Gaussian Process Approximations of Stochastic Differential Equations

Cédric Archambeau C.ARCHAMBEAU@CS.UCL.AC.UK
Department of Computer Science, University College London
Gower Street, London WC1E 6BT, United Kingdom

Dan Cornford D.CORNFORD@ASTON.AC.UK
Neural Computing Research Group, Aston University
Aston Triangle, Birmingham B4 7ET, United Kingdom

Manfred Opper OPPERM@CS.TU-BERLIN.DE
Artificial Intelligence Group, Technical University Berlin
Franklinstraße. 28/29, D-10587 Berlin, Germany

John Shawe-Taylor JST@CS.UCL.AC.UK
Department of Computer Science, University College London
Gower Street, London WC1E 6BT, United Kingdom

Editor: Neil D. Lawrence, Anton Schwaighofer and Joaquin Quiñonero Candela

Abstract

Stochastic differential equations arise naturally in a range of contexts, from financial to environmental modeling. Current solution methods are limited in their representation of the posterior process in the presence of data. In this work, we present a novel Gaussian process approximation to the posterior measure *over paths* for a general class of stochastic differential equations in the presence of observations. The method is applied to two simple problems: the Ornstein-Uhlenbeck process, of which the exact solution is known and can be compared to, and the double-well system, for which standard approaches such as the ensemble Kalman smoother fail to provide a satisfactory result. Experiments show that our variational approximation is viable and that the results are very promising as the variational approximate solution outperforms standard Gaussian process regression for non-Gaussian Markov processes.

Keywords: Dynamical Systems, Stochastic Processes, Bayesian Inference, Gaussian Processes

1. Introduction

Stochastic differential equations are used in a wide range of applications in environmental modeling, engineering and biological modeling. They typically describe the time dynamics of the evolution of a state vector, based on the (approximate) physics of the real system, together with a driving noise process. The noise process can be thought of in several ways. It often represents processes not included in the model, but present in the real system. In our work we are motivated by problems arising in numerical weather prediction models, however the methods we are developing have a far more general relevance.

To provide a brief insight into the motivation for our work, we consider numerical weather prediction models, which are based on a discretization of a coupled set of partial differential equations (*the dynamics*) which govern the time evolution of the atmosphere, described in terms of temperature, pressure, velocity, etc. (Haltiner and Williams, 1980). However, the dynamics do not include all the physical processes acting in the atmosphere such as radiation and clouds, so these are included, often as empirical parametrizations (*the model physics*). These dynamical models typically have state vectors with dimension $\mathcal{O}(10^6)$ and are currently treated as deterministic models, although there is increasing recognition that a full stochastic treatment is necessary for progress to be made on probabilistic weather forecasting (Seiffert et al., 2006). This work is the first step in a move towards methods that will be able to treat such large, complex models in a fully probabilistic framework. A particular issue we address in this paper is the use of observations, together with dynamics defined by a stochastic differential equation to infer the posterior distribution of the state of the system, a process often referred to in meteorology as data assimilation (Kalnay, 2003). We do not review data assimilation methods extensively here, but note that recently much work has been done to address the issue of the propagation of the uncertainty at initial time through the nonlinear model equations. The most popular sequential method is called the ensemble Kalman filter, a simplified Monte Carlo approach (Evensen and van Leeuwen, 2000; Whitaker and Hamill, 2002), whereas more advanced techniques include Bayesian sequential MCMC methods (Golightly and Wilkinson, 2006). A widely used alternative to sequential treatments of the data assimilation problem is the so called 4DVAR method which seeks to find the most probable *model trajectory* over a given time window, typically using the model equations as a strong constraint, using a simple variational approach (Courtier et al., 1994).

In this work we seek a variational Bayesian treatment of the dynamic data assimilation problem. In particular, we focus on the issue of defining a Gaussian process approximation to the temporal evolution of the solution of a general stochastic differential equation with additive noise, and the posterior approximation given the observations. We expect the variational nature of this approximation to make it possible for us to apply these methods to very large models, by exploiting localization, hierarchical models and sparse representations (Seeger et al.). We present the results of our initial work, which focuses on theoretical developments. These are applied to two commonly used stochastic processes, the Ornstein-Uhlenbeck process and the noisy double well system, to illustrate their application. Further work is required before these ideas can be applied to complex models, such as those used in weather forecasting.

The issues and contributions raised by this work are as follows. If we want to do the modeling properly, we have to take into account that in general the prior process is a non-Gaussian process. Therefore, we cannot deal with the prior in an efficient way just as we cannot deal with the posterior. Thus this work is significantly different from the Gaussian process methods recently extensively studied in Machine Learning (see for example Csató and Opper, 2002; Seeger, 2004). To be more specific, we cannot compute any prior moment or marginal exactly. When the process is Markovian (not necessarily time-homogeneous), any marginal can be expressed as the product of the transition probabilities. Even for the prior, this would require the solution of a Fokker-Planck equation, which is a partial differential equation. For almost all realistic problems, the solution of the corresponding exact Fokker-Planck equation is in practice impossible, so we need to make approximations (Risken, 1989). Making approximations to solve very difficult problems is not a new idea in Machine Learning. However, because we can always explicitly compute all prior marginals (at observations and test points) for a Gaussian process (Csató and Opper, 2002), essentially nearly all current work in this direction boils down to the approximation of a multivariate (but finite dimensional) posterior density. Thus, the important feature that a process is infinite dimensional almost never plays a role in inference. In this work, things are different.

As in many areas of Machine Learning, we are using the variational method of approximating an intractable probability distribution by a tractable one (Jaakkola, 2001; Beal, 2003; Winn, 2003). In contrast to most other works, a factorizing density does not seem to make sense in an infinite setting. We are thus working with Gaussian variational densities. This has mostly been ignored in the Machine Learning literature. The *Kullback-Leibler* (KL) divergence (Kullback and Leibler, 1951) between the approximating posterior process and the exact one is one between processes (i.e., between probability measures over paths) which makes the computation non-trivial. In the field of data assimilation such a setting is not new and there has been some work done for computing approximate predictions (Eyink et al., 2004; Apte et al., 2006). These papers, however, do not provide a natural framework for estimating unknown model parameters, while we can attempt this by a variational bound for the probability of observed data which can be used within a maximum likelihood framework.

We start, in Section 2, with a review of the basic setting in which we are working. In Section 3 we develop the variational approximation methods for the general class of problems introduced in Section 2. We show how to compute the divergence between the true processes, and the approximate Gaussian process and derive expressions for the posterior moments of the approximating process. We use these expressions to derive the relevant Euler-Lagrange equations for the problem. In Section 3.2 we show how the variational approximation can be used in a smoother algorithm to approximate the conditional distribution of the state given a series of observations within a certain time frame, and the process stochastic differential equation. Section 4 show results of applying the method to two example problems: an Ornstein-Uhlenbeck process and the noisy double well system. We conclude in Section 5.

2. Basic setting

Consider a finite set of d-dimensional noisy observations $\{\mathbf{y}_n\}_{n=1}^N$ of a D-dimensional hidden state $\mathbf{x}(t)$. It is assumed that the time evolution of $\mathbf{x}(t)$ is described by an (Ito) stochastic differential equation (SDE):

$$d\mathbf{x}(t) = \mathbf{f}(\mathbf{x})dt + \sqrt{\mathbf{\Sigma}}\, d\mathbf{W}(t) \tag{1}$$

where for simplicity, we assume that $\mathbf{\Sigma} = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_D^2\}$ is diagonal, with $\mathbf{f}(\mathbf{x})$ a nonlinear function and $d\mathbf{W}(t)$ a standard (multivariate) *Wiener process*.

We do not attempt any rigorous presentation of such processes (which would involve proper Ito-calculus) in this paper, but rather resort to an intuitive picture where we understand such a process as an appropriate limit of a discrete time process $\mathbf{x}_k$. To be precise, we use the Euler-Maruyama representation of (1):

$$\Delta\mathbf{x}_k \equiv \mathbf{x}_{k+1} - \mathbf{x}_k = \mathbf{f}(\mathbf{x}_k)\Delta t + \sqrt{\Delta t\, \mathbf{\Sigma}}\, \boldsymbol{\epsilon}_k \ . \tag{2}$$

Δt is the time increment and $\boldsymbol{\epsilon}_k$ denotes a sequence of independent Gaussian random vectors $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Note that the the noise scales with $\sqrt{\Delta t}$, which is necessary to obtain the non-trivial limit of a diffusion process. Such stochastic processes are widely used in physics and finance to model continuous random systems that evolve continuously over time (e.g., Brownian motion and other diffusions). In fact, it can be viewed as a limiting case of a random walk as the time increment goes to zero. The non-trivial scaling also prohibits us from writing the SDE using ordinary derivatives, because sample paths are continuous but not differentiable with probability one. Note that this form can be used for approximate (in the sense of discretizing a differential equation) generation of samples from the (prior) process (Kloeden and Platen, 1992, ch. 9).

The process $\mathbf{x}(t)$ is a continuous time, but, unfortunately, non-Gaussian (if $\mathbf{f}(\mathbf{x})$ is nonlinear) Markov process. It is not required that this process is time-homogeneous. It defines a probability measure p_{sde} *over paths* $\{\mathbf{x}(t)\}_{t \in I}$, where $I = [0, T]$ is a time interval over which we would like to perform our inference.

As usual, in the presence of observations, the posterior measure is given by

$$\frac{dp_{post}}{dp_{sde}} = \frac{1}{Z} \times \prod_{n=1}^{N} p(\mathbf{y}_n | \mathbf{x}(t_n)), \tag{3}$$

where $\mathbf{y}_{1:N} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ are observed at the discrete times $\{t_1, \ldots, t_N\}$ and Z is the normalizing constant. The likelihood $p(\mathbf{y}_n | \mathbf{x}_n)$ is assumed to have the form of a multivariate Gaussian density:

$$p(\mathbf{y}_n | \mathbf{x}(t_n)) = \mathcal{N}(\mathbf{y}_n | \mathbf{H}\mathbf{x}(t_n), \mathbf{R}), \tag{4}$$

where $\mathbf{H} \in \mathbb{R}^{d \times D}$ defines a linear transformation and $\mathbf{R} \in \mathbb{R}^{d \times d}$ is the noise covariance matrix. In future work we will generalize this to arbitrary nonlinear observations operators $\mathbf{h}(\mathbf{x}(t))$.

3. Variational approximation

We consider the approximation of the true posterior measure by a Gaussian measure (i.e., by a Gaussian process), such that the KL divergence between the two is minimized. We will construct such a measure by the following idea: since the posterior process is Markovian, we will also use a Gaussian Markov process as its approximation. The assumption of Gaussianity implies that such a process must be governed by a linear SDE:

$$d\mathbf{x}(t) = \mathbf{f}_L(\mathbf{x}, t)dt + \sqrt{\boldsymbol{\Sigma}}\, d\mathbf{W}(t), \tag{5}$$

where

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

The matrix $\mathbf{A}(t) \in \mathbb{R}^{D \times D}$ and the vector $\mathbf{b}(t) \in \mathbb{R}^{D}$ are functions to be optimized in the variational approach. They must be time dependent to account for the non-stationarity in the process caused by the observations. Note, that the use of the same noise variance $\boldsymbol{\Sigma}$

for both processes is not a restriction, because a different choice would lead to *infinite* KL divergences.

The KL divergence of the two measures over the time interval $[0, T]$ is computed in Appendix A, giving

$$\mathrm{KL}\left[q \| p_{post}\right] = \int dq \ln \frac{dq}{dp} = \int_0^T E(t) dt + \frac{Nd}{2} \ln(2\pi) + \frac{N}{2} \ln |\mathbf{R}| + \ln Z, \tag{7}$$

with

$$E(t) = E_{sde}(t) + E_{obs}(t)$$

and

$$E_{sde}(t) = \frac{1}{2} \left\langle (\mathbf{f}(\mathbf{x}) - \mathbf{f}_L(\mathbf{x}, t))^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{f}(\mathbf{x}) - \mathbf{f}_L(\mathbf{x}, t)) \right\rangle_{q_t}, \tag{8}$$

$$E_{obs}(t) = \frac{1}{2} \sum_{n=1}^{N} \left\langle (\mathbf{y}_n - \mathbf{H}\mathbf{x}(t))^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y}_n - \mathbf{H}\mathbf{x}(t)) \right\rangle_{q_t} \delta(t - t_n), \tag{9}$$

where $\delta(t)$ is the Dirac function and $\langle \cdot \rangle_{q_t}$ indicates the expectation with respect to the marginal distribution q_t of the process at time t. As usual, the fact that $\mathrm{KL}\left[q \| p_{post}\right] \geq 0$ gives an upper bound on $-\ln Z$. So far, we have not used the assumption that $\mathbf{f}_L$ is linear.

3.1 Gaussian Process posterior moments

To evaluate the expression of the KL divergence and to permit its subsequent minimization, we need the functional dependency on the variational parameter functions $\mathbf{A}(t)$ and $\mathbf{b}(t)$) of the marginal distributions of the process q at any time t. For a Gaussian (i.e. fixed form) variational q, we can write

$$q_t(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \mathbf{m}(t), \mathbf{S}(t)). \tag{10}$$

One might expect that the time evolution of $q_t(\mathbf{x})$ can be described by a set of ordinary differential equations (ODEs) for the mean $\mathbf{m}(t)$ and the covariance matrix $\mathbf{S}(t)$. As shown in Appendix B these are given by

$$\frac{d\mathbf{m}}{dt} = -\mathbf{A}(t)\mathbf{m}(t) + \mathbf{b}(t), \tag{11}$$

$$\frac{d\mathbf{S}}{dt} = -\mathbf{A}(t)\mathbf{S}(t) - \mathbf{S}(t)\mathbf{A}^{\mathrm{T}}(t) + \boldsymbol{\Sigma}. \tag{12}$$

3.2 Variational approximation of the posterior

In order to compute the parameters and the required moments, we minimize the KL divergence (7) subject to the constraints (11) and (12), with respect to independent variations of $\mathbf{A}(t)$, $\mathbf{b}(t)$, $\mathbf{m}(t)$ and $\mathbf{S}(t)$. Therefore, we look for the stationary points of the following Lagrangian:

$$\mathcal{L} = \int_0^T \left\{ E - \mathrm{tr}\left\{ \boldsymbol{\Psi} \left(\tfrac{d\mathbf{S}}{dt} + \mathbf{A}\mathbf{S} + \mathbf{S}\mathbf{A}^{\mathrm{T}} - \boldsymbol{\Sigma} \right) \right\} - \boldsymbol{\lambda}^{\mathrm{T}} \left(\tfrac{d\mathbf{m}}{dt} + \mathbf{A}\mathbf{m} - \mathbf{b} \right) \right\} dt \tag{13}$$

where $\mathbf{\Psi}(t) \in \mathbb{R}^{D\times D}$ and $\boldsymbol{\lambda}(t) \in \mathbb{R}^{D}$ are Lagrange multipliers. Note that matrix $\mathbf{\Psi}$ is symmetric. To allow for an explicit variation, we perform an integration by parts, which gives

$$\mathcal{L} = \int_0^T \left\{E - \operatorname{tr}\left\{\mathbf{\Psi}\left(\mathbf{AS} + \mathbf{SA}^{\mathrm{T}} - \mathbf{\Sigma}\right)\right\} - \boldsymbol{\lambda}^{\mathrm{T}}\left(\mathbf{Am} - \mathbf{b}\right) + \operatorname{tr}\left\{\tfrac{d\mathbf{\Psi}}{dt}\mathbf{S}\right\} + \tfrac{d\boldsymbol{\lambda}^{\mathrm{T}}}{dt}\mathbf{m}\right\} dt$$
$$- \operatorname{tr}\left\{\mathbf{\Psi}(T)\mathbf{S}(T)\right\} + \operatorname{tr}\left\{\mathbf{\Psi}(0)\mathbf{S}(0)\right\} - \boldsymbol{\lambda}^{\mathrm{T}}(T)\mathbf{m}(T) + \boldsymbol{\lambda}^{\mathrm{T}}(0)\mathbf{m}(0). \tag{14}$$

At the latest time we only take variations with respect to $\mathbf{A}(T)$ and $\mathbf{b}(T)$, such that $\mathbf{\Psi}(T) = \mathbf{0}$ and $\boldsymbol{\lambda}(T) = \mathbf{0}$. For simplicity, we also fix the values of $\mathbf{S}(0)$ and $\mathbf{m}(0)$ rather than optimizing them. Hence, taking the derivatives of $\mathcal{L}$ with respect to $\mathbf{A}$, $\mathbf{b}$, $\mathbf{S}$ and $\mathbf{m}$ leads respectively to the following Euler-Lagrange equations:

$$\frac{\partial E}{\partial \mathbf{A}} - 2\mathbf{\Psi}\mathbf{S} - \boldsymbol{\lambda}\mathbf{m}^{\mathrm{T}} = 0, \tag{15}$$
$$\frac{\partial E}{\partial \mathbf{b}} + \boldsymbol{\lambda} = 0, \tag{16}$$
$$\frac{\partial E}{\partial \mathbf{S}} - 2\mathbf{\Psi}\mathbf{A} + \frac{d\mathbf{\Psi}}{dt} = 0, \tag{17}$$
$$\frac{\partial E}{\partial \mathbf{m}} - \mathbf{A}^{\mathrm{T}}\boldsymbol{\lambda} + \frac{d\boldsymbol{\lambda}}{dt} = 0. \tag{18}$$

It follows from (15) and (16) that the *variational functions* $\widetilde{\mathbf{A}}(t)$ and $\tilde{\mathbf{b}}(t)$ at each time can be expressed as a function of the Lagrange multipliers $\mathbf{\Psi}(t)$ and $\boldsymbol{\lambda}(t)$, as well as the moments $\mathbf{m}(t)$ and $\mathbf{S}(t)$:

$$\widetilde{\mathbf{A}}(t) = -\left\langle \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right\rangle_{q_t} + 2\mathbf{\Sigma}\mathbf{\Psi}(t), \tag{19}$$
$$\tilde{\mathbf{b}}(t) = \left\langle \mathbf{f}(\mathbf{x}) \right\rangle_{q_t} + \widetilde{\mathbf{A}}(t)\mathbf{m}(t) - \mathbf{\Sigma}\boldsymbol{\lambda}(t), \tag{20}$$

where we have used the fact that $\left\langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \right\rangle_{q_t} = \mathbf{m}\mathbf{m}^{\mathrm{T}} + \mathbf{S}$ and $\left\langle \mathbf{f}(\mathbf{x})(\mathbf{x}-\mathbf{m})^{\mathrm{T}} \right\rangle_{q_t} = \left\langle \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right\rangle_{q_t} \mathbf{S}$ for any Gaussian q_t, along with the following results:

$$\frac{\partial E}{\partial \mathbf{A}} = \mathbf{\Sigma}^{-1}\left(\left\langle \mathbf{f}(\mathbf{x})\mathbf{x}^{\mathrm{T}} \right\rangle_{q_t} + \mathbf{A}\left\langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \right\rangle_{q_t} - \mathbf{b}\left\langle \mathbf{x}^{\mathrm{T}} \right\rangle_{q_t}\right), \tag{21}$$
$$\frac{\partial E}{\partial \mathbf{b}} = -\mathbf{\Sigma}^{-1}\left(\left\langle \mathbf{f}(\mathbf{x}) \right\rangle_{q_t} + \mathbf{A}\left\langle \mathbf{x} \right\rangle_{q_t} - \mathbf{b}\right). \tag{22}$$

Smoothing algorithm

We propose the following smoothing algorithm. Make some initial guesses for $\mathbf{A}(t)$ and $\mathbf{b}(t)$ and choose a sufficiently small relaxation parameter ω.

Repeat until the KL reaches its minimum value:

1. Solve (11) and (12) forward in time for fixed variational parameters $\mathbf{A}(t)$ and $\mathbf{b}(t)$, as well as fixed $\mathbf{\Psi}(t)$ and $\boldsymbol{\lambda}(t)$.

2. With $\mathbf{m}(t)$ and $\mathbf{S}(t)$ found for $0 \leq t \leq T$, solve backward in time with $\boldsymbol{\Psi}(T) = \mathbf{0}$ and $\boldsymbol{\lambda}(T) = \mathbf{0}$:

$$\frac{d\boldsymbol{\Psi}}{dt} = 2\boldsymbol{\Psi}(t)\mathbf{A}(t) - \frac{\partial E_{sde}}{\partial \mathbf{S}}, \tag{23}$$

$$\frac{d\boldsymbol{\lambda}}{dt} = \mathbf{A}^{\mathrm{T}}(t)\boldsymbol{\lambda}(t) - \frac{\partial E_{sde}}{\partial \mathbf{m}}. \tag{24}$$

When there is an observation, use the following jump-conditions:

$$\boldsymbol{\Psi}(t_n^+) = \boldsymbol{\Psi}(t_n^-) - \frac{1}{2}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}, \tag{25}$$

$$\boldsymbol{\lambda}(t_n^+) = \boldsymbol{\lambda}(t_n^-) + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}_n - \mathbf{H}\mathbf{m}(t_n)). \tag{26}$$

The amplitude of the jumps are found by evaluating the derivatives of E_{obs} with respect to $\mathbf{S}$ and $\mathbf{m}$ at time t.

Meanwhile, update the variational parameters as follows:

$$\mathbf{A}(t) \leftarrow \mathbf{A}(t) - \omega\{\mathbf{A}(t) - \widetilde{\mathbf{A}}(t)\}, \tag{27}$$

$$\mathbf{b}(t) \leftarrow \mathbf{b}(t) - \omega\{\mathbf{b}(t) - \tilde{\mathbf{b}}(t)\}, \tag{28}$$

where $0 < \omega \leq 1$.

The underlying motivation for using the updates (27) and (28) rather than directly using (19) and (20) is to avoid numerical instabilities due to possibly too large update steps.

If $d \leq D$, then the complexity of the smoothing algorithm is $\mathcal{O}(KMD^3)$, where K is the number of sweeps (or iterations) and $M = T/dt$.

4. Experiments

In this section, the approach is validated on two 1-dimensional examples: the Ornstein-Uhlenbeck (OU) process and the double-well (DW) system. The OU process is a mathematical model of the velocity of a particle undergoing Brownian motion (Uhlenbeck and Ornstein, 1930). It is here considered as a reference example. Actually, we know the exact solution for the kernel covariance function, which is induced by the corresponding prior Gaussian Markov process. One can thus perform standard Gaussian Process (GP) regression (Rasmussen and Williams, 2006, ch. 2) for computing the posterior process. The variational approximation leads in this case to the same exact result. By contrast, the DW system is a standard data assimilation benchmark (see Miller et al., 1994; Eyinck and Restrepo, 2000). Its prior Markov process is non-Gaussian as there are two equally likely equilibria, resulting in p_{sde} being bimodal. Therefore, the posterior process is necessarily non-Gaussian, but can be well approximated by a Gaussian one given appropriate observations.

4.1 Ornstein-Uhlenbeck process

The SDE of the Ornstein-Uhlenbeck (OU) process is defined as follows:

$$dx = -\gamma x dt + \sigma^2 dW, \tag{29}$$

where γ is the drift parameter. The induced stationary covariance kernel is given by

$$K(t,t') = \frac{\sigma^2}{2\gamma} e^{-\gamma|t-t'|}. \tag{30}$$

This kernel can be plugged into the GP regression formulae to compute the exact posterior process and make prediction on unseen data.

Next, let us consider the approximation of the scalar SDE defined by the linear function $f_L(x,t) = -\alpha x + \beta$. The variational fixed points parameters are given by

$$\alpha = \gamma + 2\psi\sigma^2, \tag{31}$$

$$\beta = (\alpha - \gamma)m - \lambda\sigma^2, \tag{32}$$

where $\psi(t)$ and $\lambda(t)$ denote the scalar Lagrange multipliers. The parameters are updated after each set of forward and backward passes.

The forward pass consists in propagating the mean and the variance using the discretized Euler-Lagrange equations corresponding to (11) and (12). The backward pass uses the discretized ODEs of the Lagrange multipliers, along with the jump conditions:

$$\psi(t_n^+) = \psi(t_n^-) - \frac{1}{2\sigma^2}, \tag{33}$$

$$\lambda(t_n^+) = \lambda(t_n^-) + \frac{1}{\sigma^2}(y_n - m(t_n)). \tag{34}$$

The specific OU ODEs are given by

$$\frac{d\psi}{dt} = 2\alpha\psi - \frac{dE_{sde}}{ds^2}, \tag{35}$$

$$\frac{d\lambda}{dt} = \alpha\lambda - \frac{dE_{sde}}{dm}, \tag{36}$$

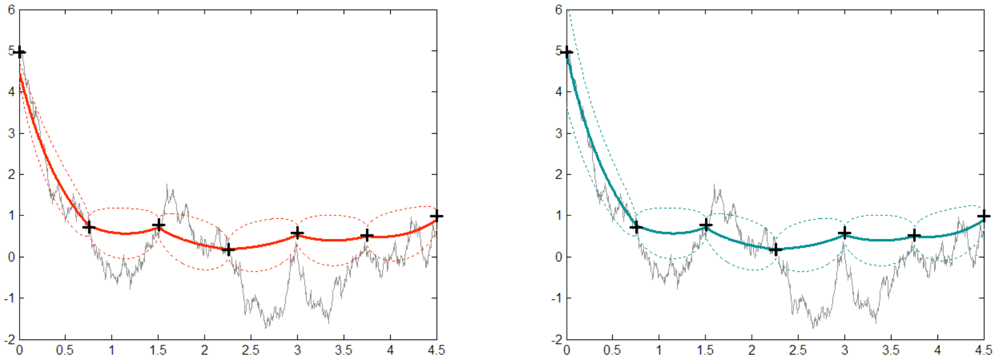
where $E_{sde} = \frac{(\alpha-\gamma)^2}{2\sigma^2}\left\langle x^2 \right\rangle_{q_t} - \frac{\beta(\alpha-\gamma)}{\sigma^2} m + \frac{\beta^2}{2\sigma^2}$.

Figure 1 shows a realization of the OU process and compares the posterior solution found by GP regression and the variational method. The true states are corrupted by zero-mean Gaussian noise. The noise levels are assumed to be known. Observe how the resulting smoothers are identical, except at the first observation where we have set $m(0) = y_0$ and $s(0) = \sigma_y^2$ for initializing the smoothing algorithm. Figure 2 shows the time evolution of the variational parameters and the Lagrange multipliers after convergence. Due to the jump conditions, the value of the Lagrange multipliers jumps when there are observations, as does the value of the variational parameters. Nevertheless, the posterior process is smooth and continuous over time (but not differentiable at times where there are observations).

4.2 Double-well system

The second example that we consider is the double-well system, which is highly nonlinear. The force f arises from a double-well potential $u(x) = -2x^2 + x^4$. The SDE is given by

$$dx = f dt + \sigma^2 dW, \tag{37}$$



(a) GP regression.

(b) Variational approximation.

Figure 1: Ornstein-Uhlenbeck example. The true process is indicated in grey and the noisy observations are marked by crosses. The left panel shows the expected posterior process (solid) and the 1-standard deviation tube (dashed) obtained by GP regression, while the right one shows the same quantities obtained by the Gaussian variational approximation.

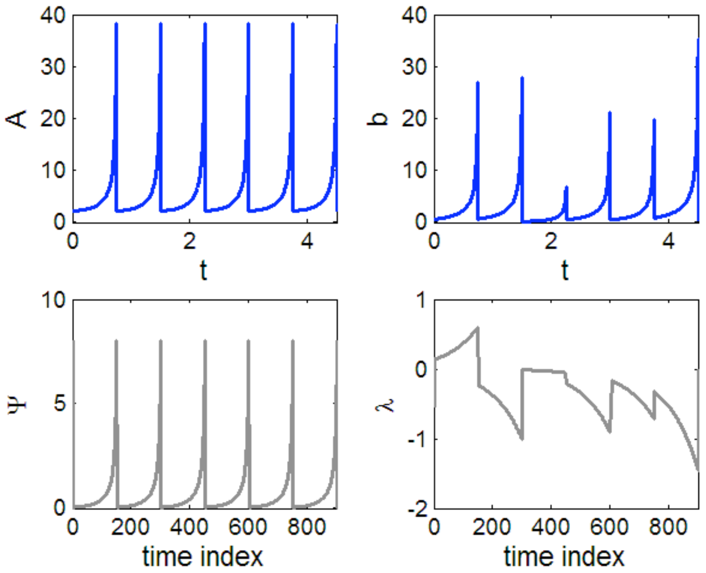


Figure 2: Variational parameters and Lagrange multipliers *vs.* time.

where $f(x) = \frac{du(x)}{dx} = 4x(1-x^2)$. Due to the driving noise, the solution $x(t)$ fluctuates around one of the two minima located at ±1. Occasionally, however, larger fluctuations arise, possibly leading to the transition to the other well. Therefore, the associated Markov process is non-Gaussian.

For the fixed point variational parameters of the approximate SDE, we have

$$\alpha = -4(1 - 3m^2 - 3s^2) + 2\psi\sigma^2, \tag{38}$$

$$\beta = -4\left\langle x^3 \right\rangle_{qt} + (4+\alpha)m - \lambda\sigma^2. \tag{39}$$

The ODEs describing the time evolution of the Lagrange multipliers are then

$$\frac{d\psi}{dt} = 2\alpha\psi - \frac{dE_{sde}}{ds^2}, \tag{40}$$

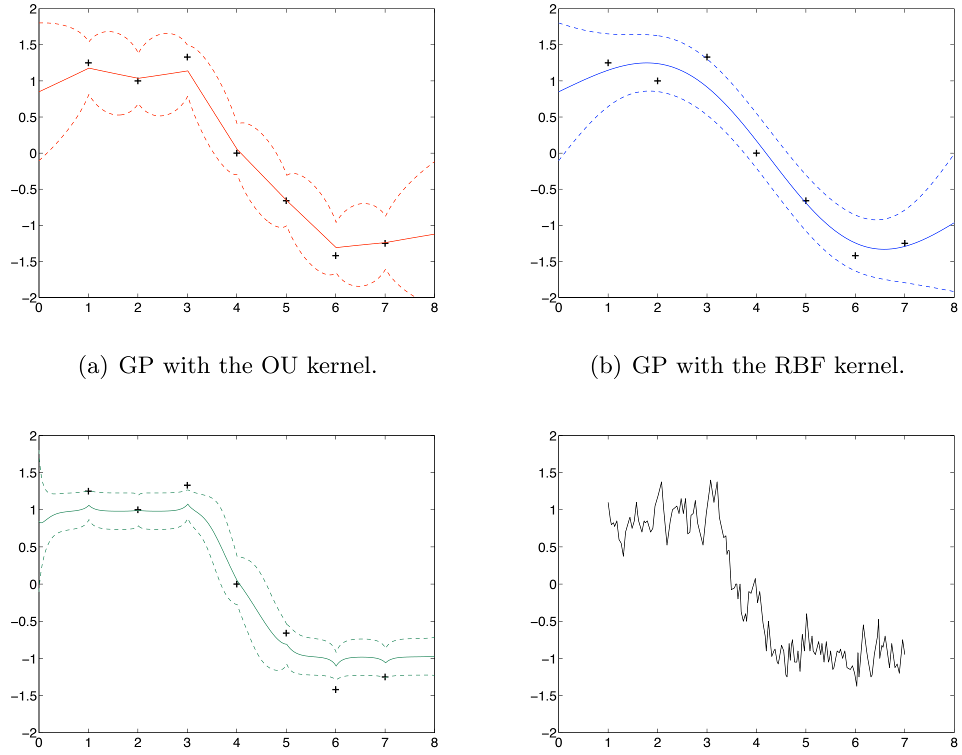
$$\frac{d\lambda}{dt} = \alpha\lambda - \frac{dE_{sde}}{dm}, \tag{41}$$

where $E_{sde} = \frac{8}{\sigma^2}\left\langle x^6\right\rangle_{qt} - \frac{4(4+\alpha)}{\sigma^2}\left\langle x^4\right\rangle_{qt} + \frac{4\beta}{\sigma^2}\left\langle x^3\right\rangle_{qt} + \frac{(4+\alpha)^2}{2\sigma^2}\left\langle x^2\right\rangle_{qt} - \frac{\beta(4+\alpha)}{\sigma^2}m + \frac{\beta^2}{2\sigma^2}$. The jump conditions are given by (33) and (34).

In the experiments we consider the same sample as the one considered by Miller et al. (1994); Eyink et al. (2004) and the parametrization is identical: the time step Δt is equal to 0.01; there are 7 noisy observations in the time window; the variance of the observation noise is equal to 0.04 and the driving noise σ^2 is equal to 0.5. Also, only one transition occurs in the time window.

First, we compare the variational solution to the one obtained by GP regression using the OU kernel and the standard RBF kernel (see Figure 3). The observation noise is set to its true value. The OU drift parameter and the RBF kernel width are selected as the ones maximizing the evidence, that is the marginal probability of the observations. Both regressors are able to locate the transition. However, they are not able to estimate the wells accurately. Note also how the GP using the OU kernel overestimates the posterior covariance. These results are not surprising as the GPs do not make use of the knowledge of the dynamics and assume the observation noise is small. By contrast, the variational solution is expected to be much closer to the true posterior process. Indeed, a good solution needs not necessarily to be closer to the true history, but to its most probable value. The noise tube is also more informative in this case. Qualitatively, this solution is the same as the one obtained by Eyink et al. (2004), who recently introduced an alternative mean field approximation approach to data assimilation. These authors also noted that the popular ensemble Kalman smoother fails to correctly locate the transition (Eyinck et al., 2006). As a matter of fact, it lags by one measurement and this failure is not cured by the backward pass.

Figure 4 shows the evolution of the KL divergence as a function of the number of sweeps for different values of the under-relaxation parameter ω, which has an effect on the convergence rate. The higher it is the faster the algorithm converges. However, when too large, numerical instabilities may occur leading to the smoothing algorithm to diverge. For reasonable values of ω the algorithm converges relatively fast, that is in less than 100 sweeps. Finally, note that the evolution of the Lagrangian is identical to the evolution of the KL as a function of the number of sweeps. This can be understood by noticing that the constraints (11) and (12) are satisfied by construction after each forward sweep.



(a) GP with the OU kernel.

(b) GP with the RBF kernel.

(c) Variational approximation.

(d) Sample path.

Figure 3: Double-well system. (a) and (b) show the GP regression solution for two different kernels while (c) is the optimal variational solution. (d) Shows the true history (or sample path). The solid lines are the posterior means and the dashed ones the posterior means ± the posterior 1-standard deviation.

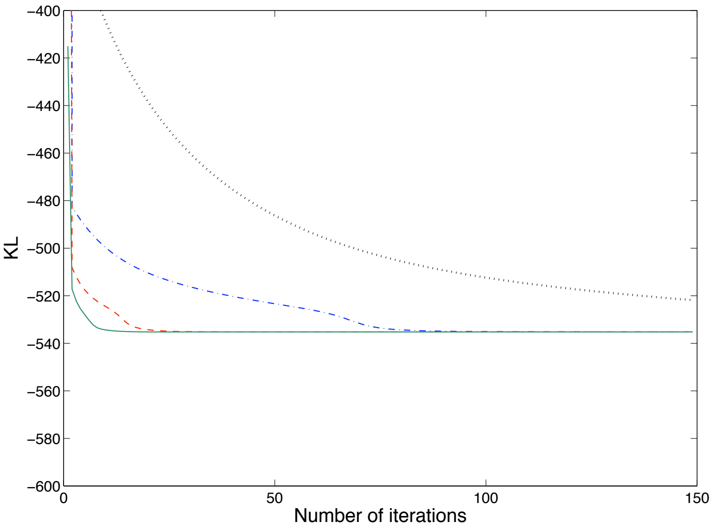


Figure 4: Double-well system. Evolution of the KL divergence as a function of the number of iterations (i.e. sweeps) for different values of the under-relaxation parameter (solid: $\omega = 0.5$; dash: $\omega = 0.25$; dash-dot: $\omega = 0.1$; dot: $\omega = 0.05$).

5. Conclusion

In this paper we have introduced a novel variational approximation to the posterior distribution of a system governed by a general stochastic differential equation. The main innovation of the work is that the posterior distribution is over paths, rather than a finite dimensional multivariate posterior as in standard Gaussian process inference. We have shown how to incorporate observations into the approximation scheme, proposing a smoothing algorithm. Results of applying the method to two example systems are very promising. On the one hand, the approach is consistent, i.e. the variational solution is identical to the exact solution when the stochastic process is a Gaussian one. On the other hand, the method is able to cope with strongly nonlinear systems (for example inducing multimodal probability measures), in contrast to most approximate state-of-the-art techniques.

This work represents an initial step towards the application of variational Bayesian inference methods to general stochastic differential equation based models. Much remains to be done to enable us the apply these methods to the very large, complicated models used in weather forecasting (see for example Dance, 2004, for a discussion of current issues). Future work will explore further the links between our methods and the cutting edge data assimilation methods developped by Eyink et al. (2004); Apte et al. (2006) and will focus on a better smoothing algorithm. One area we expect to make progress in is the application of our methods to spatially distributed systems since we are able to control the complexity of the posterior approximation by defining the class and representations of linear models, $\mathbf{f}_L(\mathbf{x}, t)$ used in our method, for example by constraining $-\mathbf{A}(t)$ to have some simplified form, such as a sparse representation or a tri-diagonal form. Another interesting application of our methods would be to situations where the system model $\mathbf{f}(\mathbf{x}, t)$ is known only approximately, but the magnitude of the model errors is not well known; employing our variational framework would allow us to make inference of the model error, represented by $\mathbf{\Sigma}$, which would produce better estimates of the posterior uncertainty after data assimilation.

Acknowledgments

We would like to thank the organizers of the Gaussian Processes in Practice meeting for putting together such an interesting workshop and providing us with the opportunity to present this work. We would also like to thank Hauke Tschach for checking the equations in the experimental part. This work has been funded by the EPSRC as part of the Variational Inference for Stochastic Dynamic Environmental Models (VISDEM) project (EP/C005848/1).

Appendix A. Kullback-Leibler divergence along a state path

To derive the result for the KL divergence, we use the following discrete time heuristics (a derivation in continuous time using *Girsanovs* theorem (Kloeden and Platen, 1992) will be given elsewhere): Consider the discretized version (2) of the SDE (1) and the corresponding version for the approximate linear one (5):

$$\Delta\mathbf{x}_k \equiv \mathbf{x}_{k+1} - \mathbf{x}_k = \mathbf{f}(\mathbf{x}_k)\Delta t + \sqrt{\mathbf{\Sigma}\Delta t}\;\boldsymbol{\epsilon}_k, \tag{42}$$

$$\Delta\mathbf{x}_k \equiv \mathbf{x}_{k+1} - \mathbf{x}_k = \mathbf{f}_L(\mathbf{x}_k, t_k)\Delta t + \sqrt{\mathbf{\Sigma}\Delta t}\boldsymbol{\epsilon}_k, \tag{43}$$

where $\boldsymbol{\epsilon}_k$ is drawn from a multivariate Gaussian density with identity covariance.

Hence, the probability density of a discrete time path, i.e. a sequence $\mathbf{x}_{1:K} = (\mathbf{x}_1, \ldots, \mathbf{x}_K)$ generated by the true prior process (i.e., without observations) and the approximate process posterior are respectively given by

$$p(\mathbf{x}_{1:K}) = \prod_{k=1}^{K-1} \mathcal{N}(\mathbf{x}_{k+1}|\mathbf{x}_k + \mathbf{f}(\mathbf{x}_k)\Delta t, \mathbf{\Sigma}\Delta t), \tag{44}$$

$$q(\mathbf{x}_{1:K}) = \prod_{k=1}^{K-1} \mathcal{N}(\mathbf{x}_{k+1}|\mathbf{x}_k + \mathbf{f}_L(\mathbf{x}_k, t_k)\Delta t, \mathbf{\Sigma}\Delta t). \tag{45}$$

The KL divergence between $q(\mathbf{x}_{1:K})$ and $p(\mathbf{x}_{1:K})$ is given by

$$\begin{aligned}\mathrm{KL}\left[q(\mathbf{x}_{1:K})\|p_{sde}(\mathbf{x}_{1:K})\right] &= \int d\mathbf{x}_{1:K}\; q(\mathbf{x}_{1:K}) \ln \frac{q(\mathbf{x}_{1:K})}{p(\mathbf{x}_{1:K})} \\ &= \sum_{k=1}^{K-1} \int d\mathbf{x}_k\; q(\mathbf{x}_k) \int d\mathbf{x}_{k+1}\; q(\mathbf{x}_{k+1}|\mathbf{x}_k) \ln \frac{q(\mathbf{x}_{k+1}|\mathbf{x}_k)}{p(\mathbf{x}_{k+1}|\mathbf{x}_k)} \\ &= \frac{1}{2}\sum_{k=1}^{K-1} \int d\mathbf{x}_k\; q(\mathbf{x}_k)\; (\mathbf{f} - \mathbf{f}_L)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{f} - \mathbf{f}_L)\Delta t,\end{aligned} \tag{46}$$

where we have used the fact that $\langle \Delta \mathbf{x}_k | \mathbf{x}_k \rangle_q = \mathbf{f}_L(\mathbf{x}_k, t_k)\Delta t$. It is possible to pass to the continuum limit $\Delta t \to 0$ on the time interval $[0, T]$, because all terms have the ordinary linear scaling with Δt of Riemann sums. It can be shown that if we had not assumed that both processes have the same noise variance $\mathbf{\Sigma}$, the corresponding sum would have diverged.

Hence, in the limit, we obtain the KL divergence between the two probability measures for state paths in this time interval:

$$\mathrm{KL}\left[q \| p_{sde}\right] = \frac{1}{2} \int_0^T \left\langle (\mathbf{f} - \mathbf{f}_L)^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{f} - \mathbf{f}_L) \right\rangle_{q_t} dt, \tag{47}$$

where $\langle \cdot \rangle_{q_t}$ indicates the expectation with respect to $q_t(\mathbf{x})$, which is the marginal density of $\mathbf{x}$ at time t.

Appendix B. Ordinary differential equations for the GP parameters

The ODEs of the means (11) and the covariance matrices (12) follow from (5) when neglecting terms beyond first order in dt:

$$\begin{aligned} d\mathbf{m}(t) &= \mathrm{E}\{\mathbf{x} + d\mathbf{x}\} - \mathrm{E}\{\mathbf{x}\} \\ &= -\mathbf{A}(t)\mathbf{m}(t)\ dt + \mathbf{b}(t)\ dt, && (48) \\ d\mathbf{S}(t) &= \mathrm{E}\{(\mathbf{x} - \mathbf{m} + d\mathbf{x} - d\mathbf{m})(\mathbf{x} - \mathbf{m} + d\mathbf{x} - d\mathbf{m})^{\mathrm{T}}\} - \mathrm{E}\{(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^{\mathrm{T}}\} \\ &= -\mathbf{A}(t)\mathbf{S}(t)\ dt - \mathbf{S}(t)\mathbf{A}^{\mathrm{T}}(t)\ dt + \mathbf{\Sigma}\ dt + \mathcal{O}(dt^2), && (49) \end{aligned}$$

where we have used the fact that $\mathbf{W}(t)$ is a Wiener process, such that $\mathrm{E}\{d\mathbf{W}(t)\} = \mathbf{0}$ and $\mathrm{E}\{d\mathbf{W}(t)d\mathbf{W}^{\mathrm{T}}(t)\} = dt\mathbf{I}$.

References

Amit Apte, Martin Hairer, Andrew Stuart, and Jochen Voss. Sampling the posterior: An approach to non-Gaussian data assimilation. *Physica D*, 2006. Submitted, available from http://www.maths.warwick.ac.uk/∼stuart/sample.html.

Matthew J. Beal. *Variational Algorithms for Approximate Bayesian Inference.* PhD thesis, Gatsby Computational Neuroscience Unit, University College, London, UK, 2003.

P. Courtier, J. N. Thepaut, and A. Hollingsworth. A strategy for operational implementation of 4D-VAR, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120:1367–1387, 1994.

Lehel Csató and Manfred Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3):641–669, 2002.

Sarah L. Dance. Issues in high resolution limited area data assimilation for quantitative precipitation forecasting. *Physica D*, 196:1–27, 2004.

Geir Evensen and Peter J. van Leeuwen. An ensemble Kalman smoother for nonlinear dynamics. *Monthly Weather Review*, 128:1852–1867, 2000.

Gregory L. Eyinck and Juan R. Restrepo. Most probable histories for nonlinear dynamics: tracking climate transitions. *Journal of Statistical Physics*, 101:459–472, 2000.

Gregory L. Eyinck, Juan L. Restrepo, and Francis J. Alexander. A statistical-mechanical approach to data assimilation for nonlinear dynamics ii. Evolution approximations. *Journal of Statistical Physics*, 2006. Accepted.

Gregory L. Eyink, Juan L. Restrepo, and Francis J. Alexander. A mean field approximation in data assimilation for nonlinear dynamics. *Physica D*, 194:347–368, 2004.

Andrew Golightly and Darren J. Wilkinson. Bayesian sequential inference for nonlinear multivariate diffusions. *Statistics and Computing*, 2006. To appear.

George J. Haltiner and Roger T. Williams. *Numerical Prediction and Dynamic Meteorology.* John Wiley and Sons, Chichester, 1980.

Tommi Jaakkola. Tutorial on variational approximation methods. In Manfred Opper and David Saad, editors, *Advanced Mean Field Methods: Theory and Practice.* The MIT Press, 2001.

Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability.* Cambridge University Press, Cambridge, 2003.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations.* Springer-Verlag, Berlin, 1992.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Robert N. Miller, Michael Ghil, and Francois Gauthiez. Advanced data assimilation in strongly nonlinear dynamical systems. *Journal of the Atmospheric Sciences*, 51:1037–1056, 1994.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, Cambridge, Massachusetts, 2006.

H. Risken. *The Fokker-Planck equation: methods of solutions and applications.* Springer-Verlag, Berlin, 1989.

Matthias Seeger. Gaussian processes for Machine Learning. *International Journal of Neural Systems*, 14(2):1–38, 2004.

Matthias Seeger, Neil Lawrence, and Ralf Herbrich. Efficient nonparametric Bayesian modelling with sparse Gaussian process approximations. Submitted for Journal Publication (2007).

Rita Seiffert, Richard Blender, and Klaus Fraedrich. Subscale forcing in a global atmospheric circulation model and stochastic parameterisation. *Quarterly Journal of the Royal Meteorological Society*, 2006. accepted.

G. E. Uhlenbeck and L. S. Ornstein. On the theory of Brownian motion. *Physical Review*, 36:823–841, 1930.

Jeffrey S. Whitaker and Thomas M. Hamill. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130:1913–1924, 2002.

John Winn. *Variational Message Passing and its Applications*. PhD thesis, Department of Physics, University of Cambridge, UK, 2003.