PLOS | ONE

RESEARCH ARTICLE

Applying diffusion-based Markov chain Monte Carlo

Radu Herbei[1]*, **Rajib Paul**[2], **L. Mark Berliner**[1]

1 The Ohio State University, Department of Statistics, Columbus, OH, United States of America, **2** Western Michigan University, Department of Statistics, Kalamazoo, MI, United States of America

* herbei@stat.osu.edu

Abstract

We examine the performance of a strategy for Markov chain Monte Carlo (MCMC) developed by simulating a discrete approximation to a stochastic differential equation (SDE). We refer to the approach as *diffusion MCMC*. A variety of motivations for the approach are reviewed in the context of Bayesian analysis. In particular, implementation of diffusion MCMC is very simple to set-up, even in the presence of nonlinear models and non-conjugate priors. Also, it requires comparatively little problem-specific tuning. We implement the algorithm and assess its performance for both a test case and a glaciological application. Our results demonstrate that in some settings, diffusion MCMC is a faster alternative to a general Metropolis-Hastings algorithm.

Citation: Herbei R, Paul R, Berliner LM (2017) Applying diffusion-based Markov chain Monte Carlo. PLoS ONE 12(3): e0173453. https://doi.org/10.1371/journal.pone.0173453

Editor: Gui-Quan Sun, Shanxi University, CHINA

Received: November 7, 2016

Accepted: February 22, 2017

Published: March 16, 2017

Data Availability Statement: The surface, basal and velocity data used in the current manuscript are available for download as a MATLAB file in the supporting information section. Data used in this paper are from the study below whose corresponding author may be contacted at mb@stat.osu.edu: Berliner, L. M., Jezek, K., Cressie, N., Kim, Y., Lam, C. Q., and van der Veen, C.J., Modeling dynamic controls on ice streams: a Bayesian statistical approach. Journal of Glaciology 54 (2008), pp. 705714. Available at: http://www.ingentaconnect.com/content/igsoc/jog/2008/00000054/00000187/art00014.

Introduction

The advent of Markov Chain Monte Carlo (MCMC) has led to major advances in the application of Bayesian analysis in complex problems. The idea is simply put: faced with a posterior distribution too complicated to compute or simulate from directly (i.e., we cannot readily obtain the normalizer or denominator appearing in Bayes' Theorem), one develops a Markov chain whose stationary distribution is known to coincide with the target posterior distribution. One then runs that chain, knowing that eventually realizations from the chain form an approximate dependent sample from the posterior. Those realizations are then used to estimate features of the posterior (i.e., posterior expectations of interesting quantities, predictive densities, etc.) [1–3].

For example, in some settings, nonlinearity and/or nonconjugacy of certain components of a large model render the standard Gibbs Sampler unusable. Metropolis-Hastings algorithms and Gibbs-Metropolis hybrids can be suggested, though these approaches can be taxing and may require substantial tuning.

In response to such difficulties, we explore diffusion based strategies for MCMC analysis. That is, one develops a diffusion (a solution, in the sense of Itô, to a stochastic differential equation) whose stationary distribution is the target posterior, see Chapter 5 of [4]. The key idea is certainly not new. Indeed, Langevin MCMC procedures are often suggested for

Funding: RH has been supported in part by the National Science Foundation (www.nsf.gov) grant number DMS-1209142. RP has been supported in part by the National Science Foundation (www.nsf.gov) grant NSF-CMMI-1537379. LMB was supported in part by the National Science Foundation (www.nsf.gov) grants ATM-07-3024403 and DMS-10-49064. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Competing interests: The authors have declared that no competing interests exist.

generating candidate states in Metropolis steps in MCMC. In this article, we suggest diffusion MCMC as a stand-alone algorithm.

In part, our motivation for suggesting diffusion MCMC is its simplicity in terms of set-up. There are no probability calculations to perform, as in Gibbs' Sampling, nor any need for choosing and updating distributions for generating candidate states. Indeed, the approach is recommended as an "off-the-shelf" strategy that can be readily implemented. However, as indicated below, it is not a panacea. Further, issues such as burn-in, mixing, convergence rates, and output analysis remain challenging.

Consider a Bayesian analysis for an unknown quantity θ (after the introduction, we allow vector-valued unknowns) based on observational data $\mathbf{Y}$, having conditional density $g(\mathbf{y} \mid \theta)$. Let $\pi(\theta)$ denote our prior distribution for θ. We are to obtain the posterior distribution for θ based on the fixed observation $\mathbf{Y} = \mathbf{y}$,

$$p(\theta) \overset{\text{def}}{=} p(\theta \mid \mathbf{y}) = C(\mathbf{y})^{-1} g(\mathbf{y} \mid \theta)\pi(\theta), \tag{1}$$

where $C(\mathbf{y})$ is the normalizing constant. Consider a one-dimensional stochastic differential equation (SDE)

$$d\theta(t) = b(\theta)dt + \sigma(\theta)dW(t), \quad \theta(0) = \theta_0, \quad t \geq 0, \tag{2}$$

where θ_0 is some fixed initial value and the *drift* $b(\cdot)$ and *diffusion* $\sigma(\cdot) > 0$ are specified functions, such that Eq (2) admits a unique weak solution. The initial state θ_0 is a random variable with specified density $p(\theta, 0)$; and $dW(t)$ represents *white noise*. Specifically, $\{W(t): t \geq 0\}$ is a standard Brownian motion process or a Wiener process. Consider the temporal evolution of the probability density function, $p(\theta, t)$, of a solution $\theta(t)$. Under regularity conditions requiring b and σ to be differentiable and satisfy a Lipschitz condition

$$|b(\theta) - b(\theta')| + |\sigma(\theta) - \sigma(\theta')| \leq K|\theta - \theta'|,$$

for some constant K and for all θ, θ', $p(\theta, t)$ is the solution to the ordinary partial differential equation, known as the *Fokker-Planck* or *Kolmogorov Forward* Equation,

$$\frac{\partial p}{\partial t} = 0.50 \frac{\partial^2}{\partial \theta^2}(\sigma^2 p) - \frac{\partial}{\partial \theta}(b p), \tag{3}$$

subject to the initial condition $p(\theta, 0)$ and the assumption that $\theta(0)$ and $\{W(t): t \geq 0\}$ are independent.

Our interest is in stationary solutions, i.e., solutions that are functionally independent of time, so the partial derivative with respect to t is zero. Setting the right-hand side of Eq (3) equal to zero and integrating the result once w.r.t. θ, it suffices to find a stationary density $p(\theta)$ such that

$$0.50 \frac{\partial}{\partial \theta}(\sigma^2(\theta) p(\theta)) = b(\theta) p(\theta) \ , \tag{4}$$

for all values of θ in the parameter space. The general solution of Eq (4) is

$$p(\theta) = (c\sigma^2(\theta))^{-1} \exp\left(\int_0^\theta \frac{2b(z)}{\sigma^2(z)} dz\right),$$

where the constant c is a normalizer.

We let $p(\theta)$ be the target posterior Eq (1) for our Bayesian model and find appropriate functions $b(\theta)$ and $\sigma(\theta)$ such that $p(\cdot)$ satisfies the Eq (4). That is,

$$b(\theta) \quad = 0.50\left(\sigma^2(\theta)' + \sigma^2(\theta)\frac{p(\theta)'}{p(\theta)}\right) = 0.5\sigma^2(\theta)\left(\frac{\sigma^2(\theta)'}{\sigma^2(\theta)} + \frac{p(\theta)'}{p(\theta)}\right)$$

where the derivatives are taken with respect to θ. When $p(\theta) = g(\mathbf{y} \mid \theta)\pi(\theta)$ as in Eq (1), we look for functions $b(\cdot)$ and $\sigma^2(\cdot)$ satisfying

$$\begin{aligned} b(\theta) &= 0.50\sigma^2(\theta)\left(\frac{d}{d\theta}\log\left(g(\mathbf{y} \mid \theta)\pi(\theta)\sigma^2(\theta)\right)\right) \\ &= 0.50\sigma^2(\theta)\left(\frac{g(\mathbf{y} \mid \theta)'}{g(\mathbf{y} \mid \theta)} + \frac{\pi(\theta)'}{\pi(\theta)} + \frac{\sigma^2(\theta)'}{\sigma^2(\theta)}\right), \end{aligned} \tag{5}$$

Having completed this step, we can simulate the diffusion and proceed as in MCMC. This is typically accomplished by forming a discrete-time approximation to Eq (2), that is, a Markov chain approximation to the continuous-time process. There may be many pairs $(b(\cdot), \sigma^2(\cdot))$ that work for a fixed Bayesian model. It is important to note that the core of a *diffusion MCMC* (DMCMC) implementation has been completely described.

In this article we discretize the diffusion Eq (2) using the Euler scheme. This method has been extensively discussed in the literature, see for example [5] for a comprehensive overview and [6–8] for recent developments. The solution of the stochastic differential Eq (2) is approximated using a discrete time Markov chain $\{\theta_m\}_{m \geq 0}$,

$$\theta_{m+1} = \theta_m + hb(\theta_m) + \sigma(\theta_m)h^{1/2}Z_{m+1}, \quad m \geq 0 \tag{6}$$

where $\theta_0 = \theta(0)$, $h > 0$ is the discretization step-size, and Z_{m+1} is a realization from a standard Gaussian distribution. Abusing notation, we use θ to denote both the continuous-time defined in Eq (2) and the discrete-time process from Eq (6). It is typical to extend $\{\theta_m\}_{m \geq 0}$ to a continuous time process via interpolation; however this step is not necessary for this paper. From a practical perspective, we are interested in the process $\{\theta_m\}_{m \geq 0}$. Two critical questions arise:

1. Does the discrete stochastic process converge to a stationary, ergodic distribution?
2. If so, is that stationary distribution "close" enough to the target posterior distribution to justify the use of conventional output analysis to enable approximate Bayesian inference?

Unfortunately, there are situations where for any choice of the time step $h > 0$, the Markov chain described by Eq (6) will behave drastically different than the continuous time version Eq (2), see [9] for a discussion on this issue. Nevertheless, in many cases the ergodic properties of the discretized process $\{\theta_m\}_{m \geq 0}$ are similar to those of $\{\theta(t)\}_{t \geq 0}$. In particular, in [10] the author shows that under regularity conditions, the Euler discretization scheme does have a stationary measure which converges at an appropriate rate to the unique stationary measure of the continuous-time SDE. Furthermore, in [9] the authors provide conditions under which the continuous-time Langevin diffusion (defined below in Eq (7)) as well as the discretized version Eq (6) are geometrically ergodic. In [7], the authors study the asymptotic properties of time averages $(1/M)\sum_{m=1}^{M} F(\theta_m)$, where F is a given function. This statistic is the natural estimate for the expected value $E(F(\theta)) = \int F(\theta)dp$. They use Poisson equations to show that under mild regularity conditions, any stationary measure of the Euler-discretized process Eq (6) will be close to the unique stationary measure of the underlying SDE. Their Theorem 5.1 also shows that the time average estimator is of order $O(h + 1/M)$.

To illustrate *diffusion-MCMC* and indicate its potential value and simplicity, examples are reviewed in the next section. Henceforth, we set $\sigma \equiv 1$, and then find the function $b(\cdot)$; this approach often has the tag *Langevin*. That is, we restrict ourselves to diffusion processes of the form

$$d\theta(t) = \frac{1}{2}\frac{\partial}{\partial\theta}\log p(\theta(t))dt + dW(t), \quad t \geq 0, \tag{7}$$

where $p(\cdot)$ is the posterior distribution Eq (1). We note that application of Eq (6) yields the corresponding transition distribution as

$$\theta_{m+1}|\theta_m \sim N(\theta_m + 0.5h\nabla\log p(\theta_m), h).$$

We emphasize again that our goal is to present the benefits of a procedure that has been present in the literature for over a decade. Due to its wild behavior even in some simple cases, it has received little attention, especially from practitioners and applied scientists. It has been proposed (see for example the MALA algorithm presented in [9]) that an additional Metropolis-Hastings step will correct the explosive behavior. The MALA algorithm has been further studied and extended in [11] and [12]. In both cases the improvement comes with an increase in computational complexity. We stress that our goal is to avoid a Metropolis-Hastings accept-reject step and this work is motivated by recent theoretical advances in this direction, see [7]. We explore the efficiency and applicability of DMCMC to high-dimensional problems arising in a Bayesian framework, **without** performing the Metropolis-Hastings correction step. When classical (or adaptive) MCMC fails (for example, due to computational time restrictions or inability to select good proposals), we show that diffusion MCMC is a viable alternative which requires little input from the user and can be computationally more efficient.

Motivating examples

The multivariate form of the diffusion Eq (2) is written as

$$d\boldsymbol{\theta}(t) = b(\boldsymbol{\theta}(t))dt + \sigma(\boldsymbol{\theta}(t))d\mathbf{W}(t), \quad t > 0, \tag{8}$$

where $\{\boldsymbol{\theta}(t): t \geq 0\}$ is a q-dimensional stochastic process. The initial state is $\boldsymbol{\theta}(0)$ and $\{\mathbf{W}(t), t \geq 0\}$ is a q-dimensional vector whose elements are each independent standard Brownian motions. Except for the first one, the examples below were chosen to be suggestive of realistic problems for which other MCMC methods can be difficult. As in Eq (1), we use $g(\mathbf{y} \mid \boldsymbol{\theta})$ to denote the likelihood function, where $\mathbf{y}$ is the observed data and $\pi(\boldsymbol{\theta})$ to denote the prior density.

One class of problems in which diffusion MCMC may be useful involve nonlinearity. For example, suppose the likelihood function g depends on $\boldsymbol{\theta}$ via a "link" function $k(\cdot)$, that is $g = g(\cdot \mid k(\boldsymbol{\theta}))$. Nonlinear structures may also arise in hierarchically specified priors. Nonlinearity may make both Gibbs sampling and Metropolis algorithms difficult. However, if the nonlinearity does not disable the required differentiation, diffusion MCMC may be comparatively simple. We remark that in such cases, the drift function $b(\cdot)$ may be unruly. If necessary, selection of the diffusion coefficient $\sigma(\cdot)$ may be used to control $b(\cdot)$. However, for the balance of the article we restrict to Langevin diffusions ($\sigma = 1$).

Example 1. Assume that for τ^2 known, $Y \mid \theta \sim N(\theta, \tau^2)$ and $\theta \sim N(\mu, \eta^2)$. Of course, we know that $\theta \mid Y = y$ is normally distributed with easily computed mean and variance (these will appear below). Applying Eq (5) with the choice of $\sigma(\theta) = 1$, yields

$$b(\theta) = 0.50\left(\frac{y}{\tau^2} + \frac{\mu}{\eta^2} - \theta\left(\frac{1}{\tau^2} + \frac{1}{\eta^2}\right)\right).$$

Let $\alpha = 0.50(\tau^{-2}\, y + \eta^{-2}\, \mu)$ and $\beta = 0.50(\tau^{-2} + \eta^{-2})$. The solution to

$$d\theta(t) = (\alpha - \beta\theta(t))dt + dW(t)$$

is

$$\theta(t) = \int_0^t \alpha e^{-\beta(t-s)} ds + \int_0^t e^{-\beta(t-s)} dW(s) + \theta(0)e^{-\beta t}.$$

It follows that

$$E(\theta(t)) = \int_0^t \alpha e^{-\beta(t-s)} ds + E(\theta(0))e^{-\beta t} \tag{9}$$

$$= \frac{\alpha}{\beta}(1 - e^{-\beta t}) + E(\theta(0))e^{-\beta t}. \tag{10}$$

It can be shown that

$$\mathrm{var}(\theta(t)) = \int_0^t e^{-2\beta(t-s)} ds + \mathrm{var}(\theta(0)e^{-\beta t}) \tag{11}$$

$$= (2\beta)^{-1}(1 - e^{-2\beta t}) + \mathrm{var}(\theta(0))e^{-2\beta t}. \tag{12}$$

Returning to the original parameterization, we conclude that as $t \to \infty$,

$$E(\theta(t)) \to \left(\frac{1}{\tau^2} + \frac{1}{\eta^2}\right)^{-1} \frac{y}{\tau^2} + \frac{\mu}{\eta^2}$$

and

$$\mathrm{var}(\theta(t)) \to \left(\frac{1}{\tau^2} + \frac{1}{\eta^2}\right)^{-1},$$

which are the usual posterior mean and variance for this Bayesian model. If the initial condition $\theta(0)$ is normally distributed, then for each t, $\theta(t)$ is also normally distributed. Note that the convergence rate to the stationary distribution is exponentially fast.

Example 2. Diffusion MCMC is useful in combining data from highly different likelihoods. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and assume that $Y_{ij}|\theta_i \sim g_i(\cdot|\theta_i)$ where $i = 1, \ldots, K$ and $j = 1, \ldots, r_i$ and all Y_{ij} are conditionally independent. For example, let g_i be the Gaussian pdf with mean θ_i and variance τ^2, and the prior for θ_i be a Cauchy distribution with median μ and scale parameter A. For a Langevin setting (i.e. $\sigma = 1$), the i^{th} component of the drift coefficient $\nabla \log p(\boldsymbol{\theta})$ is

$$\frac{\partial}{\partial\theta_i} \nabla \log p(\boldsymbol{\theta}) = -\sum_{j=1}^{r_i} \frac{(\theta_i - Y_{ij})}{\tau^2} - \sum_{i=1}^{K} \frac{2(\theta_i - \mu)}{A^2 + (\theta_i - \mu)^2}.$$

Note that conjugacy plays no direct role in this approach, though the presence of the Cauchy distribution makes a Gibbs sampler infeasible. This example is further analyzed in the next Section.

Example 3. (Mixture Models) Suppose $Y_1, \ldots, Y_n$ are conditionally independent and identically distributed given $\boldsymbol{\theta}$ according to a finite mixture of m probability density functions $g_i(\cdot \mid \theta)$. For example, assume that the conditional distribution of the data is

$$g(\mathbf{y} \mid \theta) = \prod_{j=1}^{n} \left(\sum_{i=1}^{m} \alpha_i g_i(y_j \mid \theta) \right),$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$, $\alpha_i > 0$, $i = 1, \ldots, m$ and $\sum_i \alpha_i = 1$. Diffusion MCMC is easily formulated if the derivatives of the g_i with respect to θ are easily available, either via formal calculations or by using symbolic software, such as Mathematica. We note that similar steps can be used to treat mixture priors.

Example 4. **(Hierarchical Models)**. In many models g and π are products of a variety of terms, e.g., for conditionally independent observations, g is a product; π is often represented as a product of hierarchical components. In such cases, we have that

$$\frac{\partial \log (g\pi)}{\partial \theta_i} = \frac{\partial \log (g^{(i)} \pi^{(i)})}{\partial \theta_i},$$

where the superscripts indicate that only those components of g and π that explicitly depend on θ_i are involved in the calculation. This parallels the familiar step in computing full conditionals in setting up a Gibbs Sampler. Namely, for each i, one computes the distributions

$$[\theta_i \mid \text{all other } \theta_j] = \frac{g^{(i)} \pi^{(i)}}{\int g^{(i)} \pi^{(i)} d\theta_i}. \tag{13}$$

Suppose that the Bayesian model takes the form $\mathbf{Y} \mid \theta_1, \ldots \theta_q \sim g(\mathbf{y} \mid \theta_1, \ldots \theta_q)$ and

$$\pi(\theta_1, \ldots \theta_q) = \pi^1(\theta_1 \mid \theta_2, \ldots \theta_q) \pi^2(\theta_2 \mid \theta_3, \ldots \theta_q) \cdots \pi^q(\theta_q).$$

We adapt the notation in Eq (5) as follows: for a function $f(\theta_1, \ldots \theta_q)$ define

$$f_{(i)} = \frac{\partial f}{\partial \theta_i}, \; i = 1, \ldots, q.$$

Hence,

$$\frac{\partial \log (g\pi)}{\partial \theta_i} = \frac{g_{(i)}}{g} + \sum_{j=1}^{i} \frac{\pi^j_{(i)}}{\pi^j} \tag{14}$$

We note that Gibbs sampling is useful when the full conditionals Eq (13) are readily obtained and simulated. This typically arises when the full conditionals actually depend on a small subset of the parameters in the conditions. This is not necessary in diffusion MCMC.

Applications

To provide insight into diffusion MCMC (DMCMC), we present a standard test case and a real-data example. Our goal is to assess the performance of DMCMC, especially in comparison with the current *state-of-the art* adaptive MCMC approach, see [13, 14]. The DMCMC methodology is compared to a multivariate adaptive Metropolis sampler (AM). For the AM algorithm, the proposal distribution at iteration m is given by

$$(1 - \beta) N(x, (2.38)^2 \Sigma_m / q) + \beta N(0, (0.1)^2 I_q / q)$$

where Σ_m is the current estimate of the covariance matrix of the target distribution and β is a small positive constant (we take $\beta = 0.05$). The AM algorithm is widely accepted as one of the best sampling algorithms, especially for complex target distributions where dependencies among parameters make it difficult to select proposal distributions. We refer the reader to [15] for several comparisons between MCMC algorithms. The scaling factor $(2.38)^2$ can also be "adapted"; in this case we refer to the procedure as *adaptive scaling within adaptive MCMC*. However, user input is not eliminated completely as there remain tuning parameters to be specified.

For comparisons, we inspect trace-plots to assess convergence and compare algorithms via their *averaged squared jumping distance*

$$ASJD = E((X_m - X_{m-1})^2).$$

This quantity is estimated by $\frac{1}{M}\sum_{m=1}^{M}(X_m - X_{m-1})^2$ for both AM and DMCMC algorithms. Comparatively large *ASJD* indicates the desirable property of fast mixing.

We also add a computational constraint for our examples. We limit ourselves to relatively short runs of the Markov chains (AM and DMCMC). This can be very dangerous for classical MCMC since one will have difficulty assessing whether the chains have reached stationarity. Our examples will show that the diffusion approach quickly finds regions with high posterior probability and explores them thoroughly.

Synthetic example

Assume that $Y_{i1}, \ldots, Y_{ir_i}|\theta_i, \gamma$ are an iid sample from a *Gaussian*$(\theta_i, V(\gamma))$ distribution, where $1 \leq i \leq 1000$ and $1 \leq j \leq r_i$. We specify the variance $V(\gamma)$ to be

$$V(\gamma) = \frac{a + b\, e^{\gamma}}{1 + e^{\gamma}}\ , \qquad \text{for } \gamma \in \mathbb{R}, \tag{15}$$

where $0 < a < b < \infty$ are specified constants. The reason behind Eq (15) is twofold: (1) we require that all the parameters of the model be supported on the entire real line, hence a transformation is required for all variances, and (2) we aim for a *Uniform*(a, b) prior distribution for the variance $V(\gamma)$. Certainly, other distributions (such as Gamma or Inverse Gamma) can be considered. Using an inverse transformation, this is equivalent to specifying the prior density for γ as

$$f(\gamma) = \frac{e^{\gamma}}{(1 + e^{\gamma})^2}\ , \quad \gamma \in \mathbb{R}\ .$$

We let the sample sizes r_i vary between 5 and 500. For $\theta_1, \theta_2, \ldots, \theta_{500}$ we specify independent prior distributions, $\theta_i|\mu, A \sim$ *Cauchy*(μ, A), with density proportional to $[1 + ((\theta_i - \mu)/A)^2]^{-1}$. The parameter A is held fixed for this example, although it can be treated similarly to the data variance $V(\gamma)$. For the hyperparameter μ we specify a *Gaussian*$(0, 1)$ prior distribution. Using the independence assumption, the likelihood function is written as

$$\begin{aligned} g_y(\boldsymbol{\theta}) &\equiv g_y(\theta_1, \ldots, \theta_{1000}, \gamma, \mu) \\ &\propto \prod_{i=1}^{1000}\left(V(\gamma)^{-r_i/2}\right)\exp\left\{-\frac{\sum_{j=1}^{r_i}(Y_{ij} - \theta_i)^2}{2V(\gamma)}\right\}, \end{aligned} \tag{16}$$

and the prior density is proportional to

$$\begin{aligned}\pi(\boldsymbol{\theta}) &\equiv \pi(\theta_1,\ldots,\theta_{1000},\gamma,\mu)\\ &\propto \left[\prod_{i=1}^{1000}\frac{1}{1+\left(\dfrac{\theta_i-\mu}{A}\right)^2}\right]\cdot\frac{e^{\gamma}}{(1+e^{\gamma})^2}\cdot\exp\{-\mu^2/2\}\ .\end{aligned} \tag{17}$$

Selection of the time step. The selection of "good" time steps for DMDMC is challenging. First, time steps that are too large may result in explosive (transient) processes. In general, the user faces a conundrum: a very small time step typically results in chain whose dynamics are similar to those of the target continuous-time diffusion, but is slowly mixing.

In our experiments we observed that selecting $h = O(1/q)$ results in a good performance of the DMCMC algorithm for this example. In Fig 1 we display the trace plots and autocorrelation functions for three parameters (results for all parameters where quite similar) from the

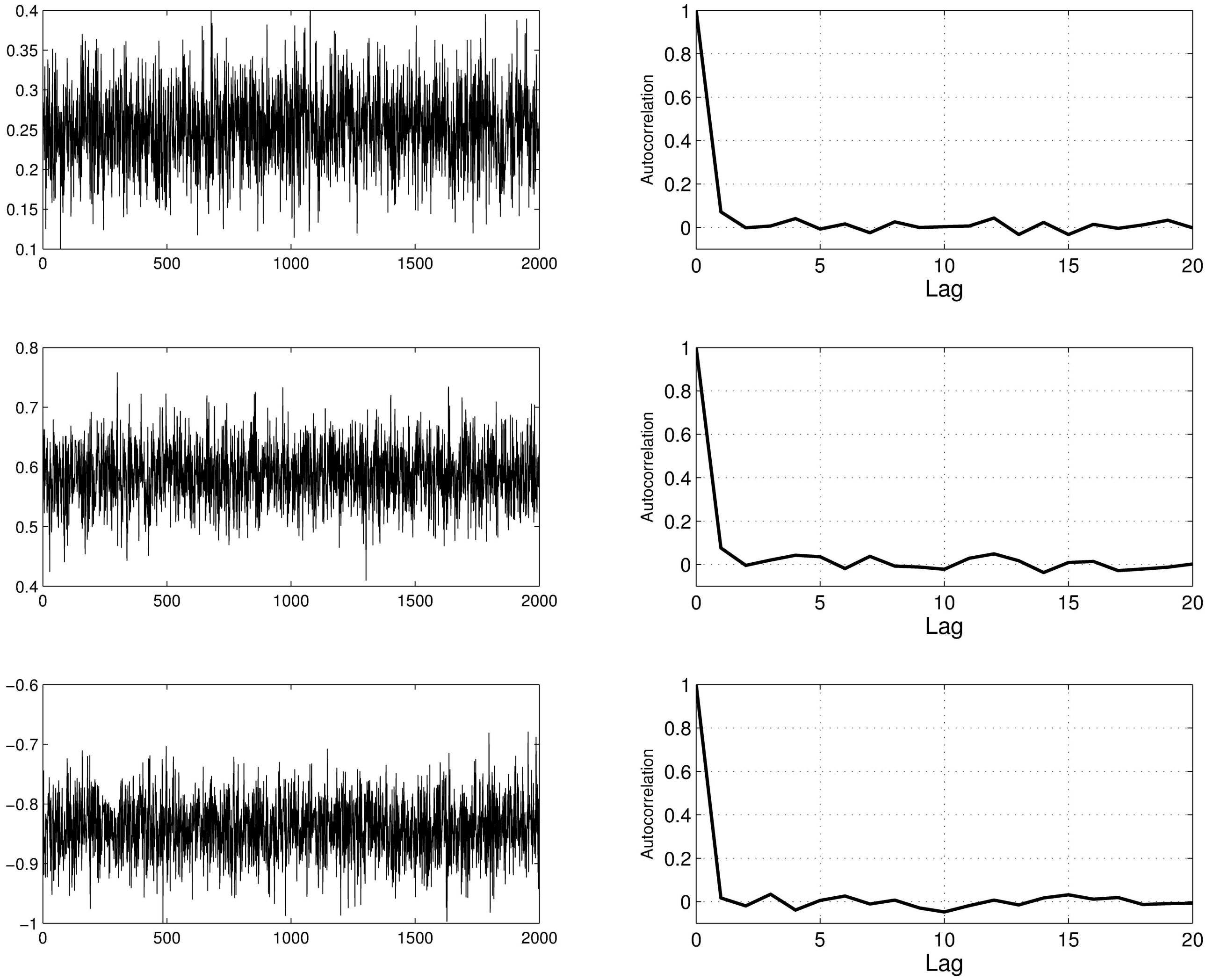

Fig 1. Trace plots and autocorrelation functions for three parameters: θ_1, θ_{101}, θ_{201}.

https://doi.org/10.1371/journal.pone.0173453.g001

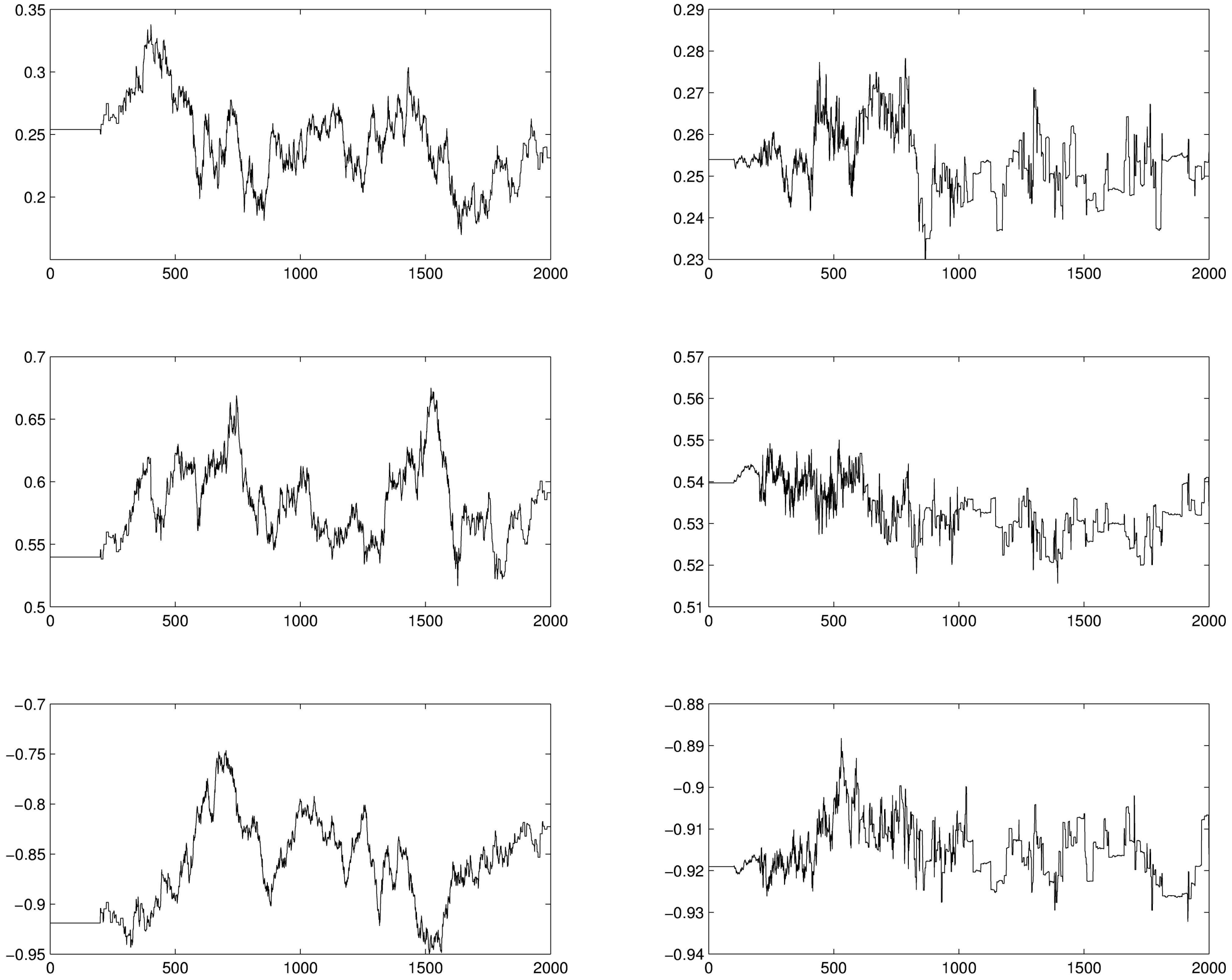

Fig 2. Trace plots for three parameters: θ_1, θ_{101}, θ_{201}. The left panels show the results from the adaptive MCMC sampler. The right panels show the *adaptive scaling within adaptive MCMC* sampler.

https://doi.org/10.1371/journal.pone.0173453.g002

DMCMC approach. Fig 2 shows trace plots for the same parameters resulting from two adaptive approaches as described above. In each case the chains were run for 20000 iterations and thinned by ten. It is evident from these figures that the adaptive approach has difficulty exploring the state space. The key issue is the dimension of the state space (1000+ in this example). The adaptive approach will not "learn" a one thousand dimensional covariance matrix properly. In Table 1 we summarize the target values and estimates (posterior means) for two parameters, θ_1 and θ_{201}. Table 2 contains corresponding estimates of ASJD. We see that the estimated ASJD indicates that the mixing rates of the diffusion MCMC algorithm is much higher than the adaptive case.

Table 1. Selected DMCMC and AM point estimates.

Parameter	Target value	DMCMC estimate	AM estimate
θ_1	0.2508	0.2517	0.2451
θ_{201}	-0.9333	-0.8403	-0.8637

https://doi.org/10.1371/journal.pone.0173453.t001

Table 2. Average squared jumping distance for θ_1 and θ_{201}.

Parameter	AJSD$_{DMCMC}$	ASJD$_{AM}$
θ_1	0.0042	0.17×10^{-4}
θ_{201}	0.0044	0.16×10^{-4}

https://doi.org/10.1371/journal.pone.0173453.t002

Glacial dynamics

In [16], the authors present a hierarchical Bayesian analysis for inferring features of the dynamics of the Northeast Ice Stream in Greenland. For our purposes, we use a subset of their data and a simplified version of their model. Glaciers flow under the influence of gravity moderated by resistive forces at its base and sides. For a path roughly along the center of the glacier as it flows toward the sea, physical reasoning and simplifying approximations lead to a model for surface velocity of the stream as a function of ice thickness and the shape of the surface. Let x (m) denote a spatial location. The model for the surface velocity $u(x)$ (ms^{-1}) used here is

$$u(x) = u_{bx} + (0.50)\, A(x)\, (\rho g)^3\, H^4(x) \left(\frac{ds(x)}{dx}\right)^3, \tag{18}$$

where u_{bx} is sliding velocity, $s(x)$ (m) is ice-surface elevation, $H(x)$ (m) is ice thickness, $\rho = 911$ kgm^{-3} is the density of ice, and $g = 9.81$ ms^{-2} is the gravity constant. Though the quantity $A(x)$ depends on temperature, it is often treated as a constant *flow parameter*. In this article we model A using a Fourier expansion

$$A(x) = a_0 + \sum_{k=1}^{3} a_k \cos(kx\omega) + b_k \sin(kx\omega). \tag{19}$$

We assume the following quadratic model for the surface:

$$s(x) = \beta_1 x^2 + \beta_0, \tag{20}$$

where β_0 and β_1 are unknown parameters. The authors in [16] use a different, more complicated functional form for s. We found Eq (20) sufficient for our purposes. Let $B(x)$ be the elevation of the base of the glacier so that

$$H(x) = s(x) - B(x).$$

The dataset consists of vectors **S** observed at fill-in spatial locations covering approximately 200 km of the ice stream, the observed surface topography; **B**, the observed basal topography; and **U**, surface velocities. Additional description of the data is given in [16].

Data models. Let $\boldsymbol{\theta}$ represent the set of all parameters introduced in the modeling. We assume that **S**, **B**, **U** are conditionally independent given $\boldsymbol{\theta}$.

Surface Data. The data model for **S** is a conventional Gaussian measurement error model:

$$\mathbf{S} \mid \boldsymbol{\theta} \sim N(\mathbf{s}, \sigma_S^2 I), \tag{21}$$

where **s** is the vector of values of Eq (20) at the observation locations; σ_S^2 is an unknown measurement error variance; and I is the identity matrix.

Basal Data. It is argued (see [16]) that the basal data must be smoothed to be useful in Eq (18). Following their approach we partition the domain of the data into $2^{10} = 1024$ bins of equal length (189.5 m). All basal observations within each bin are averaged, leading to a data vector $\bar{\mathbf{B}}$ of length 1024. As in [16], we use a wavelet model with two sets of wavelets to provide smoothing. The first group of wavelets captures a smooth signal; the second captures fine-scale or detail signals. Based on the results of [16], we used four smooth and 28 detail wavelets. Specifically, we assume that

$$\bar{\mathbf{B}} \mid \boldsymbol{\theta} \ \sim N(\mathbf{WC}, \sigma_B^2 \, \text{diag}\{n_i^{-1}\}), \tag{22}$$

where **W** is the 1024 × 32 matrix of discretized wavelet basis functions; **C** is the 32-dimensional vector of wavelet coefficients; σ_B^2 is an error variance; and $\text{diag}\{n_i^{-1}\}$ is a 1024 × 1024 diagonal matrix with diagonal elements equal to n_i^{-1} where n_i is the number of observations averaged in bin i (all n_i are either one or two). We selected Daubechies wavelets, see [17] and [18] for discussion.

Velocity Model. We assume that

$$\mathbf{U} \mid \boldsymbol{\theta} \sim N(\mathbf{u}, \sigma_U^2 I), \tag{23}$$

where **u** is the vector of values given by Eq (18) at the observation locations; σ_U^2 is the unknown measurement error variance; and I is the identity matrix. Note that the sliding velocity is assumed to be a constant over the study range.

Priors for parameters. **Error Variances**. The measurement error variances σ_S^2, σ_B^2, and σ_U^2 were assigned independent, inverse gamma distributions with means and standard deviations (100, 10), (2500, 500), and (9, 3), respectively.

Surface Model Parameters. The prior distributions for β_0 and β_1 were specified to be independent normal distributions with large variances: 10,000 for β_0 and 10 for β_1. The means of these normal distributions were set to be equal to the least squares estimates of β_0 and β_1 derived from a traditional analysis fitting the model in Eq (20) to the surface observations.

Basal Model Parameters. The prior used for the four coefficients of the smooth-signal wavelets is

$$\mathbf{C}_s \sim N(\boldsymbol{\mu}_s\,, \sigma_c^2\, I_4)$$

where $\boldsymbol{\mu}_s$ is the vector of conventional least squares estimates of Haar-wavelet coefficients. The prior for the remaining 28 coefficients is

$$\mathbf{C}_d \sim N(\mathbf{0}, \sigma_d^2\, I_{28}).$$

We set $\sigma_c^2 = 2000$ and $\sigma_d^2 = 10000$.

Velocity Model Parameters. The prior for the sliding velocity is

$$u_b \sim N(35, 14^2).$$

To develop reasonable priors for the Fourier coefficients in Eq (19), we first obtained the least squares fits to the surface and the basal models (20) and (22). These fitted models were substituted into the velocity model (23). We then fitted the result via least squares. As above, the least squares estimates of the Fourier coefficients were used as prior means for the corresponding parameters. These values were on the order of 10^{-16} (which is consistent with the theoretical value of the parameter A) except for the frequency parameter ω which was estimated to be roughly 10^{-5}. These parameters were all assumed to be independent, normal random variables with prior variances equal to 10.

Performance. Fig 3 shows trace plots for various parameters. We ran the algorithm for 100000 iterations and thinned it by fifty steps. The diffusion MCMC algorithm performs very well. It appears that it explores the state space properly and mixes very fast. In Fig 4 we show posterior means for the surface, velocity and basal processes. For comparison, we added the

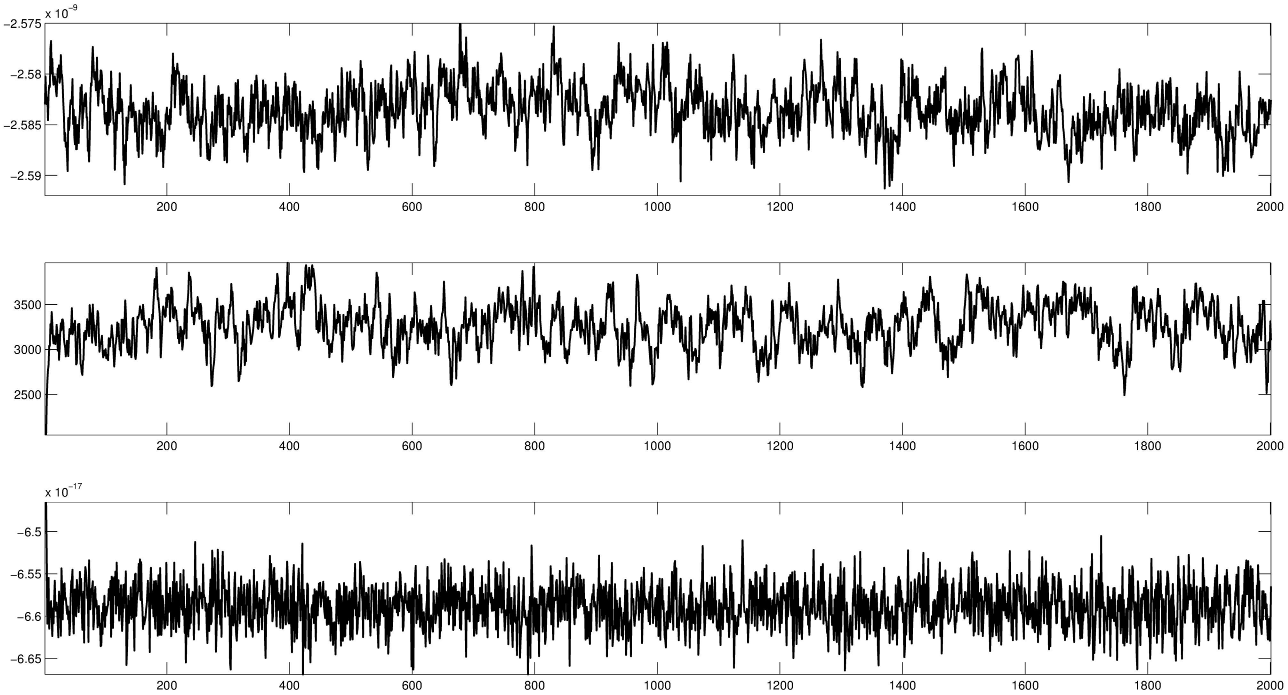

Fig 3. Trace plots for three parameters: Parameter β_1 from the surface model (20)—Top plot; a wavelet coefficient from the basal model—Middle plot and a Fourier coefficient from the flow model (19).

https://doi.org/10.1371/journal.pone.0173453.g003

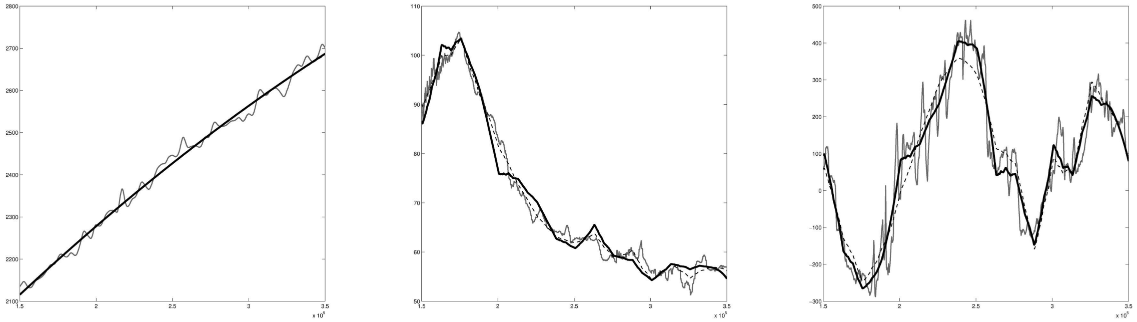

Fig 4. Data (light grey) and posterior means for the diffusion MCMC (solid black lines) and adaptive MCMC (dashed lines). Left panel shows the surface data, middle panel show the velocity data, right panel shows the basal data.

https://doi.org/10.1371/journal.pone.0173453.g004

posterior means (dashed lines) for the three processes from a much longer adaptive MCMC run. This plot confirms that diffusion MCMC performs as expected, giving similar results to the adaptive MCMC approach with the added benefit of a much shorter computing time.

Conclusions

Simulation of a diffusion formulated to have stationary distribution coinciding with a target posterior distribution is a viable MCMC method. The approach is comparatively simple to implement since it requires no probability computations such as those needed in Gibbs' sampling nor any accept-reject steps as in Metropolis algorithms. These advantages can be significant in a variety of settings including mixture likelihoods and/or priors, hierarchical models, nonconjugate priors, and nonlinear models.

The key problem that arises in diffusion MCMC is the approximation of the desired continuous time diffusion by a discrete time Markov chain. Our implementations use Euler discretizations. As reviewed in the Introduction, there are results in the literature providing sufficient conditions under which the discrete approximation has a stationary distribution that approximates that of the target, continuous-time diffusion. Though beyond our scope here, selection of the time-step h can be done adaptively, see [6] for some recent theoretical developments in this area.

We implemented diffusion MCMC for a familiar test problem and compared it to an adaptive MCMC procedure. We found that diffusion MCMC out-performed the "state-of-the-art" adaptive MCMC. Next, we implemented the diffusion MCMC approach in a complicated, nonlinear model involving glacial dynamics. Again, we found that our suggested approach performs well, mixing very fast.

In summary, we believe that diffusion MCMC is a valuable addition to the MCMC toolbox. By construction, the DMCMC algorithm has the ability to quickly find important regions of the target distribution, while a classical, even adaptive MCMC, may require longer exploration times (as seen in the glaciological example). It can be applied in great generality and with ease in some complicated contexts for which other MCMC methods are difficult or very time-consuming to implement. DMCMC does carry the baggage of temporal discretization and concern for the quality of the resulting approximation. Nevertheless, the potential power of diffusion MCMC justifies its application and further development.

Supporting information

S1 Dataset. The data set used in this analysis are available in the file S1_Dataset.zip. We provide the surface, basal and velocity data used in this manuscript.
(ZIP)

Author Contributions

Conceptualization: RH RP LMB.

Data curation: RH.

Formal analysis: RH RP LMB.

Investigation: RH RP LMB.

Methodology: RH RP LMB.

Project administration: RH RP LMB.

Software: RH.

Supervision: RH RP LMB.

Validation: RH RP LMB.

Visualization: RH.

Writing – original draft: RH RP LMB.

Writing – review & editing: RH RP LMB.

References

1. Bernardo J. M. and Smith A. F. M., *Bayesian Theory*, Wiley, 1994.
2. Gilks W. R., Richardson S., and Spiegelhalter D. J. (ed.), *Markov chain Monte Carlo in practice*, Chapman and Hall, New York, 1995.
3. Robert C.P. and Casella G., *Monte Carlo Statistical Methods (second edition)*. New York: Springer-Verlag, 2004.
4. Karatzas I. and Shreve S. E., *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York, 1991.
5. Kloeden P. E. and Platen E., *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, Berlin, 1992.
6. Lamba H., Mattingly J.C., Stuart A.M., An adaptive Euler-Maruyama scheme for SDEs: convergence and stability. IMA Journal of Numerical Analysis, 27 (2007), pp. 479–506. https://doi.org/10.1093/imanum/drl032
7. Mattingly J. C., Stuart A. M., Tretyakov M. V., Convergence of numerical time-averaging and stationary measures via Poisson equations. SIAM J. Numer. Anal., 48 (2010), pp. 552–577. https://doi.org/10.1137/090770527
8. Roberts G. O. and Stramer O., Langevin Diffusions and Metropolis-Hastings Algorithms. Methodology and Computing in Applied Probability, 4 (2002), pp. 337–357. https://doi.org/10.1023/A:1023562417138
9. Roberts G. O. and Tweedie R. L., Exponential Convergence of Langevin Distributions and Their Discrete Approximations. Bernoulli, 4 (1996) pp. 341–363. https://doi.org/10.2307/3318418
10. Talay D., Second order discretization schemes of stochastic differential systems for the computation of the invariant law. Stochastics and Stochastics Reports, 29 (1990), pp. 13–36. https://doi.org/10.1080/17442509008833606
11. Girolami M. and Calderhead B., Riemann manifold Langevin and Hamiltonian Monte Carlo methods, J. R. Statist. Soc. B, 73 (2011), pp. 123–214. https://doi.org/10.1111/j.1467-9868.2010.00765.x
12. Bou-Rabee N., Hairer M., Vanden-Eijnden E., Non-asymptotic mixing of the MALA algorithm, IMA Journal of Numerical Analysis, 33, (2012), 80–110. https://doi.org/10.1093/imanum/drs003
13. Gilks W. R., Roberts G. O. and Sahu S. K., Adaptive Markov chain Monte Carlo. J. Amer. Statist. Assoc., 93, (1998), pp. 1045–1054. https://doi.org/10.1080/01621459.1998.10473766
14. Haario H., Saksman E. and Tamminen J., An adaptive Metropolis algorithm. Bernoulli, 7 (2001), pp. 223–242. https://doi.org/10.2307/3318737
15. Roberts G. O. and Rosenthal J. S., Examples of Adaptive MCMC. J. Comp. Graph. Stat., 18 (2009), pp. 349–367. https://doi.org/10.1198/jcgs.2009.06134
16. Berliner L. M., Jezek K., Cressie N., Kim Y., Lam C. Q., and van der Veen C.J., Modeling dynamic controls on ice streams: a Bayesian statistical approach. Journal of Glaciology 54 (2008), pp. 705–714. https://doi.org/10.3189/002214308786570917
17. Bruce A. and Gao H. Y., *Applied wavelet analysis with S-PLUS*, New York, Springer-Verlag, 1996.
18. Vidakovic B. *Statistical Modeling by Wavelets*. John Wiley & Sons, New York, 1999.