual $U_{gg}/(2\beta)$ in (15) for GA-SRMC, the analog of the residual term in Hu et al. [34, Section 3, Eq. (15), and Appendix C.10] for SRMC; this limit is a lower bound when the cross term in (40) vanishes, e.g., under independent sampling. Two points, listed below, distinguish the GA-SRMC form. First, *scalarization*: the M-dependent component vanishes uniformly in direction at rate $O(\alpha^{-1})$, rather than direction-dependently as under $M = I_d$. Second, *a common floor*: U_{gg} is the score-orthogonal long-run covariance of the equilibrium kernel, set entirely by the test function and that kernel, so it depends on

M only if the kernel does; at a common equilibrium kernel GA-SRMC and SRMC share the same floor (compare (46) with (47)).

Genealogy of the residual decomposition. Turning from the *SRMC line* to GA-SRMC's other intellectual ancestor: the score-aligned / score-orthogonal split $f = CS^{-1}s + g$ used in Step 1 of Appendix C.2 is the multivariate version of the variance-reducing projection identified in three classical pieces of the Monte Carlo control-variate literature. Assaraf and Caffarel [5] introduced score-based control variates as a *zero-variance principle*, augmenting an estimator by terms in the linear span of the score so that the residual variance vanishes under the target. Mira et al. [43] extended the idea to general Bayesian estimators, designing zero-variance MCMC estimators by regressing the test function onto the score. Oates et al. [45] formalized the construction within a reproducing kernel Hilbert space (RKHS) framework as *control functionals*, in which Stein operators are applied to an appropriate function space to generate variance-reducing controls. When these controls are restricted to the *linear span of the score*, the procedure removes precisely the score-aligned component $CS^{-1}s$, leaving the residual term g; more expressive Stein-type controls of the form $\phi' + s\phi$ are also permitted and can further reduce the residual contribution. For example, when $\pi = \mathcal{N}(0,1)$ and $f(x) = x^2$, one has $C = 0$ and hence $g = f$, yet choosing $\phi(x) = x$ yields the correction $1 - x^2$ and produces a zero-variance estimator. The present analogy, however, pertains exclusively to the linear-score component. GA-SRMC implements this idea *dynamically* via the surrogate target $\pi_\theta^{(M)}$, rather than through a static, post-hoc adjustment. Specifically, the score-aligned component is attenuated by the running score average θ_n at rate $O(\alpha^{-1})$, while the residual term induces the same asymptotic floor $U_{gg}/(2\beta)$. This decomposition characterizes the dynamic estimator for every fixed M: for a common equilibrium kernel, the limiting value $U_{gg}/(2\beta)$ is independent of M (cf. (46) and (47)). Theorem 3.1 further identifies an optimal choice of metric: M_{opt} is the unique trace-constrained metric for which $D_M = (\beta + \alpha\tau/d)I_d$ is scalar, thereby ensuring that the score-aligned component is damped isotropically, i.e., at the same rate in every direction, rather than at direction-dependent rates $\beta + \alpha\lambda_i(SM)$.

Why a closed form was not previously available. Both lines stop short of an optimal metric on the score space, but for opposite reasons. The *SRMC line* [34] fixes $M = I_d$ a priori and obtains scalar (Frobenius) bounds. The *control-variate line* [5, 43, 45] constructs a residual decomposition but uses the Fisher matrix in a static post-hoc estimator rather than as a dynamic damping geometry, leaving the metric question unmotivated. Theorem 3.1 bridges the two by combining three ingredients: surrogate-target machinery delivers an $O(\alpha^{-1})$ damping on the score-aligned component, the residual decomposition isolates the floor it leaves, and the trace-normalized inverse Fisher is the unique metric that scalarizes the resulting score-history Lyapunov equation. The result is a closed-form joint (θ, μ) CLT for every $f \in L^2(\pi)$, the strongest such statement currently available in either line.

C.5 Proof of Theorem 3.2

The argument follows Steps 1–8 in the proof of Proposition 2.2 (Appendix B.5), applied to the augmented iterate z_n; accordingly, we only document the points at which the proof requires modification. The two adjustments stated after Theorem 3.2, namely, the use of a moving-metric Lyapunov function and the identity $\pi_0^{(M)} = \pi$ for all M, both enter in Step 4. Because the stability of $\widehat{S}_n$ is postulated via Assumption 3, rather than established through an ODE@∞ argument, we instead verify the conditions of Borkar et al. [12] for a localized Markov chain that agrees with z_n as long as z_n remains in a given compact set, and subsequently pass to the limit by allowing this compact set to expand.

Augmented state and mean-field notation. Extending the fixed-M notation of Section 2, we write $z = (\theta, \mu, \widehat{S}) \in \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{S}^d$ (identifying symmetric $d \times d$ matrices $\mathbb{S}^d$ with their vectorized representations in $\mathbb{R}^{d(d+1)/2}$) and denote the target equilibrium by $z^\star \triangleq (\vartheta^\star, S) = (0, \mu, S)$. For any $\widehat{S} \succ 0$, define the trace-normalized inverse-Fisher metric map

$$M_{\widehat{S}} \;\triangleq\; \frac{\mathrm{tr}(\widehat{S})}{d}\,\widehat{S}^{-1}, \tag{56}$$

so that the online metric in (16) is $M_n = M_{\widehat{S}_n}$ and its value at $\widehat{S} = S$ is $M_S = M_{\mathrm{opt}}$ from (14) with scale $\tau = \mathrm{tr}\, S$. Matching $S_M(\theta)$ and $F_M(\theta)$ in (5), define the surrogate second-moment map

$$Q_M(\theta) \;\triangleq\; \mathbb{E}_{\pi_\theta^{(M)}}\big[s(X)s(X)^\top\big]. \tag{57}$$

The one-step SA residual and the associated mean-field drift on $\mathbb{R}^d \times \mathbb{R}^m \times \mathbb{S}^d_{++}$ are then

$$\begin{aligned} H(z,x) \;&\triangleq\; \big(s(x)-\theta,\; f(x)-\mu,\; s(x)s(x)^\top - \widehat{S}\big), \\ h(z) \;&\triangleq\; \big(S_M(\theta)-\theta,\; F_M(\theta)-\mu,\; Q_M(\theta)-\widehat{S}\big)\Big|_{M=M_{\widehat{S}}}, \end{aligned} \tag{58}$$

and the online recursion of Algorithm 1 takes the coupled SA form

$$X_{n+1} \;\sim\; K_{\theta_n,M_n}(X_n,\cdot), \qquad z_{n+1} \;=\; z_n + \gamma_{n+1}\, H(z_n, X_{n+1}). \tag{59}$$

Under Assumption 1, the super-linear tail bound $U(x) \geq c\|x\|^p$ ($p>1$) implies that π has finite exponential moments of all linear orders (Lemma B.2), while $\|s(x)\| \leq L\|x\| + c_0$ grows at most linearly. By the dominated convergence theorem [23, Theorem 2.27] applied as in Lemma B.1, we may differentiate under the integral sign to arbitrary order for the test functions s, f, and $ss^\top$, where the $\pi_\theta^{(M)}$-integrability of $f \in L^2(\pi)$ and its derivatives follows from the Cauchy–Schwarz inequality:

$$\mathbb{E}_\pi\big[\|f(X)\|\,\|s(X)\|^k\, e^{-\alpha\theta^\top M s(X)}\big] \;\leq\; \|f\|_{L^2(\pi)} \Big(\mathbb{E}_\pi\big[\|s(X)\|^{2k}\, e^{-2\alpha\theta^\top M s(X)}\big]\Big)^{1/2} \;<\; \infty.$$

Since $\pi_\theta^{(M)} = \pi_{M\theta}^{(I_d)}$, the maps $S_M(\theta)$, $F_M(\theta)$, and $Q_M(\theta)$ are C^∞ functions of the product $\eta \triangleq M\theta \in \mathbb{R}^d$. Furthermore, $Q_M(\theta) \succ 0$ for every $\theta \in \mathbb{R}^d$ and $M \succ 0$: if $v^\top Q_M(\theta) v = \mathbb{E}_{\pi_\theta^{(M)}}[(v^\top s(X))^2] = 0$ for some $v \in \mathbb{R}^d$, then because $\pi_\theta^{(M)}$ has a strictly positive continuous density on $\mathbb{R}^d$ and $x \mapsto v^\top s(x)$ is continuous, we must have $v^\top \nabla U(x) = -v^\top s(x) = 0$ for all $x \in \mathbb{R}^d$. By the fundamental theorem of calculus along lines parallel to v, $U(x+tv) = U(x)$ for all $t \in \mathbb{R}$, which contradicts the coercive growth $U(x) \geq c\|x\|^p$ as $|t| \to \infty$ unless $v = 0$. In particular, $S = Q_M(0) \succ 0$.

We next establish two structural lemmas before introducing the localization.

Lemma C.2 (Shared-gain dominance of θ_n by $\widehat{S}_n$)**.** *Suppose that Algorithm 1 is initialized with $\theta_0 = 0$ and $\widehat{S}_0 \succ 0$. Then, along every sample path and for every $n \geq 0$, the Schur complement matrix $\widehat{S}_n - \theta_n\theta_n^\top$ is strictly positive definite: $\widehat{S}_n - \theta_n\theta_n^\top \succ 0$. Consequently, $\widehat{S}_n \succ 0$ (so M_n is well-defined), $\theta_n^\top \widehat{S}_n^{-1}\theta_n < 1$, and $\|\theta_n\|^2 < \lambda_{\max}(\widehat{S}_n)$.*

Proof. Fix $n \geq 0$ and abbreviate $\gamma \triangleq \gamma_{n+1} \in (0,1)$ and $s_{n+1} \triangleq s(X_{n+1})$. Because θ_n and $\widehat{S}_n$ share the step size γ, their recursions read

$$\theta_{n+1} \;=\; (1-\gamma)\theta_n + \gamma s_{n+1}, \qquad \widehat{S}_{n+1} \;=\; (1-\gamma)\widehat{S}_n + \gamma s_{n+1}s_{n+1}^\top.$$

Expanding the outer product $\theta_{n+1}\theta_{n+1}^\top = (1-\gamma)^2\theta_n\theta_n^\top + \gamma(1-\gamma)(\theta_n s_{n+1}^\top + s_{n+1}\theta_n^\top) + \gamma^2 s_{n+1}s_{n+1}^\top$ and substituting $(1-\gamma)^2 = (1-\gamma) - \gamma(1-\gamma)$ and $\gamma - \gamma^2 = \gamma(1-\gamma)$, we obtain the exact one-step decomposition

$$\begin{aligned} \widehat{S}_{n+1} - \theta_{n+1}\theta_{n+1}^\top &= (1-\gamma)\widehat{S}_n - (1-\gamma)^2\theta_n\theta_n^\top + \gamma(1-\gamma)\big(s_{n+1}s_{n+1}^\top - \theta_n s_{n+1}^\top - s_{n+1}\theta_n^\top\big) \\ &= (1-\gamma)\big(\widehat{S}_n - \theta_n\theta_n^\top\big) + \gamma(1-\gamma)\big(\theta_n\theta_n^\top - \theta_n s_{n+1}^\top - s_{n+1}\theta_n^\top + s_{n+1}s_{n+1}^\top\big) \\ &= (1-\gamma)\big(\widehat{S}_n - \theta_n\theta_n^\top\big) + \gamma(1-\gamma)(s_{n+1}-\theta_n)(s_{n+1}-\theta_n)^\top. \end{aligned}$$

We now induct on n. At $n=0$, $\widehat{S}_0 - \theta_0\theta_0^\top = \widehat{S}_0 \succ 0$ by initialization. If $\widehat{S}_n - \theta_n\theta_n^\top \succ 0$, then since $\gamma \in (0,1)$, the first term on the right-hand side is strictly positive definite and the second term is rank-one positive semidefinite; hence $\widehat{S}_{n+1} - \theta_{n+1}\theta_{n+1}^\top \succ 0$, completing the induction.

For the stated consequences, $\widehat{S}_n \succ \theta_n\theta_n^\top \succeq 0$ immediately gives $\widehat{S}_n \succ 0$, so $\widehat{S}_n^{-1}$ and M_n in (16) are well-defined. Congruently conjugating $\widehat{S}_n - \theta_n\theta_n^\top \succ 0$ by the nonsingular matrix $\widehat{S}_n^{-1/2}$ preserves strict positive definiteness [32, Observation 7.1.8], yielding

$$I_d - \widehat{S}_n^{-1/2}\theta_n\theta_n^\top\widehat{S}_n^{-1/2} \;\succ\; 0.$$

Since $\widehat{S}_n^{-1/2}\theta_n\theta_n^\top\widehat{S}_n^{-1/2}$ has rank at most one with largest eigenvalue equal to its trace $\theta_n^\top\widehat{S}_n^{-1}\theta_n$, it follows that $\theta_n^\top\widehat{S}_n^{-1}\theta_n < 1$. Finally, for $\theta_n \neq 0$, evaluating the quadratic form of $\widehat{S}_n - \theta_n\theta_n^\top \succ 0$ at the unit vector $v \triangleq \theta_n/\|\theta_n\|$ and applying the Rayleigh–Ritz theorem [32, Theorem 4.2.2] gives

$$\|\theta_n\|^2 \;=\; v^\top(\theta_n\theta_n^\top)v \;<\; v^\top\widehat{S}_n v \;\leq\; \lambda_{\max}(\widehat{S}_n),$$

while for $\theta_n = 0$ the inequality $0 < \lambda_{\max}(\widehat{S}_n)$ is immediate from $\widehat{S}_n \succ 0$. □

Lemma C.2 explains why Assumption 3(a) only needs to postulate boundedness of $\widehat{S}_n$: the score-history iterate θ_n is automatically bounded by $\lambda_{\max}(\widehat{S}_n)^{1/2}$, and the moving-metric Lyapunov function $\theta_n^\top \widehat{S}_n^{-1}\theta_n$ remains strictly bounded by 1 along every trajectory.

Lemma C.3 (Regularity and algebraic properties of the metric map)**.** *The metric map $\widehat{S} \mapsto M_{\widehat{S}}$ in* (56) *satisfies:*

- ***(i)*** *$\widehat{S} \mapsto M_{\widehat{S}}$ is C^∞ on the open positive-definite cone $\mathbb{S}^d_{++} \triangleq \{\widehat{S} \in \mathbb{S}^d : \widehat{S} \succ 0\}$, and hence Lipschitz on every compact subset of $\mathbb{S}^d_{++}$.*
- ***(ii)*** *At the true Fisher matrix $\widehat{S} = S \succ 0$, $M_S = M_{\mathrm{opt}}$ with scale $\tau = \operatorname{tr} S$.*
- ***(iii)*** *$SM_{\mathrm{opt}} = \frac{\operatorname{tr} S}{d} I_d$.*
- ***(iv)*** *For every $\widehat{S} \succ 0$ and $M = M_{\widehat{S}}$, $dM/\operatorname{tr}(M) = d\,\widehat{S}^{-1}/\operatorname{tr}(\widehat{S}^{-1})$, which at $\widehat{S} = \widehat{S}_n$ coincides with the unregularized ($\delta = 0$) FisherMALA proposal preconditioner A_n of Appendix D.2.*

Proof. **(i)** By Cramer's rule [32, Section 0.8.2], each entry of $\widehat{S}^{-1}$ is a rational function of the entries of $\widehat{S}$ with nonvanishing denominator $\det(\widehat{S}) > 0$ on $\mathbb{S}^d_{++}$, and $\widehat{S} \mapsto \operatorname{tr}(\widehat{S})$ is linear; hence $\widehat{S} \mapsto M_{\widehat{S}}$ is C^∞ on $\mathbb{S}^d_{++}$. Using the standard Fréchet derivative of the matrix inverse $D(\widehat{S}^{-1})[E] = -\widehat{S}^{-1}E\widehat{S}^{-1}$ [14, Appendix A.4.1], the directional derivative of $M_{\widehat{S}}$ along $E \in \mathbb{S}^d$ is

$$DM_{\widehat{S}}[E] \;=\; \frac{\operatorname{tr}(E)}{d}\,\widehat{S}^{-1} - \frac{\operatorname{tr}(\widehat{S})}{d}\,\widehat{S}^{-1}E\widehat{S}^{-1}.$$

Because $\mathbb{S}^d_{++}$ is convex, the convex hull of any compact subset $\mathcal{C} \subset \mathbb{S}^d_{++}$ is again a compact subset of $\mathbb{S}^d_{++}$. By continuity, $\|DM_{\widehat{S}}\|$ is bounded on $\operatorname{conv}(\mathcal{C})$, and the mean value inequality [52, Theorem 9.19] implies that $\widehat{S} \mapsto M_{\widehat{S}}$ is Lipschitz on $\mathcal{C}$.

(ii)–(iii) Evaluating (56) at $\widehat{S} = S \succ 0$ gives $M_S = \frac{\operatorname{tr} S}{d} S^{-1}$, which matches $M_{\mathrm{opt}} = (\tau/d)S^{-1}$ in (14) with $\tau = \operatorname{tr} S$, and multiplying on the left by S gives $SM_{\mathrm{opt}} = \frac{\operatorname{tr} S}{d} SS^{-1} = \frac{\operatorname{tr} S}{d} I_d$.

(iv) Taking the trace of $M = M_{\widehat{S}} = \frac{\operatorname{tr}(\widehat{S})}{d}\widehat{S}^{-1}$ gives $\operatorname{tr}(M) = \frac{\operatorname{tr}(\widehat{S})}{d}\operatorname{tr}(\widehat{S}^{-1}) > 0$. Canceling the scalar factor $\operatorname{tr}(\widehat{S})/d$ between numerator and denominator yields

$$\frac{dM}{\operatorname{tr}(M)} \;=\; \frac{d\,\big(\operatorname{tr}(\widehat{S})/d\big)\,\widehat{S}^{-1}}{\big(\operatorname{tr}(\widehat{S})/d\big)\,\operatorname{tr}(\widehat{S}^{-1})} \;=\; \frac{d\,\widehat{S}^{-1}}{\operatorname{tr}(\widehat{S}^{-1})},$$

which at $\widehat{S} = \widehat{S}_n$ recovers $A_n = (d/\operatorname{tr}(\widehat{S}_{n,\delta}^{-1}))\widehat{S}_{n,\delta}^{-1}$ from Appendix D.2 at $\delta = 0$. □

Why localization on $(\theta, \widehat{S})$ is needed, and construction of the localized chain. Unlike the fixed-M recursion of Proposition 2.2, the unlocalized online drift $h(z)$ in (58) cannot satisfy the global Lipschitz and ODE@∞ radial-scaling hypotheses of Borkar et al. [12, Assumptions (A2), (A3), and (A5a)] over the entire open domain $\mathbb{R}^d \times \mathbb{R}^m \times \mathbb{S}^d_{++}$, for two structural reasons:

- **(a)** *Boundary singularity on $\partial\mathbb{S}^d_{++}$:* As $\lambda_{\min}(\widehat{S}) \downarrow 0$, the inverse-Fisher metric map $\widehat{S} \mapsto M_{\widehat{S}} = (\operatorname{tr}\widehat{S}/d)\widehat{S}^{-1}$ is unbounded, so $h(z)$ is only locally Lipschitz on $\mathbb{S}^d_{++}$.
- **(b)** *Super-linear growth of $Q_M(\theta)$ at infinity:* When $\alpha > 0$, the surrogate second moment $Q_M(\theta)$ grows super-linearly as $\|\theta\| \to \infty$. Even for the standard Gaussian target $\pi = \mathcal{N}(0, I_d)$ with $s(x) = -x$ and $z = (\theta, 0, I_d)$ ($\theta \neq 0$), completing the square in (3) gives $\pi_{r\theta}^{(I_d)} = \mathcal{N}(\alpha r\theta, I_d)$, so $M_{rI_d} = I_d$ and $Q_{I_d}(r\theta) = \mathbb{E}_{\pi_{r\theta}^{(I_d)}}[XX^\top] = I_d + \alpha^2 r^2 \theta\theta^\top$. The $\widehat{S}$-block of the rescaled drift $r^{-1}h(rz)$ is therefore $r^{-1}I_d + \alpha^2 r\theta\theta^\top - I_d$, which diverges as $r \to \infty$.

This is precisely why Theorem 3.2 postulates sample-path boundedness of $\widehat{S}_n$ in compact subsets of $\mathbb{S}^d_{++}$ via Assumption 3(a) (the standard localization route of Borkar et al. [12, Section 3]). Specifically, since $K_{\theta,M}$ and $(S_M(\theta), F_M(\theta), Q_M(\theta))$ depend only on $(\theta, \widehat{S})$ and not on μ, we define for each integer $j \geq 1$ the compact subset of $\mathbb{R}^d \times \mathbb{S}^d_{++}$:

$$\mathcal{K}_j \;\triangleq\; \big\{(\theta, \widehat{S}) \in \mathbb{R}^d \times \mathbb{S}^d :\; \|\theta\| \leq j,\;\; j^{-1}I_d \preceq \widehat{S} \preceq jI_d\big\}, \tag{60}$$

which is closed and bounded in $\mathbb{R}^d \times \mathbb{S}^d$ by continuity of the extreme eigenvalues [32, Corollary 4.3.15] and hence compact by the Heine–Borel theorem. Fix a smooth cutoff function $\chi_j \in C^\infty(\mathbb{R}^d \times \mathbb{S}^d; [0,1])$ satisfying $\chi_j \equiv 1$ on $\mathcal{K}_j$ and $\mathrm{supp}(\chi_j) \subset \mathcal{K}_{2j}$ [40, Proposition 2.25], and define the localized metric and score-history maps (viewed as functions of $z = (\theta, \mu, \widehat{S})$ via $(\theta, \widehat{S})$):

$$\bar{M}_j(z) \triangleq M_{\chi_j(\theta,\widehat{S})\widehat{S}+(1-\chi_j(\theta,\widehat{S}))I_d}, \quad \bar{\theta}_j(z) \triangleq \begin{cases} \chi_j(\theta,\widehat{S})\, \bar{M}_j(z)^{-1} M_{\widehat{S}}\theta, & (\theta,\widehat{S}) \in \mathrm{supp}(\chi_j), \\ 0, & (\theta,\widehat{S}) \notin \mathrm{supp}(\chi_j). \end{cases} \tag{61}$$

The prefactor $\bar{M}_j(z)^{-1} M_{\widehat{S}}$ in $\bar{\theta}_j(z)$ is chosen so that multiplying $\bar{\theta}_j(z)$ on the left by $\bar{M}_j(z)$ cancels $\bar{M}_j(z)^{-1}$ and preserves the exact alignment

$$\bar{M}_j(z)\bar{\theta}_j(z) \;=\; \chi_j(\theta,\widehat{S})\, M_{\widehat{S}}\theta \qquad \text{for all } z \in \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{S}^d_{++}, \tag{62}$$

including in the transition shell $\mathcal{K}_{2j} \setminus \mathcal{K}_j$ where $0 < \chi_j(\theta, \widehat{S}) < 1$. Consequently, the localized surrogate invariant density $\pi^{(\bar{M}_j(z))}_{\bar{\theta}_j(z)} = \pi^{(I_d)}_{\bar{M}_j(z)\bar{\theta}_j(z)} = \pi^{(M_{\widehat{S}})}_{\chi_j\theta}$ always tilts along $M_{\widehat{S}}\theta \propto \widehat{S}^{-1}\theta$, which will allow the moving-metric Lyapunov function $V(\theta, \widehat{S}) = \theta^\top \widehat{S}^{-1}\theta$ to decrease globally along the localized ODE flow in Step 4a. Moreover, because convex combinations of matrices in $\mathcal{K}_{2j}$ and I_d satisfy $(2j)^{-1} I_d \preceq \chi_j \widehat{S} + (1-\chi_j) I_d \preceq 2j I_d$, Lemma C.3(i) guarantees that $z \mapsto (\bar{\theta}_j(z), \bar{M}_j(z))$ is C^∞ on the entire vector space $\mathbb{R}^d \times \mathbb{R}^m \times \mathbb{S}^d$, coincides with $(\theta, M_{\widehat{S}})$ on $\mathcal{K}_j$, equals $(0, I_d)$ outside $\mathcal{K}_{2j}$, and takes values in a compact subset of $\mathbb{R}^d \times \mathbb{S}^d_{++}$.

Let $(\tilde{z}_n, \tilde{X}_n)$ be the localized chain initialized at $(\tilde{z}_0, \tilde{X}_0) = (z_0, X_0)$ and driven by the recursion

$$\tilde{X}_{n+1} \;\sim\; K_{\bar{\theta}_j(\tilde{z}_n), \bar{M}_j(\tilde{z}_n)}(\tilde{X}_n, \cdot), \qquad \tilde{z}_{n+1} \;=\; \tilde{z}_n + \gamma_{n+1}\, H(\tilde{z}_n, \tilde{X}_{n+1}), \tag{63}$$

with associated mean-field drift

$$h_j(z) \;\triangleq\; \big(S_{\bar{M}_j(z)}(\bar{\theta}_j(z)) - \theta,\; F_{\bar{M}_j(z)}(\bar{\theta}_j(z)) - \mu,\; Q_{\bar{M}_j(z)}(\bar{\theta}_j(z)) - \widehat{S}\big). \tag{64}$$

By construction, $h_j(z) = h(z)$ whenever $(\theta, \widehat{S}) \in \mathcal{K}_j$, and outside $\mathcal{K}_{2j}$ we have $\bar{\theta}_j(z) = 0$, so $\pi^{(\bar{M}_j(z))}_{\bar{\theta}_j(z)} = \pi$ and

$$g_j(z) \;\triangleq\; h_j(z) + z \;=\; \big(S_{\bar{M}_j(z)}(\bar{\theta}_j(z)),\; F_{\bar{M}_j(z)}(\bar{\theta}_j(z)),\; Q_{\bar{M}_j(z)}(\bar{\theta}_j(z))\big) \tag{65}$$

is equal to the constant $z^\star = (0, \mu, S)$ outside $\mathcal{K}_{2j}$. Since g_j is C^∞ and constant outside a compact set in $(\theta, \widehat{S})$ (and independent of μ), g_j is globally bounded by some constant $M_j < \infty$ and globally Lipschitz on $\mathbb{R}^d \times \mathbb{R}^m \times \mathbb{S}^d$.

Finally, coupling $(\tilde{z}_n, \tilde{X}_n)$ with (z_n, X_n) using identical driving innovations up to the first exit time of $(\theta_n, \widehat{S}_n)$ from $\mathcal{K}_j$, the two chains coincide for all $n \geq 0$ on the event

$$\mathsf{E}_j \;\triangleq\; \big\{(\theta_n, \widehat{S}_n) \in \mathcal{K}_j \;\text{ for all } n \geq 0\big\}. \tag{66}$$

By Assumption 3(a) and Lemma C.2 ($\|\theta_n\|^2 < \lambda_{\max}(\widehat{S}_n)$), the events E_j are nondecreasing in j and satisfy $\lim_{j\to\infty} \Pr(\mathsf{E}_j) = \Pr(\bigcup_{j\geq 1} \mathsf{E}_j) = 1$. Throughout the proof below, we fix j large enough that $(0, S)$ lies in the interior of $\mathcal{K}_j$.

Proof of Theorem 3.2. **Step 1: Step-size conditions (A1, A5b).** The step-size schedule $\gamma_n = (n + n_0)^{-\rho}$ ($n_0 \geq 1$) with $\rho \in (\frac{1}{2}, 1]$ is shared across all coordinates of $\tilde{z}_n$, so Borkar et al. [12, Assumptions (A1) and (A5b)] hold verbatim as in Step 1 of Appendix B.5, with $\lim_{n\to\infty}(\gamma_{n+1}^{-1} - \gamma_n^{-1}) = \mathbf{1}_{\{\rho=1\}}$.

Step 2: Lipschitz continuity and W-weighted growth of H via coordinate rescaling (A2, A4a). By (58), $z \mapsto H(z, x) = H(0, x) - z$ is 1-Lipschitz in z for every x, and Assumption 3(b) gives $\|H(0,x)\| \leq \|s(x)\| + \|f(x)\| + \|s(x)\|^2 \leq C\, W(x)$ for a constant $C > 0$. Because the empirical Fisher term $\|s(x)s(x)^\top\| = \|s(x)\|^2$ is $O(W(x))$ rather than $o(W(x))$, we use the non-$o(W)$ stability criterion of Borkar et al. [12, Assumption (A4a) and Proposition 2], which requires the W-weighted Lipschitz envelope ratio $\delta_L \triangleq \sup_{x\in\mathbb{R}^d} L(x)/W(x)$ from Borkar et al. [12, Assumption (A2)] to be sufficiently small (specifically, $\delta_L < 1/(4T_r)$ for the ODE@∞ relaxation time T_r determined in Step 3). To make δ_L arbitrarily small without altering the recursion, we pass to the linearly rescaled coordinate $y_n \triangleq \varepsilon \tilde{z}_n$ ($\varepsilon > 0$), which satisfies $y_{n+1} = y_n + \gamma_{n+1} H_\varepsilon(y_n, \tilde{X}_{n+1})$

with $H_\varepsilon(y,x) \triangleq \varepsilon H(\varepsilon^{-1}y,x) = \varepsilon H(0,x) - y$. The rescaled map $H_\varepsilon(\cdot,x)$ remains 1-Lipschitz in y and satisfies $\|H_\varepsilon(0,x)\| \le \varepsilon C\, W(x)$, so its Lipschitz/growth envelope in Borkar et al. [12, Assumption (A2)] is $L_\varepsilon(x) = \max\{1, \varepsilon C\, W(x)\}$. By Borkar et al. [12, Lemma 2(i)], replacing $W(x)$ with the truncated Lyapunov function $W_3(x) \triangleq \max\{3, W(x)\}$ preserves the (DV3) drift condition and yields

$$\delta_L \;=\; \sup_{x\in\mathbb{R}^d} \frac{L_\varepsilon(x)}{W_3(x)} \;\le\; \max\Big\{\frac{1}{3},\, \varepsilon C\Big\} \;=\; \frac{1}{3} \qquad \text{for all } \varepsilon \in \big(0, (3C)^{-1}\big]. \tag{67}$$

By linearity of $y_n = \varepsilon \tilde{z}_n$, almost-sure convergence and the CLT for y_n transfer immediately back to $\tilde{z}_n$.

Step 3: ODE@∞ stability for the localized drift (A3) and completion of (A4a). For the localized drift $h_j(z) = -z + g_j(z)$ in (64), the map g_j in (65) is globally bounded by $M_j < \infty$. Consequently, as $r \to \infty$,

$$\sup_{\|z\|=1} \big\| r^{-1} h_j(rz) - (-z) \big\| \;=\; \sup_{\|z\|=1} r^{-1} \big\| g_j(rz) \big\| \;\le\; \frac{M_j}{r} \;\longrightarrow\; 0.$$

Thus the limiting ODE@∞ is $\dot{z}_\infty = -z_\infty$, whose flow $\phi_t^\infty(z_0) = e^{-t} z_0$ contracts the unit ball to radius $1/2$ at the relaxation time $T_r = \log 2 \approx 0.693 < 3/4$. Combining $T_r < 3/4$ with $\delta_L \le 1/3$ from (67) immediately gives $\delta_L \le 1/3 < 1/(4\log 2) = 1/(4T_r)$, completing the verification of Borkar et al. [12, Assumptions (A3) and (A4a) and Proposition 2].

Step 4: Moving-metric Lyapunov function, Hurwitz Jacobian, and global contraction (A5a, A5b). Because $g_j(z) = h_j(z) + z$ is C^∞ and constant outside $\mathcal{K}_{2j}$, h_j is globally Lipschitz and smooth, verifying the regularity part of Borkar et al. [12, Assumption (A5a)]. We establish global exponential stability of $\dot{z} = h_j(z)$ in three sub-steps:

Step 4a: Uniform positive definiteness of $\widehat{S}_t$, moving-metric Lyapunov decay, and global asymptotic stability. Since $\eta = \bar{M}_j(z)\bar{\theta}_j(z)$ ranges over a compact set $\mathcal{C}_j \subset \mathbb{R}^d$ and $\eta \mapsto \lambda_{\min}(Q_{I_d}(\eta))$ and $\eta \mapsto \lambda_{\max}(Q_{I_d}(\eta))$ are continuous [32, Corollary 4.3.15] with $Q_{I_d}(\eta) \succ 0$ on $\mathbb{R}^d$, Weierstrass's extreme value theorem [52, Theorem 4.16] yields constants $0 < q_j \le Q_j^{\max} < \infty$ such that $q_j I_d \preceq Q_{\bar{M}_j(z)}(\bar{\theta}_j(z)) \preceq Q_j^{\max} I_d$ for all z. Along any trajectory of $\dot{z} = h_j(z)$, the $\widehat{S}$-component satisfies $q_j I_d - \widehat{S}_t \preceq \dot{\widehat{S}}_t = Q_{\bar{M}_j}(\bar{\theta}_j) - \widehat{S}_t \preceq Q_j^{\max} I_d - \widehat{S}_t$; multiplying by e^t and integrating over $[0,t]$ (variation of constants) gives

$$e^{-t}\widehat{S}_0 + (1-e^{-t}) q_j I_d \;\preceq\; \widehat{S}_t \;\preceq\; e^{-t}\widehat{S}_0 + (1-e^{-t}) Q_j^{\max} I_d. \tag{68}$$

Hence, from any initial $\widehat{S}_0 \in \mathbb{S}^d$, $\widehat{S}_t$ enters the open cone $\mathbb{S}^d_{++}$ at a finite time $t_0 \ge 0$ ($t_0 = 0$ when $\widehat{S}_0 \succ 0$) and satisfies $\underline{\lambda} I_d \preceq \widehat{S}_t \preceq \overline{\lambda} I_d$ for all $t \ge t_0$ for some $0 < \underline{\lambda} \le \overline{\lambda} < \infty$. For $t \ge t_0$, define the moving-metric Lyapunov function

$$V(\theta, \widehat{S}) \;\triangleq\; \theta^\top \widehat{S}^{-1} \theta. \tag{69}$$

Using the inverse-derivative identity $\frac{d}{dt}(\widehat{S}_t^{-1}) = -\widehat{S}_t^{-1}\dot{\widehat{S}}_t\widehat{S}_t^{-1}$ [14, Appendix A.4.1] together with $\dot{\theta}_t = S_{\bar{M}_j}(\bar{\theta}_j) - \theta_t$ and $\dot{\widehat{S}}_t = Q_{\bar{M}_j}(\bar{\theta}_j) - \widehat{S}_t$, the time derivative of V along the flow for $t \ge t_0$ is

$$\begin{aligned}
\dot{V}(\theta_t, \widehat{S}_t) &= 2\theta_t^\top \widehat{S}_t^{-1}\big(S_{\bar{M}_j}(\bar{\theta}_j) - \theta_t\big) - \theta_t^\top \widehat{S}_t^{-1}\big(Q_{\bar{M}_j}(\bar{\theta}_j) - \widehat{S}_t\big)\widehat{S}_t^{-1}\theta_t \\
&= 2\theta_t^\top \widehat{S}_t^{-1} S_{\bar{M}_j}(\bar{\theta}_j) - \theta_t^\top \widehat{S}_t^{-1} Q_{\bar{M}_j}(\bar{\theta}_j)\widehat{S}_t^{-1}\theta_t - V(\theta_t, \widehat{S}_t) \\
&\le 2\theta_t^\top \widehat{S}_t^{-1} S_{\bar{M}_j}(\bar{\theta}_j) - V(\theta_t, \widehat{S}_t) \;\le\; -V(\theta_t, \widehat{S}_t).
\end{aligned} \tag{70}$$

Here the second equality combines $-2\theta_t^\top \widehat{S}_t^{-1}\theta_t + \theta_t^\top \widehat{S}_t^{-1}\widehat{S}_t\widehat{S}_t^{-1}\theta_t = -V(\theta_t, \widehat{S}_t)$; the first inequality drops $-\theta_t^\top \widehat{S}_t^{-1} Q_{\bar{M}_j}(\bar{\theta}_j)\widehat{S}_t^{-1}\theta_t \le 0$ (since $Q_{\bar{M}_j} \succeq 0$); and the second inequality uses $\theta_t^\top \widehat{S}_t^{-1} S_{\bar{M}_j}(\bar{\theta}_j) \le 0$, which follows from the alignment (62): whenever $\chi_j(\theta_t, \widehat{S}_t) > 0$, we have $S_{\bar{M}_j}(\bar{\theta}_j) = S_{M_{\widehat{S}_t}}(\chi_j \theta_t)$ with $M_{\widehat{S}_t} = (\operatorname{tr} \widehat{S}_t/d)\widehat{S}_t^{-1}$ a positive scalar multiple of $\widehat{S}_t^{-1}$, so the log-partition convexity bound (29) evaluated at $\chi_j\theta_t$ gives $\theta_t^\top \widehat{S}_t^{-1} S_{M_{\widehat{S}_t}}(\chi_j\theta_t) \le 0$; whenever $\chi_j(\theta_t, \widehat{S}_t) = 0$, $S_{\bar{M}_j}(0) = \mathbb{E}_\pi[s(X)] = 0$. By Grönwall's inequality [13, Appendix B], $V(\theta_t, \widehat{S}_t) \le e^{-(t-t_0)} V(\theta_{t_0}, \widehat{S}_{t_0}) \to 0$, and since $\widehat{S}_t \preceq \overline{\lambda} I_d$, we obtain $\|\theta_t\|^2 \le \overline{\lambda}\, V(\theta_t, \widehat{S}_t) \to 0$ as

$t \to \infty$. Because $\bar{M}_j(z_t)$ is bounded on $\mathcal{K}_{2j}$, $\theta_t \to 0$ implies $\eta_t \triangleq \bar{M}_j(z_t)\bar{\theta}_j(z_t) \to 0$, so by continuity $Q_{\bar{M}_j(z_t)}(\bar{\theta}_j(z_t)) = Q_{I_d}(\eta_t) \to Q_{I_d}(0) = S$ and $F_{\bar{M}_j(z_t)}(\bar{\theta}_j(z_t)) = F_{I_d}(\eta_t) \to F_{I_d}(0) = \mu$. Feeding these converging forcing terms into the linear inhomogeneous ODEs $\dot{\widehat{S}}_t = -\widehat{S}_t + Q_{I_d}(\eta_t)$ and $\dot{\mu}_t = -\mu_t + F_{I_d}(\eta_t)$ yields $\widehat{S}_t \to S$ and $\mu_t \to \mu$ as $t \to \infty$ by variation of constants.

Step 4b: Block-lower-triangular Hurwitz Jacobian via $\pi_0^{(M)} = \pi$. Because $(0, S) \in \operatorname{int}(\mathcal{K}_j)$, we have $\chi_j \equiv 1$ and $h_j(z) = h(z)$ in a neighborhood of $z^\star = (0, \mu, S)$. There, the maps $S_{M_{\widehat{S}}}(\theta) = S_{I_d}(M_{\widehat{S}}\theta)$, $F_{M_{\widehat{S}}}(\theta) = F_{I_d}(M_{\widehat{S}}\theta)$, and $Q_{M_{\widehat{S}}}(\theta) = Q_{I_d}(M_{\widehat{S}}\theta)$ depend on $\widehat{S}$ only through the product $\eta = M_{\widehat{S}}\theta$. By the chain rule, the directional derivative of $S_{M_{\widehat{S}}}(\theta)$ with respect to $\widehat{S}$ along any $E \in \mathbb{S}^d$ evaluated at $(\theta, \widehat{S}) = (0, S)$ carries a factor of $\theta = 0$ and therefore vanishes identically:

$$\frac{\partial}{\partial \widehat{S}}\big[S_{I_d}(M_{\widehat{S}}\theta)\big]\Big|_{(\theta,\widehat{S})=(0,S)}[E] \;=\; \frac{\partial S_{I_d}}{\partial \eta}(0)\, DM_S[E]\,\theta\Big|_{\theta=0} \;=\; 0,$$

and identically $\frac{\partial}{\partial \widehat{S}}[F_{M_{\widehat{S}}}(\theta)]_{(0,S)} = 0$ and $\frac{\partial}{\partial \widehat{S}}[Q_{M_{\widehat{S}}}(\theta)]_{(0,S)} = 0$. Combining Lemma B.3 for the (θ, μ) derivatives with Lemma B.1 applied to $h = ss^\top$ at $\theta = 0$ for the $\widehat{S}$ block therefore yields the block-lower-triangular Jacobian

$$J_\star \;\triangleq\; \nabla h_j(z^\star) \;=\; \begin{bmatrix} J_\star^{(M_{\mathrm{opt}})} & 0 \\ \big[-\alpha\, \mathrm{Cov}_\pi(ss^\top, s)\, M_{\mathrm{opt}} \quad 0\big] & -I_{d(d+1)/2} \end{bmatrix}. \tag{71}$$

By Lemma C.3(iii), $SM_{\mathrm{opt}} = (\operatorname{tr} S/d)I_d$, so the eigenvalues of $J_\star$ are $-1 - \alpha \operatorname{tr} S/d$ (with multiplicity d) and -1 (with multiplicity $m + d(d+1)/2$). Consequently, $\frac{1}{2}\mathbf{1}_{\{\rho=1\}}I + J_\star$ is Hurwitz and $z^\star$ is the unique globally asymptotically stable equilibrium of $\dot{z} = h_j(z)$.

Step 4c: Global exponential contraction bound (A5b). Because $J_\star = \nabla h_j(z^\star)$ is Hurwitz with spectral abscissa -1, local exponential stability [35, Theorem 4.7] provides a positively invariant neighborhood $U_\delta = \{\|z - z^\star\| \le \delta_0\}$ ($\delta_0 > 0$) and constants $c_0 \ge 1$, $\rho_e \in (0, 1)$ such that $\|\phi_t(z_0) - z^\star\| \le c_0\|z_0 - z^\star\|e^{-\rho_e t}$ for all $z_0 \in U_\delta$ and $t \ge 0$. For trajectories starting outside U_δ ($\|z_0 - z^\star\| > \delta_0$), writing $\dot{z}_t = -z_t + g_j(z_t)$ with $\|g_j(z)\| \le M_j$ and applying variation of constants gives

$$\|\phi_t(z_0) - z^\star\| \;\le\; e^{-t}\|z_0 - z^\star\| + (1 - e^{-t})\big(M_j + \|z^\star\|\big) \qquad \text{for all } t \ge 0.$$

Hence every trajectory enters the compact ball $B_R \triangleq \{\|z - z^\star\| \le R\}$ with $R \triangleq 2(M_j + \|z^\star\|)$ within time $T_1(z_0) \triangleq \max\{0, \log(\|z_0 - z^\star\|/(M_j + \|z^\star\|))\}$. By global asymptotic stability (Step 4a) and continuity of the flow ϕ_t, compactness of B_R yields a uniform finite transit time $T_2 < \infty$ such that $\phi_t(B_R) \subset U_\delta$ for all $t \ge T_2$. Using $1 < \delta_0^{-1}\|z_0 - z^\star\|$ on $\{z_0 \notin U_\delta\}$ to absorb the bounded transient $(M_j + \|z^\star\|)$ over $[0, T_1(z_0) + T_2]$ and splicing with the exponential decay inside U_δ for $t \ge T_1(z_0) + T_2$ yields $\|\phi_t(z_0) - z^\star\| \le c_e\|z_0 - z^\star\|e^{-\rho_e t}$ for all z_0 and all $t \ge 0$, verifying Borkar et al. [12, Assumption (A5b), Eq. (43)].

Step 5: Base kernel regularity (DV3 and global Lipschitzness in A2). Because $(\bar{\theta}_j(z), \bar{M}_j(z))$ takes values in a compact subset of $\mathbb{R}^d \times \mathbb{S}^d_{++}$, Assumption 3(b) provides the uniform (DV3) drift condition on the sublevel sets of W (and hence of W_3) and Lipschitz continuity of $(\theta, M) \mapsto K_{\theta,M}$ on that compact range. Composing with the smooth map $z \mapsto (\bar{\theta}_j(z), \bar{M}_j(z))$—which is constant outside $\mathcal{K}_{2j}$ and globally Lipschitz by Lemma C.3(i)—makes $z \mapsto K_{\bar{\theta}_j(z), \bar{M}_j(z)}$ globally Lipschitz in the W_3^δ-norm across the entire state space, fulfilling Borkar et al. [12, Assumption (A2)].

Step 6: Conclusion via the SA CLT and transfer to the unlocalized chain. Having verified Borkar et al. [12, Assumptions (A1)–(A5)] for the localized chain $\tilde{z}_n$ in Steps 1–5, Borkar et al. [12, Theorem 2] bounds the moments of $\tilde{z}_n$ and of the exponential (DV3) Lyapunov function uniformly in n. Because (DV3) dominates W_3 by the exponential Lyapunov function plus a constant, applying Borkar et al. [12, Lemma 1 and Propositions 7(ii) and 8] to the Markovian noise increments $\Delta_{k+1} \triangleq H(\tilde{z}_k, \tilde{X}_{k+1}) - h_j(\tilde{z}_k)$ shows that the noise series $R_n \triangleq \sum_{k=0}^{n-1} \gamma_{k+1}\Delta_{k+1}$ converges almost surely (its martingale part is bounded in L^2 and converges by Doob's martingale convergence theorem, while the Poisson-equation boundary and drift-difference terms are absolutely summable). Writing $\tilde{z}_{n+1} = (1 - \gamma_{n+1})\tilde{z}_n + \gamma_{n+1}g_j(\tilde{z}_n) + (R_{n+1} - R_n)$ with g_j globally bounded by M_j and R_n convergent, the Kushner–Clark / ODE-method lemma [13, Chapter 2, Lemma 1 and Theorem 2] implies that $\tilde{z}_n$ is almost surely bounded and converges almost surely to the unique globally asymptotically stable equilibrium $z^\star$ of $\dot{z} = h_j(z)$ from Step 4: $\tilde{z}_n \xrightarrow{a.s.} z^\star$.

Simultaneously, Borkar et al. [12, Theorem 4] applied to $\tilde{z}_n$ yields the joint CLT

$$\gamma_n^{-1/2}(\tilde{z}_n - z^\star) \xrightarrow{dist.} \mathcal{N}(0, \Sigma_z), \tag{72}$$

where Σ_z is the unique symmetric positive-semidefinite solution of the Lyapunov equation

$$\big(\tfrac{1}{2}\mathbf{1}_{\{\rho=1\}} I + J_\star\big)\Sigma_z + \Sigma_z\big(\tfrac{1}{2}\mathbf{1}_{\{\rho=1\}} I + J_\star\big)^\top + U_{zz} \;=\; 0, \tag{73}$$

and U_{zz} is the long-run covariance (9) of the centered residual $H(z^\star, \cdot)$ along a stationary trajectory of the equilibrium kernel $K_{0,M_{\mathrm{opt}}}$. Crucially, neither $J_\star$ nor U_{zz} depends on the localization index j.

To transfer both conclusions from $\tilde{z}_n$ to the true online iterate z_n, recall from (66) that $z_n = \tilde{z}_n$ for all $n \geq 0$ on the event E_j, and that $\lim_{j\to\infty} \Pr(\mathsf{E}_j) = 1$ by Assumption 3(a) and Lemma C.2. Therefore,

$$\Pr(z_n \to z^\star) \;\geq\; \Pr\big(\{z_n = \tilde{z}_n \;\forall n\} \cap \{\tilde{z}_n \to z^\star\}\big) \;=\; \Pr(\mathsf{E}_j) \xrightarrow{j\to\infty} 1,$$

proving $z_n \xrightarrow{a.s.} z^\star$. Likewise, for any bounded continuous test function ψ,

$$\Big|\mathbb{E}\big[\psi\big(\gamma_n^{-1/2}(z_n - z^\star)\big)\big] - \mathbb{E}\big[\psi\big(\gamma_n^{-1/2}(\tilde{z}_n - z^\star)\big)\big]\Big| \;\leq\; 2\|\psi\|_\infty \Pr(\mathsf{E}_j^c).$$

Taking $\limsup_{n\to\infty}$ via (72) and then letting $j \to \infty$ shows that $\mathbb{E}[\psi(\gamma_n^{-1/2}(z_n - z^\star))] \to \mathbb{E}[\psi(Z)]$ with $Z \sim \mathcal{N}(0, \Sigma_z)$, which by the portmanteau theorem [11, Theorem 2.1] yields $\gamma_n^{-1/2}(z_n - z^\star) \xrightarrow{dist.} \mathcal{N}(0, \Sigma_z)$. (When the initial state X_0 is random, applying the above argument conditionally on X_0 and taking expectations via the bounded convergence theorem transfers weak convergence unconditionally.)

Step 7: Decoupling the (θ, μ) marginal. Partitioning $z = (\vartheta, \widehat{S})$ into the $(d+m)$-dimensional parameter block $\vartheta = (\theta, \mu)$ and the $d(d+1)/2$-dimensional Fisher block $\widehat{S}$, the shifted Jacobian in (71), the asymptotic covariance Σ_z, and the long-run noise covariance U_{zz} take the 2×2 block forms

$$\tfrac{1}{2}\mathbf{1}_{\{\rho=1\}} I + J_\star = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \qquad \Sigma_z = \begin{bmatrix} \Sigma_{\vartheta\vartheta} & \Sigma_{\vartheta S} \\ \Sigma_{S\vartheta} & \Sigma_{SS} \end{bmatrix}, \qquad U_{zz} = \begin{bmatrix} U_{HH} & U_{\vartheta S} \\ U_{S\vartheta} & U_{SS} \end{bmatrix},$$

where $A_{11} \triangleq \frac{1}{2}\mathbf{1}_{\{\rho=1\}} I_{d+m} + J_\star^{(M_{\mathrm{opt}})}$ and the upper-left block of U_{zz} is U_{HH} from Section 2 because the first two components of $H(z^\star, \cdot)$ coincide with the fixed-M residual $H(\vartheta^\star, \cdot)$ under the equilibrium kernel $K_{0,M_{\mathrm{opt}}}$. Expanding the $(1,1)$ block of the full Lyapunov equation (73) gives

$$\big(A_{11}\Sigma_{\vartheta\vartheta} + 0 \cdot \Sigma_{S\vartheta}\big) + \big(\Sigma_{\vartheta\vartheta} A_{11}^\top + \Sigma_{\vartheta S} \cdot 0^\top\big) + U_{HH} \;=\; A_{11}\Sigma_{\vartheta\vartheta} + \Sigma_{\vartheta\vartheta} A_{11}^\top + U_{HH} \;=\; 0,$$

which is precisely the fixed-M Lyapunov equation (10) evaluated at $M = M_{\mathrm{opt}}$. By uniqueness of the solution to (10), we obtain $\Sigma_{\vartheta\vartheta} = \Sigma_\vartheta^{(M_{\mathrm{opt}})}$, establishing part (ii). (Although the off-diagonal cross-covariance $\Sigma_{\vartheta S}$ between (θ_n, μ_n) and $\widehat{S}_n$ is generically nonzero, the zero upper-right block of $J_\star$ prevents $\Sigma_{\vartheta S}$ from entering the $(1,1)$ block equation.)

Step 8: Asymptotic unbiasedness and the learned metric. Since $\widehat{S}_n \xrightarrow{a.s.} S \succ 0$ by Step 6 and $\widehat{S} \mapsto M_{\widehat{S}}$ is continuous at S by Lemma C.3(i), the continuous mapping theorem [56, Theorem 2.3] gives $M_n = M_{\widehat{S}_n} \xrightarrow{a.s.} M_S = M_{\mathrm{opt}}$, completing part (i). Consequently, $M_n\theta_n \xrightarrow{a.s.} 0$, so the surrogate density $\pi_{\theta_n}^{(M_n)} = \pi_{M_n\theta_n}^{(I_d)}$ converges pointwise to π and $\mu_n \xrightarrow{a.s.} \mu = \mathbb{E}_\pi[f(X)]$. Because the (θ, μ) marginal CLT in Step 7 matches the fixed-M CLT at $M = M_{\mathrm{opt}}$, the optimal-M variance bounds (13)–(15) of Theorem 3.1 hold at the learned metric under the equilibrium kernel $K_{0,M_{\mathrm{opt}}}$. □

The same localization argument applies verbatim to the Tikhonov-regularized online metric $M_{\widehat{S}_{n,\delta}}$ with $\widehat{S}_{n,\delta} = \widehat{S}_n + \delta I_d$ ($\delta > 0$) used in Section 4, replacing $\widehat{S}^{-1}$ by $(\widehat{S} + \delta I_d)^{-1}$ in the moving-metric Lyapunov function (69). In that case, only the upper bound $\sup_{n\geq 0} \lambda_{\max}(\widehat{S}_n) < \infty$ in Assumption 3(a) is needed (since $\widehat{S}_{n,\delta} \succeq \delta I_d$ is automatically bounded away from zero), and the learned metric converges almost surely to $M_{S+\delta I_d}$, which recovers M_{opt} as $\delta \downarrow 0$ (or whenever S is isotropic). When $\widehat{S}_n$ is instead frozen after a finite warmup period (as in the `freeze_after_L` option of Appendix D.2), the subsequent chain evolves with a constant positive-definite metric M and is governed directly by Proposition 2.2. Finally, the same localization also relaxes the kernel hypotheses

of Proposition 2.2 to compact sets: for any fixed $M \succ 0$, omitting the $\widehat{S}$ coordinate and using the quadratic Lyapunov function (23) in Step 4 yields the conclusions of Proposition 2.2 whenever $\sup_{n\geq 0} \|\theta_n\| < \infty$ a.s. and Assumption 2 holds on compact subsets of θ with $\|s(x)\| + \|f(x)\| \leq c\,W(x)$.

D Additional Discussion

Summary of Appendix D. This appendix collects two threads of discussion that did not fit into the main body. Appendix D.1 explains why two natural-looking alternatives to the bilinear repulsion exponent $\theta^\top M s(x)$, i.e., cosine similarity and score-norm normalization, are not viable substitutes: cosine similarity destroys the asymptotic theory at the equilibrium $\theta^\star = 0$, while score-norm normalization either breaks the strong-Lyapunov argument that underlies the global convergence guarantee or, if the history recursion is also normalized, breaks Stein's identity and biases the Monte Carlo estimator. Appendix D.2 gives a self-contained description of the two online Fisher estimators introduced in Section 3 (diagonal and full), each with the explicit FISHERUPDATE and INVFISHER recursions, the trace-normalized inverse used in (16), and the per-iteration cost.

D.1 Why not cosine similarity or score-norm normalization?

A reasonable starting question is whether one could simply 'normalize away' the geometry mismatch motivating GA-SRMC by replacing the bilinear inner product $\theta^\top s(x)$ in the SRMC surrogate (2) with a scale-invariant alternative. Two candidates come to mind. One is the cosine similarity

$$\cos(\theta, s(x)) \;\triangleq\; \frac{\theta^\top s(x)}{\|\theta\|\,\|s(x)\|}, \tag{74}$$

i.e. the (Euclidean) inner product normalized by both $\|\theta\|$ and $\|s(x)\|$, so that it depends only on the angle between θ and $s(x)$ in $\mathbb{R}^d$. The other is score-norm normalization with a weaker scaling that only divides by $\|s(x)\|$,

$$\theta^\top u(x) \;\triangleq\; \frac{\theta^\top s(x)}{\|s(x)\|}, \qquad u(x) \triangleq \frac{s(x)}{\|s(x)\|}, \tag{75}$$

which is still linear in θ but uses the unit-norm direction $u(x)$ in place of the raw score. We argue that neither yields the matched structure used by GA-SRMC. The bilinear form $\theta^\top M s(x)$, with M symmetric positive definite, is the unique replacement that (a) preserves the linear-in-θ structure that makes the surrogate an exponential family in θ, (b) keeps the equilibrium $\theta^\star = 0$ at which $\pi^{(M)}_{\theta^\star} = \pi$, and (c) admits the Hessian-of-log-partition derivative identity (21) that drives the SA CLT.

D.1.1 Cosine similarity destroys the asymptotic theory at $\theta^\star = 0$

Cosine similarity introduces a nonlinear and singular dependence on θ: the surrogate is undefined at $\theta^\star = 0$, which is precisely the equilibrium to which SRMC must converge. Moreover, the corresponding tilt is no longer an exponential family with a sufficient statistic independent of θ, so the clean derivation of the log-partition function in Appendix B.2 no longer applies. We make these failures precise as a single proposition with two formal claims.

Proposition D.1 (Two failures of the cosine surrogate)**.** *Let $\pi^{\cos}_\theta(x) \propto \pi(x)\exp\{-\alpha\cos(\theta, s(x))\}$ be the cosine-similarity surrogate, with $\cos(\theta, s(x))$ defined in (74) and $\alpha > 0$ fixed. Then:*

- ***(i) Discontinuous surrogate at the equilibrium.*** *$\pi^{\cos}_\theta$ does not extend continuously to $\theta^\star = 0$: for any deterministic convention $\cos(0, s(x)) \triangleq c_0(x) \in \mathbb{R}$, the directional limits $\lim_{t\to 0}\pi^{\cos}_{tv} \neq \lim_{t\to 0}\pi^{\cos}_{tw}$ disagree for any two non-collinear $v, w \in \mathbb{R}^d$.*
- ***(ii) Non-existent Jacobian at the equilibrium.*** *The mean-field map $h^{\cos}(\theta) \triangleq \mathbb{E}_{\pi^{\cos}_\theta}[s(X)] - \theta$ is not differentiable at $\theta^\star = 0$: any candidate Jacobian carries a $1/\|\theta\|$ singularity along every approach direction.*

Proof. **Step 1: Direction-dependence of the limit (claim (i)).** Cosine similarity is scale-invariant in θ: for every $c > 0$, $\cos(c\theta, s(x)) = \cos(\theta, s(x))$, so the surrogate depends only on the direction $\theta/\|\theta\|$.

Fix two unit vectors $\hat{v} \neq \hat{w}$ with $\hat{v}^\top \hat{w} \in (-1, 1)$, and consider the directional limits along the rays $\theta_t^v \triangleq t\hat{v}$ and $\theta_t^w \triangleq t\hat{w}$ as $t \to 0$. By scale-invariance,

$$\lim_{t\to 0} \cos(\theta_t^v, s(x)) = \frac{\hat{v}^\top s(x)}{\|s(x)\|}, \qquad \lim_{t\to 0} \cos(\theta_t^w, s(x)) = \frac{\hat{w}^\top s(x)}{\|s(x)\|},$$

and these two functions of x are pointwise distinct since $\hat{v} \neq \hat{w}$. Hence, the surrogate distributions $\lim_{t\to 0} \pi^{\cos}_{\theta_t^v}$ and $\lim_{t\to 0} \pi^{\cos}_{\theta_t^w}$ are different probability measures, and no scalar convention $c_0(x)$ at the origin can reconcile them.

Step 2: Singularity of the gradient (claim (ii)). Differentiating cosine similarity in θ for $\theta \neq 0$ yields

$$\nabla_\theta \cos(\theta, s(x)) = \frac{1}{\|\theta\| \, \|s(x)\|} \Big(I - \frac{\theta\theta^\top}{\|\theta\|^2} \Big) s(x). \tag{76}$$

The factor $\|\theta\|^{-1}$ blows up as $\theta \to 0$, so $\nabla_\theta \cos(\theta, s(x))$ is unbounded in any neighborhood of the origin and the dominated-convergence step that justifies differentiating $\mathbb{E}_{\pi^{\cos}_\theta}[s]$ under the integral sign fails. The map $\theta \mapsto \theta/\|\theta\|$ is itself non-differentiable at 0, hence so is $h^{\cos}$. □

The cosine surrogate is *designed* to be scale-invariant in θ, which seems desirable as a way to remove the α-tuning burden. However, every surrogate must reduce to π at $\theta = 0$ in a path-independent way, since the score-history average θ_n converges to zero from a stochastic path. By contrast, the bilinear form $\theta^\top M s(x)$ is homogeneous of degree one in θ: if $\theta_n \to 0$, then $\theta_n^\top M s(x) \to 0$ pointwise for every x, so $\pi^{(M)}_{\theta_n} \to \pi$ regardless of the path of approach, and the Jacobian ∇h_M exists and is given in closed form by Lemma B.1.

D.1.2 Score-norm normalization breaks the strong-Lyapunov argument or breaks Stein's identity

Consider score-norm normalization (75), with the surrogate

$$\pi^{\mathrm{dir}}_\theta(x) \propto \pi(x) \exp\{-\alpha \, \theta^\top u(x)\}, \qquad u(x) \triangleq s(x)/\|s(x)\|.$$

Unlike cosine similarity, this form *is* an exponential family in θ, with sufficient statistic u. Differentiating its log-partition function under the integral sign (Appendix B.2) gives the analog of Lemma B.1,

$$\nabla_\theta \mathbb{E}_{\pi^{\mathrm{dir}}_\theta}[h(X)] \;=\; -\alpha \, \mathrm{Cov}_{\pi^{\mathrm{dir}}_\theta}(h, u) \qquad \text{for any integrable } h, \tag{77}$$

so the surrogate itself is well-defined. We now show that downstream of this identity, the construction faces a dichotomy.

Proposition D.2 (Score-norm dichotomy)**.** *The score-norm-normalized surrogate $\pi^{\mathrm{dir}}_\theta$ admits exactly two natural pairings with the SRMC score-history recursion, and each fails:*

- ***(i) Raw-score history.*** *If the history recursion is the original $\theta_{n+1} = \theta_n + \gamma_{n+1}(s(X_{n+1}) - \theta_n)$, then in $d \geq 2$ the mean field $h_{\mathrm{dir}}(\theta) \triangleq \mathbb{E}_{\pi^{\mathrm{dir}}_\theta}[s(X)] - \theta$ generically fails to be the gradient of any scalar potential, so the strongly-convex Lyapunov argument used in the proof of Proposition 2.1 no longer applies.*
- ***(ii) Normalized-score history.*** *If the history recursion is replaced by $\theta_{n+1} = \theta_n + \gamma_{n+1}(u(X_{n+1}) - \theta_n)$ to restore consistency with the sufficient statistic, then there exist targets π for which Stein's identity for u fails ($\mathbb{E}_\pi[u(X)] \neq 0$), and consequently $\theta^\star = 0$ is not the SA fixed point: the surrogate $\pi^{\mathrm{dir}}_{\theta^\star} \neq \pi$ and the Monte Carlo estimator μ_n is asymptotically biased.*

Proof. **Step 1: Cross-covariance Jacobian fails to be a Hessian (claim (i)).** Applying (77) with $h = s$ gives the Jacobian

$$\nabla_\theta S_{\mathrm{dir}}(\theta) \;=\; -\alpha \, \mathrm{Cov}_{\pi^{\mathrm{dir}}_\theta}(s, u), \qquad S_{\mathrm{dir}}(\theta) \triangleq \mathbb{E}_{\pi^{\mathrm{dir}}_\theta}[s(X)]. \tag{78}$$

At $\theta = 0$ the cross-covariance collapses to $\mathbb{E}_\pi[ss^\top/\|s\|]$ (since $\mathbb{E}_\pi[s] = 0$ by Stein's identity), which is symmetric PSD; so the linearisation $\nabla h_{\mathrm{dir}}(0) = -I_d - \alpha \, \mathbb{E}_\pi[ss^\top/\|s\|] \prec 0$ is symmetric negative definite and $\theta^\star = 0$ is locally exponentially stable. Away from the origin, however, $\mathrm{Cov}_{\pi^{\mathrm{dir}}_\theta}(s, u)$ is generically non-symmetric in $d \geq 2$: there is no symmetry condition forcing $\partial S^i_{\mathrm{dir}}/\partial\theta_j = \partial S^j_{\mathrm{dir}}/\partial\theta_i$ on the off-diagonal entries, since s and u are different statistics. By Poincaré's lemma, h_{dir} is not the gradient of any scalar potential, so the deterministic ODE $\dot{\theta}_t = h_{\mathrm{dir}}(\theta_t)$ is not a gradient flow and the

strongly-convex Lyapunov argument of Step 3 in the proof of Proposition 2.1 no longer goes through. This is genuinely a $d \geq 2$ phenomenon: in 1-D, ∇S_{dir} is a scalar and trivially symmetric, so the drift always integrates.

Step 2: Stein's identity fails for u on a concrete target (claim (ii)). Lemma D.3 below exhibits a 1-D target with $\mathbb{E}_\pi[u(X)] \neq 0$, so the normalized-score recursion has nonzero drift at $\theta = 0$ and is pushed away from the origin. By the argument in Step 3 below, the resulting SA fixed point $\theta^\star \neq 0$ defines a surrogate $\pi^{\mathrm{dir}}_{\theta^\star} \neq \pi$, biasing the Monte Carlo estimator.

Step 3: Bias of the Monte Carlo estimator. Under the normalized-score recursion of (ii), the mean field is $h(\theta) = \mathbb{E}_{\pi^{\mathrm{dir}}_\theta}[u(X)] - \theta$, and any SA fixed point satisfies $\theta^\star = \mathbb{E}_{\pi^{\mathrm{dir}}_{\theta^\star}}[u(X)]$. Whenever $\mathbb{E}_\pi[u(X)] \neq 0$, the candidate $\theta = 0$ does not solve this equation (since at $\theta = 0$ the right-hand side is the nonzero value $\mathbb{E}_\pi[u(X)]$), so $\theta^\star \neq 0$. The surrogate at the fixed point is then $\pi^{\mathrm{dir}}_{\theta^\star}(x) \propto \pi(x) \exp\{-\alpha(\theta^\star)^\top u(x)\} \neq \pi(x)$ pointwise, and the empirical Monte Carlo average $\mu_n \to \mathbb{E}_{\pi^{\mathrm{dir}}_{\theta^\star}}[f(X)] \neq \mathbb{E}_\pi[f(X)]$ is asymptotically biased. □

Lemma D.3 (Stein's identity fails for the unit-norm direction)**.** *There exists a 1-D target π for which $\mathbb{E}_\pi[s(X)] = 0$ but $\mathbb{E}_\pi[u(X)] \neq 0$, where $u = s/\|s\| = \mathrm{sign}(s)$ in dimension one.*

Proof. Consider the asymmetric two-sided exponential

$$\pi(x) = \frac{ab}{a+b}\, e^{-ax}\, \mathbf{1}_{\{x>0\}} + \frac{ab}{a+b}\, e^{bx}\, \mathbf{1}_{\{x<0\}}, \qquad a > b > 0, \tag{79}$$

which has full support and finite normalizing constant. The score is piecewise constant: $s(x) = -a$ for $x > 0$ and $s(x) = b$ for $x < 0$. Stein's identity for s holds:

$$\mathbb{E}_\pi[s(X)] = -a\ \Pr(X>0) + b\ \Pr(X<0) = -a \cdot \tfrac{b}{a+b} + b \cdot \tfrac{a}{a+b} = 0.$$

However, the unit-norm direction $u(x) = \mathrm{sign}(s(x)) = -\mathrm{sign}(x)$ has

$$\mathbb{E}_\pi[u(X)] = -\Pr(X>0) + \Pr(X<0) = -\tfrac{b}{a+b} + \tfrac{a}{a+b} = \tfrac{a-b}{a+b} \ \neq\ 0,$$

since $a > b > 0$. At the canonical setting $a = 2$, $b = 1$, $\alpha = 1$, a numerical sweep of the fixed-point equation $\theta^\star = \mathbb{E}_{\pi^{\mathrm{dir}}_{\theta^\star}}[u(X)]$ gives $\theta^\star \approx 0.17 \neq 0$, so that the surrogate target no longer converges to the true target π. □

Stein's identity $\mathbb{E}_\pi[s(X)] = 0$ holds because $s = \nabla \log \pi$ has the integration-by-parts symmetry built into the score function. The bilinear $\theta^\top s(x)$ inherits this property: $\mathbb{E}_\pi[\theta^\top s(X)] = \theta^\top \mathbb{E}_\pi[s(X)] = 0$ for every θ, which is what makes $\theta^\star = 0$ the natural SA fixed point. Replacing s with any nonlinear function of the score (e.g., $u = s/\|s\|$) breaks this linearity, and only special targets satisfy a corresponding identity. GA-SRMC's matrix-weighted form $\theta^\top M s(x)$ keeps s as the inner statistic and weights it by a fixed matrix M, which preserves $\mathbb{E}_\pi[\theta^\top M s(X)] = 0$ regardless of M.

Proposition D.1 and Proposition D.2 together exhaust the natural 'scale-free' alternatives to the bilinear repulsion exponent. Cosine similarity (full normalization) destroys the equilibrium structure entirely; score-norm normalization (partial normalization) preserves the equilibrium but either loses the gradient-flow potential (case (i)) or loses Stein-correctness (case (ii)). The matrix-weighted bilinear $\theta^\top M s(x)$ is the *minimal* extension of SRMC that admits a Fisher-aware geometry while preserving both exactness (Stein) and the gradient-flow Lyapunov structure (Proposition 2.1); it is in this precise sense that GA-SRMC is the unique geometry-aware scale-equivariant lift of SRMC.

D.2 Online Fisher estimation

This subsection describes the two online Fisher estimators corresponding to the diagonal and full fidelity options of Section 3 and Algorithm 1. For each estimator, we describe the state $\mathcal{F}_n$ that is stored on the chain, the recursion that ingests a fresh score s_n and updates $\mathcal{F}_n \to \mathcal{F}_{n+1}$, the routine that applies the regularized inverse $\widehat{S}^{-1}_{n,\delta}$ to a vector, the trace-normalization that produces the metric M_n used in Eq. (16), and the resulting per-iteration cost. The recursion on $\widehat{S}_n$, the trace-normalized inverse Fisher used as the preconditioner A_n, and the diagonal-vs-full alternative are not new to GA-SRMC: they originate in the FisherMALA construction of Titsias [54] and are surveyed at length in the recent analysis of linear MCMC preconditioning by Hird and Livingstone [30]. What is new here is only the dual use of these objects on the repulsion side (the weight matrix M_n in addition

to the proposal-side A_n) and the small modifications described below; we attribute every borrowed component explicitly at the point where it is introduced.

Both estimators share the regularized matrix introduced in Section 3,

$$\widehat{S}_{n,\delta} \;=\; \widehat{S}_n + \delta I_d, \qquad \delta > 0, \tag{80}$$

which replaces the eigenvalue clipping $\lambda_i \leftarrow \max(\lambda_i, \eta)$ used in some earlier preconditioned-MCMC implementations. Three properties make the additive shift convenient. It is differentiable; it commutes with the linear recursion, so the recursion on $\widehat{S}_n$ propagates to $\widehat{S}_{n,\delta}$ without further bookkeeping; and it caps the eigenvalues of $\widehat{S}_{n,\delta}^{-1}$ uniformly at $1/\delta$, which bounds the proposal step and the surrogate-target tilt in directions where $\widehat{S}_n$ becomes small. The choice of δ turns out to be the dominant practical lever for samplers in this family: the sweep in Appendix A.3 shows that $\delta = 0.01$ improves the worst-direction ESS by 1.5 to 2.5 times over the customary $\delta = 10$ [54] on every target we considered, and we use this value throughout. With $\tau_n \triangleq \mathrm{tr}(\widehat{S}_{n,\delta})/d$, the proposal preconditioner is $A_n = (d/\,\mathrm{tr}(\widehat{S}_{n,\delta}^{-1}))\,\widehat{S}_{n,\delta}^{-1}$ and the repulsion-side metric is $M_n = \tau_n \widehat{S}_{n,\delta}^{-1}$. The proposal-side form A_n is exactly the trace-normalized inverse-Fisher preconditioner of Titsias [54, Section 3]; the repulsion-side M_n is its GA-SRMC counterpart introduced in Section 3. Both depend on $\widehat{S}_{n,\delta}^{-1}$ through a single matrix-vector product, so the cost of inverting $\widehat{S}_{n,\delta}$ on demand is the only piece that needs to be optimized estimator-by-estimator.

The step size that drives the Fisher recursion is $\omega_n = c_\theta/(n + n_0)^{\beta_\theta}$ with $c_\theta = 1$, $n_0 = 1{,}000$, and $\beta_\theta = 0.6$ throughout. For the Bayesian logistic-regression and Gaussian targets, we freeze the Fisher state after the first iteration (`freeze_after_L=1`) because the Hessian is approximately constant near the MAP.

D.2.1 Diagonal Fisher: $O(d)$ memory and per-step cost

The diagonal estimator below is the diagonal restriction of the FisherMALA recursion of Titsias [54] and is the same diagonal-preconditioning style discussed in the broader survey of Hird and Livingstone [30, Section 3.5]: we adopt it verbatim and reuse it on the GA-SRMC repulsion side M_n in addition to the FisherMALA-style proposal side A_n. Concretely, this estimator stores only the per-coordinate squared-score average $\widehat{r}_n \in \mathbb{R}^d$ and updates it element-wise as $\widehat{r}_{n+1} = (1-\omega_{n+1})\widehat{r}_n + \omega_{n+1}\, s_n \odot s_n$, where $\odot$ is the Hadamard (elementwise) product. The quantity that this estimator preserves is the diagonal of $\widehat{S}_n$, which equals $\mathbb{E}_\pi[s \odot s]$ and, by Stein's identity, the diagonal of the Fisher information whenever $\mathbb{E}_\pi[s] = 0$. The regularized inverse is applied element-wise as $\widehat{S}_{n,\delta}^{-1} v = v \oslash (\widehat{r}_n + \delta \mathbf{1})$ at cost $O(d)$, where $\oslash$ denotes Hadamard (elementwise) division, i.e., $(a \oslash b)_i = a_i/b_i$, the natural pair of $\odot$; and the trace and inverse trace are sums over $\widehat{r}_n + \delta$ in $O(d)$. The proposal-noise factor $A_n^{1/2}\xi$ is similarly diagonal, so no factorization is required at any stage, and the entire estimator runs in linear memory and linear per-step time.

This estimator is most useful when the posterior is approximately axis-aligned, a regime characterized in detail by Hird and Livingstone [30, Section 3.5]. The ablation in Appendix A.6 confirms this on the full 8-target benchmark and reveals a three-way split by sampler family. On the axis-aligned Gauss-2D target, GA-SRMC remains the within-family winner under both diagonal MALA and diagonal FisherMALA, with the GA-SRMC vs SRMC ratio shrunk from $4.4\times$ (full) to $2.4\times$ (diagonal) on MALA, but the ranking is preserved. On the BLR posteriors, the diagonal-FisherMALA columns retain GA-SRMC as the within-family winner on five of seven rows (Australian, Heart, Pima, Ripley, Gauss-2D), with German and Caravan/MNIST flipping to SRMC by 0–5%, because the proposal-side preconditioner $A_n = (d/\,\mathrm{tr}\,\widehat{S}_{n,\delta}^{-1})\widehat{S}_{n,\delta}^{-1}$ is itself nearly diagonal on these targets and the geometry-aware mechanism survives the same approximation. The diagonal-MALA columns, however, flip GA-SRMC to a 1.4–$2.4\times$ loss against SRMC on Australian, Heart, Pima, Ripley, and Caravan: when both the proposal kernel is Euclidean and M_n is coordinate-wise, but the dominant directions of the BLR Gram matrix $\Phi^\top\Phi$ in the Hessian (19) of Appendix A.2 are not coordinate-aligned, GA-SRMC's metric-weighted tilt amplifies the wrong directions and the geometry-blind SRMC choice $M = I_d$ becomes uniformly safer, an effect predicted by Hird and Livingstone [30, Theorem 3.7] for highly correlated targets. The diagonal preconditioning style is also the default in mainstream automatic-differentiation samplers [1, 15], so a diagonal-Fisher GA-SRMC fits naturally into existing tooling whenever the target (or the proposal-side preconditioner) is approximately axis-aligned.

D.2.2 Full Fisher with Cholesky-of-inverse maintenance: $O(d^2)$ per step amortized

The recursion on $\widehat{S}_n$ in (81) below, the Sherman–Morrison rank-one update for the inverse in (82), and the use of the Cholesky factor of $\widehat{S}_{n,\delta}^{-1}$ to apply both $\widehat{S}_{n,\delta}^{-1}$ and the proposal noise factor $\widehat{S}_{n,\delta}^{-1/2}$ all originate in the FisherMALA implementation of Titsias [54, Section 4] and are also covered as the canonical online full-Fisher template in Hird and Livingstone [30, Sections 3.3–3.4]. Our only deviation is the periodic-refresh strategy for the Cholesky factor, which we adopt to handle the regularization shift $\delta > 0$ in $\widehat{S}_{n,\delta}$ that makes the direct Sherman–Morrison maintenance of Titsias [54] no longer exact (the precise reason is given after (82) below). With this single modification, the full estimator stores the recursion Fisher $\widehat{S}_n \in \mathbb{R}^{d\times d}$ and a lower-triangular Cholesky factor L_n satisfying $L_n L_n^\top = \widehat{S}_{n,\delta}^{-1} = (\widehat{S}_n + \delta I_d)^{-1}$. The rank-one recursion

$$\widehat{S}_{n+1} = (1-\omega_{n+1})\,\widehat{S}_n + \omega_{n+1}\, s_n s_n^\top \tag{81}$$

is the same recursion used by Titsias [54, Eq. (11)] and runs at every step at cost $O(d^2)$. We do not attempt to maintain L_n incrementally; instead, we recompute it from scratch every K iterations (default $K = 500$) by Cholesky-decomposing $\widehat{S}_{n,\delta}$, at a cost of $O(d^3)$ per refresh and $O(d^3/K)$ amortized per step.

The reason for the periodic-refresh strategy, rather than the more obvious Woodbury identity of $\widehat{S}_{n,\delta}^{-1}$ used in the unregularized FisherMALA recursion of Titsias [54], is structural. By the Woodbury matrix identity [14, Section A.5.4], with $a=1-\omega_{n+1}$ and $b=\omega_{n+1}$, we have

$$\widehat{S}_{n+1,\delta}^{-1} = \frac{1}{a}\,\widehat{S}_{n,\delta}^{-1} - \frac{b/a^2\,\widehat{S}_{n,\delta}^{-1} s_n s_n^\top \widehat{S}_{n,\delta}^{-1}}{1+(b/a)\,s_n^\top \widehat{S}_{n,\delta}^{-1} s_n}, \tag{82}$$

which is the exact identity that Titsias [54] uses to maintain their Cholesky-of-inverse incrementally; it holds for the recursion on $\widehat{S}_n$ only in the unregularized case $\delta = 0$ analyzed by them. When $\delta > 0$, the recursion shrinks the data-dependent part of $\widehat{S}_{n,\delta} = \widehat{S}_n + \delta I$ by $(1-\omega)$ at every step while leaving the additive shift fixed, which is a rank-d perturbation of $\widehat{S}_{n,\delta}$ rather than a rank-one one, and (82) is no longer exact. Decoupling the two updates keeps the per-step cost $O(d^2)$ for the recursion on $\widehat{S}_n$ and confines the cubic cost to a periodic refresh; on $d \le 100$ the refresh cadence is invisible against the per-iteration HVP and proposal costs, and on larger d, one can reduce K in proportion to the mixing time.

The downstream uses of L_n are again standard: the action of $\widehat{S}_{n,\delta}^{-1}$ on a vector involves two batched triangular solves, $\widehat{S}_{n,\delta}^{-1} v = L_n(L_n^\top v)$ at a cost of $O(d^2)$, and the proposal-noise factor reuses the same factor as $\widehat{S}_{n,\delta}^{-1/2}\xi = L_n\xi$. Both are exactly the operations executed in the FisherMALA implementation of Titsias [54, Algorithm 1] on the proposal side, and we reuse the same L_n on the GA-SRMC repulsion side at no extra factorization cost. The trace of the inverse is the Frobenius norm squared, $\mathrm{tr}(\widehat{S}_{n,\delta}^{-1}) = \|L_n\|_F^2$, which is read off the maintained L_n in $O(d^2)$, and the trace of $\widehat{S}_{n,\delta}$ itself is $\mathrm{tr}(\widehat{S}_n) + d\delta$, available in $O(d)$ from the diagonal of $\widehat{S}_n$. The proposal preconditioner A_n and the repulsion-side metric M_n both depend on $\widehat{S}_{n,\delta}^{-1}$ through this same L_n, with only scalar multiplications added.

The full estimator is the closest practical realization of the optimal $M_{\mathrm{opt}} \propto S^{-1}$ identified in Theorem 3.1, and we use it whenever the d^2 memory is acceptable. In our simulation, with d up to 785, this is feasible everywhere, and all main-body Table 1 numbers are produced with this estimator.

Table 7 summarizes the per-iteration costs of the estimators. The diagonal estimator updates and inverts in $O(d)$; the full estimator updates $\widehat{S}_n$ in $O(d^2)$ and refreshes its Cholesky-of-inverse every K steps for an amortized $O(d^3/K)$ overhead, which strictly improves on the $O(d^2)$ Sherman–Morrison maintenance of Titsias [54] only in constants but is what allows us to keep the regularization shift $\delta = 0.01$ exact. The regularization shift contributes no additional asymptotic cost. On our experimental scales ($d \le 785$), the per-step cost of the full estimator is dominated by the Hessian-vector product and the proposal evaluation themselves; at substantially larger dimensions, the diagonal or low-rank-r estimator becomes preferable, with the caveats noted above.

Table 7: Per-iteration cost of the two online Fisher estimators. Both apply the regularization shift $\widehat{S}_{n,\delta} = \widehat{S}_n + \delta I$ before inversion at no additional asymptotic cost. The full-Fisher implementation maintains the Cholesky-of-inverse $L_n L_n^\top = \widehat{S}_{n,\delta}^{-1}$ via periodic Cholesky every K steps (default $K = 500$), so the $O(d^3)$ refresh is amortized to $O(d^3/K)$ per step.

Estimator	Memory	FISHERUPDATE	$\widehat{S}_{n,\delta}^{-1} v$ apply	Trace & trace-inv
Diagonal	$O(d)$	$O(d)$	$O(d)$	$O(d)$
Full (Cholesky-inv)	$O(d^2)$	$O(d^2) + O(d^3/K)^*$	$O(d^2)$	$O(d^2)$ for $\mathrm{tr}(\widehat{S}_{n,\delta}^{-1})$, $O(d)$ for $\mathrm{tr}(\widehat{S}_{n,\delta})$

* Amortized cost: the rank-one recursion on $\widehat{S}_n$ is $O(d^2)$, plus a full Cholesky $O(d^3)$ every K iterations; `K_reortho`$= 500$ in production.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the contribution list at the end of Section 1 state three claims that exactly match the paper's content.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 explicitly discusses scope and limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All formal results are numbered and accompanied by complete proofs in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the exact simulation setup and results in Section 4 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: To preserve double-blind anonymity at submission time, the implementation is not released with this version.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the exact simulation setup and results in Section 4 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Every numeric ESS reported in Table 1, Table 4, and Table 5 is given as per-chain mean$\pm$standard deviation across 40 independent replicate Markov chains; the markers in Figure 2 are the same means with one-standard-deviation error bars across the same 40 chains. The factor of variability is the chain-to-chain randomness (independent random initial states, independent proposal noise, and independent stochastic-approximation paths). Section 4 states the protocol is the single-chain replication protocol of Titsias [54].

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper characterizes the computational cost of each algorithmic component analytically — $O(d)$ per step for the diagonal Fisher estimator and $O(d^2)$ per step for the full estimator with an amortized $O(d^3/K)$ Cholesky-of-inverse refresh every $K = 500$ iterations (Appendix D.2, Table 7), and notes that GA-SRMC adds exactly one Hessian-vector product per iteration over a standard preconditioned MALA step (Section 3). However, we do not report wall-clock execution times or the specific CPU/GPU hardware used to produce Tables 1–5, since the 40-chain experiments fit comfortably on a standard CPU workstation and timing comparisons are not the contribution. We will add a wall-clock summary in a future revision.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The paper develops a generic MCMC sampling algorithm and evaluates it on standard public benchmark posteriors. It involves no human subjects, no personal data, and no models with high misuse potential. The work conforms to the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The contribution is a foundational MCMC sampler refinement (a Riemannian-metric extension of a recent score-repellent surrogate), evaluated on standard Bayesian-inference benchmarks. It is generic Bayesian-inference infrastructure rather than an application-specific deployment, with no direct path to either positive or negative societal applications beyond what already follows from improved MCMC sampling in scientific computing.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper releases no data, no models, no scraped content, and no high-risk artifacts.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The seven BLR benchmark datasets (RIPLEY, PIMA, HEART, AUSTRALIAN, GERMAN, CARAVAN, MNIST) are standard public benchmarks. All third-party software dependencies used to produce the reported numbers are credited at the point of use: ArviZ [36], TensorFlow [1] and Stan [15].

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper introduces no new datasets, no new pretrained models, and no released code at submission time (the implementation will be released upon acceptance, as noted under "Open access to data and code" above).

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper involves no crowdsourcing and no human-subject research; all experiments are simulations of MCMC chains on fixed analytic or standard public-benchmark posteriors.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper involves no human subjects, so no IRB approval is applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Large language models are not part of the methodology of this paper. LLM is only used to polish the paper writing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.