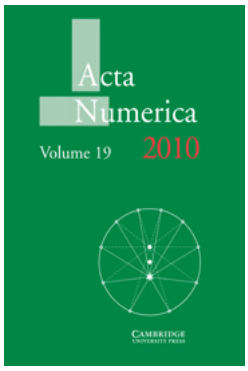

Inverse problems: A Bayesian perspective

A. M. Stuart

Acta Numerica / Volume 19 / May 2010, pp 451 - 559
DOI: 10.1017/S0962492910000061, Published online: 10 May 2010

Acta Numerica (2010), pp. 451–559
doi:10.1017/S0962492910000061

Inverse problems: A Bayesian perspective

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

The subject of inverse problems in differential equations is of enormous practical importance, and has also generated substantial mathematical and computational innovation. Typically some form of regularization is required to ameliorate ill-posed behaviour. In this article we review the Bayesian approach to regularization, developing a function space viewpoint on the subject. This approach allows for a full characterization of all possible solutions, and their relative probabilities, whilst simultaneously forcing significant modelling issues to be addressed in a clear and precise fashion. Although expensive to implement, this approach is starting to lie within the range of the available computational resources in many application areas. It also allows for the quantification of uncertainty and risk, something which is increasingly demanded by these applications. Furthermore, the approach is conceptually important for the understanding of simpler, computationally expedient approaches to inverse problems.

We demonstrate that, when formulated in a Bayesian fashion, a wide range of inverse problems share a common mathematical framework, and we highlight a theory of well-posedness which stems from this. The well-posedness theory provides the basis for a number of stability and approximation results which we describe. We also review a range of algorithmic approaches which are used when adopting the Bayesian approach to inverse problems. These include MCMC methods, filtering and the variational approach.

CONTENTS

1. Introduction

A significant challenge facing mathematical scientists is the development of a coherent mathematical and algorithmic framework enabling researchers to blend complex mathematical models with the (often vast) data sets which are now routinely available in many fields of engineering, science and technology. In this article we frame a range of inverse problems, mostly arising from the conjunction of differential equations and data, in the language of Bayesian statistics. In so doing our aim is twofold: (i) to highlight common mathematical structure arising from the numerous application areas where significant progress has been made by practitioners over the last few decades and thereby facilitate exchange of ideas between different application domains; (ii) to develop an abstract function space setting for the problems in order to evaluate the efficiency of existing algorithms, and to develop new algorithms. Applications are far-reaching and include fields such as the atmospheric sciences, oceanography, hydrology, geophysics, chemistry and biochemistry, materials science, systems biology, traffic flow, econometrics, image processing and signal processing.

The guiding principle underpinning the specific development of the subject of Bayesian inverse problems in this article is to *avoid discretization until the last possible moment*. This principle is enormously empowering throughout numerical analysis. For example, the first-order wave equation is not controllable to a given final state in arbitrarily small time because of finite speed of propagation. Yet every finite difference spatial discretization of the first-order wave equation gives rise to a linear system of ordinary differential equations which is controllable, in any finite time, to a given final state; asking the controllability question *before* discretization is key to understanding (Zuazua 2005). As another example consider the heat equation. If this is discretized in time by the theta method (with $\theta \in [0, 1]$ and $\theta = 0$ being explicit Euler, $\theta = 1$ implicit Euler), but left undiscretized in space, the resulting algorithm on function space is only defined if $\theta \in [\frac{1}{2}, 1]$; thus it is possible to deduce that there *must* be a Courant restriction if $\theta \in [0, \frac{1}{2})$ (Richtmyer and Morton 1967) before even introducing spatial discretization. Yet another example may be found in the study of Newton methods: conceptual application of this algorithm on function space, before discretization, can yield considerable insight when applying it as an iterative method for boundary value problems in nonlinear differential equations (Deuflhard 2004). The list of problems where it is beneficial to defer discretization to the very end of the algorithmic formulation is almost endless. It is perhaps not surprising, therefore, that the same idea yields insight in the solution of inverse problems and we substantiate this idea in the Bayesian context.

The article is divided into five parts. The next section, Section 2, is devoted to a description of the basic ideas of Bayesian statistics as applied to inverse problems in the finite-dimensional setting. It also includes a pointer to the common structure that we will highlight in the remainder of the article when developing the Bayesian viewpoint in function space. Section 3 contains a range of inverse problems arising in differential equations, showing how the Bayesian approach may be applied to inverse problems for functions; in particular, we discuss the problem of recovering a field from noisy pointwise data, recovering the diffusion coefficient from a boundary value problem, given noisy pointwise observations of the solution, recovering the wave speed from noisy observations of solutions of the wave equation and recovering the initial condition of the heat equation from noisy observation of the solution at a positive time. We also describe a range of applications, involving similar but more complex models, arising in weather forecasting, oceanography, subsurface geophysics and molecular dynamics. In Section 4 we describe, and exploit, the common mathematical structure which underlies *all* of these Bayesian inverse problems for functions. In that section we prove a form of well-posedness for these inverse problems, by showing Lipschitz continuity of the posterior measure with respect to changes in the data; we also prove an approximation theorem which exploits this well-posedness to show that approximation of the forward problem (by spectral or finite element methods, for example) leads to similar approximation results for the posterior probability measure. Section 5 is devoted to a survey of the existing algorithmic tools used to solve the problems highlighted in the article. In particular, Markov chain Monte Carlo (MCMC) methods, variational methods and filtering methods are surveyed. When discussing variational methods we show, in the setting of Section 4, that posterior probability maximizers can be characterized through solution of an optimal control problem, and that this optimal control problem has a minimizer under the same conditions that lead to a well-posed Bayesian inverse problem. Section 6 contains the background probability required to read the article; the presentation in this section is necessarily terse and the reader is encouraged to follow up references in the bibliography for further detail.

A major theme of the article is thus to confront the infinite-dimensional nature of many inverse problems. This is important because, whilst all computational algorithms work on finite-dimensional approximations, these approximations are typically in spaces of very high dimension and many significant challenges stem from this fact. By formulating inverse problems in an infinite-dimensional setting we build these challenges into the fabric of the problem setting. We provide a clear concept of *the ideal solution to the inverse problem* when blending a forward mathematical model with observational data. This concept can be used to test the practical algorithms used in applications which, in many cases, use crude approximations for

reasons of computational efficiency. Furthermore, it is also possible that the function space Bayesian setting will also lead to the development of improved algorithms which exploit the underlying mathematical structure common to a diverse range of applications. In particular, the theory of (Bayesian) well-posedness which we describe forms the cornerstone of many perturbation theories, including finite-dimensional approximations.

Kaipio and Somersalo (2005) provide a good introduction to the Bayesian approach to inverse problems, especially in the context of differential equations. Furthermore, Calvetti and Somersalo (2007*b*) provide a useful introduction to the Bayesian perspective in scientific computing. Another overview of the subject of inverse problems in differential equations, including a strong argument for the philosophy taken in this article, namely to formulate and study inverse problems in function space, is the book by Tarantola (2005) (see, especially, Chapter 5); however, the mathematics associated with this philosophical standpoint is not developed there to the same extent that it is in this article, and the focus is primarily on Gaussian problems. A frequentist viewpoint for inverse problems on function space is contained in the book by Ramsay and Silverman (2005); however, we adopt a different, Bayesian, perspective here, and study more involved differential equation models than those arising in Ramsay and Silverman (2005). These books indicate that the development that we undertake here is a natural one, which builds upon the existing literature.

The subject known as *data assimilation* provides a number of important applications of the material presented here. Its development has been driven, to a large extent, by practitioners working in the atmospheric and oceanographic sciences and in the geosciences, resulting in a plethora of algorithmic approaches and a number of significant algorithmic innovations. A good source for an understanding of data assimilation in the context of the atmospheric sciences, and weather prediction in particular, is the book by Kalnay (2003). A book motivated by applications in oceanography, which simultaneously highlights some of the underlying function space structure of data assimilation for linear, Gaussian problems, is that of Bennett (2002). The book by Evensen (2006) provides a good overview of many computational aspects of the subject, reflecting the author's experience in geophysical applications and related areas. The recent special edition of *Physica D* devoted to data assimilation provides a good entry point to some of the current research in this area (Ide and Jones 2007). Another application that fits the mathematical framework developed here is molecular dynamics. The problems of interest do not arise from Bayesian inverse problems, as such, but rather from conditioned diffusion processes. However, the mathematical structure has much in common with that arising in Bayesian inverse problems, and so we include a description of this problem area.

Throughout the article we use standard notation for Banach and Hilbert space norm and inner products, $\|\cdot\|, \langle\cdot,\cdot\rangle$, and the following notation for the finite-dimensional Euclidean norm and inner product: $|\cdot|, \langle\cdot,\cdot\rangle$. We also use the concept of weighted inner products and norms in any Hilbert space. For any self-adjoint positive operator $\mathcal{A}$, we define

$$\langle\cdot,\cdot\rangle_{\mathcal{A}} = \langle\mathcal{A}^{-1/2}\cdot, \mathcal{A}^{-1/2}\cdot\rangle, \quad \|\cdot\|_{\mathcal{A}} = \|\mathcal{A}^{-1/2}\cdot\|$$

in the general setting and, in finite dimensions,

$$|\cdot|_{\mathcal{A}} = |\mathcal{A}^{-1/2}\cdot|.$$

For any $a, b \in \mathcal{H}$, a Hilbert space, we define the operator $a \otimes b$ by the identity $(a \otimes b)c = \langle b, c\rangle a$ for any $c \in \mathcal{H}$. We use * to denote the adjoint of a linear operator between two Hilbert spaces. In particular, we may view $a, b \in \mathcal{H}$ as linear operators from $\mathbb{R}$ to $\mathcal{H}$ and then $a \otimes b = ab^*$.

In order to highlight the common structure arising in many of the problems in this book, we will endeavor to use the same notation repeatedly in the different contexts. A Gaussian measure will be denoted as $\mathcal{N}(m, \mathcal{C})$ with m the *mean* and $\mathcal{C}$ the *covariance operator/matrix*. The mean of the prior Gaussian measure will be m_0 and its covariance matrix/operator will be Σ_0 or $\mathcal{C}_0$ (we will drop the subscript 0 on the prior where no confusion arises in doing so). We will use the terminology *precision operator* for the (densely defined) $\mathcal{L} := \mathcal{C}^{-1}$. For inverse problems the operator mapping the unknown vector/field to the observations will be denoted by $\mathcal{G}$ and termed the *observation operator*, and the *observational noise* will be denoted by η.

We emphasize that in this article we will work for the most part with Gaussian priors. In terms of the classical theory of regularization this means that we are limiting ourselves to quadratic regularization terms, typically a Sobolev-type Hilbert space norm. We recognize that there are many applications of importance where other regularizations are natural, especially in image processing (Rudin, Osher and Fatemi 1992, Scherzer, Grasmair, Grossauer, Haltmeier and Lenzen 2009). A significant challenge is to take the material in this article and generalize it to these other settings, and there is some recent interesting work in this direction (Lassas, Saksman and Siltanen 2009).

There are other problem areas which lead to the need for computation of random functions. For example, there is a large body of work concerned with *uncertainty quantification* (DeVolder *et al.* 2002, Kennedy and O'Hagan 2001, Mohamed, Christie and Demyanov 2010, Efendiev, Datta-Gupta, Ma and Mallick 2009). In this field the input data to a differential equation is viewed as a random variable and the interest is in computing the resulting variability in the solution, as the input data varies. This is currently an active area of research in the engineering community (Spanos and Ghanem 1989, 2003). The work is thus primarily concerned with approximating

measures which are the push forward, under a nonlinear map, of a Gaussian measure; in contrast, the inverse problem setting which we study here is concerned with the approximation of non-Gaussian measures whose Radon–Nikodym derivative with respect to a Gaussian measure is defined through a related nonlinear map. A rigorous numerical analysis underpinning the work of Spanos and Ghanem (1989, 2003) is an active area of research: see in particular Schwab and Todor (2006) and Todor and Schwab (2007), where the problem is viewed as an example of Schwab's more general program of tensor product approximation for high-(infinite)-dimensional problems (Gittelson and Schwab 2011). A different area where tensor products are used to form approximations of functions of many variables is computational quantum mechanics and approximation of the Schrödinger equation (Lubich 2008); this work may also be seen in the more general context of tensor product approximations in linear algebra (Kolda and Bader 2009). It would be interesting to investigate whether any of these tensor product ideas can be transferred to the approximation of probability density functions in high-dimensional spaces, as arise naturally in Bayesian inverse problems.

More generally speaking, this article is concerned with a research area which is at the interface of applied mathematics and statistics. This is a rich research interface, where there is currently significant effort. Examples include work in compressed sensing, which blends ideas from statistics, probability, approximation theory and harmonic analysis (Candès and Wakin 2008, Donoho 2006), and research aimed at efficient sampling of Gaussian random fields combining numerical linear algebra and statistics (Rue and Held 2005).

2. The Bayesian framework

2.1. Overview

This section introduces the Bayesian approach to inverse problems and outlines the common structure that we will develop in the remainder of the article. In Section 2.2 we introduce finite-dimensional inverse problems and describe the Bayesian approach to their solution, highlighting the role of observational noise which pollutes the data in many problems of practical interest. We show how to construct a formula for the posterior measure on the unknown of interest, from the data and from a prior measure incorporating structural knowledge about the problem which is present prior to the acquisition of the data. In Section 2.3 we study the effect on the posterior of small observational noise, in order to connect the Bayesian viewpoint with the classical perspective on inverse problems. We first study problems where the dimensions of the data set and the unknown match; we show that the prior measure is asymptotically irrelevant and that, in the limit of zero noise, the posterior measure converges weakly to a Dirac measure

centred on the solution of the noise-free equation. We next study the special structure which arises when the mathematical model and observations are described through linear operators, and when the prior and the noise are Gaussian; this results in a Gaussian posterior measure. In this Gaussian setting we first study the limit of vanishing observational noise in the case where the dimension of the data set is greater than that of the unknown, showing that the prior is asymptotically irrelevant, and that the posterior measure approaches a Dirac concentrated on the solution of a natural least-squares problem. We then study the situation where the dimension of the data set is smaller than that of the unknown. We show that, in the limit of small observational noise, the prior remains important and we characterize this effect explicitly. Section 2.4 completes the introductory material by describing the common framework which we will illustrate and exploit in the remainder of the article when developing the Bayesian viewpoint on function space.

2.2. Linking the classical and Bayesian approaches

In applications it is frequently of interest to solve *inverse problems*: to find u, an input to a mathematical model, given y an observation of (some components of, or functions of) the solution of the model. We have an equation of the form

$$y = \mathcal{G}(u) \tag{2.1}$$

to solve for $u \in X$, given $y \in Y$, where X, Y are Banach spaces. We will refer to $\mathcal{G}$ as the *observation operator*.[1] We refer to y as *data*. It is typical of inverse problems that they are *ill-posed*: there may be no solution, or the solution may not be unique and may depend sensitively on y. One approach to the problem in this situation is to replace it by the *least-squares* optimization problem of finding, for the norm $\|\cdot\|_Y$ on Y,

$$\text{argmin}_{u\in X}\ \frac{1}{2}\|y - \mathcal{G}(u)\|_Y^2. \tag{2.2}$$

This problem, too, may be difficult to solve as it may possess minimizing sequences $u^{(n)}$ which do not converge to a limit in X, or it may possess multiple minima and sensitive dependence on the data y. These issues can be somewhat ameliorated by solving a *regularized* minimization problem of the form, for some Banach space $(E, \|\cdot\|_E)$ contained in X, and point $m_0 \in E$,

$$\text{argmin}_{u\in E}\left(\frac{1}{2}\|y - \mathcal{G}(u)\|_Y^2 + \frac{1}{2}\|u - m_0\|_E^2\right). \tag{2.3}$$

[1] This operator is often denoted with the letter $\mathcal{H}$ in the atmospheric sciences community; because we need $\mathcal{H}$ for Hilbert space later on, we use the symbol $\mathcal{G}$.

However, the choice of norms $\|\cdot\|_E, \|\cdot\|_Y$ and the point m_0 are somewhat arbitrary, without making further modelling assumptions. We will adopt a statistical approach to the inverse problems, in which these issues can be articulated and addressed in an explicit fashion. Roughly speaking, the Bayesian approach will lead to the notion of finding a *probability measure* μ^y on X, containing information about the relative probability of different states u, given the data y. For example, in the case where X, Y are both finite-dimensional, the noise polluting (2.1) is additive and Gaussian, and the prior measure is Gaussian, the posterior measure will have density π^y given by

$$\pi^y(u) \propto \exp\left(-\frac{1}{2}\|y - \mathcal{G}(u)\|_Y^2 - \frac{1}{2}\|u - m_0\|_E^2\right). \tag{2.4}$$

The properties of a measure μ^y with such a density π^y are intimately related to the minimization problem (2.3): the density is largest at minimizers. But the probabilistic approach is far richer. For example, the derivation of the probability measure μ^y will force us to confront various modelling and mathematical issues which, together, will guide the choice of norms $\|\cdot\|_E$, $\|\cdot\|_Y$ and the point m_0. Furthermore, the probabilistic approach enables us to answer questions such as: 'What is the relative probability that the unknown function u is determined by the different local minimizers of (2.3)?' 'How certain can we be that a prediction made by a mathematical model will lie in certain specified regimes?'

We now outline a probabilistic framework which will include the specific probability measure with density given by (2.4) as a special case. This framework starts from the observation that a deeper understanding of the source of data often reveals that the observations y are subject to noise and that a more appropriate model equation is often of the form

$$y = \mathcal{G}(u) + \eta, \tag{2.5}$$

where η is a mean zero random variable, whose statistical properties we might know, but whose actual value is unknown to us; we refer to η as the *observational noise*. In this context it is natural to adopt a *Bayesian* approach to the problem of determining u from y: see Section 6.6. We describe our prior beliefs about u, in terms of a probability measure μ_0, and use Bayes' formula (see (6.24)) to calculate the posterior probability measure μ^y, for u given y.

To be concrete, in the remainder of this subsection and in the next subsection we consider the case where $u \in \mathbb{R}^n, y \in \mathbb{R}^q$ and we let π_0 and π^y denote the p.d.f.s (see Section 6.1) of measures μ_0 and μ^y. We assume that $\eta \in \mathbb{R}^q$ is a random variable with density ρ. Then the probability of y given u has density

$$\rho(y|u) := \rho(y - \mathcal{G}(u)).$$

This is often referred to as the *data likelihood*. By Bayes' formula (6.24) we obtain

$$\pi^y(u) = \frac{\rho(y - \mathcal{G}(u))\pi_0(u)}{\int_{\mathbb{R}^n} \rho(y - \mathcal{G}(u))\pi_0(u)\,\mathrm{d}u}. \tag{2.6}$$

Thus

$$\pi^y(u) \propto \rho(y - \mathcal{G}(u))\pi_0(u) \tag{2.7}$$

with constant of proportionality depending only on y. Abstractly (2.7) expresses the fact that the posterior measure μ^y (with density π^y) and prior measure μ_0 (with density π_0) are related through the Radon–Nikodym derivative (see Theorem 6.2)

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \rho(y - \mathcal{G}(u)). \tag{2.8}$$

Since ρ is a density and thus non-negative, without loss of generality we may write the right-hand side as the exponential of the negative of a potential $\Phi(u; y)$, to obtain

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp(-\Phi(u; y)). \tag{2.9}$$

It is this form which generalizes naturally to situations where X, and possibly Y, is infinite-dimensional. We show in Section 3 that many inverse problems can be formulated in a Bayesian fashion and that the posterior measure takes this form.

In general it is hard to obtain information from a probability measure in high dimensions. One useful approach to extracting information is to find a *maximum a posteriori estimator*, or *MAP estimator*: a point u which maximizes the posterior p.d.f. π^y; such *variational* methods are surveyed in Section 5.3. Another commonly used method for interrogating a probability measure in high dimensions is *sampling*: generating a set of points $\{u_n\}_{n=1}^N$ distributed (perhaps only approximately) according to $\pi^y(u)$. In this context formula (2.7) (or (2.9) in the general setting), in which the posterior density is known only up to a constant, is useful because *MCMC methods* may be used to sample from it: MCMC methods have the advantage of sampling from a probability measure only known up to a normalizing constant; we outline these methods in Section 5.2. Time-dependent problems, where the data is acquired sequentially, also provide a class of problems where useful approximations can be developed; these *filtering* methods are outlined in Section 5.4.

We will often be interested in problems where prior μ_0 and observational noise η are Gaussian. If $\eta \sim \mathcal{N}(0, B)$ and $\mu_0 = \mathcal{N}(m_0, \Sigma_0)$, then we obtain

from (2.7) the formula[2]

$$\pi^y(u) \propto \exp\left(-\frac{1}{2}\left|B^{-1/2}\big(y-\mathcal{G}(u)\big)\right|^2 - \frac{1}{2}\left|\Sigma_0^{-1/2}\big(u-m_0\big)\right|^2\right)$$
$$= \exp\left(-\frac{1}{2}\big|\big(y-\mathcal{G}(u)\big)\big|_B^2 - \frac{1}{2}\big|\big(u-m_0\big)\big|_{\Sigma_0}^2\right). \tag{2.10}$$

In terms of measures this is the statement that

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2}\big|\big(y-\mathcal{G}(u)\big)\big|_B^2\right). \tag{2.11}$$

The *maximum a posteriori estimator*, or *MAP estimator*, is then

$$\mathrm{argmin}_{u\in\mathbb{R}^n}\left(\frac{1}{2}\big|y-\mathcal{G}(u)\big|_B^2 + \frac{1}{2}\big|u-m_0\big|_{\Sigma_0}^2\right). \tag{2.12}$$

This is a specific instance of the regularized minimization problem (2.3). Note that in the Bayesian framework the norms $\|\cdot\|_Y, \|\cdot\|_E$ and the point m_0 all have a clear interpretation in terms of the statistics of the observational noise and the prior measure. In contrast, these norms and point are somewhat arbitrary in the classical approach.

In general the posterior probability measure (2.10) is not itself Gaussian. However, if $\mathcal{G}$ is linear then the posterior μ^y is also Gaussian. Identifying the mean and covariance (or precision) matrix can be achieved by *completing the square*, as formalized in Theorem 6.20 and Lemma 6.21 (see also Examples 6.22 and 6.23). The following simple examples illustrate this. They also show further connections between the Bayesian and classical approaches to inverse problems, a subject we develop further in the following subsection.

Example 2.1. Let $q = 1$ and $\mathcal{G}$ be linear so that

$$y = \langle g, u\rangle + \eta$$

for some $g \in \mathbb{R}^n$. Assume further that $\eta \sim \mathcal{N}(0,\gamma^2)$ and that we place a prior Gaussian measure $\mathcal{N}(0,\Sigma_0)$ on u. Then

$$\pi^y(u) \propto \exp\left(-\frac{1}{2\gamma^2}|y-\langle g,u\rangle|^2 - \frac{1}{2}\langle u,\Sigma_0^{-1}u\rangle\right). \tag{2.13}$$

As the exponential of a quadratic form, this is the density of a Gaussian measure.

[2] The notation for weighted norms and inner products is defined at the end of Section 1.

From Theorem 6.20 we find that the posterior mean and covariance are given by

$$m = \frac{(\Sigma_0 g)y}{\gamma^2 + \langle g, \Sigma_0 g\rangle},$$
$$\Sigma = \Sigma_0 - \frac{(\Sigma_0 g)(\Sigma_0 g)^*}{\gamma^2 + \langle g, \Sigma_0 g\rangle}.$$

If we consider the case where observational noise disappears from the system, then we find that

$$m^+ := \lim_{\gamma\to 0} m = \frac{(\Sigma_0 g)y}{\langle g, \Sigma_0 g\rangle}, \quad \Sigma^+ := \lim_{\gamma\to 0} \Sigma = \Sigma_0 - \frac{(\Sigma_0 g)(\Sigma_0 g)^*}{\langle g, \Sigma_0 g\rangle}.$$

Notice that $\Sigma^+ g = 0$ and $\langle m^+, g\rangle = y$. This states the intuitively reasonable fact that, as the observational noise decreases, knowledge of u in the direction of g becomes certain. In directions not aligned with g, uncertainty remains, with magnitude determined by an interaction between properties of the prior and of the observation operator. Thus the prior plays a central role, even as observational noise disappears, in this example where the solution is underdetermined. ◇

Example 2.2. Assume that $q \geqslant 2$ and $n = 1$, and let $\mathcal{G}$ be nonlinear with the form

$$y = g(u + \beta u^3) + \eta,$$

where $g \in \mathbb{R}^q \backslash \{0\}$, $\beta \in \mathbb{R}$ and $\eta \sim \mathcal{N}(0, \gamma^2 I)$. Assume further that we place a Gaussian measure $\mathcal{N}(0, 1)$ as a prior on u. Then

$$\pi^y(u) \propto \exp\left(-\frac{1}{2\gamma^2}|y - g(u + \beta u^3)|^2 - \frac{1}{2}u^2\right).$$

This measure is not Gaussian unless $\beta = 0$.

Consider the linear case where $\beta = 0$. The posterior measure is then Gaussian:

$$\pi^y(u) \propto \exp\left(-\frac{1}{2\gamma^2}|y - gu|^2 - \frac{1}{2}|u|^2\right).$$

By Theorem 6.20, using the identity

$$\left(\gamma^2 I + gg^*\right)^{-1} g = \left(\gamma^2 + |g|^2\right)^{-1} g,$$

we deduce that the posterior mean and covariance are given by

$$m = \frac{\langle g, y\rangle}{\gamma^2 + |g|^2},$$
$$\sigma^2 = \frac{\gamma^2}{\gamma^2 + |g|^2}.$$

In the limit where observational noise disappears, we find that

$$m^+ = \lim_{\gamma\to 0} m = \frac{\langle g, y\rangle}{|g|^2}, \quad (\sigma^+)^2 = \lim_{\gamma\to 0}\sigma^2 = 0.$$

The point m^+ is the least-squares solution of the overdetermined linear equation $y = gu$ found from the minimization problem

$$\text{argmin}_{u\in\mathbb{R}}|y - gu|^2.$$

This is a minimization problem of the form (2.2). In this case, where the system is overdetermined, the prior plays no role in the limit of zero observational noise. ◇

2.3. Small noise limits of the posterior measure

We have shown that the Bayesian and classical perspectives are linked through the relationship between the posterior probability density given by (2.10) and the MAP estimator (2.12). This directly connects minimization of a regularized least-squares problem with the Bayesian perspective. Our aim now is to further the link between the Bayesian and classical approaches by considering the limit of small observational noise.

The small observational noise limit is illustrated in the two examples concluding the previous subsection. In the first, where the underlying noise-free problem is underdetermined, the prior provides information about the posterior mean, and uncertainty remains in the posterior, even as observational noise disappears; furthermore, that uncertainty is related to the choice of prior. In the second example, where the underlying noise-free problem is overdetermined, uncertainty disappears and the posterior converges to a Dirac measure centred on the least-squares solution of the limiting deterministic equation. The intuition obtained from these two examples, concerning the behaviour of the posterior measure in the small noise limit, is important. The first example suggests that in the underdetermined case the prior measure plays a role in determining the posterior measure, even as the observational noise disappears; in contrast, the second example suggests that, in the overdetermined case, the prior plays no role in the small noise limit. Many of the inverse problems for functions that we study later in this paper are underdetermined. For these problems *the prior measure plays an important role in the solution, even when observational noise is small.* A significant advantage of the Bayesian framework over classical approaches is that it makes the modelling assumptions which underly the prior both clear and explicit.

In the remainder of this subsection we demonstrate that the intuition obtained from the two examples can be substantiated on a broad class of finite-dimensional inverse problems. We first concentrate on the general case which lies between these two examples, where $q = n$ and, furthermore,

equation (2.1) has a unique solution. We then restrict our attention to Gaussian problems, studying the over- and underdetermined cases in turn. We state the results first, and provide proofs at the end of the subsection. The results are stated in terms of weak convergence of probability measures, denoted by $\Rightarrow$; see the end of Section 6.1 for background on this concept. Throughout this subsection we consider the data y to be fixed, and we study the limiting behaviour of the posterior measure μ^y as the observational noise tends to zero. Other limits, where y is a random variable, depending on the observational noise, are also of interest, but we stick to the simpler setting where y is fixed, for expository purposes.

We start with the case $q = n$ and assume that equation (2.1) has a unique solution

$$u = \mathcal{F}(y) \tag{2.14}$$

for every $y \in \mathbb{R}^n$. Intuitively this unique solution should dominate the Bayesian solution to the problem (which is a probability distribution on $\mathbb{R}^n$, not a single point). We show that this is indeed the case: the probability distribution concentrates on the single point given by (2.14) as observational noise disappears.

We assume that there is a positive constant C such that, for all $y, \delta \in \mathbb{R}^n$,

$$|y - \mathcal{G}(\mathcal{F}(y) + \delta)|^2 \geqslant C \min\{1, |\delta|^2\}. \tag{2.15}$$

This condition implies that the derivative $DG(u)$ is invertible at $u = \mathcal{F}(y)$, so that the implicit function theorem holds; the condition also excludes the possibility of attaining the minimum 0 of $\frac{1}{2}|y - \mathcal{G}(u)|^2$ along a sequence $u_n \to \infty$. We then have the following.

Theorem 2.3. Assume that $k = n$, that $\mathcal{G} \in C^2(\mathbb{R}^n, \mathbb{R}^n)$ and that equation (2.1) has a unique solution given by (2.14), for every $y \in \mathbb{R}^n$. We place a Gaussian prior $\mu_0 = \mathcal{N}(m_0, \Sigma_0)$ on u and assume that the observational noise η in (2.5) is distributed as $\mathcal{N}(0, \gamma^2 I)$. Then the posterior measure μ^y, with density given by (2.10) and $B = \gamma^2 I$, satisfies $\mu^y \Rightarrow \delta_{\mathcal{F}(y)}$ as $\gamma \to 0$. $\diamond$

The preceding theorem concerns problems where the underlying equation (2.1) relating data to model is uniquely solvable. This situation rarely arises in practice, but is of course important for building links between the Bayesian and classical perspectives.

We now turn to problems which are either over- or underdetermined and, for simplicity, confine our attention to purely Gaussian problems. We again work in arbitrary finite dimensions and study the small observational noise limit and its relation to the the underlying noise-free problem (2.1). In Theorem 2.4 we show that the posterior measure converges to a Dirac measure concentrated on minimizers of the least-squares problem (2.2). Of course, when (2.1) is uniquely solvable this will lead to a Dirac measure as its solution, as in Theorem 2.3; but more generally there may be no solution to (2.1)

and least-squares minimizers provide a natural generalized solution concept. In Theorem 2.5 we study the Gaussian problem in the undetermined case, showing that the posterior measure converges to a Gaussian measure whose support lies on a hyperplane embedded in the space where the unknown u lies. The structure of this Gaussian measure is determined by an interplay between the prior, the forward model and the data. In particular, *prior information remains in the small noise limit.* This illustrates the important idea that for (frequently occurring) underdetermined problems the prior plays a significant role, even when noise is small, and should therefore be treated very carefully from the perspective of mathematical modelling.

If the observational noise η is Gaussian, if the prior μ_0 is Gaussian and if $\mathcal{G}$ is a linear map, then the posterior measure μ^y is also Gaussian. This follows immediately from the fact that the logarithm of π^y given by (2.6) is quadratic in u under these assumptions. We now study the properties of this Gaussian posterior.

We assume that

$$\eta \sim \mathcal{N}(0, B), \quad \mu_0 = \mathcal{N}(m_0, \Sigma_0), \quad \mathcal{G}(u) = Au$$

and that B and Σ_0 are both invertible.

Then, since $y|u \sim \mathcal{N}(Au, B)$, Theorem 6.20 shows that the posterior measure μ^y is Gaussian $\mathcal{N}(m, \Sigma)$ with

$$m = m_0 + \Sigma_0 A^*(B + A\Sigma_0 A^*)^{-1}(y - Am_0), \tag{2.16a}$$

$$\Sigma = \Sigma_0 - \Sigma_0 A^*(B + A\Sigma_0 A^*)^{-1} A\Sigma_0. \tag{2.16b}$$

In the case where $k = n$ and A, Σ_0 are invertible, we see that, as $B \to 0$,

$$m \to A^{-1}y, \quad \Sigma \to 0.$$

From Lemma 6.5 we know that convergence of all characteristic functions implies weak convergence. Furthermore, the characteristic function of a Gaussian is determined by the mean and covariance: see Theorem 6.4. Hence, for a finite-dimensional family of Gaussians, convergence of the mean and covariance to a limit implies weak convergence to the Gaussian with that limiting mean and covariance. For this family of measures the limiting covariance is zero and thus the $B \to 0$ limit recovers a Dirac measure on the solution of the equation $Au = y$, in accordance with Theorem 2.3. It is natural to ask what happens in the limit of vanishing noise, more generally. The following two theorems provide an answer to this question.

Theorem 2.4. Assume that B and Σ_0 are both invertible. The posterior mean and covariance can be rewritten as

$$m = (A^*B^{-1}A + \Sigma_0^{-1})^{-1}(A^*B^{-1}y + \Sigma_0^{-1}m_0), \tag{2.17a}$$

$$\Sigma = (A^*B^{-1}A + \Sigma_0^{-1})^{-1}. \tag{2.17b}$$

If $\mathrm{Null}(A) = \{0\}$ and $B = \gamma^2 B_0$ then, in the limit $\gamma^2 \to 0$, $\mu^y \Rightarrow \delta_{m^+}$, where m^+ is the solution of the least-squares problem

$$m^+ = \mathrm{argmin}_{u \in \mathbb{R}^n} |B_0^{-1/2}(y - Au)|^2. \qquad \diamond$$

The preceding theorem shows that, in the overdetermined case where A^*BA is invertible, the small observational noise limit leads to a posterior which is a Dirac, centred on the solution of a least-squares problem determined by the observation operator and the relative weights on the observational noise. Uncertainty disappears, and the prior plays no role in this limit. Example 2.2 illustrates this situation.

We now assume that $y \in \mathbb{R}^q$ and $u \in \mathbb{R}^n$ with $q < n$, so that the problem is underdetermined. We assume that $\mathrm{rank}(A) = q$, so that we may write

$$A = (A_0 \;\; 0)Q^* \tag{2.18}$$

with $Q \in \mathbb{R}^{n \times n}$ an orthogonal matrix so that $Q^*Q = I$, $A_0 \in \mathbb{R}^{q \times q}$ an invertible matrix and $0 \in \mathbb{R}^{q \times (n-q)}$ a zero matrix. We also let $L_0 = \Sigma_0^{-1}$, the precision matrix for the prior, and write

$$Q^* L_0 Q = \begin{pmatrix} L_{11} & L_{12} \\ L_{12}^* & L_{22} \end{pmatrix}. \tag{2.19}$$

Here $L_{11} \in \mathbb{R}^{q \times q}$, $L_{12} \in \mathbb{R}^{q \times (n-q)}$ and $L_{22} \in \mathbb{R}^{(n-q) \times (n-q)}$; both L_{11} and L_{22} are positive definite symmetric, because Σ_0 is.

If we write

$$Q = (Q_1 \;\; Q_2) \tag{2.20}$$

with $Q_1 \in \mathbb{R}^{n \times q}$ and $Q_2 \in \mathbb{R}^{n \times (n-q)}$, then Q_1^* projects onto a q-dimensional subspace $\mathcal{O}$ and Q_2^* projects onto an $(n-q)$-dimensional subspace $\mathcal{O}^\perp$; here $\mathcal{O}$ and $\mathcal{O}^\perp$ are orthogonal.

Assume that $y = Au$ for some $u \in \mathbb{R}^n$. This identity is at the heart of the inverse problem in the small γ limit. If we define $z \in \mathbb{R}^q$ to be the unique solution of the system of equations $A_0 z = y$, then $z = Q_1^* u$. On the other hand, $Q_2^* u$ is not determined by the identity $y = Au$. Thus, intuitively we expect to determine z without uncertainty, in the limit of small noise, but for uncertainty to remain in other directions. With this in mind we define $w \in \mathbb{R}^q$ and $w' \in \mathbb{R}^{n-q}$ via the equation

$$\Sigma_0^{-1} m_0 = Q \begin{pmatrix} w \\ w' \end{pmatrix}, \tag{2.21}$$

and then set

$$z' = -L_{22}^{-1} L_{12}^* z + L_{22}^{-1} w' \in \mathbb{R}^{n-q}.$$

Theorem 2.5. Assume that B and Σ_0 are both invertible and let $B = \gamma^2 B_0$. Then, in the limit $\gamma^2 \to 0$, $\mu^y \Rightarrow \mathcal{N}(m^+, \Sigma^+)$, where

$$m^+ = Q \begin{pmatrix} z \\ z' \end{pmatrix}, \tag{2.22a}$$

$$\Sigma^+ = Q_2 L_{22}^{-1} Q_2^*. \tag{2.22b}$$

◇

We now interpret this theorem. Since $Q_2^* Q_1 = 0$, the limiting measure may be viewed as a Dirac measure, centred at z in $\mathcal{O}$, and a Gaussian measure $\mathcal{N}(z', L_{22}^{-1})$ in $\mathcal{O}^\perp$. These measures are independent, so that the theorem states that

$$\mu^y \Rightarrow \delta_z \otimes \mathcal{N}(z', L_{22}^{-1}),$$

viewed as a measure on $\mathcal{O} \oplus \mathcal{O}^\perp$. Thus, in the small observational noise limit, we determine the solution without uncertainty in $\mathcal{O}$, whilst in $\mathcal{O}^\perp$ uncertainty remains. Furthermore, the prior plays a role in the posterior measure in the limit of zero observational noise; specifically it enters the formulae for z' and L_{22}.

We finish this subsection by providing proofs of the preceding three results.

Proof of Theorem 2.3. Define $\delta := u - \mathcal{F}(y)$ and let

$$f(\delta) = -\frac{1}{2\gamma^2}|y - \mathcal{G}(\mathcal{F}(y) + \delta)|^2 - \frac{1}{2}|\Sigma_0^{-1/2}(\mathcal{F}(y) + \delta - m_0)|^2.$$

Fix $\ell \in \mathbb{R}^n$. Then, with $\mathbb{E}$ denoting expectation under μ^y,

$$\mathbb{E}\exp\bigl(\mathrm{i}\langle \ell, u\rangle\bigr) = \frac{1}{Z}\exp\bigl(\mathrm{i}\langle \ell, \mathcal{F}(y)\rangle\bigr)\int_{\mathbb{R}^n} \exp\bigl(\mathrm{i}\langle \ell, \delta\rangle + f(\delta)\bigr)\,\mathrm{d}\delta,$$

where

$$Z = \int_{\mathbb{R}^n} \exp(f(\delta))\,\mathrm{d}\delta.$$

Thus, by Lemma 6.5, it suffices to prove that, as $\gamma \to 0$,

$$\frac{1}{Z}\int_{\mathbb{R}^n} \exp\bigl(\mathrm{i}\langle \ell, \delta\rangle + f(\delta)\bigr)\,\mathrm{d}\delta \to 1.$$

Define

$$I(\ell) = \int_{\mathbb{R}^n} \exp\bigl(\mathrm{i}\langle \ell, \delta\rangle + f(\delta)\bigr)\,\mathrm{d}\delta,$$

noting that $Z = I(0)$. For $a \in (2/3, 1)$, we split $I(\ell)$ into $I(\ell) = I_1(\ell) + I_2(\ell)$ where

$$I_1(\ell) = \int_{|\delta| \leqslant \gamma^a} \exp\bigl(\mathrm{i}\langle \ell, \delta\rangle + f(\delta)\bigr)\,\mathrm{d}\delta,$$

$$I_2(\ell) = \int_{|\delta| > \gamma^a} \exp\bigl(\mathrm{i}\langle \ell, \delta\rangle + f(\delta)\bigr)\,\mathrm{d}\delta.$$

We consider $I_1(\ell)$ first so that $|\delta| \leqslant \gamma^a$. By Taylor-expanding $f(\delta)$ around $\delta = 0$, we obtain

$$f(\delta) = -\frac{1}{2\gamma^2}|B\delta|^2 - \frac{1}{2}\big|\Sigma_0^{-1/2}\big(\mathcal{F}(y) + \delta - m_0\big)\big|^2 + \mathcal{O}\bigg(\frac{\delta^3}{\gamma^2}\bigg),$$

where $B = D\mathcal{G}(\mathcal{F}(y))$. Thus, for $b = a \wedge (3a-2) = 3a-2$,

$$\mathrm{i}\langle \ell, \delta\rangle + f(\delta) = -\frac{1}{2}\big|\Sigma_0^{-1/2}\big(\mathcal{F}(y) - m_0\big)\big|^2 - \frac{1}{2\gamma^2}|B\delta|^2 + \mathcal{O}(\gamma^b).$$

Thus

$$I_1(\ell) = \exp\bigg(-\frac{1}{2}\big|\Sigma_0^{-1/2}\big(\mathcal{F}(y) - m_0\big)\big|^2\bigg)\int_{|\delta|\leqslant\gamma^a} \exp\bigg(-\frac{1}{2\gamma^2}|B\delta|^2 + \mathcal{O}(\gamma^b)\bigg)\,\mathrm{d}\delta.$$

It follows that

$$\begin{aligned} I_1(\ell) &= \exp\bigg(-\frac{1}{2}\big|\Sigma_0^{-1/2}\big(\mathcal{F}(y) - m_0\big)\big|^2\bigg)\int_{|\delta|\leqslant\gamma^a} \exp\bigg(-\frac{1}{2\gamma^2}|B\delta|^2\bigg) \\ &\qquad \times \big(1 + \mathcal{O}(\gamma^b)\big)\,\mathrm{d}\delta \\ &= \gamma^n \exp\bigg(-\frac{1}{2}\big|\Sigma_0^{-1/2}\big(\mathcal{F}(y) - m_0\big)\big|^2\bigg)\int_{|z|\leqslant\gamma^{a-1}} \exp\bigg(-\frac{1}{2}|Bz|^2\bigg) \\ &\qquad \times \big(1 + \mathcal{O}(\gamma^b)\big)\,\mathrm{d}z. \end{aligned}$$

We now estimate $I_2(\ell)$ and show that it is asymptotically negligible compared with $I_1(\ell)$. Note that, by (2.15),

$$\begin{aligned} f(\delta) &\leqslant -\frac{C\min\{1, |\delta|^2\}}{2\gamma^2} - \frac{1}{2}\big|\Sigma_0^{-1/2}\big(\mathcal{F}(y) + \delta - m_0\big)\big|^2 \\ &\leqslant -\frac{C\min\{1, |\delta|^2\}}{2\gamma^2}. \end{aligned}$$

Thus

$$\begin{aligned} |I_2(\ell)| \leqslant{}& \int_{1\geqslant|\delta|>\gamma^a} \exp\bigg(-\frac{C\delta^2}{\gamma^2}\bigg)\mathrm{d}\delta \\ &+ \int_{|\delta|>1} \exp\bigg(-\frac{C}{\gamma^2}\bigg)\exp\bigg(-\frac{1}{2}\big|\Sigma_0^{-1/2}\big(\mathcal{F}(y) + \delta - m_0\big)\big|^2\bigg)\mathrm{d}\delta. \end{aligned}$$

Since $a < 1$, it follows that $I_2(\ell)$ is exponentially small in $\gamma \to 0$. As $I_1(\ell)$ is, to leading order, $\mathcal{O}(\gamma^n)$ and independent of ℓ, we deduce that

$$\frac{1}{Z}\int_{\mathbb{R}^n} \exp\big(\mathrm{i}\langle \ell, \delta\rangle + f(\delta)\big)\,\mathrm{d}\delta = \frac{I(\ell)}{I(0)} \to 1$$

as $\gamma \to 0$, and the result follows. □

Proof of Theorem 2.4. We first note the identity

$$A^*B^{-1}(B + A\Sigma_0 A^*) = (A^*B^{-1}A + \Sigma_0^{-1})\Sigma_0 A^*,$$

which follows since Σ_0 and B are both positive definite. Since $A^*B^{-1}A+\Sigma_0^{-1}$ and $B + A\Sigma_0 A^*$ are also positive definite, we deduce that

$$(A^*B^{-1}A + \Sigma_0^{-1})^{-1}A^*B^{-1} = \Sigma_0 A^*(B + A\Sigma_0 A^*)^{-1}.$$

Thus the posterior mean may be written as

$$\begin{aligned} m &= m_0 + (A^*B^{-1}A + \Sigma_0^{-1})^{-1}A^*B^{-1}(y - Am_0) \\ &= (A^*B^{-1}A + \Sigma_0^{-1})^{-1}(A^*B^{-1}y + A^*B^{-1}Am_0 + \Sigma_0^{-1}m_0 - A^*B^{-1}Am_0) \\ &= (A^*B^{-1}A + \Sigma_0^{-1})^{-1}(A^*B^{-1}y + \Sigma_0^{-1}m_0), \end{aligned}$$

as required. A similar calculation establishes the desired property of the posterior covariance.

If $B = \gamma^2 B_0$ then we deduce that

$$\begin{aligned} m &= (A^*B_0^{-1}A + \gamma^2\Sigma_0^{-1})^{-1}(A^*B_0^{-1}y + \gamma^2\Sigma_0^{-1}m_0), \\ \Sigma &= \gamma^2(A^*B_0^{-1}A + \gamma^2\Sigma_0^{-1})^{-1}. \end{aligned}$$

Since $\mathrm{Null}(A) = \{0\}$, we deduce that there is $\alpha > 0$ such that

$$\langle \xi, A^*B_0^{-1}A\xi\rangle = |B_0^{-1/2}A\xi|^2 \geqslant \alpha|\xi|^2, \quad \forall \xi \in \mathbb{R}^n.$$

Thus $A^*B_0^{-1}A$ is invertible and it follows that, as $\gamma \to 0$, the posterior mean converges to

$$m^+ = (A^*B_0^{-1}A)^{-1}A^*B_0^{-1}y$$

and the posterior covariance converges to zero. By Lemma 6.5 we deduce the desired weak convergence of μ^y to δ_{m^+}. It remains to characterize m^+.

Since the null space of A is empty, minimizers of

$$\phi(u) := \frac{1}{2}\|B_0^{-1/2}(y - Au)\|^2$$

are unique and satisfy the normal equations

$$A^*B_0^{-1}Au = A^*B_0^{-1}y.$$

Hence m^+ solves the desired least-squares problem and the proof is complete. □

Proof of Theorem 2.5. By Lemma 6.5 we see that it suffices to prove that the mean m and covariance Σ given by the formulae in Theorem 2.4 converge to m^+ and Σ^+ given by (2.22). We start by studying the covariance matrix which, by Theorem 2.4, is given by

$$\Sigma = \left(\frac{1}{\gamma^2}A^*B_0^{-1}A + L_0\right)^{-1}.$$

Using the definition (2.18) of A, we see that

$$A^*B_0^{-1}A = Q\begin{pmatrix} A_0^*B_0^{-1}A_0 & 0 \\ 0 & 0 \end{pmatrix}Q^*.$$

Then, by (2.19) we have

$$\Sigma^{-1} = Q\begin{pmatrix} \frac{1}{\gamma^2}A_0^*B_0^{-1}A_0 + L_{11} & L_{12} \\ L_{12}^* & L_{22} \end{pmatrix}Q^*.$$

Applying the Schur complement formula for the inverse of a matrix as in Lemma 6.21, we deduce that

$$\Sigma = Q\begin{pmatrix} \gamma^2(A_0^*B_0^{-1}A_0)^{-1} & 0 \\ -\gamma^2 L_{22}^{-1}L_{12}^*(A_0^*B_0^{-1}A_0)^{-1} & L_{22}^{-1} \end{pmatrix}Q^* + \Delta, \tag{2.23}$$

where

$$\frac{1}{\gamma^2}\big(|\Delta_{11}| + |\Delta_{21}|\big) \to 0$$

as $\gamma \to 0$, and there is a constant $C > 0$ such that

$$|\Delta_{12}| + |\Delta_{22}| \leqslant C\gamma^2$$

for all γ sufficiently small. From this it follows that, as $\gamma \to 0$,

$$\Sigma \to Q\begin{pmatrix} 0 & 0 \\ 0 & L_{22}^{-1} \end{pmatrix}Q^* := \Sigma^+,$$

writing Q as in (2.20). We see that $\Sigma^+ = Q_2 L_{22}^{-1}Q_2^*$, as required.

We now return to the mean. By Theorem 2.4 this is given by the formula

$$m = \Sigma\big(A^*B^{-1}y + \Sigma_0^{-1}m_0\big).$$

Using the expression $A = (A_0\ 0)Q^*$, we deduce that

$$m \to \Sigma\left(\frac{1}{\gamma^2}Q\begin{pmatrix} A_0^*B_0^{-1} \\ 0 \end{pmatrix}y + \Sigma_0^{-1}m_0\right).$$

By definition of w, w', we deduce that

$$m = \Sigma Q\begin{pmatrix} \frac{1}{\gamma^2}A_0^*B_0^{-1}y + w \\ w' \end{pmatrix}.$$

Using equation (2.23), we find that

$$m = Q\begin{pmatrix} z \\ -L_{22}^{-1}L_{12}^*z + L_{22}^{-1}w' \end{pmatrix} = Q\begin{pmatrix} z \\ z' \end{pmatrix} := m^+.$$

This completes the proof. □

2.4. Common structure

In the previous subsection we showed that, for finite-dimensional problems, Bayes' rule gives the relationship (2.6) between the prior and posterior p.d.f.s π_0 and π^y respectively. Expressed in terms of the measures μ^y and μ_0 corresponding to these densities, the formula may be written as in (2.9):

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{Z(y)} \exp(-\Phi(u;y)). \tag{2.24}$$

The normalization constant $Z(y)$ is chosen so that μ^y is a probability measure:

$$Z(y) = \int_X \exp(-\Phi(u;y))\, \mathrm{d}\mu_0(u). \tag{2.25}$$

It is this form which generalizes readily to the setting on function space where there are no densities π^y and π_0 with respect to Lebesgue measure, but where μ^y has a Radon–Nikodym derivative (see Theorem 6.2) with respect to μ_0.

In Section 3 we will describe a range of inverse problems which can be formulated in terms of finding, and characterizing the properties of, a probability measure μ^y on a separable Banach space $(X, \|\cdot\|_X)$, specified via its Radon–Nikodym derivative with respect to a reference measure μ_0 as in (2.24) and (2.25). In this subsection we highlight the common framework into which many of these problems can be placed, by studying conditions on Φ which arise naturally in a wide range of applications. This framework will then be used to develop a general theory for inverse problems in Section 4. It is important to note that, when studying inverse problems, the properties of Φ that we highlight in this section are typically determined by the *forward PDE problem*, which maps the unknown function u to the data y. In particular, probability theory does not play a direct role in verifying these properties of Φ. Probability becomes relevant when choosing the prior measure so that it charges the Banach space X, on which the desired properties of Φ hold, with full measure. We illustrate how to make such choices of prior in Section 3.

We assume that the data y is in a separable Banach space $(Y, \|\cdot\|_Y)$. When applying the framework outlined in this article we will always assume that the prior measure is Gaussian: $\mu_0 \sim \mathcal{N}(m_0, \mathcal{C}_0)$. The properties of Gaussian random measures on Banach space, and Gaussian random fields in particular, may be found in Sections 6.3, 6.4 and 6.5. The two key properties of the prior that we will use repeatedly are the tail properties of the measure as encapsulated in the Fernique Theorem (Theorem 6.9), and the ability to establish regularity properties from the covariance operator: see Theorem 6.24 and Lemmas 6.25 and 6.27. It is therefore possible to broaden the scope of this material to non-Gaussian priors, for any measures

for which analogues of these two key properties hold. However, Gaussian priors do form an important class of priors for a number of reasons: they are relatively simple to define through covariance operators defined as fractional inverse powers of differential operators; they are relatively straightforward to sample from; and the Hölder and Sobolev regularity properties of functions drawn from the prior are easily understood.

The properties of Φ may be formalized through the following assumptions, which we verify on a case-by-case basis for many of the PDE inverse problems encountered in Section 3.

Assumption 2.6. The function $\Phi : X \times Y \to \mathbb{R}$ has the following properties.

(i) For every $\varepsilon > 0$ and $r > 0$ there is an $M = M(\varepsilon, r) \in \mathbb{R}$ such that, for all $u \in X$ and all $y \in Y$ with $\|y\|_Y < r$,

$$\Phi(u; y) \geqslant M - \varepsilon \|u\|_X^2.$$

(ii) For every $r > 0$ there is a $K = K(r) > 0$ such that, for all $u \in X$ and $y \in Y$ with $\max\{\|u\|_X, \|y\|_Y\} < r$,

$$\Phi(u; y) \leqslant K.$$

(iii) For every $r > 0$ there is an $L(r) > 0$ such that, for all $u_1, u_2 \in X$ and $y \in Y$ with $\max\{\|u_1\|_X, \|u_2\|_X, \|y\|_Y\} < r$,

$$|\Phi(u_1; y) - \Phi(u_2; y)| \leqslant L \|u_1 - u_2\|_X.$$

(iv) For every $\varepsilon > 0$ and $r > 0$ there is a $C = C(\varepsilon, r) \in \mathbb{R}$ such that, for all $y_1, y_2 \in Y$ with $\max\{\|y_1\|_Y, \|y_2\|_Y\} < r$, and for all $u \in X$,

$$|\Phi(u; y_1) - \Phi(u; y_2)| \leqslant \exp\bigl(\varepsilon \|u\|_X^2 + C\bigr) \|y_1 - y_2\|_Y. \qquad \diamond$$

These assumptions are, in turn, a lower bound, an upper bound and Lipschitz properties in u and in y. When Y is finite-dimensional and the observational noise is $\mathcal{N}(0, \Gamma)$, then Φ has the form

$$\begin{aligned} \Phi(u; y) &= \frac{1}{2} \bigl| \Gamma^{-1/2} \bigl(y - \mathcal{G}(u) \bigr) \bigr|^2 \\ &= \frac{1}{2} \bigl| \bigl(y - \mathcal{G}(u) \bigr) \bigr|_\Gamma^2. \end{aligned} \tag{2.26}$$

It is then natural to derive the bounds and Lipschitz properties of Φ from properties of $\mathcal{G}$.

Assumption 2.7. The function $\mathcal{G} : X \to \mathbb{R}^q$ satisfies the following.

(i) For every $\varepsilon > 0$ there is an $M = M(\varepsilon) \in \mathbb{R}$ such that, for all $u \in X$,

$$|\mathcal{G}(u)|_\Gamma \leqslant \exp\bigl(\varepsilon \|u\|_X^2 + M \bigr).$$

(ii) For every $r > 0$ there is a $K = K(r) > 0$ such that, for all $u_1, u_2 \in X$ with $\max\{\|u_1\|_X, \|u_2\|_X\} < r$,

$$|\mathcal{G}(u_1) - \mathcal{G}(u_2)|_\Gamma \leqslant K \|u_1 - u_2\|_X. \qquad \diamond$$

It is straightforward to see the following.

Lemma 2.8. Assume that $\mathcal{G} : X \to \mathbb{R}^q$ satisfies Assumption 2.7. Then $\Phi : X \times \mathbb{R}^q \to \mathbb{R}$ given by (2.26) satisfies Assumption 2.6 with $(Y, \|\cdot\|_Y) = (\mathbb{R}^q, |\cdot|_\Gamma)$. ◇

Many properties follow from these assumptions concerning the density between the posterior and the prior. Indeed, the fact that μ^y is well-defined is typically established by using the continuity properties of $\Phi(\cdot; y)$. Further properties following from these assumptions include continuity of μ^y with respect to the data y, and desirable perturbation properties of μ^y based on finite-dimensional approximation of Φ or $\mathcal{G}$. All these properties will be studied in detail in Section 4. We emphasize that many variants on the assumptions above could be used to obtain similar, but sometimes weaker, results than those appearing in this article. For example, we work with Lipschitz continuity of Φ in both arguments; similar results can be proved under the weaker assumptions of continuity in both arguments. However, since Lipschitz continuity holds for most of the applications of interest to us, we work under these assumptions.

We re-emphasize that the properties of Φ encapsulated in Assumption 2.6 are properties of the forward PDE problem, and they do not involve inverse problems and probability at all. The link to Bayesian inverse problems comes through the choice of prior measure μ_0 which, as we will see in Sections 3 and 4, should be chosen so that $\mu_0(X) = 1$; this means that functions drawn at random from the prior measure should be sufficiently regular that they lie in X with probability one, so that the properties of Φ from Assumptions 2.6 apply to it. In the function space setting, regularity of the mean function, together with the spectral properties of the covariance operator, determines the regularity of random draws. In particular, the rate of decay of the eigenvalues of the covariance operator plays a central role in determining the regularity properties. These issues are discussed in detail in Section 6.5. For simplicity we will work throughout with covariance operators which are defined through (possibly fractional) negative powers of the Laplacian, or operators that behave like the Laplacian in a sense made precise below.

To make these ideas precise, consider a second-order differential operator $\mathcal{A}$ on a bounded open set $D \subset \mathbb{R}^d$, with domain chosen so that $\mathcal{A}$ is positive definite and invertible. Let $\mathcal{H} \subset L^2(D)$. For example, $\mathcal{H}$ may be restricted to the subspace where

$$\int_D u(x)\,\mathrm{d}x = 0 \tag{2.27}$$

holds, in order to enforce positivity for an operator with Neumann or periodic boundary conditions, which would otherwise have constants in its

kernel; or it may be restricted to divergence-free fields when incompressible fluid flow is being modelled.

We let $\{(\phi_k, \lambda_k)\}_{k\in\mathbb{K}}$ denote a complete orthonormal basis for $\mathcal{H}$, comprising eigenfunctions/eigenvalues of $\mathcal{A}$. Then $\mathbb{K} \subseteq \mathbb{Z}^d\backslash\{0\}$. For Laplacian-like operators we expect that the eigenvalues will grow like $|k|^2$ and that, in simple geometries, the ϕ_k will be bounded in L^∞ and the gradient of the ϕ_k will grow like $|k|$ in L^∞. We make these ideas precise below. For all infinite sums over $\mathbb{K}$ in the following we employ standard orderings.

For any $u \in \mathcal{H}$ we may write

$$u = \sum_{k\in\mathbb{K}} \langle u, \phi_k\rangle \phi_k.$$

We may then define fractional powers of $\mathcal{A}$ as follows, for any $\alpha \in \mathbb{R}$:

$$\mathcal{A}^\alpha u = \sum_{k\in\mathbb{K}} \lambda_k^\alpha \langle u, \phi_k\rangle \phi_k. \tag{2.28}$$

For any $s \in \mathbb{R}$ we define the separable Hilbert spaces $\mathcal{H}^s$ by

$$\mathcal{H}^s = \left\{ u : \sum_{k\in\mathbb{K}} \lambda_k^s |\langle u, \phi_k\rangle|^2 < \infty \right\}. \tag{2.29}$$

These spaces have norm $\|\cdot\|_s$ defined by

$$\|u\|_s^2 = \sum_{k\in\mathbb{K}} \lambda_k^s |\langle u, \phi_k\rangle|^2.$$

If $s \geqslant 0$ then these spaces are contained in $\mathcal{H}$, but for $s < 0$ they are larger than $\mathcal{H}$. The following assumptions characterize a 'Laplacian-like' operator. These operators will be useful to us when constructing Gaussian priors, as they will enable us to specify regularity properties of function drawn from the prior in a transparent fashion.

Assumption 2.9. The operator $\mathcal{A}$, densely defined on a Hilbert space $\mathcal{H} \subset L^2(D; \mathbb{R}^n)$, satisfies the following properties.

(i) $\mathcal{A}$ is positive definite, self-adjoint and invertible.

(ii) The eigenfunctions/eigenvalues $\{\phi_k, \lambda_k\}_{k\in\mathbb{K}}$ of $\mathcal{A}$, indexed by $k \in \mathbb{K} \subset \mathbb{Z}^d\backslash\{0\}$, form an orthonormal basis for $\mathcal{H}$.

(iii) There exist $C^\pm > 0$ such that the eigenvalues satisfy, for all $k \in \mathbb{K}$,

$$C^- \leqslant \frac{\lambda_k}{|k|^2} \leqslant C^+.$$

(iv) There exists $C > 0$ such that

$$\sup_{k\in\mathbb{K}} \left(\|\phi_k\|_{L^\infty} + \frac{1}{|k|} \|D\phi_k\|_{L^\infty} \right) \leqslant C. \qquad \diamond$$

Note that if $\mathcal{A}$ is the Laplacian with Dirichlet or Neumann boundary conditions, then the spaces $\mathcal{H}^s$ are contained in the usual Sobolev spaces H^s. In the case of periodic boundary conditions they are identical to the Sobolev spaces H^s_{per}. Thus the final assumption (v) above is a generalization of the following Sobolev Embedding Theorem for the Laplacian.

Theorem 2.10. (Sobolev Embedding Theorem) Assume that $\mathcal{A} := -\triangle$ is equipped with periodic, Neumann or Dirichlet boundary conditions on the unit cube. If $u \in \mathcal{H}^s$ and $s > d/2$, then $u \in C(\overline{D})$, and there is a constant $C > 0$ such that $\|u\|_{L^\infty} \leqslant C\|u\|_s$. ◇

2.5. Discussion and bibliography

An introduction to the Bayesian approach to statistical problems in general is Bernardo and Smith (1994). The approach taken to Bayesian inverse problems as outlined in Kaipio and Somersalo (2005) is to first discretize the problem and then secondly apply the Bayesian methodology to a finite-dimensional problem. This is a commonly adopted methodology. In that approach, the idea of trying to capture the limit of infinite resolution is addressed by use of statistical extrapolation techniques based on modelling the error from finite-dimensional approximation (Kaipio and Somersalo 2007*b*). The approach that is developed in this article reverses the order of these two steps: we first apply the Bayesian methodology to an infinite-dimensional problem, and then discretize. There is some literature concerning the Bayesian viewpoint for linear inverse problems on function space, including the early study by Franklin (1970), and the subsequent papers by Mandelbaum (1984), Lehtinen, Paivarinta and Somersalo (1989) and Fitzpatrick (1991); the paper by Lassas *et al.* (2009) contains a good literature review of this material, and further references. The papers of Lassas *et al.* (2009) and Lassas and Siltanen (2004) also study Bayesian inversion for linear inverse problems on function space; they introduce the notion of *discretization invariance* and investigate the question of whether it is possible to derive regularizations of families of finite-dimensional problems, in a fashion which ensures that meaningful limits are obtained; this idea also appears somewhat earlier in the data assimilation literature, for a particular PDE inverse problem, in the paper of Bennett and Budgell (1987). In the approach taken in this article, discretization invariance is guaranteed for finite-dimensional approximations of the function space Bayesian inverse problem. Furthermore, our approach is not limited to problems in which a Gaussian posterior measure appears; in contrast, existing work on discretization invariance is confined to the linear, Gaussian observational noise setting in which the posterior is Gaussian if the prior is Gaussian.

The least-squares approach to inverse problems encapsulated in (2.3) is often termed *Tikhonov regularization* (Engl, Hanke and Neubauer 1996) and,

more generally, the *variational method* in the applied literature (Bennett 2002, Evensen 2006). The book by Engl *et al.* (1996) discusses regularization techniques in the Hilbert space setting and the Banach space setting is discussed in, for example, the recent papers of Kaltenbacher, Schöpfer and Schuster (2009), Neubauer (2009) and Hein (2009). As we demonstrated, regularization is closely related to finding the MAP estimator as defined in Kaipio and Somersalo (2005). As such it is clear that, from the Bayesian standpoint, regularization is intimately related to the choice of prior. Another classical regularization method for linear inverse problems is through iterative solution (Engl *et al.* 1996); this topic is related to the Bayesian approach to inverse problems in Calvetti (2007) and Calvetti and Somersalo (2005*a*).

Although we concentrate in this paper on Gaussian priors, and hence on regularization via addition of a quadratic penalization term, there is active research in the use of different regularizations (Kaltenbacher *et al.* 2009, Neubauer 2009, Hein 2009, Lassas and Siltanen 2004). In particular, the use of total variation-based regularization, and related wavelet-based regularizations, is central in image processing (Rudin *et al.* 1992, Scherzer *et al.* 2009). We will not address such regularizations in this article, but note that the development of a function space Bayesian viewpoint on such problems, along the lines developed here for Gaussian priors, is an interesting research direction (Lassas *et al.* 2009).

Theorem 2.4 concerns the small noise limit for Gaussian noise. This topic has been studied in greater detail in the papers by Engl, Hofinger and Kindermann (2005), Hofinger and Pikkarainen (2007, 2009) and Neubauer and Pikkarainen (2008), where the convergence of the posterior distribution is quantified by use of the Prokohorov and Ky Fan metrics. Gaussian problems are often amenable to closed-form analysis, as illustrated in this section, and are hence useful for illustrative purposes. Furthermore, there are many interesting applications where Gaussian structure prevails. Thus we will, on occasion, exploit Gaussianity throughout the article, for both these reasons.

The common structure underlying a wide range of Bayesian inverse problems for functions, and which we highlight in Section 2.4, is developed in Cotter, Dashti, Robinson and Stuart (2009, 2010*b*) and Cotter, Dashti and Stuart (2010*a*).

In the general framework established at the start of this section we have implicitly assumed that the observation operator $\mathcal{G}(\cdot)$ is known to us. In practice it is often approximated by some computer code $\mathcal{G}(\cdot;h)$ in which h denotes a mesh parameter, or parameter controlling missing physics. In this case (2.5) can be replaced by the equation

$$y = \mathcal{G}(u;h) + \varepsilon + \eta, \tag{2.30}$$

where $\varepsilon := \mathcal{G}(u) - \mathcal{G}(u;h)$. Whilst it is possible to lump ε and η together

into one single error term, and work again with equation (2.1), this can be misleading because the observation error η and the computational model error ε are very different in character. The latter is typically not mean zero, and depends upon u; in contrast it is frequently realistic to model η as a mean zero random variable, independent of u. Attempts to model the effects of ε and η separately may be found in a number of publications, including Kaipio and Somersalo (2005, Chapter 7), Kaipio and Somersalo (2007*a*), Kaipio and Somersalo (2007*b*), Glimm, Hou, Lee, Sharp and Ye (2003), Orrell, Smith, Barkmeijer and Palmer (2001), Kennedy and O'Hagan (2001), O'Sullivan and Christie, (2006*b*, 2006*a*), Christie, Pickup, O'Sullivan and Demyanov (2008) and Christie (2010). A different approach to dealing with model error is to extend the variable u to include model terms which represent missing physics or lack of resolution in the model and to try to learn about such systematic error from the data; this approach is undertaken in Cotter, Dashti, Robinson and Stuart (2009).

3. Examples

3.1. Overview

In this section we study a variety of inverse problems arising from boundary value problems and initial-boundary value problems. Our goal is to enable application of the framework for Bayesian inverse problems on function space that is developed in Section 4, in order to justify a formula of the form (2.24) for a posterior measure μ^y on a function space, and to establish properties of the measure μ^y.

In order to carry this out it is desirable to establish that, for a wide range of problems, the common structure encapsulated in Assumptions 2.6 or 2.7 may be shown to hold. These assumptions concern properties of the *forward problem* underlying the inverse problem, and have no reference to the inverse problem, its Bayesian formulation or to probability. The link between the forward problem and the Bayesian inverse problem is provided in this section, and in the next section. In this section we show that choosing the prior measure so that $\mu_0(X) = 1$, where X is the space in which Assumptions 2.6 or 2.7 may be shown to hold, ensures that the posterior measure is well-defined; this may often be done by use of Theorem 6.31. The larger the space X, the fewer restrictions the condition $\mu_0(X) = 1$ places on the choice of prior, since it is equivalent to asking that draws from μ_0 are almost surely in the space X; the larger X is, the easier this is to satisfy. The next section is concerned with ramifications of Assumptions 2.6 or 2.7 for various stability properties of the posterior measure μ^y with respect to perturbations.

We will work in a Banach space setting and will always specify the prior measure as a Gaussian. The required background material on Gaussian

measures in Banach space, and Gaussian random fields, may be found in Section 6. We also make regular use of the key Theorem 6.31, from Section 6.6, to show that the posterior is well-defined and absolutely continuous with respect to the prior. For simplicity we work with priors whose covariance operator is a fractional negative power of an operator such as the Laplacian. The reader should be aware that much greater generality than this is possible and that the simple setting for choice of priors is chosen for expository purposes. Other Gaussian priors may be chosen so long as the constraint $\mu_0(X) = 1$ is satisfied.

We start in Section 3.2 by studying the inverse problems of determining a field from *direct* pointwise observations. We use this example to illustrate our approach to identifying the Radon–Nikodym derivative between posterior and prior measures. All of the subsequent subsections in this chapter involve Bayesian inference for random fields, but in contrast to the first subsection they are based on *indirect* measurements defined through solution of a differential equation. In Section 3.3 we study the problem of finding the diffusion coefficient in a two-point boundary value problem, from observations of the solution. In Section 3.4 we consider the problem of determining the wave speed for the scalar wave equation from observations of the solution. Section 3.5 concerns the problem of recovering the initial condition for the heat equation, from observation of the entire solution at a positive time, when polluted by an additive Gaussian random field. We then describe several more involved examples arising in applications such as fluid mechanics, geophysics and molecular dynamics, all of which can be placed in the common framework developed here, but for which space precludes a full development of the details; see Sections 3.6, 3.7 and 3.8. The problems in fluid mechanics are natural extensions of the inverse problem for the initial condition of the heat equation, and those arising in subsurface geophysics generalize the inverse problem for the diffusion coefficient in a two-point boundary value problem. The problem in molecular dynamics is somewhat different, as it does not arise from a Bayesian inverse problem but rather from a conditioned diffusion process. However, the resulting mathematical structure shares much with the inverse problems and we include it for this reason. References to some of the relevant literature on these applications are given in Section 3.9.

3.2. Pointwise data for a random field

Let $D \subset \mathbb{R}^d$ be a bounded open set. Consider a field $u : D \to \mathbb{R}^n$. We view u as an element of the Hilbert space $\mathcal{H} = L^2(D)$. Assume that we are given noisy observations $\{y_k\}_{k=1}^q$ of a function $g : \mathbb{R}^n \to \mathbb{R}^\ell$ of the field at a set of points $\{x_k\}_{k=1}^q$. Thus

$$y_k = g(u(x_k)) + \eta_k, \tag{3.1}$$

where the $\{\eta_k\}_{k=1}^q$ describe the observational noise. Concatenating data, we have

$$y = \mathcal{G}(u) + \eta, \tag{3.2}$$

where $y = (y_1^*, \dots, y_q^*)^* \in \mathbb{R}^{\ell q}$ and $\eta = (\eta_1^*, \dots, \eta_q^*)^* \in \mathbb{R}^{\ell q}$. The observation operator $\mathcal{G}$ maps $X := C(\overline{D}) \subset \mathcal{H}$ into $Y := \mathbb{R}^{\ell q}$. The inverse problem is to reconstruct the field u from the data y.

We assume that the observational noise η is Gaussian $\mathcal{N}(0, \Gamma)$. We specify a prior measure μ_0 on u which is Gaussian $\mathcal{N}(m_0, \mathcal{C}_0)$ and determine the posterior measure μ^y for u given y. Since $\mathbb{P}(\mathrm{d}y|u) = \mathcal{N}(\mathcal{G}(u), \Gamma)$, informal application of Bayes' rule leads us to expect that the Radon–Nikodym derivative of μ^y with respect to μ_0 is

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2}|y - \mathcal{G}(u)|_\Gamma^2\right). \tag{3.3}$$

Below we deduce appropriate choices of prior measure which ensure that this measure is well-defined and does indeed determine the desired posterior distribution for u given y.

If $g : \mathbb{R}^n \to \mathbb{R}^\ell$ is linear, so that $\mathcal{G}(u) = Au$ for some linear operator $A : X \to \mathbb{R}^{\ell q}$, then the calculations in Example 6.23 show that the posterior measure μ^y is also Gaussian with $\mu^y = \mathcal{N}(m, \mathcal{C})$ where

$$m = m_0 + \mathcal{C}_0 A^* (\Gamma + A\mathcal{C}_0 A^*)^{-1}(y - Am_0), \tag{3.4a}$$

$$\mathcal{C} = \mathcal{C}_0 - \mathcal{C}_0 A^* (\Gamma + A\mathcal{C}_0 A^*)^{-1} A\mathcal{C}_0. \tag{3.4b}$$

Let $\triangle$ denote the Laplacian on D, with domain chosen so that Assumptions 2.9(i)–(iv) hold. Recall the (Sobolev-like) spaces $\mathcal{H}^s$ from (2.29). The following theorem is proved by application of Theorem 6.31, which the reader is encouraged to study before continuing in this section.

Theorem 3.1. Assume that the domain of $-\triangle$ is chosen so that Assumptions 2.9(i)–(v) hold. Let $g : \mathbb{R}^n \to \mathbb{R}^\ell$ be continuous. Assume that $\mathcal{C}_0 \propto (-\triangle)^{-\alpha}$ with $\alpha > d/2$ and assume that $m_0 \in \mathcal{H}^\alpha$. Then $\mu^y(\mathrm{d}u) = \mathbb{P}(\mathrm{d}u|y)$ is absolutely continuous with respect to $\mu_0(\mathrm{d}u) = \mathcal{N}(m_0, \mathcal{C}_0)$ with Radon–Nikodym derivative given by (3.3). Furthermore, when g is linear, so that $\mathcal{G}(u) = Au$ for some linear $A : X \to \mathbb{R}^{\ell q}$, then the posterior is Gaussian with mean and covariance given by (3.4).

Proof. The formulae for the mean and covariance of the Gaussian posterior measure $\mu^y = \mathcal{N}(m, \mathcal{C})$, which arises when g is linear, follow from Example 6.23. We now proceed to determine the posterior measure in the non-Gaussian case. Define $L_z : X \to \mathbb{R}^n$ to be the pointwise evaluation operator at $z \in D$. Notice that

$$|L_z u - L_z v| = |u(z) - v(z)| \leqslant \|u - v\|_{L^\infty}$$

so that $L_z : X \to \mathbb{R}^n$ is continuous. The function $\mathcal{G}$ is found by composing the continuous function g with the operator $L.$ at a finite set of points and is thus itself continuous from X into $\mathbb{R}^{\ell q}$. To apply Theorem 6.31 it suffices to show that $\mu_0(X) = 1$. Since $\mathcal{H}^\alpha$ is the Cameron–Martin space for $\mathcal{C}_0$ and since $m_0 \in \mathcal{H}^\alpha$, we deduce that $\mu_0 = \mathcal{N}(m_0, \mathcal{C}_0)$ and $\mathcal{N}(0, \mathcal{C}_0)$ are equivalent as measures, by Theorem 6.13. Thus $\mu_0(X) = 1$ since, by Lemma 6.25, draws from $\mathcal{N}(0, \mathcal{C}_0)$ are a.s. s-Hölder for all $s \in (0, \min\{1, \alpha - d/2\})$. □

In Section 2.4 we indicated that obtaining bounds and Lipschitz properties of $\mathcal{G}$ or Φ, the mappings appearing in the Radon–Nikodym derivative between μ^y and μ_0, will be important to us below. The following lemma studies this issue.

Lemma 3.2. In the setting of Theorem 3.1 assume, in addition, that $g \in C^1(\mathbb{R}^n, \mathbb{R}^\ell)$ and that g is polynomially bounded. Then $\mathcal{G}$ satisfies Assumption 2.7 with $X = C(\overline{D})$ and $Y = \mathbb{R}^{\ell q}$. Furthermore, if Dg is polynomially bounded then $K(r)$ is polynomially bounded.

Proof. Since g is polynomially bounded and $\mathcal{G}$ is found by pointwise evaluation at a finite number of points, it follows that

$$|\mathcal{G}(u)| \leqslant p(\|u\|_X)$$

for some polynomial $p : \mathbb{R} \to \mathbb{R}$. The bound (i) of Assumption 2.7 follows. By the mean-value theorem (Taylor theorem with remainder) we have that

$$|\mathcal{G}(u) - \mathcal{G}(v)|_\Gamma \leqslant \max_{1 \leqslant k \leqslant K} \left| \int_0^1 Dg\big(su(x_k) + (1-s)v(x_k)\big)\, \mathrm{d}s \big(u(x_k)) - v(x_k)\big) \right|.$$

Thus, for all u, v satisfying $\max\{\|u\|_X, \|v\|_X\} < r$,

$$|\mathcal{G}(u) - \mathcal{G}(v)|_\Gamma \leqslant K(r)\|u - v\|_X.$$

Furthermore, K may be bounded polynomially if Dg is bounded polynomially. The result follows. □

3.3. Inverse problem for a diffusion coefficient

The previous example illustrated the formulation of an inverse problem for a function, using the Bayesian framework. However, the observations of the function comprised direct measurements of the function at points in its domain D. We now consider a problem where the measurements are more indirect, and are defined through the solution of a differential equation.

We consider the inverse problem of determining the diffusion coefficient from observations of the solution of the two-point boundary value problem

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(k(x)\frac{\mathrm{d}p}{\mathrm{d}x}\right) = 0, \tag{3.5a}$$

$$p(0) = p^-, \quad p(1) = p^+. \tag{3.5b}$$

We assume that $p^+ > p^- > 0$ and we make observations of $\{p(x_k)\}_{k=1}^q$, at a set of points $0 < x_1 < \cdots < x_q < 1$ subject to Gaussian measurement error. We write the observations as

$$y_j = p(x_j) + \eta_j, \quad j = 1, \ldots, q \tag{3.6}$$

and, for simplicity, assume that the η_k form an i.i.d. sequence with $\eta_1 \sim \mathcal{N}(0, \gamma^2)$. Our interest is in determining the diffusion coefficient k from y. To ensure that k is strictly positive on $[0,1]$, we introduce $u(x) = \ln(k(x))$ and view $u \in L^2((0,1))$ as the basic unknown function.

The forward problem (3.5) for p given u is amenable to considerable explicit analysis, and we now use this to write down a formula for the observation operator $\mathcal{G}$ and to study its properties. We first define $J_x : L^\infty((0,1)) \to \mathbb{R}$ by

$$J_x(w) = \int_0^x \exp(-w(z))\,\mathrm{d}z. \tag{3.7}$$

The solution of (3.5) may be written as

$$p(x) = (p^+ - p^-)\frac{J_x(u)}{J_1(u)} + p^- \tag{3.8}$$

and is monotonic increasing; furthermore, $p(x)$ is unchanged under $u(x) \to u(x) + \lambda$ for any $\lambda \in \mathbb{R}$. The observation operator is then given by the formula

$$\mathcal{G}(u) = \big(p(x_1), \ldots, p(x_q)\big)^*. \tag{3.9}$$

Lemma 3.3. The observation operator $\mathcal{G} : C([0,1]) \to \mathbb{R}^q$ is Lipschitz and satisfies the bound

$$|\mathcal{G}(u)| \leqslant \sqrt{q}p^+. \tag{3.10}$$

Indeed, $\mathcal{G}$ satisfies Assumption 2.7 with $X = C([0,1])$ and $K(\cdot)$ exponentially bounded: there are $a, b > 0$ such that $K(r) \leqslant a\exp(br)$.

Proof. The fact that $\mathcal{G}$ is defined on $C([0,1])$ follows from the explicit solution given in equation (3.8). The bound on $\mathcal{G}$ follows from the monotonicity of the solution. For the Lipschitz property it suffices to consider the case $q = 1$ and, without loss of generality, take $x_1 = 1/2$. Note that then

$$\begin{aligned}
\frac{|\mathcal{G}(u) - \mathcal{G}(v)|}{p^+ - p^-} &= \frac{1}{J_1(u)J_1(v)}\big|J_{\frac12}(u)J_1(v) - J_{\frac12}(v)J_1(u)\big| \\
&= \frac{1}{J_1(u)J_1(v)}\big|J_{\frac12}(u)\big(J_1(v) - J_1(u)\big) + J_1(u)\big(J_{\frac12}(u) - J_{\frac12}(v)\big)\big| \\
&\leqslant J_1(v)^{-1}|J_1(v) - J_1(u)| + J_1(v)^{-1}\big|J_{\frac12}(u) - J_{\frac12}(v)\big|.
\end{aligned}$$

But

$$J_1(v)^{-1} \leqslant \exp\big(\|v\|_\infty\big)$$

and

$$|J_x(u) - J_x(v)| \leqslant x \exp\bigl(\max\{\|u\|_\infty, \|v\|_\infty\}\bigr)\|u - v\|_\infty.$$

Thus we deduce that

$$|\mathcal{G}(u) - \mathcal{G}(v)| \leqslant \frac{3}{2}(p^+ - p^-)\exp\bigl(\|v\|_\infty + \max\{\|u\|_\infty, \|v\|_\infty\}\bigr)\|u - v\|_\infty. \quad \square$$

We place a Gaussian prior measure $\mu_0 \sim \mathcal{N}(u_0, \mathcal{C}_0)$ on u. We say that k is *log-normal*. Since changing u by an arbitrary additive constant does not change the solution of (3.5), we cannot expect to determine any information about the value of this constant from the data. Thus we must build our assumptions about this constant into the prior. To do this we assume that u integrates to zero on $(0,1)$ and define the prior measure μ_0 on the space

$$\mathcal{H} = \left\{ u \in L^2((0,1)) \middle| \int_0^1 u(x)\,\mathrm{d}x = 0 \right\}. \tag{3.11}$$

We define $A = -\mathrm{d}^2/\mathrm{d}x^2$ to be a densely defined operator on $\mathcal{H}$ with

$$D(A) = \left\{ u \in H^2_{\mathrm{per}}((0,1)) \middle| \int_0^1 u(x)\,\mathrm{d}x = 0 \right\}.$$

Then A is positive definite self-adjoint and, for any $\beta > 0$ and $\alpha > 1/2$ (which ensures that the covariance operator is trace-class), we may define the Gaussian measure $\mathcal{N}(m_0, \beta A^{-\alpha})$ on $\mathcal{H}$.

We have

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

where $y = (y_1, \ldots, y_q)^* \in \mathbb{R}^q$ and $\eta \in \mathbb{R}^q$ is distributed as $\mathcal{N}(0, \gamma^2 I)$. The probability of y given u (the data likelihood) is

$$\mathbb{P}(y|u) \propto \exp\left(-\frac{1}{2\gamma^2}|y - \mathcal{G}(u)|^2\right).$$

We wish to find $\mu^y(\mathrm{d}u) = \mathbb{P}(\mathrm{d}u|y)$. Informal use of Bayes' rule suggests that

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2\gamma^2}|y - \mathcal{G}(u)|^2\right). \tag{3.12}$$

We now justify this formula. Since $\mathcal{G}(\cdot)$ is Lipschitz on $X := C([0,1])$, by Lemma 3.3, the basic idea underlying the justification of (3.12) in the next theorem is to choose α so that $\mu_0(X) = 1$, so that we may apply Theorem 6.31.

Theorem 3.4. Consider the Bayesian inverse problem for $u(x) = \ln(k(x))$ subject to observation in the form (3.6), with p solving (3.5), and prior measure $\mu_0 = \mathcal{N}(m_0, \mathcal{C}_0)$ with $m_0 \in \mathcal{H}^\alpha$ and $\mathcal{C}_0 = \beta A^{-\alpha}$. If $\beta > 0$ and $\alpha > 1/2$ then $\mu^y(\mathrm{d}u) = \mathbb{P}(\mathrm{d}u|y)$ is absolutely continuous with respect to $\mu_0(\mathrm{d}u)$ with Radon–Nikodym derivative given by (3.12), with $\mathcal{G}$ defined in (3.9).

Proof. We apply Theorem 6.31. The function $\mathcal{G}$ is continuous from X into $\mathbb{R}^q$. Hence it suffices to show that $\mu_0(X) = 1$. Since $\mathcal{H}^\alpha$ is the Cameron–Martin space for $\mathcal{C}_0$ and since $m_0 \in \mathcal{H}^\alpha$, we deduce that $\mu_0 = \mathcal{N}(m_0, \mathcal{C}_0)$ and $\mathcal{N}(0, \mathcal{C}_0)$ are equivalent as measures, by Theorem 6.13. Thus $\mu_0(X) = 1$ since, by Lemma 6.25, draws from $\mathcal{N}(0, \mathcal{C}_0)$ are a.s. s-Hölder for all $s \in (0, \min\{1, \alpha - 1/2\})$. □

3.4. Wave speed for the wave equation

Consider the equation

$$\frac{\partial v}{\partial t} + c(x)\frac{\partial v}{\partial x} = 0, \quad (x,t) \in \mathbb{R} \times (0,\infty) \tag{3.13a}$$

$$v = f, \quad (x,t) \in \mathbb{R} \times \{0\} \tag{3.13b}$$

We assume that the wave speed $c(x)$ is known to be a positive, one-periodic function, and that we are interested in the inverse problem of determining c given the observations

$$y_j = v(1, t_j) + \eta_j, \quad j = 1, \dots, q. \tag{3.14}$$

We assume that the observational noise $\{\eta_j\}_{j=1}^q$ is mean zero Gaussian. Since c is positive, we write $c = \exp(u)$ and view the inverse problem as being the determination of u. We thus concatenate the data and write

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

where $\eta \sim \mathcal{N}(0, \Gamma)$ and $\mathcal{G} : X \to \mathbb{R}^q$ where $X = C^1(\mathbb{S})$; here $\mathbb{S}$ denotes the unit circle $[0, 1)$ with end points identified to enforce periodicity. We equip X with the norm

$$\|u\|_X = \sup_{x \in \mathbb{S}} |u(x)| + \sup_{x \in \mathbb{S}} \left| \frac{\mathrm{d}u}{\mathrm{d}x}(x) \right|.$$

Note that we may also view u as a function in $X_{\mathrm{per}} := C^1_{\mathrm{per}}(\mathbb{R})$, the space of 1-periodic C^1 functions on $\mathbb{R}$. Before defining the inverse problem precisely, we study the properties of the forward operator $\mathcal{G}$.

Lemma 3.5. Assume that $f \in C^1(\mathbb{R}; \mathbb{R})$ and f is polynomially bounded: there are constants $K > 0$ and $p \in \mathbb{Z}^+$ such that

$$|f(x)| \leqslant K\big(1 + |x|^p\big).$$

Then $\mathcal{G} : X \to \mathbb{R}^q$ satisfies the following conditions.

- There is a constant $C > 0$:

$$|\mathcal{G}(u)| \leqslant C\big(1 + \exp(p\|u\|_X)\big).$$

- For all $u, w \in X : \|u\|_X, \|w\|_X < r$ there exists $L = L(r)$:

$$|\mathcal{G}(u) - \mathcal{G}(w)| \leqslant L\|u - w\|_\infty.$$

Proof. It suffices to consider the case $q = 1$ and take $t_1 = 1$ for simplicity. Let $\Psi(\cdot; t, u) : \mathbb{R} \to \mathbb{R}$ denote the one-parameter group given by the solution operator for the equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\exp(u(x)), \tag{3.15}$$

where we view u as an element of X_{per} in order to define the solution of this equation. Then v solving (3.13) with $c = \exp(u)$ is given by the formula

$$v(x,t) = f\big(\Psi(x;t,u)\big).$$

Thus

$$\mathcal{G}(u) = v(1,1) = f\big(\Psi(1;1,u)\big) \tag{3.16}$$

and

$$|\mathcal{G}(u)| = |v(1,1)| \leqslant K\big(1 + |\Psi(1;1,u)|^p\big).$$

But the solution of (3.15) subject to the condition $x(0) = 1$ satisfies

$$\begin{aligned} |x(1)| &\leqslant 1 + \int_0^1 \exp\big(u(x(s))\big)\,\mathrm{d}s \\ &\leqslant 1 + \exp(\|u\|_X). \end{aligned}$$

Hence

$$\Psi(1;1,u) \leqslant 1 + \exp(\|u\|_X), \tag{3.17}$$

and the first result follows.

For the second result let $x(t) = \Psi(1;t,u)$ and $y(t) = \Psi(1;t,w)$ so that, by (3.15),

$$\begin{aligned} x(t) &= 1 - \int_0^t \exp\big(u(x(s))\big)\,\mathrm{d}s, \\ y(t) &= 1 - \int_0^t \exp\big(u(y(s))\big)\,\mathrm{d}s + \int_0^t \exp\big(u(y(s))\big)\,\mathrm{d}s - \int_0^t \exp\big(w(y(s))\big)\mathrm{d}s. \end{aligned}$$

Thus, using (3.17),

$$|x(t) - y(t)| \leqslant C(\|u\|_X, \|w\|_X)\left(\int_0^1 |x(s) - y(s)|\,\mathrm{d}s + \|u - w\|_\infty\right).$$

Application of the Gronwall inequality gives

$$\sup_{t\in[0,1]} |x(t) - y(t)| \leqslant C(\|u\|_X, \|w\|_X)\|u - w\|_\infty.$$

Thus

$$|\Psi(1;1,u) - \Psi(1,1,w)| \leqslant C(\|u\|_X, \|w\|_X)\|u - w\|_\infty.$$

Hence, using (3.16), the fact that f is C^1 and the bound (3.17), we deduce that

$$\begin{aligned}|\mathcal{G}(u)-\mathcal{G}(w)| &= \big|f\big(\Psi(1;1,u)\big)-f\big(\Psi(1;1,w)\big)\big| \\ &\leqslant L(r)\|u-w\|_\infty \\ &\leqslant L(r)\|u-w\|_X. \end{aligned}$$ □

We wish to find $\mu^y(\mathrm{d}u)=\mathbb{P}(\mathrm{d}u|y)$. Informal use of Bayes' rule gives us

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2}|y-\mathcal{G}(u)|_\Gamma^2\right). \tag{3.18}$$

We now justify this formula by choice of prior and by application of Theorem 6.31.

We place a prior measure μ_0 on the space X by assuming that $u\sim\mu_0$ is Gaussian and that

$$u'(x)=\frac{\mathrm{d}u}{\mathrm{d}x}(x)\sim\mathcal{N}(0,\beta\mathcal{A}^{-\alpha}),$$

where $\mathcal{A}=-\mathrm{d}^2/\mathrm{d}x^2$ is a densely defined operator on $\mathcal{H}=L^2(\mathbb{S})$ with

$$D(\mathcal{A})=\left\{u\in H^2(\mathbb{S})\Big|\int_0^1 u(x)\,\mathrm{d}x=0\right\}.$$

If $\beta>0$ and $\alpha>1/2$ then u' is almost surely a continuous function, by Lemma 6.25. Defining

$$u(x)=u_0+\int_0^x u'(s)\,\mathrm{d}s,$$

where $u_0\sim\mathcal{N}(0,\sigma^2)$, determines the distribution of u completely. Furthermore, for $\beta>0$ and $\alpha>1/2$ we have that u drawn from this measure is in X with probability 1: $\mu_0(X)=1$. Hence we deduce the following result.

Theorem 3.6. Consider the Bayesian inverse problem for $u(x)=\ln(c(x))$ subject to observation in the form (3.14), with v solving (3.13), and prior measure μ_0 as constructed immediately preceding this theorem. If $\beta>0$ and $\alpha>1/2$ then $\mu^y(\mathrm{d}u)=\mathbb{P}(\mathrm{d}u|y)$ is absolutely continuous with respect to $\mu_0(\mathrm{d}u)$, with Radon–Nikodym derivative given by (3.18).

Proof. To apply Theorem 6.31 it suffices to show that $\mu_0(X)=1$, since the function $\mathcal{G}$ is continuous from X into $\mathbb{R}^q$. The fact that $\mu_0(X)=1$ is established immediately prior to this theorem. □

3.5. Initial condition for the heat equation

We now study an inverse problem where the data y is a function, and is hence infinite-dimensional, in contrast to preceding examples where the

data has been finite-dimensional. We assume that our observation is the solution of the heat equation at some fixed positive time $T > 0$, with an added Gaussian random field as observational noise, and that we wish to determine the initial condition.

To be concrete we consider the heat equation on a bounded open set $D \subset \mathbb{R}^d$, with Dirichlet boundary conditions, and written as an ODE in Hilbert space $\mathcal{H} = L^2(D)$:

$$\frac{\mathrm{d}v}{\mathrm{d}t} + Av = 0, \quad v(0) = u. \tag{3.19}$$

Here $A = -\triangle$ with $D(A) = H_0^1(D) \bigcap H^2(D)$. We assume sufficient regularity conditions on D and its boundary ∂D to ensure that the operator A is positive and self-adjoint on $\mathcal{H}$, and is the generator of an analytic semigroup. We define the (Sobolev-like) spaces $\mathcal{H}^s$ as in (2.29).

Assume that we observe the solution v at time T, subject to error which has the form of a Gaussian random field, and that we wish to recover the initial condition u. This problem is classically ill-posed, because the heat equation is smoothing, and inversion of this operator is not continuous on $\mathcal{H}$. Nonetheless, we will construct a well-defined Bayesian inverse problem.

We place a prior measure on u, which is a Gaussian $\mu_0 \sim \mathcal{N}(m_0, \mathcal{C}_0)$ with $\mathcal{C}_0 = \beta A^{-\alpha}$, for some $\beta > 0$ and $\alpha > d/2$; consequently $u \in \mathcal{H}$ μ_0-a.s. by Lemma 6.27. Our aim is to determine conditions on α, and on m_0, which ensure that the Bayesian inverse problem is well-defined. In particular, we would like conditions under which the posterior measure is equivalent (as a measure, see Section 6) to the prior measure.

We model the observation y as

$$y = \mathrm{e}^{-AT}u + \eta, \tag{3.20}$$

where $\eta \sim \mathcal{N}(0, \mathcal{C}_1)$ is independent of u. The observation operator $\mathcal{G} : \mathcal{H} \to \mathcal{H}$ is given by $\mathcal{G}(u) = \mathrm{e}^{-AT}u$ and, in fact, $\mathcal{G} : \mathcal{H} \to \mathcal{H}^\ell$ for any $\ell > 0$. If we assume that $\mathcal{C}_1 = \delta A^{-\gamma}$ for some $\gamma > d/2$ and $\delta > 0$, we then have that, almost surely, $\eta \in \mathcal{H}$ by Lemma 6.27.

Consider the Gaussian random variable $(u, y) \in \mathcal{H} \times \mathcal{H}$. We have

$$\mathbb{E}(u, y) := (\overline{u}, \overline{y}) = (m_0, \mathrm{e}^{-AT}m_0).$$

Straightforward calculation shows that

$$\begin{aligned}
\mathbb{E}(u - \overline{u}) \otimes (u - \overline{u}) &= \mathcal{C}_0, \\
\mathbb{E}(u - \overline{u}) \otimes (y - \overline{y}) &= \mathcal{C}_0 \, \mathrm{e}^{-AT}, \\
\mathbb{E}(y - \overline{y}) \otimes (y - \overline{y}) &= \mathrm{e}^{-AT}\mathcal{C}_0 \, \mathrm{e}^{-AT} + \mathcal{C}_1.
\end{aligned}$$

By Theorem 6.20 we find that the posterior measure μ^y for u given y is also

Gaussian, with mean

$$m = m_0 + \frac{\beta}{\delta}\,\mathrm{e}^{-AT}A^{\gamma-\alpha}\left(I + \frac{\beta}{\delta}\,\mathrm{e}^{-2AT}A^{\gamma-\alpha}\right)^{-1}(y - \mathrm{e}^{-AT}m_0) \tag{3.21}$$

and covariance operator

$$\mathcal{C} = \mathcal{C}_0\big(I + \mathrm{e}^{-2AT}\mathcal{C}_0\mathcal{C}_1^{-1}\big)^{-1} \tag{3.22}$$

$$= \beta A^{-\alpha}\left(I + \frac{\beta}{\delta}\,\mathrm{e}^{-2AT}A^{\gamma-\alpha}\right)^{-1}. \tag{3.23}$$

We now show that the posterior (Gaussian) measure on $\mathcal{H}$ is indeed equivalent to the prior. We will assume $\alpha > d/2$ since this ensures that samples from the prior are continuous functions, by Lemma 6.25.

Theorem 3.7. Consider an initial condition for the heat equation (3.19) with prior Gaussian measure $\mu_0 \sim \mathcal{N}(m_0, \beta A^{-\alpha})$, $m_0 \in \mathcal{H}^\alpha$, $\beta > 0$ and $\alpha > d/2$. If an observation is given in the form (3.20) then the posterior measure μ^y is also Gaussian, with mean and variance determined by (3.21) and (3.23). Furthermore, μ^y and the prior measure μ_0 are equivalent Gaussian measures.

Proof. Let $\{\phi_k, \lambda_k\}_{k\in\mathbb{K}}$, $\mathbb{K} = \mathbb{Z}^d\backslash\{0\}$, denote the eigenvalues of A and define $\kappa := \frac{\beta}{\delta}\sup_{k\in\mathbb{K}}\mathrm{e}^{-2\lambda_k T}\lambda_k^{\gamma-\alpha}$ which is finite since $T > 0$ and A generates an analytic semigroup. Furthermore, the operator

$$K = \left(I + \frac{\beta}{\delta}\,\mathrm{e}^{-2AT}A^{\gamma-\alpha}\right)^{-1}$$

is diagonalized in the same basis as A, and is a bounded and invertible linear operator with all eigenvalues lying in $[(1+\kappa)^{-1}, 1]$. Now, from (3.22), for any $h \in \mathcal{H}$,

$$\frac{1}{1+\kappa}\langle h, \mathcal{C}_0 h\rangle \leqslant \langle h, \mathcal{C}h\rangle = \langle h, \mathcal{C}_0 K h\rangle \leqslant \langle h, \mathcal{C}_0 h\rangle.$$

Thus, by Lemma 6.15, we deduce that condition (i) of Theorem 6.13 is satisfied, with $E = \mathcal{H}^\alpha = \mathrm{Im}(\mathcal{C}_0^{1/2})$.

From (3.21) we deduce that

$$m - m_0 = \frac{\beta}{\delta}\,\mathrm{e}^{-AT}A^{\gamma-\alpha}K(y - \mathrm{e}^{-AT}m_0).$$

Since A generates an analytic semigroup and since K is bounded, we deduce that $m - m_0 \in \mathcal{H}^r$ for any $r \in \mathbb{R}$. Hence condition (ii) of Theorem 6.13 is

satisfied. To check the remaining condition (iii), define

$$\begin{aligned}\mathcal{T} &= \mathcal{C}_0^{-1/2}\mathcal{C}\mathcal{C}_0^{-1/2} - I \\ &= -\frac{\beta}{\delta}\left(I + \frac{\beta}{\delta}\,\mathrm{e}^{-2AT}A^{\gamma-\alpha}\right)^{-1} A^{\gamma-\alpha}\mathrm{e}^{-2AT}.\end{aligned}$$

The operator T is clearly Hilbert–Schmidt because its eigenvalues μ_k satisfy

$$|\mu_k| \leqslant \frac{\beta}{\delta}\lambda_k^{\gamma-\alpha}\mathrm{e}^{-2\lambda_k T}$$

and hence decay exponentially fast. This establishes (iii) of Theorem 6.13 and the proof is complete. □

The preceding theorem uses the Gaussian structure of the posterior measure explicitly. To link the presentation to the other examples in this section, it is natural to ask whether a similar result can be obtained less directly.

We define $\Phi : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ by

$$\Phi(u; y) = \frac{1}{2}\|\mathrm{e}^{-AT}u\|_{\mathcal{C}_1}^2 - \langle \mathrm{e}^{-AT}u, y\rangle_{\mathcal{C}_1},$$

and use this function to derive Bayes' formula for the measure $\mu^y(\mathrm{d}u) = \mathbb{P}(\mathrm{d}u|y)$. We will show that $\mu^y(\mathrm{d}u)$ is absolutely continuous with respect to the prior $\mu_0(\mathrm{d}u)$ with density

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\bigl(-\Phi(u; y)\bigr). \tag{3.24}$$

Remark 3.8. It would be tempting to define a potential

$$\begin{aligned}\Psi(u; y) &= \frac{1}{2}\|y - \mathcal{G}(u)\|_{\mathcal{C}_1}^2 \\ &= \frac{1}{2}\|y - \mathrm{e}^{-AT}u\|_{\mathcal{C}_1}^2\end{aligned}$$

in analogy with the examples in the two previous sections: this Ψ is a least-squares functional measuring model/data mismatch. However, this quantity is almost surely infinite, with respect to the random variable y, since draws from a Gaussian measure in infinite dimensions do not lie in the corresponding Cameron–Martin space $\mathrm{Im}(\mathcal{C}_1^{1/2})$: see Lemma 6.10. This undesirable property of Ψ stems directly from the fact that the data y is a function rather than a finite-dimensional vector. To avoid the problem we work with $\Phi(u; y)$ which, informally, may be viewed as being given by the identity

$$\Phi(u; y) = \Psi(u; y) - \frac{1}{2}\|y\|_{\mathcal{C}_1}^2.$$

Thus we 'subtract off' the infinite part of Ψ. Since Bayes' formula in the form (3.24) only gives the density up to a y-dependent constant, we see intuitively

why this subtraction of a term involving y is reasonable. The issues outlined in this remark arise quite generally when the data y is infinite-dimensional and the observational noise η is Gaussian. ◇

The form of Φ arising in this problem, and the fact that the data is infinite-dimensional, precludes us from using Theorem 6.31 to establish that (3.24) is correct; however, the method of proof is very similar to that used to prove Theorem 6.31.

Before proving that (3.24) is correct, we state and prove some properties of the potential Φ.

Lemma 3.9. The function $\Phi : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ satisfies Assumptions 2.6 with $X = Y = \mathcal{H}$ and $L(r)$ linearly bounded.

Proof. We may write

$$\Phi(u;y) = \frac{1}{2}\big\|\mathcal{C}_1^{-1/2}\mathrm{e}^{-AT}u\big\|^2 - \big\langle \mathcal{C}_1^{-1/2}\mathrm{e}^{-\frac{1}{2}AT}u, \mathcal{C}_1^{-1/2}\mathrm{e}^{-\frac{1}{2}AT}y\big\rangle.$$

Since $\mathcal{C}_1^{-1} = \delta A^\gamma$ we deduce that $\mathcal{K}_\lambda := \mathcal{C}_1^{-1/2}\mathrm{e}^{-\lambda AT}$ is a compact operator on $\mathcal{H}$ for any $\lambda > 0$. By the Cauchy–Schwarz inequality we have, for any $a > 0$,

$$\Phi(u;y) \geqslant -\frac{a^2}{2}\big\|\mathcal{C}_1^{-1/2}\mathrm{e}^{-\frac{1}{2}AT}u\big\|^2 - \frac{1}{2a^2}\big\|\mathcal{C}_1^{-1/2}\mathrm{e}^{-\frac{1}{2}AT}y\big\|^2.$$

By the compactness of $\mathcal{K}_{1/2}$ and by choosing a arbitrarily small, we deduce that Assumption 2.6(i) holds. Assumption 2.6(ii) holds by a similar Cauchy–Schwarz argument. Since Φ is quadratic in u, and using the compactness of $\mathcal{K}_{1/2}$ and $\mathcal{K}_1$, we see that

$$\begin{aligned}
|\Phi(u_1;y) - \Phi(u_2;y)| &\leqslant C\big(\|\mathcal{K}_1u_1\| + \|\mathcal{K}_1u_2\| + \|\mathcal{K}_{1/2}y\|\big)\|\mathcal{K}_{1/2}(u_1-u_2)\| \\
&\leqslant C\big(\|u_1\| + \|u_2\| + \|y\|\big)\|\mathrm{e}^{-\frac{1}{4}AT}(u_1-u_2)\| \qquad (3.25)\\
&\leqslant C\big(\|u_1\| + \|u_2\| + \|y\|\big)\|u_1-u_2\|, \qquad (3.26)
\end{aligned}$$

and similarly

$$|\Phi(u;y_1) - \Phi(u;y_2)| \leqslant C\|u\|\|y_1-y_2\|,$$

so that Assumptions 2.6(iii) and (iv) hold. □

Theorem 3.10. Consider the inverse problem for the initial condition u in (3.19), subject to observation in the form (3.20) and with prior Gaussian measure $\mu_0 = \mathcal{N}(m_0, \beta A^{-\alpha})$. If $m_0 \in \mathcal{H}^\alpha, \beta > 0$ and $\alpha > d/2$, then the posterior measure $\mu^y(\mathrm{d}u) = \mathbb{P}(\mathrm{d}u|y)$ and the prior $\mu_0(\mathrm{d}u)$ are equivalent with Radon–Nikodym derivative given by (3.24).

Proof. Recall that $\mathcal{C}_1 = \delta A^{-\gamma}$ and that $\mathcal{C}_0 = \beta A^{-\alpha}$. Define the measures

$$\begin{aligned}\mathbb{Q}_0(\mathrm{d}y) &= \mathcal{N}(0, \mathcal{C}_1),\\ \mathbb{Q}(\mathrm{d}y|u) &= \mathcal{N}(\mathrm{e}^{-AT}u, \mathcal{C}_1),\\ \mu_0(\mathrm{d}u) &= \mathcal{N}(m_0, \mathcal{C}_0),\end{aligned}$$

and then define

$$\begin{aligned}\nu_0(\mathrm{d}y, \mathrm{d}u) &= \mathbb{Q}_0(\mathrm{d}y) \otimes \mu_0(\mathrm{d}u),\\ \nu(\mathrm{d}y, \mathrm{d}u) &= \mathbb{Q}(\mathrm{d}y|u)\mu_0(\mathrm{d}u).\end{aligned}$$

By Theorem 6.14 we deduce that

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{Q}_0}(y|u) = \exp\left(-\frac{1}{2}\|\mathrm{e}^{-AT}u\|_{\mathcal{C}_1}^2 + \langle \mathrm{e}^{-AT}u, y\rangle_{\mathcal{C}_1}\right).$$

The measure ν is well-defined because the function $\Phi(\cdot; y) : \mathcal{H} \to \mathbb{R}$ is continuous and hence μ_0-measurable if $\mu_0(\mathcal{H}) = 1$. This last fact follows from Lemma 6.27, which shows that draws from μ_0 are almost surely in $\mathcal{H}$. Hence

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu_0}(y, u) = \exp\left(-\frac{1}{2}\|\mathrm{e}^{-AT}u\|_{\mathcal{C}_1}^2 + \langle \mathrm{e}^{-AT}u, y\rangle_{\mathcal{C}_1}\right).$$

By applying Theorem 6.29, noting that under ν_0 the random variables y and u are independent with $u \sim \mu_0$, we deduce that

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2}\|\mathrm{e}^{-AT}u\|_{\mathcal{C}_1}^2 + \langle \mathrm{e}^{-AT}u, y\rangle_{\mathcal{C}_1}\right),$$

with constant of proportionality independent of u. □

3.6. Fluid mechanics

The preceding four subsections provide a range of examples where somewhat explicit calculations, using the solution of various forward linear PDE problems, establish that the associated inverse problems may be placed in the general framework that we outlined in Section 2.4 and will study further in Section 4. However, it is by no means necessary to have explicit solutions of the forward problem to use the framework developed in this article, and the examples of this subsection, and the two subsections which follow it, illustrate this.

Fluid mechanics provides an interesting range of applications where the technology of inverse problems is relevant. We outline examples of such problems and sketch their formulation as Bayesian inverse problems for functions. We also show that these problems may be formulated to satisfy Assumptions 2.7. Unlike the previous three sections, however, we do not provide full details; we refer to other works for these, in the bibliography subsection.

In *weather forecasting* a variety of instruments are used to measure the velocity of the air in the atmosphere. Examples include weather balloons, data from commercial and military aircraft, as well as special-purpose aircraft, and satellites. An important inverse problem is to determine the global velocity field, and possibly other fields, from the *Eulerian data* comprising the various noisy measurements described above.

As a concrete, and simplified, model of this situation we consider the linearized shallow water equations on a two-dimensional torus. The equations are a coupled pair of PDEs for the two-dimensional velocity field v and a scalar height field h, with the form

$$\frac{\partial v}{\partial t} = Sv - \nabla h, \quad (x,t) \in \mathbb{T}^2 \times [0,\infty), \tag{3.27a}$$

$$\frac{\partial h}{\partial t} = -\nabla \cdot v, \qquad (x,t) \in \mathbb{T}^2 \times [0,\infty). \tag{3.27b}$$

The two-dimensional unit torus $\mathbb{T}^2$ is shorthand for the unit square with periodic boundary conditions imposed. The skew matrix S is given by

$$S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

and the term involving it arises from the Coriolis effect.

The objective is to find the initial velocity and height fields $(v(0), h(0)) = (u,p) \in \mathcal{H}$, where

$$\mathcal{H} := \left\{ u \in L^2(\mathbb{T}^2; \mathbb{R}^3) \Big| \int_{\mathbb{T}^2} u \, \mathrm{d}x \right\}.$$

We assume that we are given noisy observations of the velocity field at positions $\{x_j\}_{j=1}^{\mathrm{J}}$ and times $\{t_k\}_{k=1}^{\mathrm{K}}$, all positive. Concatenating data, we write

$$y = \mathcal{G}(u,p) + \eta. \tag{3.28}$$

Here $\mathcal{G}$ maps a dense subset of $\mathcal{H}$ into $\mathbb{R}^{2\mathrm{JK}}$ and is the *observation operator*. Because the PDE (3.27) is linear, so too is $\mathcal{G}$. We assume that $\eta \sim \mathcal{N}(0,\Gamma)$ is independent of u and we consider the Bayesian inverse problem of finding the posterior measure $\mu^y(\mathrm{d}u) = \mathbb{P}(\mathrm{d}u|y)$ from the prior μ_0. We let $A = -\triangle$ on $\mathbb{T}^2$ with domain

$$D(A) = \left\{ H^2(\mathbb{T}^2) \Big| \int_{\mathbb{T}^2} u \, \mathrm{d}x = 0 \right\}$$

and define the prior through fractional powers of A.

Theorem 3.11. Consider an initial condition for the shallow water equations (3.27) with prior Gaussian measure $\mu_0 = \mathcal{N}(m_0, \beta A^{-\alpha})$ with $m_0 \in \mathcal{H}^\alpha$, $\beta > 0$ and $\alpha > 2$. If a noisy observation is made in the form (3.28), then the

posterior measure μ^y is also Gaussian, and is absolutely continuous with respect to the prior measure μ_0, with Radon–Nikodym derivative

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u,p) \propto \exp\left(-\frac{1}{2}|y-\mathcal{G}(u,p)|_\Gamma^2\right), \tag{3.29}$$

where $\mathcal{G}$ is given by (3.28). Furthermore, the observation operator $\mathcal{G}$ satisfies Assumptions 2.7 with $X = \mathcal{H}^s$ and K globally bounded, for any $s > 1$. $\diamond$

In *oceanography* a commonly used method of gathering data about ocean currents, temperature, salinity and so forth is through the use of Lagrangian instruments which are transported by the fluid velocity field and transmit positional information using GPS. An important inverse problem is to determine the velocity field in the ocean from the Lagrangian data comprising the GPS information about the position of the instruments. As an idealized model consider the incompressible Stokes ($\iota = 0$) or Navier–Stokes ($\iota = 1$) equations written in the form:

$$\frac{\partial v}{\partial t} + \iota v.\nabla v = \nu\Delta v - \nabla p + f, \quad (x,t) \in \mathbb{T}^2 \times [0,\infty), \tag{3.30a}$$

$$\nabla \cdot v = 0, \qquad (x,t) \in \mathbb{T}^2 \times [0,\infty), \tag{3.30b}$$

$$v = u, \qquad (x,t) \in \mathbb{T}^2 \times \{0\}. \tag{3.30c}$$

As in the preceding example we impose periodic boundary conditions, here on the velocity field v and the pressure p. We assume that f has zero average over D, noting that this implies the same for $v(x,t)$, provided that $u(x) = v(x,0)$ has zero initial average. We define the Stokes operator A and Leray projector P in the standard fashion, together with the Sobolev spaces $\mathcal{H}^s = D(A^{s/2})$ as in (2.29). The equations (3.30) can be written as an ODE in the Hilbert space $\mathcal{H}$:

$$\frac{\mathrm{d}v}{\mathrm{d}t} + \iota B(v,v) + \nu Au = \psi, \tag{3.31}$$

where $\psi = Pf$ and $B(v,v)$ represents the projection, under P, of the nonlinear convective term.

We assume that we are given noisy observations of Lagrangian tracers with position z solving the integral equation

$$z_j(t) = z_{j,0} + \int_0^t v(z_j(s),s)\,\mathrm{d}s. \tag{3.32}$$

These equations have a unique solution if $u \in \mathcal{H}$ and $\psi \in L^2((0,T);\mathcal{H})$.

For simplicity assume that we observe all the tracers z at the same set of times $\{t_k\}_{k=1}^{\mathrm{K}}$, and that the initial particle tracer positions $z_{j,0}$ are known to us:

$$y_{j,k} = z_j(t_k) + \eta_{j,k}, \quad j = 1,\ldots,\mathrm{J} \ \text{ and } \ k = 1,\ldots,\mathrm{K}, \tag{3.33}$$

where the $\eta_{j,k}$ are zero mean Gaussian random variables. The times $\{t_k\}$ are assumed to be positive. Concatenating data, we may write

$$y = \mathcal{G}(u) + \eta, \tag{3.34}$$

with $y = (y^*_{1,1}, \ldots, y^*_{\mathrm{J,K}})^*$ and $\eta \sim \mathcal{N}(0, \Gamma)$ for some covariance matrix Γ capturing the correlations present in the noise. The function $\mathcal{G}$ maps a dense subspace of $\mathcal{H}$ into $\mathbb{R}^{2\mathrm{JK}}$. The objective is to find the initial velocity field u, given y. We start by stating a result concerning the observation operator.

Lemma 3.12. Assume that $\psi \in C([0,T]; \mathcal{H}^\gamma)$ for some $\gamma > 0$. Then $\mathcal{G}$ given by (3.34) satisfies Assumptions 2.7 with $X = \mathcal{H}^\ell$ for any $\ell > 0$. ◇

These properties of the observation operator $\mathcal{G}$ lead to the following result.

Theorem 3.13. Let $\mu_0 = \mathcal{N}(m_0, \beta A^{-\alpha})$ denote a prior Gaussian measure on μ_0. If $m_0 \in \mathcal{H}^\alpha$, $\beta > 0$ and $\alpha > 1$ then the measure $\mu^y(\mathrm{d}u) = \mathbb{P}(\mathrm{d}u|y)$ is absolutely continuous with respect to μ_0, with Radon–Nikodym derivative given by

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2}|y - \mathcal{G}(u)|^2_\Gamma\right), \tag{3.35}$$

with $\mathcal{G}$ defined by (3.34). ◇

Notice that the required lower bound on the exponent α in the preceding theorem is lower than that appearing in Theorem 3.11. This is because the (Navier–)Stokes equation is smoothing, and hence less regularity is required on the initial condition in order to define the observation operator $\mathcal{G}$ than for the linearized shallow water equations.

3.7. Subsurface geophysics

Determining the permeability of subsurface rock is enormously important in a range of different applications. Among these applications are the prediction of transport of radioactive waste from underground waste repositories, and the optimization of oil recovery from underground fields. We give an overview of some inverse problems arising in this area. As in the previous subsection we do not give full details, leaving these to the cited literature in the bibliography subsection.

The permeability tensor K is a central component of *Darcy's law*, which relates the velocity field v to the gradient of the pressure p in porous media flow:

$$v = -K\nabla p. \tag{3.36}$$

In general K is a tensor field. However, the problem is often simplified by

assuming that $K = kI$, where k is a scalar field and I the identity tensor; we make this simplification.

In many subsurface flow problems it is reasonable to model the velocity field as incompressible. Combining this constraint with Darcy's law (3.36) shows that the pressure p is governed by the PDE

$$\nabla \cdot (-k\nabla p) = 0, \quad x \in D, \tag{3.37a}$$

$$p = h, \quad x \in \partial D. \tag{3.37b}$$

This model is a widely used simplified model in groundwater flow modelling. The inverse problem is to find the permeability k from observations of the pressure at points in the interior of D; this information can be found by measuring the height of the water table. For simplicity we work in two or three dimensions d and assume that $D \subset \mathbb{R}^d$ is bounded and open. As in Section 3.3 it is physically and mathematically important that k be positive, in order that the elliptic equation for the pressure is well-posed. Hence we write $k = \exp(u)$ and consider the problem of determining u.

We assume that we observe

$$y_j = p(x_j) + \eta_j, \quad j = 1, \ldots, \mathrm{J}, \tag{3.38}$$

and note that this may be written as

$$y = \mathcal{G}(u) + \eta \tag{3.39}$$

for some implicitly defined function $\mathcal{G}$. We assume that $\eta \sim \mathcal{N}(0, \Gamma)$ is independent of u. Before formulating the Bayesian inverse problem, we state the following result concerning the forward problem.

Lemma 3.14. Assume that the boundary of D, ∂D, is C^1-regular and that the boundary data h may be extended to a function $h \in W^{1,2r}(D)$ with $r > d/2$. The function $\mathcal{G}$ satisfies Assumptions 2.7 with $X = C(\overline{D})$. ◇

We define the prior Gaussian measure through fractional powers of the Laplacian $\mathcal{A} = -\triangle$ with

$$D(\mathcal{A}) = \left\{ u \in H^2(D) \middle| \nabla u \cdot n = 0, \int_D u(x)\,\mathrm{d}x = 0 \right\}.$$

Here n denotes the unit outward normal on the boundary of D.

Theorem 3.15. Let the assumptions of Lemma 3.14 hold and let $\mu_0 = \mathcal{N}(0, \beta A^{-\alpha})$ denote a prior Gaussian measure on μ_0. If $\beta > 0$ and $\alpha > d - 1/2$, then the measure $\mu^y(\mathrm{d}u) = \mathbb{P}(\mathrm{d}u|y)$ is absolutely continuous with respect to μ_0, with Radon–Nikodym derivative

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(x) \propto \exp\left(-\frac{1}{2}|y - \mathcal{G}(u)|_\Gamma^2\right) \tag{3.40}$$

and $\mathcal{G}$ given by (3.39). ◇

Once the posterior measure on u is known it can be used to quantify uncertainty in predictions made concerning the Lagrangian transport of radioactive particles under the velocity field v given by (3.36). In particular, the push forward of the measure μ^y onto v, and hence onto particle trajectories z obeying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = v(z),$$

will define a measure on the possible spread of radioactive contaminants, enabling risk assessment to be undertaken.

The oil industry routinely confronts an inverse problem similar to but more complex than that arising in the nuclear waste industry. Again, uncertainty quantification is important as it enables more effective decision making concerned with the substantial investment of resources required to extract oil from increasingly complex environments. The primary difference between the simple model we have described for nuclear waste management and that which we are about to describe for oil extraction arises because the subsurface fluid flow for oil extraction is multiphase (gas, water, oil) and significant on much shorter time scales than in the nuclear waste management scenario. We study a simplified two-phase problem, for oil and water alone. The physical model contains two unknown scalar fields, the water saturation S (volume fraction of water in an oil–water mixture) and pressure p, and is posed in a bounded open set $D \subset \mathbb{R}^d$. Darcy's law now takes the form

$$v = -\lambda(S)k\nabla p. \tag{3.41}$$

Mass conservation and transport, respectively, give the equations

$$\begin{aligned} -\nabla \cdot (\lambda(S)k\nabla p) &= h_1, \qquad (x,t) \in D \times [0,\infty), \\ \frac{\partial S}{\partial t} + v \cdot \nabla f(S) &= \eta \triangle S, \quad (x,t) \in D \times [0,\infty), \qquad (3.42) \\ p &= h_2, \qquad (x,t) \in \partial D \times [0,\infty). \qquad (3.43) \end{aligned}$$

The flux function f is known (typically the Buckley–Leverett form is used) and the source/boundary terms h_1, h_2 are also both assumed known. The scalar η is the (also assumed known) diffusivity of the multiphase flow, typically very small. Initial conditions for S are specified on D at time $t = 0$. There are additional boundary conditions on S which we now describe. We partition $\partial D = \partial D^{\mathrm{out}} \bigcup \partial D^{\mathrm{in}}$. We think of pumping water in on the boundary ∂D^{in}, so that $S = 1$ there, and specify a Robin boundary condition on $\partial D^{\mathrm{out}}$, determining the flux of fluid in terms of S the water saturation.

We assume that we have access to noisy measurements of the *fractional flow* $F(t)$, which quantifies the fraction of oil produced on a subset $\partial D^{\mathrm{meas}}$

of the outflow boundary ∂D^{out}. This measurement is via the function

$$F(t) = 1 - \frac{\int_{\partial D^{\text{meas}}} f(S) v_n \, \mathrm{d}l}{\int_{\partial D^{\text{meas}}} v_n \, \mathrm{d}l},$$

where v_n is the component of the velocity v which is normal to the boundary and $\mathrm{d}l$ denotes integration along the boundary. Assume that we make measurements of F at times $\{t_k\}_{k=1}^{\mathrm{K}}$, polluted by Gaussian noise. Then the data are as follows:

$$y_k = F(t_k) + \eta_k, \quad k = 1, \ldots, \mathrm{K},$$

where the η_k are zero mean Gaussian random variables. Concatenating data, we may write

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

where, as before, $k(x) = \exp(u(x))$. We assume that $\eta \sim \mathcal{N}(0, \Gamma)$ for some covariance matrix Γ encapsulating measurement errors. The prior μ_0 is a Gaussian measure on u, specified as in the previous section. We once again anticipate that

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(x) \propto \exp\left(-\frac{1}{2}|y - \mathcal{G}(u)|_B^2\right). \tag{3.44}$$

This is similar to the nuclear waste problem, but the observation operator $\mathcal{G}$ is now more complicated. However, similar analyses of the properties of the forward problem, and the resulting Bayesian inverse problem, can be undertaken.

3.8. Molecular dynamics

Consider a molecule described by the positions x of N atoms moving in $\mathbb{R}^d$, with $d = 1, 2$ or 3. If we assume that the particles interact according to a potential $V : \mathbb{R}^d \to \mathbb{R}$ and are subject to thermal activation, then, in the *over-damped* limit where the inertial relaxation time is fast, we obtain the *Brownian dynamics* model for the position of x:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\nabla V(x) + \sqrt{\frac{2}{\beta}} \frac{\mathrm{d}W}{\mathrm{d}t}. \tag{3.45}$$

Here W is a standard $\mathbb{R}^{Nd}$-valued Brownian motion and β the inverse temperature. One of the key challenges in molecular dynamics is to understand how molecules rearrange themselves to change from one configuration to another: in some applications this may represent a chemical reaction, and in others a conformational change such as seen in biomolecules. When the temperature is small ($\beta \gg 1$), the solutions of (3.45) spend most of their time near the minima of the potential V. Transitions between different minima of the potential are rare events. Simply solving the SDE starting from one of the minima will be a computationally infeasible way of

generating sample paths which jump between minima, since the time to make a transition is exponentially small in β. Instead we may condition on this rare event occurring. This may be viewed as an inverse problem to determine the control W which drives the system from one configuration to another. However, we will work directly with the functions x which result from this control, as these constitute the more physically interesting quantity. Because the Brownian motion W is a random function, this leads naturally to the question of determining the probability measure on functions x undergoing the desired transition between configurations. The desired transition can be defined by conditioning the dynamics given by (3.45) to satisfy the boundary conditions

$$x(0) = x^-, \quad x(T) = x^+. \tag{3.46}$$

We view x as an element of $L^2((0,T);\mathbb{R}^{Nd})$ and denote the Nd-dimensional *Brownian bridge measure* arising from (3.45) and (3.46) in the case $V \equiv 0$ by μ_0. We also define μ to be the desired bridge diffusion measure arising from the given V. We may view both μ_0 and μ as measures on $L^2((0,T);\mathbb{R}^{Nd})$; the measure μ_0 is Gaussian but, unless V is quadratic, the measure μ is not. We now proceed to determine the Radon–Nikodym derivative of μ with respect to the Gaussian bridge diffusion μ_0.

Theorem 3.16. Assume $V \in C^2(\mathbb{R}^{Nd};\mathbb{R})$ and that the stochastic initial value problem, found from (3.45) and (3.46) without the condition $x(T) = x^+$, has solutions which do not explode almost surely on $t \in [0,T]$. Then the measure μ defined by the bridge diffusion problem (3.45) and (3.46) is absolutely continuous with respect to the Brownian bridge measure μ_0 found from (3.45) and (3.46) in the case $V \equiv 0$. Furthermore, the Radon–Nikodym derivative is given by

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(x) \propto \exp(-\Phi(x)), \tag{3.47}$$

where the potential Φ is defined by

$$\Phi(x) = \frac{\beta}{2}\int_0^T G(x(t))\,\mathrm{d}t, \tag{3.48a}$$

$$G(x) = \frac{1}{2}\|\nabla V(x)\|^2 - \frac{1}{\beta}\Delta V(x). \tag{3.48b}$$

◇

In addition, we find that a large class of problems leads to the common structure of Section 2.4. There is no explicit data $y \in Y$ in this problem, but we can let $y \in \mathbb{R}^p$ denote the parameters appearing in the potential V, and hence in G. (Note that β is not such a parameter, as it appears in G but not in V; more fundamentally it appears in μ_0 and so is not simply a parameter in the potential Φ). We thus write $V(x;y)$ and $G(x;y)$.

Lemma 3.17. Consider the function Φ defined by (3.48a) and (3.48b) with $V : \mathbb{R}^{Nd} \times \mathbb{R}^p \to \mathbb{R}$. Assume that for any $\varepsilon, r > 0$ there exists $M = M(\varepsilon, r) \in \mathbb{R}$ such that, for all $\|y\| < r$,

$$G(x; y) \geqslant -\varepsilon^2 |x|^2 + M;$$

assume also that $G \in C^1(\mathbb{R}^{Nd} \times \mathbb{R}^p, \mathbb{R})$ with derivative $D_y G(x; y)$ which is polynomially bounded in x. Then Φ satisfies Assumptions 2.6 with $X = H^1((0, T))$. ◇

3.9. Discussion and bibliography

We started this section by studying the problem of determining a field from observation. This is intimately related to the study of interpolation of data by splines, a subject comprehensively developed and reviewed in Wahba (1990). The link between spline interpolation and inverse problems using Gaussian fields is surveyed in Gu (2002).

The inverse problem for the diffusion coefficient in Section 3.3 is a one-dimensional analogue of the inverse problems arising in the geophysics community, which we outline in Section 3.7; these problems, which arise in the study of groundwater flow and are hence of interest to the burial of (radioactive nuclear and other) waste, are discussed in Zimmerman *et al.* (1998). A related inverse problem for the diffusion coefficient of an elliptic PDE is that arising in electrical impedence tomography; this widely studied inverse problem requires recovery of the diffusion coefficient from measurements of the boundary flux. It is of central importance in the medical sciences, and also has a rich mathematical structure; see Borcea (2002) and Uhlmann (2009) for reviews.

Inverse problems for the heat equation, the subject of Section 3.5, are widely studied. See, for example, the cited literature in Beck, Blackwell and Clair (2005) and Engl *et al.* (1996). An early formulation of this problem in a Bayesian framework appears in Franklin (1970).

We study applications to fluid dynamics in Section 3.6: the subject known as data assimilation. Kalnay (2003) and Bennett (2002) survey inverse problems in fluid mechanics from the perspective of weather prediction and oceanography respectively; see also Apte, Jones, Stuart and Voss (2008*b*), Lorenc (1986), Ide, Kuznetsov and Jones (2002), Kuznetsov, Ide and Jones (2003), Nichols (2003*a*) and Nodet (2006) for representative examples, some closely related to the specific model problems that we study in this article. Theorem 3.11, arising in our study of Eulerian observations and integration into a wave equation model, is proved in Dashti, Pillai and Stuart (2010*b*). Lemma 3.12 and Theorem 3.13, arising in the study of Lagrangian observations, are proved in Cotter *et al.* (2009) (Navier–Stokes case) and Cotter *et al.* (2010*a*) (Stokes case). A major question facing the research community in data assimilation for fluid mechanics applications is to determine

whether future increase in available computer resources is used to increase resolution of the computational models, or to improve estimates of uncertainty. (The question is discussed, in the context of climate modelling, in Palmer *et al.* (2009).) The framework developed in Section 3.6 allows for a systematic treatment of uncertainty, as quantified by the variability in the posterior measure; furthermore, the framework may be extended to make inference not only about the initial condition but also about forcing to the model, thereby enabling model error to be uncovered in a systematic fashion. In this context we define model error to be an error term in the dynamical model equations, as in Hagelberg, Bennett and Jones (1996). Note, however, that in practical data assimilation, model errors are sometimes combined with the observation errors (Cohn 1997). Further discussion of model error for problems arising in the atmospheric sciences may be found in the papers of Nichols (2003*b*) and Fang *et al.* (2009*b*). In Cotter *et al.* (2009) we discuss both Eulerian and Lagrangian data assimilation with and without model error, with fluid flow model given by the Navier–Stokes equations (3.30) with $\iota = 1$.

The subject of minimal regularity required to define Lagrangian trajectories (3.32) in a Navier–Stokes velocity field is covered in Chemin and Lerner (1995) and Dashti and Robinson (2009). This theory is easily extended to cover the case of the Stokes equations.

The systematic treatment of Lagrangian data assimilation is developed in the sequence of papers by Ide *et al.* (2002), Kuznetsov *et al.* (2003), Salman, Kuznetsov, Jones and Ide (2006) and Salman, Ide and Jones (2008) with recent application in Vernieres, Ide and Jones (2010). Although the subject had been treated in an applied context, these were the first papers to develop a clear dynamical systems framework in which the coupled (skew-product) dynamical system for the fluid and the Lagrangian particles was introduced as the fundamental object of study.

The papers by Pimentel, Haines and Nichols (2008*a*, 2008*b*), Bell, Martin and Nichols (2004), Huddleston, Bell, Martin and Nichols (2004) and Martin, Bell and Nichols (2002) describe a variety of applications of ideas from data assimilation to problems in oceanography. The paper by Wlasak, Nichols and Roulstone (2006) discusses data assimilation in the atmospheric sciences, using a potential vorticity formulation. In Bannister, Katz, Cullen, Lawless and Nichols (2008), forecast errors are studied for data assimilation problems in fluid flow. The paper by Alekseev and Navon (2001) uses a wavelet-based approach to study the inverse problem of determining inflow fluid properties from outflow measurements.

Some of the earliest work concerning the statistical formulation of inverse problems was motivated by geophysical applications (Backus 1970*a*, 1970*b*, 1970*c*), such as those introduced in Section 3.7. The interpolation of a random field, observed at a finite set of points, is outlined in Gu (2008) and

is often referred to as 'kriging' (Cressie 1993). Overviews of issues arising in oil reservoir simulation may be found in Farmer (2005, 2007). The mathematical statement of the oil reservoir simulation problem as outlined here is formulated in Ma, Al-Harbi, Datta-Gupta and Efendiev (2008) and further discussion of numerical methods is undertaken in Dostert, Efendiev, Hou and Luo (2006). Lemma 3.14 and Theorem 3.15, concerning the elliptic inverse problem for subsurface flow, are proved in Dashti, Harris and Stuart (2010*a*).

The formulation of problems from molecular dynamics in terms of probability measures on time-dependent functions has a long history. On the mathematical side this is intimately related to the theory of rare events (Freidlin and Wentzell 1984) and an overview of some of the sampling techniques used for this problem may be found in Bolhuis, Chandler, Dellago and Geissler (2002). The particular formulation of the problem that we undertake here, in which the length of the transition T is specified *a priori*, can be found in Dashti *et al.* (2010*b*); see also Reznikoff and Vanden Eijnden (2005), Hairer, Stuart, Voss and Wiberg (2005) and Hairer, Stuart and Voss (2007). A generalization to second-order Newtonian dynamics models, in place of the over-damped Brownian dynamics model (3.45) may be found in Hairer, Stuart and Voss (2010*a*).

4. Common structure

4.1. Overview

It is natural to view the posterior measure μ^y given by (2.24) as *the ideal solution* to the problem of combining a mathematical model with data y. However, obtaining a formula such as this is only the beginning: we are confronted with the formidable task of extracting information from this formula. At a high level this entire section is devoted to the question of the *stability* of measures μ^y to perturbations of various kinds, under Assumptions 2.6 or 2.7. These stability results help to create firm foundations for the *algorithms* designed to obtain information from the measure μ^y; these algorithms are summarized in the next section.

In this section, then, we study the well-posedness of problems with respect to parameters, or data, entering the definition of the measure: we show Lipschitz properties of the posterior measure with respect to changes in the data. We also study the related issue of approximating the measure, in particular the approximation by measures defined over a finite-dimensional space. Section 4.2 concerns well-posedness in the setting where the data is in the form of a function: it is infinite-dimensional. In practical applications the data will always be finite, but when the data is very dense it is a useful abstraction to consider the data as being a function, and so this situation is conceptually important. However, when the data is sparse it is best

viewed as finite, as a number of mathematical simplifications follow from this. The well-posedness of the posterior measure in this finite data situation is studied in Section 4.3. In Section 4.4 we study the effect of approximating the potential Φ and the effect of this approximation on the measure μ^y given by (2.24).

A key idea throughout this section is the use of metrics to study distances between probability measures. This topic is discussed in Section 6.7 and, in particular, the Hellinger metric which we use throughout this section is introduced. The primary message concerning the Hellinger metric is this: consider two measures which are absolutely continuous with respect to a common Gaussian reference measure and which are distance ε apart in the Hellinger metric. Then the expectations of polynomially bounded functions under these two measures are also $\mathcal{O}(\varepsilon)$ apart. In particular, the mean and covariance operator are $\mathcal{O}(\varepsilon)$ apart.

4.2. Well-posedness

The probability measure of interest is typically defined through a density with respect to a Gaussian reference measure $\mu_0 = \mathcal{N}(0, \mathcal{C})$ on a Hilbert space $\mathcal{H}$ which, by shift of origin, we have taken to have mean zero. We assume that, for some separable Banach space X, we have $\mu_0(X) = 1$. We let $\{\phi_k, \gamma_k\}_{k=1}^{\infty}$ denote the eigenfunctions and eigenvalues of $\mathcal{C}$.

As in our previous developments, μ^y denotes the measure of interest, with y denoting parameters, or data, entering its definition. As in (2.24) we assume that

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{Z(y)} \exp\bigl(-\Phi(u;y)\bigr). \tag{4.1}$$

Recall that $\Phi(u;y)$ is the *potential* and that the *normalization constant* $Z(y)$ is chosen so that μ^y is a probability measure:

$$Z(y) = \int_H \exp\bigl(-\Phi(u;y)\bigr)\, \mathrm{d}\mu_0(u). \tag{4.2}$$

Both for this integral, and for others below, we observe that if $\mu_0(X) = 1$ we may write

$$Z(y) = \int_X \exp\bigl(-\Phi(u;y)\bigr)\, \mathrm{d}\mu_0(u),$$

and hence use properties of $\Phi(\cdot;y)$ which hold on X.

In the preceding section we showed that a number of inverse problems give rise to a probability measure μ^y of the form (4.1), where $\Phi : X \times Y \to \mathbb{R}$ satisfies Assumptions 2.6. The data (or parameters) y is (are) assumed to lie in a Banach space $(Y, \|\cdot\|_Y)$. We allow for the case where Y is infinite-dimensional and the data is in the form of a function. The four Assumptions 2.6(i)–(iv) play different roles, indicated by the following two

theorems. The third assumption is important for showing that the posterior probability measure is well-defined, whilst the fourth is important for showing continuity with respect to data. The first and second assumptions lead to bounds on the normalization constant Z from above and below, respectively.

Theorem 4.1. Let Φ satisfy Assumptions 2.6(i), (ii) and (iii) and assume that μ_0 is a Gaussian measure satisfying $\mu_0(X) = 1$. Then μ^y given by (4.1) is a well-defined probability measure on H.

Proof. Assumption 2.6(ii) may be used to show that Z is bounded below, as shown in the proof of Theorem 4.2 below. Under Assumption 2.6(iii) it follows that Φ is μ_0-measurable, and hence the measure μ^y is well-defined by (4.1). By Assumption 2.6(i) we have that, for $\|y\|_Y < r$ and all ε sufficiently small,

$$\begin{aligned} Z(y) &= \int_X \exp\bigl(-\Phi(u;y)\bigr)\,\mathrm{d}\mu_0(u) \\ &\leqslant \int_X \exp\bigl(\varepsilon\|u\|_X^2 - M(\varepsilon,r)\bigr)\,\mathrm{d}\mu_0(u) \\ &\leqslant C\exp\bigl(-M(\varepsilon,r)\bigr) < \infty, \end{aligned}$$

since μ_0 is a Gaussian probability measure and we may choose ε sufficiently small so that the Fernique Theorem (Theorem 6.9) applies. Thus the measure is normalizable and the proof is complete. □

This proof directly shows that the posterior measure is a well-defined probability measure, without recourse to a conditioning argument. The conditioning argument used in Theorem 6.31 provides the additional fact that $\mu^y(\mathrm{d}u) = \mathbb{P}(\mathrm{d}u|y)$.

Now we study continuity properties of the measure μ^y with respect to $y \in Y$, under Assumptions 2.6(i), (ii) and (iv). This establishes the robustness of many of the problems introduced in the preceding section to changes in data.

Theorem 4.2. Let Φ satisfy Assumptions 2.6(i), (ii) and (iv). Assume also that μ_0 is a Gaussian measure satisfying $\mu_0(X) = 1$ and that the measure $\mu^y \ll \mu_0$ with Radon–Nikodym derivative given by (4.1), for each $y \in Y$. Then μ^y is Lipschitz in the data y, with respect to the Hellinger distance: if μ^y and $\mu^{y'}$ are two measures corresponding to data y and y' then there exists $C = C(r) > 0$ such that, for all y, y' with $\max\{\|y\|_Y, \|y'\|_Y\} < r$,

$$d_{\mathrm{Hell}}(\mu^y, \mu^{y'}) \leqslant C\|y - y'\|_Y.$$

Consequently the expectation of any polynomially bounded function $f : X \to E$ is continuous in y. In particular the mean and, in the case where X is a Hilbert space, the covariance operator, are continuous in y.

Proof. Throughout the proof, all integrals are over X, unless specified otherwise. The constant C may depend on ε and r and changes from occurrence to occurrence. Let $Z = Z(y)$ and $Z' = Z(y')$ denote the normalization constants for μ^y and $\mu^{y'}$ so that

$$Z = \int \exp\bigl(-\Phi(u;y)\bigr)\,\mathrm{d}\mu_0(u),$$
$$Z' = \int \exp\bigl(-\Phi(u;y')\bigr)\,\mathrm{d}\mu_0(u).$$

Using Assumption 2.6(ii) gives, for any $r > 0$,

$$Z \geqslant \int_{\{\|u\|_X \leqslant r\}} \exp(-K(r))\,\mathrm{d}\mu_0(u) = \exp(-K(r))\mu_0\{\|u\|_X \leqslant r\}.$$

This lower bound is positive because μ_0 has full measure on X and is Gaussian, so that all balls in X have positive probability. We have an analogous lower bound for $|Z'|$.

Using Assumptions 2.6(i) and (iv) and using the Fernique Theorem, for μ_0,

$$\begin{aligned}|Z - Z'| &\leqslant \left(\int \exp\bigl(\varepsilon\|u\|_X^2 - M\bigr)\exp\bigl(\varepsilon\|u\|_X^2 + C\bigr)\,\mathrm{d}\mu_0(u)\right)\|y - y'\|_Y \\ &\leqslant C\|y - y'\|_Y.\end{aligned}$$

From the definition of Hellinger distance, we have

$$\begin{aligned}2d_{\mathrm{Hell}}(\mu^y,\mu^{y'})^2 &= \int\left(Z^{-1/2}\exp\left(-\frac{1}{2}\Phi(u;y)\right)\right. \\ &\qquad \left. - (Z')^{-1/2}\exp\left(-\frac{1}{2}\Phi(u;y')\right)\right)^2 \mathrm{d}\mu_0(u) \\ &\leqslant I_1 + I_2\end{aligned}$$

where

$$I_1 = \frac{2}{Z}\int\left(\exp\left(-\frac{1}{2}\Phi(u;y)\right) - \exp\left(-\frac{1}{2}\Phi(u;y')\right)\right)^2 \mathrm{d}\mu_0(u),$$
$$I_2 = 2\bigl|Z^{-1/2} - (Z')^{-1/2}\bigr|^2 \int \exp\bigl(-\Phi(u;y')\bigr)\,\mathrm{d}\mu_0(u).$$

Now, again using Assumptions 2.6(i) and (iv) and the Fernique Theorem,

$$\begin{aligned}\frac{Z}{2}I_1 &\leqslant \int \frac{1}{4}\exp\bigl(\varepsilon\|u\|_X^2 - M\bigr)\exp\bigl(2\varepsilon\|u\|_X^2 + 2C\bigr)\|y - y'\|_Y^2\,\mathrm{d}\mu_0(u) \\ &\leqslant C\|y - y'\|_Y^2.\end{aligned}$$

A similar use of the Fernique Theorem and Assumption 2.6(i) shows that

the integral in I_2 is finite. Also, using the bounds on Z, Z' from below,

$$\begin{aligned}\left|Z^{-1/2} - (Z')^{-1/2}\right|^2 &\leqslant C\big(Z^{-3} \vee (Z')^{-3}\big)|Z - Z'|^2 \\ &\leqslant C\|y - y'\|_Y^2.\end{aligned}$$

Combining gives the desired continuity result in the Hellinger metric.

Finally all moments of u in X are finite under μ^y and $\mu^{y'}$ because the change of measure from Gaussian μ_0 involves a term which may be bounded by use of Assumption 2.6(i). The Fernique Theorem may then be applied. The desired result concerning the continuity of moments follows from Lemma 6.37. □

Example 4.3. An example in which the data is a function is given in Section 3.5, where we study the inverse problem of determining the initial condition for the heat equation, given noisy observation of the solution at a positive time; in Lemma 3.9 we establish that Assumptions 2.6 hold in this case. ◇

4.3. Well-posedness: finite data

For Bayesian inverse problems in which a finite number of observations are made, the potential Φ has the form

$$\Phi(u; y) = \frac{1}{2}|y - \mathcal{G}(u)|_\Gamma^2, \tag{4.3}$$

where $y \in \mathbb{R}^q$ is the data, $\mathcal{G} : X \to \mathbb{R}^q$ is the observation operator and $|\cdot|_\Gamma$ is a covariance weighted norm on $\mathbb{R}^q$. In this case it is natural to express conditions on the potential Φ in terms of $\mathcal{G}$. Recall that this is undertaken in Assumptions 2.7. By Lemma 2.8 we know that Assumptions 2.7 imply Assumptions 2.6 for Φ given by (4.3). The following corollary of Theorem 4.2 is hence automatic.

Corollary 4.4. Assume that $\Phi : X \times \mathbb{R}^q \to \mathbb{R}$ is given by (4.3) and let $\mathcal{G}$ satisfy Assumptions 2.7. Assume also that μ_0 is a Gaussian measure satisfying $\mu_0(X) = 1$. Then the measure μ^y given by (4.1) is a well-defined probability measure and is Lipschitz in the data y, with respect to the Hellinger distance: if μ^y and $\mu^{y'}$ are two measures corresponding to data y and y', then there is $C = C(r) > 0$ such that, for all y, y' with $\max\{|y|_\Gamma, |y'|_\Gamma\} < r$,

$$d_{\text{Hell}}(\mu^y, \mu^{y'}) \leqslant C|y - y'|_\Gamma.$$

Consequently the expectation of any polynomially bounded function $f : X \to E$ is continuous in y. In particular the mean and, in the case where X is a Hilbert space, the covariance operator are continuous in y. ◇

Example 4.5. The first example of a problem with the structure of Assumptions 2.7 may be found in the discussion of finite-dimensional inverse

problems in Section 2.2, and formula (2.8) in the case where ρ is a Gaussian density; if, for example, $\mathcal{G}$ is differentiable and polynomially bounded, then Assumptions 2.7 hold: see Example 2.2 for an explicit illustration. All the examples in Section 3, with the exception of the heat equation example, for which the data is infinite, and the oil reservoir problem, for which the appropriate analysis and choice of X has not yet been carried out, fit the framework of Corollary 4.4. ◇

4.4. Approximation of measures in the Hellinger metric

To implement algorithms designed to sample the posterior measure μ^y given by (4.1), we need to make finite-dimensional approximations. We study this issue here. Since the dependence on y is not relevant in this section, we study measures μ given by

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(u) = \frac{1}{Z}\exp\bigl(-\Phi(u)\bigr), \tag{4.4}$$

where the normalization constant Z is given by

$$Z = \int_X \exp\bigl(-\Phi(u)\bigr)\,\mathrm{d}\mu_0(u). \tag{4.5}$$

We approximate μ by approximating Φ. In particular, we define μ^N by

$$\frac{\mathrm{d}\mu^N}{\mathrm{d}\mu_0}(u) = \frac{1}{Z^N}\exp\bigl(-\Phi^N(u)\bigr), \tag{4.6}$$

where

$$Z^N = \int_X \exp\bigl(-\Phi^N(u)\bigr)\,\mathrm{d}\mu_0(u). \tag{4.7}$$

Our interest is in translating approximation results for Φ (determined by the forward problem) into approximation results for μ (which describes the inverse problem).

The following theorem proves such a result, bounding the Hellinger distance, and hence the total variation distance, between measures μ and μ^N in terms of the error in approximating Φ.

Theorem 4.6. Assume that the measures μ and μ^N are both absolutely continuous with respect to μ_0, satisfying $\mu_0(X) = 1$, with Radon–Nikodym derivatives given by (4.4) and (4.6) and that Φ and Φ^N satisfy Assumptions 2.6(i) and (ii) with constants uniform in N. Assume also that for any $\varepsilon > 0$ there exists $K = K(\varepsilon) > 0$ such that

$$|\Phi(u) - \Phi^N(u)| \leqslant K\exp\bigl(\varepsilon\|u\|_X^2\bigr)\psi(N), \tag{4.8}$$

where $\psi(N) \to 0$ as $N \to \infty$. Then the measures μ and μ^N are close with

respect to the Hellinger distance: there is a constant C, independent of N, and such that

$$d_{\mathrm{Hell}}(\mu, \mu^N) \leqslant C\psi(N). \tag{4.9}$$

Consequently the expectation under μ and μ^N of any polynomially bounded function $f : X \to E$ are $\mathcal{O}(\psi(N))$ close. In particular, the mean and, in the case where X is a Hilbert space, the covariance operator are $\mathcal{O}(\psi(N))$ close.

Proof. Throughout the proof, all integrals are over X. The constant C changes from occurrence to occurrence. The normalization constants Z and Z^N satisfy lower bounds which are identical to that proved for Z in the course of establishing Theorem 4.2.

From Assumption 2.6(i) and (4.8), using the fact that μ_0 is a Gaussian probability measure so that the Fernique Theorem 6.9 applies,

$$\begin{aligned} |Z - Z^N| &\leqslant \int K\psi(N) \exp\bigl(\varepsilon\|u\|_X^2 - M\bigr) \exp\bigl(\varepsilon\|u\|_X^2\bigr)\, \mathrm{d}\mu_0(u) \\ &\leqslant C\psi(N). \end{aligned}$$

From the definition of Hellinger distance, we have

$$\begin{aligned} 2d_{\mathrm{Hell}}(\mu, \mu^N)^2 &= \int \biggl(Z^{-1/2} \exp\biggl(-\frac{1}{2}\Phi(u)\biggr) \\ &\qquad - (Z^N)^{-1/2} \exp\biggl(-\frac{1}{2}\Phi^N(u)\biggr)\biggr)^2 \mathrm{d}\mu_0(u) \\ &\leqslant I_1 + I_2, \end{aligned}$$

where

$$\begin{aligned} I_1 &= \frac{2}{Z} \int \biggl(\exp\biggl(-\frac{1}{2}\Phi(u)\biggr) - \exp\biggl(-\frac{1}{2}\Phi^N(u)\biggr)\biggr)^2 \mathrm{d}\mu_0(u), \\ I_2 &= 2\bigl|Z^{-1/2} - (Z^N)^{-1/2}\bigr|^2 \int \exp\bigl(-\Phi^N(u)\bigr)\, \mathrm{d}\mu_0(u). \end{aligned}$$

Now, again using Assumption 2.6(i) and equation (4.8), together with the Fernique Theorem,

$$\begin{aligned} \frac{Z}{2} I_1 &\leqslant \int K^2 \exp\bigl(3\varepsilon\|u\|_X^2 - M\bigr)\psi(N)^2\, \mathrm{d}\mu_0(u) \\ &\leqslant C\psi(N)^2. \end{aligned}$$

A similar use of the Fernique Theorem and Assumption 2.6(i) shows that the integral in I_2 is finite. Thus, using the bounds on Z, Z^N from below,

$$\begin{aligned} \bigl|Z^{-1/2} - (Z^N)^{-1/2}\bigr|^2 &\leqslant C\bigl(Z^{-3} \vee (Z^N)^{-3}\bigr)|Z - Z^N|^2 \\ &\leqslant C\psi(N)^2. \end{aligned}$$

Combining gives the desired continuity result in the Hellinger metric.

Finally, all moments of u in X are finite under μ and μ^N because the change of measure from Gaussian μ_0 involves a term which may be controlled by the Fernique Theorem. The desired results follow from Lemma 6.37. □

Example 4.7. Consider the inverse problem for the heat equation, from Section 3.5, in the case where $D = (0, 1)$. Approximate the Bayesian inverse problem by use of a spectral approximation of the forward map $\mathrm{e}^{-AT} : \mathcal{H} \to \mathcal{H}$. Let P^N denote the orthogonal projection in $\mathcal{H}$ onto the first N eigenfunctions of A. Then, for any $T > 0$ and $r \geqslant 0$,

$$\|\mathrm{e}^{-AT} - \mathrm{e}^{-AT}P^N\|_{\mathcal{L}(\mathcal{H},\mathcal{H}^r)} = \mathcal{O}(\exp(-cN^2)).$$

From (3.25) we have the Lipschitz property that

$$|\Phi(u) - \Phi(v)| \leqslant C\big(\|u\| + \|v\| + \|y\|\big)\|\mathrm{e}^{-\frac{1}{4}AT}(u - v)\|.$$

If we define $\Phi^N(u) = \Phi(P^N u)$, then the two preceding estimates combine to give, for some $C, c > 0$ and independent of (u, y),

$$|\Phi(u) - \Phi^N(u)| \leqslant C\big(\|u\| + \|y\|\big)\|u\| \exp(-cN^2).$$

Thus (4.8) holds and Theorem 4.6 shows that the posterior measure is perturbed by a quantity with order of magnitude $\mathcal{O}(\exp(-cN^2))$ in the Hellinger metric. ◇

Remark 4.8. Approximation may come from two sources: (i) from representing the target function u in a finite-dimensional basis; and (ii) from approximating the forward model, and hence the potential Φ, by a numerical method such as a finite element or spectral method. In general these two sources of approximation error are distinct and must be treated separately. An important issue is to balance the two sources of error to optimize workload. In the case where u is a subset of, or the entire, initial condition for a dynamical system and $\mathcal{G}$ is defined through composition of some function with the solution operator, then (i) and (ii) will overlap if a spectral approximation is employed for (ii), using the finite-dimensional basis from (i). This is the situation in the preceding example. ◇

For Bayesian inverse problems with finite data, the potential Φ has the form given in (4.3), where $y \in \mathbb{R}^q$ is the data, $\mathcal{G} : X \to \mathbb{R}^q$ is the observation operator and $|\cdot|_\Gamma$ is a covariance weighted norm on $\mathbb{R}^q$. If $\mathcal{G}^N$ is an approximation to $\mathcal{G}$ and we define

$$\Phi^N := \frac{1}{2}|y - \mathcal{G}^N(u)|_\Gamma^2, \tag{4.10}$$

then we may define an approximation μ^N to μ as in (4.6). The following corollary relating μ and μ^N is useful.

Corollary 4.9. Assume that the measures μ and μ^N are both absolutely continuous with respect to μ_0, with Radon–Nikodym derivatives given by (4.4), (4.3) and (4.6), (4.10) respectively. Assume also that $\mathcal{G}$ is approximated by a function $\mathcal{G}^N$ with the property that, for any $\varepsilon > 0$, there is $K' = K'(\varepsilon) > 0$ such that

$$|\mathcal{G}(u) - \mathcal{G}^N(u)| \leqslant K' \exp\bigl(\varepsilon \|u\|_X^2\bigr)\psi(N), \tag{4.11}$$

where $\psi(N) \to 0$ as $N \to \infty$. If $\mathcal{G}$ and $\mathcal{G}^N$ satisfy Assumption 2.7(i) uniformly in N, then there is a constant C, independent of N, and such that

$$d_{\text{Hell}}(\mu, \mu^N) \leqslant C\psi(N). \tag{4.12}$$

Consequently the expectation under μ and μ^N of any polynomially bounded function $f : X \to E$ is $\mathcal{O}(\psi(N))$ close. In particular, the mean and, in the case where X is a Hilbert space, the covariance operator are $\mathcal{O}(\psi(N))$ close.

Proof. We simply show that the conditions of Theorem 4.6 hold. That (i) and (ii) of Assumptions 2.6 hold follows as in the proof of Lemma 2.8. Also (4.8) holds since (for some $K : \mathbb{R}^+ \to \mathbb{R}^+$ defined in the course of the following chain of inequalities)

$$\begin{aligned}
|\Phi(u) - \Phi^N(u)| &\leqslant \frac{1}{2}|2y - \mathcal{G}(u) - \mathcal{G}^N(u)|_\Gamma |\mathcal{G}(u) - \mathcal{G}^N(u)|_\Gamma \\
&\leqslant \bigl(|y|_\Gamma + \exp\bigl(\varepsilon\|u\|_X^2 + M(\varepsilon)\bigr)\bigr) \times K'(\varepsilon)\exp\bigl(\varepsilon\|u\|_X^2\bigr)\psi(N) \\
&\leqslant K(2\varepsilon)\exp\bigl(2\varepsilon\|u\|_X^2\bigr)\psi(N),
\end{aligned}$$

as required. □

A notable fact concerning Theorem 4.6 is that the rate of convergence attained in the solution of the forward problem, encapsulated in approximation of the function Φ by Φ^N, is transferred into the rate of convergence of the related inverse problem for measure μ given by (4.4) and its approximation by μ^N. Key to achieving this transfer of rates of convergence is the dependence of the constant in the forward error bound (4.8) on u. In particular it is necessary that this constant is integrable by use of the Fernique Theorem. In some applications it is not possible to obtain such dependence. Then convergence results can sometimes still be obtained, but at weaker rates. We state a theorem applicable in this situation.

Theorem 4.10. Assume that the measures μ and μ^N are both absolutely continuous with respect to μ_0, satisfying $\mu_0(X) = 1$, with Radon–Nikodym derivatives given by (4.4) and (4.6), and that Φ and Φ^N satisfy Assumptions 2.6(i) and (ii) with constants uniform in N. Assume also that for any $R > 0$ there is a $K = K(R) > 0$ such that, for all u with $\|u\|_X \leqslant R$,

$$|\Phi(u) - \Phi^N(u)| \leqslant K\psi(N), \tag{4.13}$$

where $\psi(N) \to 0$ as $N \to \infty$. Then the measures μ and μ^N are close with respect to the Hellinger distance:

$$d_{\mathrm{Hell}}(\mu, \mu^N) \to 0 \tag{4.14}$$

as $N \to \infty$. Consequently the expectation of any polynomially bounded function $f : X \to E$ under μ^N converges to the corresponding expectation under μ as $N \to \infty$. In particular, the mean and, in the case where X is a Hilbert space, the covariance operator converge. ◇

4.5. Discussion and bibliography

The idea of placing a number of inverse problems within a common Bayesian framework, and studying general properties in this abstract setting, is developed in Cotter *et al.* (2009). That paper contains Theorems 4.1 and 4.2 under Assumptions 2.6 in the case where (i) is satisfied trivially because Φ is bounded from below by a constant; note that this case occurs whenever the data is finite-dimensional. Generalizing the theorems to allow for (i) as stated here was undertaken in Hairer, Stuart and Voss (2010*b*), in the context of signal processing for stochastic differential equations.

Theorem 4.2 is a form of well-posedness. Recall that, in the approximation of forward problems in differential equations, well-posedness and a local approximation property form the key concepts that underpin the equivalence theorems of Dahlquist (Hairer, Nørsett and Wanner 1993, Hairer and Wanner 1996), Lax (Richtmyer and Morton 1967) and Sanz-Serna and Palencia (Sanz-Serna and Palencia 1985). It is also natural that the well-posedness that we have exhibited for inverse problems should, when combined with forward approximation, give rise to approximation results for the inverse problem. This is the basic idea underlying Theorem 4.6. That result, Corollary 4.9 and Theorem 4.10 are all stated and proved in Cotter *et al.* (2010*a*).

The underlying well-posedness of properly formulated Bayesian inverse problems has a variety of twists and turns which we do not elaborate fully here. The interested reader should consult Dashti *et al.* (2010*b*).

5. Algorithms

5.1. Overview

We have demonstrated that a wide range of inverse problems for functions u given data y give rise to a posterior measure μ^y with the form (2.24). This formula encapsulates neatly the ideal information that we have about a function, formed from conjunction of model and data. Furthermore, for many applications, the potential Φ satisfies Assumptions 2.6. From this we have shown in Section 4 that the formula (2.24) indeed leads to a well-defined

posterior μ^y and that this measure enjoys nice robustness properties with respect to changes in the data or approximation of the forward problem. However, we have not yet addressed the issue of how to obtain information from the formula (2.24) for the posterior measure. We devote this section to an overview of the computational issues which arise in this context.

If the prior measure is Gaussian and the potential $\Phi(\cdot; y)$ is quadratic, then the posterior is also Gaussian. This situation arises, for example, in the inverse problem for the heat equation described in Section 3.5. The measure μ^y is then characterized by a function (the mean) and an operator (the covariance) and formulae can be obtained for these quantities by completing the square using Theorem 6.20: see the developments for the heat equation, or Example 6.23, for an illustration of this.

However, in general there is no explicit way of characterizing the measure μ^y as can be done in the Gaussian case. Thus approximations and computational tools are required to extract information from the formula (2.24). One approach to this problem is to employ sampling techniques which (approximately) generate sample functions according to the probability distribution implied by (2.24). Among the most powerful generic tools for sampling are the *Markov chain Monte Carlo* (MCMC) methods, which we review in Section 5.2. However, whilst these methods can be very effective when tuned carefully to the particular problem at hand, they are undeniably costly and, for many applications, impracticable at current levels of computer resources. For this reason we also devote two subsections to *variational* and *filtering* methods, which are widely used in practice because of their computational expedience. When viewed in terms of their relation to (2.24) these methods constitute approximations. Furthermore, these approximations are, in many cases, not well understood. In the near future we see the main role of MCMC methods as providing controlled approximations to the true posterior measure μ^y, against which variational and filtering methodologies can be tested, on well-designed model problems. In the longer term, as computational power and algorithmic innovation grows, we also anticipate increasing use of MCMC methods in their own right to approximate (2.24).

From a Bayesian perspective, the variational methods of Section 5.3 start from the premise that variability in the posterior measure is small and that most of the information resides in a single peak of the probability distribution, which can be found by optimization techniques. We view this problem from the standpoint of optimal control, showing that a minimizer exists whenever the common framework of Section 2.4 applies; we also review algorithmic practice in the area. Section 5.4 describes the widely used filtering methods which approximate the posterior measure arising in time-dependent data assimilation problems by a sequence of probability measures in time, updated sequentially. The importance of this class of algorithms

stems from the fact that, in many applications, solutions are required online, with updates required as more data is acquired; thus sequential updating of the posterior measure at the current time is natural. Furthermore, sequential updates are computationally efficient as they reduce the dimensionality of the desired posterior measure, breaking a correlated measure at a sequence of times into a sequence of conditionally independent measures at each time, provided there is an underlying Markovian structure. We conclude, in Section 5.5, with references to the literature concerning algorithms.

When discussing MCMC methods and variational methods, the dependence of the potential Φ appearing in (2.24) will not be relevant and we will consider the problem for the posterior measure written in the form

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(u) = \frac{1}{Z}\exp\bigl(-\Phi(u)\bigr), \tag{5.1}$$

with normalization constant

$$Z = \int_X \exp\bigl(-\Phi(u)\bigr)\,\mathrm{d}\mu_0(u). \tag{5.2}$$

We refer to μ as the *target distribution*. For the study of both MCMC and variational methods, we will also find it useful to define

$$I(u) = \frac{1}{2}\|u\|_{\mathcal{C}}^2 + \Phi(u). \tag{5.3}$$

This is, of course, a form of regularized least-squares functional as introduced in Section 2.

5.2. Markov chain Monte Carlo

The basic idea of MCMC methods is simple: design a Markov chain with the property that a single sequence of output from the chain $\{u_n\}_{n=0}^{\infty}$ is distributed according to μ given by (5.1). This is a *very* broad algorithmic prescription and allows for significant innovation in the design of methods tuned to the particular structure of the desired target distribution. We will focus on a particular class of MCMC methods known as *Metropolis–Hastings* (MH) methods.

The key ingredient of these methods is a probability measure on X, parametrized by $u \in X$: a Markov transition kernel $q(u, dv)$. This kernel is used to propose moves from the current state of the Markov chain u_n to a new point distributed as $q(u_n, \cdot)$. This proposed point is then accepted or rejected according to a criterion which uses the target distribution μ. The resulting Markov chain has the desired property of preserving the target distribution. Key to the success of the method is the choice of q. We now give details of how the method is constructed.

Given $q(u,\cdot)$ and the target μ we define a new measure on $X\times X$ defined by

$$\nu(\mathrm{d}u,\mathrm{d}v) = q(u,dv)\mu(\mathrm{d}u).$$

We define the same measure, with the roles of u and v reversed, by

$$\nu^\top(\mathrm{d}u,\mathrm{d}v) = q(v,\mathrm{d}u)\mu(\mathrm{d}v).$$

Provided that $\nu^\top$ is absolutely continuous with respect to ν, we may define

$$\alpha(u,v) = \min\left\{1, \frac{\mathrm{d}\nu^\top}{\mathrm{d}\nu}(u,v)\right\}.$$

Now define a random variable $\gamma(u,v)$, independent of the probability space underlying the transition kernel q, with the property that

$$\gamma(u,v) = \begin{cases} 1 & \text{with probability } \alpha(u,v), \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

We now create a random Markovian sequence $\{u_n\}_{n=0}^\infty$ as follows. Given a proposal $v_n \sim q(u_n,\cdot)$, we set

$$u_{n+1} = \gamma(u_n,v_n)v_n + (1-\gamma(u_n,v_n))u_n. \tag{5.5}$$

If we choose the randomness in the proposal v_n and the binary random variable $\gamma(u_n,v_n)$ independently of each other for each n, and independently of their values for different n, then this construction gives rise to a Markov chain with the desired property.

Theorem 5.1. Under the given assumptions, the Markov chain defined by (5.5) is invariant for μ: if $u_0 \sim \mu$ then $u_n \sim \mu$ for all $n \geqslant 0$. Furthermore, if the resulting Markov chain is ergodic then, for any continuous bounded function $f : X \to \mathbb{R}$, any $M \geqslant 0$, and for u_0 μ-a.s.,

$$\frac{1}{N}\sum_{n=1}^{N} f(u_{n+M}) \to \int_X f(u)\mu(\mathrm{d}u) \quad \text{as} \quad N\to\infty. \tag{5.6}$$

◇

In words, this theorem states that the empirical distribution of the Markov chain converges weakly to that of the target measure μ. However, this nice abstract development has not addressed the question of actually constructing an MH method. If $X = \mathbb{R}^n$ and the target measures have positive density with respect to Lebesgue measure, then this is straightforward: any choice of kernel $q(u,dv)$ will suffice, provided it too has positive density with respect to Lebesgue measure, for every u. It then follows that $\nu^\top \ll \nu$. From this wide range of admissible proposal distributions, the primary design choice is to identify proposals which lead to low correlation in the resulting Markov chain, as this increases efficiency.

Example 5.2. A widely used proposal kernel is simply that of a *random walk*; for example, if $\mu_0 = \mathcal{N}(0, \mathcal{C})$ it is natural to propose

$$v = u + \sqrt{2\delta}\xi, \tag{5.7}$$

where $\xi \sim \mathcal{N}(0, \mathcal{C})$. A straightforward calculation shows that

$$\alpha(u, v) = \min\{1, \exp\big(I(u) - I(v)\big)\}$$

where I is given by (5.3). Thus, if the proposed state corresponds to a lower value of the regularized least-squares functional I, then the proposal is automatically accepted; otherwise it will be accepted with a probability depending on $I(u) - I(v)$.

The parameter δ is a scalar which controls the size of the move. Large values lead to proposals which are hence unlikely to be accepted, leading to high correlation in the Markov chain. On the other hand small moves do not move very far, again leading to high correlation in the Markov chain. Identifying appropriate values of δ between these extremes is key to making effective algorithms. More complex proposals use additional information about $D\Phi$ in an attempt to move into regions of high probability (low Φ). $\diamond$

In infinite dimensions things are not so straightforward: a random walk will not typically deliver the required condition $\nu^\top \ll \nu$. For example, if $\mu_0 = \mathcal{N}(0, \mathcal{C})$ and X is infinite-dimensional, then the proposal (5.7) will not satisfy this constraint. However, a little thought shows that appropriate modifications are possible.

Example 5.3. The random walk can be modified to obtain the desired absolute continuity of $\nu^\top$ with respect to ν. The proposal

$$v = (1 - 2\delta)^{1/2}u + \sqrt{2\delta}\xi, \tag{5.8}$$

where $\xi \sim \mathcal{N}(0, \mathcal{C})$, will satisfy the desired condition for any $\delta \in \mathbb{R}$. The acceptance probability is

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

Thus, if the proposed state corresponds to a lower value of Φ than does the current state, it will automatically be accepted.

The proposal in (5.8) should be viewed as an appropriate analogue of the random walk proposal in infinite-dimensional problems. Intuition as to why this proposal works in the infinite-dimensional setting can be obtained by observing that, if $u \sim \mathcal{N}(0, \mathcal{C})$ and v is constructed using (5.8), then $v \sim \mathcal{N}(0, \mathcal{C})$; thus the proposal preserves the underlying reference measure (prior) μ_0. In contrast, the proposal (5.7) does not: if $u \sim \mathcal{N}(0, \mathcal{C})$ then $v \sim \sqrt{(1 + 2\delta)}\mathcal{N}(0, \mathcal{C})$.

Note that the choice $\delta = 1/2$ in (5.8) yields an *independence sampler* where proposals v are made from the prior measure μ_0, independently of the current state of the Markov chain u. As in finite dimensions, improved proposals can be found by including information about $D\Phi$ in the proposal. ◇

In computational practice, of course, we always implement a sampling method in finite dimensions. The error incurred by doing so may be quantified by use of Theorem 4.6. It is natural to ask whether there is any value in deriving MH methods on function space, especially since this appears harder than doing so in finite dimensions. The answer, of course, is 'yes'. Any MH method in finite dimensions which does not correspond to a well-defined limiting MH method in the function space (infinite-dimensional) limit will degenerate as the dimension of the space increases. This effect can be quantified and compared with what happens when proposals defined on function space are used. In conclusion, then, the function space viewpoint on MCMC methods is a useful one which leads to improved algorithms, and an understanding of the shortcomings of existing algorithms.

5.3. Variational methods

Variational methods attempt to answer the following question: 'How do we find the most likely function u under the posterior measure μ given by (5.1)?' To understand this consider first the case where $X = \mathbb{R}^n$ and $\mu_0 = \mathcal{N}(0, \mathcal{C})$ is a Gaussian prior. Then μ has density with respect to Lebesgue measure and the negative logarithm of this density is given by (5.3).[3] Thus the Lebesgue density of μ is maximized by minimizing I over $\mathbb{R}^n$. Another way of looking at this is as follows: if $\overline{u}$ is such a minimizer then the probability of a small ball of radius ε and centred at u will be maximized, asymptotically as $\varepsilon \to 0$, by choosing $u = \overline{u}$.

If X is an infinite-dimensional Hilbert space then there is no Lebesgue measure on X, and we cannot directly maximize the density. However, we may again consider the probability of small balls at $u \in X$, of radius ε. We may then ask how u should be chosen to maximize the probability of the ball, asymptotically as $\varepsilon \to 0$. Again taking $\mu_0 = \mathcal{N}(0, \mathcal{C})$ this question leads to the conclusion that u should be chosen as a global minimizer of I given by (5.3) over the Cameron–Martin space E with inner product $\langle \cdot, \cdot \rangle_{\mathcal{C}}$ and norm $\| \cdot \|_{\mathcal{C}}$.

Recall that Φ measures model/data mismatch, in the context of applications to inverse problems. In the case where y is finite-dimensional it has the form (4.3). It is thus natural to minimize Φ directly, as in (2.2). However, when X is infinite-dimensional, this typically leads to minimizing sequences

[3] Recall that for economy of notation we drop explicit reference to the y dependence of Φ in this subsection, as it plays no role.

which do not converge in any reasonable topology. The addition of the quadratic penalization in E may be viewed as a *Tikhonov regularization* to overcome this problem. Minimization of I is thus a regularized *nonlinear least-squares problem* as in (2.3). Of course this optimization approach can be written down directly, with no reference to probability. The beauty of the Bayesian approach is that it provides a rational basis for the choice of norms underlying the objective functional Φ, as well as the choice of norm in the regularization term proportional to $\|u\|_{\mathcal{C}}^2$. Furthermore, the Bayesian viewpoint gives an interpretation of the resulting optimization problem as a probability maximizer. And finally the framework of Section 2.4, which leads to well-posed posterior measures, also leads directly to an existence theory for probability maximizers. We now describe this theory.

Theorem 5.4. Let Assumptions 2.6(i) and (iii) hold, and assume that $\mu_0(X) = 1$. Then there exists $\overline{u} \in E$ such that

$$I(\overline{u}) = \overline{I} := \inf\{I(u) : u \in E\}.$$

Furthermore, if $\{u_n\}$ is a minimizing sequence satisfying $I(u_n) \to I(\overline{u})$, then there is a subsequence $\{u_{n'}\}$ that converges strongly to $\overline{u}$ in E.

Proof. First we show that I is weakly lower semicontinuous on E. Let $u_n \rightharpoonup \overline{u}$ in E. By the compact embedding of E in X, which follows from Theorem 6.11 since $\mu_0(X) = 1$, we deduce that $u_n \to \overline{u}$, strongly in X. By the Lipschitz continuity of Φ in X (Assumption 2.6(iii)) we deduce that $\Phi(u_n) \to \Phi(\overline{u})$. Thus Φ is weakly continuous on E. The functional $J(u) := \frac{1}{2}\|u\|_{\mathcal{C}}^2$ is weakly lower semicontinuous on E. Hence $I(u) = J(u) + \Phi(u)$ is weakly lower semicontinuous on E.

Now we show that I is coercive on E. Again using the fact that E is compactly embedded in X, we deduce that there is a $K > 0$ such that

$$\|u\|_X^2 \leqslant K\|u\|_{\mathcal{C}}^2.$$

Hence, by Assumption 2.6(i), it follows that, for any $\varepsilon > 0$, there is an $M(\varepsilon) \in \mathbb{R}$ such that

$$\left(\frac{1}{2} - K\varepsilon\right)\|u\|_{\mathcal{C}}^2 + M(\varepsilon) \leqslant I(u).$$

By choosing ε sufficiently small, we deduce that there is an $M \in \mathbb{R}$ such that, for all $u \in E$,

$$\frac{1}{4}\|u\|_{\mathcal{C}}^2 + M \leqslant I(u). \tag{5.9}$$

This establishes coercivity.

Consider a minimizing sequence. For any $\delta > 0$ there is an $N_1 = N_1(\delta)$:

$$M \leqslant \overline{I} \leqslant I(u_n) \leqslant \overline{I} + \delta, \quad \forall n \geqslant N_1.$$

Using (5.9) we deduce that the sequence $\{u_n\}$ is bounded in E and, since E is a Hilbert space, there exists $\overline{u} \in E$ such that (possibly along a subsequence) $u_n \rightharpoonup \overline{u}$ in E. From the weak lower semicontinuity of I it follows that, for any $\delta > 0$,

$$\overline{I} \leqslant I(\overline{u}) \leqslant \overline{I} + \delta.$$

Since δ is arbitrary the first result follows.

Now consider the subsequence $u_n \rightharpoonup \overline{u}$. Then there is an $N_2 = N_2(\delta) > 0$ such that, for $n, \ell \geqslant N_2$,

$$\begin{aligned}
\frac{1}{4}\|u_n - u_\ell\|_{\mathcal{C}}^2 &= \frac{1}{2}\|u_n\|_{\mathcal{C}}^2 + \frac{1}{2}\|u_\ell\|_{\mathcal{C}}^2 - \|\frac{1}{2}(u_n + u_\ell)\|_{\mathcal{C}}^2 \\
&= I(u_n) + I(u_\ell) - 2I\left(\frac{1}{2}(u_n + u_\ell)\right) - \Phi(u_n) \\
&\quad - \Phi(u_\ell) + 2\Phi\left(\frac{1}{2}(u_n + u_\ell)\right) \\
&\leqslant 2(\overline{I} + \delta) - 2\overline{I} - \Phi(u_n) - \Phi(u_\ell) + 2\Phi\left(\frac{1}{2}(u_n + u_\ell)\right) \\
&\leqslant 2\delta - \Phi(u_n) - \Phi(u_\ell) + 2\Phi\left(\frac{1}{2}(u_n + u_\ell)\right).
\end{aligned}$$

But u_n, u_ℓ and $\frac{1}{2}(u_n + u_\ell)$ all converge strongly to $\overline{u}$ in X. Thus, by continuity of Φ, we deduce that, for all $n, \ell \geqslant N_3(\delta)$,

$$\frac{1}{4}\|u_n - u_\ell\|_{\mathcal{C}}^2 \leqslant 3\delta.$$

Hence the sequence is Cauchy in E and converges strongly, and the proof is complete. □

Example 5.5. Recall the inverse problem for the diffusion coefficient of the one-dimensional elliptic problem described in Section 3.3. The objective is to find $u(x)$ appearing in

$$\begin{aligned}
-\frac{\mathrm{d}}{\mathrm{d}x}\left(\exp(u(x))\frac{\mathrm{d}p}{\mathrm{d}x}\right) &= 0, \\
p(0) = p^- \quad p(1) &= p^+,
\end{aligned}$$

where $p^+ > p^-$. The observations are

$$y_k = p(x_k) + \eta_k, \quad k = 1, \ldots, q$$

written succinctly as

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

where $\eta \in \mathbb{R}^q$ is distributed as $\mathcal{N}(0, \gamma^2 I)$. The function $\mathcal{G}$ is Lipschitz in the space of continuous functions $X = C([0,1])$ by Lemma 3.3.

Recall that changing u by an arbitrary additive constant does not change the solution of (3.5), and so we assume that u integrates to zero on $(0,1)$. We define

$$\mathcal{H} = \left\{ u \in L^2((0,1)) \Big| \int_0^1 u(x)\,\mathrm{d}x = 0 \right\}.$$

We take $\mathcal{A} = -\mathrm{d}^2/\mathrm{d}x^2$ with

$$D(\mathcal{A}) = \left\{ u \in H^2_{\mathrm{per}}((0,1)) \Big| \int_0^1 u(x)\,\mathrm{d}x = 0 \right\}.$$

Then $\mathcal{A}$ is positive definite self-adjoint, and we may define the prior Gaussian measure $\mu_0 = \mathcal{N}(0, \mathcal{A}^{-1})$ on $\mathcal{H}$. By Lemma 6.25 we deduce that $\mu_0(X) = 1$. The Cameron–Martin space

$$E = \mathrm{Im}(\mathcal{A}^{-1/2}) = \left\{ u \in H^1_{\mathrm{per}}((0,1)) \Big| \int_0^1 u(x)\,\mathrm{d}x = 0 \right\}$$

is compactly embedded into $C([0,1])$ by Theorem 2.10; this is also a consequence of the general theory of Gaussian measures since $\mu_0(X) = 1$. By the Lipschitz continuity of $\mathcal{G}$ in X and Theorem 5.4, we deduce that

$$I(u) := \frac{1}{2}\|u\|^2_{H^1_{\mathrm{per}}} + \frac{1}{2\gamma^2}|y - \mathcal{G}(u)|^2$$

attains its infimum at $\overline{u} \in E$. ◇

In summary, the function space Bayesian viewpoint on inverse problems is instructive in developing an understanding of variational methods. In particular it implicitly guides choice of the regularization that will lead to a well-posed minimization problem.

5.4. Filtering

There are two key ideas underlying filtering: the first is to build up knowledge about the posterior sequentially, and hence perhaps more efficiently; the second is to break up the unknown u and build up knowledge about its constituent parts sequentially, hence reducing the computational dimension of each sampling problem. Thus the first idea relies on decomposing the *data* sequentially, whilst the second relies on decomposing the *unknown* sequentially.

The first basic idea is to build up information about μ^y sequentially as the size of the data set increases. For simplicity assume that the data is finite-dimensional and can be written as $y = \{y_j\}_{j=1}^{\mathrm{J}}$. Assume also that each data point y_j is found from a mapping $\mathcal{G}_j : X \to \mathbb{R}^\ell$ and subject to independent Gaussian observational noises $\eta_j \sim \mathcal{N}(0, \Gamma_j)$ so that

$$y_j = \mathcal{G}_j(u) + \eta_j. \tag{5.10}$$

Thus the data is in $\mathbb{R}^q$ for $q = \ell \mathrm{J}$. The posterior measure has the form

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2}\sum_{j=1}^{\mathrm{J}} |y_j - \mathcal{G}_j(u)|^2_{\Gamma_j}\right). \tag{5.11}$$

Now let μ^y_i denote the posterior distribution given only the data $y = \{y_j\}_{j=1}^i$. Then

$$\frac{\mathrm{d}\mu^y_i}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2}\sum_{j=1}^{i} |y_j - \mathcal{G}_j(u)|^2_{\Gamma_j}\right). \tag{5.12}$$

Furthermore, setting $\mu^y_0 = \mu_0$, we have

$$\frac{\mathrm{d}\mu^y_{i+1}}{\mathrm{d}\mu^y_i}(u) \propto \exp\left(-\frac{1}{2}|y_{i+1} - \mathcal{G}_{i+1}(u)|^2_{\Gamma_j}\right). \tag{5.13}$$

Compare formulae (5.11) and (5.13). When J is large, it is intuitive that μ^y_{i+1} is closer to μ^y_i than $\mu^y = \mu^y_{\mathrm{J}}$ is to μ_0. This suggests that formula (5.13) may be used as the basis for obtaining μ^y_{i+1} from μ^y_i, and thereby to approach $\mu^y = \mu^y_{\mathrm{J}}$ by iterating this over i. In summary, the first key idea enables us to build up our approximation to μ^y incrementally over an ordered set of data.

The second key idea involves additional structure. Imagine that we have $y_j = y(t_j)$ for some set of times

$$0 \leqslant t_1 < t_2 < \cdots < t_{\mathrm{J}} < \infty.$$

Assume furthermore that u is also time-dependent and can be decomposed as $u = \{u_j\}_{j=1}^{\mathrm{J}}$, where $u_j = u(t_j)$, and that (5.10) simplifies to

$$y_j = \mathcal{G}_j(u_j) + \eta_j. \tag{5.14}$$

Then it is reasonable to seek to find the conditional measures

$$\nu_{i|1:i}(\mathrm{d}u_i) := \mathbb{P}(\mathrm{d}u_i|\{y_j\}_{j=1}^i). \tag{5.15}$$

Notice that each of these measures lives on a smaller space than does μ^y and this dimension reduction is an important feature of the methodology. Assuming that the sequence $u = \{u_j\}_{j=1}^{\mathrm{J}}$ is governed by a Markovian evolution, the measure (5.15) uniquely determines the measure

$$\nu_{i+1|1:i}(\mathrm{d}u_{i+1}) := \mathbb{P}(\mathrm{d}u_{i+1}|\{y_j\}_{j=1}^i).$$

Incorporating the $(i+1)$st data point, we find that

$$\frac{\mathrm{d}\nu_{i+1|1:i+1}}{\mathrm{d}\nu_{i+1|1:i}}(u_{i+1}) \propto \exp\left(-\frac{1}{2}|y_{i+1} - \mathcal{G}_{i+1}(u_{i+1})|^2_{\Gamma_j}\right). \tag{5.16}$$

Thus we have a way of building the measures given by (5.15) incrementally in i.

Clearly, by definition, $\nu_{\mathrm{J}|1:\mathrm{J}}(\mathrm{d}u_{\mathrm{J}})$ agrees with the marginal distribution of $\mu^y(\mathrm{d}u)$ on the coordinate $u_{\mathrm{J}} = u(t_{\mathrm{J}})$; however, the distribution of $\nu_{i|1:i}(\mathrm{d}u_i)$ for $i < \mathrm{J}$ does not agree with the marginal distribution of $\mu^y(\mathrm{d}u)$ on coordinate $u_i = u(t_i)$. Thus the algorithm is potentially very powerful at updating the current state of the system given data up to that time; but it fails to update previous states of the system, given data that subsequently becomes available. We discuss the implications of this in Section 5.5.

5.5. Discussion and bibliography

We outline the methods described in this section, highlight some relevant related literature, and discuss inter-relations between the methodologies. A number of aspects concerning computational methods for inverse problems, both classical and statistical, are reviewed in Vogel (2002). An important conceptual algorithmic distinction to make in time-dependent data assimilation problems is between *forecasting* methods, which are typically used online to make predictions as data is acquired sequentially, and *hindcasting* methods which are used offline to obtain improved understanding (this is also called *reanalysis*) and, for example, may be used for the purposes of parameter estimation to obtain improved models. MCMC methods are natural for hindcasting and reanalysis; filtering is natural in the forecasting context. Filtering methods update the estimate of the state based only on data from the past, whereas the full posterior measure estimates the state at any given time based on both past and future observations; methods based on this full posterior measure are known as *smoothing methods* and include MCMC methods based on the posterior and variational methods which maximize the posterior probability.

The development of MCMC methods was initiated with the paper by Metropolis, Rosenbluth, Teller and Teller (1953), in which a symmetric random walk proposal was used to determine thermodynamic properties, such as the equation of state, from a microscopic statistical model. Hastings (1970) demonstrated that the idea could be generalized to quite general families of proposals, providing the seed for the study of these methods in the statistics community (Gelfand and Smith 1990, Smith and Roberts 1993, Bernardo and Smith 1994). The paper of Tierney (1998) provides the infinite-dimensional framework for MH methods that we outline here; in particular, Theorem 5.1 follows from the work in that paper. Ergodic theorems, such as the convergence of time averages as in (5.6), can in many cases be proved for much wider classes of functions than continuous bounded functions. The general methodology is described in Meyn and Tweedie (1993) and an application to MH methods is given in Roberts and Tweedie (1996).

The degeneration of many MH methods on state spaces of finite but growing dimension is a well-known phenomenon to many practitioners. An analysis and quantification of this effect was first undertaken in Roberts,

Gelman and Gilks (1997), where random walk proposals were studied for an i.i.d. target, and subsequently in Roberts and Rosenthal (1998, 2001), Beskos and Stuart (2009) and Beskos, Roberts and Stuart (2009) for other target distributions and proposals; see Beskos and Stuart (2010) for an overview. The idea of using proposals designed to work in the infinite-dimensional context to overcome this degeneration is developed in Stuart, Voss and Wiberg (2004) and Beskos, Roberts, Stuart and Voss (2008) in the context of sampling conditioned diffusions, and is described more generally in Beskos and Stuart (2009), Beskos *et al.* (2009), Beskos and Stuart (2010) and Cotter, Dashti, Robinson and Stuart (2010*b*).

The use of MCMC methods for sampling the posterior distribution arising in the Bayesian approach to inverse problems is highlighted in Kaipio and Somersalo (2000, 2005), Calvetti and Somersalo (2006) and Calvetti, Kuceyeski and Somersalo (2008). When sampling complex high-dimensional posterior distributions, such as those that arise from finite-dimensional approximation of measures μ^y given by (2.24), can be extremely computationally challenging. It is, however, starting to become feasible; recent examples of work in this direction include Calvetti and Somersalo (2006), Dostert *et al.* (2006), Kaipio and Somersalo (2000), Heino, Tunyan, Calvetti and Somersalo (2007) and Calvetti, Hakula, Pursiainen and Somersalo (2009). In Cotter *et al.* (2010*b*) inverse problems such as those in Section 3.6 are studied by means of the MH technology stemming from the proposal (5.8). Examples of application of MCMC techniques to the statistical solution of inverse problems arising in oceanography, hydrology and geophysics may be found in Efendiev *et al.* (2009), Cui, Fox, Nicholls and O'Sullivan (2010), McKeague, Nicholls, Speer and Herbei (2005), Herbei, McKeague and Speer (2008), McLaughlin and Townley (1996), Michalak and Kitanidis (2003) and Mosegaard and Tarantola (1995). The paper by Herbei and McKeague (2009) studies the geometric ergodicity properties of the resulting Markov chains, employing the framework developed in Meyn and Tweedie (1993).

The idea of using proposals more general than (5.7), and in particular proposals that use derivative information concerning Φ, is studied in Roberts and Tweedie (1996). A key concept here is the *Langevin equation*: a stochastic differential equation for which μ is an invariant measure. Discretizing this equation, which involves the derivative of Φ, is the basis for good proposals. This is related to the fact that, for small discretization parameter, the proposals nearly inherit this invariance under μ. Applying this idea in the infinite-dimensional context is described in Apte, Hairer, Stuart and Voss (2007) and Beskos and Stuart (2009), based on the idea of Langevin equations in infinite dimensions (Hairer *et al.* 2005, Hairer *et al.* 2007, Hairer, Stuart and Voss 2009).

Characterizing the centres of small balls with maximum probability has been an object of interest in the theory of stochastic differential equations for

some time. See Ikeda and Watanabe (1989) and Dürr and Bach (1978) for the simplest setting, and Zeitouni and Dembo (1987) for a generalization to signal processing problems. Our main Theorem 5.4 concerning the existence of probability maximizers provides a nice link between Bayesian inverse problems and optimal control. The key ingredients are continuity of the forward mapping from the unknown function to the data, in the absence of observational noise, in a space X, and choice of a prior measure which has the properties that draws from it are almost surely in X: $\mu_0(X) = 1$; this then guarantees that the Tikhonov regularization, which is in the Cameron–Martin norm for the prior measure, is sufficient to prove existence of a minimizer for the variational method.

The idea concluding the proof of the first part of Theorem 5.4 is standard in the theory of calculus of variations: see Dacarogna (1989, Chapter 3, Theorem 1.1). The strong convergence argument generalizes an argument from Kinderlehrer and Stampacchia (1980, Theorem II.2.1). The PhD thesis of Nodet (2005) contains a specific instance of Theorem 5.4, for a model of Lagrangian data assimilation in oceanography, and motivated the approach that we take here; related work is undertaken in White (1993) for Burgers' equation. An alternative approach to the existence of minimizers is to study the Euler–Lagrange equations. The paper of Hagelberg *et al.* (1996) studies existence by this approach for a minimization problem closely related to the MAP estimator. The paper studies the equations of fluid mechanics, formulated in terms vorticity–streamfunction variables. Their approach has the disadvantage of requiring a derivative to define the Euler–Lagrange equations, a short time interval to obtain existence of a solution, and also requires further second-derivative information to distinguish between minimizers and saddle points. However, it does form the basis of a numerical approach to find the MAP estimator. For linear differential equations subject to Gaussian noise there is a beautiful explicit construction of the MAP estimator, using the Euler–Lagrange equations, known as the *representer method*. This method is described in Bennett (2002).

Variational methods in image processing are reviewed in Scherzer *et al.* (2009) and the Bayesian approach to this field is exemplified by Calvetti and Somersalo (2005*b*, 2007*a*, 2008) and, implicitly, in Ellerbroek and Vogel (2009). Variational methods are known in the atmospheric and oceanographic literature as *4DVAR* methods (Derber 1989, Courtier and Talagrand 1987, Talagrand and Courtier 1987, Courtier 1997) and, as we have shown, they are linked to probability maximizers. In the presence of model error the method is known as *weak constraint 4DVAR* (Zupanski 1997). There are also variational methods for sequential problems which update the probability maximizer at a sequence of times; this methodology is known as *3DVAR* (Courtier *et al.* 1998) and is closely related to filtering. Indeed, although filtering and variational methods may be viewed as competing methodologies,

they are, in fact, not distinct methodologies, and hybrid methods are sought which combine the advantages of both; see Kalnay, Li, Miyoshi, Yang and Ballabrera-Poy (2007), for example.

Although we strongly advocate the function space viewpoint on variational methods, a great deal of work is carried out by first discretizing the problem and then defining the variational problem. Some representative papers which take this approach for large-scale applications arising in fluid mechanics include Bennett and Miller (1990), Bennett and Chua (1994), Eknes and Evensen (1997), Chua and Bennett (2001), Yu and O'Brien (1991), Watkinson, Lawless, Nichols and Roulstone (2007), Gratton, Lawless and Nichols (2007), Johnson, Hoskins, Nichols and Ballard (2006), Lawless and Nichols (2006), Johnson, Hoskins and Nichols (2005), Lawless, Gratton and Nichols (2005*b*, 2005*a*), Stanton, Lawless, Nichols and Roulstone (2005) and Wlasak and Nichols (1998). The paper of Griffith and Nichols (1998) contains an overview of adjoint methods, used in the solution of data assimilation problems with model error, primarily in the context of variational methods. A discussion of variational methods for the Lorenz equations, and references to the extensive literature in this area, may be found in Evensen (2006).

The regularized nonlinear least-squares or Tikhonov approach to inverse problems is widely studied, including in the infinite-dimensional setting of Hilbert spaces – see the book by Engl *et al.* (1996) and the references therein – and Banach spaces – see the papers by Kaltenbacher *et al.* (2009), Neubauer (2009) and Hein (2009) and the references therein. Although we have concentrated on Bayesian priors, and hence on regularization via addition of a quadratic penalization term, there is active research in the use of different regularizations (Kaltenbacher *et al.* 2009, Neubauer 2009, Hein 2009, Lassas and Siltanen 2004). In particular, the use of total variation-based regularization, and related wavelet-based regularizations, is central in image processing (Rudin *et al.* 1992).

Solving the very high-dimensional optimization problems which arise from discretizing the minimization problem (5.3) is extremely challenging and, as with filtering methods, ideas from model reduction (Antoulas, Sorensen and Gugerrin 2001) are frequently used to obtain faster algorithms. Some applications of model reduction techniques, mainly to data assimilation problems arising in fluid mechanics, may be found in Lawless, Nichols, Boess and Bunse-Gerstner (2008*a*, 2008*b*), Griffith and Nichols (1998, 2000), Akella and Navon (2009), Fang *et al.* (2009*a*, 2009*b*) and the references therein. Another approach to dealing with the high-dimensional problems that arise in data assimilation is to use ideas from machine learning (Mitchell *et al.* 1990) to try to find good quality low-dimensional approximations to the posterior measure; see, for example, Shen *et al.* (2008*b*), Shen, Cornford, Archambeau and Opper (2010), Vrettas, Cornford and Shen (2009), Shen,

Archambeau, Cornford and Opper (2008*a*), Archambeau, Opper, Shen, Cornford and Shawe-Taylor (2008) and Archambeau, Cornford, Opper and Shawe-Taylor (2007).

There are some applications where the objective functional may not be differentiable. This can arise for two primary reasons. Firstly the PDE model itself may have discontinuous features arising from switches, or shock-like solutions; and secondly the method of observing the PDE may have switches at certain threshold values of the physical parameters. In this case it is of interest to find computational algorithms to identify MAP estimators which do not require derivatives of the objective functional; see Zupanski, Navon and Zupanski (2008).

An overview of the algorithmic aspects of particle filtering, for non-Gaussian problems, is contained in the edited volume by Doucet and Gordon (2001) and a more mathematical treatment of the subject may be found in Bain and Crisan (2009). An introduction to filtering in continuous time, and a derivation of the Kalman–Bucy filter in particular, which exploits the Gaussian structure of linear problems with additive Gaussian noise, is undertaken in Oksendal (2003). It should be emphasized that these methods are all developed primarily in the context of low-dimensional problems. In practice filtering in high-dimensional systems is extremely hard. This is because the iterative formulae (5.13) and (5.16) do not express the density of the target measure with respect to an easily understood Gaussian measure, as happens in (2.24). To overcome this issue, particle approximations of the reference measures are used, corresponding to approximation by Dirac masses; thus the algorithms build up sequential approximations based on Dirac masses. In high dimensions this can be extremely computationally demanding and various forms of approximation are employed to deal with the curse of dimensionality. See Bengtsson, Bickel and Li (2008) and Bickel, Li and Bengtsson (2008) for discussion of the fundamental difficulties arising in high-dimensional filtering, and Snyder, Bengtsson, Bickel and Anderson (2008) for a development of these ideas in the context of applications. A review of some recent mathematical developments in the subject of high-dimensional filtering, especially in the context of the modelling or turbulent atmospheric flows, may be found in Majda, Harlim and Gershgorin (2010). A review of filtering from the perspective of geophysical applications, may be found in Van Leeuwen (2009). A widely used approach is that based on the ensemble Kalman filter (Burgers, Van Leeuwen and Evensen 1998, Evensen and Van Leeuwen 2000, Evensen 2006), which uses an ensemble of particles to propagate the dynamics, but incorporates data using a Gaussian approximation which is hard to justify in general; see also Berliner (2001) and Ott *et al.* (2004). Further approaches based on the use of ensembles to approximate error covariance propagation may be found in Chorin and Krause (2004) and Livings, Dance and Nichols (2008). The

paper of Bengtsson, Snyder and Nychka (2003) describes a generalization of the ensemble Kalman filter, based on mixtures of Gaussians, motivated by the high-dimensional systems arising in fluid dynamics data assimilation problems. The paper of Bennett and Budgell (1987) studies the use of filtering techniques in high dimensions, motivated by oceanographic data assimilation, and contains a study of the question of how to define families of finite-dimensional filters which converge to a function-space-valued limit as the finite-dimensional computation is refined; it is thus related to the concept of discretization invariance referred to in Section 2.5. However, the methodology for proving limiting behaviour in Bennett and Budgell (1987), based on Fourier analysis, is useful only for linear Gaussian problems; in contrast, the approach developed here, namely formulation of the inverse problem on function space, gives rise to algorithms which are robust under discretization even in the non-Gaussian case.

In Apte *et al.* (2007) and Apte, Jones and Stuart (2008*a*), studies of the ideal solution obtained from applying MCMC methods to the posterior (2.24) are compared with ensemble Kalman filter methods. The context is a Lagrangian data assimilation problem driven by a low-dimensional truncation of the linearized shallow water equations (3.27) and the results demonstrate pitfalls in the ensemble Kalman filter approach. An unambiguous and mathematically well-defined definition of the *ideal solution*, as given by (2.24), plays an important role in underpinning such computational studies.

A study of particle filters for Lagrangian data assimilation is undertaken in Spiller, Budhiraja, Ide and Jones (2008), and another application of filtering to oceanographic problems can be found in Brasseur *et al.* (2005). Recent contributions to the study of filtering in the context of the high-dimensional systems of interest in geophysical applications include Bergemann and Reich (2010), Cui *et al.* (2010), Chorin and Krause (2004), Chorin and Tu (2009, 2010), Majda and Grote (2007), Majda and Gershgorin (2008), Majda and Harlim (2010) and Van Leeuwen (2001, 2003). A comparison of various filtering methods, for the Kuramoto–Sivashinsky equation, may be found in Jardak, Navon and Zupanski (2010). In the paper of Pikkarainen (2006), filtering is studied in the case where the the state space for the dynamical variable is infinite-dimensional, and modelled by an SPDE. An attempt is made to keep track of the error made when approximating the infinite-dimensional system by a finite-dimensional one. In this regard, a useful approximation is introduced in Huttunen and Pikkarainen (2007), building on ideas in Kaipio and Somersalo (2007*a*). Parameter estimation in the context of filtering can be problematic, and smoothing should ideally be used when parameters are also to be estimated. However, there is some activity to try and make parameter estimation feasible in online scenarios; see Hurzeler and Künsch (2001) for a general discussion and Vossepoel and Van Leeuwen (2007) for an application.

We conclude this bibliography by highlighting an important question confronting many applied disciplines for which data assimilation is important. It is typically the case that models in fields such as climate prediction, oceanography, oil reservoir simulation and weather prediction are not fully resolved and various subgrid-scale models are used to compensate for this fact. This then raises the question: 'Should future increased computer resources be invested in further model resolution, or in more detailed study of uncertainty?' In the language of this section a stark version of this question is as follows: 'Should we employ only variational methods which identify probability maximizers, but do not quantify risk, investing future computer power in resolving the function space limit more fully? Or should we use MCMC methods, which quantify risk and uncertainty very precisely, but whose implementation is very costly and will preclude further model resolution?' This is a hard question. An excellent discussion in the context of climate models may be found in Palmer *et al.* (2009).

6. Probability

6.1. Overview

This section contains an overview of the probabilistic ideas used throughout the article. The presentation is necessarily terse and the reader is referred to the bibliography subsection at the end for references to material containing the complete details. Section 6.2 describes a number of basic definitions from the theory of probability that we will use throughout the article. In Section 6.3 we introduce Gaussian measures on Banach spaces and describe the central ideas of the Cameron–Martin space and the Fernique Theorem. Section 6.4 describes some explicit calculations concerning Gaussian measures on Hilbert space. In particular, we discuss the Karhunen–Loève expansion and conditioned Gaussian measures. The Karhunen–Loève expansion is a basic tool for constructing random draws from a Gaussian measure on Hilbert space, and for analysing the regularity properties of such random draws. Conditioned measures are key to the Bayesian approach to inverse problems and the Gaussian setting provides useful examples which help to build intuition. In Section 6.5 we introduce random fields and, in the Gaussian case, show how these may be viewed as Gaussian measures on vector fields. The key idea that we use from this subsection is to relate the properties of the covariance operator to sample function regularity. In Section 6.6 we describe Bayesian probability and a version of Bayes' theorem appropriate for function space. This will underpin the approach to inverse problems that we take in this article. We conclude, in Section 6.7, with a discussion of metrics on probability measures, and describe properties of the Hellinger metric in particular. This will enable us to measure

distance between pairs of probability measures, and is a key ingredient in the definition of well-posed posterior measures described in this article.

In this section, and indeed throughout the article, we will use the following notational conventions. The measure μ_0 will denote a prior measure, and π_0 its density with respect to Lebesgue measure when the state space is $\mathbb{R}^n$. Likewise the measure μ^y will denote a posterior measure, given data y, and π^y its density with respect to Lebesgue measure when the state space is $\mathbb{R}^n$; occasionally we will drop the y dependence and write μ and π. Given a density $\rho(u, y)$ on a pair of jointly distributed random variables, we will write $\rho(u|y)$ (resp. $\rho(y|u)$) for the density of the random variable u (resp. y), given a single observation of y (resp. u). We also write $\rho(u)$ for the marginal density found by integrating out y, and similarly $\rho(y)$ for the marginal density found by integrating out u. We will use similar conventions for other densities, and the densities arising from conditioning and marginalization.

6.2. Basic concepts

A measure (resp. probability) space is a triplet $(\Omega, \mathcal{F}, \mu)$, where Ω is the sample space, $\mathcal{F}$ the σ-algebra of events and μ the measure (resp. probability measure). In this article we will primarily be concerned with situations in which Ω is a separable Banach space $(X, \|\cdot\|_X)$ and $\mathcal{F}$ is the Borel σ-algebra $\mathcal{B}(X)$ generated by the open sets, in the strong topology. We are interested in *Radon measures* on X which are characterized by the property

$$\mu(A) = \sup\{\mu(B) \,\big|\, B \subset A,\ B \,\text{compact}\}, \quad A \in \mathcal{B}(X).$$

We use $\mathbb{E}$ and $\mathbb{P}$ to denote expectation and probability, respectively, and $\mathbb{E}(\cdot|\cdot)$ and $\mathbb{P}(\cdot|\cdot)$ for conditional expectation and probability; on occasion we will use the notation $\mathbb{E}^\mu$ or $\mathbb{P}^\mu$ if we wish to indicate that the expectation (or probability) in question is with respect to a particular measure μ. We use $\sim$ as shorthand for *is distributed as*; thus $x \sim \mu$ means that x is drawn from a probability measure μ. A real-valued measurable function on the measure space $(\Omega, \mathcal{F}, \mu)$ is one for which the pre-image of every Borel set in $\mathbb{R}$ is in $\mathcal{F}$ (is μ-measurable).

A function $m \in X$ is called *the mean* of μ on Banach space X if, for all $\ell \in X^*$, where X^* denotes the dual space of linear functionals on X,

$$\ell(m) = \int_X \ell(x)\mu(\mathrm{d}x).$$

If $m = 0$ the measure is called *centred*. In the Hilbert space setting we have that, for $x \sim \mu$, $m = \mathbb{E}x$. A linear operator $K : X^* \to X$ is called the covariance operator if, for all $k, \ell \in X^*$,

$$k(K\ell) = \int_X k(x - m)\ell(x - m)\mu(\mathrm{d}x).$$

In the Hilbert space setting where $X = X^*$, the covariance operator is characterized by the identity

$$\langle k, K\ell \rangle = \mathbb{E}\langle k, (x-m)\rangle\langle (x-m), \ell\rangle, \tag{6.1}$$

for $x \sim \mu$ and for all $k, \ell \in X$. Thus

$$K = \mathbb{E}(x-m) \otimes (x-m). \tag{6.2}$$

If μ and ν are two measures on the same measure space, then μ is *absolutely continuous* with respect to ν if $\nu(A) = 0$ implies $\mu(A) = 0$. This is sometimes written $\mu \ll \nu$. The two measures are *equivalent* if $\mu \ll \nu$ and $\nu \ll \mu$. If the measures are supported on disjoint sets then they are *mutually singular* or *singular*.

A family of measures $\mu^{(n)}$ on Banach space X is said to *converge weakly* to measure μ on X if

$$\int_X f(x)\mu^{(n)}(\mathrm{d}x) \to \int_X f(x)\mu(\mathrm{d}x)$$

for all continuous bounded $f : E \to \mathbb{R}$. We write $\mu^{(n)} \Rightarrow \mu$.[4]

The characteristic function of a probability distribution μ on a separable Banach space X is, for $\ell \in X^*$,

$$\varphi_\mu(\ell) = \mathbb{E}\exp(\mathrm{i}\ell(x)).$$

Theorem 6.1. If μ and ν are two Radon measures on a separable Banach space X and if $\varphi_\mu(\ell) = \varphi_\nu(\ell)$ for all $\ell \in X^*$, then $\mu = \nu$. ◇

The following *Radon–Nikodym Theorem* plays an important role in this article.

Theorem 6.2. (Radon–Nikodym Theorem) Let μ and ν be two measures on the same measure space $(\Omega, \mathcal{F})$. If $\mu \ll \nu$ and ν is σ-finite then there exists ν-measurable function $f : \Omega \to [0, \infty]$ such that, for all ν-measurable sets $A \in \mathcal{F}$,

$$\mu(A) = \int_A f(x)\,\mathrm{d}\nu(x).$$ ◇

The function f is known as the *Radon–Nikodym derivative* of μ with respect to ν. The derivative is written as

$$\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) = f(x). \tag{6.3}$$

We will sometimes simply refer to $f = \mathrm{d}\mu/\mathrm{d}\nu$ as the *density* of μ with

[4] This should not be confused with weak convergence of functions.

respect to ν. If μ is also a probability measure then

$$1 = \mu(\Omega) = \int_\Omega f(x)\,\mathrm{d}\nu(x).$$

Thus, if ν is a probability measure, $\mathbb{E}^\nu f(x) = 1$.

We give an example which illustrates a key idea underlying the material we develop in this section. We work in finite dimensions but highlight what can be transferred to probability measures on a Banach space.

Example 6.3. For a probability measure μ on $\mathbb{R}^d$ which is absolutely continuous with respect to Lebesgue measure λ, we use the shorthand p.d.f. for the probability density function, or *density*, ρ defined so that

$$\mu(A) = \int_A \rho(x)\,\mathrm{d}x \tag{6.4}$$

for $A \in \mathcal{F}$, where $\mathcal{F}$ is the sigma algebra generated by the open sets in $\mathbb{R}^d$. Strictly speaking this is the p.d.f. *with respect to Lebesgue measure*, as we integrate the density against Lebesgue measure to find the probability of a set A. Note that

$$\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x) = \rho(x).$$

It is also possible to find the density of μ with respect to a Gaussian measure. To illustrate this, let $\mu_0 = \mathcal{N}(0, I)$ denote a standard unit Gaussian in $\mathbb{R}^d$. Then

$$\mu_0(\mathrm{d}x) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}|x|^2\right)\mathrm{d}x.$$

Thus the density of μ with respect to μ_0 is

$$\rho_{\mathrm{g}}(x) = (2\pi)^{d/2} \exp\left(\frac{1}{2}|x|^2\right)\rho(x).$$

We then have the identities

$$\mu(A) = \int_A \rho_{\mathrm{g}}(x)\mu_0(\mathrm{d}x) \tag{6.5}$$

and

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(x) = \rho_{\mathrm{g}}(x).$$

It turns out that, in the infinite-dimensional setting, the formulation (6.5) generalizes much more readily than does (6.4). This is because infinite-dimensional Gaussian measure is well-defined, and because many measures have a density (Radon–Nikodym derivative) with respect to an infinite-dimensional Gaussian measure. In contrast, infinite-dimensional Lebesgue measure does not exist. ◇

We conclude this subsection with two definitions of operators, both important for definitions associated with Gaussian measures on Hilbert space. Let $\{\phi_k\}_{k=1}^{\infty}$ denote an orthonormal basis for a separable Hilbert space $\mathcal{H}$. A linear operator $A : \mathcal{H} \to \mathcal{H}$ is *trace-class* or *nuclear* if

$$\mathrm{Tr}(A) := \sum_{k=1}^{\infty} \langle A\phi_k, \phi_k \rangle < \infty. \tag{6.6}$$

The sum is independent of the choice of basis. The operator A is *Hilbert–Schmidt* if

$$\sum_{k=1}^{\infty} \|A\phi_k\|^2 < \infty. \tag{6.7}$$

If A is self-adjoint and we choose the $\{\phi_k\}$ to be the eigenfunctions of A, then the sum in (6.6) is simply the sum of the eigenvalues of A. A weaker condition is that the eigenvalues are square-summable, which is (6.7).

6.3. Gaussian measures

We will primarily employ Gaussian measures in the Hilbert space setting. However, they can also be defined on Banach spaces and, on occasion, we will employ this level of generality. Indeed, when studying Gaussian random fields in Section 6.5, we will show that, for a Gaussian measure μ on a Hilbert space $\mathcal{H}$, there is often a Banach space X which is continuously embedded in $\mathcal{H}$ and has the property that $\mu(X) = 1$. We would then like to define the measure μ on the Banach space X. We thus develop Gaussian measure theory on separable Banach spaces here.

Having defined Gaussian measure, we describe its characteristic function and we state the Fernique Theorem, which exploits tail properties of Gaussian measure. We follow this with definition and discussion of the Cameron–Martin space. We then describe the basic tools required to study the absolute continuity of two Gaussian measures.

A measure μ on $(X, \mathcal{B}(X))$ is *Gaussian* if, for any $\ell \in X^*$, $\ell(x) \sim \mathcal{N}(m_\ell, \sigma_\ell^2)$ for some $m_\ell \in \mathbb{R}, \sigma_\ell \in \mathbb{R}$. Note that $\sigma_\ell = 0$ is allowed, so that the induced measure on $\ell(x)$ may be a Dirac mass at m_ℓ. Note also that it is expected that $m_\ell = \ell(m)$, where m is the mean defined above, and $\sigma_\ell^2 = \ell(K\ell)$, where K is the covariance operator. The mean m and covariance operator K are indeed well-defined by this definition of covariance operator.

Theorem 6.4. A Gaussian measure on $(X, \mathcal{B}(X))$ has a mean m and covariance operator K. Further, the characteristic function of the measure is

$$\varphi(\ell) = \exp\left(\mathrm{i}\ell(m) - \frac{1}{2}\ell(K\ell) \right). \qquad \diamond$$

Hence, by Theorem 6.1 we see that the mean and covariance completely

characterize the Gaussian measure, and so we are justified in denoting it by $\mathcal{N}(m, K)$. The following lemma demonstrates an important role for characteristic functions in studying weak convergence.

Lemma 6.5. Consider a family of probability measures $\mu^{(n)}$. Assume that, for all $\ell \in X^*$,

$$\varphi_{\mu^{(n)}}(\ell) \to \exp\left(\mathrm{i}\ell(m^+) - \frac{1}{2}\ell(K^+\ell)\right).$$

Then $\mu^{(n)} \Rightarrow \mathcal{N}(m^+, K^+)$. ◇

In the Hilbert space setting we refer to the inverse of the covariance operator $\mathcal{C}$ as the *precision operator* and denote it by $\mathcal{L}$. It is natural to ask what conditions an operator must satisfy in order to be a covariance operator. Good intuition can be obtained by thinking of the precision operator as a (possibly) fractional differential operator of sufficiently high order. To pursue this issue a little further we confine ourselves to the Hilbert space setting. The following theorem provides a precise answer to the question concerning properties of the covariance operator.

Theorem 6.6. If $\mathcal{N}(0, \mathcal{C})$ is a Gaussian measure on a Hilbert space $\mathcal{H}$, then $\mathcal{C}$ is a self-adjoint, positive semi-definite trace-class operator on $\mathcal{H}$. Furthermore, for any integer p, there is a constant $C = C_p \geqslant 0$ such that, for $x \sim \mathcal{N}(0, \mathcal{C})$,

$$\mathbb{E}\|x\|^{2p} \leqslant C_p(\mathrm{Tr}(\mathcal{C}))^p.$$

Conversely, if $m \in \mathcal{H}$, and $\mathcal{C}$ is a self-adjoint, positive semi-definite, trace-class linear operator on a Hilbert space $\mathcal{H}$, then there is a Gaussian measure $\mu = \mathcal{N}(m, \mathcal{C})$ on $\mathcal{H}$. ◇

Example 6.7. Unit Brownian bridge on $J = (0, 1)$ may be viewed as a Gaussian process on $L^2(J)$ with precision operator $\mathcal{L} = -\mathrm{d}^2/\mathrm{d}x^2$ and $D(\mathcal{L}) = H^2(J) \bigcap H_0^1(J)$. Thus the eigenvalues of $\mathcal{C}$ are $\gamma_k = (k^2\pi^2)^{-1}$ and are summable. ◇

If $x \sim \mathcal{N}(0, \mathcal{C})$, then $\mathbb{E}\|x\|^2 = \mathrm{Tr}(\mathcal{C})$. Combining this fact with the previous theorem we have the following generalization of the well-known property concerning the moments of finite-dimensional Gaussian measures.

Corollary 6.8. If $\mathcal{N}(0, \mathcal{C})$ is a Gaussian measure on a Hilbert space $\mathcal{H}$ then, for any positive integer p, there exists $C_p \geqslant 0$ such that $\mathbb{E}\|x\|^{2p} \leqslant C_p(\mathbb{E}\|x\|^2)^p$. ◇

In fact, as in finite dimensions, the exponentials of certain quadratic functionals are bounded for Gaussian measures. This is the Fernique Theorem, which we state in the Banach space context.

Theorem 6.9. (Fernique Theorem) If $\mu = \mathcal{N}(0, K)$ is a Gaussian measure on Banach space X, so that $\mu(X) = 1$, then there exists $\alpha > 0$ such that

$$\int_X \exp\bigl(\alpha \|x\|_X^2\bigr)\mu(\mathrm{d}x) < \infty. \qquad \diamond$$

We define the *Cameron–Martin space* E associated with a Gaussian measure $\mu = \mathcal{N}(0, K)$ on Banach space X to be the intersection of all linear spaces of full measure under μ.[5]

Lemma 6.10. Let E be the Cameron–Martin space of Gaussian measure $\mu = \mathcal{N}(0, K)$ on Banach space X. In infinite dimensions it is necessarily the case that $\mu(E) = 0$. Furthermore, E can be endowed with a Hilbert space structure. Indeed, for Gaussian measures $\mathcal{N}(0, \mathcal{C})$ on the Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$, the Cameron–Martin space is the Hilbert space $E := \mathrm{Im}(\mathcal{C}^{1/2})$ with inner product

$$\langle\cdot,\cdot\rangle_{\mathcal{C}} = \langle \mathcal{C}^{-1/2}\cdot, \mathcal{C}^{-1/2}\cdot\rangle. \qquad \diamond$$

Note that the covariance operator $\mathcal{C}$ of a Gaussian probability measure on a Hilbert space $\mathcal{H}$ is necessarily compact because $\mathcal{C}$ is trace-class, so that the eigenvalues of $\mathcal{C}^{1/2}$ decay at least algebraically. Thus the Cameron–Martin space $\mathrm{Im}(\mathcal{C}^{1/2})$ is compactly embedded in $\mathcal{H}$. In fact we have the following more general result.

Theorem 6.11. The Cameron–Martin space E associated with a Gaussian measure $\mu = \mathcal{N}(0, K)$ on Banach space X is compactly embedded in all separable Banach spaces X' with full measure ($\mu(X') = 1$) under μ. $\diamond$

Example 6.12. Consider a probability measure ν on $\mathbb{R}^2$ which is a product measure of the form $\delta_0 \otimes \mathcal{N}(0, 1)$. Introduce coordinates (x_1, x_2) so that the marginal on x_1 is δ_0 and the marginal on x_2 is $\mathcal{N}(0, 1)$. The intersection of all linear spaces with full measure is the subset of $\mathbb{R}^2$ defined by the line

$$E = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = 0\}.$$

Note, furthermore, that this subset is characterized by the property that the measures $\nu(\cdot)$ and $\nu(a + \cdot)$ are equivalent (as measures) if and only if $a \in E$. Thus, for this example, the Cameron–Martin space defines the space of allowable shifts, under which equivalence of the measures holds. $\diamond$

We now generalize the last observation in the preceding example: we show that the Cameron–Martin space characterizes precisely those shifts in the mean of a Gaussian measure which preserve equivalence.

[5] In most developments of the subject this characterization is given after a more abstract definition of the Cameron–Martin space. However, for our purposes this level of abstraction is not needed.

Theorem 6.13. Two Gaussian measures $\mu_i = \mathcal{N}(m_i, \mathcal{C}_i)$, $i = 1, 2$, on a Hilbert space $\mathcal{H}$ are either singular or equivalent. They are equivalent if and only if the following three conditions hold:

(i) $\mathrm{Im}(\mathcal{C}_1^{1/2}) = \mathrm{Im}(\mathcal{C}_2^{1/2}) := E$,
(ii) $m_1 - m_2 \in E$,
(iii) the operator $T := \big(\mathcal{C}_1^{-1/2}\mathcal{C}_2^{1/2}\big)\big(\mathcal{C}_1^{-1/2}\mathcal{C}_2^{1/2}\big)^* - I$ is Hilbert–Schmidt in $\overline{E}$. ◇

In particular, choosing $\mathcal{C}_1 = \mathcal{C}_2$ we see that shifts in the mean give rise to equivalent Gaussian measures if and only if the shifts lie in the Cameron–Martin space E. It is of interest to characterize the Radon–Nikodym derivative arising from such shifts in the mean.

Theorem 6.14. Consider two measures $\mu_i = \mathcal{N}(m_i, \mathcal{C})$, $i = 1, 2$, on Hilbert space $\mathcal{H}$, where $\mathcal{C}$ has eigenbasis $\{\phi_k, \lambda_k\}_{k=1}^{\infty}$. Denote the Cameron–Martin space by E. If $m_1 - m_2 \in E$, then the Radon–Nikodym derivative is given by

$$\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}(x) = \exp\left(\langle m_1 - m_2, x - m_2\rangle_{\mathcal{C}} - \frac{1}{2}\|m_1 - m_2\|_{\mathcal{C}}^2\right). \qquad \diamond$$

Since $m_1 - m_2 \in \mathrm{Im}(\mathcal{C}^{1/2})$, the quadratic form $\|m_1 - m_2\|_{\mathcal{C}}^2$ is defined; the random variable $x \to \langle m_1 - m_2, x - m_2\rangle_{\mathcal{C}}$ is defined via a limiting procedure as follows. By the the Karhunen–Loève expansion (6.9) below, we have the representation of $x \sim \mathcal{N}(0, \mathcal{C})$ as

$$x - m_2 = \sum_{k=1}^{\infty} \sqrt{\lambda_k}\omega_k\phi_k,$$

where $\omega = \{\omega_k\}_{k=1}^{\infty} \in \Omega$ is an i.i.d. sequence of $\mathcal{N}(0, 1)$ random variables. Then $\langle m_1 - m_2, x - m_2\rangle_{\mathcal{C}}$ is defined as the $L^2(\Omega; \mathcal{H})$ limit in n of the series

$$\sum_{k=1}^{n} \frac{1}{\sqrt{\lambda_k}}\langle m_1 - m_2, \phi_k\rangle\omega_k.$$

In establishing the first of the conditions in Theorem 6.13, the following lemma is often useful.

Lemma 6.15. For any two positive definite, self-adjoint, bounded linear operators $\mathcal{C}_i$ on a Hilbert space $\mathcal{H}$, $i = 1, 2$, the condition $\mathrm{Im}(\mathcal{C}_1^{1/2}) \subset \mathrm{Im}(\mathcal{C}_2^{1/2})$ holds if and only if there exists a constant $K > 0$ such that

$$\langle h, \mathcal{C}_1 h\rangle \leqslant K\langle h, \mathcal{C}_2 h\rangle, \quad \forall h \in \mathcal{H}. \qquad \diamond$$

Example 6.16. Consider two Gaussian measures μ_i on $\mathcal{H} = L^2(J)$, $J = (0, 1)$ both with precision operator $\mathcal{L} = -\mathrm{d}^2/\mathrm{d}x^2$ and the domain of $\mathcal{L}$ being $H_0^1(J) \cap H^2(J)$. (Informally $-\mathcal{L}$ is the Laplacian on J with homogeneous

Dirichlet boundary conditions.) The mean of μ_1 is a function $m \in \mathcal{H}$ and the mean of μ_2 is 0. Thus $\mu_1 \sim \mathcal{N}(m, \mathcal{C})$ and $\mu_2 \sim \mathcal{N}(0, \mathcal{C})$, where $\mathcal{C} = \mathcal{L}^{-1}$. Here $\mathcal{C}_1 = \mathcal{C}_2 = \mathcal{C}$ and $T = 0$, so that (i) and (iii) in Theorem 6.13 are satisfied with $E = \mathrm{Im}\,(\mathcal{C}^{1/2}) = H_0^1(J)$. It follows that the measures are equivalent if and only if $m \in E$. If this condition is satisfied then, from Theorem 6.14, the Radon–Nikodym derivative between the two measures is given by

$$\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}(x) = \exp\left(\langle m, x \rangle_{H_0^1} - \frac{1}{2}\|m\|_{H_0^1}^2 \right). \qquad \diamond$$

Example 6.17. Consider two mean zero Gaussian measures μ_i on $\mathcal{H} = L^2(J)$, $J = (0,1)$ with norm $\|\cdot\|$ and precision operators $\mathcal{L}_1 = -\mathrm{d}^2/\mathrm{d}x^2 + I$ and $\mathcal{L}_2 = -\mathrm{d}^2/\mathrm{d}x^2$ respectively, both with domain $H_0^1(J) \cap H^2(J)$.

The operators $\mathcal{L}_1, \mathcal{L}_2$ share the same eigenfunctions,

$$\phi_k(x) = \sqrt{2}\sin(k\pi x),$$

and have eigenvalues

$$\lambda_k(1) = \lambda_k(2) + 1, \quad \lambda_k(2) = k^2\pi^2,$$

respectively. Thus $\mu_1 \sim \mathcal{N}(0, \mathcal{C}_1)$ and $\mu_2 \sim \mathcal{N}(0, \mathcal{C}_2)$ where, in the basis of eigenfunctions, $\mathcal{C}_1$ and $\mathcal{C}_2$ are diagonal with eigenvalues

$$\frac{1}{k^2\pi^2 + 1}, \quad \frac{1}{k^2\pi^2},$$

respectively. We have, for $h_k = \langle h, \phi_k \rangle$,

$$\frac{\pi^2}{\pi^2 + 1} \leqslant \frac{\langle h, \mathcal{C}_1 h \rangle}{\langle h, \mathcal{C}_2 h \rangle} = \frac{\sum_{k \in \mathbb{Z}^+} (1 + k^2\pi^2)^{-1} h_k^2}{\sum_{k \in \mathbb{Z}^+} (k\pi)^{-2} h_k^2} \leqslant 1.$$

Thus, by Lemma 6.15, Theorem 6.13(i) is satisfied. Part (ii) holds trivially. Notice that

$$T = \mathcal{C}_1^{-1/2} \mathcal{C}_2 \mathcal{C}_1^{-1/2} - I$$

is diagonalized in the same basis as the $\mathcal{C}_i$, and has eigenvalues

$$\frac{1}{k^2\pi^2}.$$

These are square-summable, and so part (iii) of Theorem 6.13 holds and the two measures are absolutely continuous with respect to one another. $\diamond$

A Hilbert space $(X, \langle \cdot, \cdot \rangle_{\mathrm{X}})$ of functions $f : D \subset \mathbb{R}^d \to \mathbb{R}$ is called a *reproducing kernel Hilbert space*, RKHS for short, if pointwise evaluation is a continuous linear functional in the Hilbert space. If $f(y) = \langle f, r_{\mathrm{y}} \rangle_{\mathrm{X}}$, then r_{y} is called the *representer* of the RKHS.

Example 6.18. Let $J = (0,1)$. Note that $\mathcal{H} = L^2(J;\mathbb{R})$ is not an RKHS. Consider $X = H^1(J;\mathbb{R})$ equipped with the inner product

$$(a,b) = a(0)b(0) + \int_0^1 a'(x)b'(x)\,\mathrm{d}x. \tag{6.8}$$

If $r_{\mathrm{y}}(x) = 1 + x \wedge y$ then $f(y) = (f, r_{\mathrm{y}})$. Notice that $r_{\mathrm{y}} \in X$. Thus, by the Cauchy–Schwarz inequality,

$$\begin{aligned} |f(y) - g(y)| &\leqslant |(f-g, r_{\mathrm{y}})| \\ &\leqslant \|f-g\|_X \|r_{\mathrm{y}}\|_X, \end{aligned}$$

demonstrating that pointwise evaluation is a continuous linear functional on X. Notice, furthermore, that the expression $f(y) = (f, r_{\mathrm{y}})$ is an explicit statement of the Riesz Representation Theorem. ◇

In the literature there is often an overlap of terminology surrounding the RKHS and the Cameron–Martin space. This is related to the fact that the representer of an RKHS can often be viewed as the covariance function (see Section 6.5 below) of a covariance operator associated to a Gaussian measure on $L^2(D;\mathbb{R})$.

6.4. Explicit calculations with Gaussian measures

In this section we confine our attention to Gaussian measures on Hilbert space. We provide a number of explicit formulae that are helpful throughout the article, and which also help to build intuition about measures on infinite-dimensional spaces.

We can construct random draws from a Gaussian measure on Hilbert space $\mathcal{H}$ as follows, using the *Karhunen–Loève expansion*.

Theorem 6.19. Let $\mathcal{C}$ be a self-adjoint, positive semi-definite, nuclear operator in a Hilbert space $\mathcal{H}$ and let $m \in \mathcal{H}$. Let $\{\phi_k, \gamma_k\}_{k=1}^{\infty}$ be an orthonormal set of eigenvectors/eigenvalues for $\mathcal{C}$ ordered so that

$$\gamma_1 \geqslant \gamma_2 \geqslant \cdots.$$

Take $\{\xi_k\}_{k=1}^{\infty}$ to be an i.i.d. sequence with $\xi_1 \sim \mathcal{N}(0,1)$. Then the random variable $x \in \mathcal{H}$ given by the *Karhunen–Loève expansion*

$$x = m + \sum_{k=1}^{\infty} \sqrt{\gamma}_k \xi_k \phi_k \tag{6.9}$$

is distributed according to $\mu = \mathcal{N}(m, \mathcal{C})$. ◇

In applications the eigenvalues and eigenvectors of $\mathcal{C}$ will often be indexed over a different countable set, say $\mathbb{K}$. In this context certain calculations are

cleaner if we write the Karhunen–Loève expansion (6.9) in the form

$$x = m + \sum_{k \in \mathbb{K}} \sqrt{\gamma_k} \xi_k \phi_k. \tag{6.10}$$

Here the $\{\xi_k\}_{k \in \mathbb{K}}$ are an i.i.d. set of random variables all distributed as $\mathcal{N}(0,1)$. Of course, the order of summation does, in general, matter; whenever we use (6.10), however, the ordering will not be material to the outcome and will streamline the calculations to use (6.10).

The next theorem concerns conditioning of Gaussian measures.

Theorem 6.20. Let $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ be a separable Hilbert space with projectors $\Pi_i \colon \mathcal{H} \to \mathcal{H}_i$. Let $(x_1, x_2) \in \mathcal{H}_1 \oplus \mathcal{H}_2$ be an $\mathcal{H}$-valued Gaussian random variable with mean $m = (m_1, m_2)$ and positive definite covariance operator $\mathcal{C}$. Define

$$\mathcal{C}_{ij} = \mathbb{E}(x_i - m_i) \otimes (x_j - m_j).$$

Then the conditional distribution of x_1 given x_2 is Gaussian with mean

$$m' = m_1 + \mathcal{C}_{12}\mathcal{C}_{22}^{-1}(x_2 - m_2) \tag{6.11}$$

and covariance operator

$$\mathcal{C}' = \mathcal{C}_{11} - \mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{21}. \tag{6.12}$$

◇

To understand this theorem it is useful to consider the following finite-dimensional result concerning block matrix inversion.

Lemma 6.21. Consider a positive definite matrix C with the block form

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{12}^* & C_{22} \end{pmatrix}.$$

Then C_{22} is positive definite symmetric and the *Schur complement* S defined by $S = C_{11} - C_{12}C_{22}^{-1}C_{12}^*$ is positive definite symmetric. Furthermore,

$$C^{-1} = \begin{pmatrix} S^{-1} & -S^{-1}C_{12}C_{22}^{-1} \\ -C_{22}^{-1}C_{12}^*S^{-1} & C_{22}^{-1} + C_{22}^{-1}C_{12}^*S^{-1}C_{12}C_{22}^{-1} \end{pmatrix}.$$

Now let (x, y) be jointly Gaussian with distribution $\mathcal{N}(m, C)$ and $m = (m_1^*, m_2^*)^*$. Then the conditional distribution of x given y is Gaussian with mean m' and covariance matrix C' given by

$$m' = m_1 + C_{12}C_{22}^{-1}(y - m_2),$$

$$C' = C_{11} - C_{12}C_{22}^{-1}C_{12}^*. \qquad \diamond$$

Example 6.22. Consider a random variable u with Gaussian prior probability distribution $\mathcal{N}(0,1)$, and hence associated p.d.f.

$$\pi_0(u) \propto \exp\left(-\frac{1}{2}u^2\right).$$

Let y be the random variable $y = u + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2)$ is independent of u. Then the likelihood of y given u has p.d.f. proportional to

$$\exp\left(-\frac{1}{2\sigma^2}|y-u|^2\right).$$

The joint probability of (u, y) thus has p.d.f. proportional to

$$\exp\left(-\frac{1}{2\sigma^2}|y-u|^2 - \frac{1}{2}|u|^2\right).$$

Since

$$\frac{1}{2\sigma^2}|y-u|^2 + \frac{1}{2}|u|^2 = \left(\frac{\sigma^2+1}{2\sigma^2}\right)\left|u - \frac{1}{\sigma^2+1}y\right|^2 + c_y,$$

where c_y is independent of u, we see that the $u|y$ is a Gaussian $\mathcal{N}(m, \gamma^2)$ with

$$m = \frac{1}{\sigma^2+1}y, \quad \gamma^2 = \frac{\sigma^2}{\sigma^2+1}.$$

This technique for deriving the mean and covariance of a Gaussian measure is often termed *completing the square*; it may be rigorously justified by Theorem 6.20 as follows. First we observe that $m_1 = m_2 = 0$, that $\mathcal{C}_{11} = \mathcal{C}_{12} = \mathcal{C}_{21} = 1$ and that $\mathcal{C}_{22} = 1 + \sigma^2$. The formulae (6.11) and (6.12) then give identical results to those found by completing the square. ◇

We now study an infinite-dimensional version of the previous example.

Example 6.23. Consider a random variable u on a Hilbert space $\mathcal{H}$ distributed according to a measure $\mu_0 \sim \mathcal{N}(m_0, \mathcal{C}_0)$. We assume that $m_0 \in \mathrm{Im}(\mathcal{C}_0^{1/2})$. Assume that $y \in \mathbb{R}^m$ is also Gaussian and is given by

$$y = Au + \eta,$$

where $A : X \to \mathbb{R}^m$ is linear and continuous on a Banach space $X \subseteq \mathcal{H}$ with $\mu_0(X) = 1$. The adjoint of A, denoted A^*, is hence the operator from $\mathbb{R}^m \to X^*$ defined by the identity

$$\langle Au, v\rangle = (A^*v)(u),$$

which holds for all $v \in \mathbb{R}^m, u \in X$, and where $A^*v \in X^*$ is a linear functional on X. We also assume that $\eta \sim \mathcal{N}(0, \Gamma)$ is independent of u and that Γ is positive definite. Thus $y|u$ is Gaussian with density proportional to $\exp\left(-\frac{1}{2}|y - Au|_\Gamma^2\right)$. We would like to characterize the Gaussian measure μ^y

for $u|y$. Let $\mu^y = \mathcal{N}(m, \mathcal{C})$. To calculate $\mathcal{C}$ and m we first use the idea of completing the square, simply computing formally as if the Hilbert space for u were finite-dimensional and had a density with respect to Lebesgue measure; we will then justify the resulting formulae after the fact by means of Theorem 6.20. The formal Lebesgue density for $u|y$ is proportional to

$$\exp\left(-\frac{1}{2}|y - Au|_\Gamma^2 - \frac{1}{2}\|u - m_0\|_{\mathcal{C}_0}^2\right).$$

But

$$\frac{1}{2}|y - Au|_\Gamma^2 + \frac{1}{2}\|u - m_0\|_{\mathcal{C}_0}^2 = \frac{1}{2}\|u - m\|_{\mathcal{C}}^2 + c_y$$

with c_y independent of u, and hence completing the square gives

$$\mathcal{C}^{-1} = A^*\Gamma^{-1}A + \mathcal{C}_0^{-1}, \tag{6.13a}$$

$$m = \mathcal{C}(A^*\Gamma^{-1}y + \mathcal{C}_0^{-1}m_0). \tag{6.13b}$$

We now justify this informal calculation.

The pair (u, y) is jointly Gaussian with $\mathbb{E}u = m_0$ and $\mathbb{E}y = Am_0$. We define $\overline{u} = u - m_0$ and $\overline{y} = y - Am_0$. Note that $\overline{y} = A\overline{u} + \eta$. The pair (u, y) has covariance operator with components

$$\begin{aligned}
\mathcal{C}_{11} &= \mathbb{E}\overline{u}\overline{u}^* = \mathcal{C}_0,\\
\mathcal{C}_{22} &= \mathbb{E}\overline{y}\overline{y}^* = A\mathcal{C}_0A^* + \Gamma,\\
\mathcal{C}_{21} &= \mathbb{E}\overline{y}\overline{u}^* = A\mathcal{C}_0.
\end{aligned}$$

Thus, by Theorem 6.20, we deduce that the mean m and covariance operator $\mathcal{C}$ for u conditional on y are given, respectively, by

$$m = m_0 + \mathcal{C}_0A^*(\Gamma + A\mathcal{C}_0A^*)^{-1}(y - Am_0) \tag{6.14}$$

and

$$\mathcal{C} = \mathcal{C}_0 - \mathcal{C}_0A^*(\Gamma + A\mathcal{C}_0A^*)^{-1}A\mathcal{C}_0. \tag{6.15}$$

We now demonstrate that the formulae (6.14) and (6.15) agree with (6.13). To check agreement with the formula for the inverse of $\mathcal{C}$ found by completing the square, we show that the product is indeed the identity. Note that

$$\begin{aligned}
&\big(\mathcal{C}_0 - \mathcal{C}_0A^*(\Gamma + A\mathcal{C}_0A^*)^{-1}A\mathcal{C}_0\big)(\mathcal{C}_0^{-1} + A^*\Gamma^{-1}A)\\
&\quad= \big(I - \mathcal{C}_0A^*(\Gamma + A\mathcal{C}_0A^*)^{-1}A\big)(I + \mathcal{C}_0A^*\Gamma^{-1}A)\\
&\quad= I + \mathcal{C}_0A^*\Gamma^{-1}A - \mathcal{C}_0A^*(\Gamma + A\mathcal{C}_0A^*)^{-1}(A + A\mathcal{C}_0A^*\Gamma^{-1}A)\\
&\quad= I + \mathcal{C}_0A^*\Gamma^{-1}A - \mathcal{C}_0A^*\Gamma^{-1}A\\
&\quad= I.
\end{aligned}$$

To check agreement with the two formulae for the mean, we proceed as follows. We have

$$\Gamma^{-1} - (\Gamma + A\mathcal{C}_0A^*)^{-1}A\mathcal{C}_0A^*\Gamma^{-1} = (\Gamma + A\mathcal{C}_0A^*)^{-1}. \tag{6.16}$$

The formula for the mean derived by completing the square gives

$$\begin{aligned} m &= \mathcal{C}\big((\mathcal{C}^{-1} - A^*\Gamma^{-1}A)m_0 + A^*\Gamma^{-1}y\big) \\ &= m_0 + \mathcal{C}A^*\Gamma^{-1}(y - Am_0). \end{aligned}$$

To get agreement with the formula (6.14) it suffices to show that

$$\mathcal{C}A^*\Gamma^{-1} = \mathcal{C}_0A^*(\Gamma + A\mathcal{C}_0A^*)^{-1}.$$

By (6.15) and (6.16),

$$\begin{aligned} \mathcal{C}A^*\Gamma^{-1} &= \mathcal{C}_0A^*\Gamma^{-1} - \mathcal{C}_0A^*(\Gamma + A\mathcal{C}_0A^*)^{-1}A\mathcal{C}_0A^*\Gamma^{-1} \\ &= \mathcal{C}_0A^*(\Gamma + A\mathcal{C}_0A^*)^{-1}, \end{aligned}$$

and we are done. ◇

6.5. Gaussian random fields

Our aim in this subsection is to construct, and study the properties of, Gaussian random functions. We first consider the basic construction of random functions, then Gaussian random functions, following this by a study of the regularity properties of Gaussian random functions.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $D \subseteq \mathbb{R}^d$ an open set. A *random field* on D is a measurable mapping $u : D \times \Omega \to \mathbb{R}^n$. Thus, for any $x \in D$, $u(x;\cdot)$ is an $\mathbb{R}^n$-valued random variable; on the other hand, for any $\omega \in \Omega$, $u(\cdot;\omega) : D \to \mathbb{R}^n$ is a vector field. In the construction of random fields it is commonplace to first construct the *finite-dimensional distributions*. These are found by choosing any integer $q \geqslant 1$, and any set of points $\{x_k\}_{k=1}^q$ in D, and then considering the random vector $(u(x_1;\cdot)^*, \ldots, u(x_q;\cdot)^*)^* \in \mathbb{R}^{nq}$. From the finite-dimensional distributions of this collection of random vectors we would like to be able to make sense of the probability measure μ on X, a Banach space, via the formula

$$\mu(A) = \mathbb{P}(u(\cdot;\omega) \in A), \quad A \in \mathcal{B}(X), \tag{6.17}$$

where ω is taken from a common probability space on which the random element $u \in X$ is defined. It is thus necessary to study the joint distribution of a set of q $\mathbb{R}^n$-valued random variables, all on a common probability space. Such $\mathbb{R}^{nq}$-valued random variables are, of course, only defined up to a set of zero measure. It is desirable that all such finite-dimensional distributions are defined on a common subset $\Omega_0 \subset \Omega$ with full measure, so that u may be viewed as a function $u : D \times \Omega_0 \to \mathbb{R}^n$; such a choice of random field is termed a *modification*. In future developments, statements about almost sure (regularity) properties of a random field should be interpreted as statements concerning the existence of a modification possessing the stated almost sure regularity property. We will often simply write $u(x)$, suppressing the explicit dependence on the probability space.

A *Gaussian random field* is one where, for any integer $q \geqslant 1$, and any set of points $\{x_k\}_{k=1}^q$ in D, the random vector $(u(x_1;\cdot)^*, \ldots, u(x_q;\cdot)^*)^* \in \mathbb{R}^{nq}$ is a Gaussian random vector. The *mean function* of a Gaussian random field is $m(x) = \mathbb{E}u(x)$. The *covariance function* is $c(x,y) = \mathbb{E}(u(x) - m(x))(u(y) - m(y))^*$. For Gaussian random fields this function, together with the mean function, completely specify the joint probability distribution for $(u(x_1;\cdot)^*, \ldots, u(x_q)^*)^* \in \mathbb{R}^{nq}$. Furthermore, if we view the Gaussian random field as a Gaussian measure on $L^2(D;\mathbb{R}^n)$, then the covariance operator can be constructed from the covariance function as follows. Without loss of generality we consider the mean zero case; the more general case follows by shift of origin. Since the field has mean zero we have, from (6.1),

$$\begin{aligned}
\langle h_1, \mathcal{C}h_2\rangle &= \mathbb{E}\langle h_1, u\rangle\langle u, h_2\rangle \\
&= \mathbb{E}\int_D \int_D h_1(x)^*(u(x)u(y)^*)h_2(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \mathbb{E}\int_D h_1(x)^*\left(\int_D (u(x)u(y)^*)h_2(y)\,\mathrm{d}y\right)\mathrm{d}x \\
&= \int_D h_1(x)^*\left(\int_D c(x,y)h_2(y)\,\mathrm{d}y\right)\mathrm{d}x
\end{aligned}$$

and we deduce that

$$(\mathcal{C}\phi)(x) = \int_D c(x,y)\phi(y)\,\mathrm{d}y. \tag{6.18}$$

Thus the covariance operator of a Gaussian random field is an integral operator with kernel given by the covariance function.

If we view the Gaussian random field as a measure on the space $X = C(\overline{D};\mathbb{R}^n)$, then the covariance operator $K : X^* \to X$ may also be written as an integral operator as follows. For simplicity we consider the case $n = 1$. We note that $\ell = \ell_\mu \in X^*$ may be identified with a signed measure μ on D. Then similar arguments to those used in the Hilbert space case show that

$$(K\ell_\mu)(x) = \int_D c(x,y)\mu(\mathrm{d}y). \tag{6.19}$$

This may be extended to the case of random fields taking values in $\mathbb{R}^n$.

A mean zero Gaussian random field is termed *stationary* if $c(x,y) = s(x-y)$ for some matrix-valued function s, so that shifting the field by a fixed random vector does not change the statistics. It is *isotropic* if, in addition, $s(x-y) = \iota(|x-y|)$, for some matrix-valued function ι.

An important general question concerning random fields is to find criteria to establish their regularity, expressed in terms of the covariance function or operator. An important tool in this context is the *Kolmogorov Continuity Theorem*, which follows below. This theorem expresses sample function regularity in terms of the covariance function of the random field. Another

key tool in establishing regularity is the Karhunen–Loève expansion (6.10), which expresses a random draw from a Gaussian measure in terms of the eigenfunctions and eigenvalues of the covariance operator and may be used to express sample function regularity in terms of the decay of the eigenvalues of the covariance operator. Both these approaches to sample function regularity, one working from the covariance functions, and one from eigenvalues of the covariance operator, are useful in practice when considering Bayesian inverse problems for functions; this is because prior Gaussian measures may be specified via either the covariance function or the covariance operator.

Theorem 6.24. (Kolmogorov Continuity Theorem) Consider an $\mathbb{R}^n$-valued random field u on a bounded open set $D \subset \mathbb{R}^d$. Assume that there are constants $K, \varepsilon > 0$ and $\delta \geqslant 1$ such that

$$\mathbb{E}|u(x) - u(y)|^\delta \leqslant K|x - y|^{2d+\varepsilon}.$$

Then u is almost surely Hölder-continuous on D with any exponent smaller than $\min\{1, \varepsilon/\delta\}$. ◇

In this article we mainly work with priors specified through the covariance operator on a simple geometry, as this makes the exposition more straightforward. Specifically, we consider covariance operators constructed as fractional powers of operators $\mathcal{A}$ whose leading-order behaviour is like that of the Laplacian on a rectangle. Precisely, we will assume that Assumptions 2.9 hold.

By using the Kolmogorov Continuity Theorem we can now prove the following.

Lemma 6.25. Let $\mathcal{A}$ satisfy Assumptions 2.9(i)–(iv). Consider a Gaussian measure $\mu = \mathcal{N}(0, \mathcal{C})$ with $\mathcal{C} = \mathcal{A}^{-\alpha}$ with $\alpha > d/2$. Then $u \sim \mu$ is almost surely s-Hölder-continuous for any exponent $s < \min\{1, \alpha - d/2\}$.

Proof. The Karhunen–Loève expansion (6.10) shows that

$$u(x) = \sum_{k \in \mathbb{K}} \frac{1}{|\lambda_k|^{\alpha/2}} \xi_k \phi_k(x).$$

Thus, for any $\iota > 0$ and for C a (possibly changing) constant independent of t, x and ξ,

$$\begin{aligned}
\mathbb{E}|u(x+h) - u(x)|^2 &\leqslant C \sum_{k \in \mathbb{K}} \frac{1}{|k|^{2\alpha}} |\phi_k(x+h) - \phi_k(x)|^2 \\
&\leqslant C \sum_{k \in \mathbb{K}} \frac{1}{|k|^{2\alpha}} \min\{|k|^2|h|^2, 1\} \\
&\leqslant C \int_{|k| \geqslant 1} \frac{1}{|k|^{2\alpha}} \min\{|k|^2|h|^2, 1\} \, \mathrm{d}k
\end{aligned}$$

$$\begin{aligned}
&\leqslant C\int_{1\leqslant|k|\leqslant|h|^{-\iota}}|k|^{2(1-\alpha)}|h|^2\,\mathrm{d}k + C\int_{|k|\geqslant|h|^{-\iota}}|k|^{-2\alpha}\,\mathrm{d}k\\
&\leqslant C|h|^2\int_1^{|h|^{-\iota}} r^{d-1+2(1-\alpha)}\,\mathrm{d}r + C\int_{|h|^{-\iota}}^{\infty} r^{d-1-2\alpha}\,\mathrm{d}r\\
&= C\big(|h|^{2-\iota(d+2(1-\alpha))} + |h|^{-\iota(d-2\alpha)}\big).
\end{aligned}$$

Making the optimal choice $\iota = 1$ gives

$$\mathbb{E}|u(x+h) - u(x)|^2 \leqslant C|h|^{2\alpha-d}.$$

Thus, by Corollary 6.8 with $\mathcal{H} = \mathbb{R}^n$,

$$\mathbb{E}|u(x) - u(y)|^{2p} \leqslant C|x-y|^{(2\alpha-d)p}$$

for any $p \in \mathbb{N}$. Choosing the exponents $\delta = 2p$ and $\varepsilon = (2\alpha-d)p - 2d$ and letting $p \to \infty$, we deduce from Theorem 6.24 that the function is s-Hölder with any exponent s as specified. □

Example 6.26. Assume that a Gaussian random field with measure μ has the property that, for $X = C(\overline{D}; \mathbb{R}^n)$, $\mu(X) = 1$. Then the Cameron–Martin space for this measure, denoted by $(E, \langle\cdot,\cdot\rangle_E)$, is compactly embedded in X, by Theorem 6.11, and hence there is a constant $C > 0$ such that

$$\|\cdot\|_X \leqslant C\|\cdot\|_E.$$

Thus pointwise evaluation is a continuous linear functional on the Cameron–Martin space so that this space may be viewed as an RKHS.

As an example consider the Gaussian measure $\mathcal{N}(0, \beta\mathcal{A}^{-\alpha})$ on $\mathcal{H}$, with $\mathcal{A}$ satisfying Assumptions 2.9(i)–(iv). Then $\mu(X) = 1$ for $\alpha > d/2$ by Lemma 6.25. The Cameron–Martin space is just $\mathcal{H}^\alpha$. This shows that the space $\mathcal{H}^\alpha$ is compactly embedded in the space of continuous functions, for $\alpha > d/2$. (Of course, a related fact follows more directly from the Sobolev Embedding Theorem, Theorem 2.10.) ◇

We now turn to Sobolev regularity, again using the Karhunen–Loève expansion. Recall the Sobolev-like spaces (2.29) defining $\mathcal{H}^s = D(\mathcal{A}^{s/2})$.

Lemma 6.27. Consider a Gaussian measure $\mu = \mathcal{N}(0, \mathcal{A}^{-\alpha})$, where $\mathcal{A}$ satisfies Assumptions 2.9(i)–(iii) and $\alpha > d/2$. Then $u \sim \mu$ is in $\mathcal{H}^s$ almost surely for any $s \in [0, \alpha - d/2)$.

Proof. The Karhunen–Loève expansion (6.10) shows that

$$u = \sum_{k\in\mathbb{K}} \sqrt{\gamma}_k \xi_k \phi_k,$$

with $\{\xi_k\}$ an i.i.d. $\mathcal{N}(0,1)$-sequence and $\gamma_k = \lambda_k^{-\alpha}$. Thus

$$\mathbb{E}\|u\|_s^2 = \sum_{k\in\mathbb{K}} \gamma_k \lambda_k^s.$$

If the sum is finite then $\mathbb{E}\|u\|_s^2 < \infty$ and $u \in \mathcal{H}^s$ μ-a.s. We have

$$\sum_{k\in\mathbb{K}} \gamma_k \lambda_k^s = \sum_{k\in\mathbb{K}} \lambda_k^{s-\alpha}.$$

Since the eigenvalues λ_k of $\mathcal{A}$ grow like $|k|^2$, we deduce that this sum is finite if and only if $\alpha > s + d/2$, by comparison with an integral. □

It is interesting that the Hölder and Sobolev exponents predicted by Lemmas 6.25 and Lemma 6.27 agree for $d/2 < \alpha < d/2+1$. The proof of Hölder regularity uses Gaussianity in a fundamental way to obtain this property. In particular, in the proof of Lemma 6.25, we use the fact that the second moment of Gaussians can be used to bound arbitrarily high moments. Note that using the Sobolev Embedding Theorem, together with Lemma 6.27, to determine Hölder properties does not, of course, give results which are as sharp as those obtained from Lemma 6.25. For example, using Lemma 6.27 and Theorem 2.10 shows that choosing $\alpha > d$ ensures that u is almost surely continuous. On the other hand Lemma 6.25 shows that choosing $\alpha = d$ ensures that u is almost surely Hölder-continuous with any exponent less than $d/2$; in particular, u is almost surely continuous.

Example 6.28. Consider the case $d = 2, n = 1$ and $D = [0,1]^2$. Define the Gaussian random field through the measure $\mu = \mathcal{N}(0, (-\triangle)^{-\alpha})$, where $\triangle$ is the Laplacian with domain $H_0^1(D) \cap H^2(D)$. Then Assumptions 2.9 are satisfied by $-\triangle$. By Lemma 6.27 it follows that choosing $\alpha > 1$ suffices to ensure that draws from μ are almost surely in $L^2(D)$. Then, by Lemma 6.25, it follows that, in fact, draws from μ are almost surely in $C(D)$. ◇

In many applications in this article we will be interested in constructing a probability measure μ on a Hilbert space $\mathcal{H}$ which is absolutely continuous with respect to a given reference Gaussian measure μ_0. We can then write, via Theorem 6.2,

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(x) \propto \exp\bigl(-\Phi(x)\bigr). \tag{6.20}$$

The Theorem 6.14 provides an explicit example of this structure when μ and μ_0 are both Gaussian. For expression (6.20) to make sense we require that the potential $\Phi : \mathcal{H} \mapsto \mathbb{R}$ is μ_0-measurable. Implicit in the statement of Theorem 6.14 is just such a measurability property of the logarithm of the density between the two Gaussian measures. We return to the structure (6.20) again, in the case where μ is not necessarily Gaussian, in the next subsection.

6.6. Bayesian probability

Bayesian probability forms the underpinnings of the approach to inverse problems taken in this article. In this subsection we first discuss the general

concept of conditioned measures. We then turn to Bayesian probability in the finite-dimensional case, and finally generalize Bayes' theorem to the function space setting. The following theorem is of central importance.

Theorem 6.29. Let μ, ν be probability measures on $S \times T$, where $(S, \mathcal{A})$ and $(T, \mathcal{B})$ are measurable spaces. Let (x, y), with $x \in S$ and $y \in T$, denote an element of $S \times T$. Assume that $\mu \ll \nu$ and that μ has Radon–Nikodym derivative ϕ with respect to ν. Assume further that the conditional distribution of $x|y$ under ν, denoted by $\nu^y(\mathrm{d}x)$, exists. Then the conditional distribution of $x|y$ under μ, denoted $\mu^y(\mathrm{d}x)$, exists and $\mu^y \ll \nu^y$. The Radon–Nikodym derivative is given by

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\nu^y}(x) = \begin{cases} \frac{1}{c(y)}\phi(x,y) & \text{if } c(y) > 0, \text{ and} \\ 1 & \text{else,} \end{cases} \tag{6.21}$$

with $c(y) = \int_S \phi(x,y)\,\mathrm{d}\nu^y(x)$ for all $y \in T$. ◇

Given a probability triplet $(\Omega, \mathcal{F}, \mathbb{P})$ and two sets $A, B \in \mathcal{F}$ with $\mathbb{P}(A) > 0, \mathbb{P}(B) > 0$, we define the probabilities of A given B and B given A by

$$\mathbb{P}(A|B) = \frac{1}{\mathbb{P}(B)}\mathbb{P}(A \cap B),$$
$$\mathbb{P}(B|A) = \frac{1}{\mathbb{P}(A)}\mathbb{P}(A \cap B).$$

Combining gives Bayes' formula:

$$\mathbb{P}(A|B) = \frac{1}{\mathbb{P}(B)}\mathbb{P}(B|A)\mathbb{P}(A). \tag{6.22}$$

If $(u, y) \in \mathbb{R}^d \times \mathbb{R}^\ell$ is a jointly distributed pair of random variables with Lebesgue density $\rho(u, y)$, then the infinitesimal version of the preceding formula tells us that

$$\rho(u|y) \propto \rho(y|u)\rho(u), \tag{6.23}$$

and where the normalization constant depends only on y. Thus

$$\rho(u|y) = \frac{\rho(y|u)\rho(u)}{\int_{\mathbb{R}^d} \rho(y|u)\rho(u)\,\mathrm{d}u}. \tag{6.24}$$

This gives an expression for the probability of a random variable u, given a single observation of a random variable y, which requires knowledge of only the *prior* (unconditioned) probability density $\rho(u)$ and the conditional probability density $\rho(y|u)$ of y given u. Both these expressions are readily available in many modelling scenarios, as we demonstrate in Section 3. This observation is the starting point for the Bayesian approach to probability. Furthermore, there is a wide range of sampling methods which are designed

to sample probability measures known only up to a multiplicative constant (see Section 5), and knowledge of the normalization constant is not required in this context: the formula (6.23) may be used directly to implement these algorithms. Recall that in the general Bayesian framework introduced in Section 1 we refer to the observation y as *data* and to $\rho(y|u)$ as the *likelihood* of the data.

Example 6.30. Consider Example 6.22. The random variable (u, y) is distributed according to a measure $\mu_0(u, y)$, which has density with respect to Lebesgue measure given by

$$\pi_0(u, y) = \frac{1}{2\pi\sigma} \exp\left(-\frac{1}{2}u^2 - \frac{1}{2\sigma^2}|y - u|^2\right).$$

By completing the square we showed that the posterior probability measure for u given y is $\mu_0(u|y)$ with density

$$\pi_0(u|y) = \sqrt{\left(\frac{1+\sigma^2}{2\pi\sigma^2}\right)} \exp\left(-\left(\frac{\sigma^2+1}{2\sigma^2}\right)\left|u - \frac{1}{\sigma^2+1}y\right|^2\right).$$

This result also follows from (6.23), which shows that

$$\pi_0(u|y) = \frac{\pi(u, y)}{\int_{\mathbb{R}^d} \pi(u, y)\,\mathrm{d}u}.$$

Now consider a random variable (u, y) distributed according to measure $\mu(u, y)$, which has density $\rho(u, y)$ with respect to $\mu_0(u, y)$. We assume that $\rho > 0$ everywhere on $\mathbb{R}^d \times \mathbb{R}^\ell$. By Theorem 6.29 the random variable found by conditioning u from μ on y has density

$$\rho(u|y) = \frac{\rho(u, y)}{\int_{\mathbb{R}^d} \rho(u, y)\pi_0(u|y)\,\mathrm{d}u}$$

with respect to $\pi_0(u|y)$. ◇

The expression (6.23) may be rewritten to give an expression for the *ratio* of the posterior and prior p.d.f.s:

$$\frac{\rho(u|y)}{\rho(u)} \propto \rho(y|u), \tag{6.25}$$

with constant of proportionality which depends only on y, and not on u. Stated in this way, the formula has a natural generalization to infinite dimensions, as we now explain.

Let u be a random variable distributed according to measure μ_0 on a separable Banach space $(X, \|\cdot\|)$. We assume that the *data* $y \in \mathbb{R}^m$ is given in terms of the *observation operator* $\mathcal{G}$ by the formula $y = \mathcal{G}(u)+\eta$, where $\eta \in \mathbb{R}^m$ is independent of u and has density ρ with respect to Lebesgue measure;

for simplicity we assume that the support of ρ is $\mathbb{R}^m$. Define $\Phi(u;y)$ to be any function which differs from $-\log(\rho(y-\mathcal{G}(u)))$ by an additive function of y only. Hence it follows that

$$\frac{\rho(y-\mathcal{G}(u))}{\rho(y)} \propto \exp\bigl(-\Phi(u;y)\bigr),$$

with constant of proportionality independent of u. Use of Bayes' rule in the form (6.25) suggests that the probability measure for u given y, denoted $\mu^y(\mathrm{d}u)$, has Radon–Nikodym derivative with respect to μ_0 given by

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\bigl(-\Phi(u;y)\bigr). \tag{6.26}$$

We refer to such an argument as *informal application of Bayes' rule*. We now justify the formula rigorously.

Theorem 6.31. Assume that $\mathcal{G} : X \to \mathbb{R}^m$ is continuous, that ρ has support equal to $\mathbb{R}^m$ and that $\mu_0(X) = 1$. Then $u|y$ is distributed according to the measure $\mu^y(\mathrm{d}u)$, which is absolutely continuous with respect to $\mu_0(\mathrm{d}u)$ and has Radon–Nikodym derivative given by (6.26).

Proof. Throughout the proof $C(y)$ denotes a constant depending on y, but not on u, and possibly changing between occurrences. Let $\mathbb{Q}_0(\mathrm{d}y) = \rho(y)\,\mathrm{d}y$ and $\mathbb{Q}(\mathrm{d}y|u) = \rho(y-\mathcal{G}(u))\,\mathrm{d}y$. By construction,

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{Q}_0}(y|u) = C(y)\exp\bigl(-\Phi(u;y)\bigr),$$

with constant of proportionality independent of u. Now define

$$\begin{aligned}\nu_0(\mathrm{d}y,\mathrm{d}u) &= \mathbb{Q}_0(\mathrm{d}y)\otimes\mu_0(\mathrm{d}u),\\ \nu(\mathrm{d}y,\mathrm{d}u) &= \mathbb{Q}(\mathrm{d}y|u)\mu_0(\mathrm{d}u).\end{aligned}$$

Note that ν_0 is a product measure under which u and y are independent random variables. Since $\mathcal{G} : X \to \mathbb{R}^m$ is continuous we deduce that $\Phi : X \to \mathbb{R}$ is continuous and hence, since $\mu_0(X) = 1$, μ_0-measurable. Thus ν is well-defined and is absolutely continuous with respect to ν_0 with Radon–Nikodym derivative

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu_0}(y,u) = C(y)\exp\bigl(-\Phi(u;y)\bigr);$$

again the constant of proportionality depends only on y. Note that

$$\int_X \exp\bigl(-\Phi(u;y)\bigr)\mu_0(\mathrm{d}u) = C(y)\int_X \rho(y-\mathcal{G}(u))\mu_0(\mathrm{d}u) > 0$$

since $\rho > 0$ everywhere on $\mathbb{R}^m$ and since $\mathcal{G} : X \to \mathbb{R}^m$ is continuous. By Theorem 6.29 we have the desired result, since $\nu_0(\mathrm{d}u|y) = \mu_0(\mathrm{d}u)$. □

Remark 6.32. Finally we remark that, if μ^y is absolutely continuous with respect to μ_0 then any property which holds almost surely under μ_0 will also hold almost surely under μ^y. The next example illustrates how useful this fact is. ◇

Example 6.33. Let μ_0 denote the Gaussian random field constructed in Example 6.28, with $\alpha > 1$ so that draws from μ_0 are almost surely continuous. Now imagine that we observe y, the $L^2(D)$-norm of u drawn from μ_0, subject to noise η:

$$y = \|u\|^2 + \eta.$$

We assume that $\eta \sim \mathcal{N}(0, \gamma^2)$, independently of u. The $L^2(D)$-norm is a continuous function on $X = C(D)$ and $\mu_0(X) = 1$; hence evaluation of the $L^2(D)$-norm is μ_0-measurable, and the measure $\mu^y(\mathrm{d}u) = \mathbb{P}(\mathrm{d}u|y)$ is absolutely continuous with respect to μ_0, with Radon–Nikodym derivative given by

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2\gamma^2}\left|y - \|u\|^2\right|^2\right).$$

Note that the probability measure μ^y is not Gaussian. Nonetheless, any function drawn from μ^y is almost surely in $C(D)$. ◇

6.7. Metrics on measures

In Section 4 it will be important to estimate the distance between two probability measures and thus we will be interested in metrics which measure distance between probability measures.

In this section we introduce two useful metrics on measures: the *total variation distance* and the *Hellinger distance*. We discuss the relationships between the metrics and indicate how they may be used to estimate differences between expectations of random variables under two different measures.

Assume that we have two probability measures μ and μ', both absolutely continuous with respect to the same reference measure ν. The following definitions give two concepts of distance between μ and μ'.

Definition 6.34. The *total variation distance* between μ and μ' is

$$d_{\mathrm{TV}}(\mu, \mu') = \frac{1}{2}\int \left|\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - \frac{\mathrm{d}\mu'}{\mathrm{d}\nu}\right| \mathrm{d}\nu. \qquad \diamond$$

In particular, if μ' is absolutely continuous with respect to μ, then

$$d_{\mathrm{TV}}(\mu, \mu') = \frac{1}{2}\int \left|1 - \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right| \mathrm{d}\mu. \tag{6.27}$$

Definition 6.35. The *Hellinger distance* between μ and μ' is

$$d_{\mathrm{Hell}}(\mu,\mu') = \sqrt{\left(\frac{1}{2}\int\left(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}} - \sqrt{\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}}\right)^2 \mathrm{d}\nu\right)}. \qquad \diamond$$

In particular, if μ' is absolutely continuous with respect to μ, then

$$d_{\mathrm{Hell}}(\mu,\mu') = \sqrt{\left(\frac{1}{2}\int\left(1 - \sqrt{\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}}\right)^2 \mathrm{d}\mu\right)}. \tag{6.28}$$

The total variation distance as defined is invariant under the choice of ν in that it is unchanged if a different reference measure, with respect to which μ and μ' are absolutely continuous, is used. Furthermore, it follows from the definition that

$$0 \leqslant d_{\mathrm{TV}}(\mu,\mu') \leqslant 1.$$

The Hellinger distance is also unchanged if a different reference measure, with respect to which μ and μ' are absolutely continuous, is used. Furthermore, it follows from the definition that

$$0 \leqslant d_{\mathrm{Hell}}(\mu,\mu') \leqslant 1.$$

The Hellinger and total variation distances are related as follows:

Lemma 6.36. Assume that two probability measures μ and μ' are both absolutely continuous with respect to a measure ν. Then

$$\frac{1}{\sqrt{2}} d_{\mathrm{TV}}(\mu,\mu') \leqslant d_{\mathrm{Hell}}(\mu,\mu') \leqslant d_{\mathrm{TV}}(\mu,\mu')^{1/2}. \qquad \diamond$$

The Hellinger distance is particularly useful for estimating the difference between expectation values of functions of random variables under different measures. This idea is encapsulated in the following lemma.

Lemma 6.37. Assume that two probability measures μ and μ' on a Banach space $(X, \|\cdot\|_X)$ are both absolutely continuous with respect to a measure ν. Assume also that $f : X \to E$, where $(E, \|\cdot\|)$ is a Banach space, has second moments with respect to both μ and μ'. Then

$$\|\mathbb{E}^{\mu} f - \mathbb{E}^{\mu'} f\| \leqslant 2\big(\mathbb{E}^{\mu}\|f\|^2 + \mathbb{E}^{\mu'}\|f\|^2\big)^{1/2} d_{\mathrm{Hell}}(\mu,\mu').$$

Furthermore, if $(E, \langle\cdot,\cdot\rangle, \|\cdot\|)$ is a Hilbert space and $f : X \to E$ has fourth moments, then

$$\|\mathbb{E}^{\mu} f\otimes f - \mathbb{E}^{\mu'} f\otimes f\| \leqslant 2\big(\mathbb{E}^{\mu}\|f\|^4 + \mathbb{E}^{\mu'}\|f\|^4\big)^{1/2} d_{\mathrm{Hell}}(\mu,\mu'). \qquad \diamond$$

Remark 6.38. Note, in particular, that choosing $X = E$, and with f chosen to be the identity mapping, we deduce that the differences in mean and covariance operators under two measures are bounded above by the Hellinger distance between the two measures. $\diamond$

6.8. Discussion and bibliography

For a general classical introduction to probability theory see Breiman (1992), and for a modern treatment of the subject see Grimmett and Stirzaker (2001). For a concise, modern (and more advanced) treatment of the subject see Williams (1991). The text by Chorin and Hald (2006) provides an overview of tools from probability and stochastic processes aimed at applied mathematicians.

The discussion of Gaussian measures in a Hilbert space, and proofs of Lemma 6.15 and Theorems 6.6, 6.2, 6.13 and 6.14 may be found in Da Prato and Zabczyk (1992). Theorem 6.14 is also proved in Bogachev (1998). The lecture notes by Hairer (2009) are also a good source, and contain a proof of Theorem 6.1 as well as the Fernique Theorem. Bogachev (1998), Hairer (2009) and Lifshits (1995) all discuss Gaussian measures in the Banach space setting. In particular, Theorem 6.4 is proved in Lifshits (1995), and Hairer (2009) has a nice exposition of the Fernique Theorem.

The Karhunen–Loève expansion is described in Loève (1977, 1978) and a modern treatment of Gaussian random fields is contained in Adler (1990). Recent work exploiting the Karhunen–Loève expansion to approximate solutions of differential equations with random coefficients may be found in Schwab and Todor (2006) and Todor and Schwab (2007).

Theorem 6.29 is proved in Dudley (2002, Section 10.2). For a general discussion of Bayes' rule in finite dimensions see, for example, Bickel and Doksum (2001). The approach to Bayes' rule in infinite dimensions that we adopt in Theorem 6.31 was used to study a specific problem arising in signal processing in Hairer *et al.* (2007). The topic of metrics on probability measures, and further references to the literature, may be found in Gibbs and Su (2002). Note that the choice of normalization constants in the definitions of the total variation and Hellinger metrics differs in the literature.

Acknowledgements

The material in this article is developed in greater detail in the lecture notes of Dashti *et al.* (2010*b*). These notes are freely available for download from: http://www.warwick.ac.uk/~masdr/inverse.html. The author is grateful to his co-authors Masoumeh Dashti and Natesh Pillai for their input into this article. The author also thanks Sergios Agapiou, Andrew Duncan, Stephen Harris, Sebastian Reich and Sebastian Vollmer for numerous comments which improved the presentation, and to Daniella Calvetti and Erkki Somersalo for useful pointers to relevant literature. Finally, the author is grateful to have received financial support from the Engineering and Physical Sciences Research Council (UK), the European Research Council and from the US Office of Naval Research during the writing of this article. This funded research has helped shape much of the material presented.

REFERENCES

R. J. Adler (1990), *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*, Vol. 12 of *Institute of Mathematical Statistics Lecture Notes: Monograph Series*, Institute of Mathematical Statistics, Hayward, CA.

S. Akella and I. Navon (2009), 'Different approaches to model error formulation in 4D-Var: A study with high resolution advection schemes', *Tellus* **61A**, 112–128.

A. Alekseev and I. Navon (2001), 'The analysis of an ill-posed problem using multiscale resolution and second order adjoint techniques', *Comput. Meth. Appl. Mech. Engrg* **190**, 1937–1953.

A. Antoulas, D. Soresen and S. Gugerrin (2001), *A Survey of Model Reduction Methods for Large Scale Dynamical Systems*, AMS.

A. Apte, M. Hairer, A. Stuart and J. Voss (2007), 'Sampling the posterior: An approach to non-Gaussian data assimilation', *Physica D* **230**, 50–64.

A. Apte, C. Jones and A. Stuart (2008*a*), 'A Bayesian approach to Lagrangian data assimilation', *Tellus* **60**, 336–347.

A. Apte, C. Jones, A. Stuart and J. Voss (2008*b*), 'Data assimilation: Mathematical and statistical perspectives', *Internat. J. Numer. Methods Fluids* **56**, 1033–1046.

C. Archambeau, D. Cornford, M. Opper and J. Shawe-Taylor (2007), Gaussian process approximations of stochastic differential equations. In *JMLR Workshop and Conference Proceedings 1: Gaussian Processes in Practice* (N. Lawrence, ed.), The MIT Press, pp. 1–16.

C. Archambeau, M. Opper, Y. Shen, D. Cornford and J. Shawe-Taylor (2008), Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems 20* (J. Platt, D. Koller, Y. Singer and S. Roweis, eds), The MIT Press, Cambridge, MA, pp. 17–24.

G. Backus (1970*a*), 'Inference from inadequate and inaccurate data I', *Proc. Nat. Acad. Sci.* **65**, 1–7.

G. Backus (1970*b*), 'Inference from inadequate and inaccurate data II', *Proc. Nat. Acad. Sci.* **65**, 281–287.

G. Backus (1970*c*), 'Inference from inadequate and inaccurate data III', *Proc. Nat. Acad. Sci.* **67**, 282–289.

A. Bain and D. Crisan (2009), *Fundamentals of Stochastic Filtering*, Springer.

R. Bannister, D. Katz, M. Cullen, A. Lawless and N. Nichols (2008), 'Modelling of forecast errors in geophysical fluid flows', *Internat. J. Numer. Methods Fluids* **56**, 1147–1153.

J. Beck, B. Blackwell and C. Clair (2005), *Inverse Heat Conduction: Ill-Posed Problems*, Wiley.

M. Bell, M. Martin and N. Nichols (2004), 'Assimilation of data into an ocean model with systematic errors near the equator', *Quart. J. Royal Met. Soc.* **130**, 873–894.

T. Bengtsson, P. Bickel and B. Li (2008), 'Curse of dimensionality revisited: The collapse of importance sampling in very large scale systems', *IMS Collections: Probability and Statistics: Essays in Honor of David Freedman* **2**, 316–334.

T. Bengtsson, C. Snyder and D. Nychka (2003), 'Toward a nonlinear ensemble filter for high-dimensional systems', *J. Geophys. Res.* **108**, 8775.

A. Bennett (2002), *Inverse Modeling of the Ocean and Atmosphere*, Cambridge University Press.

A. Bennett and W. Budgell (1987), 'Ocean data assimilation and the Kalman filter: Spatial regularity', *J. Phys. Oceanography* **17**, 1583–1601.

A. Bennett and B. Chua (1994), 'Open ocean modelling as an inverse problem', *Monthly Weather Review* **122**, 1326–1336.

A. Bennett and R. Miller (1990), 'Weighting initial conditions in variational assimilation schemes', *Monthly Weather Review* **119**, 1098–1102.

K. Bergemann and S. Reich (2010), 'A localization technique for ensemble transform Kalman filters', *Quart. J. Royal Met. Soc.* To appear.

L. Berliner (2001), 'Monte Carlo based ensemble forecasting', *Statist. Comput.* **11**, 269–275.

J. Bernardo and A. Smith (1994), *Bayesian Theory*, Wiley.

A. Beskos and A. Stuart (2009), MCMC methods for sampling function space. In *Invited Lectures: Sixth International Congress on Industrial and Applied Mathematics, ICIAM07* (R. Jeltsch and G. Wanner, eds), European Mathematical Society, pp. 337–364.

A. Beskos and A. M. Stuart (2010), Computational complexity of Metropolis–Hastings methods in high dimensions. In *Monte Carlo and Quasi-Monte Carlo Methods 2008* (P. L'Ecuyer and A. B. Owen, eds), Springer, pp. 61–72.

A. Beskos, G. O. Roberts and A. M. Stuart (2009), 'Optimal scalings for local Metropolis–Hastings chains on non-product targets in high dimensions', *Ann. Appl. Probab.* **19**, 863–898.

A. Beskos, G. O. Roberts, A. M. Stuart and J. Voss (2008), 'MCMC methods for diffusion bridges', *Stochastic Dynamics* **8**, 319–350.

P. Bickel and K. Doksum (2001), *Mathematical Statistics*, Prentice-Hall.

P. Bickel, B. Li and T. Bengtsson (2008), 'Sharp failure rates for the bootstrap particle filter in high dimensions', *IMS Collections: Pushing the Limits of Contemporary Statistics* **3**, 318–329.

V. Bogachev (1998), *Gaussian Measures*, AMS.

P. Bolhuis, D. Chandler, D. Dellago and P. Geissler (2002), 'Transition path sampling: Throwing ropes over rough mountain passes', *Ann. Rev. Phys. Chem.* **53**, 291–318.

L. Borcea (2002), 'Electrical impedence tomography', *Inverse Problems* **18**, R99–R136.

P. Brasseur, P. Bahurel, L. Bertino, F. Birol, J.-M. Brankart, N. Ferry, S. Losa, E. Remy, J. Schroeter, S. Skachko, C.-E. Testut, B. Tranchat, P. Van Leeuwen and J. Verron (2005), 'Data assimilation for marine monitoring and prediction: The Mercator operational assimilation systems and the Mersea developments', *Quart. J. Royal Met. Soc.* **131**, 3561–3582.

L. Breiman (1992), *Probability*, Vol. 7 of *Classics in Applied Mathematics*, SIAM, Philadelphia, PA. Corrected reprint of the 1968 original.

G. Burgers, P. Van Leeuwen and G. Evensen (1998), 'On the analysis scheme in the ensemble Kalman filter', *Monthly Weather Review* **126**, 1719–1724.

D. Calvetti (2007), 'Preconditioned iterative methods for linear discrete ill-posed problems from a Bayesian inversion perspective', *J. Comput. Appl. Math.* **198**, 378–395.

D. Calvetti and E. Somersalo (2005*a*), 'Priorconditioners for linear systems', *Inverse Problems* **21**, 1397–1418.

D. Calvetti and E. Somersalo (2005*b*), 'Statistical elimination of boundary artefacts in image deblurring', *Inverse Problems* **21**, 1697–1714.

D. Calvetti and E. Somersalo (2006), 'Large-scale statistical parameter estimation in complex systems with an application to metabolic models', *Multiscale Modeling and Simulation* **5**, 1333–1366.

D. Calvetti and E. Somersalo (2007*a*), 'Gaussian hypermodel to recover blocky objects', *Inverse Problems* **23**, 733–754.

D. Calvetti and E. Somersalo (2007*b*), *Introduction to Bayesian Scientific Computing*, Vol. 2 of *Surveys and Tutorials in the Applied Mathematical Sciences*, Springer.

D. Calvetti and E. Somersalo (2008), 'Hypermodels in the Bayesian imaging framework', *Inverse Problems* **24**, #034013.

D. Calvetti, H. Hakula, S. Pursiainen and E. Somersalo (2009), 'Conditionally Gaussian hypermodels for cerebral source location', *SIAM J. Imag. Sci.* **2**, 879–909.

D. Calvetti, A. Kuceyeski and E. Somersalo (2008), 'Sampling based analysis of a spatially distributed model for liver metabolism at steady state', *Multiscale Modeling and Simulation* **7**, 407–431.

E. Candès and M. Wakin (2008), 'An introduction to compressive sampling', *IEEE Signal Processing Magazine*, March 2008, 21–30.

J.-Y. Chemin and N. Lerner (1995), 'Flot de champs de veceurs non lipschitziens et équations de Navier–Stokes', *J. Diff. Equations* **121**, 314–328.

A. Chorin and O. Hald (2006), *Stochastic Tools in Mathematics and Science*, Vol. 1 of *Surveys and Tutorials in the Applied Mathematical Sciences*, Springer, New York.

A. Chorin and P. Krause (2004), 'Dimensional reduction for a Bayesian filter', *Proc. Nat. Acad. Sci.* **101**, 15013–15017.

A. Chorin and X. Tu (2009), 'Implicit sampling for particle filters', *Proc. Nat. Acad. Sci.* **106**, 17249–17254.

A. Chorin and X. Tu (2010), 'Interpolation and iteration for nonlinear filters', *Math. Model. Numer. Anal.* To appear.

M. Christie (2010), Solution error modelling and inverse problems. In *Simplicity, Complexity and Modelling*, Wiley, New York, to appear.

M. Christie, G. Pickup, A. O'Sullivan and V. Demyanov (2008), Use of solution error models in history matching. In *Proc. European Conference on the Mathematics of Oil Recovery XI*, European Association of Geoscientists and Engineers.

B. Chua and A. Bennett (2001), 'An inverse ocean modelling system', *Ocean. Meteor.* **3**, 137–165.

S. Cohn (1997), 'An introduction to estimation theory', *J. Met. Soc. Japan* **75**, 257–288.

S. Cotter, M. Dashti, J. Robinson and A. Stuart (2009), 'Bayesian inverse problems for functions and applications to fluid mechanics', *Inverse Problems* **25**, #115008.

S. Cotter, M. Dashti and A. Stuart (2010*a*), 'Approximation of Bayesian inverse problems', *SIAM J. Numer. Anal.* To appear.

S. Cotter, M. Dashti, J. Robinson and A. Stuart (2010*b*). In preparation.

P. Courtier (1997), 'Dual formulation of variational assimilation', *Quart. J. Royal Met. Soc.* **123**, 2449–2461.

P. Courtier and O. Talagrand (1987), 'Variational assimilation of meteorological observations with the adjoint vorticity equation II: Numerical results', *Quart. J. Royal Met. Soc.* **113**, 1329–1347.

P. Courtier, E. Anderson, W. Heckley, J. Pailleux, D. Vasiljevic, M. Hamrud, A. Hollingworth, F. Rabier and M. Fisher (1998), 'The ECMWF implementation of three-dimensional variational assimilation (3D-Var)', *Quart. J. Royal Met. Soc.* **124**, 1783–1808.

N. Cressie (1993), *Statistics for Spatial Data*, Wiley.

T. Cui, C. Fox, G. Nicholls and M. O'Sullivan (2010), 'Using MCMC sampling to calibrate a computer model of a geothermal field'. Submitted.

G. Da Prato and J. Zabczyk (1992), *Stochastic Equations in Infinite Dimensions*, Vol. 44 of *Encyclopedia of Mathematics and its Applications*, Cambridge University Press.

B. Dacarogna (1989), *Direct Methods in the Calculus of Variations*, Springer, New York.

M. Dashti and J. Robinson (2009), 'Uniqueness of the particle trajectories of the weak solutions of the two-dimensional Navier–Stokes equations', *Nonlinearity* **22**, 735–746.

M. Dashti, S. Harris and A. M. Stuart (2010*a*), Bayesian approach to an elliptic inverse problem. In preparation.

M. Dashti, N. Pillai and A. Stuart (2010*b*), *Bayesian Inverse Problems in Differential Equations*. Lecture notes, available from: http://www.warwick.ac.uk/~masdr/inverse.html.

J. Derber (1989), 'A variational continuous assimilation technique', *Monthly Weather Review* **117**, 2437–2446.

P. Deuflhard (2004), *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*, Springer.

B. DeVolder, J. Glimm, J. Grove, Y. Kang, Y. Lee, K. Pao, D. Sharp and K. Ye (2002), 'Uncertainty quantification for multiscale simulations', *J. Fluids Engrg* **124**, 29–42.

D. Donoho (2006), 'Compressed sensing', *IEEE Trans. Inform. Theory* **52**, 1289–1306.

P. Dostert, Y. Efendiev, T. Hou and W. Luo (2006), 'Coarse-grain Langevin algorithms for dynamic data integration', *J. Comput. Phys.* **217**, 123–142.

N. Doucet, A. de Frietas and N. Gordon (2001), *Sequential Monte Carlo in Practice*, Springer.

R. Dudley (2002), *Real Analysis and Probability*, Cambridge University Press, Cambridge.

D. Dürr and A. Bach (1978), 'The Onsager–Machlup function as Lagrangian for the most probable path of a diffusion process', *Comm. Math. Phys.* **160**, 153–170.

Y. Efendiev, A. Datta-Gupta, X. Ma and B. Mallick (2009), 'Efficient sampling techniques for uncertainty quantification in history matching using nonlinear error models and ensemble level upscaling techniques', *Water Resources Res.* **45**, #W11414.

M. Eknes and G. Evensen (1997), 'Parameter estimation solving a weak constraint variational formulation for an Ekman model', *J. Geophys. Res.* **12**, 479–491.

B. Ellerbroek and C. Vogel (2009), 'Inverse problems in astronomical adaptive optics', *Inverse Problems* **25**, #063001.

H. Engl, M. Hanke and A. Neubauer (1996), *Regularization of Inverse Problems*, Kluwer.

H. Engl, A. Hofinger and S. Kindermann (2005), 'Convergence rates in the Prokhorov metric for assessing uncertainty in ill-posed problems', *Inverse Problems* **21**, 399–412.

G. Evensen (2006), *Data Assimilation: The Ensemble Kalman Filter*, Springer.

G. Evensen and P. Van Leeuwen (2000), 'An ensemble Kalman smoother for nonlinear dynamics', *Monthly Weather Review* **128**, 1852–1867.

F. Fang, C. Pain, I. Navon, M. Piggott, G. Gorman, P. Allison and A. Goddard (2009*a*), 'Reduced order modelling of an adaptive mesh ocean model', *Internat. J. Numer. Methods Fluids* **59**, 827–851.

F. Fang, C. Pain, I. Navon, M. Piggott, G. Gorman, P. Farrell, P. Allison and A. Goddard (2009*b*), 'A POD reduced-order 4D-Var adaptive mesh ocean modelling approach', *Internat. J. Numer. Methods Fluids* **60**, 709–732.

C. Farmer (2005), Geological modelling and reservoir simulation. In *Mathematical Methods and Modeling in Hydrocarbon Exploration and Production* (A. Iske and T. Randen, eds), Springer, Heidelberg, pp. 119–212.

C. Farmer (2007), Bayesian field theory applied to scattered data interpolation and inverse problems. In *Algorithms for Approximation* (A. Iske and J. Levesley, eds), Springer, pp. 147–166.

B. Fitzpatrick (1991), 'Bayesian analysis in inverse problems', *Inverse Problems* **7**, 675–702.

J. Franklin (1970), 'Well-posed stochastic extensions of ill-posed linear problems', *J. Math. Anal. Appl.* **31**, 682–716.

M. Freidlin and A. Wentzell (1984), *Random Perturbations of Dynamical Systems*, Springer, New York.

A. Gelfand and A. Smith (1990), 'Sampling-based approaches to calculating marginal densities', *J. Amer. Statist. Soc.* **85**, 398–409.

A. Gibbs and F. Su (2002), 'On choosing and bounding probability metrics', *Internat. Statist. Review* **70**, 419–435.

C. Gittelson and C. Schwab (2011), Sparse tensor discretizations of high-dimensional PDEs. To appear in *Acta Numerica*, Vol. 20.

J. Glimm, S. Hou, Y. Lee, D. Sharp and K. Ye (2003), 'Solution error models for uncertainty quantification', *Contemporary Mathematics* **327**, 115–140.

S. Gratton, A. Lawless and N. Nichols (2007), 'Approximate Gauss–Newton methods for nonlinear least squares problems', *SIAM J. Optimization* **18**, 106–132.

A. Griffith and N. Nichols (1998), Adjoint methods for treating model error in data assimilation. In *Numerical Methods for Fluid Dynamics VI*, ICFD, Oxford, pp. 335–344.

A. Griffith and N. Nichols (2000), 'Adjoint techniques in data assimilation for treating systematic model error', *J. Flow, Turbulence and Combustion* **65**, 469–488.

G. Grimmett and D. Stirzaker (2001), *Probability and Random Processes*, Oxford University Press, New York.

C. Gu (2002), *Smoothing Spline ANOVA Models*, Springer.

C. Gu (2008), 'Smoothing noisy data via regularization', *Inverse Problems* **24**, #034002.

C. Hagelberg, A. Bennett and D. Jones (1996), 'Local existence results for the generalized inverse of the vorticity equation in the plane', *Inverse Problems* **12**, 437–454.

E. Hairer and G. Wanner (1996), *Solving Ordinary Differential Equations II*, Vol. 14 of *Springer Series in Computational Mathematics*, Springer, Berlin.

E. Hairer, S. P. Nørsett and G. Wanner (1993), *Solving Ordinary Differential Equations I*, Vol. 8 of *Springer Series in Computational Mathematics*, Springer, Berlin.

M. Hairer (2009), *Introduction to Stochastic PDEs*. Lecture notes.

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

M. Hairer, A. M. Stuart and J. Voss (2009), Sampling conditioned diffusions. In *Trends in Stochastic Analysis*, Vol. 353 of *London Mathematical Society Lecture Notes*, Cambridge University Press, pp. 159–186.

M. Hairer, A. Stuart and J. Voss (2010*a*), 'Sampling conditioned hypoelliptic diffusions'. Submitted.

M. Hairer, A. Stuart and J. Voss (2010*b*), Signal processing problems on function space: Bayesian formulation, stochastic PDEs and effective MCMC methods. In *Oxford Handbook of Nonlinear Filtering* (D. Crisan and B. Rozovsky, eds), Oxford University Press, to appear.

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

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

T. Hein (2009), 'On Tikhonov regularization in Banach spaces: Optimal convergence rate results', *Applicable Analysis* **88**, 653–667.

J. Heino, K. Tunyan, D. Calvetti and E. Somersalo (2007), 'Bayesian flux balance analysis applied to a skeletal muscle metabolic model', *J. Theor. Biol.* **248**, 91–110.

R. Herbei and I. McKeague (2009), 'Geometric ergodicity of hybrid samplers for ill-posed inverse problems', *Scand. J. Statist.* **36**, 839–853.

R. Herbei, I. McKeague and K. Speer (2008), 'Gyres and jets: Inversion of tracer data for ocean circulation structure', *J. Phys. Oceanography* **38**, 1180–1202.

A. Hofinger and H. Pikkarainen (2007), 'Convergence rates for the Bayesian approach to linear inverse problems', *Inverse Problems* **23**, 2469–2484.

A. Hofinger and H. Pikkarainen (2009), 'Convergence rates for linear inverse problems in the presence of an additive normal noise', *Stoch. Anal. Appl.* **27**, 240–257.

M. Huddleston, M. Bell, M. Martin and N. Nichols (2004), 'Assessment of wind stress errors using bias corrected ocean data assimilation', *Quart. J. Royal Met. Soc.* **130**, 853–872.

M. Hurzeler and H. Künsch (2001), Approximating and maximizing the likelihood for a general state space model. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. Gordon, eds), Springer, pp. 159–175.

J. Huttunen and H. Pikkarainen (2007), 'Discretization error in dynamical inverse problems: One-dimensional model case', *J. Inverse and Ill-posed Problems* **15**, 365–386.

K. Ide and C. Jones (2007), 'Data assimilation', *Physica D* **230**, vii–viii.

K. Ide, L. Kuznetsov and C. Jones (2002), 'Lagrangian data assimilation for point-vortex system', *J. Turbulence* **3**, 53.

N. Ikeda and S. Watanabe (1989), *Stochastic Differential Equations and Diffusion Processes*, second edn, North-Holland, Amsterdam.

M. Jardak, I. Navon and M. Zupanski (2010), 'Comparison of sequential data assimilation methods for the Kuramoto–Sivashinsky equation', *Internat. J. Numer. Methods Fluids* **62**, 374–402.

C. Johnson, B. Hoskins and N. Nichols (2005), 'A singular vector perspective of 4DVAR: Filtering and interpolation', *Quart. J. Royal Met. Soc.* **131**, 1–20.

C. Johnson, B. Hoskins, N. Nichols and S. Ballard (2006), 'A singular vector perspective of 4DVAR: The spatial structure and evolution of baroclinic weather systems', *Monthly Weather Review* **134**, 3436–3455.

J. Kaipio and E. Somersalo (2000), 'Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography', *Inverse Problems* **16**, 1487–1522.

J. Kaipio and E. Somersalo (2005), *Statistical and Computational Inverse problems*, Vol. 160 of *Applied Mathematical Sciences*, Springer.

J. Kaipio and E. Somersalo (2007*a*), 'Approximation errors in nonstationary inverse problems', *Inverse Problems and Imaging* **1**, 77–93.

J. Kaipio and E. Somersalo (2007*b*), 'Statistical inverse problems: Discretization, model reduction and inverse crimes', *J. Comput. Appl. Math.* **198**, 493–504.

E. Kalnay (2003), *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press.

E. Kalnay, H. Li, S. Miyoshi, S. Yang and J. Ballabrera-Poy (2007), '4D-Var or ensemble Kalman filter?', *Tellus* **59**, 758–773.

B. Kaltenbacher, F. Schöpfer and T. Schuster (2009), 'Iterative methods for nonlinear ill-posed problems in Banach spaces: Convergence and applications to parameter identification problems', *Inverse Problems* **25**, #065003.

M. Kennedy and A. O'Hagan (2001), 'Bayesian calibration of computer models', *J. Royal Statist. Soc.* **63B**, 425–464.

D. Kinderlehrer and G. Stampacchia (1980), *An Introduction to Variational Inequalities and their Applications*, SIAM.

T. Kolda and B. Bader (2009), 'Tensor decompositions and applications', *SIAM Review* **51**, 455–500.

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

M. Lassas and S. Siltanen (2004), 'Can one use total variation prior for edge-preserving Bayesian inversion?', *Inverse Problems* **20**, 1537–1563.

M. Lassas, E. Saksman and S. Siltanen (2009), 'Discretization-invariant Bayesian inversion and Besov space priors', *Inverse Problems and Imaging* **3**, 87–122.

A. Lawless and N. Nichols (2006), 'Inner loop stopping criteria for incremental four-dimensional variational data assimilation', *Monthly Weather Review* **134**, 3425–3435.

A. Lawless, S. Gratton and N. Nichols (2005*a*), 'Approximate iterative methods for variational data assimilation', *Internat. J. Numer. Methods Fluids* **47**, 1129–1135.

A. Lawless, S. Gratton and N. Nichols (2005*b*), 'An investigation of incremental 4D-Var using non-tangent linear models', *Quart. J. Royal Met. Soc.* **131**, 459–476.

A. Lawless, N. Nichols, C. Boess and A. Bunse-Gerstner (2008*a*), 'Approximate Gauss–Newton methods for optimal state estimation using reduced order models', *Internat. J. Numer. Methods Fluids* **56**, 1367–1373.

A. Lawless, N. Nichols, C. Boess and A. Bunse-Gerstner (2008*b*), 'Using model reduction methods within incremental four-dimensional variational data assimilation', *Monthly Weather Review* **136**, 1511–1522.

M. Lehtinen, L. Paivarinta and E. Somersalo (1989), 'Linear inverse problems for generalized random variables', *Inverse Problems* **5**, 599–612.

M. Lifshits (1995), *Gaussian Random Functions*, Vol. 322 of *Mathematics and its Applications*, Kluwer, Dordrecht.

D. Livings, S. Dance and N. Nichols (2008), 'Unbiased ensemble square root filters', *Physica D: Nonlinear Phenomena* **237**, 1021–1028.

M. Loève (1977), *Probability Theory I*, fourth edn, Vol. 45 of *Graduate Texts in Mathematics*, Springer, New York.

M. Loève (1978), *Probability Theory II*, fourth edn, Vol. 46 of *Graduate Texts in Mathematics*, Springer, New York.

A. Lorenc (1986), 'Analysis methods for numerical weather prediction', *Quart. J. Royal Met. Soc.* **112**, 1177–1194.

C. Lubich (2008), *From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis*, European Mathematical Society.

X. Ma, M. Al-Harbi, A. Datta-Gupta and Y. Efendiev (2008), 'Multistage sampling approach to quantifying uncertainty during history matching geological models', *Soc. Petr. Engrg J.* **13**, 77–87.

A. Majda and B. Gershgorin (2008), 'A nonlinear test model for filtering slow–fast systems', *Comm. Math. Sci.* **6**, 611–649.

A. Majda and M. Grote (2007), 'Explicit off-line criteria for stable accurate filtering of strongly unstable spatially extended systems', *Proc. Nat. Acad. Sci.* **104**, 1124–1129.

A. Majda and J. Harlim (2010), 'Catastrophic filter divergence in filtering nonlinear dissipative systems', *Comm. Math. Sci.* **8**, 27–43.

A. Majda, J. Harlim and B. Gershgorin (2010), 'Mathematical strategies for filtering turbulent dynamical systems', *Disc. Cont. Dyn. Sys.* To appear.

A. Mandelbaum (1984), 'Linear estimators and measurable linear transformations on a Hilbert space', *Probab. Theory Rel. Fields* **65**, 385–397.

M. Martin, M. Bell and N. Nichols (2002), 'Estimation of systematic error in an equatorial ocean model using data assimilation', *Internat. J. Numer. Methods Fluids* **40**, 435–444.

I. McKeague, G. Nicholls, K. Speer and R. Herbei (2005), 'Statistical inversion of south Atlantic circulation in an abyssal neutral density layer', *J. Marine Res.* **63**, 683–704.

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

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

S. P. Meyn and R. L. Tweedie (1993), *Markov Chains and Stochastic Stability*, Communications and Control Engineering Series, Springer, London.

A. Michalak and P. Kitanidis (2003), 'A method for enforcing parameter nonnegativity in Bayesian inverse problems with an application to contaminant source identification', *Water Resources Res.* **39**, 1033.

T. Mitchell, B. Buchanan, G. DeJong, T. Dietterich, P. Rosenbloom and A. Waibel (1990), 'Machine learning', *Annual Review of Computer Science* **4**, 417–433.

L. Mohamed, M. Christie and V. Demyanov (2010), 'Comparison of stochastic sampling algorithms for uncertainty quantification', *Soc. Petr. Engrg J.* To appear. http://dx.doi.org/10.2118/119139-PA

K. Mosegaard and A. Tarantola (1995), 'Monte Carlo sampling of solutions to inverse problems', *J. Geophys. Research* **100**, 431–447.

A. Neubauer (2009), 'On enhanced convergence rates for Tikhonov regularization of nonlinear ill-posed problems in Banach spaces', *Inverse Problems* **25**, #065009.

A. Neubauer and H. Pikkarainen (2008), 'Convergence results for the Bayesian inversion theory', *J. Inverse and Ill-Posed Problems* **16**, 601–613.

N. Nichols (2003*a*), Data assimilation: Aims and basic concepts. In *Data Assimilation for the Earth System* (R. Swinbank, V. Shutyaev and W. A. Lahoz, eds), Kluwer Academic, pp. 9–20.

N. Nichols (2003*b*), Treating model error in 3-D and 4-D data assimilation. In *Data Assimilation for the Earth System* (R. Swinbank, V. Shutyaev and W. A. Lahoz, eds), Kluwer Academic, pp. 127–135.

M. Nodet (2005), Mathematical modeling and assimilation of Lagrangian data in oceanography. PhD thesis, University of Nice.

M. Nodet (2006), 'Variational assimilation of Lagrangian data in oceanography', *Inverse Problems* **22**, 245–263.

B. Oksendal (2003), *Stochastic Differential Equations: An Introduction with Applications*, sixth edn, Universitext, Springer.

D. Orrell, L. Smith, J. Barkmeijer and T. Palmer (2001), 'Model error in weather forecasting', *Non. Proc. in Geo.* **8**, 357–371.

A. O'Sullivan and M. Christie (2006*a*), 'Error models for reducing history match bias', *Comput. Geosci.* **10**, 405–405.

A. O'Sullivan and M. Christie (2006*b*), 'Simulation error models for improved reservoir prediction', *Reliability Engineering and System Safety* **91**, 1382–1389.

E. Ott, B. Hunt, I. Szunyogh, A. Zimin, E. Kostelich, M. Corazza, E. Kalnay, D. Patil and J. Yorke (2004), 'A local ensemble Kalman filter for atmospheric data assimilation', *Tellus A* **56**, 273–277.

T. Palmer, F. Doblas-Reyes, A. Weisheimer, G. Shutts, J. Berner and J. Murphy (2009), 'Towards the probabilistic earth-system model', *J. Climate* **70**, 419–435.

H. Pikkarainen (2006), 'State estimation approach to nonstationary inverse problems: Discretization error and filtering problem', *Inverse Problems* **22**, 365–379.

S. Pimentel, K. Haines and N. Nichols (2008*a*), 'The assimilation of satellite derived sea surface temperatures into a diurnal cycle model', *J. Geophys. Research: Oceans* **113**, #C09013.

S. Pimentel, K. Haines and N. Nichols (2008*b*), 'Modelling the diurnal variability of sea surface temperatures', *J. Geophys. Research: Oceans* **113**, #C11004.

J. Ramsay and B. Silverman (2005), *Functional Data Analysis*, Springer.

M. Reznikoff and E. Vanden Eijnden (2005), 'Invariant measures of SPDEs and conditioned diffusions', *CR Acad. Sci. Paris* **340**, 305–308.

D. Richtmyer and K. Morton (1967), *Difference Methods for Initial Value Problems*, Wiley.

G. Roberts and J. Rosenthal (1998), 'Optimal scaling of discrete approximations to Langevin diffusions', *J. Royal Statist. Soc. B* **60**, 255–268.

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

G. Roberts and R. Tweedie (1996), 'Exponential convergence of Langevin distributions and their discrete approximations', *Bernoulli* **2**, 341–363.

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

L. Rudin, S. Osher and E. Fatemi (1992), 'Nonlinear total variation based noise removal algorithms', *Physica D* **60**, 259–268.

H. Rue and L. Held (2005), *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall.

H. Salman, K. Ide and C. Jones (2008), 'Using flow geometry for drifter deployment in Lagrangian data assimilation', *Tellus* **60**, 321–335.

H. Salman, L. Kuznetsov, C. Jones and K. Ide (2006), 'A method for assimilating Lagrangian data into a shallow-water equation ocean model', *Monthly Weather Review* **134**, 1081–1101.

J. M. Sanz-Serna and C. Palencia (1985), 'A general equivalence theorem in the theory of discretization methods', *Math. Comp.* **45**, 143–152.

O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier and F. Lenzen (2009), *Variational Methods in Imaging*, Springer.

C. Schwab and R. Todor (2006), 'Karhunen–Loeve approximation of random fields in domains by generalized fast multipole methods', *J. Comput. Phys.* **217**, 100–122.

Y. Shen, C. Archambeau, D. Cornford and M. Opper (2008*a*), Variational Markov chain Monte Carlo for inference in partially observed nonlinear diffusions. In *Proceedings of the Workshop on Inference and Estimation in Probabilistic*

Time-Series Models (D. Barber, A. T. Cemgil and S. Chiappa, eds), Isaac Newton Institute for Mathematical Sciences, Cambridge, pp. 67–78.

Y. Shen, C. Archambeau, D. Cornford, M. Opper, J. Shawe-Taylor and R. Barillec (2008*b*), 'A comparison of variational and Markov chain Monte Carlo methods for inference in partially observed stochastic dynamic systems', *J. Signal Processing Systems*. In press (published online).

Y. Shen, D. Cornford, C. Archambeau and M. Opper (2010), 'Variational Markov chain Monte Carlo for Bayesian inference in partially observed non-linear diffusions', *Comput. Statist.* Submitted.

A. Smith and G. Roberts (1993), 'Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods', *J. Royal Statist. Soc. B* **55**, 3–23.

T. Snyder, T. Bengtsson, P. Bickel and J. Anderson (2008), 'Obstacles to high-dimensional particle filtering', *Monthly Weather Review* **136**, 4629–4640.

P. Spanos and R. Ghanem (1989), 'Stochastic finite element expansion for random media', *J. Engrg Mech.* **115**, 1035–1053.

P. Spanos and R. Ghanem (2003), *Stochastic Finite Elements: A Spectral Approach*, Dover.

E. Spiller, A. Budhiraja, K. Ide and C. Jones (2008), 'Modified particle filter methods for assimilating Lagrangian data into a point-vortex model', *Physica D* **237**, 1498–1506.

L. Stanton, A. Lawless, N. Nichols and I. Roulstone (2005), 'Variational data assimilation for Hamiltonian problems', *Internat. J. Numer. Methods Fluids* **47**, 1361–1367.

A. Stuart, J. Voss and P. Wiberg (2004), 'Conditional path sampling of SDEs and the Langevin MCMC method', *Comm. Math. Sci* **2**, 685–697.

P. Talagrand and O. Courtier (1987), 'Variational assimilation of meteorological observations with the adjoint vorticity equation I: Theory', *Quart. J. Royal Met. Soc.* **113**, 1311–1328.

A. Tarantola (2005), *Inverse Problem Theory*, SIAM.

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

R. Todor and C. Schwab (2007), 'Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients', *IMA J. Numer. Anal.* **27**, 232–261.

G. Uhlmann (2009), Visibility and invisibility. In *Invited Lectures, Sixth International Congress on Industrial and Applied Mathematics, ICIAM07* (R. Jeltsch and G. Wanner, eds), European Mathematical Society, pp. 381–408.

P. Van Leeuwen (2001), 'An ensemble smoother with error estimates', *Monthly Weather Review* **129**, 709–728.

P. Van Leeuwen (2003), 'A variance minimizing filter for large-scale applications', *Monthly Weather Review* **131**, 2071–2084.

P. Van Leeuwen (2009), 'Particle filtering in geophysical systems', *Monthly Weather Review* **137**, 4089–4114.

G. Vernieres, K. Ide and C. Jones (2010), 'Lagrangian data assimilation, an application to the Gulf of Mexico', *Physica D*. Submitted.

C. Vogel (2002), *Computational Methods for Inverse Problems*, SIAM.

F. Vossepoel and P. Van Leeuwen (2007), 'Parameter estimation using a particle method: Inferring mixing coefficients from sea-level observations', *Monthly Weather Review* **135**, 1006–1020.

M. Vrettas, D. Cornford and Y. Shen (2009), A variational basis function approximation for diffusion processes. In *Proceedings of the 17th European Symposium on Artificial Neural Networks*, D-side publications, Evere, Belgium, pp. 497–502.

G. Wahba (1990), *Spline Models for Observational Data*, SIAM.

L. Watkinson, A. Lawless, N. Nichols and I. Roulstone (2007), 'Weak constraints in four dimensional variational data assimilation', *Meteorologische Zeitschrift* **16**, 767–776.

L. White (1993), 'A study of uniqueness for the initialization problem for Burgers' equation', *J. Math. Anal. Appl.* **172**, 412–431.

D. Williams (1991), *Probability with Martingales*, Cambridge University Press, Cambridge.

M. Wlasak and N. Nichols (1998), Application of variational data assimilation to the Lorenz equations using the adjoint method. In *Numerical Methods for Fluid Dynamics VI*, ICFD, Oxford, pp. 555–562.

M. Wlasak, N. Nichols and I. Roulstone (2006), 'Use of potential vorticity for incremental data assimilation', *Quart. J. Royal Met. Soc.* **132**, 2867–2886.

L. Yu and J. O'Brien (1991), 'Variational estimation of the wind stress drag coefficient and the oceanic eddy viscosity profile', *J. Phys. Ocean.* **21**, 1361–1364.

O. Zeitouni and A. Dembo (1987), 'A maximum *a posteriori* estimator for trajectories of diffusion processes', *Stochastics* **20**, 221–246.

D. Zimmerman, G. de Marsily, C. Gotway, M. Marietta, C. Axness, R. Beauheim, R. Bras, J. Carrera, G. Dagan, P. Davies, D. Gallegos, A. Galli, J. Gomez-Hernandez, P. Grindrod, A. Gutjahr, P. Kitanidis, A. Lavenue, D. McLaughlin, S. Neuman, B. RamaRao, C. Ravenne and Y. Rubin (1998), 'A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow', *Water Resources Res.* **6**, 1373–1413.

E. Zuazua (2005), 'Propagation, observation, control and numerical approximation of waves approximated by finite difference method', *SIAM Review* **47**, 197–243.

D. Zupanski (1997), 'A general weak constraint applicable to operational 4DVAR data assimilation systems', *Monthly Weather Review* **125**, 2274–2292.

M. Zupanski, I. Navon and D. Zupanski (2008), 'The maximum likelihood ensemble filter as a non-differentiable minimization algorithm', *Quart. J. Royal Met. Soc.* **134**, 1039–1050.