Moment-Based Variational Inference for Stochastic Differential Equations

Christian Wildner
Technische Universität Darmstadt
christian.wildner@bcs.tu-darmstadt.de

Heinz Koeppl
Technische Universität Darmstadt
heinz.koeppl@bcs.tu-darmstadt.de

Abstract

Existing deterministic variational inference approaches for diffusion processes use simple proposals and target the marginal density of the posterior. We construct the variational process as a controlled version of the prior process and approximate the posterior by a set of moment functions. In combination with moment closure, the smoothing problem is reduced to a deterministic optimal control problem. Exploiting the path-wise Fisher information, we propose an optimization procedure that corresponds to a natural gradient descent in the variational parameters. Our approach allows for richer variational approximations that extend to state-dependent diffusion terms. The classical Gaussian process approximation is recovered as a special case.

1 INTRODUCTION

Itô processes governed by a stochastic differential equation (SDE) are an important class of time series models involving uncertainty. Originating from the statistical physics of diffusion, SDEs have become an important modeling tool in areas as diverse as biology, finance and engineering. However, applying SDEs as a predictive tool requires learning model parameters from real data. Usually, such data is corrupted by noise and only available at discrete sampling times. In such a scenario, likelihood-based parameter inference requires estimation of the posterior over the latent process. Computing this posterior requires the solution of a PDE that is only computationally tractable for very low-dimensional state spaces or for linear systems (see Särkkä and Solin (2019) for an accessible introduction). Thus, standard approximations linearize the system dynamics or use a discrete time approximation. In a Bayesian setting, Monte Carlo methods such as MCMC, SMC or particle MCMC methods are a common (Golightly and Wilkinson, 2011). In practice, sampling-based methods often struggle with high dimensional settings or with highly informative observations (Del Moral and Murray, 2014). In such a scenario, variational inference (Blei et al., 2017) may provide a more scalable alternative.

Related Work The variational formulation of Bayesian inference of latent stochastic processes and its connection to stochastic control have been observed first by Mitter and Newton (2003). Archambeau et al. (2007a) introduced variational inference for SDEs to the machine learning community. Their core idea is to compute the best linear Gaussian process approximation of the posterior. While this approach has been refined and extended several times over the years (e. g. Vrettas et al., 2011; Ruttor et al., 2013; Duncker et al., 2019), it is limited to state independent diffusion terms. An alternative approach presented by Sutter et al. (2016) constructs the variational process such that the marginal density belongs to a prespecified exponential family. While overcoming the Gaussian limitation, the construction is also mathematically involved. Cseke et al. (2016) suggested an approximation of the posterior in terms of moments rather than the marginal density within an expectation propagation framework for smoothing. Another moment-based approximation, albeit in the context of Markov jump processes, was proposed by Wildner and Koeppl (2019). However, the key idea of transition space partitioning for complexity reduction cannot be applied to SDEs. The main drawback of the deterministic approaches above is that they rely on model-specific derivations. Sampling-based variational inference does not require such computations and can also be applied to SDEs (Ryder et al., 2018). However, this comes at the price of much longer training times. More recently,

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

a promising neural SDE framework based on a stochastic adjoint method has been proposed (Li et al., 2020).

Contributions In this work, we propose a new sampling-free structured variational approach to latent diffusion processes that mitigates some drawbacks of earlier methods. Similarly to the approach of Cseke et al. (2016), we construct the proposal process as a controlled version of the prior process and reduce complexity by projecting the stochastic process onto a collection of summary statistics. To solve the variational problem, we adapt a strategy proposed by Wildner and Koeppl (2019). Using the Markov property in combination with moment closure, we map the full smoothing problem to a deterministic optimal control problem. Exploiting the path-wise Fisher information, we construct an effective natural gradient descent in the variational parameters. To keep model-specific derivations at a minimum, we implement our method in the PyTorch framework. Thus, we can circumvent a large part of the model-specific computations by exploiting Pytorch's automatic differentiation capabilities. Exploiting the structural similarity to the moment-based approach to Markov jump processes, we provide a unified framework capable of handling both SDEs and MJPs. The accompanying code is available at `https://git.rwth-aachen.de/bcs/projects/cw/public/mbvi_sde`.

2 PRELIMINARIES

This section summarizes material on SDEs, the inference problem for noisy observations in discrete time and the general variational formulation.

2.1 Stochastic Differential Equations

Let $\mathcal{X} \subset \mathbb{R}^n$. We consider a stochastic processes X on $\mathbb{R}^n$ over a finite time interval $[0, t_f]$ given by the Itô SDE

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t + b(X_t)\mathrm{d}W_t\,. \tag{1}$$

Here, W is an n-dimensional Wiener process and $a : \mathbb{R}^n \to \mathbb{R}^n$, $b : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ are functions of suitable regularity, i.e. satisfying a Lipschitz condition. Additionally, we will focus on cases where $b(x)$ has full rank for all $x \in \mathcal{X}$. The solution of an SDE of the form (1) is a Markov process and the corresponding marginal density satisfies the Fokker-Planck equation. In practice, one is often not interested in the full density but rather certain summary statistics $S : \mathbb{R}^n \to \mathbb{R}^l$. Often, S will correspond to first and second order monomials but other choices are possible as well. Now define the moment functions $\varphi(t) := \mathsf{E}[S(X_t)]$. The idea is now to propagate φ in time rather than the density. One can show that the moment functions φ_i satisfy a system of differential equations

$$\dot{\varphi}_i(t) = \mathsf{E}[A^\dagger S_i(X_t)] \tag{2}$$

where the backward generator $A^\dagger$ is the L_2-adjoint of the Fokker-Planck operator and given by

$$[A^\dagger f](x) = \sum_{i=1}^{n} a_i(x)\partial_i f(x) + \frac{1}{2}\sum_{i,j=1}^{N} D_{ij}(x)\partial_i\partial_j f(x) \tag{3}$$

for $f \in \mathcal{C}^2(\mathbb{R}^n)$ (Ethier and Kurtz, 2005). The diffusion tensor D is determined by the SDE (1) through the relation $D = bb^\top$. In general, the system (2) is not closed in φ, i.e. it will be of the form

$$\dot{\varphi}(t) = B\varphi(t) + B'\mathsf{E}[S'(X_t)]\,. \tag{4}$$

Here, B and B' are matrices of suitable dimension and S' corresponds to a collection of higher order moments. Thus, Eq. (4) still depends on the full process X. In order to obtain a closed form description, one can employ moment closure (Kuehn, 2016). A general closure is given by a function h that approximates the higher order moments S' such that (4) reduces to

$$\dot{\varphi}(t) = B\varphi(t) + B'h(\varphi(t))\,. \tag{5}$$

Two common methods to obtain closure schemes are via extensions and truncation of the summary statistics and by assuming an underlying distribution. In this work, we focus on the latter approach as it has been shown to correspond to a projection of the stochastic process onto a parametric family of distributions (Bronstein and Koeppl, 2018).

2.2 Posterior Path Estimation

We consider a scenario where the underlying process X is not observed directly. Instead, we have access to sparse and noisy observations $Y = (Y_1, \ldots, Y_n)^\top$ obtained at sample times $0 \le t_1 \le \ldots \le t_n \le t_f$. We assume that the observations are conditionally independent given the latent path of X and follow a noise distribution $Y_i \sim P_{\mathrm{obs}}(\cdot \mid X(t_i))$. The smoothing problem refers to evaluating expectations of the form $\mathsf{E}[f(X_t) \mid \sigma(Y)]$ where $\sigma(Y)$ denotes the history of the observation process Y up to the terminal time t_f. Under mild conditions, $\mathsf{E}[f(X_t) \mid \sigma(Y)]$ can be represented by a conditional probability density $\pi(x, t \mid y_1, \ldots, y_n)$. Now π can be understood as the marginal density of a posterior process $\bar{X}$. The posterior process $\bar{X}$ obeys an SDE with the same diffusion term as the prior process (1) and a modified drift

$$\bar{a}(x,t) = a(x) + D(x)\nabla \log(\beta(x,t)) \tag{6}$$

where the source term β satisfies a backward equation (Archambeau and Opper, 2011)

$$\beta(x,t) = -A^\dagger \beta(x,t)\,.$$

Intuitively, (6) corresponds to a controlled version of the prior process where the second term steers the process towards future observations. This analogy is the main motivation underlying our structured variational approximation introduced in Sec. 3.1.

2.3 Variational Smoothing

Let μ and ν be probability measures on a common probability space such that μ is absolutely continuous with respect to ν. Recall that the Kullback-Leibler divergence or relative entropy between μ and ν is defined as

$$D_{\mathrm{KL}}[\,\mu\,||\,\nu] = \int \log\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\mu\,.$$

Now consider two diffusions Z, X with drifts a^Z, a^X respectively and a shared diffusion tensor D that is invertible for almost every $x \in \mathcal{X}$. Then the Kullback-Leibler divergence on the level of sample paths is given by

$$\begin{aligned} D_{\mathrm{KL}}[P^Z \,||\, P^X] = \int_0^{t_f} \mathsf{E}\Big[\big(a^Z(Z_t) - a^X(Z_t)\big)^\top \\ \times\, D(Z_t)^{-1}\,\big(a^Z(Z_t) - a^X(Z_t)\big)\Big]\mathrm{d}t\,, \end{aligned} \tag{7}$$

where P^Z, P^X denote the measures over sample paths induced by the processes Z and X, respectively. A rigorous exposition on the relative entropy of diffusion processes is given in Mitter and Newton (2003). More intuitively, the path divergence (7) can be derived by considering the divergence of a corresponding discrete time system and taking the continuum limit (Archambeau et al., 2007a,b). For variational smoothing, we aim to find an approximate process Z within a class $\mathfrak{Z}$ of simpler processes. Following the usual variational inference framework (Blei et al., 2017), the best approximation Z^* within $\mathfrak{Z}$ is given by

$$Z^* = \arg\min_{Z\in\mathfrak{Z}} D_{\mathrm{KL}}[P^Z \,||\, P^{\bar{X}}]\,.$$

By inserting the true posterior drift (6), one can show that this objective function decomposes into

$$\begin{aligned} D_{\mathrm{KL}}[P^Z \,||\, P^{\bar{X}}] = D_{\mathrm{KL}}[P^Z \,||\, P^X] \\ - \sum_{k=1}^{n} \mathsf{E}[\log p(y_k \mid Z_{t_k})] + \log C \end{aligned} \tag{8}$$

where X is the prior process and $C = \mathsf{E}[p(y_1,\dots,y_n \mid X(t_1),\dots,X(t_n))]$ is the evidence.

3 VARIATIONAL SMOOTHING

3.1 Structured Variational Approximation

From the posterior drift (6), we observe that the true posterior process $\bar{X}$ is a controlled version of the prior process X. The idea is now to approximate the driving term in (6) by a feedback control. This leads to a drift of the form

$$a^Z(z,t) = a^X(z) + R(x)v(t)T(x)\,. \tag{9}$$

Here $v : [0,t_f] \to \mathbb{R}^{n\times m}$ is a deterministic, matrix-valued function corresponding to the variational parameters while $T : \mathbb{R}^n \times [0,t_f] \to \mathbb{R}^m$ represents a collection of control features and $R : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is a rescaling matrix. Typically, R will be set to the identity, the diffusion term b or the diffusion tensor D. Suitable choices of the rescaling factor can simplify the resulting equations and also reduce the computational complexity of the algorithm. A more detailed discussion is given in App. A2.2. In general, the control features T will be different from the summary statistics S. In the simple case where T is the identity map, (9) corresponds to a linear feedback control. For the following discussion, we also introduce $u : [0,t_f] \to \mathbb{R}^{nm}$ as a vectorized control obtained by stacking the columns of v.

Lemma 1. *Under the variational drift (9), the KL-term in the objective function (8) becomes a quadratic form in the vectorized controls u and can be represented as*

$$D_{\mathrm{KL}}[P^Z \,||\, P^X] = \frac{1}{2}\int_0^{t_f} u(t)^\top g(\varphi(t))u(t)\,\mathrm{d}t\,, \tag{10}$$

where the matrix valued function $g : \mathbb{R}^l \to R^{nm\times nm}$ is determined by the diffusion tensor D, the rescaling matrix R, the control features T and the summary statistics S.

Proof sketch. First, we show by direct calculation that under the variational drift (9) the KL term can be written as

$$D_{\mathrm{KL}}[P^Z \,||\, P^X] = \frac{1}{2}\int_0^{t_f} u(t)^\top \mathsf{E}[\psi(x(t))]u(t)\,\mathrm{d}t$$

with $\psi : \mathbb{R}^n \to \mathbb{R}^{nm\times nm}$ such that

$$\psi(x) = \begin{pmatrix} T_1T_1\tilde{D}^{-1} & \dots & T_1T_m\tilde{D}^{-1} \\ \vdots & \dots & \vdots \\ T_mT_1\tilde{D}^{-1} & \dots, & T_mT_m\tilde{D}^{-1} \end{pmatrix}(x)$$

with $\tilde{D}^{-1} = R^\top D^{-1} R$. Under a suitable choice of the summary statistics S, one can express the expectation as $\mathsf{E}[\psi(Z_t)] = g(\varphi(t))$. Such a g can always be found, e.g. by augmenting the summary statistics S accordingly. The details are given in App. A1.1. □

The full variational inference problem now corresponds to minimizing the objective function (8) with respect to u and φ subject to the moment equation (2). Since (2) still depends on the full stochastic process Z, we consider instead a relaxed variational inference problem by replacing the exact moment constraint with

$$\dot{\varphi}(t) = f(u(t), \varphi(t))\,. \tag{11}$$

where f is obtained from a closure scheme. The relaxation of the moment constraint simplifies the variational inference problem considerably as summarized in the following proposition.

Proposition 1. *The relaxed variational inference problem corresponds to a finite dimensional deterministic optimal control problem of the form*

$$\begin{aligned} \min_{u,\varphi} \quad & J[u,\varphi] \\ \text{s.t.} \quad & \dot{\varphi}(t) = f(u(t),\varphi(t)) \end{aligned} \tag{12}$$

with

$$J[u,\varphi] = \int_0^{t_f} L(u(t),\varphi(t))\,\mathrm{d}t - \sum_{k=1}^{n} F_k(\varphi(t_k)) \tag{13}$$

where

$$F_k(\varphi(t_k))) = \mathsf{E}[\log p(y_k \mid Z_{t_k})]$$

represents the contributions of the observations in (8) expressed in terms of φ. The cost function L is given by

$$L(u(t),\varphi(t)) = \frac{1}{2} u(t)^\top g(\varphi(t)) u(t)\,. \tag{14}$$

Proposition 1 is a consequence of Lemma 1 in combination with the moment closure relaxation. A detailed discussion is given in App. A1.2.

3.2 Gradient-Based Optimization

A standard approach to solve control problems of the form (12) is a gradient descent in the controls u (see e.g. Stengel, 1994). While such a gradient descent may work in principle, it often suffers from slow convergence. We can do better in our scenario by exploiting the probabilistic nature of the objective function. The key insight here is that the variational family induces a statistical manifold on the sample path space parametrized by the controls. This allows us in a first step to construct the path-wise Fisher information which we then use to derive a natural gradient descent (Amari, 1998) in the controls u.

Lemma 2. *Let Z and Z' be two members of the variational process family parametrized by u and u' respectively. We then have*

$$D_{\mathrm{KL}}[P^Z \,||\, P^{Z'}] = \frac{1}{2} G(u)[u-u', u-u'] \tag{15}$$

where $G(u)[\cdot,\cdot]$ for fixed u is a symmetric positive semidefinite bilinear form given by

$$\begin{aligned} G(u)[u'-u, u'-u] = \int_0^{t_f} & (u'(t)-u(t))^\top \\ & \times g(\varphi(t))(u'(t)-u(t))\,\mathrm{d}t\,. \end{aligned}$$

Lemma 2 can be proved very similarly to Lemma 1. For completeness, the proof is provided in App. A1.3. Now the Fisher information corresponds to the second order approximation of $D_{\mathrm{KL}}[P^Z \,||\, P^{Z'}]$ as u' approaches u. Since the divergence is already a quadratic form, it follows immediately from Lemma 2 that $G(u)[\cdot,\cdot]$ is the path-wise Fisher information at u. This allows us to construct natural gradient updates to solve the control problem (12). Both optimization algorithms featured in this work are summarized in the following proposition. An algorithmic representation is given in Alg. 1.

Proposition 2. *The regular (RGD) and natural (NGD) gradient descent updates of the control problem (12) with respect to the statistical manifold induced by u and step size h are given by*

$$\begin{aligned} u^{(i+1)}(t) &= u^{(i)}(t) \\ &\quad - h\left(g(\varphi^{(i)}(t))u^{(i)}(t) - {f_u^{(i)}}^\top(t)\cdot\eta^{(i)}(t)\right), \end{aligned} \tag{16}$$

$$\begin{aligned} u^{(i+1)}(t) &= u^{(i)}(t) \\ &\quad - h\left(u^{(i)}(t) - g(\varphi^{(i)}(t))^{-1}{f_u^{(i)}}^\top(t)\cdot\eta^{(i)}(t)\right), \end{aligned} \tag{17}$$

where $\varphi^{(i)}$ is the solution of the forward equation

$$\dot{\varphi}^{(i)}(t) = f(u^{(i)}(t),\varphi^{(i)}(t))$$

and $\eta^{(i)}$ is the solution of the adjoint equation

$$\dot{\eta}^{(i)}(t) = L_\varphi^{(i)}(t) - {f_\varphi^{(i)}(t)}^\top\cdot\eta^{(i)}(t)\,. \tag{18}$$

The notation $(\cdot)_\varphi^{(i)}$ and $(\cdot)_u^{(i)}$ denote the Jacobians with respect to $\varphi^{(i)}$ and $u^{(i)}$, respectively.

Proof sketch. Since the control u fully defines the moments φ, we can understand the control problem (12) as the minimization of a functional $J[u]$. Steepest descent with respect to a local metric G corresponds solving the constrained optimization problem

$$\begin{aligned} u^{(i+1)} &= \arg\min_u J[u] \\ \text{s.t.} \quad & \frac{1}{2} G(u^{(i)})[u-u^{(i)}, u-u^{(i)}] = \epsilon \end{aligned}$$

for small ϵ and then taking the limit $\epsilon \to 0$. For small ϵ, one can expand $J[u]$ around $u^{(i)}$. Keeping only the first order term leads to a quadratic problem that can be solved with variational calculus. For RGD, we use

that gradient descent corresponds to a steepest descent w.r.t. to the Euclidean metric. The result follows therefore by an identical computation with $G(u)$ replaced by the L_2 inner product. For the details, we refer to App. A1.4. □

If the dynamic equation (11) is obtained via moment closure, the summary statistics φ will not correspond to a globally valid stochastic process. Thus, the gradient has to be understood as an approximation as well. It is then advisable to check the results empirically by creating samples from the variational process with optimized control u^*.

3.3 A Note on Implementation

Solving the backward equation and computing the gradient updates requires the derivation of a number of model-specific functions. To reduce this overhead, we exploit the automatic differentiation capabilities of PyTorch which allows to effectively compute gradients and Jacobian-vector products. The most general version of our implementation only requires the specification of two functions: the r.h.s. of the forward equation f and either g or L. For certain subclasses, the implementation can be further simplified. In particular, we construct a general purpose method by fixing the control features and summary statistics as

$$\begin{aligned} T(x) &= (1, x_1, \dots, x_n)^\top , \\ S(x) &= (x_1, \dots, x_n, x_1^2, x_1 x_2, x_2^2, \dots, x_n^2)^\top . \end{aligned} \tag{19}$$

Intuitively, the choice of features T corresponds to a linear feedback control. The summary statistics S consist of first and second order moments and thus directly correspond to the mean and covariance of the approximate posterior, which is in line with many approximate non-linear filtering techniques Särkkä (2013). Here, it is also convenient to represent the control in terms of functions u_0, u_1 such that we can write

$$v(t)T(x) = u_0(t) + u_1(t)x\,. \tag{20}$$

With $m(t) \equiv \mathsf{E}[Z_t]$ and $M(t) \equiv \mathsf{E}[(Z_t - m(t))(Z_t - m(t))^\top]$, we obtain from (2)

$$\begin{aligned} \dot{m}(t) &= \mathsf{E}[a(Z_t)] + u_0(t) + u_1(t)m(t)] \\ \dot{M}(t) &= E[a(Z_t)Z_t^\top] + E[Z_t a(Z_t)^\top] + \mathsf{E}[D(Z_t)] \\ &\quad + u_1(t)M(t) + M(t)u_1(t)^\top \\ &\quad - \mathsf{E}[a(Z_t)]m(t)^\top - m(t)\mathsf{E}[a(Z_t)]^\top \end{aligned}$$

Under the choice (19), f and g can be constructed automatically by specifying $\mathsf{E}[a(Z_t)]$, $\mathsf{E}[a(Z_t)Z_t^\top]$ and $\mathsf{E}[D(Z_t)]$ in terms of the first and second order moments. This will typically require a moment closure. We include two standard closure schemes that lead to a reduction to moments of first and second order: a Gaussian closure for processes defined on the whole $\mathbb{R}^n$ and a log-normal closure for processes defined on $\mathbb{R}_+^n$ (see App. A2.1). We conclude this section by commenting on the relation to the standard Gaussian process approximation Archambeau et al. (2007a,b). As shown in App. A1.5, by a suitable choice of the control features the GP approximation arises as a special case within our framework.

Algorithm 1 Robust Natural Gradient Descent for Moment-Based Variational Smoothing

Input: Initial guess $u^{(0)}$, initial condition $\varphi(0)$, learning rate h, step size modifiers α, β.
for i = 0, ..., maxiter **do**
  Given $u^{(i)}$, $\varphi(0)$, compute $\varphi^{(i)}$ using (11).
  Given $u^{(i)}$, $\varphi^{(i)}$, compute $\eta^{(i)}$ using (18).
  Set u' according to (17).
  if $J[u'] < J[u^{(i)}]$ **then**
    $h \to \alpha \cdot h$, $u^{(i+1)} \to u'$
  else
    $h \to \beta \cdot h$, $u^{(i+1)} \to u^{(i+1)}$
  end if
end for
Output: Variational control u^* .

3.4 Online Variational Smoothing

The optimization based on Alg. 1 processes the full sequence of observations at once. This can be problematic for some dynamical systems as the initial estimate might be far away from the observations or when the variance of the prior process is very large. For such cases, we employ an online version of the variational smoother. For this online version, Alg. 1 is run for a number of steps on the first observation only. Then, the second observation is included and the smoother is initialized with the last control of the previous step. This procedure is repeated until all observations are processed.

4 PARAMETER INFERENCE

Variational smoothing algorithms can be straightforwardly extended to inference of model parameters. Let θ be a collection of real-valued parameters and extend the prior model such that the drift and diffusion terms are understood as functions of θ. More explicitly, replace $a(x) \to a(x, \theta)$ and $b(x) \to b(x, \theta)$ in the model given by (1). We can now proceed along the line of Sec. 3 to derive a relaxed variational inference problem (see App. A2.5.1). Again the result can be phrased as a

control problem

$$\min_{\theta,u,\varphi} \quad \int_0^{t_f} L(\theta, u(t), \varphi(t))\,\mathrm{d}t - \sum_{k=1}^{n} F_i(\varphi(t_k)) \qquad (21)$$
$$\text{s.t.} \quad \dot{\varphi}(t) = f(\theta, u(t), \varphi(t))$$

Solving the control problem (21) is equivalent to maximizing an approximate evidence lower bound. We discuss three ways to solve (21). In the first approach, θ and u are optimized interchangeably corresponding to the usual variational expectation maximization framework. The second idea is to construct a joint gradient descent in the parameters θ and the controls u. In practice, we observed that a combination of both approaches works well, where we alternately take a number of gradient steps for θ and u.

Finally, we consider a scenario were we have several independent time series samples $Y^1, \ldots, Y^N$ from the same underlying model. The standard variational inference procedure in this case requires computing $u_n^*(t)$ for each time series Y^n to perform a single parameter update. This becomes intractable for larger data sets. We therefore consider an amortized approach based on an inference network. The idea is to model the controls as a parametric function of the observations. In our case, we set $u_n(t) = h(y_n, \phi)$ where h is a feed-forward neural network parametrized by ϕ. As shown in App. A2.5.2, the corresponding optimization problem becomes

$$\min_{\theta,\phi,\varphi} \quad \sum_{i=1}^{N} \int_0^{t_f} L(\theta, h(y_i, \phi), \varphi_i(t))\,\mathrm{d}t - \sum_{k=1}^{n} F_i(\varphi_i(t_k))$$
$$\text{s.t.} \quad \dot{\varphi}_i(t) = f(\theta, h(y_n), \varphi_i(t)) \quad i = 1, \ldots, N \qquad (22)$$

For an implementation in PyTorch, we can exploit that our approach is gradient-based. Prop. 2 allows us to compute the gradient of the objective function with respect to an arbitrary control $u(t)$. We can thus backpropagate through the variational smoothing code such that it supports automatic differentiation. Conceptually, this is similar to neural ODE framework (Chen et al., 2018) which allows to backpropagate through an ODE solver. Using the resulting module as the loss function, the inference network can be trained end-to-end using standard optimizers based on back-propagation. For a simple conceptual demonstration of the inference network, we refer to Sec. 5.3.

5 EXPERIMENTS

In this section, we present four examples chosen to illustrate the versatility of our approach. For more details regarding the model equations and implementation, we refer to App. A3.

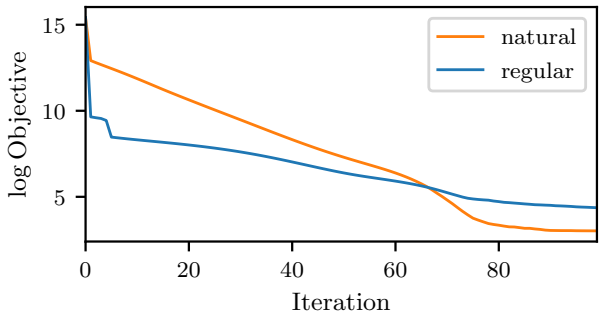

Figure 1: Evolution of the objective function under natural gradient descent (red) and regular gradient descent (blue). The lines correspond to the log of the objective function averaged over 10 runs started with randomly initialized controls.

5.1 Regular Gradient vs. Natural Gradient

We are interested in comparing the performance of the natural and regular gradient descent. We investigate this using the non-linear diffusion given by

$$\mathrm{d}X_t = 4X_t(1 - X_t^2)\mathrm{d}t + \sigma \mathrm{d}W_t \qquad (23)$$

that was also featured in the original work on Gaussian process approximations Archambeau et al. (2007a,b). The drift of the system has two stable stationary points at $x = \pm 1$. On occasion, the process noise may drive the system from one stationary point to another. We pick one fixed trajectory for which such a switch occurs. We then generate 10 different initial controls at random. For each of these initial controls, we perform the optimization with regular gradient descent and with natural gradient descent. The averaged log-transformed objective functions over gradient iterations are shown in Fig. 1. We observe that the natural gradient descent is more effective then the regular gradient descent, in particular in the middle part of the optimization. Also note that for small to medium dimensions, the computation time per gradient step is approximately equal for both methods. This is because the Fisher information is required for both (see Prop. 2). Only for larger system, the matrix inversion in (17) may become prohibitive compared to the forward and backward ODE solution.

5.2 Joint Smoothing and Inference

Geometric Brownian motion is a simple example of a process with a state dependent diffusion term and thus cannot be treated in the linear gaussian process framework. Here, we consider a simple multivariate extension given by the SDE system

$$\mathrm{d}X_{i,t} = r_i X_{i,t}\mathrm{d}t + X_{i,t}\mathrm{d}\tilde{W}_{i,t} \qquad (24)$$

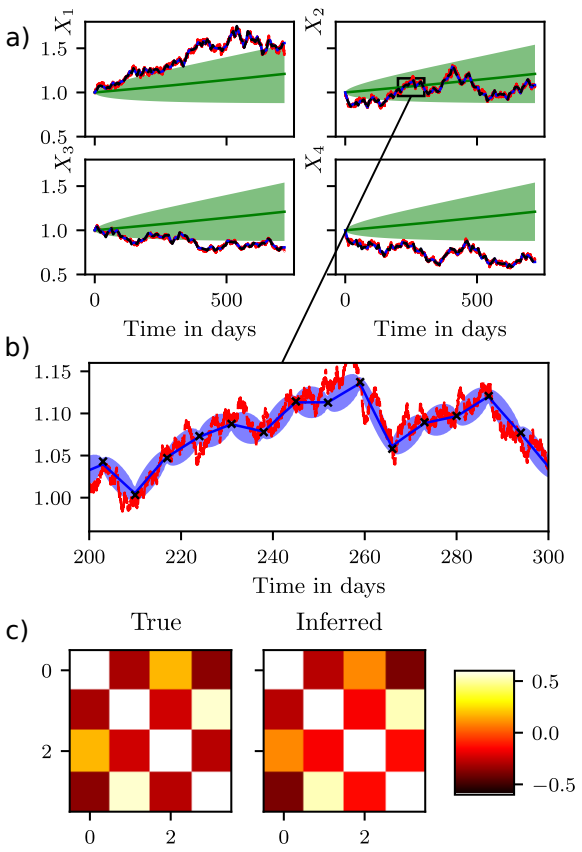

Figure 2: Joint smoothing and inference for a multivariate geometric Brownian with $n = 4$. a) Noisy observations, ground truth and smoothed state for the process components. For comparison, we show the prior process initialized with uninformative parameters. The shaded region indicates the standard deviation of the prior. b) A zoom-in showing the posterior compared to the noisy observations and the ground truth. The shaded blue region indicates the standard deviation of the variational posterior. c) Ground truth of the correlation matrix compared to reconstructed correlation matrix.

for each component i. Here $\tilde{W}_{i,t}$ is a collection of correlated Brownian motions. Similar as for a multivariate normal distribution, a correlated Brownian motion can be constructed as $\tilde{W}_t = RW_t$ where W_t is a vector of independent standard Brownian motions and the matrix R encodes the correlations. We consider a noise-dominant scenario and thus treat R as the parameter to be inferred. To test joint inference and smoothing, we simulated a trajectory over an interval of $[0, 720]$ with independent Gaussian observations every 7 units. For optimization, we use the alternating gradient descent. As demonstrated qualitatively in Fig. 2, state and correlation structure can be inferred quite well. Note that we show the correlation matrix $RR^\top$ since many R may give rise to the same process. The details of the experiment and a more quantitative evaluation are given in App. A3.1.

5.3 Amortized Smoothing

We explore the possibility of amortized smoothing (Sec. 4). To keep it simple, we consider a two-dimensional Ornstein-Uhlenbeck process given by the SDE

$$\mathrm{d}X_t = -\gamma(X_t - \mu)\mathrm{d}t + \sigma \mathrm{d}W_t \,.$$

where $\mu \in \mathbb{R}^2$, $\gamma, \sigma \in \mathbb{R}^{2\times 2}$. We generated 1000 trajectories of a two-dimensional model with fixed parameters and initial conditions. Each sample was observed over 20 s with 9 evenly spaced observation. The inference network was trained over 50 epochs using the Adam optimizer with default parameters, a weight decay of 0.001 and a batch size of one. Fig. 3 shows the prediction of the smoothing network on a previously unseen sample compared to the exact solution. This demonstrates, in principle, that the controls for variational smoothing can be learned and that the inference network generalizes to unseen trajectories.

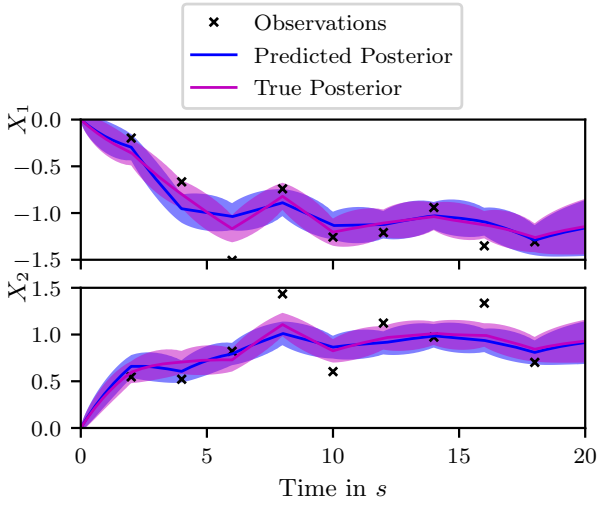

Figure 3: Two-dimensional Ornstein-Uhlenbeck process. For both components, the graph shows the smoothing predicted by the inference network on a previously unseen example. For comparison, we also show the simulated ground truth and the true posterior. Shaded regions indicate the standard deviations of the corresponding process.

5.4 Population Models

Population models describe the time evolution of a number of species over time. A convenient way to represent a population model is via the language of

chemical reactions. More precisely, let there be species $\mathcal{X}_1, \mathcal{X}_d$ and r reactions of the form

$$s_{i1}\mathcal{X}_1 + \ldots + s_{id}\mathcal{X}_d \xrightarrow{c_i} p_{i1}\mathcal{X}_1 + \ldots + p_{id}\mathcal{X}_d$$

with the matrices S and P encoding the number of molecules before and after a certain reaction event and the rate constants c_i determining the time scale of each reaction. In addition, let $V \equiv P - S$. Then the j-th row v_j of V encodes the net change caused by reaction j. Under certain conditions, the concentrations of the species is governed by the chemical Langevin equation Gillespie (2000). This leads to an SDE of the form

$$\mathrm{d}X_t = V^\top h(X_t)dt + \sqrt{V^\top \mathrm{diag}(h(X_t))V}\mathrm{d}W_t \quad (25)$$

where $\sqrt{\cdot}$ indicates a matrix square root and the mass-action propensity $h : \mathbb{R}^d \to \mathbb{R}^r$ is defined component wise by

$$h_i(x) = c_i \prod_{k=1}^{d} \frac{x_k!}{s_{ik}!(x_k - s_{ik})!}.$$

We combine a linear control and a multivariate log-normal closure to derive a general method for (25) (see App. A2.4). As a test system, we use the stochastic Lotka-Volterra model that describes the interaction of a prey species and a predator species. The corresponding matrices are given by

$$S = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \quad , \quad P = \begin{pmatrix} 2 & 0 \\ 0 & 2 \\ 0 & 0 \end{pmatrix}$$

We stress, however, that our code is not specific to the predator prey dynamics but takes general S and P as input. To study the behavior of our approach we recreate a scenario from Ryder et al. (2018). We generate a synthetic trajectory starting from the initial $X_0 = (71, 79)^\top$ and take four observations within the interval $[0, 50]$. As shown Fig. 4, the variational smoothing can reconstruct the true trajectory quite accurately. We also observe that only four observations restrict the variance of the process significantly.

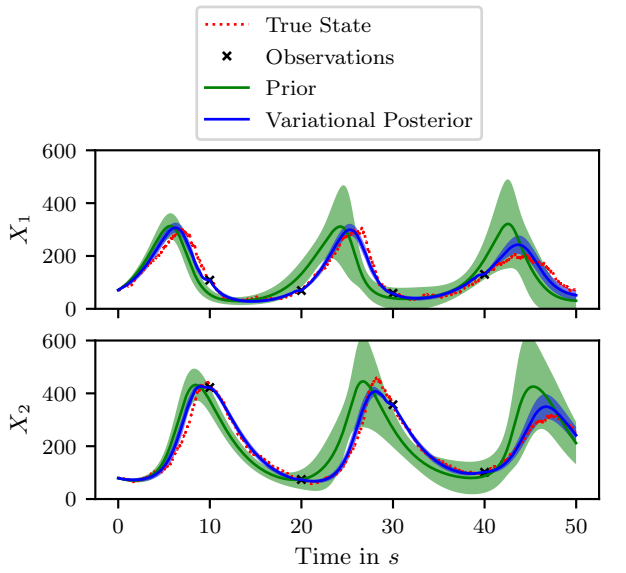

Figure 4: Smoothing for the stochastic Lotka-Volterra system. Solid lines indicate the mean, the shaded area indicates the corresponding standard deviation. Results are shown for the prior process and the variational smoothing. For reference, the simulated ground truth and the noisy observations are also provided.

6 DISCUSSION

We provide an ODE-based approach to variational smoothing that extends classical Gaussian process regression to models with state-dependent diffusion and allows for more versatile variational families. To achieve this, we understand the variational process as a controlled modification of the prior process and project the marginal posterior to a set of selected moment functions. In comparison to earlier work, we apply a refined optimization algorithm based on the natural gradient descent. Conceptually, our work extends a previous moment-based variational approach from MJPs to SDEs. Due to the structural similarity of both approaches, the moment-based variational method provides a unified inference framework for both process classes. In interesting future direction is to extend the moment-based variational framework to other Markov processes, in particular to jump-diffusions. In this work, we have used two simple closure schemes that work sufficiently well in the considered examples. Future work may consider more advanced clsoure schemes and also investigate the effect of different closures on the inference quality.

While previous ODE-based approaches have required manual derivations of the backward equation and gradients with respect to the parameters, we exploit automatic differentiation to construct these quantities automatically. In general, our approach only requires to provide two model-specific functions. For certain subclasses, these functions can be constructed automatically as well. Since our method is gradient-based, it can implemented as an automatically differentiable function. This allows straightforward integration with deep models. As a first example, we train an amortized inference network on a toy model with known model parameters. A promising future direction is to extend this to a full variational autoencoder for time series.

Acknowledgements

This work was supported by the European Research Council (ERC) within the CONSYN project, grant agreement number 773196.

References

S.-I. Amari. Natural Gradient Works Efficiently in Learning. *Neural Comput.*, 10(2):251–276, 1998.

C. Archambeau and M. Opper. Approximate inference for continuous-time Markov processes. *Bayesian Time Series Models*, Jan. 2011.

C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor. Gaussian process approximations of stochastic differential equations. 1:1–16, 2007a.

C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor. Variational Inference for Diffusion Processes. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 17–24. Curran Associates Inc., 2007b.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518): 859–877, 2017.

L. Bronstein and H. Koeppl. A variational approach to moment-closure approximations for the kinetics of biomolecular reaction networks. *The Journal of Chemical Physics*, 148(1):014105, 2018.

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6571–6583. Curran Associates, Inc., 2018.

B. Cseke, D. Schnoerr, M. Opper, and G. Sanguinetti. Expectation propagation for continuous time stochastic processes. *Journal of Physics A: Mathematical and Theoretical*, 49(49):494002, Nov. 2016. ISSN 1751-8113.

P. Del Moral and L. Murray. Sequential Monte Carlo with Highly Informative Observations. *SIAM/ASA Journal on Uncertainty Quantification*, 3, May 2014.

L. Duncker, G. Bohner, J. Boussard, and M. Sahani. Learning interpretable continuous-time models of latent stochastic dynamical systems. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1726–1734, Long Beach, California, USA, June 2019. PMLR.

S. N. Ethier and T. G. Kurtz. *Markov Processes : Characterization and Convergence*. Wiley Series in Probability and Statistics. 2005.

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

A. Golightly and D. J. Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1(6):807–820, 2011.

L. Isserlis. On a Formula for the Product-Moment Coefficient of any Order of a Normal Frequency Distribution in any Number of Variables. *Biometrika*, 12 (1/2):134–139, 1918. ISSN 00063444. doi: 10.2307/2331932.

C. Kuehn. Moment Closure—A Brief Review. In E. Schöll, S. H. L. Klapp, and P. Hövel, editors, *Control of Self-Organizing Nonlinear Systems*, pages 253–271. 2016.

X. Li, T.-K. L. Wong, R. T. Q. Chen, and D. Duvenaud. Scalable gradients for stochastic differential equations. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3870–3882. PMLR, 26–28 Aug 2020.

A. Meurer, C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel, v. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, and A. Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, Jan. 2017.

S. Mitter and N. Newton. A Variational Approach to Nonlinear Estimation. *SIAM J. Control and Optimization*, 42:1813–1833, Jan. 2003.

A. Ruttor, P. Batz, and M. Opper. Approximate Gaussian process inference for the drift function in stochastic differential equations. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2040–2048. Curran Associates, Inc., 2013.

T. Ryder, A. Golightly, A. S. McGough, and D. Prangle. Black-box variational inference for stochastic differential equations. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4423–4432, Stockholmsmässan, Stockholm Sweden, July 2018. PMLR.

S. Särkkä. *Bayesian Filtering and Smoothing*, volume 3 of *Institute of Mathematical Statistics Textbooks*. Cambridge, 2013.

S. Särkkä and A. Solin. *Applied Stochastic Differential Equations*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge, 2019.

R. F. Stengel. *Optimal Control and Estimation*. Dover Books on Advanced Mathematics. New York, unabridged, corr. republ. edition, 1994.

T. Sutter, A. Ganguly, and H. Koeppl. A Variational Approach to Path Estimation and Parameter Inference of Hidden Diffusion Processes. *Journal of Machine Learning Research*, 17(190):1–37, 2016.

M. D. Vrettas, D. Cornford, and M. Opper. Estimating parameters in stochastic systems: A variational Bayesian approach. *Physica D: Nonlinear Phenomena*, 240(23):1877–1900, Nov. 2011.

M. J. Wainwright and M. I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1 (1–2):1–305, 2008.

C. Wildner and H. Koeppl. Moment-Based Variational Inference for Markov Jump Processes. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6766–6775. PMLR, 2019.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized Belief Propagation. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pages 668–674. MIT Press, 2000.