Sampling Conditioned Diffusions

Martin Hairer Andrew Stuart Jochen Voss

2007-11-12

Abstract

For many practical problems it is useful to be able to sample conditioned diffusions on a computer (*e.g.* in filtering/smoothing to sample from the conditioned distribution of the unknown signal given the known observations). We present a recently developed, SPDE-based method to tackle this problem. The method is an infinite dimensional generalisation of the Langevin sampling technique.

1 Introduction

In many situations, understanding the behaviour of a stochastic system is greatly aided by understanding its behaviour conditioned on certain events. This allows, for example, to study rare events by conditioning on the event happening or to analyse the behaviour of a composite system when only some of its components can be observed. Since properties of conditional distributions are often difficult to obtain analytically, it is desirable to be able to study these distributions numerically. This allows to develop meaningful conjectures about the distribution in question or, in a more applied context, to derive quantitative information about it. In this text we present a general technique to generate samples from conditional distributions on infinite dimensional spaces. We give several examples to illustrate how this technique can be applied.

Sampling, i.e. finding a mechanism which produces random values distributed according to a prescribed target distribution, is generally a difficult problem. There exist many 'tricks' to sample from specific distributions, ranging from very specialised methods, like the Box-Müller method for generating one-dimensional standard Gaussian distributed values, to generic methods, like rejection sampling, which can be applied to whole classes of distributions. In situations where none of the direct methods apply in a useful way, Markov Chain Monte Carlo (MCMC) methods are commonly applied. These techniques work by constructing a Markov chain (or, more generally, a Markov process) which has the target distribution as its stationary distribution. Assuming that the process converges to stationarity fast enough, the states of the Markov chain at 'large' times can be used as approximate samples from the target distribution. While MCMC methods are only approximate methods, they can be used in many situations where no other methods are available. This is particularly true in high dimensional problems and thus it is natural to employ MCMC methods for infinite dimensional sampling problems. Indeed, the main tool described in this text is an MCMC method for distributions on infinite dimensional spaces.

The stochastic systems of interest here are diffusion processes described by stochastic differential equations. The trajectories of these processes can be considered to be random functions and thus the probability distributions we consider typically live on function spaces like $L^2\big(I, \mathbb{R}^d\big)$ or $C\big(I, \mathbb{R}^d\big)$ where $I \subseteq \mathbb{R}$ is some interval. Thus, in order to construct an MCMC method for these distributions, we have to find Markov processes which have prescribed distributions on these function spaces as their invariant measures. In the context of our framework these Markov processes are given as solutions of stochastic partial differential equations (SPDEs), where the interval I is the 'space' direction of the SPDE.

Throughout this text we give several concrete examples of conditioned diffusions and how to sample from them. A simple case is to condition the process on its value at a fixed time, so that the resulting paths are bridges. Sampling bridges could, for example, be interesting when studying transitions between meta-stable states of some physical system: while these transitions will eventually happen, the times between transitions might be so big that they 'never' occur during an unconditioned numerical simulation. By conditioning on a transition actually happening, one can numerically study the transition mechanism.

A second application presented here will be "smoothing", i.e. reconstructing a signal from a noisy observation. Since all information which is available about such a signal is contained in the conditional distribution of the signal given the observation, one can solve smoothing problems by understanding this conditioned distribution.

The text is structured as follows: we start by presenting some well-known sampling techniques in section 2, namely Metropolis sampling and the Langevin method. In section 3 we introduce an infinite dimensional generalisation of Langevin sampling. Section 4 explains how this technique can be used to study conditioned diffusions in general and section 5 considers the special case of smoothing problems. Finally, in section 6, we present how the infinite dimensional Langevin method can be combined with Metropolis sampling to obtain numerically efficient methods. The conclusion in section 7 contains some pointers to extensions of the method and open problems.

2 Sampling Techniques

Sampling is the process of constructing random values, distributed according to a prescribed target distribution. Since our aim is to derive a numerically useful method, we are specifically interested in constructions which can be implemented on a computer. Generating random values in a computer program is usually done in two steps: first one uses a pseudo-random number generator to generate 'random' values for some simple distribution (usually the uniform distribution on the unit interval) and then, in a second step, these values are transformed to obtain the desired target distribution. In this text we will only consider the second step, i.e. we will assume the availability of a source of uniform or Gaussian distributed random numbers and describe methods to transform given random values in order to obtain values with the correct distribution.

We give an overview of some established sampling techniques which we will use later in the text. Since our aim is to sample distributions on infinite dimensional spaces, we restrict the presentation to techniques which can be applied in this context.

2.1 The Metropolis-Hastings Algorithm

A commonly used sampling technique is based on the *Metropolis-Hastings algorithm*. The idea behind this method is to modify a given Markov chain, using a rejection mechanism, in order to obtain a Markov chain with a given stationary distribution. This new Markov chain can then be used as the basis of an MCMC algorithm.

Theorem 1 *Let P be the transition kernel of a Markov chain taking values in some measurable space $(\mathcal{X}, \mathcal{F}, \mu)$. Let μ be a probability measure on $\mathcal{X}$. Assume that $\mu(dy)P(y,dx)$ is absolutely continuous w.r.t. $\mu(dx)P(x,dy)$ on $\mathcal{X}\times\mathcal{X}$. Inductively construct a process $(X_n)_{n\in\mathbb{N}}$ as follows: for $n\in\mathbb{N}$ let $Y_n \sim P(X_{n-1},\,\cdot\,)$ and U_n be uniformly distributed on $[0,1]$, where Y_n, given X_{n-1}, is conditionally independent of $X_1,\dots,X_{n-2}$ and U_n is independent of everything else, and let*

$$X_n = \begin{cases} Y_n, & \text{if } U_n \leq \alpha(X_{n-1},Y_n), \text{ and} \\ X_{n-1} & \text{else,} \end{cases}$$

where α is the (truncated) Radon-Nikodym derivative

$$\alpha(x,y) = 1 \wedge \frac{\mu(dy)P(y,dx)}{\mu(dx)P(x,dy)}.$$

Then $(X_n)_{n\in\mathbb{N}}$ is a Markov chain with stationary distribution μ.

The value $\alpha(X_{n-1},Y_n)$ is called the *acceptance probability* at step n, the value Y_n is called a *proposal*.

This theorem allows to change the distribution of any Markov chain which visits a large enough part of the state space, by rejecting some of the steps, in order to obtain a given stationary distribution. Then, assuming the resulting Markov chain is ergodic, one can compute expectations w.r.t. the stationary distribution μ, by taking ergodic averages:

$$\mathbb{E}_\mu(f) = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f(X_n).$$

The usefulness of this method depends strongly on the magnitude of the acceptance rates: if $\alpha(X_{n-1},Y_n)$ is often very small, convergence of the ergodic average will be very slow. For practical use, the transition kernel P has to be chosen in a way such that the acceptance probabilities are reasonably large.

A special case of the Metropolis-Hastings algorithm is when the transition kernel P does not depend on X_{n-1}. This corresponds to the case when the proposals are generated from an i.i.d. sequence. Because the acceptance probability at step n depends on the value X_{n-1}, the resulting Markov chain is no longer i.i.d. This method is called the *independence sampler*.

The independence sampler can for example be used to sample bridges of diffusion processes: If the target distribution μ is absolutely continuous w.r.t. Brownian bridges, one can use independent Brownian bridges as proposals. The independence sampler then gives a Markov chain with the bridge-distribution μ as its stationary distribution. See [CPS04] for a discussion of this method.

2.2 Langevin Sampling

Another method to obtain samples from a distribution on $\mathbb{R}^d$ with a density w.r.t. Lebesgue measure, called *Langevin sampling*, is given in the next theorem.

Theorem 2 *Let $\varphi \in C^2\big(\mathbb{R}^d, \mathbb{R}\big)$ be a strictly positive probability density w.r.t. the Lebesgue measure λ. Then the SDE*

$$dX_t = \nabla \log \varphi(X_t)\, dt + \sqrt{2}\, dW_t,$$

where W is a standard Brownian motion, has $\varphi\, d\lambda$ as its stationary distribution.

The SDE in the theorem is called the *Langevin equation*. One observation which often turns out to be very useful in practice is the fact that, similar to the situation for the Metropolis-Hastings algorithm, the density φ needs to be known only up to a multiplicative constant: changing the constant does not change the resulting Langevin equation.

While this method is known to work well in high dimensions, at first it seems difficult to extend this technique to more general spaces, since the theorem uses a densities w.r.t. Lebesgue measure; the latter does not exist in infinite dimensions. But it transpires that there is a variant of the idea which can be generalised.

Theorem 3 *Let $L \in \mathbb{R}^{d\times d}$ be a symmetric matrix such that the SDE*

$$dZ_t = LZ_t\, dt + \sqrt{2}\, dW_t,$$

has a stationary distribution ν. Let $\varphi \in C^2\big(\mathbb{R}^d, \mathbb{R}\big)$ be a strictly positive probability density w.r.t. ν. Then the SDE

$$dX_t = \big(LX_t + \nabla \log \varphi(X_t)\big)\, dt + \sqrt{2}\, dW_t,$$

where W is standard Brownian motion, has $\varphi\, d\nu$ as its stationary distribution.

A generalisation of this theorem to infinite dimensional spaces, presented in the next section, forms the basis of our sampling framework. Later, in section 6, we will see how a discretised version of the Langevin equation can be used to generate proposals for the Metropolis-Hastings algorithm, thus combining the two methods presented in this section.

3 Langevin Equations on Path Space

In this section we introduce the infinite dimensional analogue of the Langevin equation from section 2.2. The abstract setting is as follows: the SDEs in theorem 3 are replaced by stochastic evolution equations taking values in a real Banach space E, continuously embedded into a real separable Hilbert space $\mathcal{H}$. In our applications the space $\mathcal{H}$ will mostly be the space $L^2\big([0,1], \mathbb{R}^d\big)$ and E will be some subspace of $C\big([0,1], \mathbb{R}^d\big)$.

3.1 Linear Equations

In this section we derive a Hilbert space valued, linear SDE to sample from Gaussian distributions on $\mathcal{H}$. The results of this section can all be stated and proved in $\mathcal{H}$ without reference to the embedded Banach space E. A more detailed analysis can be found in [HSVW05].

Recall that a random variable X taking values in a separable Hilbert space $\mathcal{H}$ is said to be *Gaussian* if the law of $\langle y, X\rangle$ is Gaussian for every $y \in \mathcal{H}$. It is called *centred* if $\mathbb{E}\langle y, X\rangle = 0$ for every $y \in \mathcal{H}$. Gaussian random variables are determined by their mean $m = \mathbb{E}X \in \mathcal{H}$ and their covariance operator $\mathcal{C}\colon \mathcal{H} \to \mathcal{H}$ defined by

$$\langle y, \mathcal{C}x\rangle = \mathbb{E}\big(\langle y, X-m\rangle\langle X-m, x\rangle\big).$$

For details see *e.g.* [Bog98]. We denote the Gaussian measure with mean m and covariance operator $\mathcal{C}$ by $\mathcal{N}(m, \mathcal{C})$.

We consider the $\mathcal{H}$-valued SDE

$$dz_t = \mathcal{L}z_t\,dt + \sqrt{2}\,dw_t \tag{1}$$

where w is a cylindrical Wiener process on $\mathcal{H}$ and $\mathcal{L} = -\mathcal{C}^{-1}$. A process z is a mild solution of (1), if it satisfies

$$z_t = \mathrm{e}^{\mathcal{L}t}z_0 + \sqrt{2}\int_0^t \mathrm{e}^{\mathcal{L}(t-s)}\,dw_s.$$

Since this equation is linear, solutions are Gaussian processes and its invariant measure is a Gaussian measure on $\mathcal{H}$:

Theorem 4 *Let $\mu = \mathcal{N}(0, \mathcal{C})$ be a centred Gaussian measure on a separable Hilbert space $\mathcal{H}$. Then the corresponding evolution equation* (1) *with $\mathcal{L} = -\mathcal{C}^{-1}$ has continuous $\mathcal{H}$-valued mild solutions. Furthermore, it has μ as the unique invariant measure and there exists a constant K such that for every initial condition $x_0 \in \mathcal{H}$ one has*

$$\big\|\mathcal{L}(z_t) - \mu\big\|_{\mathrm{TV}} \le K\,\big(1 + \|x_0\|_{\mathcal{H}}\big)\exp\big(-\|\mathcal{C}\|^{-1}_{\mathcal{H}\to\mathcal{H}}t\big),$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance between measures.

By the theorem, equation (1) can be used to sample from centred Gaussian measures and by considering the process $(z_t+m)_{t\ge0}$ we have a sampling equation for arbitrary Gaussian measures $\mathcal{N}(m, \mathcal{C})$ on $\mathcal{H}$. To implement this method one has to identify the operator $\mathcal{L}$. The following example shows how this can be done in the cases which are the focus of our interest here.

Example 1. Consider the $\mathbb{R}^d$-valued, linear SDE

$$dZ_u = AZ_u\,du + B\,dW_u, \quad Z_0 = z^- \tag{2}$$

on the time interval $[0,1]$, where $A, B \in \mathbb{R}^{d\times d}$ are matrices and $x^- \in \mathbb{R}^d$ is the starting point. The solution is a Gaussian process with mean $m(u) = \mathbb{E}(Z_u) = \mathrm{e}^{uA}x^-$ and covariance function

$$C(u,v) = \mathrm{Cov}(X_u, X_v) = \mathrm{e}^{uA}\Big(\int_0^{u\wedge v} \mathrm{e}^{-rA}BB^*\mathrm{e}^{-rA^*}\,dr\Big)\mathrm{e}^{vA^*}$$

(see *e.g.* [KS91, section 5.6] for reference). It is easy to check that the corresponding covariance operator $\mathcal{C}$ is given by

$$(\mathcal{C}x)(u) = \int_0^1 C(u,v)x(v)\,dv$$

for all $u \in [0,1]$, $x \in L^2([0,1],\mathbb{R}^d)$ and, assuming BB^* is invertible, the negative of its inverse $\mathcal{L} = -\mathcal{C}^{-1}$ is the restriction of the distributional differential operator

$$L = (\partial_u + A^*)(BB^*)^{-1}(\partial_u - A) \tag{3}$$

to the domain

$$\mathcal{D}(\mathcal{L}) = \bigl\{ f \in H^2([0,1],\mathbb{R}^d) \bigm| f(0) = 0, \partial_u f(1) = Af(1) \bigr\}.$$

Thus, the stationary distribution of

$$dz_t = \mathcal{L}(z_t - m)\,dt + \sqrt{2}\,dw_t \tag{4}$$

is $\mathcal{N}(m,\mathcal{C})$.

Since $\mathcal{L}$ is a differential operator, we can write (4) as an SPDE. Using the fact that $Lm = 0$ on $(0,1)$, this formally leads to the equation

$$\begin{aligned} \partial_t z(t,u) &= Lz(t,u) + \sqrt{2}\,\partial_t w(t,u) \qquad \forall (t,u) \in (0,\infty) \times (0,1) \\ z(t,0) &= z^-, \quad \partial_u z(t,1) = Az(t,1) \qquad \forall t \in (0,\infty) \end{aligned}$$

where $\partial_t w$ is space-time white noise. By theorem 4, the stationary distribution of this SPDE coincides with the distribution of the process Z.

3.2 Semilinear Equations

In this subsection we will derive the infinite dimensional analogue of theorem 3. Here, the process $(z_t)_{t\geq 0}$ from (1) will correspond the $(Z_t)_{t\geq 0}$ in theorem 3. The equation for $(X_t)_{t\geq 0}$ will be replaced by a semilinear equation of the form

$$dx_t = \mathcal{L}x_t\,dt + F(x_t)\,dt + \sqrt{2}\,dw_t, \tag{5}$$

where $\mathcal{L}$ is a linear operator on $\mathcal{H}$, the drift F maps E into E^*, w is a cylindrical Wiener process on $\mathcal{H}$, and the process x takes values in E. As in the previous subsection, we consider mild solutions of this equation.

For our application of sampling conditioned diffusions, presented in the next section, we will have a distribution-valued drift function F which is only defined on the Banach space of continuous functions. Thus we need the setting described above and cannot use the Hilbert space based theory as found *e.g.* in [DPZ92]. Proofs of the results presented here can be found in [HSV07].

We start the presentation by giving the assumptions which we will require for our results. There are two assumptions on the linear operator $\mathcal{L}$:

(A1) The operator $\mathcal{L}$ is a self-adjoint, strictly dissipative operator on $\mathcal{H}$ which generates an analytic semigroup $S(t)$. The semigroup $S(t)$ can be restricted to a C_0-semigroup of contraction operators on E.

(A2) Let $\mathcal{H}^\alpha$ be the domain of $(-\mathcal{L})^\alpha$, equipped with the inner product $\langle x, y\rangle_\alpha = \langle(-\mathcal{L})^\alpha x, (-\mathcal{L})^\alpha y\rangle$. Then there exists an $\alpha \in (0, 1/2)$ such that $\mathcal{H}^\alpha \subset E$ densely, $(-\mathcal{L})^{-2\alpha}$ is nuclear in $\mathcal{H}$, and the Gaussian measure $\mathcal{N}\big(0, (-\mathcal{L})^{-2\alpha}\big)$ is concentrated on E.

We write $\mathcal{H}^{-\alpha}$ for the dual of $\mathcal{H}^\alpha$ and identify $\mathcal{H}^*$ with $\mathcal{H}$ in the usual way to get the following chain of inclusions:

$$\mathcal{H}^{1/2} \hookrightarrow \mathcal{H}^\alpha \hookrightarrow E \hookrightarrow \mathcal{H} \hookrightarrow E^* \hookrightarrow \mathcal{H}^{-\alpha} \hookrightarrow \mathcal{H}^{-1/2}.$$

To formulate our conditions on the drift F we will also use the subdifferential of the norm $\|\cdot\|_E$, defined as

$$\partial\|x\|_E = \big\{x^* \in E^* \bigm| x^*(x) = \|x\|_E \text{ and } x^*(y) \le \|y\|_E \ \forall y \in E\big\}$$

for every $x \in E$. We require the following conditions.

(A3) The nonlinearity $F\colon E \to E^*$ is Fréchet differentiable with

$$\|F(x)\|_{E^*} \le C(1 + \|x\|_E)^N, \quad \text{and} \quad \|DF(x)\|_{E\to E^*} \le C(1 + \|x\|_E)^N.$$

for every $x \in E$.

(A4) There exists a sequence of Fréchet differentiable functions $F_n\colon E \to E$ such that

$$\lim_{n\to\infty} \|F_n(x) - F(x)\|_{-\alpha} = 0$$

for all $x \in E$. For every $C > 0$ there exists a $K > 0$ such that for all $x \in E$ with $\|x\|_E \le C$ and all $n \in \mathbb{N}$ we have $\|F_n(x)\|_{-\alpha} \le K$. Furthermore, there is a $\gamma > 0$ such that

$$\langle x^*, F_n(x + y)\rangle \le -\gamma\|x\|_E$$

holds for every $x^* \in \partial\|x\|_E$ and every $x, y \in E$ with $\|x\|_E \ge C(1+\|y\|_E)^N$.

Our results currently require another, quite technical condition, on the drift F which is given here as (A5). While this condition looks quite artificial, it is easy to verify that it holds for all applications discussed in this text.

(A5) For every $R > 0$, there exists a Fréchet differentiable function $F_R\colon E \to E^*$ such that

$$F_R(x) = \begin{cases} F(x), & \text{for } \|x\|_E \le R, \text{ and} \\ 0, & \text{for } \|x\|_E \ge 2R, \end{cases} \tag{6}$$

and such that there exist constants C and N with

$$\|F_R(x)\|_{E^*} + \|DF_R(x)\|_{E\to E^*} \le C(1 + R)^N,$$

for every $x \in E$.

Definition 5 *An E-valued and $(\mathcal{F}_t)$-adapted process x is called a* mild solution *of equation* (5)*, if almost surely*

$$x_t = S(t)x_0 + \int_0^t S(t - s)F\big(x_s\big)\, ds + z_t \qquad \forall t \ge 0$$

holds where z is the solution of the linear equation (1)*.*

The drift F in (5) takes only values in E^* while the operator $\mathcal{L}$ will have a smoothing effect. There is a balance between these two effects and it is not *a priori* clear in which space the resulting process takes its values. The following theorem asserts that our assumptions are strong enough so that the solution is continuous with values in E.

Theorem 6 *Let $\mathcal{L}$ and F satisfy assumptions (A1)–(A4). Then for every initial value $x_0 \in E$ the equation* (5) *has a global, E-valued, unique mild solution.*

From theorem 4 we know that the linear equation (1) has stationary distribution $\nu = \mathcal{N}(0, -\mathcal{L}^{-1})$. The following theorem, which is the infinite-dimensional analogue of theorem 3, shows that we can again get an equation to sample from $\varphi\, d\nu$ by adding $\nabla \log \varphi$ to the drift of the linear equation.

Theorem 7 *Let $U \colon E \to \mathbb{R}$ be bounded from above and Fréchet differentiable. Assume that $\mathcal{L}$ and $F = U'$ satisfy assumptions (A1)–(A5), let $\nu = \mathcal{N}(0, -\mathcal{L}^{-1})$. Then the probability measure μ given by*

$$d\mu(x) = c\, \mathrm{e}^{U(x)}\, d\nu(x),$$

where c is a normalisation constant, is the unique invariant measure for (5).

The following result helps to convert the preceding theorem into useful numerical methods: properties of the target distribution μ can be found by considering ergodic averages of the solution of the SDE (5).

Theorem 8 *Assume that (A1)–(A5) hold and let μ be the invariant measure for* (5)*. Then one has*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \varphi(x_t)\, dt = \int_E \varphi(x)\, \mu(dx), \quad \textit{almost surely}$$

for every initial condition x_0 in the support of μ and for every bounded measurable function $\varphi \colon E \to \mathbb{R}$.

While these theorems are formulated in a way that helps to identify the stationary distribution of a given stochastic evolution equation, we will use the equations in the reverse way: starting with a target distribution μ with a known density $\varphi = \mathrm{e}^U$ w.r.t. a Gaussian measure ν we will construct semilinear SDEs with invariant measure μ. From theorem 7 we know that a possible choice for the drift is $F = (\log \varphi)'$, in direct analogy with the finite dimensional result from Theorem 3. This procedure is illustrated in the following example.

Example 2. Consider the $\mathbb{R}^d$-valued SDE

$$dX_u = AX_u\, du + f(X_u)\, du + B\, dW_u, \quad X_0 = x^- \tag{7}$$

on the time interval $[0, 1]$, where $A, B \in \mathbb{R}^{d\times d}$ are matrices, $x^- \in \mathbb{R}^d$ is the starting point and W is a standard Brownian motion on $\mathbb{R}^d$. In this situation we can apply the following form of the Girsanov formula.

Lemma 9 *Let $\nu = \mathcal{L}(Z)$ be the distribution of the solution of the linear SDE* (2) *and $\mu = \mathcal{L}(X)$ be the distribution of the solution of* (7)*. Assume that* (7) *has a.s. no explosions until time 1 and that B is invertible. Then μ has a density φ w.r.t. ν on $C\big([0,1],\mathbb{R}^d\big)$ which is given by*

$$\varphi(X) = \exp\Big(\int_0^1 (BB^*)^{-1} f(X_u)\, dX_u - \int_0^1 \big\langle AX_u + \frac{1}{2} f(X_u), (BB^*)^{-1} f(X_u) \big\rangle \, du \Big).$$

If $f = -BB^*\nabla V$ for some potential $V\colon \mathbb{R}^d \to \mathbb{R}$, then φ can be written as

$$\varphi(X) = \exp\Big(V(X_0) - V(X_1) - \int_0^1 \big\langle AX_u + \frac{1}{2} f(X_u), (BB^*)^{-1} f(X_u) \big\rangle + \frac{1}{2} \operatorname{div} f(X_u)\, du \Big).$$

Proof. Since X (by assumption) and Z (since it solves a linear SDE) have no explosions, we can apply Girsanov's theorem [Elw82, Theorem 11A] to find the densities of $\mathcal{L}(X)$ and $\mathcal{L}(Z)$ w.r.t. the distribution of the Brownian motion $\mathcal{L}(BW)$. Taking the ratio of these two densities gives the first expression for φ. The second form of φ can be found by applying Ito's formula to $V(X)$ and substituting the result into the first part. ∎

In the following we will assume that f has the required gradient form so we can use the second form of φ from the lemma (without the stochastic integral). From example 1 we obtain a second-order differential operator $\mathcal{L}$ on $L^2\big([0,1],\mathbb{R}^d\big)$ such that

$$dz_t = \mathcal{L}(z_t - m)\, dt + \sqrt{2}\, dw_t,$$

where m is the mean of Z, has stationary distribution ν. From theorem 7 we see, assuming (A1)–(A5) are satisfied, that we can add the drift $F = (\log\varphi)'$ to this equation to obtain a $C\big([0,1],\mathbb{R}^d\big)$-valued SDE with stationary distribution μ. A simple calculation shows

$$F(x) = (BB^*)^{-1} f(x_1)\delta_1 - \nabla\Psi(x) \qquad \forall x \in C\big([0,1],\mathbb{R}^d\big)$$

where δ_1 is a Dirac mass at $u = 1$ and Ψ is given by

$$\Psi(\xi) = \big\langle A\xi + \frac{1}{2} f(\xi), (BB^*)^{-1} f(\xi) \big\rangle + \frac{1}{2} \operatorname{div} f(\xi) \qquad \forall \xi \in \mathbb{R}^d. \tag{8}$$

Under mild assumptions on A, B and f, the conditions for theorems 6, 7 and 8 are satisfied and the stationary distribution of

$$dx_t = \mathcal{L}(x_t - m)\, dt + F(x_t)\, dt + \sqrt{2}\, dw_t$$

coincides with the distribution of the process X. An explicit set of assumptions on A, B, and f for the result to hold can be found in [HSV07].

Again, we would like to write this equation as a stochastic partial differential equation. In order to do so, we should just add the drift F to the SPDE from example 1. One complication is the presence of the Dirac-term in F. Since, assuming smooth w for this argument, the source term $(BB^*)^{-1}f(x_1)\delta_1$ will lead to a jump of size $f(x_1)$ in the u-derivative of the solution, we can incorporate the Dirac term in the boundary condition by formally writing the SPDE as

$$\begin{aligned}\partial_t x(t,u) = Lx(t,u) - \nabla\Psi(x(t,u)) + \sqrt{2}\,\partial_t w(t,u) \qquad &\forall (t,u)\in(0,\infty)\times(0,1)\\ x(t,0)=x^-, \quad \partial_u x(t,1) = Ax(t,1) + f(x(t,1)) \qquad &\forall t\in(0,\infty).\end{aligned}$$

4 Conditioned Diffusions

In the previous sections we have seen how the Langevin sampling method can be generalised to infinite dimensional situations and how this can be used to construct SPDEs which sample from the distribution of a finite dimensional diffusion process. In this section we focus on our main interest of this text, namely on applying the presented techniques to sample from conditioned diffusion processes.

Consider the following $\mathbb{R}^d$-valued SDE on the time interval $[0,1]$:

$$dX_u = AX_u\,du + f(X_u)\,du + B\,dW_u, \quad X_0 = x^-. \tag{9}$$

As before, $A, B \in \mathbb{R}^{d\times d}$ are matrices, $x^- \in \mathbb{R}^d$ is the starting point, W is a standard Brownian motion on $\mathbb{R}^d$ and we assume that $f = -BB^*\nabla V$ for some potential $V\colon \mathbb{R}^d \to \mathbb{R}$ and that B is invertible. Our aim is to construct an SPDE which has the distribution of X, conditioned on some event C, as its stationary distribution.

Let Z be the solution of the linear SDE

$$dZ_u = AZ_u\,du + B\,dW_u, \quad Z_0 = x^- \tag{10}$$

and set $m(u) = \mathbb{E}(Z(u)|C)$ for all $u\in[0,1]$. In the cases we consider here, the event C is such that $\mathcal{L}(Z|C)$ is still Gaussian. The general idea is to perform a construction consisting of the following steps.

1. Use the results of section 3.1 to obtain an L^2-valued SDE which has the centred Gaussian measure $\mathcal{L}(Z-m|C)$ as its stationary distribution.
2. Use the Girsanov formula and results about conditional distributions to derive the density of the conditional distribution $\mathcal{L}(X|C)$ w.r.t. $\mathcal{L}(Z|C)$. Using substitution, this gives the density of the shifted distribution $\mathcal{L}(X-m|C)$ w.r.t. the centred measure $\mathcal{L}(Z-m|C)$.
3. Use the results of section 3.2 and the density from step 2 to derive a $C\big([0,1],\mathbb{R}^d\big)$-valued SDE with stationary distribution $\mathcal{L}(X-m|C)$. Shifting the process by m reverses the centring from step 2 and gives the required sampling equation. Optionally write the L^2-valued SDE as an SPDE.

Combining all these steps leads to an SPDE which samples from the conditional distribution $\mathcal{L}(X|C)$ in its stationary measure. The details of the above steps depend on the specific situation under consideration. We will study one special case in detail in the next section, where we develop a method for nonlinear filtering by using the Langevin method to sample from the distribution of some signal given the observations. In the remainder of this section we illustrate the technique in a simpler setting.

Example 3. We can use the technique described above to construct an SPDE which samples bridges from

$$dX_u = AX_u\,du + f(X_u)\,du + B\,dW_u, \quad X_0 = x^-, \quad X_1 = x^+, \tag{11}$$

that is, the stationary distribution of the SPDE coincides with the distribution of solutions of the SDE (9), conditioned on $X_1 = x^+$.

Step 1: We need to find an SPDE with stationary distribution $\mathcal{L}(Z|Z_1 = x^+)$. Mean and covariance of the conditioned process can be found by conditioning the random variable $\big(Z(u), Z(v), Z(1)\big)$ for $u \leq v \leq 1$ on the value of $Z(1)$. Since this is a finite dimensional Gaussian random variable, mean and covariance of the conditional distribution can be explicitly calculated. Let m and C be the mean and covariance function of $\mathcal{L}(Z)$. Then $\mathcal{L}(Z|Z_1 = x^+)$ is a Gaussian measure with mean

$$\tilde{m}(u) = m(u) + C(u,1)C(1,1)^{-1}\big(x^+ - m(1)\big)$$

and covariance operator $\tilde{\mathcal{C}}$ with $\tilde{\mathcal{C}}x = \int \tilde{C}(\,\cdot\,,v)x(v)\,dv$ where the covariance function is given by

$$\tilde{C}(u,v) = C(u,v) - C(u,1)C(1,1)^{-1}C(1,v).$$

A simple calculation shows that $\tilde{\mathcal{L}} = -\tilde{\mathcal{C}}^{-1}$ is again the differential operator L from (3), but this time on the domain

$$\mathcal{D}(\tilde{\mathcal{L}}) = \big\{ f \in H^2([0,1], \mathbb{R}^d) \bigm| f(0) = 0, f(1) = 0 \big\}.$$

Thus the stationary distribution of

$$dz_t = \tilde{\mathcal{L}} z_t\,dt + \sqrt{2}\,dw_t$$

is $\mathcal{L}(Z - \tilde{m}|Z_1 = x^+)$ by theorem 4.

Step 2: We have already seen in example 2 that the density of $\mathcal{L}(X)$ w.r.t. $\mathcal{L}(Z)$ is given by

$$\varphi(X) = \exp\Big(V(x^-) - V(X_1) - \int_0^1 \Psi(X_u)\,du \Big)$$

with the Ψ from equation (8). The following lemma shows that the density of $\mathcal{L}(X|X_1 = x^+)$ w.r.t. $\mathcal{L}(Z|Z_1 = x^+)$ coincides, up to a multiplicative constant, with φ.

Lemma 10 *Let P, Q be probability measures on $S \times T$ where $(S, \mathcal{A})$ and $(T, \mathcal{B})$ are measurable spaces and let $X\colon S \times T \to S$ and $Y\colon S \times T \to T$ be the canonical projections. Assume that P has a density φ w.r.t. Q and that the conditional distribution $Q_{X|Y=y}$ exists. Then the conditional distribution $P_{X|Y=y}$ exists and is given by*

$$\frac{dP_{X|Y=y}}{dQ_{X|Y=y}}(x) = \begin{cases} \frac{1}{c(y)}\varphi(x,y), & \text{if } c(y) > 0, \text{ and} \\ 1 & \text{else} \end{cases}$$

where $c(y) = \int_S \varphi(x,y)\,dQ_{X|Y=y}(x)$ for all $y \in T$.

Thus, the density of $\mathcal{L}(X - m|X_1 = x^+)$ w.r.t. $\mathcal{L}(Z - m|Z_1 = x^+)$ is

$$\tilde{\varphi}(X) = c \exp\Bigl(\int_0^1 \Psi(X_u + m_u)\, du\Bigr)$$

for some normalisation constant c where Ψ is given by (8).

Step 3: Assuming that the conditions for theorems 6, 7 and 8 are satisfied, the stationary distribution of

$$d\tilde{x} = \mathcal{L}\tilde{x}\, dt - \nabla\Psi(\tilde{x} + \tilde{m})\, dt + \sqrt{2}\, dw_t$$

is then $\mathcal{L}(X - \tilde{m}|X_1 = x^+)$. Thus the process $x = \tilde{x} + \tilde{m}$, solving

$$dx_t = \mathcal{L}(x_t - \tilde{m})\, dt - \nabla\Psi(x_t)\, dt + \sqrt{2}\, dw_t, \tag{12}$$

can be used to sample from the target distribution $\mathcal{L}(X|X_1 = x^+)$.

Finally, we can rewrite this evolution equation as an SPDE: Since the mean $\tilde{m}$ satisfies $\tilde{m}(0) = x^-$, $\tilde{m}(1) = x^+$ and $L\tilde{m} = 0$ on $(0, 1)$, we can formally write (12) in the form

$$\begin{gathered}\partial_t x(t,u) = Lx(t,u) - \nabla\Psi(x(t,u)) + \sqrt{2}\, \partial_t w(t,u), \qquad \forall (t,u) \in (0,\infty) \times (0,1) \\ x(t,0) = x^-, \quad x(t,1) = x^+ \qquad \forall t \in (0,\infty).\end{gathered}$$

Note that use of this formulation no longer requires knowledge of the conditioned mean $\tilde{m}$.

Figure 1 shows the result of a numerical simulation which implements the method derived in example 3 to sample bridges of the process (11). For the simulation we use the drift

$$f(x) = -\Bigl(\frac{(x-1)^2(x+1)^2}{1+x^2}\Bigr)' = x\Bigl(\frac{8}{(1+x^2)^2} - 2\Bigr), \tag{13}$$

$A = 0$, $B = I$ and the end-points $x^- = -1$ and $x^+ = +1$. To allow the process to transition a few times between the stable equilibrium points, we chose $u \in [0, 100]$. The upper panel illustrates how one can get an approximation to a typical sample path of (11): it displays $u \mapsto x(t,u)$ for a big value of t. Assuming that the sampling process is already close to equilibrium, this path should closely resemble a typical bridge path. The second panel illustrates how statistical properties of the bridges can be approximated by taking ergodic averages using theorem 8. The line in the centre of the shaded band shows

$$\bar{m}(u) = \frac{1}{T}\int_0^T x(t,u)\, du$$

as a function of u for a big value of T. By theorem 8 we have $\bar{m}(u) \approx \tilde{m}(u)$. The width of the band is given by

$$\bar{\sigma}(u) = \Bigl(\frac{1}{T}\int_0^T \bigl(x(t,u) - \bar{m}(u)\bigr)^2\, du\Bigr)^{1/2}.$$

Again by theorem 8, $\bar{\sigma}(u)$ is approximately equal to the standard deviation of the bridge at position u.

5 Nonlinear Smoothing

In this section we will give a more challenging application of the method developed in the previous sections: we will describe how nonlinear smoothing problems can be formulated as a problem of sampling conditioned diffusions and how it can be solved using Langevin sampling.

Let $d = m + n$ with $m, n \in \mathbb{N}$ and consider a d-dimensional diffusion process given by

$$dX_u = AX_u\, du + f(X_u)\, du + B\, dW_u, \quad X_0 = x^-$$

where B is invertible and $(BB^*)^{-1} f$ is a gradient. Assume that only the last n components of this process can be observed and that we want to gain as much knowledge as possible about the unobserved m components from one observation of the last n components. We write $X_u = (X_u^{(1)}, X_u^{(2)}) \in \mathbb{R}^m \times \mathbb{R}^n$ and call $X^{(1)}$ the "signal" and $X^{(2)}$ the "observation".

While the problem is formally very easy to solve, the solution is just the conditional distribution $\mathcal{L}(X^{(1)}|X^{(2)})$, the task of actually algorithmically computing this solution is quite challenging. There are two commonly used ways of solving this problem: the traditional method, employed for example in particle filters, is to use the Zakai equation to construct an approximation to the density of $\mathcal{L}(X_u^{(1)}|X_v^{(2)}, 0 \leq v \leq u)$. The solution we propose here is to construct an SPDE which samples from the distribution $\mathcal{L}(X^{(1)}|X^{(2)})$. Questions about this conditional distribution can then be answered by considering ergodic averages. It transpires that this way of solving the smoothing problem can be derived as a special case of the general technique of sampling from conditioned diffusions which we presented in section 4.

Commonly, finding $\mathcal{L}(X_u^{(1)}|X_v^{(2)}, 0 \leq v \leq u)$ is called "filtering" and finding $\mathcal{L}(X^{(1)}|X^{(2)})$ is called "smoothing". The standard methods, like the Kalman filter and particle filter based approaches, proceed by first solving the filtering problem and then, optionally, solving the smoothing problem in a second, backward sweep over the data. The method we propose here directly solves the smoothing problem and thus all observations must be present from the start of the computation.

5.1 Construction of the Smoothing SPDE

The construction of the SPDE to sample from the conditional distribution of $X^{(1)}$ given $X^{(2)}$ follows the steps outlined in section 4. We start the construction by considering the linear, $\mathbb{R}^{m+n}$-valued SDE

$$dZ_u = AZ_u\, du + B\, dW_u, \qquad Z_0 = x^- \tag{14}$$

which will give our reference measure as before. Since this SDE is linear, its solution is a Gaussian process, and thus the distribution $\mathcal{L}(Z^{(1)}|Z^{(2)})$ is also Gaussian. First we have to identify the mean and covariance of this distribution. The abstract mechanism we use here is given in the following lemma.

Lemma 11 *Let $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ be a separable Hilbert space with projectors $\Pi_i \colon \mathcal{H} \to \mathcal{H}_i$. Let $(Z^{(1)}, Z^{(2)})$ be an $\mathcal{H}$-valued Gaussian random variable with mean $m = (m_1, m_2)$ and positive definite covariance operator $\mathcal{C}$ and define*

$\mathcal{C}_{ij} = \Pi_i \mathcal{C} \Pi_j^*$. *Then the conditional distribution of $Z^{(1)}$ given $Z^{(2)}$ is Gaussian with mean*

$$m_{1|2} = m_1 + \mathcal{C}_{12}\mathcal{C}_{22}^{-1}\bigl(Z^{(2)} - m_2\bigr)$$

and covariance operator

$$\mathcal{C}_{1|2} = \mathcal{C}_{11} - \mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{21}.$$

If we define as above $\mathcal{L} = (-\mathcal{C})^{-1}$ and formally define $\mathcal{L}_{ij} = \Pi_i \mathcal{L} \Pi_j^*$, then a simple formal calculation shows that $m_{1|2}$ and $\mathcal{C}_{1|2}$ are expected to be given by

$$m_{1|2} = m_1 - \mathcal{L}_{11}^{-1}\mathcal{L}_{12}\bigl(Z^{(2)} - m_2\bigr), \qquad \mathcal{C}_{1|2} = -\mathcal{L}_{11}^{-1}. \tag{15}$$

In contrast to the lemma above, the relations (15) do not hold in general (consider for example the case $\mathcal{C}_{1|2} = 0$), but in our situation it can be shown that domains for the operators $\mathcal{L}_{ij}$ can be chosen so that all of the given expressions are defined and that the conditional mean and expectation really have the form given in (15). Details of this construction can be found in [HSVW05, lemma 4.6]. By theorem 4, the $L^2\bigl([0,1],\mathbb{R}^d\bigr)$-valued SDE

$$dz_t = \mathcal{L}_{11} z_t\, dt + \sqrt{2}\, dw_t$$

has $\mathcal{L}(Z^{(1)} - m_{1|2} | Z^{(2)})$ as its stationary distribution. We have already identified the differential operator $\mathcal{L}$ in section 3.1.

Now we can just follow the programme outlined in section 4: the version of Girsanov formula from lemma 9 gives the density φ of $\mathcal{L}(X)$ w.r.t. $\mathcal{L}(Z)$. From lemma 10 we know that the conditional density $\varphi_{1|2}$ of $X^{(1)}$ given $X^{(2)}$ differs from $x \mapsto \varphi(x, X^{(2)})$ only by a multiplicative constant which depends only on $X^{(2)}$. Thus we have $\nabla \log \varphi_{1|2} = \nabla_1 \log \varphi(\,\cdot\,, X^{(2)})$ where ∇ denotes the Fréchet derivative on $C\bigl([0,1],\mathbb{R}^d\bigr)$ and ∇_1 denotes the Fréchet derivative w.r.t. the first m components. By theorem 7 the equation

$$dx_t = \mathcal{L}_{11}(x_t - m_{1|2})\, dt + \nabla_1 \log \varphi(x_t, X^{(2)})\, dt + \sqrt{2} dw_t$$

has $\mathcal{L}(X^{(1)} | X^{(2)})$ as its stationary distribution and thus can be used as a Monte Carlo method to solve the smoothing problem.

Example 4. In the standard smoothing setup the signal $X^{(1)}$ evolves on its own without reference to the observation. The observation depends both on the signal and on additional noise. To fit this situation in the framework described above we consider the following case:

$$A = \begin{pmatrix} 0 & 0 \\ A_{21} & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix},$$

with $A_{21} \in \mathbb{R}^{n\times m}$, $B_{11} \in \mathbb{R}^{m\times m}$ and $B_{22} \in \mathbb{R}^{n\times n}$. Furthermore let $V(x,y) = V_1(x) + V_2(y)$ and $f = -BB^*\nabla V$. In this situation, equation (14) can be written as

$$\begin{aligned} dX^{(1)} &= f_1(X^{(1)})\, du + && B_{11}\, dW^{(1)} \\ dX^{(2)} &= f_2(X^{(2)})\, du + A_{21}X^{(1)}\, du + && B_{22}\, dW^{(2)} \end{aligned} \tag{16}$$

with $f_1 = -B_{11}B_{11}^*\nabla V_1$ and $f_2 = -B_{22}B_{22}^*\nabla V_2$.

For this choice of the matrices A and B the differential operator $\mathcal{L}$ is

$$\begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} = \begin{pmatrix} \partial_u & A_{21}^* \\ 0 & \partial_u \end{pmatrix} \begin{pmatrix} B_{11}B_{11}^* & 0 \\ 0 & B_{22}B_{22}^* \end{pmatrix}^{-1} \begin{pmatrix} \partial_u & 0 \\ -A_{21} & \partial_u \end{pmatrix}$$
$$= \begin{pmatrix} \partial_u(B_{11}B_{11}^*)^{-1}\partial_u - A_{21}^*(B_{22}B_{22}^*)^{-1}A_{21} & A_{21}^*(B_{22}B_{22}^*)^{-1}\partial_u \\ -\partial_u(B_{22}B_{22}^*)^{-1}A_{21} & \partial_u(B_{22}B_{22}^*)^{-1}\partial_u \end{pmatrix},$$

defined on some appropriate domain. A more detailed analysis, as found in [HSVW05, section 4], shows that $\mathcal{L}_{11}$ in (15) can be taken to be L_{11} on the domain

$$\mathcal{D}(\mathcal{L}_{11}) = \bigl\{ f \in H^2([0,1],\mathbb{R}^d) \bigm| f(0) = 0, \partial_u f(1) = 0 \bigr\}.$$

From (15) we find that $m_{1|2}$ is the solution of

$$\mathcal{L}_{11}(m_{1|2} - m_1) = -A_{21}^*(B_{22}B_{22}^*)^{-1}\bigl(\frac{dZ^{(2)}}{du} - m_2\bigr).$$

Here $\frac{dZ^{(2)}}{du}$ only exists as a distribution, but since $\mathcal{L}_{11}$ is a second order differential operator, the solution $m_{1|2}$ is a smooth function.

The density of $\mathcal{L}(X^{(1)}|X^{(2)})$ w.r.t. $\mathcal{L}(Z^{(1)}|Z^{(2)})$ can be simplified because of the simple structure of the matrices A and B: we get

$$\varphi(X^{(1)}|X^{(2)}) = c\exp\Bigl(- V_1(X_1^{(1)}) - \frac{1}{2}\int_0^1 |B_{11}^{-1}f_1(X_u^{(1)})|^2 + \operatorname{div} f_1(X_u^{(1)})\,du$$
$$- \int_0^1 \langle X_u^{(1)}, A_{21}^*(B_{22}B_{22}^*)^{-1} f_2(X_u^{(2)})\rangle\,du\Bigr)$$

for some normalisation constant c and the density of the target distribution $\mu = \mathcal{L}(X^{(1)} - m_{1|2}|X^{(2)})$ w.r.t. $\nu = \mathcal{L}(Z^{(1)} - m_{1|2}|Z^{(2)})$ is $\varphi(X - m_{1|2}|Y)$. Thus, for given $X^{(2)}$, the Fréchet derivative of $\log\varphi(X^{(1)}|X^{(2)})$ is

$$F(x) = \nabla_1 \log\varphi(x|X^{(2)}) = -\nabla V_1(x_1)\delta_1 - \nabla\Phi(x) - A_{21}^*(B_{22}B_{22}^*)^{-1}f_2(X^{(2)})$$

for all $x \in C\bigl([0,1],\mathbb{R}^m\bigr)$, where δ_1 is a Dirac mass at $u = 1$ and

$$\Phi(\xi) = \frac{1}{2}\bigl(|B_{11}^{-1}f_1(\xi)|^2 + \operatorname{div} f_1(\xi)\bigr) \qquad \forall \xi \in \mathbb{R}^m.$$

With F we have found the drift to be used in theorem 7: the equation

$$d\tilde{x}_t = \mathcal{L}_{11}\tilde{x}_t\,dt - \nabla\Phi(\tilde{x}_t + m_{1|2})\,dt - A_{21}^*(B_{22}B_{22}^*)^{-1}f_2(X^{(2)})\,dt$$
$$- \nabla V_1(\tilde{x}_t(1) + m_{1|2}(1))\delta_1\,dt + \sqrt{2}\,dw_t$$

is ergodic and has $\mathcal{L}(X^{(1)} - m_{1|2}|X^{(2)})$ as its stationary distribution. Defining $x_t = \tilde{x}_t + m_{1|2}$ for all $t \geq 0$ and formally writing the equation for x as an SPDE again, we find that the SPDE

$$\partial_t x(t,u) = (B_{11}B_{11}^*)^{-1}\partial_u^2 x(t,u) - \nabla\Phi\bigl(x(t,u)\bigr)$$
$$+ A_{21}^*(B_{22}B_{22}^*)^{-1}\Bigl(\frac{dX^{(2)}}{du}(u) - f_2(X^{(2)}(u)) - A_{21}x(t,u)\Bigr) + \sqrt{2}\,\partial_t w(t,u) \quad (17)$$

with boundary conditions

$$x(t,0) = 0, \quad \partial_u x(t,1) = f_1\big(x(t,1)\big)$$

for all $t \geq 0$ is the Langevin equation on $C\big([0,1],\mathbb{R}^m\big)$ to sample from the distribution $\mathcal{L}(X^{(1)}|X^{(2)})$. In the derivation above we did not check whether the conditions (A1), ..., (A5) which are required for our sampling method are satisfied. In general this depends on the specific choice of f, A and B. A (quite technical) set of conditions such that the theorems apply can be found in [HSV07].

A comparison between the sampling equation derived here and the equation derived in example 2 to sample from the unconditional distribution $\mathcal{L}(X^{(1)})$ reveals that the only difference caused by the conditioning is the presence of the term

$$A_{21}^*(B_{22}B_{22}^*)^{-1}\Big(\frac{dX^{(2)}}{du}(u) - f_2(X^{(2)}(u)) - A_{21}x(t,u)\Big).$$

The presence of this additional drift term moves the solution of the sampling SPDE towards paths $X^{(1)}$ which minimise the 'energy' of the noise required for the second equation in (16) to hold.

Figure 2 illustrates the resulting smoothing method for the system

$$\begin{aligned} dX_u^{(1)} &= f(X_u^{(1)})\,du + dW_u^{(1)}, \qquad X_0^{(1)} = -1 \\ dX_u^{(2)} &= X_u^{(1)}\,du + dW_u^{(2)}, \qquad X_0^{(2)} = 0 \end{aligned}$$

where f is the double-well drift from (13). The upper panel shows the 'true' signal $X^{(1)}$ (unknown to the algorithm), together with a reconstruction obtained by the smoothing method described above. The displayed band was obtained again as in example 3. Since the observation (not displayed) contains not only information about the signal, but also unknown additional noise, a perfect reconstruction is not possible. But the figure shows that the reconstruction captures the main features of the signal. Other statistical quantities of the conditional distribution of the signal, given the observation, like the number of transitions between the two equilibrium points can be computed similarly by taking ergodic averages. The lower panel shows a typical path of the conditional distribution for comparison with the 'true' signal in the upper panel.

5.2 Some Remarks about Smoothing

While the sampling technique developed in the previous section solves the same problem as traditional filters/smoothers do, it does so in a very different way: instead of trying to obtain the density of the conditional distribution, our method constructs samples from the conditional distribution which can be used as the basis of an MCMC algorithm.

Filtering and smoothing are sometimes used in high dimensional situations. For example applications in weather prediction, where filtering is used to incorporate the observed weather data into a model, now use values of d which are as big as 10^7 or 10^8. When d is big, a map from $\mathbb{R}^d$ to $\mathbb{R}$ like the density of $\mathcal{L}(X_u^{(1)}|X_v^{(2)}, 0 \leq v \leq u)$ is a complex object which is very hard to accurately represent in a computer. A standard way to deal with this problem, used in particle filter methods, is to approximate the conditional distribution as a sum of weighted Dirac masses. Another approach is to approximate the conditional distribution by a Gaussian, but in high-dimensional situations even storing the covariance matrix of this Gaussian has a non-negligible cost and sometimes even further approximations are necessary. In comparison, a map from $\mathbb{R}$ to $\mathbb{R}^d$, like the paths obtained by the smoothing method discussed here, is a much more manageable object. Thus the discussed method might be advantageous in high dimensions when smoothing is required and not just filtering.

Another observation to note is that the situation considered in example 4 is just one of many possible situations where a Langevin sampling based filtering method can be derived. Similar constructions are possible in many situations, for example it is easy to derive a sampling SPDE to sample from a diffusion conditioned on discrete noisy observations. See [AJSV07] for further examples.

More information about filtering and pointers into the literature can be found in [AHSV07].

6 Metropolis-Hastings Algorithm on Path Space

In the previous sections we showed how an infinite dimensional analogue of the Langevin equation can be used to sample from the distribution of conditioned diffusions. One of the main motivations behind this approach is that it directly translates into an implementable algorithm to solve these sampling problems. In this section we will discuss some issues which arise in this context. When implementing the method for practical use one has to numerically solve the sampling SPDE (5) and thus one has to discretise this equation in both 'space' u and time t. The two kinds of discretisation raise different issues and here we will mostly focus on the effects of discretising time.

There are two constraints which affect the choice of time step size Δt. Firstly, we are only interested in the stationary distribution of the sampling SPDE and thus, for our purposes, it doesn't matter if the numerical simulation accurately represents the trajectories of the solution but we require the invariant measure of the discretised equation to be close to the invariant measure of the exact equation. And, secondly, we will use the numerical solution to approximate ergodic averages as in theorem 8 and thus we need to simulate the solution over long time intervals. This leads to a trade-off in the choice of the step size Δt: small Δt requires many steps to cover big time intervals and thus makes the resulting method computationally expensive whereas big Δt leads to big discretisation error and makes the results less accurate.

One solution to this dilemma is the following idea, described in more detail in [BRSV07]: one can use a discretisation with a big step size Δt, but then use a rejection mechanism to compensate for the resulting discretisation error. More specifically, given an approximation $\hat{x}(t)$ to the exact solution x_t, a discretised version of the evolution equation gives an approximation to the solution at time $t + \Delta t$. But instead of directly using the computed value $\hat{y}(t + \Delta t)$ for the numerical solution, one can use it as the proposal in a Metropolis-Hastings algorithm and either accept or reject it as described in theorem 1.

A (partially implicit) Euler method for solving the equation

$$dx_t = \mathcal{L}x_t \, dt + F(x_t) \, dt + \sqrt{2} \, dw_t, \tag{18}$$

from section 3.2 can be formulated as

$$X_{n+1} = X_n + \mathcal{L}\big(\theta X_{n+1} + (1-\theta)X_n\big)\,\Delta t + F(X_n)\,\Delta t + \sqrt{2}\,\xi_n,$$

where the ξ_n have the same distribution as the increments of the cylindrical Wiener process w. The parameter $\theta \in [0,1]$ controls the implicitness of the method. We did not include implicitness in the evaluation of the non-linear part F of the drift, to make it easy to solve the iteration equation for X_{n+1}: One gets

$$\begin{aligned} X_{n+1} = \big(I - \Delta t\,\theta\mathcal{L}\big)^{-1}\big(I + \Delta t\,(1-\theta)\mathcal{L}\big)X_n \\ + \Delta t\,\big(I - \Delta t\,\theta\mathcal{L}\big)^{-1}F(X_n) + \sqrt{2}\big(I - \Delta t\,\theta\mathcal{L}\big)^{-1}\xi_n. \end{aligned} \tag{19}$$

It is not *a priori* clear what space this equation takes values in, since the cylindrical Wiener process w, and thus its increments, do not live in the Hilbert space $\mathcal{H}$. However, since $-\mathcal{L}^{-1}$ is trace class (it is the covariance of a Gaussian measure, see section 3.1), for $\theta > 0$ the operator $A = (I - \Delta t\,\theta\mathcal{L})^{-1}$ is Hilbert-Schmidt and thus the random increments $A\xi_n$ take values in $\mathcal{H}$. For this reason we restrict ourselves to the case $\theta > 0$ here.

When trying to use X_{n+1} as the proposal in a Metropolis algorithm, there is the following surprising dichotomy.

Theorem 12 *Let $\mathcal{H} = L^2\big([0,1],\mathbb{R}^d\big)$ and let $\mathcal{L}$ be a symmetric, negative definite operator on $\mathcal{H}$ as in section 3.2. Let μ be the invariant measure of* (18)*. Let $\theta > 0$ and define the transition kernel P on $\mathcal{H}$ by*

$$P(x,\,\cdot\,) = \mathcal{L}(X_{n+1}|X_n = x) \qquad \forall x \in \mathcal{H}$$

where X_{n+1} is defined by equation (19)*. Then there are two cases:*

a) If $\theta \neq 1/2$, then the distributions $\mu(dy)P(y,dx)$ and $\mu(dx)P(x,dy)$ on $\mathcal{H}\times\mathcal{H}$ are singular w.r.t. each other and thus the Metropolis algorithm cannot be used.

b) If $\theta = 1/2$, then the distributions $\mu(dy)P(y,dx)$ and $\mu(dx)P(x,dy)$ on $\mathcal{H}\times\mathcal{H}$ are equivalent and thus the Metropolis algorithm can be used.

Proof. For $X \in \mathcal{H}$ let $\langle X\rangle_u$ be the quadratic variation of X until time u. Then, by imitating the proof of [BRSV07, proposition 4.1], for $(X,Y) \sim \mu(dx)P(x,dy)$ we have

$$\langle Y\rangle_u = \frac{(1-\theta)^2}{\theta^2}\langle X\rangle_u \qquad \forall u \in [0,1]$$

almost surely. Since under μ the quadratic variation is a.s. constant, this shows that the measures in part a) are singular whenever $(1-\theta)^2/\theta^2 \neq 1$, i.e. when $\theta \neq 1/2$. A proof for part b) when $\mathcal{L}$ is a second derivative operator with Dirichlet boundary conditions can be found in [BRSV07, theorem 4.1]. An inspection of this proof reveals that it still holds in the more general situation considered here. ■

To implement the methods described in this text, the Langevin SPDE needs to be discretised in 'space' as well as in time. Some remarks about the required space discretisation can be found in [BRSV07]. For the space-discretised equation the dichotomy described in theorem 12 does not exist, every value of θ is possible there. But the effect from the theorem is still visible: for $\theta \neq 1/2$ one needs to decrease Δt when Δu gets smaller in order to retain large enough acceptance probabilities. For $\theta = 1/2$ one can decrease Δu without decreasing Δt. This effect, as it occurs for the smoothing problem from example 4, is displayed in figure 3.

7 Conclusion

In this text we have seen how an infinite dimensional generalisation of Langevin sampling can be used to generate samples from conditioned diffusions. We have seen that the presented method can be used as a common framework to solve very different kinds of sampling problems, such as generating bridge paths from SDEs and solving smoothing problems. The same framework can be applied to many more kinds of problems. For example one can apply the same kind of technique to processes indexed by a two-dimensional parameter instead of a single time variable. This might give rise to techniques which could be applied in image analysis, for example. It will be interesting to see what future applications will be developed based on this.

Throughout this text, we concentrated on sampling techniques which were direct generalisations of the finite dimensional result from theorem 3. But of course, since we are only interested in the stationary distribution, the sampling equation is not uniquely determined, many choices are possible. For example in the finite dimensional case the SDE

$$dX_t = LX_t\,dt + \nabla \log \varphi(X_t)\,dt + \sqrt{2}\,dW_t$$

and the "preconditioned" SDE

$$dX_t = GLX_t\,dt + G\nabla \log \varphi(X_t)\,dt + \sqrt{2G}\,dW_t,$$

where G is a symmetric, positive matrix, share the same invariant measure. This relation carries over to the infinite dimensional situation. By taking *e.g.* $G = -L^{-1}$ one obtains a new equation with very different properties: the cylindrical noise is now replaced by a significantly more regular noise, but the smoothing effect from the operator L is no longer present. This technique is discussed in [HSV07] and [BRSV07]. Other choices of sampling equations, including second order equations, are discussed in [AHSV07].

In the further development of the presented sampling techniques, several open problems remain. For example, in this text we always assumed that the densities we obtained from the Girsanov formula can be rewritten without resorting to a stochastic integral. This restricted the choice of drift functions for the underlying diffusion processes to functions which are a gradient plus a linear function. It transpires that this restriction is not easily lifted: the theorems presented here no longer apply and while it is easy to formally derive sampling equations, it is very difficult to even give sense to the resulting equations. A conjecture about the results in the non-gradient case can be found in [HSV07].

Other open problems include questions about efficient implementation of the method. This requires numerical solutions of the resulting SPDEs and a careful choice of step sizes for discretisation is required.

References

[AHSV07] A. Apte, M. Hairer, A. M. Stuart, and J. Voss. Sampling the posterior: An approach to non-gaussian data assimilation. *Physica D: Nonlinear Phenomena*, 230(1–2):50–64, 2007.

[AJSV07] A. Apte, C. K. R. T. Jones, A. M. Stuart, and J. Voss. Data assimilation: Mathematical and statistical perspectives. To appear in the International Journal for Numerical Methods in Fluids, 2007.

[Bog98] V. I. Bogachev. *Gaussian measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998.

[BRSV07] A. Beskos, G. Roberts, A. M. Stuart, and J. Voss. MCMC methods for diffusion bridges. Submitted, 2007.

[CPS04] S. Chib, M. Pitt, and N. Shephard. Likelihood based inference for diffusion driven models. Working paper, 2004.

[DPZ92] G. Da Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 44 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1992.

[Elw82] K. D. Elworthy. *Stochastic differential equations on manifolds*, volume 70 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1982.

[HSV07] M. Hairer, A. M. Stuart, and J. Voss. Analysis of SPDEs arising in path sampling, part II: The nonlinear case. *Annals of Applied Probability*, 17(5/6):1657–1706, 2007.

[HSVW05] M. Hairer, A. M. Stuart, J. Voss, and P. Wiberg. Analysis of SPDEs arising in path sampling, part I: The Gaussian case. *Communications in Mathematical Sciences*, 3(4):587–603, 2005.

[KS91] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, second edition, 1991.

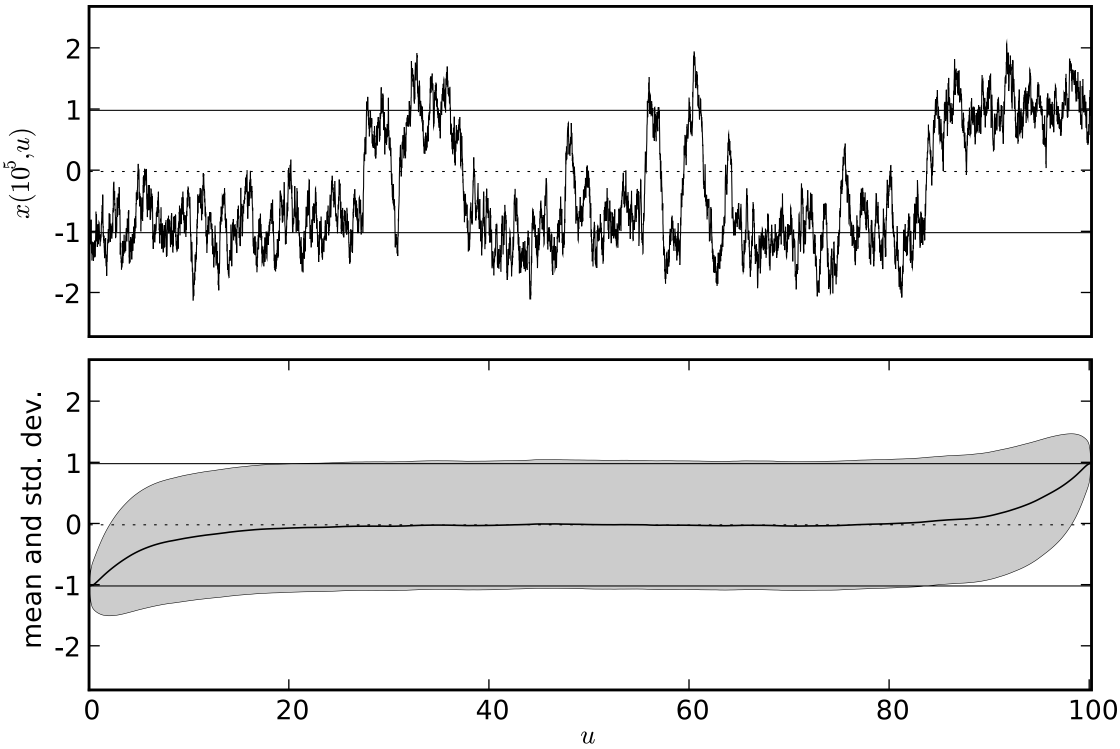

Figure 1: *Illustration of the bridge sampling method from example 3. The drift f in* (11) *is chosen to be the gradient of a double-well potential with stable equilibrium points at -1 and 1 and an unstable equilibrium point at 0 (see* (13)*), the process starts in $x^- = -1$ and is conditioned on ending up in $x^+ = +1$. The upper panel shows the value of the Langevin SPDE at time $t = 10^5$ as a function of u. This is an approximation to a typical bridge path. The lower panel shows a one-standard-deviation band around the mean of the solution as a function of u, obtained by taking averages over the interval $t \in [0, 10^5]$. This gives an approximation for the mean and standard deviation of the bridge process* (11).

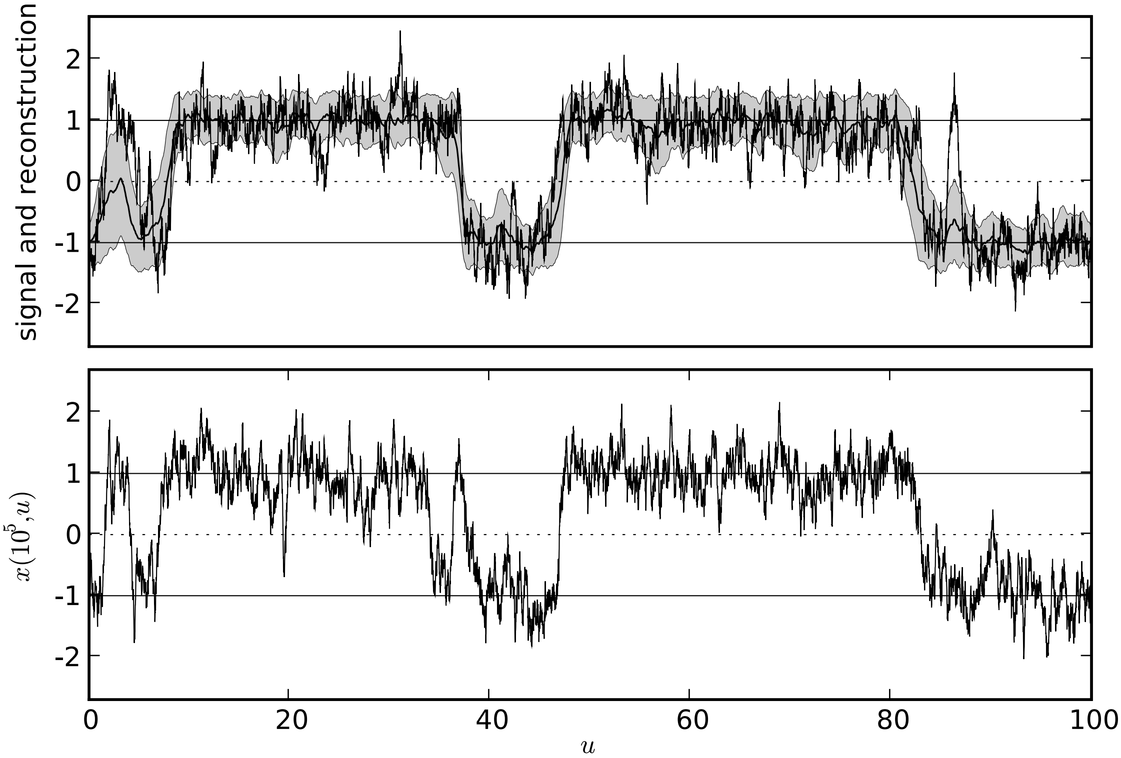

Figure 2: *Illustration of the smoothing method from example 4. The upper panel shows the true signal (unknown to the algorithm) together with a one-standard-deviation band around the mean of the sampling SPDE. This band can be seen as a reconstruction of the signal, but since the observation (not displayed) incorporates additional noise, a perfect reconstruction is not possible. The lower panel shows a typical path of the conditional distribution of the signal, given the observation, obtained by taking the value of the sampling SPDE at large t.*

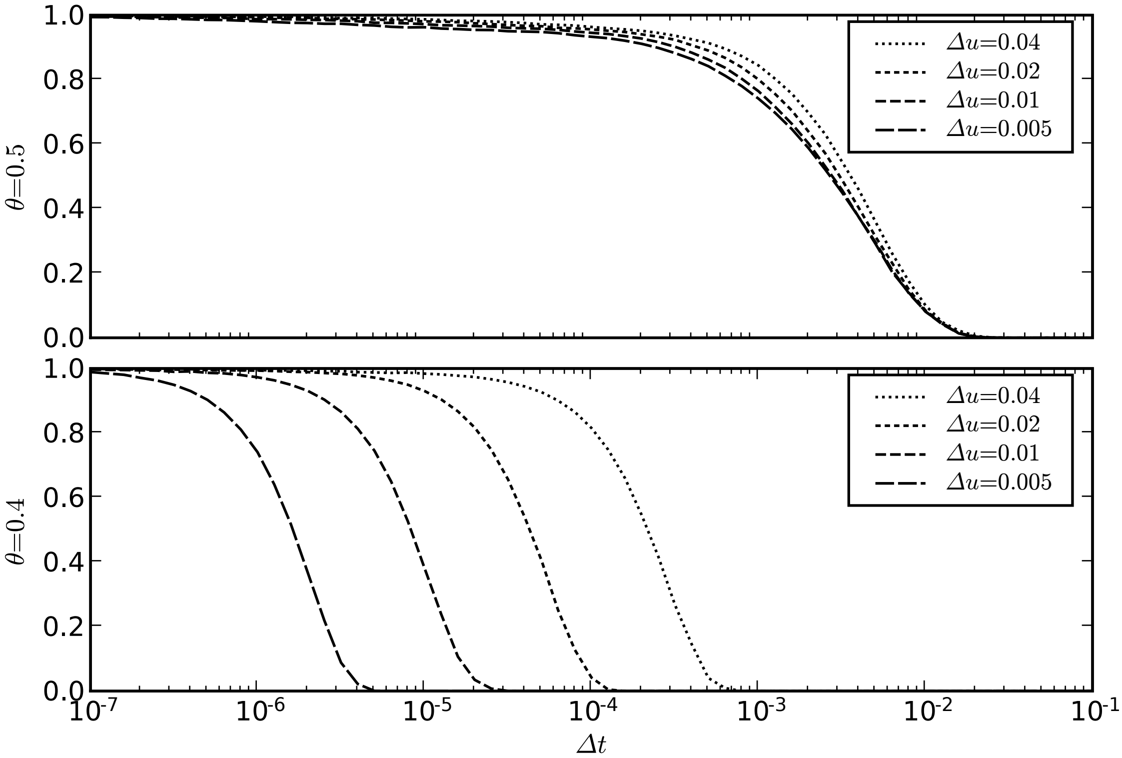

Figure 3: *This figure illustrates how the acceptance rates of the Metropolised algorithm for a discretised version of the smoothing problem from example 4 depend on the time discretisation step size* Δt*. The different curves correspond to different space discretisations* Δu*. The upper panel gives the average acceptance probabilities in equilibrium for* $\theta = 1/2$*. In this case the Metropolis-Hastings algorithm can also be applied to the infinite dimensional problem. The lower panel illustrates the case* $\theta = 0.4$ *which only makes sense for the discretised equation. One can see that the method degenerates as* $\Delta u \to 0$*.*