Signal Processing Problems on Function Space: Bayesian Formulation, Stochastic PDEs and Effective MCMC Methods

March 4, 2009

M. Hairer[1,2], A. Stuart[1], J. Voss[1,3]

[1] Mathematics Institute, Warwick University, Coventry, UK
[2] Courant Institute, NYU, New York, USA
[3] Department of Statistics, University of Leeds, Leeds, UK

Abstract

In this chapter we overview a Bayesian approach to a wide range of signal processing problems in which the goal is to find the signal, which is a solution of an ordinary or stochastic differential equation, given noisy observations of its solution. In the case of ordinary differential equations (ODEs) this gives rise to a finite dimensional probability measure for the initial condition, which then determines the measure on the signal. In the case of stochastic differential equations (SDEs) the measure is infinite dimensional, on the signal itself, a time-dependent solution of the SDE.

We derive the posterior measure for these problems, applying the ideas to ODEs and SDEs, with discrete or continuous observations, and with coloured or white noise. We highlight common structure inherent in all of the problems, namely that the posterior measure is absolutely continuous with respect to a Gaussian prior. This structure leads naturally to the study of Langevin equations which are invariant for the posterior measure and we highlight the theory and open questions relating to these S(P)DEs. We then describe the construction of effective Metropolis-based sampling methods for the posterior measure, based on proposals which can be interpreted as approximations of the Langevin equation.

Contents

1 Overview

Many applied problems concerning the integration of data and mathematical model arise naturally in dynamically evolving systems. These may be formulated in the general framework of Bayesian inference. There is particular structure inherent in these problems, arising from the underlying dynamical models, that can be exploited. In this chapter we highlight this structure in the context of continuous time dynamical models in finite dimensions. We set-up a variety of Bayesian inference problems, some finite dimensional, for the initial condition of the dynamics, and some infinite dimensional, for a time-dependent path of an SDE. All of the problems share a common mathematical structure namely that the posterior measure μ^y, given data y, has a density with respect to a Gaussian reference measure μ_0 so that

$$\frac{d\mu^y}{d\mu_0}(x) = Z(y)^{-1}\exp\Big(-\Phi(x;y)\Big) \tag{1.1}$$

for some potential function $\Phi(\cdot;y)$ and normalisation constant $Z(y)$ both parameterized by the data. We denote the mean and covariance operator for μ_0 by m_0 and $\mathcal{C}_0$, and we use $\mathcal{L} = \mathcal{C}_0^{-1}$ to denote the precision operator. Thus $\mu_0 = \mathcal{N}(m_0, \mathcal{C}_0)$.

The content of this chapter is centred around three main themes:

- Theme A. To exhibit a variety of problems arising in data assimilation which share the common structure (1.1).
- Theme B. To introduce a class of stochastic PDEs (SPDEs) which are reversible with respect to μ_0 or μ^y respectively.
- Theme C. To introduce a range of Metropolis-Hastings MCMC methods which sample from μ^y, based on the SPDEs discussed in Theme B.

A central aspect of this chapter will be to exhibit, in Theme A, common properties of the potential $\Phi(x;y)$ which then form the key underpinnings of the theories outlined in Themes B and C. These common properties of Φ include bounds from above and below, continuity in both x and y, and differentiability in x.

The continuous time nature of the problems means that, in some cases, the probability measures constructed are on infinite dimensional spaces: paths of continuous-time, vector valued processes. We sometimes refer to this as *pathspace*. Working with probability measures on function space is a key idea throughout this Chapter. We will show that this viewpoint leads to the notion of a well-posed signal processing problem, in which the target measure is continuous with respect to data. Furthermore a proper mathematical formulation of the problems on pathspace leads to efficient sampling techniques, defined on pathspace, and therefore robust under the introduction of discretization. In contrast, sampling techniques which first discretize, to obtain a finite dimensional sampling problem, and then apply standard MCMC techniques, will typically lead to algorithms which perform poorly under refinement of the discretization.

In section 2 we describe some general properties of measures defined through (1.1). Sections 3.1–3.6 are concerned with Theme A. In section 3.1 we initiate our study by considering a continuous time deterministic ODE observed noisily at discrete times; the objective is to determine the initial condition. Sections 3.2 and 3.3 generalize this set-up to the situation where the solution is observed continuously in time, and subject to white noise and coloured noise respectively. In section 3.4 we return to discrete observations, but assume that the underlying model dynamics is subject to noise – or *model error*; in particular we assume that the dynamics is forced by an Ornstein-Uhlenbeck process. This section, and the remaining sections in Theme A, all contain situations where the posterior measure is on an infinite dimensional space. Section 3.5 generalizes the ideas in section 3.4 to the situation where the model error is described by *white noise* in time. Finally, in section 3.6, we consider the situation with model error (in the form of white noise) and continuous time observations. In section 4 we address Theme B, whilst section 5 is concerned with Theme C. Some notational conventions, and background theory, are outlined in the Appendix.

We emphasize that, throughout this chapter, all the problems discussed are formulated as *smoothing* problems, not *filtering* problems. Thus time distributed data on a given time-interval $[0,T]$ is used to update knowledge about the entire state of the system on the same time interval. For a discussion of filtering methods we refer to [DdFG01].

Acknowledgements

The authors gratefully acknowledge the support of EPSRC grant EP/E002269/1.

2 General Properties of the Posterior

In the next section we will exhibit a wide variety of signal processing problems which, when tackled in a Bayesian framework, lead to a posterior probability measure μ on a Banach space

$(E, \|\cdot\|_E)$, specified via its Radon-Nikodym derivative with respect to a prior Gaussian measure μ_0. Specifically we have (1.1) where $\mu_0 = \mathcal{N}(m_0, \mathcal{C}_0)$ is the prior Gaussian measure and $\Phi(x; y)$ is a *potential*. We assume that $y \in Y$, a separable Banach space with norm $\|\cdot\|_Y$. The normalization constant $Z(y)$ is chosen so that μ is a probability measure:

$$Z(y) = \int_E \exp(-\Phi(x; y))\, d\mu_0(x). \tag{2.1}$$

For details about Gaussian measures on infinite dimensional spaces we refer to the monograph [Bog98].

In many of the applications considered here, Φ satisfies the following four properties:

Assumption 2.1 *The function $\Phi\colon E \times Y \to \mathbb{R}$ has the following properties:*

1. *For every $r > 0$ and every $\varepsilon > 0$, there exists $M = M(r, \varepsilon)$ such that, for all $x \in E$ and all y such that $\|y\|_Y \le r$, $\Phi(x; y) \ge M - \varepsilon \|x\|_E^2$.*
2. *There exists $p \ge 0$ and, for every $r > 0$, there exists $C = C(r) > 0$ such that, for all $x \in E$ and all $y \in Y$ with $\|y\|_Y \le r$, $\Phi(x; y) \le C(1 + \|x\|_E^p)$.*
3. *For every $r > 0$ and every $R > 0$ there exists $L = L(r, R) > 0$ such that, for all $x_1, x_2 \in E$ with $\|x_1\|_E \vee \|x_2\|_E \le R$ and all $y \in Y$ with $\|y\|_Y \le r$,*
$$|\Phi(x_1; y) - \Phi(x_2; y)| \le L \|x_1 - x_2\|_E\ .$$
4. *There exists $q > 0$ and, for every $r > 0$, there exists $K = K(r) > 0$ such that, for all $x \in E$ and all $y_1, y_2 \in E$ with $\|y_1\|_Y \vee \|y_2\|_Y \le r$,*
$$|\Phi(x; y_1) - \Phi(x; y_2)| \le K(1 + \|x\|_E^q) \|y_1 - y_2\|_Y\ .$$

We show that, under these assumptions, the posterior measure μ^y is continuous with respect to the data y in the total variation distance. This is a well-posedness result for the posterior measure. The result, and proof, is similar to that in [CDRS] which concerns Bayesian inverse problems for the Navier-Stokes equations, but where the Hellinger metric is used to measure distance.

Theorem 2.2 *Let μ^y and μ_0 be measures on a separable Banach space E such that μ_0 is a Gaussian probability measure, μ^y is absolutely continuous w.r.t. μ_0, and the log density $\Phi = -\log\Big(\frac{d\mu^y}{d\mu_0}\Big)\colon E \times Y \to \mathbb{R}$ satisfy Assumption 2.1. Then μ^y is a probability measure and the map $y \mapsto \mu^y$ is locally Lipschitz continuous in total variation distance: if μ and μ' are two measures given by (1.1) with data y and y' then, for every $r > 0$ and for all y, y' with $\|y\|_Y \vee \|y'\|_Y \le r$, there exists a constant $C = C(r) > 0$ such that*

$$\|\mu - \mu'\|_{\mathrm{TV}} \le C \|y - y'\|_Y.$$

Proof. Since the reference measure μ_0 is Gaussian, $\|x\|_E$ has Gaussian tails under μ_0. The lower bound (i) therefore immediately implies that $\exp(-\Phi)$ is integrable, so that $Z(y)$ is indeed finite for every y.

We now turn to the continuity of the measures with respect to y. Throughout the proof, all integrals are over E. We fix a value $r > 0$ and use the notation C to denote a strictly positive constant that may depend upon r and changes from occurrence to occurrence. As in the statement, we fix $y, y' \in Y$ and we write $\mu = \mu^y$ and $\mu' = \mu^{y'}$ as a shorthand. Let Z and Z' denote the normalization constants for μ and μ', so that

$$Z = \int \exp\Big(-\Phi(x; y)\Big) d\mu_0(x)\ , \qquad Z' = \int \exp\Big(-\Phi(x; y')\Big) d\mu_0(x)\ .$$

Since μ_0 is Gaussian, assumption (1) yields the upper bound $|Z| \vee |Z'| \le C$. In addition, since Φ is bounded above by a polynomial by (2), we have a similar lower bound $|Z| \wedge |Z'| \ge C$. Using again the Gaussianity of μ_0, the bound (4) yields

$$|Z - Z'| \le C \int \|y - y'\|_Y (1 + \|x\|_E^q) \exp(-(\Phi(x; y) \vee \Phi(x; y'))) \, d\mu_0(x)$$

$$\le C\|y-y'\|_Y \int (1+\|x\|_E^q)\exp(\varepsilon\|x\|_E^2 - M)\,d\mu_0(x)$$
$$\le C\|y-y'\|_Y\ . \tag{2.2}$$

From the definition of the total variation distance, we then have

$$\|\mu-\mu\|_{\mathrm{TV}} = \int \Big|Z^{-1}\exp(-\Phi(x;y)) - (Z')^{-1}\exp(-\Phi(x;y'))\Big|\,d\mu_0(x)$$
$$\le I_1 + I_2\ ,$$

where

$$I_1 = \frac{1}{Z}\int \Big|\exp(-\Phi(x;y)) - \exp(-\Phi(x;y'))\Big|\,d\mu_0(x)\ ,$$
$$I_2 = \frac{|Z-Z'|}{ZZ'}\int \exp(-\Phi(x;y'))\,d\mu_0(x)\ .$$

Since Z is bounded from below, we have $I_1 \le C\|y-y'\|_Y$ just as in (2.2). The second term is bounded similarly by (2.2) and the lower bound (i) on Φ, thus concluding the proof. □

Remark 2.3 We only ever use the Gaussianity of μ_0 to deduce that there exists $\varepsilon > 0$ such that $\int_E \exp(\varepsilon\|x\|_E^2)\,\mu_0(dx) < \infty$. Therefore, the statement of Theorem 2.2 extends to any reference measure μ_0 with this property.

3 Theme A. Bayesian Inference for Signal Processing

It is instructive to summarize the different cases treated in this section in a table. The choice of column determines whether or not model error is present, and when present whether it is white or coloured; the choice of row determines whether or not the observations are discrete, and when continuous whether or not the observational noise is white or coloured. There are three possibilities not covered here; however the reader should be able to construct appropriate models in these three cases after reading the material herein.

	model error		
observation	no	white	coloured
discrete	3.1	3.5	3.4
white	3.2	3.6	
coloured	3.3		

3.1 No Model Error, Discrete Observations

Here we address the question of making inference concerning the initial condition for an ODE, given noisy observation of its trajectory at later times. Thus the basic unknown quantity, which we wish to find the (posterior) distribution of, is a finite-dimensional vector.

Let $v \in C^1([0,T],\mathbb{R}^n)$ solve the ODE

$$\frac{dv}{dt} = f(v), \quad v(0) = u. \tag{3.1}$$

We assume that f is sufficiently nice (say locally Lipschitz and satisfying a coercivity condition) that the equation defines a semigroup $\varphi^t\colon \mathbb{R}^n \to \mathbb{R}^n$ with $v(t) = \varphi^t(u)$. We assume that we observe the solution in discrete time, at times $\{t_k\}_{k=1}^K$. Specifically, for some function $g\colon \mathbb{R}^n \to \mathbb{R}^l$ we observe

$$y_k = g(v(t_k)) + \eta_k, \quad k = 1,\dots,K, \tag{3.2}$$

where the $\eta_k \sim \mathcal{N}(0,B_k)$ are a sequence of Gaussian random variables, not necessarily independent. We assume that

$$0 < t_1 \le t_2 \le \cdots \le t_K \le T. \tag{3.3}$$

Concatenating the data we may write

$$y = \mathcal{G}(u) + \eta, \tag{3.4}$$

where $y = (y_1, \ldots, y_K)$ are the observations, $\mathcal{G}(u) = (g(\varphi^{t_1}(u)), \ldots, g(\varphi^{t_K}(u)))$ maps the state of the system and $\eta = (\eta_1, \ldots, \eta_K)$ is the observational noise. Thus $\eta \sim \mathcal{N}(0, B)$ for some matrix B capturing the correlations amongst the $\{\eta_k\}_{k=1}^K$.

We will now construct the distribution of the initial condition u given an observation y, using Bayes formula (see (A.1) in the Appendix). We assume that the prior measure on u is a Gaussian $\mu_0 \sim \mathcal{N}(m_0, \mathcal{C}_0)$, with mean m_0 and covariance matrix $\mathcal{C}_0$. Given the initial condition u, the observations y are distributed according to the Gaussian measure with density

$$\mathbf{P}(y|u) \propto \exp(-\Phi(u; y)), \quad \Phi(u; y) = \frac{1}{2}|y - \mathcal{G}(u)|_B^2\ , \tag{3.5}$$

where we define $|y|_B^2 = \langle y, B^{-1}y\rangle$ (see Appendix). By Bayes rule we deduce that the posterior measure ν on u, given y, has Radon-Nikodym derivative

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp\Big(-\Phi(u; y)\Big). \tag{3.6}$$

Thus the measure μ^y has density π with respect to Lebesgue measure which is given by

$$\pi(u) \propto \exp\Big(-\frac{1}{2}|y - \mathcal{G}(u)|_B^2 - \frac{1}{2}|u - m_0|_{\mathcal{C}_0}^2\Big). \tag{3.7}$$

Example 3.1 *Let $n = 1$ and consider the ODE*

$$\frac{dv}{dt} = av, \quad v(0) = u.$$

Thus $\varphi^t(u) = \exp(at)u$. As our prior measure we take the Gaussian $\mathcal{N}(m_0, \sigma^2)$ Assume that we observe the solution itself at times t_k and subject to mean zero i.i.d. Gaussian noises with variance γ^2, resulting in observations $\{y_k\}_{k=1}^K$. We have observations $y = Au + \eta$ where $\eta \sim \mathcal{N}(0, B)$

$$\begin{aligned} A &= (\exp(at_1), \ldots, \exp(at_K)) \\ B &= \gamma^2 I. \end{aligned}$$

The posterior measure is then Gaussian with density

$$\pi(u) \propto \exp\Big(-\frac{1}{2\gamma^2}\sum_{k=1}^K |y_k - \exp(at_k)u|^2 - \frac{1}{2\sigma^2}|u - m_0|^2\Big).$$

By completing the square we find that the posterior mean is

$$\frac{m_0 + \frac{\sigma^2}{\gamma^2}\sum_{k=1}^K \exp(at_k)y_k}{1 + \frac{\sigma^2}{\gamma^2}\sum_{k=1}^K \exp(2at_k)}$$

and the posterior variance is

$$\frac{\sigma^2}{1 + \frac{\sigma^2}{\gamma^2}\sum_{k=1}^K \exp(2at_k)}.$$

Since $y_k = \exp(at_k)u + \eta_k$ we may write $y_k = \exp(at_k)u + \gamma\xi_k$ for $\xi \sim \mathcal{N}(0, 1)$. We set

$$y = (y_1, \ldots, y_K)\ , \qquad \xi = (\xi_1, \ldots, \xi_K)\ .$$

The posterior mean m and covariance Σ can then be written succinctly as

$$\Sigma = \frac{\sigma^2}{1 + \frac{\sigma^2}{\gamma^2}|A|^2},$$

$$m = \frac{m_0 + \frac{\sigma^2}{\gamma^2}\langle A, y\rangle}{1 + \frac{\sigma^2}{\gamma^2}|A|^2} = \frac{m_0 + \frac{\sigma^2}{\gamma^2}\langle A, Av(0) + \gamma\xi\rangle}{1 + \frac{\sigma^2}{\gamma^2}|A|^2}.$$

We now consider the limits of small observational noise, and of large data sets, respectively.

First consider small noise. As $\gamma^2 \to 0$, the posterior variance converges to zero and the posterior mean to $\langle A, y\rangle/|A|^2$, solution of the least squares problem

$$\mathrm{argmin}_x |y - Ax|^2.$$

Now consider large data sets where $K \to \infty$. If $|A|^2 \to \infty$ as $K \to \infty$ then the posterior mean converges almost surely to the correct initial condition $v(0)$, and the posterior variance converges to zero. Thus the posterior approaches a Dirac supported on the correct initial condition. Otherwise, if $|A|^2$ approaches a finite limit, then uncertainty remains in the posterior, and the prior has significant effect in determining both the mean and variance of the posterior. □

Example 3.2 *Consider the Lorenz equations*

$$\frac{dv_1}{dt} = \sigma(v_2 - v_1), \qquad \frac{dv_2}{dt} = \rho v_1 - v_2 - v_1 v_3, \qquad \frac{dv_3}{dt} = v_1 v_2 - \beta v_3,$$

started at $v(0) = u \in \mathbb{R}^3$. In this case, as a consequence of the chaoticity of the equations, observing a trajectory over long intervals of time does not allow one to gain more information on the initial condition. See Figure 1 for an illustration. We refer to [Eve06] for further discussion of this example.

We now return to the general problem and highlight a program that we will carry out in earnest for a number of more complex infinite-dimensional posterior measures in later sections. We work within the context of ODEs with globally Lipschitz continuous drifts. This condition ensures global existence of solutions, as well as enabling a straightforward explicit bound on the solution in terms of its initial data. However other, less stringent, assumptions could also be used, provided global existence is known. For example, generalisations to locally Lipschitz drifts satisfying a suitable dissipativity condition are straightforward.

Theorem 3.3 *Assume that $f\colon \mathbb{R}^n \to \mathbb{R}^n$ and $g\colon \mathbb{R}^n \to \mathbb{R}^l$ are globally Lipschitz. Let $E = \mathbb{R}^n$, $Y = \mathbb{R}^{\ell K}$. Then $\mu^y \ll \mu_0$ with $d\mu^y/d\mu_0 \propto \exp(-\Phi)$ where Φ is given by (3.5). The map Φ satisfies Assumptions 2.1 and $y \mapsto \mu^y$ is locally Lipschitz continuous in the total variation metric.*

Proof. It will be useful to introduce the notation $g^k\colon \mathbb{R}^n \to \mathbb{R}^l$ defined by $g^k = g \circ \varphi^{t_k}$. Recall from (3.4) that

$$\mathcal{G}(u) = (g^1(u), \ldots, g^K(u)). \tag{3.8}$$

This is linearly bounded since g and φ^t are linearly bounded; clearly $\Phi \geq 0$ by construction. Thus (1) and (2) of Assumption 2.1 are satisfied. Furthermore, by the Lipschitz continuity of f, $\varphi^t\colon \mathbb{R}^n \to \mathbb{R}^n$ is well-defined and Lipschitz. As the composition of Lipschitz functions, $\mathcal{G}$ is itself a Lipschitz function. Hence $\Phi(\cdot; y)$ is Lipschitz and (3) holds. Also $\Phi(x; \cdot)$ is quadratic in y and hence (4) holds. □

For the purpose of studying algorithms that sample from μ^y, it is also of interest to show that the derivative of $\Phi(\cdot\,; y)$ is sufficiently regular.

Theorem 3.4 *Let $k > 0$. Assume, on top of the assumptions of Theorem 3.3, that $f \in C^k(\mathbb{R}^n, \mathbb{R}^n)$ and $g \in C^k(\mathbb{R}^n, \mathbb{R}^l)$. Then the potential $\Phi(\cdot; y)$ given by (3.5) is in $C^k(\mathbb{R}^n, \mathbb{R})$.*

Proof. As in the previous proof, the observation vector $\mathcal{G}(u)$ is given by (3.8). By standard ODE theory $\varphi^t \in C^k(\mathbb{R}^n, \mathbb{R}^n)$. As the composition of C^k functions, $\mathcal{G}$ is itself a C^k function. Since Φ is quadratic in $\mathcal{G}$, the result follows for Φ. □

3.2 No Model Error, Continuous White Observational Noise

We now consider the preceding problem in the limit where $K \to \infty$ and the set $\{t_i\}_{i=1}^{\infty}$ is dense in $[0,T]$. Once again $v(t)$ solves (3.1):

$$\frac{dv}{dt} = f(v), \quad v(0) = u\ , \tag{3.9}$$

but now we assume that we observe a function of the solution in continuous time, and subject to white noise. Specifically we assume that we observe the time-integrated data y solving the SDE

$$\frac{dy}{dt} = g(v) + \sqrt{\Sigma}\frac{dW}{dt}, \quad y(0) = 0. \tag{3.10}$$

Here $g\colon \mathbb{R}^n \to \mathbb{R}^\ell$ and $\Sigma \in \mathbb{R}^{\ell\times\ell}$ is positive-definite. Using as before the semigroup φ^t solving (3.9), this may be rewritten as

$$\frac{dy}{dt} = g(\varphi^t(u)) + \sqrt{\Sigma}\frac{dW}{dt}, \quad y(0) = 0.$$

The precise interpretation of the data $\{y(t)\}_{t\in[0,T]}$ is that we observe the function $y(t)$ defined by

$$y(t) = \int_0^T g(\varphi^t(u))\,dt + \sqrt{\Sigma}W(t).$$

Let $\mathbb{Q}_0$ denote the Gaussian measure on $L^2([0,T],\mathbb{R}^\ell)$ given by the law of $y(t) = \sqrt{\Sigma}W(t)$. Now place the prior measure $\mu_0 \sim \mathcal{N}(m_0, \mathcal{C}_0)$ on the initial condition in $\mathbb{R}^n$. Then take ν_0 to be the product measure on $L^2([0,T],\mathbb{R}^\ell)\times\mathbb{R}^n$ given by $\mathbb{Q}_0 \otimes \mu_0$. Note that $\nu_0(du|y) = \mu_0(du)$ since u and y are independent under ν_0.

Let $\mathbb{Q}^u$ denote the measure on $L^2([0,T],\mathbb{R}^\ell)$ for y solving (3.10), with u given. By the Girsanov Theorem A.2 we have that

$$\frac{d\mathbb{Q}^u}{d\mathbb{Q}_0}(y) = \exp\Big(-\frac{1}{2}\int_0^T |g(\varphi^t(u))|_\Sigma^2 dt + \int_0^T \langle g(\varphi^t(u)), dy\rangle_\Sigma\Big).$$

Thus if ν is the measure on $L^2([0,T],\mathbb{R}^\ell)\times\mathbb{R}^n$ given by (3.10) with u drawn from μ_0, then we have

$$\frac{d\nu}{d\nu_0}(u,y) = \exp\Big(-\frac{1}{2}\int_0^T |g(\varphi^t(u))|_\Sigma^2 dt + \int_0^T \langle g(\varphi^t(u)), dy\rangle_\Sigma\Big).$$

Let $\mu^y(du)$ denote $\nu(du|y)$. By Theorem A.1 we have

$$\begin{aligned}&\frac{d\mu^y}{d\mu_0}(u) \propto \exp(-\Phi(u;y))\ ,\\ &\Phi(u;y) = \frac{1}{2}\int_0^T |g(\varphi^t(u))|_\Sigma^2 dt - \int_0^T \langle g(\varphi^t(u)), dy\rangle_\Sigma\ .\end{aligned} \tag{3.11}$$

Integrating the second term by parts and using the fact that φ^t solves (3.9), we find that

$$\Phi(u;y) = \frac{1}{2}\int_0^T \Big(|g(\varphi^t(u))|_\Sigma^2 + 2\langle Dg(\varphi^t(u))f(\varphi^t(u)), y\rangle_\Sigma\Big) - \langle g(\varphi^T(u)), y(T)\rangle_\Sigma\ . \tag{3.12}$$

Theorem 3.5 *Assume that $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz continuous and linearly bounded, and that $g \in C^2(\mathbb{R}^n, \mathbb{R}^l)$ is globally Lipschitz continuous. Let $E = \mathbb{R}^n$ and $Y = C([0,T],\mathbb{R}^l)$. Then $\mu^y \ll \mu_0$ with $d\mu^y/d\mu_0 \propto \exp(-\Phi)$ where Φ is given by* (3.12). *The function Φ satisfies Assumptions 2.1 and $y \mapsto \mu^y$ is locally Lipschitz continuous in the total variation metric.*

Proof. Since $ab \geq -\frac{\varepsilon}{2}a^2 - \frac{1}{2\varepsilon}b^2$ for any $\varepsilon > 0$ and $a, b \in \mathbb{R}$, it follows from (3.12) that Φ is bounded from below by

$$\Phi(u;y) \geq -\frac{C}{\varepsilon}\|y\|_{L^\infty}^2 - \varepsilon|g(\varphi^T(u))|^2 - \varepsilon T\int_0^T |Dg(\varphi^t(u))f(\varphi^t(u))|^2\,dt\ ,$$

for some constant C depending only on Σ. Since the assumption that f grows linearly implies the existence of a constant C such that $|\varphi^t(u)| \le C|u|$ for every $t \in [0,T]$, the requested lower bound on Φ follows from the linear growth assumptions on f and g as well as the boundedness of Dg.

The polynomial upper bound on $\Phi(\cdot; y)$ follows in exactly the same way, yielding

$$|\Phi(u;y)| \le C(\|y\|^2_{L^\infty} + |u|^2) ,$$

for some constant C. Conditions (3) and (4) in Assumption 2.1 follow similarly, using the Lipschitz continuity of $\varphi^t(\cdot)$ as in the proof of Theorem 3.3. □

Remark 3.6 Note that it is the need to prove continuity in y which requires us to work in the function space $C([0,T],\mathbb{R}^l)$, since computation of Φ requires the evaluation of y at time T. Note also that it is possible to weaken the growth conditions on f and g, at the expense of strengthening the dissipativity assumptions on f.

We mention an insightful, unrigorous but useful, way of writing the potential Φ. If we pretend that y is differentiable in time (which it almost surely isn't), then we may write Φ from (3.11) as

$$\Phi(u;y) = \frac{1}{2}\int_0^T \Big|g(\varphi^t(u)) - \frac{dy}{dt}\Big|_\Sigma^2 dt - \frac{1}{2}\int_0^T \Big|\frac{dy}{dt}\Big|_\Sigma^2 dt .$$

The Gaussian reference measure μ_0, again only at a formal level, has density $\exp(-\frac{1}{2}\int_0^T |\frac{dy}{dt}|_\Sigma^2 dt)$ with respect to (the of course nonexistent) 'Lebesgue measure'. This suggests that μ^y has density $\exp(-\frac{1}{2}\int_0^T |g(\varphi^t(u)) - \frac{dy}{dt}|_\Sigma^2\, dt)$.

In this form we see that, as in the previous section, the potential Φ for the Radon-Nikodym derivative between posterior and prior, written in the general form (1.1), simply measures the mismatch between data and observation operator. This nonrigorous rewrite of the potential Φ is useful precisely because it highlights this fact, easily lost in the mathematically correct formulation (3.11).

The following theorem is proved similarly to Theorem 3.4.

Theorem 3.7 *Let $k > 0$. Assume that we are in the setting of Theorem 3.5 and that furthermore $f \in C^k(\mathbb{R}^n, \mathbb{R}^n)$, and $g \in C^{k+1}(\mathbb{R}^n, \mathbb{R}^l)$. Then the potential $\Phi(\cdot; y)$ given by (3.12) belongs to $C^k(\mathbb{R}^n, \mathbb{R})$.*

3.3 No Model Error, Continuous Coloured Observational Noise

We now consider the discrete observations problem in the limit where $K \to \infty$ and the set $\{t_i\}_{i=1}^\infty$ is dense in $[0,T]$, but we assume that the observational noise is correlated. We model this situation by assuming that $u(t)$ solves (3.1), and that we observe a function of the solution in continuous time, subject to noise drawn from the distribution of a stationary Ornstein-Uhlenbeck process. In other words, we observe the process

$$y(t) = g(\varphi^t(u)) + \psi(t) , \tag{3.13}$$

where

$$\frac{d\psi}{dt} = -R\psi + \sqrt{2\Lambda}\frac{dW}{dt}, \quad \psi(0) \sim \mathcal{N}(0, R^{-1}\Lambda) . \tag{3.14}$$

Here $g\colon \mathbb{R}^n \to \mathbb{R}^\ell$, and the matrices $R, \Lambda \in \mathbb{R}^{\ell\times\ell}$ are symmetric positive-definite and are assumed to commute for simplicity (in particular, this ensures that the process ψ is reversible).

Once again we adopt a Bayesian framework to find the posterior probability for u given y. For economy of notation we set $\theta(t) = g(\varphi^t(u))$ and denote by $\dot\theta(t)$ the time-derivative of θ. We deduce from Itô's formula that

$$\begin{aligned}\frac{dy}{dt} &= \dot\theta - R\psi + \sqrt{2\Lambda}\frac{dW}{dt} = \dot\theta - R(y-\theta) + \sqrt{2\Lambda}\frac{dW}{dt}\\ &= \dot\theta + R\theta - Ry + \sqrt{2\Lambda}\frac{dW}{dt}.\end{aligned}$$

Furthermore
$$y(0) \sim \mathcal{N}(\theta(0), R^{-1}\Lambda) = \mathcal{N}(g(u), R^{-1}\Lambda) \,,$$
independently of W.

Let $\mathbb{Q}_0$ denote the measure on $L^2([0,T],\mathbb{R}^\ell)$ generated by the Gaussian process (3.14) and place the prior measure $\mu_0 \sim \mathcal{N}(m_0, \mathcal{C}_0)$ on the initial condition u in $\mathbb{R}^n$. Then take ν_0 to be the measure on $L^2([0,T],\mathbb{R}^\ell) \times \mathbb{R}^n$ given by $\mathbb{Q}_0 \otimes \mu_0$. Note that $\nu_0(du|y) = \mu_0(du)$ since u and y are independent under ν_0.

We let ν denote the probability measure for y and u, with u distributed according to μ_0. By the Girsanov theorem
$$\begin{aligned}\frac{d\nu}{d\nu_0}(u,y) &\propto \exp(-\Phi(u;y))\\ \Phi(u;y) &= \frac{1}{4}\int_0^T (|h(\varphi^t(u))|_\Lambda^2\, dt - 2\langle h(\varphi^t(u)), dy + Ry\, dt\rangle_\Lambda)\\ &\quad + \frac{1}{2}|g(u)|^2_{R^{-1}\Lambda} - \langle y(0), g(u)\rangle_{R^{-1}\Lambda}.\end{aligned}$$
Here and below we use the shortcut
$$\begin{aligned}h(u) &= \dot\theta + R\theta,\\ &= Dg(u)f(u) + Rg(u).\end{aligned}$$

We let $\mu^y(u)$ denote the posterior distribution for u given y. By applying Bayes formula in the guise of Theorem A.1 we obtain
$$\begin{aligned}\frac{d\mu^y}{d\mu_0}(u) &\propto \exp(-\Phi(u;y)) \\ \Phi(u;y) &= \frac{1}{4}\int_0^T (|h(\varphi^t(u))|_\Lambda^2\, dt - 2\langle h(\varphi^t(u)), dy + Ry\, dt\rangle_\Lambda)\\ &\quad + \frac{1}{2}|g(u)|^2_{R^{-1}\Lambda} - \langle y(0), g(u)\rangle_{R^{-1}\Lambda}.\end{aligned} \tag{3.15}$$
As in the previous section, the posterior measure involves a Riemann integral and a stochastic integral, both parameterized by u, but the stochastic integral can be converted to a Riemann integral, by means of an integration by parts. Setting
$$\tilde h(u) = Dh(u)f(u) = Dg(u)Df(u)\,f(u) + D^2g(u)(f(u),f(u)) + (RDg(u))\,f(u)\,,$$
we find that
$$\begin{aligned}\Phi(u;y) &= \frac{1}{4}\int_0^T (|h(\varphi^t(u))|_\Lambda^2 + 2\langle \tilde h(\varphi^t(u)), y - Ry\rangle_\Lambda)dt\\ &\quad + \frac{1}{2}\langle h(u), y(0)\rangle_\Lambda - \frac{1}{2}\langle h(\varphi^T(u)), y(T)\rangle_\Lambda\\ &\quad + \frac{1}{2}|g(u)|^2_{R^{-1}\Lambda} - \langle y(0), g(u)\rangle_{R^{-1}\Lambda}.\end{aligned} \tag{3.16}$$

Proof of the following two theorems is very similar to those for Theorems 3.5 and 3.7.

Theorem 3.8 *Assume that $f \in C^2(\mathbb{R}^n,\mathbb{R}^n)$ is linearly bounded in* (3.1)*, that $g \in C^3(\mathbb{R}^n,\mathbb{R}^l)$ is globally Lipschitz continuous in* (3.13)*, and that $\tilde h$ is linearly bounded. Let $E = \mathbb{R}^n$ and $Y = C([0,T],\mathbb{R}^l)$. Then $\mu^y \ll \mu_0$ with $d\mu^y/d\mu_0 \propto \exp(-\Phi)$ where Φ is given by* (3.16)*. The map Φ satisfies Assumptions 2.1 and $y \mapsto \mu^y$ is locally Lipschitz continuous in the total variation metric.*

Remark 3.9 The condition that $\tilde h$ is linearly bounded follows for example if we assume that $D^2g(u)$ is bounded by $C/(1+|u|)$ for some constant C.

Theorem 3.10 *Let $k \geq 1$. Assume that, further to satisfying the assumptions of Theorem 3.8, one has $f \in C^{k+1}(\mathbb{R}^n,\mathbb{R}^n)$ and $g \in C^{k+2}(\mathbb{R}^n,\mathbb{R}^l)$. Then the potential $\Phi(\cdot;y)$ given by* (3.12) *belongs to $C^k(\mathbb{R}^n,\mathbb{R})$.*

3.4 Coloured Model Error, Discrete Observations

The posterior probability measures on the initial condition u in the previous three examples can be very complicated objects from which it is hard to extract information. This is particularly true in cases where the semigroup φ^t exhibits sensitive dependence on initial conditions, there is data over a large time-interval, and the system is sufficiently ergodic and mixing. The posterior is then essentially flat, with small random fluctuations superimposed, and contains little information about the initial condition. In such situations it is natural to relax the *hard constraint* that the dynamical model is satisfied exactly and to seek to explain the observations through a *forcing* to the dynamics: we allow equation (3.1) to be forced by an extraneous driving noise, known as *model error*. Thus we view the dynamics (3.1) as only being enforced as a weak constraint, in the sense that the equation need not be satisfied exactly. We then seek a posterior probability measure on both the initial condition and a driving noise process which quantifies the sense in which the dynamics is not exactly satisfied. Since we are working with continuous time, the driving noise process is a *function* and thus the resulting posterior measure is a measure on an *infinite dimensional* space of functions. This section is the first of several where the desired probability measure lives on an infinite dimensional space.

To be concrete we consider the case where the driving noise is correlated in time and governed by an Ornstein-Uhlenbeck process. We thus consider the model equations

$$\begin{aligned} \frac{dv}{dt} &= f(v) + \frac{1}{\sqrt{\delta}}\psi\ , \quad v(0) = u\ , \\ \frac{d\psi}{dt} &= -\frac{1}{\delta}R\psi + \sqrt{\frac{\Lambda}{\delta}}\frac{dW}{dt}, \quad \psi(0) \sim \mathcal{N}\Big(0, \frac{1}{2}R^{-1}\Lambda\Big)\ . \end{aligned} \tag{3.17}$$

We assume as before that R and Λ commute. The parameter δ sets a correlation time for the noise; in the next section we will let $\delta \to 0$ and recover white noise forcing. Equation (3.17) specifies our prior model for the noise process ψ. We assume that $\psi(0)$ is chosen independently of W, and then (3.17) describes a stationary OU process ψ. As our prior on the initial condition we take $u \sim \mathcal{N}(m_0, \mathcal{C}_0)$, independently of ψ. We have thus specified a prior Gaussian probability measure $\mu_0(u_0, \psi)$ on $\mathbb{R}^n \times L^2([0,T], \mathbb{R}^n)$.

As observations we take, as in section 3.1,

$$y_k = g(v(t_k)) + \eta_k, \quad k = 1, \dots, K,$$

where $\eta_k \sim \mathcal{N}(0, B_k)$ are a sequence of Gaussian random variables, not necessarily independent, and the observation times satisfy (3.3). Concatenating the data we may write

$$y = \mathcal{G}(u, \psi) + \eta, \tag{3.18}$$

where $y = (y_1, \dots, y_K)$ are the observations, $\mathcal{G}(u,\psi) = (g(v(t_1)), \dots, g(v(t_K)))$ maps the state of the system and $\eta = (\eta_1, \dots, \eta_K)$ is the observational noise. Thus $\eta \sim \mathcal{N}(0, B)$ for some matrix B capturing the correlations amongst the $\{\eta_k\}_{k=1}^K$. Here $\mathcal{G}$ is a map from $\mathbb{R}^n \times L^2([0,T], \mathbb{R}^n)$ to $\mathbb{R}^{lK}$. The likelihood for the observations is then

$$\mathbf{P}(dy|u,\psi) \propto \exp\Big(-\frac{1}{2}|y - \mathcal{G}(u,\psi)|_B^2\Big)dy.$$

Let ν denote the measure on $\mathbb{R}^n \times L^2([0,T], \mathbb{R}^n) \times \mathbb{R}^{lK}$ given by

$$\nu(du, d\psi, dy) = \mathbf{P}(dy|u,\psi)\mu_0(du, d\psi)\ ,$$

and let ν_0 denote the measure on $\mathbb{R}^n \times L^2([0,T], \mathbb{R}^n) \times \mathbb{R}^{lK}$ given by

$$\nu_0(du, d\psi, dy) \propto \exp\Big(-\frac{1}{2}|y|_B^2\Big)\, \mu_0(du, d\psi)dy\ .$$

Since (u, ψ) and y are independent under $\nu_0(du, d\psi)dy$, Theorem A.1 shows that the posterior probability measure $\mu^y(du, d\psi)$ is given by

$$\frac{d\mu^y}{d\mu_0}(u,\psi) \propto \exp(-\Phi(u,\psi;y))\ , \quad \Phi(u,\psi;y) = \frac{1}{2}|y - \mathcal{G}(u,\psi)|_B^2. \tag{3.19}$$

Example 3.11 *Consider again the Lorenz equation from Example 3.2. Using the setup from this section, we get a posterior distribution on the pairs (u, ψ) where u is the initial condition of the Lorenz ODE as in Example 3.2 above, and ψ is the additional forcing (model error) from (3.17).*

We consider again the setting from Figure 1, but this time with the additional forcing term ψ. Now the posterior for u can be obtained by averaging (3.19) over ψ. This leads to a smoothing of the posterior distribution. The effect is illustrated in Figure 2.

Theorem 3.12 *Assume that $f\colon \mathbb{R}^n \to \mathbb{R}^n$ and $g\colon \mathbb{R}^n \to \mathbb{R}^l$ are globally Lipschitz continuous. Let $E = \mathbb{R}^n \times L^2([0,T], \mathbb{R}^n)$ and $Y = \mathbb{R}^{lK}$. Then $\mu^y \ll \mu_0$ with $d\mu^y/d\mu_0 \propto \exp(-\Phi)$ where Φ is given by (3.19). The map Φ satisfies Assumptions 2.1 and $y \mapsto \mu^y$ is locally Lipschitz continuous in the total variation metric.*

Proof. Assumption 2.1(1) follows with $M = 0$ and $\varepsilon = 0$. To establish (2) we note that, for $0 \le t \le T$,

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}|v|^2 &\le \alpha + \beta|v|^2 + \frac{1}{2\sqrt{\delta}}(\|\psi\|^2 + |v|^2) \\
&\le \alpha + \beta|v|^2 + \frac{1}{2\sqrt{\delta}}(\|\psi\|^2_{L^2([0,T],\mathbb{R}^n)} + |v|^2).
\end{aligned}$$

Application of the Gronwall inequality shows that

$$\|v(t)\| \le C(t)\Big(\|\psi\|_{L^2([0,T],\mathbb{R}^n)} + |u|\Big)$$

and hence that, since g is linearly bounded,

$$|\mathcal{G}(u,\psi)| \le C\Big(\|\psi\|_{L^2([0,T],\mathbb{R}^n)} + |u|\Big).$$

From this, assumption (2) follows.

To establish (3) it suffices to show that $\mathcal{G}\colon \mathbb{R}^n \times L^2([0,T],\mathbb{R}^n) \to \mathbb{R}^{lK}$ is locally Lipschitz continuous; this follows from the stated hypotheses on f and g since the mapping $(u,\psi) \in \mathbb{R}^n \times L^2([0,T],\mathbb{R}^n) \to v \in C([0,T],\mathbb{R}^n)$ is locally Lipschitz continuous, as may be shown by a Gronwall argument similar to that used to establish (2). Assumption 2.1(4) follows from the fact that $\Phi(x;\cdot)$ is quadratic, together with the polynomial bounds on $\mathcal{G}$ from the proof of (2). □

Again, we obtain more regularity on Φ by imposing more stringent assumptions on f and g:

Theorem 3.13 *For $k > 0$, if both f and g are C^k, then Φ is also C^k.*

3.5 White Model Error, Discrete Observations

In the preceding example we described the model error as on OU process. In some situations it is natural to describe the model error as white noise. Formally this can be obtained from the preceding example by taking the limit $\delta \to 0$ so that the correlation time tends to zero. Heuristically we have

$$\frac{1}{\sqrt{\delta}}\psi = R^{-1}\sqrt{2\Lambda}\frac{dW}{dt} + \mathcal{O}(\sqrt{\delta})$$

from the OU process (3.17). Substituting this heuristic into (3.17) and setting $\delta = 0$ gives the white noise driven model

$$\frac{dv}{dt} = f(v) + \sqrt{\Gamma}\frac{dW}{dt}, \quad v(0) = u, \tag{3.20}$$

where $\sqrt{\Gamma} = R^{-1}\sqrt{2\Lambda}$.

Again we assume that we are given observations in the form (3.2). There are now two ways to proceed to define an inverse problem. We can either make inference concerning the pair (u, W), or we can make inference concerning the function v itself. We consider the two approaches in turn. Note that (u, W) uniquely define v and so a probability measure on (u, W) implies a probability measure on v.

First we consider the formulation of the problem where we make inference about (u, W). We construct the prior measure $\mu_0(du, dW)$ by assuming that u and W are independent, by taking $u \sim \mathcal{N}(m_0, \mathcal{C}_0)$ and by taking standard $n-$dimensional Wiener measure for W. Now consider the integral equation

$$v(t) = u + \int_0^t f(v(s))\, ds = \sqrt{\Gamma} W(t)\,. \tag{3.21}$$

The solution of this equation defines a map

$$\begin{aligned} \mathcal{V}\colon \mathbb{R}^n \times C([0,T], \mathbb{R}^n) &\to C([0,T], \mathbb{R}^n) \\ (u, W) &\mapsto \mathcal{V}(u, W) = v\,. \end{aligned}$$

Thus we may write the equation (3.2) for the observations as

$$y = \mathcal{G}(u, W) + \eta\,, \tag{3.22}$$

with η as in (3.18) and $\mathcal{G}_k(u, W) = g(\mathcal{V}(u, W)(t_k))$. The likelihood of y is thus

$$\mathbf{P}(dy|u, W) \propto \exp\Big(-\frac{1}{2}|y - \mathcal{G}(u, W)|_B^2\Big)\, dy\,.$$

This leads to a probability measure

$$\nu(du, dW, dy) = \mathbf{P}(y|u, \psi)\mu_0(du, d\psi)dy$$

on the space $\mathbb{R}^n \times C([0,T], \mathbb{R}^n) \times \mathbb{R}^{lK}$. By ν_0 we denote the measure on $\mathbb{R}^n \times C([0,T], \mathbb{R}^n) \times \mathbb{R}^{lK}$ given by

$$\nu_0(du, dW, dy) \propto \exp\Big(-\frac{1}{2}|y|_B^2\Big)\, \mu_0(du, dW)dy\,.$$

Since (u, ψ) and y are independent under $\nu_0(du, d\psi)dy$, Theorem A.1 shows that the posterior probability measure is given by

$$\begin{aligned} \frac{d\mu^y}{d\mu_0}(u, W) &\propto \exp\Big(-\Phi(u, W; y)\Big) \\ \Phi(u, W; y) &= \frac{1}{2}|y - \mathcal{G}(u, W)|_B^2\,. \end{aligned} \tag{3.23}$$

Theorem 3.14 *Assume that $f\colon \mathbb{R}^n \to \mathbb{R}^n$ and $g\colon \mathbb{R}^n \to \mathbb{R}^l$ are globally Lipschitz continuous. Let $E = \mathbb{R}^n \times C([0,T], \mathbb{R}^n)$ and $Y = \mathbb{R}^{lK}$. Then $\mu^y \ll \mu_0$ with $d\mu^y/d\mu_0 \propto \exp(-\Phi)$ where Φ is given by (3.23). The map Φ satisfies Assumptions 2.1 and $y \mapsto \mu^y$ is locally Lipschitz continuous in the total variation metric.*

Proof. Note that $\mu_0(\mathbb{R}^n \times C([0,T], \mathbb{R}^n)) = 1$, because Wiener measure charges continuous functions with probability 1. Assumption 2.1(1) follows with $M = \varepsilon = 0$. To establish (2) we note that, for $0 \le t \le T$,

$$\begin{aligned} |v(t)| &\le |u| + \int_0^t (\alpha + \beta|v(s)|)\, ds + |W(t)| \\ &\le |u| + \int_0^t (\alpha + \beta|v(s)|)\, ds + \|W\|_{C([0,T],\mathbb{R}^n)}. \end{aligned}$$

Application of the Gronwall inequality shows that

$$\|v(t)\| \le C(t)(\|W\|_{C([0,T],\mathbb{R}^n)} + |u|)\,.$$

Since g is polynomially bounded we have

$$|\mathcal{G}(u, W)| \le C(\|W\|_{C([0,T],\mathbb{R}^n)} + |u|)$$

and (2) follows. To establish (3) it suffices to show that $\mathcal{G}\colon \mathbb{R}^n \times C([0,T], \mathbb{R}^n) \to \mathbb{R}^{lK}$ is continuous; this follows from the stated hypotheses on f and g since the mapping $(u, W) \in \mathbb{R}^n \times C([0,T], \mathbb{R}^n) \to v \in C([0,T], \mathbb{R}^n)$ is continuous, as may be shown by a Gronwall argument, similar to that used to establish (2). Assumption 2.1(4) follows from the fact that $\Phi(x; \cdot)$ is quadratic and the bound on $\mathcal{G}$ derived to establish (2). $\square$

Again, higher-order differentiability is obtained in a straightforward manner:

Theorem 3.15 *For $k > 0$, if both f and g are C^k, then Φ is also C^k.*

We have expressed the posterior measure as a measure on the initial condition u for v and the driving noise W. However, one can argue that it is natural to take the alternative approach of making direct inference about $\{v(t)\}_{t=0}^T$ rather than indirectly through (u, W). We illustrate how this may be done. To define a prior, we first let μ_0 denote the Gaussian measure on $L^2([0,T],\mathbb{R}^n)$ defined by the equation

$$\frac{dv}{dt} = \sqrt{\Gamma}\frac{dW}{dt}, \quad u \sim \mathcal{N}(m_0, \mathcal{C}_0).$$

By the Girsanov theorem, the law of the solution v to (3.20), with $u \sim \mathcal{N}(m_0, \mathcal{C}_0)$, yields a measure ν_0 on $L^2([0,T],\mathbb{R}^n)$ which has Radon-Nikodym derivative

$$\frac{d\nu_0}{d\mu_0}(v) = \exp\Big(-\frac{1}{2}\int_0^T |f(v)|_\Gamma^2\,dt + \int_0^T \langle f(v), dv\rangle_\Gamma\Big)\,, \tag{3.24}$$

where the second integral is an Itô stochastic integral. The data is again assumed to be of the form (3.2). We have

$$\mathbf{P}(dy|v) \propto \exp\Big(-\frac{1}{2}|y - \mathcal{G}(v)|_B^2\Big)dy\,,$$

where $\mathcal{G}(v) = (g(v(t_1)), \dots, g(v(t_K)))$ and B is the correlation in the noise. Here $\mathcal{G}$ is a map from $L^2([0,T],\mathbb{R}^n)$ to $\mathbb{R}^{lK}$.

Thus we may define a probability measure on $\nu(dv, dy)$ on $L^2([0,T],\mathbb{R}^n) \times \mathbb{R}^{lK}$ given by $\mathbf{P}(y|v)\nu_0(dv)dy$. Since v and y are independent under $\nu_0(dv)dy$, Theorem A.1 shows that the posterior probability measure is defined by

$$\frac{d\mu^y}{d\nu_0}(v) \propto \exp(-\widehat{\Phi}(v;y))\,, \qquad \widehat{\Phi}(v;y) = \frac{1}{2}|y - \mathcal{G}(v)|_B^2\,. \tag{3.25}$$

This expresses the posterior measure in terms of a non-Gaussian prior (for the pathspace of a non-Gaussian SDE with Gaussian initial data).

Theorem 3.16 *Assume that $f\colon \mathbb{R}^n \to \mathbb{R}^n$ and $g\colon \mathbb{R}^n \to \mathbb{R}^l$ are globally Lipschitz continuous. Then $\mu^y \ll \nu_0$ with $d\mu^y/d\nu_0 \propto \exp(-\widehat{\Phi})$ where $\widehat{\Phi}$ is given by (3.25). The map $\widehat{\Phi}$ satisfies Assumptions 2.1 and $y \mapsto \mu^y$ is locally Lipschitz continuous in the total variation metric.*

Proof. Note that the reference measure ν_0 is not Gaussian. Thus, by Remark 2.3, we need to make sure that ν_0 has Gaussian tails. This follows immediately from the fact that the solution map $(u, W) \mapsto v$ to the model equations (3.21) is globally Lipschitz continuous from $\mathbb{R}^n \times C([0,T],\mathbb{R}^n)$ into $C([0,T],\mathbb{R}^n)$. As the push-forward of a Gaussian measure under a Lipschitz continuous map, ν_0 therefore has Gaussian tails.

The function

$$\widehat{\Phi}(v;y) = \frac{1}{2}|(y - \mathcal{G}(v))|_B^2$$

is obviously bounded from below. It is furthermore locally Lipschitz continuous in both y and v, since the solution map $\mathcal{G}$ is Lipschitz continuous from $C([0,T],\mathbb{R}^n)$ into $\mathbb{R}^{lK}$. □

Obtaining differentiability results on the density with respect to the Gaussian prior μ_0 is much more tricky, because of the appearance of the stochastic integral in (3.24). We return to this topic below, after making the following observation. It is frequently of interest to express the target measure as change of measure from a Gaussian, for example to implement sampling algorithms as in section 5. This may be achieved by the Girsanov theorem: by the properties of change of measure we have

$$\begin{aligned}\frac{d\mu^y}{d\mu_0}(v) &= \frac{d\mu^y}{d\nu_0}(v) \times \frac{d\nu_0}{d\mu_0}(v)\\ &\propto \exp\Big(-\frac{1}{2}|(y - \mathcal{G}(v))|_B^2\Big)\exp\Big(-\frac{1}{2}\int_0^T (|f(v)|_\Gamma^2 dt - 2\langle f(v), dv\rangle_\Gamma)\Big)\,.\end{aligned}$$

Thus

$$
\begin{aligned}
&\frac{d\mu^y}{d\mu_0}(v) \propto \exp\Big(-\Phi(v;y)\Big) \\
&\Phi(v;y) = \frac{1}{2}|y-\mathcal{G}(v)|_B^2 + \frac{1}{2}\int_0^T \Big(|f(v)|_\Gamma^2 dt - 2\langle f(v), dv\rangle_\Gamma\Big).
\end{aligned} \tag{3.26}
$$

There is a naturally arising case where it is possible to study differentiability: when f has a gradient structure. Specifically, if $f = -\Gamma\nabla F$, then Itô's formula yields

$$
dF(v) = -\langle f(v), dv\rangle_\Gamma + \frac{1}{2}Tr(\Gamma D^2 F)\, dt\ .
$$

Substituting this into (3.26) yields the expression

$$
\begin{aligned}
\Phi(v;y) = &\frac{1}{2}|(y-\mathcal{G}(v))|_B^2 + \frac{1}{2}\int_0^T \Big(|f(v)|_\Gamma^2 - Tr(\Gamma D^2 F)\Big)dt \\
&+ F(v(T)) - F(v(0))\ .
\end{aligned} \tag{3.27}
$$

We thus obtain the following result:

Theorem 3.17 *For $k \geq 1$, if $g \in C^k$, f is a gradient, and $f \in C^{k+1}$, then $\Phi\colon C([0,T],\mathbb{R}^n) \times \mathbb{R}^{lK} \to \mathbb{R}$ is C^k.*

3.6 White Model Error, Continuous White Observational Noise

It is interesting to consider the preceding problem in the limiting case where the observation is in continuous time. Specifically we consider underlying stochastic dynamics governed by (3.20) with observations given by (3.10). We assume that $v(0) \sim \mathcal{N}(m_0, \mathcal{C}_0)$ and hence obtain the prior model equation for v, together with the equation for the continuous time observation y, in the form:

$$
\frac{dv}{dt} = f(v) + \sqrt{\Gamma}\frac{dW_1}{dt}, \quad v(0) \sim \mathcal{N}(u_0, \mathcal{C}_0), \tag{3.28a}
$$

$$
\frac{dy}{dt} = g(v) + \sqrt{\Sigma}\frac{dW_2}{dt}, \quad y(0) = 0. \tag{3.28b}
$$

Here $v \in \mathbb{R}^n, y \in \mathbb{R}^\ell$, $f\colon \mathbb{R}^n \to \mathbb{R}^n$ and $g\colon \mathbb{R}^n \to \mathbb{R}^\ell$. Furthermore, $\Gamma \in \mathbb{R}^{n\times n}$ and $\Sigma \in \mathbb{R}^{\ell\times\ell}$ are assumed positive-definite. The Brownian motions W_1, W_2 are assumed independent.

Our aim is to find the probability distribution for $v \in C([0,T],\mathbb{R}^n)$ given $y \in C([0,T],\mathbb{R}^\ell)$. This is a classical problem in continuous time signal processing for SDEs, known as the *smoothing problem*. This differs from the *filtering problem* where the aim is to find a time-indexed family of probability measures ν_t on $\mathbb{R}^n$ for $v(t)$ given $y \in C([0,t],\mathbb{R}^\ell)$.

First we consider the unconditioned case. We let $\nu_0(dv,dy)$ denote the Gaussian measure on $C([0,T],\mathbb{R}^n) \times C([0,T],\mathbb{R}^\ell)$ obtained in the case where f and g are identically zero, and ν the same measure when f and g are not zero. By the Girsanov Theorem A.2 we have, assuming that trajectories do not explode,

$$
\frac{d\nu}{d\nu_0}(v,y) = \exp\Big(-\frac{1}{2}\int_0^T |f(v)|_\Gamma^2 dt + |g(v)|_\Sigma^2 dt - 2\langle f(v), dv\rangle_\Gamma - 2\langle g(v), dy\rangle_\Sigma\Big)\ . \tag{3.29}
$$

Now we consider the measures found by conditioning v on y. Under ν_0 the random variables v and y are independent. Thus $\nu_0(dv|y)$ is simply the Gaussian measure $\mu_0(dv)$ on $C([0,T],\mathbb{R}^n)$ found from the equation

$$
\frac{dv}{dt} = \sqrt{\Gamma}\frac{dW_1}{dt}\ , \qquad v(0) \sim \mathcal{N}(m_0, C_0)\ .
$$

Now let $\mu^y(u)$ denote the measure on $C([0,T],\mathbb{R}^n)$ found from $\nu(u|y)$. Integrating the last integral in (3.29) by parts (we can do this because v and y are independent under ν_0 so that no Itô correction appears) and then applying Theorem A.1, we deduce that

$$
\frac{d\mu^y}{d\mu_0}(v) \propto \exp\Big(-\Phi(v;y)\Big) \tag{3.30}
$$

$$\Phi(v;y) = \frac{1}{2}\int_0^T \Big(|f(v)|_\Gamma^2 dt + |g(v)|_\Sigma^2 dt - 2\langle f(v), dv\rangle_\Gamma + 2\langle y, Dg(v)\,dv\rangle_\Sigma\Big)$$
$$+ \frac{1}{2}(\langle g(v(0)), y(0)\rangle - \langle g(v(T)), y(T)\rangle)\;.$$

Here, both integrals are stochastic integrals in the sense of Itô and (3.30) is valid for ν_0-almost every $y \in C([0,T],\mathbb{R}^\ell)$. We have therefore shown that:

Theorem 3.18 *Assume that equations* (3.28a) *and* (3.28b) *have solution on* $t \in [0,T]$ *which do not explode, almost surely. Then, the family of measures* μ^y *as defined by equation* (3.30) *provides the conditional distribution for* v *given* y*.*

In this case Assumptions 2.1(1)–(4) do not hold in general. The statement obtained in this case is much weaker than previously: while integration by parts allows us to establish a conditional law for v given y for *every* realisation of the observation process y, we do not obtain Lipschitz continuity of μ^y as a function of y.

One situation where it is possible to establish Assumptions 2.1(1)–(3) for a continuous time model of the form (3.28) is the particular case when g is linear and f is a gradient of the form $f = -\Gamma\nabla F$. In this case, we may rewrite (3.28) as

$$\frac{dv}{dt} = -\Gamma\nabla F(v) + \sqrt{\Gamma}\frac{dW_1}{dt}, \quad v(0) \sim \mathcal{N}(u_0, \mathcal{C}_0), \tag{3.31a}$$
$$\frac{dy}{dt} = Av + \sqrt{\Sigma}\frac{dW_2}{dt}, \quad y(0) = 0\;. \tag{3.31b}$$

The key to what follows is that we choose a slightly different unconditioned measure from before. We let $\nu_0(dv, dy)$ denote the Gaussian measure obtained in the case where F is identically zero (but A is *not* identically 0), and denote as before by ν the measure obtained when f is not zero. By Girsanov's Theorem A.2 we have

$$\frac{d\nu}{d\nu_0}(v,y) = \exp\Big(-\frac{1}{2}\int_0^T |\nabla F(v)|_\Gamma^2 dt + 2\langle \nabla F(v), dv\rangle\Big).$$

Now we consider the measures found by conditioning v on y. Under ν_0 the random variables v and y are now dependent. The Gaussian measure $\mu_0^y := \nu_0(dv|y)$ does therefore depend on y in this case but only via its mean, not its covariance. An explicit expression for the covariance and the mean of μ_0^y is given in [HSVW05, Theorem 4.1]. Now let $\mu^y(dv)$ denote the measure on $C([0,T],\mathbb{R}^n)$ given by $\nu(dv|y)$.

By applying Theorem A.1, and then integrating by parts (Itô formula) as we did in the previous section, we find that

$$\frac{d\mu^y}{d\mu_0^y}(v) \propto \exp\Big(-\Phi(v)\Big) \tag{3.32a}$$
$$\Phi(v) = \frac{1}{2}\int_0^T \Big(|\Gamma\nabla F(v)|_\Gamma^2 dt - \mathrm{Tr}(\Gamma D^2 F(v))\Big)dt$$
$$+ F(v(T)) - F(v(0))\;. \tag{3.32b}$$

Note that $\Phi(v)$ does *not* depend on y. In this particular case, the y-dependence comes entirely from the *reference measure.*

Theorem 3.19 *Assume that* $F \in C^3(\mathbb{R}^n,\mathbb{R}^+)$ *with globally bounded first, second and third derivatives. Let* $E = C([0,T],\mathbb{R}^n)$*. Then* $\mu^y \ll \mu_0^y$ *with* $d\mu^y/d\mu_0^y \propto \exp(-\Phi)$ *where* Φ *is given by* (3.32b)*. The map* Φ *satisfies Assumptions 2.1(1)–(3) and* $y \mapsto \mu^y$ *is locally Lipschitz continuous from* $C([0,T],\mathbb{R}^\ell)$ *into the space of probability measures on* E *endowed with the total variation metric.*

Proof. Satisfaction of parts (1)–(3) of Assumptions 2.1 follow from the definition (3.32b) of $\Phi(v)$.

The covariance operator of μ_0^y, which we denote by $\mathcal{C}_0$, does not depend on y, and is the resolvent of a second order differential operator. Thus the Cameron-Martin space for the reference measure is $H^1(0,T;\mathbb{R}^\ell)$. It follows from the results in section 4 of [HSVW05] that, for almost every observation y, the mean of the Kalman-Bucy smoother belongs to the Sobolev space $H^{3/2-\varepsilon}$ for any $\varepsilon > 0$. Even better, the map $y \mapsto m$ is continuous from $C([0,T],\mathbb{R}^\ell)$ into H^1, so that there exists a constant C satisfying $|m - m'|_{\mathcal{C}_0} \le C\|y - y'\|_{L^\infty}$.

Hence μ_0^y and $\mu_0^{y'}$ are equivalent Gaussian measures denoted $\mathcal{N}(m,\mathcal{C}_0)$ and $\mathcal{N}(m',\mathcal{C}_0)$. From this it follows that

$$\begin{aligned}
\|\mu^y - \mu^{y'}\|_{\mathrm{TV}} &= \int\Big|\frac{d\mu^y}{d\mu_0^y}(v) - \frac{d\mu^{y'}}{d\mu_0^{y'}}(v)\frac{d\mu_0^{y'}}{d\mu_0^y}(v)\Big| d\mu_0^y(v) \\
&\le \int \exp(-\Phi(v))\Big|1 - \frac{d\mu_0^{y'}}{d\mu_0^y}(v)\Big| d\mu_0^y(v) \\
&\le \int \exp(\varepsilon\|v\|_E^2 - M)\Big|1 - \frac{d\mu_0^{y'}}{d\mu_0^y}(v)\Big| d\mu_0^y(v) \\
&\le \Big(\int \exp(2\varepsilon\|v\|_E^2 - 2M) d\mu_0^y(v)\Big)^{\frac12}\Big(\int\Big|1 - \frac{d\mu_0^{y'}}{d\mu_0^y}(v)\Big|^2 d\mu_0^y(v)\Big)^{\frac12} \\
&\le C\Big(\int\Big(\frac{d\mu_0^{y'}}{d\mu_0^y}(v)\Big)^2 d\mu_0^y(v) - 1\Big)^{\frac12} .
\end{aligned}$$

On the other hand, one has the identity

$$\frac{d\mu_0^{y'}}{d\mu_0^y} = \exp(\langle \mathcal{C}_0^{-\frac12}(m'-m), \mathcal{C}_0^{-\frac12}(v-m)\rangle - \frac12|\mathcal{C}_0^{-\frac12}(m'-m)|^2) ,$$

so that

$$\begin{aligned}
\int\Big(\frac{d\mu_0^{y'}}{d\mu_0^y}(v)\Big)^2 d\mu_0^y(v) &= \int \exp\Big(2\langle m'-m, v-m\rangle_{\mathcal{C}_0} - |m'-m|_{\mathcal{C}_0}^2\Big) d\mu_0^y(v) \\
&= \exp(|m-m'|_{\mathcal{C}_0}^2)\int \exp\Big(\langle 2(m'-m), v-m\rangle_{\mathcal{C}_0} - \frac12|2(m'-m)|_{\mathcal{C}_0}^2\Big) d\mu_0^y(v) \\
&= \exp(|m-m'|_{\mathcal{C}_0}^2) .
\end{aligned}$$

It follows that

$$\|\mu^y - \mu^{y'}\|_{\mathrm{TV}} \le C\Big(\exp(|m-m'|_{\mathcal{C}_0}^2) - 1\Big)^{\frac12} \le C\Big(\exp(C\|y-y'\|_{L^\infty}^2) - 1\Big)^{\frac12} ,$$

and the desired result follows. □

Theorem 3.20 *Let $k \ge 1$. Assume that, further to satisfying the assumptions of Theorem 3.19, $F \in C^{k+2}(\mathbb{R}^n,\mathbb{R}^+)$ Then the potential Φ given by* (3.32b) *is $C^k(\mathbb{R}^n,\mathbb{R})$.*

4 Theme B. Langevin Equations

In this section we construct S(P)DEs which are reversible with respect to the measure μ^y introduced in section 3. These equations are interesting in their own right; they also form the basis of efficient Metropolis-Hastings methods for sampling μ^y, the topic of section 5. In this context and in the finite dimensional case, the SDEs are often referred to as *Langevin* equations in the statistics and statistical physics literature [RC99] and we will use this terminology.

For economy of notation, in this and the next section, we drop explicit reference to the data y and consider the measure

$$\frac{d\mu}{d\mu_0}(x) = Z^{-1}\exp(-\Phi(x)) , \tag{4.1}$$

for some potential function $\Phi(\cdot)$ and normalisation constant Z. We will assume that the reference measure μ_0 is a entred Gaussian measure with covariance operator $\mathcal{C}_0 \colon E^* \to E$ on some separable Banach space E. Note that this includes the case where μ_0 is not centred, provided that its mean m_0 belongs to the Cameron-Martin space. We may then simply shift coordinates so that the new reference measure has mean zero, and change the potential to $\Phi^m(x) := \Phi(m_0 + x)$.

Hence in this section we simply work with (1.1), assume that $\mu_0 = \mathcal{N}(0, \mathcal{C}_0)$ and that the four conditions on Φ hold as stated in Assumptions 2.1. We furthermore use the notation $\mathcal{L}$ to denote the inverse $\mathcal{C}_0^{-1}$ of the covariance, sometimes called the 'precision operator'. Note that while $\mathcal{C}_0$ is always a bounded operator, $\mathcal{L}$ is usually an unbounded operator. Additional assumptions on the structure of Φ, the covariance $\mathcal{C}_0$ and the space E will be stated when required.

4.1 The finite-dimensional case

Consider the finite dimensional probability measures on the initial condition u for (3.1) which we have constructed in sections 3.1, 3.2 and 3.3. Recall that the posterior measure is μ given by (4.1). By use of the Fokker-Planck equation it is straightforward to check that, for every strictly positive-definite symmetric matrix $\mathcal{A}$ the following SDE is μ^y invariant

$$dx = -\mathcal{A}\mathcal{L}x\,dt - \mathcal{A}D\Phi(x)\,dt + \sqrt{2\mathcal{A}}\,dW(t)\;. \tag{4.2}$$

Actually one has even more: the Markov semigroup generated by (4.2) consists of operators that are selfadjoint in $L^2(\mathbb{R}^n, \mu^y)$. One of the most general theorems covering this situation is given by [Che73, Li92]:

Theorem 4.1 *Let $\Phi(\cdot)$ belong to $C^2(\mathbb{R}^n)$ and be such that* $\exp(-\Phi)$ *is integrable with respect to the symmetric Gaussian measure μ_0 with covariance $\mathcal{C}_0 = \mathcal{L}^{-1}$. Then,* (4.2) *has a unique global strong solution which admits μ^y as an invariant measure. Furthermore, the semigroup*

$$\mathcal{P}_t\varphi(x) = \mathbb{E}(\varphi(x(t))\,:\,x(0) = x)$$

can be extended to a semigroup consisting of selfadjoint contraction operators on $L^2(E, \mu^y)$.

One approach to approximately sampling from μ given by (4.1) is thus to solve this equation numerically and to rely on ergodicity of the numerical method, as well as approximate preservation of μ under discretization, to obtain samples from μ, see [Tal90]. Typical numerical methods will require draws from $\mathcal{N}(0, \mathcal{A})$ in order to simulate (4.2). In low dimensions, a good choice for $\mathcal{A}$ is $\mathcal{A} = \mathcal{C}_0$ as this equalizes the convergence rates to equilibrium in the case $\Phi \equiv 0$, $\mathcal{C}_0$ is cheap to calculate, and draws from $\mu_0 = \mathcal{N}(0, \mathcal{C}_0)$ are easily made. We refer to this as 'preconditioning'.

For a given accuracy, there will in general be an optimal stepsize that provides a suitable approximation to the desired i.i.d. sequence at minimal computational cost. Too large stepsizes will result in an inaccurate approximation to (4.2), whereas small stepsizes will require many steps before approximate independence is achieved.

4.2 The Infinite Dimensional Case

The problems in sections 3.4, 3.5 and 3.6 give rise to measures on infinite dimensional spaces. The infinite dimensional prior reference measure involves either a stationary OU process or Wiener measure. Thus draws from μ_0 (or rather from an approximation thereof) are relatively straightforward to make. Furthermore, in a number of situations, the precision operator $\mathcal{L} = \mathcal{C}_0^{-1}$ is readily characterized as a second order differential operator, see for example [HSVW05].

Even though there exists no infinite-dimensional analogue of Lebesgue measure, we have seen in the previous section that it happens in many situations that the posterior μ^y possesses a density with respect to some fixed Gaussian measure μ_0. It is therefore tempting to carry over (4.2) *mutatis mutandis* to the infinite-dimensional case. It is however much less clear in general what classes of drifts result in (4.2) being a well-posed stochastic PDE (or infinite-dimensional SDE) and, if it is well-posed, whether μ^y is indeed an invariant measure for it. The remainder of this section is devoted to a survey of some rigorous results that have been obtained in this direction.

Remark 4.2 In principle, some of these questions could be answered by invoking the theory of symmetric Dirichlet forms, as described in [RM92] or [FOT94]. However, we stay away from this course for two reasons. First, it involves a heavy technical machinery that does not seem to be justified in our case since the resulting processes are not that difficult to understand. Second, and more importantly, while the theory of Dirichlet forms allows to 'easily' construct a large family of μ^y-reversible processes (that contains as special cases the SDE's described in (4.2)), it is more difficult to characterize them as solutions to particular SDEs or SPDEs. Therefore, if we wish to approximate them numerically, we are back to the kind of analysis performed here.

We are going to start with a survey of the results obtained for the Gaussian case (that is when Φ vanishes or is itself quadratic in the x variable), before turning to the nonlinear case.

4.2.1 The Gaussian Case

In this section, we consider the situation of a Gaussian measure μ with covariance operator $\mathcal{C}_0$ and mean m on a separable Hilbert space $\mathcal{H}$. At a formal level, the 'density' of μ with respect to the (non-existent, of course) Lebesgue measure is proportional to

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

so that one would expect the evolution equation

$$dx = \mathcal{L}m\,dt - \mathcal{L}x\,dt + \sqrt{2}\,dW(t)\;, \tag{4.3}$$

where, recall, we set $\mathcal{L} = \mathcal{C}_0^{-1}$, to have μ as its invariant measure. Since, if $\mathcal{H}$ is infinite-dimensional, $\mathcal{L}$ is always an unbounded operator, it is not clear *a priori* how to interpret solutions to (4.3). The traditional way of interpreting (4.3) is to solve it by the variation of constants formula and to *define* the solution to (4.3) as being the process given by

$$x(t) = S(t)x_0 + (1 - S(t))m + \sqrt{2}\int_0^t S(t-s)\,dW(s)\;,$$

(here $S(t)$ denotes the semigroup on $\mathcal{H}$ generated by $-\mathcal{L}$; see [Hen81], [Paz83] and [Rob01] for background on semigroups) provided that the stochastic integral appearing on the right hand side takes values in $\mathcal{H}$.

This turns out to be always the case in the situation at hand. Furthermore, one has the stronger statement that this process is also the unique weak solution to (4.3) in the sense that it is the only $\mathcal{H}$-valued process such that the identity

$$d\langle x(t), h\rangle = \langle \mathcal{L}h, m - x(t)\rangle\,dt + \sqrt{2}\langle h, dW(t)\rangle\;, \tag{4.4}$$

holds for every h in the domain of $\mathcal{L}$. Combining the results from [IMM$^+$90] and [DPZ92], one obtains that:

Lemma 4.3 *Let $\mathcal{L}$ and μ be as above. Then the evolution equation* (4.3) *has continuous $\mathcal{H}$-valued mild solutions. Furthermore, it has μ as its unique invariant measure and there exists a constant K such that for every initial condition $x_0 \in \mathcal{H}$ one has*

$$\|\mathrm{Law}(x(t)) - \mu\|_{\mathrm{TV}} \le K\,(1 + \|x_0 - m\|_{\mathcal{H}})\exp(-\|\mathcal{C}_0\|^{-1}_{\mathcal{H}\to\mathcal{H}}t),$$

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

Remark 4.4 The convergence in total variation obtained in Lemma 4.3 is very strong and does *not* hold in general if one replaces (4.3) by its 'preconditioned' version as in (4.2). For example, in the particular case

$$dx = m\,dt - x\,dt + \sqrt{2\mathcal{C}_0}\,dW(t)\;, \tag{4.5}$$

it is known that convergence in total variation does not hold, unless x_0 belongs to the Cameron-Martin space of μ, that is unless $\|\mathcal{L}^{1/2}x_0\| < \infty$. However, one does still have convergence of arbitrary solutions to (4.5) to μ in the p-Wasserstein distance for arbitrary p.

4.2.2 The Nonlinear Case

This case is much less straightforward than the Gaussian case and we will not give a complete treatment here. One problem that often occurs in the infinite-dimensional case is that Φ is naturally defined on a Banach space E (typically the space of continuous functions) rather than on a Hilbert space $\mathcal{H}$. It is then tempting to work with a scale of spaces

$$E \hookrightarrow \mathcal{H} = \mathcal{H}^* \hookrightarrow E^*\;,$$

where all inclusions are dense. We are going to make the following assumptions for the precision operator $\mathcal{L}$ of our reference Gaussian measure μ_0:

(A1) The semigroup $S(t) = e^{-\mathcal{L}t}$ generated by $\mathcal{L}$ on $\mathcal{H}$ can be restricted to a strongly continuous semigroup of contraction operators on E.

(A2) There exists $\alpha \in (0, 1/2)$ such that $\mathcal{D}(\mathcal{L}^\alpha) \subset E$ (densely), $\mathcal{L}^{-2\alpha}$ is trace class in $\mathcal{H}$, and the measure $\mathcal{N}(0, \mathcal{L}^{-2\alpha})$ is concentrated on E.

The first assumption ensures that E is a 'good choice' of a space to work with. The second assumption is a slight strengthening of the statement that $\mu_0(E) = 1$, since the statement with $\alpha = \frac{1}{2}$ is implied by $\mu_0(E) = 1$. Regarding the density $\Phi\colon E \to \mathbb{R}$, we make the following assumptions:

(A3) For every $\varepsilon > 0$, there exists $M > 0$ such that $\Phi(x) \geq M - \varepsilon\|x\|_E^2$ for every $x \in E$.

(A4) The function $\Phi\colon E \to \mathbb{R}$ is twice Fréchet differentiable and its derivatives are polynomially bounded.

(A5) There exists a sequence of Fréchet differentiable functions $F_n\colon E \to E$ such that

$$\lim_{n\to\infty} \|\mathcal{L}^{-\alpha}(F_n(x) - D\Phi(x))\|_{\mathcal{H}} = 0$$

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

$$\langle x^*, F_n(x + y)\rangle \leq -\gamma\|x\|_E \tag{4.6}$$

holds for every $x^* \in \partial\|x\|_E \subset E^*$ and every $x, y \in E$ with $\|x\|_E \geq C(1 + \|y\|_E)^N$. Here, $\partial\|x\|_E$ denotes the subdifferential of the norm at x (see for example [DPZ92]).

Assumption (A3) just makes sure that $\exp(-\Phi)$ is integrable with respect to μ_0. The next assumption (A4) provides a minimum of regularity so that the equation

$$dx = -\mathcal{L}x\,dt - D\Phi(x)\,dt + \sqrt{2}\,dW(t)\;, \tag{4.7}$$

is well-posed (in its mild formulation). The last condition seems rather complicated, but it should just be thought of as a version of the dissipativity condition

$$\langle x^*, D\Phi(x + y)\rangle \leq -\gamma\|x\|_E$$

that survives approximating $D\Phi$ by E-valued functions. With these conditions at hand, we have the following result from [HSV07]:

Theorem 4.5 *Assume that conditions (A1)–(A5) hold and define the probability measure*

$$\mu(dx) = Z^{-1}\exp(-\Phi(x))\,\mu_0(dx)\;,$$

for a suitable normalisation constant Z. Then the stochastic PDE (4.7) *has a unique continuous E-valued global mild solution for every initial condition $x_0 \in E$. Furthermore, this solution admits μ as its unique invariant probability measure.*

Under a very weak additional assumption (essentially, Φ should admit approximations that have bounded support and that are Fréchet differentiable, which is not completely automatic if the norm on E is not differentiable), it is again possible to show that transition probabilities converge to the invariant measure at exponential speed and that the law of large numbers holds.

5 Theme C. MCMC Methods

In this section we describe a range of effective Metropolis-Hastings based (see [Has70, MRTT53]) MCMC methods (see [Liu01, RC99]) for sampling the target distributions constructed in Section 3. As in the previous section we drop explicit reference to the data y and work with a posterior distribution μ given by (4.1). The methods we describe are motivated by the μ-reversible stochastic evolution equations derived in the previous section. We work with measures μ given by (1.1). We assume that $m_0 = 0$ which can always be achieved by a shift of origin, provided the mean of μ_0 belongs to its Cameron-Martin space. Our aim is to draw samples from the measure μ on E given by (4.1).

The idea of MCMC methods for target μ is to construct a discrete time Markov chain $\{x_n\}$ on E that has μ as its invariant measure and that has good mixing properties. One can then take as an approximation to i.i.d. samples the sequence $k \mapsto x_{N_0+kN_1}$ with $k \geq 0$ and N_0, N_1 'sufficiently large'. In order to compute integrals of the form $I = \int f(x)\,\mu(dx)$ for some test function f, one can then use the fact that, by Birkhoff's ergodic theorem, one has the almost sure identity

$$I = \lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} f(x_k)\;.$$

Metropolis-Hastings methods work by proposing a move from the current state x_k to y from a Markov transition kernel on E, and then accepting or rejecting in a fashion which ensures that the resulting composite Markov chain is μ reversible.

In section 5.1 we first explain the idea in finite dimensions, with application to the problems formulated in sections 3.1–3.3 and, of course, to finite dimensional approximation of the problems formulated in sections 3.4–3.6. We focus on the theory related to *random walk* and *Langevin* proposals for these problems, building on the material in the previous section. Then, in section 5.2, we generalize these methods to the infinite dimensional setting.

5.1 Metropolis-Hastings in Finite Dimensions

In finite dimensions the measure μ given by (1.1) on $E = \mathbb{R}^d$ has density π with respect to Lebesque measure which is given by

$$\pi(x) \propto \exp\Big(-\frac{1}{2}|x|_{\mathcal{C}_0}^2 - \Phi(x)\Big).$$

The Metropolis-Hastings method of constructing a μ reversible Markov chain is the following. Fix a Markov transition kernel $P(x,dy)$ with density $q(x,y)dy$. The measure $\mu(dx)P(x,dy)$ on $E \times E$ then has density $\pi(x)q(x,y)$. Define the function

$$a(x,y) = 1 \wedge \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\;,$$

where we write $a \wedge b$ for the minimum between a and b.

Assume that x_k is known (start for example with $x_0 = 0$). To determine x_{k+1} draw $y \sim q(x_k, y)$ and proceed as follows:

1. $x_{k+1} = y$ (step accepted) with probability $a(x_k, y)$.
2. $x_{k+1} = x_k$ (step rejected) otherwise.

The resulting Markov chain is π-reversible.

A frequently used class of methods are the *symmetric random walk* proposals where

$$y = x_k + \sqrt{2\Delta t}\xi_k \tag{5.1}$$

where the ξ_k are i.i.d. symmetric random variables (for example $\mathcal{N}(0,I)$) on $\mathbb{R}^d$. This may be viewed as a discretization of the Brownian motion

$$dx = \sqrt{2}dW.$$

As such the proposal contains no information about the target However, by symmetry,

$$a(x,y) = 1 \wedge \frac{\pi(y)}{\pi(x)}.$$

This has the advantage of being simple to implement.

Typically, the mixing time for a Metropolis-Hastings chain will depend both on the proportion of steps that are rejected and on the variance of $y - x_k$: the higher the number of rejections and the lower the variance, the longer it will take for the Markov chain to explore the whole state space. In general, there is a competition between both effects: steps with a large variance have a high probability of rejections; on the other hand small moves are less likely to be rejected but explore the state space slowly. Roughly speaking the competition between these two effect is measured by the mean square jumping distance of the Markov chain in stationarity. Specifically, if we define

$$S_{d,i} = \mathbb{E}^{\mu}|x_{k+1,i} - x_{k,i}|^2 \tag{5.2}$$

then this quantity measures the mean square jumping distance in the i^{th} component of the vector x_k. Maximizing this quantity will enhance the mixing of functionals heavily dependent on the i^{th} component of $x \sim \mu$. We will optimize algorithms according to this criterion.

In an attempt to maximize (5.2), proposals which contain information about the target distribution can be useful. A class of proposals which does contain such information arises from discretizing the Langevin SDE (4.2). A linearly implicit Euler discretization gives rise to the following family of proposals:

$$y - x_k = -\Delta t\mathcal{A}\mathcal{L}\Big(\theta y + (1-\theta)x_k\Big) - \alpha\Delta t\mathcal{A}D\Phi(x_k) + \sqrt{2\Delta t\mathcal{A}}\xi_k \tag{5.3}$$

where the ξ_k are i.i.d $\mathcal{N}(0, I)$ random variables on $\mathbb{R}^d$, $\theta \in [0, 1]$ and $\alpha \in \{0, 1\}$. If $\alpha = 0$ the proposal contains information only about the reference measure μ_0, via its precision operator $\mathcal{L}$. If $\alpha = 1$ it contains information information about μ itself. Two natural choices for $\mathcal{A}$ are I and $\mathcal{C}_0$.

The formula for the proposal rearranges to give

$$y = \Big(I + \Delta t\theta\mathcal{A}\mathcal{L}\Big)^{-1}\Big((I - \Delta t(1-\theta)\mathcal{A}\mathcal{L})x_k - \alpha\Delta t\mathcal{A}D\Phi(x) + \sqrt{2\Delta t\mathcal{A}}\xi_k\Big). \tag{5.4}$$

In the case $\alpha = 0$ this generalizes the symmetric Random Walk to allow y to be a more complex linear combination of x_k and ξ_k. When $\alpha = 1$ the proposal also contains information which tends to make proposals which decrease Φ. Roughly speaking we expect proposals with $\alpha = 1$ to explore the state space more rapidly than those with $\alpha = 0$. However there is a cost involved in evaluating $D\Phi$ and the trade-off between cost-per-step and number of steps will be different for different problems.

A natural question of interest for these algorithms is how to choose the time-step Δt. We now study this question in the limit where the state space dimension $d \to \infty$. We define the norm

$$|x|_s = \Big(\sum_{i=i}^{d} i^{2s}x_i^2\Big)^{\frac{1}{2}}.$$

We make the following assumptions:

Assumption 5.1 *The following hold for the family of reference measures $\mu_0 = \mu_0(d)$, the family of target measures $\mu = \mu(d)$ and their interrelations.*

1. *There are constants $c^{\pm} \in (0, \infty)$ such that the eigenvalues $\lambda_{i,d}^2$ of the covariance $\mathcal{C}_0$ satisfy*

$$c^- i^{-k} \le \lambda_{i,d} \le c^+ i^{-k} \quad \forall 1 \le i \le d. \tag{5.5}$$

2. *Assumptions 2.1 (1)–(3) hold, generalized to include the case $\varepsilon = 0$, with $E = (\mathbb{R}^d, |\cdot|_s)$, $s < k - \frac{1}{2}$ and with constants independent of dimension d.*

We now state three theorems, all proved in [BRS], which quantify the efficiency of the various proposals described above in the high dimensional setting, and under the preceding assumptions.

The first theorem shows that, for the symmetric random walk the optimal choice of Δt is of $\mathcal{O}(d^{-(2k+1)})$, giving rise to a maximal mean square jump of the same magnitude.

Theorem 5.2 *Consider the symmetric random walk proposal* (5.1). *Assume that $\Delta t = \ell^2 d^{-\rho}$. Then the following dichotomy holds, for any fixed i:*

- *If $\rho \ge 2k + 1$ then*

$$\liminf_{d\to\infty} d^{\rho} S_{d,i} > 0, \quad \limsup_{d\to\infty} d^{\rho} S_{d,i} < \infty.$$

- *If* $\rho < 2k+1$ *then*
$$\limsup_{d\to\infty} d^q S_{d,i} = 0$$
for any $q \geq 0$.

The next theorem shows that, for the basic version of the Langevin proposal, the optimal choice of Δt is of $\mathcal{O}(d^{-(2k+1/3)})$, giving rise to a maximal mean square jump of the same magnitude. The improvement in the exponent by 2/3 comes as the price of evaluating the application of the precision operator $\mathcal{L}$ at each step; the cost of doing this will be problem dependent.

Theorem 5.3 *Consider the proposal* (5.4) *with* $\theta = \alpha = 0$ *and* $\mathcal{A} = I$. *Assume that* $\Delta t = \ell^2 d^{-\rho}$. *Then the following dichotomy holds, for any fixed* i*:*

- *If* $\rho \geq 2k + \frac{1}{3}$ *then*
$$\liminf_{d\to\infty} d^\rho S_{d,i} > 0, \quad \limsup_{d\to\infty} d^\rho S_{d,i} < \infty.$$
- *If* $\rho < 2k + \frac{1}{3}$ *then*
$$\limsup_{d\to\infty} d^q S_{d,i} = 0$$
for any $q \geq 0$.

The final theorem shows that, at the cost of making samples $\sqrt{\mathcal{C}_0}\xi_k$ from the prior measure at each step of the algorithm, the optimal choice of $\Delta t = \mathcal{O}(d^{-1/3})$, gives rise to a maximal mean square jump of the same magnitude.

Theorem 5.4 *Consider the proposal* (5.4) *with* $\theta = \alpha = 0$ *and* $\mathcal{A} = \mathcal{C}$. *Assume that* $\Delta t = \ell^2 d^{-\rho}$. *Then the following dichotomy holds, for any fixed* i*:*

- *If* $\rho \geq \frac{1}{3}$ *then*
$$\liminf_{d\to\infty} d^\rho S_{d,i} > 0, \quad \limsup_{d\to\infty} d^\rho S_{d,i} < \infty.$$
- *If* $\rho < \frac{1}{3}$ *then*
$$\limsup_{d\to\infty} d^q S_{d,i} = 0$$
for any $q \geq 0$.

The previous two theorems, concerning proposals of the form (5.4), concern only the cases where $\theta = \alpha = 0$. It is expected that the scaling results will be identical for $\theta = 0, \alpha = 1$. However the choice of θ can make significant differences. In the next section we show how, by working in infinite dimensions, we can in some cases eliminate dimension dependence in Metropolis Hastings algorithms, by choosing $\theta = \frac{1}{2}$.

5.2 Metropolis-Hastings in Infinite Dimensions

The ideas of the previous section can be generalized to infinite dimensions as follows [Tie98]. Assume that we are given a Polish (i.e. complete, separable, metric) space E (since we want to allow for the possibility of sampling from a measure on a space of paths, E should be thought of as a space of functions in general) and a probability measure μ on E. Assume furthermore that we are given a Markov transition kernel P over E with the property that the measures $\mu(dx)P(x,dy)$ and $\mu(dy)P(y,dx)$ are equivalent so that the quantity
$$\frac{\mu(dy)P(y,dx)}{\mu(dx)P(x,dy)},$$
which should be interpreted as the Radon-Nikodym derivative of the two aforementioned measures evaluated at the point (x,y), is well-defined. With these notations in place, we can construct a new Markov chain in the following way. Assume again that x_k is known and draw a random sample y from the probability distribution $P(x_k,\cdot)$. Now let
$$\alpha(x,y) = 1 \wedge \frac{\mu(dy)P(y,dx)}{\mu(dx)P(x,dy)}. \tag{5.6}$$
The algorithm proceeds as follows

1. $x_{k+1} = y$ (step accepted) with probability $\alpha(x_k, y)$.
2. $x_{k+1} = x_k$ (step rejected) otherwise.

If we denote by Q the transition probabilities of the process x_n, it can be checked that one has

$$Q(x,dy) = c(x)\delta_x(dy) + P(x,dy) \wedge \frac{P(y,dx)}{\mu(dx)}\,\mu(dy)\;, \tag{5.7}$$

for some function c that makes Q a Markov transition kernel. If we define a map $\Delta\colon E \to E^2$ by $\Delta(x) = (x,x)$ and denote by $\Delta^*\mu$ the push-forward of μ by Δ, one can check that (5.7) implies that

$$\mu(dx)Q(x,dy) = \sqrt{c(x)c(y)}(\Delta^*\mu)(dx,dy) + \mu(dx)P(x,dy) \wedge \mu(dy)P(y,dx)\;.$$

This expression is symmetric in $x \leftrightarrow y$, so that the Markov kernel Q (or equivalently the Markov chain generated from it) is reversible with respect to the measure μ. In particular, the measure μ is invariant for Q.

Thus key to making this idea work is the construction of proposals for which the measure $\mu(dy)P(y,dx)$ is absolutely continuous with respect to the measure $\mu(dx)P(x,dy)$. We consider this question in the context of (5.4), basing ideas on the paper [BRSV08]. Let $P_\alpha(x,dy)$ denote the transition kernel of this proposal. Then define measures η and η_0 by

$$\eta(dx,dy) = \mu(dx)P_\alpha(x,dy)$$

and

$$\eta_0(dx,dy) = \mu_0(dx)P_0(x,dy).$$

It is straightforward to show that $\eta_0(dx,dy) = \eta_0(dy,dx)$ iff $\theta = \frac{1}{2}$. We work with this assumption henceforth as it enables us to define the MCMC method on function space.

Using the fact that $P_\alpha(x,\cdot)$ is absolutely continuous with respect to $P_0(x,\cdot)$ for both $\alpha = 0$ and $\alpha = 1$ we deduce that η is absolutely continuous with respect to η_0 and that, for some $\rho(x,y) = \rho(x,y;\alpha,\mathcal{A})$, we have

$$\frac{d\eta}{d\eta_0}(x,y) \propto \exp(-\rho(x,y)).$$

Thus the acceptance probability for the Metropolis algorithm is

$$a(x,y) = 1 \wedge \exp(\rho(x,y) - \rho(y,x)). \tag{5.8}$$

For the proposals (5.4) the function $\rho(x,y)$ is given, up to an additive constant which we ignore, by the following expressions:

- for $\mathcal{A} = I$ we have

$$\begin{aligned}\rho(x,y) = \Phi(x) &+ \frac{\alpha^2\Delta t}{4}\|D\Phi(x)\|^2 + \frac{\alpha}{2}\langle D\Phi(x), y-x\rangle \\ &+ \frac{\alpha\Delta t}{4}\langle D\Phi(x), \mathcal{C}_0^{-1}(y+x)\rangle;\end{aligned}$$

- for $\mathcal{A} = \mathcal{C}_0$ we have

$$\begin{aligned}\rho(x,y) = \Phi(x) &+ \frac{\alpha^2\Delta t}{4}\|\mathcal{C}_0^{1/2}D\Phi(x)\|^2 + \frac{\alpha}{2}\langle D\Phi(x), y-x\rangle \\ &+ \frac{\alpha\Delta t}{4}\langle D\Phi(x), y+x\rangle.\end{aligned}$$

The four algorithms defined in this section (two choices for both $\alpha \in \{0,1\}$ and $\mathcal{A} \in \{I, \mathcal{C}_0\}$) all lead to well-defined Metropolis-Hastings chains on Banach space. Thus they give rise to mean square jumping distances which are bounded independently of dimension d as they are, in particular, non-zero in the infinite dimensional case.

It is straightforward to prove [BS09] that, for any $c > 0$, the acceptance probability (5.8) satisfies

$$\mathbb{E}a(x_k,y) \geq \exp(-c)\Big(1 - \frac{\mathbb{E}|\rho(x_k,y) - \rho(y,x_k)|}{c}\Big).$$

Thus if we can show that

$$\mathbb{E}|\rho(x_k,y) - \rho(y,x_k)| \to 0$$

as $\Delta t \to 0$ then we deduce that we can make the acceptance probability arbitarily close to 1. In the case $\alpha = 0$, proving this may be shown by using Assumption 2.1.

6 Discussion and Bibliography

There are several useful sources for background material relevant to both the problems studied, and methods developed, in this chapter. A general reference concerned with stochastic modelling is [CH06]. Several technical tools are required to develop the methods described in this chapter. An exhaustive treatment of Gaussian measures can be found in [Bog98] and moment bounds for SDEs can be found in [Mao97]. The book [DdFG01] is an excellent source for material concerned with sequential filtering problems, including the use of particle filters for non-Gaussian problems. The filtering and smoothing problems for SDEs with continuous time observations, as arising in section 3.6, is introduced in [Øks03], and developed in detail in the Gaussian context (f and g linear) giving rise to the Kalman-Bucy filter and smoother. This method uses an approach based on first filtering ($0 \to T$), and then reversing the process ($T \to 0$) to incorporate data from time $t > s$ into the probability distribution at me s. Good sources of signal processing problems arising from data assimilation are [Eve06] and the volume [JI07]; these problems have motivated a lot of our research in this general area. Finally note that signal processing may be viewed as an inverse problem to find a signal from partial, noisy, observations. The Bayesian approach to inverse problems in general is discussed in [KS05].

In Theme A we considered a range of differing problems arising in signal processing, constructing and deriving properties of the posterior distribution. The posterior distributions constructed in sections 3.2 and 3.3 can both be viewed as *parameter estimation* problems for SDEs. They have particular structure, inherited from the way in which the parameter u enters the expression $\varphi^t(u)$ appearing in the SDE for y. The general subject of parameter estimation for SDEs is considered in [BPR80, Kut04]. Incorporating discrete time data into a continuous time model, as undertaken in sections 3.4 and 3.5, is studied in [AHSV07, HSV09]. Carrying out this program and, at the same time estimating parameters in the dynamical model, is discussed in [RHCC07], in a non-Bayesian setting. The relationship between the coloured noise model appearing in section 3.4, and the white noise model appearing in 3.5, in the limit $\delta \to 0$, is part of the theory of homogenization for stochastic processes; see [BLP78, PS08]. The filtering and smoothing problems for SDEs with continuous time observations, as arising in section 3.6, is introduced in [Øks03], as mentioned above. In the Gaussian case the mean is characterized by the solution of a two point value problem, defined through inversion of the precision operator. The approach to smoothing outlined in [Øks03] corresponds to a continuous time analogue of LU factorization, here for the inverse of the covariance operator, facilitating its action on the data to compute the mean. The particular formulation of the smoothing problem described here is developed in [AHSV07].

In Theme B we studied the derivation of Langevin equations (stochastic partial differential equations) which are invariant with respect to a given invariant measure. This is straightforward in finite dimensions, but is an emerging subject area in infinite dimensions. The idea is developed in a fairly general setting in [HSV07], building on the Gaussian case described in [HSVW05]. The first use of the Langevin equation to solve signal processing problems may be found in [SVW04] and further applications may be found in [HSV, AJSV08]. On the theoretical side many open questions remain concerning the derivation of Langevin equations. In particular the paper [HSV07] deals with elliptic diffusions with gradient drift and additive noise. An initial analysis of a particular hypoelliptic problem may be found in [HSV]. questions relating to the derivation of Langevin equations for nongradient drifts, and for multiplicative noise, remain open.

Theme C is concerned with the design and analysis of effective MCMC methods in high dimension, motivated by the approximation of infinite dimensionsal problems. This subject is overviewed in the review articles [BS09, BS10], and full details of the analysis and application of the methods may be found in [BRS, BRSV08].

Appendix A Some Results from Probability

In this appendix, we collect miscellaneous results from stochastic analysis that were used in this chapter. Throughout the chapter we use the following notation: given a Hilbert space $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$, for any positive-definite C we define the second inner-product and norm

$$\langle a, b\rangle_C = \langle a, C^{-1}b\rangle, \quad \|a\|_C^2 = \langle a, a\rangle_C.$$

A.1 Conditional Probabilities

Throughout Theme A of this paper we will be generalizing Bayes formula to an infinite dimensional setting. There are two components to this: Bayes formula in finite dimensions, and then the generalization to the Hilbert space setting. We start in finite dimensions. Assume that we are given a random variable u on $\mathbb{R}^d$ about which we have some *prior* information in terms of a probability distribution $\mathbf{P}(u)$. Imagine that we now define a random variable y on $\mathbb{R}^l$, which depends upon u, and for which we have the probability distribution of y given u, namely $\mathbf{P}(y|u)$. By the elementary rules of probability we have

$$\mathbf{P}(u|y) = \frac{1}{\mathbf{P}(y)}\mathbf{P}(u \cap y),$$
$$\mathbf{P}(y|u) = \frac{1}{\mathbf{P}(u)}\mathbf{P}(u \cap y).$$

Combining these two formulae shows that the *posterior* probability distribution for u, given a single observation of y, is given by Bayes formula

$$\mathbf{P}(u|y) = \frac{1}{\mathbf{P}(y)}\mathbf{P}(y|u)\mathbf{P}(u). \tag{A.1}$$

In this Chapter there are many instances where we are interested in conditioning probability measures on function space. In this context the following theorem will be of central importance in constructing the appropriate generalization of Bayes formula.

Theorem A.1 *Let μ, ν 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 μ has a density φ w.r.t. ν and that the conditional distribution $\nu_{x|y}$ exists. Then the conditional distribution $\mu_{x|y}$ exists and is given by*

$$\frac{d\mu_{x|y}}{d\nu_{x|y}}(x) = \begin{cases} \frac{1}{c(y)}\varphi(x,y), & \text{if } c(y) > 0, \text{ and} \\ 1 & \text{else,} \end{cases} \tag{A.2}$$

with $c(y) = \int_S \varphi(x,y)\, d\nu_{x|y}(x)$ for all $y \in T$.

A.2 A version of Girsanov's theorem

SDEs which do not have the same diffusion coefficient generate measures which are mutually singular on pathspace; the same is true of SDEs starting from different deterministic initial conditions. However, if these two possibilities are ruled out, then two different SDEs generate measures which are absolutely continuous with respect to one another. The Girsanov formula provides an explicit expression for the Radon-Nikodym derivative between two such measures on $\mathcal{H} = L^2([0,T], \mathbb{R}^d)$.

Consider the SDE

$$\frac{dv}{dt} = A(t)v + h(v,t) + \gamma(v,t)\frac{dW}{dt}, \quad v(0) = u. \tag{A.3}$$

and the same equation with the function h set to zero, namely

$$\frac{dv}{dt} = A(t)v + \gamma(v,t)\frac{dW}{dt}, \quad v(0) = v_0. \tag{A.4}$$

The measures generated by these two equations are absolutely continuous. Define $\Gamma(\cdot, t) = \gamma(\cdot, t)\gamma(\cdot, t)^T$. We then have the following version of Girsanov's theorem, that can be found in [Elw82]:

Theorem A.2 *Assume that both equations* (A.3) *and* (A.4) *have solutions on $t \in [0,T]$ which do not explode almost surely. Then the measures μ and μ_0 on $\mathcal{H}$, generated by the two equations* (A.3) *and* (A.4) *respectively, are equivalent with Radon-Nikodym derivative*

$$\frac{d\mu}{d\mu_0}(v) = \exp\Big(-\int_0^T \frac{1}{2}\|h(v,t)\|^2_{\Gamma(v,t)}dt - \langle h(v,t), dv - A(t)v dt\rangle_{\Gamma(v(t))}\Big).$$

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*, 230:50–64, 2007.

[AJSV08] A. Apte, C. K. R. T. Jones, A. M. Stuart, and J. Voss. Data assimilation: Mathematical and statistical perspectives. *International Journal for Numerical Methods in Fluids*, 56:1033–1046, 2008.

[BLP78] A. Bensoussan, J.-L. Lions, and G. Papanicolaou. *Asymptotic Analysis for Periodic Structures*, volume 5 of *Studies in Mathematics and its Applications*. North Holland, 1978.

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

[BPR80] I. V. Basawa and B. L. S. Prakasa Rao. *Statistical Inference for Stochastic Processes*. Academic Press, 1980.

[BRS] A. Beskos, G. O. Roberts, and A. M. Stuart. Scalings for local Metropolis-Hastings chains on non-product targets. To appear in Ann. Appl. Prob.

[BRSV08] A. Beskos, G. O. Roberts, A. M. Stuart, and J. Voss. MCMC methods for diffusion bridges. *Stochastic Dynamics*, 8(3):319–350, Sep 2008.

[BS09] A. Beskos and A. M. Stuart. MCMC methods for sampling function space. In *Proceedings of the International Congress of Industrial and Applied Mathematicians, (Zurich, 2007)*, 2009.

[BS10] A. Beskos and A. M. Stuart. Computational complexity of Metropolis-Hastings methods in high dimensions. In *Proceedings of MCQMC08, 2008*, 2010.

[CDRS] S. L. Cotter, M. Dashti, J. C. Robinson, and A. M. Stuart. Data assimilation problems in fluid mechanics: Bayesian formulation on function space. Submitted to J. Inverse Problems.

[CH06] A. J. Chorin and O. H. Hald. *Stochastic Tools in Mathematics and Science*, volume 1 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer, 2006.

[Che73] P. R. Chernoff. Essential self-adjointness of powers of generators of hyperbolic equations. *J. Functional Analysis*, 12:401–414, 1973.

[DdFG01] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo methods in practice*. Springer, 2001.

[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, 1982.

[Eve06] G. Evensen. *Data Assimilation: The Ensemble Kalman Filter*. Springer, 2006.

[FOT94] M. Fukushima, Y. Oshima, and M. Takeda. *Dirichlet forms and symmetric Markov processes*, volume 19 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co, Berlin, 1994.

[Has70] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[Hen81] Daniel Henry. *Geometric theory of semilinear parabolic equations*, volume 840 of *Lecture Notes in Mathematics*. Springer, 1981.

[HSV] M. Hairer, A. M. Stuart, and J. Voss. Sampling conditioned hypoelliptic diffusions. Manuscript in preparation.

[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*, 17:1657–1706, 2007.

[HSV09] M. Hairer, A. M. Stuart, and J. Voss. Sampling conditioned diffusions. In *Trends in Stochastic Analysis*, volume 353 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, 2009.

[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.

[IMM+90] I. Iscoe, M. B. Marcus, D. McDonald, M. Talagrand, and J. Zinn. Continuity of l^2-valued Ornstein-Uhlenbeck processes. *Ann. Probab.*, 18(1):68–84, 1990.

[JI07] C. K. R. T. Jones and K. Ide, editors. *Data Assimilation*, volume 230 of *Physica D: Nonlinear Phenomena*. Elsevier, 2007.

[KS05] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160 of *Applied Mathematical Sciences*. Springer, 2005.

[Kut04] Yury A. Kutoyants. *Statistical inference for ergodic diffusion processes*. Springer Series in Statistics. Springer, 2004.

[Li92] Xue-Mei Li. *Stochastic Flows on Noncompact Manifolds*. PhD thesis, University of Warwick, 1992.

[Liu01] Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer, 2001.

[Mao97] Xuerong Mao. *Stochastic differential equations and their applications*. Horwood Publishing Series in Mathematics & Applications. Horwood Publishing Limited, Chichester, 1997.

[MRTT53] N. Metropolis, A. W. Rosenbluth, M. N. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[Øks03] Bernt Øksendal. *Stochastic differential equations*. Universitext. Springer, sixth edition, 2003. An introduction with applications.

[Paz83] A. Pazy. *Semigroups of linear operators and applications to partial differential equations*, volume 44 of *Applied Mathematical Sciences*. Springer, 1983.

[PS08] Grigorios A. Pavliotis and Andrew M. Stuart. *Multiscale methods*, volume 53 of *Texts in Applied Mathematics*. Springer, 2008. Averaging and homogenization.

[RC99] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 1999.

[RHCC07] J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(5):741–796, 2007. With discussions and a reply by the authors.

[RM92] M. Röckner and Z. M. Ma. *Introduction to the theory of (nonsymmetric) Dirichlet forms*. Springer, 1992.

[Rob01] James C. Robinson. *Infinite-dimensional dynamical systems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2001.

[SVW04] A. M. Stuart, J. Voss, and P. Wiberg. Conditional path sampling of SDEs and the Langevin MCMC method. *Communications in Mathematical Sciences*, 2(4):685–697, 2004.

[Tal90] D. Talay. Second order discretization schemes of stochastic differential systems for the computation of the invariant law. *Stochastics and Stochastic Reports*, 29(1):13–36, 1990.

[Tie98] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9, 1998.

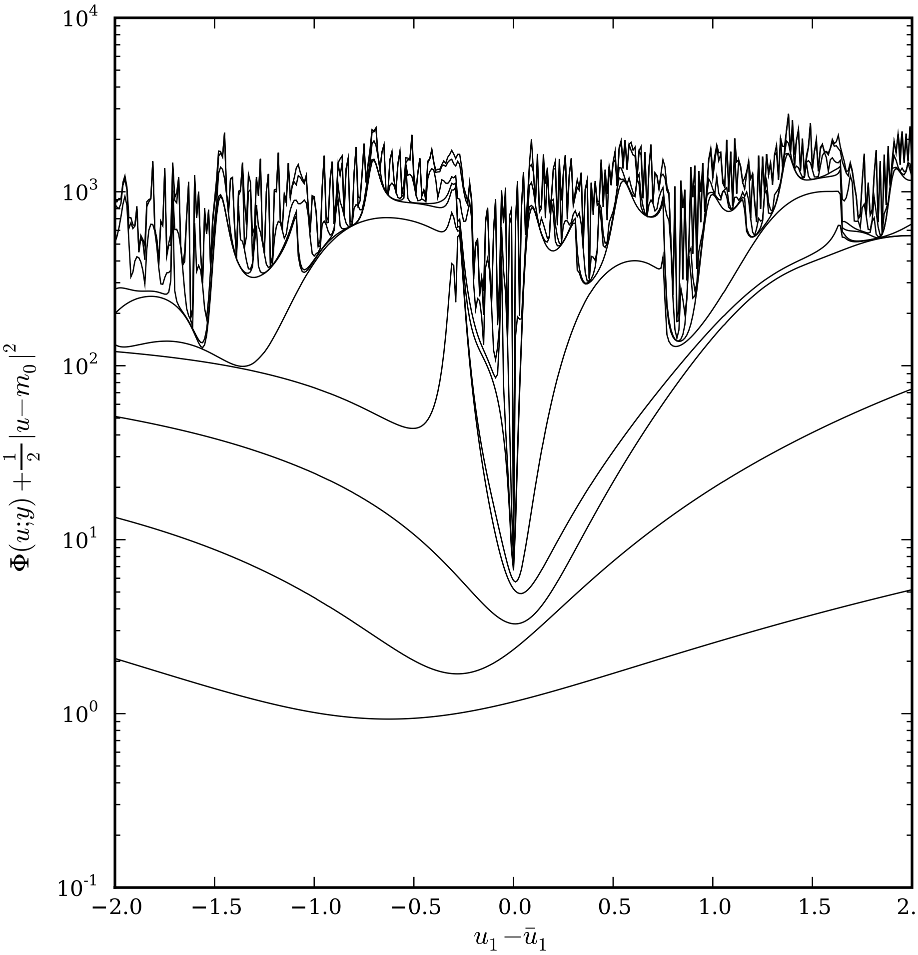

Figure 1: *Illustration of the posterior density for the Lorenz system from Example 3.2. Observations y are generated at times $1, 2, 3, \dots, 10$ for a trajectory starting at $v(0) = \bar{u} \in \mathbb{R}^3$. Then $u_1 \mapsto \Phi(u; y) + \frac{1}{2}|u - m_0|^2$ is plotted, where the last two components of u are fixed to the "exact" values $u_2 = \bar{u}_2$ and $u_3 = \bar{u}_3$. The different lines, from bottom to top, correspond to considering only the first $K = 1, \dots, 10$ observations. Up to a constant, the plotted value is $-\log \pi$ where π is the posterior density of μ^y. The figure illustrates that the effect of adding more observations is twofold: Firstly, the additional information allows to get better estimates of $\bar{u}$, the posterior distribution concentrates around this value. Second, as more observations are added, the shape of the posterior density gets more irregular and many local extrema appear.*

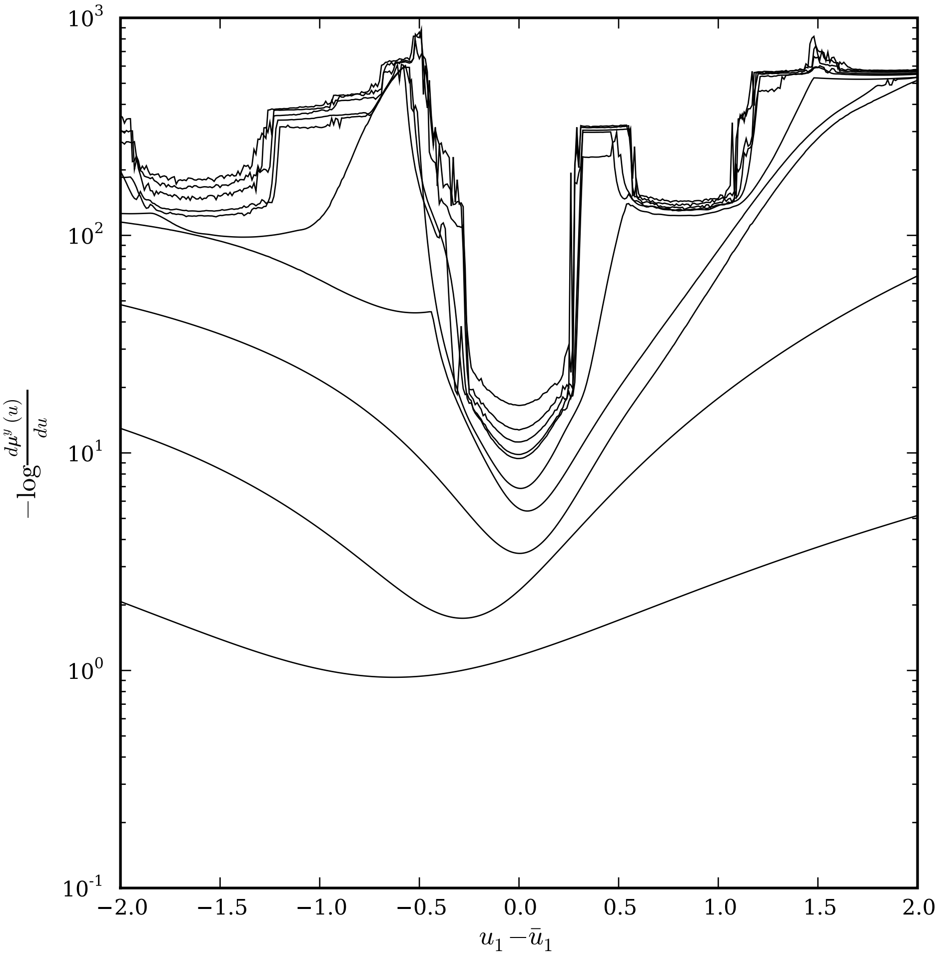

Figure 2: *Illustration of the posterior density for the initial condition in Example 3.11. The system is exactly the same as in Figure 1, except for the presence of an additional random forcing ψ in the Lorenz equation. The plotted posterior density of u is obtained by averaging $\Phi(u,\psi;y)+\frac{1}{2}|u-m_0|^2$ over ψ. The sampling is now on an infinite dimensional space, but the figure illustrates that the posterior for u is much smoother than in the situation without model error from Figure 1.*