A comparison of variational and Markov Chain Monte Carlo methods for inference in partially observed stochastic dynamic systems.

Yuan Shen
Neural Computing Research Group
Aston University
Birmingham, United Kingdom
sheny2@aston.ac.uk

Cedric Archambeau
Department of Computer Science
University College London
London, United Kingdom
c.archambeau@cs.ucl.ac.uk

Dan Cornford
Neural Computing Research Group
Aston University
Birmingham, United Kingdom
d.cornford@aston.ac.uk

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

John Shawe-Taylor
Department of Computer Science
University College London
London, United Kingdom
jst@cs.ucl.ac.uk

Remi Barillec
Neural Computing Research Group
Aston University
Birmingham, United Kingdom
barillrl@aston.ac.uk

March 16, 2009

Abstract

In recent work we have developed a novel variational inference method for partially observed systems governed by stochastic differential equations. In this paper we provide a comparison of the Variational Gaussian Process Smoother with an exact solution computed using a Hybrid Monte Carlo approach to path sampling, applied to a stochastic double well potential model. It is demonstrated that the variational smoother provides us a very accurate estimate of mean path while conditional variance is slightly underestimated. We conclude with some remarks as to the advantages and disadvantages of the variational smoother.

1 Introduction

Stochastic dynamical systems [1] have been used for modelling of real-life systems in various areas ranging from physics [1] to system biology [2] to environmental science [3]. Such systems are often only partially observed, which makes statistical inference in those systems difficult. The inference problem for stochastic dynamical systems usually includes both state- and parameter estimation. In this paper, we focus on state estimation and assume that the system equation and its parameters are both known a priori. This is known as filtering and/or smoothing problems in statistical signal processing [4]. It is known that the Kushner-Stratonovich-Pardoux (KSP) equations are the optimal solution to a general filtering/smoothing problem [5, 6, 7]. For linear systems, the filtering part of KSP equations is reduced to the well-known Kalman-Bucy filter[8] which is computationally very efficient. For non-linear dynamics in general, however, filtering/smoothing is still a challenging problem because a numerical solution to the KSP equations is not feasible for high-dimensional systems. Recently, a variational smoothing algorithm has been proposed in [10]. This paper is to illustrate the performance of that computationally efficient algorithm by comparing with Markov Chain Monte Carlo (MCMC) smoother.

Mathematically, a stochastic dynamical system is often represented by stochastic differential equation (SDE) [11]:

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

where $\mathbf{X}(t) \in \mathcal{R}^d$ is state vector, $\mathbf{D} \in \mathcal{R}^{d \times d}$ is so-called diffusion matrix, $\mathbf{f}$ represents a deterministic dynamical process. The driving noise process is represented by a Wiener process $\mathbf{W}(t)$. Eq. 1 is also referred to as a diffusion process. Note that the diffusion matrix $\mathbf{D}$ is assumed to be state-independent. The state is observed via some measurement function $\mathbf{H}(\cdot)$ at discrete times, say $\{t_k\}_{k=1,...,M}$. The observations are contaminated by i.i.d Gaussian noise:

$$\mathbf{y}_k = \mathbf{H}(\mathbf{X}(t_k)) + \mathbf{R}^{\frac{1}{2}} \cdot \xi_k \tag{2}$$

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 ξ_k represents multivariate white noise. A Bayesian approach to filtering/smoothing is typically adopted in which the posterior distribution $p(\mathbf{X}(t)|\{\mathbf{y}_1, ..., \mathbf{y}_k, t_k < t\})$ and $p(\mathbf{X}(t)|\{\mathbf{y}_1, ..., \mathbf{y}_M\})$, respectively, are to be formulated and estimated. Theoretically, an optimal estimate of $p(\cdot)$ is the solution to the corresponding KSP equations. Computational approaches are either based on a variety of approximation schemes or achieved by MCMC sampling methods.

Using Markov Chain Monte Carlo [12], one is able to sample from a posterior distribution exactly. At each step of a MCMC simulation, a new state is proposed and will be accepted or rejected in a probabilistic way. For applications to stochastic dynamical systems, it is also referred to path sampling. A path sampling approach to discrete-time state-space models has been addressed in [9] and many references therein. In those works, a Gibbs-sampler with single-site update was used. To achieve better mixing, several algorithms using multiple-site update are explored in [13]. Recently, a Hybrid Monte Carlo (HMC) algorithm for path sampling is proposed in [14]. The HMC method updates the entire sample path at each step of path sampling while keeping the acceptance of new paths high. In this work, we first scrutinise the use of HMC for non-linear smoothing and then assess the performance of the variational smoother proposed in [10] by comparing its results with those of HMC.

In contrast to MCMC, all other approaches to non-linear filtering/smoothing, including the one proposed in [10], are based on a particular approximation scheme. The extended Kalman filter is the first attempt to tackle the non-linearity by linearising the dynamics around the currently available state estimate [15]. However, unstable error growth is observed in such linearisation methods [16]. To alleviate this difficulty, the Ensemble Kalman Filter (EnKF) was introduced in [17]. An ensemble of states are integrated forward in time. Therefore, the Kalman gain can be estimated by using the error covariances which are not propagated but calculated from the ensemble of states at each time step. Note that this method drops the linear approximation of non-linear dynamics while keeping the Gaussian assumption of error statistics. Particle filter (PF) proposed in [18] represents a different direction of approximation strategies. Essentially, the posterior density of filtering variables in PF is approximated by a discrete distribution with random support. Each one in the discrete support is called particle and its probability mass is considered as weight.

It will be seen that the approximation strategy implemented in [10] is distinct from those in the above methods.

In essence, the variational smoother in [10] makes a global linear approximation of non-linear dynamics. This implies a Gaussian approximation of the posterior process

$$p(\mathbf{X}(t)|\{\mathbf{y}_1, ..., \mathbf{y}_M\}).$$

The quality of approximation is measured by Kullback-Leibler (KL) divergence [19] between the true and approximate posterior, and the optimal approximate posterior is obtained by minimising the KL divergence. Following this, any statistical inference in the true system is based on the approximate posterior. This method is within the framework of variational approximation for Bayesian inference, which is computationally very efficient and popular in machine learning community [20].

The structure of this paper is as follows: First, we present a Bayesian treatment of non-linear smoothing. In Sec. 3, the MCMC method is described in details while we give a summary of the variational smoother in Sec. 4. For detailed proofs, we refer to [10]. After that, we compare both methods in Sec. 5 by numerical experiments with a double-well potential system. The paper concludes with a discussion.

2 Bayesian approach to non-linear smoothing

Both for the MCMC method in [14] and for the variational smoother in [10], stochastic differential equations are discretized by using an explicit Euler-Maruyama scheme [11]. The discretized version of Eq. 1 is given by

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

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. An initial state $\mathbf{x}_0$ needs to be set. There are M observations within the smoothing window, and they are denoted by

$$(t_{k_j}, \mathbf{y}_j)_{j=1,...,M} \quad \text{with} \quad \{t_{k_1}, ..., t_{k_M}\} \subseteq \{t_0, ..., t_N\}.$$

In the following, we formulate the posterior distribution step by step.

The prior of a diffusion process can be written down 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 of initial states and $p(\mathbf{x}_{k+1}|\mathbf{x}_k)$ with $k = 0, ..., N-1$ are transition densities of the diffusion process. For small enough δt, those transition densities can be well approximated by a Gaussian density [21]. Accordingly,

$$p(\mathbf{x}_{k+1}|\mathbf{x}_k) = \mathcal{N}(\mathbf{x}_{k+1}|\mathbf{x}_k + \mathbf{f}(\mathbf{x}_k)\delta t, 2\mathbf{D}\delta_t).$$

Therefore, the prior is given by

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

where

$$\mathcal{H}_{\text{dynamics}} = \sum_{k=0}^{N-1} \quad \frac{\delta t}{4} \left[\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\delta t} - \mathbf{f}(\mathbf{x}_k, t_k) \right]^\top \mathbf{D}^{-1} \left[\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\delta t} - \mathbf{f}(\mathbf{x}_k, t_k) \right].$$

As we assume that measurement noises are i.i.d. Gaussian random variables, the likelihood is simply given by

$$p(\mathbf{y}_1, ..., \mathbf{y}_M | \mathbf{x}_0, ..., \mathbf{x}_N) = \exp(-\mathcal{H}_{\text{obs}}),$$

where

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

In summary, we have the posterior

$$p(\mathbf{x}_0, ..., \mathbf{x}_N | \{\mathbf{y}_1, ..., \mathbf{y}_M\}) \propto p(\mathbf{x_0}) \cdot \exp(-1(\mathcal{H}_{\text{dynamics}} + \mathcal{H}_{\text{obs}})).$$

3 Markov Chain Monte Carlo (MCMC) Smoother

In Hybrid Monte Carlo, the molecular dynamics simulation algorithm is applied to make proposals in a Metropolis-Hastings 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 $\mathcal{X}^{k+1}$, one simulates a fictitious deterministic system 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 non-linear smoothing in Sec. 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 leapfrog as follows:

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

where A denotes so-called preconditioning matrix which accelerates the convergence of matrix iterations. Further, the matrix A is a circulant matrix which is constructed from the vector

$$\{1, \exp(-\alpha), ..., \exp(-\alpha \cdot T)\}$$

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

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

A reasonably high acceptance rate can be achieved by tuning the parameter $\delta\tau$, J and α. If $\delta\tau$ is too large, then the leapfrog algorithm gives us a poor approximation to the true dynamics of the fictitious system. If J is too large, small discretisation errors could be accumulated so that the simulated trajectory shifts away from the true one. Both lead to low acceptance rate. With too small $\delta\tau$ and J, however, the change of sample paths at each step is too small to improve mixing significantly.

4 Variational Gaussian Process Approximation (VGPA) Smoother

The starting point of the variational Gaussian Process approximation method is to approximate Eq. 1 by a linear SDE:

$$d\mathbf{X}(t) = \mathbf{f}_L(\mathbf{X}, t)dt + (2\mathbf{D})^{1/2}(t)d\mathbf{W}(t), \tag{5}$$

where

$$\mathbf{f}_L(\mathbf{X}, t) = -\mathbf{A}(t)\mathbf{X}(t) + \mathbf{b}(t). \tag{6}$$

Note that $\mathbf{D}$ must not be state-dependent so that $\mathbf{X}(t)$ of the approximate SDE is a Gaussian process. The matrix $\mathbf{A}(t) \in \mathcal{R}^{d\times d}$ and the vector $\mathbf{b}(t) \in \mathcal{R}^d$ are two variational parameters to be optimised.

The approximation made by Eq. 6 implies that the true posterior process, i.e. $p(\mathbf{X}(t)|\mathbf{y}_1, ..., \mathbf{y}_M)$ and say $p(t)$, is approximated by a Gaussian Markov process, say $q(t)$. If we discretise the linear SDE in the same way as the true SDE, the approximate posterior can be written down as

$$q(\mathbf{x}_0, ..., \mathbf{x}_N) = q(\mathbf{x}_0) \cdot \prod_{k=0}^{N-1} \mathcal{N}(\mathbf{x}_{k+1}|\mathbf{x}_k + \mathbf{f}_L(\mathbf{x}_k)\delta t, 2\mathbf{D}\delta_t).$$

In [10], $q(\mathbf{x}_0)$ is fixed to $\mathcal{N}(\mathbf{x}_0|\mathbf{m}_0, \mathbf{S}_0)$, and the prior on initial states $p(\mathbf{x}_0)$ is a uniform distribution.

The optimal $\mathbf{A}(t)$ and $\mathbf{b}(t)$ are obtained by minimising the KL divergence of $q(t)$ and $p(t)$ which is given by

$$\mathrm{KL}[q||p] = \int dq \ln \frac{dq}{dp} = \int_0^T E(t)dt + const. \tag{7}$$

with $E(t) = E_{sde}(t) + E_{obs}(t)$, $E_{obs}(t) = \left\langle \mathcal{H}^{obs} \right\rangle_{q_t}$ and

$$E_{sde}(t) = \frac{1}{4} \left\langle \mathbf{f}(\mathbf{X}) - \mathbf{f}_L(\mathbf{X}))^\top \mathbf{D}^{-1} (\mathbf{f}(\mathbf{X}) - \mathbf{f}_L(\mathbf{X})) \right\rangle_{q_t}$$

where $\mathcal{H}^{obs}$ is defined in Eq. 4 and q_t denotes the marginal distribution of the approximate posterior process $q(t)$ at time t.

To compute the KL divergence, we introduce two auxiliary variational parameters $\mathbf{m}(t)$ and $\mathbf{S}(t)$ which are the mean and covariance matrix of the marginal distribution q_t. However, the pair $(\mathbf{m}(t), \mathbf{S}(t))$ is not independent of $(\mathbf{A}(t), \mathbf{b}(t))$. There exists two constraints between them:

$$\frac{d\mathbf{m}(t)}{dt} = -\mathbf{A}(t)\mathbf{m}(t) + \mathbf{b}(t), \tag{8}$$

and

$$\frac{d\mathbf{S}(t)}{dt} = -\mathbf{A}(t)\mathbf{S}(t) - \mathbf{S}(t)\mathbf{A}^\top(t) + 2\mathbf{D}. \tag{9}$$

Accordingly, we find optimal ($\mathbf{A}(t)$, $\mathbf{b}(t)$), ($\mathbf{m}(t)$, and $\mathbf{S}(t)$) by looking for the stationary points of the following Lagrangian

$$\begin{aligned} \mathcal{L} = \int \quad & \{E - tr\{\mathbf{\Psi}(\frac{\mathbf{dS}}{\mathbf{dt}} + \mathbf{AS} + \mathbf{SA}^\top - \mathbf{2D})\} \\ & -\lambda^\top(\frac{d\mathbf{m}}{dt} + \mathbf{Am}) - \mathbf{b}\}dt \end{aligned}$$

where $\mathbf{\Psi}(t) \in \mathcal{R}^{d\times d}$ and $\lambda(t) \in \mathcal{R}^d$ are Lagrange multipliers. By definition, $\mathbf{\Psi}(T) = 0$ and $\lambda(T) = 0$.

By taking the derivatives of $\mathcal{L}$ with respect to $\mathbf{m}$, $\mathbf{S}$, $\mathbf{A}$ and $\mathbf{b}$, we obtain the following Euler-Lagrange equations:

$$\frac{\partial E}{\partial \mathbf{A}} - 2\mathbf{\Psi S} - \lambda \mathbf{m}^\top = 0 \tag{10}$$

$$\frac{\partial E}{\partial \mathbf{b}} + \lambda = 0 \tag{11}$$

$$\frac{\partial E}{\partial \mathbf{m}} - \mathbf{A}^\top \lambda + \frac{d\lambda}{dt} = 0 \tag{12}$$

$$\frac{\partial E}{\partial \mathbf{S}} - 2\mathbf{\Psi A} + \frac{d\mathbf{\Psi}}{dt} = 0 \tag{13}$$

Note that the optimal $\mathbf{m}$, $\mathbf{S}$, $\mathbf{A}$, $\mathbf{b}$, $\mathbf{\Psi}$ and λ should fulfil the above equations and Eq. (8-9) as well. Hence, the non-linear smoothing problem is reduced to solving a system of first-order differential equations.

The equation system above is solved iteratively. We start with an initial guess of $\mathbf{m}$, $\mathbf{S}$, $\mathbf{A}$, $\mathbf{b}$, $\mathbf{\Psi}$ and λ. First, we compute $\mathbf{m}(t)$ and $\mathbf{S}(t)$ by performing standard Gaussian Process regression [22]. Then, we set $\mathbf{\Psi}(t) = 0$ and $\lambda(t) = 0$ for all t. Finally, $\mathbf{A}$ and $\mathbf{b}$ are initialised by

$$\mathbf{A}(t) = \left\langle \frac{\partial \mathbf{f}}{\partial \mathbf{X}} \right\rangle_{q_t} + \mathbf{D}\mathbf{\Psi}(t) \tag{14}$$

$$\mathbf{b}(t) = < \mathbf{f}(\mathbf{X}) >_{q_t} + \mathbf{A}(t)\mathbf{m}(t) - 2\mathbf{D}\lambda(t). \tag{15}$$

Note that Eq. (14-15) are derived from Eq. (10-11).

At iteration i, we first update $\mathbf{m}$ and $\mathbf{S}$ by solving Eq. (8-9) forward in time where $\mathbf{A}^i$ and $\mathbf{b}^i$ are used. Next, $\mathbf{\Psi}$ and λ are updated by solving Eq. (12-13) with final condition $\mathbf{\Psi}(T) = 0$ and $\lambda(T) = 0$ where $\mathbf{m}^{i+1}$ and $\mathbf{S}^{i+1}$ are used. Note that the data are assimilated at this step. To clarify it, this can be split into two steps:

1. Between two successive observations, we update $\mathbf{\Psi}$ and λ by solving

$$\frac{d\mathbf{\Psi}(t)}{dt} = 2\mathbf{\Psi}(t)\mathbf{A}(t) - \frac{\partial E_{sde}}{\partial \mathbf{S}} \tag{16}$$

$$\frac{d\lambda(t)}{dt} = \mathbf{A}^\top(t)\lambda(t) - \frac{\partial E_{sde}}{\partial \mathbf{m}} \tag{17}$$

2. When there is an observation at t_{k_j}, $j = 1, ..., M$, the following jump-conditions apply

$$\mathbf{\Psi}(t_{k_j}^+) = \mathbf{\Psi}(t_{k_j}^-) - \frac{1}{2}\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H} \tag{18}$$

$$\lambda(t_{k_j}^+) = \lambda(t_{k_j}^-) + \mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}(\mathbf{y}_j - \mathbf{H}\mathbf{m}(t_{k_j})). \tag{19}$$

Finally, we compute

$$\mathbf{A}(t; \mathbf{m}^{i+1}, \mathbf{S}^{i+1}, \mathbf{\Psi}^{i+1}, \lambda^{i+1})$$

and

$$\mathbf{b}(t; \mathbf{m}^{i+1}, \mathbf{S}^{i+1}, \mathbf{\Psi}^{i+1}, \lambda^{i+1})$$

by using Eq. (14-15). To keep the algorithm stable, the update of $A(t)$ and $b(t)$ is done by

$$\begin{aligned} \mathbf{A}^{i+1}(t) &= \mathbf{A}^i(t) \\ &- \omega\{\mathbf{A}^i(t) - \mathbf{A}(t; \mathbf{m}^{i+1}, \mathbf{S}^{i+1}, \mathbf{\Psi}^{i+1}, \lambda^{i+1})\} \\ \mathbf{b}^{i+1}(t) &= \mathbf{b}^i(t) \\ &- \omega\{\mathbf{b}^i(t) - \mathbf{b}(t; \mathbf{m}^{i+1}, \mathbf{S}^{i+1}, \mathbf{\Psi}^{i+1}, \lambda^{i+1})\} \end{aligned}$$

where $0 < \omega < 1$. The iteration stops when $\mathcal{L}$ has converged.

5 Numerical Experiments

The MCMC and variational algorithms are compared on a double-well potential system which is given by

$$\dot{x}(t) = f(x(t)) + \kappa \cdot \xi(t), \tag{20}$$

where

$$f(x) = 4x(1 - x^2)$$

and $\xi(t)$ is white-noise [1]. The parameter κ corresponds to $(2\mathbf{D})^{\frac{1}{2}}$ in Eq. 1 and determines the strength of random fluctuations within the system. This system has two stable states, namely $x = +1$ and $x = -1$. However, random fluctuations could cause a transition of the system from one stable state into another. The average time needed for the occurrence of such an event is called exit time [1]. In this study, we set $\kappa = 0.5$ and the corresponding exit time is about 4000 time units [23]. This provides us some prior knowledge on initial states.

In the numerical experiments, we consider a smoothing window ranging from $t = 0$ to $t = 12.0$. Further, we assume that states x can be observed directly, which makes $\mathbf{H}$ an identity function. Within the smoothing window, we generate three data sets, say A, B, and C, from a sample path which was considered in [23] and [25]. The variance R of measurement errors are 0.04, 0.09, and 0.36, respectively. Each data set consists of seven data points which are "measured" at times $t_{k_1} = 1.0$, ..., $t_{k_M} = 7.0$. Although multiple data sets are generated and analysed for each of those R-values, the results of data set A, B, and C are representative and chosen for illustration.

For the MCMC method, Eq. 20 is discretized with time increment $\delta t = 0.1$. The prior on initial states is set to a Gaussian density with mean at $x = +1$ and variance equal to 0.05. This choice is strongly based on our prior knowledge of the system. The tuning parameters of Hybrid Monte Carlo are chosen as follows: $J = 2$, $\delta\tau = 0.005$ and $\alpha = 0.02$. The use of preconditioning matrix A keeps the necessary J small, which makes the simulation computationally more efficient. However, the multiplication of the matrix A with various vectors would cost extra computational time. Because of the circulant property of A, this part of computational burden can be reduced [24].

For each of 3 data sets, we run a Markov chain of length 5,000,000 and subsample from this chain with sampling interval equal to 1,000. The first 1,000 samples are discarded as burn-in period. It turns out that it is insufficient to determine burn-in only by monitoring a summary statistic like energy $\hat{\mathcal{H}}$. On the contrary, one has to monitor the traces of state x at different time

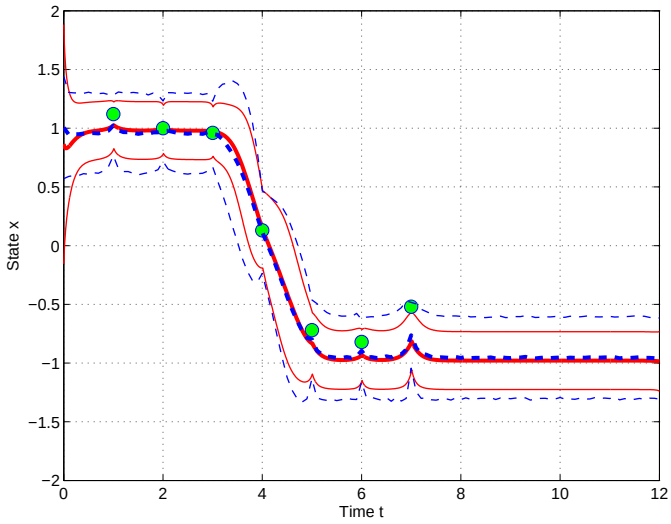

Figure 1: Comparison of mean-path and conditional-variance estimates between the MCMC- (dashed) and variational (solid) method with a double-well potential system. Filled circles represent 7 observations from data set A, with measurement noise variance equal to 0.04. The mean paths are displayed by thick lines, while each pair of thin lines indicates an envelope of mean path with 2 $\times$ standard deviation.

points. Particularly, those time points must be chosen from different phases of the smoothing window, for example, transition phases, stationary phases, and the phase before/after the first/last observation.

For the variational GP approximation method, Eq. 20 is discretized with time increment $\delta t = 0.01$. The only tuning parameter ω is set to 0.15 as in [10]. The number of iterations required for the convergence of VGPA may increase when we extend the smoothing window or add more measurement noise. This is because of the poor initial states estimated by standard GP regression.

In Fig. 1, Fig. 2 and Fig. 3, the estimates of both mean path and conditional variance are displayed for data set A, B, and C, respectively. In each figure, the results of VGPA are compared with the MCMC results.

For the data sets with relatively small measurement noise, the estimated mean paths of both methods agree with each other very well whereas the estimated conditional covariance of VGPA is overall but only slightly smaller than that of MCMC. It is also seen that the estimated mean path is slightly biased towards zero during both stationary phases. This can be explained by the fact that although the posterior of x has a distinct mode at $x = +1.0$ before the

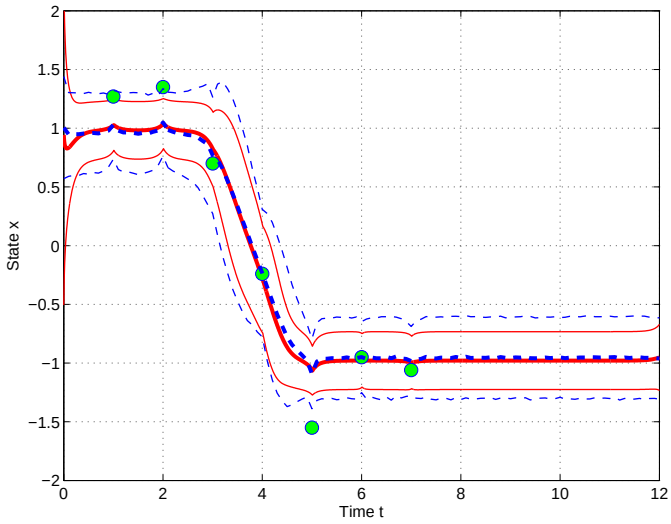

Figure 2: The same as in Fig. 1 but with data set B (measurement noise variance $= 0.09$).

transition or $x = -1.0$ after the transition, the mode at another stable state is not vanishing. Moreover, we see that the mean path stays at the left well after the last observation. This is also in accordance with the large exit time of the system we consider.

A small dip of mean paths is evident in the results of the VGPA smoother when we look into the initial period of the smoothing window. This is accompanied with large conditional variance $\mathbf{S}(0)$. To explain this observation, we run MCMC simulations with increasingly larger prior variance of initial states. As expected, the posterior variance of $\mathbf{x}_0$ increases with its prior variance. Further, it turns out that a similar dip of mean paths appears when the prior variance becomes sufficiently large. It can be understood as follows: Without any data, a double-well systems does show a bimodal probabilistic structure. With a posterior mean of $\mathbf{x}_0$ close to $+1$ and a large value of its posterior variance, the mean path could be further biased towards zero in the initial period where the first observation has little influence. Note that the approximate posterior variance $\mathbf{S}(0)$ is not optimised, but held fixed.

Finally, we turn our attention to data set C with very large measurement noise. Note that it is difficult to identify where the transition starts by visual inspection of the data themselves. In contrast, this is possible with data set A and B. From Fig. 3, we can see that there is significant difference both in mean path and in conditional variance between the MCMC and VGPA smoother, particularly in the period before $t = 5.0$. Due to the ambiguity shown by the

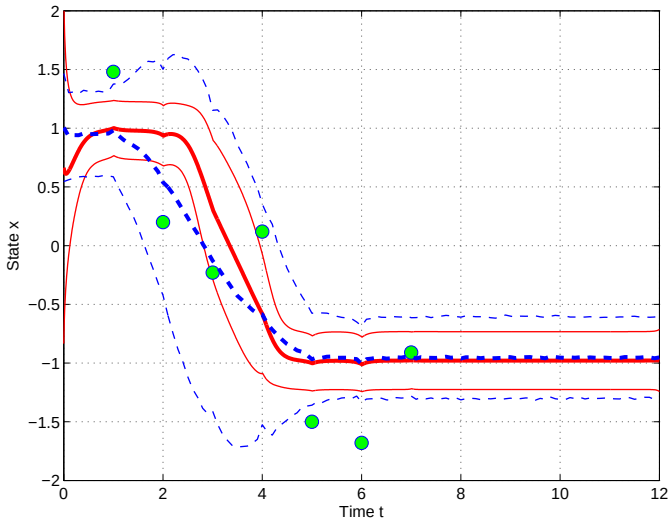

Figure 3: The same as in Fig. 1 but with data set C (measurement noise variance = 0.36).

data between $t = 2.0$ and $t = 4.0$, the MCMC sampler seems to be exploring the bimodal structure of the posterior distribution. In contrast, the approximate posterior of the VGPA method is fixed to one particular mode at any time. This may explain the difference in mean path between two methods and a significant underestimation of conditional variance for the VGPA smoother.

6 Discussions

By comparing with Markov Chain Monte Carlo , we scrutinise a variational method for non-linear smoothing which is recently proposed in [10]. Both methods are tested on a double-well potential system. Three data sets with different measurement noise are used to find out the strength and weakness of the novel smoother.

Our investigation is based on the fact that MCMC methods provide an exact inference tool for comparison. For data sets with small or moderate measurement noise, it turns out that the VGPA method does produce a very accurate estimate of mean path while the conditional variance is slightly under-estimated. As expected, the variational method is computationally more efficient than MCMC. Regarding other approximation-based smoothers, it has been reported that Ensemble Kalman smoother fails to reconstruct the transition of a double-well system accurately from a sparse data set [25]. As stated in [25], the failure is due to the fact that in KF and EnKF the propagated states are corrected by a linear interpolation scheme when new data are assimilated.

However, the weakness of the VGPA method is also evident when the ambiguity of data becomes significant. As many other variational approximation methods, the novel smoother is not good at exploring the multi-mode structure of some probability measures. In this paper, the role of prior on initial state is also investigated. It turns out that the estimates of mean path could be biased in the initial phase of the smoothing window, where the first observation has little influence, unless the prior on initial states is incorporated by the variational smoother.

In this paper, the focus of the comparison is on the accuracy of the variational smoother when compared to MCMC. Future work will focus on a comprehensive assessment of its relative performance when compared with other approximation-based algorithms. As application of so-called "statistical linearisation"-strategy, the ensemble Kalman smoother [17] and unscented Kalman smoother [26] are of most interest. For multimodal systems, the Gaussian sum smoother proposed in [27] is particularly promising, as it does propagate a Gaussian sum approximation of true marginal posterior [28] .

Many MCMC algorithms suffer from poor mixing when high-dimensional stochastic complex systems are concerned. Development of efficient MCMC algorithms is always a challenging task. A combination of variational approximation methods and sampling methods would offer a new promising direction to improve the efficiency of MCMC algorithms.

References

[1] J. Honerkamp, Stochastic Dynamical Systems, VCH Publishers Inc., 1994.

[2] D. J.Wilkinson, Stochastic Modelling for System Biology, Chapman & Hall/CRC Press, 2006.

[3] E. Kalnay, Atmospheric modeling, data assimilation and predictability, Cambridge University Press, 2003.

[4] B. D. O. Anderson and J. B. Moore, Optimal Filtering, Dover Publications Inc., 2005.

[5] H. J. Kushner, Dynamical equations for optimal filter, J. Diff. Eq. **3** (1967), 179-190.

[6] R. L. Stratonovich, Conditional Markov Processes, Theor. Prob. Appl. **5** (1960), 156-178.

[7] E. Pardoux, Équations du filtrage non linéaire de la prédiction et du lissage, Stochastics **6** (1982), 193-231.

[8] R. E. Kalman and R. S. Bucy, New results in linear filtering and prediction theory, J. Basic Eng. **83D** (1961), 95-108.

[9] R. H. Shumway and D. S. Stoffer, Time Series Analysis and Its Applications, Springer, New York, 2000.

[10] C. Archambeau and D. Cornford and M. Opper and J. Shawe-Tayler, Gaussian Process Approximations of Stochastic Differential Equations, Journal of Machine Learning Research Workshop and Conference Proceedings **1** (2007), 1-16.

[11] P. E. Klöden and Eckhard Platen. Numerical Solution of Stochastic Differential Equations. Spinger-Verlag Berlin, 1992.

[12] C. Andrieu and N. De Freitas and A. Doucet and M. I. Jordan, An introduction to MCMC for machine learning, Machine Learning **50** (2003), 5-43.

[13] M. Hürzler, Statistical Methods for General State-Space Models, PhD Thesis Nr. 12674, ETH Zürich, 1998

[14] F.J. Alexander and G. L. Eyink and J.M. Restrepo, Accelerated Monte Carlo for optimal estimation of time series, Journal of Statistical Physics **119** (2005), 1331-1345.

[15] A. Gelb, Applied Optimal Estimation, MIT Press, Cambridge, MA, 1974.

[16] G. Evensen, Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model, J. Geophys. Res. **97** (1992), 17905-17924.

[17] G. Evensen, Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res. **99** (1994), 10143-10162.

[18] G. Kitagawa, Non-Gaussian state space modelling of non-stationary time series, J. Am. Statist. Assoc. **82** (1987), 503-514.

[19] S. Kullback and R. A. Leibler, On information and sufficiency, Annal of Mathematical Statistics, **22** (1951), 79-86.

[20] T. S. Jaakkola, Tutorial on variational approximation methods, in Advanced Mean Field Methods (Eds D. Saad and M. Opper), The MIT Press, 2001.

[21] D. Crisan, P. Del Moral and T. J. Lyons, Interacting particle systems approximations of the Kushner-Stratonovich equation, Advances in Applied Probability, **31** (1999), 819-838.

[22] C. E. Rasmussen and C. K. I. Williams, Gaussian Process for Machine Learning, The MIT Press, Cambridge MA, 2006.

[23] R.N. Miller and E. F. Carter and S. T. Blue, Data assimilation into nonlinear stochastic models, Tellus **51A** (1999), 167-194.

[24] G. Chan and A. T. A. Wood, Simulation of stationary Gaussian vector fields, Statistics and Computing, **22** (1999), 265-268.

[25] G. L. Eyink and J. M. Restrepo and F.J. Alexander, A mean-field approximation in data assimilation for nonlinear dynamics, Physica D **194** (2004), 347-368.

[26] S. J. Julier, J. Uhlmann, H. .F Durrant-Whyte, A New Method for the Nonlinear Tranformation of Means and Covariances in Filters and Estimators, IEEE Trans. on Automatic Control, **45** (2000), 477-482.

[27] G. Kitagawa, The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother, Annals of the Institute of Statistical Mathematics, **46(4)** (1994), 605-623.

[28] D. L. Alspach and H. W. Sorenson, Nonlinear Bayesian Estimation Using Gaussian Sum Approximations, IEEE Transactions On Automatic Control, Vol **17(4)** (1972), 439-448.