Variational Gaussian Process Diffusion Processes

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.

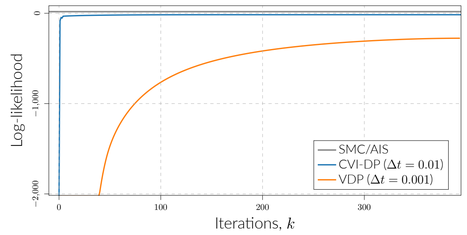

Fig. 1: Evolution of ELBO over iterations for the Double-Well process. CVI-DP (ours) converges faster than VDP (Archmabeau *et al.*(2007)) and is much closer to the SMC/AIS baseline.

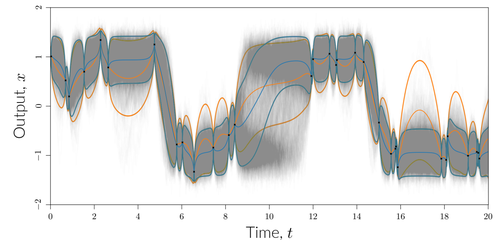

Fig. 2: Mean and 95% confidence interval of the approximate posterior processes for CVI-DP (ours) and VDP (Archambeau *et al.*(2007)) overlaid on the SMC ground-truth samples for the Double-Well process.

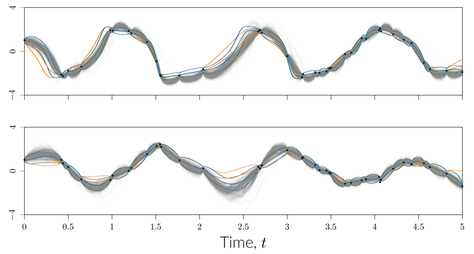

Fig. 3: Mean and 95% confidence interval of the approximate posterior processes for CVI-DP (ours) and VDP (Archambeau *et al.*(2007)) overlaid on the SMC ground-truth samples for the van der Pol process.

Motivation

Latent Diffusion Process Model drift Brownian motion

Diffusion process: $\mathrm{d}\mathbf{x}_t = f_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}_t$

Sparse observations: $y_i \,|\, \mathbf{x} \sim p(y_i \,|\, \mathbf{x}(t_i))$

Probabilistic Inference

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

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

- 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 (VDP): 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: $\boldsymbol{\eta}_q = \boldsymbol{\eta}_p + \boldsymbol{\lambda}$,

$$q^{k+1} = \arg\max_{q \in Q} \left(\langle \nabla_{\boldsymbol{\mu}} L(q)|_{\boldsymbol{\mu}=\boldsymbol{\mu}^k}, \boldsymbol{\mu} \rangle - \tfrac{1}{\rho} \mathrm{D}_{\mathrm{KL}}\left[q \,\|\, q^k\right] \right)$$

$$\boldsymbol{\lambda}^{k+1} = (1-\rho)\boldsymbol{\lambda}^k + \rho \nabla_{\boldsymbol{\mu}} \mathbb{E}_{q(\mathbf{X})}[\log p(\mathbf{y} \,|\, \mathbf{x})]$$

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

Proposed Method (CVI-DP): Algorithm

- We discretize the prior process and introduce the change of measure as

$$\frac{p(\{\mathbf{x}\}_\tau)}{p_L(\{\mathbf{x}\}_\tau)} = \exp\left(\sum_{m=0}^{M-1} \log \frac{p(\mathbf{x}_{m+1} \,|\, \mathbf{x}_m)}{p_L(\mathbf{x}_{m+1} \,|\, \mathbf{x}_m)} \right).$$

- Therefore, an alternative expression for the posterior distribution is obtained

$$p(\{\mathbf{x}\}_\tau \,|\, \mathcal{D}) = p_L(\{\mathbf{x}\}_\tau)\, p(\mathbf{y} \,|\, \{\mathbf{x}\}_\tau) \exp\left(\sum_{m=0}^{M-1} V(\tilde{\mathbf{x}}_m)\, \Delta t \right),$$

where $V(\tilde{\mathbf{x}}_m) = \frac{1}{2}\|f_L(\mathbf{x}_m) - f_p(\mathbf{x}_m)\|^2_{\mathbf{Q}_c^{-1}} + \Delta t \left\langle \frac{\mathbf{x}_{m+1} - \mathbf{x}_m}{\Delta t} - f_L(\mathbf{x}_m),\, f_p(\mathbf{x}_m) - f_L(\mathbf{x}_m) \right\rangle_{\mathbf{Q}_c^{-1}}$.

- The ELBO in CVI-DP is

$$\mathcal{L}(q) = -\mathrm{D}_{\mathrm{KL}}\left[q(\{\mathbf{x}\}_\tau) \,\|\, p_L(\{\mathbf{x}\}_\tau)\right] + \sum_{i=1}^n \mathbb{E}_{q(\mathbf{x}_i^{\mathrm{d}})} \log p(y_i \,|\, \mathbf{x}_i^{\mathrm{d}}) + \sum_{m=0}^{M-1} \mathbb{E}_{q(\tilde{\mathbf{x}}_m)} V(\tilde{\mathbf{x}}_m)\, \Delta t.$$

- Mirror Descent leads to the updates

$$\boldsymbol{\lambda}_i^{(k+1)} = (1-\rho)\boldsymbol{\lambda}_i^{(k)} + \rho\, \phi_i^{-1}(\nabla_{\boldsymbol{\mu}_i} \mathbb{E}_{q^{(k)}} \log p(y_i \,|\, \mathbf{x}_i^{\mathrm{d}}))$$

$$\boldsymbol{\Lambda}_m^{(k+1)} = (1-\rho)\boldsymbol{\Lambda}_m^{(k)} + \rho\, \psi_m^{-1}(\nabla_{\tilde{\boldsymbol{\mu}}_m} \mathbb{E}_{q^{(k)}} V(\tilde{\mathbf{x}}_m))$$

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

Double-Well process prior: The Double-Well process (Fig. 1–2)

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

- The marginal state distributions have two modes that sample state trajectories keep visiting through time.

Stochastic van der Pol prior: The stochastic van der Pol process (Fig. 3)

$$\mathrm{d}\mathbf{x}_t = \theta_0(\theta_1\, x_t^{(1)} - (1/3) x_t^{(1)} - x_t^{(2)}\ \mathrm{d}t,\ (1/\theta_1) x_t^{(1)}\ \mathrm{d}t)^\top + \mathrm{d}\boldsymbol{\beta}_t$$

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.

More details…

See the paper:

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

[3] 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.

AISTATS 2024 — Valencia, Spain.

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