Sparse Gaussian Processes for Stochastic Differential Equations


Prakhar Verma


Vincent Adam


Arno Solin

Aalto University · Secondmind · Neural Information Processing Systems

TL;DR

We address the problem of learning SDE from noisy observations and

- derive an approximate (variational) inference algorithm
- propose a novel parameterization of the approximate distribution over paths using a sparse Markovian Gaussian process

The approximation is efficient in storage and computation, allowing the usage of well-established optimizing algorithms such as natural gradient descent.

Background & Motivation

SDE

- An observed dynamical system on a time interval $[0, \tau]$ can be modeled using an SDE [1]

$$\mathrm{d}\mathbf{x}_t = f_\theta(\mathbf{x}_t, t)\,\mathrm{d}t + L\,\mathrm{d}\beta_t\,,$$

  where $f_\theta(\mathbf{x}_t, t)$ is the drift function, $LL^\top = \mathbf{\Sigma}$ is the (time-invariant) diffusion coefficient, and $\mathrm{d}\beta_t$ is the standard Brownian motion.
- We focus on systems where the diffusion term is constant, and the state $\mathbf{x}$ is indirectly observed at n discrete time points t_i via an observation model providing the likelihood $\{p(\mathbf{y}_i \mid \mathbf{x}_i)\}_{i=t_1}^{t_n}$.
- Aim is to learn the θ parameter(s) of the drift $f_\theta(\mathbf{x}_t, t)$ given observations by maximizing the marginal likelihood $p_\theta(\mathbf{y}_{t_1,\ldots,t_n})$.
- Model has arbitrary likelihood and the drift of the SDE is non-linear.

Inference with SDE priors

- The process $\mathbf{x}_t$ is continuous over time but not necessarily Gaussian.
- It defines a probability measure over paths $\mathbf{x}_t$

$$p(\mathbf{x}(\cdot) \mid \mathbf{y}_{1\ldots n}) = \frac{1}{Z} \times \prod p(\mathbf{y}_i \mid \mathbf{x}_i) \times p(\mathbf{x}(\cdot)),$$

  where Z is the normalization constant.
- Computing the posterior distribution over state paths and the marginal likelihood is intractable, we thus resort to approximate inference.

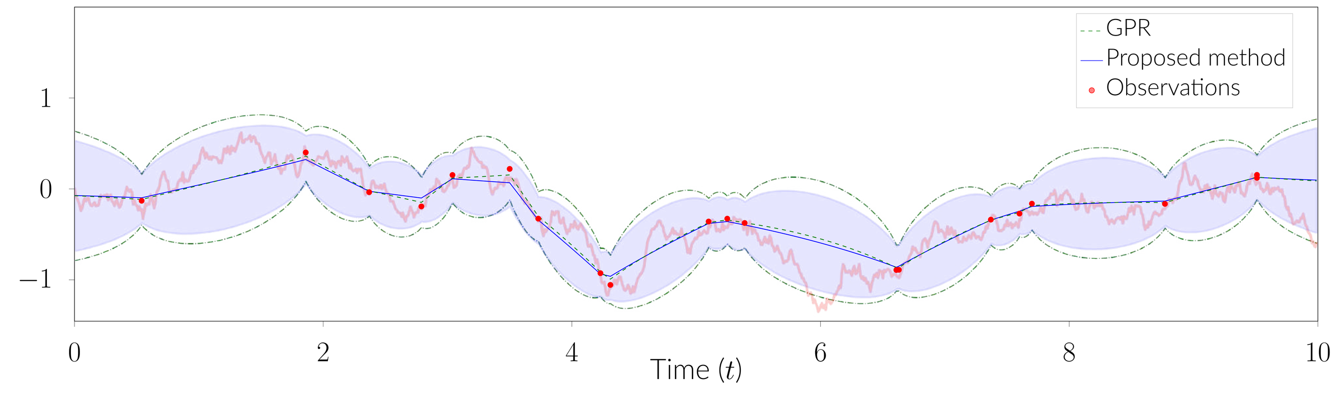

Fig. 1: GPR posterior and approximated posterior mean and 95% confidence interval of the proposed method along with the simulated trajectory and the noisy observations.

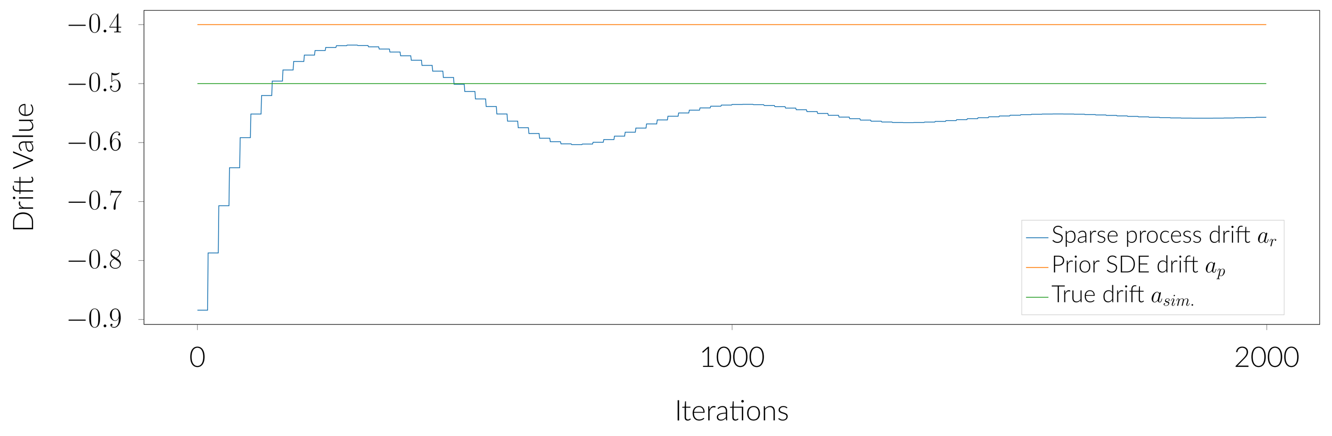

Fig. 2: The evolution of the drift of the sparse Markovian Gaussian process over iterations along with the prior SDE and the true SDE drift.

Variational Inference

VI turns an inference problem into an optimization problem with the optimal approximate posterior as $q^* = \arg\min_{q \in Q} \mathcal{L}(q)$, where $\mathcal{L}(q)$ is the ELBO: $\mathcal{L}(q) = \mathbb{E}_q \log p(\mathbf{y} \mid \mathbf{x}) - \mathrm{D}_{\mathrm{KL}}[q(\mathbf{x}) \| p(\mathbf{x})]$

Archambeau's method

Markovian Gaussian process is used as Q,

$$q(\mathbf{x}(\cdot)) \;:\; \mathrm{d}\mathbf{x}_t = f_l(\mathbf{x}_t, t) + L\,\mathrm{d}\beta_t$$

where $f_l(\mathbf{x}_t, t) = -A_t\,\mathbf{x}_t + b_t$, and A_t, b_t are functions of time [2].

Proposed Method

Conditioned Markovian GP as Q, by conditioning states of a stationary Markovian GP r_ϕ to Gaussian variable with distribution w_ψ

$$q_{\{\phi,\psi\}}(\mathbf{x}(\cdot)) = r_\phi(\bar{\mathbf{x}}(\cdot) \mid \mathbf{x}(z))\, w_\psi(\mathbf{x}(z))\,.$$

ELBO for the proposed model is

$$\mathcal{L} = \sum_{i=0}^{n} \mathbb{E}_{q(\mathbf{x}(t_i))}[l(\mathbf{x}_i)] + \int_{t=0}^{\tau} \mathbb{E}_{q(\mathbf{x}_t)}[g(\mathbf{x}_t)]\,\mathrm{d}t - \mathrm{D}_{\mathrm{KL}}[w(\mathbf{x}(z)) \| r(\mathbf{x}(z))]\,,$$

where $g(\mathbf{x}_t) = -\frac{1}{2}(f_\theta(\mathbf{x}_t) - f_\phi\,\mathbf{x}_t)^\top \mathbf{\Sigma}^{-1} (f_\theta(\mathbf{x}_t) - f_\phi\,\mathbf{x}_t)$, and $l(\mathbf{x}_i) = \log p(\mathbf{y}_i \mid \mathbf{x}_i)$, with the observations assumed i.i.d.

Inference and Learning

Two-step iterative algorithm, following the variational EM algorithm [3].

Learning

- Gradient descent to learn the θ parameters of the prior SDE, Step 1.

Inference

- Gradient descent for ϕ parameters of pseudo-prior r, Step 2.
- Natural gradient descent for parameters ψ of the distribution w_ψ, Step 3.

Algorithm 1: Optimization

```
η, ν, γ ← learning rates
while not converged do
    θ_{n+1} ← θ_n + ν ∇_θ L_sde ;                // Step 1(Learn θ)
    while not converged do
        Hyperparameter gradient step:
        φ_{n+1} ← φ_n + η ∇_φ L ;                // Step 2(Learn r)
        while not converged do
            Natural gradient step:
            λ̄_{n+1} ← γ_t ∇_μ α + (1 − γ_t) λ̄_n ; // Step 3(Learn w)
        end
    end
end
```

- Natural gradient updates, following [4]

$$\boldsymbol{\lambda}_{t+1} = \gamma_t \nabla_{\boldsymbol{\mu}}\,\alpha + (1 - \gamma_t)\,\boldsymbol{\lambda}_t\,,$$

  where $\alpha = \int_{t=0}^{\tau} \left(\mathbb{E}_{q(\mathbf{x}_t)}[g(\mathbf{x}_t)] + \Sigma_{i=0}^{n} \delta(t - t_n) \mathbb{E}_{q(\mathbf{x}(t_i))}[l(\mathbf{x}_i)]\right) \mathrm{d}t$, and $\gamma_t = \frac{1}{1+\rho_t}$ with $\boldsymbol{\mu}$ being the mean parameter, $\boldsymbol{\lambda}$ the natural parameter of w, and δ is the dirac function.

Experiment with Ornstein–Uhlenbeck Process

We consider the OU process driven by SDE,

$$\mathrm{d}\mathbf{x}(t) = -a\,\mathbf{x}(t)\,\mathrm{d}t + \sigma\,\mathrm{d}\beta(t).$$

The proposed method is applied to approximate the posterior with

$$q(\mathbf{x}(\cdot)) = r(\mathbf{x}(\cdot) \mid \mathbf{x}(z))\, w(\mathbf{x}(z))\,,$$

where the kernel of r is the modified Matérn-1/2; whose diffusion coefficient matches that of the prior SDE.

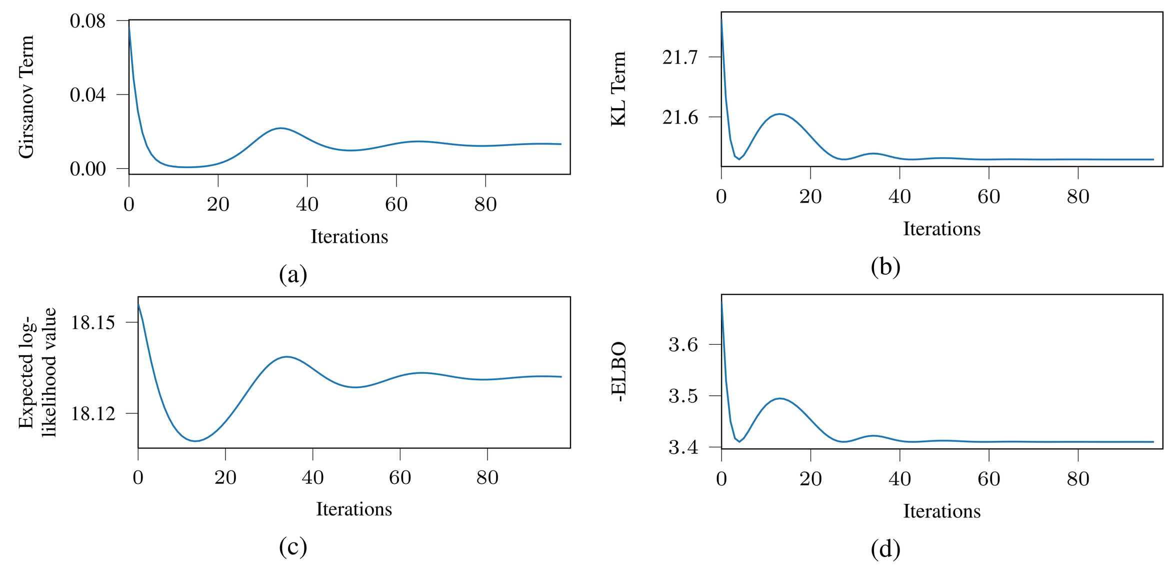

Fig. 3: Ornstein–Uhlenbeck process: The evolution of the (a) Girsanov value; (b) Kullbeck–Liebler divergence value; (c) Expected log-likelihood value; (d) Negative ELBO; over training iterations.

Conclusion

The method can be summarized as performing GP regression with a *pseudo* Markovian GP prior, while ensuring that the drift of this *pseudo* prior matches that of the prior SDE.

Limitations & Extensions

- Stationary GP has a linear drift and can not be expected to approximate well a non-linear drift.
- A natural extension is to use a piecewise stationary Markovian GP whose drift coefficient is different in between each consecutive pair of inducing points.
- Alternatively, a mixture of Markovian GPs could be used which would automatically cluster the state-space to provide a global approximation to the prior drift.

References

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

[2] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor, "Gaussian process approximations of stochastic differential equations," in *Gaussian Processes in Practice*, vol. 1 of *Proceedings of Machine Learning Research*, pp. 1–16, PMLR, 2007.

[3] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*, pp. 355–368, Springer, 1998.

[4] M. E. Khan, "Decoupled variational Gaussian inference," in *Advances in Neural Information Processing Systems 27 (NIPS)*, pp. 1547–1555, Curran Associates, Inc., 2014.

prakhar.verma@aalto.fi