A BASIS FUNCTION APPROACH TO BAYESIAN INFERENCE IN DIFFUSION PROCESSES.

Y. Shen[*]*, D. Cornford*[*]*, and M. Opper*[‡]

[*]Neural Computing Research Group, Aston University, Birmingham, United Kingdom

[‡]Artificial Intelligence Group, Technical University Berlin, Berlin, Germany

ABSTRACT

In this paper, we present a framework for Bayesian inference in continuous-time diffusion processes. The new method is directly related to the recently proposed variational Gaussian Process approximation (VGPA) approach to Bayesian smoothing of partially observed diffusions. By adopting a basis function expansion (BF-VGPA), both the time-dependent control parameters of the approximate GP process and its moment equations are projected onto a lower-dimensional subspace. This allows us both to reduce the computational complexity and to eliminate the time discretisation used in the previous algorithm. The new algorithm is tested on an Ornstein-Uhlenbeck process. Our preliminary results show that BF-VGPA algorithm provides a reasonably accurate state estimation using a small number of basis functions.

Index Terms— Stochastic differential equations, data assimilation, variational approximation, model reduction

1. INTRODUCTION

Diffusion processes are widespread models in physics, biology and environmental science. There is increasing interest in developing computational methods for Bayesian inference in partially observed diffusions. Examples are, to name but few, four-dimensional variational data assimilation (4DVAR) [1], Hybrid Monte Carlo (HMC) [2], Ensemble Kalman Smoothing (EnKS) [3], and Variational Gaussian Process Approximation (VGPA) [4].

For variational methods, a major concern about their efficiency is the dimension of the control space. To reduce computational complexity, so-called reduced-order strategy is often adopted through the projection of the state space onto a linear space spanned by a small set of basis functions [5, 6]. In this paper, we adopt a similar strategy to VGPA.

Most of inferential methods for diffusion processes are formulated in continuous time but do make use of time discretization in their implementation [2, 4]. Recently, an exact algorithm for the simulation of a particular class of diffusion processes has been proposed [7]. This leads to the development of likelihood estimation procedures and particle filtering algorithms that are both free from discretization error [7, 8] In this paper, we adopt a basis function approach to eliminate time discretization.

In Section 2, we first formulate a mathematical setting of Bayesian inference in diffusion processes. Next, the VGPA framework is briefly introduced in Section 3. Following this, we outline a basis function approach to VGPA in details (Section 4). In Section 5, the new method is validated by numerical experiment with an Ornstein-Uhlenbeck process. We conclude with some discussion (Section 6).

2. BAYESIAN SMOOTHING

Mathematically, a diffusion process is often represented by a stochastic differential equation (SDE) [9]:

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

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 $\mathbf{W}(t)$. 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:

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

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 represents standard multivariate Gaussian white noise.

A Bayesian approach to smoothing is typically adopted in which the posterior distribution $p(\mathbf{x}([0,T])|\{\mathbf{y}_1, ..., \mathbf{y}_M\})$ is formulated as follows:

- The prior [10]

  $$p(\mathbf{x}(t)) = p(\mathbf{x}_0) \cdot e^{\int_0^T -\frac{1}{2}\cdot\left[\frac{d\mathbf{x}}{dt}-\mathbf{f}(\mathbf{x})\right]\mathbf{D}^{-1}\left[\frac{d\mathbf{x}}{dt}-\mathbf{f}(\mathbf{x})\right]^\top\cdot dt}$$

  where $p(\mathbf{x}_0)$ is the prior on the initial state;

- The likelihood

  $$p(\mathbf{y}_1, ..., \mathbf{y}_M|\mathbf{x}(t)) \propto \prod_{j=1}^{M} e^{\frac{-\left[\mathbf{h}(\mathbf{x}(t_{k_j})-\mathbf{y}_j\right]\mathbf{R}^{-1}\left[\mathbf{h}(\mathbf{x}(t_{k_j})-\mathbf{y}_j\right]^\top}{2}}.$$

3. THE VGPA FRAMEWORK

The starting point of the Variational Gaussian Process Approximation (VGPA) method is to approximate Eq. 1 by a linear SDE:

$$d\mathbf{x}(t) = \mathbf{f}_L(\mathbf{x}, t)dt + \mathbf{D}^{1/2}(t)d\mathbf{W}(t), \tag{3}$$

where

$$f_L(\mathbf{x}, t) = -\mathbf{A}(t)\mathbf{x}(t) + \mathbf{b}(t). \tag{4}$$

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. 4 implies that the true posterior process, i.e. $p(\mathbf{x}(t)|\mathbf{y}_1, ..., \mathbf{y}_M)$, is approximated by a Gaussian Markov process, $q(\mathbf{x}(t))$. Its moment equations are [11]

$$\frac{d\mathbf{m}(t)}{dt} = -\mathbf{A}(t)\mathbf{m}(t) + \mathbf{b}(t), \tag{5}$$

and

$$\frac{d\mathbf{S}(t)}{dt} = -\mathbf{A}(t)\mathbf{S}(t) - \mathbf{S}(t)\mathbf{A}^\top(t) + \mathbf{D}. \tag{6}$$

As in [12], $q(\mathbf{x}(t = 0))$ is fixed to $\mathcal{N}(\mathbf{x}_0|\mathbf{m}_0, \mathbf{S}_0)$. Note that $\mathbf{m}_0$ and $\mathbf{S}_0$ are also the initial values of the above equations.

The optimal $\mathbf{A}(t)$ and $\mathbf{b}(t)$ are obtained by minimising the KL divergence [13] of $q(\cdot)$ and $p(\cdot)$ which is given by

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

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

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

and

$$E_{obs}(t) = \left\langle -\log(p(\mathbf{y}_1, ..., \mathbf{y}_M|\mathbf{x}(t))\right\rangle_{q_t}.$$

Note that $<>_{q_t}$ denotes the expectation w. r. t. the marginal distribution of the approximate posterior process $q(\cdot)$ at time t, i.e., $\mathcal{N}(\mathbf{x}|\mathbf{m}(t), \mathbf{S}(t))$.

The moment equations imply that the pair $(\mathbf{m}(t), \mathbf{S}(t))$ is not independent of $(\mathbf{A}(t), \mathbf{b}(t))$. Accordingly, we find optimal ($\mathbf{A}(t)$ and $\mathbf{b}(t)$), ($\mathbf{m}(t)$, 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 $\mathbf{m}$ and $\mathbf{S}$, we obtain a system of ODEs for $\mathbf{\Psi}$ and λ, so-called adjoint equations as follows [12],

$$\frac{d\mathbf{\Psi}(t)}{dt} = 2\mathbf{\Psi}(t)\mathbf{A}(t) - \frac{\partial E_{sde}}{\partial \mathbf{S}} \tag{8}$$

$$\frac{d\lambda(t)}{dt} = \mathbf{A}^\top(t)\lambda(t) - \frac{\partial E_{sde}}{\partial \mathbf{m}}. \tag{9}$$

With the moment and adjoint equations, the non-linear smoothing is reduced to a non-linear optimisation problem with its control parameter $\mathbf{A}$ and $\mathbf{b}$.

4. THE BASIS FUNCTION APPROACH TO VGPA

For clarity, we here consider one-dimensional systems only. An extension to multivariate systems is straightforward.

First, we project the space of control functions, $\mathbf{A}(t)$ and $\mathbf{b}(t)$, onto a linear subspace spanned by a set of basis functions, $\{\phi_k(t)\}_{k=1}^N$ say. This means

$$\mathbf{A}(t) = \sum_{k=1}^N a_k \cdot \phi_k(t) \quad \text{and} \quad \mathbf{b}(t) = \sum_{k=1}^N b_k \cdot \phi_k(t),$$

where N is the number of basis functions. Note that the coefficients $\vec{A} = (a_1, ..., a_N)^\top$ and $\vec{b} = (b_1, ..., b_N)^\top$ are now control parameters.

Similarly, we approximate the auxiliary functions $\mathbf{m}(t)$ and $\mathbf{S}(t)$ by another set of basis functions, $\{\psi_k(t)\}_{k=1}^N$ say. For simplicity, the number of these basis functions is N too. To account for the fact that $\mathbf{m}(t)$ and $\mathbf{S}(t)$ are fixed at t = 0, we require $\psi_k(0)$ = 0 for all k. Accordingly, we have

$$\mathbf{m}(t) = \mathbf{m}_0 + \sum_{k=1}^N m_k \cdot \psi_k(t) \quad \text{and} \quad \mathbf{S}(t) = \mathbf{S}_0 + \sum_{k=1}^N S_k \cdot \psi_k(t).$$

where the coefficients $\vec{m} = (m_1, ..., m_N)^\top$ and $\vec{S} = (S_1, ..., S_N)^\top$ are the corresponding auxiliary parameters.

Following this, we project the moment equations onto each of those basis functions, $\{\psi_k(t)\}_{k=1}^N$, which are assumed to be linearly independent with each other. This gives us N constraint equations for $\vec{m}$ and N constraint equations for $\vec{S}$. Matrix-wise, they are

$$\mathbf{K}\vec{m} = \vec{B} \quad \text{and} \quad \mathbf{P}\vec{S} = \vec{D}$$

with

$$(\vec{B})_i = \sum_j (b_j - \mathbf{m}_0 \cdot a_j) \cdot \int_0^T \phi_j(t)\varphi_i(t)dt$$

$$(\vec{D})_i = D \cdot \int_0^T \varphi_i(t)dt - 2\mathbf{S}_0 \cdot \sum_j \cdot a_j \cdot \int_0^T \phi_j(t)\varphi_i(t)dt$$

$$\mathbf{K} = \mathcal{M}^1 + \mathcal{M}^2 \quad \text{and} \quad \mathbf{P} = \mathcal{M}^1 + 2 \cdot \mathcal{M}^2$$

where

$$\mathcal{M}^1_{ij} = \int_0^T \varphi_j(t) \cdot \frac{d\varphi_i(t)}{dt}\, dt$$

and

$$\mathcal{M}^2_{ij} = \sum_k a_k \cdot \int_0^T \phi_i(t)\phi_j(t)\varphi_k(t)dt.$$

Now, the Lagrangian approximating $\mathcal{L}$ can be defined as

$$\begin{aligned}\mathcal{L}^{BF} &= \mathbf{E}^{BF}_{sde}(\vec{A},\vec{b},\vec{m},\vec{S}) + \mathbf{E}^{BF}_{obs}(\vec{m},\vec{S}) \\ &+ \vec{\lambda}\cdot(\mathbf{K}\vec{M}-\vec{B}) + \vec{\Psi}\cdot(\mathbf{P}\vec{S}-\vec{D})\end{aligned}$$

with Lagrangian parameters $\vec{\lambda} = (\lambda_1, ..., \lambda_N)^\top$ and $\vec{\Psi} = (\Psi_1, ..., \Psi_N)^\top$. Note that $\mathbf{E}^{BF}_{sde}$, $\mathbf{E}^{BF}_{obs}$, and $\mathbf{E}^{BF} = \mathbf{E}^{BF}_{sde} + \mathbf{E}^{BF}_{obs}$ are the approximation to the integrals of $E_{sde}(t)$, $E_{obs}(t)$, and $E(t)$ over $[0, T]$, respectively.

For computing $\vec{\lambda}$ and $\vec{\Psi}$, we derive

$$\mathbf{K}^\top\vec{\lambda} = \nabla_{\vec{m}}\mathbf{E}^{BF} \quad \text{and} \quad \mathbf{P}^\top\vec{\Psi} = \nabla_{\vec{S}}\mathbf{E}^{BF}$$

by taking the derivative of $\mathcal{L}^{BF}$ w. r. t. $\vec{m}$ and $\vec{S}$. Note that the matrix $\mathbf{K}$ and $\mathbf{P}$ here are the same as those in the approximate constraint equations. The transpose of $\mathbf{K}$ and $\mathbf{P}$ plays the same role as the backward integration of the adjoint equations for $\mathbf{\Psi}$ and λ.

The algorithm so far can be outlined as follows:

step 1 compute $\vec{D}$ and initialise ($\vec{A}$, $\vec{b}$);

step 2 compute $\vec{B}$, $\mathbf{K}$ and $\mathbf{P}$ using ($\vec{A}$, $\vec{b}$);

step 3 compute $\vec{m}$ and $\vec{S}$ by solving $\mathbf{K}\vec{m} = \vec{B}$ and $\mathbf{P}\vec{S} = \vec{D}$;

step 4 compute $\mathbf{E}^{BF}$, $\nabla_{\vec{m}}\mathbf{E}^{BF}$, and $\nabla_{\vec{S}}\mathbf{E}^{BF}$ using ($\vec{A}$, $\vec{b}$) and ($\vec{m}$, $\vec{S}$);

step 5 compute $\vec{\lambda}$ and $\vec{\Psi}$ by solving $\mathbf{K}^\top\vec{\lambda} = \nabla_{\vec{m}}\mathbf{E}^{BF}$ and $\mathbf{P}^\top\vec{\Psi} = \nabla_{\vec{S}}\mathbf{E}^{BF}$

step 6 compute gradients ($\nabla_{\vec{A}}\mathcal{L}^{BF}$, $\nabla_{\vec{b}}\mathcal{L}^{BF}$) and update ($\vec{A}$, $\vec{b}$) by gradient-based optimisation algorithm,

step 7 return to step 2.

In the above algorithm, we need to compute a number of summations which can generally be expressed as follows:

$$\sum_{I(1)}...\sum_{I(n_A)}\sum_{J(k)}...\sum_{J(n_b)}\sum_{K(k)}...\sum_{K(n_m)}\sum_{L(k)}...\sum_{L(n_S)}\chi_1\cdot\chi_2$$

with

$$\chi_1 = \prod_{k=1}^{n_A} a_{I(k)} \prod_{k=1}^{n_b} b_{J(k)} \prod_{k=1}^{n_m} m_{K(k)} \prod_{k=1}^{n_S} s_{L(k)}$$

and

$$\chi_2 = \int_0^T \prod_{k=1}^{n_A} \phi_{I(k)} \prod_{k=1}^{n_b} \phi_{J(k)} \prod_{k=1}^{n_m} \psi_{K(k)} \prod_{k=1}^{n_S} \psi_{L(k)} dt.$$

Note that all index $I(\cdot)$, $J(\cdot)$, $K(\cdot)$, and $L(\cdot)$ run from 1 to N. Moreover, n_A, n_b, n_m and n_S are non-negative integer. For a Ornstein-Uhlenbeck process and a double-well potential system, we have $n_A + n_b + n_m + n_S \leq 4$ and $n_A + n_b + n_m + n_S \leq 6$, respectively. We can compute χ_2 for all possible index configurations prior to the optimization but need to update χ_1 at each step of the optimization for each index configuration. This makes the above algorithm not feasible if the number of basis functions, N, is not very small.

The above problem can be sorted out in two ways. First, we can make use of sparsity. If a set of localised basis functions is used, χ_2 is in fact vanishingly small for most of index configurations. Secondly, we can approximate the power of those control and auxiliary functions directly by a linear combination of basis functions, for example,

$$\mathbf{m}^2(t) = \sum_{k=1}^{N} \kappa_k \cdot \phi_k(t).$$

To guarantee that the projection of $\mathbf{m}^2(t)$ is consistent with the square of the projection of $\mathbf{m}(t)$ itself, we get N additional constraint equations by $\mathbf{Q}\vec{\kappa} = \vec{C}$ with $\vec{\kappa} = (\kappa_1, ..., \kappa_N)^\top$ and

$$\vec{C} = \left(\sum_{ij} m_i m_j \cdot \int_0^T \phi_i(t)\phi_j(t)\phi_1(t)dt, ...\right)^\top.$$

In doing so, we can guarantee $n_A + n_b + n_m + n_S \leq 3$ for all diffusion processes with a polynomial drift.

5. NUMERICAL EXPERIMENTS

In this work, we validate the method outlined in Section 4 with an Ornstein-Uhlenbeck process, i.e.

$$dx = (A \cdot x + b)dt + \sigma dw,$$

with $A = -3$, $b = 0.5$, $\sigma = 0.3$, and $x_0 = 0.17$. For smoothing, the window size T is set to 5. We have one observation per time unit and the observation error variance is 0.01. For this result, 30 Gaussian RBF are used. Their centre are equally located within the window and the width is set to 0.125. Fig. 1 shows that we obtain very accurate estimate of mean path from the RBF-VGPA with a small number of RBF. The ratio of control parameters between the original- and RBF-VGPA is 15:1. However, we also observe that there is some discrepancy in estimating marginal variance.

6. DISCUSSION

In this paper, we have presented an extension to the variational Gaussian Process approximation framework for Bayesian inference in partially observed diffusions. The method is validated by numerical experiments with an Ornstein-Uhlenbeck process.

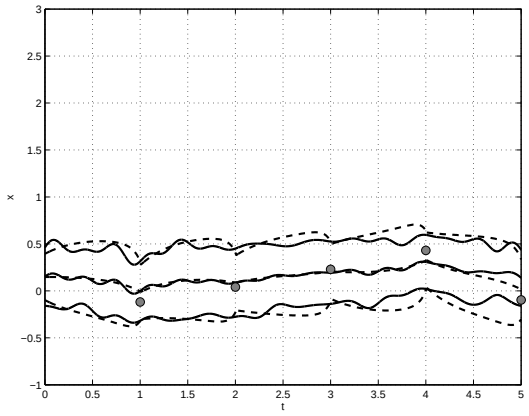

Fig. 1. Comparison of mean-path and conditional-variance estimates between the original VGPA (Dashed) and RBF-VGPA (Solid) method with a Ornstein-Uhlenbeck process. Filled circles represent 5 observations with measurement noise variance equal to 0.01. The mean paths are displayed by solid lines, while each pair of dashed lines indicates an envelope of mean path with $2 \times$ standard deviation.

Compared to the original approach, the new method can significantly reduce the number of control parameters. In the original framework, both the control functions and the moment equations must be computed numerically by time discretisation. Numerical experiments show that for many systems a time increment of 0.01 is often required. Therefore, an improvement of computational efficiency can be achieved if the number of basis functions required is much less than 100 per time unit which our numerical experiments show it is the case.

As seen in the previous sections, the new method does not need time discretisation. From a practical point of view, the continuous-time treatment can improve the stability of the VGPA framework. Although the framework is formulated in continuous time, its implementation is however based on the Euler discretisation scheme of a stochastic differential equation. If the time discretisation is not sufficiently small, the Gaussian assumption of transition probabilities of the discretised true and approximate SDE could fail. This could lead to negative value of marginal covariance $\mathbf{S}(t)$.

The key to computational efficiency of the proposed method is to reduce the number of basis functions needed for a given accuracy. Localised polynomial basis functions may be more suitable than Gaussian RBF. In future work, we adopt a spline approximation to control- and auxiliary function. An extensive comparison of computational efficiency between the original VGPA and the basis function approach to VGPA will be presented in a longer paper.

7. REFERENCES

[1] J. Derber, "A variational continuous assimilation technique," *Mon. Wea. Rev.*, vol. 117, pp. 2437–2446, 1989.

[2] A. Apte, M. Hairer, A. M. Stuart, and J. Voss, "Sampling the posterior: An approach to non-Gaussian data assimilation," *Physica. D*, vol. 230, pp. 50–64, 2007.

[3] G. Evensen, "An ensemble Kalman smoother for nonlinear dynamics," *Mon. Wea. Rev.*, vol. 128, pp. 1852–1867, 2000.

[4] C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Tayler, "Variational inference for diffusion processes," in *Neural Information Processing Systems (NIPS)*, C. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. 2008, vol. 20, pp. 17–24, The MIT Press, Cambridge MA.

[5] Y. Cao, J. Zhu, I. M. Navon, and Z. Luo, "A reduced-order approach to four-dimensional variational data assimilation using proper orthogonal decomposition," *Int. J. Num. Meth. Flu.*, vol. 53, pp. 1571–1583, 2007.

[6] B. F. Farrell and P. J. Ioannou, "State estimation using a reduced-order kalman filter," *J. Atmos. Sci.*, vol. 58, pp. 3666–3680, 2001.

[7] G. Q. Roberts, A. Beskos, O. Papaspiliopoulous, and P. Fearnhead, "Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes," *J. R. Statist. Soc. B*, vol. 68, pp. 333–382, 2006.

[8] P. Fearnhead, O. Papaspiliopoulous, and G. O. Roberts, "Particle filters for partially observed diffusions," *J. R. Stat. Soc. B*, vol. 70, pp. 755–777, 2008.

[9] P. E. Klöden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Spinger, Berlin, 1992.

[10] A. M. Stuart, J. Voss, and P. Winberg, "Conditional path sampling of SDEs and the langevin MCMC method," *Comm. Math. Sci.*, vol. 2, pp. 685–697, 2004.

[11] J. Honerkamp, *Stochastic Dynamical Systems*, VCH, Weinheim, 1994.

[12] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Tayler, "Gaussian Process approximations of stochastic differential equations," *J. Mach. Learn. Res. Workshop and Conference Proceedings*, vol. 1, pp. 1–16, 2007.

[13] S. Kullbackand and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79–86, 1951.