MCMC Methods for Sampling Function Space

Alexandros Beskos and Andrew Stuart*

Abstract. Applied mathematics is concerned with developing models with predictive capability, and with probing those models to obtain qualitative and quantitative insight into the phenomena being modelled. Statistics is data-driven and is aimed at the development of methodologies to optimize the information derived from data. The increasing complexity of phenomena that scientists and engineers wish to model, together with our increased ability to gather, store and interrogate data, mean that the subjects of applied mathematics and statistics are increasingly required to work in conjunction in order to significantly progress understanding.

This article is concerned with a research program at the interface between these two disciplines, aimed at problems in differential equations where profusion of data and the sophisticated model combine to produce the mathematical problem of obtaining information from a probability measure on function space. In this context there is an array of problems with a common mathematical structure, namely that the probability measure in question is a change of measure from a Gaussian. We illustrate the wide-ranging applicability of this structure. For problems whose solution is determined by a probability measure on function space, information about the solution can be obtained by sampling from this probability measure. One way to do this is through the use of Markov chain Monte-Carlo (MCMC) methods. We show how the common mathematical structure of the aforementioned problems can be exploited in the design of effective MCMC methods.

Mathematics Subject Classification (2000). Primary 35R30; Secondary 65C40.

Keywords. Bayes's formula, inverse problem, change of measure from Gaussian, Langevin SPDEs, MCMC.

1. Introduction

The Bayesian approach to inverse problems is natural in many situations where data and model must be integrated with one another to provide maximal information about the system. When the object of interest is a function then the posterior measure from Bayes's formula is a measure on a function space. In this article we introduce a range of applied problems where this viewpoint is natural, and which all possess a common mathematical framework: the posterior measure on function space, π, has density with respect to a Gaussian reference measure;

*Zeeman Building, University of Warwick, Coventry CV4 7AL, UK

see Section 2. In Section 3 we describe a general approach for writing down π-invariant stochastic partial differential equations (SPDEs). It is important to be able to sample the posterior measure to get information about it. This is the topic of Section 4 where we introduce Markov chain Monte-Carlo (MCMC) methods and describe the Metropolis-Hastings variant. Section 5 contains statements of theoretical results concerning the complexity of these MCMC methods, when applied to (finite-dimensional approximations of) the target measures of interest in this article; proofs are contained in the Appendix. Section 6 contains a summary and directions for further research.

2. Measures on Function Space

In this section we give several illustrations of problems whose solution requires sampling of a measure on function space. For simplicity we confine our analysis to the situation where the functions are in a Hilbert space $\mathcal{H}$. In all cases we will see that the target measure π has Radon-Nikodym derivative with respect to a reference Gaussian measure π_0, so that we can write

$$\frac{d\pi}{d\pi_0}(x) \propto \exp\Big(-\Phi(x)\Big). \tag{1}$$

For future reference we will assume that π_0 has *mean* m and *covariance operator* $\mathcal{C}$. Adopting standard notation we will write $\pi_0 \sim \mathcal{N}(m, \mathcal{C})$. For expression (1) to make sense we require that the potential $\Phi : \mathcal{H} \mapsto \mathbb{R}$ is defined π_0-almost surely. Informally[1], it is instructive to write the density for the Gaussian measure as

$$\pi_0(x) \propto \exp\Big(-\frac{1}{2}\langle x-m, \mathcal{C}^{-1}(x-m)\rangle\Big). \tag{2}$$

The inverse of $-\mathcal{C}$ is known as the *precision operator* and will be denoted by $\mathcal{L}$. Using this notation and combining (1) and (2) we get the following informal expression for the density $\pi(x)$:

$$\pi(x) \propto \exp\Big(-\Phi(x) + \frac{1}{2}\langle x-m, \mathcal{L}(x-m)\rangle\Big). \tag{3}$$

In many of our applications $\mathcal{L}$ will be a differential operator. Note that the density (3) is maximized at solutions of the equation

$$\mathcal{L}(x-m) - D\Phi(x) = 0.$$

This is a first hint at the difficulties inherent in sampling measures on function space: even locating places of high probability involves the solution of differential equations. Sampling the entire measure will typically be even more difficult.

[1] In finite dimensions formula (2) gives the density of a Gaussian measure $\mathcal{N}(m, \mathcal{C})$ with respect to Lebesgue measure. On a general Hilbert space there is no analogue of Lebesgue measure, so the formula should be viewed simply as a useful heuristic, which is helpful for understanding the ideas in this article. For economy of notation we use the symbol π for both a measure and its density.

2.1. Molecular Dynamics. In the mathematical description of molecules a commonly used model is that of *Brownian dynamics* in which the atomic positions x undergo thermally activated motion in a potential V:

$$\frac{dx}{dt} = -\nabla V(x) + \sqrt{\frac{2}{\beta}}\frac{dB}{dt}. \tag{4}$$

Here x denotes the vector of atomic positions in $\mathbb{R}^{Nd}$ where N is the number of atoms and d the spatial dimension. The process B is a standard Brownian motion in $\mathbb{R}^{Nd}$ and β the inverse temperature. When the temperature is small ($\beta \gg 1$) the solution of this stochastic differential equation (SDE) spends most of its time near the minima of the potential V. Transitions between different minima are then rare events. Simply solving the SDE starting from one of the minima will be a computationally infeasible way of generating sample paths which jump between minima since the time to make a transition is exponentially small in β [12]. Instead we may condition on this rare event occurring. Let $x^{\pm}$ denote two minima of the potential and consider the boundary conditions

$$x(0) = x^- \quad \text{and} \quad x(T) = x^+. \tag{5}$$

If we now view the Brownian motion as a control, we see that it may be chosen to drive the solution of (4) from one minimum to the other. Since paths of Brownian motion carry a probability measure, which induces a measure on paths x, we have a mechanism to construct a probability measure on a function space of paths which respect (5). We now make these ideas more precise.

The probability measure π governing the stochastic boundary value problem (4), (5) has density with respect to the Brownian bridge (Gaussian) measure π_0 arising in the case $V \equiv 0$. Girsanov's theorem, together with Itô's formula [17, 25], gives that

$$\frac{d\pi}{d\pi_0}(x) \propto \exp\left(-\frac{\beta}{2}\int_0^T G(x;\beta)dt\right)$$

where

$$G(x;\beta) = \frac{1}{2}|\nabla V(x)|^2 - \frac{1}{\beta}\Delta V(x).$$

We have thus established a particular instance of (1). It is useful conceptually to write the Brownian bridge probability density function with respect to an infinite dimensional Lebesgue measure as is frequently done in the physics literature [7]; the desired expression, which may be found by discretization and passage to the limit (see [31] for example) is

$$\exp\left(-\frac{\beta}{4}\int_0^T \left|\frac{dx}{dt}\right|^2 dt\right),$$

together with boundary conditions enforcing (5). The rigorous interpretation of this expression for π_0 is that in this case the precision operator is $\mathcal{L} = \frac{\beta}{2}\frac{d^2}{dt^2}$

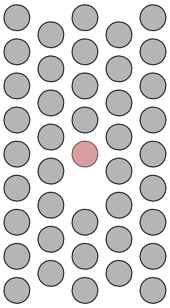

Figure 1. Crystal lattice with vacancy. We condition on the red atom moving into the vacancy.

equipped with homogeneous Dirichlet boundary conditions on $t \in [0,T]$, and the mean is the function

$$m(t) = \frac{t}{T}\,x^{+} + \frac{T-t}{T}\,x^{-}.$$

Thus, we may think of the probability density for π as being proportional to

$$\exp\left(-\frac{\beta}{4}\int_0^T \left|\frac{dx}{dt}\right|^2 dt - \frac{\beta}{2}\int_0^T G(x;\beta)dt\right)$$

with the boundary conditions (5) enforced. This is an explicit example of the general structure (3).

A typical application from molecular dynamics is illustrated in Figure 2.1. The figure shows a crystal lattice of atoms in two dimensions, with an atom removed from one site. The potential is a sum of pairwise potentials between atoms which has an r^{-12} repulsive singularity, r being the distance between a pair of atoms. The lattice should be viewed as spatially extended to the whole of $\mathbb{Z}^2$ by periodicity. Removal of an atom creates a vacancy which, under thermal activation as in (4), will diffuse around the lattice: the vacancy will move to a different lattice site whenever one of the neighboring atoms moves into the current vacancy position. This motion of the vacancy is a rare event; we can now condition our model on this event occurring. The solution of such rare event problems arising in chemistry and physics is an active area of research. See [4] for an overview of the subject and [11] for an approach which is useful in the zero temperature limit or close to it.

In summary, we have defined a probability measure for $x = x(t)$ in the Hilbert space $\mathcal{H} = L^2([0,T],\mathbb{R}^{Nd})$ which we term the *diffusion bridge* measure. This measure describes the distribution of sample paths of the SDE (4) conditioned to link two points in phase space $\mathbb{R}^{Nd}$ within a specified time period, as in (5). Solving problems of this form has wide application, not only in chemistry and physics, but also in areas such as econometrics where it is frequently of interest to augment discrete time data driven by an SDE [5, 6].

2.2. Signal Processing. It is often of interest to identify an underlying *signal* $\{x(t)\}_{0\leq t\leq T}$, given some *observation* $\{y(t)\}_{0\leq t\leq T}$. In the context of SDEs this can be formulated via a pair of coupled equations:

$$\frac{dx}{dt} = f(x) + \frac{dB_1}{dt}, \quad X(0) \sim \zeta \tag{6}$$

$$\frac{dy}{dt} = g(x,y) + \sigma\frac{dB_2}{dt}, \quad Y(0) = 0. \tag{7}$$

The *filtering problem* [24] is to find, for each $t \in [0,T]$, the probability distribution of $x(t) \in \mathbb{R}^m$ given y only at times up to t: $\{y(s)\}_{0\leq s\leq t}$. In contrast, the *smoothing problem* is to find the distribution of $x(t)$ given all observations $\{y(s)\}_{0\leq s\leq T}$; the smoothing problem can be viewed as finding the probability measure on the entire path $\{x(s)\}_{0\leq s\leq T}$, conditioned on $\{y(s)\}_{0\leq s\leq T}$. The filtering and smoothing distributions on $x(T)$ are the same but differ on $x(t)$ for any $t \in (0,T)$.

The smoothing problem can be formulated as determining a probability measure on $L^2([0,T],\mathbb{R}^m)$ of the form

$$\frac{d\pi}{d\pi_0}(x) \propto \exp\left(-\int_0^T G(x;y)dt\right)$$

where the observation y appears as fixed data in the probability measure for x. Here π_0 is again a Gaussian measure, known as the Kalman-Bucy smoother, derived from the original problem in the case where f and g are set to zero and ζ is Gaussian. The inverse of the covariance operator is again a second order differential operator, as for the bridge diffusion in the previous example; details may be found in [15, 17]. Once again we have established a particular instance of the general framework (1). Figure 2.2 illustrates the set-up.

2.3. Lagrangian Data Assimilation. Understanding oceans is fundamental in the atmospheric and environmental sciences, and for both commercial and military purposes. One way of probing the oceans is by placing “floats” (at a specified depth) or “drifters” (on the surface) in the ocean and allowing them to act as Lagrangian tracers in the flow. These tracers broadcast GPS data concerning their positions which can be used to make inference about the oceans themselves. The natural mathematical formulation is that of an inverse problem. We derive such a formulation, providing at the same time a straightforward illustration of the Bayesian approach to inverse problems. In so doing we show that Lagrangian data assimilation is yet another example of a problem which inherits the structure (1).

As a concrete model of this situation we consider the incompressible forced Navier-Stokes equations written in the form:

$$\frac{\partial v}{\partial t} + v\cdot\nabla v = \nu\Delta v - \nabla p + f, \quad (x,t) \in \Omega\times[0,\infty),$$
$$\nabla\cdot v = 0, \quad (x,t)\in\Omega\times[0,\infty).$$

Here Ω is the unit square and ν the viscosity. Also, we impose periodic boundary conditions on the velocity field v and the pressure p. We assume that f has zero

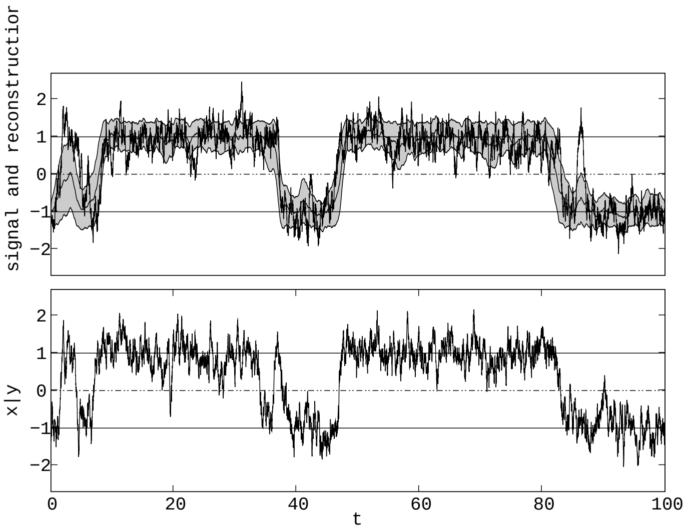

Figure 2. The upper panel shows the original signal (a sample path from (6)) together with the mean and standard deviation of the posterior measure on x given y (shaded band). The lower panel shows a single draw from the posterior measure on x given y. (The posterior measure is sampled using the SPDE defined later in Section 3.2.)

average over Ω; note that this implies the same for $v(x,t)$, provided that we require that the initial velocity field $u(x) = v(x,0)$ has zero average.

Our objective is to find the initial velocity field $u(x) \in \mathcal{H}$ where $\mathcal{H}$ is here the Hilbert space found as the closure in $L^2(\mathbb{T}^2, \mathbb{R}^2)$ of the space of periodic divergence-free, smooth functions on $\mathbb{T}^2$, with zero average. We assume that we are given noisy observations of Lagrangian tracers with position z solving

$$\frac{dz}{dt} = v(z,t).$$

The issue of minimal regularity assumptions on u and f so that Lagrangian tracers are well defined is discussed in [9]. For simplicity assume that we observe a single tracer z at a set of times $\{t_k\}_{k=1}^K$:

$$y_k = z(t_k) + \xi_k, \quad k = 1, \dots, K,$$

where the ξ_k's are zero mean Gaussian random variables. Concatenating data we may write

$$y = \tilde{z} + \xi$$

where $y = (y_1, \dots, y_K)$, $\tilde{z} = (z(t_1), \dots, z(t_K))$ and $\xi \sim \mathcal{N}(0, \Sigma)$ for some covariance matrix Σ. Figure 2.3 illustrates the set-up, showing a snap-shot of the flow field streamlines for $v(x,t)$ and the tracer particles $z(t)$ for some fixed time instance t.

We now construct the probability measure of interest, namely the probability of u given y. The first step is to assign a *prior* measure on u. We choose this to

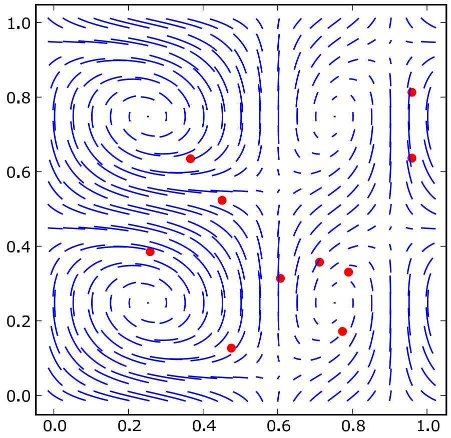

Figure 3. An example configuration of the velocity field at a given time instance. The small circles correspond to a number of Langangian tracers.

be the Gaussian measure with mean zero and precision operator which is minus the square of the Stokes operator $\mathcal{A}$ on $\mathcal{H}$ [29]. We now condition this prior on the observations, to find the *posterior* measure on u. We observe that $\tilde{z}$ is a (complicated) function $\mathcal{G}$ of u, the initial condition, so we may write

$$y = \mathcal{G}(u) + \xi.$$

Thus the probability of y given u is

$$\mathbb{P}\left(y \mid u\right) \propto \exp\Big(-\frac{1}{2}|y - \mathcal{G}(u)|_{\Sigma}^{2}\Big)$$

where $|\cdot|_{\Sigma}^{2} = |\Sigma^{-\frac{1}{2}} \cdot|^{2}$ and $|\cdot|$ is the standard finite-dimensional Euclidean norm. By Bayes's rule we deduce that

$$\frac{d\pi}{d\pi_0}(u) \propto \exp\Big(-\frac{1}{2}|y - \mathcal{G}(u)|_{\Sigma}^{2}\Big)$$

where π_0 is the prior Gaussian measure. We have now determined another example of the probability density structure (1).

Informally we may write

$$\pi(u) \propto \exp\Big(-\frac{1}{2}|y - \mathcal{G}(u)|_{\Sigma}^{2} - \frac{1}{2}\|\mathcal{A}u\|_{\mathcal{H}}^{2}\Big),$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm induced by the inner-product on $\mathcal{H}$. This expression provides another example of the general structure (3). The model is a very simple one, but more realistic models, in complex geometries and for coupled evolution of velocity, temperature and other fields (with multiple also observations) have a similar mathematical structure.

2.4. Geophysics. An important problem in subsurface geophysical applications, of interest to both petroleum engineers and hydrologists, is the determination of subsurface properties, in particular the permeability field (also known as the hydraulic conductivity). Making direct subsurface measurements is hard, so the primary observation is via indirect measurements of flow and transport through the medium. The following model for this set-up is taken from [10, 22]. The forward problem contains two unknown scalar fields: the water saturation S (volume fraction of water in an oil-water mixture) and pressure p. We study the problem in a bounded open set $\Omega \subset \mathbb{R}^d$ (typically $d = 2$ or 3). By means of Darcy's law we define the velocity field $v = -\lambda(S)K\nabla p$, where K is a permeability tensor field and the scalar $\lambda(S)$ determines the effect of saturation on permeability. In terms of the velocity field v, mass conservation and scalar advection respectively give the equations

$$\begin{aligned} -\nabla \cdot v = h, &\quad (x,t) \in \Omega \times [0,\infty), \\ \frac{\partial S}{\partial t} + v \cdot \nabla f(S) = 0, &\quad (x,t) \in \Omega \times [0,\infty), \end{aligned}$$

Here h is a source term and f the flux function. Boundary conditions are given for the pressure, or its gradient in the normal direction, on $\partial\Omega$. One way to understand the equations is as follows: Darcy's Law determines p, given S; the mass conservation equation is then a non-local hyperbolic conservation law for S and boundary conditions are specified on the inflow boundary $\partial\Omega^{\mathrm{in}} \subset \partial\Omega$. We set $\partial\Omega^{\mathrm{out}} = \partial\Omega \backslash \partial\Omega^{\mathrm{in}}$. The initial condition for the saturation is $S = 0$ and the boundary conditions on the inflow boundary are $S = 1$. In physical terms, the subsurface rock is assumed to be saturated entirely with oil at time $t = 0$, and water is then pumped in at the boundaries.

For simplicity we assume that the tensor K has the simple form $K = kI$, were k is the scalar *permeability field*, The inverse problem is to find the permeability field k from noisy measurements of what is known as the *fractional flow* or *oil cut* $F(t)$, a measurement which quantifies the fraction of oil produced at the outflow boundary $\partial\Omega^{\mathrm{out}}$ as water is pumped in through $\partial\Omega^{\mathrm{in}}$. Specifically

$$F(t) = 1 - \frac{\int_{\partial\Omega^{\mathrm{out}}} f(S) v_n dl}{\int_{\partial\Omega^{\mathrm{out}}} v_n dl}$$

where v_n is the component of v normal to the boundary and dl denotes integration along the boundary. Assume that we make measurements of F at times $\{t_k\}_{k=1}^K$ subject to Gaussian noise. So, the data are as follows

$$y_k = F(t_k) + \xi_k, \quad k = 1, \ldots, K,$$

where the ξ_k's are zero mean Gaussian random variables. Concatenating data we may write

$$y = \tilde{F} + \xi$$

where $y = (y_1, \ldots, y_K)$, $\tilde{F} = (F(t_1), \ldots, F(t_K))$ and $\xi \sim \mathcal{N}(0, \Sigma)$ for some covariance matrix Σ encapsulating measurement errors.

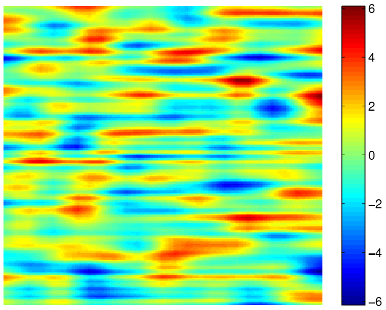

Figure 4. A realization from the prior distribution on the log-permeability field.

It is physically and mathematically important that k be positive in order to ensure that the elliptic equation for the pressure is well-posed. We thus write $k = \exp(u)$ and consider the problem of determining u. We observe that $\tilde{F}$ is a (complicated) function of u and so we may write

$$y = \mathcal{G}(u) + \xi$$

as in the previous data assimilation application. The prior is a zero mean Gaussian measure on u, usually specified through a covariance function $c(x, y)$ concerning which there is direct experimental information. The covariance operator $\mathcal{C}$ is defined by

$$\left(\mathcal{C}u\right)(x) = \int_{\Omega} c(x, y)u(y)dy.$$

Applying a zero mean Gaussian prior on u with this covariance operator gives rise to what is termed a *log-normal* permeability. We then have

$$\frac{d\pi}{d\pi_0}(u) \propto \exp\Big(-\frac{1}{2}|y - \mathcal{G}(u)|_{\Sigma}^{2}\Big)$$

where π_0 is the prior Gaussian measure $\mathcal{N}(0, \mathcal{C})$. This provides another explicit example of the structure (1). A typical sample from the prior distribution on a permeability field is shown in Figure 2.4.

For this problem the precision operator $\mathcal{L}$ is not necessarily a differential operator; in fact, it is typically a non-local operator. Informally we may write the desired probability via a density of the form

$$\pi(u) \propto \exp\Big(-\frac{1}{2}|y - \mathcal{G}(u)|_{\Sigma}^{2} - \frac{1}{2}\|(-\mathcal{L})^{\frac{1}{2}}u\|_{L^2(\Omega)}\Big)$$

providing another explicit example of the structure (3). We have only described a simplistic model. However, more involved models, in complex geometries and for the coupled evolution of multiple phases of oil, water and gas, and with multiple injection and production sites (for generation and measurement of data), and determination of a tensor permeability, share a similar mathematical structure.

2.5. Mathematical Structure of the Posterior Measure. The examples given in this section suggest a general approach to the rigorous mathematical formulation of a range of problems defined through a probability measure on function space. A fundamental step in such a formulation is a choice of prior measure π_0 for which Φ in (1) is π_0-measurable and the Radon-Nikodym derivative (1) is π_0–integrable. In the data assimilation application this is intimately connected with the question of determining sufficient regularity on the initial velocity field u so that Lagrangian tracers are well-defined. Similarly, in the geophysics application, it is necessary to specify sufficient regularity on the log-permeability field to ensure that the coupled equations for pressure and water saturation have a unique solution. The regularity of samples from a Gaussian measure on function space can be understood in terms of the rate of decay of eigenvalues of the covariance operator $\mathcal{C}$, via the Karhunen-Loève expansion. In this context it is natural in many applications to specify $\mathcal{C}$ through a precision operator $\mathcal{L} = -\mathcal{C}^{-1}$ which is a differential operator as then the full power of spectral theory for differential equations can be used. In many applications the primary role of the prior measure will indeed be to specify regularity information. However in the geophysics application the situation is somewhat different as there exists direct experimental evidence concerning the covariance function $c(\cdot,\cdot)$ which must also be combined with regularity issues to determine the prior. In both the data assimilation and geophysics applications this complete rigorous mathematical formulation is not carried out in this article, but is left for future study. It is our belief that there are a wide range of problems which will benefit from such an analytical investigation. A rigorous formulation of the first two examples, from molecular dynamics and signal processing, is undertaken in [17].

3. Langevin Stochastic PDEs

Underpinning the probability measure π on $\mathcal{H}$ given by (1) is a stochastic partial differential equation for which π is invariant. This is an infinite dimensional Langevin equation. In terms of the precision operator $\mathcal{L} = -\mathcal{C}^{-1}$ this Langevin equation may be written as an SDE on Hilbert space with the form

$$\frac{dx}{ds} = \mathcal{L}(x-m) - D\Phi(x) + \sqrt{2}\,\frac{dW}{ds} \tag{8}$$

where W is an $\mathcal{H}$-valued Brownian motion. This equation is written down in [16] and can be given a rigorous interpretation in many concrete situations: see [15, 17]. It corresponds to a noisy gradient flow for the functional found as the logarithm of the formal expression (3) for the probability density on function space. We now give several explicit instances of this Langevin SPDE connected with the examples introduced in the previous section. For general background concerning SPDEs see [8, 13, 30].

3.1. Molecular Dynamics. In the bridge diffusion case of (4), (5), arising in Brownian dynamics models of thermally activated atomic motion, the Langevin equation takes the form

$$\frac{\partial x}{\partial s} = \frac{\beta}{2}\frac{\partial^2 x}{\partial t^2} - \frac{\beta}{2}\nabla G(x;\beta) + \sqrt{2}\,\frac{\partial W}{\partial s}, \tag{9a}$$

$$x(0,s) = x^- \quad \text{and} \quad x(T,s) = x^+, \tag{9b}$$

$$x(t,0) = x_0(t). \tag{9c}$$

Here the last term in (9a) is space-time white noise. This SPDE is derived in [17, 25, 31]. Notice that t, the spatial variable in the SPDE, represents the real time in (4) whereas s, the time-like variable in the SPDE, is an artificial "algorithmic" time.

3.2. Signal Processing. In the signal processing case the objective is to sample a path of x from (6) given a single realization of the observation y from (7). The SPDE which is invariant with respect to this conditional distribution of x is as follows:

$$\begin{aligned}\frac{\partial x}{\partial s} = \frac{\partial^2 x}{\partial t^2} - (\nabla f(x) - \nabla f(x)^\top)\frac{\partial x}{\partial s} - \nabla_x F(x) + \sqrt{2}\,\frac{\partial W}{\partial s}\\ + dg(x,y)^\top(\sigma\sigma^\top)^{-1}\Big(\frac{dy}{dt} - g(x,y)\Big) - \frac{1}{2}\nabla_x\big(\nabla_y \cdot g(x,y)\big)\end{aligned}$$

$$\frac{\partial x}{\partial t} = \big(f(x) - \nabla_x \ln\zeta(x)\big), \quad t=0, \qquad \frac{\partial x}{\partial t} = f(x), \quad t=1,$$
$$x = x_0, \quad s=0.$$

Here

$$F(x) = \frac{1}{2}|f(x)|^2 + \frac{1}{2}\nabla\cdot f(x).$$

This SPDE is derived in [17, 31].

3.3. Lagrangian Data Assimilation. Recall that in this case we take $\mathcal{L}$ to be the square of the Stokes operator and

$$\Phi(u) = \frac{1}{2}|y - \mathcal{G}(u)|_\Sigma^2$$

where $\mathcal{G}$ maps the initial data for the velocity field into the positions of a Lagrangian tracer. We have

$$D\Phi(u) = -D\mathcal{G}(u)^\top\Sigma^{-1}\Big(y - \mathcal{G}(u)\Big).$$

Note that $D\mathcal{G}$ requires knowledge of the derivative of the Navier-Stokes equations with respect to initial data. The Langevin stochastic PDE is

$$\frac{\partial u}{\partial s} = -\nu^2\Delta^2 u - \nabla p + D\mathcal{G}(u)^\top\Sigma^{-1}\left(y - \mathcal{G}(u)\right) + \sqrt{2}\,\frac{\partial W}{\partial s}, \tag{11a}$$

$$\nabla\cdot u = 0, \tag{11b}$$

together with periodic boundary conditions on Ω and a divergence free initial condition. Here the last term in (11a) is space-time white noise in $\mathcal{H}$. As in the Navier-Stokes equations themselves, the pressure p is a Lagrange multiplier which acts to enforce the incompressibility condition.

3.4. Geophysics. The geophysical application and the Lagrangian data assimilation problem share a common mathematical structure, with the exception of the choice of the precision operator $\mathcal{L}$. Consequently the Langevin stochastic differential equation in this case is

$$\frac{\partial u}{\partial s} = \mathcal{L}u + D\mathcal{G}(u)^{\top}\Sigma^{-1}\left(y - \mathcal{G}(u)\right) + \sqrt{2}\,\frac{\partial W}{\partial s} \tag{12}$$

together with an initial condition. Here $\mathcal{G}$ maps the log-permeability into the fractional flow at the boundary, hence its derivative will be a complex object. The operator $\mathcal{L}$ is not necessarily a differential operator in this application: it may be a non-local operator. So, in this case equation (12) is not necessarily an SPDE.

3.5. Mathematical Structure of the Langevin Equation. Many outstanding questions remain concerning the rigorous formulation of the above Langevin SDEs. Such questions have been resolved for the bridge diffusion measure arising in the molecular dynamics example in [17], and the signal processing problems for some limited choice of vector fields (f, g): the pair should be the sum of a linear function plus a gradient [17]. For the general signal processing problem there are still open questions [16]. Similarly, checking that the SPDEs for data assimilation and for the geophysics application are well-posed remains an open question. As we will see, discretizations of the Langevin SPDE provide good proposals for MCMC methods and in this context development of the rigorous underpinnings of the subject revolve around showing that the MCMC methods can be defined on function space. Doing so is intimately bound up with the construction of effecient MCMC methods, as shown in [2]. It is to the subject of MCMC methods that we now turn.

4. Metropolis-Hastings Methods

We have illustrated that a wide range of problems can be written in a single unifying framework: that of a probability measure on Hilbert space with Radon-Nikodym derivative with respect to a Gaussian measure. Formulating the problems in this way is, of course, simply the first step in their resolution. The second step is to develop methods to interrogate the probability measure and thereby extract information from it. In practice we must discretize the function space (via finite differences, finite elements or spectral methods for example) leading to a high dimensional measure on $\mathbb{R}^n$ with $n \gg 1$. Sampling probability measures in high dimensions is notoriously hard. A generic approach to sampling that has seen spectacular success in recent years is the Markov chain Monte-Carlo (MCMC) methodology [21, 26]. A particular variant of this approach, which we will employ for our problems, is the Metropolis-Hastings method [23, 19]. In the next section we overview the analysis of such algorithms, when applied to measures arising from discretization of the structure (1), and show how our set-up fits into a broader context concerning the analysis of Metropolis-Hastings methods in high or infinite dimensions. In this section we give the necessary background concerning the MCMC methodology.

We start by discussing a variety of forms of target measure that have been studied in the literature, introducing a hierarchy of increasing complexity which eventually leads to discretizations of (1). We then explain how the Metropolis-Hastings method works in general, illustrating that the key tunable parameters arise through the choice of the *proposal distribution.* Finally, we introduce a range of proposal distributions appropriate for sampling measures such as (1) and its discretizations.

4.1. Structure of the Target. The following hierarchy of target measures will be central in our discussion of the computational complexity of Metropolis-Hastings methods in high dimensions.

- **IID Product** in $\mathbb{R}^n$. The earliest attempts to understand the behaviour of MCMC methods in $\mathbb{R}^n, n \gg 1$, concentrated on measures of product form in which each component is independent and identically distributed with density proportional to f (see [14] and references therein to the physics literature which preceded that work). Clearly, such measures are not intrinsically high dimensional as only one component need be sampled accurately to determine the entire measure. However the Metropolis-Hastings algorithm couples the different components, through the proposal, so study of these measures does provide an interesting starting point for analysis of MCMC methods in high dimensions. The structure of the target distribution π is now

$$\pi(x) = \Pi_{i=1}^n f(x_i).$$

- **Scaled Product** in $\mathbb{R}^n$. An interesting variant of the IID product is the case where independence is retained but the independent components are no longer identical. Specifically they are all derived by scaling a single measure on $\mathbb{R}$ with density f. The target measure is now

$$\pi(x) = \Pi_{i=1}^n \frac{1}{\lambda_i} f\Big(\frac{x_i}{\lambda_i}\Big).$$

  Assuming for simplicity that the measure on $\mathbb{R}$ has mean 0 and unit variance, the variance of each component is λ_i^2.

- **Change of Measure from Product** in $\mathbb{R}^n$. Product measures are intrinsically limiting for applications. Change of measure from product is a far more general setting. We will now consider target measures of the form

$$\pi(x) \propto \exp\Big(-\Phi_n(x)\Big) \Pi_{i=1}^n \frac{1}{\lambda_i} f\Big(\frac{x_i}{\lambda_i}\Big). \tag{13}$$

  Here we allow for dependency among the different components of x via the presence of Φ_n. We will show that, under certain conditions on Φ_n as $n \to \infty$, the behavior of Metropolis-Hastings methods on targets like (13) can be very similar to that arising in the scaled product case. We will give some motivation for these results in the sequel.

- **Change of Measure from Gaussian** in $\mathbb{R}^n$. If $f(x) = \exp(-x^2/2)$ then the product measure is Gaussian and the form (13) becomes

$$\pi(x) \propto \exp\Big(-\Phi_n(x) + \frac{1}{2}\langle x, \mathcal{L}_n x\rangle\Big) \tag{14}$$

with $\mathcal{L}_n$ a diagonal matrix with entries $-1/\lambda_i^2$. More generally the structure (14) is of interest for any negative definite *precision matrix* $\mathcal{L}_n$. Viewed in this context, we see that the structure (14) is exactly what will arise from an approximation of the measure (1) which is of interest to us in this article.

4.2. Metropolis-Hastings Algorithm.

The basic idea of MCMC is to generate a sequence $\{x_j\}_{j=1}^J$ which, for large J, produces a set of approximate draws from a given target measure π. This is done by creating a Markov chain for which π is invariant. The approximate samples x_j from π are correlated. The MCMC method is very flexible allowing for the construction of a wide range of methods with the aforementioned properties. A key issue is the construction of methods which minimize correlation amongst samples, thereby increasing efficiency.

The Metropolis algorithm, a particular MCMC method, was introduced in [23] where it was used by physicists aiming at calculating averages under the Boltzmann distribution. It was later generalized by Hastings in [19]. The algorithm has proven particularly effective in a range of applications; we will concentrate on this variant of MCMC methods here.

The goal is to sample $\pi : \mathbb{R}^n \mapsto \mathbb{R}^+$. The idea of the method is, given an approximate sample x_j, to *propose* a new sample y from some Markov chain with transition kernel $q(x_j, \cdot)$. This proposal is then accepted ($x_{j+1} = y$) with probability $a(x_j, y)$ and rejected ($x_{j+1} = x_j$) otherwise. The composition of proposal from a Markov kernel and the accept-reject criteria gives a modified Markov chain. If

$$a(x, y) = \min\Big(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\Big) \tag{15}$$

then the resulting Markov chain for the sequence $\{x_j\}_{j=1}^J$ is π-invariant and will, for large J, generate samples from π under mild ergodicity hypotheses [21, 26]. The following piece of pseudo-code defines the algorithm:

Algorithm 1.

```
1. Set $j = 0$.  Pick $x_0 \in \mathbb{R}^n$.
2. Given $x_j$ propose $y \sim q(x_j, \cdot)$.
3. Calculate $a(x_j, y)$.
```

```
4. Set $x_{j+1} = y$ with probability $a(x_j, y)$.
5. Otherwise set $x_{j+1} = x_j$.
6. Set $j = j + 1$ and return to 2.
```

We mentioned that key to success of the algorithm is minimizing correlation in the generated sequence. From this point of view, the acceptance probability is clearly a key object of interest: if it is small (on average) then the sequence will be highly correlated. In the high dimensional case that we study here our focus will be on defining appropriate proposals which ensure that the acceptance probability is bounded away from zero, on average, as the dimension grows $n \to \infty$. We now turn to the class of proposals which effect this.

4.3. Proposals for Metropolis-Hastings.

Consider a target density $\pi : \mathbb{R}^n \mapsto \mathbb{R}^+$. A commonly used family of proposals are *random-walks* for which $q(x, y)$ is the transition kernel associated with the proposal

$$y = x + \sqrt{2\Delta s}\,\xi \tag{16}$$

where $\xi \sim \mathcal{N}(0, I)$ is a standard Gaussian random variable in $\mathbb{R}^n$. These proposals are very simple to implement but, as we will see, can suffer from (relatively) high rejection rate due to the fact that they contain no information about π. For what comes next it is instructive to note that the proposal (16) can be seen as a discretization of the SDE

$$\frac{dx}{ds} = \sqrt{2}\,\frac{dW}{ds}.$$

This SDE contains no information about the target π. In contrast, the Langevin SDE

$$\frac{dx}{ds} = \nabla \log \pi(x) + \sqrt{2}\,\frac{dW}{ds} \tag{17}$$

is π-invariant if W is an $\mathbb{R}^n$-valued Brownian motion; a straightforward calculation with the Fokker-Planck equation will show this. Equation (8) is an infinite dimensional version of this SDE, applied to the formal density (3). If we could sample exactly from the transition density for equation (17) over some time-increment Δs, we would obtain a perfect proposal: it would be accepted with probability 1, and a large enough choise of Δs would ensure lack of correlation among samples. Unfortunately it is not possible, in general, to sample from this transition density. However we can discretize the equation in s to obtain proposals which approximate this distribution and hence, for small Δs, should deliver reasonable acceptance probability. We now pursue this idea further.

It turns out that there is a whole family of equations, including (17) as a special case, which are π-invariant. For any positive-definite self-adjoint matrix $\mathcal{A}$ the SDE

$$\frac{dx}{ds} = \mathcal{A}\nabla \log \pi(x) + \sqrt{2\mathcal{A}}\,\frac{dW}{ds} \tag{18}$$

is π-invariant[2]. Many of the proposals we consider below arise from discetization of equations of this type.

[2] Making these assertions about π-invariance rigorous in infinite dimensions requires being

5. Computational Complexity

We now explain a heuristic approach for selecting the time-step Δs in the proposals mentioned above, with a view toward optimizing the acceptance probability. We will choose the time-step as an inverse power of the dimension n of the state-space so that

$$\Delta s = n^{-\gamma}. \tag{19}$$

Note that the proposal y is now a function of: (i) the current state x; (ii) the parameter γ through the time-step scaling above; and (iii) the noise ξ which will appear in all the proposals that we consider. Thus $y = y(x, \xi; \gamma)$. We would like γ to be as small as possible, so that the chain will be making large steps and decorrelation amongst samples will be maximised. However, we would additionally like to ensure that the acceptance probability does not degenerate to 0 as $n \to \infty$, also to prevent high correlation amongst samples. To that end we define γ_0 as follows:

$$\gamma_0 = \min\nolimits_{\gamma_c \geq 0} \Big\{ \gamma_c : \liminf_{n\to\infty} \mathbb{E} a(x,y) > 0 \; \forall \gamma \in [\gamma_c, \infty) \Big\}.$$

Here the expectation is with respect to x distributed according to π and y chosen from the proposal distribution. In other words, we take the largest possible time-steps, as a function of n, constrained by asking that the average acceptance probability is bounded away from zero, uniformly in n. The resulting time-step restriction (19) is reminiscent of a Courant restriction arising in the numerical solution of PDEs.

Carrying this analogy further, we introduce the heuristic that the number of steps required to reach stationarity is given by

$$M(n) = n^{\gamma_0}.$$

As we will discuss below, this heuristic can be given a firm foundation in a number of cases. Here we simply note that, in these cases, the Markov chain arising from the Metropolis-Hastings method approximates a Langevin SDE; one could think of the Markov chain as traveling with time-step Δs on the paths of the Langevin SDE. It takes $\mathcal{O}(1)$ for the limiting SDE to reach stationarity, so in terms of the time-step Δs we obtain the expression for $M(n)$ above. We give more details on this point in the sequel.

Our goal now is to understand how $M(n)$ depends on the structure of the target distribution and the choice of proposal distribution. At our disposal are the form of the discretization and the form of $\mathcal{A}$. We will carry out such a study for the hierarchy of target distributions introduced in subsection 4.1. We require a regularity condition on the density f.

Condition 1. *(i) All moments of f are finite. (ii) $\log f$ is infinitely differentiable; $\log f$ and all its derivatives have a polynomial growth bound.*

much more specific about the problem; for the set-up of subsections 2.1 and 2.2. such a task is carried out in [17].

All results are obtained under Condition 1 which we assume to hold throughout without further mention. For clarity of exposition all the proofs are collected in the Appendix; within this section we confine ourselves to a brief discussion of the results. In this article we make strong conditions on the scalings λ_i and the change of measure Φ_n in order to simplify the proofs. Weaker conditions, and stronger theoretical results, are given in [3].

5.1. IID Products. Here we consider the case of target density with the form

$$\pi(x) = \Pi_{i=1}^{n} f(x_i).$$

We discuss two different proposals $y = y(x, \xi)$ found by setting $\beta = 0$ and $\beta = 1$ in the following formula:

$$\frac{y - x}{\Delta s} = \beta \, \nabla \log \pi(x) + \sqrt{\frac{2}{\Delta s}}\, \xi, \quad \xi \sim \mathcal{N}(0, I).$$

The choice $\beta = 0$ corresponds to the random walk proposal (16) whereas $\beta = 1$ corresponds to an Euler-Maruyama discretization of the Langevin SDE (17).

Theorem 1.

- *If* $\beta = 0$ *then* $M(n) = \mathcal{O}(n)$.
- *If* $\beta = 1$ *then* $M(n) = \mathcal{O}(n^{1/3})$.

We provide a direct proof of Theorem 1 only for completeness, since these results are implicit in the pair of papers [14, 27] (see also the survey [28]). In fact in these papers the much stronger result of convergence, as $n \to \infty$, of any scalar component of the n-dimensional Markov chain to that of a Langevin diffusion, is demonstrated. To be more precise, if $x_1^{(i)}, x_2^{(i)}, \ldots$ is the trajectory of the i^{th} scalar component, then by appropriately tuning $\Delta s \propto n^{-\gamma_0}$, the continuous-time process $\{x_{[s\, n^{\gamma_0}]}^{(i)}; s \geq 0\}$ converges to a Langevin diffusion. Such a result justifies the statement that the number of steps to reach stationarity is of the order $M(n) = n^{\gamma_0}$.

The basic takehome message of Theorem 1 is that using steepest ascents information in the proposal which, for small Δs, suggests moves in the direction of modes of the distribution, positively impacts the computational complexity of Metropolis-Hastings algorithms for iid target densities in high dimension. We now take this idea further.

5.2. Scaled Products. Now consider the target density of the form

$$\pi(x) = \Pi_{i=1}^{n} \frac{1}{\lambda_i} f\Big(\frac{x_i}{\lambda_i}\Big) \tag{20}$$

with $\lambda_i = i^{-\kappa}$ for some $\kappa > 0$. Thus, the target measure is of product form with the i^{th} component having variance $i^{-2\kappa}$ times a constant. We saw in the previous theorem that including steepest descent information improves complexity. For this reason we will henceforth work only with proposals arising from discetizations which include the $\nabla \log \pi$ term. Specifically, we employ discretizations of the Langevin equation in the form (18) giving

$$\frac{y-x}{\Delta s} = \mathcal{A}\,\nabla \log \pi(x) + \sqrt{\frac{2\mathcal{A}}{\Delta s}}\,\xi, \quad \xi \sim \mathcal{N}(0,I). \tag{21}$$

We define the diagonal matrix $\mathcal{C}_n = \mathrm{diag}\{\lambda_1^2, \cdots, \lambda_n^2\}$.

Theorem 2.

- *If $\mathcal{A} = I$ then $M(n) = \mathcal{O}(n^{2\kappa+1/3})$.*
- *If $\mathcal{A} = \mathcal{C}_n$ then $M(n) = \mathcal{O}(n^{1/3})$.*

Matrix $\mathcal{A}$ can be viewed as a preconditioner which, in the case $\mathcal{A} = \mathcal{C}_n$, acts by placing different components on the same scale. By doing so, it is possible to optimize the time-step Δs for all components of the proposal, resulting in a substantial impovement in computational complexity. Thus, the takehome message from this theorem is that preconditioning positively impacts complexity of Metropolis-Hastings algorithms. The proof of this result is given in the Appendix. It should be noted however that the result can be proved by a straightforward generalization of the ideas in [27]. The theorem is readily extended to the case where the λ_i are replaced by $\lambda_{i,n}$ satisfying algebraic upper and lower bounds in i, uniformly in n – see [3]. Related results, for scalings somewhat different in nature from those considered here, may be found in [1].

5.3. Change of Measure.

In both of the previous sections the target measure was of product type and hence not fundamentally high dimensional as each component could be considered separately. We now move away from this restrictive assumption and consider targets of the form

$$\pi(x) \propto \exp\Big(-\Phi_n(x)\Big)\pi_0(x),$$
$$\pi_0(x) = \Pi_{i=1}^n \frac{1}{\lambda_i} f\Big(\frac{x_i}{\lambda_i}\Big).$$

Similarly to the previous section, we assume that $\lambda_i = i^{-\kappa}$. We use a family of proposals which, in the product case, coincides with the proposal (21):

$$\frac{y-x}{\Delta s} = \mathcal{A}\,\nabla \log \pi_0(x) + \sqrt{\frac{2\mathcal{A}}{\Delta s}}\,\xi, \quad \xi \sim \mathcal{N}(0,I). \tag{22}$$

We also assume the following uniform bound on Φ_n:

$$\sup_{n\in\mathcal{Z}^+, x\in\mathbb{R}^n} \Big(|\Phi_n(x)|\Big) < \infty.$$

Defining $\mathcal{C}_n$ as in the previous section we have the following result.

Theorem 3.

- *If $\mathcal{A} = I$ then $M(n) = \mathcal{O}(n^{2\kappa+1/3})$.*
- *If $\mathcal{A} = \mathcal{C}_n$ then $M(n) = \mathcal{O}(n^{1/3})$.*

We prove this theorem in the Appendix. The takehome message from this theorem is that the change of measure does not affect the computational complexity. The boundedness assumption on Φ_n is very severe and mostly considered for clarity of exposition. Weaker and more pragmatic conditions, based on Lipschitz properties of a limiting Φ on Hilbert space, may be found in [3]. The intution behind all the results concerning change of measure, both here and in [3], is that we work under conditions on the Φ_n under which the reference product measure structure dominates in the tails; such a situation arises naturally when approximating infinite dimensional measures with Radon-Nikodym derivative (1) with respect to a product measure π_0.

Note that a proposal derived from the discretization of the Langevin SDE (18) would take the form

$$\frac{y-x}{\Delta s} = \mathcal{A}\,\nabla \log \pi_0(x) - \nabla\Phi_n(x) + \sqrt{\frac{2\mathcal{A}}{\Delta s}}\,\xi, \quad \xi \sim \mathcal{N}(0,I) \tag{23}$$

instead of (22). However we have omitted the term $\nabla\Phi_n(x)$ in (22) to simplify the proof of the complexity results in the above theorem, and because the resulting proposal suffices (under the stated conditions on Φ_n) to deliver an algorithm which has the same computational complexity as the one corresponding to product targets in Theorem 2; this is in some sense (and apart from extraordinary choices of Φ_n) the best one can expect. However, whilst use of the proposal (23) might not improve the asymptotic computational complexity in n, when compared with the results obtained for the proposal (22), it can have a significant positive effect in terms of the constant in the asymptotic cost, and in other measures of efficiency.

5.4. Change of Measure from Gaussian.

In the previous section we made the useful step of considering settings which are no longer of product form, taking us into a family of problems with practical application. Here we take a further step in the direction of applicability, by assuming that the reference measure π_0 is Gaussian so that the target has the form

$$\pi(x) \propto \exp\Big(-\Phi_n(x) + \frac{1}{2}\langle x, \mathcal{L}_n x\rangle\Big). \tag{24}$$

We have used $\mathcal{L}_n = -\mathcal{C}_n^{-1} = \mathrm{diag}\{-\lambda_1^{-2}, \cdots, -\lambda_n^{-2}\}$. We consider a family of proposals parameterised by a $\theta \in [0,1]$ which, in the Gaussian reference measure case, is identical to that from the previous section, in the case $\theta = 0$. When $\theta \in (0,1)$ the family corresponds to using an implicit discretization of (18):

$$\frac{y-x}{\Delta t} = \mathcal{A}\left(\theta\mathcal{L}y + (1-\theta)\mathcal{L}x\right) + \sqrt{\frac{2\mathcal{A}}{\Delta t}}\,\xi, \quad \xi \sim \mathcal{N}(0,I).$$

We make the same assumption on Φ_n as in the previous section.

Theorem 4.

- *If* $\theta = \frac{1}{2}$ *and* $\mathcal{A} = I$ *or* $\mathcal{A} = \mathcal{C}_n$ *then* $M(n) = \mathcal{O}(1)$.
- *If* $\theta = 0$ *and* $\mathcal{A} = \mathcal{C}_n$ *then* $M(n) = \mathcal{O}(n^{1/3})$.
- *If* $\theta = 0$ *and* $\mathcal{A} = I$ *then* $M(n) = \mathcal{O}(n^{2\kappa+1/3})$.

Thus the takehome message from this theorem is that implicitness in the proposal can positively impact computational complexity. It turns out that the choice $\theta = \frac{1}{2}$ is crucial to obtaining n-independent estimates on $M(n)$. This is due to the fact that $\theta = \frac{1}{2}$ is the unique choice of θ for which the Metropolis-Hastings method is well-defined on the limiting (for $n \to \infty$) infinite dimensional Hilbert space $\mathcal{H}$. This result is proved in [2]; for numerical illustrations of the effect of θ see that paper and [18].

The results of Theorem 4 are directly relevant to the infinite dimensional models of interest in this paper characterised by the general density structure π in (3) and the π-invariant SPDE (8). The target π_n in (24) should be viewed as an approximation of π. One can readily obtain such a structure for a finite dimensional approximation of π by truncating the spectral expansion corresponding to eigenbasis of the covariance operator $\mathcal{C}$ of the reference Gaussian measure appearing at the definition of π. Equivalently, this corresponds to an n-dimensional projection of the Karhunen-Loève expansion for Gaussian measures. Other methods, like finite differences or finite elements can deliver a similar structure. In these cases note that appropriate orthogonal transformations can force a diagonal structure for the approximation of the covariance operator, thus granting the structure (24). In terms of results, Theorem 4 dictates that one should use a θ-method for the discretization of the SPDE (8) in the algorithmic time s-direction, with the particular choice $\theta = \frac{1}{2}$.

6. Conclusions

In this article we have studied a class of problems that lie at the interface of applied mathematics and statistics. We have illustrated the following:

- **Applications.** Measures which have density with respect to a Gaussian arise naturally in many applications where the solution is a measure on functions.
- **SPDEs.** There is a natural notion of Langevin equations on function space for these measures. These Langevin equations are often stochastic partial differential equations (SPDEs).
- **Algorithms.** Using these SPDEs, and their finite dimensional analogues, natural MCMC methods can be constructed to sample function space.
- **Numerical Analysis.** Ideas such as steepest descents, preconditioning and implicitness have crucial impact on the complexity of MCMC algorithms.

Many interesting issues remain open for further study:

- **Mathematical Formulation.** As indicated in subsection 2.5, providing a rigorous formulation of many problems which require a measure on function space, especially inverse problems, is an open and interesting area for analysis.

- **Algorithms.** In theory it is advantageous to incorporate information concerning $\nabla\Phi_n(x)$ (as in (23)) in the proposal. In practice, calculation of this derivative may be very expensive: study of the data assimilation [20] and geophysical applications [22] will illustrate this. Thus, it is important to find cheaper surrogates for $\nabla\Phi$ which result in improved acceptance probabilities.

- **Applications.** As we have shown these are numerous in chemistry, physics, data assimilation, signal processing and econometrics. Realizing the potential for the methodology studied here remains a significant challenge.

- **Stochastic Analysis.** The existing theory of π-invariant SPDEs would benefit from extension, in the case of conditioned diffusions, to non-gradient vector fields, state-dependent noise, degenerate noise and non-Gaussian noise. More generally, in particular for inverse problems, making sense of the resulting SPDEs remains an open and interesting problem - see subsection 3.5.

- **Numerical Analysis.** It is important to develop an approximation theory for the S(P)DEs and MCMC methods on function space written down in this article. Challenging issues include nonlinear boundary conditions, nonlinear Dirac sources, and preserving symmetry of the inverse covariance matrix.

- **Statistics.** Incorporation of this function space sampling into the (Gibbs) sampler to estimate parameters as well as functions. Study of optimal scaling of proposals in various singular limits, such as small diffusion in the case of bridge diffusions or signal processing, or rapidly varying permeability in the case of geophysical applications.

Apart from the intrinsic interest in the class of problems studied here, and the specific conlclusions listed, the work presented here is perhaps also of interest because it highlights an important general trend, namely that applied mathematics and statistics are increasingly required to work in tandem in order to tackle significant problems in science, engineering and beyond.

Acknolwedgements. We are very grateful to Yalchin Efendiev, Frank Pinski, Gareth Roberts and Jochen Voss for comments on early drafts of this article, and to Efendiev (Figure 2.4), Pinski (Figure 2.1) and Voss (Figures 2.2 and 2.3) for providing the illustrations in the article.

References

[1] M. Bédard, *Weak Convergence of Metropolis algorithms for non-i.i.d. target distributions.* Ann. Appl. Probab. **17**(2007), 1222-1244.

[2] A. Beskos, G.O. Roberts, A.M. Stuart and J. Voss. *An MCMC Method for diffusion bridges.* Submitted.

[3] A. Beskos, G. Roberts and A.M. Stuart. *Scalings for local Metropolis-Hastings chains on non-product targets.* Submitted.

[4] P.G. Bolhuis, D. Chandler, C. Dellago and P.L. Geissler *Transition path sampling: Throwing ropes over rough mountain passes, in the dark.* Ann. Rev. Phys. Chem. **53**(2002), 291-318.

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

[6] O. Elerian, S. Chib, N. Shephard. *Likelihood inference for discretely observed nonlinear diffusions.* Econometrica **69**(2001), 959–993.

[7] A.J. Chorin and O.H. Hald. *Stochastic tools in mathematics and science*, volume 1 of *Surveys and Tutorials in the Applied Mathematical Sciences.* Springer, New York, 2006.

[8] G. Da Prato and J. Zabczyk. *Stochastic equations in infinite dimensions.* Cambridge University Press, 1992.

[9] M. Dashti and J. Robinson. *Uniqueness of the particle trajectories of the weak solutions of the two-dimensional Navier-Stokes equations.* Arch. Rat. Mech. Anal. (2007). Submitted.

[10] P. Dostert, Y. Efendiev, T.Y. Hou and W. Luo, *Coarse-grain Langevin algorithms for dynamic data integration and uncertainty quantification.* J.Comp. Phys, **217**(2006), 123–142.

[11] W. E, W.Q. Ren and E. Vanden-Eijnden. *String method for the study of rare events.* Phys. Rev. B **66**, 2002, p. 052301.

[12] M.I. Freidlin and A.D. Wentzell. *Random Perturbations of Dynamical Systems.* Springer-Verlag, New York, 1998.

[13] G. Garcia-Ojalvo and J.M. Sancho, *Noise in Spatially Extended Systems* Springer(1999).

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

[15] M. Hairer, A.M.Stuart, J. Voss and P. Wiberg. *Analysis of SPDEs Arising in Path Sampling. Part 1: The Gaussian Case.* Comm. Math. Sci. **3**(2005), 587–603

[16] M. Hairer, A.M.Stuart and J. Voss. *Sampling the posterior: an approach to non-Gaussian data assimilation.* PhysicaD, **230**(2007), 50–64.

[17] M. Hairer, A.M.Stuart and J. Voss. *Analysis of SPDEs Arising in Path Sampling. Part 2: The Nonlinear Case.* Ann. Appl. Prob., **17**(2007), 1657–1706.

[18] M. Hairer, A.M.Stuart and J. Voss. *Sampling conditioned diffusions.* To appear in "Trends in Stochastic Analysis", Cambridge University Press, 20 pages (2008).

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

[20] F.-X. Le Dimet and O. Talagrand, *Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects.* Tellus A, **38**(1986), 97–110.

[21] J. Liu, *Monte Carlo Strategies in Scientific Computing.* Springer Texts in Statistics, Springer-Verlag, New York, 2001.

[22] X. Ma, M. Al-Harbi, A. Datta-Gupta, and Y. Efendiev. *Multistage sampling approach to quantifying uncertainty during history matching geological models*, Soc. Petr. Eng. Journal (2007), to appear.

[23] N. Metropolis, A.W. Rosenbluth, M.N. Teller and E. Teller, *Equations of state calculations by fast computing machines.* J. Chem. Phys. **21**(1953), 1087–1092.

[24] B. Oksendal, Stochastic differential equations, Springer, New York, 1998.

[25] M. Reznikoff and E. Vanden Eijnden. Invariant measures of SPDEs and conditioned diffusions. C.R. Acad. Sci. Paris, **340**(2005), 305 - 308

[26] C.P. Robert and G.C. Casella, *Monte Carlo Statistical Methods.* Springer Texts in Statistics, Springer-Verlag, 1999.

[27] G.O. Roberts and J. Rosenthal, *Optimal scaling of discrete approximations to Langevin diffusions.* JRSSB **60**(1998), 255–268.

[28] G.O. Roberts and J. Rosenthal, *Optimal scaling for various Metropolis-Hastings algorithms.* Statistical Science **16**(2001), 351–367.

[29] J.C. Robinson, *Infinite-Dimensional Dynamical Systems.* Camrbridge University Press, Cambridge, 2001.

[30] B. Rozovskii, *Stochastic evolution systems: linear theory and applications to non-linear filtering.* Kluwer Academic Publishers, 1990.

[31] A.M.Stuart, J. Voss and P. Wiberg. *Conditional Path Sampling of SDEs and the Langevin MCMC Method.* Comm. Math. Sci. **2**(2004), 685–697.

Appendix. Proof of Theorems

The following generic result will allow us to obtain estimates for the Metropolis-Hastings acceptance probability.

Lemma 1. *Let T be a real-valued random variable.*

i) For any $c > 0$:

$$\mathbb{E}\,[\,1 \wedge e^{T}\,] \geq e^{-c}\left(1 - \frac{\mathbb{E}\,|\,T\,|}{c}\right) .$$

ii) If $\mathbb{E}\,[\,T\,] < 0$*, then:*

$$\mathbb{E}\,[\,1 \wedge e^{T}\,] \leq e^{\mathbb{E}\,[\,T\,]/2} + 2\,\frac{\mathbb{E}\,|\,T - \mathbb{E}\,[\,T\,]\,|}{(-\mathbb{E}\,[\,T\,])}\ .$$

Proof. For the first result note that:

$$\mathbb{E}\,[\,1 \wedge e^{T}\,] \geq \mathbb{E}\,[\,(1 \wedge e^{T}) \cdot \mathbb{I}\{|\,T| \leq c\}\,] \geq e^{-c}\,\mathbb{P}\,[\,|\,T| \leq c\,]\ .$$

The Markov inequality now gives the required result. For the second result, we set $\mu := -\mathbb{E}\,[\,T\,]$, $T_0 := T - \mathbb{E}[\,T\,]$. Then:

$$\mathbb{E}\,[\,(1 \wedge e^{T}) \cdot \mathbb{I}\{|\,T_0| \leq \frac{\mu}{2}\}\,] + \mathbb{E}\,[\,(1 \wedge e^{T}) \cdot \mathbb{I}\{|\,T_0| > \frac{\mu}{2}\}\,] \leq e^{-\mu/2} + \mathbb{P}\,[\,|\,T_0|\, > \frac{\mu}{2}\,]\ .$$

The result follows from Markov inequality. □

For simplicity in the proofs that follow, we set:

$$g(x) = \log f(x)$$

and we use $g^{(j)}$ to denote the j^{th} derivative of g.

Proof of Theorem 1:

- $\beta = 0$

The acceptance probability $a(x, y)$ in (15) is now determined as follows:

$$a(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)} = 1 \wedge e^{R_n}; \quad R_n := \sum_{i=1}^{n} \Big(g(y_i) - g(x_i) \Big)\ .$$

Recall that since $\beta = 0$:

$$y_i = x_i + \sqrt{2\Delta s}\,\xi_i\ .$$

Case A: $\Delta s = n^{-\gamma}$ with $\gamma \geq 1$.

We take a second order Taylor expansion of $R_n = R_n(\sqrt{\Delta s})$ around $\sqrt{\Delta s} = 0$. So:

$$R_n = \mathcal{A}_{1,n} + \mathcal{A}_{2,n} + \mathcal{U}_n\ ,$$

with individual components:

$$\mathcal{A}_{1,n} = \sqrt{\Delta s}\,\sum_{i=1}^{n} C_{1,i} \quad \mathcal{A}_{2,n} = \Delta s\,\sum_{i=1}^{n} C_{2,i}, \quad \mathcal{U}_n = \Delta_s^{3/2} \sum_{i=1}^{n} U_{i,n};$$

$$C_{1,i} = \sqrt{2}\,g^{'}(x_i)\xi_i, \quad C_{2,i} = g^{''}(x_i)\xi_i^2, \quad U_{i,n} = \frac{\sqrt{2}}{3} g^{(3)}(x_i + \sqrt{2}\Delta_i^{*}\xi_i)\,\xi_i^3\ ,$$

for some $\Delta_i^* \in [0, \sqrt{\Delta s}]$, $i = 1, \ldots, n$. Notice that $\{C_{1,i}\}_i$ and $\{C_{2,i}\}_i$ are both sequences of iid random variables, so we will ignore reference to the index i when considering expectations w.r.t. $C_{1,i}$ or $C_{2,i}$. Using Condition 1(ii), we find that:

$$|U_{i,n}| \leq M_1(x_i)\, M_2(\xi_i)\, M_3(\Delta_i^*) \ , \tag{25}$$

for some positive polynomials M_1, M_2, M_3. Using Condition 1(i), $\mathbb{E}\,[\, M_1(x_i)\,] < \infty$, $\mathbb{E}\,[\, M_2(\xi_i)\,] < \infty$, both expectations not depending on i. Since Δ_i^* is bounded above uniformly in i, n, so is $M_3(\Delta_i^*)$. Since the x_i and ξ are independent of one another, it is now clear that $\mathbb{E}\,|U_{i,n}| \leq K_0$ for some constant K_0 not depending on i, n, and subsequently:

$$\lim_{n\to\infty} \mathbb{E}\,|\,\mathcal{U}_n\,| = 0 \ .$$

Note now that, since $\mathbb{E}\,[\,C_{1,\cdot}\,] = 0$, Jensen's inequality gives:

$$\mathbb{E}\,[\,|\mathcal{A}_{1,n}|\,] \leq \sqrt{\Delta s}\,\sqrt{n}\,\mathbb{E}\,[\,C_{1,\cdot}^2\,]^{1/2} \ .$$

Also,

$$\mathbb{E}\,[\,|\mathcal{A}_{2,n}|\,] \leq \Delta s\, n\, \mathbb{E}\,[\,|C_{2,\cdot}|\,] \ .$$

Since $\Delta s = n^{-\gamma}$ with $\gamma \geq 1$, we deduce that $\limsup_n \mathbb{E}\,|\,R_n\,| < \infty$. Lemma 1(i) now implies that:

$$\liminf_{n\to\infty} \mathbb{E}\, a(x,y) > 0 \ .$$

Case B: $\Delta s = n^{-\gamma}$ with $\gamma \in (0,1)$.

We select an integer m such that $(m+1)\gamma > 2$ and use the m^{th}-order Taylor expansion:

$$R_n = \sum_{j=1}^{m} \mathcal{A}_{j,n} + \mathcal{U}_n^{'} \ ,$$

with terms specified as follows:

$$\mathcal{A}_{j,n} := (\sqrt{\Delta s})^j \sum_{i=1}^{n} C_{j,i}, \quad \mathcal{U}_n^{'} = (\sqrt{\Delta s})^{m+1} \sum_{i=1}^{n} U_{i,n}^{'};$$

$$C_{j,i} = \frac{(\sqrt{2})^j}{j!} g^{(j)}(x_i)\xi^j, \quad U_{i,n}^{'} = \frac{(\sqrt{2})^{m+1}}{(m+1)!} g^{(m+1)}(x_i + \sqrt{2}\Delta_i^*\xi_i) \ .$$

for some corresponding $\Delta_i^* \in [0, \sqrt{\Delta s}]$. The residual terms $U_{i,n}^{'}$ can be bounded by a constant as in (25), so the particular choice of m gives that:

$$\lim_{n\to\infty} \mathbb{E}\,|\mathcal{U}_n^{'}| = 0 \ .$$

Also, since $\mathbb{E}\,[\,\mathcal{A}_{1,n}\,] = 0$:

$$\mathbb{E}\,[\,R_n\,] = \sum_{j=2}^{m} \mathbb{E}\,[\,\mathcal{A}_{j,n}\,] + \mathcal{O}(1); \quad \mathbb{E}\,[\,\mathcal{A}_{j,n}\,] = (\sqrt{\Delta s})^j\, n\, \mathbb{E}\,[\,C_{j,\cdot}\,] \ .$$

From the analytical expression for $C_{2,i}$:

$$\mathbb{E}\,[\,C_{2,\cdot}\,] = -\int_{\mathbb{R}} \{g^{'}(x)\}^2 \exp\{g(x)\}dx < 0.$$

All other $C_{j,\cdot}$ satisfy $\mathbb{E}|C_{j,\cdot}| < \infty$. So, $\mathbb{E}\,[\,R_n\,] \rightarrow -\infty$ as fast as $-n^{1-\gamma}$. For the expectation $\mathbb{E}\,|\,R_n - \mathbb{E}\,[\,R_n\,]\,|$, we use Jensen's inequality to get the following upper bound:

$$\mathbb{E}\,|\,R_n - \mathbb{E}\,[\,R_n\,]\,| \leq \sum_{j=1}^{m} (\sqrt{\Delta s})^j \,\sqrt{n}\,\mathrm{Var}\,[\,C_{j,\cdot}\,]^{1/2} + \mathcal{O}(1)\ .$$

So, $\mathbb{E}\,|R_n - \mathbb{E}\,[\,R_n\,]\,|$ does not grow faster than $(\,|\,\mathbb{E}[R_n]\,|\,)^{1/2}$. From Lemma 1(ii):

$$\lim_{n\rightarrow\infty} \mathbb{E}\,a(x,y) = 0\ .$$

- $\beta = 1$

The proof follows the same lines as the case $\beta = 0$. The acceptance probability $a(x,y)$ can be written again as $1 \wedge e^{R_n}$ for some corresponding R_n. We will again consider Taylor expansions of $R_n = R_n(\sqrt{\Delta s})$ around $\sqrt{\Delta s} = 0$. So, considering an m^{th}-order expansion we obtain the following structure:

$$R_n(\sqrt{\Delta s}) = \sum_{j=1}^{m} \mathcal{A}_{j,n} + \mathcal{U}_n; \qquad (26)$$

$$A_{j,n} = (\sqrt{\Delta}_s)^j \sum_{i=1}^{n} C_{j,i}, \quad \mathcal{U}_n = (\sqrt{\Delta s})^{m+1} \sum_{i=1}^{n} G(x_i,\xi_i,\Delta_i^*) \qquad (27)$$

for some $C_{j,i}$, G involving g and it's derivatives, and some $\Delta_i^* \in [0,\sqrt{\Delta s}]$, $1 \leq i \leq n$. For the explicit expressions for $C_{j,i}$ and G see [27]. We will only exploit the following characteristics:

$$C_{j,i} = C_{j,\cdot}(x_i,\xi_i)\ , \qquad (28)$$
$$C_{1,i} = C_{2,i} \equiv 0,\ i = 1,\dots n; \quad \mathbb{E}\,[\,C_{3,\cdot}\,] = \mathbb{E}\,[\,C_{4,\cdot}\,] = \mathbb{E}\,[\,C_{5,\cdot}\,] = 0; \quad \mathbb{E}\,[\,C_{6,\cdot}\,] < 0\ ,$$
$$G \text{ has a polynomial growth bound}\ .$$

Since the first two terms in the expansion cancel out, a larger step-size $\sqrt{\Delta s}$ can now control the remaining term compared with the case $\beta = 0$. Working as above, we can show that:

$$\mathbb{E}\,|\,\mathcal{A}_{j,n}\,| \leq (\sqrt{\Delta s})^j \,\sqrt{n}\,\mathbb{E}\,[\,C_{j,\cdot}^2\,]^{1/2}, \quad j = 3,4,5\ ,$$
$$\mathbb{E}\,[\,|\mathcal{A}_{6,n}|\,] \leq (\Delta s)^3\, n\,\mathbb{E}\,|\,C_{6,\cdot}\,|\ .$$

So, when $\Delta s = n^{-\gamma}$ with $\gamma \geq 1/3$ all terms in a sixth-order Taylor expansion of $R_n(\sqrt{\Delta s})$ will have n-bounded absolute expectation, and Lemma 1(i) will again give the bound $\liminf_n \mathbb{E}\,a(x,y) > 0$. Using the same arguments as in the case when $\beta = 0$, one can also prove that $\lim_{n\rightarrow\infty} \mathbb{E}\,a(x,y) = 0$ if $\gamma \in (0,1/3)$. We avoid further details.

□

Proof of Theorem 2:

- $\mathcal{A} = I$

The proof is a slight modification of the proof of Theorem 1. Again, we consider the exponent $R_n = R_n(\sqrt{\Delta s})$ from the expression $1 \wedge e^{R_n}$ for the acceptance probability $a(x,y)$, and consider Taylor expansions of it around $\sqrt{\Delta s} = 0$. The formulae are similar to the ones for the iid case given in (26) and (27). Analytically:

$$R_n = \sum_{j=3}^{m} \mathcal{A}_{j,n} + \mathcal{U}_n;$$

$$A_{j,n} = (\sqrt{\Delta s})^j \sum_{i=1}^{n} C_{j,i}/\lambda_i^j, \quad \mathcal{U}_n = (\sqrt{\Delta s})^{m+1} \sum_{i=1}^{n} G(x_i/\lambda_i, \xi_i, \Delta_i^*/\lambda_i)/\lambda_i^{m+1} \ .$$

for some $\Delta_i^* \in [0, \sqrt{\Delta s}]$, $i = 1, \dots n$. The functional G is the same as in (27), whereas $C_{j,i} = C_{j,\cdot}(x_i/\lambda_i)$ for the functions $C_{j,\cdot}$ in (28); in particular $\{C_{j,i}\}_i$ are again iid for all $j \geq 1$.

We work as before. For $\gamma \geq 2\kappa + 1/3$, we consider the sixth-order expansion ($m = 6$), and find that:

$$|G(x_i/\lambda_i, \xi_i, \Delta_i^*/\lambda_i)| \leq M_1(x_i/\lambda_i)\, M_2(\xi_i)\, M_3(\Delta_i^*/\lambda_i) \ ,$$

for some positive polynomials M_1, M_2, M_3. One can now easily check that:

$$\lim_{n\to\infty} \mathbb{E}\,|\mathcal{U}_n| = 0 \ .$$

We then obtain the bounds:

$$\mathbb{E}\,|\,\mathcal{A}_{j,n}\,| \leq (\sqrt{\Delta s})^j \Big(\sum_{i=1}^{n} \lambda_i^{-2j}\Big)^{1/2} \mathbb{E}\,[\,C_{j,\cdot}^2\,]^{1/2}, \quad j = 3, 4, 5 \ ,$$

$$\mathbb{E}\,|\mathcal{A}_{6,n}\,| \leq (\Delta s)^3 \Big(\sum_{i=1}^{n} \lambda_i^{-6}\Big)\, \mathbb{E}\,|\,C_{6,\cdot}\,| \ .$$

Recall that $\lambda_i = i^{-\kappa}$. So, when $\Delta s = n^{-\gamma}$ with $\gamma \geq 2\kappa + 1/3$, then one can easily verify that $\limsup_n \mathbb{E}\,|\,R_n\,| < \infty$. So, from Lemma 1(i), $\liminf_{n\to\infty} \mathbb{E}\, a(x,y) > 0$.

When $\gamma \in (2\kappa, 2\kappa{+}1/3)$, we consider an m-th order expansion, for $(m{+}1)\gamma > 2$ and work as in Theorem 1, taking into consideration the scalings λ_i as above. We avoid further details.

- $\mathcal{A} = \mathcal{C}_n$

One can easily check that, on the transformed space $x \mapsto \mathcal{C}_n^{-1/2} x$ the original algorithm with target distribution (20) and proposal (21) coincides with the algorithm of the iid case given in section 5.1. So the result follows from Theorem 1, with $\beta = 1$.

□

Proof of Theorem 3:
The acceptance probability will now be:

$$a(x,y) = 1 \wedge e^{R_n - \Phi_n(y) + \Phi_n(x)} \ ,$$

for R_n as in the product case. Note now that:

$$\limsup_n \mathbb{E}_\pi \left| R_n - \Phi_n(y) + \Phi_n(x) \right| \le K_1 + K_2 \limsup_n \mathbb{E}_{\pi_0} |R_n| \ ,$$
$$\mathbb{E}_\pi \Big[1 \wedge e^{R_n - \Phi_n(y) + \Phi_n(x)} \Big] \le K \, \mathbb{E}_{\pi_0} [1 \wedge e^{R_n}] \ ,$$

for some constants $K, K_1, K_2 > 0$, where we have used the assumption of a uniform bound on Φ_n. Consider the case $\mathcal{A} = I$ with $\Delta s = n^{-\gamma}$. We have already showed in the proof for Theorem 2 above that if $\gamma \ge 2\kappa + 1/3$ then $\limsup_n \mathbb{E}_{\pi_0} |\, R_n \,| < \infty$. The first inequality above implies that also $\limsup_n \mathbb{E}_\pi |R_n - \Phi_n(y) + \Phi_n(x)\,| < \infty$, and Lemma 1(i) gives a lower bound for the average acceptance probability in stationarity. When $\gamma \in (2\kappa, 2\kappa + 1/3)$, we showed that $\mathbb{E}_{\pi_0} [\, 1 \wedge e^{R_n} \,] \to 0$, so also $\mathbb{E}_\pi [\, 1 \wedge e^{R_n - \Phi_n(y) + \Phi_n(x)} \,] \to 0$.

A similar argument gives the required result for $\mathcal{A} = \mathcal{C}_n$.

□

Proof of Theorem 4:

- $\theta = 0$.

The required results for $\theta = 0$ are special cases of Theorem 3.

- $\theta = \frac{1}{2}$, $\mathcal{A} = I$.

After carrying out some calculations, the acceptance probability can be written as $1 \wedge e^{T_n}$ where:

$$T_n = \Phi_n(x) - \Phi_n(y) + \frac{1}{2}(\theta - \frac{1}{2})\, \Delta s \sum_{i=1}^{n} \lambda_i^{-2} \big((y_i/\lambda_i)^2 - (x_i/\lambda_i)^2 \big) \ .$$

So, when $\theta = 1/2$, the average acceptance probability in stationarity is lower bounded even for constant $\Delta s = c$. A similar simplification of the acceptance probability expression arises also in the case $\mathcal{A} = \mathcal{C}_n$.

□

Department of Statistics, University of Warwick, Coventry CV4 7AL, UK
E-mail: a.beskos@warwick.ac.uk

Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK
E-mail: A.M.Stuart@warwick.ac.uk