A Bayesian Approach to Data Assimilation

M. Hairer[1], A. M. Stuart[1], and J. Voss[1]

August 30, 2005

Abstract

Data assimilation is formulated in a Bayesian context. This leads to a sampling problem in the space of continuous time paths. By writing down a density in path space, and conditioning on observations, it is possible to define a range of Markov Chain Monte Carlo (MCMC) methods which sample from the desired distribution in path space, and thereby solve the data assimilation problem. The basic building-blocks for the MCMC methods that we concentrate on in this paper are stochastic partial differential equations which are ergodic and whose invariant measure gives the desired distribution in path space.

Two examples are given to show how data assimilation can be formulated in a Bayesian fashion. The first is weather prediction, and the second is Lagrangian data assimilation for oceanic velocity fields. Furthermore the relationship between the Bayesian approach outlined here and the commonly used Kalman filter-based techniques, prevalent in practice, is discussed. A simple paedogogical example is studied to illustrate the application of Bayesian sampling to data assimilation concretely. Finally a range of open mathematical and computational issues, arising from the Bayesian approach, are outlined.

1 Introduction

In this paper we describe a Bayesian approach to data assimilation, in which a continuous time *path* (time-dependent solution of a differential equation) is viewed as a random object whose distribution, conditional on observations, solves the data assimilation problem. In this context the key concept which needs elucidation is that of a probability density in the space of paths. Once this density is defined, and a conditional density is written down which incorporates observations, the complete Bayesian framework can be employed to sample in the space of continuous time paths.

We believe that this viewpoint may be useful for two primary reasons: firstly the Bayesian approach gives, in some sense, the correct theoretical answer to the data assimilation problem and other approaches which have been adopted,

[1]Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK

such as ensemble Kalman filtering, should be evaluated by their ability to approximate the posterior distribution in the Bayesian approach; secondly it will be necessary to use the Bayesian approach given here to sample any data assimilation problems which are bi- or multi-modal (for which Kalman-based methods will necessarily fail – see [12].)

The paper is organized as follows. In section 2 we formulate a number of variants of the data assimilation problem abstractly in the language of stochastic differential equations (SDEs). We give two concrete examples, arising in oceanic and atmospheric science, to motivate the abstract setting. Section 3 outlines the Bayesian approach to data assimilation that we will highlight in this paper, and introduces the central idea of probability density in path space. In subsection 3.2 we describe a generalization of the Langevin equation to path space, leading to nonlinear parabolic stochastic PDEs (SPDEs) which, when statistically stationary, sample from the distribution which solves the data assimilation problem; we also look at a second order Langevin equation, leading to a nonlinear damped stochastic wave equation. Subsection 3.3 describes another sampling strategy that might be used to sample path space, namely a Hybrid Monte Carlo technique. In section 4 we discuss MCMC methods in path space in general terms, discusing how Metropolis-Hastings ideas might be used to improve the Langevin and Hybrid methods from the previous section, and more generally to explore a wide range of sampling techniques. In section 5 we relate the Bayesian approach adopted here to other commonly used methods of data assimilation. Section 6 contains a paedogogical example of Lagrangian data assimilation, based on a Gaussian random field model of a velocity field, included to illustrate the Bayesian methodology. Section 7 concludes with a description of a number of open mathematical and computational questions arising from adopting the Bayesian viewpoint on data assimilation.

The SPDE based approach to sampling continuous time paths was introduced in [24] and is subsequently analyzed in [8] and [9], building on analysis in [26]. (For paths conditioned only on knowing the value of the path at two points in time – *bridges* – the SPDE based approach was simultaneously written down in [18].) The SPDE approach generalizes the Langevin equation to sampling in infinite dimensions. The Langevin approach to sampling in finite dimensions is outlined in the book [19] where it is shown how to use a discretization of the Langevin equation, in conjunction with a Metropolis-Hastings accept-reject criterion, to create a Markov chain Monte Carlo (MCMC) method. The infinite dimensional version of this MCMC method, arising when sampling the space of paths, is studied in [21]. Hybrid Monte Carlo methods, which are widely used in molecular dynamics, were generalized to sample in path space in [1], as were Langevin based methods; however that paper proceeded by discretizing the evolution equations to be sampled and then applying a finite dimensional sampling method. It is our view that it is conceptually and algorithmically preferable to formulate the sampling problem in infinite dimensions (the space of paths). It is conceptually important to know that the infnite dimensional problem makes sense mathematically. Once this infinite dimensional problem is defined, it is algorithmically important to find an efficient way of approx-

imating it by discretization. Discretizing first, so that the sampling problem is never written down in continuous time, and then sampling, may lead to a non-optimal approximation of the desired infinite dimensional problem; see the end of section 3.

The subject of Brownian motion and stochastic calculus is described in [10], whilst texts on SDEs include [5] and [16]. The subject of SPDEs is covered in the text [3].

2 The Framework

In this section we write down a precise mathematical framework into which a variety of data assimilation problems can be cast. We start with two motivational examples, and then express them precisely in the language of SDEs. We finish with some technical assumptions and notation that will be used in the remainder of the paper.

2.1 Motivation

When discretized in space, a typical model for numerical weather prediction is an ODE system with dimension of order 10^8. If modelling error and external forcing are modelled as temporal white noise then an equation of the form (2.1) below is obtained. In this context the state x represents the nodal values of the unknown quantities such as velocity, temperature, pressure and so forth. The observations which we wish to assimilate are then various projections of the state x, possibly different at different times, and may be viewed as subject to independent Gaussian white noises. We thus obtain observations y of the form (2.3) below.

A second motivational example is that of Lagrangian data assimilation in the ocean (see [12] for work in this direction). For expository purposes consider trying to make inference about a 2D velocity field governed by the noisy incompressible Navier-Stokes equations, by means of Lagrangian particle trajectories. We write the velocity field $v(z,t)$ as an (incompressible) trigonometric series

$$v(z,t) = \sum_{k \in \mathcal{K}} ik^{\perp} x^k(t) \exp\{ik \cdot z\}.$$

The vector x made up of the x^k then satisies an equation like (2.1) below. Now imagine a set of Lagrangian drifters, indexed by j, and with positions $y_j(t)$ governed by

$$\frac{dy_j}{dt} = v(y_j,t) + \sigma_j \frac{dW_j}{dt}.$$

From the representation of the velocity field it is clear that

$$v(z,t) = \chi(x(t),z)$$

for some function χ and hence that the collection of Lagrangian drifters satisfy an equation of the form (2.2) below. If data from the drifters (obtained by GPS

for example) is assumed to be essentially continuous in time then we may view (2.2) as giving the observational data y which is to be assimilated. (It is also possible to formulate Lagrangian data assimilation in the case where the drifters are observed only at discrete times).

2.2 Mathematical Setting

We now abstract these two examples of data assimilation. The *signal* that we wish to determine, and into which we wish to assimilate observational data, is assumed to satisfy the SDE

$$\frac{dx}{dt} = f(x) + \gamma \frac{dW_x}{dt}, \tag{2.1}$$

where f determines the systematic part of the evolution, and dW_x/dt is Gaussian white noise perturbing it. We assume that $x(0)$ is distributed with density ζ. Because W_x is a random function, the SDE defines a probability distribution on the space of continuous time paths. In data assimilation the ultimate objective is to probe this distribution, conditional on some form of *observation*.

If the observation is in continuous time then we denote it by $y(t)$ and assume that it too satisfies an SDE. This has the form

$$\frac{dy}{dt} = g(x, y) + \sigma \frac{dW_y}{dt}, \tag{2.2}$$

where g determines the systematic evolution of the observation, which depends on the signal x, and dW_y/dt is a standard Gaussian white noise perturbing it, independent of the white noise dW_x/dt.

If the observation is in discrete time then we assume that we observe $y = (y_1, \ldots, y_J)$ satisfying

$$y_j = h_j(x(t_j)) + \sigma_j \xi_j, \quad j = 1, \ldots, J. \tag{2.3}$$

Here h_j determines which function of the signal x is observed, the ξ_j are standard i.i.d. Gaussian random variables and the σ_j determine their covariances; both the h_j and σ_j are indexed by j because the nature of the observations may differ at different times. We assume that the ξ_j are independent of the white noise driving (2.1). The times $\{t_i\}$ are ordered and assumed to satisfy

$$0 < t_1 < t_2 < \cdots < t_J \leq T.$$

(Any observation at $t = 0$ is incorporated into ζ.)

2.3 Assumptions and Notation

In equation (2.1) we have $f \colon \mathbb{R}^d \to \mathbb{R}^d$, $\gamma \in \mathbb{R}^{d \times d}$ and W_x is standard $d-$dimensional Brownian motion. We assume that γ is invertible and we define $\Gamma = \gamma\gamma^T$. In equation (2.2) we have $g \colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^m$, $\sigma \in \mathbb{R}^{m \times m}$ and W_y is standard

$m-$dimensional Brownian motion, independent of W_x. We assume that σ is invertible and we define $\Sigma = \sigma\sigma^T$. In equation (2.3) we have $h_j : \mathbb{R}^d \to \mathbb{R}^m$ and $\sigma_j \in \mathbb{R}^{m\times m}$. The ξ_j are assumed independent of W_x. We also assume that σ_j is invertible and define $\Sigma_j = \sigma_j\sigma_j^T$.

For any positive-definite $n \times n$ covariance matrix A we define the inner-product on $\mathbb{R}^n$ given by

$$\langle a, b\rangle_A = a^T A^{-1} b$$

and the induced norm $\|\cdot\|_A^2 = \langle\cdot,\cdot\rangle_A$. This notation is used immediately in the next section, with A equal to Γ, Σ or Σ_j, in order to define a probability density in path space.

3 Path Space Sampling and (S)PDEs

3.1 Density in Path Space

In order to develop a Bayesian approach to path sampling for $\{x(t)\}_{t\in[0,T]}$, conditional on observations, we need to define a probability density in path space. To this end we define the following functionals:

$$\begin{aligned} I(x) &= \int_0^T [\frac{1}{2}\|\frac{dx}{dt} - f(x)\|_\Gamma^2 + \frac{1}{2}\nabla_x \cdot f(x)]dt, \\ J(x,y) &= \int_0^T [\frac{1}{2}\|\frac{dy}{dt} - g(x,y)\|_\Sigma^2 + \frac{1}{2}\nabla_y \cdot g(x,y)]dt, \\ J_D(x,y) &= \sum_{j=1}^J \frac{1}{2}\|y_j - h_j(x(t_j))\|_{\Sigma_j}^2. \end{aligned}$$

(Note that where the observation y appears in J it is a function, and where it appears in J_D it is a finite vector.)

Here $I(x)$ is known as the Onsager-Machlup functional for (2.1) and the unconditional density for paths x solving (2.1) may be thought of as being proportional to (see [7])

$$Q(x) := q(x)\zeta(x(0))$$

where

$$q(x) := \exp\{-I(x)\}$$

and ζ is the density of the initial condition for $x(t)$. Similarly $I(x) + J(x,y)$ is the Onsager-Machlup functional for (2.1) and (2.2), with unconditional density for paths x, y found by exponentiating the negative of this functional. Hence, by Bayes rule, the conditional density for paths x solving (2.1), given observation of y solving (2.2), may be thought of as being proportional to $Q(x) := q(x)\zeta(x(0))$ where

$$q(x) := \exp\{-I(x) - J(x,y)\}.$$

Similarly the conditional density for paths x solving (2.1), given observation of y solving (2.3), may be thought of as being proportional to $Q(x) := q(x)\zeta(x(0))$ where

$$q(x) := \exp\{-I(x) - J_D(x,y)\}.$$

Note that, in all cases, q maps the Sobolev space of functions with square integrable first derivative $H^1([0,T])$ into the positive reals $\mathbb{R}^+$. The observations y parameterize $q(x)$.

In the following two sections we will introduce continuous and discrete time Markov chains whose invariant measure samples from densities on path space such as the functionals $Q(x)$ defined above. This will lead to SPDEs in subsection 3.2 and a Markov chain constructed through a PDE with random initial data in subsection 3.3.

Defining these SPDEs will require calculation of the variational derivatives of $I(x), J(x,y)$ and $J_D(x,y)$ with respect to x. We list these derivatives here. To this end it is useful to define

$$\begin{aligned}\mathcal{F}(x) &= \frac{1}{2}\|f(x)\|_\Gamma^2 + \frac{1}{2}\nabla_x \cdot f(x)\\ \mathcal{H}(x) &= \Gamma^{-1}df(x) - df(x)^T\Gamma^{-1},\end{aligned}$$

where $df : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is the Jacobian of f. We also use $dg : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^{m\times d}$ to denote the Jacobian of g with respect to x and $dh_j : \mathbb{R}^d \to \mathbb{R}^{m\times d}$ to denote the Jacobian of h_j with respect to x. Then the required variational derivatives are:

$$\begin{aligned}\frac{\delta I}{\delta x} &= -\Gamma^{-1}\frac{d^2x}{dt^2} + \mathcal{H}(x)\frac{dx}{dt} + \nabla_x\mathcal{F}(x)\\ \frac{\delta J}{\delta x} &= -dg(x,y)^T\Sigma^{-1}[\frac{dy}{dt} - g(x,y)] + \frac{1}{2}\nabla_x\{\nabla_y \cdot g(x,y)\},\\ \frac{\delta J_D}{\delta x} &= -\sum_{j=1}^{J} dh(x(t_j))^T\Sigma_j^{-1}[y_j - h_j(x(t_j))]\delta(t-t_j).\end{aligned}$$

Notice that the last derivative is made up of point sources at the t_j. If $t_J = T$ then the jump induced by the delta function modifies the boundary condition at $t = T$ in the (S)PDEs that we write down in the next two sections – compare (3.3) and (3.4). Otherwise the delta jumps are in the interior of the domain for the (S)PDEs.

One important observation here is that the presence of the second term in $\mathcal{F}$, namely the divergence of f, is something which has caused some controversy in the physics literature. A least squares definition of the density, based on Gaussian white noise, misses the term. Even if it is included, its magnitude – the factor $\frac{1}{2}$ – has been queried [13]. The analysis in [9, 18] and numerical experiments [24] are unequivocal that its presence is necessary and that the pre-factor of $\frac{1}{2}$ is the correct choice.

It is also because of this second term in $\mathcal{F}$ that we have concerns about sampling methods which first discretize the SDE (2.1) and then apply standard

finite dimensional sampling techniques ([1]). Such an approach can lead to a very indirect and numerically unsatisfactory approximation of the second term (see [24]). For this reason we strongly recommend employing the methodology outlined in this paper: namely to formulate an infinite dimensional sampling method in path space, and then approximate it.

3.2 Langevin SPDEs Which Sample Path Space

The basic idea of **Langevin methods** is to construct a potential given by the gradient of the logarithm of the target density and to consider motion in this potential, driven by noise [19, 20]. In our case the desired target density is $Q(x) \propto q(x)\zeta(x(0))$. Ignoring the boundary conditions (i.e. ζ) for a moment, we obtain the following (proposal) SPDE for $x(t,s)$:

$$\frac{\partial x}{\partial s} = \frac{\delta \ln q(x)}{\delta x} + \sqrt{2}\frac{\partial W}{\partial s}, \quad (s,t) \in (0,\infty) \times (0,T). \tag{3.1}$$

Here s is an algorithmic time introduced to to facilitate sampling in the space of paths, parameterized by real time t, and $\frac{\partial W}{\partial s}$ is a white noise in (t,s). The variational derivative of $\ln q(x)$ gives a second order differential operator in t and so the PDE is of reaction-diffusion type, subject to noise. The details of the SPDE depend upon whether the sampling of x is unconditional, or subject to observations y; the latter may be in discrete or continuous time. The previous section implicitly calculates the derivative of $\ln q(x)$ in each of these three cases, through the variational derivatives of $I(x), J(x)$ and $J_D(x)$.

To find boundary conditions for the SPDE we argue in the standard fashion adopted in the calculus of variations. Notice that

$$\ln Q(x+\Delta x) = \ln Q(x) + (\frac{\delta}{\delta x}\ln Q(x), \Delta x) + \mathcal{O}(\|\Delta x\|^2)$$

where $(\cdot,\cdot)$ is the $L^2([0,T])$ inner-product and $\|\cdot\|$ an appropriate norm. Now

$$\begin{aligned}(\frac{\delta}{\delta x}\ln Q(x), \Delta x) &= (\frac{\delta}{\delta x}\ln q(x), \Delta x) \\ &+\langle \frac{dx(0)}{dt} - f(x(0)) + \Gamma\nabla_x \ln\zeta(x(0), \Delta x(0)\rangle_\Gamma - \langle \frac{dx(T)}{dt} - f(x(T)), \Delta x(T)\rangle_\Gamma.\end{aligned}$$

The first term on the right hand side gives the contribution to the derivative of $Q(x)$ appearing in the interior of the SPDE. Equating the second and third terms to zero, for all possible variations Δx, we obtain the following boundary conditions for the SPDE:

$$\frac{\partial x}{\partial t} - f(x) + \Gamma\nabla_x \ln\zeta(x) = 0, \quad t = 0, \tag{3.2}$$

$$\frac{\partial x}{\partial t} - f(x) = 0, \quad t = T. \tag{3.3}$$

The resulting SPDE (3.1)–(3.3) then has the desired equilibrium distribution.

When the observations are in discrete time and the last observsation coincides with the last point at which we wish to sample x (so that $t_J = T$) the delta function at $t = t_J$ in the variational derivative of $\ln q(x)$ does not appear in the interior $t \in (0, T)$ and instead modifies the second boundary condition to read

$$\frac{\partial x}{\partial t} - f(x) - \Gamma dh_J(x)^T \Sigma_J^{-1}[y_J - h_J(x)] = 0, \quad t = T. \tag{3.4}$$

Note that the case $h(x) = x$ and $y_J = x^+$ gives, in the limit where $\Sigma_J \to 0$, the Dirichlet boundary condition $x = x^+$ at $t = T$. Choosing ζ to be a Gaussian centred at x^-, and taking the limit of variance to zero, will also give a Dirichlet boundary condition $x = x^-$ at $t = 0$. These Dirichlet boundary conditions arise naturally in some applications of path sampling when bridges are studied [18, 24].

By generalizing the **second order Langevin method** we obtain the following (proposal) SPDE for $x(t, s)$:

$$\frac{\partial^2 x}{\partial s^2} + \iota \frac{\partial x}{\partial s} = \frac{\delta \ln q(x)}{\delta x} + \sqrt{2\iota} \frac{\partial W}{\partial s}, \quad (s, t) \in (0, \infty) \times (0, T), \tag{3.5}$$

with boundary conditions (3.2), (3.3). Here $\iota > 0$ is an arabitry positive parameter whose value may be optimized to improve sampling. This SPDE is a damped driven wave equation which yields the desired equilbrium distribution, when marginalized to x. The equilibrium distribution gives white noise for the momentum variable $\frac{\partial x}{\partial s}$ and this is hence natural initial data for the momentum variable.

It is also of interest to discuss **preconditioned Langevin equations**. Let $\mathcal{G}$ denote an arbitrary positive definite self-adjoint operator on the space of paths and consider the following SPDEs derived from (3.1) and (3.5) respectively:

$$\frac{\partial x}{\partial s} = \mathcal{G} \frac{\delta \ln q(x)}{\delta x} + \sqrt{2\mathcal{G}} \frac{\partial W}{\partial s}, \quad (s, t) \in (0, \infty) \times (0, T)$$

and

$$\mathcal{G}^{-1} \frac{\partial^2 x}{\partial s^2} + \iota \frac{\partial x}{\partial s} = \mathcal{G} \frac{\delta \ln q(x)}{\delta x} + \sqrt{2\iota\mathcal{G}} \frac{\partial W}{\partial s}, \quad (s, t) \in (0, \infty) \times (0, T).$$

The boundary conditions depend on $\mathcal{G}$ – some examples are given in [9] and [21]. Formally both these SPDEs preserve the desired invariant measure, for any choice of $\mathcal{G}$.

The simplest way to use any of the Langevin SPDEs described above to probe the desired (conditional) distribution on path space is as follows. Given some function $\varphi : C([0, T], \mathbb{R}^d) \to \mathbb{R}$ (such as the maximum value along the path, or the value of $|x(t)|^2$ at some time point $t = \tau$) solve one of the Langevin SPDEs numerically thereby generating a sequence $x_n(t) \approx x(t, n\Delta s)$ (in practice this will need to be discretized along the path in t as well as in s). For M sufficiently large, the collection $\{x_n(t)\}_{n \geq M}$ form approximate samples from the desired

distribution in path space. Hence, as $N \to \infty$, the average

$$\frac{1}{N}\sum_{n=0}^{N-1}\varphi(x_n(t)). \tag{3.6}$$

will converge, by ergodicity, to an approximation of the average of φ in the desired conditional distribution. (The fact that we obtain an approximation, rather than the exact stationary value, results from discretization of the SPDE in t, s – see [23, 25].) The role of $\mathcal{G}$ is to accelerate convergence as $N \to \infty$.

3.3 Hybrid Monte Carlo Methods Which Sample Path Space

By generalizing the **Hybrid Monte Carlo method** we obtain the following Markov chain $x_n(t)$. Setting $\iota = 0$ in the SPDE (3.5) gives the PDE

$$\frac{\partial^2 x}{\partial s^2} = \frac{\delta \ln q(x)}{\delta x}, \quad (s,t) \in (0,\infty) \times (0,T). \tag{3.7}$$

The boundary conditions are again (3.2), (3.3). This equation defines a solution operator

$$\mathcal{M} : (x_0, y_0) \to (x(\tau), \frac{\partial x}{\partial s}(\tau))$$

mapping initial conditions to the solution at time τ. With the notation

$$P_x : (x, y) \to x$$

we construct the Markov chain

$$x_{n+1} = P_x \mathcal{M}(x_n, \xi_n)$$

where the ξ_n are chosen to be i.i.d. spatial white noises. This yields the desired equilibrium distribution. The formula (3.6) can again be used to probe the desired conditional distribution. Each step of the Markov chain requires the solution of a nonlinear wave equation over an interval of length τ in s. Because numerical approximation of the wave equation (and hence $\mathcal{M}$) can lead to errors the formula (3.6) will in practice again only give an approximation of the true ergodic limit as $N \to \infty$. Pre-conditioning can also be used in the context of the Hybrid Monte Carlo method, replacing (3.7) by

$$\frac{\partial^2 x}{\partial s^2} = \mathcal{G}^2 \frac{\delta \ln q(x)}{\delta x}, \quad (s,t) \in (0,\infty) \times (0,T).$$

Again, $\mathcal{G}$ is used to accelerate convergence to stationarity.

4 Path Space Sampling and Other MCMC Methods

The Langevin SPDEs and the Hybrid Monte Carlo methods both give rise to Markov chains which, if solved exactly (which is impossible in almost all practical situations), sample exactly from the desired distribution in their stationary measure. They are all examples of MCMC methods. But there is no reason to restrict sampling methods to these particular MCMC methods and in this section we briefly outline directions which might be fruitfully pursued to get improved sampling.

4.1 Metropolis-Hastings

In practice the MCMC methods in the previous section require numerical approximation of an (S)PDE in (s,t). This will incur errors and hence the stationary distribution will only be sampled approximately. The errors arising from integration in s can be corrected by means of a Metropolis-Hastings accept-reject criterion (see [15, 19]). Furthermore, optimizing the choice of time-step in s can improve efficiency of the algorithm – we outline this below.

To apply the Metropolis-Hastings idea in path space, first discretize the path $\{x(t)\}$ giving rise to a vector x at the grid points. In the case of discrete observations this grid should ideally be chosen to include the observation times $\{t_j\}$. The signal $\{y(t)\}$ in the case of continuous time observations should also be discretized on the same grid.

The target density $Q(x)$ can then be approximated, using finite differences on the integrals, to define a finite-dimensional target density $Q_D(x)$. By discretizing the (S)PDEs in the previous section on the same grid of points in t, as well as discretizing in s, we obtain a *proposal distribution*. Moves according to this proposal distribution (discretized (S)PDE) are then accepted or rejected with the Metropolis-Hastings probability leading to a Markov chain with invariant density $Q_D(x)$. Thus the effect of error introduced by integrating in s is removed; and the error due to approximation in t is controlled by the approximation of $Q(x)$ by $Q_D(x)$.

If a small time-step is used in s then the proposal distribution is not far from the current position of the Markov chain. This is known as a local proposal and for these there is a well-developed theory of optimality for the resulting MCMC methods [20]. The variance of an estimator in a Markov chain is given by the integrated autocorrelation function. Roughly speaking, very small steps in s are undesirable because the correlation in the resulting Markov chain is high, leading to high variance in estimators, which is inefficient; on the other hand, large steps in s lead to frequent rejections, which is also inefficient, again because correlation between steps is high when rejections are included. Choosing the optimal scaling of the step in s, with respect to the number of discretization points used along the path $\{x(t)\}$, is an area of current research activity [21], building on the existing studies of MCMC methods in high dimensions [20]. In

the context of Metropolis-Hastings, good choices for the preconditioner $\mathcal{G}$ are ones which approximately equilibrate the convergence rates in different Fourier modes of the distribution. With this in mind, an interesting choice for $\mathcal{G}$ is a Green's operator for $-\frac{d^2}{dt^2}$ with homogeneous boundary conditions (see [9], [21], [1]).

If τ is small in the Hybrid Monte Carlo method then it too gives rise to a local proposal distribution. However, larger τ will lead to better decorrelation, and hence efficiency, if the rejection rate is not too large. Hence it is of interest to study optimal choices for τ, as a function of the number of discretization points, for this problem. The Hybrid Monte Carlo method was introduced and studied for discretizations of the path sampling problem in [1] where choices for the operator $\mathcal{G}$ were also discussed.

4.2 Global Moves

Langevin methods have a potential problem for the sampling of multi-modal distributions, namely that they can get stuck in a particular mode of the distribution for long times, because of the local (in state space) nature of the proposals. The Hybrid Monte Carlo method goes some way to ameliorating this issue as it allows free vibrations in the Hamiltonian given by the logarithm of the target density, and this is known to be beneficial in many finite dimensional sampling problems. However it is undoubtedly the case that sampling in path space will frequently be accelerated if problem specific global moves are incorporated into the proposal disctributions. This is an open area for investigation. In the context of bridges the paper [11] contains some ideas that might form the basis of global proposal moves; but these are not likely to extend to data assimilation directly.

5 Relationship to Other Approaches

The first observation to make in this context is that, in the language of signal-processing, the Bayesian method proposed here is performing *smoothing*, not *filtering*. This is because we sample from $x(s), s \in [0,T]$ given the entire set of observations on $[0,T]$, whereas filtering would sample from $x(s)$ given only observations in $[0,s]$. Filtering is appropriate in applications where the data is on-line. But for off-line data, smoothing is quite natural. Off-line situations arise when performing parameter estimation, for example, and also in Lagrangian data assimilation for oceanic velocity fields.

The standard method for performing filtering for nonlinar SDEs conditional on observations is via the *Zakai equation* and its generalizations. This is a linear partial differential equation for the probability density of the signal, conditional on observations. It is thus in the form of a Fokker-Planck equation, driven by noise (the observation). Informally it may be derived by employing the unconditional Fokker-Planck equation for (2.1) as a prior, and incorporating the observations via Bayes law; the Markovian structure of the signal and

observations allows filtering to be performed sequentially $0 \to T$. Smoothing can then be performed by means of a backward sweep, using a similar linear SPDE, incorporating data in reverse time $T \to 0$. See [22], Chapter 6, and the bibliographical Notes on Chapter 6, for further details and references.

A significant problem with use of the Zakai equation in the context of high dimensional problems ($d \gg 1$) is that the *independent* variables are in $\mathbb{R}^d$ and it is notoriously difficult to solve PDEs in high dimensions. Particle filters are a good tool for approximation of the Zakai equation in moderate dimension, but perform poorly in very high dimension.

Weather prediction leads to d of order 10^8 and solution of the Zakai equation by particle filters is impractical. In this context two simplifications are usually introduced. The first is to use the *extended Kalman filter* [2, 14] which proceeds by linearising the system and propagating a Gaussian model for the uncertainty; it is hence necessary to update the mean in $\mathbb{R}^d$ and the covariance matrix in $\mathbb{R}^{d\times d}$ sequentially, a task which is significantly easier than solving the Zakai equation. However even this approximation is impractical for large d and further approximations, primarily to effect dimension reduction on the covariance matrix, are performed; this leads to the *local ensemble Kalman filter* [17].

The approach we advocate in this paper is conceptually quite different from those based on the Zakai equation, and its Gaussian approximations. Instead of trying to sample from the probability distribution of the signal, at each point in time, by sequential means, we try to sample an entire path of the signal, from a distribution on path space. This leads to a nonlinear SPDE in one space dimension (t) and one time-like dimension indexing the sampling (s). The high dimension d enters as dimension of the *dependent* variable $x(t, s)$ which solves the SPDE; in contrast the Zakai equation has dimension d in the *independent* variable. The nonlinear SPDE proposed here hence has a considerable computational advantage over methods based on the Zakai equation, at least for problems which cannot be approximated in a Gaussian fashion.

6 Paedagogical Example

We discuss a simple example motivated by Lagrangian data assimilation. We use the example to illustrate the use of the (first order) Langevin SPDE for sampling conditional paths of (2.1). Consider a one dimensional velocity field of the form

$$v(y,t) = x_1(t) + x_2(t)\sin(y) + x_3(t)\cos(y)$$

where the $x_i(t)$ are Ornstein-Uhlenbeck processes solving

$$\frac{dx_i}{dt} = -\alpha x_i + \gamma \frac{dW_{x,i}}{dt}. \tag{6.1}$$

We assume that the particles are initially stationary and independent so that each $x_i(0)$ is distributed as $\mathcal{N}(0, \gamma^2/2\alpha)$, with density $\zeta(x) \propto \exp\{-\alpha x^2/\gamma^2\}$.

We study the question of making inference about the paths $x_i(t)$ from the observation of m drifters $\{y_i\}_{i=1}^m$ moving in the velocity field, and subject to random forcing idealized as white noise (e.g. molecular diffusion):

$$\frac{dy_i}{dt} = v(y_i, t) + \sigma \frac{dW_{y,i}}{dt}. \tag{6.2}$$

Here the $W_{x,i}$ and $W_{y,i}$ are independent standard Brownian motions. The initial conditions for the y_i are i.i.d. random variables drawn from the distribution $\mathcal{N}(0, 2\pi)$.

Writing $y = (y_1, \ldots, y_m)^T$ and $W_y = (W_{y,1}, \ldots, W_{y,m})^T$ we obtain

$$\frac{dy}{dt} = h(y)x + \sigma \frac{dW_y}{dt}, \tag{6.3}$$

where $h : \mathbb{R}^m \to \mathbb{R}^{m\times 3}$, $\sigma \in \mathbb{R}^+$.

In this case the Langevin SPDE (3.1)–(3.3) is hence

$$\begin{aligned}
\frac{\partial x}{\partial s} =& \frac{1}{\gamma^2}\frac{\partial^2 x}{\partial t^2} - \frac{\alpha^2}{\gamma^2}x + \frac{1}{\sigma^2}h(y)^T[\frac{dy}{dt} - h(y)x] \\
&- \frac{1}{2}\nabla_y \cdot h(y)^T + \sqrt{2}\frac{\partial W}{\partial s}, \quad (s,t) \in (0,\infty)\times(0,T) \\
\frac{\partial x}{\partial t} =& + \alpha x, \quad (s,t) \in (0,\infty)\times\{0\} \\
\frac{\partial x}{\partial t} =& - \alpha x, \quad (s,t) \in (0,\infty)\times\{T\} \\
x =& x_0, \ (s,t) \in 0 \times [0,T].
\end{aligned}$$

Because the SPDE is linear, the mean $\bar{x}$ in the stationary measure is found by removing the derivative in s and the noise to obtain

$$\begin{aligned}
\frac{1}{\gamma^2}\frac{d^2\bar{x}}{dt^2} - \frac{\alpha^2}{\gamma^2}\bar{x} - \frac{1}{\sigma^2}h(y)^T h(y)\bar{x} = -\frac{1}{\sigma^2}h(y)^T\frac{dy}{dt} + \frac{1}{2}\nabla_y \cdot h(y)^T, \quad t \in (0,T), \\
\frac{d\bar{x}}{dt} = +\alpha\bar{x}, \quad t = 0, \\
\frac{d\bar{x}}{dt} = -\alpha\bar{x}, \quad t = T.
\end{aligned}$$

Note that if $\sigma \ll \min(\gamma, 1)$ then, formally, the equation for the mean is dominated by the *normal equations*

$$h(y)^T[\frac{dy}{dt} - h(y)\bar{x}] \approx 0$$

which arise from trying to solve the over-determined equation (6.3) for x, when the noise is ignored. But when noise is present, however small, $\frac{dy}{dt}$ exists only as a distribution (it has the regularity of white noise) and so the second order differential operator in x, which incorporates prior information on x, is required to make sense of the mean.

Our numerical experiments are conducted as follows. We set $\alpha = \gamma = \sigma = 1$ and generated a single path for each x_i, $i = 1, 2, 3$ solving (6.1) on the interval $t \in [0, 10]$, using stationary initial conditions as described above. We also generated the trajectories of 500 drifters y_i moving according to (6.2), with initial conditions drawn from a Gaussian distrbution as described above. We then chose m drfiter paths, with $m = 5, 50$ and 500 respectively, and solved the Langevin SPDE to sample from the distribution of the x_i. We integrated over 100 algorithmic time units in s and approximated the mean of the x_i, together with one standard deviation, using (3.6). We also calculated the mean directly by solving the boundary value problem for $\bar{x}$. We emphasize that the signals x_i are not available to the Langevin SPDE or the boundary value problem: only information about the drifters y_i is used to reconstruct the x_i. The signals are shown in the following figures so that the reconstruction of the signal may be judged.

The results are shown in Figures 1, 2 and 3, corresponding to $m = 5, 50$ and 500 respectively. In each figure we consider x_1 in the top panel, x_2 in the middle and x_3 at the bottom. The actual signal x_i is the non-smooth curve whilst the mean of the desired conditional distribution, found by solving the equation for $\bar{x}$, is the smooth curve. The shaded bands show an estimate of one standard deviation about the mean, with both mean and standard deviation estimated by time averaging solution of the Langevin SPDE in s.

The figures illustrate two facts, one a property of the path sampling procedure we propose in this paper, the second a property of the desired conditional distribution for this data assimilation problem. The first fact is this: because the true mean $\bar{x}$ lies in the middle of the shaded band, it is clear that the estimate of the mean, calculated through time-averaging, is accurate at $s = 100$. The second fact is this: as m is increased our ability to recover the actual signal increases; this is manifest in the fact that the mean gets closer to the signal, and the standard deviation bounds get tighter.

To give some insight into how long the Langevin SPDE has to be integrated to obtain accurate time averages, we generated data analogous to that in Figure 1, but only integrated to time $s = 10$. The results are shown in Figure 4. The fact that $\bar{x}$ no longer lies in the middle of the shaded bands, at least for some parts of the paths, indicates that the time average of the path has not converged to the mean value in the stationary distribution.

7 Challenges

The Bayesian framework for data assimilation outlined here presents a number of significant scientific challenges. We outline some of these here, breaking the challenges down into three categories: applications, mathematical and computational.

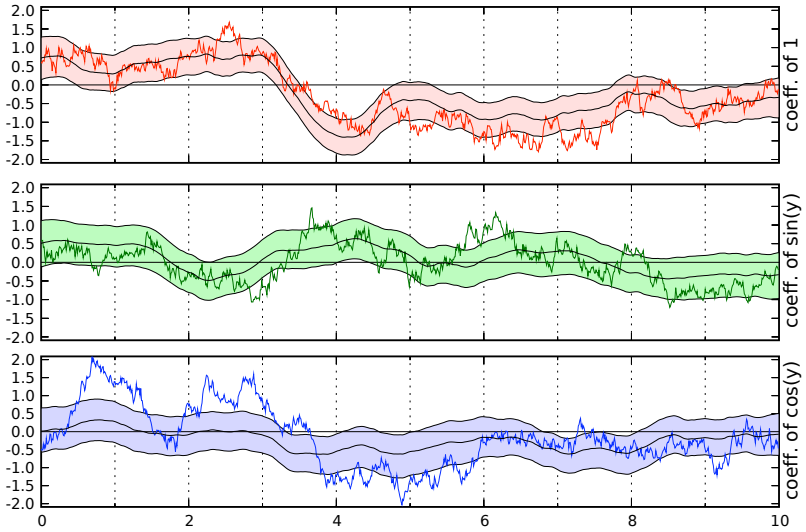

Figure 1: Reconstruction of the x_i solving (6.1), together with one standard deviation bounds, on $s \in [0, 100]$; 5 drifters are used.

7.1 Applications

- In the context of short term weather prediction, the Gaussian Kalman filter approximation appears quite effective; it would be interesting to quantify this by comparing it with the Bayesian approach described here.
- In the context of Lagrangian data assimilation for oceans, it would be of interest to use the methodology proposed here to study the multi-modal problems which often arise quite naturally, and for which the extended Kalman filter diverges.
- For both weather prediction and ocean modelling it would be of interest to incorporate the methodology proposed here for the purposes of parameter estimation. In this context the paths of (2.1) are treated as missing data which are sampled to enable estimation of parameters appearing in (2.1) itself. A Gibbs sampler ([19]) could be used to alternate between the missing data and the parameters.
- There are many other potential applications of this methodology in chemistry, physics, electrical engineering and econometrics, for example.

7.2 Mathematical

- The SPDEs which arise as the formal infinite dimensional Langevin equations, and the related PDE which arises in the hybrid Monte Carlo method,

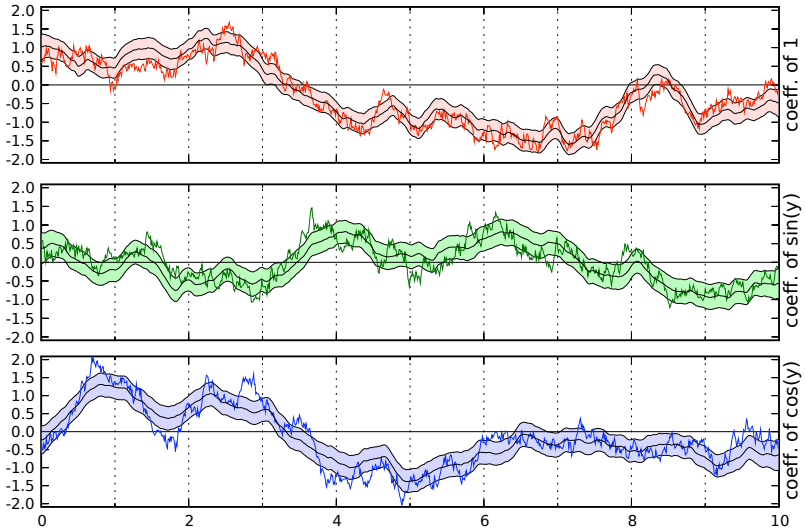

Figure 2: Reconstruction of the x_i solving (6.1), together with one standard deviation bounds, on $s \in [0, 100]$; 50 drifters are used.

all lead to significant problems in analysis concerned with the existence, uniqueness, ergodicity and rate of convergence to stationarity. Some of these issues have been resolved for particular forms of nonlinearity in (2.1) and (2.2) (see [8], [9]) primarily for vector fields f, and g in the case of continuous time observations, which are combinations of gradients and linear vector fields.

- For non-gradient vector fields the presence of the term $\mathcal{H}(x)\frac{\partial x}{\partial t}$ causes particlar problems in the development of a theory for the SPDE as, when the solution operator for the linear part of the Langevin SPDE is applied to it, a definition of stochastic integral is required. Numerical evidence as well as the derivation of $I(x)$ by means of the Girsanov formula, suggests that this should be a Stratonovich-type centred definition, but the mathematical analysis remains to be developed.

- In some applications the underlying path to be sampled arises from an SPDE itself: i.e. equation (2.1) is itself an SPDE; it would be of interest to derive the relevant Langevin SPDE here, in which the variable t would appear as a spatial variable, in addition to the spatial derivatives already appearing in (2.1).

- We have assumed for simplicity that white noise affects all components of the signal and observation equations; relaxing this assumption is natural in some applications, and it would be of interest to find the relevant SPDEs for sampling in this case.

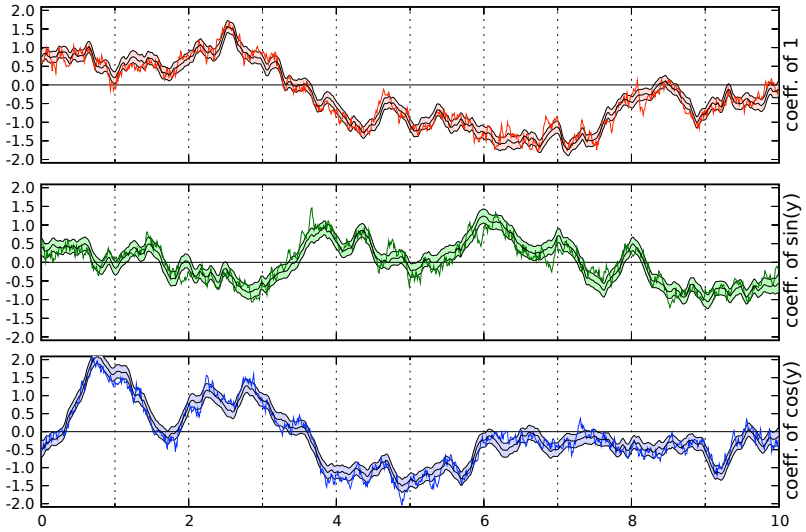

Figure 3: Reconstruction of the x_i solving (6.1), together with one standard deviation bounds, on $s \in [0, 100]$; 500 drifters are used.

7.3 Computational

- Within the context of Langevin algorithms it would be of interest to study choices of the pre-conditioner $\mathcal{G}$, and discretization method for the SPDE, which lead to efficient algorithms; efficiency in this context should be measured through the integrated auto-correlation function which quantifies the fluctuations in estimates of the form (3.6), for expectations of $\varphi(x(\cdot))$ with respect to the desired conditional measure [20].

- Similar considerations apply to Hybrid Monte Carlo methods, and the choice of pre-conditioner.

- It is also of interest to compare first order and second order Langevin based methods with one another and with the Hybrid Monte Carlo method, once good pre-conditioners have been found. See [1] for a step in this direction.

- The use of other MCMC methods to sample the desired probability measures on path space should also be explored. It is common practical experience that, whilst Langevin type methods are provably efficient within the context of methods using local (in state space) proposals [20], greater speed-ups can often be obtained by incorporating additional global moves, based on problem specific knowledge.

- The issue of how to discretize the SPDE is also non-trivial. In particular for non-gradient vector fields in (2.1), (2.2), the term $\mathcal{H}(x)\frac{\partial x}{\partial t}$ needs to be

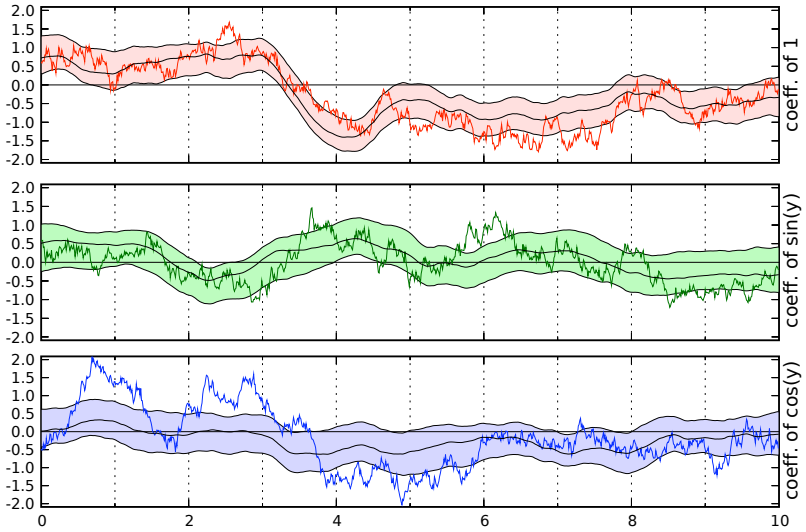

Figure 4: Reconstruction of the x_i solving (6.1), together with one standard deviation bounds, on $s \in [0, 10]$; 5 drifters are used. Note that the estimate of the mean (the middle of the shaded bands) is not always close to the actual mean (the smooth curve). This should be contrasted with Figure 1 which is on a longer interval in algorithmic time s.

discretized carefully (numerical evidence suggest that centred differencing is necessary) essentially for the same reasons that the SPDE theory is hard to develop in this case.

- If the dimension d is high then, since the number of dependent variables in the (S)PDEs proposed here will scale like d, techniques are required to reduce the dimensionality for sampling; multiscale methods are likely to be useful in this context [6]. Some interesting work in this direction, using relative entropy, may be found in [4].

Acknowledgements The authors are grateful to Greg Eyink, Chris Jones and Arthur Krener for helpful discussions.

References

[1] F. Alexander, G. Eyink and J. Restrepo, *Accelerated Monte-Carlo for optimal estimation of time series*, submitted.

[2] D.E. Caitlin *Estimation, Control and the Discrete Kalman Filter.* Springer, New York, 1989.

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

[4] G. Eyink. *In preparation*, 2005.

[5] C.W. Gardiner *Handbook of stochastic methods*. Springer, Berlin, 1985.

[6] D. Givon, R. Kupferman and A.M. Stuart. *Extracting macroscopic dynamics: model problems and algorithms*. Nonlinearity **17**(2004) R55–R127.

[7] R. Graham *Path integral formulation of general diffusion processes*. Z. Physik B**26**(1977), 281–290.

[8] M. Hairer, A. M. Stuart, J. Voss, and P. Wiberg. *Analysis of SPDEs arising in path sampling, part I: The Gaussian case.* In preparation.

[9] M. Hairer, A. M. Stuart and J. Voss. *Analysis of SPDEs arising in path sampling, part II: The nonlinear case.* In preparation.

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

[11] A. Krener. *Reciprocal diffusions in flat space*. Prob. Theor. Relat. Fields **107**(1997), 243–281.

[12] L. Kuznetsov, K. Ide and C.K.R.T. Jones, *A method for assimilation of Lagrangian data*. Monthly Weather Review, **131**(2003), 2247–2260.

[13] F. Langouche, D. Roekaerts and E. Tirapegui. *Functional integral methods for stochastic fields*. Physica **95A**(1979), 252–274.

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

[15] J. Liu *Monte Carlo Strategies in Scientific Computing*, Springer, New York, 2001.

[16] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin, fifth edition, 1998.

[17] E. Ott, B.R. Hunt, I Szunyogh A.V. Zimin, E.J. Kostelich, M. Corazza, E. Kalnay, D.J. Patil, J.A. Yorke. *A local ensemble Kalman filter for atmospheric data assimilation*. Tellus A**56**(2004), 415–428.

[18] M. Reznikoff and E. Vanden-Eijnden. *Invariant measures of stochastic PDEs*. C.R. Acad. Sci. Paris **340**(2005), 305–308.

[19] C. P. Robert and G. Casella. *Monte Carlo statistical methods.* Springer, New York, 1999.

[20] G. Roberts and J.S. Rosenthal, *Optimal scaling for various Metropolis-Hastings algorithms*, Statistical Science, 16, 4, 351-367, 2001.

[21] G. O. Roberts, A. M. Stuart, and J. Voss. *Langevin sampling in path space: implicit methods and pre-conditioning.* In preparation.

[22] B.L. Rozovskii. *Stochastic Evolution Systems: Linear Theory and Applications to Nonlinear Filtering.* Kluwer, The Netherlands, 1990.

[23] T. Shardlow and A.M. Stuart *A perturbation theory for ergodic Markov chains with application to numerical approximation.* SIAM J. Num. Anal. **37**(2000), 1120–1137.

[24] 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.

[25] D. Talay, *Second-order discretization schemes for stochastic differential systems for the computation of the invariant law.* Stochastics and Stochastics Reports **29**(1990), 13–36.

[26] J. Zabczyk. Symmetric solutions of semilinear stochastic equations. In G. Da Prato and L. Tubaro, editors, *Stochastic Partial Differential Equations and Applications II*, volume 1390 of *Lecture Notes in Mathematics*, pages 237–256. Springer, 1988. Proceedings, Trento 1988.