Black-Box Variational Inference for Stochastic Differential Equations

Thomas Ryder [* 1 2] **Andrew Golightly** [1] **A. Stephen McGough** [2] **Dennis Prangle** [* 1]

Abstract

Parameter inference for stochastic differential equations is challenging due to the presence of a latent diffusion process. Working with an Euler-Maruyama discretisation for the diffusion, we use variational inference to jointly learn the parameters and the diffusion paths. We use a standard mean-field variational approximation of the parameter posterior, and introduce a recurrent neural network to approximate the posterior for the diffusion paths conditional on the parameters. This neural network learns how to provide Gaussian state transitions which bridge between observations as the conditioned diffusion process does. The resulting black-box inference method can be applied to any SDE system with light tuning requirements. We illustrate the method on a Lotka-Volterra system and an epidemic model, producing accurate parameter estimates in a few hours.

1. Introduction

A stochastic differential equation (SDE) defines a *diffusion process*, which evolves randomly over time, by describing its instantaneous behaviour. As such, SDEs are powerful modelling tools used extensively in fields such as econometrics (Black & Scholes, 1973; Eraker, 2001), biology (Gillespie, 2000; Golightly & Wilkinson, 2011), physics (van Kampen, 2007) and epidemiology (Fuchs, 2013).

It is only possible to work with analytic solutions to SDEs in special cases. Therefore it is common to use a numerical approximation, such as the Euler-Maruyama scheme. Here the diffusion process is defined only on a grid of time points, and the transition density between successive diffusion states is approximated as Gaussian. The approximation

error involved converges to zero as the grid becomes finer.

Even under discretisation, statistical inference for SDEs observed at discrete times is challenging. The difficulty is that, along with unknown parameters θ in the description of the SDE, there is an unknown *latent path* of the diffusion process, x. An inference method must somehow deal with these high dimensional, highly structured latent variables.

Our proposed method uses recent advances in variational inference to jointly infer θ and x. We introduce a flexible family of approximations to the posterior distribution and select the member closest to the true posterior. We use a standard mean-field approximation for the θ posterior, and introduce a novel recurrent neural network (RNN) approximation for the posterior of x conditional on θ. The RNN learns how to supply Gaussian state transitions between successive time points which closely match those for the intractable conditioned diffusion process.

Our black-box variational inference method is a simple and fast way to produce approximate inference for any SDE system. We illustrate our method on Lotka-Volterra and epidemic examples, achieving accurate parameter estimates in just a few hours under default tuning choices. Although our parameter posteriors are over-concentrated, as in most variational methods, our approximation of the conditioned diffusion process is close to the true posterior. In comparison, existing Markov chain Monte Carlo (MCMC) methods (see Section 1.1) require more tuning choices and can take days to run (Whitaker et al., 2017a).

1.1. Related Work

Variational inference Several authors have looked at variational inference for SDEs (Archambeau et al., 2008) or related problems such as Markov jump processes (Ruttor et al., 2010) and state space models (Archer et al., 2016; Quiroz et al., 2018). The novelty of our approach is to use: (1) stochastic optimisation rather than variational calculus; (2) a RNN-based variational approximation for the latent states instead of a mean-field or multivariate Gaussian approach. We expect (2) is especially relevant to sparsely observed SDEs, where the latent states between observations may have a particularly complex dependency structure.

Another approach (Moreno et al., 2016) is to perform vari-

*Equal contribution [1]School of Mathematics, Statistics and Physics, Newcastle University, Newcastle, UK [2]School of Computing, Newcastle University, Newcastle, UK. Correspondence to: Tom Ryder <t.ryder2@newcastle.ac.uk>, Dennis Prangle <dennis.prangle@newcastle.ac.uk>.

ational inference for the parameters only, using latent variables drawn from their prior in the ELBO estimate. Such latent states are typically a poor match to the observed data and so make a negligible contribution to the ELBO. To deal with the problem, close matches are upweighted. Our approach avoids this extra approximation by instead learning the posterior distribution of the latent variables.

Our method can also be related to recent work on *normalising flows* as variational approximations (Rezende & Mohamed, 2015). As in that work, our variational approximation can be viewed as transforming a $N(0, I)$ sample vector by successive linear transformations to an approximate posterior sample (of the diffusion states in our case). Our work uses SDE theory to select simple and cheap transformations which produce a particularly good approximation.

Monte Carlo A popular approach in the Monte Carlo literature on SDEs is to introduce a *bridge construct*: an approximation to the discretised diffusion process conditional on the parameters and observations at a single time, derived using probability theory and various simplifying approximations. The goal is to produce a path bridging between two observation times. Combining successive bridges forms a complete diffusion path. Bridge constructs can be used to produce proposals within Monte Carlo algorithms such as MCMC (see e.g. Roberts & Stramer 2001; Golightly & Wilkinson 2008; Fuchs 2013; van der Meulen et al. 2017). However, designing a bridge construct with desirable features for a particular problem is a challenging and time consuming tuning choice. (Some particularly difficult regimes for bridge constructs are discussed in Section 5.) From this point of view, our contribution is to use machine learning to effectively automate the design of a bridge construct.

Another Monte Carlo approach is to perform approximate inference based on low dimensional summary statistics of the observations (Picchini, 2014). This results in a loss of information, which our approach avoids.

2. Stochastic Differential Equations

Consider an Itô process $\{X_t, t \geq 0\}$ satisfying the SDE

$$dX_t = \alpha(X_t, \theta)dt + \sqrt{\beta(X_t, \theta)}dW_t, \quad X_0 = x_0. \quad (1)$$

Here X_t is a p-dimensional vector of random variables, α is a p-dimensional *drift vector*, β is a $p \times p$ positive definite *diffusion matrix* (with $\sqrt{\beta}$ representing a matrix square root) and W_t is a p-vector of standard and uncorrelated Brownian motion processes. The drift and diffusion depend on $\theta = (\theta_1, \theta_2, ..., \theta_c)'$, a vector of unknown parameters (which may also include the initial condition x_0).

We assume that $\alpha(\cdot)$ and $\beta(\cdot)$ are sufficiently regular that (1) has a weak non-explosive solution (Øksendal, 2003). In

this case, (1) defines a *diffusion process*. Such processes are always Markovian (i.e. memoryless).

We further assume partial noisy observations of the latent process. Suppose that there are $d + 1$ observation times $t_0, t_1, \ldots, t_d = T$. In the simplest case, these times are *equally spaced*, separated by a time-step of Δt. Let y_{t_j} be a vector of p_0 observations at time t_j, for some $p_0 \leq p$. Following Golightly & Wilkinson (2008), among others, we assume that

$$y_{t_j} = F'X_{t_j} + \omega_{t_j}, \quad \omega_{t_j} \overset{indep}{\sim} N(0, \Sigma), \quad (2)$$

where F is a constant $p \times p_0$ matrix, and Σ is a $p_0 \times p_0$ matrix which may be assumed known or the object of inference. For the latter case Σ should be a specified function of θ.

Upon choosing a prior density $p(\theta)$, Bayesian inference proceeds via the parameter posterior $p(\theta|y)$, or alternatively the joint posterior $p(\theta, x|y)$.

Discretisation Few SDEs permit analytical solutions to (1) and instead it is common to use an approximation based on time discretisation. We therefore introduce intermediate time-points between observation times. For concreteness, we present our methods for the case of equally spaced observations with $t_0 = 0$. (It is easy to adapt them to alternative specifications of time points, such as those required by irregularly-spaced observation times.) We introduce $k - 1$ time-points between successive observations, giving a regular grid of times $\tau_i = i\Delta\tau$ for $i = 0, 1, 2, \ldots, m = dk$, with time-step $\Delta\tau = \Delta t/k$. Note that $i = 0, k, 2k, \ldots, dk$ give the observation times. The role of k is to ensure the discretisation can be made arbitrarily accurate, at the expense of increased computational cost.

We work with the simplest discretisation, the Euler-Maruyama scheme, in which transition densities between states at successive times are approximated as Gaussian

$$p\left(x_{\tau_{i+1}}|x_{\tau_i}, \theta\right) = \varphi\big(x_{\tau_{i+1}} - x_{\tau_i}; \alpha(x_{\tau_i}, \theta)\Delta\tau, \\ \beta(x_{\tau_i}, \theta)\Delta\tau\big), \quad (3)$$

where $\varphi(\cdot; \mu, S)$ is the Gaussian density with mean μ and variance matrix S. A generative expression of this is

$$x_{\tau_{i+1}} = x_{\tau_i} + \alpha(x_{\tau_i}, \theta)\Delta\tau + \sqrt{\beta(x_{\tau_i}, \theta)\Delta\tau}\, z_{i+1}, \quad (4)$$

where z_{i+1} is an independent $N(0, I_p)$ realisation.

Discretisation is not guaranteed to preserve properties of the underlying SDE. An issue which is particularly relevant later is positivity. In many SDEs, such as population models, it is guaranteed that some components of X_t are always positive. However, in (4) $x_{\tau_{i+1}}$ is sampled from a Gaussian, which has unbounded support. Consequently, there is a non-zero probability of sampling negative values. This is problematic

as the drift or diffusion function may be poorly behaved or undefined for such input. A simple solution to this problem is the use of a reflecting boundary (Skorokhod, 1961), for example by projecting invalid $x_{\tau_{i+1}}$ values back to the valid region (Dangerfield et al., 2012).

Posterior   The joint posterior under the Euler-Maruyama discretisation is

$$p(\theta, x|y) \propto p(\theta)p(x|\theta)p(y|x, \theta), \tag{5}$$

$$p(x|\theta) = \prod_{i=0}^{m-1} \varphi\big(x_{\tau_{i+1}} - x_{\tau_i}; \alpha(x_{\tau_i}, \theta)\Delta\tau, \tag{6}$$
$$\beta(x_{\tau_i}, \theta)\Delta\tau\big)$$

$$p(y|x, \theta) = \prod_{i=0}^{d} \varphi\left(y_{t_i}; F'x_{t_i}, \Sigma\right). \tag{7}$$

In principle Monte Carlo algorithms can sample from (5). However this is difficult in practice due to its high dimension and complex dependency structure.

Conditioned processes   Consider the process defined by conditioning (1) on an initial state, x_0 and an exactly observed future state, x_{t_1}. This conditioned process itself satisfies an SDE (see e.g. Rogers & Williams, 2013) with drift and diffusion

$$\hat{\alpha}(x_t, \theta) = \alpha(x_t, \theta) + \beta(x_t, \theta)\nabla_{x_t} \log \pi\left(x_{t_1}|x_t, \theta\right), \tag{8}$$

$$\hat{\beta}(x_t, \theta) = \beta(x_t, \theta), \tag{9}$$

where $\pi\left(x_{t_1}|x_t, \theta\right)$ is the transition density of the unconditioned process. While this is intractable in most cases, the result motivates our choice of variational approximation later.

In some simple situations a discretised approximation of this conditioned process can be derived (see e.g. Papaspiliopoulos et al. 2013) in which the diffusion matrix is scaled down as the observation time is approached. Intuitively this is appealing: conditioned paths converge towards the observation, so nearby random deviations are smaller in scale. This motivates us to use a variational approximation in which the diffusion matrix is not constrained to follow (9), and instead is allowed to shrink.

3. Approximate Bayesian Inference

Suppose we have a likelihood $p(y|\theta)$ for parameters θ under observations y. Given a prior density $p(\theta)$ we wish to infer the posterior density $p(\theta|y) = p(\theta)p(y|\theta)/p(y)$. It is typically possible to numerically evaluate the unnormalised posterior $p(\theta, y) = p(\theta)p(y|\theta)$. Estimating the normalising constant $p(y) = \int p(\theta, y)d\theta$, known as the *evidence*, is useful for Bayesian model selection.

3.1. Variational Inference

Variational inference (VI) (see e.g. Blei et al., 2017) introduces a family of approximations to the posterior indexed by ϕ, $q(\theta; \phi)$. Optimisation is then used to find ϕ minimising the Kullback-Leibler divergence $KL(q(\theta; \phi)||p(\theta|y))$. This is equivalent to maximising the ELBO (evidence lower bound) (Jordan et al., 1999),

$$\mathrm{E}_{\theta \sim q(\cdot; \phi)}[\log p(\theta, y) - \log q(\theta; \phi)]. \tag{10}$$

The optimal $q(\theta; \phi)$ is an approximation to the posterior distribution. This is typically overconcentrated, unless the approximating family is rich enough to allow particularly close matches to the posterior.

The optimisation required by VI can be performed efficiently using the *reparameterisation trick* (Kingma & Welling, 2014; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014). This requires expressing $\theta \sim q(\cdot; \phi)$ as a *non-centred parameterisation* (Papaspiliopoulos et al., 2003). That is, writing θ as the output of an invertible deterministic function $g(\epsilon, \phi)$ for some random variable ϵ with a fixed distribution. Then the ELBO can be written as

$$\mathcal{L}(\phi) = \mathrm{E}_\epsilon[\log p(\theta, y) - \log q(\theta; \phi)], \tag{11}$$

with an unbiased Monte-Carlo estimate

$$\hat{\mathcal{L}}(\phi) = \frac{1}{n} \sum_{i=1}^{n} [\log p(\theta^{(i)}, y) - \log q(\theta^{(i)}; \phi)], \tag{12}$$

where $\theta^{(i)} = g(\epsilon^{(i)}, \phi)$ and $\epsilon^{(1)}, \ldots, \epsilon^{(n)}$ are independent ϵ samples. Assuming $\hat{\mathcal{L}}$ is differentiable with respect to ϕ, the gradient of (12) can be calculated using automatic differentiation, and the resulting unbiased estimator of $\nabla\mathcal{L}(\phi)$ used in stochastic gradient descent or similar algorithms.

3.2. Importance Sampling

When variational inference outputs a good match to the posterior distribution, *importance sampling* (IS) (see e.g. Robert, 2004) can correct remaining inaccuracies and provide near-exact posterior inference. In more detail, select an *importance density* $q(\theta)$ which can easily be sampled from, and satisfies $\mathrm{supp}\, q(\theta) \supseteq \mathrm{supp}\, p(\theta|y)$. IS samples $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$ from q and calculates weights $w_i = p(\theta^{(i)}, y)/q(\theta^{(i)})$. Then, for any function h, an estimate of $\mathrm{E}_{\theta \sim p(\cdot|y)}[h(\theta)]$ is

$$\sum_{i=1}^{N} h(\theta^{(i)})w_i \bigg/ \sum_{i=1}^{N} w_i. \tag{13}$$

This is consistent for large N, but in practice q should approximate the posterior for accurate estimation at a feasible cost. Also note that $N^{-1}\sum_{i=1}^{N} w_i$ is an unbiased and consistent estimate of the evidence.

A diagnostic for the quality of IS results is the *effective sample size* (ESS). This is defined as

$$N_{\text{eff}} = \left(\sum_{i=1}^{N} w_i \right)^2 \Big/ \sum_{i=1}^{N} w_i^2. \tag{14}$$

For most functions h, the variance of (13) approximately equals that of an idealised Monte Carlo estimate based on N_{eff} independent samples from $p(\theta|y)$ (Liu, 1996). In practice we will use a variational approximation as the importance density, and the ESS to assess whether this is sufficiently good to produce accurate estimates. However, ESS values can be an unstable for poor importance densities (Vehtari et al., 2017) so later we also consider other problem-specific evidence for the quality of our results.

4. Variational Inference for SDEs

Our variational approximation to the posterior (5) is

$$q(\theta, x; \phi) = q(\theta; \phi_\theta) q(x|\theta; \phi_x). \tag{15}$$

These factors represent approximations to $p(\theta|y)$ and $p(x|\theta, y)$ respectively, which are described below. Here ϕ_θ and ϕ_x are the variational parameters for the two factors, and ϕ is the collection of all variational parameters. Note our eventual choice of $q(x|\theta; \phi_x)$ depends on several features of the data and model (see list in Section 4.5), but we suppress this in our notation for simplicity.

4.1. Approximate Parameter Posterior

For $q(\theta; \phi_\theta)$ we use the *mean-field Gaussian* approximation

$$q(\theta; \phi_\theta) = \prod_{i=1}^{c} \varphi(\theta_i; \mu_i, s_i^2), \tag{16}$$

with $\phi_\theta = (\mu_1, \ldots, \mu_c, s_1, \ldots, s_c)$. Hence the components of θ are independent Gaussians. To express θ using a non-centred parameterisation, we write

$$\theta = g_\theta(\epsilon_1, \phi_\theta) = S\epsilon_1 + \mu. \tag{17}$$

where $\epsilon \sim N(0, I_c)$, $S = \text{diag}(s_1, \ldots, s_c)$ and $\mu = (\mu_1, \ldots, \mu_c)$.

It may be necessary to transform θ to an alternative parameterisation ϑ so that a Gaussian approximation is appropriate e.g. log-transforming parameters constrained to be positive.

Mean-field approximations are imperfect, often producing underdispersed estimates of the posterior (see e.g. Blei et al., 2017), and more sophisticated approximations (e.g. Rezende & Mohamed, 2015) could be used here instead. However mean-field approximations suffice to give good parameter estimation in our examples.

4.2. Approximate Conditioned Diffusion Process

Motivated by the result that a diffusion process conditioned on an exact observation is itself an SDE (see Section 2), we base $q(x|\theta; \phi_x)$ upon a discretised diffusion. A generative definition is

$$x_{\tau_{i+1}} = h\bigg(x_{\tau_i} + \tilde{\alpha}(x_{\tau_i}, y, \theta, \tau_i; \phi_x)\Delta\tau \\ + \sqrt{\tilde{\beta}(x_{\tau_i}, y, \theta, \tau_i; \phi_x)\Delta\tau} z_{i+1} \bigg), \tag{18}$$

where $\tilde{\alpha}$ and $\tilde{\beta}$ are drift and diffusion functions. Taking h as the identity function gives a discretised diffusion process. However often we use h to impose positivity constraints on some components of x – see Section 4.3.

The resulting variational density $q(x|\theta; \phi_x)$ is

$$\prod_{i=0}^{m-1} \varphi\bigg(x_{\tau_{i+1}} - x_{\tau_i}; \tilde{\alpha}(x_{\tau_i}, y, \theta, \tau_i; \phi_x)\Delta\tau, \\ \tilde{\beta}(x_{\tau_i}, y, \theta, \tau_i; \phi_x)\Delta\tau \bigg) |\det J_i|. \tag{19}$$

where J_i is the Jacobian matrix associated with the transformation h in (18). To express x with a non-centred parameterisation, let $\epsilon_2 \sim N(0, I_{pm})$ be the flattened vector of $(z_1, z_2, \ldots, z_m)$ realisations. Then apply (18) repeatedly. Let the outcome be represented by the function

$$x = g_x(\epsilon_2, \theta, \phi_x). \tag{20}$$

We use a neural network, with parameters ϕ_x, to serve as our functions $\tilde{\alpha}$ and $\tilde{\beta}$. At time τ_i it acts as follows. The network's input is several features (described in Section 4.5) computed from: the current diffusion state x_{τ_i}, the observations y, the parameters θ and the current time τ_i. The network outputs a drift vector and diffusion matrix (see Section 4.5 for details of the latter), which are used to sample $x_{\tau_{i+1}}$ from (18). This state forms part of the neural network input at time τ_{i+1}. So the network just discussed forms a *cell* of an overall recurrent neural network (RNN) structure for $q(x|\theta; \phi_x)$. Note that long-term memory features are not required as we wish to produce a diffusion process, which is memoryless.

4.3. Ensuring Positivity

In practice, we take h in (18) to be a function which applies the identity function to unconstrained components of the diffusion state, and the function $\text{softplus}(z) = \log(1 + e^z)$ to components with positivity constraints. This function produces strictly positive outputs while having little effect on positive inputs above 2. The latter property means our variational approximation usually remains similar to a discretised

diffusion process. Alternative transformations could be used if state values below 2 were believed to be common, potentially with tuning parameters so a suitable shape could be learned. However, this was not necessary for our examples.

Preliminary work found this transformation approach to enforcing positivity was much easier to implement in the variational framework than reflection methods. It can be interpreted as constraining the variational approximation based on prior beliefs about positivity of diffusion paths.

4.4. Algorithm

Our overall inference algorithm is given in Algorithm 1. This aims to maximise the ELBO

$$\mathcal{L}(\phi) = \mathrm{E}_{\theta, x \sim q(\cdot; \phi)} \left[\log \frac{p(\theta)p(x|\theta)p(y|x, \theta)}{q(\theta; \phi_\theta)q(x|\theta; \phi_x)} \right], \quad (21)$$

by differentiating the Monte Carlo estimate

$$\hat{\mathcal{L}}(\phi) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(\theta^{(i)})p(x^{(i)}|\theta^{(i)})p(y|x^{(i)}, \theta^{(i)})}{q\left(\theta^{(i)}; \phi_\theta\right) q\left(x^{(i)}|\theta^{(i)}; \phi_x\right)}, \quad (22)$$

where $\theta^{(i)} = g_\theta(\epsilon_1^{(i)}, \phi_\theta)$ and $x^{(i)} = g_x(\epsilon_2^{(i)}, \theta^{(i)}, \phi_x)$.

Algorithm 1 Black-box variational inference for SDEs

    Initialise ϕ_0 and $k = 0$.
  loop
      Sample $\epsilon_1^{(i)}, \epsilon_2^{(i)}$ for $1 \le i \le n$.
      Calculate $\nabla \hat{\mathcal{L}}(\phi_k)$ using automatic differentiation of (22).
      Calculate ϕ_{k+1} using stochastic gradient descent, or a similar algorithm, and increment k.
  end loop

4.5. Implementation Details

Our RNN cell input at time τ_i, with $t_j \le \tau_i < t_{j+1}$, is:

- The parameters θ.
- The most recent latent state, $x_{\tau_{i-1}}$.
- The time until the next observation, $t_{j+1} - \tau_i$.
- The next observation time, t_{j+1}.
- The difference between the next observation and what the mean observation would be at the most recent latent state, $y_{t_{j+1}} - F' x_{\tau_{i-1}}$.

Exploratory work showed that the RNN produces a much better approximation of the conditioned process with these features as input rather than simply x_{τ_i}, y, θ and τ_i.

Our RNN cell outputs a vector $\tilde{\alpha}$ and the coefficients of a lower-triangular matrix, M. In order to return a Cholesky factor of $\tilde{\beta}$, the diagonal elements of M are transformed

using the softplus function to ensure positivity. We also regularise to avoid $\tilde{\beta}$ matrices with very small determinants.

Algorithm 1 requires automatic differentiation of (22). This can be achieved using the standard tool-kit of backpropagation after *rolling-out* the RNN i.e. stacking m copies of the RNN cell to form a deep feed-forward network. The canonical challenge in training such networks, known as the exploding-gradient problem (Bengio et al., 1994), often necessitates the use of gradient clipping to control for numerical instability. We follow Pascanu et al. (2013) and perform gradient clipping using the L_1 norm.

To initialise ϕ_0 in Algorithm 1, we select ϕ_θ so the margins of the variational approximation are based on those of the parameter priors. Standard choices from the neural network literature – random Gaussian weights and constant biases – are used for ϕ_x.

5. Experiments

We implement our method for two examples: (1) analysing synthetic data from a Lotka-Volterra SDE; (2) analysing real data from an SDE model of a susceptible-infectious-removed (SIR) epidemic. Our experiments include challenging regimes such as: (A) low-variance observations; (B) conditioned diffusions with non-linear dynamics; (C) unobserved time series; (D) widely spaced observation times; (E) data which is highly unlikely under the unconditioned model. Many of these violate the assumptions used by existing diffusion bridge constructs (Whitaker et al., 2017b).

In all our experiments below similar tuning choices worked well. We use batch size $n = 50$ in (22). Our RNN cell has four hidden layers each with 20 hidden units and rectified-linear activation. We implement our algorithms in Tensorflow using the Adam optimiser (Kingma & Ba, 2015) and report results using an 8-core CPU. The code is available at https://github.com/Tom-Ryder/VIforSDEs.

5.1. Lotka-Volterra

Lotka-Volterra models describe simple predator-prey population dynamics combining three types of event: prey reproduction, predation (in which prey are consumed and predators have the resources to reproduce) and predator death. A SDE Lotka-Volterra model (for derivation see e.g. Golightly & Wilkinson, 2011) is defined by

$$\alpha(X_t, \theta) = \begin{pmatrix} \theta_1 U_t - \theta_2 U_t V_t \\ \theta_2 U_t V_t - \theta_3 V_t \end{pmatrix}, \quad (23)$$

$$\beta(X_t, \theta) = \begin{pmatrix} \theta_1 U_t + \theta_2 U_t V_t & -\theta_2 U_t V_t \\ -\theta_2 U_t V_t & \theta_3 V_t + \theta_2 U_t V_t \end{pmatrix}, \quad (24)$$

where $X_t = (U_t, V_t)'$ represents the populations of prey and predators at time t. The parameters $\theta = (\theta_1, \theta_2, \theta_3)'$ control the rates of the three events described above.

In the experiments below, we use discretisation time step $\Delta\tau = 0.1$ and observation variance $\Sigma = I$, which is small relative to the typical population sizes (see e.g. Figure 1.)

A single observation time with known parameters We begin with the case of a single observation time and known parameter values, where we follow Boys et al. (2008) by taking $\theta = (0.5, 0.0025, 0.3)'$ and $x_0 = (71, 79)'$. This setting solely investigates our ability to learn x, without uncertainty in θ: essentially the same problem as creating a bridge construct (described in Section 1.1.)

We consider four different observations at time $t = 10$, listed in Table 1. For each example we train our variational approximation until convergence (assessed manually throughout this paper), which takes roughly 20 minutes for the first 3 examples, and 90 minutes for the last. We then perform importance sampling using 500,000 samples from the fitted approximation. Table 1 shows the resulting ESS values. The first 3 rows in the table are typical observations under the model given our θ, while the final row represents highly unlikely observations (double those in the previous row). Figure 1 shows fitted diffusion paths for this case.

Table 1. Summary of importance sampling performance for the Lotka-Volterra example with a single observation time and known θ. Each row shows the prey (U_{10}) and predator (V_{10}) observations at time $t = 10$. For each, 500,000 iterations of importance sampling are performed, using variational inference output as the importance density, and the effective sample size (14) is shown.

U_{10}	V_{10}	ESS
15.3	298.2	184,329
46.7	389.1	212,313
108.7	503.4	196,956
217.4	1006.9	95,711

This example contains several challenging features: all those listed at the start of Section 5 except (C). While these features make it hard to use existing bridge constructs, our variational method produces a close approximation to the true posterior, as illustrated by the high ESS values.

The case of highly unlikely observations takes longer to train and receives a lower ESS, reflecting that a more complicated diffusion path must be learned here. (To check this we found that a simpler RNN cell suffices for good performance in the other examples but not this one.)

Multiple observation times with known parameters We now extend the previous example to multiple observation times, $t = 0, 10, 20, 30, 40$. We analyse synthetic data, produced using the parameters specified previously (including observation noise with $\Sigma = I$). Here convergence takes 6 hours, and importance sampling with 500,000 samples produces an ESS of 96,812. The resulting diffusion
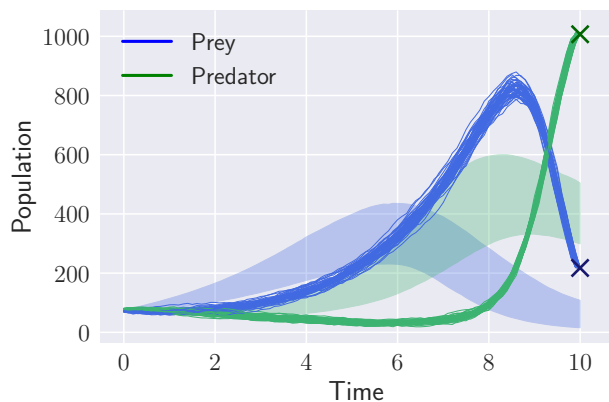

Figure 1. 50 conditioned diffusion paths for the Lotka-Volterra example with a single observation time and known parameters. These are sampled from the trained variational approximation. The observations used are from the last row of Table 1, representing highly unlikely observations. The shaded regions show typical paths of the unconditioned process (95% intervals estimated from repeated simulations.)

paths are not shown as they are very similar visually to the next example (see Figure 2).

This example shows that our method learns the conditioned process well even when there are several observation times.

Multiple observation times with unknown parameters We now analyse the same synthetic data with unknown θ parameters. As these must take positive values, we work with the log-transformed parameters ϑ and assume they have independent $N(0, 3^2)$ priors. Our results are shown after transforming back to the original parameterisation.

Convergence takes 2 hours, and importance sampling with 500,000 iterations produces an ESS of 635.4. Figure 2 shows 50 diffusion paths sampled from the fitted variational approximation. Figure 3 shows two estimates of the marginal parameter posteriors: the variational inference output, and a kernel density estimate based on importance sampling results.

The estimated posteriors give accurate estimates of the true parameter values. However, the low ESS here shows that both estimates of the parameter posteriors are imperfect approximations (also illustrated by the variational posterior estimates appearing overconcentrated compared to the importance sampling results.) Achieving good point estimates but imperfect posteriors is typical for variational inference (Blei et al., 2017).

5.2. Epidemic Model

An SIR epidemic model (Andersson & Britton, 2000) describes the spread of an infectious disease. The population is
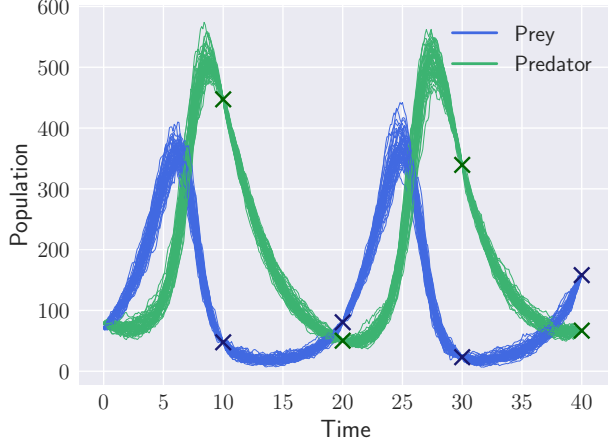
Figure 2. 50 conditioned diffusion paths for the Lotka-Volterra example with multiple observation times and unknown parameters. They are sampled from the fitted variational approximation.
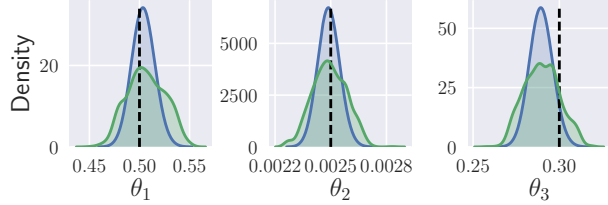

Figure 3. Marginal posterior density estimates for the Lotka-Volterra example from variational inference (blue) and importance sampling (green). Black dashed lines show true parameter values.

split into those susceptible (S), infectious (I) and removed (R). Two types of event take place: susceptibles can be infected by the infectious, and the infectious eventually become removed. Constant population size is assumed. Hence only the S and I population sizes need to be modelled.

An SIR epidemic model using SDEs is defined by

$$\alpha(X_t, \theta) = \begin{pmatrix} -\theta_1 S_t I_t \\ \theta_1 S_t I_t - \theta_2 I_t, \end{pmatrix} \tag{25}$$

$$\beta(X_t, \theta) = \begin{pmatrix} \theta_1 S_t I_t & -\theta_1 S_t I_t \\ -\theta_1 S_t I_t & \theta_1 S_t I_t + \theta_2 I_t \end{pmatrix}, \tag{26}$$

where $X_t = (S_t, I_t)'$ is the state of the system at time t, θ_1 is an infection parameter and θ_2 is a removal parameter. For a detailed derivation see Fuchs (2013).

Our data is taken from an outbreak of influenza at a boys boarding school in 1978 (Jackson et al., 2013). Influenza was introduced to the population by a student returning from holiday in Hong Kong. Of the 763 boys at the school, 512 were infected within 14 days. Hence we assume $x_0 = (762, 1)'$. Observations of the number infectious are provided daily by those students confined to bed. We

assume Gaussian observation error with unknown variance σ^2. Our analyses use a discretisation time step of $\Delta\tau = 0.1$.

We also consider an alternative model with a time-varying infection parameter. Here we let $\vartheta_1 = \log\theta_1$ follow an Ornstein-Uhlenbeck process

$$d\vartheta_{t,1} = \theta_3\left(\theta_4 - \vartheta_{t,1}\right)dt + \theta_5 dW_t, \tag{27}$$

where θ_3, θ_4 and θ_5 are the mean-reversion rate, process mean and volatility, respectively, and $\theta_{t,1}$ is the infection parameter at time t. Previous related work has focused on ODE epidemic models with time-varying parameters following SDEs (Dureau et al., 2013; Del Moral & Murray, 2015). In contrast, our approach can easily be applied to a full SDE system.

Time-invariant infection parameter   We infer the log-transformed parameters $\vartheta = (\log\theta_1, \log\theta_2, \log\sigma^2)'$ under independent $N(0, 3^2)$ priors. Our results are shown after transforming back to the original parameterisation.

Convergence takes 2.5 hours, and importance sampling with 500,000 iterations produces an ESS of 718.2. Figure 4 shows two estimates of the marginal parameter posteriors: variational inference output, and a kernel density estimate based on importance sampling results. Figure 6 shows 50 diffusion paths sampled from the variational approximation.

The small ESS indicates there is some approximation error. However, the marginal posteriors for θ_1 and θ_2 are very similar to those from the MCMC analysis of Fuchs (2013, pg 293), despite some modelling differences (that analysis fixed $\sigma^2 = 0$ and used exponential priors for θ_1 and θ_2).

Time-variant infection parameter   We infer the log-transformed parameters $\vartheta = (\log\theta_{0,1}, \log\theta_2, \log\theta_3, \log\theta_4, \log\theta_5, \log\sigma^2)'$ under independent $N(0, 3^2)$ priors. Results are shown after transforming back to the original parameterisation.

Convergence now takes 3 hours, and 500,000 iterations of importance sampling produces an ESS of 256.1. Figure 5 shows estimates of the marginal parameter posteriors, using variational inference and importance sampling outputs as before. Figures 7 (SIR) and 8 (Ornstein-Uhlenbeck) show 50 diffusion paths sampled from the variational approximation. Again the low ESS indicates some approximation error.

Model comparison   The two models produce visually similar diffusion paths, but close inspection shows some differences. The time-invariant model paths for I_t appear smooth but slightly miss some of the observation points. The time-variant model paths for I_t are less smooth and more accurately capture the shape of the data. Correspondingly, the time-varying model infers a smaller σ^2 value. The most obvious difference in I_t paths occurs for the $t = 7, 8, 9$

observations, shown in the zoomed-in inset of Figures 6 and 7. Figure 8 shows that shortly before this the time-variant θ_1 values become constrained to smaller values.

Although the time-varying model appears to fit the data better, this is at the cost of increased model complexity, and could simply reflect overfitting. A better estimate of the parameter posteriors would allow formal model comparison based on importance sampling evidence estimates.

6. Conclusion

We provide a black-box variational approach to inference for SDES which is simple, practical and fast (relative to existing methods). This performs inference for a broad class of SDEs with minimal tuning requirements. Empirical investigation shows we obtain close matches to the posterior of the conditioned diffusion paths. Approximate parameter inference is also possible, with our results recovering known parameters for synthetic data (Section 5.1), and previous results for real data (Section 5.2), using only a few hours of computation for a desktop PC. An interesting future direction is develop choices of $q(x|\theta; \phi_x)$ more efficient than standard RNNs, to further reduce computing time and enable real-time applications of this methodology.
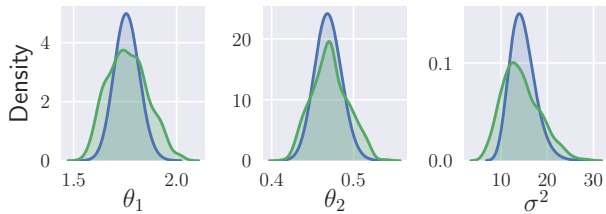
Figure 4. Marginal posterior density estimates from variational inference (blue) and importance sampling (green) for the SIR epidemic model with constant θ_1.
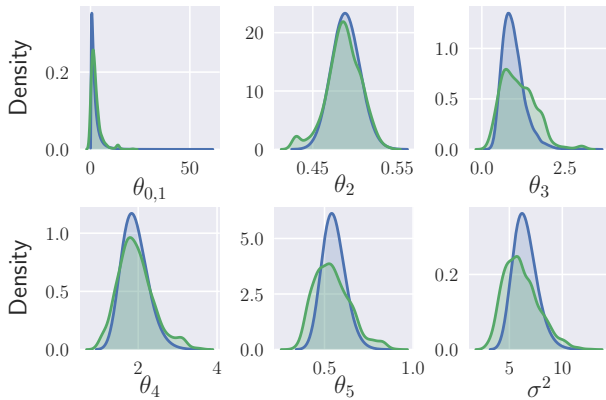
Figure 5. Marginal posterior density estimates from variational inference (blue) and importance sampling (green) for the SIR epidemic model with time varying θ_1.
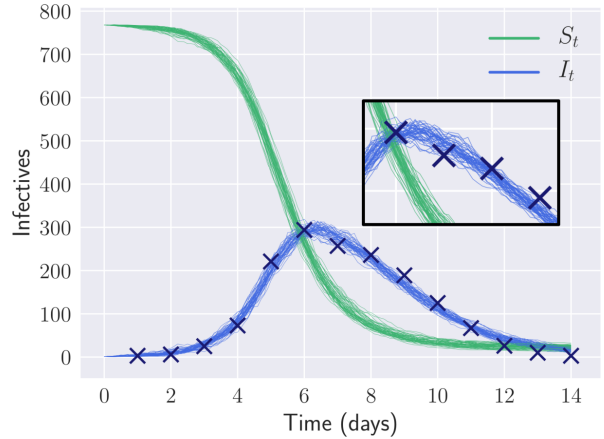
Figure 6. 50 conditioned diffusion paths for the SIR epidemic model with constant θ_1, sampled from the trained variational approximation. Observations are represented by dark blue crosses.
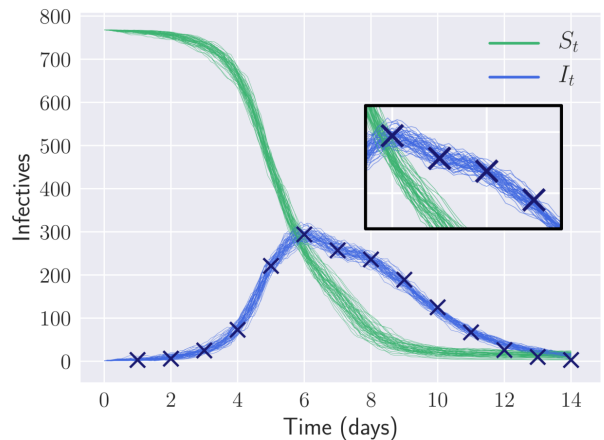
Figure 7. 50 conditioned diffusion paths for the SIR epidemic model with time varying θ_1, sampled from the trained variational approximation. Observations are represented by dark blue crosses.
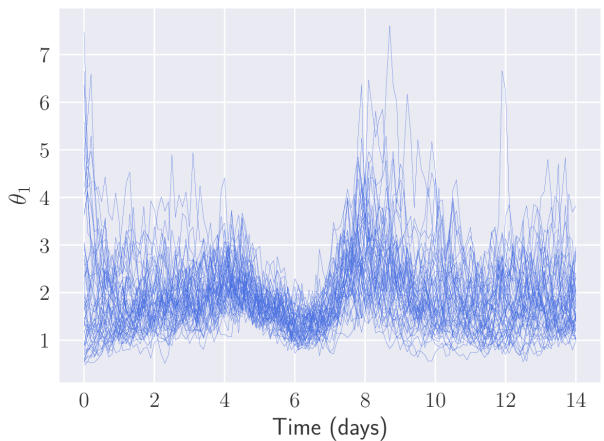
Figure 8. 50 conditioned diffusion paths for θ_1 for the SIR epidemic model, sampled from the trained variational approximation.

Acknowledgements

References

Andersson, H. and Britton, T. *Stochastic Epidemic Models and Their Statistical Analysis*. Springer-Verlag, 2000.

Archambeau, C., Opper, M., Shen, Y., Cornford, D., and Shawe-Taylor, J. S. Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems*, pp. 17–24, 2008.

Archer, E., Park, I. M., Buesing, L., Cunningham, J., and Paninski, L. Black box variational inference for state space models. In *International Conference on Learning Representations (ICLR) 2016, Workshops*, 2016.

Bengio, Y., Simard, P., and Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *Trans. Neur. Netw.*, 5(2):157–166, 1994.

Black, F. and Scholes, M. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3): 637–654, 1973.

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

Boys, R. J., Wilkinson, D. J., and Kirkwood, T. B. L. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18(2):125–135, 2008.

Dangerfield, C. E., Kay, D., and Burrage, K. Modeling ion channel dynamics through reflected stochastic differential equations. *Physical Review E*, 85(5):051907, 2012.

Del Moral, P. and Murray, L. M. Sequential Monte Carlo with highly informative observations. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):969–997, 2015.

Dureau, J., Kalogeropoulos, K., and Baguelin, M. Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics*, 14(3):541–555, 2013.

Eraker, B. MCMC analysis of diffusion models with application to finance. *Journal of Business & Economic Statistics*, 19(2):177–191, 2001.

Fuchs, C. *Inference for Diffusion Processes: With Applications in Life Sciences*. Springer Science & Business Media, 2013.

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

Golightly, A. and Wilkinson, D. J. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3): 1674–1693, 2008.

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

Jackson, C., Vynnycky, E., Hawker, J., Olowokure, B., and Mangtani, P. School closures and influenza: systematic review of epidemiological studies. *BMJ Open*, 3(2), 2013.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR) 2015*, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR) 2014*, 2014.

Liu, J. S. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119, 1996.

Moreno, A., Adel, T., Meeds, E., Rehg, J. M., and Welling, M. Automatic variational ABC. *arXiv preprint arXiv:1606.08549*, 2016.

Øksendal, B. *Stochastic Differential Equations: An Introduction with Applications*. Hochschultext / Universitext. Springer, 2003.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. Non-centred parameterisations for hierarchical models and data augmentation. In Bernardo, J. M., Bayarri, M. J., Berger, J., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 7*, pp. 307–326, 2003.

Papaspiliopoulos, O., Roberts, G. O., and Stramer, O. Data augmentation for diffusions. *Journal of Computational and Graphical Statistics*, 22(3):665–688, 2013.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International*

Conference on Machine Learning - Volume 28, ICML'13, pp. 1310–1318. JMLR.org, 2013.

Picchini, U. Inference for SDE models via approximate Bayesian computation. *Journal of Computational and Graphical Statistics*, 23(4):1080–1100, 2014.

Quiroz, M., Nott, D. J., and Kohn, R. Gaussian variational approximation for high-dimensional state space models. *arXiv preprint arXiv:1801.07873*, 2018.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 1530–1538, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Bejing, China, 2014. PMLR.

Robert, C. P. *Monte Carlo methods*. Wiley Online Library, 2004.

Roberts, G. O. and Stramer, O. On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, 88(3):603–621, 2001.

Rogers, L. C. G. and Williams, D. *Diffusions, Markov processes and martingales*. Cambridge University Press, 2013.

Ruttor, A., Sanguinetti, G., and Opper, M. Approximate inference for stochastic reaction processes. In Lawrence, N. D., Girolami, M., Rattray, M., and Sanguinetti, G. (eds.), *Learning and Inference in Computational Systems Biology*, pp. 189–205. MIT Press, 2010.

Skorokhod, A. V. Stochastic equations for diffusion processes in a bounded region. *Theory of Probability & Its Applications*, 6(3):264–274, 1961.

Titsias, M. and Lázaro-Gredilla, M. Doubly stochastic variational Bayes for non-conjugate inference. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1971–1979. JMLR Workshop and Conference Proceedings, 2014.

van der Meulen, F., Schauer, M., and van Zanten, H. Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals. *Electronic Journal of Statistics*, 11(1):2358–2396, 2017.

van Kampen, N. G. *Stochastic processes in physics and chemistry*. Elsevier, 2007.

Vehtari, A., Gelman, A., and Gabry, J. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2017.

Whitaker, G. A., Golightly, A., Boys, R. J., and Sherlock, C. Bayesian inference for diffusion-driven mixed-effects models. *Bayesian Analysis*, 12(2):435–463, 2017a.

Whitaker, G. A., Golightly, A., Boys, R. J., and Sherlock, C. Improved bridge constructs for stochastic differential equations. *Statistics and Computing*, 27(4):885–900, 2017b.