A variational mean field algorithm for efficient inference in large systems of stochastic differential equations

Michail D. Vrettas,[1, a)] Manfred Opper,[2, b)] and Dan Cornford[3, c)]
[1)] *University of California, Berkeley - Berkeley, CA-94720, U.S.A.*
[2)] *Technical University Berlin - Berlin, D-10587, Germany.*
[3)] *Aston University - Birmingham, B4-7ET, UK.*

(Dated: 9 January 2015)

This work introduces a new Gaussian variational mean field approximation for inference in dynamical systems which can be modeled by ordinary stochastic differential equations. This new approach allows one to express the variational free energy as a functional of the marginal moments of the approximating Gaussian process. A restriction of the moment equations to piecewise polynomial functions, over time, dramatically reduces the complexity of approximate inference for stochastic differential equation models and makes it comparable to that of discrete time hidden Markov models. The algorithm is demonstrated on state and parameter estimation for non-linear problems with up to one thousand dimensional state vectors and compares the results empirically with various well known inference methodologies.

PACS numbers: 02.50.-r, 02.60.-x, 05.10.-a
Keywords: stochastic differential equations, approximate inference, variational mean field

I. INTRODUCTION

Stochastic differential equations (SDEs) arise naturally as descriptions of continuous time dynamical systems, such as geophysical, chemical and biological systems[1]. The continuous time interpretation can be important in both the details of the dynamics and the ability to employ physical understanding in developing such dynamic models. Such models are commonly used in a range of applications from system biology[2], to data assimilation in weather prediction[3]. Despite the increasing computational power of modern computers, efficient inference in high dimensional dynamical systems still poses as a major challenge with significant implications in scientific and engineering applications. Exact inference is computationally hard and often infeasible, therefore a range of approximation methods have been developed over the years[4].

From a Bayesian perspective, inference in dynamical systems essentially consists of updating ones beliefs about the distribution of state or parameters within the SDEs, conditioning on observations. Inference in dynamical systems is often considered at two levels: direct inference on the state which is variously described as data assimilation, filtering or smoothing; and inference on the parameters within the SDE. Inference on the state is widely used in applications such as weather and environmental forecasting, where knowledge of the current state of the system approximated by the SDEs can be used to predict future conditions. For longer term predictions, such as climate modeling or in systems biology applications it is the parameters in the SDEs that are often of more interest, requiring inference to estimate not just the system state, but also distributions or point estimates of the system parameters.

Existing methods for state and parameter estimation in non-linear systems have explored a range of approaches including a variety of Markov chain Monte Carlo (MCMC) methods[5], sequential Monte Carlo[6], a range of Kalman smoothers[7], Maximum A-Posteriori (MAP) approaches based on classical variational techniques[8] and mean field conditional analysis of the SDEs obtained with moment closure methods[9]. These inference and estimation methods introduce a range of approximations and computational complexities. Monte Carlo based methods introduce sampling error, have issues with convergence and assessment of convergence[10–12]. The computational cost of Monte Carlo means such methods are only really applicable to relatively low dimensional systems, but many physical systems are described using a large number of state variables, often discretised over some spatial context. However recent advances[13], show that the utilization of graphics processing units GPUs can speed up these methods by several orders of magnitude. Kalman filters/smoothers on the other hand are fast, but can suffer from numerical instabilities and have approximation errors due the linearisation or statistical linearisation of the non-linear differential equations[14,15].

Other hybrid techniques have been proposed, based primarily on the aforementioned methods, by combining the ensemble and variational approaches[16,17], with the hope to bring more non–linearity into the data assimilation problem. All methods are very challenging to apply to parameter estimation, mainly due the inherent coupling of state and parameters, which in particular makes state augmentation approaches require very careful tuning.

New techniques of approximate inference, originally developed in the field of statistical physics, have been

[a)] m.vrettas@berkeley.edu (corresponding author)
[b)] manfred.opper@tu-berlin.de
[c)] d.cornford@aston.ac.uk

applied to the problem of inference for such models. As shown in[18,19], the *Variational Gaussian Process Approximation* (VGPA) approximates the posterior process over states by a non-stationary Gaussian process, which is induced by a SDE with a time-varying linear drift. The method minimizes a variational free energy using similar approximations to path integrals in quantum statistics[20,21]. The parameters in the drift function, which are the variational (functional) parameters of the approximation, are a D dimensional time dependent 'mean' vector and a $D \times D$ dimensional time dependent 'covariance' matrix, where D is the dimensionality of the state vector.

Although this approximation reduces inference in SDEs to solving ordinary differential equations (ODEs), rather than partial differential equations (PDEs) that arise for exact inference, the matrices contribute a factor of $\mathcal{O}(D^2)$ to the overall computational complexity. This makes the algorithm slow for larger systems. One also has to deal with the *infinite dimensional* nature of the variational parameters by discretising the ODEs in time which introduces an additional factor $\mathcal{O}(N)$ into the complexity of the algorithm, N being the number of time points used in the discretization, which is typically large for numerical stability.

A. Main ideas and contribution

In this work a new framework is presented that extends the VGPA developed by[18], allowing for a significant speed-up of approximate inference for SDEs. It is based on the great advantage of variational approximations over other ad-hoc approximation schemes; one can tune their computational complexity to the computational resources available. The three key ideas of this new approach are as follows:

- First, we adopt to a Gaussian Mean Field (MF) approximation. This means that we further restrict the approximating Gaussian measure to be factorizing in the individual components of the D dimensional state vector. This reduces the complexity from $\mathcal{O}(D^2)$ to $\mathcal{O}(D)$.

- Secondly, within the new MF approximation, we show that the original variational parameters can be eliminated analytically and expressed in terms of the marginal moments of the approximating Gaussian, thus removing the need of *forward–backward* integrations.

- Finally, since the free energy is now expressed as a functional of the marginal moments alone (which are the new functional variational parameters), we can further reduce the complexity of the approximation by a restriction to the subclass of functions defined by a *finite set of parameters*. We choose piecewise (low order) polynomials where the parameters are fixed between two subsequent observations. This reduces the complexity from $\mathcal{O}(N)$[1] to $\mathcal{O}(K)$, where K is the number of observations, with $K \ll N$ typically. Moreover this parameterisation removes any errors associated with time discretisation.

The overall complexity of the proposed approximate inference scheme for a *continuous time* Markov process is now comparable to inference for a *discrete time* hidden Markov model, thus making the new algorithm practically applicable to higher dimensional systems where the previous variational algorithm was realistically infeasible.

B. Outline

The rest of the paper is structured as follows: Section II introduces inference for stochastic differential equations. The basic setup is provided along with the appropriate notation. In addition we briefly review the VGPA scheme, as the new mean field framework extends this algorithm. In Section III, we develop the new mean field approximation and discuss its implementation details in Section IV. Numerical experiments and comparisons with other inference approaches are presented in Section V and we conclude in Section VI with a discussion.

II. STATISTICAL INFERENCE FOR DIFFUSIONS

Diffusion processes are a special class of continuous time Markov processes with continuous sample paths[22]. The time evolution of a general, D dimensional, diffusion process $\{\mathbf{x}_t\}_{t \in T}$ can be described by a stochastic differential equation (here to be interpreted in the Itō sense):

$$d\mathbf{x}_t = \mathbf{f}(t, \mathbf{x}_t; \boldsymbol{\theta})\ dt + \boldsymbol{\Sigma}(t, \mathbf{x}_t; \boldsymbol{\theta})^{\frac{1}{2}}\ d\mathbf{w}_t\ , \qquad (1)$$

where $d\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, dt\mathbf{I})$, $\mathbf{x}_t = [x_t^1, \ldots, x_t^D]^\top$ is the latent state vector, $\mathbf{f}(t, \mathbf{x}_t; \boldsymbol{\theta}) \in \mathbb{R}^D$ is the (typically) non-linear drift function, that models the deterministic part of the system, $\boldsymbol{\Sigma}(t, \mathbf{x}_t; \boldsymbol{\theta}) \in \mathbb{R}^{D \times D}$ is the diffusion or system noise covariance matrix and $d\mathbf{w}_t$ is the derivative of a D dimensional Wiener process $\{\mathbf{w}_t\}_{t \in T}$, which often models the effect of faster dynamical modes not explicitly represented in the drift function but present in the real system, or 'model discrepancy'. $T = [t_0, t_f]$ is a fixed time window of inference, with t_0 and t_f denoting the initial and final times respectively. The vector $\boldsymbol{\theta} \in \mathbb{R}^m$ is a set of parameters within the drift and diffusion functions.

[1] $N = |t_f - t_0|/\delta t$, is the number of discrete time points in a predefined time interval $[t_0, t_f]$, with δt time step.

Equation (1) defines a system with multiplicative (i.e. state dependent) system noise. The VGPA framework considers diffusion processes with additive system noise[23,24]. At first this might seem restrictive, however re-parametrization makes it possible to map a class of multiplicative noise models into this additive class[22]. Moreover there are many examples of models of physical systems (e.g. the atmosphere), where constant diffusion SDEs are considered as an advanced representation of the system, and reasonable approximations to 'model error' or 'discrepancy' component. Hence, the following SDE is considered:

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t;\boldsymbol{\theta})\ dt + \boldsymbol{\Sigma}^{\frac{1}{2}}\ d\mathbf{w}_t\ ,\quad d\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, dt\mathbf{I})\ , \tag{2}$$

where the notation is the same as in (1), with the only exception being the noise covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ which for simplicity is assumed state independent and diagonal (i.e. $\boldsymbol{\Sigma} = \text{diag}\{\sigma_i^2\}$ for $i = 1, 2, \dots, D$). Also the explicit dependency of the drift function $\mathbf{f}(\mathbf{x}_t;\boldsymbol{\theta})$ on time 't' has been suppressed for notational convenience.

A. Observation model

The stochastic process $\{\mathbf{x}_t\}_{t\in T}$ is assumed to be observed at a finite set of discrete times $\{t_k\}_{k=1}^K$, leading to a set of discrete time observations $\{\mathbf{y}_k \in \mathbb{R}^d\}_{k=1}^K$. In addition the observations are corrupted by i.i.d. Gaussian white noise, as is typically the case in the applications we consider. Therefore:

$$\mathbf{y}_k = \mathbf{h}(\mathbf{x}_{t_k}) + \boldsymbol{\epsilon}_k\ ,\quad \boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R})\ , \tag{3}$$

with $\mathbf{h}(\cdot) : \mathbb{R}^D \to \mathbb{R}^d$ representing the (potentially) nonlinear observation operator and $\mathbf{R} \in \mathbb{R}^{d\times d}$ being the observation noise covariance matrix. Moreover, it is further assumed (unless stated otherwise) that the dimensionality of the observation vector is equal to the state's vector (i.e. $d = D$) and that the discrete time measurements are "direct observations" of the state variables (i.e. $\mathbf{y}_k = \mathbf{x}_{t_k} + \boldsymbol{\epsilon}_k$). This assumption simplifies the presentation of the algorithm and is the most common case in practice. Adding arbitrary observation operators or incomplete observation to the equations only affects the system in the observation energy term in (6) and can be readily included if required.

B. Approximate inference

Inference in our algorithm is performed on the *posterior distribution*, i.e. the distribution of the state variables conditioned on the observations. Setting $\mathcal{Y} = \mathbf{y}_{1:K}$, for notational convenience, this posterior measure *over paths* $\boldsymbol{X} \equiv \{\mathbf{x}_t\}_{t\in T}$, is given by:

$$p(\boldsymbol{X}|\mathcal{Y},\boldsymbol{\theta},\boldsymbol{\Sigma}) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\Sigma})}{p(\mathcal{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma})}\prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}_{t_k})\ , \tag{4}$$

where K denotes the total number of noisy observations, $p(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\Sigma})$ represents the prior measure over paths, as defined by (2) and $p(\mathbf{y}_k|\mathbf{x}_{t_k})$ is the likelihood for the observation at time t_k from (3). $p(\mathcal{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma})$ is the normalizing marginal likelihood or partition function of the problem, computed as an integral:

$$p(\mathcal{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma}) = \int dp(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\Sigma})\prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}_{t_k})\ , \tag{5}$$

over all paths of the diffusion process. Path integrals of this type can be explicitly represented in terms of the *Wiener measure* and can be reduced (when the drift function is the gradient of a potential) to expressions well known in quantum statistical mechanics for which variational approximations have been successfully applied.

In the variational approximation of statistical physics, an exact probability measure (e.g. the 'true' posterior) is approximated by another that belongs to a family of tractable ones, in our case a Gaussian. This is done by minimizing the so called "*variational free energy*", defined as:

$$\begin{aligned}\mathcal{F}(q(\boldsymbol{X}|\boldsymbol{\Sigma}),\boldsymbol{\theta},\boldsymbol{\Sigma}) = &-\left\langle \ln\frac{p(\boldsymbol{X}|\mathcal{Y},\boldsymbol{\theta},\boldsymbol{\Sigma})}{q(\boldsymbol{X}|\boldsymbol{\Sigma})}\right\rangle_{q(\boldsymbol{X}|\boldsymbol{\Sigma})}\\ &- \ln p(\mathcal{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma})\ ,\end{aligned} \tag{6}$$

where q is the *approximate* posterior measure and $\langle\cdot\rangle_{q(\boldsymbol{X}|\boldsymbol{\Sigma})}$ denotes the expectation with respect to $q(\boldsymbol{X}|\boldsymbol{\Sigma})$. The first term, on the right hand side, is the relative entropy (or Kullback–Leibler divergence) between q and the posterior p. This can be brought into a more standard form, better known in statistical physics, if we define:

$$p(\boldsymbol{X}|\mathcal{Y},\boldsymbol{\theta},\boldsymbol{\Sigma}) = \frac{1}{Z}\mu(\boldsymbol{X})e^{-H(\boldsymbol{X})}\ , \tag{7}$$

$$q(\boldsymbol{X}|\boldsymbol{\Sigma}) = \frac{1}{Z_0}\mu(\boldsymbol{X})e^{-H_0(\boldsymbol{X})}\ , \tag{8}$$

$$Z = p(\mathcal{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma})\ . \tag{9}$$

Here $\mu(\boldsymbol{X})$ is a reference measure over paths, which we take to be the Wiener measure and $H(\boldsymbol{X})$ is a Hamiltonian which can be derived from *Girsanov's* change of measure theorem[25,26]. Inserting these definitions into (6), and using the fact that the relative entropy is nonnegative we get the well known variational bound:

$$-\ln Z = -\ln p(\mathcal{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma}) \leq \mathcal{F}(q(\boldsymbol{X}|\boldsymbol{\Sigma}),\boldsymbol{\theta},\boldsymbol{\Sigma}) \tag{10}$$

$$= -\ln Z_0 + \langle H(\boldsymbol{X})\rangle_q - \langle H_0(\boldsymbol{X})\rangle_q\ , \tag{11}$$

for the exact free energy of the model. Note, that for this bound to be finite the system noise covariance (i.e. $\boldsymbol{\Sigma}$), for both measures p and q must be the same[18].

1. Optimal approximate posterior Gaussian process

Gaussian variational approximations for path integrals of the form (11) have been extensively studied in statistical physics and quantum mechanics since Feynman's celebrated work on the Polaron problem[20,21]. The Hamiltonian H can be explicitly computed using *Girsanov's* change of measure theorem for diffusion processes[25], as the sum of an ordinary integral over time and an Ito–integral and is given by:

$$H(\boldsymbol{X}) = \frac{1}{2}\int_{t_0}^{t_f}\Big\{\mathbf{f}(\mathbf{x}_t;\boldsymbol{\theta})^\top\boldsymbol{\Sigma}^{-1}\mathbf{f}(\mathbf{x}_t;\boldsymbol{\theta})\;dt \\ -\,2\mathbf{f}(\mathbf{x}_t;\boldsymbol{\theta})^\top\boldsymbol{\Sigma}^{-1}\;d\mathbf{x}_t\Big\} - \sum_{k=1}^{K}\ln p(\mathbf{y}_k|\mathbf{x}_{t_k})\;. \tag{12}$$

If the "trial Hamiltonian" H_0 is chosen to be a combination of a linear and a quadratic functionals in $\boldsymbol{X}$, $q(\boldsymbol{X}|\boldsymbol{\Sigma})$ becomes a Gaussian measure over paths. The usual approach would then be to choose a simple parametric form of the trial 'harmonic oscillator' Hamiltonian H_0 and optimize these parameters by minimizing the variational bound (11).

In[18] a different, non-parametric approach was chosen, in order to accommodate the fact that through the observations the posterior is non-stationary and thus should be approximated by a non-stationary Gaussian process. This method also avoids an explicit introduction of Hamiltonians H and H_0. We have thus assumed that the trial Gaussian measure q is generated from an approximating *linear* SDE which is defined as:

$$d\mathbf{x}_t = \mathbf{g}(\mathbf{x}_t)\;dt + \boldsymbol{\Sigma}^{1/2}\;d\mathbf{w}_t\;, \tag{13}$$

where $\mathbf{g}(\mathbf{x}_t) = -\mathbf{A}_t\mathbf{x}_t + \mathbf{b}_t$, with $\mathbf{A}_t \in \mathbb{R}^{D\times D}$ and $\mathbf{b}_t \in \mathbb{R}^D$ define the time varying linear drift in the approximating process, and $\{\mathbf{w}_t\}_{t\in T}$ is a D dimensional Wiener process. Both of these variational parameters $\mathbf{A}_t$ and $\mathbf{b}_t$ are time dependent *functions* that need to be optimized as part of the estimation procedure. We work directly with the expression (6) for the variational free energy and use the fact that the prior process and the Gaussian approximation to the posterior are Markov processes. Using a time discretisation of paths and an explicit representation of path probabilities as products of transition probabilities (which have a simple Gaussian form for short times), we have been able to derive the following expression of the variational free energy as a sum of three cost functions (which we will also term 'energies' in the following):

$$\mathcal{F}(q(\boldsymbol{X}|\boldsymbol{\Sigma}),\boldsymbol{\theta},\boldsymbol{\Sigma}) = E_0 + \int_{t_0}^{t_f} E_{sde}(t)dt + \sum_{k=1}^{K} E_{obs}(t_k)\;, \tag{14}$$

where t_0 and t_f again define the initial and final times of the total time window (i.e. $T = [t_0, t_f]$). These three energies can be interpreted in the following manner.

- Initial state ($t = 0$) cost:

$$E_0 = \left\langle \ln\frac{q(\mathbf{x}_0)}{p(\mathbf{x}_0)}\right\rangle_{q(\mathbf{x}_0)} \tag{15}$$

- Prior process (SDE) cost:

$$E_{sde}(t) = \frac{1}{2}\left\langle\|\mathbf{f}(\mathbf{x}_t;\boldsymbol{\theta}) - \mathbf{g}(\mathbf{x}_t)\|^2_{\boldsymbol{\Sigma}}\right\rangle_{q_t} \tag{16}$$

- Discrete time observations (likelihood) cost:

$$E_{obs}(t_k) = \frac{1}{2}\left\langle\|\mathbf{y}_k - \mathbf{x}_{t_k}\|^2_{\mathbf{R}}\right\rangle_{q_t} + \frac{D}{2}\ln(2\pi) \\ + \frac{1}{2}\ln|\mathbf{R}|\;, \tag{17}$$

where the state of the system at initial time ($t = 0$) is assumed to have a prior $\mathbf{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \tau_0\mathbf{I})$, with $\mathbf{I} \in \mathbb{R}^{D\times D}$ being the identity matrix and the notation $\|\mathbf{u} - \mathbf{v}\|^{\mathbf{2}}_{\mathbf{M}}$ denotes the squared distance of vectors $\mathbf{u}$ and $\mathbf{v}$, weighted by the inverse of matrix $\mathbf{M}$. To keep the paper self-contained, we have included the main arguments of the derivation in Appendix A.

2. Gaussian process posterior moments

The averages in the above cost functions are over the Gaussian *marginal* densities of the approximating process at given times t defined as:

$$q(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t|\mathbf{m}_t, \mathbf{S}_t)\;, \text{with } t \in T\;, \tag{18}$$

where $\mathbf{m}_t \in \mathbb{R}^D$ and $\mathbf{S}_t \in \mathbb{R}^{D\times D}$, are respectively the marginal mean and covariance at time 't'. The time evolution of this general time varying linear system in (13), is described by two well known ordinary differential equations (ODEs), one for the marginal means $\mathbf{m}_t$ and one for the marginal covariances $\mathbf{S}_t$[1,22]:

$$\dot{\mathbf{m}}_t = -\mathbf{A}_t\mathbf{m}_t + \mathbf{b}_t\;, \tag{19}$$

$$\dot{\mathbf{S}}_t = -\mathbf{A}_t\mathbf{S}_t - \mathbf{S}_t\mathbf{A_t}^\top + \boldsymbol{\Sigma}\;, \tag{20}$$

and thus become functionals of $\mathbf{A}_t$ and $\mathbf{b}_t$, where $\dot{\mathbf{m}}_t \in \mathbb{R}^D$ and $\dot{\mathbf{S}}_t \in \mathbb{R}^{D\times D}$ denote the time derivatives. We emphasize that Eq. (19) and (20) are hard constraints to be satisfied ensuring consistency in the variational approximation[18,23]. One way to enforce these constraints, within a predefined time window $[t_0, t_f]$, is to introduce additional Lagrange multipliers $\boldsymbol{\lambda}_t \in \mathbb{R}^D$, $\boldsymbol{\Psi}_t \in \mathbb{R}^{D\times D}$; one for (19) and one for (20) respectively. These are also time dependent functions that need to be optimized during the estimation process. For more information on the formulation of this approach and its algorithmic solution see[27].

III. MEAN FIELD APPROXIMATION

The computational complexity of this approach caused by the matrices $\mathbf{S}_t$ and $\mathbf{A}_t$ leads to an impractical algorithm, when the SDEs are high-dimensional. Hence, here we make a mean field assumption that further restricts the approximating posterior process q. This approximation is equivalent to assuming a diagonal matrix $\mathbf{A}_t = \mathrm{diag}\{a_1(t), \dots, a_D(t)\}$. Since $\mathbf{\Sigma}$ is also assumed to be diagonal, the marginal covariance inherits the same property (i.e. $\mathbf{S}_t = \mathrm{diag}\{s_1(t), \dots, s_D(t)\}$). The corresponding factorization of the approximating measure q allows the individual processes of the dimensions (x_t^i ,$\forall\, i = 1, \dots, D$) to be treated independently, although critically dependencies among the *mean and variances* of the state variables are still maintained through the drift function $\mathbf{f}(\mathbf{x}_t; \boldsymbol{\theta})$.

Setting $\mathbf{m}_t = [m_1(t), \dots, m_D(t)]^\top$, as in the original variational approach and $\mathbf{s}_t = [s_1(t), \dots, s_D(t)]^\top$, a vector containing only the diagonal elements of the covariance matrix $\mathbf{S}_t$, the ODEs (19) and (20) simplify to:

$$\begin{aligned} \dot{m}_i(t) &= -a_i(t) m_i(t) + b_i(t)\ , \\ \dot{s}_i(t) &= -2a_i(t) s_i(t) + \sigma_i^2\ , \quad \forall\ i = 1, \dots, D\ . \end{aligned} \tag{21}$$

Hence, we can expect that the dimensionality enters the complexity linearly, as $\mathcal{O}(D)$. This allows us to express the initial variational problem in terms of the functions $\mathbf{m}_t$ and $\mathbf{s}_t$ alone.

A. Mean field free energy

To make this more clear consider the ODEs (19 and 20), for $\mathbf{A}_t$ diagonal:

$$\dot{\mathbf{m}}_t = -\mathbf{A}_t \mathbf{m}_t + \mathbf{b}_t\ , \tag{22}$$

$$\dot{\mathbf{S}}_t = -2\mathbf{A}_t \mathbf{S}_t + \mathbf{\Sigma}\ . \tag{23}$$

Solving (23) for $\mathbf{A}_t$ and replacing the result in (22) leads to:

$$\mathbf{A}_t = \frac{1}{2}(\mathbf{\Sigma} - \dot{\mathbf{S}}_t)\mathbf{S}_t^{-1}\ , \tag{24}$$

$$\mathbf{b}_t = \frac{1}{2}(\mathbf{\Sigma} - \dot{\mathbf{S}}_t)\mathbf{S}_t^{-1}\mathbf{m}_t + \dot{\mathbf{m}}_t\ . \tag{25}$$

Substituting (24) and (25) into the linear drift of (13) provides a new form of the drift function that depends only on the marginal values of the means $\mathbf{m}_t$ and variances $\mathbf{S}_t$, at each time 't':

$$\mathbf{g}(\mathbf{x}_t) = \dot{\mathbf{m}}_t - \frac{1}{2}(\mathbf{\Sigma} - \dot{\mathbf{S}}_t)\mathbf{S}_t^{-1}(\mathbf{x}_t - \mathbf{m}_t)\ , \tag{26}$$

where $\mathbf{\Sigma}$ and $\mathbf{S}_t$ are now both diagonal matrices. This expression of the linear approximation gives rise to a new formulation of the $E_{sde}(t)$ function (16), hence the variational free energy (14). This is given by

$$E_{sde}(t) = \frac{1}{2}\sum_{i=1}^{D} \frac{1}{\sigma_i^2}\left\{ \left\langle (f_i(\mathbf{x}_t) - \dot{m}_i(t))^2 \right\rangle_{q_t} + \frac{(\dot{s}_i(t) - \sigma_i^2)^2}{4s_i^2(t)} + (\sigma_i^2 - \dot{s}_i(t))\left\langle \frac{\partial f_i(\mathbf{x}_t)}{\partial x_t^i} \right\rangle_{q_t} \right\}. \tag{27}$$

It is clear from the above that since the constraints (Equations 22 and 23) are eliminated in the optimization problem there is no need to introduce additional Lagrange parameters. Hence the problem reduces to estimating the optimal mean and variance functions $m_i(t)$ and $s_i(t)$, $\forall i = 1, \dots, D$.

A direct functional minimization of the free energy with respect to $\mathbf{m}_t$ and $\mathbf{S}_t$ would lead to Euler-Lagrange equations which are ODEs of second order. These would be of mixed boundary type: while $m_i(t)$ and $s_i(t)$ are given at initial time $t = t_0$ (assuming that the density q_0 of the initial state is optimized later in an outer loop), their first derivatives are not. On the other hand, stationarity of the functional imposes conditions on $m_i(t)$ and $s_i(t)$ at the final time $t = t_f$. We will not pursue this route here, but introduce a further approximation which entirely avoids ODEs and the need to deal with ODEs and their time discretization.

IV. POLYNOMIAL APPROXIMATION OF THE MARGINAL MOMENTS

Instead of using direct discretization of the mean and variance functions over time, which would prohibit the use of the algorithm in very high dimensional systems, we take an important step to speed up the variational approximation by suggesting a further restriction of the variational parameters. We minimize the free energy functional in the subspace of functions $m_i(t)$ and $s_i(t)$ defined by a finite number of parameters. Note, that this approach strongly relies on the remarkable fact that the ODEs (21) are hard-coded in our new approach: a finite parametrization of the original variational functions $\mathbf{A}_t$ and $\mathbf{b}_t$ alone, would not have led to a finite parametrization of the resulting $\mathbf{m}_t$ and $\mathbf{s}_t$[19]. Since $m_i(t)$ and $s_i(t)$ must be continuous[18], but their time derivatives jump at the observations, we will use piecewise low-order (in time t) local polynomial approximations.

Here, we assume third order polynomials for the mean functions and second order polynomials for the variance functions, i.e.:

$$\begin{aligned} m_i(t) &= m_{i,0} + m_{i,1}t + m_{i,2}t^2 + m_{i,3}t^3\ , \\ s_i(t) &= s_{i,0} + s_{i,1}t + s_{i,2}t^2\ . \end{aligned}$$

There is no theoretical constraint on the order of the polynomials, for each function. However, if we restrict the solution to families of low orders, then the desired integrals of the new cost function (27) are easier to compute analytically. When drift functions $\mathbf{f}(\mathbf{x}_t; \boldsymbol{\theta})$ are polynomials in $\mathbf{x}_t$, expectations over Gaussian marginals can be performed in closed form and finally all time integrals can be computed analytically.

A. Practical implementation

The aforementioned polynomial approach has two obvious constraints that need to be satisfied for all times 't'. These are: (a) the functions $m_i(t)$ and $s_i(t)$, must be continuous (even though they may not be differentiable at observation times t_k) and (b) the variance functions must stay positive over the whole time window of inference (i.e. $s_i(t) > 0\ , \forall\, t \in [t_0, t_f]$).

To fulfill these constraints simultaneously, avoiding the use of additional parameters that would increase the complexity of our algorithm, we followed the approach of representing the piecewise polynomials by their *Lagrange formula* analogue (i.e. using four points for the 3'rd order mean and three points for the 2'nd order variance functions, per time-interval). Lagrange's interpolation formula[28] provides us with an explicit expression of the polynomials in terms of the function values at given points. In numerical analysis this is also known as the polynomial of the *least degree*; that means given a finite set of $\mathcal{C}$ points there exists a *unique polynomial* of least degree $(\mathcal{C}-1)$, that interpolates exactly these $\mathcal{C}$ points. Therefore, the optimization now consists of estimating the optimal positions of these points rather than the coefficients of the functions.

B. Lagrange polynomial form of the mean functions

For the mean function $m_i^j(t)$ which is defined on the j'th interval $[t_j, t_{j+1}]$, since we have chosen a 3'rd order polynomial, we need at least four points to represent this polynomial uniquely. These are:

$$M_i^j = \left\{ m_i^j(t_j), m_i^j(t_j+h), m_i^j(t_j+2h), m_i^j(t_{j+1}) \right\}\ ,$$

where $h = \frac{t_{j+1}-t_j}{3}$ is the spacing between the points. Here, without loss of generality, we assume that the points are evenly spread within the time interval $[t_j, t_{j+1}]$, although this is only to simplify the presentation of the algorithm.

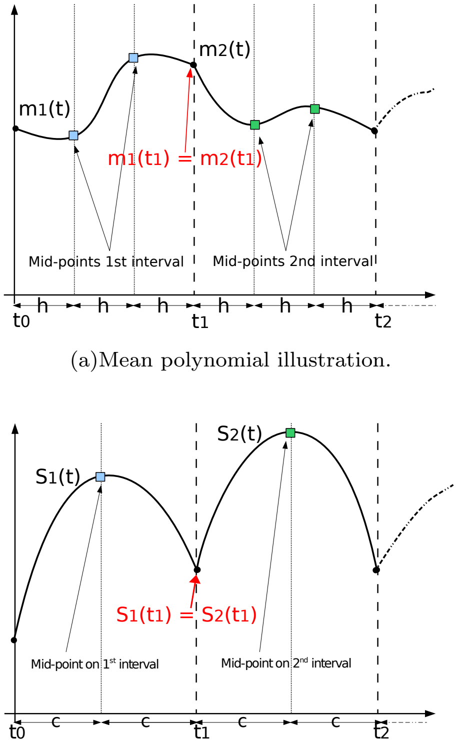
The Lagrange polynomial formula that exactly interpolates the above points is given by:

$$m_i^j(t) = \sum_{k=0}^{3} \left\{ m_i^j(t_j+kh) \left(\prod_{\substack{0\leq l\leq 3\\ l\neq k}} \frac{t-(t_j+lh)}{t_j-(t_j+lh)} \right) \right\}\ , \tag{28}$$

where t_{j+1} has been replaced with $(t_j + 3h)$.

C. Lagrange polynomial form of the variance functions

In a similar manner the assumption of the 2'nd order for the variance function $s_i^j(t)$, implies that we need at least three points to represent this polynomial uniquely within the predefined time interval $[t_j, t_{j+1}]$. These are:



(a)Mean polynomial illustration.

(b)Variance polynomial illustration.

FIG. 1. Construction of the mean (a) and variance (b) functions using local polynomials. Notice how the end-point of the first polynomial coincides with the start-point of the subsequent polynomial (pointed red arrows), ensuring continuity over the whole time domain.

$$S_i^j = \left\{ s_i^j(t_j), s_i^j(t_j+c), s_i^j(t_{j+1}) \right\}\ ,$$

where $c = \frac{t_{j+1}-t_j}{2}$ is the spacing between the points (in this case the mid-point). The Lagrange polynomial formula that exactly interpolates the above points is given by:

$$s_i^j(t) = \sum_{k=0}^{2} \left\{ s_i^j(t_j+kc) \left(\prod_{\substack{0\leq l\leq 2\\ l\neq k}} \frac{t-(t_j+lc)}{t_j-(t_j+lc)} \right) \right\}\ , \tag{29}$$

where t_{j+1} has been replaced with $(t_j + 2c)$.

D. The algorithm

For these parameterizations the *infinite dimensional* inference problem reduces to optimizing the positions of $4 \times D \times J$ points for the mean functions, together with $3 \times D \times J$ points for the variance functions (with $J = K + 1$, the total number of time intervals defined by the K observations). For real problems where the true underlying process $\{\mathbf{x}_t\}_{t \in T}$ is sparsely observed we anticipate that the number of optimized variables $(7 \times D \times J) \ll N$. The minimization is performed using a scaled conjugate gradient algorithm, although a Newton-type minimization could also be applied taking advantage of the sparsity of the Hessian matrix. Hence, the optimization becomes fast and efficient exploiting the fact that the integrals on each dimension i can be computed in parallel.

The benefits of our approach are twofold. First, as shown in Fig. (1), the continuity constraint is satisfied by requiring that the last point of each function on the j'th time-interval will be identical with the first point of the function of the following $(j+1)$'th time interval (e.g. $m_i^j(t_k) = m_i^{j+1}(t_k)$, where i represents the spatial dimension, j the time interval, in which the function exists and t_k is an observation time).

Second, the positivity constraint is satisfied by optimizing the logarithm of the variance points (e.g. $\log(s_i^j(t_k))$ instead of $s_i^j(t_k)$), making the appropriate adjustments to the gradient functions. Since the variance functions are parabolas if the three points that define the function are positive, then all other points within the time interval will also be positive.

V. SIMULATION RESULTS

This section explores experimentally the properties of the new MF approximation in comparison with a range of other approximate inference techniques used in stochastic processes. The new approach is validated on two nonlinear dynamical systems. The first system considered is a stochastic version of the three dimensional chaotic Lorenz '63 (hereafter L3D) system[29], described by the following SDE:

$$d\mathbf{x}_t = \begin{bmatrix} \sigma(y_t - x_t) \\ \rho x_t - y_t - x_t z_t \\ x_t y_t - \beta z_t \end{bmatrix} dt + \mathbf{\Sigma}^{\frac{1}{2}} \ d\mathbf{w}_t \ , \qquad (30)$$

where $\mathbf{x}_t = [x_t, \ y_t, \ z_t]^\top \in \mathbb{R}^3$ is the state vector representing all three dimensions, $\boldsymbol{\theta} = [\sigma, \ \rho, \ \beta]^\top \in \mathbb{R}^3$, is the drift parameter vector, $\mathbf{\Sigma} \in \mathbb{R}^{3\times 3}$ is a (diagonal) covariance matrix and $\mathbf{w}_t \in \mathbb{R}^3$ is an uncorrelated multivariate Wiener process. The parameters used in the simulations are the standard settings that produce the chaotic behavior[2]. Additional noise is added to the original deterministic equations with noise coefficient $\mathbf{\Sigma} = \text{diag}\{\sigma_i^2 = 10 \mid i = 1, 2, 3\}$ and the process is observed fully every $\Delta\tau = 0.2$ time units, with error covariance $\mathbf{R} = \text{diag}\{\rho_i^2 = 2 \mid i = 1, 2, 3\}$. The simulation time used was $T = [0, 20]$. This system was chosen since it is both challenging in terms of non-linearity and exhibits chaotic behavior often seen in real physical systems, but has sufficiently low dimension that a range of methods including MCMC approaches can be contrasted on it.

The second system considered is a stochastic version of the Lorenz '96 system, with drift function:

$$\mathbf{f}(\mathbf{x}_t; \theta) = [f_1(\mathbf{x}_t; \theta), \ldots, f_D(\mathbf{x}_t; \theta)]^\top \ , \qquad (31)$$

where

$$f_i(\mathbf{x}_t; \theta) = (x_t^{i+1} - x_t^{i-2}) x_t^{i-1} - x_t^i + \theta \ ,$$

with cyclic index $i \in \{1, 2, \ldots, D\}$ and $\theta \in \mathbb{R}$, as the forcing (drift) parameter. The diffusion is again an uncorrelated multivariate Wiener process, with $\mathbf{\Sigma} = \text{diag}\{\sigma_i^2 = 10 \mid i = 1, \ldots, D\}$ and $\mathbf{R} = \text{diag}\{\rho_i^2 = 2 \mid i = 1, \ldots, D\}$.

These equations simulate advection, damping and forcing of some atmospheric variable x_t^i, therefore it can be seen as a simplistic, yet manageable "weather forecasting like" model[30]. When the forcing parameter $\theta < 0.895$, solutions decay to a steady state solution, i.e. $x_t^1 = \cdots = x_t^D = \theta$; when $0.895 \leq \theta < 4.0$; solutions are periodic and when $\theta \geq 4.0$, solutions are chaotic[31]. Finally, to test the efficiency of the new MF approach on a higher dimensional system, where sampling approaches such as the MCMC are not efficient, the model (31) is extended to $D = 1000$ (or L1000D). Table (I) summarizes the setup for the systems considered in the simulations.

A. State estimation

In the first set of experiments we focus on state inference. We compare the results to Hybrid Monte Carlo (HMC) path sampling[5], the variational Gaussian process approximation (VGPA)[18], an unscented Kalman smoother (UnKS)[7], an ensemble version of the forward-backward Kalman smoother (with very large number of ensemble members $M_{ens} = 1000$) and a weak constraint 4DVar method[8], which is a popular variational approach to MAP estimation in diffusion processes, widely used in weather forecasting.

Figure 2 shows an example of smoothing (state inference) for the x_t variable of the L3D system, over a central part of the time window considered $[0, 20]$, applying all the algorithms to the same set of simulated observations. It is clear that visually the results appear rather similar. There are subtle differences, but qualitatively the differences are minor, the main issue being that the 4DVar does not provide an estimate of posterior uncertainty as it is a MAP estimate. Also shown on the figures is the elapsed CPU time in seconds for each algorithm to run on the full 20 time unit window. Evidently HMC (25,000 samples, 5,000 burn-in, hand tuned to best performance), which one might consider a *reference solution*, takes orders of magnitude longer to run. The MF algorithm on

[2] the values are $\boldsymbol{\theta} = [10, \ 28, \ 2.6667]^\top$.

System	D	d	t_0	t_f	δt	$\boldsymbol{\theta}$	σ_i^2	ρ_i^2	N_o
Lorenz'63	3	3	0	20	0.01	[10, 28, 2.66667]	10	2	5 ($\Delta\tau = 0.2$)
Lorenz'96	1000	350	0	4	N/A	8	4	1	8 ($\Delta\tau = 0.125$)

TABLE I. Experimental setup that generated the data (trajectories and observations). System dimension is denoted by D, while observation dimension by d. The time windows are defined by the initial times (t_0) and final times (t_f), whilst δt is the time discretization step that was used by the other algorithms with which we compare our new method. The vector $\boldsymbol{\theta}$ contains the parameters related to the drift function, while σ_i^2 and ρ_i^2 represent the noise variances of the driving noise and the observations accordingly, per i'th dimension. In this example the variances are identical for all dimensions. N_o defines the number of available i.i.d. observations *per time unit* (i.e. observation density), which without loss of generality is measured at equidistant time instants.

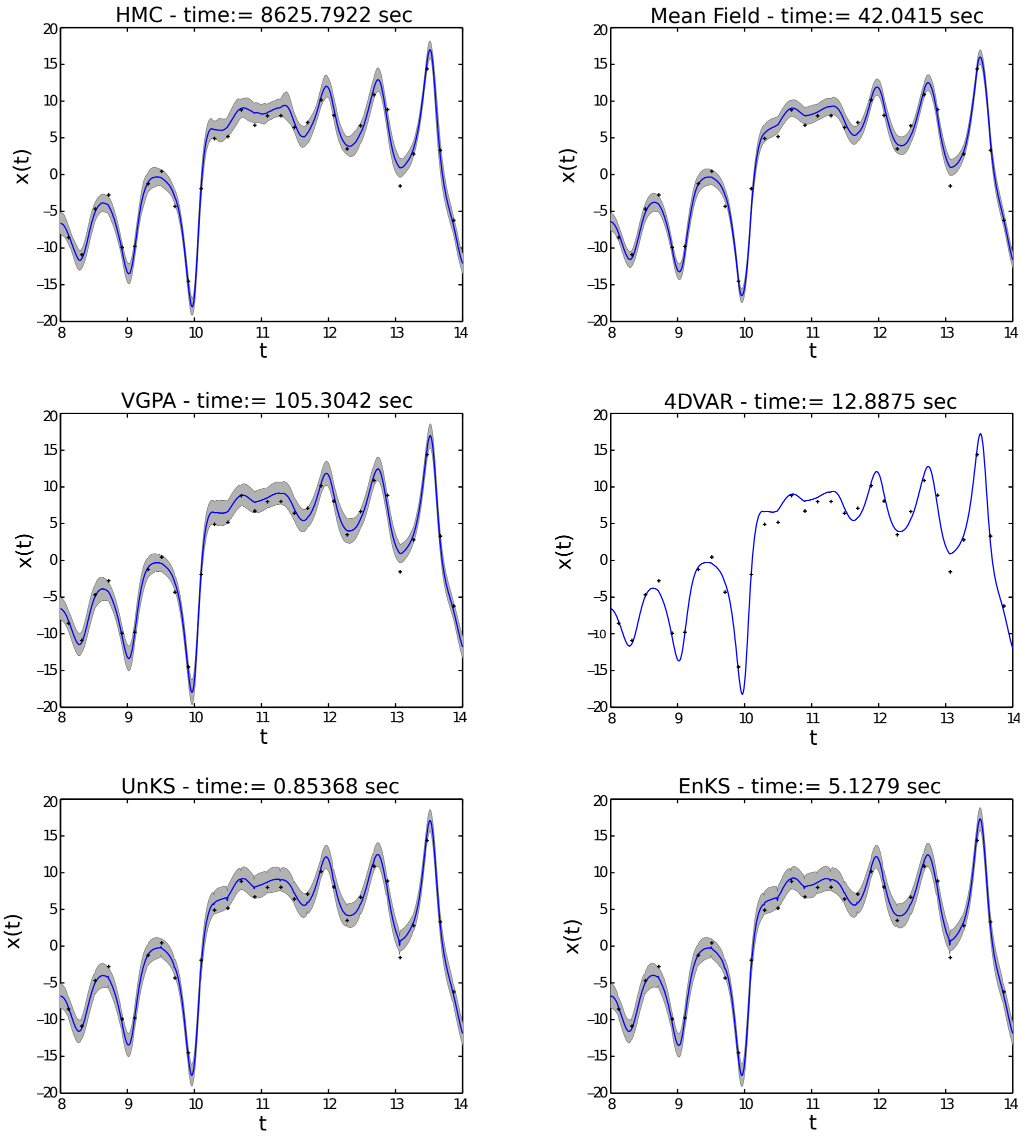


FIG. 2. The mean (solid line) and marginal variance (shaded region, plotted at +/- 2×std) for state estimation of x_t variable for an example of the L3D system for the methods HMC, MF, VGPA, 4DVar, UnKS and EnKS (from top left to right). Observations are plotted as black crosses.

this example is 3 times faster than VGPA, and only twice as slow as 4DVar, although the results depend on the available computational resources (for this set of experiments we used an eight core desktop PC). The UnKS is significantly faster than all methods on this low dimensional system, since a very small number of particles (Ens $= 2 \times D + 1 = 7$) are needed in the low dimensional system, while the EnKS took a little longer but we should emphasize the unusually high number of ensemble members that is used here ($M_{ens} = 1000$) for the forward filtering pass. A coarse time discretization of $\delta t = 0.01$ was used, for all but the MF algorithm, to permit HMC to run in reasonable time.

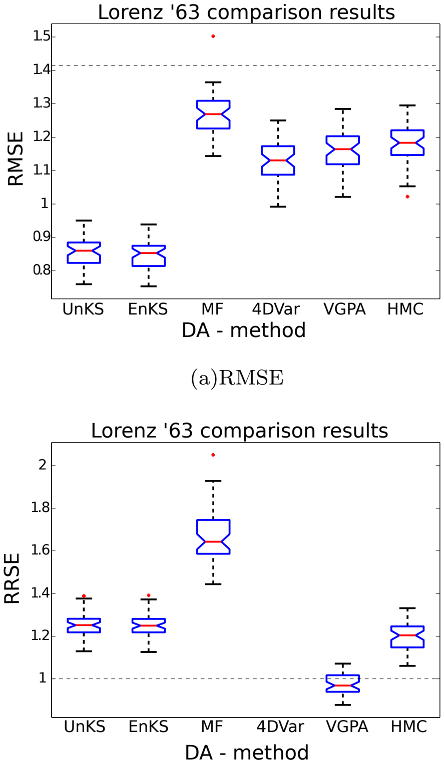
To quantify the differences in the methods we generate 50 random trajectories of the L3D system, with simulated noisy observations. The time window is $[0, 20]$, with observation density $N_o = 5$ per time unit, which is representative of the observation frequency in realistic settings. This observation frequency was chosen to be similar to that expected in operational weather forecasting applications. We apply the smoothing algorithms (UnKS, EnKS, MF, 4DVar, VGPA and HMC) and compare the Root Mean Square Errors (RMSE) and the Root Residual Square Errors (RRSE) in Figure 3(a) and 3(b) respectively, averaged over all the system dimensions. The errors are defined as follows:

$$RMSE = \frac{1}{D}\sum_{j=1}^{D}\sqrt{\frac{1}{K}\sum_{k=1}^{K}(y_j(t_k) - m_j(t_k))^2}\ , \tag{32}$$

$$RRSE = \frac{1}{D}\sum_{j=1}^{D}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{(x_j(t_k) - m_j(t_k))^2}{s_j(t_k)}}\ , \tag{33}$$

where D is the system dimension, K is the total number of observations, $y_j(t_k)$ is the noisy observation on the j'th dimension at time t_k and $x_j(t_k)$, $m_j(t_k)$ and $s_j(t_k)$ are the true signal, the marginal mean and variance at the same time instant.

We note that we are comparing the estimated mean (mode for 4DVar) state with noisy observations to compute the RMSE, so with observation variance set at 2, we would expect a value of around 1.4 for the RMSE. The plot shows that both the UnKS and EnKS systematically over-fit to the noisy observations. The other methods also show some over-fitting but are in essence very similar, although the MF method did show a rather poor fit in one simulation. The MF method is both robust and fast, producing an uncertainty estimate that is not available from 4DVar. One issue we have observed, and requires further investigation, is that the MF approach appears to underestimate the marginal variances with respect to HMC, as shown in the left plot of Figure 3(b), where for the RRSE we would expect a value close to one. Nevertheless, this underestimation does not seem to affect the mean estimates significantly.



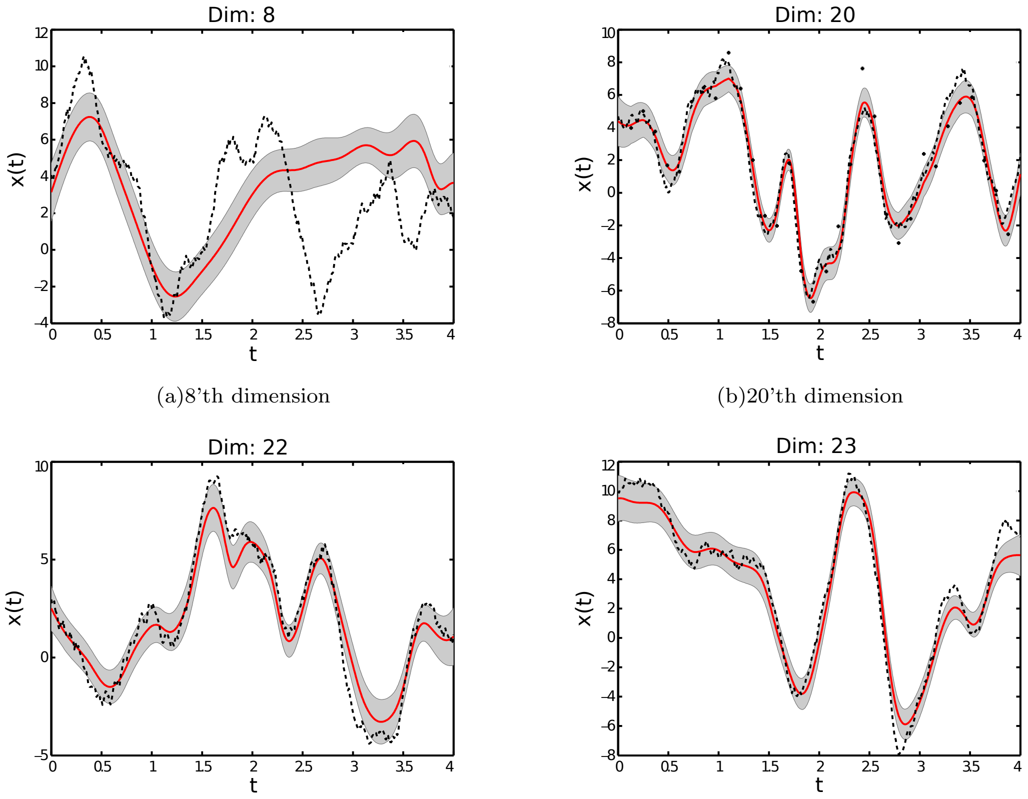
(a)RMSE

(b)RRSE

FIG. 3. A comparison of the root mean square error (a) and the root residual square error (b) of the smoothing methods on 50 realizations of the L63 system. Note the box plots describe the statistics of the 50 different error estimates.

Since one of the goals of this new MF approach is the application of the proposed variational framework to high dimensional systems as a first step we applied the new method on the Lorenz'96 system with $D = 1000$. To make the simulation more challenging, but at the same time more realistic, unlike the previous experiments where for simplicity we assumed that the observed vector has the same dimensions as the state vector (i.e. $D = d$), here we assume that we measure only $d = 350$ from the total $D = 1000$ with the locations picked at random. The partial observation in addition to the discrete time nature of the observation process makes inference for such systems a very difficult task. In this case we apply a linear operator $\mathbf{H}$ with number of rows and columns $[350 \times 1000]$ such as, $\mathbf{y}_k = \mathbf{H}\mathbf{x}_{t_k} + \boldsymbol{\epsilon}_k$. This matrix has zero elements everywhere except from the predefined observed locations, in the diagonal, which are set to one. This way the expression for the energy term from the observations $E_{obs}(t_k)$ (Eq. 17) remains unchanged. In the case of an arbitrary (non-linear) operator, the expectation $\left\langle \|\mathbf{y}_k - h(\mathbf{x}_{t_k})\|^2_{\mathbf{R}} \right\rangle_{q_t}$ would have to be approximated (e.g. with unscented transformation methods[32]).

Figures 4(a) to 4(d) show the variational marginal mean and variance, of the MF algorithm, applied on a typical example of the $L1000D$ system. As expected, when the MF algorithm "observes" a dimension it pro-



(a)8'th dimension

(b)20'th dimension

(c)22'nd dimension

(d)23'rd dimension

FIG. 4. The marginal mean (solid red line) and variance (shaded region, plotted at +/- 2×std) for state estimation of the 8'th, 20'th, 22'nd and 23'rd dimension of the L1000D system. Dashed lines show the true trajectory $\{\mathbf{x}_t, t \in T\}$ that generated the observations (crosses). Note that only (b) includes observations.

vides a good match between the posterior mean with the observations and the path that generated them and the variance is tighter than in dimensions where we do not have observations. When the algorithm does not have observations then it can either fail to track the true trajectory, as shown in Figure 4(a), or it can successfully track the true trajectory, see Figures 4(c) and 4(d). Because of the couplings of the Lorenz'96 model that occur in space domain (i.e. $(x_t^{i+1} - x_t^{i-2})x_t^{i-1} - x_t^i)$, when many subsequent dimensions are not observed the MF algorithm can fail to follow the true signal. In this example, the closest observed state variable to the 8'th dimension was the 17'th; which within the given time-window it was not close enough to recover the "truth". However, close to measured dimensions the algorithm recovers the "reality" quickly, with quite broad error-bars, and a better initialization can improve tracking of unobserved dimensions.

B. Parameter estimation

The main benefit of our MF approach is not in the state estimation, but in the stability of the hyper-parameter estimation. By *hyper-parameters* we mean the set of parameters that exist in the model equations (drift), the diffusion coefficient matrix (or function) and also in the observation process model. In real world problems we might have prior beliefs over their values, but ideally we would like to estimate, or update our beliefs about them, using information from the available observations. This section presents results displayed as marginal profiles, although a gradient based estimation has also been implemented. We compute the free energy, from both MF and VGPA algorithms, at convergence of the (state inference) smoothing procedure, varying only one parameter at a time to plot the bound on the marginal likelihood. In this work we do not provide a comparison with other estimation techniques because this has already been presented in[19] for the original VGPA algorithm.

1. Drift parameters

Most of the results presented here are from simulations of the L3D system, mainly because of its low dimensionality which allows the application of other techniques such as the HMC sampling algorithm. Figures 5(a) to 5(c), contrast the profiles from both MF and VGPA when estimating the drift parameters of the L3D system. It is obvious that both approaches provide smooth results and there are subtle differences in their final minimum found (indicated with $\star$), which we ascribe to the the different nature of the MF approximation compared to VGPA. However, as Figure 10 shows, MF was on average three times faster than VGPA, for the given settings and available computational resources.

To illustrate the difficulty of estimating parameters in stochastic chaotic models, even in low dimensional systems such as the L3D, we include here results of posterior estimates obtained by the HMC algorithm which can be assumed to provide reference solution. The approach we followed here augments the state vector with the parameter that is estimated and then the sampling is performed jointly. Figure 6 (from left to right) shows the histograms of the posterior samples of σ, ρ and β parameters of the L3D drift function. It is apparent that even though the parameters are sampled marginally (i.e. the other parameters that are not sampled are fixed to their true values) there are still biases that shift the distributions away from the correct values.

2. Diffusion parameters

Unlike the previous section where the estimation for the drift parameters was achieved, within reasonable uncertainty, by both algorithms, when estimating the diffusion coefficients the message is not so clear. As Figures 7(a) to 7(c) show, there are cases where both VGPA and MF fail to provide a clear minimum (inside the test range) and other cases where one algorithm or the other does provide a minimum close to the true value. One reason for this behavior can be the relative low observation density that we used in this example ($N_o = 5$, per time unit). As shown in[19], to estimate this crucial parameter one needs very dense observations ($N_o > 10$) regardless of the inference method used. Another explanation is that these profiles were generated by a single realization, therefore we cannot generalize any conclusions definitively, that is the results are only illustrative. Nevertheless, one thing we can argue is that the MF is consistently faster than VGPA, as shown in Figure 7(d). However, the problem of estimating diffusion coefficients is far from easy, as even MCMC sampling approaches fail to estimate these noise parameters for the settings used in this paper.

3. Observation process

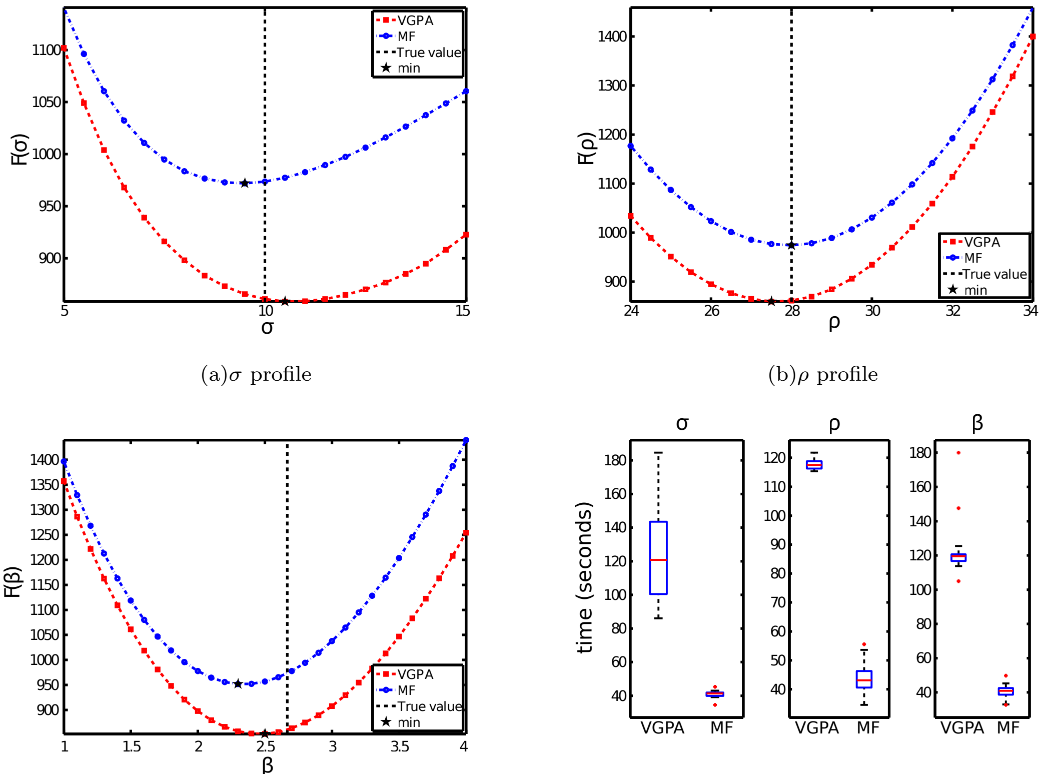
Even though estimation of the parameters related to the observation process is a natural extension of any parameter estimation framework, these results are the first time that both MF and VGPA are put to the test. For this example we use a relatively low, but realistic, observation noise variance ($\rho^2 = 1$) and test the performance of both algorithms with two different observation densities, keeping all the other model parameters to their 'correct' values. Figures 8(a) and 8(b) present the marginal profiles with $N_o = 5$, for the x_t and y_t dimensions respectively, of the L3D, whereas Figures 9(a) and 9(b) illustrate the same experiment with $N_o = 10$.

Both algorithms are able to identify the correct minimum quite accurately, although MF seems to be a bit more confident (narrower profile around the minimum) especially when the observation density increases to 10 per time unit. Observation noise estimation on z_t dimension had similar behavior and was not included here. What is more impressive here is that the speed-up of the MF approximation, in obtaining these profiles, was much higher than the other estimation simulations. Figure 10 shows that MF was seven times faster than VGPA with $N_o = 5$ and roughly four times faster with $N_o = 10$. This can be explained by the fact that by increasing the observation density we actually increase the number of iterations in the parallel loop that the MF uses to compute the free energy (when the parallelization of the algorithm is on the time domain). Therefore we do expect the algorithm to slow down slightly (given fixed computational resource).

VI. DISCUSSION

The new MF algorithm presented herein provides robust approximate inference for the state and parameters in SDE models. We show how the variational framework enables us to control the computational complexity of the algorithm. The inclusion of a mean field approximation is critical in allowing us to completely re-cast the algorithm without a forward-backward approach, producing a significantly different algorithm compared to the original VGPA algorithm[18], which is both more scalable and more robust. This introduces the potential to undertake approximate likelihood based parameter inference in SDEs using very large (long time window) data sets in high dimensional systems. The ability to work in continuous time removes the need for time discretization and thus eliminates any associated discretisation errors. The mean field method is inherently parallelizable and scales linearly with respect to state vector dimension opening a way for treating very large systems which have previously not been amenable to Bayesian approaches.

Experimental results show the method to be both robust and comparable to the computationally more expensive MCMC methods in terms of the accuracy of state inference given sufficient observations. In particular compared to the original VGPA algorithm[18], the results are more robust and can be computed at lower computational expense. We are able to run the MF approximation on very long time windows, which is important if the aim is estimation of the parameters in the SDE. The profiles of free energy, which bound the marginal likelihood for the parameters in the SDE, are very smooth and can be



(a)σ profile

(b)ρ profile

(c)β profile

(d)timings (seconds)

FIG. 5. (a) to (c) are the marginal profiles of the drift parameters. The vertical dashed line indicates the true parameter value, while the star symbol ($\star$) denotes the minimum found by each algorithm. (d) summarizes the timings (in seconds) for each algorithm to generate the profiles.

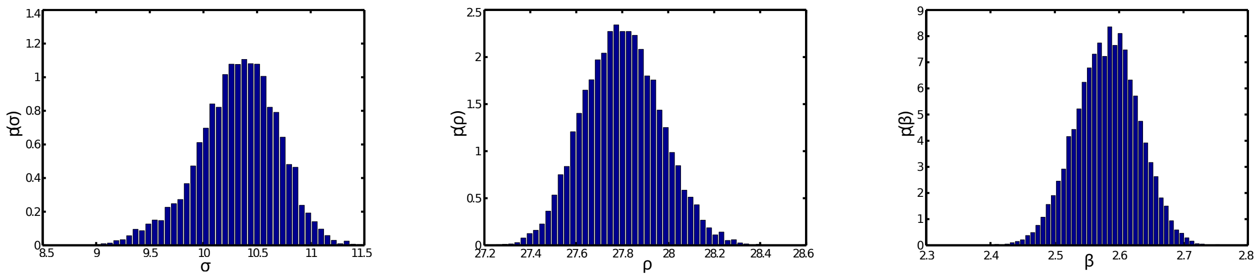
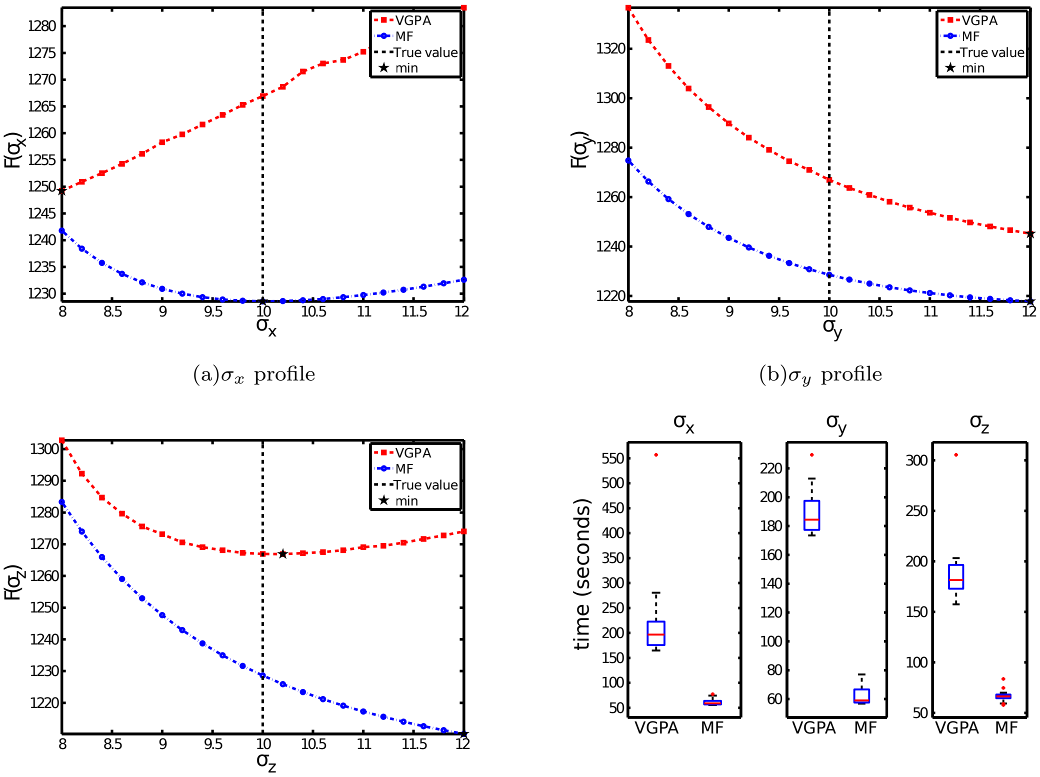


FIG. 6. Histograms of posterior samples for the σ, ρ and β (from left to right) drift parameters of the L3D system obtained with the HMC algorithm.

obtained without any algorithm tuning, something that is not true for the other methods employed in this paper, which required careful tuning to produce reliable results.

The MF approximation suffers from the same limitations as the original VGPA and is most suitable for inference in systems where there are sufficient observations to uniquely identify the state of the system, since the approximation will break down for multi-modal posterior distributions. In the case of multi-modal posteriors over paths in very large systems we expect that no methods would work practically, although recent developments in particle filtering claim some success[15]. Also the fact that the new MF was not able to consistently estimate diffusion parameters requires further investigation. One possibility to improve on the MF assumption could be the so–called *linear response* corrections[33,34], which can yield a useful approximation to the neglected correlations. Such correction would be computed *after*



(a)σ_x profile

(b)σ_y profile

(c)σ_z profile

(d)timings (seconds)

FIG. 7. (a) to (c) are the marginal profiles of the diffusion coefficients in each dimension. The vertical dashed line indicates the true parameter value, while the star symbol ($\star$) denotes the minimum found by each algorithm. (d) summarizes the timings (in seconds) for each algorithm to generate the profiles.

convergence of the MF algorithm and would require an extra linearization of the MF equations.

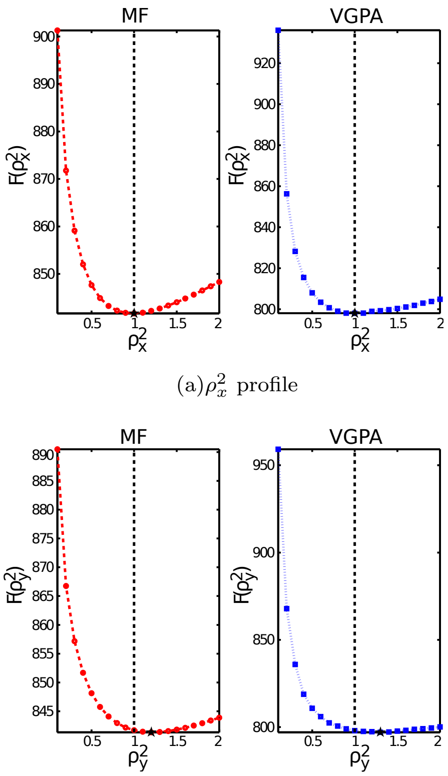
In the experiments described in this paper the observation density was rather low, compared to the timescales in the system, and thus it is not surprising that the MF approximation is unable to estimate the variance of the driving noise process. All inference methods find this challenging in practice. Implementing the algorithm is also quite complex[3], but can be largely automated using symbolic manipulation tools.

There are several interesting directions for further research. Parallelization is only partially exploited using 8 cores on a desktop machine, and is possible in several places. Which to use depends on the dimension of the state vector, the frequency of the observations and the order of polynomials used in the approximation. For very long time windows it might be possible to further approximate the marginal likelihood using a factorizing assumption on the free energy, splitting the long time window into sub-time intervals. Given sufficiently long time sub-intervals the approximation will not be significantly affected. It is also interesting to consider the practical application of the method, and the degree to which it can be applied to really high dimensional systems such as those used in weather forecasting, where the state vector is of the order of 10^7 dimensions. In these applications the ability to use parallel computation is essential, and the flexibility of the variational framework we put forward, which allows us to match computational complexity to available computational resource, makes the method particularly attractive.

ACKNOWLEDGMENTS

This work was partially funded from a European grant: EC-FP7 under the QeoViQua project (ENV.2010.4.1.2-2; 265178), and the European Community's Seventh Framework Programme (FP7, 2007-2013) under the grant agreement 270327 (CompLACS).

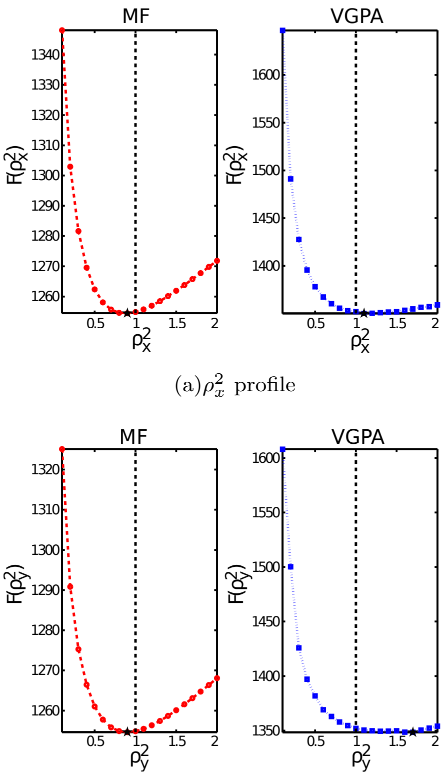
[3] An implementation of the algorithm in MATLAB is available upon request.



(a)ρ_x^2 profile

(b)ρ_y^2 profile

FIG. 8. With $N_o = 5$, (a) presents the marginal profile of the noise estimation (variance component) on x_t dimension. The vertical dashed line shows the true value of the parameter that generated the observations. Star symbol ($\star$) denotes the minimum found by each algorithm. (b) same as (a) but for the y_t dimension.



(a)ρ_x^2 profile

(b)ρ_y^2 profile

FIG. 9. Same as 8(a) and 8(b), but with increased observation density ($N_o = 10$).

Appendix A: Derivation of the Variational Gaussian Approximation

We aim at approximating the posterior measure over paths $\mathbf{X} = \{\mathbf{x}_t\}_{t \in T}$, for the posterior process, with another process which is governed by an SDE:

$$d\mathbf{x}_t = \mathbf{g}(\mathbf{x}_t)\ dt + \mathbf{\Sigma}^{1/2}\ d\mathbf{w}_t\ ,$$

with different drift function $\mathbf{g}(\mathbf{x}_t)$ and with measure q, which belongs to a family of tractable ones. The "goodness" of fit between the true posterior p and the approximating one q is given by the variational free energy

$$\begin{aligned}\mathcal{F}(q(\boldsymbol{X}|\mathbf{\Sigma}), \boldsymbol{\theta}, \mathbf{\Sigma}) &= -\left\langle \ln \frac{p(\boldsymbol{X}|\mathcal{Y}, \boldsymbol{\theta}, \mathbf{\Sigma})}{q(\boldsymbol{X}|\mathbf{\Sigma})} \right\rangle_q - \ln p(\mathcal{Y}|\boldsymbol{\theta}, \mathbf{\Sigma}) \\ &= \left\langle \ln \frac{q(\mathbf{X})}{p(\mathbf{X})} \right\rangle_q - \langle \ln p(\mathcal{Y}|\mathbf{X}) \rangle_q \quad \text{(A1)}\end{aligned}$$

where q is a shorthand notation for $q(\mathbf{X}|\mathbf{\Sigma})$ and the dependence on the drift and diffusion parameters $\boldsymbol{\theta}$ and $\mathbf{\Sigma}$ has been suppressed for brevity. $p(\mathcal{Y}|\mathbf{X})$ is the probability density of the discrete time observations and $p(\mathbf{X})$ denotes the prior measure over paths generated from the SDE (1). We will compute the variational free energy by a discretization of the sample paths in time (i.e. $\mathbf{X} = \{\mathbf{x}_k\}_{k=0,\dots,N}$) and then taking appropriate limits.

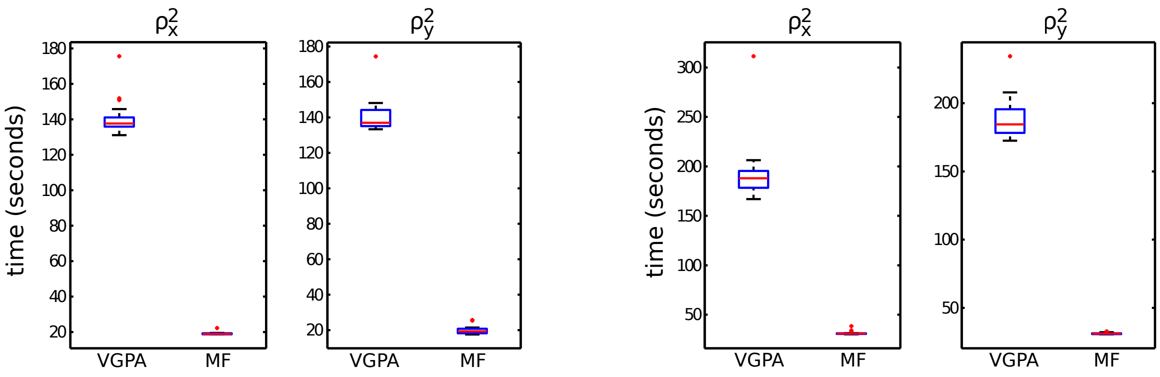
Using an Euler–Maruyama discretisation of the prior SDE and the posterior approximation we get:

$$\begin{aligned}\delta\mathbf{x}_{k+1} &= \mathbf{f}(\mathbf{x}_k; \boldsymbol{\theta})\delta t + \sqrt{\mathbf{\Sigma}\delta t}\ \boldsymbol{\epsilon}_k\ , \\ \delta\mathbf{x}_{k+1} &= \mathbf{g}(\mathbf{x}_k)\delta t + \sqrt{\mathbf{\Sigma}\delta t}\ \boldsymbol{\epsilon}_k\ , \qquad \text{(A2)}\end{aligned}$$

where $\delta\mathbf{x}_{k+1} = \mathbf{x}_{k+1} - \mathbf{x}_k$, δt is a small time step and $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Since both processes are Markovian, the joint probability densities of the discretized paths can be written as products of their transition densities:

$$\begin{aligned}p(\mathbf{x}_{0:N}) &= p(\mathbf{x}_0) \prod_{k=0}^{N-1} p(\mathbf{x}_{k+1}|\mathbf{x}_k)\ , \\ q(\mathbf{x}_{0:N}) &= q(\mathbf{x}_0) \prod_{k=0}^{N-1} q(\mathbf{x}_{k+1}|\mathbf{x}_k)\ , \qquad \text{(A3)}\end{aligned}$$

where $\mathbf{x}_{0:N}$ is shorthand notation for $(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_N)$.



(a)Timings with $N_o = 5$

(b)Timings with $N_o = 10$

FIG. 10. (a) and (b) summarize the timing results for both algorithms to attain the profiles, with different observation densities. All results are presented in seconds.

Using the factorization of Eq. (A3), we get:

$$\left\langle \ln \frac{q(\mathbf{x}_{0:N})}{p(\mathbf{x}_{0:N})} \right\rangle_q = \left\langle \ln \frac{q(\mathbf{x}_0)}{p(\mathbf{x}_0)} \right\rangle_q + \left\langle \sum_{k=0}^{N-1} \ln \frac{q(\mathbf{x}_{k+1}|\mathbf{x}_k)}{p(\mathbf{x}_{k+1}|\mathbf{x}_k)} \right\rangle_q \tag{A4}$$

For short times δt the transitions densities for both processes can be approximated by Gaussian densities:

$$\begin{aligned} p(\mathbf{x}_{k+1}|\mathbf{x}_k) &\simeq Z_p^{-1} \exp\left\{ -\frac{1}{2\delta t} \|\delta \mathbf{x}_{k+1} - \mathbf{f}(\mathbf{x}_k; \boldsymbol{\theta})\delta t\|^2_{\boldsymbol{\Sigma}} \right\} , \\ q(\mathbf{x}_{k+1}|\mathbf{x}_k) &\simeq Z_q^{-1} \exp\left\{ -\frac{1}{2\delta t} \|\delta \mathbf{x}_{k+1} - \mathbf{g}(\mathbf{x}_k)\delta t\|^2_{\boldsymbol{\Sigma}} \right\} . \end{aligned} \tag{A5}$$

Because the noise covariances $\boldsymbol{\Sigma}$ are identical for both processes the normalization constants are equivalent $Z_p = Z_q$. Note that the same is true even if the noises were time dependent. Therefore, using Eq. (A5) and taking the limit of $\delta t \to 0$, Eq. (A4) reduces to:

$$\left\langle \ln \frac{q(\mathbf{x}_{0:N})}{p(\mathbf{x}_{0:N})} \right\rangle_q = \left\langle \ln \frac{q(\mathbf{x}_0)}{p(\mathbf{x}_0)} \right\rangle_q + \frac{1}{2} \int_{t_0}^{t_f} \left\langle \|\mathbf{f}(\mathbf{x}_t; \boldsymbol{\theta}) - \mathbf{g}(\mathbf{x}_t)\|^2_{\boldsymbol{\Sigma}} \right\rangle_{q_t} dt. \tag{A6}$$

The last term in Eq. (A1), assuming that the discrete time observations have a Gaussian error distribution, i.e. $p(\mathbf{y}_k|\mathbf{x}_{t_k}) = \mathcal{N}(\mathbf{y}_k|\mathbf{x}_{t_k}, \mathbf{R})$, becomes:

$$\begin{aligned} -\langle \ln p(\mathcal{Y}|\mathbf{X}) \rangle_q &= -\left\langle \ln \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}_{t_k}) \right\rangle_q \\ &= -\left\langle \sum_{k=1}^{K} \ln \mathcal{N}(\mathbf{y}_k|\mathbf{x}_{t_k}, \mathbf{R}) \right\rangle_q \\ &= \frac{1}{2} \sum_{k=1}^{K} \left\langle \|\mathbf{y}_k - \mathbf{x}_{t_k}\|^2_{\mathbf{R}} \right\rangle_{q_t} + \text{const} . \end{aligned} \tag{A7}$$

[1] C. W. Gardiner, *Handbook of Stochastic Methods: for physics, chemistry and the natural sciences*, 3rd ed. (Springer Series in Synergetics, 2003).

[2] A. Golightly and D. J. Wilkinson, Computational Statistics and Data analysis **52**, 1674 (2007).

[3] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability* (Cambridge University Press, 2003).

[4] M. Opper and D. Saad, eds., *Advanced Mean Field Methods: Theory and Practice* (MIT Press, Cambridge, MA, 2001).

[5] F. J. Alexander, G. Eyink, and J. Restrepo, Journal of Statistical Physics **119**, 1331 (2005).

[6] P. Fearnhead, O. Papaspiliopoulos, and G. O. Roberts, Journal of the Royal Statistical Society **70**, 755 (2008).

[7] A. S. Paul and E. A. Wan, in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.* (2008) pp. 3621–3624.

[8] Y. Tremolet, Quarterly Journal of the Royal Meteorological Society **132**, 2483 (2006).

[9] G. L. Eyink, J. M. Restrepo, and F. J. Alexander, Physica D: Nonlinear Phenomena **195**, 347 (2004).

[10] M. Gilorami and B. Calderhead, Journal of Royal Statistical Society - B **73**, 123 (2011).

[11] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, Statistical Science **28**, 424 (2013).

[12] S. Brooks, *Handbook of Markov Chain Monte Carlo*, 1st ed., edited by S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, Handbooks of Modern Statistical Methods (Chapman and Hall / CRC, 2011).

[13] J. C. Quin and H. D. I. Abarbanel, Journal of Computational Physics **230**, 8168 (2011).

[14] G. Evensen, *Data assimilation: The Ensemble Kalman Filter*, 2nd ed. (Springer, 2009).

[15] P. J. van Leeuwen, Quarterly Journal of the Royal Meteorological Society **136**, 1991 (2010).

[16] F. Zhang, M. Zhang, and J. A. Hansen, Advances in Atmospheric Science **26**, 1 (2009).

[17] M. Zhang and F. Zhang, Monthly Weather Review **140**, 587 (2012).

[18] C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor, in *Advances in Neural Information Processing Systems 20*, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (2007) pp. 17–24.

[19] M. D. Vrettas, D. Cornford, and M. Opper, Physica D: Nonlinear Phenomena **240**, 1877 (2011).

[20] R. P. Feynman, *Statistical Mechanics: A Set Of Lectures*, 2nd ed., Advanced Books Classics (Westview Press, 1998).

[21] H. Kleinert, *Path Integral in Quantum Mechanics, Statistics,*

Polymer Physics and Financial Markets, 5th ed. (World Scientific, 2009).

[22] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, 3rd ed. (Springer, Applications of Mathematics, 1999).

[23] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor, Journal of Machine Learning Research (JMLR), Workshop and Conference Proceedings **1**, 1 (2007).

[24] A. Beskos, O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead, Journal of Royal Statistical Society **68**, 333 (2006).

[25] I. V. Girsanov, Theory of Probability and its Applications **V**, 285 (1960).

[26] B. Øksendal, *Stochastic Differential Equations*, 5th ed., An Introduction with Applications (Springer-Verlag, 2005).

[27] M. D. Vrettas, *Approximate Bayesian techniques for inference in stochastic dynamical systems*, Ph.D. thesis, Aston University, School of Engineering and Applied Science (2010).

[28] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Dover Publications, New York, 1964).

[29] E. N. Lorenz, Journal of Atmospheric Science **20**, 130 (1963).

[30] E. N. Lorenz, Journal of Atmospheric Science **62**, 1574 (2005).

[31] E. N. Lorenz and K. Emanuel, Journal of Atmospheric Science **55**, 399 (1998).

[32] J. Simon and U. Jeffrey, IEEE Transaction on Automatic Control. **45**, 477 (2000).

[33] G. Parisi, *Statistical Field Theory*, Advanced Book Classics (Westview Press, 1998).

[34] M. Opper and O. Winther, in *Advances in Neural Information Processing Systems*, Vol. 16, edited by S. Thrun, L. Saul, and B. Schölkopf (MIT Press, Cambridge, MA, 2004).