Path-wise, spectral, and geometric perspectives on Gaussian processes

Alexander Terenin
Imperial College London

Joint work with James T. Wilson,* Viacheslav Borovitskiy,* Peter Mostowsky,* and Marc Deisenroth

Talk for GE Research

July 29th, 2020

HTTPS://AVT.IM/ · @AVT_IM

Imperial College London

UCL

*Equal contribution

Brief review of Multivariate Gaussians

Let $\boldsymbol{f} \in \mathbb{R}^d$ be a finite-dimensional random vector. We say that $\boldsymbol{f}$ is *Gaussian* if either one of the following equivalent properties hold.

1. For all $\boldsymbol{\ell} \in \mathbb{R}^d$, the quantity $\boldsymbol{\ell}^T \boldsymbol{f}$ is a univariate Gaussian.
2. There is a matrix $\mathbf{L}$ and vector $\boldsymbol{\mu}$ such that $\boldsymbol{f} = \mathbf{L}\boldsymbol{z} + \boldsymbol{\mu}$, where $\boldsymbol{z}$ is a standard multivariate Gaussian $(p(\boldsymbol{z}) \propto \exp(-\boldsymbol{z}^T \boldsymbol{z}/2))$.

We call $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{f})$ and $\mathbf{K} = \operatorname{Cov}(\boldsymbol{f}) = \mathbf{L}\mathbf{L}^T$ the mean vector and covariance matrix, respectively. These uniquely characterize $\boldsymbol{f}$.

Conditional distributions have a nice analytic form.

Brief review of Gaussian processes

In infinite dimensions, one can tell a similar story in different ways.

Let $f : M \to \mathbb{R}$ be a random function. We say that f is a *Gaussian process* if the following property holds.

1. For any finite set $\boldsymbol{x} \in \bigtimes_{i=1}^{n} M$, $f(\boldsymbol{x})$ is a multivariate Gaussian.

We call $\mu(\cdot) = \mathbb{E}(f(\cdot))$ and $k(\cdot, \cdot) = \operatorname{Cov}(f(\cdot), f(\cdot))$ the mean function and covariance kernel. These uniquely characterize f.

Conditioned processes have nice analytic marginals.

Property 2 also generalizes, but is substantially more subtle.

Efficiently sampling functions from Gaussian process posteriors

James T. Wilson,[*] Viacheslav Borovitskiy,[*] Alexander Terenin,[*]
Peter Mostowsky,[*] and Marc Deisenroth

[*]Equal contribution

Sampling in Gaussian processes

Consider rollouts in model-based reinforcement learning

$$x_{t+1} = x_t + f(x_t, u(x_t))$$
$$x_{t+2} = x_{t+1} + f(x_{t+1}, u(x_{t+1}))$$

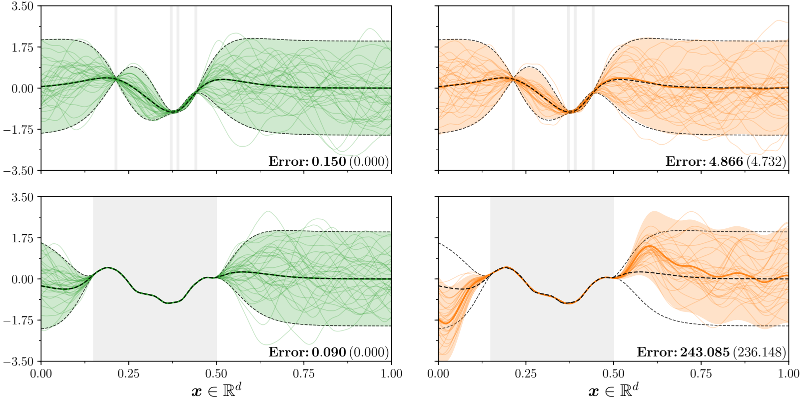
✓ Gaussian processes: excellent data-efficiency

☒ Gaussian process rollouts: $\mathcal{O}(T^3)$

Same issue occurs in Bayesian optimization
when minimizing GPs on a large grid

This work: address this without sacrificing accuracy

Sampling with sparse GPs

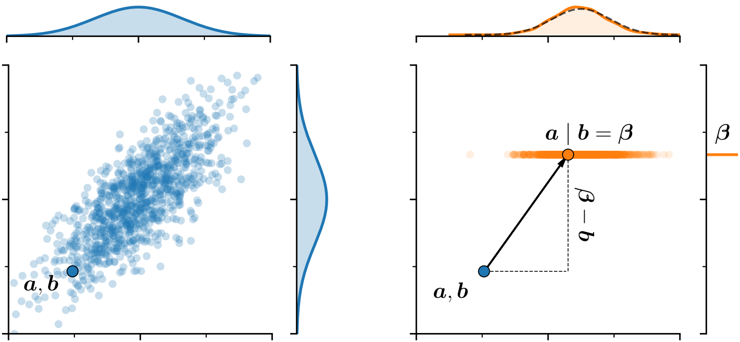


Random feature methods are fast, but introduce approximation error which can manifest as variance starvation

Matheron's update rule

For $\begin{bmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \right)$ we have

$$(\boldsymbol{f}_1 \mid \boldsymbol{f}_2 = \boldsymbol{u}) \stackrel{\mathrm{d}}{=} \boldsymbol{f}_1 + \mathbf{K}_{12}\mathbf{K}_{22}^{-1}(\boldsymbol{u} - \boldsymbol{f}_2)$$



Path-wise sampling with sparse GPs

This expression *lifts* to a path-wise characterization of posterior GPs

$$\underbrace{(f \mid \boldsymbol{y})(\cdot)}_{\text{posterior}} \overset{\mathrm{d}}{=} \underbrace{f(\cdot)}_{\text{prior}} + \underbrace{\mathbf{K}_{(\cdot)\boldsymbol{x}}\mathbf{K}_{\boldsymbol{x}\boldsymbol{x}}^{-1}(\boldsymbol{y} - f(\boldsymbol{x}))}_{\text{update}}$$

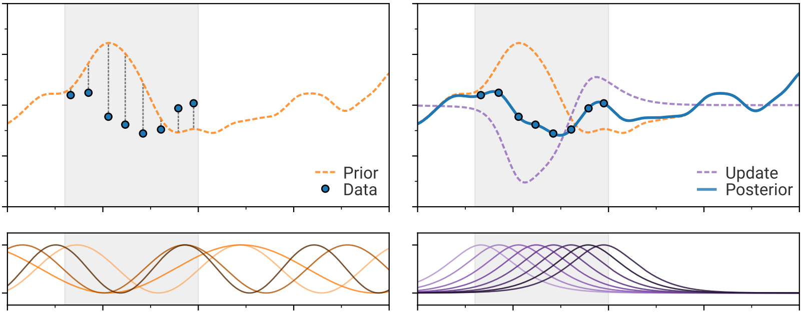
Prior term: discretize with random Fourier features

Data term: approximate with sparse GPs

$$\underbrace{(f \mid \boldsymbol{y})(\cdot)}_{\substack{\text{approximate} \\ \text{posterior}}} \overset{\mathrm{d}}{\approx} \underbrace{\sum_{i=1}^{\ell} w_i \phi_i(\cdot)}_{\substack{\text{RFF basis for} \\ \text{stationary prior}}} + \underbrace{\sum_{i=1}^{m} v_i k(\cdot, z_i)}_{\substack{\text{canonical basis} \\ \text{for sparse update}}} \qquad \boldsymbol{v} = \mathbf{K}_{\boldsymbol{z}\boldsymbol{z}}^{-1}(\boldsymbol{u} - \boldsymbol{\Phi}^T \boldsymbol{w})$$

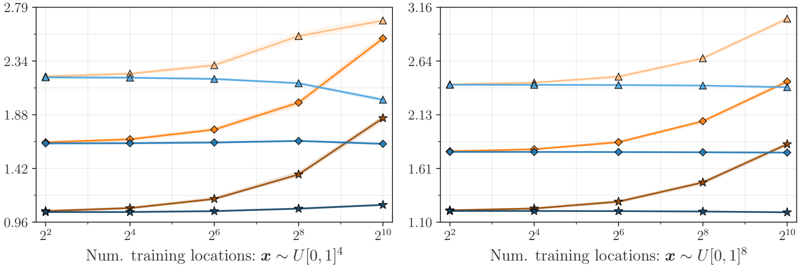
Visualizing path-wise sampling

$$\underbrace{(f \mid \boldsymbol{y})(\cdot)}_{\text{approximate posterior}} \overset{\mathrm{d}}{\approx} \underbrace{\sum_{i=1}^{\ell} w_i \phi_i(\cdot)}_{\text{RFF basis for stationary prior}} + \underbrace{\sum_{i=1}^{m} v_i k(\cdot, z_i)}_{\text{canonical basis for sparse update}} \qquad \boldsymbol{v} = \mathbf{K}_{\boldsymbol{zz}}^{-1}(\boldsymbol{u} - \boldsymbol{\Phi}^T \boldsymbol{w})$$



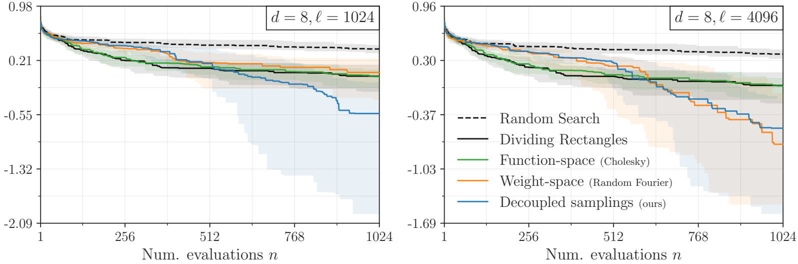
Error analysis

$$\underbrace{W_{2,L^2(\mathcal{X})}(f^{(d)}, f \mid \boldsymbol{y})}_{\text{total approximation error}} \leq \underbrace{W_{2,L^2(\mathcal{X})}(f^{(s)}, f \mid \boldsymbol{y})}_{\text{error in sparse posterior}} + \underbrace{C\,W_{2,L^2(\mathcal{X})}(f^{(w)}, f)}_{\text{error in approximate prior}}$$



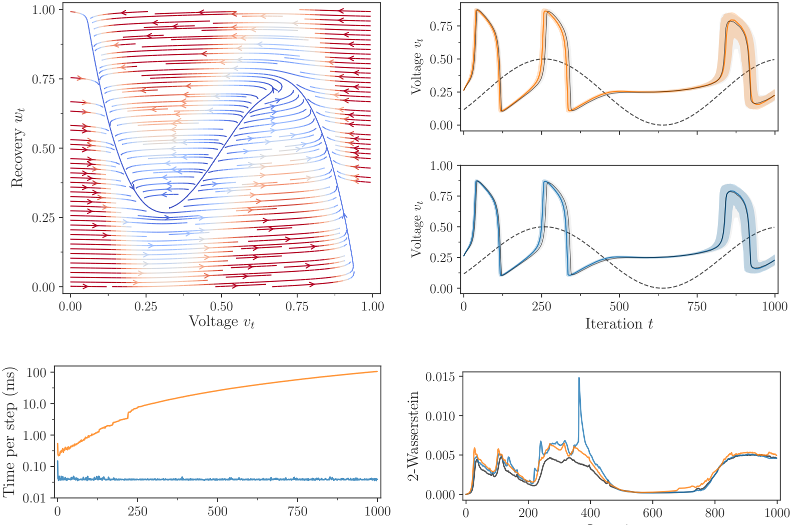
Empirical Wasserstein error smaller than for RFF

Bayesian optimization: Thompson sampling



Improved performance owing to smaller error

FitzHugh-Nagumo model neuron dynamical system



Significantly more efficient time-stepping

Matérn Gaussian processes on Riemannian manifolds

Viacheslav Borovitskiy,* Alexander Terenin,*
Peter Mostowsky,* and Marc Deisenroth

*Equal contribution

Matérn Gaussian processes on Riemannian manifolds

Matérn Gaussian processes are a popular GP model class

$$k(x,x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{\|x-x'\|}{\kappa} \right)^{\nu} K_\nu \left(\sqrt{2\nu} \frac{\|x-x'\|}{\kappa} \right)$$

σ^2: variance κ: length scale ν: smoothness

$\nu \to \infty$: recovers square exponential kernel

This defines GPs $f : \mathbb{R}^d \to \mathbb{R}$

What about $f : M \to \mathbb{R}$ where M is a Riemannian manifold?

A candidate generalization via geodesics

Let's consider the $\nu \to \infty$ case, and try extending via geodesics

$$k_{\text{naïve}}(x, x') = \sigma^2 \exp\left(-\frac{d_g(x, x')^2}{2\kappa^2}\right)$$

Theorem. (Feragen et al.) Let M be a complete Riemannian manifold without boundary. If $k_{\text{naïve}}$ is a positive semi-definite kernel for all κ, then M is isometric to a Euclidean space.

$\implies$ need a different candidate generalization

Matérn GPs as solutions of stochastic PDEs

Matérn and squared exponential GPs are solutions of stochastic partial differential equations

$$\left(\frac{2\nu}{\kappa^2} - \Delta\right)^{\frac{\nu}{2}+\frac{d}{4}} f = \mathcal{W} \qquad\qquad e^{-\frac{\kappa^2}{4}\Delta} f = \mathcal{W}$$

Δ: Laplacian $\quad \mathcal{W}$: (rescaled) white noise

$e^{-\frac{\kappa^2}{4}\Delta}$: (rescaled) heat semigroup

- This is the GP analog of $\boldsymbol{f} = \mathbf{L}\boldsymbol{z}$

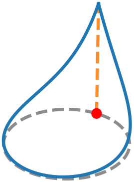
✓ Generalizes well to the Riemannian setting

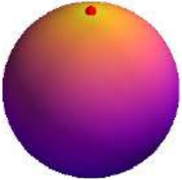
☒ Not very constructive, requires solving SPDEs

This work: compute the kernel, enable training via standard methods

A candidate generalization via stochastic PDEs

What do these kernels look like? $(\nu = 1/2)$





$$\left(\frac{2\nu}{\kappa^2} - \Delta_g\right)^{\frac{\nu}{2}+\frac{d}{4}} f = \mathcal{W}_g$$

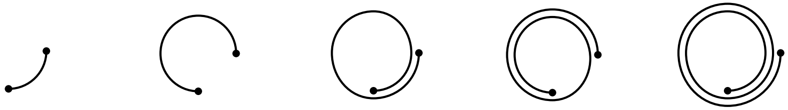
Δ_g: Laplace–Beltrami operator
$\mathcal{W}_g$: (rescaled) Riemannian white noise

What's wrong with the geodesic definition?

Consider the special case of the torus $\mathbb{T}^d = \mathbb{S}^1 \times .. \times \mathbb{S}^1$

Proposition. Up to a pair of additive and multiplicative constants, the kernel of the squared exponential GP on $\mathbb{T}^d$ is given by

$$k(x, x') = \sum_{n \in \mathbb{Z}^d} \sigma^2 \exp\left(-\frac{\|x - x' + n\|^2}{2\kappa^2}\right).$$



Similar to naïve generalization, but with extra terms

Riemannian Matérn kernels on compact spaces

Theorem. The kernel of Riemannian Matérn Gaussian processes is

$$k(x, x') = \frac{\sigma^2}{C} \sum_{n=0}^{\infty} \left(\frac{2\nu}{\kappa^2} - \lambda_n \right)^{\nu - \frac{d}{2}} f_n(x) f_n(x')$$

where λ_n and f_n are Laplace–Beltrami eigenpairs.

For the sphere, this is given by

$$k_\nu(x, x') = \frac{\sigma^2}{C_\nu} \sum_{n=0}^{\infty} c_{n,d}\, \rho_\nu(n)\, \mathcal{C}_n^{(d-1)/2}\Big(\cos\big(d_g(x, x')\big)\Big)$$

where $c_{n,d}$ are explicit constants, $\mathcal{C}_n^{(\cdot)}$ are the Gegenbauer polynomials, and $\rho_\nu(n)$ is the generalized spectral measure.

Posterior samples from Riemannian Matérn GPs

How do posterior samples look?

(a) Ground Truth (b) Posterior Mean (c) Standard Deviation

✓ Train by sampling from the prior and using pathwise formula

✓ Does not require repeated numerical SPDE solves

Concluding remarks

Thank you for your attention!

HTTPS://AVT.IM/

@AVT_IM

Imperial College London

UCL

J. T. Wilson,* V. Borovitskiy,* A. Terenin,* P. Mostowsky,* M. P. Deisenroth. Efficiently sampling functions from Gaussian process posteriors. International Conference on Machine Learning, 2020. *Equal contribution. Available at: HTTPS://ARXIV.ORG/ABS/2002.09309.

V. Borovitskiy,* A. Terenin,* P. Mostowsky,* M. P. Deisenroth. Matérn Gaussian processes on Riemannian manifolds, 2020. *Equal contribution. Available at: HTTPS://ARXIV.ORG/ABS/2006.10160.