Expectation propagation for continuous time stochastic processes

Expectation propagation for continuous time stochastic processes

Botond Cseke*†[1], David Schnoerr*‡[2,3], Manfred Opper§[4], and Guido Sanguinetti¶[2]

[1]Microsoft Research, Cambridge, UK
[2]School of Informatics, University of Edinburgh, UK
[3]School of Biological Sciences, University of Edinburgh, UK
[4]Fakultät für Elektrotechnik und Informationstechnik, Technische Universität Berlin, DE

Abstract

We consider the inverse problem of reconstructing the posterior measure over the trajectories of a diffusion process from discrete time observations and continuous time constraints. We cast the problem in a Bayesian framework and derive approximations to the posterior distributions of single time marginals using variational approximate inference. We then show how the approximation can be extended to a wide class of discrete-state Markov jump processes by making use of the chemical Langevin equation. Our empirical results show that the proposed method is computationally efficient and provides good approximations for these classes of inverse problems.

1 Introduction

Physical and technological processes frequently exhibit intrinsic stochasticity. The main mathematical framework to describe and reason about such systems is provided by the theory of continuous time (Markovian) stochastic processes. Such processes have been well studied in chemical physics for several decades as models of chemical reactions at very low concentrations [Gardiner, 1985, e.g.]. More recently, the theory has found novel and diverse areas of application including systems biology at the single cell level [Wilkinson, 2011], ecology [Volkov et al., 2007] and performance modelling in computer systems [Hillston, 2005], to name but a few. The popularity of the approach has been greatly enhanced by the availability of efficient and accurate simulation algorithms [Gillespie, 1977, Gillespie et al., 2013], which permit a numerical solution of medium-sized systems within a reasonable time frame.

As with most of science, many of the application domains of continuous time stochastic processes are becoming increasingly data-rich, creating a critical demand for inference algorithms which can use data to calibrate the models and analyse the uncertainty in the predictions. This raises new challenges and opportunities for statistics and machine learning, and has motivated the development of several algorithms for efficient inference in these systems. In this paper, we focus on the Bayesian approach, and formulate the inverse problem in terms of obtaining an approximation to a posterior distribution over the stochastic process, given observations of the system and using existing scientific information to build a prior model of the process.

*Authors B. C. and D. S. contributed equally.
†Email: botcse@microsoft.com
‡Email: david.schnoerr@ed.ac.uk
§Email: opperm@cs.tu-berlin.de
¶Email: gsanguin@inf.ed.ac.uk

The data scenario which has attracted most attention within the Bayesian framework is the discretely (partially) observed case. In this scenario, the experiment returns noisy observations of the state (or some components) of the system at a precise set of time points. To proceed with Bayesian inference over the trajectories/parameters of the process, one needs to approximate the likelihood function, i.e. the conditional probability of the observations given the parameters. This is challenging, as likelihood computations are generally analytically intractable for continuous time processes, and has motivated the development of several approximation strategies based on sampling, variational approximations or system approximations [e.g. Beskos et al., 2006, Opper and Sanguinetti, 2008, Ruttor and Opper, 2009, Vrettas et al., 2015, Golightly et al., 2014, Zechner et al., 2014, Georgoulas et al., 2016]. An alternative data scenario which has received much less attention consists of qualitative observations of the trajectorys' behaviour. These are not uncommon: for example, in a biochemical experiment, we may observe that a protein level is above a certain detection threshold over a certain interval of time, without being able to precisely quantify its abundance at any time. In a computer science application, observations may be error logs which report whether the system's trajectory has violated e.g. some safety threshold over its run. This type of observations cannot be localised in time, but it is concerned with global properties of the system's trajectories: we term them continuous time constraints.

Evaluating the probability of a system satisfying some specific trajectory constraints is a highly non-trivial problem; in computer science, this is called the *model checking* problem (not to be confused with the problem of model checking in statistics, i.e. assessing the statistical fit to a data set). Evaluating this probability as a function of parameters, providing a likelihood for Bayesian inference, is even more challenging. Recent work, based on Gaussian process emulation and optimisation [Bortolussi and Sanguinetti, 2013, Bortolussi et al., 2015], has provided a practical solution for maximum likelihood parameter identification for small and medium scale systems; however, Gaussian process optimisation can only provide a point estimate of the parameters, and does not provide a posterior measure over the space of trajectories.

In this paper, we present a flexible approximate scheme for posterior inference in a wide class of stochastic processes from both discrete time observations and continuous time observations/trajectory constraints. The method can be seen as an extension to continuous time of the Expectation-Propagation (EP) approximate inference algorithm [Opper and Winther, 2000, Minka, 2001]. The algorithm was already presented in Cseke et al. [2013] for latent linear diffusion processes; in this paper, we extend that work in several ways. We extend the approach to a wider class of processes, including Markov jump processes (MJP), by applying moment closure of the corresponding chemical Langevin Equation (CLE). Furthermore, we present a novel derivation of the approach based on optimisation of a variational free energy [Opper and Winther, 2005, Heskes et al., 2005]. We demonstrate the approach on new numerical examples, demonstrating the effectiveness and efficiency of the approach on realistic models.

2 Models

In this paper we consider Bayesian models for diffusion processes that are observed in discrete and continuous time, where the continuous time observation model can be represented in a specific time integral form. This class of models for continuous time observations can represent a wide range of phenomena such as continuously observed state space models or path constraints.

We consider a diffusion process $\{\boldsymbol{x}_t\}$ with known dynamics defined on the time interval $[0, 1]$. We define the process $\{\boldsymbol{x}_t\}$ through the stochastic differential equation (SDE)

$$d\boldsymbol{x}_t = \boldsymbol{a}(\boldsymbol{x}_t, t, \boldsymbol{\theta})dt + \boldsymbol{b}(\boldsymbol{x}_t, t, \boldsymbol{\theta})^{1/2} d\boldsymbol{W}_t, \tag{1}$$

where $\{\boldsymbol{W}_t\}$ is the standard Wiener process [Gardiner, 1985] and $\boldsymbol{a}(\boldsymbol{x}_t, t, \boldsymbol{\theta})$ and $\boldsymbol{b}(\boldsymbol{x}_t, t, \boldsymbol{\theta})$ are vector and matrix valued functions respectively with $\boldsymbol{b}(t, \boldsymbol{x}_t, \boldsymbol{\theta})$ being positive semi-definite for all $t \in [0, 1]$. The functions $\boldsymbol{a}(\boldsymbol{x}_t, t, \boldsymbol{\theta})$ and $\boldsymbol{b}(\boldsymbol{x}_t, t, \boldsymbol{\theta})$ are referred to as *drift* and *diffusion* functions, respectively. Note that the SDE in (1) is an informal way to represent the process defined by an Itô

integral formulation [Gardiner, 1985]. Alternatively, the process $\{\boldsymbol{x}_t\}$ can also be defined through the Markovian transition probabilities satisfying the Fokker-Planck equation

$$\partial_t p(\boldsymbol{x}_t|\boldsymbol{x}_s) = -\sum_i \partial_{x_i}[a_i(\boldsymbol{x}_t, t, \boldsymbol{\theta})p(\boldsymbol{x}_t|\boldsymbol{x}_s)] + \frac{1}{2}\sum_{ij} \partial_{x_i}\partial_{x_j}[b_{ij}(\boldsymbol{x}_t, t, \boldsymbol{\theta})p(\boldsymbol{x}_t|\boldsymbol{x}_s)]. \tag{2}$$

Even though the process does not possess a formulation through density functions (with respect to the Lebesgue measure), we use the alias $p(\boldsymbol{x})$ to denote the "probability density" of the path/trajectory $\{\boldsymbol{x}_t\}$ in order to be able to symbolically represent and manipulate the process (or its variables) in the Bayesian formalism. Alternatively, we also use this notation as a reference to a process.

We assume that the process can be observed (noisily) both at discrete time points and for continuous time intervals; we use the $\boldsymbol{y}_i$ to denote the discrete time observations at times $t_i \in [0, 1]$ and $\boldsymbol{y}_t$ to denote the continuous time observations for all $t \in [0, 1]$. We also use $\boldsymbol{y} = \{\{\boldsymbol{y}_i\}, \{\boldsymbol{y}_t\}\}$ to shorten notation where necessary. In this paper we consider models where the observation likelihood admits the general formulation

$$p(\{\boldsymbol{y}_i\}, \{\boldsymbol{y}_t\}|\boldsymbol{x}, \boldsymbol{\theta}) \propto \prod_i p(\boldsymbol{y}_i|\boldsymbol{x}_{t_i}, \boldsymbol{\theta}) \times \exp\left\{-\int_0^1 dt\, U(\boldsymbol{y}_t, \boldsymbol{x}_t, t, \boldsymbol{\theta})\right\}. \tag{3}$$

The term $p(\boldsymbol{y}_i|\boldsymbol{x}_{t_i}, \boldsymbol{\theta})$ is the conditional probability of the discrete observation $\boldsymbol{y}_i$ given the state of the process at time t_i, while the function $U(\boldsymbol{y}_t, \boldsymbol{x}_t, t, \boldsymbol{\theta})$ is the negative log-likelihood of the continuous time observations (also referred to as loss function). We choose this class of models because they are sufficiently expressive to model a wide range of phenomena and because the Markovian structure of the probabilistic model is preserved, thus making Bayesian inference accessible. For example, the well known linear dynamical systems with the continuous time observation model $d\boldsymbol{y}_t = \boldsymbol{C}_t\boldsymbol{x}_t dt + \boldsymbol{R}_t^{1/2} d\boldsymbol{W}_t$ can be be formulated as follows: (1) we choose a linear drift $\boldsymbol{a}(\boldsymbol{x}_t, t, \boldsymbol{\theta}) = \boldsymbol{a}_t\boldsymbol{x}_t$ and a time-only dependent diffusion $\boldsymbol{b}(\boldsymbol{x}_t, t, \boldsymbol{\theta}) = \boldsymbol{b}_t$ (2) we choose the continuous time loss as the quadratic function $U(\boldsymbol{y}_t, \boldsymbol{x}_t, t, \theta) = -\boldsymbol{x}_t^T[\boldsymbol{C}_t^T\boldsymbol{R}_t^{-1}(d\boldsymbol{y}_t/dt)] - \boldsymbol{x}_t^T[\boldsymbol{C}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{C}_t]\boldsymbol{x}_t/2$. Another class of models where such time integral representations are useful are temporal log Gaussian Cox process models [e.g. Cseke et al., 2013, Harel et al., 2015].

Bayesian inference in diffusion process models is generally understood as the inference of the posterior process

$$p(\boldsymbol{x}|\{\boldsymbol{y}_i\}, \{\boldsymbol{y}_t\}, \boldsymbol{\theta}) \propto p(\boldsymbol{x}|\boldsymbol{\theta}) \times p(\{\boldsymbol{y}_i\}|\boldsymbol{x}, \boldsymbol{\theta}) \times p(\{\boldsymbol{y}_t\}|\boldsymbol{x}, \boldsymbol{\theta}), \tag{4}$$

and model evidence

$$p(\boldsymbol{y}\,|\,\boldsymbol{\theta}) \propto \int d\boldsymbol{x}\, p(\boldsymbol{x}|\boldsymbol{\theta}) \times p(\{\boldsymbol{y}_i\}|\boldsymbol{x}, \boldsymbol{\theta}) \times p(\{\boldsymbol{y}_t\}|\boldsymbol{x}, \boldsymbol{\theta}). \tag{5}$$

Clearly, the existence of a closed form specification of the posterior process (4) as, say, defined by a new pair of drift and diffusion functions, would be desirable. However, this is only possible in a few special cases such as Ornstein-Uhlenbeck/Gaussian-Markov processes with Gaussian discrete time observations and quadratic loss functions. For this reason we aim to approximate the time marginals $p(\boldsymbol{x}_t|\,\boldsymbol{y}, \boldsymbol{\theta})$ (referred to as marginals hereafter) and the model evidence (5).

In this paper we address the case where the intractability is mainly due to the likelihood terms; intractability in the prior (e.g. due to nonlinear drift/ diffusion terms) can be handled as long as efficient methods for approximating marginal moments exist (see Section 3.2). We address this problem in two steps. First, we approximate the likelihood terms by terms that have exponential forms such as

$$p(\boldsymbol{y}_i|\boldsymbol{x}_{t_i}, \boldsymbol{\theta}) \approx \exp\{\boldsymbol{\xi}_i^T\boldsymbol{f}(\boldsymbol{x}_{t_i})\} \quad \text{and} \quad U(\boldsymbol{y}_t, \boldsymbol{x}_t, t, \boldsymbol{\theta}) \approx \boldsymbol{\xi}_t^T\boldsymbol{f}(\boldsymbol{x}_t) + \text{constant}. \tag{6}$$

Second, we propose approximations to the posterior marginals $p(\boldsymbol{x}_t|\,\boldsymbol{y}, \boldsymbol{\theta})$ and the model evidence $p(\boldsymbol{y}|\,\boldsymbol{\theta})$ given by (6). Here, the choice of the function $\boldsymbol{f}$ typically follows from the model and

the approximation method. For example, when the prior $p(\boldsymbol{x}|\boldsymbol{\theta})$ is Gaussian-Markov or when we can obtain good Gaussian approximations to $p(\boldsymbol{x}_t|\boldsymbol{\theta})$ we choose $\boldsymbol{f}$ to be linear and quadratic $\boldsymbol{f}(\boldsymbol{x}_t) = (\boldsymbol{x}_t, -\boldsymbol{x}_t\boldsymbol{x}_t^T/2)$, thus corresponding to a Gaussian likelihood approximation. In some cases, however, the resulting computations can still be intractable and a factorising Gaussian corresponding to $\boldsymbol{f}(\boldsymbol{x}_t) = (\{x_t^j\}, \{-x_t^j x_t^j/2\})$ is chosen to make computations tractable. Throughout this paper we consider $\boldsymbol{f}$ to correspond to a multivariate Gaussian, however, to simplify notation and to emphasise that the results hold for any suitably chosen $\boldsymbol{f}$ or any restricted class of Gaussian, we opt for this exponential form representation.

3 Approximate inference

In this section we introduce an *expectation propagation* (EP) based method to approximate the marginals $p(\boldsymbol{x}_t|\boldsymbol{y}, \boldsymbol{\theta})$. In its original derivation [Opper and Winther, 2000, Minka, 2001], expectation propagation is an algorithm for approximating an intractable distribution. One assumes that the distribution is written as a product of certain terms, where typically one term is the prior and the other terms correspond to the likelihoods of individual observations. In the approximating distribution, terms are replaced by tractable 'likelihood' proxies. The EP algorithm aims at achieving consistency between the moments of the approximating distribution and a set of auxiliary distributions. Each auxiliary distribution is obtained by replacing a *single* likelihood proxy by its original counterpart. For many problems auxiliary distributions are still tractable.

EP has been applied to dynamical models [Heskes and Zoeter, 2002, Ypma and Heskes, 2005, Barber, 2006] in discrete time. However, a generalisation to continuous time processes is not straightforward. It is not immediately clear what an addition or removing of a 'single' likelihood term at a given time means for the case of continuous time observations/constraints.

We will show in the following that by first applying EP to a time discretised process and by taking a subsequent continuous time limit, we obtain a well defined algorithm, which treats discrete and continuous time likelihoods in different ways.

Our derivation of EP for stochastic processes will not follow the original derivation of the EP algorithm in [Minka, 2001] but we will use an equivalent free energy approximation instead. We thereby follow a line of arguments similar to Opper and Winther [2005] where the fixed points of the EP algorithm were derived as the stationary points of a specific free energy function. This strategy is similar to the derivation of belief propagation algorithms from a Bethe free energy [e.g. Yedidia et al., 2000, Heskes, 2003].

Our approach will proivde us with an approximation to the log partition function

$$\begin{aligned}\log Z(\boldsymbol{\theta}) &= \log p(\boldsymbol{y}|\boldsymbol{\theta}) \\ &= \log \int d\boldsymbol{x}\, p(\boldsymbol{x}) \exp\Big\{ -\int_0^1 dt U(\boldsymbol{y}_t, \boldsymbol{x}_t, t, \boldsymbol{\theta})\Big\} \prod_i p(\boldsymbol{y}_i|\boldsymbol{x}_{t_i}, \boldsymbol{\theta}),\end{aligned} \tag{7}$$

which normalises the posterior distribution, and also approximations to marginal posterior moments

$$\langle \boldsymbol{f}(\boldsymbol{x}_t)\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y})} = \int d\boldsymbol{x}_t\, p(\boldsymbol{x}_t|\boldsymbol{y}) \boldsymbol{f}(\boldsymbol{x}_t), \tag{8}$$

where $\boldsymbol{f}$ corresponds to moments up to second order. As mentioned above, in this paper we focus on the Gaussian case where $\boldsymbol{f}(\boldsymbol{x}_t) = (\boldsymbol{x}_t, -\boldsymbol{x}_t\boldsymbol{x}_t^2/2)$ or $\boldsymbol{f}(\boldsymbol{x}_t) = (\{x_t^j\}, \{-x_t^j x_t^j/2\})$. However, our approach is sufficiently general to accommodate other choices of $\boldsymbol{f}$. Note that the integral in (7) is over the path $\boldsymbol{x}$.

Notation. Since we focus on approximating $\log p(\boldsymbol{y}|\boldsymbol{\theta})$ and the moments $\langle \boldsymbol{f}(\boldsymbol{x}_t)\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y},\boldsymbol{\theta})}$ for a fixed $\boldsymbol{\theta}$, we omit $\boldsymbol{\theta}$ from our notation in the following. Moreover, we use the shorthand notation $U(\boldsymbol{x}_t, t) = U(\boldsymbol{y}_t, \boldsymbol{x}_t, t, \boldsymbol{\theta})$ and use $\boldsymbol{z}_1 \cdot \boldsymbol{z}_2 = \boldsymbol{z}_1^T \boldsymbol{z}_2$ as an alternative notation for vector inner product.

We refer to the exponential family of distributions defined by the sufficient statistic $\boldsymbol{f}$ as the family $\mathcal{F}$ of distributions having the form $r_{\boldsymbol{\xi}}(\boldsymbol{z}) = \exp\{\boldsymbol{\xi} \cdot \boldsymbol{f}(\boldsymbol{z}) - \log Z_f(\boldsymbol{\xi})\}$ with $\log Z_f(\boldsymbol{\xi}) =$

$\log \int d\boldsymbol{z} \exp\{\boldsymbol{\xi} \cdot \boldsymbol{f}(\boldsymbol{z})\}$. We refer to the parameters $\boldsymbol{\xi}$ as canonical parameters. The mean and covariance of $\boldsymbol{f}(\boldsymbol{z})$ can be computed as $\langle \boldsymbol{f}(\boldsymbol{z}) \rangle_{r_{\boldsymbol{\xi}}} = \partial_{\boldsymbol{\xi}} \log Z_f(\boldsymbol{\xi})$ and $\langle \boldsymbol{f}(\boldsymbol{z}), \boldsymbol{f}(\boldsymbol{z}) \rangle_{r_{\boldsymbol{\xi}}} = \partial^2_{\boldsymbol{\xi}} \log Z_f(\boldsymbol{\xi})$. To avoid inconsistent notation, we use $\log Z$ to denote the log partition function in (7). In the following we assume that for any distribution $s(\boldsymbol{z})$ for which $\langle \boldsymbol{f}(\boldsymbol{z}) \rangle_s$ exists, there exists a unique canonical parameter vector $\boldsymbol{\xi}$ such that $\langle \boldsymbol{f}(\boldsymbol{x}) \rangle_s = \langle \boldsymbol{f}(\boldsymbol{z}) \rangle_{r_{\boldsymbol{\xi}}}$. This canonical parameter can be defined formally as

$$\boldsymbol{\xi} = \operatorname*{argmin}_{\boldsymbol{\xi}} \mathrm{KL}[\, s(\boldsymbol{x}) \,||\, \exp\{\boldsymbol{\xi} \cdot \boldsymbol{f}(\boldsymbol{x}_t) - \log Z_f(\boldsymbol{\xi})\}\,], \tag{9}$$

where KL denotes the Kullback-Leibler divergence $\mathrm{KL}[s_1(\boldsymbol{z})||s_2(\boldsymbol{z})] = \langle \log(s_1(\boldsymbol{z})/s_2(\boldsymbol{z})) \rangle_{s_1(\boldsymbol{z})}$. Throughout this paper we assume that given $\langle \boldsymbol{f}(\boldsymbol{z}) \rangle_s$, we can compute $\boldsymbol{\xi}$ efficiently; we denote it as $\boldsymbol{\xi} = \mathrm{Project}[s(\boldsymbol{x}); \boldsymbol{f}]$ and refer to this operation as "moment-to-canonical parameter transformation".

3.1 Approximating $\log Z$ and the posterior marginal moments

In order to approximate $\log Z$ in (7), we first introduce an auxiliary approximating process

$$q_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \frac{1}{Z_q(\boldsymbol{\lambda})} p(\boldsymbol{x}) \exp\Big\{ \int_0^1 dt \boldsymbol{\lambda}_t \cdot \boldsymbol{f}(\boldsymbol{x}_t) + \sum_i \boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i}) \Big\}, \tag{10}$$

where the likelihoods are replaced by simpler likelihood proxies. We assume that the partition function $\log Z_q(\boldsymbol{\lambda})$ and the marginal moments $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t)$ of the process in (10) are computationally tractable or can be approximated with reasonable accuracy. We will present a suitable approximation method in Section 3.2. Here the parameter $\boldsymbol{\lambda} = \{\{\boldsymbol{\lambda}_t\}, \{\boldsymbol{\lambda}_i\}\}$ is a variational parameter to be optimised later. Using the process (10) and its partition function, we can represent (7) as

$$\begin{aligned} \log Z = \log \int d\boldsymbol{x}\, p(\boldsymbol{x}) \exp\Big\{ \int_0^1 dt \boldsymbol{\lambda}_t \cdot \boldsymbol{f}(\boldsymbol{x}_t) + \sum_i \boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i}) \Big\} \\ \times \exp\Big\{ -\int_0^1 dt U(\boldsymbol{x}_t, t) - \int_0^1 dt \boldsymbol{\lambda}_t \cdot \boldsymbol{f}(\boldsymbol{x}_t) \Big\} \times \prod_i p(\boldsymbol{y}_i | \boldsymbol{x}_{t_i}) e^{-\boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i})}. \end{aligned} \tag{11}$$

Rewriting this using (10) we obtain

$$\begin{aligned} \log Z = \log Z_q(\boldsymbol{\lambda}) \\ + \log \Big\langle \exp\Big\{ -\int_0^1 dt U(\boldsymbol{x}_t, t) - \int_0^1 dt \boldsymbol{\lambda}_t \cdot \boldsymbol{f}(\boldsymbol{x}_t) \Big\} \times \prod_i p(\boldsymbol{y}_i | \boldsymbol{x}_{t_i}) e^{-\boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i})} \Big\rangle_{q_{\boldsymbol{\lambda}}}. \end{aligned} \tag{12}$$

This yields an expression of $\log Z$ as the sum of a tractable log–partition function $\log Z_q(\boldsymbol{\lambda})$ and a correction term accounting for the "error" resulting from replacing the likelihoods by simpler proxies. A popular approach to simplify the correction would be to use Jensen's inequality in (11) and move the expectation from inside to the outside of the logarithm. This would give the approximating bound

$$\begin{aligned} \log Z \geq \log Z_q(\boldsymbol{\lambda}) \\ - \int_0^1 dt \big\langle U(\boldsymbol{x}_t, t) \big\rangle_{q_{\boldsymbol{\lambda}}} + \sum_i \big\langle \log p(\boldsymbol{y}_i | \boldsymbol{x}_{t_i}) \big\rangle_{q_{\boldsymbol{\lambda}}} - \int_0^1 dt \boldsymbol{\lambda}_t \cdot \big\langle \boldsymbol{f}(\boldsymbol{x}_t) \big\rangle_{q_{\boldsymbol{\lambda}}} - \sum_i \boldsymbol{\lambda}_i \cdot \big\langle \boldsymbol{f}(\boldsymbol{x}_{t_i}) \big\rangle_{q_{\boldsymbol{\lambda}}}. \end{aligned} \tag{13}$$

This approximation would be followed by an optimisation of the resulting lower bound with respect to the variational parameters $\boldsymbol{\lambda}$. For recent applications to inference in diffusion processes, see [Archambeau et al., 2007, Ala-Luhtala et al., 2014, Vrettas et al., 2015, Sutter et al., 2015].

The EP approximation to the expectation term in (12) proceeds in a different way. To this end, we define a set of marginals $\mathcal{E}(\boldsymbol{\eta}_t) = \{q_{\boldsymbol{\eta}_t}(\boldsymbol{x}_t); t \in [0,1]\}$ with $q_{\boldsymbol{\eta}_t}(\boldsymbol{x}_t) = \exp\{\boldsymbol{\eta}_t \cdot \boldsymbol{f}(\boldsymbol{x}_t) - \log Z_f(\boldsymbol{\eta}_t)\}$ that satisfy the marginal moment matching constraints

$$\langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\lambda}}} = \langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\eta}_t}} \text{ for all } t \in [0,1]. \tag{14}$$

Note that for a Gaussian approximation to the posterior process, these would simply be the Gaussian marginal densities. Following a similar route as in Opper and Winther [2005] in their derivation of EP, our goal will be to approximate the intractable average over the process $q_{\boldsymbol{\lambda}}$ by a distribution which *factorises in time* and is given by the product of the densities $q_{\boldsymbol{\eta}_t}(\boldsymbol{x}_t)$. Hence, we retain the information about marginal statistics, but lose the dependencies between different time points. This type of approximation—approximating the expectation of a product with the product of expectations given certain moment constraints—is used in a variety of successful approximation methods such as expectation propagation in latent Gaussian models [Opper and Winther, 2000, Minka, 2001] or belief propagation in pairwise graphical models [e.g. Yedidia et al., 2000]. The approximation we propose here can be viewed as the extension of this approach to continuous time (infinite dimensional) models. Of course, it is not trivial to define such a factorising density (corresponding to a delta–correlated process) directly in continuous time. Hence, it will be useful to introduce a discretisation in time into slices of size Δt first and then proceed to the limit $\Delta t \to 0$. We use the Euler discretisation $\int_0^1 dt U(\boldsymbol{x}_t, t) \backsimeq \sum_k \Delta t U(\boldsymbol{x}_{k\Delta t}, k\Delta t)$ and $\int_0^1 dt \boldsymbol{\lambda}_t \cdot \boldsymbol{f}(\boldsymbol{x}_t) \backsimeq \sum_k \Delta t\, \boldsymbol{\lambda}_{k\Delta t} \cdot \boldsymbol{f}(\boldsymbol{x}_{k\Delta t})$ and approximate the expectation of a product, see (12), as

$$\begin{aligned}
&\Big\langle \prod_k \exp\Big\{ -\Delta t\, U(\boldsymbol{x}_{k\Delta t}, k\Delta t) - \Delta t\, \boldsymbol{\lambda}_{k\Delta t} \cdot \boldsymbol{f}(\boldsymbol{x}_{k\Delta t}) \Big\} \times \prod_i p(\boldsymbol{y}_i|\boldsymbol{x}_{t_i}) e^{-\boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i})} \Big\rangle_{q_{\boldsymbol{\lambda}}} \\
&\quad \approx \prod_k \Big\langle \exp\Big\{ -\Delta t\, U(\boldsymbol{x}_{k\Delta t}, k\Delta t) - \Delta t\, \boldsymbol{\lambda}_{k\Delta t} \cdot \boldsymbol{f}(\boldsymbol{x}_{k\Delta t}) \Big\} \Big\rangle_{q_{\boldsymbol{\eta}_{k\Delta t}}} \times \prod_i \Big\langle p(\boldsymbol{y}_i|\boldsymbol{x}_{t_i}) e^{-\boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i})} \Big\rangle_{q_{\boldsymbol{\eta}_{k\Delta t}}}.
\end{aligned}$$

Taking the limit $\Delta t \to 0$ results in the approximation

$$\begin{aligned}
&\log\Big\langle \exp\Big\{ -\int_0^1 dt U(\boldsymbol{x}_t, t) - \int_0^1 dt \boldsymbol{\lambda}_t \cdot \boldsymbol{f}(\boldsymbol{x}_t) \Big\} \times \prod_i p(\boldsymbol{y}_i|\boldsymbol{x}_{t_i}) e^{-\boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i})} \Big\rangle_{q_{\boldsymbol{\lambda}}} \\
&\quad \approx -\int_0^1 dt \big\langle U(\boldsymbol{x}_t, t) \big\rangle_{q_{\boldsymbol{\eta}_t}} - \int_0^1 dt \boldsymbol{\lambda}_t \cdot \big\langle \boldsymbol{f}(\boldsymbol{x}_t) \big\rangle_{q_{\boldsymbol{\eta}_t}} + \sum_i \log \Big\langle p(\boldsymbol{y}_i|\boldsymbol{x}_{t_i}) e^{-\boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i})} \Big\rangle_{q_{\boldsymbol{\eta}_{t_i}}},
\end{aligned} \tag{15}$$

where the first two terms follow from $\log[\langle \exp\{-\Delta t\, U(\boldsymbol{x}_t, t)\}\rangle_{q_{\boldsymbol{\eta}_t}}] \backsimeq -\Delta t \langle U(\boldsymbol{x}_t, t)\rangle_{q_{\boldsymbol{\eta}_t}}$. A comparison with (13) shows that the first two terms in (15) corresponding to continuous time likelihoods would equal their counterparts in the variational bound if the densities $q_{\boldsymbol{\eta}_t}(\boldsymbol{x}_t)$ are the correct marginals of the process $q_{\boldsymbol{\lambda}}(\boldsymbol{x})$. This is the case for a Gaussian posterior approximation. However, the second term is different.

By introducing the variables $\boldsymbol{\eta}_i = \boldsymbol{\eta}_{t_i} - \boldsymbol{\lambda}_i$, $\boldsymbol{\eta} = \{\{\boldsymbol{\eta}_t\}, \{\boldsymbol{\eta}_i\}\}$ and applying (15) to (12) we obtain the approximation

$$\begin{aligned}
\ln Z \approx L(\boldsymbol{\lambda}, \boldsymbol{\eta}) \equiv \log Z_q(\boldsymbol{\lambda}) &+ \sum_i \log \int dx_{t_i}\, p(\boldsymbol{y}_i|\boldsymbol{x}_{t_i}) e^{\boldsymbol{\eta}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i})} - \sum_i \log Z_f(\boldsymbol{\lambda}_i + \boldsymbol{\eta}_i) \\
&- \int_0^1 dt \big\langle U(\boldsymbol{x}_t, t) \big\rangle_{q_{\boldsymbol{\eta}_t}} - \int_0^1 dt \boldsymbol{\lambda}_t \cdot \big\langle \boldsymbol{f}(\boldsymbol{x}_t) \big\rangle_{q_{\boldsymbol{\eta}_t}},
\end{aligned} \tag{16}$$

where we require the marginal moment matching constraints in (14) to hold. This approximation contains the two sets of parameters $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$. Following similar arguments as given in Opper and Winther [2005] to derive EP for Gaussian latent variable models, we will now argue that it makes sense to optimise the approximation by computing the stationary points of $L(\boldsymbol{\lambda}, \boldsymbol{\eta})$ with respect to the variation of these parameters. In fact, we can show that variation w.r.t. $\boldsymbol{\lambda}_t$ leads to the moment matching condition (14). Since the exact partition function does not depend on $\boldsymbol{\lambda}$ it also makes sense to set the variation w.r.t. $\boldsymbol{\lambda}$ to zero, thereby making the approximation least sensitive w.r.t. to variation of $\boldsymbol{\lambda}$.

Using straightforward calculus one can show that the differentials of (16) w.r.t. $\boldsymbol{\eta}_i$ and $\boldsymbol{\lambda}_i$ are given by

$$\partial_{\boldsymbol{\eta}_i} L = \langle \boldsymbol{f}(\boldsymbol{x}_{t_i}) \rangle_{\tilde{q}_{\boldsymbol{\eta}_i}} - \langle \boldsymbol{f}(\boldsymbol{x}_{t_i}) \rangle_{q_{\boldsymbol{\eta}_i + \boldsymbol{\lambda}_i}} \qquad \partial_{\boldsymbol{\lambda}_i} L = \langle \boldsymbol{f}(\boldsymbol{x}_{t_i}) \rangle_{q_\lambda} - \langle \boldsymbol{f}(\boldsymbol{x}_{t_i}) \rangle_{q_{\boldsymbol{\eta}_i + \boldsymbol{\lambda}_i}}, \tag{17}$$

where we use $\tilde{q}_{\boldsymbol{\eta}_i}$ to denote the distribution

$$\tilde{q}_{\boldsymbol{\eta}_i}(\boldsymbol{x}_t) \propto p(\boldsymbol{y}_i | \boldsymbol{x}_{t_i}) e^{\boldsymbol{\eta}_i \cdot \boldsymbol{f}(x_{t_i})}. \tag{18}$$

The variations of (16) w.r.t. $\boldsymbol{\eta}_t$ and $\boldsymbol{\lambda}_t$ are

$$\delta_{\boldsymbol{\eta}_t} L = -\partial_{\boldsymbol{\eta}_t} \big\langle U(\boldsymbol{x}_t, t) \big\rangle_{q_{\boldsymbol{\eta}_t}} - \partial^2_{\boldsymbol{\eta}_t} \log Z_f(\boldsymbol{\eta}_t) \boldsymbol{\lambda}_t \quad \text{and} \quad \delta_{\boldsymbol{\lambda}_t} L = \langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\lambda}}} - \langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\eta}_t}}, \tag{19}$$

where we use of the $\langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\eta}_t}} = \partial_{\boldsymbol{\eta}_t} \log Z_f(\boldsymbol{\eta}_t)$ property of the exponential family distributions. Note that from $\delta_{\boldsymbol{\lambda}_t} L = 0$ we recover the marginal moment matching constraints postulated in (14) and $\boldsymbol{\eta}_i = \boldsymbol{\eta}_{t_i} - \boldsymbol{\lambda}_i$ is guaranteed to hold when setting $\partial_{\boldsymbol{\lambda}_i} L = 0$ and $\delta_{\boldsymbol{\lambda}_t} L = 0$.

Optimisation

Except $\delta_{\boldsymbol{\eta}_t} L = 0$, all other stationary conditions corresponding to (17) and (19) can be viewed as moment matching conditions. Since $q_{\boldsymbol{\eta}_t}$ is from the exponential family, $\partial_{\boldsymbol{\eta}_i} L = 0$ and $\delta_{\boldsymbol{\lambda}_i} L = 0$ can be expressed in terms of canonical parameters as

$$\boldsymbol{\lambda}_i + \boldsymbol{\eta}_i = \text{Project}[\tilde{q}_{\boldsymbol{\eta}_i}(\boldsymbol{x}_{t_i}); \boldsymbol{f}]. \quad \text{and} \quad \boldsymbol{\lambda}_i + \boldsymbol{\eta}_i = \text{Project}[q_{\boldsymbol{\lambda}}(\boldsymbol{x}_{t_i}); \boldsymbol{f}]. \tag{20}$$

We then use (20) to define the fixed point updates

$$\boldsymbol{\lambda}_i^{new} = \text{Project}[\tilde{q}_{\boldsymbol{\eta}_i}(\boldsymbol{x}_{t_i}); \boldsymbol{f}] - \boldsymbol{\eta}_i \quad \text{and} \quad \boldsymbol{\eta}_i^{new} = \text{Project}[q_{\boldsymbol{\lambda}}(\boldsymbol{x}_{t_i}); \boldsymbol{f}] - \boldsymbol{\lambda}_i. \tag{21}$$

Similarly, from (19), we obtain updates

$$\boldsymbol{\lambda}_t^{new} = -[\partial^2_{\boldsymbol{\eta}_t} \log Z_f(\boldsymbol{\eta}_t)]^{-1} \partial_{\boldsymbol{\eta}_t} \big\langle U(\boldsymbol{x}_t, t) \big\rangle_{q_{\boldsymbol{\eta}_t}} \quad \text{and} \quad \boldsymbol{\eta}_t^{new} = \text{Project}[q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t); \boldsymbol{f}]. \tag{22}$$

Readers familiar with the expectation propagation frameworks proposed in [Opper and Winther, 2000], [Minka, 2001] and [Heskes et al., 2005] can identify $\boldsymbol{\lambda}_i$ as the canonical parameters of the term approximations. The distributions $\tilde{q}_{\boldsymbol{\eta}_i}$ are the tilted distributions and $\boldsymbol{\eta}_i$ are the parameters of the so-called cavity distributions. The updates in (21) for the discrete time likelihood proxies correspond to expectation propagation updates. The updates in (22) correspond to non-conjugate variational updates [Knowles and Minka, 2011]. A similar fixed point iteration for latent Gaussian-Markov models has been derived in Cseke et al. [2013] by applying expectation propagation to the Euler discretisation of the posterior process.

3.2 Moment approximations by moment closures

In order to run the fixed point iteration we need to (approximately) compute the canonical parameters corresponding to the updates in (21) and (22). Since we use exponential family distributions, this simplifies to (approximately) computing $\langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{\tilde{q}_{\boldsymbol{\eta}_i}}$, $\langle U(\boldsymbol{x}_t, t) \rangle_{q_{\boldsymbol{\eta}_t}}$, and $\langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\lambda}}}$ and computing the corresponding canonical parameters. As postulated in Section 1, $\boldsymbol{f}(\boldsymbol{x}_t)$ is chosen to correspond to a Gaussian approximation (linear and quadratic).

We further assume that good numerical approximation for $\langle U(\boldsymbol{x}_t, t) \rangle_{q_{\boldsymbol{\eta}_t}}$ and $\langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{\tilde{q}_{\boldsymbol{\eta}_i}}$ exist and the (computational) bottleneck of the proposed method is the computation or approximation of $\langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\lambda}}}$. The latter corresponds to computing the marginal moments of the process $q_{\boldsymbol{\lambda}_t}$ in (10) and thus requires the solution of a finite set of ODEs. If the assumptions for $\langle U(\boldsymbol{x}_t, t) \rangle_{q_{\boldsymbol{\eta}_t}}$ and $\langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{\tilde{q}_{\boldsymbol{\eta}_i}}$ do not hold (e.g. $\tilde{q}_{\boldsymbol{\eta}_i}$ is a complicated multivariate density because of $q_{\boldsymbol{\eta}_t}$), we can relax the problem by choosing a restricted family of approximations $\mathcal{E}(\boldsymbol{\eta})$ corresponding to "weaker" sufficient statistics, say, $\boldsymbol{f}(\boldsymbol{x}_t) = (\{x_t^i\}, \{-[x_t^i]^2/2\})$.

Now we introduce approximations to $\langle \boldsymbol{f}(\boldsymbol{x}_t)\rangle_{q_{\boldsymbol{\lambda}}}$. Due to the Markovian nature of the process $q_{\boldsymbol{\lambda}}(\boldsymbol{x})$, the exact marginals $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t)$ can be expressed in terms of the distributions $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t|\,\boldsymbol{\lambda}_{i:t_i\leq t}, \boldsymbol{\lambda}_{s\leq t})$ and the conditional likelihoods $q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_{i:t_i>t}, \boldsymbol{\lambda}_{s>t}\,|\,\boldsymbol{x}_t)$ of $\{\boldsymbol{\lambda}_{i:t_i>t}, \boldsymbol{\lambda}_{s>t}\}$ as

$$q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t) \propto q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_{i:t_i>t}, \boldsymbol{\lambda}_{s>t}\,|\,\boldsymbol{x}_t)\, q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t|\,\boldsymbol{\lambda}_{i:t_i\leq t}, \boldsymbol{\lambda}_{s\leq t}), \tag{23}$$

where we define

$$q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t\,|\,\boldsymbol{\lambda}_{i:t_i\leq t}, \boldsymbol{\lambda}_{s\leq t}) \propto \int d\boldsymbol{x}_{s<t}\, p(\boldsymbol{x}_{s\leq t}) \exp\Big\{ \int_0^t ds\, \boldsymbol{\lambda}_s \cdot \boldsymbol{f}(\boldsymbol{x}_s) + \sum_{i:t_i\leq t} \boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i})\Big\}, \tag{24}$$

and

$$q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_{i:t_i>t}, \boldsymbol{\lambda}_{s>t}\,|\,\boldsymbol{x}_t) \propto \int d\boldsymbol{x}_{s>t}\, p(\boldsymbol{x}_{s>t}\,|\,\boldsymbol{x}_t) \exp\Big\{ \int_t^1 ds\, \boldsymbol{\lambda}_s \cdot \boldsymbol{f}(\boldsymbol{x}_s) + \sum_{i:t_i>t} \boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i})\Big\}. \tag{25}$$

Generally, there are two ways to compute (23):

(i) We independently compute $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t\,|\,\boldsymbol{\lambda}_{i:t_i\leq t}, \boldsymbol{\lambda}_{s\leq t})$ and $q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_{i:t_i>t}, \boldsymbol{\lambda}_{s>t}\,|\,\boldsymbol{x}_t)$ by combining the solutions of the forward and backward Fokker-Planck equations—corresponding to prior (2)— with iterative Bayesian updates corresponding to the likelihood proxies in (24) and (25). We then multiply these quantities to obtain the marginals in (23).

(ii) Instead of computing $q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_{i:t_i>t}, \boldsymbol{\lambda}_{s>t}\,|\,\boldsymbol{x}_t)$ we compute $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t)$ directly by making use of the *smoothing equation* for $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t)$ [Striebel, 1965, Leondes et al., 1970]. The smoothing equation depends on $\boldsymbol{\lambda}$ only through $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t\,|\,\boldsymbol{\lambda}_{i:t_i\leq t}, \boldsymbol{\lambda}_{s\leq t})$ which are computed as in (i).

In the following we use the latter approach. We do this because the approximation method we introduce in the next section is not well defined for conditional likelihoods like $q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_{i:t_i>t}, \boldsymbol{\lambda}_{s>t}\,|\,\boldsymbol{x}_t)$.

When the prior process $p(\boldsymbol{x}_t)$ is linear, that is, $d\boldsymbol{x} = \boldsymbol{a}_t\boldsymbol{x}_t dt + \boldsymbol{b}_t^{1/2} d\boldsymbol{W}_t$, and $\boldsymbol{f}$ is linear and quadratic, the computations result in solving the Kalman-Bucy forward and backward equations [e.g. Särkkä and Sarmavuori, 2013]. These are a set ODEs for the mean and covariance of the corresponding Gaussian distributions for which efficient numerical methods exist. However, when the prior $p(\boldsymbol{x}_t)$ is non-linear we have to resort to approximations as the computational cost of solving the forward/backward Fokker-Plank equations is generally excessive. Most approximations proposed in the literature assume a parametric form for $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t\,|\,\boldsymbol{\lambda}_{i:t_i\leq t}, \boldsymbol{\lambda}_{s\leq t})$ and $q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_{i:t_i>t}, \boldsymbol{\lambda}_{s>t}\,|\,\boldsymbol{x}_t)$ and derive ODEs for the parameters of the corresponding approximations. For example, there is a variety of different approximation methods using (multivariate) Gaussian approximations presented in Särkkä [2010] and Särkkä and Sarmavuori [2013].

3.2.1 Forward moment approximations

As mentioned above, to compute $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t\,|\,\boldsymbol{\lambda}_{i:t_i\leq t}, \boldsymbol{\lambda}_{s\leq t})$ we need to solve a non-linear Fokker Planck equation (2) for which generally no analytic solutions are known. However, note that in our approach we only require the moments $\langle \boldsymbol{f}(\boldsymbol{x}_t)\rangle_{q_{\boldsymbol{\lambda}}}$ and we hence aim at approximating these directly. If we multiply (2) with $\boldsymbol{f}(\boldsymbol{x}_t)$ and take the expectation, we obtain the following ODEs for the moments $\langle \boldsymbol{f}(\boldsymbol{x}_t)\rangle_p$ of the solution p of (2):

$$\partial_t \langle f_l(\boldsymbol{x}_t)\rangle_p = \sum_j \big\langle a_j(\boldsymbol{x}_t, t)\partial_{x_j} f_l(\boldsymbol{x}_t)\big\rangle_p + \frac{1}{2}\sum_{j,k} \big\langle b_{jk}(\boldsymbol{x}_t, t)\partial_{x_j}\partial_{x_k} f_l(\boldsymbol{x}_t)\big\rangle_p . \tag{26}$$

The moments $\langle \boldsymbol{f}(\boldsymbol{x}_t)\rangle_{q_{\boldsymbol{\lambda}}}$ fulfil similar ODEs which additionally take the measurements $\exp\{\boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i})\}$ and $\exp\{\int dt \boldsymbol{\lambda}_t \cdot \boldsymbol{f}(\boldsymbol{x}_t)\}$ into account and which are given in Appendix B. Unfortunately, the ODEs in (26) are not closed (equations for the moments of order n depend on higher order moments), leading to an infinite hierarchy of ODEs [Gardiner, 1985]. Nonetheless, there are well established moment-closure methods to approximate the solutions of these ODEs. One popular

class of moment-closure approximations breaks this infinite hierarchy by expressing moments above a certain order as functions of lower order moments [Goodman, 1953, Whittle, 1957, McQuarrie et al., 1964, Lakatos et al., 2015, Schnoerr et al., 2015]. One is thus left with a finite system of coupled ODEs for which efficient numerical integration schemes exist. In this article we use the *normal* or *cumulant-neglect* moment closure which sets all cumulants above a certain order to zero. Setting all cumulants above order two to zero corresponds to choosing a multivariate Gaussian distribution as approximation [Gomez-Uribe and Verghese, 2007, Goutsias, 2007, Schnoerr et al., 2014b]. We then combine the resulting equations with iterative Bayesian updates corresponding to $\exp\{\boldsymbol{\lambda}_i \cdot \boldsymbol{f}(\boldsymbol{x}_{t_i})\}$ and $\exp\{\int dt \boldsymbol{\lambda}_t \cdot \boldsymbol{f}(\boldsymbol{x}_t)\}$ to approximate the moments of $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t | \, \boldsymbol{\lambda}_{i:t_i \leq t}, \boldsymbol{\lambda}_{s \leq t})$. We denote these moments as $\hat{\boldsymbol{\mu}}_t^{\mathrm{fw}}$ and the corresponding canonical parameters as $\hat{\boldsymbol{\eta}}_t^{\mathrm{fw}}$. More details are given in Appendix A.

3.2.2 Smoothed moment approximations

As detailed above, in order to approximate $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t)$ we can either approximate $q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_{i:t_i>t}, \boldsymbol{\lambda}_{s>t} \,|\, \boldsymbol{x}_t)$ or approximate $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t)$ directly. To compute $q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_{i:t_i>t}, \boldsymbol{\lambda}_{s>t} \,|\, \boldsymbol{x}_t)$ we would need to solve a non-linear backward Fokker-Planck equation. Since $q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_{i:t_i>t}, \boldsymbol{\lambda}_{s>t} \,|\, \boldsymbol{x}_t)$ is not a distribution, we cannot approximate it using moment closure. However, we can use moment closure on the moment smoothing equations proposed in Striebel [1965] and Leondes et al. [1970] to directly approximate the moments of $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t)$ instead. These equations compute the marginal moments $\langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\lambda}}}$ by using the (exact) $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t \,|\, \boldsymbol{\lambda}_{i:t_i \leq t}, \boldsymbol{\lambda}_{s \leq t})$. They read as

$$\begin{aligned}
\partial_t \langle f_l(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\lambda}}} = & \sum_j \left\langle a_j(\boldsymbol{x}_t, t) \partial_{x_j} f_l(\boldsymbol{x}_t) \right\rangle_{q_{\boldsymbol{\lambda}}} \\
& - \sum_{j,k} \left\langle \partial_{x_j} f_l(\boldsymbol{x}_t) \partial_{x_k} b_{jk}(\boldsymbol{x}_t, t) \right\rangle_{q_{\boldsymbol{\lambda}}} - \frac{1}{2} \sum_{j,k} \left\langle b_{jk}(\boldsymbol{x}_t, t) \partial_{x_j} \partial_{x_k} f_l(\boldsymbol{x}_t) \right\rangle_{q_{\boldsymbol{\lambda}}} \\
& - \sum_{j,k} \left\langle b_{jk}(\boldsymbol{x}_t, t) \partial_{x_j} f_l(\boldsymbol{x}_t) \partial_{x_k} \log q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t | \, \boldsymbol{\lambda}_{i:t_i \leq t}, \boldsymbol{\lambda}_{s \leq t}) \right\rangle_{q_{\boldsymbol{\lambda}}} . \qquad (27)
\end{aligned}$$

Similar to (26), this corresponds to an infinite cascade of coupled ODEs. We solve these approximately by substituting $q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t \,|\, \boldsymbol{\lambda}_{i:t_i \leq t}, \boldsymbol{\lambda}_{s \leq t}) \approx q_{\hat{\boldsymbol{\eta}}_t^{\mathrm{fw}}}(\boldsymbol{x}_t)$—as obtained in the previous section—into (27) and applying a corresponding moment closure. We denote the resulting moments by $\hat{\boldsymbol{\mu}}_t$, the canonical parameters by $\hat{\boldsymbol{\eta}}_t$, and the approximation by $q_{\hat{\boldsymbol{\eta}}_t}(\boldsymbol{x}_t) \approx q_{\boldsymbol{\lambda}}(\boldsymbol{x}_t)$.

Overall, we have introduced two levels of approximations: (i) an approximation of (7) using independence assumptions (Section 3.1); (ii) moment closure to approximate the moments required by the optimisation problem resulting from (i) (Section 3.2). The first level of approximations reduces the inference problem to moment computations, while the second level performs these moment computations approximately by conveniently combining moment closures and iterative Bayesian updates within an exponential family of distributions.

To derive an algorithm, one first has to decide what family of exponential distributions (choice of $\boldsymbol{f}$) is best for the model and data at hand. Given $\boldsymbol{f}$, the form of the moment-closure equations for (26) and (27) can be derived. It is important to point out that one can do moment closure for a wider set of moments than the ones given by $\boldsymbol{f}$, be it for computational or accuracy reasons. For example, one can opt for a linear and quadratic $\boldsymbol{f}$ but compute moments up to 4^{th} order for better accuracy in approximating $\langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\lambda}}}$. In Appendix A we provide a detailed description of the algorithm we used to implement the (approximate) fixed point iteration in (21) and (22).

The computational complexity of the algorithm scales linearly w.r.t. time (solving ODEs) while the scaling w.r.t. the dimensionality of $\boldsymbol{x}_t$ can vary according to $\boldsymbol{f}$. For the linear and quadratic $\boldsymbol{f}$ (Gaussian approximations) we consider in this paper the computational complexity of solving the ODEs resulting from the approach presented in Section 3.2 scales cubically w.r.t. the dimensionality of $\boldsymbol{x}_t$ (matrix multiplications). This computational complexity is similar to the complexity of the method presented in Särkkä and Sarmavuori [2013].

3.3 Extension to Markov jump processes

We next aim at extending the applicability of the EP method developed in the previous Section to Markov jump processes (MJPs). MJPs are used in many scientific disciplines ranging from queueing theory to epidemiology and systems biology. They constitute a convenient framework to model stochastic dynamics in discrete valued processes. Typically these are systems in which a set of species interact via various stochastic rules. The latter are implemented as transitions or "jumps" between the discrete states of the system. The state of a N-dimensional MJP is given by the vector $\boldsymbol{n}_t = (n_t^1, \ldots, n_t^N)$ where each n_t^i is typically a non-negative integer number. In this paper we consider MJPs that have a finite set of possible transitions $\boldsymbol{n} \to \boldsymbol{n} + \boldsymbol{S}_r, r = 1, \ldots, R$, and whose rates $g_r(\boldsymbol{n_t})$ depend only on the current state $\boldsymbol{n}_t$ of the system. Here, $\boldsymbol{S}_r$ is the rth column vector of the stoichiometric matrix $\boldsymbol{S}$ characterising the transitions. The single-time marginal distribution $p(\boldsymbol{n}_t|\boldsymbol{n}_0)$ of the process with initial state $\boldsymbol{n}_0$ is known to fulfil the master equation [Gillespie, 1992]

$$\partial_t p(\boldsymbol{n}_t|\boldsymbol{n}_0) = \sum_{r=1}^{R} g_r(\boldsymbol{n}_t - \boldsymbol{S}_r) p(\boldsymbol{n}_t - \boldsymbol{S}_r|\boldsymbol{n}_0) - \sum_{r=1}^{R} g_r(\boldsymbol{n}_t) p(\boldsymbol{n}_t|\boldsymbol{n}_0). \tag{28}$$

The state component n_t^i could for instance denote the molecule number of the ith species in a chemical reaction system. In this case the transitions correspond to chemical reactions between species and (28) is called the *chemical master equation* [Gillespie, 1992, Gillespie et al., 2013]. Other types of systems that can be described by a master equation of the type in (28) are for instance prey-predator [Reichenbach et al., 2006] or epidemic systems [Rozhnova and Nunes, 2009]. For all but the most simple systems, analytic solutions to the master equation in (28) are not available, and one has to rely on either stochastic simulations or analytic approximations.

Diffusion approximation

We discuss next a popular method that approximates an MJP by a non-linear diffusion process. In the chemical reaction context the equation defining the diffusion process is often called the *chemical Langevin equation* [Gillespie, 2000, Schnoerr et al., 2014a]. The approximating non-linear diffusion process is of the same form as the diffusion processes considered in the previous sections and defined in (1). The inference method proposed in this paper can hence be readily applied to MJPs by combining it with the diffusion approximation presented here.

The diffusion equation approximating an MJP described by (28) is a diffusion process with drift and diffusion given by [Gillespie, 2000]

$$\boldsymbol{a}(\boldsymbol{x}_t, t, \boldsymbol{\theta}) = \boldsymbol{S} \cdot \boldsymbol{g}(\boldsymbol{x}_t), \tag{29}$$

$$\boldsymbol{b}(\boldsymbol{x}_t, t, \boldsymbol{\theta}) = \boldsymbol{S} \cdot \mathrm{diag}(\boldsymbol{g}(\boldsymbol{x}_t)) \cdot \boldsymbol{S}^T. \tag{30}$$

Here $\boldsymbol{x}_t$ is the continuous real-valued pendant to $\boldsymbol{n}_t$, $\boldsymbol{g}(\boldsymbol{x}_t) = (g_1(\boldsymbol{x}_t), \ldots, g_R(\boldsymbol{x}_t))$, where $g_i(\boldsymbol{x}_t)$ is the rate function of the ith reaction, and $\mathrm{diag}(\boldsymbol{g}(\boldsymbol{x}_t))$ is the diagonal matrix with $\boldsymbol{g}(\boldsymbol{x}_t)$ on the diagonal. In Section 5 we apply the proposed EP method to a Lotka-Volterra system modeled as a MJP by combining it with the diffusion approximation presented here.

4 Discussion and related work

In Cseke et al. [2013] we propose an expectation propagation method for diffusion process models where the prior process is Gaussian–Markov. In this paper, we extend this method to models with non-nonlinear prior processes. Here we use expectation propagation only to approximate the likelihood terms. We avoid approximating the prior process by using moment-closure approximations on the process resulting from the prior and the likelihood approximations. When we choose a Gaussian–Markov prior process, the method proposed in this paper is identical to the one proposed in Cseke et al. [2013].

In Archambeau et al. [2007] the authors present a variational approach to approximate non-linear processes with time-only dependent diffusion terms by Orstein-Uhlenbeck/Gaussian-Markov processes. To our knowledge the extension of the approach in [Archambeau et al., 2007] to prior processes with state-dependent diffusion terms is not straightforward since a Gaussian-Markov approximation to the posterior process would lead to an ill-defined variational objective. The approach presented in this paper provides a convenient way to avoid this problem. We only obtain approximations of the posterior marginals instead of a process approximation [Archambeau et al., 2007], however, we can address inference problems where the diffusion terms are state dependent. In recent work, Sutter et al. [2015] proposed an alternative variational approach based on an approximating process with fixed marginal laws. This extends the Gaussian approximation of Archambeau et al. [2007] to cater for cases where Gaussian marginals are not appropriate, e.g. in stochastic reaction networks where concentrations are constrained positive. The constraint on the marginals however considerably limits the flexibility of their algorithm, and requires a considerable amount of user input; furthermore, it is unclear how accurate the approximation is in general.

There have been many application of EP in various discrete time models, early works include Heskes and Zoeter [2002], Ypma and Heskes [2005] and [Barber, 2006]. In these papers the joint distribution of the variables (Markov chain) is approximated by using a factorising approximation in EP. Note that this is not identical to a mean-field type variational approximation (Minka [2001]). As mentioned in Section 3.1 the EP formulation in Heskes and Zoeter [2002], Ypma and Heskes [2005] and [Barber, 2006] is not straightforward to extend to continuous time and the derivation we present here is a possible way to go around the problem. In Nodelman et al. [2005] the authors develop an EP algorithm for continuous time Bayesian networks. Their algorithm can be viewed as a generalisation of belief propagation [Yedidia et al., 2000] where each variable in belief propagation corresponds to the path of a single variable in the Bayesian network. The problems they address, like computing the marginal distribution of the whole path of a group of variables (marginalisation over paths), are not directly related to the ones we address in this paper. Their work is similar in sprit to Opper and Sanguinetti [2008] and Vrettas et al. [2015] (see below).

Inference in Markov jump processes from discrete time observations is a well studied problem, with several available algorithms employing sampling, variational or system approximations [Opper and Sanguinetti, 2008, Ruttor and Opper, 2009, Golightly et al., 2014, Zechner et al., 2014, Georgoulas et al., 2016, e.g]. The extension of our proposed method to MJPs (Section 3.3) can be viewed as an alternative way to do inference for such models, with the additional capability of performing inference from continuous time observations.

Särkkä [2010] and Särkkä and Sarmavuori [2013] propose a continuous time extension of the popular unscented transformation in [Julier et al., 2000] to obtain Gaussian state space approximations in SDE models with time-only dependent diffusion terms and both non-linear/non-Gaussian discrete and continuous time observation. In [Ala-Luhtala et al., 2014] the authors compare these approaches to the variational method in [Archambeau et al., 2007] which they then use to improve on their smoothing estimates.

In a recent work Vrettas et al. [2015] present a mean-field variational approximation where they approximate the posterior process with a set of independent univariate Gaussian processes (factorised approximation). The considered model has polynomial drift terms and state-independent diffusion terms and the observations are at discrete time-points. Due to a clever parameterisation (piecewise polynomials) of the mean and the variance function of the variational approximation the dimensionality of the state can scale to thousands.

5 Examples

As an example, we consider a classical benchmark problem, the Lotka-Volterra system. This system is a popular model describing the nonlinear interactions between a prey and a predator population. The prey X and predator Y interact via the reactions

$$\varnothing \xrightarrow{k_0} X, \quad X \xrightarrow{k_1} X + X, \quad Y + X \xrightarrow{k_2} Y + Y, \quad Y \xrightarrow{k_3} \varnothing, \tag{31}$$

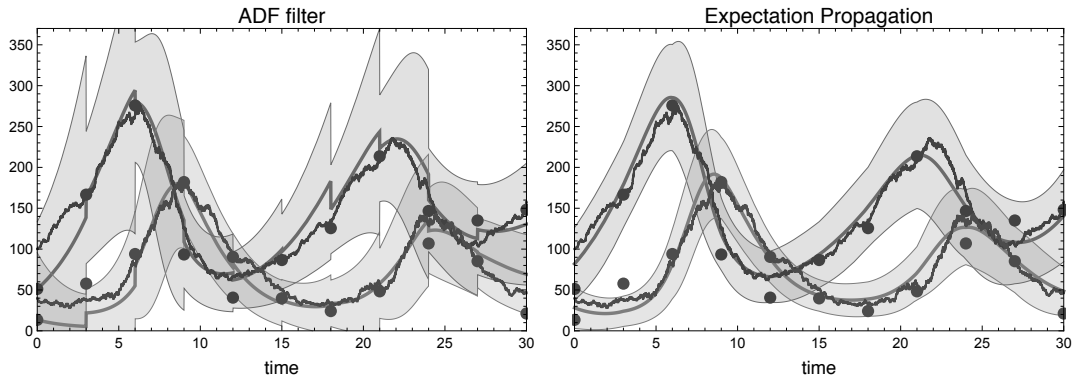

Figure 1: ADF filtering and EP for discrete measurements of the Lotka-Volterra system with reactions in (31), with log-normal measurement noise with variance 750. The non-smooth curves show the sampled path, dots correspond to sampled observation data while the smooth curves and shaded areas show the approximate posterior mean and three standard deviations.

where the first reaction corresponds to a birth process of X, the second to the reproduction of X, the third to reproduction of Y by consumption of one X, and the fourth to a death process of Y. The corresponding rate function $\boldsymbol{g}(n_1, n_2)$ and stoichiometric $\boldsymbol{S}$ matrix can be written as

$$\boldsymbol{g}(n_1, n_2) = (k_0,\ k_1 n_1,\ k_2 n_1 n_2,\ k_3 n_2)^T, \quad \boldsymbol{S} = \begin{pmatrix} 1 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \tag{32}$$

where n_1 and n_2 are the number counts of species X and Y, respectively. Depending on the parameters, single realisations of the process show oscillatory behaviour. We choose a fixed parameter set for which this is the case, namely $(k_0, k_1, k_2, k_3) = (5, 0.3, 0.004, 0.6)$.

To our knowledge, no analytic solutions are known for the master equation (28) of this system. To perform approximate inference, we first approximate the master equation by its diffusion approximation defined in (29) and (30). This allows us to apply the fixed point iteration procedure described in Section 3.1 to perform approximate inference for non-Gaussian likelihoods and continuous time constraints. The data is generated by simulating the MJP by means of the stochastic simulation algorithm [Gillespie, 1977]. In all scenarios presented in this section, the EP algorithm converged to an accuracy of 0.01 (corresponding to $\tau = 0.01$ in the Appendix A) after a few iterations, typically 10-20. We have chosen $\tau = 0.01$ because further iterations resulted in no significant changes in the relevant performance measures (see below).

We first consider discrete time measurements of the system assuming log-normal measurement noise with a variance of 750. The panels of Figure 1 show the sampled data and the inferred approximate state space marginals. The left panel shows the results of the Assumed Density Filtering (ADF) method [Maybeck, 1982, Minka, 2001] shown in Algorithm 2, of Section Appendix A, while the right panel shows the results obtained by the expectation propagation (EP) method developed in this paper. The details of the EP algorithm are presented in Algorithm 1 in Appendix A.

Next we consider again discrete log-normal measurements with variance 750 and additionally impose a continuous time constraint with loss function

$$U(\boldsymbol{x}_t), t, \boldsymbol{\theta}) = U_1(x_t^1, t, \boldsymbol{\theta}) + U_2(x_t^2, t, \boldsymbol{\theta}), \quad U_i(x, t, \boldsymbol{\theta}) = a_i(x - b_i)^4. \tag{33}$$

Figure 2 shows the results obtained by the EP algorithm, without (left panel) and with (right panel) the continuous time constraint taken into account. The constraint was chosen to limit the process close to its originally sampled path in the regions highlighted on the panel (grey area). We observe that the constraints significantly reduce the variance of the approximate posterior state space marginals in the corresponding regions.

In the following we compare EP to ADF combined with the smoothing method proposed in Section 3.2.2. We call the latter *ADF-S*. Note that ADF-S corresponds to EP with only

noise variance	RMSE observations		RMSE path	
	ADF	EP	ADF	EP
250	10.2	10.3	11.6	11.6
500	12.7	12.5	13.5	13.3
750	15.5	15.0	16.1	15.9
1000	16.1	15.9	16.8	16.5
1500	18.4	18.4	19.3	19.2

Table 1: Comparing RMSE of ADF-S and EP. The table shows the average root mean square error (RMSE) resulting from the mean of the approximate state-space marginals. The RMSE were obtained by averaging over 40 process/data samples for each noise variance value. We find that the EP method gives slightly more accurate results than ADF. See main text in Section 5 for further details.

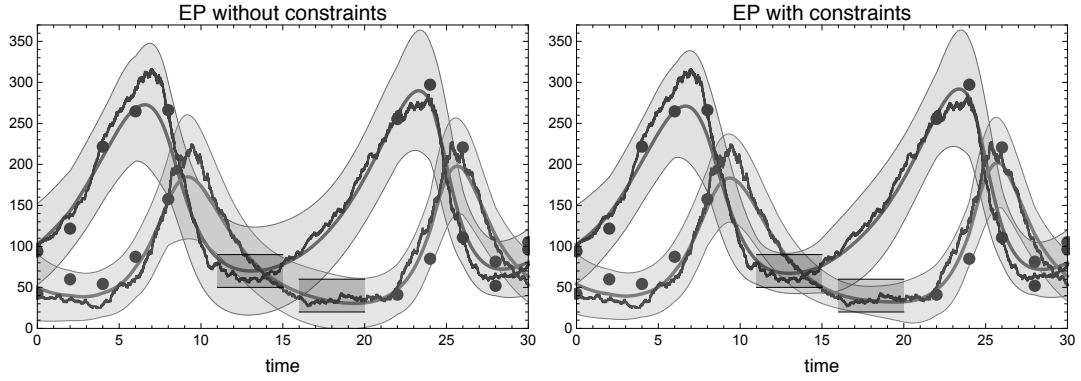

Figure 2: EP for discrete log-normal measurements with variance 750 and continuous time constraints for the Lotka-Volterra system defined in (31). The panels show the EP result with (right panel) and without (left panel) continuous time constraints taken into account. The shaded boxes correspond to continuous time constraints. (See Figure 1 for further details on figure elements.)

one iteration step. To our knowledge the proposed ADF-S method has not been reported in the literature before. For several values of the observation noise variance we sampled 40 process paths/trajectories and observation values. We then measured and averaged the RMSE between the true path and the inferred results using ADF combined with the corresponding smoothing (ADF-S)—this corresponds to the first step of EP—and the EP algorithm. The latter was iterated until convergence (a few iterations). We computed the RMSE at the data locations (RMSE observations) as well as over the whole sampled path (RMSE path). In this way we assessed both the training and the predictive performance of the approximation. The results are shown in Table 1. The results show that EP slightly improves on ADF-S on average for most parameter settings. A frequentist comparison is not feasible due to large variations of RMSE w.r.t. sampled paths, but for this model the two methods behave similarly.

6 Conclusions

In this paper, we have derived a novel approximate solution to the Bayesian inference problem for continuous time stochastic processes of diffusion type. This approach can be generalised via a Langevin approximation to Markov jump processes, providing therefore a practical solution to Bayesian inference in a wide class of models. A distinctive feature of our approach is that it can handle both discrete-time observations and trajectory constraints encoded as a continuous-time loss function. The resulting approach is therefore highly flexible. Numerical experiments on a classical benchmark, the Lotka-Volterra system, show both high accuracy and good computational

performance.

We defined an EP free energy that approximates the free energy corresponding to our inference problem defined in Section 2 and presented a self-consistent algorithm to optimise it. This formulation is closely related to other variational approaches for inference in continuous-time processes [Archambeau et al., 2007, Cseke et al., 2013], however our approach is distinct from others in that we do not seek to reconstruct an approximating process, but focus on computing accurate approximations to the marginal moments of the posterior distribution. A major advantage of the moment-based approach is that it still leads to a well-defined algorithm even in the case of state-dependent diffusion processes, when a Gaussian variational approach cannot be deployed.

References

J. Ala-Luhtala, S. Särkkä, and R. Piché. Gaussian filtering and variational approximations for Bayesian smoothing in continuous-discrete stochastic dynamic systems. *ArXiv e-prints*, 2014.

C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor. Gaussian process approximations of stochastic differential equations. *Journal of Machine Learning Research - Proceedings Track*, 1:1–16, 2007.

D. Barber. Expectation correction for smoothed inference in switching linear dynamical systems. *Journal of Machine Learning Research*, 7:2515–2540, December 2006. ISSN 1532-4435.

A. Beskos, O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):333–382, 2006.

L. Bortolussi and G. Sanguinetti. Learning and designing stochastic processes from logical constraints. In *10th International Conference on Quantitative Evaluation of SysTems, QEST 2013*, volume 8054, pages 89–105. Springer Verlag, 2013.

L. Bortolussi, D. Milios, and G. Sanguinetti. U-check: Model checking and parameter synthesis under uncertainty. In *Proceedings of QEST Conference*, 2015.

L. Csató and M. Opper. Sparse representation for Gaussian process models. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, Cambridge, MA, USA, 2001. MIT Press.

B. Cseke, M. Opper, and G. Sanguinetti. Approximate inference in latent Gaussian-Markov models from continuous time observations. In *Advances in Neural Information Processing Systems 26*, pages 971–979. 2013.

C. W. Gardiner. *Handbook of stochastic methods*, volume 4. Springer Berlin, 1985.

A. Georgoulas, Jane H., and G. Sanguinetti. Unbiased bayesian inference for population Markov jump processes via random truncations. *Statistics and Computing*, pages 1–12, 2016.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1):404–425, 1992.

D. T. Gillespie. The chemical Langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

D. T. Gillespie, A. Hellander, and L. R. Petzold. Perspective: Stochastic algorithms for chemical kinetics. *The Journal of Chemical Physics*, 138(17):170901, 2013.

A. Golightly, D. A. Henderson, and C. Sherlock. Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statistics and Computing*, pages 1–17, 2014.

C. A Gomez-Uribe and G. C. Verghese. Mass fluctuation kinetics: Capturing stochastic effects in systems of chemical reactions through coupled mean-variance computations. *The Journal of Chemical Physics*, 126(2):024109, 2007.

L. A Goodman. Population growth of the sexes. *Biometrics*, 9(2):212–225, 1953.

J. Goutsias. Classical versus stochastic kinetics modeling of biochemical reaction systems. *Biophysical Journal*, 92(7):2350–2365, 2007.

Y. Harel, R. Meir, and M. Opper. A tractable approximation to optimal point process filtering: Application to neural encoding. In *Advances in Neural Information Processing Systems 2015*, pages 1603–1611, 2015.

T. Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *Advances in Neural Information Processing Systems 15*, pages 359–366, Cambridge, MA, 2003. The MIT Press.

T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 216–223. Morgan Kaufmann Publishers Inc., 2002.

T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter. Approximate inference techniques with expectation constraints. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11015, 2005.

J. Hillston. *A compositional approach to performance modelling*, volume 12. Cambridge University Press, 2005.

S. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Trans. Automat. Contr.*, 45(3):477–482, 2000.

D. A. Knowles and T. Minka. Non-conjugate Variational Message Passing for Multinomial and Binary Regression. In *Advances in Neural Information Processing Systems*, pages 1701–1709, 2011.

E. Lakatos, A. Ale, P. D. W. Kirk, and M. P. H. Stumpf. Multivariate moment closure techniques for stochastic kinetic models. *The Journal of Chemical Physics*, 143(9):094107, 2015.

S. L. Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108, 1992.

C. Leondes, J. Peller, and E. Stear. Nonlinear Smoothing Theory. *IEEE Transactions on Systems Science and Cybernetics*, 6(1):63–71, 1970.

P. S Maybeck. *Stochastic models, estimation, and control*, volume 3. Academic press, 1982.

D. A. McQuarrie, C. J. Jachimowski, and M. E. Russell. Kinetics of small systems. II. *The Journal of Chemical Physics*, 40(10):2914–2921, 1964.

T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001.

U. Nodelman, D. Koller, and C.R. Shelton. Expectation propagation for continuous time bayesian networks. In *Proceedings of the Twenty-first Conference on Uncertainty in Artificial Intelligence*, pages 431–440, 2005.

M. Opper and G. Sanguinetti. Variational inference for Markov jump processes. In *Advances in Neural Information Processing Systems*, pages 1105–1112, 2008.

M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

M. Opper and O. Winther. Expectation consistent approximate inference. *Journal of Machine Learing Research*, 6:2177–2204, 2005.

T. Reichenbach, M. Mobilia, and E. Frey. Coexistence versus extinction in the stochastic cyclic lotka-volterra model. *Physical Review E*, 74:051907, 2006.

G. Rozhnova and A. Nunes. Fluctuations and oscillations in a simple epidemic model. *Physical Review E*, 79:041922, 2009.

A. Ruttor and M. Opper. Efficient statistical inference for stochastic reaction processes. *Physical Review Letters*, 103(23):230601, 2009.

S. Särkkä. Continuous-time and continuous–discrete-time unscented Rauch–Tung–Striebel smoothers. *Signal Processing*, 90(1):225–235, 2010.

S. Särkkä and J. Sarmavuori. Gaussian filtering and smoothing for continuous-discrete dynamic systems. *Signal Processing*, 93(2):500–510, 2013.

D. Schnoerr, G. Sanguinetti, and R. Grima. The complex chemical Langevin equation. *The Journal of Chemical Physics*, 141(2):024103, 2014a.

D. Schnoerr, G. Sanguinetti, and R. Grima. Validity conditions for moment closure approximations in stochastic chemical kinetics. *The Journal of Chemical Physics*, 141(8):084103, 2014b.

D. Schnoerr, G. Sanguinetti, and R. Grima. Comparison of different moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics*, 143(18):185101, 2015.

C. T. Striebel. Partial differential equations for the conditional distribution of a Markov process given noisy observations. *Journal of Mathematical Analysis and Applications*, 11:151–159, 1965.

T. Sutter, A. Ganguly, and H. Koeppl. A variational approach to path estimation and parameter inference of hidden diffusion processes. *arXiv:1508.00506*, 2015.

I. Volkov, J. R. Banavar, S. P. Hubbell, and A. Maritan. Patterns of relative species abundance in rainforests and coral reefs. *Nature*, 450(7166):45–49, 2007.

M. D. Vrettas, D. Cornford, and M. Opper. Variational mean-field algorithm for efficient inference in large systems of stochastic differential equations. *Physical Review E*, 91:012148, 2015.

P. Whittle. On the use of the normal approximation in the treatment of stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 268–281, 1957.

D. J. Wilkinson. *Stochastic modelling for systems biology*. CRC press, 2011.

Jo. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems 13*, volume 13, pages 689–695, 2000.

A. Ypma and T. Heskes. Novel approximations for inference in nonlinear dynamical systems using expectation propagation. *Neurocomputing*, 69(1-3):85–99, 2005.

C. Zechner, S. Unger, M.and Pelet, M. Peter, and H. Koeppl. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature methods*, 11(2):197–202, 2014.

A Algorithms and practical considerations

In Section 5 we compared our EP method with ADF, and we here give a detailed description of both algorithms. We present two algorithms: (i) the EP algorithm corresponding to the fixed point iteration in Section 3.1 and (ii) an Assumed Density Filtering (ADF) algorithm [Maybeck, 1982, Lauritzen, 1992, Csató and Opper, 2001, Minka, 2001] and the ADF-S algorithm that performs an extra smoothing step after ADF. ADF-S a can be viewed as one single step of the EP.

Before presenting the algorithms, we provide details of how moment closure and iterative Bayesian updates are combined to approximate the moments $\langle \boldsymbol{f}(\boldsymbol{x}_t)\rangle_{q_{\boldsymbol{X}}}$. Let the moment closure for the filtering (26) and the smoothing equation (27) result in

$$d\hat{\boldsymbol{\mu}}_t^{\mathrm{fw}} = \mathcal{M}_{\mathrm{fw}}(\hat{\boldsymbol{\mu}}_t^{\mathrm{fw}})dt \quad \text{and} \quad d\hat{\boldsymbol{\mu}}_t = \mathcal{M}_{\mathrm{sm}}(\hat{\boldsymbol{\mu}}_t; \hat{\boldsymbol{\mu}}_t^{\mathrm{fw}})dt. \tag{34}$$

Algorithm 1 Expectation Propagation (EP)

function ExpectationPropagationForDiffusionProcesses(ϵ, τ, K_{max})
  Initialise $\boldsymbol{\lambda}_t = 0$ and $\boldsymbol{\lambda}_i = 0$
  for $k = 1, \ldots, K_{max}$ **do**
    Compute $\hat{\boldsymbol{\mu}}_t$, $t \in [0, 1]$ and $\hat{\boldsymbol{\eta}}_{t_i}$ as described in Sections 3.2.1 and 3.2.2
    Compute $\boldsymbol{\eta}_i = \hat{\boldsymbol{\eta}}_{t_i} - \boldsymbol{\lambda}_i$
    Compute $\boldsymbol{\lambda}_i^{new} = \text{Project}[\tilde{q}_{\boldsymbol{\eta}_i} \boldsymbol{f}] - \boldsymbol{\eta}_i$ for all i
    Compute $\boldsymbol{\lambda}_t^{new} = -\partial_{\hat{\boldsymbol{\mu}}_t} \langle U(\boldsymbol{x}_t, t) \rangle_{q_{\hat{\boldsymbol{\eta}}_t}}$ for all $t \in [0, 1]$
    Update $\boldsymbol{\lambda}_t^{new} = (1 - \epsilon)\boldsymbol{\lambda}_t + \epsilon \boldsymbol{\lambda}_t^{new}$, $\boldsymbol{\lambda}_i^{new} = (1 - \epsilon)\boldsymbol{\lambda}_i + \epsilon \boldsymbol{\lambda}_t^{new}$
    if $\max(\max_t |\boldsymbol{\lambda}_t^{new} - \boldsymbol{\lambda}_t|, \max_i |\boldsymbol{\lambda}_i^{new} - \boldsymbol{\lambda}_i|) < \tau$ **then**
      break
    end if
  end for
  Approximate $\log Z$ as described in Section 3.1 and 3.2
end function

We add the contribution of the Bayesian updates in the forward computation by

$$d\hat{\boldsymbol{\mu}}_t^{\text{fw}} = \mathcal{M}_{\text{fw}}(\hat{\boldsymbol{\mu}}_t^{\text{fw}})dt + \partial^2 \log Z_f(\text{Project}[\hat{\boldsymbol{\mu}}_t^{\text{fw}}; \boldsymbol{f}])\boldsymbol{\lambda}_t dt, \tag{35}$$

$$d\hat{\boldsymbol{\mu}}_{t_i+}^{\text{fw}} = \partial \log Z_f(\text{Project}[\hat{\boldsymbol{\mu}}_{t_i}^{\text{fw}}; \boldsymbol{f}] + \boldsymbol{\lambda}_i), \tag{36}$$

where we use $\text{Project}[\hat{\boldsymbol{\mu}}_t^{\text{fw}}; \boldsymbol{f}]$ to denote the moment-to-canonical parameter transformation. We use $\partial \log Z_f$ to denote the canonical-to-moment transformation, that is, we have the identity relation $\hat{\boldsymbol{\mu}}_t^{\text{fw}} = \partial \log Z_f(\text{Project}[\hat{\boldsymbol{\mu}}_t^{\text{fw}}; \boldsymbol{f}])$. The form of the second term in the r.h.s. of (35) follows from the continuous time Bayesian updates by applying a Taylor expansion to $\partial \log Z_f(\text{Project}[\hat{\boldsymbol{\mu}}_t^{\text{fw}}; \boldsymbol{f}] + dt\boldsymbol{\lambda}_t)$. Smoothing is performed by solving the second equation in (34) backward in time. The measurements do not have to be incorporated explicitly here, they implicitly enter via the solution through $\hat{\boldsymbol{\mu}}_t^{\text{fw}}$.

Now that the approximation of $\langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\lambda}}}$ is formally fixed, we turn our attention to formulating the algorithm corresponding to the fixed point iteration defined in Section 3.1. Algorithm 1 shows an implementation of this iteration. We initialise $\boldsymbol{\lambda}_i$ by choosing a good approximation to $p(\boldsymbol{y}_i|\boldsymbol{x}_{t_i})$ and $U(\boldsymbol{x}_t, t)$. When this is not possible we simply set $\boldsymbol{\lambda}_i = 0$ and $\boldsymbol{\lambda}_t = 0$. In each step of the algorithm we proceed as follows: (i) approximate the moments $\langle \boldsymbol{f}(\boldsymbol{x}_t) \rangle_{q_{\boldsymbol{\lambda}}}$ by solving (35)-(36) and the smoothing equation in (34), (ii) compute $\boldsymbol{\eta}_i$ and the cavity distribution $\tilde{q}_{\boldsymbol{\eta}_i}$ and (iii) update $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_t$. In many cases $\langle U(\boldsymbol{x}_t, t) \rangle_{q_{\hat{\boldsymbol{\eta}}_t}}$ can be expressed as a function of of $\hat{\boldsymbol{\mu}}_t$. Therefore, we choose to compute the differentials w.r.t. $\hat{\boldsymbol{\mu}}_t$ instead of $\hat{\boldsymbol{\eta}}_t$. We perform damped updates of $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_t$ and terminate the iteration when the absolute value of the change in the updates falls below a specified threshold.

Algorithm 2 shows an implementation of the ADF algorithm. In ADF a single forward step is performed. Here the iterative Bayesian updates for the likelihood proxies are performed such that $\boldsymbol{\lambda}_t$ and $\boldsymbol{\lambda}_i$ represent the current estimates computed using $\hat{\boldsymbol{\mu}}_t^{\text{fw}}$. For the continuous time likelihoods this can be viewed as substituting the update of $\boldsymbol{\lambda}_t$ in (22) into (35)—line 4 of the algorithm. For the discrete time likelihoods the update can be viewed as choosing $\boldsymbol{\lambda}_i = 0$ when performing the first and only EP update.

Algorithm 2 Assumed Density Filtering (ADF/ADF-S)

function AssumedDensityFilteringForDiffusionProcesses
  Let $d\hat{\boldsymbol{\mu}}_t^{\text{fw}} = \mathcal{M}_{\text{fw}}(\hat{\boldsymbol{\mu}}_t^{\text{fw}})dt$ be the ODE resulting from the moment closure of (26)
  for $i = 1, \ldots, n$ **do**
    Solve $d\hat{\boldsymbol{\mu}}_t^{\text{fw}} = \mathcal{M}_{\text{fw}}(\hat{\boldsymbol{\mu}}_t^{\text{fw}})dt - \partial^2 \log Z_f(\text{Project}[\hat{\boldsymbol{\mu}}_t^{\text{fw}}; \boldsymbol{f}])\partial_{\hat{\boldsymbol{\mu}}_t^{\text{fw}}} \left\langle U(\boldsymbol{x}_t, t) \right\rangle_{q_{\boldsymbol{\eta}_t^{\text{fw}}}} dt$ on $(t_{i-1}, t_i]$
    Compute $\boldsymbol{\eta}_i = \text{Project}[\hat{\boldsymbol{\mu}}_{t_i}^{\text{fw}}; \boldsymbol{f}]$ and $\tilde{q}_{\boldsymbol{\eta}_i}(\boldsymbol{x}_{t_i})$
    Compute $\hat{\boldsymbol{\mu}}_{t_i+}^{\text{fw}} = \langle \boldsymbol{f}(\boldsymbol{x}_{t_i}) \rangle_{\tilde{q}_{\boldsymbol{\eta}_i}(\boldsymbol{x}_{t_i})}$
  end for
  (for ADF-S compute $\hat{\boldsymbol{\mu}}_t$ as described in Sections 3.2.1 and 3.2.2)
end function