Neural Computing Research Group
Dept of Computer Science & Applied Mathematics
Aston University
Birmingham B4 7ET
United Kingdom
Tel: +44 (0)121 333 4631
Fax: +44 (0)121 333 4586
http://www.ncrg.aston.ac.uk/

Variational Markov Chain Monte Carlo for Bayesian Smoothing of Non-linear Diffusions

Y. Shen
Y.Shen2@aston.ac.uk

D. Cornford
d.cornford@aston.ac.uk

M. Opper
Artificial Intelligence Group, Technical University Berlin,
opperm@cs.tu-berlin.de

C. Archambeau
Department of Computer Science, University College London
C.Archambeau@cs.ucl.ac.uk

Technical Report Aston University Research Archive Monograph #8106 January 19, 2010

Submitted to Computational Statistics

Abstract

In this paper we develop set of novel Markov chain Monte Carlo algorithms for Bayesian smoothing of partially observed non-linear diffusion processes. The sampling algorithms developed herein use a deterministic approximation to the posterior distribution over paths as the proposal distribution for a mixture of an independence and a random walk sampler. The approximating distribution is sampled by simulating an optimized time-dependent linear diffusion process derived from the recently developed variational Gaussian process approximation method. Flexible blocking strategies are introduced to further improve mixing, and thus the efficiency, of the sampling algorithms. The algorithms are tested on two diffusion processes: one with double-well potential drift and another with SINE drift. The new algorithm's accuracy and efficiency is compared with state-of-the-art hybrid Monte Carlo based path sampling. It is shown that in practical, finite sample, applications the algorithm is accurate except in the presence of large observation errors and low observation densities, which lead to a multi-modal structure in the posterior distribution over paths. More importantly, the variational approximation assisted sampling algorithm outperforms hybrid Monte Carlo in terms of computational efficiency, except when the diffusion process is densely observed with small errors in which case both algorithms are equally efficient.

1 Introduction

Stochastic dynamic systems, also often referred to as diffusion processes or stochastic differential equations (SDEs), have been used for modelling real systems in various areas ranging from physics to system biology to environmental science (Honerkamp, 1994; Wilkinson, 2006; Miller et al, 1999). The current work was motivated by the problem of data assimilation (Kalnay, 2003). In the data assimlation context dynamic systems representing the evolution of the atmosphere system are partially observed by an array of different instruments. The primary aim of data assimilation is the estimation of the current state of the system to provide initial conditions for forecasting. Such continuous time systems are often only partially observed, which makes likelihood based statistical inference difficult. In this paper, the inference problem we focus on is smoothing, that is estimation of the posterior distribution over paths in state space. From a methodological point of view, the smoothing problem for stochastic dynamic systems has been pursued in three main directions.

The first direction is based on solving the Kushner-Stratonovich-Pardoux (KSP) equations (Kushner, 1967; Pardoux, 1982) which are the most general optimal solutions to the smoothing problem. However, solution of the KSP equations is numerically intractable for even quite low-dimensional non-linear systems (Miller et al, 1999), so various approximation strategies have been developed. In the particle approximation framework (Kitagawa, 1987), the solution of the KSP equations is approximated by a discrete distribution with random support. As proposed by Kitagawa (1996), the smoothed density, namely the posterior density, is obtained from the combination of a forward filter and a backward filter. For linear, Gaussian systems, the forward KSP equation reduces to the well-known Kalman-Bucy filter (Kalman and Bucy, 1961). The smoothed estimates of mean and covariance can be calculated from the filtering results by recursion, a method often referred to as the Kalman smoother (Jazwinski, 1970). To treat non-linear systems a number of approximation strategies have extended the Kalman smoother, for example, the ensemble Kalman smoother which employs a moderately sized, randomly sampled, ensemble to estimate the predicted state covariance matrix and then applies the linear Kalman updates (Evensen, 1994, 2000) and the unscented Kalman smoother which is similar in spirit but uses an optimal deterministic sampling strategy to select the 'ensemble ' members (Julier et al, 2000; Wan and van der Merwe, 2001).

The second direction involves a variational approximation to the posterior process. In Archambeau et al (2007), a linear diffusion approximation is proposed and its time varying linear drift is optimised globally. This is explained in more detail in Sect. 2.2. In the work by Eyink et al (2004), a mean field approximation is applied to the KSP equations and the mean field representation of possible trajectories is optimised globally, to determine the most probable trajectory in a similar vein to 4DVAR data assimilation methods (Derber, 1989). 4DVAR methods are widely used in operational data assimilation (Rabier et al, 2000) and essentially seek the mode of the approximate posterior smoothing distribution but do not provide any estimate of the uncertainty about this most probably trajectory.

The third direction employs Markov Chain Monte Carlo (MCMC) methods (Andrieu et al, 2003) to sample the posterior process, which is the focus of this paper. At each step of an MCMC simulation, a new state is proposed and will be accepted or rejected in a probabilistic manner. For applications to continuous-time stochastic dynamic systems, it is also often referred to as path sampling. A single-site update approach to path sampling is adopted by Eraker (2001) in the context of parameter estimation of diffusion processes. The author reported arbitrarily poor mixing of this basic algorithm. To achieve better mixing, two closely related MCMC algorithms for path sampling, namely the Metropolis-adjusted Langevin (Stuart et al, 2004) and the Hybrid Monte Carlo (HMC) algorithm (Alexander et al, 2005) have recently been proposed. Both methods update the entire sample path at each sampling iteration while keeping the acceptance of new paths high. This is achieved by combining the basic MCMC algorithm with a fictitious dynamics so that the MCMC sampler proposes moves towards the regions of higher probability in the state space while maintaining detailed balance. Another strategy to achieve better mixing in path sampling is to update one of the sub-paths between two neighbouring observations at each Metropolis-Hastings step leading to "blocking strategies". In Golightly and Wilkinson (2008), the so-called "modified diffusion bridge" approach is used

to propose candidates for such sub-paths. This method is a further development of the Brownian bridge sampler proposed by Roberts and Stramer (2001) and Durham and Gallant (2002). Also, sequential Monte Carlo (SMC) methods are used to implement the above blocking scheme in Golightly and Wilkinson (2006) (for the use of SMC to build efficient proposal distributions, we refer to Andrieu et al (2010)). Further, a random blocking scheme is proposed by Elerian et al (2001). To the same end, Beskos et al (2006) have developed a so-called "retrospective sampling" method which can simulate a wide class of diffusion bridge processes exactly, under certain conditions on the driving noise process.

For an overview of strategies to develop efficient MCMC algorithms, we refer to the book by Liu (2001). Recently, a new strategy combining sampling methods with variational methods was introduced by de Freitas et al (2001). The starting point of this strategy is to use an optimized approximation to the posterior distribution as a proposal distribution in a Metropolis-Hastings (MH) step. The sampling scheme is implemented in an independence sampler setting. The resulting algorithm is called variational MCMC (VMC). In the presence of approximation error, as stated by de Freitas et al (2001), the acceptance rate of those proposals for a high-dimensional state space is likely to be very low. To control the acceptance rate, a block sampling algorithm is proposed by de Freitas et al (2001) to update only a subset of the components of state variables at individual MH steps. They argued that a "variational approximation"-assisted approach to sampling is helpful in exploring the regions of high probability efficiently, however the sampler could get stuck in the neighbouring regions of lower probabilities as the approximate posterior is often much more peaked than the true one. To avoid this, the algorithm is further developed by combining a Metropolis-type sampler with the independence sampler in a probabilistic manner (de Freitas et al, 2001). This methodology is illustrated in de Freitas et al (2001) using an example of Bayesian parameter estimation for logistic belief networks.

In this paper, we employ the variational approximation method developed by Archambeau et al (2007) to produce a computationally efficient sampling method. From the variational method, we obtain a time-varying linear SDE representing the optimized Gaussian process approximation to the true posterior process, where we emphasise this is over the full smoothing path density. From the approximate linear SDE, we can derive any bridging process within the smoothing window exactly, which extends the blocking strategies proposed by Golightly and Wilkinson (2008), allowing blocks of arbitrary size. To implement the mixture strategy, we split the proposal procedure into two steps: the first step generating the driving white noise and the second step simulating the process forward in time with generated noise. The white noise can be generated either by direct sampling from a standard multivariate Gaussian distribution or by a random walk MH sampler. The implementation can be seen as an adaptation of the reparameterisation strategy used in Golightly and Wilkinson (2008) for parameter estimation, however here our aim is to produce efficient proposals for the driving noise process.

The paper is organised as follows; Sect. 2 presents a Bayesian treatment of non-linear smoothing which is followed by a summary of the Markov Chain Monte Carlo smoother (Alexander et al, 2005) in Sect. 2.1 and the variational Gaussian process smoother (Archambeau et al, 2007) in Sect. 2.2. The novel algorithms are described in Sect. 3 and the performance of these algorithms is demonstrated in Sect. 4 by numerical experiments with two stochastic dynamic systems. The paper concludes with a discussion.

2 Computational methods for Bayesian smoothing of non-linear diffusions

Consider a stochastic dynamical system represented by

$$d\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, t)dt + \mathbf{D}^{1/2}(t)d\mathbf{W}(t), \tag{1}$$

where $\mathbf{x}(t) \in \mathcal{R}^d$ is the state vector, $\mathbf{D} \in \mathcal{R}^{d\times d}$ is the so-called diffusion term, and $\mathbf{f}$ represents a deterministic dynamical process, generally called the drift. The driving noise process is represented by a Wiener

process $\mathbf{W}(t)$. Note that SDE (1) is also referred to as a sub-class of diffusion processes whose diffusion term $\mathbf{D}$ is independent of state $\mathbf{x}$ (Klöden and Platen, 1992).

The state is observed via some measurement function $\mathbf{h}(\cdot)$ at discrete times, say $\{t_k\}_{k=1,...,M}$. The observations are assumed contaminated by i.i.d Gaussian noise:

$$\mathbf{y}_k = \mathbf{h}(\mathbf{x}(t_k)) + \mathbf{R}^{\frac{1}{2}} \cdot \eta \tag{2}$$

where $\mathbf{y}_k \in \mathcal{R}^{d'}$ is the k-th observation, $\mathbf{R} \in \mathcal{R}^{d' \times d'}$ is the covariance matrix of measurement errors, and η represents standard multivariate Gaussian white noise.

A Bayesian approach to smoothing is typically adopted in which the posterior distribution

$$p(\mathbf{x}([0,T])|\{\mathbf{y}_1, ..., \mathbf{y}_M, 0 < t_1 < ... < t_M < T\}),$$

is formulated and estimated, using for example the methods described in Sect. 1. In this work the continuous-time SDE is discretised using an explicit Euler-Maruyama scheme (Klöden and Platen, 1992). This discretisation induces an approximate non-linear discrete time model which describes a Markov chain. The discretised version of (1) is given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{f}(\mathbf{x}_k, t_k)\delta t + \mathbf{D}^{1/2}(t_k)\sqrt{\delta t} \cdot \xi_k, \tag{3}$$

with $t_k = k \cdot \delta t$, k = 0, 1, ..., N, and a smoothing window from t = 0 to $T = N \cdot \delta t$. Note that ξ_k are white noise random variables. An initial state, $\mathbf{x}_0$, needs to be set. There are M observations within the smoothing window chosen at a subset of discretisation times $(t_{k_j}, \mathbf{y}_j)_{j=1,...,M}$ with

$$\{t_{k_1}, ..., t_{k_M}\} \subseteq \{t_0, ..., t_N\}.$$

In the following the posterior distribution is formulated step by step. As a result of Euler-Maruyama discretisation, the prior of a diffusion process can be written as

$$p(\mathbf{x}_0, ..., \mathbf{x}_N) = p(\mathbf{x}_0) \cdot p(\mathbf{x}_1|\mathbf{x}_0) \cdot .... \cdot p(\mathbf{x}_N|\mathbf{x}_{N-1}),$$

where $p(\mathbf{x}_0)$ is the prior on the initial state and $p(\mathbf{x}_{k+1}|\mathbf{x}_k)$ with k = 0, ..., $N-1$ are the transition densities of the diffusion process. Note that in the limit of small enough δt, those transition densities can be well approximated by a Gaussian density and thus $p(\mathbf{x}_{k+1}|\mathbf{x}_k) = \mathcal{N}(\mathbf{x}_k + f(\mathbf{x}_k)\delta t, \mathbf{D}\delta t)$. Therefore, the prior over the path, defined by the SDE is given by

$$p(\mathbf{x}_0, ..., \mathbf{x}_N) \propto p(\mathbf{x}_0) \cdot \exp(-\mathcal{H}_{\text{dynamics}}),$$

where $\mathcal{H}_{\text{dynamics}} =$

$$\sum_{k=0}^{N-1} \frac{\delta t}{2} \left[\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\delta t} - \mathbf{f}(\mathbf{x}_k, t_k)\right]^\top \mathbf{D}^{-1} \left[\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\delta t} - \mathbf{f}(\mathbf{x}_k, t_k)\right].$$

Assuming the measurement noise is i.i.d. Gaussian, the likelihood is simply given by

$$p(\mathbf{y}_1, ..., \mathbf{y}_M|\mathbf{x}(t_0), ..., \mathbf{x}(t_N)) \propto \exp(-\mathcal{H}_{\text{obs}}),$$

where

$$\mathcal{H}_{obs} = \frac{1}{2} \sum_{j=1}^{M} \left[\mathbf{h}(\mathbf{x}(t_{k_j})) - \mathbf{y}_j\right]^\top \mathbf{R}^{-1} \left[\mathbf{h}(\mathbf{x}(t_{k_j})) - \mathbf{y}_j\right]. \tag{4}$$

In summary, the posterior distribution of all states, i.e. $\mathbf{p}(\{\mathbf{x}_0, ..., \mathbf{x}_N\}|\{\mathbf{y}_1, ..., \mathbf{y}_M\})$ is given by

$$\mathbf{p}(\cdot|\cdot) \propto \exp(\log(p(\mathbf{x_0})) - \mathcal{H}_{\text{dynamics}} - \mathcal{H}_{\text{obs}}). \tag{5}$$

In Fig. 1, Bayesian smoothing is illustrated using an example of a one-dimensional stochastic double-well system. This system is a diffusion process with two stable states; for details refer to Sect. 4. The reference smoothing result is obtained using a Hybrid Monte Carlo algorithm based on the work by Alexander et al (2005).

2.1 Hybrid Monte Carlo Methods

In HMC approaches, a molecular dynamics simulation algorithm is applied to make proposals in a MH algorithm, for example,

$$\mathcal{X}^k = \{\mathbf{x}_0^k, ..., \mathbf{x}_N^k\} \longrightarrow \mathcal{X}^{k+1} = \{\mathbf{x}_0^{k+1}, ..., \mathbf{x}_N^{k+1}\},$$

at step k. To make a proposal of $\mathcal{X}^{k+1}$ a fictitious deterministic system is simulated as follows

$$\begin{aligned} \frac{d\mathcal{X}}{d\tau} &= \mathbf{P} \\ \frac{d\mathbf{P}}{d\tau} &= -\nabla_{\mathcal{X}}\hat{\mathcal{H}}(\mathcal{X}, \mathbf{P}) \end{aligned}$$

where $\mathbf{P} = (\mathbf{p}_0, ..., \mathbf{p}_N)$ represents momentum and $\hat{\mathcal{H}}$ is a fictitious Hamiltonian which is the sum of potential energy $\mathcal{H}^{pot}$ and kinetic energy $\mathcal{H}^{kin} = \frac{1}{2}\sum_{k=1}^{N} \mathbf{p}_k^2$. For the posterior distribution of the non-linear smoothing problem given in Sect. 2, the potential energy is given by

$$\mathcal{H}^{pot} = -\log[p(\mathbf{x}_0)] + \mathcal{H}^{\text{dynamics}} + \mathcal{H}^{\text{obs}}.$$

The above system is initialised by setting $\mathcal{X}(\tau = 0) = \mathcal{X}_k$ and sampling a random number from $\mathcal{N}(0, 1)$ for each component of $\mathbf{P}(\tau = 0)$. After that, one integrates the system equations forward in time with time increment $\delta\tau$ by using a leapfrog scheme as follows:

$$\begin{aligned} \mathcal{X}' &= \mathcal{X} + \delta\tau\mathbf{P} + \frac{\delta\tau^2}{2}(-\nabla_{\mathcal{X}}\hat{\mathcal{H}}) \\ \mathbf{P}' &= \mathbf{P} + \frac{\delta\tau}{2}(-\nabla_{\mathcal{X}}\hat{\mathcal{H}} - \nabla_{\mathcal{X}'}\hat{\mathcal{H}}) \end{aligned}$$

After J iterations, the state $\mathcal{X}(\tau = J\delta\tau)$ is proposed as $\mathcal{X}^{k+1}$ which will be accepted with probability

$$\min\left\{1, \exp\left(-\hat{\mathcal{H}}^{k+1} + \hat{\mathcal{H}}^k\right)\right\}.$$

The sequence of states generated from this mechanism is then a sample for the posterior smoothing distribution, (5).

2.2 Variational Gaussian process approximation smoother

The starting point of the Variational Gaussian Process Approximation (VGPA) method (Archambeau et al, 2007) is to approximate (1) by a linear SDE:

$$d\mathbf{x}(t) = \mathbf{f}_L(\mathbf{x}, t)dt + \mathbf{D}^{1/2}(t)d\mathbf{W}(t), \tag{6}$$

where the time varying linear drift approximation is given by

$$f_L(\mathbf{x}, t) = -\mathbf{A}(t)\mathbf{x}(t) + \mathbf{b}(t)\ . \tag{7}$$

The matrix $\mathbf{A}(t) \in \mathcal{R}^{d\times d}$ and the vector $\mathbf{b}(t) \in \mathcal{R}^d$ are the variational parameters to be optimised in the procedure.

The approximation in (7) implies that the true posterior process, i.e. $\mathbf{p}(\mathbf{x}_t|\mathbf{y}_1, ..., \mathbf{y}_M)$, is approximated by a Gaussian Markov process, $\mathbf{q}(\mathbf{x}_t)$. Discretising the linear SDE in the same way as the true SDE, the approximate posterior can be written as

$$\mathbf{q}(\mathbf{x}_0, ..., \mathbf{x}_N) = q(\mathbf{x}_0) \cdot \prod_{k=0}^{N-1} \mathcal{N}(\mathbf{x}_{k+1}|\mathbf{x}_k + f_L(\mathbf{x}_k)\delta t, \mathbf{D}\delta_t)\ ,$$

where $q(\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0|m(0), S(0))$. The optimal $\mathbf{A}(t)$ and $\mathbf{b}(t)$, together with the optimal marginal means and covariances $\mathbf{m}(t)$ and $\mathbf{S}(t)$, are obtained by minimising the Kullback-Leibler (KL) divergence of $\mathbf{q}(\cdot)$ and $\mathbf{p}(\cdot)$ (Archambeau et al, 2007). The variational approximation can also be derived in continuous time using Girsanov's change of measure theorem, however it is then still necessary to discretise the system for computational implementation (Archambeau et al, 2008).

In Fig. 2 and Fig. 3, the VGPA smoothing method is illustrated using the same double-well example as in Fig. 1. It can be observed from Fig. 2 that the smooth variation of the parameters $\mathbf{A}$ and $\mathbf{b}$ is interrupted by jumps at observation times. This not only draws the mean path towards the observations but also reduces marginal variance around those observation times, which can be seen in the upper panel of Fig. 3. The lower panel shows that a typical realisation generated by the optimized approximate diffusion is visually similar to the original sample path shown in Fig. 1.

3 Variational MCMC methods

In a Metropolis-Hastings algorithm (Hastings, 1970) for sampling a posterior density $\pi(x)$, defined on a general state space $\mathcal{X}$, a new state $x' \in \mathcal{X}$ is proposed according to some density $q(x, x')$. The proposed state will be accepted with probability $\alpha(x, x')$, given by

$$\alpha = \min\left\{1, \frac{\pi(x')}{\pi(x)} \cdot \frac{q(x', x)}{q(x, x')}\right\}.$$

When the Metropolis-Hastings (MH) algorithm is applied to a particular Bayesian inference problem, intelligent proposal mechanisms are often required to make the algorithm efficient.

3.1 The variational independence sampler

In the new variational MCMC algorithms, proposals are made using the variational parameters $\mathbf{A}$ and $\mathbf{b}$, as well as the marginal moments $\mathbf{m}$ and $\mathbf{S}$, from the VGPA method described in Sect. 2.2. In this setting, an independence sampler to update the whole sample path at once is implemented as follows.

First, propose the initial state by sampling from a normal distribution specified by the marginal moments at $t = 0$, i.e. $\mathbf{x}_0 \sim \mathcal{N}(\cdot|\mathbf{m}_0, \mathbf{S}_0)$; then, integrate the SDE with its time-varying linear drift, $\mathbf{f_L}(\mathbf{x}, t) = -\mathbf{A}_t\mathbf{x} + \mathbf{b}_t$, forward in time using an Euler scheme. The implementation is thus:

$$\mathbf{x}_0 = \mathbf{m}_0 + \sqrt{\mathbf{S}_0} \cdot w_0$$

and

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{f_L}(\mathbf{x}, t)\delta t + \mathbf{D}^{1/2}\sqrt{\delta t} \cdot w_k \quad \text{for} \quad k = 1, ..., N,$$

where $\mathbf{w}^\top = (w_0, w_1, ..., w_N)$ is a set of realisations of white noise, which can be seen as a realisation of the driving process of the approximate SDE. The resulting sample path is taken as a new proposal. The acceptance rate of the independence sampler is given by,

$$\alpha = \min\left\{1, \frac{\pi(\mathbf{x}'|\{\mathbf{y}_1, ..., \mathbf{y}_M\})}{\pi(\mathbf{x}|\{\mathbf{y}_1, ..., \mathbf{y}_M\})} \cdot \frac{q(\mathbf{x})}{q(\mathbf{x}')}\right\}.$$

where $\pi(\cdot)$ and $q(\cdot)$ are the posterior and proposal density of sample path $\mathbf{x}$, respectively. For simplicity $\{\mathbf{y}_1, ..., \mathbf{y}_M\}$ is omitted in $\pi(\cdot)$ in the remainder of this section. Let $q(\mathbf{w})$ denote the proposal density of white noise $\mathbf{w}$. For both original and approximate SDE, there is one-to-one relationship between $\mathbf{x}$ and

$\mathbf{w}$ and the corresponding Jacobian in the MH-ratio is cancelled due to a constant diffusion coefficient. Accordingly,

$$\frac{q(\mathbf{x})}{q(\mathbf{x}')} = \frac{q(\mathbf{w})}{q(\mathbf{w}')} \quad \text{with} \quad q(\mathbf{w}) \propto \prod_{k=0}^{N} \exp\left(-\frac{w_k^2}{2}\right).$$

The method above differs from the one proposed in Roberts and Stramer (2001) which uses an Ozaki approximation (Ozaki, 1992) to construct a linear bridge between two (noise-free) observations. In contrast, the approximation scheme herein is not based on local Taylor expansion of the non-linear drift but optimized globally, conditioning on all observations. Moreover the variational approximation derived drift is time-varying where the bridge proposed by Roberts and Stramer (2001) has constant drift between two observations.

The efficiency of the independence sampler depends on how well the proposal density approximates the target measure. In this work, the proposal density is a Gaussian process approximation to the target measure. Therefore, the efficiency of the above algorithms is determined by how far the target measure deviates from a Gaussian one. In cases with a highly non-linear drift term in the diffusion *and* relatively few observations, the above proposal mechanisms need to be further refined.

3.2 Conditional block sampling - the variational diffusion bridge

The idea of blocking helps to improve the performance of independence samplers. Simply speaking, only a sub-path of the full sample path is proposed,

$$\mathbf{X}^{sub} = \{x_k, ..., x_{k+L-1}\} \subset \mathbf{X} = \{x_0, ..., x_T\}$$

say, at each MH step while the remaining parts are fixed. The index k is chosen randomly while the sub-path length L is a tuning parameter of the algorithm (for details see below). Due to the Markov property of the process $\mathbf{X}_t$, the sub-path needs only be conditioned on x_{k-1} and x_{k+L}, instead of the entire remaining path. To implement this blocking strategy, the simulation of a time-varying, linear bridging process is required to make proposals.

For the bridging simulation, the effective drift and diffusion term of a time-varying linear SDE are derived. For clarity, a one-dimensional SDE is considered and the problem of simulating a bridging process is further reduced to the following question: how to sample x_t at time t with $t = t' + \delta t$ and $t < T$ conditioning $x_{t'}$ at time t' and x_T at time T? Note that $\Delta t = T - t' >> \delta t$.

To sample x_t, first compute the conditional probability

$$p(x_t|x_{t'}, x_T) \propto p(x_t|x_{t'}) \cdot p(x_T|x_t) \ .$$

As the time increment δt is the one used to discretise both the original and approximate SDE by a Euler-Maruyama scheme

$$p(x_t|x_{t'}) \propto \exp\left\{ \underbrace{-\frac{1}{2D\delta t}(x_t - x_{t'} - f_L(x_{t'})\delta t)^2}_{I_1} \right\} . \tag{8}$$

On the other hand,

$$p(x_T|x_t) \propto p(x_t|x_T) \cdot p(x_T).$$

In fact, the right side can be understood as the marginal density of x at time t for a linear SDE represented by (6) which runs from $t = T$ to $t = 0$. The moment equations of this backward linear SDE are

$$\frac{d}{dt}\mathbf{c}_t = -A\mathbf{c}_t + b$$

and

$$\frac{d}{dt}\mathbf{d}_t = -2A\mathbf{d}_t - D \ ,$$

where $\mathbf{c}_t$ and $\mathbf{d}_t$ are the marginal mean and variance, respectively. For a bridge, the state at time T is fixed to x_T which implies $\mathbf{c}_T = x_T$ and $\mathbf{d}_T = 0$. Thus

$$p(x_T|x_t) \propto \exp\left\{\underbrace{-\frac{1}{2\mathbf{d}_t}(x_t - \mathbf{c}_t)^2}_{I_2}\right\} \ . \tag{9}$$

Re-formulating I_1 and I_2 in (8) and (9), respectively, the effective drift and diffusion terms for the bridging process are obtained as follows:

$$f_L^{eff} = -1 \cdot \underbrace{\frac{\mathbf{d}A + D}{\mathbf{d} + D\delta t}}_{A_t^{eff}} \cdot x_{t'} + \underbrace{\frac{\mathbf{c}D + b\mathbf{d}}{\mathbf{d} + D\delta t}}_{b_t^{eff}} \tag{10}$$

and

$$D_t^{eff} = D \cdot \frac{\mathbf{d}}{\mathbf{d} + D\delta t} \ . \tag{11}$$

As the forward effective SDE is equivalent to the process specified by the above backward moment equations, it is clear that its realization must hit x_T at time T.

In the continuous-time limit, the above effective drift- and diffusion terms are in accordance with those derived from the corresponding Fokker-Planck equations, which are

$$f_L^{eff} = f_L + \sigma^2 \partial_x \ln p(X_T|x_t)$$

and $D_t^{eff} = D$, respectively.

In Fig. 4, block sampling is illustrated with the same stochastic double-well example as in Fig. 1. Notice that both effective $\mathbf{A}$ and $\mathbf{b}$ rise sharply at the right end of the block (middle and lower panel, respectively). In this particular example, the rise of $\mathbf{b}$ forces the effective mean path to evolve towards the fixed end point. The proposed sub-path also approaches the fixed end point because the effective marginal variance is reducing to zero due to the increase of $\mathbf{A}$.

The general principle of constructing a bridging process from the approximate linear SDE has been outlined. To implement it efficiently and correctly, the following two technical details also need to be addressed:

1. With blocking, the MH-ratio is given by

$$\alpha = \min\left\{1, \frac{\pi(\mathbf{x}'_{sub})}{\pi(\mathbf{x}_{sub})} \cdot \frac{q(\mathbf{w}_{sub})}{q(\mathbf{w}'_{sub})}\right\} .$$

   As stated above, $\mathbf{w}'_{sub}$ is used to propose $\mathbf{x}'_{sub}$ by integrating the effective linear SDE forward in time. However, $\mathbf{w}_{sub}$ must be re-constructed from $\mathbf{x}_{sub}$ with the same SDE that generates $\mathbf{x}'_{sub}$. Note that $\mathbf{w}_{sub}$ are different from those that are actually used to propose $\mathbf{x}_{sub}$. This is because the conditioning has been changed;

2. To increase the acceptance rate, the proposal of $\mathbf{x}_0$ must be conditioned on $\mathbf{x}_1$. This is obtained by sampling from a conditional Gaussian distribution as follows

$$\mathcal{N}\left(\frac{\mathbf{m}_0 \cdot \mathbf{S}^* + \mathbf{m}^* \cdot \mathbf{S}_0}{\mathbf{S}_0 + \mathbf{S}^*}, \quad \frac{\mathbf{S}_0 \cdot \mathbf{S}^*}{\mathbf{S}_0 + \mathbf{S}^*}\right)$$

where

$$\mathbf{m}^* = \frac{\mathbf{x}_1 - \mathbf{b}_0 \cdot dt}{1 - \mathbf{A}_0 \cdot dt} \quad \text{and} \quad \mathbf{S}^* = \frac{\mathbf{D} \cdot dt}{(1 - \mathbf{A}_0 \cdot dt)^2}$$

Numerical experiments show that the above algorithm for path sampling could have poor mixing if the sampler gets stuck in the regions of lower probabilities. Therefore, a VGPA-assisted random walk sampler is developed for path sampling in the following section.

3.3 The variational random walk sampler

To augment the independence sampler, a random walk scheme for path sampling is also developed. A trivial implementation of such scheme would update the sample path directly, which leads to a vanishingly low acceptance rate. Instead, the driving process $\mathbf{w}$ is updated using a random walk sampler, i.e.

$$\mathbf{w}'_t = \mathbf{w}_t + \sigma \cdot \eta_t.$$

where η_t is white noise and σ represents the step size of the random walk. Given $\mathbf{w}'_t$, a new sample path is obtained by simulating the approximate SDE forward in time. Denote this deterministic mapping function by f_T, i.e. $\mathbf{x}'_t = f_T(\mathbf{w}'_t)$. Up to a Jacobian, the joint posterior is thus given by

$$\pi(\mathbf{x}', \mathbf{w}'|\mathbf{y}) \propto \mathbf{p}(\mathbf{x}) \cdot \mathbf{q}(\mathbf{w}) \cdot \mathbf{p}(\mathbf{y}|f_T(\mathbf{w}))$$

where $\mathbf{p}(\mathbf{x})$ and $\mathbf{q}(\mathbf{w})$ represents the prior density of the diffusion process and the driving process, respectively. As the Jacobian is cancelled due to a constant diffusion coefficient, the resulting MH ratio is given by

$$\alpha = \min\left\{1, \frac{\pi(\mathbf{x}')}{\pi(\mathbf{x})} \cdot \frac{\mathbf{q}(\mathbf{w}')}{\mathbf{q}(\mathbf{w})}\right\}.$$

Recall that $\pi(\mathbf{x}) = \mathbf{p}(\mathbf{x}) \cdot \mathbf{p}(\mathbf{y}|\mathbf{x} = f_T(\mathbf{w}))$.

The idea of a random walk in $\mathbf{w}$-space is very similar to the innovation scheme proposed by Golightly and Wilkinson (2008). In (Golightly and Wilkinson, 2008) the innovation scheme is utilised to break high dependence between unknown diffusion parameters and missing sample paths whereas in this work the random walk sampler helps an independence sampler to avoid getting stuck. Note that the random walk sampler could be very slow in exploring the posterior density. Thus, a mixture strategy needs to be adopted.

3.4 The variational diffusion sampler

Finally, the independence sampler and the random walk sampler proposed here are combined into a mixture transition kernel in the variational assisted MCMC algorithm devised for path sampling. This results in a transition kernel given by

$$\mathcal{T} = p \cdot \mathcal{T}^{ind} + (1 - p) \cdot \mathcal{T}^{rand}, \quad \text{with probability } p,$$

where $\mathcal{T}^{ind}$ and $\mathcal{T}^{rand}$ represent the transition kernel of an independence sampler and a random walk sampler, respectively. The complete algorithm is detailed in Algorithm 1.

4 Numerical Experiments

In this section, the variational MCMC (VMC) algorithms described in Sect. 3 are compared with the state-of-the-art HMC method based on the implementation developed by Alexander et al (2005). HMC methods

Algorithm 1 The VMC path sampler.

fix block length, L, apply VGPA to estimate $\mathbf{A}$ and $\mathbf{b}$.
generate a set of white noise $\{\mathbf{w}_{0,1,...,T}\}$
initialise $\mathbf{x}_0 \sim \mathcal{N}(\cdot|\mathbf{m}_0, \mathbf{S}_0)$ using $\mathbf{w}_0$
initialise $\mathbf{x}_{1,...,T}$ by using $\mathbf{w}_{1,...,T}$ to integrate $d\mathbf{x} = (-\mathbf{A}\mathbf{x} + \mathbf{b})dt + \mathbf{D}d\mathbf{w}$
repeat
  randomly choose block k from $\{1, 2, ..., T\}$
  if $k \leq T - L$ *(block fully within time window)* **then**
    compute effective $\hat{\mathbf{A}}$, $\hat{\mathbf{b}}$ and $\hat{\mathbf{D}}$ for the block conditioned on $\mathbf{x}_{k-1}$ and $\mathbf{x}_{k+L}$
    compute the effective white noise $\hat{\mathbf{w}}_k^{k+L-1}$
    choose an independence sampler or a random walk sampler with probability p
    if an independence sampler is chosen **then**
      generate new white noise $\mathbf{w}_k^{k+L-1}$
      propose $\mathbf{x}_k^{k+L-1}$ by integrating $d\mathbf{x} = (-\hat{\mathbf{A}}\mathbf{x} + \hat{\mathbf{b}})dt + \hat{\mathbf{D}}d\mathbf{w}$
      make a Metropolis-Hastings update for $\mathbf{x}_k^{k+L-1}$
    else
      generate white noise ξ_k^{k+L-1}
      propose new $\mathbf{w}_k^{k+L-1}$ by $\mathbf{w} = \hat{\mathbf{w}} + \xi \cdot$ step size
      propose $\mathbf{x}_k^{k+L-1}$ by integrating $d\mathbf{x} = (-\hat{\mathbf{A}}\mathbf{x} + \hat{\mathbf{b}})dt + \hat{\mathbf{D}}d\mathbf{w}$
      make a Metropolis-Hastings update for $\mathbf{w}_k^{k+L-1}$ and $\mathbf{x}_k^{k+L-1}$ jointly
    end if
  else
    update $\mathbf{x}_0$ and $\mathbf{x}_{k,...,T}$ applying methods similar to the above, using the original $\mathbf{A}$ and $\mathbf{b}$ from VGPA
  end if
until a sufficient number of sample paths are obtained

are used in the comparison since these are know to be amongst the most computationally efficient sampling strategies for high dimensional problems.

4.1 Experimental setup

The algorithms are tested on two one-dimensional systems: a stochastic double-well potential system (DW) and a SINE-drift diffusion process (SINE). Note that while the dynamic systems have a single state variable, the discrete time state vector (path) being sampled is very high dimensional, typically around 800 to 5000 D. The systems are described by

$$dx = 4x(1 - x^2)dt + Ddw \ ,$$

and

$$dx = 2\pi D \sin(2\pi x)dt + Ddw \ ,$$

respectively, where D is the diffusion variance and dw represents white-noise. As can be seen from Figs. (5, 7, 8), both systems have meta-stable states. They are $x = \pm 1$ for DW and $x = \pm k \cdot \frac{1}{2}$, with $k \in \mathbb{N}$ for SINE. Note that SINE systems have an infinite number of meta-stable states. Clearly, the parameter D determines how frequent the transition between those meta-stable states are. Two variants of DW are investigated, one with D = 0.25 typical of rare transitions and one with D = 1.0 typical of frequent transitions. For SINE, D is set to 0.656 so that its transition frequency is comparable with the 2nd variant of DW (about 6 transitions in a time interval of 50 units, see Fig. 7 and Fig. 8). Note that the average exit time for DW with D = 0.25 is about 3000 time units, thus we have selected a rare transition to illustrate the worst case behaviour of the new algorithms on this system. In this example, the time interval is set to 8 units to enable a large-scale Monte Carlo experiment to be carried out in order to assess the variability of the numerical results. It is clear that the mixing of the MCMC algorithms will depend on the quality of observations in terms of observation density ρ (number of observations per time unit) as well as observation error variance R. Therefore, both HMC and VMC are tested with 9 combinations of 3 different ρ-values and 3 different R-values, i.e. $\rho_1 = 1$, $\rho_2 = 2$, $\rho_3 = 4$, $R_1 = 0.04$, $R_2 = 0.09$, and $R_3 = 0.36$.

The efficiency of the different MCMC algorithms is compared using their mixing properties while also rigorously assessing the accuracy of the new algorithm. Given a fixed amount of computing time, the mixing property is a critical measure for the computational efficiency of a particular MCMC algorithm. The auto-correlation between subsequent samples is often used to quantify the mixing. Here, each sample is a realisation, or sample path, of the diffusion process conditioned on the data. The auto-correlation of a simple summary statistic of sample paths is computed as a diagnostic, namely

$$L = \int_0^T \mathbf{x}(t)dt.$$

To summarise the auto-correlation function, a so-called auto-correlation time τ is defined by

$$\tau = 1 + 2 \cdot \sum_{k=1}^{\infty} \text{ACF}(k).$$

with

$$\text{ACF}(k) = \left\langle \frac{(L_{\hat{t}} - < L_{\hat{t}} >)(L_{\hat{t}+k} - < L_{\hat{t}} >)}{\text{Var}(L)} \right\rangle$$

where $< \cdot >$ represents the expectation and $\hat{t}$ denotes algorithmic time.

The summation is truncated at lag k = 40 to minimise the impact of finite sample size errors on the estimates. When the total computing time is fixed, the length of the Markov chains generated by different algorithms varies. Thus the chains are sub-sampled so that the same number of sample paths are obtained from all algorithms.

The accuracy of the new algorithm is characterised by the integrated marginal KL-divergence between its samples and those from HMC:

$$\mathrm{KL} = \int_0^T \left\{ \int \hat{\pi}_t^{\mathrm{HMC}}(x_t) \log \frac{\hat{\pi}_t^{\mathrm{HMC}}(x_t)}{\hat{\pi}_t^{\mathrm{VMC}}(x_t)} dx_t \right\} dt \ ,$$

where $\hat{\pi}_t^{\mathrm{HMC}}$ and $\hat{\pi}_t^{\mathrm{VMC}}$ are the estimates of the marginal posterior at time t computed from samples x_t obtained by HMC and VMC algorithms respectively. It is clear that the smaller this KL-divergence estimate is, the more accurate is the algorithm. However a non-vanishing residual value of KL-divergence is expected even for two exact algorithms due to finite sample size. Therefore, the KL-divergence estimates are also computed for two independent sets of sample paths from HMC which are additionally used for convergence assessment of HMC.

As seen in previous sections, both HMC and VMC have several tuning parameters: the number of molecular dynamics steps J and the (fictitious) time increment $\delta\tau$ for HMC; sub-path length l used in block sampling, step size σ used in random walk sampling, and probability p used in the mixture transition kernel for VMC. As the computing time of a HMC algorithm increases with J, the optimal choice of J and τ is obtained by minimising the autocorrelation measure per computing time unit based on a set of pilot experiments. The corresponding acceptance rates vary between 60% and 70%. As reported in the literature, the optimal step size σ is chosen so that the algorithm considered has an acceptance rate between 20% and 40%. The sample principle applies to the choice of block size l. For the determination of the optimal p, the accuracy factor needs to be taken into account. In both extreme cases of a purely independence sampler based scheme (p = 0) and a purely random walk sampler based scheme (p = 1), the resulting estimated distribution of sample paths could be inaccurate given a sufficiently small sample.

Technical details of the numerical experiments are as follows. Both DW and SINE are discretised with time increment δt = 0.01. From pilot runs, the optimal settings for the HMC tuning parameters depend strongly on the choice of δt but varies little over different systems and different observation sets. This also applies to other tuning parameters. In the reported experiments J = 100, $\delta\tau$ = 0.01, l = 100, σ = 0.025, and p = 0.01 are the fixed values used throughout. All algorithms, including HMC, are initialised by a realisation of the approximate time-varying linear SDE from VGPA. For every data set, the amount of computing time is fixed to that of a HMC chain of 50,000 MH updates. All chains of both HMC and VMC are then sub-sampled to obtain 5,000 sample paths from each. The burn-in period is generally small as a result of good initialisation and the first 100 sample paths are discarded as burn-in. An exceptional case is the data set generated by SINE with ρ = 1 and R = 0.36. For this case, the mixing of the HMC algorithm is very poor and its burn-in is extremely large so it is necessary to run a chain of 5,000,000 MH-updates. The VMC chains are also adapted accordingly. All results are summarised in Tables 1 - 3. For illustration, Fig. 5, 7, 8 show three examples of both HMC and VMC results summarised by the mean paths, the 2 $\times$ standard deviations envelopes, and the decay of the auto-correlation function.

4.2 Discussion of numerical results

The results for DW with D = 0.25 are initially discussed. From the third and fourth columns of Table 1, it can be seen that the autocorrelation times of both MCMC algorithms, τ^{HMC} and τ^{VMC}, decrease with increasing observation densities, as might be expected. Mixing improves with reducing observation errors in line with expectations. For HMC, both conditions have the effect of increasing the information provided to the sampler from the gradients of log posterior which helps to more efficiently explore the posterior. However, the overall trend is much more marked for HMC than VMC. VMC can improve the mixing compared to HMC by an order of magnitude for low observation density and large observation error, i.e. ρ = 1 and R = 0.36. In contrast, both algorithms show a comparable efficiency in cases where the process is densely observed, i.e. ρ = 4, with lower observation errors i.e. R = 0.04 and R = 0.09. A reduction of auto-correlation times for VMC is observed with rates ranging between 3 to 8 for the observation sets between the above two extreme cases (for example, see Fig. 5). The VMC results are more stable across

the different Monte Carlo experiments as shown by the estimated standard deviation of the auto-correlation times.

The accuracy of VMC is indicated by the integrated marginal KL-divergence per time unit, KL_2, shown in the sixth column. A similar overall trend to that found in the autocorrelation times is observed. The KL-divergence estimates decrease with improving quality of observations. From the fifth column, KL_1, it seems that the residual (finite sample derived) KL-divergence value is not constant, fluctuating around 0.17 - 0.29 per time unit for different observation densities and sets. By comparing KL_1 and KL_2, we conclude that VMC does sample from the posterior exactly for the data sets with ($\rho = 2$, $R = 0.04$), ($\rho = 4$, $R = 0.04$), and ($\rho = 4$, $R = 0.09$). For other data sets, their non-zero KL-divergence values are statistically, but probably not practically, significant. This indicates that for finite sample sizes the corresponding estimated posterior distributions from VMC are slightly less accurate than those obtained from HMC. This is a result of the approximation employed in the VGPA. The VGPA clearly cannot capture non-Gaussian behaviour, and thus the approximation quality of the VGPA posterior varies, as shown by KL_3 (final column). It has been reported that the approximate posterior obtained from variational methods is often much more narrowly peaked compared to the true posterior. However such bias remains small except for the extreme case with infrequent and low quality observations ($\rho = 1$, $R = 0.36$), see Fig. 6. A range of visual diagnostics were explored to assess the reasons for the relatively poor accuracy in the VGPA and thus VMC. The main issue is in the tails of the distribution lending further support to the idea that the variational posterior is rather narrow, particularly in the regions where the true posterior is multi-modal. The presence of large observation errors leads to multi-modality of the posterior. The right panel in Fig. 6 shows that the marginal posterior seems to have three modes at $x = 0$ and $x = \pm 1$ in the transition phase ($t = 3.5$) whereas the approximate posterior from VGPA can clearly only possess a single mode, whose variance is underestimated.

For two examples of frequent transitions, i.e. DW with $D = 1.0$ and SINE with $D = 0.656$, similar results are observed, shown in Table 2 and Table 3. However, the KL-values have increased for both cases when compared with those of the previous example. The increase in residual KL-divergence values, KL_1, indicates that this is partially the result of the larger smoothing window used in the inference. Bearing this in mind, the results for DW with $D = 1.0$ are better than DW with $D = 0.25$ (compare Fig. 5 and Fig. 7). One factor that can explain this improvement is the increased diffusion coefficient ($D = 1.0$) which also drives the variationally optimised SDE and thus spreads the realisations from the variational diffusion bridge more widely in state space. It should also be noted that the VGPA results, KL_3, are also relatively improved in all cases, also helping the VMC sampler.

Compared to two DW examples, the accuracy of VMC has declined for SINE (Table 3). Recall that the SINE system has a series of meta-stable states which could make multi-modality problems more severe, and this is a problem for both HMC (with very slow mixing) and VMC (with relatively poor accuracy). However as the observation density and accuracy increases the VGPA attains good accuracy and the VMC works very well (see Fig. 8).

For all three systems, good accuracy and a significant improvement in mixing are achieved in the cases where observation densities ρ and observation errors R are either both low or both high. For higher ρ and lower R, a slight improvement in mixing is observed accompanied by a high accuracy. For lower ρ and higher R, an increase in mixing by an order of magnitude is seen but the accuracy is relatively poor. When compared with the approximation error arising from VGPA, however, the relative accuracy is still of great value, that is KL_2 is always a substantial reduction from KL_3. This means that the approximation error of VGPA is significantly reduced when employing a VMC run using only a fraction of the computing time needed for a HMC run. In practical applications this could be of great value in obtaining improved approximations in reasonable computational time. Thus the VMC method has applicability across a wide range of systems, and improves on the VGPA significantly in all cases.

5 Conclusions

This paper develops a novel MCMC algorithm which combines two particular MH algorithms, namely an independence sampler and a random walk sampler, with the recently developed VGPA for Bayesian inference in non-linear diffusions. This demonstrates that variational approximations can be combined with MCMC methods to good effect in path sampling. We stress that the variational approximation introduced in this paper is not the traditional fully factorising approximation, rather the approximation is over the joint posterior distribution of the state path, which makes the variational approximation an attractive proposal distribution for path sampling.

The basic implementation of the VMC sampling scheme is enhanced by introducing a flexible blocking strategy to improve the performance of the variational samplers. For path sampling, the implementation of blocking needs the simulation of a bridging process. The idea has already been applied to likelihood inference in diffusion processes (Durham and Gallant, 2002; Golightly and Wilkinson, 2006). However both previous papers made a relatively crude Gaussian approximation to the transition density of the process, based on a modified Brownian bridge scheme. To make proposals with reasonable acceptance probabilities the length of the blocking time interval is limited. In contrast, the novel bridging process herein is derived exactly from the approximate linear diffusion which has been optimised by the VGPA method. This sophisticated method of proposing sub-paths renders VMC an accurate and efficient sampling scheme.

As in the original VGPA framework, the VMC sampling algorithms apply only to a sub-class of diffusion processes where the diffusion term could be time-dependent but not state-dependent. In this sub-class, there are interesting systems to which the sophisticated framework of exact sampling, developed by Beskos et al (2006), doesn't apply, for instance, the stochastic double-well systems studied here.

The VMC algorithm adopts a mixture of an independence sampler and a random walk sampler. The switching between these two samplers is arranged in a probabilistic manner, with a pre-determined switching probability. In the VMC framework, the independence sampler has much better mixing than the random walk sampler so long as the former does not get stuck. If the chain becomes stuck, the random walk sampler will help to maintain mixing of the chain when it is chosen. Ideally, an adaptive mixture should be devised. Extensive experiments show that the problem of getting stuck is associated with those states which have a very low probability with respect to the proposal densities, but have relatively significant support under the true posterior. In order to trigger switching adaptively, a threshold of this probability could be determined based on a pilot run. As only information about the current state is used to make any decision, there is no risk of biasing the equilibrium distribution.

As seen in Sect. 4, the VMC algorithm outperforms HMC by mixing while obtaining good accuracy only when both observation densities and observation errors are moderate. If the observation density is low and the observation error is large, the marginal posterior clearly shows some multi-modal structure. This leads to a large approximation error in the VGPA, which in turn causes lower accuracy of VMC, although HMC methods also encounter problems with slow mixing in this case. A possible solution, which should be explored in future work, is to adopt a tempering strategy. Simply speaking, the true posterior needs to be tempered such that the VGPA can approximate the tempered posterior with desired accuracy. The essence of the idea is to use the diffusion variance D as the 'temperature' in a tempering scheme. The algorithm would be completed by constructing a ladder connecting the exact and approximate posteriors. We believe such a tempering scheme could maintain the efficiency of the VMC, while also improving the accuracy.

As illustrated in (Archambeau et al, 2008), the VGPA framework gives an upper bound of the marginal likelihood which can be used for approximate parameter estimation. A Langevin algorithm can be easily adopted to sample from the approximate marginal likelihood. The blocking strategy developed here could help improve the computational efficiency of such a parameter estimation algorithm. Also, in future work the variational sampling scheme could be embedded into a MCMC algorithm for parameter inference as the step for sampling missing paths.

We also believe these methods can be extended to cope with larger systems exploiting the variational approximation and could provide a framework for MCMC based inference in more complex, larger stochastic dynamic systems, where methods such as HMC become computationally prohibitive. The extension of the VMC algorithm to multivariate settings will require matrix inversion, of size equal to the state dimension, in the construction of the variational diffusion bridge and we are currently exploring efficient methods to implement this, including mean field approaches. In future work we plan to assess the ability of variational approximations to provide computationally efficient mechanisms for generating proposal distributions for blocked independence samplers, where we employ localisation in time and space to reduce the computational burden of sampling paths in very high dimensional spaces.

Acknowledgements

This work was funded by EPSRC as part of the VISDEM project (EP/C005848/1).

References

Alexander F, Eyink G, Restrepo J (2005) Accelerated Monte Carlo for optimal estimation of time series. J Stat Phys 119:1331–1345

Andrieu C, de Freitas D, Doucet A, Jordan M (2003) An introduction to MCMC for machine learning. Mach Learn 50:5–43

Andrieu C, Doucet A, Holenstein R (2010) Particle Markov Chain Monte Carlo methods. J R Statist Soc B 72:1–33

Archambeau C, Cornford D, Opper M, Shawe-Tayler J (2007) Gaussian Process approximations of stochastic differential equations. J Mach Learn Res Workshop and Conference Proceedings 1:1–16

Archambeau C, Opper M, Shen Y, Cornford D, Shawe-Tayler J (2008) Variational inference for diffusion processes. In: Platt C, Koller D, Singer Y, Roweis S (eds) Neural Information Processing Systems (NIPS), The MIT Press, Cambridge MA, vol 20, pp 17–24

Beskos A, Papaspiliopoulous O, Roberts GO, Fearnhead P (2006) Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. J R Statist Soc B 68:333–382

Derber J (1989) A variational continuous assimilation technique. Mon Wea Rev 117:2437–2446

Durham GB, Gallant AR (2002) Numerical techniques for maximum likelihood estimation of continuous-time diffusion process. J Bus Econom Stat 20:297–338

Elerian O, Chib S, Shephard N (2001) Likelihood inference for discretely observed nonlinear diffusions. Econometrica 69:959–993

Eraker (2001) Markov Chain Monte Carlo analysis of diffusion models with application to finance. J Bus Econ Statist 19:177–191

Evensen G (1994) Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J Geophys Res 99:10,143–10,162

Evensen G (2000) An ensemble Kalman smoother for nonlinear dynamics. Mon Wea Rev 128:1852–1867

Eyink GL, Restrepo JM, Alexander FJ (2004) A mean-field approximation in data assimilation for nonlinear dynamics. Physica D 194:347–368

de Freitas N, Hϕjen-Sϕrensen P, Jordan M, Russell S (2001) Variational MCMC. In: Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc. San Francisco, CA, pp 120–127

Golightly A, Wilkinson GJ (2006) Bayesian sequential inference for nonlinear multivariate diffusions. Stat Comput 16:323–338

Golightly A, Wilkinson GJ (2008) Bayesian inference for nonlinear multivariate diffusion models observed with error. Comput Stat Data Anal 52:1674–1693

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Honerkamp J (1994) Stochastic Dynamical Systems. VCH, Weinheim

Jazwinski AH (1970) Stochastic Processes and Filtering Theory. Academic Press

Julier SJ, Uhlmann J, Durrant-Whyte H (2000) A new method for the nonlinear tranformation of means and covariances in filters and estimators. IEEE Trans Autom Control 45:477–482

Kalman RE, Bucy R (1961) New results in linear filtering and prediction theory. J Basic Eng D 83:95–108

Kalnay E (2003) Atmospheric Modelling, Data Assimilation and Predictability. Cambridge University Press, Cambridge

Kitagawa G (1987) Non-Gaussian state space modelling of non-stationary time series. J Am Stat Assoc 82:503–514

Kitagawa G (1996) Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. J Comput Graph Stat 5:1–25

Klöden PE, Platen E (1992) Numerical Solution of Stochastic Differential Equations. Spinger, Berlin

Kushner HJ (1967) Dynamical equations for optimal filter. J Diff Eq 3:179–190

Liu JS (2001) Monte Carlo Strategies in Scientific Computing. Spinger, Berlin

Miller RN, Carter EF, Blue ST (1999) Data assimilation into nonlinear stochastic models. Tellus A 51:167–194

Ozaki T (1992) A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: A local linearization approach. Statistica Sinica 2:113–135

Pardoux E (1982) équations du filtrage non linéaire de la prédiction et du lissage. Stochastics 6:193–231

Rabier F, Jarvinen H, Klinker E, Mahfouf JF, Simmons A (2000) The ecmwf operational implementation of four-dimensional variational assimilation. part i: experimental results with simplified physics. Quart J Roy Met Soc 126:1143–1170

Roberts GQ, Stramer O (2001) On inferencee for partially observed non-linear diffusion models using Metropolis-Hasting algorithm. Biometrika 88:603–621

Stuart AM, Voss J, Winberg P (2004) Conditional path sampling of SDEs and the Langevin MCMC method. Comm Math Sci 2:685–697

Wan E, van der Merwe R (2001) The unscented Kalman filter. In: Haykin S (ed) Kalman Filtering and Neural Networks, Wiley, pp 207–219

Wilkinson D (2006) Stochastic Modelling for Systems Biology. Chapman & Hall/CRC

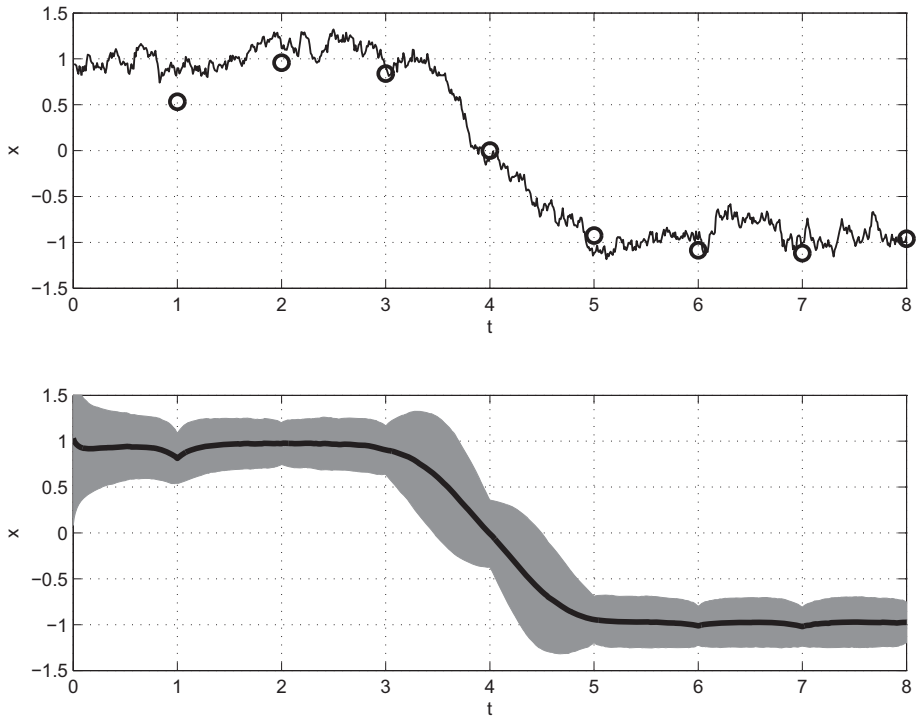

Figure 1: Upper panel: a typical realisation of a stochastic double-well system with small diffusion noise. The circles denote a set of observations obtained from this particular realisation; Lower panel: mean path (solid line) and its 2 standard deviation envelope estimated by extensive HMC sampling. For state estimation, the variance of diffusion noise is assumed known.

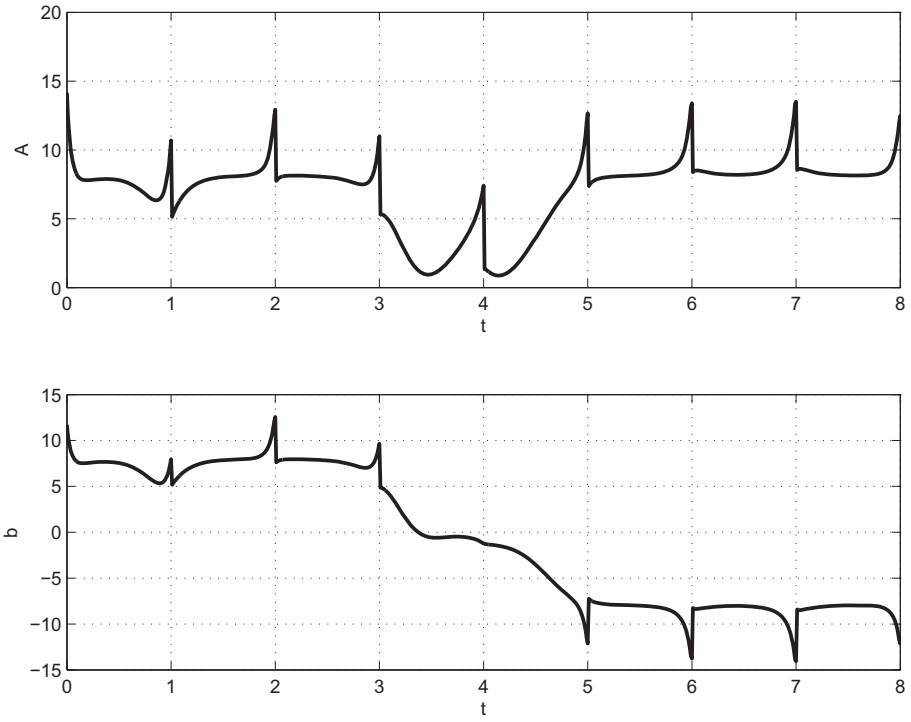

Figure 2: The temporal evolution of the trend $\mathbf{A}$ (upper panel) and offset $\mathbf{b}$ (lower panel) of the approximate diffusion process with its linear drift $f_L(\mathbf{x}) = -\mathbf{A}\mathbf{x} + \mathbf{b}$ which is optimised by the VGPA algorithm for the data set shown in Fig. 1.

Table 1: Comparison of the integrated autocorrelation times between two sets of sample paths from VMC and HMC, and comparisons of the integrated KL-divergence between two independent sets of sample paths from HMC (KL_1), between VMC and HMC (KL_2), and between the set of sample paths from HMC and the estimated marginal distributions of the state, x, from VGPA (KL_3). The results are obtained from a double-well system with diffusion coefficient with $D = 0.25$, with $T = 8$ and are shown as a function of observation density ρ^{obs} and error variance R. The variability of all above estimates are indicated by their standard deviations which were obtained by 20 Monte Carlo repetitions of the experiments.

ρ^{obs}	R	τ^{HMC}	τ^{VMC}	KL_1	KL_2	KL_3
1	0.04	3.67±0.37	1.31±0.22	0.20±0.01	0.41±0.09	6.19±0.67
	0.09	5.98±1.04	1.37±0.21	0.22±0.01	0.61±0.24	9.74±2.76
	0.36	24.23±6.54	1.91±0.46	0.29±0.02	2.46±0.83	49.46±16.72
2	0.04	1.49±0.24	1.37±0.34	0.19±0.01	0.27±0.10	2.49±0.62
	0.09	3.29±0.60	1.24±0.32	0.20±0.01	0.48±0.38	5.66±2.52
	0.36	10.77±3.37	1.61±0.27	0.24±0.01	0.98±0.32	16.88±5.38
4	0.04	1.11±0.12	1.32±0.27	0.17±0.01	0.20±0.01	1.40±0.19
	0.09	1.41±0.17	1.10±0.21	0.18±0.01	0.21±0.02	2.09 ±0.31
	0.36	5.48±1.12	1.43±0.19	0.22±0.02	0.91±0.85	12.13 ±8.25

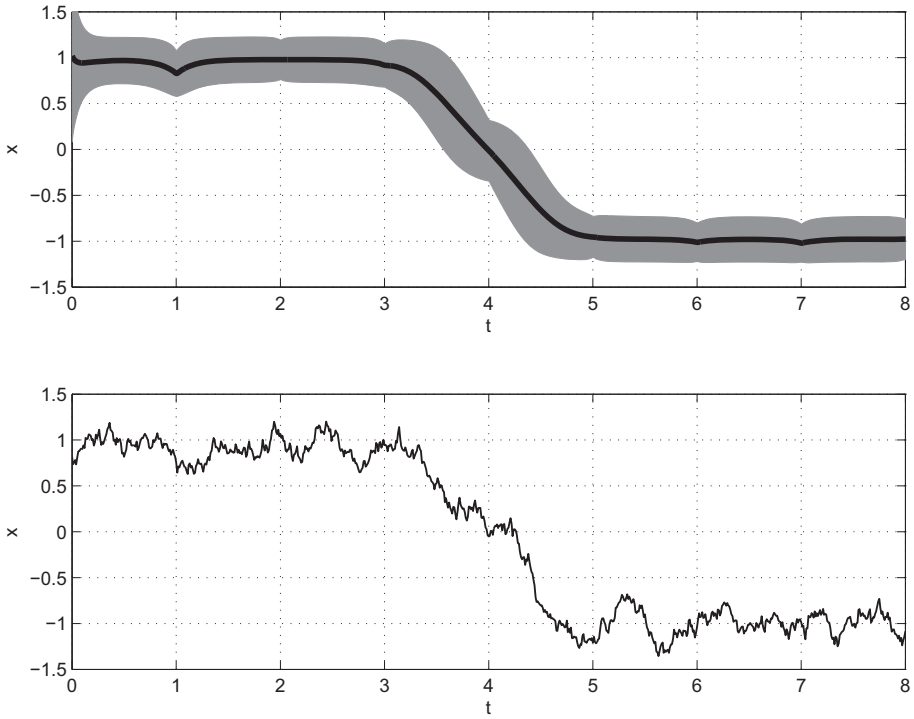

Figure 3: Upper panel: the mean path (solid line) and its 2 standard deviation envelope obtained from the VGPA for the data set shown in Fig. 1; Lower panel: a typical realisation from the VGPA posterior.

Table 2: As Table 1 but for a double-well system with diffusion coefficient with $D = 1.0$, $T = 50$.

ρ^{obs}	R	τ^{HMC}	τ^{VMC}	(KL_1)	(KL_2)	(KL_3)
1	0.04	3.64	1.70	0.41	0.59	9.18
	0.09	6.73	1.57	0.43	0.69	12.67
	0.36	24.36	2.17	0.50	1.70	55.44
2	0.04	1.31	1.26	0.36	0.38	2.77
	0.09	3.38	1.31	0.39	0.41	4.14
	0.36	9.70	1.78	0.44	0.91	14.55
4	0.04	1.02	1.16	0.31	0.33	1.00
	0.09	1.17	1.27	0.34	0.36	1.24
	0.36	2.96	1.17	0.39	0.43	5.22

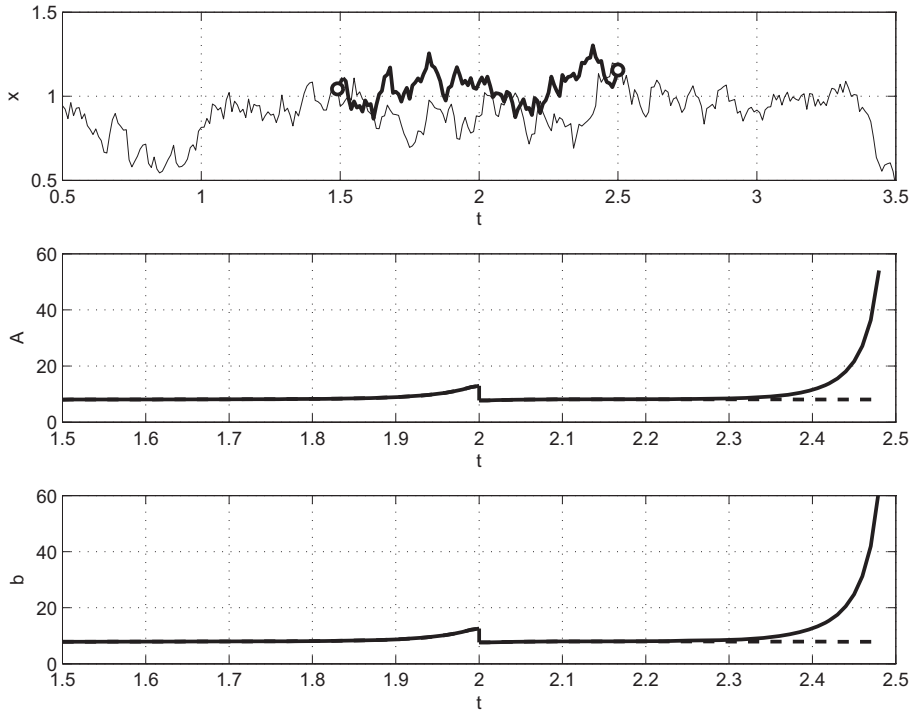

Figure 4: An illustration of the blocking strategy for path sampling. From top to bottom: choosing a sub-path for updating, whose ends are indicated by black circles. A proposal of this sub-path is highlighted by the thick black line; computing the effective trend $\mathbf{A}$ and offset $\mathbf{b}$ (solid lines) conditioned on the state, $\mathbf{x}$, at both ends of the block, compared to the original $\mathbf{A}$ and offset $\mathbf{b}$ from the VGPA (dashed lines).

Table 3: As Table 1 but for a SINE-drift system with diffusion coefficient with $D = 0.656$, $T = 50$. The figures with * are obtained from the chains which are sub-sampled with a fixed between-sample interval 10 times longer than other figures, due to the extremely poor mixing of HMC for this particular example. These two τ-estimates are scaled accordingly, which makes the values more noisy. We believe the τ^{VMC}-value is closer to 1.

ρ^{obs}	R	τ^{HMC}	τ^{VMC}	(KL_1)	(KL_2)	(KL_3)
1	0.04	18.88	1.07	0.57	1.75	61.82
	0.09	28.74	1.10	0.46	1.22	96.64
	0.36	75.7^*	2.5^*	0.49	6.95	393.64
2	0.04	4.43	1.10	0.26	0.94	13.90
	0.09	15.18	0.91	0.45	1.17	43.58
	0.36	15.51	0.77	0.47	4.83	163.44
4	0.04	1.10	1.05	0.23	0.32	4.15
	0.09	1.84	1.01	0.25	0.37	6.93
	0.36	7.59	0.97	0.33	2.93	102.42

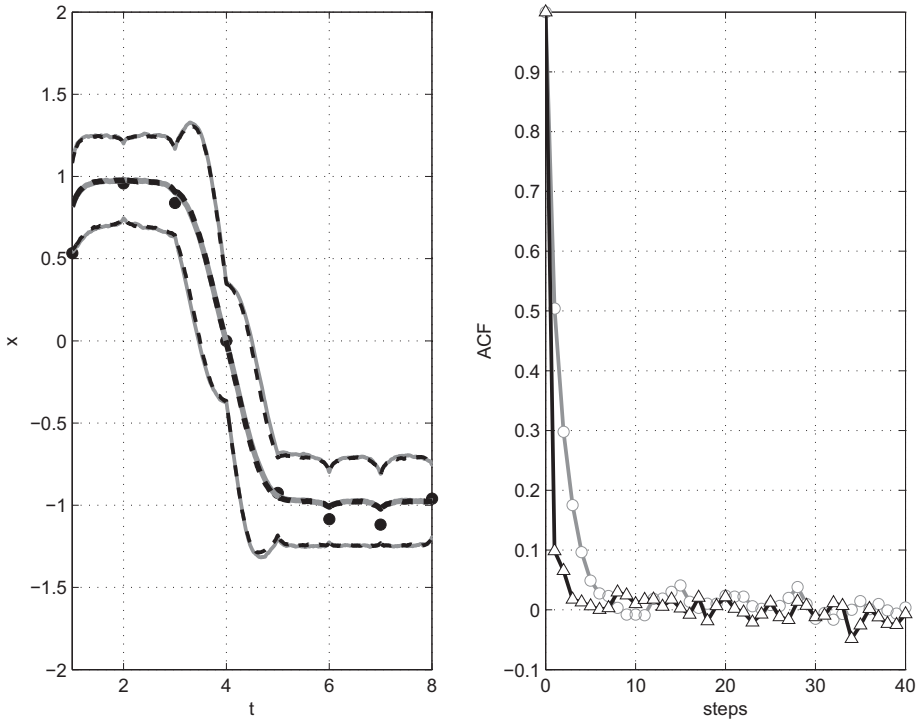

Figure 5: Comparison of marginal estimates (left) and mixing properties (right) between VMC (black) and HMC (grey) for a double-well system with diffusion coefficient $D = 0.25$ and time window $T = 8$. Left panel: the estimate of mean path and its ± 2 standard deviation envelopes for HMC and VMC. The dots denote the observations, at a density of one per time unit, and the observation error variance, $R = 0.04$. Right panel: decay of the autocorrelation function of time series of the summary statistic L, VMC (black, triangles) and HMC (grey, circles) (see text).

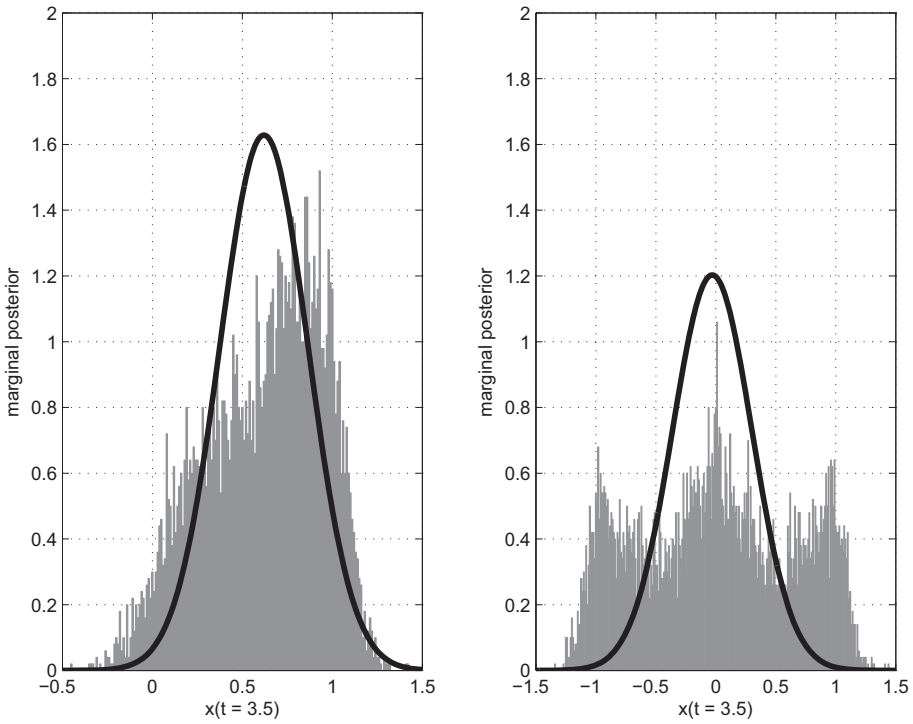

Figure 6: Comparison of HMC (Grey) and VGPA (Black) estimate of the marginal posterior of the state $\mathbf{x}$ at time $t = 3.5$ for the examples shown in Fig. 5 and an extreme example with high observation noise that shows a multi-modal marginal posterior (left and right panel, respectively).

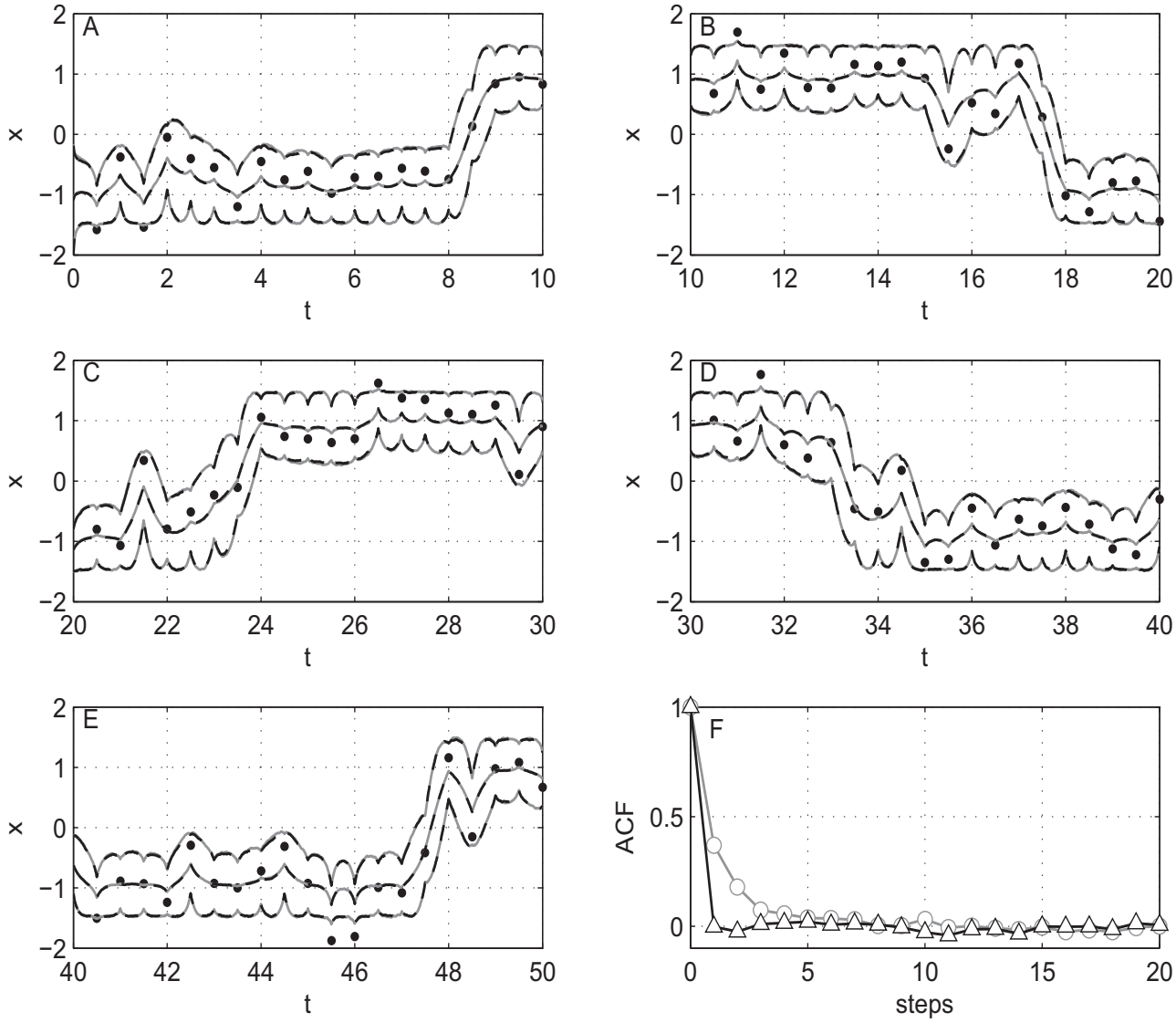

Figure 7: As Fig. 5 but for a double-well system with diffusion coefficient D = 1.0 and time window T = 50. Observations are made with density two per time unit and the observation error variance R = 0.09. For clarity, the marginal estimates and observations are displayed in separate panels (A - E) for 5 consecutive time intervals of length 10 units.

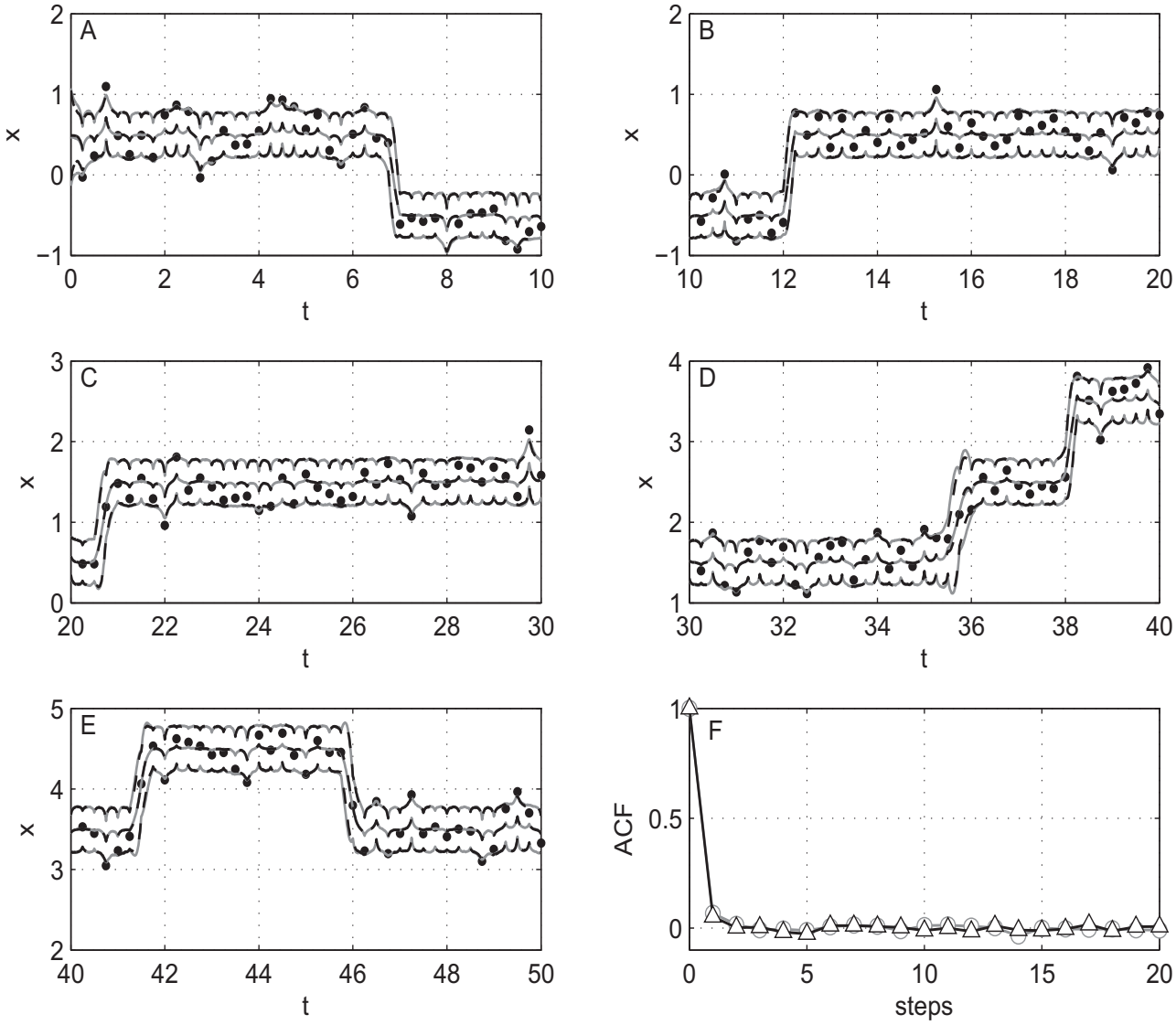

Figure 8: As Fig. 5 but for a SINE-drift system with diffusion coefficient D = 0.656 and time window T = 50. Observations are made with density four per time unit and the observation error variance R = 0.04. For clarity, the marginal estimates and observations are displayed in separate panels (A - E) for 5 consecutive time intervals of length 10 units.