Efficient Training of Neural SDEs Using Stochastic Optimal Control

Rembert Daems [1,2], Manfred Opper [3,4,5],
Guillaume Crevecoeur [1,2] and Tolga Birdal [6] *

1- D2Lab, Ghent University, Belgium

2- FlandersMake@UGent – corelab MIRO, Belgium

3- Dept. of Theor. Comp. Science, Technical University of Berlin, Germany

4- Inst. of Mathematics, University of Potsdam, Germany

5- Centre for Sys. Modelling and Quant. Biomed., University of Birmingham, UK

6- Dept. of Computing, Imperial College London, UK

Abstract. We present a hierarchical, control theory inspired method for variational inference (VI) for neural stochastic differential equations (SDEs). While VI for neural SDEs is a promising avenue for uncertainty-aware reasoning in time-series, it is computationally challenging due to the iterative nature of maximizing the ELBO. In this work, we propose to decompose the control term into linear and residual non-linear components and derive an optimal control term for linear SDEs, using stochastic optimal control. Modeling the non-linear component by a neural network, we show how to efficiently train neural SDEs without sacrificing their expressive power. Since the linear part of the control term is optimal and does not need to be learned, the training is initialized at a lower cost and we observe faster convergence.

1 Introduction

Continuous-time models of dynamical systems provide a powerful framework for capturing the intricate variations in real-world phenomena. Among these, stochastic differential equations (SDEs) extend the capabilities of deterministic models by abstracting away unaccounted factors into instantaneous noise. SDEs naturally model various processes, including the motion of small particles (e.g., molecules) and financial market dynamics. When combined with neural networks [1, 2], they become expressive tools for learning from irregular time-series observations.

Despite their promise, path-wise inference for neural SDEs remains a notorious challenge due to the complexity in fitting the non-Gaussian posterior distributions. Variational inference (VI) has become a prevalent tool with significant success in scaling inference methods [3]. Yet, computational challenges persist.

*MO acknowledges funding by Deutsche Forschungsgemeinschaft (DFG)-SFB1294/ 1-318763901. This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. Furthermore it was supported by Flanders Make under the SBO project CADAIVISION. TB was supported by a UKRI Future Leaders Fellowship [grant number MR/Y018818/1].

Existing works attempt to address these issues in VI for neural SDEs in various ways. Park *et al.* [4] introduced finite-dimensional matching for efficient path comparison to train neural SDEs. Kidger *et al.* [5] adopted a generative-adversarial approach to train these models. Course and Nair [6] proposed an amortized method for fast VI in latent neural SDEs, scaling efficiently with data size using a linear posterior. However, resorting to linear posteriors is a severe limitation in practice.

Inspired by *optimal control theory*, we propose a novel approach to efficiently perform VI in neural SDEs. Our key idea is to represent the prior as the combination of a linear model and a residual non-linear model. We leverage this decomposition to split the control function—used to compute the variational posterior—into two components. The first linear component is tractable and admits a closed-form solution, making it computationally efficient but less expressive. The residual non-linear component, modeled by a neural network, captures higher-order effects at the cost of iterative optimization. We combine the strengths of these two approaches. First, we compute the linear part in closed form, which serves as an efficient initialization for the neural network modeling the non-linear residual. This hierarchical design allows us to achieve faster and more stable inference compared to existing approaches that directly model the full control term [2, 3].

In summary, our contributions are:

1. We derive the optimal control function solution for VI of a linear SDE driven by Brownian motion (BM), or by Markov–approximated fractional BM.
2. We propose a neural SDE model with a linear and a residual non-linear (neural network) part, both for the prior SDE and the control terms, for which the linear part is optimal and does not need to be optimized or learned.
3. We show that our proposed model trains faster and more stable than a standard non-linear network model on a financial data.

We will make our implementation publicly available upon publication.

2 Variational Inference of Stochastic Differential Equations

Definition 1 (SDEs driven by BM (BMSDE))**.** *A common generative model for stochastic dynamical systems considers a set of observational data* $\mathcal{D} = \{O_1, \ldots, O_M\}$*, where the* O_i *are generated (conditionally) independent at random at discrete times* t_i *with a likelihood* $p_\theta(O_i \mid X(t_i))$*. The prior information about the unobserved path* $\{X(t); t \in [0,\ T]\}$ *of the latent process* $X(t) \in \mathbb{R}^D$ *is given by the assumption that* $X(t)$ *fulfils the SDE:*

$$\mathrm{d}X(t) = b_\theta(X(t), t)\,\mathrm{d}t + \sigma_\theta(X(t), t)\,\mathrm{d}B(t) \qquad \text{(PRIOR-SDE)}$$

The drift function $b_\theta(X(t), t) \in \mathbb{R}^D$ *models the deterministic part of the change* $\mathrm{d}X(t)$ *of the state variable* $X(t)$ *during the infinitesimal time interval* $\mathrm{d}t$*, whereas the diffusion matrix* $\sigma_\theta(X(t), t) \in \mathbb{R}^{D\times B}$ *encodes the strength of the added Gaus-*

sian white noise process, where $\mathrm{d}B(t) \sim \mathcal{N}(0, \mathrm{d}t) \in \mathbb{R}^B$ *is the infinitesimal increment of a vector of independent Wiener processes during* $\mathrm{d}t$.

Definition 2 (Posterior SDE)**.** *The paths of the* Prior*-SDE can be steered by adding a control term* $u(X(t), t)$ *that depends on all variables to be optimised and the observations, to the drift resulting in the variational posterior [2, 7]:*

$$\mathrm{d}\tilde{X}(t) = \left(b_\theta\left(\tilde{X}(t), t\right) + \sigma_\theta\left(\tilde{X}(t), t\right) u\left(\tilde{X}(t), t\right)\right)\mathrm{d}t + \sigma_\theta\left(\tilde{X}(t), t\right)\mathrm{d}B(t) \quad (1)$$

In what follows, we will assume a parametric form for the control function $u(\tilde{X}(t), t) \equiv u_\phi(\tilde{X}(t), t)$ and will recall a scheme for inferring the *variational parameters* (θ, ϕ), *i.e.* variational inference.

Proposition 1 (Variational Inference for BMSDE [2, 7])**.** *The variational parameters* ϕ *are optimised by minimising the KL–divergence between the posterior and the prior, where the corresponding evidence lower bound (ELBO) is maximized to find the most likely parameters* θ*:*

$$\sum_{i=1}^{M} \log p\left(O_i \mid \theta\right) \geq \mathbb{E}_{\tilde{X}}\left[\sum_{i=1}^{M} \log p_\theta\left(O_i \mid \tilde{X}(t_i)\right) - \int_0^T \frac{1}{2}\left\|u_\phi\left(\tilde{X}(t), t\right)\right\|^2 \mathrm{d}t\right] \quad (2)$$

where the observations $\{O_i\}$ *are included by likelihoods* $p_\theta\left(O_i \mid \tilde{X}(t_i)\right)$ *and the expectation is taken over random paths of the approximate posterior process defined by (eq.* (1)*).*

3 Optimal Control for Variational Inference for SDEs

Our approach uses optimal control to decouple the possible linear and non-linear effects in the drift. While the linear part is easier to solve in closed-form, the non-linear terms will account for the complex variations in real data. In the sequel, we describe these two parts, respectively, finally leveraging the strengths of both.

3.1 Optimal posterior control term for a linear prior SDE

The control term $u(x, t) := u_\phi(x, t)$ can be obtained explicitly from the solution of the transformed *Hamilton–Jacobi–Bellman* equation (HJBE) [8–10]:

$$u(x, t) = \sigma_\theta(x, t)^\top \nabla_x \log \mathbb{E}_{\text{prior}}\left[\prod_{i:t_i > t} p_\theta\left(O_i \mid X(t_i)\right) | X(t) = x\right]. \quad (3)$$

In general, such expectations over the paths of the Prior-SDE involve solving second order partial differential equations in the $D + 1$ variables and are intractable in closed form. However, in what follows, we will show how to compute it exactly when both the prior process $X(t)$ and the observation likelihood are Gaussian. This requires the drift $b_\theta(x, t)$ to be a linear function in x, and the diffusion $\sigma_\theta(t)$ independent of x.

Proposition 2. *For a process $X(t)$ with linear drift and state-independant diffusion $\sigma(t)$ where we have M observations $\boldsymbol{O} = [O(T_1), \ldots, O(T_M)]$ after time t, the optimal control term takes the form:*

$$u(x,t) = \sigma(t)^\top \nabla_x \log \mathcal{N}(\boldsymbol{O}; \boldsymbol{m}_x, \boldsymbol{C} + \boldsymbol{\Sigma}_0) \tag{4}$$

$$= \sigma(t)^\top \left(\nabla_x \boldsymbol{m}_x\right)^\top \left(\boldsymbol{C} + \boldsymbol{\Sigma}_0\right)^{-1} \left(\boldsymbol{O} - \boldsymbol{m}_x\right), \tag{5}$$

where $p(\boldsymbol{X}(\boldsymbol{T})|x) = \mathcal{N}(\boldsymbol{m}_x, \boldsymbol{C})$ is the joint Gaussian distribution of the solutions of the prior SDE $\boldsymbol{X}(\boldsymbol{T}) = [X(T_1), \ldots, X(T_M)]$ conditioned on $X(t) = x$ having mean vector by $\boldsymbol{m}_x$ and covariance matrix by $\boldsymbol{C}$. The observation likelihood is assumed to be of the form $\mathcal{N}(\boldsymbol{O}; 0, \boldsymbol{\Sigma}_0)$.

Sketch of the proof. Under these assumptions, the expectation in eq. (3) for $X(t)$ becomes an M dimensional Gaussian integral of the form:

$$\mathbb{E}_{\text{prior}}\left[\ldots\right] = \int p(\boldsymbol{X}(\boldsymbol{T})|x) p(\boldsymbol{O}|\boldsymbol{X}(\boldsymbol{T}))\, \mathrm{d}\boldsymbol{X}(\boldsymbol{T}) = \mathcal{N}(\boldsymbol{O}; \boldsymbol{m}_x, \boldsymbol{C} + \boldsymbol{\Sigma}_0)\,. \tag{6}$$

□

Specifically, for a one–dimensional process $X(t) \in \mathbb{R}$ parameterized by $\lambda \in \mathbb{R}_+$, $\eta \in \mathbb{R}$ and constant diffusion $\varsigma \in \mathbb{R}_+$:

$$\mathrm{d}X(t) = \left(-\lambda X(t) + \eta\right) \mathrm{d}t + \varsigma\, \mathrm{d}B(t) \tag{7}$$

we can write the solution at some later time T conditioned on the state x at current time t as [11]:

$$X(T) = x e^{-\lambda(T-t)} + \int_t^T e^{-\lambda(T-s)} \eta\, \mathrm{d}s + \int_t^T e^{-\lambda(T-s)} \varsigma\, \mathrm{d}B(s) \tag{8}$$

which leads to the mean and covariance:

$$\boldsymbol{m}_{x(i)} = \mathbb{E}\left[X(T_i)|X(t) = x\right] = x e^{-\lambda(T_i - t)} + \frac{\eta}{\lambda}\left(1 - e^{-\lambda(T_i - t)}\right) \tag{9}$$

$$\boldsymbol{C}_{(i,j)} = \mathrm{Cov}\left(X(T_i), X(T_j)\right) = \varsigma^2 \int_t^{\min(T_i, T_j)} e^{-\lambda(T_i - s)} e^{-\lambda(T_j - s)}\, \mathrm{d}s \tag{10}$$

$$= \varsigma^2 \frac{e^{-\lambda|T_i - T_j|} - e^{-\lambda(T_i + T_j - 2t)}}{2\lambda}\,. \tag{11}$$

3.2 Incorporating non-linear residual terms

We propose to define a prior SDE composed of linear and non-linear drifts as

$$\mathrm{d}X(t) = \left(-\lambda_\theta X(t) + \eta_\theta + b_\theta\left(X(t)\right)\right) \mathrm{d}t + \left(\varsigma_\theta + \sigma_\theta(X(t))\right) \mathrm{d}B(t) \tag{12}$$

where $b_\theta(\cdot)$ and $\sigma_\theta(\cdot)$ are non-linear functions (e.g. neural networks) and θ indicates learnable parameters. Equivalently, the control term is defined as

$$u(\tilde{X}(t), t) \equiv u_c(\tilde{X}(t), t) + u_\phi(\tilde{X}(t), t) \tag{13}$$

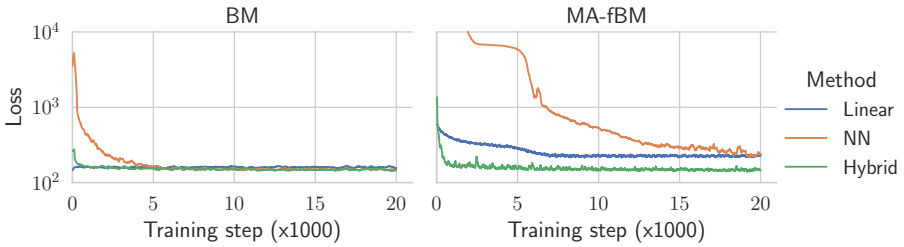

Fig. 1: We show the loss (negative ELBO) curves of the models driven by BM (left) and MA-fBM (right). For both experiments, our proposed hybrid model (green) starts training with a loss that is multiple orders of magnitude smaller and converges much faster than a standard non-linear neural network model (blue). Our hybrid model (green) also performs better than the strictly linear model (orange), especially for the MA-fBM experiment.

where $u_c(\cdot)$ is the analytical optimal control solution (eq. (5)) that depends on λ_θ, η_θ and ς_θ (eq. (9) and (11)) and u_ϕ is a residual non-linear control term, modeled e.g. by a neural network. For a purely linear model, without the non-linear components, the ELBO would be optimal by definition. However, such a model would not be expressive, i.e., not be able to capture realistic, non-linear data. The core idea of our work is to combine the linear terms with the residual non-linear terms $b_\theta(\cdot)$, $\sigma_\theta(\cdot)$ and $u_\phi(\cdot)$ such that the training of the model is more robust and fast, benefiting from the best of both worlds.

Furthermore, a crucial advantage of the linear model is the use of the tractable log-likelihood function $\log \mathcal{N}(\boldsymbol{O}; \boldsymbol{m}_x, \boldsymbol{C}+\boldsymbol{\Sigma}_0)$ to directly find λ_θ, η_θ and ς_θ, without having to solve computationally costly SDEs. This allows initialization of training where the linear component is already optimal.

Extension to fractional Brownian motion (fBM). A method for variational inference for SDEs with long-term correlation, driven by fBM, was recently proposed [3]. A Markov approximation of fBM (MA-fBM) is used, essentially enlarging the state–space by multiple processes driven by a shared BM. This allows variational inference in a similar way as explained in Section 2. Hence, SDEs driven by MA-fBM readily benefits from our proposed methods.

4 Experiments

We apply our method on the first 500 days of the 3–Month US Treasury Bills[1]. We compare the training of our proposed hybrid model with the non-linear residual part to the training of a standard non-linear model and a strictly linear model. We also apply our method to the SDEs driven by MA-fBM, presented in Section 3.2. The non-linear prior drift $b_\theta(\cdot)$, diffusion $\sigma_\theta(\cdot)$ and control term $u_\phi(\cdot)$ are neural networks. The observations are encoded by an additional neural

[1] `https://fred.stlouisfed.org/series/DTB3`

network into $u_\phi(\cdot)$, as is typically done in VI for SDEs [2,3]. All neural networks have three layers, 128 hidden neurons and the tanh activation function. The observations noise $\mathbf{\Sigma}_0 = 0.1^2\mathbf{I}$. For the MA-fBM experiment we set a Hurst index of 0.65 which is a reasonable choice for this data [12]. Figure 1 shows the loss (negative ELBO) curves of the three models, both for the models driven by BM and MA-fBM.

5 Conclusion

We present an optimal control inspired method for efficient variational inference for (neural) SDEs. Under practically reasonable assumptions, we explicitly formulate the control term with linear and residual non-linear components and derive a closed-form control term for the linear part using stochastic optimal control. This model is shown to converge faster than a standard non-linear SDE, both for SDEs driven by BM and Markov-approximate fBM.

Future work and limitations. Our work applies only to 1-d SDEs, future work will involve a multi-dimensional formulation. We also plan to cover latent SDEs [6].

References

[1] Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019.

[2] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David K Duvenaud. Scalable gradients and variational inference for stochastic differential equations. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–28. PMLR, 2020.

[3] Rembert Daems, Manfred Opper, Guillaume Crevecoeur, and Tolga Birdal. Variational inference for SDEs driven by fractional noise. In *International Conference on Learning Representations*, 2024.

[4] Sung Woo Park, Kyungjae Lee, and Junseok Kwon. Neural markov controlled sde: Stochastic optimization for continuous-time data. In *International Conference on Learning Representations*, 2021.

[5] Patrick Kidger, James Foster, Xuechen Chen Li, and Terry Lyons. Efficient and accurate gradients for neural sdes. In *Advances in Neural Information Processing Systems*, 2021.

[6] Kevin Course and Prasanth Nair. Amortized reparametrization: efficient and scalable variational inference for latent sdes. In *Advances in Neural Information Processing Systems*, 2024.

[7] Manfred Opper. Variational inference for stochastic differential equations. *Annalen der Physik*, 531(3):1800233, 2019.

[8] Hilbert J Kappen. Linear theory for control of nonlinear stochastic systems. *Physical review letters*, 95(20):200201, 2005.

[9] Cedric Archambeau and Manfred Opper. Approximate inference for continuous-time markov processes. *Bayesian time series models*, pages 125–140, 2011.

[10] Dimitra Maoutsa and Manfred Opper. Deterministic particle flows for constraining stochastic nonlinear systems. *Physical Review Research*, 4(4):043035, 2022.

[11] Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

[12] Martin Lysy and Natesh S Pillai. Statistical inference for stochastic differential equations with memory. *arXiv preprint arXiv:1307.1164*, 2013.