MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster

Cotter, Simon and Roberts, Gareth and Stuart, Andrew and White, David

2013

MIMS EPrint: **2014.71**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

Statistical Science
2013, Vol. 28, No. 3, 424–446
DOI: 10.1214/13-STS421

MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster

S. L. Cotter, G. O. Roberts, A. M. Stuart and D. White

Abstract. Many problems arising in applications result in the need to probe a probability distribution for functions. Examples include Bayesian nonparametric statistics and conditioned diffusion processes. Standard MCMC algorithms typically become arbitrarily slow under the mesh refinement dictated by nonparametric description of the unknown function. We describe an approach to modifying a whole range of MCMC methods, applicable whenever the target measure has density with respect to a Gaussian process or Gaussian random field reference measure, which ensures that their speed of convergence is robust under mesh refinement.

Gaussian processes or random fields are fields whose marginal distributions, when evaluated at any finite set of N points, are $\mathbb{R}^N$-valued Gaussians. The algorithmic approach that we describe is applicable not only when the desired probability measure has density with respect to a Gaussian process or Gaussian random field reference measure, but also to some useful non-Gaussian reference measures constructed through random truncation. In the applications of interest the data is often sparse and the prior specification is an essential part of the overall modelling strategy. These Gaussian-based reference measures are a very flexible modelling tool, finding wide-ranging application. Examples are shown in density estimation, data assimilation in fluid mechanics, subsurface geophysics and image registration.

The key design principle is to formulate the MCMC method so that it is, in principle, applicable for functions; this may be achieved by use of proposals based on carefully chosen time-discretizations of stochastic dynamical systems which exactly preserve the Gaussian reference measure. Taking this approach leads to many new algorithms which can be implemented via minor modification of existing algorithms, yet which show enormous speed-up on a wide range of applied problems.

Key words and phrases: MCMC, Bayesian nonparametrics, algorithms, Gaussian random field, Bayesian inverse problems.

S. L. Cotter is Lecturer, School of Mathematics, University of Manchester, M13 9PL, United Kingdom e-mail: simon.cotter@manchester.ac.uk. G. O. Roberts is Professor, Statistics Department, University of Warwick, Coventry, CV4 7AL, United Kingdom. A. M. Stuart is Professor e-mail: a.m.stuart@warwick.ac.uk and D. White is Postdoctoral Research Assistant, Mathematics Department, University of Warwick, Coventry, CV4 7AL, United Kingdom.

1. INTRODUCTION

The use of Gaussian process (or field) priors is widespread in statistical applications (geostatistics [48], nonparametric regression [24], Bayesian emulator modelling [35], density estimation [1] and inverse quantum theory [27] to name but a few substantial areas where they are commonplace). The success of using Gaussian priors to model an unknown function stems largely from the model flexibility they afford, together with recent advances in computational methodology (particularly MCMC for exact likelihood-based methods). In this paper we describe a wide class of statistical problems, and an algorithmic approach to their study, which adds to the growing literature concerning the use of Gaussian process priors. To be concrete, we consider a process $\{u(x); x \in D\}$ for $D \subseteq \mathbb{R}^d$ for some d. In most of the examples we consider here u is not directly observed: it is hidden (or latent) and some complicated nonlinear function of it generates the data at our disposal.

Gaussian processes or random fields are fields whose marginal distributions, when evaluated at any finite set of N points, are $\mathbb{R}^N$-valued Gaussians. Draws from these Gaussian probability distributions can be computed efficiently by a variety of techniques; for expository purposes we will focus primarily on the use of Karhunen–Loéve expansions to construct such draws, but the methods we propose simply require the ability to draw from Gaussian measures and the user may choose an appropriate method for doing so. The Karhunen–Loéve expansion exploits knowledge of the eigenfunctions and eigenvalues of the covariance operator to construct series with random coefficients which are the desired draws; it is introduced in Section 3.1.

Gaussian processes [2] can be characterized by either the covariance or inverse covariance (precision) operator. In most statistical applications, the covariance is specified. This has the major advantage that the distribution can be readily marginalized to suit a prescribed statistical use. For instance, in geostatistics it is often enough to consider the joint distribution of the process at locations where data is present. However, the inverse covariance specification has particular advantages in the interpretability of parameters when there is information about the local structure of u. (E.g., hence the advantages of using Markov random field models in image analysis.) In the context where x varies over a continuum (such as ours) this creates particular computational difficulties since we can no longer work with a projected prior chosen to reflect available data and quantities of interest [e.g., $\{u(x_i); 1 \le i \le m\}$ say]. Instead it is necessary to consider the entire distribution of $\{u(x); x \in D\}$. This poses major computational challenges, particularly in avoiding unsatisfactory compromises between model approximation (discretization in x typically) and computational cost.

There is a growing need in many parts of applied mathematics to blend data with sophisticated models involving nonlinear partial and/or stochastic differential equations (PDEs/SDEs). In particular, credible mathematical models must respect physical laws and/or Markov conditional independence relationships, which are typically expressed through differential equations. Gaussian priors arises naturally in this context for several reasons. In particular: (i) they allow for straightforward enforcement of differentiability properties, adapted to the model setting; and (ii) they allow for specification of prior information in a manner which is well-adapted to the computational tools routinely used to solve the differential equations themselves. Regarding (ii), it is notable that in many applications it may be computationally convenient to adopt an inverse covariance (precision) operator specification, rather than specification through the covariance function; this allows not only specification of Markov conditional independence relationships but also the direct use of computational tools from numerical analysis [45].

This paper will consider MCMC based computational methods for simulating from distributions of the type described above. Although our motivation is primarily to nonparametric Bayesian statistical applications with Gaussian priors, our approach can be applied to other settings, such as conditioned diffusion processes. Furthermore, we also study some generalizations of Gaussian priors which arise from truncation of the Karhunen–Loéve expansion to a random number of terms; these can be useful to prevent overfitting and allow the data to automatically determine the scales about which it is informative.

Since in nonparametric Bayesian problems the unknown of interest (a function) naturally lies in an infinite-dimensional space, numerical schemes for evaluating posterior distributions almost always rely on some kind of finite-dimensional approximation or truncation to a parameter space of dimension d_u, say. The Karhunen–Loéve expansion provides a natural and mathematically well-studied approach to this problem. The larger d_u is, the better the ap-

proximation to the infinite-dimensional *true* model becomes. However, *off-the-shelf* MCMC methodology usually suffers from a curse of dimensionality so that the numbers of iterations required for these methods to converge diverges with d_u. Therefore, we shall aim to devise strategies which are robust to the value of d_u. Our approach will be to devise algorithms which are well-defined mathematically for the infinite-dimensional limit. Typically, then, finite-dimensional approximations of such algorithms possess robust convergence properties in terms of the choice of d_u. An early specialised example of this approach within the context of diffusions is given in [43].

In practice, we shall thus demonstrate that small, but significant, modifications of a variety of standard Markov chain Monte Carlo (MCMC) methods lead to substantial algorithmic speed-up when tackling Bayesian estimation problems for functions defined via density with respect to a Gaussian process prior, when these problems are approximated on a finite-dimensional space of dimension $d_u \gg 1$. Furthermore, we show that the framework adopted encompasses a range of interesting applications.

1.1 Illustration of the Key Idea

Crucial to our algorithm construction will be a detailed understanding of the dominating reference Gaussian measure. Although prior specification might be Gaussian, it is likely that the posterior distribution μ is not. However, the posterior will at least be absolutely continuous with respect to an appropriate Gaussian density. Typically the dominating Gaussian measure can be chosen to be the prior, with the corresponding Radon–Nikodym derivative just being a re-expression of Bayes' formula

$$\frac{d\mu}{d\mu_0}(u) \propto \mathsf{L}(u)$$

for likelihood L and Gaussian dominating measure (prior in this case) μ_0. This framework extends in a natural way to the case where the prior distribution is not Gaussian, but is absolutely continuous with respect to an appropriate Gaussian distribution. In either case we end up with

$$\frac{d\mu}{d\mu_0}(u) \propto \exp(-\Phi(u)) \tag{1.1}$$

for some real-valued *potential* Φ. We assume that μ_0 is a centred Gaussian measure $\mathcal{N}(0,\mathcal{C})$.

The key algorithmic idea underlying all the algorithms introduced in this paper is to consider (stochastic or random) differential equations which preserve μ or μ_0 and then to employ as proposals for Metropolis–Hastings methods specific discretizations of these differential equations which exactly preserve the Gaussian reference measure μ_0 when $\Phi \equiv 0$; thus, the methods do not reject in the trivial case where $\Phi \equiv 0$. This typically leads to algorithms which are minor adjustments of well-known methods, with major algorithmic speed-ups. We illustrate this idea by contrasting the standard random walk method with the pCN algorithm (studied in detail later in the paper) which is a slight modification of the standard random walk, and which arises from the thinking outlined above. To this end, we define

$$I(u) = \Phi(u) + \tfrac{1}{2}\|\mathcal{C}^{-1/2}u\|^2 \tag{1.2}$$

and consider the following version of the standard random walk method:

- Set $k=0$ and pick $u^{(0)}$.
- Propose $v^{(k)} = u^{(k)} + \beta\xi^{(k)}, \xi^{(k)} \sim N(0,\mathcal{C})$.
- Set $u^{(k+1)} = v^{(k)}$ with probability $a(u^{(k)}, v^{(k)})$.
- Set $u^{(k+1)} = u^{(k)}$ otherwise.
- $k \to k+1$.

The acceptance probability is defined as

$$a(u,v) = \min\{1, \exp(I(u) - I(v))\}.$$

Here, and in the next algorithm, the noise $\xi^{(k)}$ is independent of the uniform random variable used in the accept–reject step, and this pair of random variables is generated independently for each k, leading to a Metropolis–Hastings algorithm reversible with respect to μ.

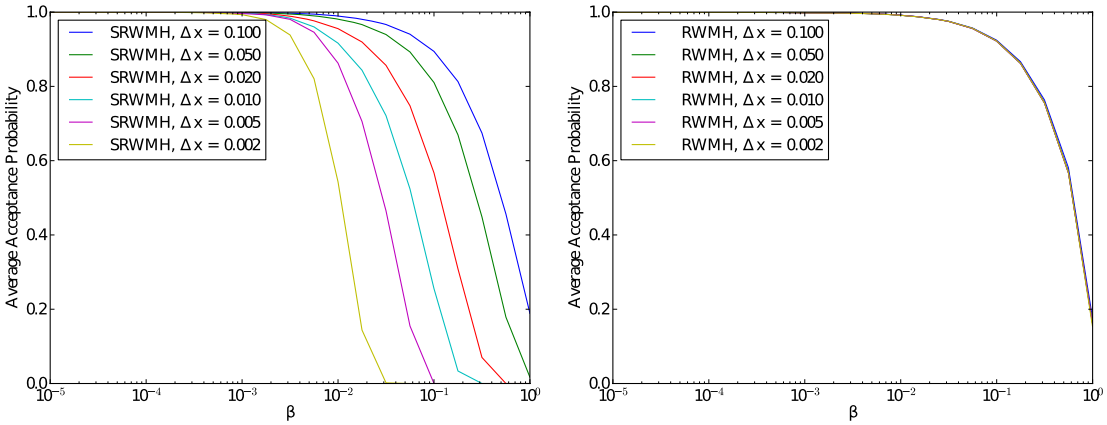
The pCN method is the following modification of the standard random walk method:

- Set $k=0$ and pick $u^{(0)}$.
- Propose $v^{(k)} = \sqrt{(1-\beta^2)}u^{(k)} + \beta\xi^{(k)}, \xi^{(k)} \sim N(0,\mathcal{C})$.
- Set $u^{(k+1)} = v^{(k)}$ with probability $a(u^{(k)}, v^{(k)})$.
- Set $u^{(k+1)} = u^{(k)}$ otherwise.
- $k \to k+1$.

Now we set

$$a(u,v) = \min\{1, \exp(\Phi(u) - \Phi(v))\}.$$

The pCN method differs only slightly from the random walk method: the proposal is not a centred random walk, but rather of AR(1) type, and this results in a modified, slightly simpler, acceptance probability. As is clear, the new method accepts the proposed move with probability one if the potential $\Phi = 0$; this is because the proposal is reversible with respect to the Gaussian reference measure μ_0.



(a) (b)

FIG. 1. *Acceptance probabilities versus mesh-spacing, with* (a) *standard random walk and* (b) *modified random walk (pCN).*

This small change leads to significant speed-ups for problems which are discretized on a grid of dimension d_u. It is then natural to compute on sequences of problems in which the dimension d_u increases, in order to accurately sample the limiting infinite-dimensional problem. The new pCN algorithm is robust to increasing d_u, whilst the standard random walk method is not. To illustrate this idea, we consider an example from the field of data assimilation, introduced in detail in Section 2.2 below, and leading to the need to sample measure μ of the form (1.1). In this problem $d_u = \Delta x^{-2}$, where Δx is the mesh-spacing used in each of the two spatial dimensions.

Figure 1(a) and (b) shows the average acceptance probability curves, as a function of the parameter β appearing in the proposal, computed by the standard and the modified random walk (pCN) methods. It is instructive to imagine running the algorithms when tuned to obtain an average acceptance probability of, say, 0.25. Note that for the standard method, Figure 1(a), the acceptance probability curves shift to the left as the mesh is refined, meaning that smaller proposal variances are required to obtain the same acceptance probability as the mesh is refined. However, for the new method shown in Figure 1(b), the acceptance probability curves have a limit as the mesh is refined and, hence, as the random field model is represented more accurately; thus, a fixed proposal variance can be used to obtain the same acceptance probability at all levels of mesh refinement. The practical implication of this difference in acceptance probability curves is that the number of steps required by the new method is independent of the number of mesh points d_u used to represent the function, whilst for the old random walk method it grows with d_u. The new method thus mixes more rapidly than the standard method and, furthermore, the disparity in mixing rates becomes greater as the mesh is refined.

In this paper we demonstrate how methods such as pCN can be derived, providing a way of thinking about algorithmic development for Bayesian statistics which is transferable to many different situations. The key transferable idea is to use proposals arising from carefully chosen discretizations of stochastic dynamical systems which exactly preserve the Gaussian reference measure. As demonstrated on the example, taking this approach leads to new algorithms which can be implemented via minor modification of existing algorithms, yet which show enormous speed-up on a wide range of applied problems.

1.2 Overview of the Paper

Our setting is to consider measures on function spaces which possess a density with respect to a Gaussian random field measure, or some related non-Gaussian measures. This setting arises in many applications, including the Bayesian approach to inverse problems [49] and conditioned diffusion processes (SDEs) [20]. Our goals in the paper are then fourfold:

- to show that a wide range of problems may be cast in a common framework requiring samples to be drawn from a measure known via its density with respect to a Gaussian random field or, related, prior;

- to explain the principles underlying the derivation of these new MCMC algorithms for functions, leading to desirable d_u-independent mixing properties;
- to illustrate the new methods in action on some nontrivial problems, all drawn from Bayesian nonparametric models where inference is made concerning a function;
- to develop some simple theoretical ideas which give deeper understanding of the benefits of the new methods.

Section 2 describes the common framework into which many applications fit and shows a range of examples which are used throughout the paper. Section 3 is concerned with the reference (prior) measure μ_0 and the assumptions that we make about it; these assumptions form an important part of the model specification and are guided by both modelling and implementation issues. In Section 4 we detail the derivation of a range of MCMC methods on function space, including generalizations of the random walk, MALA, independence samplers, Metropolis-within-Gibbs' samplers and the HMC method. We use a variety of problems to demonstrate the new random walk method in action: Sections 5.1, 5.2, 5.3 and 5.4 include examples arising from density estimation, two inverse problems arising in oceanography and groundwater flow, and the shape registration problem. Section 6 contains a brief analysis of these methods. We make some concluding remarks in Section 7.

Throughout we denote by $\langle\cdot,\cdot\rangle$ the standard Euclidean scalar product on $\mathbb{R}^m$, which induces the standard Euclidean norm $|\cdot|$. We also define $\langle\cdot,\cdot\rangle_C := \langle C^{-1/2}\cdot, C^{-1/2}\cdot\rangle$ for any positive-definite symmetric matrix C; this induces the norm $|\cdot|_C := |C^{-1/2}\cdot|$. Given a positive-definite self-adjoint operator $\mathcal{C}$ on a Hilbert space with inner-product $\langle\cdot,\cdot\rangle$, we will also define the new inner-product $\langle\cdot,\cdot\rangle_{\mathcal{C}} = \langle\mathcal{C}^{-1/2}\cdot,\mathcal{C}^{-1/2}\cdot\rangle$, with resulting norm denoted by $\|\cdot\|_{\mathcal{C}}$ or $|\cdot|_{\mathcal{C}}$.

2. COMMON STRUCTURE

We will now describe a wide-ranging set of examples which fit a common mathematical framework giving rise to a probability measure $\mu(du)$ on a Hilbert space X,[1] when given its density with respect to a random field measure μ_0, also on X. Thus, we have the measure μ as in (1.1) for some *potential* $\Phi : X \to \mathbb{R}$. We assume that Φ can be evaluated to any desired accuracy, by means of a numerical method. *Mesh-refinement* refers to increasing the resolution of this numerical evaluation to obtain a desired accuracy and is tied to the number d_u of basis functions or points used in a finite-dimensional representation of the target function u. For many problems of interest Φ satisfies certain common properties which are detailed in Assumptions 6.1 below. These properties underlie much of the algorithmic development in this paper.

A situation where (1.1) arises frequently is nonparametric density estimation (see Section 2.1), where μ_0 is a random process prior for the unnormalized log density and μ the posterior. There are also many inverse problems in differential equations which have this form (see Sections 2.2, 2.3 and 2.4). For these inverse problems we assume that the data $y \in \mathbb{R}^{d_y}$ is obtained by applying an operator[2] $\mathcal{G}$ to the unknown function u and adding a realisation of a mean zero random variable with density ρ supported on $\mathbb{R}^{d_y}$, thereby determining $\mathbb{P}(y|u)$. That is,

$$y = \mathcal{G}(u) + \eta, \quad \eta \sim \rho. \tag{2.1}$$

After specifying $\mu_0(du) = \mathbb{P}(du)$, Bayes' theorem gives $\mu(dy) = \mathbb{P}(u|y)$ with $\Phi(u) = -\ln\rho(y - \mathcal{G}(u))$. We will work mainly with Gaussian random field priors $\mathcal{N}(0,\mathcal{C})$, although we will also consider generalisations of this setting found by random truncation of the Karhunen–Loéve expansion of a Gaussian random field. This leads to non-Gaussian priors, but much of the methodology for the Gaussian case can be usefully extended, as we will show.

2.1 Density Estimation

Consider the problem of estimating the probability density function $\rho(x)$ of a random variable supported on $[-\ell,\ell]$, given d_y i.i.d. observations y_i. To ensure positivity and normalisation, we may write

$$\rho(x) = \frac{\exp(u(x))}{\int_{-\ell}^{\ell}\exp(u(s))\,ds}. \tag{2.2}$$

If we place a Gaussian process prior μ_0 on u and apply Bayes' theorem, then we obtain formula (1.1) with $\Phi(u) = -\sum_{i=1}^{d_y}\ln\rho(y_i)$ and ρ given by (2.2).

[1] Extension to Banach space is also possible.

[2] This operator, mapping the unknown function to the measurement space, is sometimes termed the observation operator in the applied literature; however, we do not use that terminology in the paper.

2.2 Data Assimilation in Fluid Mechanics

In weather forecasting and oceanography it is frequently of interest to determine the initial condition u for a PDE dynamical system modelling a fluid, given observations [3, 26]. To gain insight into such problems, we consider a model of incompressible fluid flow, namely, either the Stokes ($\gamma = 0$) or Navier–Stokes equation ($\gamma = 1$), on a two-dimensional unit torus $\mathbb{T}^2$. In the following $v(\cdot, t)$ denotes the velocity field at time t, u the initial velocity field and $p(\cdot, t)$ the pressure field at time t and the following is an implicit nonlinear equation for the pair (v, p):

$$\begin{aligned} \partial_t v - \nu \triangle v + \gamma v \cdot \nabla v + \nabla p &= \psi \\ &\forall (x,t) \in \mathbb{T}^2 \times (0,\infty), \\ \nabla \cdot v &= 0 \quad \forall t \in (0,\infty), \\ v(x,0) &= u(x), \quad x \in \mathbb{T}^2. \end{aligned} \tag{2.3}$$

The aim in many applications is to determine the initial state of the fluid velocity, the function u, from some observations relating to the velocity field v at later times.

A simple model of the situation arising in weather forecasting is to determine v from *Eulerian data* of the form $y = \{y_{j,k}\}_{j,k=1}^{N,M}$, where

$$y_{j,k} \sim \mathcal{N}(v(x_j, t_k), \Gamma). \tag{2.4}$$

Thus, the inverse problem is to find u from y of the form (2.1) with $\mathcal{G}_{j,k}(u) = v(x_j, t_k)$.

In oceanography *Lagrangian data* is often encountered: data is gathered from the trajectories of particles $z_j(t)$ moving in the velocity field of interest, and thus satisfying the integral equation

$$z_j(t) = z_{j,0} + \int_0^t v(z_j(s), s)\, ds. \tag{2.5}$$

Data is of the form

$$y_{j,k} \sim \mathcal{N}(z_j(t_k), \Gamma). \tag{2.6}$$

Thus, the inverse problem is to find u from y of the form (2.1) with $\mathcal{G}_{j,k}(u) = z_j(t_k)$.

2.3 Groundwater Flow

In the study of groundwater flow an important inverse problem is to determine the permeability k of the subsurface rock from measurements of the head (water table height) p [30]. To ensure the (physically required) positivity of k, we write $k(x) = \exp(u(x))$ and recast the inverse problem as one for the function u. The head p solves the PDE

$$\begin{aligned} -\nabla \cdot (\exp(u) \nabla p) &= g, \quad x \in D, \\ p &= h, \quad x \in \partial D. \end{aligned} \tag{2.7}$$

Here D is a domain containing the measurement points x_i and ∂D its boundary; in the simplest case g and h are known. The forward solution operator is $\mathcal{G}(u)_j = p(x_j)$. The inverse problem is to find u, given y of the form (2.1).

2.4 Image Registration

In many applications arising in medicine and security it is of interest to calculate the distance between a curve Γ_{obs}, given only through a finite set of noisy observations, and a curve Γ_{db} from a database of known outcomes. As we demonstrate below, this may be recast as an inverse problem for two functions, the first, η, representing reparameterisation of the database curve Γ_{db} and the second, p, representing a momentum variable, normal to the curve Γ_{db}, which initiates a dynamical evolution of the reparameterized curve in an attempt to match observations of the curve Γ_{obs}. This approach to inversion is described in [7] and developed in the Bayesian context in [9]. Here we outline the methodology.

Suppose for a moment that we know the entire observed curve Γ_{obs} and that it is noise free. We parameterize Γ_{db} by q_{db} and Γ_{obs} by q_{obs}, $s \in [0,1]$. We wish to find a path $q(s,t)$, $t \in [0,1]$, between Γ_{db} and Γ_{obs}, satisfying

$$q(s,0) = q_{\text{db}}(\eta(s)), \quad q(s,1) = q_{\text{obs}}(s), \tag{2.8}$$

where η is an orientation-preserving reparameterisation. Following the methodology of [17, 31, 52], we constrain the motion of the curve $q(s,t)$ by asking that the evolution between the two curves results from the differential equation

$$\frac{\partial}{\partial t} q(s,t) = v(q(s,t), t). \tag{2.9}$$

Here $v(x,t)$ is a time-parameterized family of vector fields on $\mathbb{R}^2$ chosen as follows. We define a metric on the "length" of paths as

$$\int_0^1 \frac{1}{2} \|v\|_B^2 \, dt, \tag{2.10}$$

where B is some appropriately chosen Hilbert space. The dynamics (2.9) are defined by choosing an appropriate v which minimizes this metric, subject to the end point constraints (2.8).

In [7] it is shown that this minimisation problem can be solved via a dynamical system obtained from the Euler–Lagrange equation. This dynamical

system yields $q(s,1) = G(p,\eta,s)$, where p is an initial momentum variable normal to Γ_{db}, and η is the reparameterisation. In the perfectly observed scenario the optimal values of $u = (p,\eta)$ solve the equation $G(u,s) := G(p,\eta,s) = q_{\text{obs}}(s)$.

In the partially and noisily observed scenario we are given observations

$$\begin{aligned} y_j &= q_{\text{obs}}(s_j) + \eta_j \\ &= G(u,s_j) + \eta_j \end{aligned}$$

for $j = 1,\ldots,J$; the η_j represent noise. Thus, we have data in the form (2.1) with $\mathcal{G}_j(u) = G(u,s_j)$. The inverse problem is to find the distributions on p and η, given a prior distribution on them, a distribution on η and the data y.

2.5 Conditioned Diffusions

The preceding examples all concern Bayesian nonparametric formulation of inverse problems in which a Gaussian prior is adopted. However, the methodology that we employ readily extends to any situation in which the target distribution is absolutely continuous with respect to a reference Gaussian field law, as arises for certain conditioned diffusion processes [20]. The objective in these problems is to find $u(t)$ solving the equation

$$du(t) = f(u(t))\,dt + \gamma\,dB(t),$$

where B is a Brownian motion and where u is conditioned on, for example, (i) end-point constraints (bridge diffusions, arising in econometrics and chemical reactions); (ii) observation of a single sample path $y(t)$ given by

$$dy(t) = g(u(t))\,dt + \sigma\,dW(t)$$

for some Brownian motion W (continuous time signal processing); or (iii) discrete observations of the path given by

$$y_j = h(u(t_j)) + \eta_j.$$

For all three problems use of the Girsanov formula, which allows expression of the density of the path-space measure arising with nonzero drift in terms of that arising with zero-drift, enables all three problems to be written in the form (1.1).

3. SPECIFICATION OF THE REFERENCE MEASURE

The class of algorithms that we describe are primarily based on measures defined through density with respect to random field model $\mu_0 = \mathcal{N}(0,\mathcal{C})$, denoting a centred Gaussian with covariance operator $\mathcal{C}$. To be able to implement the algorithms in this paper in an efficient way, it is necessary to make assumptions about this Gaussian reference measure. We assume that information about μ_0 can be obtained in at least one of the following three ways:

1. the eigenpairs (ϕ_i,λ_i^2) of $\mathcal{C}$ are known so that exact draws from μ_0 can be made from truncation of the Karhunen–Loéve expansion and that, furthermore, efficient methods exist for evaluation of the resulting sum (such as the FFT);
2. exact draws from μ_0 can be made on a mesh, for example, by building on exact sampling methods for Brownian motion or the stationary Ornstein–Uhlenbeck (OU) process or other simple Gaussian process priors;
3. the precision operator $\mathcal{L} = \mathcal{C}^{-1}$ is known and efficient numerical methods exist for the inversion of $(I + \zeta\mathcal{L})$ for $\zeta > 0$.

These assumptions are not mutually exclusive and for many problems two or more of these will be possible. Both precision and Karhunen–Loéve representations link naturally to efficient computational tools that have been developed in numerical analysis. Specifically, the precision operator $\mathcal{L}$ is often defined via differential operators and the operator $(I + \zeta\mathcal{L})$ can be approximated, and efficiently inverted, by finite element or finite difference methods; similarly, the Karhunen–Loéve expansion links naturally to the use of spectral methods. The book [45] describes the literature concerning methods for sampling from Gaussian random fields, and links with efficient numerical methods for inversion of differential operators. An early theoretical exploration of the links between numerical analysis and statistics is undertaken in [14]. The particular links that we develop in this paper are not yet fully exploited in applications and we highlight the possibility of doing so.

3.1 The Karhunen–Loéve Expansion

The book [2] introduces the Karhunen–Loéve expansion and its properties. Let $\mu_0 = \mathcal{N}(0,\mathcal{C})$ denote a Gaussian measure on a Hilbert space X. Recall that the orthonormalized eigenvalue/eigenfunction pairs of $\mathcal{C}$ form an orthonormal basis for X and solve the problem

$$\mathcal{C}\phi_i = \lambda_i^2\phi_i, \quad i = 1,2,\ldots.$$

Furthermore, we assume that the operator is *trace-class*:

$$\sum_{i=1}^{\infty} \lambda_i^2 < \infty. \tag{3.1}$$

Draws from the centred Gaussian measure μ_0 can then be made as follows. Let $\{\xi_i\}_{i=1}^{\infty}$ denote an independent sequence of normal random variables with distribution $\mathcal{N}(0, \lambda_i^2)$ and consider the random function

$$u(x) = \sum_{i=1}^{\infty} \xi_i \phi_i(x). \tag{3.2}$$

This series converges in $L^2(\Omega; X)$ under the trace-class condition (3.1). It is sometimes useful, both conceptually and for purposes of implementation, to think of the unknown function u as being the infinite sequence $\{\xi_i\}_{i=1}^{\infty}$, rather than the function with these expansion coefficients.

We let $\mathcal{P}_d$ denote projection onto the first d modes[3] $\{\phi_i\}_{i=1}^{d}$ of the Karhunen–Loéve basis. Thus,

$$\mathcal{P}^{d_u} u(x) = \sum_{i=1}^{d_u} \xi_i \phi_i(x). \tag{3.3}$$

If the series (3.3) can be summed quickly on a grid, then this provides an efficient method for computing exact samples from truncation of μ_0 to a finite-dimensional space. When we refer to *mesh-refinement* then, in the context of the prior, this refers to increasing the number of terms d_u used to represent the target function u.

3.2 Random Truncation and Sieve Priors

Non-Gaussian priors can be constructed from the Karhunen–Loéve expansion (3.3) by allowing d_u itself to be a random variable supported on $\mathbb{N}$; we let $p(i) = \mathbb{P}(d_u = i)$. Much of the methodology in this paper can be extended to these priors. A draw from such a prior measure can be written as

$$u(x) = \sum_{i=1}^{\infty} \mathbb{I}(i \le d_u) \xi_i \phi_i(x), \tag{3.4}$$

where $\mathbb{I}(i \in E)$ is the indicator function. We refer to this as *random truncation prior*. Functions drawn from this prior are non-Gaussian and almost surely C^{∞}. However, expectations with respect to d_u will be Gaussian and can be less regular: they are given by the formula

$$\mathbb{E}^{d_u} u(x) = \sum_{i=1}^{\infty} \alpha_i \xi_i \phi_i(x), \tag{3.5}$$

where $\alpha_i = \mathbb{P}(d_u \ge i)$. As in the Gaussian case, it can be useful, both conceptually and for purposes of implementation, to think of the unknown function u as being the infinite vector $(\{\xi_i\}_{i=1}^{\infty}, d_u)$ rather than the function with these expansion coefficients.

Making d_u a random variable has the effect of switching on (nonzero) and off (zero) coefficients in the expansion of the target function. This formulation switches the basis functions on and off in a fixed order. Random truncation as expressed by equation (3.4) is not the only variable dimension formulation. In dimension greater than one we will employ the *sieve prior* which allows every basis function to have an individual on/off switch. This prior relaxes the constraint imposed on the order in which the basis functions are switched on and off and we write

$$u(x) = \sum_{i=1}^{\infty} \chi_i \xi_i \phi_i(x), \tag{3.6}$$

where $\{\chi_i\}_{i=1}^{\infty} \in \{0, 1\}$. We define the distribution on $\chi = \{\chi_i\}_{i=1}^{\infty}$ as follows. Let ν_0 denote a reference measure formed from considering an i.i.d. sequence of Bernoulli random variables with success probability one half. Then define the prior measure ν on χ to have density

$$\frac{d\nu}{d\nu_0}(\chi) \propto \exp\left(-\lambda \sum_{i=1}^{\infty} \chi_i\right),$$

where $\lambda \in \mathbb{R}^+$. As for the random truncation method, it is both conceptually and practically valuable to think of the unknown function as being the pair of random infinite vectors $\{\xi_i\}_{i=1}^{\infty}$ and $\{\chi_i\}_{i=1}^{\infty}$. Hierarchical priors, based on Gaussians but with random switches in front of the coefficients, are termed "sieve priors" in [54]. In that paper posterior consistency questions for linear regression are also analysed in this setting.

4. MCMC METHODS FOR FUNCTIONS

The transferable idea in this section is that design of MCMC methods which are defined on function spaces leads, after discretization, to algorithms which are robust under mesh refinement $d_u \to \infty$. We demonstrate this idea for a number of algorithms,

[3]Note that "mode" here, denoting an element of a basis in a Hilbert space, differs from the "mode" of a distribution.

generalizing random walk and Langevin-based Metropolis–Hastings methods, the independence sampler, the Gibbs sampler and the HMC method; we anticipate that many other generalisations are possible. In all cases the proposal exactly preserves the Gaussian reference measure μ_0 when the potential Φ is zero and the reader may take this key idea as a design principle for similar algorithms.

Section 4.1 gives the framework for MCMC methods on a general state space. In Section 4.2 we state and derive the new *Crank–Nicolson* proposals, arising from discretization of an OU process. In Section 4.3 we generalize these proposals to the Langevin setting where steepest descent information is incorporated: *MALA proposals*. Section 4.4 is concerned with *Independence Samplers* which may be derived from particular parameter choices in the random walk algorithm. Section 4.5 introduces the idea of randomizing the choice of δ as part of the proposal which is effective for the random walk methods. In Section 4.6 we introduce *Gibbs samplers* based on the Karhunen–Loéve expansion (3.2). In Section 4.7 we work with non-Gaussian priors specified through random truncation of the Karhunen–Loéve expansion as in (3.4), showing how Gibbs samplers can again be used in this situation. Section 4.8 briefly describes the HMC method and its generalisation to sampling functions.

4.1 Set-Up

We are interested in defining MCMC methods for measures μ on a Hilbert space $(X, \langle\cdot,\cdot\rangle)$, with induced norm $\|\cdot\|$, given by (1.1) where $\mu_0 = \mathcal{N}(0,\mathcal{C})$. The setting we adopt is that given in [51] where Metropolis–Hastings methods are developed in a general state space. Let $q(u,\cdot)$ denote the transition kernel on X and $\eta(du,dv)$ denote the measure on $X\times X$ found by taking $u\sim\mu$ and then $v|u\sim q(u,\cdot)$. We use $\eta^\perp(u,v)$ to denote the measure found by reversing the roles of u and v in the preceding construction of η. If $\eta^\perp(u,v)$ is equivalent (in the sense of measures) to $\eta(u,v)$, then the Radon–Nikodym derivative $\frac{d\eta^\perp}{d\eta}(u,v)$ is well-defined and we may define the *acceptance probability*

$$a(u,v) = \min\left\{1, \frac{d\eta^\perp}{d\eta}(u,v)\right\}. \tag{4.1}$$

We accept the proposed move from u to v with this probability. The resulting Markov chain is μ-reversible.

A key idea underlying the new variants on random walk and Langevin-based Metropolis–Hastings algorithms derived below is to use discretizations of stochastic partial differential equations (SPDEs) which are invariant for either the reference or the target measure. These SPDEs have the form, for $\mathcal{L} = \mathcal{C}^{-1}$ the precision operator for μ_0, and $D\Phi$ the derivative of potential Φ,

$$\frac{du}{ds} = -\mathcal{K}(\mathcal{L}u + \gamma D\Phi(u)) + \sqrt{2\mathcal{K}}\frac{db}{ds}. \tag{4.2}$$

Here b is a Brownian motion in X with covariance operator the identity and $\mathcal{K} = \mathcal{C}$ or I. Since $\mathcal{K}$ is a positive operator, we may define the square-root in the symmetric fashion, via diagonalization in the Karhunen–Loéve basis of $\mathcal{C}$. We refer to it as an SPDE because in many applications $\mathcal{L}$ is a differential operator. The SPDE has invariant measure μ_0 for $\gamma = 0$ (when it is an infinite-dimensional OU process) and μ for $\gamma = 1$ [12, 18, 22]. The target measure μ will behave like the reference measure μ_0 on high frequency (rapidly oscillating) functions. Intuitively, this is because the data, which is finite, is not informative about the function on small scales; mathematically, this is manifest in the absolute continuity of μ with respect to μ_0 given by formula (1.1). Thus, discretizations of equation (4.2) with either $\gamma = 0$ or $\gamma = 1$ form sensible candidate proposal distributions.

The basic idea which underlies the algorithms described here was introduced in the specific context of conditioned diffusions with $\gamma = 1$ in [50], and then generalized to include the case $\gamma = 0$ in [4]; furthermore, the paper [4], although focussed on the application to conditioned diffusions, applies to general targets of the form (1.1). The papers [4, 50] both include numerical results illustrating applicability of the method to conditioned diffusion in the case $\gamma = 1$, and the paper [10] shows application to data assimilation with $\gamma = 0$. Finally, we mention that in [33] the algorithm with $\gamma = 0$ is mentioned, although the derivation does not use the SPDE motivation that we develop here, and the concept of a nonparametric limit is not used to motivate the construction.

4.2 Vanilla Local Proposals

The *standard random walk* proposal for $v|u$ takes the form

$$v = u + \sqrt{2\delta\mathcal{K}}\xi_0 \tag{4.3}$$

for any $\delta\in[0,\infty)$, $\xi_0\sim\mathcal{N}(0,I)$ and $\mathcal{K} = I$ or $\mathcal{K} = \mathcal{C}$. This can be seen as a discrete skeleton of (4.2) after

ignoring the drift terms. Therefore, such a proposal leads to an infinite-dimensional version of the well-known random walk Metropolis algorithm.

The random walk proposal in finite-dimensional problems always leads to a well-defined algorithm and rarely encounters any reducibility problems [46]. Therefore, this method can certainly be applied for arbitrarily fine mesh size. However, taking this approach does not lead to a well-defined MCMC method for *functions*. This is because $\eta^{\perp}$ is singular with respect to η so that all proposed moves are rejected with probability 1. (We prove this in Theorem 6.3 below.) Returning to the finite mesh case, algorithm mixing time therefore increases to ∞ as $d_u \to \infty$.

To define methods with convergence properties robust to increasing d_u, alternative approaches leading to well-defined and irreducible algorithms on the Hilbert space need to be considered. We consider two possibilities here, both based on Crank–Nicolson approximations [38] of the linear part of the drift. In particular, we consider discretization of equation (4.2) with the form

$$v = u - \tfrac{1}{2}\delta\mathcal{K}\mathcal{L}(u+v) - \delta\gamma\mathcal{K}D\Phi(u) + \sqrt{2\mathcal{K}\delta}\,\xi_0 \tag{4.4}$$

for a (spatial) white noise ξ_0.

First consider the discretization (4.4) with $\gamma = 0$ and $\mathcal{K} = I$. Rearranging shows that the resulting *Crank–Nicolson proposal* (CN) for $v|u$ is found by solving

$$(I + \tfrac{1}{2}\delta\mathcal{L})v = (I - \tfrac{1}{2}\delta\mathcal{L})u + \sqrt{2\delta}\,\xi_0. \tag{4.5}$$

It is this form that the proposal is best implemented whenever the prior/reference measure μ_0 is specified via the precision operator $\mathcal{L}$ and when efficient algorithms exist for inversion of the identity plus a multiple of $\mathcal{L}$. However, for the purposes of analysis it is also useful to write this equation in the form

$$(2\mathcal{C} + \delta I)v = (2\mathcal{C} - \delta I)u + \sqrt{8\delta\mathcal{C}}\,w, \tag{4.6}$$

where $w \sim \mathcal{N}(0,\mathcal{C})$, found by applying the operator $2\mathcal{C}$ to equation (4.5).

A well-established principle in finite-dimensional sampling algorithms advises that proposal variance should be approximately a scalar multiple of that of the target (see, e.g., [42]). The variance in the prior, $\mathcal{C}$, can provide a reasonable approximation, at least as far as controlling the large d_u limit is concerned. This is because the data (or change of measure) is typically only informative about a finite set of components in the prior model; mathematically, the fact that the posterior has density with respect to the prior means that it "looks like" the prior in the large i components of the Karhunen–Loéve expansion.[4]

The CN algorithm violates this principle: the proposal variance operator is proportional to $(2\mathcal{C}+\delta I)^{-2}\cdot\mathcal{C}^2$, suggesting that algorithm efficiency might be improved still further by obtaining a proposal variance of $\mathcal{C}$. In the familiar finite-dimensional case, this can be achieved by a standard *reparameterisation* argument which has its origins in [23] if not before. This motivates our final local proposal in this subsection.

The *preconditioned CN* proposal (pCN) for $v|u$ is obtained from (4.4) with $\gamma = 0$ and $\mathcal{K} = \mathcal{C}$ giving the proposal

$$(2+\delta)v = (2-\delta)u + \sqrt{8\delta}\,w, \tag{4.7}$$

where, again, $w \sim \mathcal{N}(0,\mathcal{C})$. As discussed after (4.5), and in Section 3, there are many different ways in which the prior Gaussian may be specified. If the specification is via the precision $\mathcal{L}$ and if there are numerical methods for which $(I + \zeta\mathcal{L})$ can be efficiently inverted, then (4.5) is a natural proposal. If, however, sampling from $\mathcal{C}$ is straightforward (via the Karhunen–Loéve expansion or directly), then it is natural to use the proposal (4.7), which requires only that it is possible to draw from μ_0 efficiently. For $\delta \in [0,2]$ the proposal (4.7) can be written as

$$v = (1-\beta^2)^{1/2}u + \beta w, \tag{4.8}$$

where $w \sim \mathcal{N}(0,\mathcal{C})$, and $\beta \in [0,1]$; in fact, $\beta^2 = 8\delta/(2+\delta)^2$. In this form we see very clearly a simple generalisation of the finite-dimensional random walk given by (4.3) with $\mathcal{K} = \mathcal{C}$.

The numerical experiments described in Section 1.1 show that the pCN proposal significantly improves upon the naive random walk method (4.3), and similar positive results can be obtained for the CN method. Furthermore, for both the proposals (4.5) and (4.7) we show in Theorem 6.2 that $\eta^{\perp}$ and η are equivalent (as measures) by showing that they are both equivalent to the same Gaussian reference measure η_0, whilst in Theorem 6.3 we show that the

[4] An interesting research problem would be to combine the ideas in [16], which provide an adaptive preconditioning but are only practical in a finite number of dimensions, with the prior-based fixed preconditioning used here. Note that the method introduced in [16] reduces exactly to the preconditioning used here in the absence of data.

proposal (4.3) leads to mutually singular measures $\eta^\perp$ and η. This theory explains the numerical observations and motivates the importance of designing algorithms directly on function space.

The accept–reject formula for CN and pCN is very simple. If, for some $\rho : X \times X \to \mathbb{R}$, and some reference measure η_0,

$$\begin{aligned} \frac{d\eta}{d\eta_0}(u,v) &= Z\exp(-\rho(u,v)), \\ \frac{d\eta^\perp}{d\eta_0}(u,v) &= Z\exp(-\rho(v,u)), \end{aligned} \tag{4.9}$$

it then follows that

$$\frac{d\eta^\perp}{d\eta}(u,v) = \exp(\rho(u,v)-\rho(v,u)). \tag{4.10}$$

For both CN proposals (4.5) and (4.7) we show in Theorem 6.2 below that, for appropriately defined η_0, we have $\rho(u,v) = \Phi(u)$ so that the acceptance probability is given by

$$a(u,v) = \min\{1, \exp(\Phi(u)-\Phi(v))\}. \tag{4.11}$$

In this sense the CN and pCN proposals may be seen as the *natural generalisations of random walks* to the setting where the target measure is defined via density with respect to a Gaussian, as in (1.1). This point of view may be understood by noting that the accept/reject formula is defined entirely through differences in this log density, as happens in finite dimensions for the standard random walk, if the density is specified with respect to the Lebesgue measure. Similar random truncation priors are used in non-parametric inference for drift functions in diffusion processes in [53].

4.3 MALA Proposal Distributions

The CN proposals (4.5) and (4.7) contain no information about the potential Φ given by (1.1); they contain only information about the reference measure μ_0. Indeed, they are derived by discretizing the SDE (4.2) in the case $\gamma = 0$, for which μ_0 is an invariant measure. The idea behind the Metropolis-adjusted Langevin (MALA) proposals (see [39, 44] and the references therein) is to discretize an equation which is invariant for the measure μ. Thus, to construct such proposals in the function space setting, we discretize the SPDE (4.2) with $\gamma = 1$. Taking $\mathcal{K} = I$ and $\mathcal{K} = \mathcal{C}$ then gives the following two proposals.

The *Crank–Nicolson Langevin proposal* (CNL) is given by

$$\begin{aligned} (2\mathcal{C}+\delta)v &= (2\mathcal{C}-\delta)u - 2\delta\mathcal{C}\mathcal{D}\Phi(u) \\ &\quad + \sqrt{8\delta\mathcal{C}}w, \end{aligned} \tag{4.12}$$

where, as before, $w \sim \mu_0 = \mathcal{N}(0,\mathcal{C})$. If we define

$$\begin{aligned} \rho(u,v) = \Phi(u) &+ \frac{1}{2}\langle v-u, \mathcal{D}\Phi(u)\rangle \\ &+ \frac{\delta}{4}\langle \mathcal{C}^{-1}(u+v), \mathcal{D}\Phi(u)\rangle \\ &+ \frac{\delta}{4}\|\mathcal{D}\Phi(u)\|^2, \end{aligned}$$

then the acceptance probability is given by (4.1) and (4.10). Implementation of this proposal simply requires inversion of $(I+\zeta\mathcal{L})$, as for (4.5). The CNL method is the special case $\theta = \frac{1}{2}$ for the IA algorithm introduced in [4].

The *preconditioned Crank–Nicolson Langevin proposal* (pCNL) is given by

$$(2+\delta)v = (2-\delta)u - 2\delta\mathcal{C}\mathcal{D}\Phi(u) + \sqrt{8\delta}w, \tag{4.13}$$

where w is again a draw from μ_0. Defining

$$\begin{aligned} \rho(u,v) = \Phi(u) &+ \frac{1}{2}\langle v-u, \mathcal{D}\Phi(u)\rangle \\ &+ \frac{\delta}{4}\langle u+v, \mathcal{D}\Phi(u)\rangle \\ &+ \frac{\delta}{4}\|\mathcal{C}^{1/2}\mathcal{D}\Phi(u)\|^2, \end{aligned}$$

the acceptance probability is given by (4.1) and (4.10). Implementation of this proposal requires draws from the reference measure μ_0 to be made, as for (4.7). The pCNL method is the special case $\theta = \frac{1}{2}$ for the PIA algorithm introduced in [4].

4.4 Independence Sampler

Making the choice $\delta = 2$ in the pCN proposal (4.7) gives an *independence sampler*. The proposal is then simply a draw from the prior: $v = w$. The acceptance probability remains (4.11). An interesting generalisation of the independence sampler is to take $\delta = 2$ in the MALA proposal (4.13), giving the proposal

$$v = -\mathcal{C}\mathcal{D}\Phi(u) + w \tag{4.14}$$

with resulting acceptance probability given by (4.1) and (4.10) with

$$\rho(u,v) = \Phi(u) + \langle v, \mathcal{D}\Phi(u)\rangle + \tfrac{1}{2}\|\mathcal{C}^{1/2}\mathcal{D}\Phi(u)\|^2.$$

4.5 Random Proposal Variance

It is sometimes useful to randomise the proposal variance δ in order to obtain better mixing. We discuss this idea in the context of the pCN proposal (4.7). To emphasize the dependence of the proposal kernel on δ, we denote it by $q(u, dv; \delta)$. We show in Section 6.1 that the measure $\eta_0(du, dv) = q(u, dv; \delta)\mu_0(du)$ is well-defined and symmetric in u, v for every $\delta \in [0, \infty)$. If we choose δ at random from any probability distribution ν on $[0, \infty)$, independently from w, then the resulting proposal has kernel

$$q(u, dv) = \int_0^\infty q(u, dv; \delta)\nu(d\delta).$$

Furthermore, the measure $q(u, dv)\mu_0(du)$ may be written as

$$\int_0^\infty q(u, dv; \delta)\mu_0(du)\nu(d\delta)$$

and is hence also symmetric in u, v. Hence, both the CN and pCN proposals (4.5) and (4.7) may be generalised to allow for δ chosen at random independently of u and w, according to some measure ν on $[0, \infty)$. The acceptance probability remains (4.11), as for fixed δ.

4.6 Metropolis-Within-Gibbs: Blocking in Karhunen–Loéve Coordinates

Any function $u \in X$ can be expanded in the Karhunen–Loéve basis and hence written as

$$u(x) = \sum_{i=1}^\infty \xi_i \phi_i(x). \tag{4.15}$$

Thus, we may view the probability measure μ given by (1.1) as a measure on the coefficients $u = \{\xi_i\}_{i=1}^\infty$. For any index set $I \subset \mathbb{N}$ we write $\xi^I = \{\xi_i\}_{i \in I}$ and $\xi^I_- = \{\xi_i\}_{i \notin I}$. Both ξ^I and ξ^I_- are independent and Gaussian under the prior μ_0 with diagonal covariance operators $\mathcal{C}^I$ $\mathcal{C}^I_-$, respectively. If we let μ_0^I denote the Gaussian $\mathcal{N}(0, \mathcal{C}^I)$, then (1.1) gives

$$\frac{d\mu}{d\mu_0^I}(\xi^I | \xi^I_-) \propto \exp(-\Phi(\xi^I, \xi^I_-)), \tag{4.16}$$

where we now view Φ as a function on the coefficients in the expansion (4.15). This formula may be used as the basis for Metropolis-within-Gibbs samplers using blocking with respect to a set of partitions $\{I_j\}_{j=1,\dots,J}$ with the property $\bigcup_{j=1}^J I_j = \mathbb{N}$. Because the formula is defined for functions this will give rise to methods which are robust under mesh refinement when implemented in practice. We have found it useful to use the partitions $I_j = \{j\}$ for $j = 1, \dots, J-1$ and $I_J = \{J, J+1, \dots\}$. On the other hand, standard Gibbs and Metropolis-within-Gibbs samplers are based on partitioning via $I_j = \{j\}$, and do not behave well under mesh-refinement, as we will demonstrate.

4.7 Metropolis-Within-Gibbs: Random Truncation and Sieve Priors

We will also use Metropolis-within-Gibbs to construct sampling algorithms which alternate between updating the coefficients $\xi = \{\xi_i\}_{i=1}^\infty$ in (3.4) or (3.6), and the integer d_u, for (3.4), or the infinite sequence $\chi = \{\chi_i\}_{i=1}^\infty$ for (3.6). In words, we alternate between the coefficients in the expansion of a function and the parameters determining which parameters are active.

If we employ the non-Gaussian prior with draws given by (3.4), then the negative log likelihood Φ can be viewed as a function of (ξ, d_u) and it is natural to consider Metropolis-within-Gibbs methods which are based on the conditional distributions for $\xi | d_u$ and $d_u | \xi$. Note that, under the prior, ξ and d_u are independent with $\xi \sim \mu_{0,\xi} := \mathcal{N}(0, \mathcal{C})$ and $d_u \sim \mu_{0,d_u}$, the latter being supported on $\mathbb{N}$ with $p(i) = \mathbb{P}(d_u = i)$. For fixed d_u we have

$$\frac{d\mu}{d\mu_{0,\xi}}(\xi | d_u) \propto \exp(-\Phi(\xi, d_u)) \tag{4.17}$$

with $\Phi(u)$ rewritten as a function of ξ and d_u via the expansion (3.4). This measure can be sampled by any of the preceding Metropolis–Hastings methods designed in the case with Gaussian μ_0. For fixed ξ we have

$$\frac{d\mu}{d\mu_{0,d_u}}(d_u | \xi) \propto \exp(-\Phi(\xi, d_u)). \tag{4.18}$$

A natural biased random walk for $d_u | \xi$ arises by proposing moves from a random walk on $\mathbb{N}$ which satisfies detailed balance with respect to the distribution $p(i)$. The acceptance probability is then

$$a(u, v) = \min\{1, \exp(\Phi(\xi, d_u) - \Phi(\xi, d_v))\}.$$

Variants on this are possible and, if $p(i)$ is monotonic decreasing, a simple random walk proposal on the integers, with local moves $d_u \to d_v = d_u \pm 1$, is straightforward to implement. Of course, different proposal stencils can give improved mixing properties, but we employ this particular random walk for expository purposes.

If, instead of (3.4), we use the non-Gaussian sieve prior defined by equation (3.6), the prior and posterior measures may be viewed as measures on $u = (\{\xi_i\}_{i=1}^{\infty}, \{\chi_j\}_{j=1}^{\infty})$. These variables may be modified as stated above via Metropolis-within-Gibbs for sampling the conditional distributions $\xi|\chi$ and $\chi|\xi$. If, for example, the proposal for $\chi|\xi$ is reversible with respect to the prior on ξ, then the acceptance probability for this move is given by

$$a(u,v) = \min\{1, \exp(\Phi(\xi_u, \chi_u) - \Phi(\xi_v, \chi_v))\}.$$

In Section 5.3 we implement a slightly different proposal in which, with probability $\frac{1}{2}$, a nonactive mode is switched on with the remaining probability an active mode is switched off. If we define $N_{\text{on}} = \sum_{i=1}^{N} \chi_i$, then the probability of moving from ξ_u to a state ξ_v in which an extra mode is switched on is

$$a(u,v) = \min\left\{1, \exp\left(\Phi(\xi_u, \chi_u) - \Phi(\xi_v, \chi_v) + \frac{N - N_{\text{on}}}{N_{\text{on}}}\right)\right\}.$$

Similarly, the probability of moving to a situation in which a mode is switched off is

$$a(u,v) = \min\left\{1, \exp\left(\Phi(\xi_u, \chi_u) - \Phi(\xi_v, \chi_v) + \frac{N_{\text{on}}}{N - N_{\text{on}}}\right)\right\}.$$

4.8 Hybrid Monte Carlo Methods

The algorithms discussed above have been based on proposals which can be motivated through discretization of an SPDE which is invariant for either the prior measure μ_0 or for the posterior μ itself. HMC methods are based on a different idea, which is to consider a Hamiltonian flow in a state space found from introducing extra "momentum" or "velocity" variables to complement the variable u in (1.1). If the momentum/velocity is chosen randomly from an appropriate Gaussian distribution at regular intervals, then the resulting Markov chain in u is invariant under μ. Discretizing the flow, and adding an accept/reject step, results in a method which remains invariant for μ [15]. These methods can break random-walk type behaviour of methods based on local proposal [32, 34]. It is hence of interest to generalise these methods to the function sampling setting dictated by (1.1) and this is undertaken in [5]. The key novel idea required to design this algorithm is the development of a new integrator for the Hamiltonian flow underlying the method; this integrator is exact in the Gaussian case $\Phi \equiv 0$, on function space, and for this reason behaves well for nonparametric where d_u may be arbitrarily large infinite dimensions.

5. COMPUTATIONAL ILLUSTRATIONS

This section contains numerical experiments designed to illustrate various properties of the sampling algorithms overviewed in this paper. We employ the examples introduced in Section 2.

5.1 Density Estimation

Section 1.1 shows an example which illustrates the advantage of using the function-space algorithms highlighted in this paper in comparison with standard techniques; there we compared pCN with a standard random walk. The first goal of the experiments in this subsection is to further illustrate the advantage of the function-space algorithms over standard algorithms. Specifically, we compare the Metropolis-within-Gibbs method from Section 4.6, based on the partition $I_j = \{j\}$ and labelled MwG here, with the pCN sampler from Section 4.2. The second goal is to study the effect of prior modelling on algorithmic performance; to do this, we study a third algorithm, RTM-pCN, based on sampling the randomly truncated Gaussian prior (3.4) using the Gibbs method from Section 4.7, with the pCN sampler for the coefficient update.

5.1.1 *Target distribution* We will use the true density

$$\rho \propto \mathcal{N}(-3,1)\mathbb{I}(x \in (-\ell, +\ell)) + \mathcal{N}(+3,1)\mathbb{I}(x \in (-\ell, +\ell)),$$

where $\ell = 10$. [Recall that $\mathbb{I}(\cdot)$ denotes the indicator function of a set.] This density corresponds approximately to a situation where there is a 50/50 chance of being in one of the two Gaussians. This one-dimensional multi-modal density is sufficient to expose the advantages of the function spaces samplers pCN and RTM-pCN over MwG.

5.1.2 *Prior* We will make comparisons between the three algorithms regarding their computational performance, via various graphical and numerical measures. In all cases it is important that the reader appreciates that the comparison between MwG and

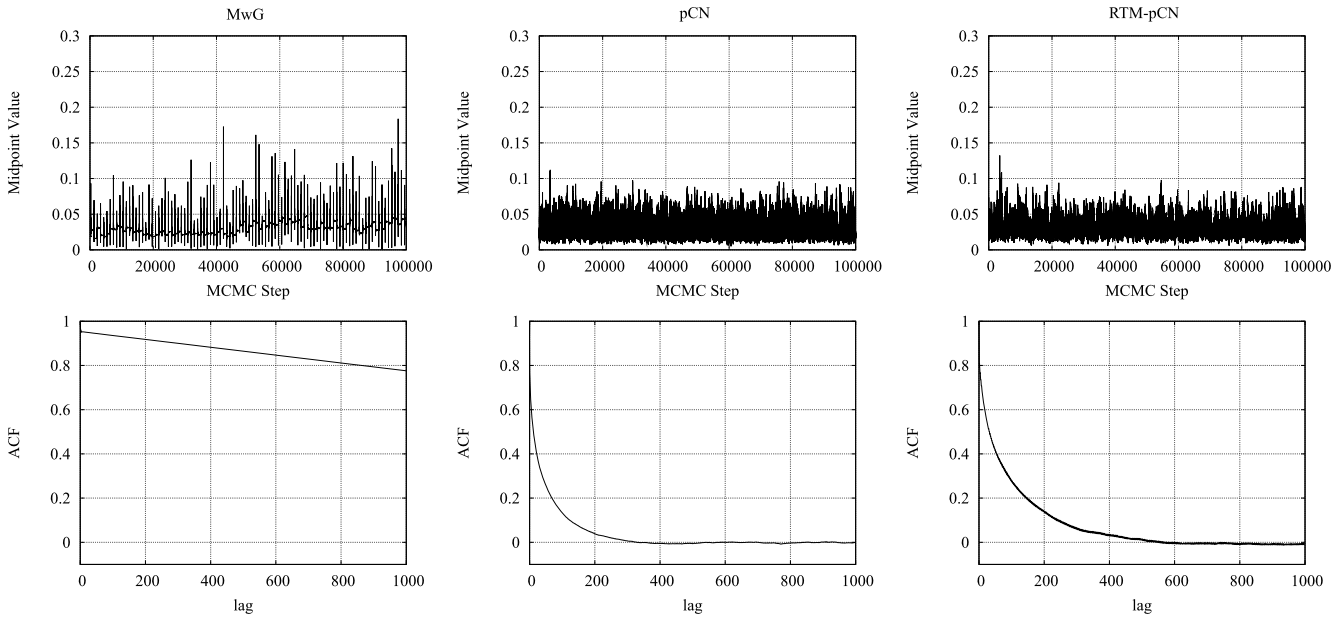


FIG. 2. *Trace and autocorrelation plots for sampling posterior measure with true density ρ using MwG, pCN and RTM-pCN methods.*

pCN corresponds to sampling from the same posterior, since they use the same prior, and that all comparisons between RTM-pCN and other methods also quantify the effect of prior modelling as well as algorithm.

Two priors are used for this experiment: the Gaussian prior given by (3.2) and the randomly truncated Gaussian given by (3.4). We apply the MwG and pCN schemes in the former case, and the RTM-pCN scheme for the latter. The prior uses the same Gaussian covariance structure for the independent ξ, namely, $\xi_i \sim \mathcal{N}(0, \lambda_i^2)$, where $\lambda_i \propto i^{-2}$. Note that the eigenvalues are summable, as required for draws from the Gaussian measure to be square integrable and to be continuous. The prior for the number of active terms d_u is an exponential distribution with rate $\lambda = 0.01$.

5.1.3 *Numerical implementation* In order to facilitate a fair comparison, we tuned the value of δ in the pCN and RTM-pCN proposals to obtain an average acceptance probability of around 0.234, requiring, in both cases, $\delta \approx 0.27$. (For RTM-pCN the average acceptance probability refers only to moves in $\{\xi\}_{i=1}^{\infty}$ and not in d_u.) We note that with the value $\delta = 2$ we obtain the independence sampler for pCN; however, this sampler only accepted 12 proposals out of 10^6 MCMC steps, indicating the importance of tuning δ correctly. For MwG there is no tunable parameter, and we obtain an acceptance of around 0.99.

TABLE 1
Approximate integrated autocorrelation times for target ρ

Algorithm	**IACT**
MwG	894
pCN	73.2
RTM-pCN	143

5.1.4 *Numerical results* In order to compare the performance of pCN, MwG and RTM-pCN, we show, in Figure 2 and Table 1, trace plots, correlation functions and integrated auto-correlation times (the latter are notoriously difficult to compute accurately [47] and displayed numbers to three significant figures should only be treated as indicative). The autocorrelation function decays for ergodic Markov chains, and its integral determines the asymptotic variance of sample path averages. The integrated autocorrelation time is used, via this asymptotic variance, to determine the number of steps required to determine an independent sample from the MCMC method. The figures and integrated autocorrelation times clearly show that the pCN and RTM-pCN outperform MwG by an order of magnitude. This reflects the fact that pCN and RTM-pCN are function space samplers, designed to mix independently of the mesh-size. In contrast, the MwG method is

TABLE 2
Comparison of computational timings for target ρ

Algorithm	Time for 10^6 steps (s)	Time to draw an indep sample (s)
MwG	262	0.234
pCN	451	0.0331
RTM-pCN	278	0.0398

heavily mesh-dependent, since updates are made one Fourier component at a time.

Finally, we comment on the effect of the different priors. The asymptotic variance for the RTM-pCN is approximately double that of pCN. However, RTM-pCN can have a reduced runtime, per unit error, when compared with pCN, as Table 2 shows. This improvement of RTM-pCN over pCN is primarily caused by the reduction in the number of random number generations due to the adaptive size of the basis in which the unknown density is represented.

5.2 Data Assimilation in Fluid Mechanics

We now proceed to a more complex problem and describe numerical results which demonstrate that the function space samplers successfully sample nontrivial problems arising in applications. We study both the Eulerian and Lagrangian data assimilation problems from Section 2.2, for the Stokes flow forward model $\gamma = 0$. It has been demonstrated in [8, 10] that the pCN can successfully sample from the posterior distribution for such problems. In this subsection we will illustrate three features of such methods: convergence of the algorithm from different starting states, convergence with different proposal step sizes, and behaviour with random distributions for the proposal step size, as discussed in Section 4.5.

5.2.1 *Target distributions* In this application we aim to characterize the posterior distribution on the initial condition of the two-dimensional velocity field u_0 for Stokes flow [equation (2.3) with $\gamma = 0$], given a set of either Eulerian (2.4) or Lagrangian (2.6) observations. In both cases, the posterior is of the form (1.1) with $\Phi(u) = \frac{1}{2}\|\mathcal{G}(u) - y\|_\Gamma^2$, with $\mathcal{G}$ a nonlinear mapping taking u to the observation space. We choose the observational noise covariance to be $\Gamma = \sigma^2 I$ with $\sigma = 10^{-2}$.

5.2.2 *Prior* We let A be the Stokes operator defined by writing (2.3) as $dv/dt + Av = 0, v(0) = u$ in the case $\gamma = 0$ and $\psi = 0$. Thus, A is ν times the negative Laplacian, restricted to a divergence free space; we also work on the space of functions whose spatial average is zero and then A is invertible. For the numerics that follow, we set $\nu = 0.05$. It is important to note that, in the periodic setting adopted here, A is diagonalized in the basis of divergence free Fourier series. Thus, fractional powers of A are easily calculated. The prior measure is then chosen as

$$\mu_0 = \mathcal{N}(0, \delta A^{-\alpha}), \tag{5.1}$$

in both the Eulerian and Lagrangian data scenarios. We require $\alpha > 1$ to ensure that the eigenvalues of the covariance are summable (a necessary and sufficient condition for draws from the prior, and hence the posterior, to be continuous functions, almost surely). In the numerics that follow, the parameters of the prior were chosen to be $\delta = 400$ and $\alpha = 2$.

5.2.3 *Numerical implementation* The figures that follow in this section are taken from what are termed *identical twin* experiments in the data assimilation community: the same approximation of the model described above to simulate the data is also used for evaluation of Φ in the statistical algorithm in the calculation of the likelihood of u_0 given the data, with the same assumed covariance structure of the observational noise as was used to simulate the data.

Since the domain is the two-dimensional torus, the evolution of the velocity field can be solved exactly for a truncated Fourier series, and in the numerics that follow we truncate this to 100 unknowns, as we have found the results to be robust to further refinement. In the case of the Lagrangian data, we integrate the trajectories (2.5) using an Euler scheme with time step $\Delta t = 0.01$. In each case we will give the values of N (number of spatial observations, or particles) and M (number of temporal observations) that were used. The observation stations (Eulerian data) or initial positions of the particles (Lagrangian data) are evenly spaced on a grid. The M observation times are evenly spaced, with the final observation time given by $T_M = 1$ for Lagrangian observations and $T_M = 0.1$ for Eulerian. The true initial condition u is chosen randomly from the prior distribution.

5.2.4 *Convergence from different initial states* We consider a posterior distribution found from data comprised of 900 Lagrangian tracers observed at 100 evenly spaced times on $[0, 1]$. The data volume is high and a form of posterior consistency is observed for low Fourier modes, meaning that the posterior is approximately a Dirac mass at the truth. Observations were made of each of these tracers up to a final

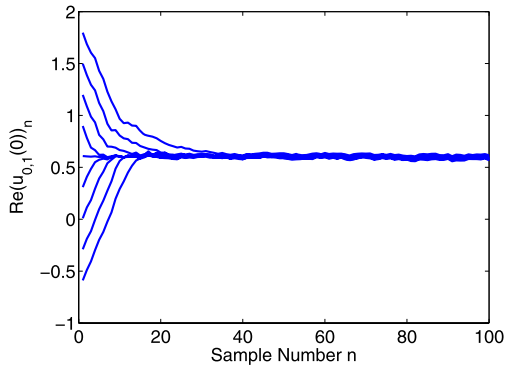


FIG. 3. *Convergence of value of one Fourier mode of the initial condition u_0 in the pCN Markov chains with different initial states, with Lagrangian data.*

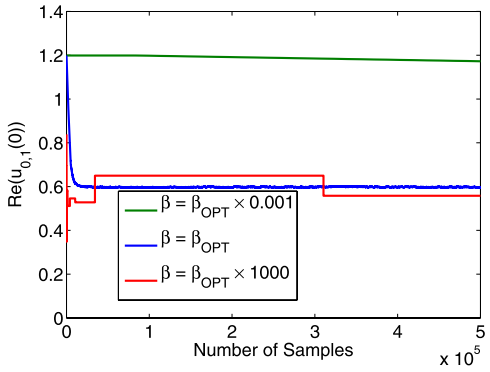


FIG. 4. *Convergence of one of the Fourier modes of the initial condition in the pCN Markov chains with different proposal variances, with Eulerian data.*

time $T = 1$. Figure 3 shows traces of the value of one particular Fourier mode[5] of the true initial conditions. Different starting values are used for pCN and all converge to the same distribution. The proposal variance β was chosen in order to give an average acceptance probability of approximately 25%.

5.2.5 *Convergence with different β* Here we study the effect of varying the proposal variance. Eulerian data is used with 900 observations in space and 100 observation times on $[0, 1]$. Figure 4 shows the different rates of convergence of the algorithm with different values of β, in the same Fourier mode coefficient as used in Figure 3. The value labelled β_{opt} here is chosen to give an acceptance rate of approximately 50%. This value of β is obtained by using an adaptive burn-in, in which the acceptance probability is estimated over short bursts and the step size β adapted accordingly. With β too small, the algorithm accepts proposed states often, but these changes in state are too small, so the algorithm does not explore the state space efficiently. In contrast, with β too big, larger jumps are proposed, but are often rejected since the proposal often has small probability density and so are often rejected. Figure 4 shows examples of both of these, as well as a more efficient choice β_{opt}.

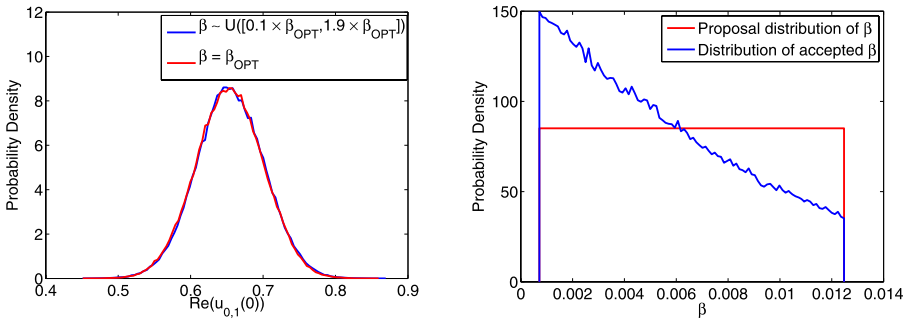
5.2.6 *Convergence with random β* Here we illustrate the possibility of using a random proposal variance β, as introduced in Section 4.5 [expressed in terms of δ and (4.7) rather than β and (4.8)]. Such methods have the potential advantage of including the possibility of large and small steps in the proposal. In this example we use Eulerian data once again, this time with only 9 observation stations, with only one observation time at $T = 0.1$. Two instances of the sampler were run with the same data, one with a static value of $\beta = \beta_{\text{opt}}$ and one with $\beta \sim U([0.1 \times \beta_{\text{opt}}, 1.9 \times \beta_{\text{opt}}])$. The marginal distributions for both Markov chains are shown in Figure 5(a), and are very close indeed, verifying that randomness in the proposal variance scale gives rise to (empirically) ergodic Markov chains. Figure 5(b) shows the distribution of the β for which the proposed state was accepted. As expected, the initial uniform distribution is skewed, as proposals with smaller jumps are more likely to be accepted.

The convergence of the method with these two choices for β were roughly comparable in this simple experiment. However, it is of course conceivable that when attempting to explore multimodal posterior distributions it may be advantageous to have a mix of both large proposal steps, which may allow large leaps between different areas of high probability density, and smaller proposal steps in order to explore a more localised region.

5.3 Subsurface Geophysics

The purpose of this section is twofold: we demonstrate another nontrivial application where function space sampling is potentially useful and we demonstrate the use of sieve priors in this context. Key to understanding what follows in this problem is to appreciate that, for the data volume we employ, the posterior distribution can be very diffuse and expensive to explore unless severe prior modelling is imposed, meaning that the prior is heavily weighted to solutions with only a small number of active Fourier modes, at low wave numbers. This is because the

[5]The real part of the coefficient of the Fourier mode with wave number 0 in the x-direction and wave number 1 in the y-direction.



(a) (b)

FIG. 5. *Eulerian data assimilation example.* (a) *Empirical marginal distributions estimated using the pCN with and without random β.* (b) *Plots of the proposal distribution for β and the distribution of values for which the pCN proposal was accepted.*

homogenizing property of the elliptic PDE means that a whole range of different length-scale solutions can explain the same data. To combat this, we choose very restrictive priors, either through the form of Gaussian covariance or through the sieve mechanism, which favour a small number of active Fourier modes.

5.3.1 *Target distributions* We consider equation (2.7) in the case $D = [0,1]^2$. Recall that the objective in this problem is to recover the permeability $\kappa = \exp(u)$. The sampling algorithms discussed here are applied to the log permeability u. The "true" permeability for which we test the algorithms is shown in Figure 6 and is given by

$$\kappa(x) = \exp(u_1(x)) = \tfrac{1}{10}. \tag{5.2}$$

The pressure measurement data is $y_j = p(x_j) + \sigma\eta_j$ with the η_j i.i.d. standard unit Gaussians, with the measurement location shown in Figure 7.

5.3.2 *Prior* The priors will either be Gaussian or a sieve prior based on a Gaussian. In both cases the Gaussian structure is defined via a Karhunen–Loéve expansion of the form

$$u(x) = \zeta_{0,0}\varphi^{(0,0)} + \sum_{(p,q)\in\mathbb{Z}^2/\{0,0\}} \frac{\zeta_{p,q}\varphi^{(p,q)}}{(p^2+q^2)^{\alpha}}, \tag{5.3}$$

where $\varphi^{(p,q)}$ are two-dimensional Fourier basis functions and the $\zeta_{p,q}$ are independent random variables with distribution $\zeta_{p,q} \sim \mathcal{N}(0,1)$ and $a \in \mathbb{R}$. To ensure that the eigenvalues of the prior covariance operator are summable (a necessary and sufficient condition for draws from it to be continuous functions, almost surely), we require that $\alpha > 1$. For target defined via κ we take $\alpha = 1.001$.

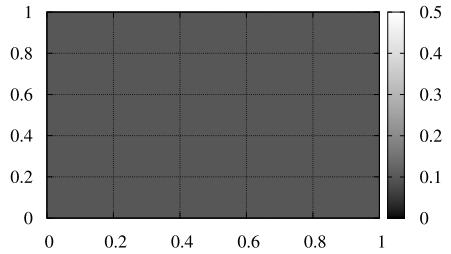


FIG. 6. *True permeability function used to create target distributions in subsurface geophysics application.*

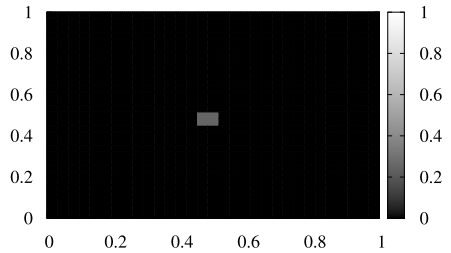


FIG. 7. *Measurement locations for subsurface experiment.*

For the Gaussian prior we employ MwG and pCN schemes, and we employ the pCN-based Gibbs sampler from Section 4.7 for the sieve prior; we refer to this latter algorithm as Sieve-pCN. As in Section 5.1, it is important that the reader appreciates that the comparison between MwG and pCN corresponds to sampling from the same posterior, since they use the same prior, but that all comparisons between Sieve-pCN and other methods also quantify the effect of prior modelling as well as algorithm.

5.3.3 *Numerical implementation* The forward model is evaluated by solving equation (2.7) on the two-dimensional domain $D = [0,1]^2$ using a finite difference method with mesh of size $J \times J$. This results in

a $J^2 \times J^2$ banded matrix with bandwidth J which may be solved, using a banded matrix solver, in $\mathcal{O}(J^4)$ floating point operations (see page 171 [25]). As drawing a sample is a $\mathcal{O}(J^4)$ operation, the grid sizes used within these experiments was kept deliberately low: for target defined via κ we take $J = 64$. This allowed a sample to be drawn in less than 100 ms and therefore 10^6 samples to be drawn in around a day. We used 1 measurement point, as shown in Figure 7.

5.3.4 *Numerical results* Since $\alpha = 1.001$, the eigenvalues of the prior covariance are only just summable, meaning that many Fourier modes will be active in the prior. Figure 8 shows trace plots obtained through application of the MwG and pCN methods to the Gaussian prior and a pCN-based Gibbs sampler for the sieve prior, denoted Sieve-pCN. The proposal variance for pCN and Sieve-pCN was selected to ensure an average acceptance of around 0.234. Four different seeds are used. It is clear from these plots that only the MCMC chain generated by the sieve prior/algorithm combination converges in the available computational time. The other algorithms fail to converge under these test conditions. This demonstrates the importance of prior modelling assumptions for these under-determined inverse problems with multiple solutions.

5.4 Image Registration

In this subsection we consider the image registration problem from Section 2.4. Our primary purpose is to illustrate the idea that, in the function space setting, it is possible to extend the prior modelling to include an unknown observational precision and to use conjugate Gamma priors for this parameter.

5.4.1 *Target distribution* We study the setup from Section 2.4, with data generated from a noisily observed truth $u = (p, \eta)$ which corresponds to a smooth closed curve. We make J noisy observations of the curve where, as will be seen below, we consider the cases $J = 10, 20, 50, 100, 200, 500$ and 1000. The noise used to generate the data is an uncorrelated mean zero Gaussian at each location with variance $\sigma^2_{\text{true}} = 0.01$. We will study the case where the noise variance σ^2 is itself considered unknown, introducing a prior on $\tau = \sigma^{-2}$. We then use MCMC to study the posterior distribution on (u, τ), and hence on (u, σ^2).

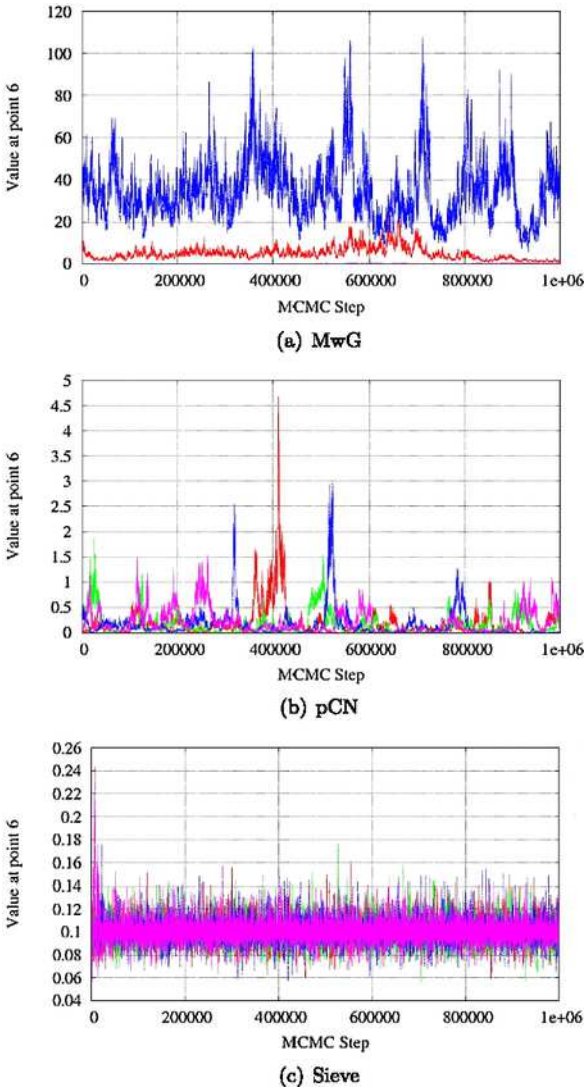


FIG. 8. *Trace plots for the subsurface geophysics application, using 1 measurement. The MwG, pCN and Sieve-pCN algorithms are compared. Different colours correspond to identical MCMC simulations with different random number generator seeds.*

5.4.2 *Prior* The priors on the initial momentum and reparameterisation are taken as

$$\begin{aligned} \mu_p(p) &= \mathcal{N}(0, \delta_1 \mathcal{H}^{-\alpha_1}), \\ \mu_\nu(\nu) &= \mathcal{N}(0, \delta_2 \mathcal{H}^{-\alpha_2}), \end{aligned} \tag{5.4}$$

where $\alpha_1 = 0.55$, $\alpha_2 = 1.55$, $\delta_1 = 30$ and $\delta_2 = 5 \cdot 10^{-2}$. Here $\mathcal{H} = (I - \Delta)$ denotes the Helmholtz operator in one dimension and, hence, the chosen values of α_i ensure that the eigenvalues of the prior covariance operators are summable. As a consequence, draws from the prior are continuous, almost surely. The prior for τ is defined as

$$\mu_\tau = \text{Gamma}(\alpha_\sigma, \beta_\sigma), \tag{5.5}$$

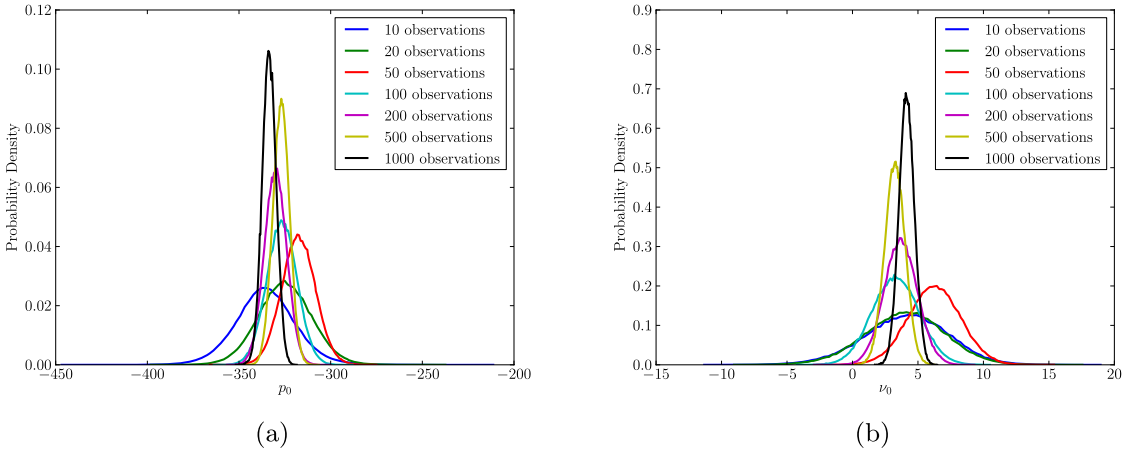


FIG. 9. *Convergence of the lowest wave number Fourier modes in* (a) *the initial momentum* P_0*, and* (b) *the reparameterisation function* ν*, as the number of observations is increased, using the pCN.*

noting that this leads to a conjugate posterior on this variable, since the observational noise is Gaussian. In the numerics that follow, we set $\alpha_\sigma = \beta_\sigma = 0.0001$.

5.4.3 *Numerical implementation* In each experiment the data is produced using the same template shape Γ_{db}, with parameterization given by

$$q_{\text{db}}(s) = (\cos(s) + \pi, \sin(s) + \pi), \qquad s \in [0, 2\pi). \tag{5.6}$$

In the following numerics, the observed shape is chosen by first sampling an instance of p and ν from their respective prior distributions and using the numerical approximation of the forward model to give us the parameterization of the target shape. The N observational points $\{s_i\}_{i=1}^N$ are then picked by evenly spacing them out over the interval $[0, 1)$, so that $s_i = (i-1)/N$.

5.4.4 *Finding observational noise hyperparameter* We implement an MCMC method to sample from the joint distribution of (u, τ), where (recall) $\tau = \sigma^{-2}$ is the inverse observational precision. When sampling u we employ the pCN method. In this context it is possible to either: (i) implement a Metropolis-within-Gibbs sampler, alternating between use of pCN to sample $u|\tau$ and using explicit sampling from the Gamma distribution for $\tau|u$; or (ii) marginalize out τ and sample directly from the marginal distribution for u, generating samples from τ separately; we adopt the second approach.

We show that, by taking data sets with an increasing number of observations N, the true values of the functions u and the precision parameter τ can both be recovered: a form of posterior consistency.

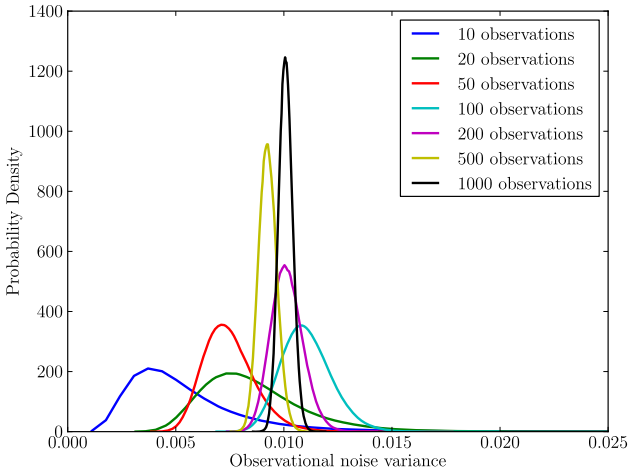


FIG. 10. *Convergence of the posterior distribution on the value of the noise variance* $\sigma^2 I$*, as the number of observations is increased, sampled using the pCN.*

This is demonstrated in Figure 9, for the posterior distribution on a low wave number Fourier coefficient in the expansion of the initial momentum p and the reparameterisation η. Figure 10 shows the posterior distribution on the value of the observational variance σ^2; recall that the true value is 0.01. The posterior distribution becomes increasingly peaked close to this value as N increases.

5.5 Conditioned Diffusions

Numerical experiments which employ function space samplers to study problems arising in conditioned diffusions have been published in a num-

ber of articles. The paper [4] introduced the idea of function space samplers in this context and demonstrated the advantage of the CNL method (4.12) over the standard Langevin algorithm for bridge diffusions; in the notation of that paper, the IA method with $\theta = \frac{1}{2}$ is our CNL method. Figures analogous to Figure 1(a) and (b) are shown. The article [19] demonstrates the effectiveness of the CNL method, for smoothing problems arising in signal processing, and figures analogous to Figure 1(a) and (b) are again shown. The paper [5] contains numerical experiments showing comparison of the function-space HMC method from Section 4.8 with the CNL variant of the MALA method from Section 4.3, for a bridge diffusion problem; the function-space HMC method is superior in that context, demonstrating the power of methods which break random-walk type behaviour of local proposals.

6. THEORETICAL ANALYSIS

The numerical experiments in this paper demonstrate that the function-space algorithms of Crank–Nicolson type behave well on a range of nontrivial examples. In this section we describe some theoretical analysis which adds weight to the choice of Crank–Nicolson discretizations which underlie these algorithms. We also show that the acceptance probability resulting from these proposals behaves as in finite dimensions: in particular, that it is continuous as the scale factor δ for the proposal variance tends to zero. And finally we summarize briefly the theory available in the literature which relates to the function-space viewpoint that we highlight in this paper. We assume throughout that Φ satisfies the following assumptions:

ASSUMPTIONS 6.1. The function $\Phi : X \to \mathbb{R}$ satisfies the following:

1. there exists $p > 0, K > 0$ such that, for all $u \in X$
$$0 \le \Phi(u; y) \le K(1 + \|u\|^p);$$
2. for every $r > 0$ there is $K(r) > 0$ such that, for all $u, v \in X$ with $\max\{\|u\|, \|v\|\} < r$,
$$|\Phi(u) - \Phi(v)| \le K(r)\|u - v\|.$$

These assumptions arise naturally in many Bayesian inverse problems where the data is finite dimensional [49]. Both the data assimilation inverse problems from Section 2.2 are shown to satisfy Assumptions 6.1, for appropriate choice of X in [11] (Navier–Stokes) and [49] (Stokes). The groundwater flow inverse problem from Section 2.3 is shown to satisfy these assumptions in [13], again for approximate choice of X. It is shown in [9] that Assumptions 6.1 are satisfied for the image registration problem of Section 2.4, again for appropriate choice of X. A wide range of conditioned diffusions satisfy Assumptions 6.1; see [20]. The density estimation problem from Section 2.1 satisfies the second item from Assumptions 6.1, but not the first.

6.1 Why the Crank–Nicolson Choice?

In order to explain this choice, we consider a one-parameter (θ) family of discretizations of equation (4.2), which reduces to the discretization (4.4) when $\theta = \frac{1}{2}$. This family is

$$v = u - \delta\mathcal{K}((1-\theta)\mathcal{L}u + \theta\mathcal{L}v) - \delta\gamma\mathcal{K}D\Phi(u) + \sqrt{2\delta\mathcal{K}}\xi_0, \tag{6.1}$$

where $\xi_0 \sim \mathcal{N}(0, I)$ is a white noise on X. Note that $w := \sqrt{\mathcal{C}}\xi_0$ has covariance operator $\mathcal{C}$ and is hence a draw from μ_0. Recall that if u is the current state of the Markov chain, then v is the proposal. For simplicity we consider only Crank–Nicolson proposals and not the MALA variants, so that $\gamma = 0$. However, the analysis generalises to the Langevin proposals in a straightforward fashion.

Rearranging (6.1), we see that the proposal v satisfies

$$v = (I - \delta\theta\mathcal{K}\mathcal{L})^{-1} \cdot ((I + \delta(1-\theta)\mathcal{K}\mathcal{L})u + \sqrt{2\delta\mathcal{K}}\xi_0). \tag{6.2}$$

If $\mathcal{K} = I$, then the operator applied to u is bounded on X for any $\theta \in (0, 1]$. If $\mathcal{K} = \mathcal{C}$, it is bounded for $\theta \in [0, 1]$. The white noise term is almost surely in X for $\mathcal{K} = I, \theta \in (0, 1]$ and $\mathcal{K} = \mathcal{C}, \theta \in [0, 1]$. The Crank–Nicolson proposal (4.5) is found by letting $\mathcal{K} = I$ and $\theta = \frac{1}{2}$. The preconditioned Crank–Nicolson proposal (4.7) is found by setting $\mathcal{K} = \mathcal{C}$ and $\theta = \frac{1}{2}$. The following theorem explains the choice $\theta = \frac{1}{2}$.

THEOREM 6.2. *Let $\mu_0(X) = 1$, let Φ satisfy Assumption 6.1(2) and assume that μ and μ_0 are equivalent as measures with the Radon–Nikodym derivative (1.1). Consider the proposal $v|u \sim q(u, \cdot)$ defined by (6.2) and the resulting measure $\eta(du, dv) = q(u, dv)\mu(du)$ on $X \times X$. For both $\mathcal{K} = I$ and $\mathcal{K} = \mathcal{C}$ the measure $\eta^\perp = q(v, du)\mu(dv)$ is equivalent to η if and only if $\theta = \frac{1}{2}$. Furthermore, if $\theta = \frac{1}{2}$, then*

$$\frac{d\eta^\perp}{d\eta}(u, v) = \exp(\Phi(u) - \Phi(v)).$$

By use of the analysis of Metropolis–Hastings methods on general state spaces in [51], this theorem

shows that the Crank–Nicolson proposal (6.2) leads to a well-defined MCMC algorithm in the function-space setting, if and only if $\theta = \frac{1}{2}$. Note, relatedly, that the choice $\theta = \frac{1}{2}$ has the desirable property that $u \sim \mathcal{N}(0, \mathcal{C})$ implies that $v \sim \mathcal{N}(0, \mathcal{C})$: thus, the prior measure is preserved under the proposal. This mimics the behaviour of the SDE (4.2) for which the prior is an invariant measure. We have thus justified the proposals (4.5) and (4.7) on function space. To complete our analysis, it remains to rule out the standard random walk proposal (4.3).

THEOREM 6.3. *Consider the proposal $v|u \sim q(u,\cdot)$ defined by (4.3) and the resulting measure $\eta(du,dv) = q(u,dv)\mu(du)$ on $X \times X$. For both $\mathcal{K} = I$ and $\mathcal{K} = \mathcal{C}$ the measure $\eta^{\perp} = q(v,du)\mu(dv)$ is not absolutely continuous with respect to η. Thus, the MCMC method is not defined on function space.*

6.2 The Acceptance Probability

We now study the properties of the two Crank–Nicolson methods with proposals (4.5) and (4.7) in the limit $\delta \to 0$, showing that finite-dimensional intuition carries over to this function space setting. We define

$$R(u,v) = \Phi(u) - \Phi(v)$$

and note from (4.11) that, for both of the Crank–Nicolson proposals,

$$a(u,v) = \min\{1, \exp(R(u,v))\}.$$

THEOREM 6.4. *Let μ_0 be a Gaussian measure on a Hilbert space $(X, \|\cdot\|)$ with $\mu_0(X) = 1$ and let μ be an equivalent measure on X given by the Radon–Nikodym derivative (1.1), satisfying Assumptions 6.1(1) and 6.1(2). Then both the pCN and CN algorithms with fixed δ are defined on X and, furthermore, the acceptance probability satisfies*

$$\lim_{\delta \to 0} \mathbb{E}^{\eta} a(u,v) = 1.$$

6.3 Scaling Limits and Spectral Gaps

There are two basic theories which have been developed to explain the advantage of using the algorithms introduced here which are based on the function-space viewpoints. The first is to prove scaling limits of the algorithms, and the second is to establish spectral gaps. The use of scaling limits was pioneered for local-proposal Metropolis algorithms in the papers [40–42], and recently extended to the hybrid Monte Carlo method [6]. All of this work concerned i.i.d. target distributions, but recently it has been shown that the basic conclusions of the theory, relating to optimal scaling of proposal variance with dimension, and optimal acceptance probability, can be extended to the target measures of the form (1.1) which are central to this paper; see [29, 36]. These results show that the standard MCMC method must be scaled with proposal variance (or time-step in the case of HMC) which is inversely proportional to a power of d_u, the discretization dimension, and that the number of steps required grows under mesh refinement. The papers [5, 37] demonstrate that judicious modifications of these standard algorithms, as described in this paper, lead to scaling limits *without* the need for scalings of proposal variance or time-step which depend on dimension. These results indicate that the number of steps required is stable under mesh refinement, for these new methods, as demonstrated numerically in this paper. The second approach, namely, the use of spectral gaps, offers the opportunity to further substantiate these ideas: in [21] it is shown that the pCN method has a dimension independent spectral gap, whilst a standard random walk which closely resembles it has spectral gap which shrinks with dimension. This method of analysis, via spectral gaps, will be useful for the analysis of many other MCMC algorithms arising in high dimensions.

7. CONCLUSIONS

We have demonstrated the following points:

- A wide range of applications lead naturally to problems defined via density with respect to a Gaussian random field reference measure, or variants on this structure.
- Designing MCMC methods on function space, and then discretizing the nonparametric problem, produces better insight into algorithm design than discretizing the nonparametric problem and then applying standard MCMC methods.
- The transferable idea underlying all the methods is that, in the purely Gaussian case when only the reference measure is sampled, the resulting MCMC method should accept with probability one; such methods may be identified by time-discretization of certain stochastic dynamical systems which preserve the Gaussian reference measure.
- Using this methodology, we have highlighted new random walk, Langevin and Hybrid Monte Carlo Metropolis-type methods, appropriate for problems where the posterior distribution has density with respect to a Gaussian prior, all of which can

be implemented by means of small modifications of existing codes.
- We have applied these MCMC methods to a range of problems, demonstrating their efficacy in comparison with standard methods, and shown their flexibility with respect to the incorporation of standard ideas from MCMC technology such as Gibbs sampling and estimation of noise precision through conjugate Gamma priors.
- We have pointed to the emerging body of theoretical literature which substantiates the desirable properties of the algorithms we have highlighted here.

The ubiquity of Gaussian priors means that the technology that is described in this article is of immediate applicability to a wide range of applications. The generality of the philosophy that underlies our approach also suggests the possibility of numerous further developments. In particular, many existing algorithms can be modified to the function space setting that is shown to be so desirable here, when Gaussian priors underlie the desired target; and many similar ideas can be expected to emerge for the study of problems with non-Gaussian priors, such as arise in wavelet based nonparametric estimation.

ACKNOWLEDGEMENTS

S. L. Cotter is supported by EPSRC, ERC (FP7/2007-2013 and Grant 239870) and St. Cross College. G. O. Roberts is supported by EPSRC (especially the CRiSM grant). A. M. Stuart is grateful to EPSRC, ERC and ONR for the financial support of research which underpins this article. D. White is supported by ERC.

REFERENCES

[1] Adams, R. P., Murray, I. and Mackay, D. J. C. (2009). The Gaussian process density sampler. In *Advances in Neural Information Processing Systems 21*.

[2] Adler, R. J. (2010). *The Geometry of Random Fields*. SIAM, Philadelphia, PA.

[3] Bennett, A. F. (2002). *Inverse Modeling of the Ocean and Atmosphere*. Cambridge Univ. Press, Cambridge. MR1920432

[4] Beskos, A., Roberts, G., Stuart, A. and Voss, J. (2008). MCMC methods for diffusion bridges. *Stoch. Dyn.* **8** 319–350. MR2444507

[5] Beskos, A., Pinski, F. J., Sanz-Serna, J. M. and Stuart, A. M. (2011). Hybrid Monte Carlo on Hilbert spaces. *Stochastic Process. Appl.* **121** 2201–2230. MR2822774

[6] Beskos, A., Pillai, N. S., Roberts, G. O., Sanz-Serna, J. M. and Stuart, A. M. (2013). Optimal tuning of hybrid Monte Carlo. *Bernoulli*. To appear. Available at http://arxiv.org/abs/1001.4460.

[7] Cotter, C. J. (2008). The variational particle-mesh method for matching curves. *J. Phys. A* **41** 344003, 18. MR2456340

[8] Cotter, S. L. (2010). Applications of MCMC methods on function spaces. Ph.D. thesis, Univ. Warwick.

[9] Cotter, C. J., Cotter, S. L. and Vialard, F. X. (2013). Bayesian data assimilation in shape registration. *Inverse Problems* **29** 045011.

[10] Cotter, S. L., Dashti, M. and Stuart, A. M. (2012). Variational data assimilation using targetted random walks. *Internat. J. Numer. Methods Fluids* **68** 403–421. MR2880204

[11] Cotter, S. L., Dashti, M., Robinson, J. C. and Stuart, A. M. (2009). Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse Problems* **25** 115008, 43. MR2558668

[12] Da Prato, G. and Zabczyk, J. (1992). *Stochastic Equations in Infinite Dimensions. Encyclopedia of Mathematics and Its Applications* **44**. Cambridge Univ. Press, Cambridge. MR1207136

[13] Dashti, M., Harris, S. and Stuart, A. (2012). Besov priors for Bayesian inverse problems. *Inverse Probl. Imaging* **6** 183–200. MR2942737

[14] Diaconis, P. (1988). Bayesian numerical analysis. In *Statistical Decision Theory and Related Topics, IV, Vol. 1 (West Lafayette, Ind., 1986)* 163–175. Springer, New York. MR0927099

[15] Duane, S., Kennedy, A. D., Pendleton, B. and Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222.

[16] Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 123–214. MR2814492

[17] Glaunes, J., Trouvé, A. and Younes, L. (2004). Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on 2* 712–718. IEEE.

[18] Hairer, M., Stuart, A. M. and Voss, J. (2007). Analysis of SPDEs arising in path sampling. II. The nonlinear case. *Ann. Appl. Probab.* **17** 1657–1706. MR2358638

[19] Hairer, M., Stuart, A. and Voss, J. (2009). Sampling conditioned diffusions. In *Trends in Stochastic Analysis. London Mathematical Society Lecture Note Series* **353** 159–185. Cambridge Univ. Press, Cambridge. MR2562154

[20] Hairer, M., Stuart, A. and Voss, J. (2011). Signal processing problems on function space: Bayesian formulation, stochastic PDEs and effective MCMC methods. In *The Oxford Handbook of Nonlinear Filtering* (D. Crisan and B. Rozovsky, eds.) 833–873. Oxford Univ. Press, Oxford. MR2884617

[21] Hairer, M., Stuart, A. M. and Vollmer, S. (2013). Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. Available at http://arxiv.org/abs/1112.1392.

[22] Hairer, M., Stuart, A. M., Voss, J. and Wiberg, P. (2005). Analysis of SPDEs arising in path sampling. I. The Gaussian case. *Commun. Math. Sci.* **3** 587–603. MR2188686

[23] Hills, S. E. and Smith, A. F. M. (1992). Parameterization issues in Bayesian inference. In *Bayesian Statistics, 4 (PeñíScola, 1991)* 227–246. Oxford Univ. Press, New York. MR1380279

[24] Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G., eds. (2010). *Bayesian Nonparametrics. Cambridge Series in Statistical and Probabilistic Mathematics* **28**. Cambridge Univ. Press, Cambridge. MR2722987

[25] Iserles, A. (2004). *A First Course in the Numerical Analysis of Differential Equations*. Cambridge Univ. Press, Cambridge.

[26] Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge Univ. Press, Cambridge.

[27] Lemm, J. C. (2003). *Bayesian Field Theory*. Johns Hopkins Univ. Press, Baltimore, MD. MR1987925

[28] Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR1842342

[29] Mattingly, J. C., Pillai, N. S. and Stuart, A. M. (2012). Diffusion limits of the random walk Metropolis algorithm in high dimensions. *Ann. Appl. Probab.* **22** 881–930. MR2977981

[30] McLaughlin, D. and Townley, L. R. (1996). A reassessment of the groundwater inverse problem. *Water Res. Res.* **32** 1131–1161.

[31] Miller, M. T. and Younes, L. (2001). Group actions, homeomorphisms, and matching: A general framework. *Int. J. Comput. Vis.* **41** 61–84.

[32] Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer, New York.

[33] Neal, R. M. (1998). Regression and classification using Gaussian process priors. Available at http://www.cs.toronto.edu/~radford/valencia.abstract.html.

[34] Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* 113–162. CRC Press, Boca Raton, FL. MR2858447

[35] O'Hagan, A., Kennedy, M. C. and Oakley, J. E. (1999). Uncertainty analysis and other inference tools for complex computer codes. In *Bayesian Statistics, 6 (Alcoceber, 1998)* 503–524. Oxford Univ. Press, New York. MR1724872

[36] Pillai, N. S., Stuart, A. M. and Thiéry, A. H. (2012). Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *Ann. Appl. Probab.* **22** 2320–2356. MR3024970

[37] Pillai, N. S., Stuart, A. M. and Thiery, A. H. (2012). On the random walk Metropolis algorithm for Gaussian random field priors and gradient flow. Available at http://arxiv.org/abs/1108.1494.

[38] Richtmyer, D. and Morton, K. W. (1967). *Difference Methods for Initial Value Problems*. Wiley, New York.

[39] Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York. MR1707311

[40] Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7** 110–120. MR1428751

[41] Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 255–268. MR1625691

[42] Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statist. Sci.* **16** 351–367. MR1888450

[43] Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. *Biometrika* **88** 603–621. MR1859397

[44] Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2** 341–363. MR1440273

[45] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications. Monographs on Statistics and Applied Probability* **104**. Chapman & Hall/CRC, Boca Raton, FL. MR2130347

[46] Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **55** 3–23. MR1210421

[47] Sokal, A. D. (1989). Monte Carlo methods in statistical mechanics: Foundations and new algorithms, Univ. Lausanne, Bâtiment des sciences de physique Troisième Cycle de la physique en Suisse romande.

[48] Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York. MR1697409

[49] Stuart, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numer.* **19** 451–559. MR2652785

[50] Stuart, A. M., Voss, J. and Wiberg, P. (2004). Fast communication conditional path sampling of SDEs and the Langevin MCMC method. *Commun. Math. Sci.* **2** 685–697. MR2119934

[51] Tierney, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *Ann. Appl. Probab.* **8** 1–9. MR1620401

[52] Vaillant, M. and Glaunes, J. (2005). Surface matching via currents. In *Information Processing in Medical Imaging* 381–392. Springer, Berlin.

[53] van der Meulen, F., Schauer, M. and van Zanten, H. (2013). Reversible jump MCMC for nonparametric drift estimation for diffusion processes. *Comput. Statist. Data Anal.* To appear.

[54] Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28** 532–552. MR1790008