

**BAYESIAN PARAMETER INFERENCE FOR STOCHASTIC
DIFFERENTIAL EQUATIONS FROM DISCRETE OBSERVATIONS:
CONVERGENCE AND BIAS ANALYSIS OF AN EULER–MARUYAMA
METROPOLIS–HASTINGS SCHEME**

Dharmendra Kumar Upadhyay

NET Qualified

Email: dharmendraupadhyaysir@gmail.com

Abstract

Parameters of a diffusion are usually learned from observations recorded at discrete times, while the transition density of the process is unknown in closed form. A common remedy replaces this density by the Gaussian transition of the Euler–Maruyama scheme and samples the resulting approximate posterior with a Metropolis–Hastings algorithm. This paper analyses the errors that such a scheme makes. For the stochastic differential equation $dX_t = b(X_t; \theta) dt + \sigma dW_t$ observed at spacing h over a fixed horizon, we prove that the random-walk Metropolis–Hastings chain targeting the Euler–Maruyama posterior is geometrically ergodic under a drift and minorization condition on the log-posterior, and we give an explicit bound on the convergence rate in total variation. We then bound the distance between the Euler–Maruyama posterior and the posterior based on the exact transition density by a constant multiple of h in the Wasserstein-1 distance, which yields an $\mathcal{O}(h)$ bias for the posterior mean. Combining the two results with the Monte Carlo error, expressed through the effective sample size, gives a bound on the total error in terms of the step size, the chain length and the burn-in. An adaptive version of the sampler, whose proposal covariance is updated from the chain history with diminishing adaptation, is proposed and shown to remain ergodic. For the Ornstein–Uhlenbeck process the Euler–Maruyama posterior under a conjugate prior and the exact-likelihood posterior can both be computed independently of the sampler, so the sampler and the bias bound are tested against reference values on simulated paths. The tests are repeated for a nonlinear drift, for which no closed form exists. The results give practical guidance on how to choose the step size and the chain length for a target accuracy.

Keywords: Stochastic differential equations; Bayesian inference; Euler–Maruyama scheme; Metropolis–Hastings algorithm; geometric ergodicity; Wasserstein distance; adaptive MCMC.

MSC 2020: 60H10, 60H35, 62F15, 65C05, 65C40, 60J22.

1. Introduction

Stochastic differential equations (SDEs) describe systems driven by noise, from interest rates and neuron voltages to molecular dynamics. Their parameters must be learned from observations recorded at discrete times, and the Bayesian approach returns a full posterior distribution together with a point estimate. Consider

$$dX_t = b(X_t; \theta) dt + \sigma dW_t, \quad t \in [0, T], \quad \theta \in \Theta \subseteq \mathbb{R}^d, \quad (1)$$

where b is the drift, $\sigma > 0$ is known and W is a Wiener process. We observe $x_i = X_{t_i}$ at the times $t_i = ih$, $i = 0, \dots, n$, with $nh = T$ fixed, so that a smaller spacing h gives more data on the same horizon [4, 2]. Let $p_h(y | x; \theta)$ be the transition density of (1) over a time step h . With a prior π_0 , the posterior is

$$\pi(\theta | x) \propto \pi_0(\theta) \prod_{i=0}^{n-1} p_h(x_{i+1} | x_i; \theta). \quad (2)$$

Except for a few linear models, p_h is not available in closed form. The Euler–Maruyama scheme [1], $X_{t+h} \approx x + b(x; \theta)h + \sigma\sqrt{h}\xi$ with $\xi \sim \mathcal{N}(0, 1)$, replaces it by the Gaussian density

$$q_h(y | x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2h}} \exp \left\{ -\frac{(y - x - b(x; \theta)h)^2}{2\sigma^2h} \right\}, \quad (3)$$

which gives the log-likelihood and the approximate posterior

$$\ell_h(\theta) = -\frac{1}{2} \sum_{i=0}^{n-1} \left[\log(2\pi\sigma^2h) + \frac{(x_{i+1} - x_i - b(x_i; \theta)h)^2}{\sigma^2h} \right], \quad \pi_h(\theta | x) \propto \pi_0(\theta) e^{\ell_h(\theta)}. \quad (4)$$

The posterior π_h is sampled by Markov chain Monte Carlo, so two errors arise: the *bias* between π_h and π , caused by the discretization, and the *sampling error* of a finite chain.

1.1 A Worked Example: Two Scales of Error

For the Ornstein–Uhlenbeck process $b(x; \theta) = -\theta x$ the exact transition is Gaussian,

$$X_{t+h} | X_t = x \sim \mathcal{N}(ax, s^2), \quad a = e^{-\theta h}, \quad s^2 = \sigma^2 \frac{1 - e^{-2\theta h}}{2\theta}, \quad (5)$$

while (3) uses mean $(1 - \theta h)x$ and variance σ^2h . The local errors are

$$a - (1 - \theta h) = \frac{\theta^2 h^2}{2} + \mathcal{O}(h^3), \quad s^2 - \sigma^2 h = -\sigma^2 \theta h^2 + \mathcal{O}(h^3), \quad (6)$$

which is of second order in h for one step. To compare the two posteriors given the *same* data, let $r_i = x_{i+1} - (1 - \theta h)x_i$ and $\Delta_h(\theta) = \ell_h(\theta) - \log \prod_i p_h(x_{i+1} | x_i; \theta)$. Expanding the logarithm of (5) and the quadratic form in h gives

$$\Delta_h(\theta) = \frac{\theta h}{2} \sum_{i=0}^{n-1} \left(\frac{r_i^2}{\sigma^2 h} - 1 \right) - \frac{\theta^2 h}{2\sigma^2} \sum_{i=0}^{n-1} x_i r_i + \mathcal{O}_P(h). \quad (7)$$

If the data come from the parameter θ_0 , then $r_i = \sigma\sqrt{h}\xi_i + (\theta - \theta_0)h x_i + \mathcal{O}(h^2)$ with ξ_i i.i.d. standard normal, and the first sum in (7) is dominated by the data-dependent fluctuation $\sum_i(\xi_i^2 - 1) \approx \sqrt{2n} Z$, so that

$$\Delta_h(\theta) = \theta \sqrt{\frac{Th}{2}} Z_h + \mathcal{O}_P(h), \quad Z_h = \frac{1}{\sqrt{2n}} \sum_{i=0}^{n-1} (\xi_i^2 - 1) \Rightarrow \mathcal{N}(0, 1). \quad (8)$$

Thus π_h is the exact posterior multiplied by the exponential tilt $e^{\theta\kappa_h}$ with $\kappa_h = \sqrt{Th/2} Z_h$, up to an $\mathcal{O}(h)$ correction. If V is the posterior variance, the posterior mean moves by about $V\kappa_h$. Two error scales follow.

$$\underbrace{|m_h(x) - m(x)|}_{\text{fixed data}} = \mathcal{O}_P(\sqrt{h}), \quad \underbrace{|\mathbb{E}_x[m_h(x) - m(x)]|}_{\text{averaged over data}} = \mathcal{O}(h), \quad (9)$$

where m and m_h are the means of π and π_h . For a given data set the discretization error is therefore of the larger order $\sqrt{h}$, with a random sign, and it is only its average over data sets that has the order h of the weak error of the Euler–Maruyama scheme [3, 1]. A preliminary simulation ($\theta_0 = \sigma = 1$, $T = 2$, 300 paths, flat prior) is consistent with (9): for h between 10^{-3} and 3×10^{-2} the root-mean-square difference of the two posterior means was between $1.4\sqrt{h}$ and $1.7\sqrt{h}$, while their average difference over the paths was of order h (about -1 to -2 times h).

1.2 Sampling by Metropolis–Hastings and Ergodicity

The posterior π_h is sampled by the random-walk Metropolis algorithm [6, 7]: from θ_k propose $\theta' = \theta_k + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \Gamma)$, and accept with probability

$$\alpha(\theta_k, \theta') = \min \left\{ 1, \frac{\pi_0(\theta') e^{\ell_h(\theta')}}{\pi_0(\theta_k) e^{\ell_h(\theta_k)}} \right\}, \quad (10)$$

otherwise $\theta_{k+1} = \theta_k$. Let P be the transition kernel. The chain is *geometrically ergodic* if there are $\rho < 1$ and $M : \Theta \rightarrow [0, \infty)$ with

$$\|P^k(\theta, \cdot) - \pi_h\|_{\text{TV}} \leq M(\theta) \rho^k, \quad k \geq 1, \theta \in \Theta. \quad (11)$$

The standard route is a Lyapunov function $V \geq 1$ and a small set C with

$$PV(\theta) \leq \lambda V(\theta) + b \mathbf{1}_C(\theta), \quad \lambda < 1, b < \infty, \quad (12)$$

together with a minorization $P(\theta, \cdot) \geq \delta \nu(\cdot)$ on C [9, 8]. For the Ornstein–Uhlenbeck model, π_h is Gaussian when σ is known and the prior is Gaussian, and the conditions hold. For general drifts, the constants λ, b, δ depend on h, n and the tails of ℓ_h , and no explicit dependence has been worked out for the Euler–Maruyama posterior.

1.3 Total Error and Adaptive Proposals

For a function f , the chain average after B burn-in steps is $\hat{f}_N = (N - B)^{-1} \sum_{k=B+1}^N f(\theta_k)$. Since $(u + v)^2 \leq 2u^2 + 2v^2$,

$$\mathbb{E}[(\hat{f}_N - \pi f)^2] \leq 2(\pi_h f - \pi f)^2 + 2\mathbb{E}[(\hat{f}_N - \pi_h f)^2], \quad (13)$$

where the first term is the squared bias of Section 1.2 and, for a chain started near stationarity, the second is approximately $\text{Var}_{\pi_h}(f) \tau_f / (N - B)$ with

$$\tau_f = 1 + 2 \sum_{j \geq 1} \varrho_j(f), \quad \text{ESS} = \frac{N - B}{\tau_f}, \quad (14)$$

and $\varrho_j(f)$ the lag- j autocorrelation. A good proposal covariance Γ is the key to a small τ_f , and the adaptive Metropolis algorithm [10] sets

$$\Gamma_k = \frac{2.4^2}{d} \left(\text{Cov}(\theta_0, \dots, \theta_{k-1}) + \epsilon I_d \right), \quad k > k_0, \quad (15)$$

for which the change $\|\Gamma_{k+1} - \Gamma_k\| = \mathcal{O}(1/k)$ is diminishing. Alternatively, the transition density can be recovered by imputing the path between observations, but then the mixing of the sampler for the diffusion parameters degrades as the number of imputed points grows [5], which is a reason to study the direct Euler–Maruyama posterior.

1.4 Research Gaps and Objectives

The pieces above are known separately: the weak error of the Euler scheme [3], frequentist estimators from discrete data [4], the Metropolis–Hastings algorithm and its ergodicity [8, 9] and adaptive proposals [10]. What is missing is (i) a bound on the distance between π_h and π that holds for fixed data and states its order, since the weak error concerns averages over paths and (9) shows that the two behave differently; (ii) geometric ergodicity for the Euler–Maruyama posterior with a rate that can be computed; and (iii) a total-error bound that turns (13) into a rule for choosing h, N and B . The **key objectives** of this paper are the following.

1. To prove geometric ergodicity (11) of the random-walk Metropolis chain targeting π_h , by verifying (12) and a minorization condition under stated assumptions on the drift, and to give a computable bound on ρ .
2. To bound the bias between π_h and π on two scales: for fixed data, $W_1(\pi_h, \pi) \leq C(x)\sqrt{h}$, with the order shown to be attained in the Ornstein–Uhlenbeck case through (8), and, averaged over the data, an $\mathcal{O}(h)$ bound on the bias of the posterior mean.
3. To combine the two results with the Monte Carlo error in (13) into a total-error bound in terms of h , N , B and the effective sample size, to check that the adaptive proposal (15) preserves ergodicity, and to translate the bound into a rule for choosing h and N for a target accuracy.
4. To validate the results numerically on the Ornstein–Uhlenbeck process, for which the two posteriors can be computed by quadrature independently of the sampler, and on a nonlinear drift, and to confirm the two orders in (9).

Table 1 shows which of these features each of the closest works treats.

Table 1 Features treated by the closest works and by the present paper (✓: treated; ×: not treated).

Work	Euler error analysed	Inference from discrete data	MCMC sampler analysed	Geometric ergodicity proved	Posterior bias bound in h	Total error (bias + Monte Carlo)	Adaptive proposal
Kloeden and Platen [1]	✓	×	×	×	×	×	×
Talay and Tubaro [3]	✓	×	×	×	×	×	×
Kessler [4]	×	✓	×	×	×	×	×
Roberts and Stramer [5]	×	✓	✓	×	×	×	×
Roberts and Tweedie [8]	×	×	✓	✓	×	×	×
Haario et al. [10]	×	×	✓	×	×	×	✓
This paper	✓	✓	✓	✓	✓	✓	✓

Each of the cited works supplies one or two ingredients: the discretization error of the scheme, an estimator from discrete data, or the behaviour of a sampler. The present paper puts them in one analysis of the Euler–Maruyama posterior, from its bias (on two scales) through the ergodicity of the sampler to a total error that can be used to choose h and N .

The rest of the paper is organized as follows. Section 2 states the assumptions and auxiliary results, Section 3 proves the ergodicity and bias results and presents the algorithm, Section 4 gives the total-error bound and treats the adaptive sampler, Section 5 reports the numerical experiments and Section 6 concludes.

1.5 Assumptions, Distances and Auxiliary Results

Throughout, $\Theta \subseteq \mathbb{R}^d$ is open, $\sigma > 0$ is known, $h \leq 1$ and $nh = T$.

(A1) For every θ in a compact set $\Theta_0 \subseteq \Theta$, $b(\cdot; \theta)$ is Lipschitz, $|b(x; \theta) - b(y; \theta)| \leq L|x - y|$, with linear growth $|b(x; \theta)|^2 \leq K_b^2(1 + |x|^2)$.

(A2) X_0 is independent of W with $\mathbb{E}|X_0|^2 < \infty$.

(A3) The prior π_0 is continuous and strictly positive on Θ .

[Distances] For probability measures μ, ν on $\mathbb{R}^d$,

$$\|\mu - \nu\|_{\text{TV}} = \sup_A |\mu(A) - \nu(A)|, \quad W_1(\mu, \nu) = \sup_{\text{Lip}(f) \leq 1} \left| \int f d\mu - \int f d\nu \right|, \quad (16)$$

and, for $d = 1$ with distribution functions F_μ, F_ν , $W_1(\mu, \nu) = \int_{\mathbb{R}} |F_\mu(\theta) - F_\nu(\theta)| d\theta$. The Kullback–Leibler divergence is $\text{KL}(\mu \| \nu) = \int \log \frac{d\mu}{d\nu} d\mu$, and Pinsker's inequality gives $\|\mu - \nu\|_{\text{TV}} \leq \sqrt{\text{KL}(\mu \| \nu)/2}$.

[Local error of the Euler step] Under (A1), let X_{t+h} solve (1) from $X_t = x$ and set

$$R_h(x) = X_{t+h} - [x + b(x; \theta)h + \sigma(W_{t+h} - W_t)] = \int_t^{t+h} (b(X_s; \theta) - b(x; \theta)) ds. \quad (17)$$

Then there is a constant K depending on L, K_b, σ and T such that

$$\mathbb{E}|R_h(x)|^2 \leq K h^3 (1 + |x|^2), \quad W_2(p_h(\cdot | x; \theta), q_h(\cdot | x; \theta)) \leq \sqrt{K} h^{3/2} \sqrt{1 + |x|^2}. \quad (18)$$

Proof. By the moment bounds for (1) [1], $\mathbb{E}|X_s - x|^2 \leq C(s - t)(1 + |x|^2)$ for $s - t \leq h \leq 1$. By the Cauchy–Schwarz inequality and the Lipschitz condition,

$$\mathbb{E}|R_h(x)|^2 \leq h \int_t^{t+h} L^2 \mathbb{E}|X_s - x|^2 ds \leq L^2 C(1 + |x|^2) h \int_0^h u du = \frac{1}{2} L^2 C(1 + |x|^2) h^3. \quad (19)$$

The law of $x + bh + \sigma(W_{t+h} - W_t)$ is $q_h(\cdot | x; \theta)$, so the coupling through the same Wiener increment gives the bound on W_2 . $\square$

[Kullback–Leibler divergence between transitions] For Gaussians,

$$\text{KL}(\mathcal{N}(m_1, v_1) \| \mathcal{N}(m_2, v_2)) = \frac{1}{2} \left[\frac{v_1}{v_2} - 1 - \log \frac{v_1}{v_2} + \frac{(m_1 - m_2)^2}{v_2} \right]. \quad (20)$$

For the Ornstein–Uhlenbeck transitions (5) and the Euler kernel, with $m_1 - m_2 = (e^{-\theta h} - 1 + \theta h)x$, $v_1 = s^2$, $v_2 = \sigma^2 h$,

$$\text{KL}_h(x; \theta) = \frac{\theta^2 h^2}{4} + \frac{\theta^4 h^3 x^2}{8\sigma^2} + \mathcal{O}(h^3), \quad (21)$$

and, for the observations $x_0, \dots, x_n$ of a stationary path, the chain rule for KL gives

$$\text{KL}(P_\theta^{\text{exact}} \| P_\theta^{\text{EM}}) = \sum_{i=0}^{n-1} \mathbb{E} \text{KL}_h(x_i; \theta) = \frac{\theta^2 Th}{4} + \mathcal{O}(h^2), \quad \|P_\theta^{\text{exact}} - P_\theta^{\text{EM}}\|_{\text{TV}} \leq \theta \sqrt{\frac{Th}{8}} + \mathcal{O}(h). \quad (22)$$

Proof. Formula (20) is a direct computation. With $\rho = v_1/v_2 = 1 - \theta h + \frac{2}{3}\theta^2 h^2 + \mathcal{O}(h^3)$ and $u = \rho - 1$, we have $\rho - 1 - \log \rho = \frac{1}{2}u^2 + \mathcal{O}(u^3) = \frac{1}{2}\theta^2 h^2 + \mathcal{O}(h^3)$. The mean term is $(\frac{1}{2}\theta^2 h^2 x)^2 / (\sigma^2 h) + \mathcal{O}(h^4)$, which gives (21). For a stationary path, $\mathbb{E}x_i^2 = \sigma^2 / (2\theta)$, so the mean term contributes $\mathcal{O}(h^3)$ per step and $\mathcal{O}(h^2)$ in total. Pinsker's inequality gives the last bound. $\square$

The total-variation distance between the laws of the data under the two models is thus of order $\sqrt{h}$, which is the scale of the fixed-data error in (9).

[Perturbation of a posterior] Let π be a probability density and $\pi_h \propto \pi e^\Delta$ with Δ measurable. Then for $f \in L^2(\pi)$,

$$\pi_h f - \pi f = \frac{\text{Cov}_\pi(f, e^\Delta)}{\mathbb{E}_\pi e^\Delta}. \quad (23)$$

If $\delta = \frac{1}{2}(\text{ess sup } \Delta - \text{ess inf } \Delta) < \infty$, then

$$\|\pi_h - \pi\|_{\text{TV}} \leq \frac{1}{2}(e^{2\delta} - 1), \quad |\pi_h f - \pi f| \leq \sqrt{\text{Var}_\pi(f)} \frac{e^{2\delta} - 1}{2}. \quad (24)$$

Proof. Write $\pi_h f = \mathbb{E}_\pi[f e^\Delta] / \mathbb{E}_\pi[e^\Delta]$ and subtract πf to get (23). A constant added to Δ does not change π_h , so we may take $|\Delta| \leq \delta$. Then $e^\Delta \in [e^{-\delta}, e^\delta]$, $\mathbb{E}_\pi e^\Delta \geq e^{-\delta}$, and by Popoviciu's inequality $\text{Var}_\pi(e^\Delta) \leq \sinh^2 \delta$. The Cauchy-Schwarz inequality gives $|\pi_h f - \pi f| \leq \sqrt{\text{Var}_\pi f} e^\delta \sinh \delta = \sqrt{\text{Var}_\pi f} (e^{2\delta} - 1)/2$. For the total variation, $d\pi_h/d\pi = e^\Delta / \mathbb{E}_\pi e^\Delta \in [e^{-2\delta}, e^{2\delta}]$, so $\|\pi_h - \pi\|_{\text{TV}} \leq \frac{1}{2} \sup |d\pi_h/d\pi - 1| \leq \frac{1}{2}(e^{2\delta} - 1)$. $\square$

[Exponential tilt of a Gaussian] If $\pi = \mathcal{N}(m, \Sigma)$ and $\pi_h \propto \pi e^{\kappa^\top \theta}$, then $\pi_h = \mathcal{N}(m + \Sigma \kappa, \Sigma)$, and

$$W_1(\pi_h, \pi) = \|\Sigma \kappa\|_2, \quad \|\pi_h - \pi\|_{\text{TV}} = 2\Phi\left(\frac{1}{2}\sqrt{\kappa^\top \Sigma \kappa}\right) - 1. \quad (25)$$

For $d = 1$, $\Sigma = V$ and $\kappa = \kappa_h = \sqrt{Th/2} Z_h$ from (8),

$$\mathbb{E} W_1(\pi_h, \pi) = V \sqrt{\frac{Th}{2}} \mathbb{E}|Z_h| \longrightarrow V \sqrt{\frac{Th}{\pi}} \quad (h \rightarrow 0), \quad (26)$$

because $\mathbb{E}|Z| = \sqrt{2/\pi}$ for $Z \sim \mathcal{N}(0, 1)$.

Proof. Completing the square in $-\frac{1}{2}\theta^\top \Sigma^{-1} \theta + \theta^\top (\Sigma^{-1} m + \kappa)$ gives the new mean, and the covariance is unchanged. A translation by s has $W_1 = \|s\|_2$: the coupling by translation gives

$\leq$ and the coordinate functions give $\geq$. The total variation between $\mathcal{N}(0, \Sigma)$ and $\mathcal{N}(s, \Sigma)$ is $2\Phi(\|\Sigma^{-1/2}s\|/2) - 1$. $\square$

[Conjugate Euler posterior] Let $b(x; \theta) = \varphi(x)^\top \theta$ with $\varphi : \mathbb{R} \rightarrow \mathbb{R}^d$ (for the Ornstein–Uhlenbeck process, $d = 1$, $\varphi(x) = -x$) and let the prior be $\mathcal{N}(\mu_0, S_0)$. Then the Euler posterior is $\pi_h = \mathcal{N}(\mu_n, \Sigma_n)$ with

$$\Sigma_n^{-1} = S_0^{-1} + \frac{h}{\sigma^2} G, \quad \mu_n = \Sigma_n \left(S_0^{-1} \mu_0 + g \right), \quad G = \sum_{i=0}^{n-1} \varphi(x_i) \varphi(x_i)^\top, \quad g = \frac{1}{\sigma^2} \sum_{i=0}^{n-1} \varphi(x_i) (x_{i+1} - x_i). \quad (27)$$

Moreover $hG \rightarrow \int_0^T \varphi(X_t) \varphi(X_t)^\top dt$ as $h \rightarrow 0$, so Σ_n stays bounded away from 0 for fixed T , and the posterior does not concentrate as $h \rightarrow 0$.

Proof. Expanding $(x_{i+1} - x_i - h\varphi_i^\top \theta)^2 / (2\sigma^2 h)$ in (4) gives $\ell_h(\theta) = \text{const} + \theta^\top g - \frac{h}{2\sigma^2} \theta^\top G \theta$. Adding the Gaussian log-prior and completing the square gives (27). The limit of hG is a Riemann sum. $\square$

[Small sets for random-walk Metropolis] Let π_h be continuous and positive on $\Theta = \mathbb{R}^d$ and let the proposal be $\mathcal{N}(0, \Gamma)$. For $R > 0$ let B_R be the ball of radius R and put

$$q_R = (2\pi)^{-d/2} |\Gamma|^{-1/2} \exp \left(-\frac{2R^2}{\lambda_{\min}(\Gamma)} \right), \quad \alpha_R = \min \left\{ 1, \frac{\min_{B_R} \pi_h}{\max_{B_R} \pi_h} \right\} > 0, \quad (28)$$

and $\delta_R = q_R \alpha_R \text{Leb}(B_R)$, with $\text{Leb}(B_R) = \pi^{d/2} R^d / \Gamma(\frac{d}{2} + 1)$. Then B_R is a small set:

$$P(\theta, A) \geq \delta_R \nu_R(A) \quad \text{for all } \theta \in B_R, A \subseteq \mathbb{R}^d, \quad (29)$$

where ν_R is the uniform distribution on B_R .

Proof. For $\theta, \theta' \in B_R$ we have $|\theta' - \theta| \leq 2R$, so the proposal density is at least q_R , and the acceptance probability is at least α_R . Hence $P(\theta, A) \geq \int_{A \cap B_R} q_R \alpha_R d\theta' = \delta_R \nu_R(A)$. $\square$

Lemma 1.5 gives the minorization constant δ_R of (12) explicitly. For the Gaussian target of Lemma 1.5, the tails are super-exponential with elliptical contours, and the sufficient conditions of [8] for geometric ergodicity of the random-walk chain hold, so that (11) is available with a computable δ_R . The drift constants λ and b are needed in addition to obtain an explicit rate ρ .

[Effective sample size for geometric autocorrelation] If the stationary autocorrelations of $f(\theta_k)$ are $\varrho_j = \rho^{|j|}$ with $|\rho| < 1$, then

$$\tau_f = 1 + 2 \sum_{j \geq 1} \rho^j = \frac{1 + \rho}{1 - \rho}, \quad \text{ESS} = (N - B) \frac{1 - \rho}{1 + \rho}, \quad (30)$$

so a chain with $\rho = 0.9$ carries only $\approx 5\%$ of the information of independent draws.

[Diminishing adaptation of the sample covariance] Let $\bar{\theta}_k = \frac{1}{k+1} \sum_{j=0}^k \theta_j$ and $C_k = \frac{1}{k} \sum_{j=0}^k (\theta_j - \bar{\theta}_k)(\theta_j - \bar{\theta}_k)^\top$. Then, with $u_{k+1} = \theta_{k+1} - \bar{\theta}_k$,

$$\bar{\theta}_{k+1} = \bar{\theta}_k + \frac{u_{k+1}}{k+2}, \quad C_{k+1} - C_k = -\frac{C_k}{k+1} + \frac{u_{k+1}u_{k+1}^\top}{k+2}, \quad (31)$$

and therefore the proposal covariance $\Gamma_k = \frac{2.4^2}{d}(C_k + \epsilon I_d)$ of (15) satisfies

$$\|\Gamma_{k+1} - \Gamma_k\| \leq \frac{2.4^2}{d} \cdot \frac{\|C_k\| + |u_{k+1}|^2}{k+1}. \quad (32)$$

Hence the adaptation is diminishing whenever the second moments of the chain stay bounded.

Proof. With $S_k = kC_k$, the Welford recursion is $S_{k+1} = S_k + \frac{k+1}{k+2}u_{k+1}u_{k+1}^\top$. Then

$$C_{k+1} - C_k = \frac{S_k}{k+1} - \frac{S_k}{k} + \frac{u_{k+1}u_{k+1}^\top}{k+2} = -\frac{C_k}{k+1} + \frac{u_{k+1}u_{k+1}^\top}{k+2}, \quad (33)$$

and (32) follows from the triangle inequality and $\frac{1}{k+2} \leq \frac{1}{k+1}$. $\square$

[Gaussian picture of the fixed-data error] By Lemma 1.5, the Euler posterior of a linear-in- θ drift is Gaussian, and by (8) the exact posterior is approximately π_h tilted by $e^{-\theta\kappa_h}$ with $\kappa_h = \sqrt{Th/2}Z_h$. Lemma 1.5 then gives $W_1(\pi_h, \pi) \approx V|\kappa_h|$, of order $\sqrt{h}$ with a random factor $|Z_h|$, whereas $\mathbb{E} \kappa_h = 0$ exactly. This is the mechanism behind the two scales in (9): the fixed-data bias is $\mathcal{O}(\sqrt{h})$ with a random sign that averages out over data sets, leaving an $\mathcal{O}(h)$ remainder.

2. Results and Discussion

2.1 Experimental Setup

All experiments use $\sigma = 1$ (known), $T = 2$ (except Section 5.6), the prior $\mathcal{N}(0.5, 4)$ and double-precision arithmetic.

Ornstein–Uhlenbeck (OU) model. Here $b(x; \theta) = -\theta x$ with $\theta_0 = 1$ and a stationary start. Paths were generated with the exact transition on a grid of width 2^{-10} and subsampled to $h = 2^{-k}$, $k = 4, \dots, 10$, giving $n = 32, \dots, 2048$ observations. The exact posterior π was computed by quadrature on 6501 points in $[-4, 9]$, using the exact transition (5), and the Euler posterior π_h is the conjugate Gaussian of Lemma 1.5. The number of simulated paths is $R = 4000$.

Nonlinear drift. Here $b(x; \theta) = -\theta x - x^3$ with $\theta_0 = 1$, for which no closed-form transition density exists. Paths ($R = 300$) were generated by an Euler scheme with step 2^{-13} after a burn-in of one time unit, and observed at $h = 2^{-k}$, $k = 3, \dots, 7$. The reference posterior uses transition densities computed from the Fokker–Planck equation, discretized by a conservative finite-volume scheme on $[-3, 3]$ with $\Delta x = 0.02$, exponentiated with a matrix exponential, and interpolated by cubic splines, on a θ -grid of 161 points in $[-2.5, 5.5]$. The reference error was measured by running the same pipeline on the OU model, where the exact answer is known: the root-mean-square error of the posterior mean was 2.2×10^{-3} at $h = 2^{-7}$ and 1.2×10^{-3} at $h = 2^{-5}$ (maximum 7×10^{-3} and 2.7×10^{-3}), which is below 2% of the bias measured below. Because the accuracy of this reference deteriorates as h decreases, $h < 2^{-7}$ was not used.

Notation. Let m and m_h be the means of π and π_h , $d = m_h - m$, and V the posterior variance of π (averaged over paths).

2.2 Two Scales of Bias: OU Process

Table 2 reports, for each h , the mean and the root-mean-square value of d over the $R = 4000$ paths, the mean Wasserstein distance $W_1(\pi_h, \pi)$, and the ratios to the leading-order predictions of Lemma 1.5: $V\sqrt{Th/2}$ for the standard deviation of d and $V\sqrt{Th/\pi}$ for $\mathbb{E} W_1$ (here $V = 1.10$).

Table 2 OU model ($R = 4000$ paths). Fixed-data error (rms of d , mean W_1) against data-averaged bias (mean of $d \pm$ standard error), and ratios to the predictions of Lemma 1.5.

h	n	mean of d ($\pm$ s.e.)	rms of d	mean W_1	$\frac{\text{rms}(d)}{V\sqrt{Th/2}}$	$\frac{\mathbb{E} W_1}{V\sqrt{Th/\pi}}$
2^{-10}	2048	−0.0002 (0.0006)	0.0403	0.0281	1.169	1.020
2^{-9}	1024	−0.0007 (0.0009)	0.0545	0.0388	1.118	0.996
2^{-8}	512	−0.0017 (0.0012)	0.0774	0.0551	1.122	1.002
2^{-7}	256	−0.0073 (0.0017)	0.1104	0.0780	1.132	1.001
2^{-6}	128	−0.0160 (0.0024)	0.1552	0.1109	1.125	1.007
2^{-5}	64	−0.0339 (0.0034)	0.2191	0.1567	1.122	1.005
2^{-4}	32	−0.0646 (0.0047)	0.3031	0.2183	1.095	0.988

The log–log slopes against h are 0.49 for $\text{rms}(d)$ and 0.50 for the mean W_1 , so that for fixed data the error is of order $\sqrt{h}$. The data-averaged bias is significant ($|\text{mean}| > 3$ s.e.) for $h \geq 2^{-7}$, where its log–log slope is 1.05 (order h) and $\text{mean}(d)/h$ lies between -0.93 and -1.08 . The predictions from the Gaussian tilt (26) are accurate: the ratio for $\mathbb{E} W_1$ is 0.99 to 1.02 for all h , and that for $\text{rms}(d)$ is 1.10 to 1.17.

2.3 Nonlinear Drift

Table 3 gives the same quantities for $b(x; \theta) = -\theta x - x^3$ against the Fokker–Planck reference ($V = 1.35$).

Table 3 Nonlinear drift ($R = 300$ paths), reference posterior from the Fokker–Planck equation.

h	n	mean of d ($\pm$ s.e.)	rms of d	mean W_1	$\frac{\text{rms}(d)}{V\sqrt{Th/2}}$	$\frac{\mathbb{E}W_1}{V\sqrt{Th/\pi}}$
2^{-7}	256	−0.015 (0.007)	0.122	0.094	1.02	0.99
2^{-6}	128	−0.026 (0.010)	0.172	0.131	1.02	0.98
2^{-5}	64	−0.051 (0.012)	0.220	0.172	0.92	0.90
2^{-4}	32	−0.113 (0.018)	0.325	0.255	0.96	0.94
2^{-3}	16	−0.214 (0.025)	0.479	0.377	0.99	0.97

The behaviour is the same as for the OU model: the log–log slopes are 0.49 for $\text{rms}(d)$ and 0.50 for $\mathbb{E}W_1$, the slope of the significant averaged bias is 1.04 (three points), and $\text{mean}(d)/h$ is between -1.6 and -1.9 . The mechanism of Section 1.2 therefore does not depend on linearity of the drift in x , and the leading constants of Lemma 1.5 are accurate to about 10%.

2.4 Ergodicity and Efficiency of the Sampler

Since the Euler posterior of the OU model is Gaussian (Lemma 1.5) and random-walk Metropolis is affine invariant, it suffices to study the standard normal target. We ran 10^5 parallel chains from $x_0 = 6$ (six posterior standard deviations from the mean) with proposal scale ς (in units of the posterior standard deviation), and measured W_1 between the empirical law of the chains after k steps and the target. In every case W_1 decays exponentially, as in (11), and Table 4 gives the fitted factor $\hat{\rho}$ per step, computed on the range $0.05 < W_1 < 1$, together with the acceptance rate and the efficiency of a stationary chain.

Table 4 Random-walk Metropolis on the standardized Gaussian target. $\hat{\rho}$: fitted per-step decay factor of W_1 ; τ_{int} : integrated autocorrelation time; $\text{ESS}/N = 1/\tau_{\text{int}}$; last column: the value $(1 - \varrho_1)/(1 + \varrho_1)$ of Lemma 1.5 computed from the lag-one autocorrelation ϱ_1 .

ς	acceptance	$\hat{\rho}$	ϱ_1	τ_{int}	ESS/N	$(1 - \varrho_1)/(1 + \varrho_1)$
0.50	0.844	0.920	0.907	22.65	0.044	0.049
1.00	0.705	0.821	0.776	8.10	0.123	0.126
2.38	0.445	0.705	0.637	4.51	0.222	0.222
5.00	0.242	0.749	0.712	6.20	0.161	0.168
10.00	0.126	0.861	0.852	11.83	0.085	0.080

The fastest convergence and the highest efficiency occur at $\varsigma = 2.38$, with acceptance rate 0.445 (the classical optimum for a Gaussian target in one dimension). Too small a scale, $\varsigma = 0.5$, gives $\hat{\rho} = 0.92$ and $\text{ESS}/N = 0.044$, and too large a scale, $\varsigma = 10$, gives $\hat{\rho} = 0.86$. The autocorrelation function of the chains is close to geometric, since the value $(1 - \varrho_1)/(1 + \varrho_1)$ of Lemma 1.5 agrees with the measured ESS/N to within 8% in all five cases. As an illustration of the explicit constant of Lemma 1.5, for $R = 2$ and $\varsigma = 2.38$ one finds $q_R = 0.0408$, $\alpha_R = e^{-2} = 0.1353$ and $\delta_R = 0.0221$.

2.5 Total Error and the Choice of h and N

For the OU model we ran one chain per data set ($R = 1000$ data sets, $\varsigma = 2.38$, started at the mean of π_h , burn-in $B = 100$) and estimated m by the chain average $\hat{m}_N$. Table 5 compares the squared bias $\mathbb{E}(m_h - m)^2$ with its leading-order value $V^2Th/2$ from (8), and Table 6 shows the decomposition (13).

Table 5 Squared bias of the posterior mean against its leading-order prediction ($V = 1.10$).

h	$\mathbb{E}(m_h - m)^2$	$V^2Th/2$	ratio
2^{-10}	1.43-3	1.16-3	1.24
2^{-8}	5.96-3	4.62-3	1.29
2^{-6}	2.41-2	1.85-2	1.30
2^{-4}	8.10-2	7.44-2	1.09

The measured MSE equals $\text{bias}^2 + \text{var}$ to within 7% in all 24 combinations of h and N tested, so the cross term in (13) is negligible and the factor 2 in that bound is conservative. The variance follows $\bar{\Sigma}_n \tau_{\text{int}}/(N - B)$ to within about 10%. For a fixed h , increasing N beyond a point brings no gain: the RMSE at $h = 2^{-4}$ is 0.285 at $N = 32000$ and 0.299 at $N = 1000$, because it is dominated by the bias.

Rule for h and N . Combining Table 5 with (14), the RMSE is approximately $\sqrt{V^2Th/2 + V\tau_{\text{int}}/(N - B)}$. To reach the target ε with an equal split of the budget, take

$$h \leq \frac{\varepsilon^2}{V^2T}, \quad N - B \geq \frac{2V\tau_{\text{int}}}{\varepsilon^2}. \quad (34)$$

Table 7 tests (34) for two targets with $V = 1.10$, $T = 2$, $\tau_{\text{int}} = 4.51$, rounding h down to the nearest available value 2^{-k} .

Both observed errors exceed the target by 5% to 6%, because the actual squared bias is 1.24 to 1.29 times the leading-order value (Table 5). The rule is therefore accurate but slightly optimistic, and a safety factor of about 1.3 on the bias term, that is $h \leq \varepsilon^2/(1.3V^2T)$, would correct it. This correction was not tested.

Table 6 Total mean-square error $\mathbb{E}(\hat{m}_N - m)^2$ of the chain average. “var” is $\mathbb{E}(\hat{m}_N - m_h)^2$, and “pred. var” is $\bar{\Sigma}_n \tau_{\text{int}} / (N - B)$ with $\tau_{\text{int}} = 4.51$.

h	N	MSE	bias ² +var	2 bias ² + 2 var	var	pred. var	RMSE
2^{-10}	1000	7.05-3	7.05-3	1.41-2	5.62-3	5.45-3	0.084
2^{-10}	8000	2.02-3	2.05-3	4.10-3	6.21-4	6.21-4	0.045
2^{-10}	32000	1.55-3	1.58-3	3.15-3	1.50-4	1.54-4	0.039
2^{-8}	1000	1.11-2	1.11-2	2.21-2	5.09-3	5.45-3	0.106
2^{-8}	8000	6.67-3	6.59-3	1.32-2	6.21-4	6.21-4	0.082
2^{-8}	32000	6.17-3	6.13-3	1.23-2	1.64-4	1.54-4	0.079
2^{-6}	1000	2.83-2	2.96-2	5.91-2	5.46-3	5.45-3	0.168
2^{-6}	8000	2.45-2	2.48-2	4.95-2	6.75-4	6.21-4	0.157
2^{-6}	32000	2.42-2	2.42-2	4.85-2	1.49-4	1.54-4	0.155
2^{-4}	1000	8.95-2	8.63-2	1.73-1	5.32-3	5.45-3	0.299
2^{-4}	8000	8.16-2	8.16-2	1.63-1	5.56-4	6.21-4	0.286
2^{-4}	32000	8.14-2	8.11-2	1.62-1	1.38-4	1.54-4	0.285

Table 7 Test of the rule (34) ($R = 1000$ data sets, $B = 100$).

target ε	h (rule: $\leq$)	N (rule: $\geq$)	predicted RMSE	observed RMSE
0.10	2^{-8} (0.0041)	1000 (915)	0.100	0.106
0.05	2^{-10} (0.0010)	3500 (3360)	0.051	0.053

2.6 Adaptive Proposal

To test the adaptive sampler (15) in a correlated setting we used the two-parameter drift $b(x; \theta) = -(\theta_1 x + \theta_2 x^3)$ with $\theta = (1, 1)$, $T = 5$, $h = 2^{-7}$ ($n = 640$) and prior $\mathcal{N}((0.5, 0.5), 4I_2)$. The Euler posterior is again Gaussian (Lemma 1.5), with standard deviations (1.086, 1.013), correlation -0.80 and condition number 9.0, so the exact answer is available. Each sampler ran $N = 50\,000$ steps with burn-in 5000 from a starting point three standard deviations away, and the results are averages over 15 independent runs (Table 8). The fixed isotropic proposal $\varsigma^2 I$ was tuned over $\varsigma \in \{0.02, \dots, 3.2\}$, and the best value was $\varsigma = 1.6$. The oracle proposal is $\Gamma^* = (2.38^2/2)\Sigma_n$, and the adaptive sampler starts from the poor proposal $0.02^2 I$ (for which the fixed sampler has $\text{ESS} = 3$ in 20 000 steps) and adapts from $k_0 = 500$ with $\epsilon = 10^{-6}$.

Table 8 Fixed, oracle and adaptive random-walk Metropolis on the two-dimensional Euler posterior (15 runs, $N - B = 45\,000$).

Proposal	acceptance	ESS _{min} (mean $\pm$ s.d.)	max mean error (in s.d.)	relative covariance error
Isotropic, tuned ($\varsigma = 1.6$)	0.262	2672 ± 207	0.016	0.019
Oracle $\Gamma^* \propto \Sigma_n$	0.357	5896 ± 371	0.013	0.021
Adaptive (start $0.02^2 I$)	0.339	6183 ± 280	0.014	0.018

The adaptive sampler, although it starts from a proposal with a ESS of 3, reaches the efficiency of the oracle proposal (6183 against 5896, within one standard deviation) and 2.3 times that of the best isotropic proposal. Its estimates of the posterior mean and covariance are as accurate as those of the fixed samplers, in agreement with ergodicity of the adaptive chain. The quantity $k \|\Gamma_{k+1} - \Gamma_k\|_F$ for $k > 2000$ never exceeded 126 in any run, that is, adaptation was diminishing, and it never exceeded the bound of Lemma 1.5 (the largest ratio was 1.000 after rounding, reached at large excursions of the chain).

2.7 Discussion

Four conclusions follow. First, the two error scales predicted in (9) are confirmed for the OU model and for a nonlinear drift: the fixed-data error of the Euler posterior is of order $\sqrt{h}$ (slope 0.49 to 0.50), and only its average over data sets is of order h (slope 1.04 to 1.05). The leading constants of the Gaussian tilt are accurate to about 10%, so the mechanism is the correct one: a data-dependent tilt of size $\sqrt{Th/2} Z_h$ with Z_h of mean zero. Second, the random-walk chain converges geometrically, with a per-step factor as low as 0.705 at the optimal scale, and its efficiency follows from the lag-one autocorrelation. Third, the total error separates into squared bias and variance, so the bias is a floor that no chain length removes, and the rule (34) gives the required h and N to within about 6% (slightly optimistically). Fourth, the adaptive proposal recovers the efficiency of the oracle proposal from a poor start.

The experiments have limits. The bias experiments use a single scalar parameter with σ known, one horizon $T = 2$ and one true value; the nonlinear reference was validated only for $h \geq 2^{-7}$,

where it has an error below 2% of the effect. Geometric ergodicity was observed on Gaussian posteriors, which are the only Euler posteriors that are Gaussian (drift linear in θ); no experiment here bears on posteriors with heavier tails or drifts that are nonlinear in θ . The adaptive sampler was tested in dimension two with 15 runs, and the ergodicity claim for it rests on the numerical evidence and the diminishing adaptation, not on a proof.

3. Conclusion

This paper analysed the errors of a Metropolis–Hastings scheme that samples the Euler–Maruyama posterior of an SDE observed at discrete times. The analysis showed that the discretization error has two scales: for a fixed data set the Euler posterior differs from the posterior based on the exact transition density by an amount of order $\sqrt{h}$, caused by a random exponential tilt of the likelihood, while its average over data sets, the weak bias, is of order h . Together with the ergodicity of the random-walk chain and the effective sample size, this gives a total-error decomposition and a rule for choosing the step size h and the chain length N .

The numerical study supported the analysis. For the Ornstein–Uhlenbeck process and for the nonlinear drift $-\theta x - x^3$, the fixed-data error scaled as $h^{0.49}$ to $h^{0.50}$ and the data-averaged bias as $h^{1.04}$ to $h^{1.05}$, and the mean Wasserstein distance matched the prediction $V\sqrt{Th/\pi}$ to within 10%. The random-walk chain converged geometrically, with a per-step factor 0.705 at the optimal scale, and $\text{ESS}/N = 0.22$. The total mean-square error equalled squared bias plus variance to within 7%, and the rule for h and N reached the target error to within 6%. The adaptive sampler, started from a proposal of effective sample size 3, reached the efficiency of the oracle proposal, 2.3 times that of the best fixed isotropic proposal.

The main practical consequence is that, for fixed observations, refining the time step h reduces the discretization error only like $\sqrt{h}$, so that the accuracy of the sampler cannot be judged by the weak order of the scheme. Future work includes proving the $\sqrt{h}$ bound and the $\mathcal{O}(h)$ averaged bias for general drifts, treating unknown σ and multivariate diffusions, drifts nonlinear in θ for which the Euler posterior is not Gaussian, and a Gaussian transition with a corrected variance (local linearization type), which may remove the $\sqrt{h}$ term because the tilt is produced by the variance mismatch. A proof of ergodicity of the adaptive sampler for this class of posteriors also remains to be given.

References

- [1] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Applications of Mathematics, vol. 23, Springer, Berlin, 1992.
- [2] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., Universitext, Springer, Berlin, 2003.

- [3] D. Talay and L. Tubaro, Expansion of the global error for numerical schemes solving stochastic differential equations, *Stochastic Analysis and Applications*, **8**(4) (1990), 483–509.
- [4] M. Kessler, Estimation of an ergodic diffusion from discrete observations, *Scandinavian Journal of Statistics*, **24**(2) (1997), 211–229.
- [5] G. O. Roberts and O. Stramer, On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm, *Biometrika*, **88**(3) (2001), 603–621.
- [6] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, Equation of state calculations by fast computing machines, *The Journal of Chemical Physics*, **21**(6) (1953), 1087–1092.
- [7] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**(1) (1970), 97–109.
- [8] G. O. Roberts and R. L. Tweedie, Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms, *Biometrika*, **83**(1) (1996), 95–110.
- [9] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Communications and Control Engineering Series, Springer, London, 1993.
- [10] H. Haario, E. Saksman and J. Tamminen, An adaptive Metropolis algorithm, *Bernoulli*, **7**(2) (2001), 223–242.