Notes on Latent Stochastic Differential Equations

Daniel Scharfstein, Charles Halberg

August 2023

1 Simplest Setting

1.1 Observed Data

Let $\{N(t) : 0 \leq t \leq T\}$ be the counting process for the assessment times. Let $\{Y(t) : 0 \leq t \leq T\}$ be the outcome process, which is only observed at the times when $N(\cdot)$ jumps. Let $W(t) = Y(t)$ if $dN(t) = 1$ and $W(t) = \emptyset$ if $dN(t) = 0$. Let $\overline{W}(t) = \{W(s) : 0 \leq s \leq t\}$.

1.2 Assumptions

Assumption 1: Prior of the latent process We assume the true latent process, denoted $Z_{true}(t)$, is a continuous time 1-dimensional stochastic process, whose dynamics are controlled by a stochastic differential equation (SDE) of the following form:

$$dZ_{true}(t) = \mu_{true}(Z(t), t)\, dt + \sigma_{true}(Z(t), t)\, dB(t), \tag{1}$$

where the drift, μ_{true}, maps to a scalar, the diffusion, σ_{true}, maps to scalar, and $B(t)$ is a 1-dimensional standard Brownian motion process.

It is widely known that given mild conditions on μ and σ, there exists a unique strong solution of (1) which is given by

$$Z(t) = Z(0) + \int_0^t \mu_{true}(Z(s), s)ds + \int_0^t \sigma_{true}(Z(s), s)dB(s) \quad \forall\, t \in [0, T],$$

where $Z(0)$ is the initial state of the process [**?**]. Note, the second integral is an Itô stochastic integral, which is defined as

$$\int_0^t \sigma(Z(s), s)dB(s) = \lim_{n\to\infty} \sum_{j=1}^{n} \sigma(Z(t_{j-1}), t_{j-1})\{B(t_j) - B(t_{j-1})\}$$

where $t_j = \frac{jt}{n}$.

Assumption 2: Outcome process generation The true outcome process $\{Y(t) \colon 0 \leq t \leq T\}$ is a Markov process given by an invertible, differentiable function of the latent process $Z(t)$ and a base process $O(t)$;

$$Y(t) = F(Z(t), O(t), t; \theta) \quad \forall t \in [0, T],$$

where θ represents the parameters of the function.

Since we do not assume to know the true value of the outcome process $Y(t)$ at all times $t \in [0, T]$, or the true latent process at any time, we also assume that the true function F is unknown. Because $W(t) = Y(t)$ whenever $dN(t) = 1$, it immediately follows that for $i \in \{1, 2, \ldots, n\}$,

$$W(t_i) = F(Z(t_i), O(t_i), t_i; \theta).$$

Assumption 3: Initial condition For simplicity, we consider the initial conditions to be samples from a standard normal Gaussian distribution:

$$Z(0), O(0) \sim N(0, \mathbf{I}).$$

1.3 Model

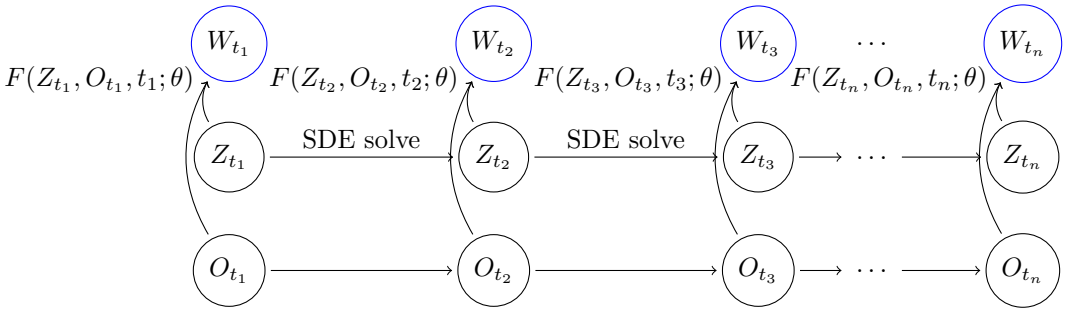

Figure 1: Diagram of the modeling process. The latent variable $Z(t)$ is generated by solving the SDE given in (2) at time t, and the base process $O(t)$ is similarly given by the Ornstein-Uhlenbeck (OU) process (6). In contrast to OU processes, a closed form solution does not exist for the $Z(t)$ process, and so we use numerical solvers to approximate the solution. At any time $t \in [0, T]$, the outcome process $Y(t)$ is a deterministic function of $Z(t)$ and $O(t)$.

We follow the work of [**?**] by modeling the latent $Z(t)$ process with a neural SDE. We then implement a continuously indexed normalizing flow as in [**?**] to decode the latent process $Z(t)$ to the outcome process $Y(t)$. Figure 1.3 is a graphical model of the modeling process for the observed values $\overline{W}(T)$.

Latent Dynamics We will begin by formulating our model for the latent process, $Z(t)$. For any $t \in (0, T]$, $Z(t)$ is given by the following SDE:

$$dZ(t) = \mu_{post}(Z(t), t; \phi_{post})\, dt + \sigma(Z(t), t; \rho) \circ dB(t), \tag{2}$$

where ϕ_{post} and ρ are parameters specifying the form of the drift and diffusion functions, respectively. We implement μ_{post} and σ as deep neural networks. For details on the exact implementation, see Appendix A. Recall that if μ_{post} and σ satisfy global Lipschitz conditions, then there exists a unique strong solution to (2) given by

$$Z(t) = Z(0) + \int_0^t \mu_{post}(Z(s), s; \phi_{post})\, ds + \int_0^t \sigma(Z(s), s; \rho) \circ dB(s),$$

Notice the value of $Z(t)$ at some time t is given by integrating over the interval $[0, t)$, making the current value of the state is implicitly dependent on the history of the process. However, since $Z(t)$ is a Markov process, it is sufficient to only condition on the most recent value of the process. That is, given arbitrary times $0 \leq t_j < t \leq T$, we have

$$\begin{aligned}
dZ(t)|\{Z(s) : 0 \leq s \leq t_j\} &= Z(0) + \int_0^t \mu_{post}(Z(s), s; \phi_{post})\, ds + \int_0^t \sigma(Z(s), s; \rho) \circ dB(s) \\
&= Z(0) + \int_0^{t_j} \mu_{post}(\cdot)\, ds + \int_0^{t_j} \sigma(\cdot) \circ dB(s) + \int_{t_j}^t \mu_{post}(\cdot)\, ds + \int_{t_j}^t \sigma(\cdot) \circ dB(s) \\
&= Z(t_j) + \int_{t_j}^t \mu_{post}(Z(s), s; \phi_{post}) ds + \int_{t_j}^t \sigma(Z(s), s; \rho) \circ dB(s) \\
&= dZ(t)|Z(t_j).
\end{aligned}$$

Thus, $P(Z(t)|\{Z(s) : 0 \leq s \leq t_j\}) = P(Z(t)|Z(t_j))$. For a single realization of the Brownian motion, the solution to (2) specifies a trajectory of the $Z(t)$ process through time. In general, an SDE defines a distribution over paths. Numerical methods, such as Monte Carlo simulations or discretization schemes like Euler-Maruyama, can be employed to approximate the distribution over paths and obtain a representative set of trajectories from the SDE.

Time-Dependent Decoding To connect the latent process to observable outcomes, we introduce a time dependent normalizing flow which serves as a decoder. That is, given the $Z(t)$ process and a *base process* $O(t)$ over the interval $[0, T]$, we model the outcome process at any $t \in [0, T]$ as

$$\hat{Y}(t) = F(O(t), Z(t), t; \theta), \tag{3}$$

where F is an invertible, differentiable function with parameters θ. Following the work of [**?**] we implement $F(\cdot)$ as an Augmented Neural ODE (ANODE), defined as the solution to the following initial value problem:

$$\frac{d}{d\tau}\begin{pmatrix} h(\tau) \\ a(\tau) \end{pmatrix} = \begin{pmatrix} f(h(\tau), a(\tau), \tau; \theta_f) \\ g(a(\tau), \tau; \theta_g) \end{pmatrix}, \quad \begin{pmatrix} h(\tau_0) \\ a(\tau_0) \end{pmatrix} = \begin{pmatrix} o(t) \\ (z(t), t)^T \end{pmatrix}, \tag{4}$$

where $o(t)$ and $z(t)$ are the values of the $O(t)$ and $Z(t)$ processes at time t, and f and g are deep neural networks. For exact implementations of f and g, see Appendix A. Then, for any time $t \in [0, T]$ we define

$$F(O(t), Z(t), t; \theta) := h(\tau_1) = h(\tau_0) + \int_{\tau_0}^{\tau_1} f(h(s), a(s), s; \theta_f)\, ds. \tag{5}$$

Note the distinction between t and τ: where t denotes the timestamp for the continuous process dynamics, τ indexes the independent variable in the decoding process given by (4) at any time t. For the sake of simplicity, we assume $\tau_0 = 0$ and $\tau_1 = 1$ for all t.

Lastly, for the base process let $O(t)$ be given by a standard Ornstein-Uhlenbeck (OU) process:

$$dO(t) = -O(t)\, dt + dB(t), \tag{6}$$

Assuming $O(0)$ is normally distributed, the transition density for any times $0 \le s < t \le T$ is given by:

$$P(O(t)|O(s)) = \frac{1}{\sqrt{2\pi \mathrm{Var}(t, s)}} \exp\left(-\frac{(O(t) - e^{-(t-s)}O(s))^2}{2\mathrm{Var}(t, s)} \right),$$

where $\mathrm{Var}(t - s) = \frac{1}{2}\left(1 - \exp(-2(t - s))\right)$ is the conditional variance of the process at time t given its value at time s. We choose the OU process as a base process for the normalizing flow decoder because it has bounded variance and offers a closed form transition density given a realization of the Brownian motion []. Unlike Brownian motion which does not have bounded variance over an arbitrary time interval, the OU process should not add to the variance of our estimate arbitrarily.

Likelihood Computation The main advantage of using the normalizing flow framework for decoding is that it enables simple distributions to be transformed into complex ones while retaining exact likelihood computation. Given f in (5) is uniformly Lipschitz continuous in $a(\tau)$ and $h(\tau)$ and continuous in τ, we can compute the instantaneous change of log likelihood of $W(t)$ using the change of variables formula [**?**]. For a single observation, the total change in log density is then given by:

$$\log P_{W(t)|Z(t),O(t)}(w(t)|z(t), o(t)) = \log P_{O(t)}(h(\tau_0)) - \int_{\tau_0}^{\tau_1} Tr\left(\frac{\partial f(h(s), a(s), s; \theta)}{\partial h(s)} \right) ds, \tag{7}$$

where $h(\tau_0) = F^{-1}(W(t), Z(t), t; \theta)$ is computed by solving the IVP in (4) backwards from τ_1 to τ_0 with the initial condition $h(\tau_1) = W(t)$. If the Jacobian is high dimensional, then computing the trace can be computationally expensive. In this case, the trace of the Jacobian can be estimated with Hutchinson's trace estimator [**?**]. Let $\epsilon \sim N(0, I)$ be a standard multivariate normal with the same dimension as the Jacobian matrix. Then Hutchinson's trace estimator is given by:

$$\begin{aligned}
\log P_{W(t)|Z(t),O(t)}(w(t)|z(t), o(t)) &= \log P_{O(t)}(h(\tau_0)) - \int_{\tau_0}^{\tau_1} Tr\left(\frac{\partial f(h(s), a(s), s; \theta)}{\partial h(s)} \right) ds \\
&= \log P_{O(t)}(h(\tau_0)) - \int_{\tau_0}^{\tau_1} \mathbb{E}_{P(\epsilon)}\left[\epsilon^T \frac{\partial f(h(s), a(s), s; \theta)}{\partial h(s)} \epsilon \right] ds \\
&= \log P_{O(t)}(h(\tau_0)) - \mathbb{E}_{P(\epsilon)}\left[\int_{\tau_0}^{\tau_1} \epsilon^T \frac{\partial f(h(s), a(s), s; \theta)}{\partial h(s)} \epsilon\, ds \right].
\end{aligned}$$

Context Encoding If observations $\{W(t_i), t_i\}_{i=1}^m$ exist before some time t^*, an RNN can encode this prior information for the approximate posterior (2). The RNN takes the sequence of observations with timestamps $\{W(t_i), t_i\}_{i=1}^m$ as input and outputs the parameters ϕ_{post}.

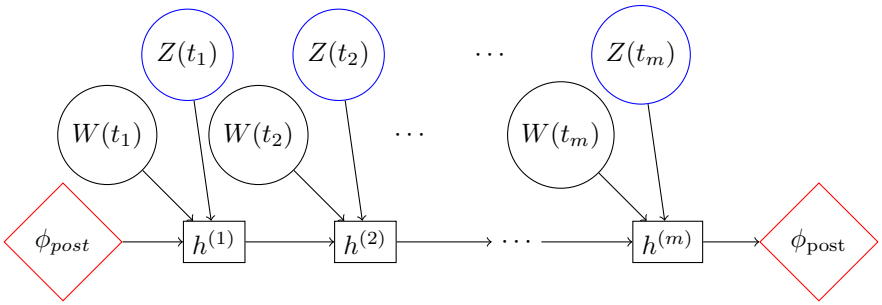

Figure 2: RNN architecture. Each layer $h^{(i)}$ of the RNN takes as input the observable values $W(t_i)$, value of the $Z(t_i)$ process as generated by the prior, and time t_i. The first layer $h^{(1)}$ takes as input the posterior drift parameters ϕ_{post} and the first values of the sequences. For each $h^{(i)}$ where $1 < i < m$, the output is the input to the next hidden layer $h^{(i+1)}$, and the last layer $h^{(m)}$ outputs the posterior drift parameters.

1.4 Inference

Our objective is to maximize the log likelihood of the observed data:

$$\mathcal{L} = \log P_{W(t_1),W(t_2),\ldots,W(t_n)}(w(t_1), w(t_2), \ldots, w(t_n)). \tag{8}$$

Given a sample of the Brownian motion over the interval $[0, T]$, denoted $b \sim B(T)$, the latent process $Z(t)$ is completely determined. Then using the fact that $W(t_i) = F(Z(t_i), O(t_i), t_i; \theta)$ for any t_i, we can expand the log-likelihood by conditioning on the latent processes:

$$\begin{aligned} \log P_{\overline{W}(T)}(w(t_1), w(t_2), \ldots, w(t_n)) &= \log \mathbb{E}_{b\sim B(T)} \left[P(\overline{w}(T) | z(t_1), z(t_2), \ldots, z(t_n)) \right] \\ &= \log \mathbb{E}_{b\sim B(T)} \left[\prod_{i=1}^{n} P_{W(t_i)|Z(t_i),O(t_i)}(w(t_i)|z(t_i), o(t_i)) \right]. \end{aligned} \tag{9}$$

Recall that $O(t)$ has closed form transition densities, so given $o(t_{i-1})$ we can determine the likelihood of $o(t_i)$. Since $o(t_{i-1}) = h(\tau_0) = F^{-1}(w(t_{i-1}), z(t_{i-1}), t_{i-1}; \theta)$, we can further simplify the likelihood using the continuous change of variables formula given in (7):

$$\mathcal{L} = \log \mathbb{E}_{b\sim B(T)} \left[\prod_{i=1}^{n} P_{O(t_i)|O(t_{i-1})}(o(t_i)|o(t_{i-1})) - \int_{\tau_0}^{\tau_1} Tr\left(\frac{\partial f(h(s), a(s), s; \theta)}{\partial h(s)} \right) ds \right]. \tag{10}$$

Variational Approximation Directly computing (10) is still intractable since the true posterior distribution is unknown. Furthermore, directly optimizing to maximize the log likelihood will lead to a collapse of the diffusion function, resulting in an over-fitted representation of uncertainty and poor generalization capabilities [**?**]. Instead, we use variational inference to approximate the posterior of the latent process $Z(t)$. Variational methods are useful for when the true posterior is difficult to compute because they only aim to approximate the true distribution with one from a tractable family. To do this, we introduce a prior on the $Z(t)$ process given by the following SDE:

$$dZ(t) = \mu_{prior}(Z(t), t; \phi_{prior}) + \sigma(Z(t), t; \rho) \circ dB(t). \tag{11}$$

Notice, the diffusion term σ is the same as that in (2). It was shown in [**?**] that the posterior and prior SDEs must share the same diffusion term, else their KL divergence will be infinite. We will see later that this KL divergence is an essential component of the Evidence Lower Bound (ELBO).

ELBO Derivation The goal of variational inference is to derive a lower bound on the log likelihood of the observable data, and maximize that lower bound with respect to the model parameters. Define

$$u(Z(t), t; \phi_{post}, \phi_{prior}, \rho) := \frac{\mu_{post}(Z(t), t; \phi_{post}) - \mu_{prior}(Z(t), t; \phi_{prior})}{\sigma(Z(t), t; \rho)},$$

and suppose u is a P measurable function which satisfies, for all $t \in [0, T]$,

$$\mathbb{E}_P \left[\exp \left(\frac{1}{2} \int_0^t \|u(Z(s), s)\|_2^2 \, ds \right) \right] < \infty.$$

For notational simplicity we will exclude the explicit dependence on parameters ϕ_{prior}, ϕ_{post}, and ρ. By Girsanov's second theorem, there exists a measure Q on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{0 \le t \le T})$ such that for any $t \in [0, T]$,

$$\frac{dQ}{dP} = \exp \left(- \int_0^t \frac{1}{2} \|u(Z(s), s)\|_2^2 \, ds - \int_0^t u(Z(s), s) \circ dB(s) \right).$$

Furthermore, Girsanov's second theorem shows the process $\hat{B}(t) = \int_0^t u(Z(s), s) \, ds + B(t)$ is a Brownian motion under the measure Q. This allows us to perform a change of measure such that the drift in the approximate posterior is replaced by the drift in the prior process:

$$\begin{aligned}
dZ(t) &= \mu_{post}(Z(t)t; \phi_{post}) \, dt + \sigma(Z(t), t; \rho) \circ dB(t) \\
&= \mu_{prior}(Z(t), t; \phi_{prior}) \, dt + \sigma(Z(t), t; \rho) \left(\frac{\mu_{post}(Z(t), t; \phi_{post}) - \mu_{prior}(Z(t), t; \phi_{prior})}{\sigma(Z(t), t; \rho)} \right) dt + \sigma(Z(t), t; \rho) \circ dB(t) \\
&= \mu_{prior}(Z(t), t; \phi_{prior}) \, dt + \sigma(Z(t), t; \rho) \circ d \left(B(t) + \int_0^t \frac{\mu_{post}(Z(s), s; \phi_{post}) - \mu_{prior}(Z(s), s; \phi_{prior})}{\sigma(Z(s), s; \rho)} \, ds \right) \\
&= \mu_{prior}(Z(t), t; \phi_{prior}) \, dt + \sigma(Z(t), t; \rho) \circ d\hat{B}(t).
\end{aligned}$$

Recall that for a realization of the Brownian motion, ω, both the prior and posterior SDEs are completely determined, so Q and P are distributions over paths in the latent space. For ease of notation, let $\mathbb{E}_Q[\cdot]$ denote the expectation induced under the approximate posterior, $Q(Z(\omega); \phi_{post}, \rho)$, and let $\mathbb{E}_P[\cdot]$ denote the expectation taken with respect to the prior $P(Z(\omega); \phi_{prior}, \rho)$. Then for any $t \in [0, T]$, let

$$M_t = \exp \left(- \int_0^t \frac{1}{2} \|u(Z(s), s)\|_2^2 \, ds - \int_0^t u(Z(s), s) \circ dB(s) \right),$$

and notice $dQ = M_t dP$. Then, by the expansion in (9), we may derive the ELBO as follows:

$$\begin{aligned}
\mathcal{L} = \log P(W(t_1), \ldots, W(t_n)) &= \log \mathbb{E}_P \left[\prod_{i=1}^n P(W(t_i) | Z(t_i), O(t_i)) \right] \\
&= \log \mathbb{E}_Q \left[\prod_{i=1}^n P(W(t_i) | Z(t_i), O(t_i)) M_{t_i} \right] \\
&\ge \mathbb{E}_Q \left[\sum_{i=1}^n \log P(W(t_i) | Z(t_i), O(t_i)) + \log M_{t_i} \right] \\
&= \mathbb{E}_Q \left[\sum_{i=1}^n \log P(W(t_i) | Z(t_i), O(t_i)) - \int_0^T \frac{1}{2} \|u(Z(t), t)\|_2^2 \, dt - \int_0^T u(Z(t), t) \circ dB(t) \right] \\
&= \mathbb{E}_Q \left[\sum_{i=1}^n \log P(W(t_i) | Z(t_i), O(t_i)) - \int_0^T \frac{1}{2} \|u(Z(t), t)\|_2^2 \, dt \right].
\end{aligned}$$

The second line follows from the identity $dQ = M_t dP$, the third follows from Jensen's inequality, and the last follows from the fact that the integral $\int_0^T u(Z(t), t) \circ dB(t)$ is a martingale. Substituting the expanded form of $P(W(t_i) | Z(t_i), O(t_i))$ from (10) we have

$$\begin{aligned}
\mathcal{L} &\ge \mathbb{E}_Q \left[\sum_{i=1}^n [\log P(W(t_i) | Z(t_i), O(t_i))] - \int_0^T \frac{1}{2} \|u(Z(s), s)\|_2^2 \, ds \right] \\
&= \mathbb{E}_Q \left[\sum_{i=1}^n \log P_{O(t_i) | O(t_{i-1})}(o(t_i) | o(t_{i-1})) - \int_{\tau_0}^{\tau_1} Tr \left(\frac{\partial f(h(s), a(s), s; \theta)}{\partial h(s)} \right) ds - \int_0^T \frac{1}{2} \|u(Z(s), s)\|_2^2 \, ds \right],
\end{aligned}$$

where

$$u(Z(t), t; \phi_{post}, \phi_{prior}, \rho) \coloneqq \frac{\mu_{post}(Z(t), t; \phi_{post}) - \mu_{prior}(Z(t), t; \phi_{prior})}{\sigma(Z(t), t; \rho)}.$$

Notice this expectation is now with respect to the approximate posterior, which is efficiently sampled using Monte-Carlo estimators. Since this is a lower bound on the log-likelihood, the objective to maximize is:

$$\mathcal{L}(\overline{W}(T); p) = \mathbb{E}_Q \left[\sum_{i=1}^{n} \log P_{O(t_i)|O(t_{i-1})}(o(t_i)|o(t_{i-1})) - \int_{\tau_0}^{\tau_1} Tr\left(\frac{\partial f(h(s), a(s), s; \theta)}{\partial h(s)} \right) ds - \int_0^T \frac{1}{2} \|u(Z(s), s)\|_2^2 \, ds \right], \tag{12}$$

where $o(t_i) = F^{-1}(W(t_i), Z(t_i); \theta)$ and $p = (\phi_{post}, \phi_{prior}, \rho, \theta_f, \theta_g)$ is a vector of all model parameters. Recall that the trace of the Jacobian can also be estimated using Hutchinson's trace estimator, where $\epsilon \sim P(\epsilon)$ is sampled outside of the integral.

1.5 Sampling and parameter optimization

Optimization To efficiently optimize (12) with respect to the model parameters p, we will use a variant of gradient ascent to iteratively update parameters until a convergence criteria is satisfied. Typical convergence criteria are termination after a pre-specified number of iterations through the dataset, or until the change in loss is less than a threshold value for a certain number of iterations. For any vector of parameters p on which the loss is dependent and a collection of observed data $\{\overline{W}^i(T)\}_{i=1}^N$, the gradient ascent update scheme is:

$$p^{(n+1)} = p^{(n)} + \eta \nabla_{p^{(n)}} \mathcal{L}(p^{(n)}; \overline{W}^i(T)),$$

where η is a hyper-parameter specifying the step size, or *learning rate* and $p^{(n)}$ denotes the parameters on the n_{th} iteration. For large datasets, it is common to use more efficient variants of gradient descent such as Stochastic Gradient Descent (SGD) or ADAM [**?**]. In stochastic gradient descent/ascent, the gradient term is replaced by an unbiased estimate of the gradient obtained from multiple individuals in the dataset.

This allows information from a batch individuals in the dataset to be used in a single training iteration, and in practice this results in faster convergence early in training. The difficulty of using this optimization scheme is in computing gradients of loss (or unbiased estimates of gradients) with respect to the parameter vectors. In continuous time models, gradient computation is less straightforward than in the discrete case, so it is covered in considerable detail in the following section.

Stochastic Adjoint Sensitivity The model we have described so far is formulated in continuous time, introducing complexities in the optimization process that are not seen in the discrete case. In discrete time models such as Recurrent Neural Networks, derivatives of loss with respect to model parameters are typically obtained by the back propagation algorithm. During backpropagation, the gradients of the model's parameters with respect to the loss function are calculated by decomposing the overall gradient with respect to the output into gradients of individual operations and layers and applying the chain rule. In continuous-time models, the concept of a discrete computational graph may not be directly applicable, as the state variable $Z(t)$ evolves continuously over time rather than in discrete steps.

Instead, we turn to a continuous time analog of back-propagation known as the *Adjoint Sensitivity Method*. The adjoint method computes gradients by defining a differential equation of the gradient with respect to time, known as the *adjoint equation*. The gradient of loss with respect to model parameters is then obtained by solving an augmented adjoint equation backwards in time from the final state to the initial state. This method for gradient computation was first applied to neural ordinary differential equations in [**?**], and a stochastic analog was later introduced in [**?**].

For any $t \in (0, T]$, define

$$A_{0,t}(z_0) = \nabla_{Z(t)} \mathcal{L}(z_0) = \frac{\partial \mathcal{L}(\Phi_{0,t}(z_0))}{\partial Z(t)}$$

to be the adjoint of the $Z(t)$ process, where $\Phi_{0,t}(z_0)$ is the unique strong solution of (2) at time t given the process started at time 0 with initial condition z_0. Similarly, define $\check{\Psi}_{0,t}(z_0) = \Phi_{0,t}^{-1}(z_0)$ to be the *backward*

flow of the $Z(t)$ process which satisfies

$$\check{\Psi}_{0,t}(z_0) = z_t - \int_0^t \mu_{post}(\check{\Psi}_{u,t}(z), u; \phi_{post})\, du - \int_0^t \sigma(\check{\Psi}_{u,t}(z), u; \rho) \circ d\check{B}(u).$$

The backwards adjoint variable is then defined as the forward adjoint variable composed with the backwards flow of the $Z(t)$ process, $\check{A}_{0,t}(z_0) \coloneqq A_{0,t}(\check{\Psi}_{0,t}(z_0))$. The backwards adjoint and backwards flow then satisfy the following system of equations [**?**]:

$$\begin{aligned}
\check{A}_{0,t}(z_0) &= \nabla_{Z(t)}\mathcal{L}(z_0) + \int_0^t \check{A}_{r,t}(z_0)\nabla_{Z(t)}\mu_{post}(\check{\Psi}_{r,t}, r; \phi_{post})\, dr + \int_0^t \check{A}_{r,t}(z_0)\nabla_{Z(t)}\sigma(\check{\Psi}_{r,t}, r; \rho) \circ d\check{B}(r) \\
\check{\Psi}_{0,t}(z_0) &= z_t - \int_0^t \mu_{post}(\check{\Psi}_{r,t}(z_0), r)\, dr - \int_0^t \sigma(\check{\Psi}_{r,t}(z_0), r) \circ d\check{B}(r).
\end{aligned} \tag{13}$$

For a more detailed description of the stochastic adjoint sensitivity method and reverse stochastic flows, see appendix C.

Gradient Computation With the adjoint equations (13) defined, it is simple to extend this to compute the gradient of loss with respect to model parameters. Let $p = [\phi_{post}\, \phi_{prior}\, \rho\, \theta_f\, \theta_g]^T$ be a vector of the concatenated model parameters. Additionally, to compute the ELBO given by (12) on the forward pass, we define $U(t) \coloneqq \int_0^t \frac{1}{2}\|u(Z(s), s)\|_2^2\, ds$ to be a variable representing the the KL-divergence term. Then we define the augmented variable $L(t) = (Z(t), p, U(t))$ which satisfies the following SDE:

$$dL(t) = \begin{pmatrix} \mu_{post}(Z(t), t; \phi_{post}) \\ \mathbf{0} \\ \frac{1}{2}\|u(Z(t), t)\|_2^2 \end{pmatrix} dt + \begin{pmatrix} \sigma(Z(t), t; \rho) \\ \mathbf{0} \\ 0 \end{pmatrix} \circ dB(t),$$

where $B(t)$ is a Brownian motion of appropriate dimension. By (13) the backward SDEs of the adjoint processes are given by the following:

$$\begin{aligned}
\check{A}^z(t) &= \check{A}^z(T) + \int_t^T \left(\check{A}^z(s)\frac{\partial \mu_{post}(\check{Z}(s), s; \phi_{post})}{\partial \check{Z}(s)} + \frac{1}{2}\check{A}^l(s)\frac{\partial ||u(\check{Z}(s), s)||_2^2}{\partial \check{Z}(s)} \right) ds + \int_t^T \check{A}^z(s)\frac{\partial \sigma(\check{Z}(s), s; \rho)}{\partial \check{Z}(s)} \circ d\check{B}(s) \\
\check{A}^p(t) &= \check{A}^p(T) + \int_t^T \left(\check{A}^z(s)\frac{\partial \mu_{post}(\check{Z}(s), s; \phi_{post})}{\partial p} + \frac{1}{2}\check{A}^l(s)\frac{\partial ||u(\check{Z}(s), s)||_2^2}{\partial p} \right) ds + \int_t^T \check{A}^z(s)\frac{\partial \sigma(\check{Z}(s), s; \rho)}{\partial p} \circ d\check{B}(s). \\
\check{A}^l(t) &= \check{A}^l(T).
\end{aligned} \tag{14}$$

Lastly, by solving (13) backwards from T to 0, we get the desired gradient,

$$\check{A}^p(0) = \frac{\partial \mathcal{L}(\Phi_{0,T}(\Psi_{0,T}(z_0)))}{\partial p} = \frac{\partial \mathcal{L}(z_0)}{\partial p}.$$

A Neural Network Implementation

B Backward Stochastic Flows

To formalize a backwards SDE, we first recall the basic definitions necessary for understanding SDEs.

Definition: Filtration
Given a probability space $(\Omega, \mathcal{F}, P)$ and stochastic process $\{X(t)\}_{t\geq 0}$, a filtration $\underline{\mathcal{F}}$ is a family of information sets, $\mathcal{F}_t = \sigma\{X_s \colon s \in [0,t]\}$, indexed by time, $\{\mathcal{F}_t \colon t \geq 0\}$, which is right continuous

$$\mathcal{F}_t = \mathcal{F}_{t+} = \bigcap_{\epsilon \geq 0} \mathcal{F}_{t+\epsilon}, \quad \forall t \geq 0.$$

We say a stochastic process $X(t)$ is *adapted* to a filtration $\underline{\mathcal{F}}$ if for all $t \geq 0$, $X(t)$ is $\mathcal{F}_t$-measurable. Filtration formalizes the idea of evolving information over time, and is a necessary component in defining a Brownian motion.

Definition: Brownian Motion
Given a probability space $(\Omega, \mathcal{F}, \underline{\mathcal{F}}, P)$, a standard Brownian motion $B = \{B(t) \colon t \geq 0\}$ with variance parameter σ^2 is a stochastic process adapted to a filtration $\underline{\mathcal{F}}$ which satisfies:
(1) $B(0) = 0$;
(2) The random increments $(B(t_{j+1}) - B(t_j))$, where $j \in \{1, \ldots, n-1\}$ are independent for all sequences of time stamps $t_1, t_2, \ldots, t_n \in [0, \infty)$;
(3) $\forall s \leq t$, the random variable $(B(t) - B(s))$ has a $\mathcal{N}(0, (t-s)\sigma^2)$ distribution;
(4) the sample paths are continuous, meaning the mapping $t \mapsto B(t)$ is continuous.

The definition of a standard multidimensional Brownian motion follows easily by requiring that components are pairwise independent scalar standard Brownian motions. With this, we may define an SDE using the Itô stochastic integral.

Definition: Itô SDE
A stochastic process $\{Z(t)\}_{t\in[0,T]}$ can be defined by a stochastic differential equation of the form

$$Z(t) = z_0 + \int_0^t f(Z_s, s)\, ds + \int_0^t \sigma(Z_s, s)\, dB(s),$$

where z_0 is an initial condition. The second integral is an Itô stochastic integral, defined

$$\int_0^t \sigma(Z_s, s)\, dB(t) \coloneqq \lim_{|\Pi|\to 0} \sum_{k=1}^{N} \sigma(Z(t_{k-1}), t_{k-1})(B(t_k) - B(t_{k-1})),$$

where $|\Pi| = \max_k(t_k - t_{k-1})$ for any partition $0 = t_1, t_2, \ldots, t_N = T$ of the time interval $[0, T]$.

From these definitions, it is clear that the usual formulation of an SDE is reliant on a filtration defined forwards in time. Fortunately, we can define a two-sided filtration which will allow for Brownian motions to be well defined both forwards and backwards in time on the same probability space.

Definition: Two-sided filtration
Given a probability space $(\Omega, \mathcal{F}, P)$ and stochastic process $\{X(t)\}_{t\geq 0}$, the two-sided filtration $\underline{\mathcal{F}} = \{\mathcal{F}_{s,t} \colon 0 \leq s \leq t\}$ is a right continuous family of information sets

$$\mathcal{F}_{s,t} = \sigma\{X_v - X_u \colon s \leq u \leq v \leq t\} \quad \forall\, 0 \leq s \leq t < \infty.$$

In particular, if we consider the time interval $[0, T]$ we can define the *forward filtration* to be

$$\{\mathcal{F}_{0,t} \colon t \in [0,T]\},$$

and the *backward filtration* to be

$$\{\mathcal{F}_{t,T} \colon t \in [0,T]\}.$$

With a forward and backward filtration formalized, we can define a backward Brownian motion. This also allows us to rigorously define both forward and backward SDEs on the same probability space. Since we ultimately wish to define a backward SDE for the adjoint process with respect to the forward latent process, it is more useful to use the Stratonovich stochastic integral because of its symmetry with respect to time.

Definition: Forward and Backward Stratonovich Integrals
Let the process $\{\sigma(Z(t),t) \colon t \in [0,T]\}$ be adapted to the forward filtration $\{\mathcal{F}_{0,t} \colon t \in [0,T]\}$, and let B be a Brownian motion adapted to the forward filtration. The *forward Stratonovich stochastic integral* is

$$\int_0^t \sigma(Z(s),s) \circ dB(s) \coloneqq \lim_{|\Pi|\to 0} \sum_{k=1}^{N} \frac{\sigma(Z(t_{k-1}),t_{k-1}) + \sigma(Z(t_k),t_k)}{2}(B(t_k) - B(t_{k-1})),$$

where $|\Pi| = \max_k(t_k - t_{k-1})$ for any partition $0 = t_1, t_2, \ldots, t_N = T$ of the time interval $[0,T]$.

Now define the *backward Brownian motion* $\check{B}(t) \coloneqq B(t) - B(T)$ for all $t \in [0,T]$ to be adapted to the backward filtration $\{\mathcal{F}_{t,T} \colon t \in [0,T]\}$. Let the process $\{\check{\sigma}(Z(t),t) \colon t \in [0,T]\}$ be adapted to the backward filtration, then the *backward Stratonovich integral* is

$$\int_t^T \check{\sigma}(Z(s),s) \circ d\check{B}(s) \coloneqq \lim_{|\Pi|\to 0} \sum_{k=1}^{N} \frac{\check{\sigma}(Z(t_{k-1}),t_{k-1}) + \check{\sigma}(Z(t_k),t_k)}{2}(\check{B}(t_k) - \check{B}(t_{k-1})).$$

The formulation of a Stratonovich SDE differs from an Itô SDE only in which stochastic integral is being taken. Also notice that a Stratonovich SDE takes the average of an interval as the approximation point, giving it a natural symmetry with respect to time. This symmetry is useful for our application because it simplifies the form of the "reverse" process $\check{\sigma}$ that almost surely reconstructs the path of the forward process σ.

C Stochastic Adjoint Sensitivity Method

The stochastic adjoint sensitivity method provides an efficient method for computing gradients $\nabla_{\phi_{prior}}\mathcal{L}$ and $\nabla_{\phi}\mathcal{L}$ for a parametric SDE by defining and solving an adjoint state SDE. Generalizing the adjoint method to the stochastic setting poses additional challenges;

- The adjoint method requires defining differential equations forward and backward in time. Since SDEs are defined using a Brownian motion adapted to a filtration forwards in time, it is not immediately clear how to define a "backwards" SDE.
- In ordinary differential equations, the end state of the forward dynamics is deterministic. However, in SDEs the end state is dependent on the realized Brownian motion path, so the stochastic adjoint method must generalize to handle a stochastic end state.

Backwards Stratonovich SDE reconstructs forward trajectory
In this part we will show the advantage of formalizing SDEs using the Stratonovich integral. The main advantage is that the form of the backward SDE which almost surely reconstructs the original trajectory is given by simply negating the drift and diffusion of the forward SDE and taking the Stratonovich integral with respect to the backward Brownian motion $\{\check{B}(t) \colon t \in [0,T]\}$. Using the fact that if the drift and diffusion functions of an SDE satisfy mild differentiability conditions then there exists a unique strong solution, the backwards process which reconstructs the forward trajectory is given in the following theorem.

Theorem 1
Let $\Phi_{s,t}(z)$ be the unique strong solution of the following SDE at time t, given the process started at the state z_s at time s:

$$\Phi_{s,t}(z) \coloneqq z_s + \int_s^t f(\Phi_{u,t}(z),u)\,du + \int_s^t \sigma(\Phi_{u,t}(z),u) \circ dB(u). \tag{15}$$

Then the backward flow $\check{\Psi}_{s,t}(z) \coloneqq \Phi_{s,t}^{-1}(z)$ satisfies

$$\check{\Psi}_{s,t}(z) = z_t - \int_s^t f(\check{\Psi}_{u,t}(z), u)\, du - \int_s^t \sigma(\check{\Psi}_{u,t}(z), u) \circ d\check{B}(u), \tag{16}$$

for all initial conditions $z \in \mathbb{R}$, and for all $s, t \in [0, T]$ such that $s \leq t$.

Now that we have well defined SDEs forward and backward in time, and we know the form of the backward process which reconstructs the forward trajectory, we can derive an SDE for the adjoint process $\frac{\partial \mathcal{L}}{\partial Z(t)}$. From this adjoint process, it is simple to define an augmented adjoint process, the solution of which will yield the desired gradients $\frac{\partial \mathcal{L}}{\partial \phi_{prior}}$.

Deriving the stochastic adjoint process
We will proceed in three steps:

1. Derive a backward SDE for the adjoint process $\left\{ \frac{\partial Z(T)}{\partial Z(t)} : t \in [0, T] \right\}$, assuming $Z(t) = \check{\Psi}_{t,T}(Z(T))$ follows the inverse flow (16) from a deterministic endpoint $Z(T)$.
2. Extend to consider an endpoint obtained from the forward flow (15): $Z(T) = \Phi_{0,T}(z_0)$.
3. Augment the $Z(t)$ process and its adjoint to include processes for the parameters of the drift and diffusion functions of the approximate posterior and prior.

The following lemma addresses (1) by providing an SDE for the Jacobian matrix of the backwards flow.

Lemma (Dynamics of $\frac{\partial Z(T)}{\partial Z(t)}$)
Consider the backwards flow, $\check{\Psi}_{s,t}(z)$, generated by (16). Define $J_{s,t}(z) = \nabla_{Z(t)} \check{\Psi}_{s,t}(z)$. Then we have

$$J_{s,t}(z) = I_d - \int_s^t \nabla_{Z(t)} f(\check{\Psi}_{r,t}(z), r) J_{r,t}(z)\, dr - \int_s^t \nabla_{Z(t)} \sigma(\check{\Psi}_{r,t}(z), r) J_{r,t}(z) \circ d\check{B}(r). \tag{17}$$

Furthermore, if we define the inverse of the Jacobian $K_{s,t}(z) = [J_{s,t}(z)]^{-1}$, then we have

$$K_{s,t}(z) = I_d + \int_s^t K_{r,t}(z) \nabla_{Z(t)} f(\check{\Psi}_{r,t}(z), r)\, dr + \int_s^t K_{r,t}(z) \nabla_{Z(t)} \sigma(\check{\Psi}_{r,t}(z), r) \circ d\check{B}(r). \tag{18}$$

This lemma follows from the definitions of $J_{s,t}(z)$ and $K_{s,t}(z)$ and Itô's formula. To consider when the endpoint is the result of the forward flow, we compose the state process and loss function to define the adjoint state process:

$$A_{s,t}(z_0) \coloneqq \frac{\partial \mathcal{L}(\Phi_{s,t}(z_0))}{\partial Z(t)}, \tag{19}$$

and notice the chain rule gives

$$A_{s,t}(z_0) = \nabla_{Z(t)} \mathcal{L}(\Phi_{s,t}(z_0)) \nabla_{Z(t)} \Phi_{s,t}(z_0).$$

For the backwards adjoint process, define $\check{A}_{s,t}(z_0) \coloneqq A_{s,t}(\check{\Psi}_{s,t}(z_0))$. Note

$$\begin{aligned} \check{A}_{s,t}(z_0) &= A_{s,t}(\check{\Psi}_{s,t}(z_0)) && (20)\\ &= \nabla \mathcal{L}(z_0) \nabla \Phi_{s,t}(\check{\Psi}_{s,t}(z_0)) && (21)\\ &= \nabla \mathcal{L}(z_0) K_{s,t}(z_0). && (22) \end{aligned}$$

This also implies that the forward adjoint process is equivalent to the backward adjoint process composed with the forward flow:

$$A_{s,t}(z_0) = \check{A}_{s,t}(\Phi_{s,t}(z_0)). \tag{23}$$

Since $\mathcal{L}(z_0)$ is a constant, we can now derive a system of SDEs for the backward adjoint process.

Theorem 2
The backwards adjoint process $\check{A}_{s,t}(z_0)$ given by (20) and the backwards flow $\check{\Psi}_{s,t}(z_0)$ given by (16) satisfy the following system of SDEs:

$$\begin{aligned}\check{A}_{s,t}(z_0) &= \nabla\mathcal{L}(z_0) + \int_s^t \check{A}_{r,t}(z_0)\nabla f(\check{\Psi}_{r,t}, r)\, dr + \int_s^t \check{A}_{r,t}(z_0)\nabla\sigma(\check{\Psi}_{r,t}, r) \circ d\check{B}(r) \\ \check{\Psi}_{s,t}(z_0) &= z_0 - \int_s^t f(\check{\Psi}_{r,t}(z_0), r)\, dr - \int_s^t \sigma(\check{\Psi}_{r,t}(z_0), r) \circ d\check{B}(r),\end{aligned} \tag{24}$$

Since we already assumed the drift and diffusion functions satisfy the conditions to guarantee a unique strong solution, the system given in (24) has a unique strong solution given a realization of the Brownian motion from the forward flow. Let $B = \{B(t) \colon t \in [0, T]\}$ be a realized trajectory of the Brownian motion from the forward flow (15), then the solution to (24) on the time interval $[0, T]$, given B and initial condition z_0, is a function

$$\check{A}_{0,T}(z_0) = F(z_0, B),$$

which computes the value of the backward adjoint process given the initial condition z_0 and Brownian motion path B. If we define $G(z, B) = \Phi_{0,T}(z_0)$ to be the solution to the forward SDE over $[0, T]$, then we have the following relation.

Theorem (Backwards adjoint process with non-deterministic end state)
For P-almost all $\omega \in \Omega$, we have

$$A_{0,T}(z_0) = \check{A}_{0,T}(G(z_0, B)) = F(G(z_0, B), B).$$

In other words, this shows we can get the value of the adjoint process by composing the solution to the backwards adjoint process (24) with the original forward SDE. Thus, the forward and backward adjoint processes for stochastic differential equations are well defined, and computable using efficient numerical solvers.