Interpretable deep Gaussian processes for geospatial tasks

Daniel Augusto de Souza, Daniel Giles, Marc Peter Deisenroth

Why deep Gaussian processes?

- Gaussian processes (GPs) are widely used in machine learning for their simple uncertainty quantification;
- GP's kernel function → Modelling and uncertainty quality;
- Common stationary kernels, like square exponential and Matérn are unsuitable for non-stationary data;
  → Many geospatial processes, such as sea surface height, bathymetry.
- By composing multiple stationary GPs, this deep Gaussian process model can learn non-stationary functions.

Stationary kernels and lengthscales

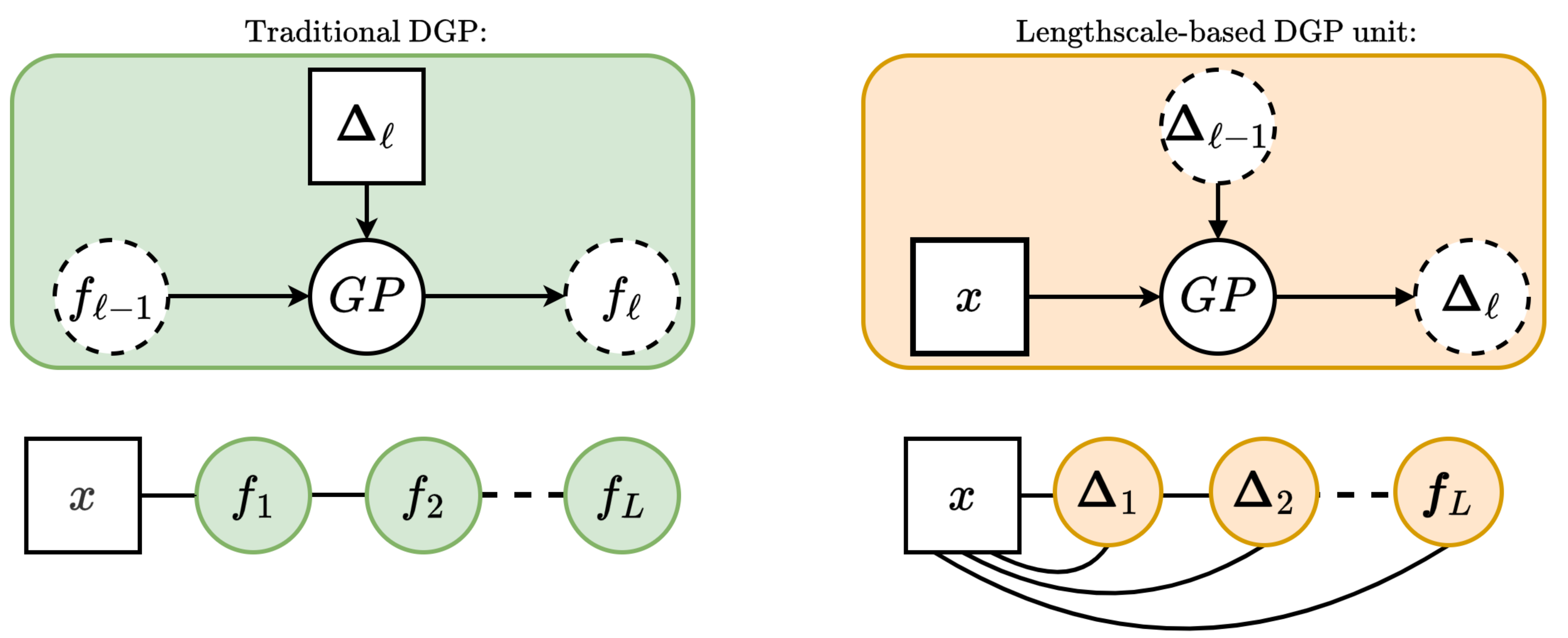
- Isotropic stationary: $\mathrm{k}(\boldsymbol{a},\boldsymbol{b}) = \pi_{\mathrm{k}}\left((\boldsymbol{a}-\boldsymbol{b})^{\mathrm{T}}\boldsymbol{\Delta}^{-1}(\boldsymbol{a}-\boldsymbol{b})\right)$;
  → Kernel is a function of the distance, weighted by the lengthscale matrix $\boldsymbol{\Delta}$.
- E.g., squared exponential kernel $\pi_{\mathrm{k}}(d^2) = \exp\left[-\frac{1}{2}d^2\right]$
- For a GP with zero mean and squared exponential kernel:

$$\mathrm{p}(\mathrm{f}(\boldsymbol{x})) = \mathrm{N}(0, k(x,x)) \rightarrow p\left(\frac{\partial \mathrm{f}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right) = \mathrm{N}(\mathbf{0}, \boldsymbol{\Delta}^{-1})$$

- So, **the inverse lengthscale matrix is the covariance of the prior covariance of the gradient**.

Interpretability of DGP models

- We focus in two broad classes of deep GPs:



- **Traditional DGPs** are repeatedly deform the outputs of previous layers with non-linear transformations. We analyze deep kernel learning [Wilson et al., 2015] and the compositional deep Gaussian process [Damianou & Lawrence, 2013].
- **Lengthscale-based DGPs** extend the notion of a lengthscale kernel to an input-dependent lengthscale field, each layer learning the next's lengthscale field. We study the deeply non-stationary Gaussian process [Salimbeni & Deisenroth, 2017] (first proposed by Gibbs [1997] and elaborated by Paciorek [2003]) and the thin and deep Gaussian process [de Souza et al., 2023].
- As the lengthscale matrix is highly interpretable, it was commonly assumed that lengthscale-based are more interpretable than DGPs. **However, we observe that the different architectures don't preserve the simple connection with the covariance of the gradient;**
- By focusing on the covariance of the gradient, **all DGP architectures can be interpreted** in the same way **regardless of their architecture type**.

Discussion and future work

- We propose moving the focus of interpretability **away from the lengthscale functions** and **instead look to the prior covariance of the gradient**, due to its closer connections to the physics of the problem in hand and uniform interpretability between architectures.
- Our preliminary experimental task on sea surface height interpolation with NATL60 data, **shows that there are differences in data-fit between models, however, these are not directly correlated with gradient matching in the prior**.
- One way to improve this is to **condition the posterior on observed variables that correlate with the gradient of the target function. Is this where lengthscale-based DGPs might have a conceptual advantage?**

By exploiting the **closed-form gradients** of different architectures of **deep Gaussian processes**, we re-exam and expand the issue of **physical interpretability** of deep GP models.

Deep Gaussian process architectures

Compositional kernels

- Non-linear warping function: $\boldsymbol{\tau}(\cdot) \rightarrow \mathrm{k}(\boldsymbol{\tau}(\boldsymbol{a}), \boldsymbol{\tau}(\boldsymbol{b}))$;
- If $\boldsymbol{\tau}(\boldsymbol{x})$ is a neural network → Deep Kernel Learning;
- If $\boldsymbol{\tau}(\boldsymbol{x})$ is a GP → Compositional deep Gaussian process;

Limitations

- The more complex $\boldsymbol{\tau}(\cdot)$ is, the harder the model is to interpret;
- For DKL, parameters are not Bayesian, thus the model overfits.

Lengthscale mixture kernels

- Positive semi-definite function field $\boldsymbol{\Delta}(\cdot)$

$$k_{\boldsymbol{\Delta}}(\boldsymbol{a},\boldsymbol{b}) = \sqrt{\frac{\sqrt{|\boldsymbol{\Delta}(\boldsymbol{a})|}\sqrt{|\boldsymbol{\Delta}(\boldsymbol{b})|}}{|\boldsymbol{\Delta}(\boldsymbol{a}) + \boldsymbol{\Delta}(\boldsymbol{b})|}}\,\pi_{\mathrm{k}}\left((a-b)^{\mathrm{T}}\left[\frac{\boldsymbol{\Delta}(\boldsymbol{a}) + \boldsymbol{\Delta}(\boldsymbol{b})}{2}\right]^{-1}(a-b)\right)$$

- If $\sigma(\boldsymbol{\Delta}(\cdot))$ is a GP w/ linking function $\sigma(\cdot)$ → Deeply Non-stationary GP

Limitations

- Does not encode latent spaces;
- Kernel scale affected by the pre-factor, giving rise to unwanted correlations.

Locally linear deformations

- A compositional kernel with $\boldsymbol{\tau}(\boldsymbol{x}) = \mathbf{W}(\boldsymbol{x}) \cdot \boldsymbol{x}$.
- Around a specific $\boldsymbol{x}$, we have $\boldsymbol{\Delta}(\boldsymbol{x}) = [\mathbf{W}^{\mathrm{T}}(\boldsymbol{x})\mathbf{W}(\boldsymbol{x})]^{-1}$;
→ Therefore, this model learns latent spaces and lengthscale fields;
- If $\mathbf{W}(\cdot)$ is a GP, then we obtain a thin and deep GP model.

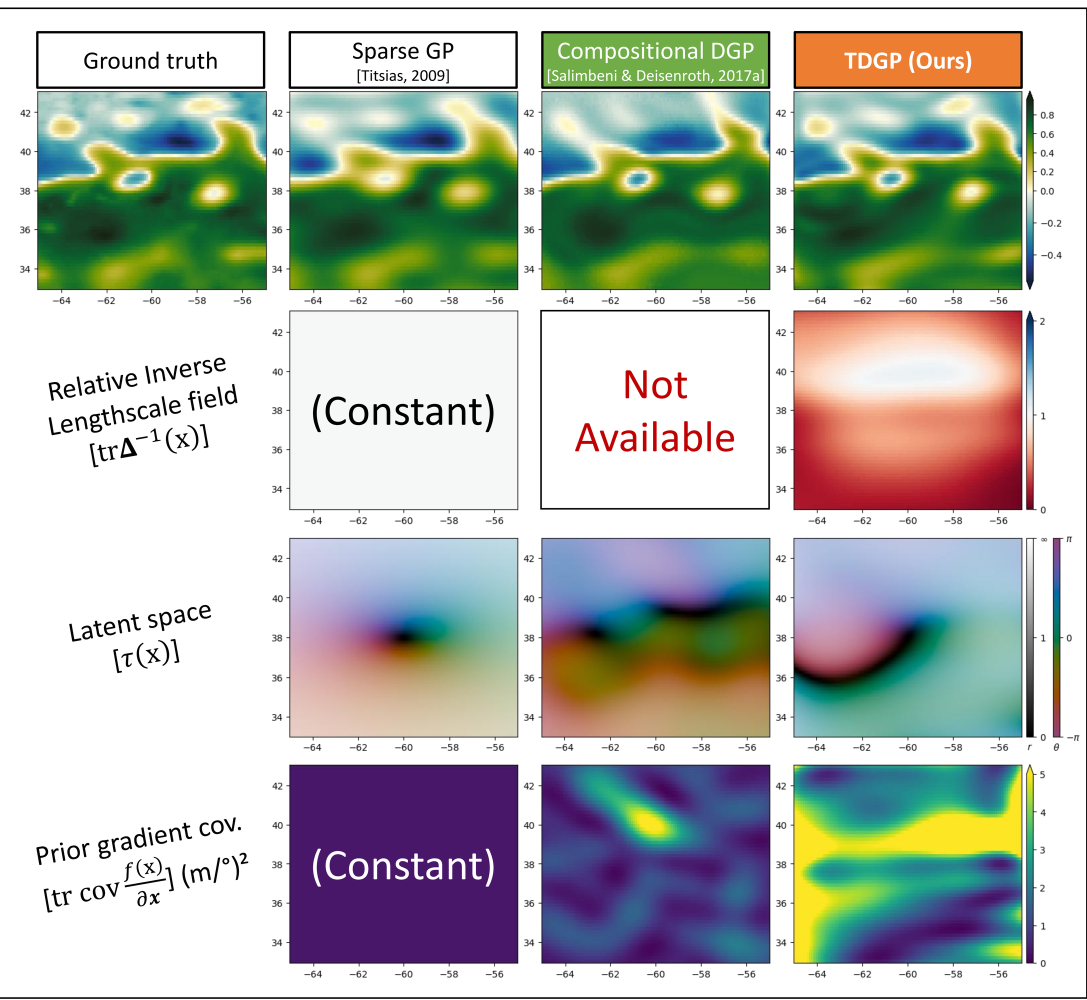
Limitations

- Adds $d \times q$ more GPs, making training harder;
- Without adding a bias dimension to input, neighborhood of 0 is unaffected;

Sea surface height experiments

- We use the simulated data from the SWOT Data Challenge NATL60 dataset [CLS/MEOM, 2020] at 2013-01-11 00:30.
- We evaluate with 10-fold train/test cross-validation splits.
- The gradient is numerically calculated in terms of the coordinates.
- We initialize the models following de Souza et al. [2023], 7000 epochs of training with Adam and 50 inducing points for the output function and 25 for the latent function.

	Data-fit NLPD	Data-fit RAE	Gradient NLPD
Sparse GP	-0.34 ± 0.11	1.05 ± 4.34	**0.56 ± 0.01**
Compositional DGP	-0.58 ± 0.21	0.46 ± 0.98	1.42 ± 0.39
Deeply Nonstationary GP	-0.69 ± 0.11	0.52 ± 1.58	---
TDGP (Ours)	-0.81 ± 0.16	**0.25 ± 0.15**	1.28 ± 0.33



UCL
The Alan Turing Institute