Gaussian Variational Inference for Diffusion Processes Revisited

Prakhar Verma[1] Vincent Adam[2] Arno Solin[1]

[1]Aalto University, Finland [2]University Pompeu Fabra, Spain

Aalto University

Universitat Pompeu Fabra Barcelona

TL;DR

What we do...

- Consider approximate Bayesian inference for generative models with diffusion process priors.
- Leverage an alternative parameterization and optimization algorithm for Gaussian variational inference, drawn from literature on linear diffusions (*i.e.*, Gaussian processes).
- Propose an approximate (variational) inference algorithm.
- Connect to posterior statistical linearization—from signal processing—and drastically improve inference time.

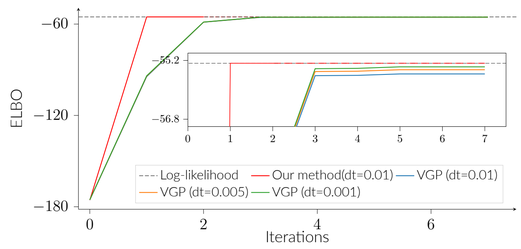

Fig. 1: Evolution of ELBO over iterations for the Ornstein–Uhlenck process. The proposed method gets to the optimal in just a one-step update, whereas Archambeau's method takes multiple steps.

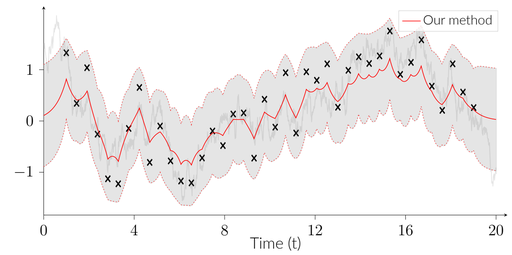

Fig. 2: Mean and 95% confidence interval of the posterior obtained on Ornstein-Uhlenbeck process by the proposed method. All the methods give identical posterior, so we plot only one.

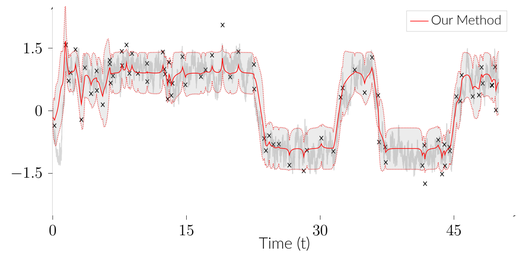

Fig. 3: Mean and 95% confidence interval of the posterior obtained on the non-linear process: Double-Well. The posterior obtained by the proposed method is able to capture the two wells.

Motivation

Latent Diffusion Process Model drift Brownian motion

$$\text{Diffusion process:} \quad \mathrm{d}\mathbf{x}_t = f_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}_t$$
$$\text{Sparse observations:} \quad \mathbf{y}_i \,|\, \mathbf{x} \sim p(\mathbf{y}_i \,|\, \mathbf{x}(t_i))$$

Probabilistic Inference

- Given a data set $\mathcal{D} = \{(t_i, \mathbf{y}_i)\}_{i=1}^N$, we are interested in the posterior process:

$$p(\mathbf{X} \,|\, \mathcal{D}) = Z^{-1} \prod_{i=1}^N p(\mathbf{y}_i \,|\, \mathbf{x}_i)\, p(\mathbf{X})$$

- Computing this posterior over state paths $\mathbf{X}$ and the marginal likelihood $p(\mathbf{y})$ is intractable. We resort to approximate inference.

Gaussian Variational Inference (VI)

Approximate Inference as Optimization

Variational Inference (VI) turns inference into the optimization problem

$$\min_{q \in Q} \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{X}) \,\|\, p(\mathbf{X} \,|\, \mathcal{D})\right] = \max_{q \in Q} L(q),$$

where $L(q)$ is the variational evidence lower bound (ELBO),

$$L(q) = \mathbb{E}_{q(\mathbf{X})}[\log p(\mathbf{y} \,|\, \mathbf{x})] - \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{X}) \,\|\, p(\mathbf{X})\right] < \log p(\mathbf{y}),$$

and Q is a set of distribution chosen to make $L(q)$ tractable and to contain a good approximation to the true posterior.

Gaussian approximate posterior

We set $Q = \{\text{diffusions with linear drift}\} = \{\text{Markovian GPs}\}$

Past work on Gaussian VI for SDE: Archambeau *et al.*

Parameterizing q: SDE with linear drift and shared diffusion (with the prior)

$$q: \ \mathrm{d}\mathbf{x}_t = (\mathbf{A}_t \mathbf{x}_t + \mathbf{b}_t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}_t,$$

Problems

- Slow fixed point optimization algorithm for inference.
- $(\mathbf{A}_t, \mathbf{b}_t)$ mix the prior and posterior (bad for learning).

Proposed Method: Inspiration

Inspiration #1: Natural gradient descent for linear diffusions

Linear diffusion = Gaussian Processes: GP are in exponential family

$$p(\mathbf{x}) = \exp(\langle T(\mathbf{x}), \boldsymbol{\eta}_p \rangle - A(\boldsymbol{\eta}_p))$$

Mirror ascent in exponential family, using the expectation parameterization $\boldsymbol{\mu} = \mathbb{E}[T(\mathbf{x})]$, with KL penalty leads to efficient inference, in the natural parameterization: $\boldsymbol{\eta}_q = \boldsymbol{\eta}_p + \underbrace{\boldsymbol{\lambda}}_{\text{sites}}$,

$$q^{k+1} = \arg\max_{q \in Q} \left(\langle \nabla_{\boldsymbol{\mu}} L(q)|_{\boldsymbol{\mu}=\boldsymbol{\mu}^k}, \boldsymbol{\mu} \rangle - \tfrac{1}{\rho} \mathrm{D}_{\mathrm{KL}}\left[q \,\|\, q^k\right] \right)$$
$$\boldsymbol{\lambda}^{k+1} = (1-\rho)\boldsymbol{\lambda}^k + \rho \underbrace{\nabla_{\boldsymbol{\mu}} \mathbb{E}_{q(\mathbf{X})}[\log p(\mathbf{y} \,|\, \mathbf{x})]}_{\text{local gradients}}$$

Inspiration #2: Statistical posterior linearisation

- Inference for linear diffusion possible in closed form.
- Many inference methods include linearization of the drift f along the ongoing approximate posterior

$$(\mathbf{A}, \mathbf{b}) = \arg\min \mathbb{E}_{q(\mathbf{x})} \|\mathbf{A}\mathbf{x} + \mathbf{b} - f(\mathbf{x})\|^2_{\mathbf{Q}^{-1}}$$

Proposed Method: Algorithm

We combine statistical linearization and mirror ascent in two steps:

- Linearize the prior diffusion $p_L^k \approx p$ at iterate q^k
- Run Mirror ascent on $L(q)$: $\boldsymbol{\eta}^{k+1} = \boldsymbol{\eta}_L^k + \boldsymbol{\lambda}^{k+1}$

Expanding the loss

$$L(q) = \mathbb{E}_{q(\mathbf{X})}[\log p(\mathbf{y} \,|\, \mathbf{x})] + \underbrace{\left(\mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{X}) \,\|\, p_L(\mathbf{X})\right] - \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{X}) \,\|\, p(\mathbf{X})\right]\right)}_{\text{linearization error}} - \underbrace{\mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{X}) \,\|\, p_L(\mathbf{X})\right]}_{\text{KL to linearized prior}},$$

MD leads to the updates

$$\boldsymbol{\lambda}_{k+1} = (1-\rho)\boldsymbol{\lambda}_k + \underbrace{\rho \nabla_{\boldsymbol{\mu}} \mathsf{VE}[q]}_{\text{data sites: sparse}} + \rho \underbrace{\left(\boldsymbol{\lambda}_k - \nabla_{\boldsymbol{\mu}} \mathrm{D}_{\mathrm{KL}}\left[q \,\|\, p\right]\right)}_{\text{error sites: dense}}$$

Properties of Our Method

- For linear diffusion, we recover the GP algorithm in [1] (no linearization error).
- For Gaussian likelihoods, the data sites reach their optimal value in a single step for $\rho = 1$.
- Decoupling the prior, data, and error contribution to the posterior leads to faster learning of the hyperparameters θ of the diffusion $\boldsymbol{\eta}^{k+1} = \boldsymbol{\eta}_L^k(\theta) + \boldsymbol{\lambda}^{k+1}$.

Experiments

Linear process prior: The Ornstein–Uhlenbeck process (Fig. 1–2)

$$\mathrm{d}\mathbf{x}_t = \alpha \mathbf{x}_t\,\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}_t$$

- Compare against the method of Archambeau *et al.* [2].
- Single-step update for inference is obtained.

Non-linear process prior: The double-well process (Fig. 3)

$$\mathrm{d}\mathbf{x}_t = 4\mathbf{x}_t(1 - \mathbf{x}_t^2)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}_t$$

Conclusion

- Alternative site-based parameterization for generative models with diffusion process priors motivated by recent advances in Gaussian processes.
- Connecting with the literature on posterior statistical linearization.
- Drastically improve inference time in processes with linear as well as non-linear diffusion process priors.

References

[1] V. Adam, P. Chang, M. E. E. Khan, and A. Solin, "Dual parameterization of sparse variational Gaussian processes," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[2] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor, "Gaussian process approximations of stochastic differential equations," in *Gaussian Processes in Practice*, Proceedings of Machine Learning Research, PMLR, 2007.

[3] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

prakhar.verma@aalto.fi · vincent.adam@upf.edu · arno.solin@aalto.fi