SIAM J. SCI. COMPUT.
Vol. 37, No. 6, pp. A2733–A2757
© 2015 Society for Industrial and Applied Mathematics

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

ALGORITHMS FOR KULLBACK–LEIBLER APPROXIMATION OF PROBABILITY MEASURES IN INFINITE DIMENSIONS*

F. J. PINSKI†, G. SIMPSON‡, A. M. STUART§, AND H. WEBER§

Abstract. In this paper we study algorithms to find a Gaussian approximation to a target measure defined on a Hilbert space of functions; the target measure itself is defined via its density with respect to a reference Gaussian measure. We employ the Kullback–Leibler divergence as a distance and find the best Gaussian approximation by minimizing this distance. It then follows that the approximate Gaussian must be equivalent to the Gaussian reference measure, defining a natural function space setting for the underlying calculus of variations problem. We introduce a computational algorithm which is well-adapted to the required minimization, seeking to find the mean as a function, and parameterizing the covariance in two different ways: through low rank perturbations of the reference covariance and through Schrödinger potential perturbations of the inverse reference covariance. Two applications are shown: to a nonlinear inverse problem in elliptic PDEs and to a conditioned diffusion process. These Gaussian approximations also serve to provide a preconditioned proposal distribution for improved preconditioned Crank–Nicolson Monte Carlo–Markov chain sampling of the target distribution. This approach is not only well-adapted to the high dimensional setting, but also behaves well with respect to small observational noise (resp., small temperatures) in the inverse problem (resp., conditioned diffusion).

Key words. MCMC, inverse problems, Gaussian distributions, Kullback–Leibler divergence, relative entropy

AMS subject classifications. 60G15, 34A55, 62G05, 65C05

DOI. 10.1137/14098171X

1. Introduction.

Probability measures on infinite dimensional spaces arise in a variety of applications, including the Bayesian approach to inverse problems [34] and conditioned diffusion processes [19]. Obtaining quantitative information from such problems is computationally intensive, requiring approximation of the infinite dimensional space on which the measures live. We present a computational approach applicable to this context: we demonstrate a methodology for computing the best approximation to the measure, from within a subclass of Gaussians. In addition we show how this best Gaussian approximation may be used to speed up Monte Carlo–Markov chain (MCMC) sampling. The measure of "best" is taken to be the Kullback–Leibler (KL) divergence, or relative entropy, a methodology widely adopted in machine learning applications [5]. In a recent paper [28], KL-approximation by Gaussians was studied using the calculus of variations. The theory from that paper provides the mathematical underpinnings for the algorithms presented here.

1.1. Abstract framework.

Assume we are given a measure μ on the separable Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$ equipped with the Borel σ-algebra, specified by its density with respect to a reference measure μ_0. We wish to find the closest element ν to μ,

*Submitted to the journal's Methods and Algorithms for Scientific Computing section August 11, 2014; accepted for publication (in revised form) April 9, 2015; published electronically November 17, 2015.

http://www.siam.org/journals/sisc/37-6/98171.html

†Department of Physics, University of Cincinnati, Cincinnati, OH 45221 (frank.pinski@uc.edu).

‡Department of Mathematics, Drexel University, Philadelphia, PA 19104 (simpson@math.drexel.edu). This author's work was supported in part by DOE Award DE-SC0002085 and NSF PIRE Grant OISE-0967140.

§Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom (a.m.stuart@warwick.ac.uk, hendrik.weber@warwick.ac.uk). The first author's work was supported by EPSRC, ERC, and ONR. The second author's work was supported by an EPSRC First Grant.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

with respect to KL divergence, from a subset $\mathcal{A}$ of the Gaussian probability measures on $\mathcal{H}$. We assume the reference measure μ_0 is itself a Gaussian $\mu_0 = N(m_0, C_0)$ on $\mathcal{H}$. The measure μ is thus defined by

$$\frac{d\mu}{d\mu_0}(u) = \frac{1}{Z_\mu} \exp\big(-\Phi_\mu(u)\big), \tag{1.1}$$

where we assume that $\Phi_\mu : X \to \mathbb{R}$ is continuous on some Banach space X of full measure with respect to μ_0, and that $\exp(-\Phi_\mu(x))$ is integrable with respect to μ_0. Furthermore, $Z_\mu = \mathbb{E}^{\mu_0} \exp\big(-\Phi_\mu(u)\big)$ ensuring that μ is indeed a *probability* measure. We seek an approximation $\nu = N(m, C)$ of μ which minimizes $D_{\mathrm{KL}}(\nu\|\mu)$, the KL divergence between ν and μ in $\mathcal{A}$. Under these assumptions it is necessarily the case that ν is equivalent[1] to μ_0 (we write $\nu \sim \mu_0$) since otherwise $D_{\mathrm{KL}}(\nu\|\mu) = \infty$. This imposes restrictions on the pair (m, C), and we build these restrictions into our algorithms. Broadly speaking, we will seek to minimize over *all* sufficiently regular functions m, whilst we will parameterize C either through operators of finite rank, or through a function appearing as a potential in an inverse covariance representation.

Once we have found the best Gaussian approximation we will use this to improve upon known MCMC methods. Here, we adopt the perspective of considering only MCMC methods that are well-defined in the infinite dimensional setting, so that they are robust to finite dimensional approximation [11]. The best Gaussian approximation is used to make Gaussian proposals within MCMC which are simple to implement, yet which contain sufficient information about Φ_μ to yield significant reduction in the autocovariance of the resulting Markov chain, when compared with the methods developed in [11].

1.2. Relation to previous work. In addition to the machine learning applications mentioned above [5], approximation with respect to KL divergence has been used in a variety of applications in the physical sciences, including climate science [15], coarse graining for molecular dynamics [22, 32], and data assimilation [3]. Our approach is formulated so as to address infinite dimensional problems.

Improving the efficiency of MCMC algorithms is a topic attracting a great deal of current interest, as many important PDE based inverse problems result in target distributions μ for which Φ_μ is computationally expensive to evaluate. One family of approaches is to adaptively update the proposal distribution during MCMC, [1, 2, 17, 31]. We will show that our best Gaussian approximation can also be used to speed up MCMC and, although we do not interweave the Gaussian optimization with MCMC in this paper, this could be done, resulting in algorithms similar to those in [1, 2, 17, 31]. Indeed, in [1, 2] the authors use KL divergence (relative entropy) in the form $D_{\mathrm{KL}}(\mu||\nu)$ to adapt their proposals, working in the finite dimensional setting. In our work, we formulate our strategy in the infinite dimensional context, and seek to minimize $D_{\mathrm{KL}}(\nu||\mu)$ instead of $D_{\mathrm{KL}}(\mu||\nu)$. Either choice of divergence measure has its own advantages, discussed below.

In [24], the authors develop a stochastic Newton MCMC algorithm, which resembles our improved preconditioned Crank–Nicolson MCMC (pCN-MCMC) Algorithm 5.2 in that it uses Gaussian approximations that are adapted to the problem within the proposal distributions. However, while we seek to find minimizers of KL in an offline computation, the work in [24] makes a quadratic approximation of Φ_μ at each step along the MCMC sequence; in this sense it has similarities with the Riemannian manifold MCMC methods of [16].

[1] Two measures are equivalent if they are mutually absolutely continuous.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

As will become apparent, a serious question is how to characterize, numerically, the covariance operator of the Gaussian measure ν. Recognizing that the covariance operator is compact, with decaying spectrum, it may be well-approximated by a low rank matrix. Low rank approximations are used in [24, 33], and in the earlier work [14]. In [14] the authors discuss how, even in the case where μ is itself Gaussian, there are significant computational challenges motivating the low rank methodology.

Other active areas in MCMC methods for high dimensional problems include the use of polynomial chaos expansions for proposals [25], and local interpolation of Φ_μ to reduce computational costs [10]. For methods which go beyond MCMC, we mention the paper [13] in which the authors present an algorithm for solving the optimal transport PDE relating μ_0 to μ.

1.3. Outline. In section 2, we examine these algorithms in the context of a scalar problem, motivating many of our ideas. The general methodology is introduced in section 3, where we describe the approximation of μ defined via (1.1) by a Gaussian, summarizing the calculus of variations framework which underpins our algorithms. We describe the problem of Gaussian approximations in general, and then consider two specific parameterizations of the covariance which are useful in practice, the first via finite rank perturbation of the covariance of the reference measure μ_0, and the second via a Schrödinger potential shift from the inverse covariance of μ_0. Section 4 describes the structure of the Euler–Lagrange equations for minimization, and recalls the Robbins–Monro algorithm for locating the zeros of functions defined via an expectation. In section 5 we describe how the Gaussian approximation found via KL minimization can be used as the basis for new MCMC methods, well-defined on function space and hence robust to discretization, but also taking into account the change of measure via the best Gaussian approximation. Section 6 contains illustrative numerical results, for a Bayesian inverse problem arising in a model of groundwater flow, and in a conditioned diffusion process, prototypical of problems in molecular dynamics. We conclude in section 7.

2. Scalar example. The main challenges and ideas of this work can be exemplified in a scalar problem, which we examine here as motivation. Consider the measure μ^ε defined via its density with respect to the Lebesgue measure:

$$\mu^\varepsilon(dx) = \frac{1}{Z_\varepsilon} \exp\left(-\varepsilon^{-1} V(x)\right) dx, \quad V : \mathbb{R} \to \mathbb{R}. \tag{2.1}$$

$\varepsilon > 0$ is a small parameter. Furthermore, let the potential V be such that μ^ε is non-Gaussian. As a concrete example, take

$$V(x) = x^4 + \tfrac{1}{2}x^2. \tag{2.2}$$

We now explain our ideas in the context of this simple example, referring to algorithms which are detailed later; additional details are given in Appendix A.

In order to link to the infinite dimensional setting, where Lebesgue measure is not defined and Gaussian measure is used as the reference measure, we write μ^ε via its density with respect to a unit Gaussian $\mu_0 = N(0, 1)$:

$$\frac{d\mu^\varepsilon}{d\mu_0} = \frac{\sqrt{2\pi}}{Z_\varepsilon} \exp\left(-\varepsilon^{-1} V(x) + \tfrac{1}{2}x^2\right).$$

We find the best fit $\nu = N(m, \sigma^2)$, optimizing $D_{\mathrm{KL}}(\nu\|\mu)$ over $m \in \mathbb{R}$ and $\sigma > 0$,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

noting that ν may be written as

$$\frac{d\nu}{d\mu_0} = \frac{\sqrt{2\pi}}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2\sigma^2}(x-m)^2 + \tfrac{1}{2}x^2\right).$$

The change of measure is then

$$\frac{d\mu^\varepsilon}{d\nu} = \frac{\sqrt{2\pi\sigma^2}}{Z_\varepsilon} \exp\left(-\varepsilon^{-1}V(x) + \tfrac{1}{2\sigma^2}(x-m)^2\right). \tag{2.3}$$

For potential (2.2), D_{KL}, defined in full generality in (3.2), can be integrated analytically, yielding

$$D_{\mathrm{KL}}(\nu||\mu^\varepsilon) = \tfrac{1}{2}\varepsilon^{-1}\left(2m^4 + m^2 + 12m^2\sigma^2 + \sigma^2 + 6\sigma^4\right) - \tfrac{1}{2} + \log Z_\varepsilon - \log\sqrt{2\pi\sigma^2}. \tag{2.4}$$

In subsection 2.1 we illustrate an algorithm to find the best Gaussian approximation numerically while subsection 2.2 demonstrates how this minimizer may be used to improve MCMC methods. See [27] for a theoretical analysis of the improved MCMC method for this problem. That analysis sheds light on the application of our methodology more generally.

2.1. Estimation of the minimizer. The Euler–Lagrange equations for (2.4) can then be solved to obtain a minimizer (m, σ) which satisfies $m = 0$ and

$$\sigma^2 = \tfrac{1}{24}\left(\sqrt{1+48\varepsilon} - 1\right) = \varepsilon - 12\varepsilon^2 + \mathrm{O}(\varepsilon^3). \tag{2.5}$$

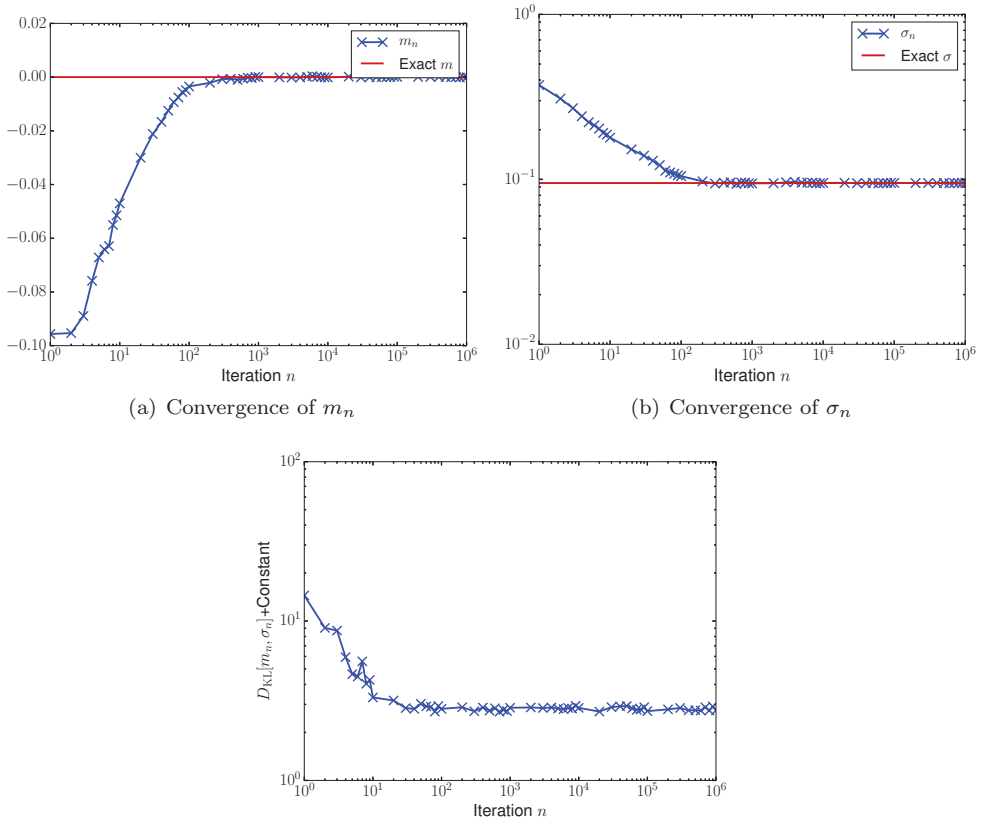
In more complex problems, $D_{\mathrm{KL}}(\nu\|\mu)$ is not analytically tractable and only defined via expectation. In this setting, we rely on the Robbins–Monro algorithm (Algorithm 4.1) to compute a solution of the Euler–Lagrange equations defining minimizers. Figure 1 depicts the convergence of the Robbins–Monro solution towards the desired root at $\varepsilon = 0.01$, $(m, \sigma) \approx (0, 0.0950)$ for our illustrative scalar example. It also shows that $D_{\mathrm{KL}}(\nu\|\mu)$ is reduced.

2.2. Sampling of the target distribution. Having obtained values of m and σ that minimize $D_{\mathrm{KL}}(\nu\|\mu)$, we may use ν to develop an improved MCMC sampling algorithm for the target measure μ^ε. We compare the performance of the standard pCN method of Algorithm 5.1, which uses no information about the best Gaussian fit ν, with the improved pCN Algorithm 5.2, based on knowledge of ν. The improved performance, gauged by acceptance rate and autocovariance, is shown in Figure 2.

All of this is summarized by Figure 3, which shows the three distributions μ^ε, μ_0, and KL optimized ν, together with a histogram generated by samples from the KL-optimized MCMC Algorithm 5.2. Clearly, ν better characterizes μ^ε than μ_0, and this is reflected in the higher acceptance rate and reduced autocovariance. Though this is merely a scalar problem, these ideas are universal. In all of our examples, we have a non-Gaussian distribution we wish to sample from, an uninformed reference measure which gives poor sampling performance, and an optimized Gaussian which better captures the target measure and can be used to improve sampling.

3. Parameterized Gaussian approximations. We start in subsection 3.1 by describing some general features of the KL distance. Then in subsection 3.2 we discuss the case where ν is Gaussian. Subsections 3.3 and 3.4 describe two particular parameterizations of the Gaussian class that we have found useful in practice.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.



(a) Convergence of m_n

(b) Convergence of σ_n

(c) Minimization of D_{KL}

FIG. 1. *Convergence of m_n and σ_n towards the values found via deterministic root finding for the scalar problem with potential* (2.2) *at $\varepsilon = 0.01$. The iterates are generated using Algorithm* 4.1, *Robbins–Monro applied to KL minimization. Also plotted are values of KL divergence along the iteration sequence. The true optimal value is recovered, and KL divergence is reduced. To ensure convergence, m_n is constrained to $[-10, 10]$ and σ_n is constrained to $[10^{-6}, 10^3]$.*

3.1. General setting. Let ν be a measure defined by

$$\frac{d\nu}{d\mu_0}(u) = \frac{1}{Z_\nu} \exp\big(-\Phi_\nu(u)\big), \tag{3.1}$$

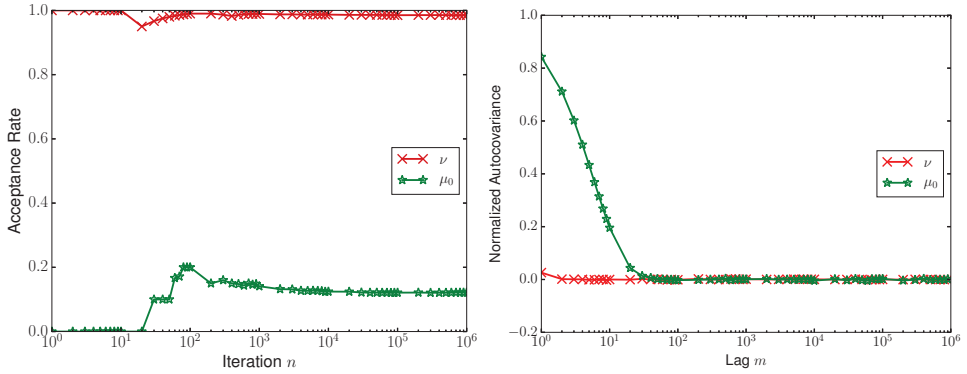
where we assume that $\Phi_\nu : X \to \mathbb{R}$ is continuous on X. We aim to choose the best approximation ν to μ given by (1.1) from within some class of measures; this class will place restrictions on the form of Φ_ν. Our best approximation is found by choosing the free parameters in ν to minimize the KL divergence between μ and ν. This is defined as

$$D_{\mathrm{KL}}(\nu\|\mu) = \int_H \log\left(\frac{d\nu}{d\mu}(u)\right)\nu(du) = \mathbb{E}^\nu \log\left(\frac{d\nu}{d\mu}(u)\right). \tag{3.2}$$

Recall that $D_{\mathrm{KL}}(\cdot\|\cdot)$ is not symmetric in its two arguments and our reason for choosing $D_{\mathrm{KL}}(\nu\|\mu)$ relates to the possibility of capturing multiple modes individually; minimizing $D_{\mathrm{KL}}(\mu\|\nu)$ corresponds to moment matching in the case where $\mathcal{A}$ is the set of all Gaussians [5, 28]. The moment matching form was also employed in the finite dimensional adaptive MCMC method of [1, 2]. An advantage of $D_{\mathrm{KL}}(\nu\|\mu)$ is that it can

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php



(a) Acceptance rate (b) Autocovariance

FIG. 2. *Acceptance rates and autocovariances for sampling from* (2.1) *with potential* (2.2) *at* $\varepsilon = 0.01$. *The curves labeled* ν *correspond to the samples generated using our improved MCMC, Algorithm* 5.2, *which uses the KL optimized* ν *for proposals. The curves labeled* μ_0 *correspond to the samples generated using Algorithm* 5.1, *which relies on* μ_0 *for proposals. Algorithm* 5.2 *shows an order of magnitude improvement over Algorithm* 5.1. *For clarity, only a subset of the data is plotted in the figures.*

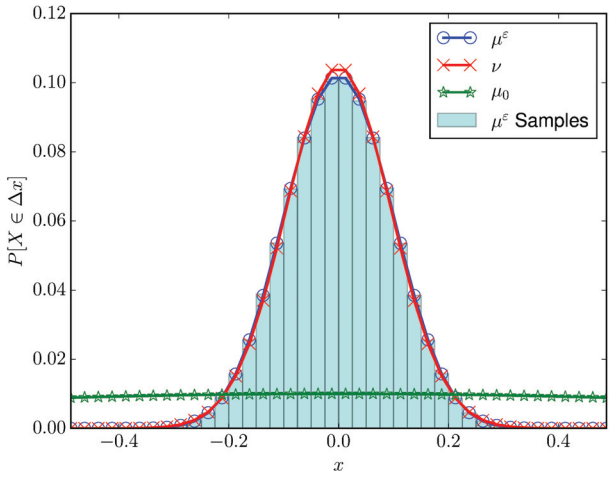


FIG. 3. *Distributions of* μ^ε *(target),* μ_0 *(reference), and* ν *(KL-optimized Gaussian) for the scalar problem with potential* (2.2) *at* $\varepsilon = 0.01$. *Posterior samples have also been plotted, as a histogram. By inspection,* ν *better captures* μ^ε, *leading to improved performance. Bins have width* $\Delta x = 0.025$.

capture detailed information about individual modes of μ, in contrast to $D_{\mathrm{KL}}(\mu\|\nu)$. See [28] for an elementary example with multiple modes.

Provided $\mu_0 \sim \nu$, we can write

$$\frac{d\mu}{d\nu}(u) = \frac{Z_\nu}{Z_\mu}\exp\bigl(-\Delta(u)\bigr), \tag{3.3}$$

where

$$\Delta(u) = \Phi_\mu(u) - \Phi_\nu(u). \tag{3.4}$$

Integrating this identity with respect to ν gives

$$\frac{Z_\mu}{Z_\nu} = \int_H \exp\bigl(-\Delta(u)\bigr)\nu(du) = \mathbb{E}^\nu \exp\bigl(-\Delta(u)\bigr). \tag{3.5}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Combining (3.2) with (3.3) and (3.5), we have

$$D_{\mathrm{KL}}(\nu\|\mu) = \mathbb{E}^{\nu}\Delta(u) + \log\Big(\mathbb{E}^{\nu}\exp\big(-\Delta(u)\big)\Big). \tag{3.6}$$

The computational task in this paper is to minimize (3.6) over the parameters that characterize our class of approximating measures $\mathcal{A}$, which for us will be subsets of Gaussians. These parameters enter Φ_ν and the normalization constant Z_ν. It is noteworthy, however, that the normalization constants Z_μ and Z_ν do not enter this expression for the distance and are, hence, not explicitly needed in our algorithms.

To this end, it is useful to find the Euler–Lagrange equations of (3.6). Imagine that ν is parameterized by θ and that we wish to differentiate $J(\theta) := D_{\mathrm{KL}}(\nu\|\mu)$ with respect to θ. We rewrite $J(\theta)$ as an integral with respect to μ, rather than ν, differentiate under the integral, and then convert back to integrals with respect to ν. From (3.3), we obtain

$$\frac{Z_\nu}{Z_\mu} = \mathbb{E}^{\mu}e^{\Delta}. \tag{3.7}$$

Hence, from (3.3),

$$\frac{d\nu}{d\mu}(u) = \frac{e^{\Delta}}{\mathbb{E}^{\mu}e^{\Delta}}. \tag{3.8}$$

Thus we obtain, from (3.2),

$$J(\theta) = \mathbb{E}^{\mu}\left(\frac{d\nu}{d\mu}(u)\log\left(\frac{d\nu}{d\mu}(u)\right)\right) = \frac{\mathbb{E}^{\mu}\left(e^{\Delta}(\Delta - \log\mathbb{E}^{\mu}e^{\Delta})\right)}{\mathbb{E}^{\mu}e^{\Delta}} \tag{3.9}$$

and

$$J(\theta) = \frac{\mathbb{E}^{\mu}\big(e^{\Delta}\Delta\big)}{\mathbb{E}^{\mu}\big(e^{\Delta}\big)} - \log\mathbb{E}^{\mu}e^{\Delta}.$$

Therefore, with D denoting differentiation with respect to θ,

$$DJ(\theta) = \frac{\mathbb{E}^{\mu}\big(e^{\Delta}\Delta\, D\Delta\big)}{\mathbb{E}^{\mu}\big(e^{\Delta}\big)} - \frac{\mathbb{E}^{\mu}\big(e^{\Delta}\Delta\big)\mathbb{E}^{\mu}\big(e^{\Delta}D\Delta\big)}{\Big(\mathbb{E}^{\mu}\big(e^{\Delta}\big)\Big)^2}.$$

Using (3.8) we may rewrite this as integration with respect to ν and we obtain

$$DJ(\theta) = \mathbb{E}^{\nu}(\Delta\, D\Delta) - (\mathbb{E}^{\nu}\Delta)(\mathbb{E}^{\nu}D\Delta). \tag{3.10}$$

Thus, this derivative is zero if and only if Δ and $D\Delta$ are uncorrelated under ν.

3.2. Gaussian approximations. Recall that the reference measure μ_0 is the Gaussian $N(m_0, C_0)$. We assume that C_0 is a strictly positive-definite trace class operator on $\mathcal{H}$ [7]. We let $\{e_j, \lambda_j^2\}_{j=1}^{\infty}$ denote the eigenfunction/eigenvalue pairs for C_0. Positive (resp., negative) fractional powers of C_0 are thus defined (resp., densely defined) on $\mathcal{H}$ by the spectral theorem and we may define $\mathcal{H}^1 := D(C_0^{-\frac{1}{2}})$, the Cameron–Martin space of measure μ_0. We assume that $m_0 \in \mathcal{H}^1$ so that μ_0 is equivalent to $N(0, C_0)$, by the Cameron–Martin theorem [7]. We seek to approximate μ given in (1.1) by $\nu \in \mathcal{A}$, where $\mathcal{A}$ is a subset of the Gaussian measures on $\mathcal{H}$. It is shown

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

in [28] that this implies that ν is equivalent to μ_0 in the sense of measures and this in turn implies that $\nu = N(m, C)$, where $m \in E$ and

$$\Gamma := C^{-1} - C_0^{-1} \tag{3.11}$$

satisfies

$$\left\| C_0^{\frac{1}{2}} \Gamma C_0^{\frac{1}{2}} \right\|^2_{\mathcal{HS}(\mathcal{H})} < \infty; \tag{3.12}$$

here $\mathcal{HS}(\mathcal{H})$ denotes the space of Hilbert–Schmidt operators on $\mathcal{H}$.

For practical reasons, we do not attempt to recover Γ itself, but instead introduce low dimensional parameterizations. Two such parameterizations are introduced in this paper. In one, we introduce a finite rank operator, associated with a vector $\phi \in \mathbb{R}^n$. In the other, we employ a multiplication operator characterized by a potential function b. In both cases, the mean m is an element of $\mathcal{H}^1$. Thus, minimization will be over either (m, ϕ) or (m, b).

In this Gaussian case the expressions for D_{KL} and its derivative, given by equations (3.6) and (3.10), can be simplified. Defining

$$\Phi_\nu(u) = -\langle u - m, m - m_0 \rangle_{C_0} + \tfrac{1}{2}\langle u - m, \Gamma(u - m)\rangle - \tfrac{1}{2}\|m - m_0\|^2_{C_0}, \tag{3.13}$$

we observe that, assuming $\nu \sim \mu_0$,

$$\frac{d\nu}{d\mu_0} \propto \exp\bigl(-\Phi_\nu(u)\bigr). \tag{3.14}$$

This may be substituted into the definition of Δ in (3.4), and used to calculate J and DJ according to (3.9) and (3.10). However, we may derive alternate expressions as follows. Let $\rho_0 = N(0, C_0)$, the centered version of μ_0, and $\nu_0 = N(0, C)$ the centered version of ν. Then, using the Cameron–Martin formula,

$$Z_\nu = \mathbb{E}^{\mu_0} \exp(-\Phi_\nu) = \mathbb{E}^{\rho_0} \exp(-\Phi_{\nu_0}) = \Bigl(\mathbb{E}^{\nu_0} \exp(\Phi_{\nu_0})\Bigr)^{-1} = Z_{\nu_0}, \tag{3.15}$$

where

$$\Phi_{\nu_0} = \tfrac{1}{2}\langle u, \Gamma u\rangle. \tag{3.16}$$

We also define a reduced Δ function which will play a role in our computations:

$$\Delta_0(u) \equiv \Phi_\mu(u + m) - \tfrac{1}{2}\langle u, \Gamma u\rangle. \tag{3.17}$$

The consequence of these calculations is that, in the Gaussian case, (3.6) is

$$\begin{aligned} D_{\mathrm{KL}}(\nu||\mu) &= \mathbb{E}^\nu \Delta - \log Z_{\nu_0} + \log Z_\mu \\ &= \mathbb{E}^{\nu_0}[\Delta_0] + \tfrac{1}{2}\|m - m_0\|^2_{C_0} + \log \mathbb{E}^{\nu_0} \exp(\tfrac{1}{2}\langle u, \Gamma u\rangle) + \log Z_\mu. \end{aligned} \tag{3.18}$$

Although the normalization constant Z_μ now enters the expression for the objective function, it is irrelevant in the minimization since it does not depend on the unknown parameters in ν. To better see the connection between (3.6) and (3.18), note that

$$\frac{Z_\mu}{Z_{\nu_0}} = \frac{Z_\mu}{Z_\nu} = \frac{\mathbb{E}^{\mu_0} \exp(-\Phi_\mu)}{\mathbb{E}^{\mu_0} \exp(-\Phi_\nu)} = \mathbb{E}^\nu \exp(-\Delta). \tag{3.19}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Working with (3.18), the Euler–Lagrange equations to be solved are

$$D_m J(m,\theta) = \mathbb{E}^{\nu_0} D_u \Phi_\mu(u+m) + C_0^{-1}(m-m_0) = 0, \tag{3.20a}$$

$$D_\theta J(m,\theta) = \mathbb{E}^{\nu_0}(\Delta_0\, D_\theta \Delta_0) - (\mathbb{E}^{\nu_0}\Delta_0)(\mathbb{E}^{\nu_0} D_\theta \Delta_0) = 0. \tag{3.20b}$$

Here, θ is any of the parameters that define the covariance operator C of the Gaussian ν. Equation (3.20a) is obtained by direct differentiation of (3.18), while (3.20b) is obtained in the same way as (3.10). These expressions are simpler for computations for two reasons. First, for the variation in the mean, we do not need the full covariance expression of (3.10). Second, Δ_0 has fewer terms to compute.

3.3. Finite rank parameterization. Let P denote orthogonal projection onto $\mathcal{H}_K := \operatorname{span}\{e_1, \dots, e_K\}$, the span of the first K eigenvectors of C_0, and define $Q = I - P$. We then parameterize the covariance C of ν in the form

$$C^{-1} = \big(QC_0Q\big)^{-1} + \chi, \quad \chi = \sum_{i,j\le K} \gamma_{ij} e_i \otimes e_j. \tag{3.21}$$

In words, C^{-1} is given by the inverse covariance C_0^{-1} of μ_0 on $Q\mathcal{H}$, and is given by χ on $P\mathcal{H}$. Because χ is necessarily symmetric it is essentially parameterized by a vector ϕ of dimension $n = \frac{1}{2}K(K+1)$. We minimize $J(m,\phi) := D_{\mathrm{KL}}(\nu\|\mu)$ over $(m,\phi) \in \mathcal{H}^1 \times \mathbb{R}^n$. This is a well-defined minimization problem as demonstrated in Example 3.7 of [28] in the sense that minimizing sequences have weakly convergent subsequences in the admissible set. Minimizers need not be unique, and we should not expect them to be, as multimodality is to be expected, in general, for measures μ defined by (1.1). Problem-specific information may also suggest better directions for the finite rank operator, but we do not pursue this here.

3.4. Schrödinger parameterization. In this subsection we assume that $\mathcal{H}$ comprises a Hilbert space of functions defined on a bounded open subset of $\mathbb{R}^d$. We then seek Γ given by (3.11) in the form of a multiplication operator so that $(\Gamma u)(x) = b(x)u(x)$. While minimization over the pair (m,Γ), with $m \in \mathcal{H}^1$ and Γ in the space of linear operators satisfying (3.12), is well-posed [28], minimizing sequences $\{m_k, \Gamma_k\}_{k\ge1}$ with $(\Gamma_k u)(x) = b_k(x)u(x)$ can behave very poorly with respect to the sequence $\{b_k\}_{k\ge1}$. For this reason we regularize the minimization problem and seek to minimize

$$J_\alpha(m,b) = J(m,b) + \tfrac{\alpha}{2}\|b\|_r^2,$$

where $J(m,b) := D_{\mathrm{KL}}(\nu\|\mu)$ and $\|\cdot\|_r$ denotes the Sobolev space H^r of functions on $\mathbb{R}^d$ with r square integrable derivatives, with boundary conditions chosen appropriately for the problem at hand. The minimization of $J_\alpha(m,b)$ over $(m,b) \in \mathcal{H} \times H^r$ is well-defined; see section 3.3 of [28].

4. Robbins–Monro algorithm. In order to minimize $D_{\mathrm{KL}}(\nu\|\mu)$ we will use the Robbins–Monro algorithm [4, 23, 26, 30]. In its most general form this algorithm calculates zeros of functions defined via an expectation. We apply it to the Euler–Lagrange equations to find critical points of a nonnegative objective function, defined

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

via an expectation. This leads to a form of gradient descent in which we seek to integrate the equations

$$\dot{m} = -D_m D_{\mathrm{KL}}, \quad \dot{\theta} = -D_\theta D_{\mathrm{KL}},$$

until they have reached a critical point. This requires two approximations. First, as (3.20) involve expectations, the right-hand sides of these differential equations are evaluated only approximately, by sampling. Second, a time discretization must be introduced. The key idea underlying the algorithm is that, provided the step length of the algorithm is sent to zero judiciously, the sampling error averages out and is diminished as the step length goes to zero. Minimization of KL by Robbins–Monro was also performed in [1, 2] for $D_{\mathrm{KL}}(\mu||\nu)$, in the case of finite dimensional problems.

4.1. Background on Robbins–Monro. In this section we review some of the structure in the Euler–Lagrange equations for the desired minimization of $D_{\mathrm{KL}}(\nu\|\mu)$. We then describe the particular variant of the Robbins–Monro algorithm that we use in practice. Suppose we have a parameterized distribution, ν_θ, from which we can generate samples, and we seek a value θ for which

$$f(\theta) \equiv \mathbb{E}^{\nu_\theta}[Y] = 0, \quad Y \sim \nu_\theta. \tag{4.1}$$

Then an estimate of the zero, $\theta_\star$, can be obtained via the recursion

$$\theta_{n+1} = \theta_n - a_n \sum_{m=1}^{M} \tfrac{1}{M} Y_m^{(n)}, \quad Y_m^{(n)} \sim \nu_{\theta_n}, \quad \text{i.i.d.} \tag{4.2}$$

(where i.i.d. is independently and identically distributed). Note that the two approximations alluded to above are included in this procedure: sampling and (Euler) time discretization. The methodology may be adapted to seek solutions to

$$f(\theta) \equiv \mathbb{E}^{\nu}[F(Y;\theta)] = 0, \quad Y \sim \nu, \tag{4.3}$$

where ν is a given, fixed, distribution independent of the parameter θ. (This setup arises, for example, in (3.20a), where ν_0 is fixed and the parameter in question is m.) Letting $Z = F(Y;\theta)$, this induces a distribution $\eta_\theta(dz) = \nu(F^{-1}(dz;\theta))$, where the preimage is with respect to the Y argument. Then $f(\theta) = \mathbb{E}^{\eta_\theta}[Z]$ with $Z \sim \eta_\theta$, and this now has the form of (4.1). As suggested in the extensive Robbins–Monro literature, we take the step sequence to satisfy

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty. \tag{4.4}$$

A suitable choice of $\{a_n\}$ is thus $a_n = a_0 n^{-\gamma}$, $\gamma \in (1/2, 1]$. The smaller the value of γ, the more "large" steps will be taken, helping the algorithm to explore the configuration space. On the other hand, once the sequence is near the root, the smaller γ is, the larger the Markov chain variance will be. In addition to the choice of the sequence a_n, (4.1) introduces an additional parameter, M, the number of samples to be generated per iteration. See [4, 8] and references therein for commentary on sample size.

The conditions needed to ensure convergence, and what kind of convergence, have been relaxed significantly through the years. In their original paper, Robbins and Monro assumed that $Y \sim \mu_\theta$ were almost surely uniformly bounded by a constant

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

independent of θ. If they also assumed that $f(\theta)$ was monotonic and $f'(\theta_\star) > 0$, they could obtain convergence in L^2. With somewhat weaker assumptions, but still requiring that the zero be simple, Blum developed convergence with probability one [6]. All of this was subsequently generalized to the arbitrary finite dimensional case; see [4, 23, 26].

As will be relevant to this work, there is the question of the applicability to the infinite dimensional case when we seek, for instance, a mean function in a separable Hilbert space. This has also been investigated; see [12, 35] along with references mentioned in the preface of [23]. In this work, we do not verify that our problems satisfy convergence criteria; this is a topic for future investigation.

A variation on the algorithm that is commonly applied is the enforcement of constraints which ensure $\{\theta_n\}$ remain in some bounded set; see [23] for an extensive discussion. We replace (4.2) by

$$\theta_{n+1} = \Pi_D \left[\theta_n - a_n \sum_{m=1}^{M} \tfrac{1}{M} Y_m^{(n)} \right], \quad Y_m^{(n)} \sim \nu_{\theta_n}, \quad \text{i.i.d.}, \tag{4.5}$$

where D is a bounded set, and $\Pi_D(x)$ computes the point in D nearest to x. This is important in our work, as the parameters must induce covariance operators. They must be positive definite, symmetric, and trace class. Our method automatically produces symmetric trace-class operators, but the positivity has to be enforced by a projection.

The choice of the set D can be set either through a priori information as we do here, or determined adaptively. In [9, 35], each time the iterate attempts to leave the current constraint set D, it is returned to a point within D, and the constraint set is expanded. It can be shown that, almost surely, the constraint is expanded only a finite number of times.

4.2. Robbins–Monro applied to KL. We seek minimizers of D_{KL} as stationary points of the associated Euler–Lagrange equations, (3.20). Before applying Robbins–Monro to this problem, we observe that we are free to precondition the Euler–Lagrange equations. In particular, we can apply bounded, positive, invertible operators so that the preconditioned gradient will lie in the same function space as the parameter; this makes the iteration scheme well-posed. For (3.20a), we have found premultiplying by C_0 to be sufficient. For (3.20b), the operator will be problem specific, depending on how θ parameterizes C, and also if there is a regularization. We denote the preconditioner for the second equation by B_θ. Thus, the preconditioned Euler–Lagrange equations are

$$0 = C_0 \mathbb{E}^{\nu_0} D_u \Phi_\mu(u+m) + (m - m_0), \tag{4.6a}$$

$$0 = B_\theta \left[\mathbb{E}^{\nu_0}(\Delta_0 \, D_\theta \Delta_0) - (\mathbb{E}^{\nu_0} \Delta_0)(\mathbb{E}^{\nu_0} D_\theta \Delta_0) \right]. \tag{4.6b}$$

We must also ensure that m and θ correspond to a well-defined Gaussian; C must be a covariance operator. Consequently, the Robbins–Monro iteration scheme is the following.

ALGORITHM 4.1.

1. *Set $n = 0$. Pick m_0 and θ_0 in the admissible set, and choose a sequence $\{a_n\}$ satisfying* (4.4).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

2. *Update m_n and θ_n according to*

(4.7a)
$$m_{n+1} = \Pi_m \left[m_n - a_n \left\{ C_0 \left(\sum_{\ell=1}^{M} \tfrac{1}{M} \cdot D_u \Phi_\mu(u_\ell) \right) + m_n - m_0 \right\} \right],$$

(4.7b)
$$\theta_{n+1} = \Pi_\theta \left[\theta_n - a_n B_\theta \left\{ \sum_{\ell=1}^{M} \tfrac{1}{M} \cdot \Delta_0(u_\ell) D_\theta \Delta_0(u_\ell) - \left(\sum_{\ell=1}^{M} \tfrac{1}{M} \cdot \Delta_0(u_\ell) \right) \left(\sum_{\ell=1}^{M} \tfrac{1}{M} \cdot D_\theta \Delta_0(u_\ell) \right) \right\} \right].$$

3. $n \to n+1$ *and return to* 2.

Typically, we have some a priori knowledge of the magnitude of the mean. For instance, $m \in H^1([0,1];\mathbb{R}^1)$ may correspond to a mean path, joining two fixed endpoints, and we know it to be confined to some interval $[\underline{m}, \overline{m}]$. In this case we choose

$$\Pi_m(f)(t) = \min\{\max\{f(t), \underline{m}\}, \overline{m}\}, \quad 0 < t < 1. \tag{4.8}$$

For Π_θ, it is necessary to compute part of the spectrum of the operator that θ induces, check that it is positive, and, if it is not, project the value to something satisfactory. In the case of the finite rank operators discussed in section 3.3, the matrix $\boldsymbol{\gamma}$ must be positive. One way of handing this, for symmetric real matrices, is to make the following choice:

$$\Pi_\theta(A) = X \operatorname{diag}\{\min\{\max\{\boldsymbol{\lambda}, \underline{\lambda}\}, \overline{\lambda}\}\} X^T, \tag{4.9}$$

where $A = X \operatorname{diag}\{\boldsymbol{\lambda}\} X^T$ is the spectral decomposition, and $\underline{\lambda}$ and $\overline{\lambda}$ are constants chosen a priori. It can be shown that this projection gives the closest, with respect to the Frobenius norm, symmetric matrix with spectrum constrained to $[\underline{\lambda}, \overline{\lambda}]$ [20].[2]

5. Improved MCMC sampling. The idea of the Metropolis–Hastings variant of MCMC is to create an ergodic Markov chain which is reversible, in the sense of Markov processes, with respect to the measure of interest; in particular, the measure of interest is invariant under the Markov chain. In our case we are interested in the measure μ given by (1.1). Since this measure is defined on an infinite dimensional space it is advisable to use MCMC methods which are well-defined in the infinite dimensional setting, thereby ensuring that the resulting methods have mixing rates independent of the dimension of the finite dimensional approximation space. This philosophy is explained in the paper [11]. The pCN algorithm is perhaps the simplest MCMC method for (1.1) meeting these requirements. It has the following form.

ALGORITHM 5.1.

Define $a_\mu(u,v) := \min\{1, \exp\big(\Phi_\mu(u) - \Phi_\mu(v)\big)\}$.

1. *Set* $k = 0$ *and Pick* $u^{(0)}$.
2. $v^{(k)} = m_0 + \sqrt{(1-\beta^2)}(u^{(k)} - m_0) + \beta \xi^{(k)}, \quad \xi^{(k)} \sim N(0, C_0)$.
3. *Set* $u^{(k+1)} = v^{(k)}$ *with probability* $a_\mu(u^{(k)}, v^{(k)})$.

[2]Recall that the Frobenius norm is the finite dimensional analog of the Hilbert–Schmidt norm.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

4. *Set* $u^{(k+1)} = u^{(k)}$ *otherwise.*
5. $k \to k+1$ *and return to* 2.

This algorithm has a spectral gap which is independent of the dimension of the discretization space under quite general assumptions on Φ_μ [18]. However, it can still behave poorly if Φ_μ, or its gradients, are large. This leads to poor acceptance probabilities unless β is chosen very small so that proposed moves are localized; either way, the correlation decay is slow and mixing is poor in such situations. This problem arises because the underlying Gaussian μ_0 used in the algorithm construction is far from the target measure μ. This suggests a potential resolution in cases where we have a good Gaussian approximation to μ, such as the measure ν. Rather than basing the pCN approximation on (1.1), we base it on (3.3); this leads to the following algorithm.

ALGORITHM 5.2.

Define $a_\nu(u,v) := \min\{1, \exp\big(\Delta(u) - \Delta(v)\big)\}$.

1. *Set* $k = 0$ *and Pick* $u^{(0)}$.
2. $v^{(k)} = m + \sqrt{(1-\beta^2)}(u^{(k)} - m) + \beta\xi^{(k)}, \quad \xi^{(k)} \sim N(0, C)$.
3. *Set* $u^{(k+1)} = v^{(k)}$ *with probability* $a_\nu(u^{(k)}, v^{(k)})$.
4. *Set* $u^{(k+1)} = u^{(k)}$ *otherwise.*
5. $k \to k+1$ *and return to* 2.

We expect Δ to be smaller than Φ, at least in regions of high μ probability. This suggests that, for given β, Algorithm 5.2 will have better acceptance probability than Algorithm 5.1, leading to more rapid sampling. We show in what follows that this is indeed the case.

6. Numerical results. In this section we describe our numerical results. These concern both a solution of the relevant minimization problem, to find the best Gaussian approximation from within a given class using Algorithm 4.1 applied to the two parameterizations given in subsections 3.3 and 3.4, together with results illustrating the new pCN Algorithm 5.2 which employs the best Gaussian approximation within MCMC. We consider two model problems: a Bayesian inverse problem arising in PDEs, and a conditioned diffusion problem motivated by molecular dynamics. Some details on the path generation algorithms used in these two problems are given in Appendix B.

6.1. Bayesian inverse problem. We consider an inverse problem arising in groundwater flow. The forward problem is modeled by the Darcy constitutive model for porous medium flow. The objective is to find $p \in V := H^1$ given by the equation

$$-\nabla \cdot \big(\exp(u)\nabla p\big) = 0, \quad x \in D, \tag{6.1a}$$

$$p = g, \quad x \in \partial D. \tag{6.1b}$$

The inverse problem is to find $u \in X = L^\infty(D)$ given noisy observations

$$y_j = \ell_j(p) + \eta_j,$$

where $\ell_j \in V^*$, the space of continuous linear functionals on V. This corresponds to determining the log permeability from measurements of the hydraulic head (height of the water table). Letting $\mathcal{G}(u) = \ell(p(\cdot;u))$, the solution operator of (6.1) is composed with the vector of linear functionals $\ell = (\ell_j)^T$. We then write, in vector form,

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

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

We assume that $\eta \sim N(0,\Sigma)$ and place a Gaussian prior $N(m_0, C_0)$ on u. Then the Bayesian inverse problem has the form (1.1) where

$$\Phi(u) := \frac{1}{2}\left\|\Sigma^{-\frac{1}{2}}\big(y - \mathcal{G}(u)\big)\right\|^2.$$

We consider this problem in dimension one, with $\Sigma = \gamma^2 I$, and employing pointwise observation at points x_j as the linear functionals ℓ_j. As prior we take the Gaussian $\mu_0 = N(0, C_0)$ with

$$C_0 = \delta\left(-\frac{d^2}{dx^2}\right)^{-1},$$

restricted to the subspace of $L^2(0,1)$ of periodic mean zero functions. For this problem, the eigenvalues of the covariance operator decay like j^{-2}. In one dimension we may solve the forward problem (6.1) on $D = (0,1)$, with $p(0) = p^-$ and $p(1) = p^+$ explicitly to obtain

$$p(x;u) = (p^+ - p^-)\frac{J_x(u)}{J_1(u)} + p^-, \quad J_x(u) \equiv \int_0^x \exp(-u(z))dz, \tag{6.2}$$

and

$$\Phi(u) = \frac{1}{2\gamma^2}\sum_{j=1}^{\ell} |p(x_j;u) - y_j|^2. \tag{6.3}$$

Following the methodology of [21], to compute $D_u\Phi(u)$ we must solve the adjoint problem for q:

$$-\frac{d}{dx}\left(\exp(u)\frac{dq}{dx}\right) = -\frac{1}{\gamma^2}\sum_{j=1}^{\ell}(p(x_j;u) - y_j)\delta_{x_j}, \quad q(0) = q(1) = 0. \tag{6.4}$$

Again, we can write the solution explicitly via quadrature:

$$\begin{aligned} q(x;u) &= K_x(u) - \frac{K_1(u)J_x(u)}{J_1(u)}, \\ K_x(u) &\equiv \sum_{j=1}^{\ell}\frac{p(x_j;u) - y_j}{\gamma^2}\int_0^x \exp(-u(z))H(z - x_j)dz. \end{aligned} \tag{6.5}$$

Using (6.2) and (6.5),

$$D_u\Phi(u) = \exp(u)\frac{dp(x;u)}{dx}\frac{dq(x;u)}{dx}. \tag{6.6}$$

For this application we use a finite rank approximation of the covariance of the approximating measure ν, as explained in subsection 3.3. In computing with the finite rank matrix (3.21), it is useful, for good convergence, to work with $B = \boldsymbol{\gamma}^{-1/2}$. The preconditioned derivatives, (4.6), also require $D_B\Delta_0$, where Δ_0 is given by (3.17). To characterize this term, if $v = \sum_i v_i e_i$, we let $\mathbf{v} = (v_1, \dots v_N)^T$ be the first N coefficients. Then for the finite rank approximation,

$$\Phi_{\nu_0}(v) = \frac{1}{2}\left\langle v, (C^{-1} - C_0^{-1})v\right\rangle = \frac{1}{2}\mathbf{v}^T(\boldsymbol{\gamma} - \mathrm{diag}(\lambda_1^{-1}, \dots \lambda_N^{-1}))\mathbf{v}. \tag{6.7}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Then using our parameterization with respect to the matrix B,

$$D_B\Delta_0(v) = D_B(\Phi(m+v) - \Phi_{\nu_0}(v)) = \frac{1}{2}\left[B^{-1}\mathbf{v}(B^{-2}\mathbf{v})^T + B^{-2}\mathbf{v}(B^{-1}\mathbf{v})^T\right]. \tag{6.8}$$

As a preconditioner for (4.6b) we found that it was sufficient to multiply by λ_N.

We solve this problem with ranks $K = 2$, 4, 6, first minimizing D_{KL}, and then running the pCN Algorithm 5.2 to sample from μ_y. The common parameters are

- $\gamma = 0.1$, $\delta = 1$, $p^- = 0$, and $p^+ = 2$;
- there are 2^7 uniformly spaced grid points in $[0, 1)$;
- (6.2) and (6.5) are solved via trapezoidal rule quadrature;
- the true value of $u(x) = 2\sin(2\pi x)$;
- the dimension of the data is four, with samples at $x = 0.2, 0.4, 0.6, 0.8$;
- $m_0 = 0$ and $B_0 = \mathrm{diag}(\lambda_n)$, $n \leq \mathrm{rank}$;
- $\int \dot{m}^2$ is estimated spectrally;
- 10^5 iterations of the Robbins–Monro algorithm are performed with 10^2 samples per iteration;
- $a_0 = .1$ and $a_n = a_0 n^{-3/5}$;
- the eigenvalues of σ are constrained to the interval $[10^{-4}, 10^0]$ and the mean is constrained to $[-5, 5]$;
- pCN Algorithms 5.1 and 5.2 are implemented with $\beta = 0.6$, and 10^6 iterations.

The results of the D_{KL} optimization phase of the problem, using the Robbins–Monro Algorithm 4.1, appear in Figure 4. This figure shows the convergence of m_n in the rank 2 case; the convergence of the eigenvalues of B for ranks 2, 4, and 6; and the minimization of D_{KL}. We only present the convergence of the mean in the rank 2 case, as the others are quite similar. At the termination of the Robbins–Monro step, the B_n matrices are

$$B_n = \begin{pmatrix} 0.0857 & 0.00632 \\ - & 0.105 \end{pmatrix}, \tag{6.9}$$

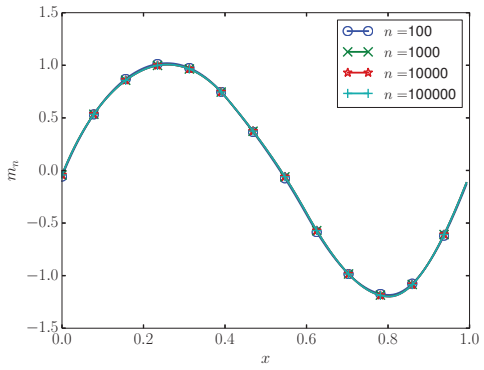
$$B_n = \begin{pmatrix} 0.0864 & 0.00500 & -0.00791 & -0.00485 \\ - & 0.106 & 0.00449 & -0.00136 \\ - & - & 0.0699 & -0.000465 \\ - & - & - & 0.0739 \end{pmatrix}, \tag{6.10}$$

$$B_n = \begin{pmatrix} 0.0870 & 0.00518 & -0.00782 & -0.00500 & -0.00179 & -0.00142 \\ - & 0.106 & 0.00446 & -0.00135 & 0.00107 & 0.00166 \\ - & - & 0.0701 & -0.000453 & -0.00244 & 9.81 \times 10^{-5} \\ - & - & - & 0.0740 & -0.00160 & 0.00120 \\ - & - & - & - & 0.0519 & -0.00134 \\ - & - & - & - & - & 0.0523 \end{pmatrix}. \tag{6.11}$$

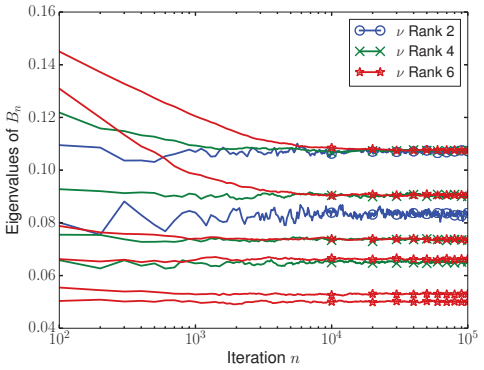
Note there is consistency as the rank increases, and this is reflected in the eigenvalues of the B_n shown in Figure 4. As in the case of the scalar problem, more iterations of Robbins–Monro are computed than are needed to ensure convergence.

The posterior sampling, by means of Algorithms 5.1 and 5.2, is described in Figure 5. There is good posterior agreement in the means and variances in all cases, and the low rank priors provide not just good means but also variances. This is reflected in the high acceptance rates and low auto covariances; there is approximately an order of magnitude in improvement in using Algorithm 5.2, which is informed by the best Gaussian approximation, and Algorithm 5.1, which is not.

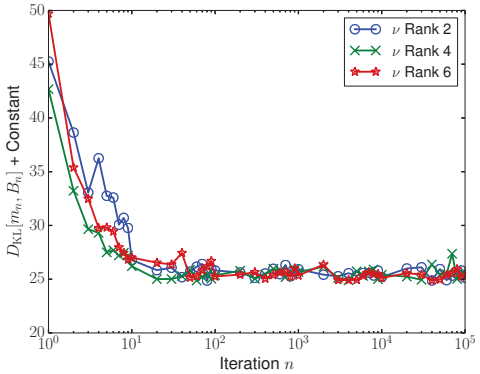
Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.



(a) $m_n(x)$ for rank 2 at particular iterations



(b) Convergence of the eigenvalues of B_n



(c) Minimization of D_{KL}

FIG. 4. *Convergence of the Robbins–Monro Algorithm* 4.1 *applied to the Bayesian inverse problem.* (a) *shows the convergence of* m_n *in the case of rank* 2*, while* (b) *shows the convergence of the eigenvalues of* B_n *for ranks* 2*,* 4 *and* 6*.* (c) *shows the minimization of* D_{KL}*. The observational noise is* $\gamma = 0.1$*. The figures indicate that rank* 2 *has converged after* 10^2 *iterations; rank* 4 *has converged after* 10^3 *iterations; and rank* 6 *has converged after* 10^4 *iterations.*

However, notice in Figure 5 that the posterior, even when ± one standard deviation is included, does not capture the truth. The results are more favorable when we consider the pressure field, and this hints at the origin of the disagreement. The values at $x = 0.2$ and 0.4, and to a lesser extent at 0.6, are dominated by the noise. Our posterior estimates reflect the limitations of what we are able to predict given our assumptions. If we repeat the experiment with smaller observational noise, $\gamma = 0.01$ instead of 0.1, we see better agreement, and also variation in performance with respect to approximations of different ranks. These results appear in Figure 6. In this smaller noise case, there is a two order magnitude improvement in performance.

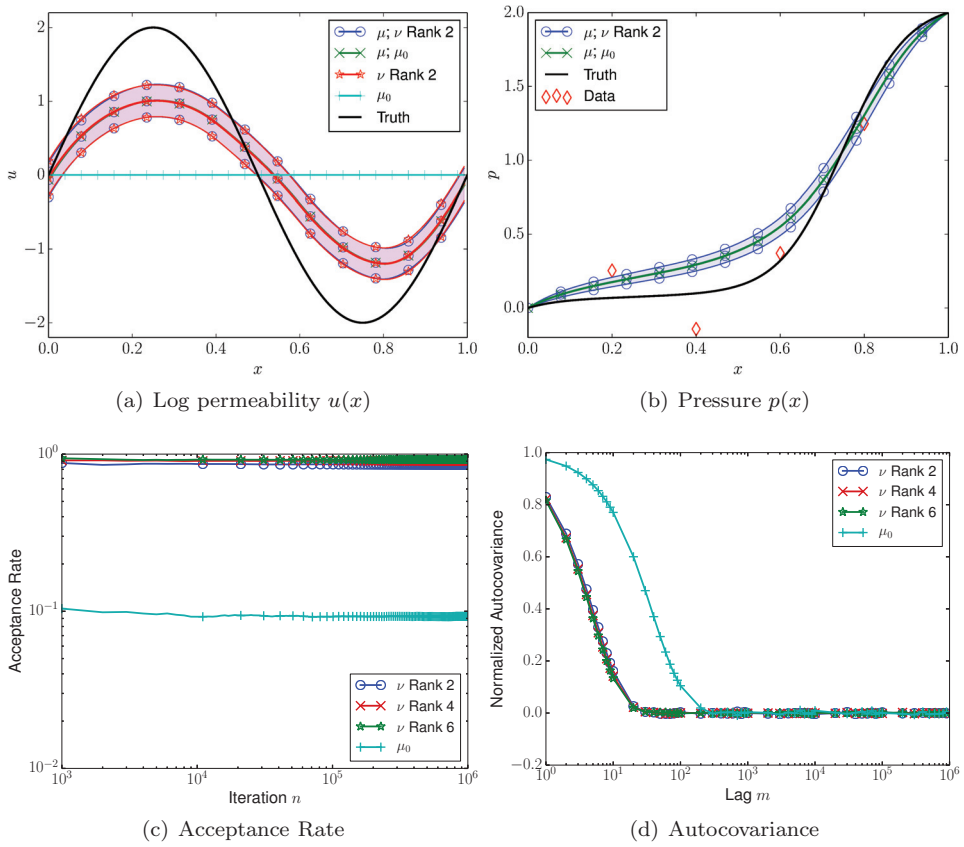
6.2. Conditioned diffusion process. Next, we consider measure μ given by (1.1) in the case where μ_0 is a unit Brownian bridge connecting 0 to 1 on the interval $(0, 1)$, and

$$\Phi = \frac{1}{4\varepsilon^2} \int_0^1 \left(1 - u(t)^2\right)^2 dt,$$

a double well potential. This also has an interpretation as a conditioned diffusion [29]. Note that $m_0 = t$ and $C_0^{-1} = -\frac{1}{2}\frac{d^2}{dt^2}$ with $D(C_0^{-1}) = H^2(I) \cap H_0^1(I)$ with $I = (0, 1)$.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.



(a) Log permeability $u(x)$ (b) Pressure $p(x)$

(c) Acceptance Rate (d) Autocovariance

FIG. 5. *Behavior of MCMC Algorithms* 5.1 *and* 5.2 *for the Bayesian inverse problem with observational noise* $\gamma = 0.1$. *The true posterior distribution,* μ, *is sampled using* μ_0 *(Algorithm* 5.1*) and* ν, *with ranks* 2, 4 *and* 6 *(Algorithm* 5.2*). The resulting posterior approximations are labeled* μ; μ_0 *(Algorithm* 5.1*) and* μ; ν *rank* 2, *(Algorithm* 5.2*). The notation* μ_0 *and* ν *rank* K *is used for the prior and best Gaussian approximations of the corresponding rank. The distributions of* $u(x)$, *in* (a), *for the optimized* ν *rank* 2 *and the posterior* μ *overlap, but are still far from the truth. The results for ranks* 4 *and* 6 *are similar.* (c) *and* (d) *compare the performance of Algorithm* 5.2 *when using* ν *rank* K *for the proposal, with* $K = 2, 4,$ *and* 6, *against Algorithm* 5.1. ν *rank* 2 *gives an order of magnitude improvement in posterior sampling over* μ_0. *There is not significant improvement when using* ν *ranks* 4 *and* 6 *over using rank* 2. *Shaded regions enclose* $\pm$ *one standard deviation.*

We seek the approximating measure ν in the form $N(m(t), C)$ with (m, B) to be varied, where

$$C^{-1} = C_0^{-1} + \tfrac{1}{2\varepsilon^2} B$$

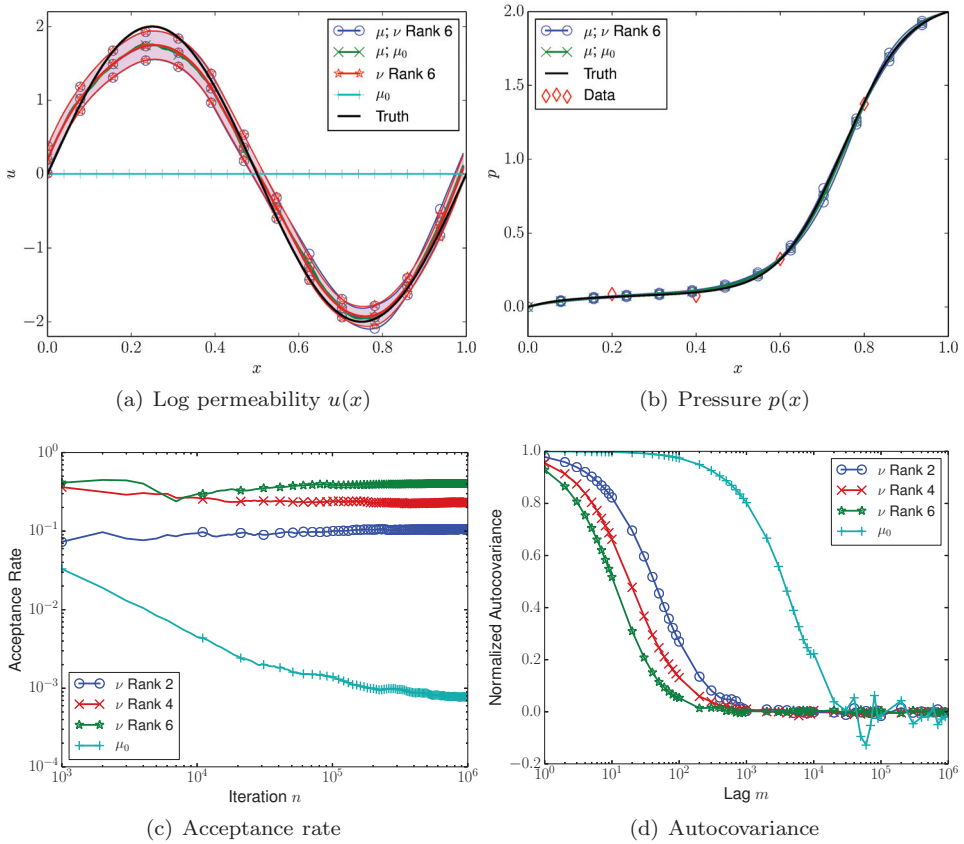
and B is either constant, $B \in \mathbb{R}$, or $B : I \to \mathbb{R}$ is a function viewed as a multiplication operator. Here, the eigenvalues of the covariance operator decay like j^{-2}.

We examine both cases of this problem, performing the optimization, followed by pCN sampling. The results were then compared against the uninformed prior, $\mu_0 = N(m_0, C_0)$. For the constant B case, no preconditioning on B was performed, and the initial guess was $B = 1$. For $B = B(t)$, a Tikhonov–Phillips regularization was introduced,

$$D_{\mathrm{KL}}^{\alpha} = D_{\mathrm{KL}} + \frac{\alpha}{2} \int \dot{B}^2 dt, \quad \alpha = 10^{-2}. \tag{6.12}$$

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.



(a) Log permeability $u(x)$ (b) Pressure $p(x)$

(c) Acceptance rate (d) Autocovariance

FIG. 6. *Behavior of MCMC Algorithms* 5.1 *and* 5.2 *for the Bayesian inverse problem with observational noise* $\gamma = 0.01$. *Notation as in* 5. *The distribution of* $u(x)$, *shown in* (a), *for both the optimized rank* 6 ν, *and the posterior* μ *overlap, and are close to the truth. Unlike the case of* $\gamma = 0.1$, (c) *and* (d) *show improvement in using* ν *rank* 6 *within Algorithm* 5.2, *over ranks* 2 *and* 4. *However, all three cases of Algorithm* 5.2 *are at least two orders of magnitude better than Algorithm* 5.1, *which uses only* μ_0. *Shaded regions enclose* $\pm$ *one standard deviation.*

For computing the gradients (4.6) and estimating D_{KL},

$$D_m\Phi(v+m) = \tfrac{1}{2\varepsilon^2}(v+m)[(v+m)^2 - 1], \tag{6.13a}$$

$$D_B\Phi_{\nu_0}(v) = \begin{cases} \frac{1}{4\varepsilon^2}\int_0^1 v^2 dt, & B \text{ constant}, \\ \frac{1}{4\varepsilon^2}v^2, & B(t). \end{cases} \tag{6.13b}$$

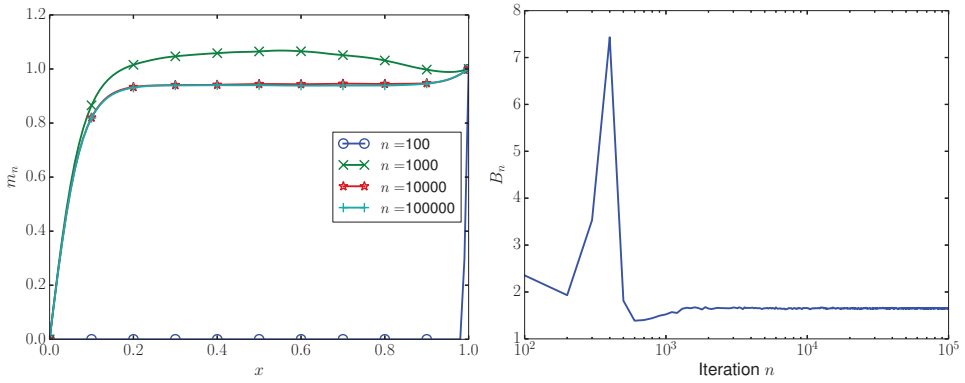
No preconditioning is applied for (6.13b) in the case that B is a constant, while in the case that $B(t)$ is variable, the preconditioned gradient in B is

$$\left\{-\alpha\tfrac{d^2}{dt^2}\right\}^{-1}(\mathbb{E}^{\nu_0}(\Delta_0 D_\theta \Delta_0) - \mathbb{E}^{\nu_0}(\Delta_0)\mathbb{E}^{\nu_0}(D_\theta\Delta_0)) + B.$$

Because of the regularization, we must invert $-d^2/dt^2$, requiring the specification of boundary conditions. By a symmetry argument, we specify the Neumann boundary condition, $B'(0) = 0$. At the other endpoint, we specify the Dirichlet condition $B(1) = V''(1) = 2$, a "far field" approximation.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php



(a) $m_n(t)$ at particular iterations (b) B_n versus n.

FIG. 7. *Convergence of the Robbins–Monro Algorithm* 4.1 *applied to the conditioned diffusion problem in the case of constant inverse covariance potential* B. (a) *shows evolution of* $m_n(t)$ *with* n; (b) *shows convergence of the* B_n *constant.*

The common parameters used are

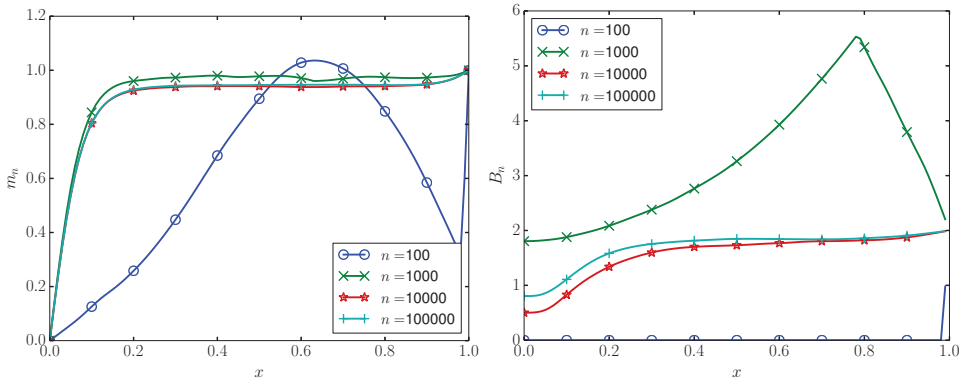
- the temperature $\varepsilon = 0.05$;
- there were 99 uniformly spaced grid points in $(0, 1)$;
- as the endpoints of the mean path are 0 and 1, we constrained our paths to lie in $[0, 1.5]$;
- B and $B(t)$ were constrained to lie in $[10^{-3}, 10^{1}]$, to ensure positivity of the spectrum;
- the standard second order centered finite difference scheme was used for C_0^{-1};
- trapezoidal rule quadrature was used to estimate $\int_0^1 \dot{m}^2$ and $\int_0^1 \dot{B}^2 dt$, with second order centered differences used to estimate the derivatives;
- $m_0(t) = t$, $B_0 = 1$, $B_0(t) = V''(1)$, the right endpoint value;
- 10^5 iterations of the Robbins–Monro algorithm are performed with 10^2 samples per iteration;
- $a_0 = 2$ and $a_n = a_0 n^{-3/5}$;
- pCN Algorithms 5.1 and 5.2 are implemented with $\beta = 0.6$, and 10^6 iterations.

Our results are favorable, and the outcome of the Robbins–Monro Algorithm 4.1 is shown in Figures 7 and 8 for the additive potentials B and $B(t)$, respectively. The means and potentials converge in both the constant and variable cases. Figure 9 confirms that in both cases, D_{KL} and D_{KL}^{α} are reduced during the algorithm.

The important comparison is when we sample the posterior using these as the proposal distributions in MCMC Algorithms 5.1 and 5.2. The results for this are given in Figure 10. Here, we compare both the prior and posterior means and variances, along with the acceptance rates. The means are all in reasonable agreement, with the exception of the m_0, which was to be expected. The variances indicate that the sampling done using μ_0 has not quite converged, which is why it is far from the posterior variances obtained from the optimized ν's, which are quite close. The optimized prior variances recover the plateau between $t = 0.2$ to $t = 0.9$, but could not resolve the peak near 0.1. Variable $B(t)$ captures some of this information in that it has a maximum in the right location, but of a smaller amplitude. However, when one standard deviation about the mean is plotted, it is difficult to see this disagreement in variance between the reference and target measures.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php



(a) $m_n(t)$ at particular iterations. (b) $B_n(t)$ at particular iterations.

FIG. 8. *Convergence of the Robbins–Monro Algorithm* 4.1 *applied to the conditioned diffusion problem in the case of variable inverse covariance potential* $B(t)$. (a) *shows* $m_n(t)$ *at particular* n. (b) *shows* $B_n(t)$ *at particular* n.

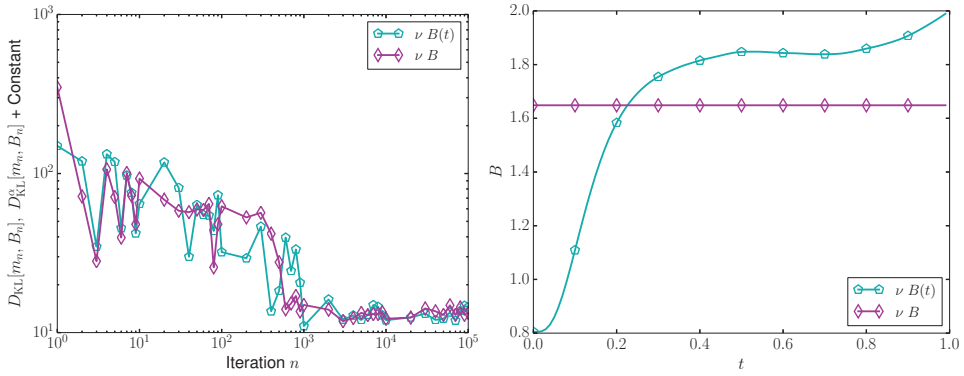


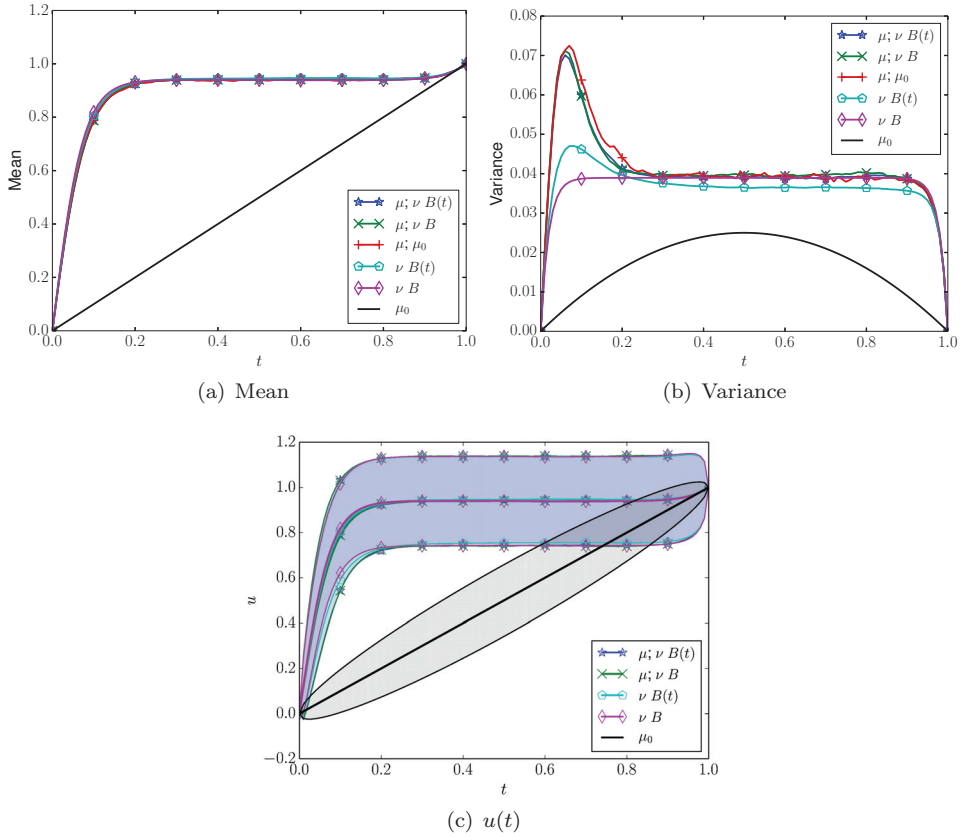
FIG. 9. *Minimization of* D^{α}_{KL} *(for* $B(t)$*) and* D_{KL} *(for* B*) during Robbins–Monro Algorithm* 4.1 *for the conditioned diffusion problem. Also plotted is a comparison of* B *and* $B(t)$ *for the optimized* ν *distributions.*

In Figure 11 we present the acceptance rate and autocovariance, to assess the performance of Algorithms 5.1 and 5.2. For both the constant and variable potential cases, there is better than an order of magnitude improvement over μ_0. In this case, it is difficult to distinguish an appreciable difference in performance between $B(t)$ and B.

7. Conclusions. We have demonstrated a viable computational methodology for finding the best Gaussian approximation to measures defined on a Hilbert space of functions, using the KL divergence as a measure of fit. We have parameterized the covariance via low rank matrices, or via a Schrödinger potential in an inverse covariance representation, and represented the mean nonparametrically, as a function; these representations are guided by knowledge and understanding of the properties of the underlying calculus of variations problem as described in [28]. Computational results demonstrate that, in certain natural parameter regimes, the Gaussian approximations are good in the sense that they give estimates of mean and covariance which are close to the true mean and covariance under the target measure of interest, and

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php



(a) Mean (b) Variance

(c) $u(t)$

FIG. 10. *Behavior of MCMC Algorithms* 5.1 *and* 5.2 *for the conditioned diffusion problem. The true posterior distribution, μ, is sampled using μ_0 (Algorithm* 5.1*) and ν, for both constant and variable potentials, B and $B(t)$, (Algorithm* 5.2*). The resulting posterior approximations are denoted by $\mu;\mu_0$ (Algorithm* 5.1*), and $\mu;\nu B$ and $\mu;\nu B(t)$ (Algorithm* 5.2*). The curves denoted μ_0, and ν B and ν $B(t)$, are the prior and best fit Gaussians. For both optimized ν's, there is good agreement between the means and the posterior mean. The variances are consistent, but the posterior shows a peak near $t = 0.1$ that is not captured by ν distributions. With the exception of μ_0, there is good general agreement amongst the distributions of $u(t)$. Shaded regions enclose $\pm$ one standard deviation.*

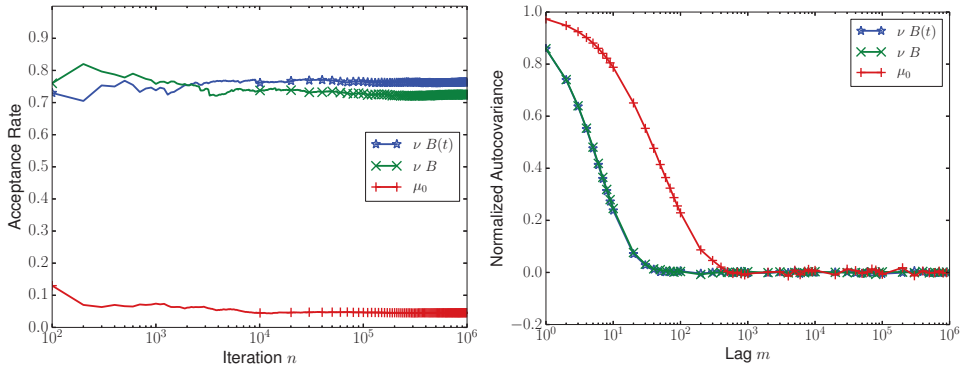
that they consequently can be used to construct efficient MCMC methods to probe the posterior distribution.

One point we highlight again is our choice to minimize $D_{\mathrm{KL}}(\nu||\mu)$ instead of $D_{\mathrm{KL}}(\mu||\nu)$. While the latter will be effective at moment matching, the former allows for the detailed approximation of individual modes. For problems where one is interested in such local structure, this is of great value. For sampling a multimodal μ, we conjecture that a Gaussian mixture proposal, with each component obtained from optimizing $D_{\mathrm{KL}}(\nu_i||\mu)$ (each ν_i corresponding to each mode) may offer greater MCMC speedup than the proposal obtained by moment matching.

Regarding the efficiency of our approach, as an MCMC sampler, consideration must be given to the cost of computing the best Gaussian fit. If n iterations of Robbins–Monro are run, each iteration will require M samples (as in (4.5)), corresponding to, approximately, M times as much work as n iterations of MCMC. The M samples are needed for the estimation of the covariance in (3.20b). This cost may

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php



(a) Acceptance rate (b) Autocovariance

FIG. 11. *Performance of MCMC Algorithms* 5.1 *and* 5.2 *for the conditioned diffusion problem. When* μ_0 *is used for proposals in Algorithm* 5.1*, the acceptance rate is far beneath either best fit Gaussian,* $\nu\ B$ *and* $\nu\ B(t)$*, within Algorithm* 5.2*. Variable* $B(t)$ *provides nominal improvement over constant* B*.*

be mitigated by the performance gain of the algorithm when sampling the target measure, μ. Moreover, when viewed as a preconditioning strategy, it will only be necessary to iterate Robbins–Monro for sufficiently long so as to improve upon the uninformed proposal algorithm. The Robbins–Monro minimization of KL could be combined with MCMC to adaptively improve the proposal, as in [1, 2, 17, 31].

There are many candidates for constructing proposal distributions with high acceptance rates. One would be to compute the MAP estimator and then to use (possibly a low rank approximation of) the Hessian of the objective function at the MAP point to form a Gaussian approximation. However, our approach benefits from being framed in terms of a variational principle. We are assured of having the optimal distribution within an admissible set. This can always be improved upon by expanding, or adapting, the admissible set.

Further work is needed to explore the methodology in larger scale applications and to develop application-specific parameterizations of the covariance operator. With respect to analysis, it would be instructive to demonstrate improved spectral gaps for the resulting MCMC methods with respect to observational noise (resp., temperature) within the context of Bayesian inverse problems (resp., conditioned diffusions).

Appendix A. Scalar example. In this section of the appendix we provide further details relating to the motivational scalar example from section 2. One of the motivations for considering such a problem is that many of the calculations for $D_{\mathrm{KL}}(\nu\|\mu)$ are explicit. Indeed if $\nu = N(m, \sigma^2)$ is the Gaussian which we intend to fit against μ, then

$$\begin{aligned} D_{\mathrm{KL}}(\nu||\mu) &= \mathbb{E}^{\nu}\left[V(x) - \tfrac{1}{2\sigma^2}|x-m|^2\right] + \log Z_\mu - \log Z_\nu \\ &= \mathbb{E}^{\nu_0}[V(y+m)] - \tfrac{1}{2} + \log Z_\mu - \log\sigma - \log\sqrt{2\pi} \\ &= \mathbb{E}^{\nu_0}[V(y+m)] - \log\sigma + \text{Constant}. \end{aligned} \tag{A.1}$$

The derivatives then take the simplified form

$$D_m D_{\mathrm{KL}} = \mathbb{E}^{\nu_0}[D_y V(y+m)], \tag{A.2a}$$

$$D_\sigma D_{\mathrm{KL}} = \mathbb{E}^{\nu_0}[V(y+m)\sigma^{-3}(y^2-\sigma^2)] - \sigma^{-1}. \tag{A.2b}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

For some choices of $V(x)$, including (2.2), the above expectations can be computed analytically, and the critical points of (A.2) can then be obtained by classical root finding. Thus, we will be able to compare the Robbins–Monro solution against a deterministic one, making for an excellent benchmark problem.

The parameters used in these computation are

- 10^6 iterations of the Robbins–Monro with 10^2 samples per iterations;
- $a_0 = .001$ and $a_n = a_0 n^{-3/5}$;
- $m_0 = 0$ and $\sigma_0 = 1$;
- m is constrained to the interval $[-10, 10]$;
- σ is constrained to the interval $[10^{-6}, 10^3]$;
- 10^6 iterations of pCN, Algorithms 5.1, 5.2, are performed with $\beta = 1$.

While 10^6 iterations of Robbins–Monro are used, Figure 1 indicates that there is good agreement after 10^3 iterations. More iterations than needed are used in all of our examples, to ensure convergence. With appropriate convergence diagnostics, it may be possible to identify a convenient termination time.

Appendix B. Sample generation. In this section of the appendix we briefly comment on how samples were generated to estimate expectations and perform pCN sampling of the posterior distributions. Three different methods were used

B.1. Bayesian inverse problem. For the Bayesian inverse problem presented in section 6.1, samples were drawn from $N(0, C)$, where C was a finite rank perturbation of C_0, $C_0^{-1} = \delta^{-1}(-d^2/dx^2)$ equipped with periodic boundary conditions on $[0, 1)$. This was accomplished using the KL series expansion (KLSE) and the fast Fourier transform (FFT). Observe that the spectrum of C_0 is

$$\varphi_n(x) = \begin{cases} \sqrt{2}\sin(2\pi\frac{n+1}{2}x), & n \text{ odd}, \\ \sqrt{2}\cos(2\pi\frac{n}{2}x), & n \text{ even}, \end{cases} \qquad \lambda_n^2 = \begin{cases} \frac{\delta}{(2\pi\frac{n+1}{2})^2}, & n \text{ odd}, \\ \frac{\delta}{(2\pi\frac{n}{2})^2}, & n \text{ even}. \end{cases} \tag{B.1}$$

Let $\mathbf{x}^n$ and μ_n^2 denote the normalized eigenvectors and eigenvalues of matrix B of rank K. Then if $u \sim N(0, C)$, $\xi_n \sim N(0, 1)$, i.i.d., the KLSE is

$$u = \sum_{\ell=1}^{K} \left\{ \sum_{n=1}^{K} \mu_n \xi_n x_\ell^n \right\} \varphi_\ell(x) + \sum_{\ell=K+1}^{\infty} \lambda_\ell \xi_\ell \varphi_\ell(x). \tag{B.2}$$

Truncating this at some index, $N > K$, we are left with a trigonometric polynomial which can be evaluated by FFT. This will readily adapt to problems posed on the d-dimensional torus.

B.2. Conditioned diffusion with constant potential. For the conditioned diffusion in section 6.2, the case of the constant potential B can easily be treated, as this corresponds to an Ornstein–Uhlenbeck (OU) bridge. Provided $B > 0$ is constant, we can associate with $N(0, C)$ the conditioned OU bridge

$$dy_t = \varepsilon^{-1}\sqrt{B} y_t dt + \sqrt{2} dw_t, \quad y_0 = y_1 = 0, \tag{B.3}$$

and the unconditioned OU process

$$dz_t = \varepsilon^{-1}\sqrt{B} z_t dt + \sqrt{2} dw_t, \quad z_0 = 0. \tag{B.4}$$

Using the relation

$$y_t = z_t - \frac{\sinh(\sqrt{B}t/\varepsilon)}{\sinh(\sqrt{B}/\varepsilon)} z_1, \tag{B.5}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

if we can generate a sample of z_t, we can then sample from $N(0, C)$. This is accomplished by picking a time step $\Delta t > 0$, and then iterating:

$$z_{n+1} = \exp\{-\varepsilon^{-1}\sqrt{B}\Delta t\}z_n + \eta_n, \quad \eta_n \sim N(0, \varepsilon/\sqrt{B}(1-\exp(-2\varepsilon^{-1}\sqrt{B}\Delta t))). \tag{B.6}$$

Here, $z_0 = 0$, and $z_n \approx z_{n\Delta t}$. This is highly efficient and generalizes to d-dimensional diffusions.

B.3. Conditioned diffusion with variable potential. Finally, for the conditioned diffusion with a variable potential $B(t)$, we observe that for the Robbins–Monro algorithm, we do not need the samples themselves, but merely estimates of the expectations. Thus, we employ a change of measure so as to sample from a constant B problem, which is highly efficient. Indeed, for any observable $\mathcal{O}$,

$$\mathbb{E}^{\nu_0}[\mathcal{O}] = \mathbb{E}^{\bar{\nu}}[\mathcal{O}\tfrac{d\nu_0}{d\bar{\nu}}] = \frac{\mathbb{E}^{\bar{\nu}}[\mathcal{O}\exp(-\Psi)]}{\mathbb{E}^{\bar{\nu}}[\exp(-\Psi)]}. \tag{B.7}$$

Formally,

$$\frac{d\nu_0}{d\bar{\nu}} \propto \exp\left\{-\frac{1}{4\varepsilon^2}\int_0^1 (B(t)-\bar{B})z_t^2 dt\right\}, \tag{B.8}$$

and we take $\bar{B} = \max_t B(t)$ for stability.

For pCN sampling we need actual samples from $N(0, C)$. We again use a KLSE, after discretizing the precision operator $C^{-1} = C_0^{-1} + B(t)$ with appropriate boundary conditions, and computing its eigenvalues and eigenvectors. While this computation is expensive, it is only done once at the beginning of the posterior sampling algorithm.

Acknowledgments. AMS is grateful to Folkmar Bornemann for helpful discussions concerning parameterization of the covariance operator. FJP would like to acknowledge the hospitality of the University of Warwick during his stay.

REFERENCES

[1] C. Andrieu and É. Moulines, *On the ergodicity properties of some adaptive MCMC algorithms*, Ann. Appl. Probab., 16 (2006), pp. 1462–1505.
[2] C. Andrieu and J. Thoms, *A tutorial on adaptive MCMC*, Stat. Comput., 18 (2008), pp. 343–373.
[3] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor, *Gaussian process approximations of stochastic differential equations*, J. Mach. Learning Res., 1 (2007), pp. 1–16.
[4] S. Asmussen and P. W. Glynn, *Stochastic Simulation*, Springer, New York, 2010.
[5] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, Vol. 1, Springer, New York, 2006.
[6] J. R. Blum, *Approximation methods which converge with probability one*, Ann. Math. Statist., 25 (1954), pp. 382–386.
[7] V. I. Bogachev, *Gaussian Measures*, Math. Surveys Monogr. 62, AMS, Providence, RI, 1998.
[8] R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu, *Sample size selection in optimization methods for machine learning*, Math. Program., 134 (2012), pp. 127–155.
[9] H.-F. Chen, L. Guo, and A.-J. Gao, *Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds*, Stochastic Process. Appl., 27 (1988), pp. 217–231.
[10] P. R. Conrad, Y. M. Marzouk, N. S. Pillai, and A. Smith, *Asymptotically Exact MCMC Algorithms via Local Approximations of Computationally Intensive Models*, preprint, arXiv:1402.1694V4, 1015.
[11] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, *MCMC methods for functions: modifying old algorithms to make them faster*, Statist. Sci., 28 (2013), pp. 424–446.
[12] A. Dvoretzky, *Stochastic approximation revisited*, Adv. Appl. Math., 7 (1986), pp. 220–227.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[13] T. A. El Moselhy and Y. M. Marzouk, *Bayesian inference with optimal maps*, J. Comput. Phys., 231 (2012), pp. 7815–7850.
[14] H. P. Flath, L. C. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas, *Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations*, SIAM J. Sci. Comput., 33 (2011), pp. 407–432.
[15] B. Gershgorin and A. J. Majda, *Quantifying uncertainty for climate change and long-range forecasting scenarios with model errors. Part* I: *Gaussian models*, J. Climate, 25 (2012), pp. 4523–4548.
[16] M. Girolami and B. Calderhead, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, J. R. Stat. Soc. Ser. B Stat. Methodol., 73 (2011), pp. 123–214.
[17] H. Haario, E. Saksman, and J. Tamminen, *An adaptive metropolis algorithm*, Bernoulli, 7 (2001), pp. 223–242.
[18] M. Hairer, A. Stuart, and S. Vollmer, *Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions*, Ann. Appl. Prob., 24 (2014), pp. 2455–2490.
[19] M. Hairer, A. Stuart, and J. Voss, *Signal processing problems on function space: Bayesian formulation, stochastic PDEs and effective MCMC methods*, in The Oxford Handbook of Nonlinear Filtering, D. Crisan and B. Rozovsky, eds., Oxford University Press, Oxford, 2011, pp. 833–873.
[20] N. J. Higham, *Computing a nearest symmetric positive semidefinite matrix*, Linear Algebra Appl., 103 (1988), pp. 103–118.
[21] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Springer, New York, 2009.
[22] M. A. Katsoulakis and P. Plecháč, *Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems*, J. Chem. Phys., 139 (2013), 074115.
[23] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, New York, 2003.
[24] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas, *A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion*, SIAM J. Sci. Comput., 34 (2012), pp. A1460–A1487.
[25] Y. M. Marzouk, H. N. Najm, and L. A. Rahn, *Stochastic spectral methods for efficient Bayesian solution of inverse problems*, J. Comput. Phys., 24 (2007), pp. 560–586.
[26] R. Pasupathy and S. Kim, *The stochastic root-finding problem*, ACM Trans. Model. Comput. Simul., 21 (2011), 19.
[27] F. Pinski, G. Simpson, A. Stuart, and H. Weber, *Algorithms for Kullback-Leibler Approximation of Probability Measures in Infinite Dimensions*, preprint, arXiv:1408.1920, 2014.
[28] F. J. Pinski, G. Simpson, A. M. Stuart, and H. Weber, *Kullback-Leibler Approximation for Probability Measures on Infinite Dimensional Spaces*, SIAM J. Math. Anal., 47 (2015), pp. 4091–4122.
[29] M. G. Reznikoff and E. Vanden-Eijnden, *Invariant measures of stochastic partial differential equations and conditioned diffusions*, C. R. Math. Acad. Sci. Paris, 340 (2005), pp. 305–308.
[30] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statist., 22 (1950), pp. 400–407.
[31] G. O. Roberts and J. S. Rosenthal, *Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms*, J. Appl. Probab., 44 (2007), pp. 458–475.
[32] M. S. Shell, *The relative entropy is fundamental to multiscale and inverse thermodynamic problems*, J. Chem. Phys., 129 (2008), 144108.
[33] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk, *Optimal low-rank approximations of Bayesian linear inverse problems*, SIAM J. Sci. Comput., 37 (2015), pp. A2451–A2487.
[34] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.
[35] G. Yin and Y. M. Zhu, *On H-valued Robbins-Monro processes*, J. Multivariate Anal., 34 (1990), pp. 116–140.

Downloaded 02/15/16 to 137.205.50.42. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.