Inter-domain Deep Gaussian Processes

Tim G. J. Rudner [1] **Dino Sejdinovic** [2] **Yarin Gal** [1]

Abstract

Inter-domain Gaussian processes (GPs) allow for high flexibility and low computational cost when performing approximate inference in GP models. They are particularly suitable for modeling data exhibiting global structure but are limited to stationary covariance functions and thus fail to model non-stationary data effectively. We propose *Inter-domain Deep Gaussian Processes*, an extension of inter-domain shallow GPs that combines the advantages of inter-domain and deep Gaussian processes (DGPs), and demonstrate how to leverage existing approximate inference methods to perform simple and scalable approximate inference using inter-domain features in DGPs. We assess the performance of our method on a range of regression tasks and demonstrate that it outperforms inter-domain shallow GPs and conventional DGPs on challenging large-scale real-world datasets exhibiting *both* global structure *as well as* a high-degree of non-stationarity.

1. Introduction

Gaussian processes (GPs) are a powerful tool for function approximation. They are Bayesian non-parametric models and as such they are flexible, robust to overfitting, and provide well-calibrated predictive uncertainty estimates (Rasmussen & Williams, 2005; Bui et al., 2016). Deep Gaussian processes (DGPs) are layer-wise compositions of GPs designed to model a larger class of functions than shallow GPs.

To scale GP and DGP models to large datasets, a wide array of approximate inference methods has been developed, with inducing points-based variational inference being the most widely used (Snelson & Ghahramani, 2006; Titsias, 2009; Wilson & Nickisch, 2015). However, conventional inducing points-based inference for GPs relies on *point* evaluations and thus, by construction, creates *local* approximations to the target function. As a result, the approximate posterior predictive distribution may fail to capture complex *global* structure in the data, severely limiting the usefulness and computational efficiency of local inducing points-based approximations.

Inter-domain GPs were designed to overcome this limitation. In order to capture global structure in the underlying data-generating process, inter-domain GPs define inducing variables as projections of the target function over the entire input space and not as mere point evaluations (Lázaro-Gredilla & Figueiras-Vidal, 2009; Rahimi & Recht, 2008; Gal & Turner, 2015). The resulting posterior predictive distribution is able to represent complex data with global structure with higher accuracy as local approximations but at the same computational cost. Unfortunately, inter-domain projections most suitable for capturing global structure (e.g., spectral transforms) are limited by the fact that they can only be used with stationary covariance functions, making them ill-suited for modeling non-stationary data and limiting their usefulness in practice.

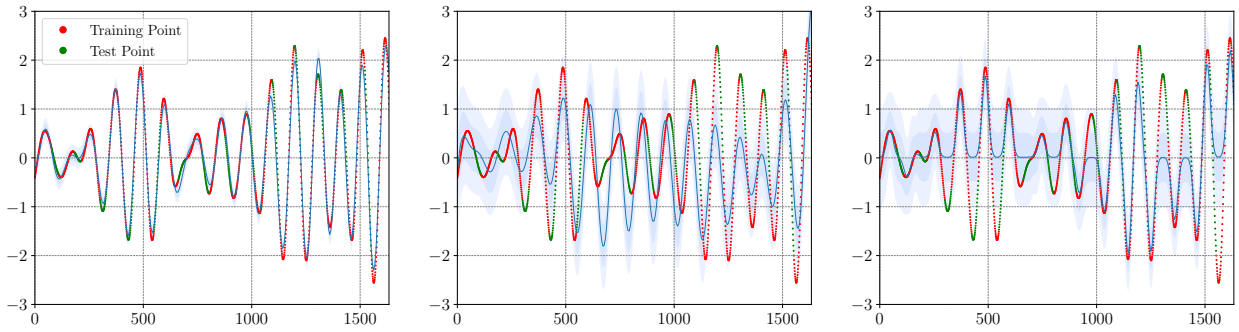
We propose *Inter-domain Deep Gaussian Processes* to overcome this limitation while retaining the benefits of inter-domain methods.[1] Specifically, we define an augmented DGP model, in which we replace local inducing variables by reproducing kernel Hilbert space (RKHS) Fourier features, and exploit the compositional structure of the variational distribution in doubly stochastic variational inference (DSVI) for DGPs. This way, we achieve simple and scalable approximate inference while efficiently capturing global structure in the underlying data-generating process. The resulting inter-domain DGP is composed of a composition of inter-domain GPs, which makes it possible to efficiently model complex, non-stationary data despite each inter-domain GP in the hierarchy being restricted to stationary covariance functions.

We establish that our method performs well on several complex real-world datasets exhibiting global structure and non-stationarity and demonstrate that inter-domain DGPs are more computationally efficient than DGPs with local approx-

[1]Department of Computer Science, University of Oxford, Oxford, United Kingdom [2]Department of Statistics, University of Oxford, Oxford, United Kingdom. Correspondence to: Tim G. J. Rudner <tim.rudner@cs.ox.ac.uk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

[1]For source code and additional results, see https://bit.ly/inter-domain-dgps.



(a) Deep GP with global approximations. **(b)** Shallow GP with global approximations. **(c)** Deep GP with local approximations.

Figure 1: Approximate posterior predictive distributions for shallow and deep GP models obtained from 20 inducing points. The blue lines denote the posterior predictive means of the models, respectively. Each shade of blue corresponds to one posterior standard deviation.

imations when modeling data with global structure. Figure 1 shows approximate posterior predictive distributions of an inter-domain deep GP (1a), an inter-domain shallow GP (1b), and a deep GP based on local approximations (1c) on a dataset with global structure.

To summarize, our main contributions are as follows:

1. We propose *Inter-domain Deep Gaussian Processes* and use RKHS Fourier features to incorporate global structure into the DGP posterior predictive distribution;
2. We present a simple approach for performing approximate inference in inter-domain DGPs by exploiting the compositional structure of the variational distribution in DSVI;
3. We show that inter-domain DGPs significantly outperform both inter-domain shallow GPs and state-of-the-art local approximate inference methods for DGPs on complex real-world datasets with global structure;
4. We demonstrate that inter-domain DGPs are more computationally efficient than local approximate inference methods for DGPs when trained on data exhibiting global structure.

2. Background

We begin by reviewing DGPs and inter-domain GPs. We will draw on this exposition in subsequent sections.

2.1. Deep Gaussian Processes

DGPs are layer-wise compositions of GPs in which the output of a previous layer is used as the input to the next layer. Similar to deep neural networks, the hidden layers of a DGP learn representations of the input data, but unlike neural networks, they allow for uncertainty to be propagated through the function compositions. This way, DGPs define probabilistic predictive distributions over the target variables and—unlike for shallow GPs—any finite collection of random variables distributed according to a DGP posterior predictive distribution does not need to be jointly Gaussian, allowing DGP models to represent a larger class of distributions over functions than shallow GPs.

Consider a set of N noisy target observations $\mathbf{y} \in \mathbb{R}^N$ at corresponding input points $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^\top \in \mathbb{R}^{N \times D}$. A DGP is defined by the composition

$$\mathbf{y} = \mathbf{f}^{(L)} + \boldsymbol{\epsilon} \stackrel{\text{def}}{=} f^{(L)}(f^{(L-1)}(...f^{(1)}(\mathbf{X}))...) + \boldsymbol{\epsilon}, \quad (1)$$

where L is the number of layers, and $\mathbf{f}^{(\ell)} = f^{(\ell)}(\mathbf{f}^{(\ell-1)})$ in the composition denotes the ℓth-layer GP, $f^{(\ell)}(\cdot)$, evaluated at $\mathbf{f}^{(\ell-1)}$. We follow previous work and absorb the noise between layers, which is assumed to be i.i.d. Gaussian, into the kernel so that $k_{\text{noisy}}(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + {\sigma^{(\ell)}}^2 \delta_{ij}$, where δ_{ij} is the Kronecker delta and ${\sigma^{(\ell)}}^2$ is the noise variance between layers (Salimbeni & Deisenroth, 2017).

A DGP with likelihood $p(\mathbf{y} \,|\, \mathbf{f}^{(L)})$ has the joint distribution

$$p(\mathbf{y}, \{\mathbf{f}^{(\ell)}\}_{\ell=1}^L) = \prod_{n=1}^N p(\mathbf{y}_i \,|\, \mathbf{f}_i^{(L)}) \prod_{\ell=1}^L p(\mathbf{f}^{(\ell)} \,|\, \mathbf{f}^{(\ell-1)}),$$

with $\mathbf{f}^0 \stackrel{\text{def}}{=} \mathbf{X}$. Unlike shallow GPs, exact inference in DGPs is not analytically tractable due to the nonlinear transformations at every layer of the composition in Equation (1).

To make posterior inference tractable, a number of approximate inference techniques for DGPs have been developed with the aim of improving performance, scalability, stability, and ease of optimization (Dai et al., 2015; Hensman & Lawrence, 2014; Bui et al., 2016; Salimbeni & Deisenroth, 2017; Cutajar et al., 2017; Mattos et al., 2015; Havasi et al., 2018; Salimbeni et al., 2019).

2.2. Inter-domain Gaussian Processes

Inter-domain GPs are centered around the idea of finding a possibly more compact representative set of input features in a domain different from the input data domain. This way, it is possible to incorporate prior knowledge about relevant characteristics of data—such as the presence of global structure—into the inducing variables.

Consider a real-valued GP $f(\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^D$ and some deterministic function $g(\mathbf{x}, \mathbf{Z})$, with M inducing points $\mathbf{Z} \in \mathbb{R}^{M \times H}$. We define the following transformation:

$$u(\mathbf{Z}) = \int_{\mathbb{R}^D} f(\mathbf{x}) g(\mathbf{x}, \mathbf{Z}) \, \mathrm{d}\mathbf{x}. \tag{2}$$

Since $u(\mathbf{Z})$ is obtained through an affine transformation of $f(\mathbf{x})$, $u(\mathbf{Z})$ is also a GP, but may lie in a different domain than $f(\mathbf{x})$ (Lázaro-Gredilla & Figueiras-Vidal, 2009). Inter-domain GPs arise when $f(\mathbf{x})$ and $u(\mathbf{Z})$ are considered *jointly* as a single, augmented GP, as is the case for local inducing points-based approximate inference. The feature extraction function $g(\mathbf{x}, \mathbf{Z})$ used in the integral then defines the transformed domain in which the inducing dataset lies. The inducing variables obtained this way can be seen as projections of the target function $f(\mathbf{x})$ on the feature extraction function over the entire input space (Lázaro-Gredilla & Figueiras-Vidal, 2009). As such, each of the inducing variables is constructed to contain information about the structure of $f(\mathbf{x})$ everywhere in the input space, making them more informative of the stochastic process than local approximations. (Hensman et al., 2018; Lázaro-Gredilla & Figueiras-Vidal, 2009).

In general, the usefulness of inducing variables mostly relies on their covariance with the remainder of the process, which, for inducing points-based approximate inference, is encoded in the vector-valued function

$$\mathbf{k}_{\mathbf{u}}(\mathbf{x}) = [k(\mathbf{z}_1, \mathbf{x}), k(\mathbf{z}_2, \mathbf{x}), ..., k(\mathbf{z}_M, \mathbf{x})].$$

The matrix $\mathbf{K}_{\mathbf{uu}} \overset{\text{def}}{=} K(\mathbf{Z}, \mathbf{Z})$ and the vector-valued function $\mathbf{k}_{\mathbf{u}}(\mathbf{x})$ are central to inducing points-based approximate inference for GPs where they are used to construct an approximate posterior distribution.

3. Inter-domain Deep Gaussian Processes

In this section, we will introduce inter-domain DGPs. First, we will present a general inter-domain DGP framework. Next, we will explain why constructing inter-domain deep GPs is more challenging than constructing inter-domain shallow GPs and how we can leverage the compositional structure of the layer-wise approximate posterior predictive distributions in doubly stochastic variational inference (Salimbeni & Deisenroth, 2017) to obtain simple and scalable inter-domain DGPs. Finally, we will draw on prior work (Hensman et al., 2018) to explicitly incorporate global structure into the inter-domain transformation.

3.1. The Augmented Inter-domain Deep Gaussian Process Model

In inducing points-based approximate inference, the GP model is augmented by a set of inducing variables, $u(\mathbf{Z})$. Unlike conventional inducing points-based approximations, inter-domain approaches do not constrain inducing points to lie in the same domain as the input data.

To distinguish between inducing points that lie in the same domain as the input data and inter-domain inducing points, we diverge from the notation in the previous section and from now on define inter-domain inducing points across DGP layers as $\{\mathbf{\Omega}^{(\ell)}\}_{\ell=0}^{L-1}$ with corresponding inducing variables $\mathbf{u}^{(\ell)} \overset{\text{def}}{=} u(\mathbf{\Omega}^{(\ell-1)})$ for $\ell = 1, ..., L$, where L is the number of DGP layers.

We can then express the augmented DGP joint distribution by

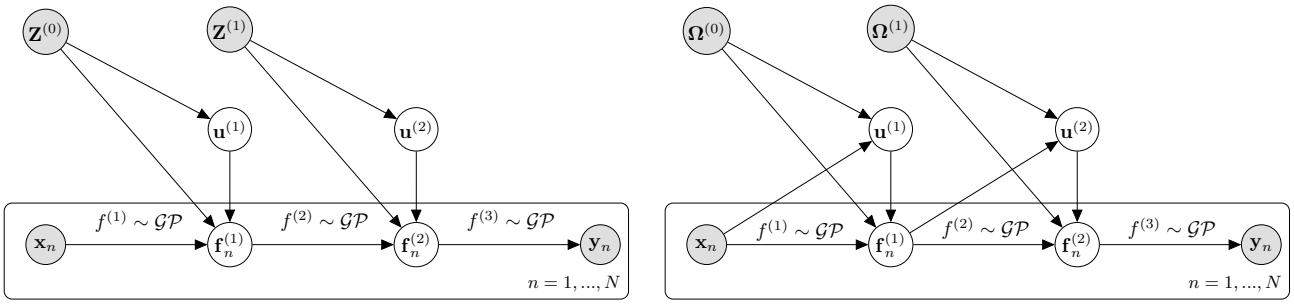
$$\begin{aligned} p(\mathbf{y}, \{\mathbf{f}^{(\ell)}, \mathbf{u}^{(\ell)}\}_{\ell=1}^{L}) = \prod_{n=1}^{N} p(\mathbf{y}_n \,|\, \mathbf{f}_n^{(L)}) \\ \cdot \prod_{\ell=1}^{L} p(\mathbf{f}^{(\ell)} \,|\, \mathbf{u}^{(\ell)}; \mathbf{f}^{(\ell-1)}, \mathbf{\Omega}^{(\ell-1)}) \\ \cdot p(\mathbf{u}^{(\ell)}; \mathbf{f}^{(\ell-1)}, \mathbf{\Omega}^{(\ell-1)}). \end{aligned} \tag{3}$$

Importantly, each $p(\mathbf{u}^{(\ell)}; \mathbf{f}^{(\ell-1)}, \mathbf{\Omega}^{(\ell-1)})$ is being evaluated at a set of inter-domain inducing points $\mathbf{\Omega}^{(\ell-1)}$ but also includes information about $\mathbf{f}^{(\ell-1)}$ via the inter-domain projections. For a graphical representation, see Figure 2b. To avoid overloading notation, we will assume that each GP layer has the same mean and covariance functions $m(\cdot)$ and $k(\cdot, \cdot)$.

For each DGP layer we thus have a transformed-domain instance of the mean function,

$$\begin{aligned} m(\mathbf{\Omega}^{(\ell-1)}) &= \mathbb{E}[u(\mathbf{\Omega}^{(\ell-1)})] \\ &= \int_{\mathbb{R}^D} \mathbb{E}[f^{(\ell)}(\mathbf{f}^{(\ell-1)})] \, g(\mathbf{f}^{(\ell-1)}, \mathbf{\Omega}^{(\ell-1)}) \, \mathrm{d}\mathbf{f}^{(\ell-1)} \\ &= \int_{\mathbb{R}^D} m(\mathbf{f}^{(\ell-1)}) \, g(\mathbf{f}^{(\ell-1)}, \mathbf{\Omega}^{(\ell-1)}) \, \mathrm{d}\mathbf{f}^{(\ell-1)} \\ &\overset{\text{def}}{=} \mathbf{m}_{\mathbf{u}^{\ell}}^{\phi}, \end{aligned} \tag{4}$$

and a transformed-domain instance of the covariance func-



(a) Graphical model representation of a DGP model with local inducing inputs, inducing variables, and two hidden layers, $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$, for $n = 1, ..., N$.

(b) Graphical model representation of an inter-domain DGP model with inducing frequencies, RKHS Fourier feature inducing variables, and two hidden layers, $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$, for $n = 1, ..., N$.

Figure 2: Graphical model representations of local inducing-points DGPs (Figure 2a) and inter-domain DGPs (Figure 2b). Greyed-out nodes denote observed data and non-greyed out nodes denote unobserved data.

tion with

$$
\begin{aligned}
&k(\mathbf{f}^{(\ell-1)}, \mathbf{\Omega}^{(\ell-1)}) \\
=&\mathbb{E}[f^{(\ell)}(\mathbf{f}^{(\ell-1)}) u(\mathbf{\Omega}^{(\ell-1)})] \\
=&\mathbb{E}\left[f^{(\ell)}(\mathbf{f}^{(\ell-1)}) \int_{\mathbb{R}^D} f^{(\ell)}(\mathbf{f}^{(\ell-1)})\, g(\mathbf{f}^{(\ell-1)}, \mathbf{\Omega}^{(\ell-1)})\, \mathrm{d}\mathbf{f}^{(\ell-1)} \right] \\
=&\int_{\mathbb{R}^D} k(\mathbf{f}^{(\ell-1)}, \mathbf{f}^{(\ell-1)'})\, g(\mathbf{f}^{(\ell-1)'}, \mathbf{\Omega}^{(\ell-1)})\, \mathrm{d}\mathbf{f}^{(\ell-1)'} \\
\overset{\text{def}}{=}&\mathbf{K}^{\phi}_{\mathbf{u}^\ell\, \mathbf{u}^\ell}
\end{aligned}
\tag{5}
$$

and

$$
\begin{aligned}
&k(\mathbf{\Omega}^{(\ell-1)}, \mathbf{\Omega}^{(\ell-1)'}) \\
=&\mathbb{E}[u(\mathbf{\Omega}^{(\ell-1)}) u(\mathbf{\Omega}^{(\ell-1)'})] \\
=&\mathbb{E}\Bigg[\int_{\mathbb{R}^D} f^{(\ell)}(\mathbf{f}^{(\ell-1)})\, g(\mathbf{f}^{(\ell-1)}, \mathbf{\Omega}^{(\ell-1)})\, \mathrm{d}\mathbf{f}^{(\ell-1)} \\
&\quad \cdot \int_{\mathbb{R}^D} f^{(\ell)}(\mathbf{f}^{(\ell-1)'})\, g(\mathbf{f}^{(\ell-1)'}, \mathbf{\Omega}^{(\ell-1)'})\, \mathrm{d}\mathbf{f}^{(\ell-1)'} \Bigg] \\
=&\int_{\mathbb{R}^D} \int_{\mathbb{R}^D} k(\mathbf{f}^{(\ell-1)}, \mathbf{f}^{(\ell-1)'})\, g(\mathbf{f}^{(\ell-1)}, \mathbf{\Omega}^{(\ell-1)}) \\
&\quad \cdot g(\mathbf{f}^{(\ell-1)'}), \mathbf{\Omega}^{(\ell-1)'})\, \mathrm{d}\mathbf{f}^{(\ell-1)'} \\
\overset{\text{def}}{=}&\mathbf{K}^{\phi}_{\mathbf{u}^\ell\, \mathbf{u}^\ell},
\end{aligned}
\tag{6}
$$

where we use the superscript ϕ to indicate that the basis functions have the form of the inter-domain instance of the covariance function (Lázaro-Gredilla & Figueiras-Vidal, 2009). Mean and covariance functions at each layer are therefore defined both by the values and domains of their arguments. We can now express the joint distribution in Equation (3) in terms of the layer-wise covariance functions given above and perform inference across domains.

3.2. Simple and Scalable Approximate Inference in Inter-domain Deep Gaussian Processes

Since exact inference in DGPs is intractable, we need approximate inference methods. Unfortunately, most approximate inference methods for DGPs require computing convolutions between $\mathbf{K}_{\mathbf{u}^\ell \mathbf{f}^\ell}$ and the distributions of the latent functions, that is,

$$
\int \mathbf{K}_{\mathbf{u}^\ell \mathbf{f}^\ell} \mathcal{N}(\mathbf{f}^{(\ell)} | \mathbf{m}_{\mathbf{f}^\ell}, \mathbf{S}_{\mathbf{f}^\ell})\, \mathrm{d}\mathbf{f}^{(\ell)}, \tag{7}
$$

where $\mathcal{N}(\mathbf{f}^{(\ell)} | \mathbf{m}_{\mathbf{f}^\ell}, \mathbf{S}_{\mathbf{f}^\ell})$ represents the variational distribution of layer ℓ with mean $\mathbf{m}_{\mathbf{f}^\ell}$ and variance $\mathbf{S}_{\mathbf{f}^\ell}$ (see, for example, pages 50-51 in Damianou (2015) or Damianou & Lawrence (2013), Dai et al. (2014), Bui et al. (2016)). While these convolutions are easy to compute in closed form for conventional inducing points-based approximations where the covariance matrix is computed from the DGP's input-domain covariance function, they are non-trivial to compute analytically for inter-domain covariance functions (Hensman et al., 2018).

To perform approximate inference in inter-domain DGPs, we exploit the fact that—in contrast to previous inducing points-based variational inference methods for DGPs—the layer-wise marginalization over each $\mathbf{f}^{(\ell)}$ in doubly stochastic variational inference (DSVI) (Salimbeni & Deisenroth, 2017) does not require computing convolutions that explicitly depend on the specific type of cross-covariance function $\mathbf{k}^{\phi}_{\mathbf{u}^{(\ell)}}(\mathbf{f}^{\ell})$. Instead, the functional form of the posterior predictive distribution $q(\mathbf{f}^{(L)})$ and the use of the reparameterization trick make marginalizing out the latent GP functions across layers straightforward and result in simple, compositional posterior predictive mean and covariance functions at each DGP layer. For further details on DSVI, see Appendix D.

This property allow us to simply use the inter-domain operators $\mathbf{K}^{\phi}_{\mathbf{u}^\ell \mathbf{f}^\ell}$ as off-the-shelf replacements for the conven-

tional inducing-point operators $\mathbf{K}_{\mathbf{u}^\ell\mathbf{f}^\ell}$ without having to analytically convolve $\mathbf{K}^\phi_{\mathbf{u}^\ell\mathbf{f}^\ell}$ with the distribution over functions at the ℓth layer, yielding the variational distribution

$$q(\{\mathbf{f}^{(\ell)}\}_{\ell=1}^L) = \prod_{\ell=1}^L q(\mathbf{f}^{(\ell)}\,|\boldsymbol{\mu}^{(\ell)}, \boldsymbol{\Sigma}^{(\ell)}; \mathbf{f}^{(\ell-1)}, \boldsymbol{\Omega}^{(\ell-1)}) = \prod_{\ell=1}^L \mathcal{N}(\mathbf{f}^{(\ell)}\,|\widetilde{\mathbf{m}}_{\mathbf{f}^\ell}, \widetilde{\mathbf{S}}_{\mathbf{f}^\ell})$$

with

$$\widetilde{\mathbf{m}}_{\mathbf{f}^\ell} \overset{\text{def}}{=} \mathbf{m}_{\mathbf{f}^\ell} - \mathbf{K}^\phi_{\mathbf{f}^\ell\mathbf{u}^\ell}\mathbf{K}^{\phi^{-1}}_{\mathbf{u}^\ell\mathbf{u}^\ell}(\boldsymbol{\mu}^{(\ell)} - \mathbf{m}^\phi_{\mathbf{u}^\ell}),$$
$$\widetilde{\mathbf{S}}_{\mathbf{f}^\ell} \overset{\text{def}}{=} \mathbf{K}_{\mathbf{f}^\ell\mathbf{f}^\ell} - \mathbf{K}^\phi_{\mathbf{f}^\ell\mathbf{u}^\ell}\mathbf{K}^{\phi^{-1}}_{\mathbf{u}^\ell\mathbf{u}^\ell}(\mathbf{K}^\phi_{\mathbf{u}^\ell\mathbf{u}^\ell} - \boldsymbol{\Sigma}^{(\ell)})\mathbf{K}^{\phi^{-1}}_{\mathbf{u}^\ell\mathbf{u}^\ell}\mathbf{K}^\phi_{\mathbf{u}^\ell\mathbf{f}^\ell}, \tag{8}$$

where $\mathbf{m}_{\mathbf{f}^\ell} \overset{\text{def}}{=} m(\mathbf{f}^{(\ell-1)})$ and $\mathbf{m}_{\mathbf{u}^\ell} \overset{\text{def}}{=} m(\boldsymbol{\Omega}^{(\ell-1)})$, and $\boldsymbol{\mu}^{(\ell)}$ and $\boldsymbol{\Sigma}^{(\ell)}$ are variational parameters. Since DSVI uses the reparameterization trick to sample functions at each layer, the inter-domain operators can be used directly to compute the posterior mean and variance for each layer, which allows for simple and scalable approximate inference in inter-domain DGPs.

3.3. RKHS Fourier Features for Approximate Inference in Gaussian Processes

In the previous section, we showed how to perform approximate inference in inter-domain DGPs with any inter-domain operators $\mathbf{k}^\phi_{\mathbf{u}}(x)$ and $\mathbf{K}^\phi_{\mathbf{uu}}$. Next, we will introduce RKHS Fourier features (Hensman et al., 2018)—an inter-domain approach able to capture global structure in data—and show how to incorporate them into inter-domain DGPs.

RKHS Fourier features use RKHS theory to construct inter-domain alternatives to the covariance matrices $\mathbf{K}_{\mathbf{uu}}$ and $\mathbf{k}_{\mathbf{u}}(x)$ used in conventional inducing points-based approximate inference methods. They are constructed by projecting the target function f onto the truncated Fourier basis

$$\boldsymbol{\phi}(x) = [1, \cos(\omega_1(x-a)), ..., \cos(\omega_M(x-a)), \sin(\omega_1(x-a)), ..., \sin(\omega_M(x-a))]^\top, \tag{9}$$

where x is a single, one-dimensional input, and $[\omega_1, ..., \omega_M]$ denote inducing *frequencies* defined by $\omega_m = \frac{2\pi m}{b-a}$ for some interval $[a, b]$. From this truncated Fourier basis, we can construct inducing variables as inter-domain projections by defining $u_m \overset{\text{def}}{=} \mathcal{P}_{\boldsymbol{\phi}_m}(f)$, which can be shown to yield transformed-domain instances of the covariance function given by

$$\text{cov}(u_m, f(x)) = \boldsymbol{\phi}_m(x), \quad \text{cov}(u_m, u_{m'}) = \langle \boldsymbol{\phi}_m, \boldsymbol{\phi}_{m'} \rangle_{\mathcal{H}},$$

for both of which there are closed-form expressions if the GP prior covariance function is given by a half-integer member of the Matérn family of kernels (Durrande et al., 2016). For further details, see Hensman et al. (2018). The resulting inter-domain operators

$$\mathbf{k}^\phi_{\mathbf{u}}(x) = \boldsymbol{\phi}_m(x), \qquad \mathbf{K}^\phi_{\mathbf{uu}} = \langle \boldsymbol{\phi}_m, \boldsymbol{\phi}_{m'} \rangle_{\mathcal{H}},$$

represent inter-domain alternatives to the $\mathbf{k}_{\mathbf{u}}(x)$ and $\mathbf{K}_{\mathbf{uu}}$ operators used in local inducing points-based approximations. By constructing linear combinations of the values of the data-generating process as projections instead of simple function evaluations, the resulting inducing variables become more informative of the underlying process and have more capacity to represent complex functions (Hensman et al., 2018; Lázaro-Gredilla & Figueiras-Vidal, 2009). Analogous to the way in which local inducing points-based approaches approximate the DGP posterior distribution through kernel functions, RKHS Fourier features approximate the posterior through sinusoids (Hensman et al., 2018). The structure imposed by the frequency domain makes RKHS Fourier features particularly well-suited to capture global structure in data. For further details on RKHS Fourier features, see Appendix C.

3.4. Inter-domain Deep Gaussian Processes with RKHS Fourier Features

To construct inter-domain DGPs that leverage global structure in data, we use approximate posterior predictive distributions based on RKHS Fourier Features at every layer. For layers $\ell = 1, ..., L$ with input dimensions $D^{(\ell-1)}$, let $\omega_m = \frac{2\pi m}{b-a}$ for $m = 1, ..., M$, and let

$$\boldsymbol{\Omega}^{(\ell-1)} \overset{\text{def}}{=} [\boldsymbol{\omega}_1^{(\ell-1)}, ..., \boldsymbol{\omega}_M^{(\ell-1)}]^\top$$

be the matrix of $M \times D^{(\ell-1)}$ inducing frequencies producing a set of $D^{(\ell-1)}$ truncated Fourier bases $\boldsymbol{\phi}^{(\ell)}(\mathbf{f}^{(\ell-1)})$, as defined in Equation (9). Each $\boldsymbol{\phi}^{(\ell)}(\mathbf{f}^{(\ell-1)})$ then maps $f^{(\ell)}$ into Fourier space by applying the RKHS inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ given by

$$u_m^{(\ell)} \overset{\text{def}}{=} \mathcal{P}_{\boldsymbol{\phi}_m^{(\ell)}}(f^{(\ell)}) = \langle \boldsymbol{\phi}_m^{(\ell)}, f^{(\ell)} \rangle_{\mathcal{H}},$$

for Fourier basis entries $\boldsymbol{\phi}_m^{(\ell)}(\mathbf{f}^{(\ell-1)})$ with $m = 1, ..., M'$ and $M' = 2M + 1$ (as in the shallow GP case), thus creating the $M' \times D^{(\ell)}$-dimensional matrix

$$\mathbf{u}^{(\ell)} = [\mathcal{P}_{\boldsymbol{\phi}_1^{(\ell)}}(f), ..., \mathcal{P}_{\boldsymbol{\phi}_{M'}^{(\ell)}}(f)]^\top.$$

We thus obtain inter-domain operators $\mathbf{k}^\phi_{\mathbf{u}^\ell}(\mathbf{f}^{(\ell-1)})$ and $\mathbf{K}^\phi_{\mathbf{u}^\ell\mathbf{u}^\ell}$ for DGP layers $\ell = 1, ..., L$.

Using the variational distribution in Equation (8), we then get a final-layer posterior predictive distribution

$$q(\mathbf{f}_n^{(L)}) = \int \prod_{\ell=1}^{L-1} q(\mathbf{f}_n^{(\ell)}\,|\boldsymbol{\mu}^{(\ell)}, \boldsymbol{\Sigma}^{(\ell)}; \mathbf{f}_n^{(\ell-1)}, \boldsymbol{\Omega}^{(\ell-1)})\,\text{d}\mathbf{f}_n^{(\ell)},$$

where $\mathbf{f}_n^{(\ell)}$ is the nth row of $\mathbf{f}^{(\ell)}$. This quantity is easy to compute using the reparameterization trick, which allows for sampling from the nth instance of the variational posteriors across layers by defining

$$\hat{\mathbf{f}}_n^{(\ell)} = \widetilde{\mathbf{m}}\left(\hat{\mathbf{f}}_n^{(\ell-1)}\right) + \boldsymbol{\epsilon}_n^{(\ell)} \odot \sqrt{\widetilde{\mathbf{S}}\left(\hat{\mathbf{f}}_n^{(\ell-1)}, \hat{\mathbf{f}}_n^{(\ell-1)}\right)} \quad (10)$$

and sampling from $\boldsymbol{\epsilon}_n^{(\ell)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D^{(\ell)}})$ (Kingma & Welling, 2014; Salimbeni & Deisenroth, 2017).

Prediction To make predictions, we sample from the approximate posterior predictive distribution of the final layer the same way as in DSVI. For a test input $\mathbf{x}^*$, we draw S samples from the posterior predictive distribution

$$q(\mathbf{f}_*^{(L)}) \approx \frac{1}{S}\sum_{s=1}^{S} q(\mathbf{f}_*^{(L)} \,|\boldsymbol{\mu}^{(L)}, \boldsymbol{\Sigma}^{(L)}; \mathbf{f}_*^{(s)(L-1)}, \boldsymbol{\Omega}^{(L-1)}) \quad (11)$$

where $q(\mathbf{f}_*^{(L)})$ is the DGPs marginal distribution at $\mathbf{x}_*$ and $\mathbf{f}_*^{(s)(L-1)}$ are draws from the penultimate layer (and thus indirectly from all previous layers) obtained via reparameterization of each layer as shown in Equation (10).

Evidence Lower Bound The evidence lower bound (ELBO) is the same as in DSVI, apart from the fact that it is computed from the inter-domain posterior predictive distributions at each DGP layer. It is given by

$$\begin{aligned}\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{q(\mathbf{f}_n^{(L)})}\left[\log p(\mathbf{y}_n \,|\, \mathbf{f}_n^{(L)})\right] \\ - \sum_{\ell=1}^{L} \mathrm{KL}(q(\mathbf{u}^{(\ell)}) \,||\, p(\mathbf{u}^{(\ell)})),\end{aligned} \quad (12)$$

which can be optimized variationally using gradient-based stochastic optimization. We include a derivation of this bound in Appendix E. To estimate the expected log-likelihood, we generate predictions at the input locations by drawing Monte Carlo samples from $q(\mathbf{f}_n^{(L)})$ as shown in Equation (11).

3.5. Further Model Details

In our implementation, we let

$$\begin{aligned}\boldsymbol{\omega}_{m,i}^{(\ell-1)} &= \boldsymbol{\omega}_{m,j}^{(\ell-1)} \,\forall i, j \le D^{(\ell-1)} \\ \boldsymbol{\omega}_m^{(\ell)} &= \boldsymbol{\omega}_m^{(\ell')} \;\forall \ell, \ell' \in \{1, ..., L\} \,\forall m \in \{1, ..., M'\},\end{aligned} \quad (13)$$

which means that we use the same inducing frequencies at every DGP layer, but this assumption can be relaxed easily. Moreover, we use additive kernels to apply RKHS Fourier features to multidimensional inputs. For for each layer, we define

$$\begin{aligned}f^{(\ell)}(\mathbf{f}^{(\ell-1)}) &= \sum_{d=1}^{D^\ell} f_d^{(\ell)}(\mathbf{f}_d^{(\ell-1)}) \\ f_d^{(\ell)} &\sim \mathcal{GP}\left(0, k_d^{(\ell)}\left(\mathbf{f}_d^{(\ell-1)}, \mathbf{f}_d^{(\ell-1)'}\right)\right),\end{aligned} \quad (14)$$

where $\mathbf{f}_d^{(\ell-1)}$ is the dth element of the multi-dimensional single input $\mathbf{f}_n^{(\ell-1)}$, and $k_p^{(\ell)}(\cdot,\cdot)$ is a kernel defined on a scalar input space (Hensman et al., 2018). This way, we obtain the DGP layer

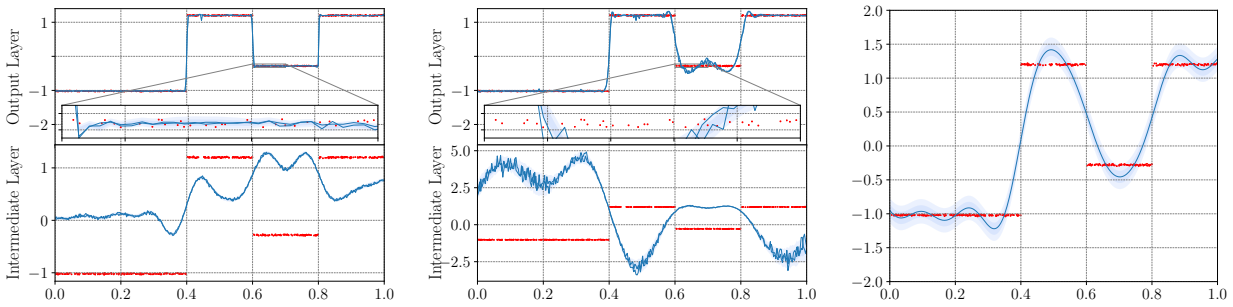
$$f^{(\ell)} \sim \mathcal{GP}\left(0, \sum_{d=1}^{D^\ell} k_d^{(\ell)}\left(\mathbf{f}_d^{(\ell-1)}, \mathbf{f}_d^{(\ell-1)'}\right)\right) \quad (15)$$

for which we are then able to construct a matrix of features with elements $u_{m,d}^{(\ell)} = \mathcal{P}_{\phi_m}(f_d^{(\ell)})$, resulting in a total of $2MD^{(\ell)} + 1$ inducing variables, independent across dimensions, i.e., $\mathrm{cov}(u_{m,d}^{(\ell)}, u_{m,d'}^{(\ell)}) = 0$. With the corresponding variational parameters estimated via gradient-based optimization, the cost per iteration when computing the posterior mean for an additive kernel is $\mathcal{O}(NM^2D)$. Using an additive kernel at each DGP layer then results in a time complexity of $\mathcal{O}(NM^2(D^{(1)} + D^{(2)} + ... + D^{(L)}))$ per iteration, which is identical to that of DSVI. In practice, however, we find that inter-domain DGPs require fewer inducing points and fewer gradient steps to achieve a given level of predictive accuracy compared to DGPs with DSVI, making them more computationally efficient.

Unlike DGPs that use conventional inducing-points based approximate inference, inter-domain DGPs have an additional hyperparameter; the frequency interval $[a, b]$. To avoid undesirable edge effects in the DGP posterior predictive distributions, we normalize all input data dimensions to lie in the interval $[0, 1]$ and define the RKHS over the interval $[a, b] = [-2, 3]$. We repeat this normalization at each DGP layer before feeding the samples into the next GP. To avoid pathologies in DGP models investigated in prior work (Duvenaud et al., 2014), we follow Salimbeni & Deisenroth (2017) and use a linear mean function $m^{(\ell)}(\mathbf{f}^{(\ell-1)}) = \mathbf{f}^{(\ell-1)}\,\mathbf{w}^{(\ell)}$, where $\mathbf{w}^{(\ell)}$ is a vector of weights, for all but the final-layer GP, for which we use a zero mean function. We used a Matérn-$\frac{3}{2}$ kernel for all experiments.

4. Related Work

Inducing points-based approximate inference has allowed GPs models to scale to large numbers of input points (Snelson & Ghahramani, 2006; Titsias, 2009; Hensman et al., 2013; Bui & Turner, 2014; Hensman et al., 2015). Our work directly builds on Hensman et al. (2018) and Salimbeni &



(a) Inter-domain DGP with DSVI (two layers). Top: DGP posterior predictive distribution. Bottom: Predictive distribution at intermediate layer.

(b) Conventional DGP with DSVI (two layers). Top: DGP posterior predictive distribution. Bottom: Predictive distribution at intermediate layer.

(c) GP with RKHS Fourier features (single layer). Posterior predictive distribution.

Figure 3: Comparison of posterior predictive distributions of different GP models on synthetic non-stationary data. The models are trained using 20 inducing frequencies and 20 inducing points, respectively. In each plot, training points are shown in red. Each shade of blue represents one standard deviation in the posterior predictive distribution. For enlarged plots, see Appendix B.

Deisenroth (2017) and adds to the literature on sparse spectrum approximations (Lázaro-Gredilla & Figueiras-Vidal, 2009; Lázaro-Gredilla et al., 2010; Gal & Turner, 2015; Wilson & Nickisch, 2015). Specifically, we extend Hensman et al. (2018) to compositions GP models by leveraging the compositional structure of the approximate posterior of Salimbeni & Deisenroth (2017). In contrast to Wilson & Nickisch (2015), Lázaro-Gredilla & Figueiras-Vidal (2009), and Gal & Turner (2015), Hensman et al. (2018) (and, by extension, our approach) combines inter-domain operators with SVI (Hensman et al., 2013) and is amenable to stochastic optimization on minibatches, which makes it possible to apply it to large datasets without facing memory constraints. Similar to our approach, random feature expansions for DGPs (Cutajar et al., 2017) use projections of each DGP layer's predictive distribution onto the spectral domain to perform approximate inference, but unlike our approach, it is not based on inducing points.

5. Empirical Evaluation

To demonstrate that inter-domain DGPs improve upon inter-domain shallow GPs in their ability to model complex, non-stationary data and to show that inter-domain DGPs improve upon local inducing points-based approximate inference methods for DGPs, we will present results from several experiments that showcase the types of prediction problems for which inter-domain DGPs are particularly well-suited. We are particularly interested in modeling complex data-generating processes which exhibit global structure as well as non-stationarity, since the former is challenging for DGPs that use local approximations, such as DSVI for DGPs, and the latter is challenging for shallow GPs with stationary covariance functions.

To illustrate the advantage of inter-domain deep GPs over inter-domain shallow GPs in modeling non-stationary data, we present a suite of qualitative and quantitative empirical evaluations on datasets that exhibit global structure and non-stationarity.

First, we present a simple, synthetic data experiment designed to demonstrate that our method is well-suited for modeling data from generating processes that exhibit both non-stationarity and global structure. Next, we illustrate that inter-domain deep GPs provide a significant gain in computational efficiency when modeling data that exhibits global structure. In particular, we compare the number of inducing frequencies and inducing points needed to attain a certain predictive accuracy when using inter-domain DGPs and local inducing points-based DGPs on a challenging real-world audio sub-band reconstruction task. Lastly, we demonstrate that our method outperforms existing state-of-the-art shallow GPs with local approximate inference, shallow GPs with global approximate inference, and deep GPs with local approximate inference on a series of challenging real-world benchmark prediction tasks. For additional experiments and more experimental details, see Appendix B.

5.1. Highly Non-Stationary Data with Global Structure

The multi-step function in Figure 3 is designed to exhibit *both* global structure *as well as* non-stationarity, providing an optimal test case to assess the performance of inter-domain DGPs vis-à-vis related methods on a simple and easily interpretable prediction task.

The plot shows the posterior predictive distributions of inter-domain DGPs, DGPs with DSVI, and inter-domain shallow GPs with RKHS Fourier features. As can be seen in the plots, inter-domain DGPs are the only method that is able to model the step locations well and to infer the global structure—that is, that the function is constant within certain intervals—

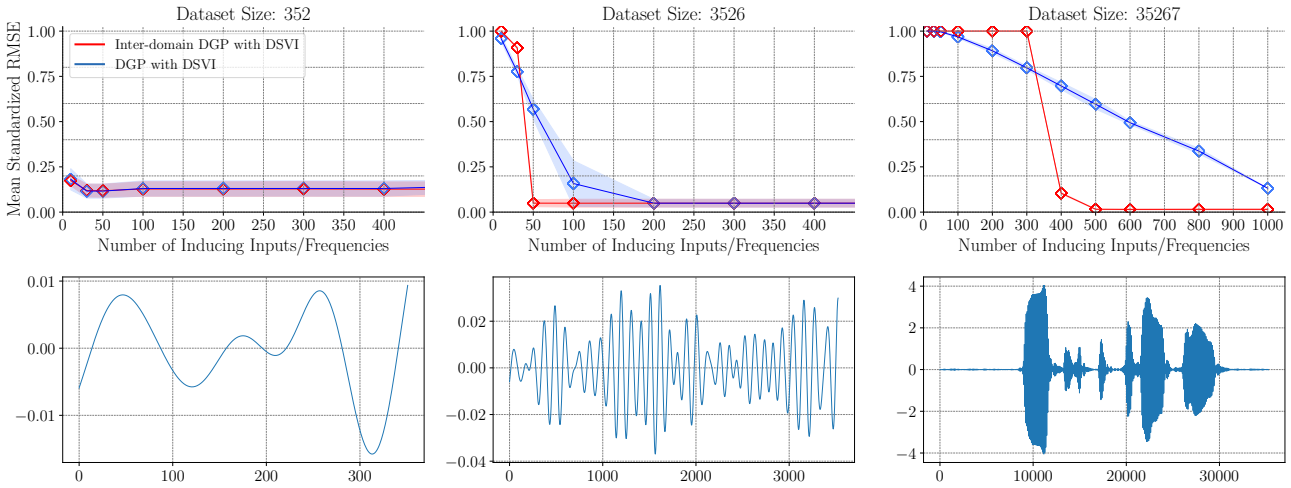


Figure 4: Comparison of average standardized root mean squared errors for varying numbers of inducing inputs on three datasets of increasing global structure and complexity. On complex datasets (center and right panel), Inter-domain deep GPs with DSVI require fewer inducing inputs than conventional DGPs with DSVI. Standardized root mean squared errors were evaluated on a test set of 40% of datapoints in each subset over 10 random seeds each.

with high accuracy and good predictive uncertainty–despite having a stationary covariance function (see Figure 3a). Inter-domain shallow GPs, in contrast, are unable to capture either the step transitions nor the global structure, reflecting their limited expressiveness (see Figure 3c). While DGPs benefit from increased expressivity, they, too, fail to fully capture the global structure and the non-stationarity (see Figure 3b). This is due to the inherently local nature of local inducing points-based inference, which requires large numbers of inducing inputs to accurately approximate complex posterior distributions. The experiment illustrates that inter-domain DGPs are in fact able to overcome key limitations of both shallow GP inter-domain approaches and outperform state-of-the-art local DGP inference methods.

Figure 3 presents another interesting insight into differences between conventional and inter-domain DGPs. In particular, the bottom plots in Figure 3a and Figure 3b, show the output of the DGP intermediate layers and are markedly different from one another. While it appears that the conventional DGP seeks to model the target function directly in intermediate-layer space, the inter-domain DGP appears to cluster datapoints from the original input space in a way such that the changes in the step function in output space become associated with smooth transitions in intermediate-layer space.

5.2. Modeling Complex Data Efficiently via Global Structure

Next, we quantitatively assess the predictive accuracy and computational efficiency of inter-domain DGPs. To do so, we use a smoothed sub-band of a speech signal taken from the TIMIT database and previously used in Bui & Turner (2014). The dataset exhibits complex global structure which is difficult to model using local approximation methods. To assess how well inter-domain DGPs are able to capture global structure in the data, we compare it to a doubly stochastic variational inference for DGPs, a state-of-the-art approximate inference method for DGPs based on local inducing points. To assess how well different approximate inference methods are able to capture the complex global structure, we look at three subset of the data: the first 352 datapoints, the first 3,526 datapoints, and the first 35,267 datapoints.

The smallest subset of only 352 datapoints does not exhibit much global structure and is small enough to be modeled with few (local) approximations, which is reflected by the left panel in Figure 4, where inter-domain DGPs and conventional DGPs perform equally well, and increasing the number of inducing frequencies/points does not lead to an improvement in performance. As we increase the size of the dataset to 3,526 datapoints, however, the global structure–measurable by a high degree of autocorrelation in the data–becomes readily apparent. As can be seen in the top row, inter-domain DGPs require relatively fewer inducing frequencies compared to conventional DGPs to achieve a test error close to zero. The difference in the number of inducing points required to model the data is most significant for the largest subset, shown in the right panel of Figure 4. As can be seen in the plot, the covariance of the process varies significantly. While this subset of the audio sub-band dataset is highly non-stationary, it does exhibit global structure in the shape of repeating patterns in output space. As a result, inter-domain deep GPs are able to attain a test error close

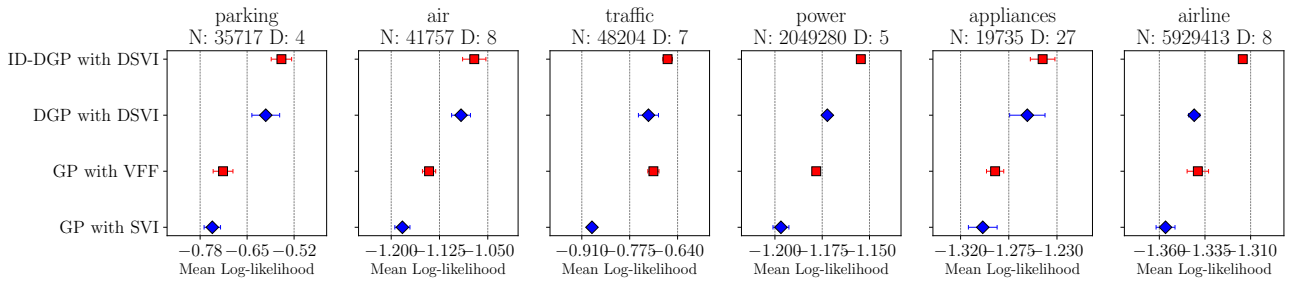


Figure 5: Average test log-likelihood (higher is better) and standard errors (over 10 random seeds) on a set of real-world datasets with global structure. All models were trained with 50 inducing points. The inter-domain DGP with DSVI has two layers and the conventional DGP with DSVI has four layers. The performance of the inter-domain DGP did not increase as additional layers were added.

to zero with fewer than half the number of inducing points needed for conventional DGPs to achieve the same level of accuracy.

Since the time complexity of DSVI scales quadratically in the number of inducing points and inter-domain and conventional DGPs have the same time complexity (that is, a single gradient step takes approximately equally long for the same number of inducing frequencies/points), inter-domain DGPs are more computationally efficient in practice when modeling data exhibiting global structure. Additionally, in Figure 1, we also show that for 20 inducing frequencies/points, inter-domain DGPs have better-calibrated posterior predictive uncertainty estimates than conventional DGPs.

5.3. Global Structure in Real-World Data

To quantitatively assess the predictive performance of inter-domain DGPs, we evaluate them on a range of real-world dataset, which exhibit global structure—usually in the form of a temporal component that induces a high autocorrelation. The experiments include medium-sized datasets ('parking', 'air', 'traffic'), two very large datasets with over two and five million datapoints each ('power' and 'airline'), and a high-dimensional dataset with 27 input dimensions ('appliances'). As can be seen in Figure 5, inter-domain DGPs consistently outperform conventional DGPs (DGPs with DSVI) as well as inter-domain shallow GPs (GPs with VFF) and significantly outperform conventional shallow GPs (SVI), suggesting that combining the increased expressivity of DGP models with the ability of inter-domain approaches to capture global structure leads to the best predictive performance. See Appendix B for a plot of the test standardized RMSEs for the experiments in Figure 5 and for additional results on datasets that do not exhibit global structure (and on which our method performs on par with existing methods).

To assess the predictive performance of inter-domain DGPs on extremely complex, non-stationary data, we test our method on the U.S. flight delay prediction problem, a large-scale regression problem that has reached a status of a standard test in GP regression due to its massive size of $5,929,413$ observations and its non-stationary nature, which makes it challenging for GPs with stationary covariance functions (Hensman et al., 2018). The data set consists of flight arrival and departure times for every commercial flight in the United States for the year 2008. We predict the delay of the aircraft at landing (in minutes) from eight covariates: the age of the aircraft (number of years since deployment), route distance, airtime, departure time, arrival time, day of the week, day of the month, and month. The non-stationarity in the data is likely due to the recurring daily, weekly, and monthly fluctuations in occupancy. In our evaluation, we find that the predictive performance of inter-domain DGPs is superior to closely-related state-of-the-art shallow and deep GPs as shown in Table 1 and Figure 5.

Table 1: Average standardized root mean squared errors and standard errors (over 10 random seeds) on the U.S. flight delay prediction task.

N	1,000,000	5,929,413
Method	RMSE $\pm$ SE	RMSE $\pm$ SE
GP with SVI (local)	0.946 ± 0.008	0.941 ± 0.005
GP with VFF (global)	0.925 ± 0.007	0.923 ± 0.006
DGP with DSVI (local)	0.932 ± 0.004	0.930 ± 0.003
DGP with DSVI (global)	$\mathbf{0.906 \pm 0.006}$	$\mathbf{0.903 \pm 0.002}$

6. Conclusion

We proposed *Inter-domain Deep Gaussian Processes* as a deep extension of inter-domain GPs that combines the advantages of inter-domain and deep GPs and allows us to model data exhibiting non-stationarity *and* global structure with high predictive accuracy and low computational overhead. We showed how to leverage the compositional nature of the approximate posterior in DSVI to perform simple and scalable approximate inference and established that inter-domain DGPs can be more computationally efficient than conventional DGPs. Finally, we demonstrated that our method significantly and consistency outperforms inter-domain shallow GPs and conventional DGPs on data exhibiting non-stationarity and global structure.

Acknowledgements

Tim G. J. Rudner is funded by the Rhodes Trust and the Engineering and Physical Sciences Research Council (EPSRC). We would like to thank Stephen Roberts, James Hensman, Andreas Damianou, Zhenwen Dai, and Neil Lawrence for helpful discussions.

References

Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Proba- bility and Statistics*. Kluwer Academic Publishers, 2004.

Bui, T., Hernandez-Lobato, D., Hernandez-Lobato, J., Li, Y., and Turner, R. Deep Gaussian processes for regression using approximate expectation propagation. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1472–1481, 20–22 Jun 2016.

Bui, T. D. and Turner, R. E. Tree-structured gaussian process approximations. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2213–2221. 2014.

Cutajar, K., Bonilla, E. V., Michiardi, P., and Filippone, M. Random feature expansions for deep Gaussian processes. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 884–893, 06–11 Aug 2017.

Dai, Z., Damianou, A., Hensman, J., and Lawrence, N. Gaussian process models with parallelization and GPU acceleration. *arXiv preprint arXiv:1410.4984*, 2014.

Dai, Z., Damianou, A. C., González, J., and Lawrence, N. D. Variational auto-encoded deep Gaussian processes. *CoRR*, abs/1511.06455, 2015.

Damianou, A. Deep Gaussian processes and variational propagation of uncertainty. *PhD Thesis, University of Sheffield*, 2015.

Damianou, A. and Lawrence, N. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pp. 207–215, 29 April 2013.

Durrande, N., Hensman, J., Rattray, M., and Lawrence, N. D. Detecting periodicities with Gaussian processes. *PeerJ Computer Science*, 4:0, 00 2016.

Duvenaud, D., Rippel, O., Adams, R., and Ghahramani, Z. Avoiding pathologies in very deep networks. In *Artificial Intelligence and Statistics*, pp. 202–210, 2 April 2014.

Gal, Y. and Turner, R. Improving the Gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs. In *International Conference on Machine Learning*, pp. 655–664, 1 June 2015.

Havasi, M., Hernández-Lobato, J. M., and Murillo-Fuentes, J. J. Inference in deep gaussian processes using stochastic gradient hamiltonian monte carlo, 2018.

Hensman, J. and Lawrence, N. D. Nested variational compression in deep Gaussian processes. *arXiv preprint arXiv:1412.1370*, 2014.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013*, 2013.

Hensman, J., de G. Matthews, A. G., and Ghahramani, Z. Scalable Variational Gaussian Process Classification. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12*, 2015.

Hensman, J., Durrande, N., and Solin, A. Variational fourier features for gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52, 2018.

Kingma, D. P. and Welling, M. Stochastic gradient VB and the variational auto-encoder. In *Second International Conference on Learning Representations, International Conference on Learning Representations*, 2014.

Lázaro-Gredilla, M. and Figueiras-Vidal, A. R. Inter-domain Gaussian processes for sparse inference using inducing features. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, pp. 1087–1095, 2009.

Lázaro-Gredilla, M., Candela, J. Q., Rasmussen, C. E., and Figueiras-Vidal, A. R. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11: 1865–1881, 2010.

Mattos, C. L. C., Dai, Z., Damianou, A., Forth, J., Barreto, G. A., and Lawrence, N. D. Recurrent gaussian processes, 2015.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pp. 1177–1184. 2008.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

Salimbeni, H. and Deisenroth, M. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.

Salimbeni, H., Dutordoir, V., Hensman, J., and Deisenroth, M. Deep Gaussian processes with importance-weighted variational inference. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5589–5598, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/salimbeni19a.html.

Snelson, E. and Ghahramani, Z. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18*, pp. 1257–1264. MIT Press, 2006.

Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 567–574, 16–18 Apr 2009.

Wilson, A. and Nickisch, H. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International Conference on Machine Learning*, pp. 1775–1784, 2015.

Supplementary Material

A. Project Website

For source code and additional results, see `https://bit.ly/inter-domain-dgps`.

B. Further Experiments & Experimental Details

In addition to the experiments presented in Section 5, we performed several qualitative and quantitative evaluations to better understand the properties and training behavior of *Inter-domain* DGPs.

In this section, we include plots that provide insights into the effect of adding additional layers to an Inter-domain DGP (see Figure 3a and Figure 3b) and plot draws from the inter-domain and conventional DGP priors. We also include plots of the posterior predictive distributions of inter-domain deep GPs, conventional DGPs, and inter-domain shallow GPs (see Figure 1) as well as standardized RMSEs for the real-world experiments presented in Figure 5 in the main paper (see Figure 6). Furthermore, we include average test root mean squared errors and log-likelihoods for a selection of datasets that do *not* exhibit global structure.

B.1. Training Details

Model For DSVI-DGPs, we set the number of hidden units per layer equal to the number of input dimensions. We used both the RBF and the Matérn-$\frac{3}{2}$ kernels with automatic relevance determination (ARD) for all experiments, but models with RBF kernel performed better.

Training For the benchmark deep GP models, we used learning rates suggested by the authors as well as 10^{-2} and 10^{-3} for all experiments. For inter-domain and conventional DGPs with DSVI, we used *Adam* optimizer with learning rates of 10^{-2} and 10^{-3} for all regression tasks. For benchmark shallow GP models, we used the BFGS algorithm. For all other models, we used the implementation default optimizer.

Parameter initializations For both inter-domain and conventional DGPs with DSVI, we initialized the inducing function value means to zero and variances to the identity and 10^{-5} for outer and inner layers, respectively. For all inducing points-based methods, we initialized the inducing inputs using the K-means algorithm on the training inputs.

B.2. Experiments

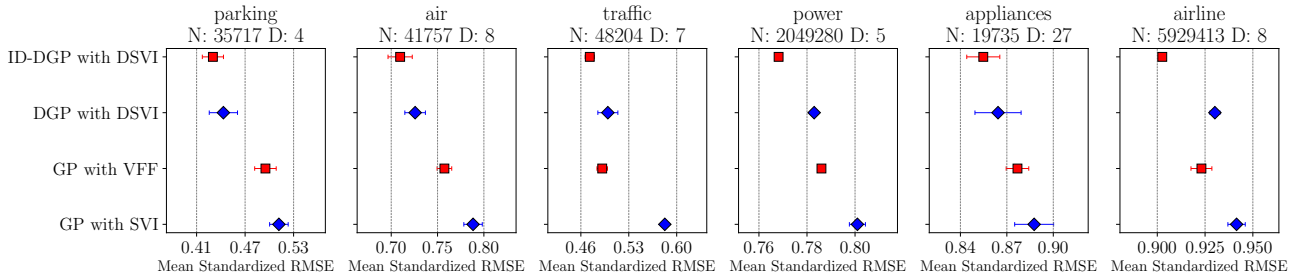


Figure 6: Average standardized root mean squared error (lower is better) and standard errors (over 10 random random seeds) on a set of real-world datasets exhibiting global structure. All models were trained with 50 inducing points.

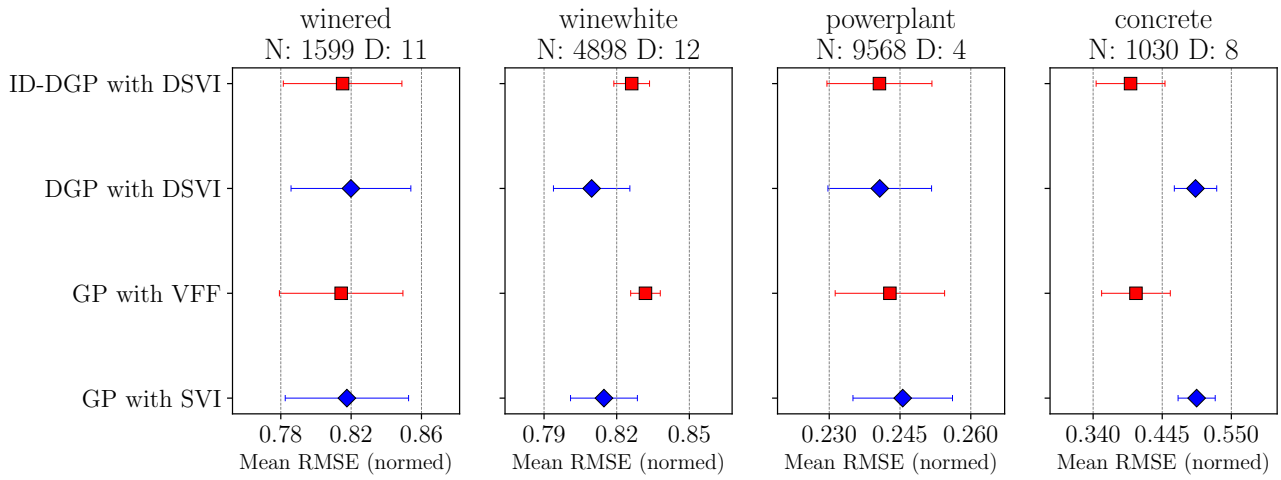


Figure 7: Average standardized root mean squared errors (lower is better) and standard errors (over 10 random seeds) on a set of small- and medium-scale regression problems. Each model was trained with 20 inducing points/inducing frequencies.

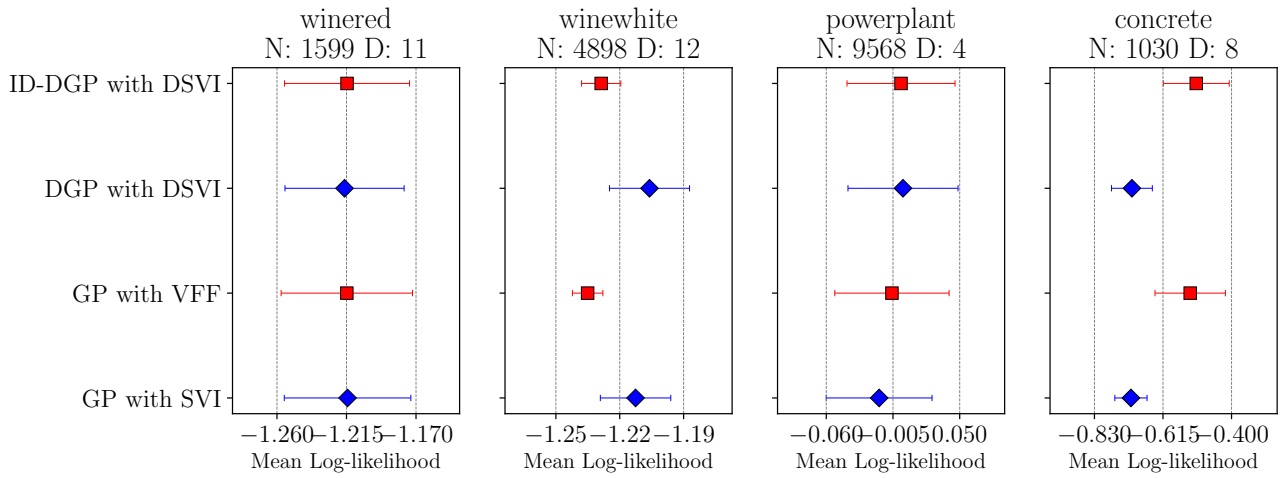


Figure 8: Average test log-likelihoods (higher is better) and standard errors (over 10 random seeds) on a set of small- and medium-scale regression problems. Each model was trained with 20 inducing points/inducing frequencies.

Figure 9 and Figure 10 are enlarged versions of the plots in the main paper. As can be seen from the plots, Inter-domain DGPs with DSVI outperform conventional DGPs with DSVI in modeling global structure, here exemplified by the different plateaus in the step function.

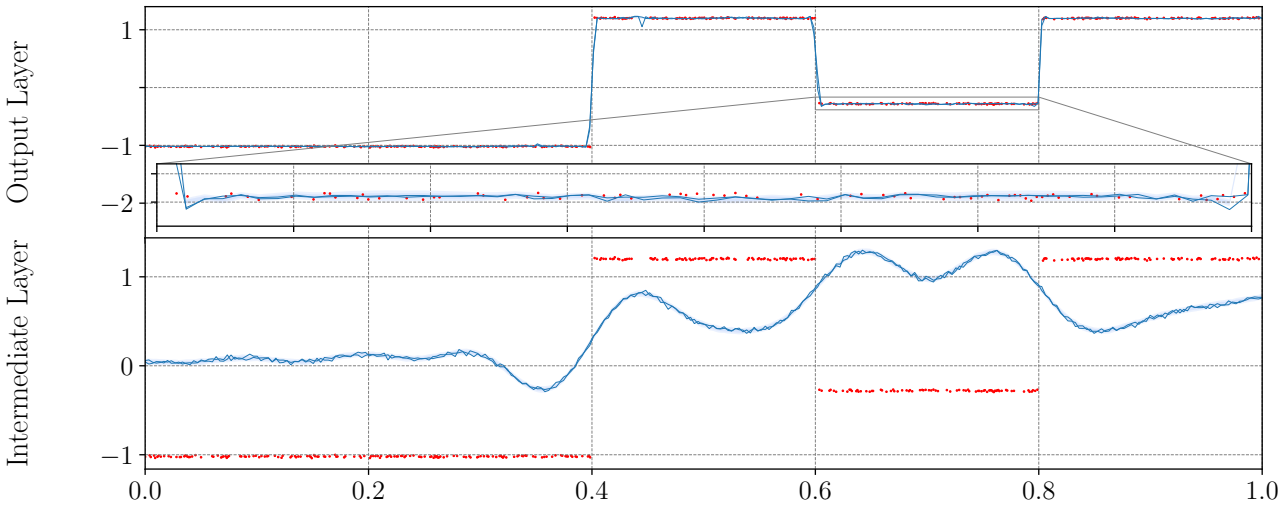


Figure 9: Inter-domain DGP with DSVI (two layers). Top: DGP posterior predictive distribution. Bottom: Marginal distribution at intermediate layer. The model is trained using 20 inducing frequencies. Training points are shown in red. Each shade of blue represents one standard deviation in the posterior predictive distribution.

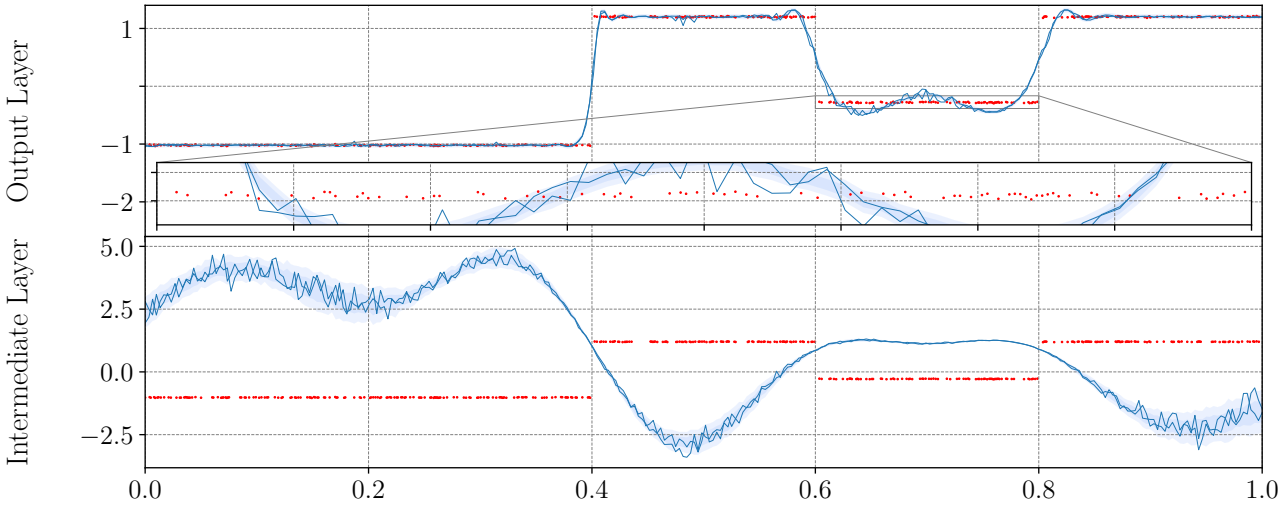


Figure 10: Conventional DGP with DSVI (two layers). Top: DGP posterior predictive distribution. Bottom: Marginal distribution at intermediate layer. The model is trained using 20 inducing frequencies. Training points are shown in red. Each shade of blue represents one standard deviation in the posterior predictive distribution.

Figure 11 and Figure 12 show the outputs of individual DGP layer for inter-domain DGPs with DSVI and conventional DGPs with DSVI, respectively. As can be seen from the plots, both penultimate layers (i.e., the 2nd layers), approximately reflect the shape of the data and of the output of the final layer. Notably, both penultimate layers appear to be (vertically) scaled versions of the output layer, which suggest that adding additional layers allow the model to 'approach' the function it is trying to model slowly with each GP composition. Comparing the output layer predictions in Figure 9 and Figure 11, however, we do not observe a significant difference. One notably difference between Figure 11 and Figure 12, however, is that the first-layer output of the inter-domain DGP is non-monotone, whereas that of the conventional DGP roughly is.

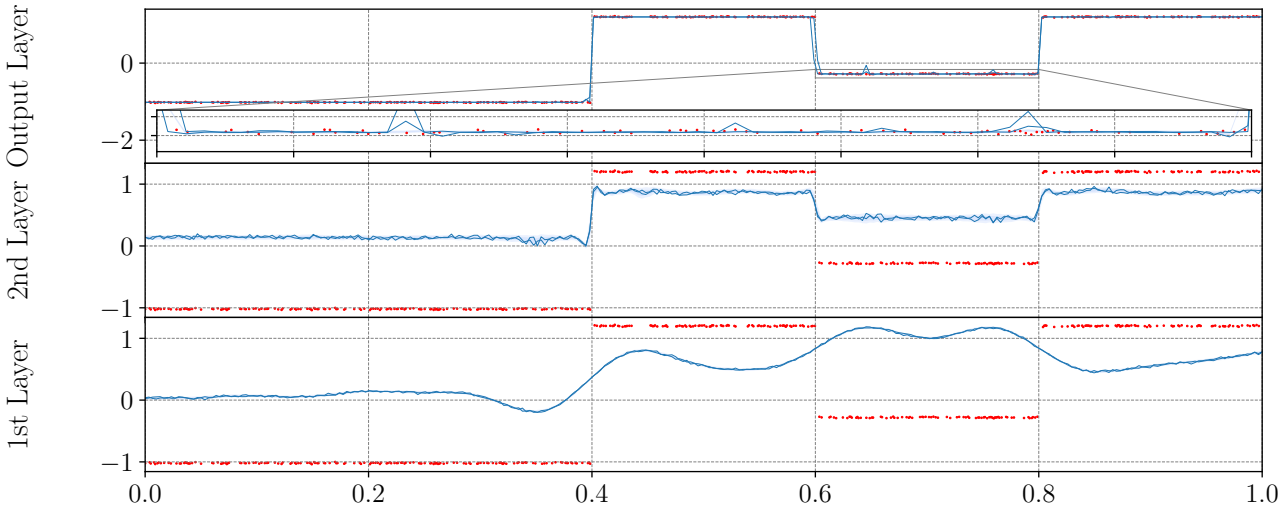


Figure 11: Inter-domain DGP with DSVI (three layers). Top: DGP posterior predictive distribution. Bottom: Marginal distribution at intermediate layer. The model is trained using 20 inducing frequencies. Training points are shown in red. Each shade of blue represents one standard deviation in the posterior predictive distribution.

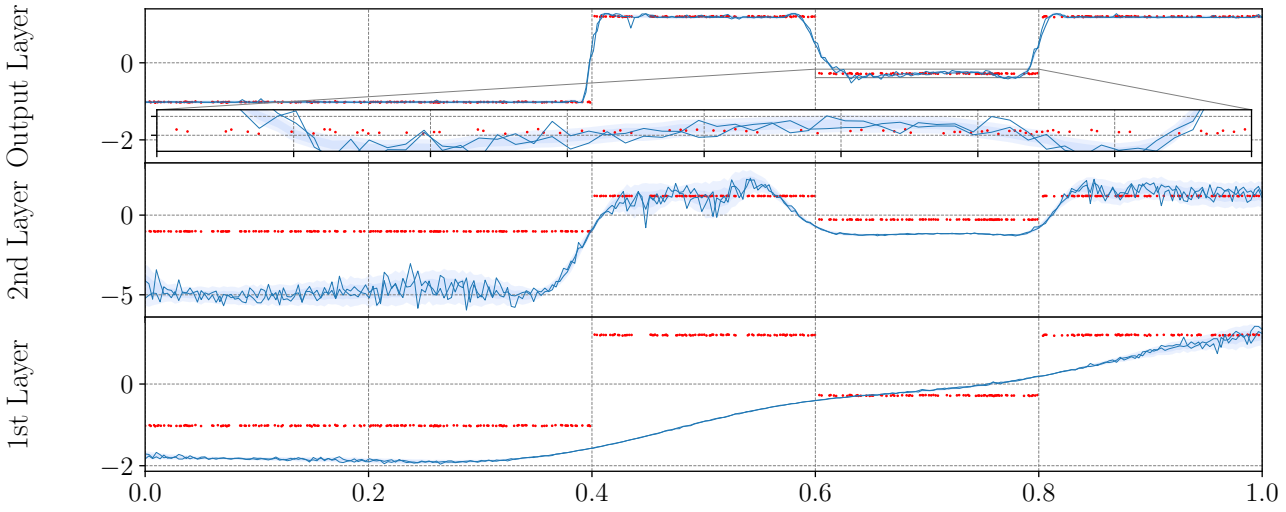


Figure 12: Conventional DGP with DSVI (three layers). Top: DGP posterior predictive distribution. Bottom: Marginal distribution at intermediate layer. The model is trained using 20 inducing frequencies. Training points are shown in red. Each shade of blue represents one standard deviation in the posterior predictive distribution.

C. RKHS Fourier features for Approximate Inference in Gaussian Processes

RKHS Fourier features were introduced as an inter-domain representation of inducing variables in variational inference for shallow GPs by Hensman et al. (2018). RKHS Fourier features use RKHS theory to construct inter-domain alternatives to the covariance matrices $\mathbf{K_{uu}}$ and $\mathbf{k_u}(\mathbf{x})$ used in inducing points-based approximate inference methods. They are constructed by projecting the target function f onto the truncated Fourier basis,

$$\boldsymbol{\phi}(x) =[1, \cos(\omega_1(x-a)), ..., \cos(\omega_M(x-a)), \sin(\omega_1(x-a)), ..., \sin(\omega_M(x-a))]^\top, \tag{C.1}$$

where x is a single, one-dimensional input, and the mth frequency ω_m is defined as

$$\omega_m = \frac{2\pi m}{b-a}$$

for some interval $[a, b]$. The specific functional form of the truncated Fourier basis is derived from the basis function used for *Random Fourier Features* (Rahimi & Recht, 2008). Berlinet & Thomas-Agnan (2004) showed that if $\mathcal{F} = \text{span}(\boldsymbol{\phi})$ is a subspace of an RKHS $\mathcal{H}$, the kernel of $\mathcal{F}$ is given by

$$k_{\mathcal{F}}(x, x') = \boldsymbol{\phi}(x)^\top \mathbf{K}_{\phi\phi}^{-1} \boldsymbol{\phi}(x'),$$

where $\mathbf{K}_{\phi\phi}[m, m'] = \langle \phi_m, \phi_{m'} \rangle_{\mathcal{H}}$ is the Gram matrix of $\boldsymbol{\phi}$ in $\mathcal{H}$ and $\boldsymbol{\phi}_m(x)$ is an entry of $\boldsymbol{\phi}(x)$ with $m = 1, ..., M'$ and $M' = 2M + 1$ for M frequencies. Furthermore, for an RKHS $\mathcal{H}$, the coordinate of the projection of a function $h \in \mathcal{H}$ onto $\boldsymbol{\phi}_m(x)$ is given by

$$\mathcal{P}_{\boldsymbol{\phi}_m}(h) = \langle h, \boldsymbol{\phi}_m \rangle_{\mathcal{H}}$$

and defines a projection between domains. Durrande et al. (2016) showed that if $\mathcal{H}$ is a Matérn RKHS of functions over $[a, b]$ with a half-integer parameter, then $\mathcal{F}$ belongs to $\mathcal{H}$. The authors also provided closed-form expressions of the inner products for the Matérn-$\frac{1}{2}$, Matérn-$\frac{3}{2}$, and Matérn-$\frac{5}{2}$ RKHS. However, in order to apply the RKHS inner product $u_m = \langle f, \phi_m \rangle_{\mathcal{H}}$ between the sinusoids and the GP sample path, which *a priori* does not belong to the RKHS, it is necessary to extend the operators $\mathcal{P}_{\phi_m} : h \mapsto \langle h, \phi_m \rangle_{\mathcal{H}}$ to square integrable functions. Hensman et al. (2018) show that this is possible for the half-integer members of the Matérn family of kernels. With these results, we can construct the inducing variables as an inter-domain projection by letting $u_m = \mathcal{P}_{\boldsymbol{\phi}_m}(f)$, which yields

$$\text{cov}(u_m, f(x)) = \boldsymbol{\phi}_m(x), \qquad \text{cov}(u_m, u_{m'}) = \langle \boldsymbol{\phi}_m, \boldsymbol{\phi}_{m'} \rangle_{\mathcal{H}},$$

for both of which there are closed-form expressions for the half-integer members of the Matérn family of kernels (Durrande et al., 2016), provided in Hensman et al. (2018). The resulting operators

$$\mathbf{k}_{\mathbf{u}}^{\phi}(x) = \boldsymbol{\phi}_m(x), \qquad \mathbf{K}_{\mathbf{uu}}^{\phi} = \langle \boldsymbol{\phi}_m, \boldsymbol{\phi}_{m'} \rangle_{\mathcal{H}},$$

represent generalized, inter-domain alternatives to the $\mathbf{k_u}(x)$ and $\mathbf{K_{uu}}$ operators used in local inducing-points approaches. Note that, as is the case for covariance matrices in local inducing-points methods, the variational Fourier feature operators $\mathbf{k}_{\mathbf{u}}^{\phi}(x)$ and $\mathbf{K}_{\mathbf{uu}}^{\phi}$ relate model inputs to the output space, but in contrast to inducing inputs in local inducing-point approaches, the inducing frequencies do not need to lie in the same space as the model inputs.

D. Doubly Stochastic Variational Inference for Deep Gaussian Processes

Inter-domain DGPs exploit the compositional structure of the approximate posterior in *doubly stochastic variational inference* for DGPs to achieve simple and scalable inference in inter-domain DGPs.

In *doubly stochastic variational inference*, proposed by Salimbeni & Deisenroth (2017), the variational posterior is defined to have the following three properties: First, conditioned on $\mathbf{u}^{(\ell)}$, the variational distribution is assumed to maintain the exact model,

$$q(\mathbf{f}^{(\ell)}, \mathbf{u}^{(\ell)}) = p(\mathbf{f}^{(\ell)} \,|\, \mathbf{u}^{(\ell)}) q(\mathbf{u}^{(\ell)});$$

second, a mean-field assumption is made so that the posterior distribution of $\{\mathbf{u}^{(\ell)}\}_{\ell=1}^{L}$ factorizes across layers (and dimensions), which implies that the variational distribution takes the form

$$\mathcal{Q} = q(\{\mathbf{f}^{(\ell)}, \mathbf{u}^{(\ell)}\}_{\ell=1}^{L}) = \prod_{\ell=1}^{L} p(\mathbf{f}^{(\ell)} \,|\, \mathbf{u}^{(\ell)}, \mathbf{f}^{(\ell-1)}) q(\mathbf{u}^{(\ell)});$$

and third, $q(\mathbf{u}^{(\ell)})$ is assumed to be Gaussian with mean $\boldsymbol{\mu}^{(\ell)}$ and variance $\boldsymbol{\Sigma}^{(\ell)}$ for $\ell = 1, ..., L$. These properties make it possible to marginalize out the set of $\mathbf{u}^{(\ell)}$ from $\mathcal{Q}$ analytically, which yields

$$\begin{aligned} q(\{\mathbf{f}^{(\ell)}\}_{\ell=1}^{L}) &= \prod_{\ell=1}^{L} q(\mathbf{f}^{(\ell)} \,|\boldsymbol{\mu}^{(\ell)}, \boldsymbol{\Sigma}^{(\ell)}; \mathbf{f}^{(\ell-1)}, \mathbf{Z}^{(\ell-1)}) \\ &= \prod_{\ell=1}^{L} \mathcal{N}(\mathbf{f}^{(\ell)} \,|\widetilde{\mathbf{m}}_{\mathbf{f}}^{(\ell)}, \widetilde{\mathbf{S}}_{\mathbf{f}}^{(\ell)}), \end{aligned} \tag{D.2}$$

where

$$\begin{aligned} \widetilde{\mathbf{m}}_{\mathbf{f}}^{(\ell)} &\overset{\text{def}}{=} \widetilde{\mathbf{m}}(\mathbf{f}^{(\ell)}) \\ &= \mathbf{m}_{\mathbf{f}^\ell} - \mathbf{K}_{\mathbf{f}^\ell\mathbf{u}^\ell}\mathbf{K}_{\mathbf{u}^\ell\mathbf{u}^\ell}(\boldsymbol{\mu}^{(\ell)} - \mathbf{m}_{\mathbf{u}^\ell}^{\phi}), \end{aligned} \tag{D.3}$$

$$\begin{aligned} \widetilde{\mathbf{S}}_{\mathbf{f}}^{(\ell)} &\overset{\text{def}}{=} \widetilde{\mathbf{S}}(\mathbf{f}^{(\ell)}, \mathbf{f}^{(\ell)}) \\ &= \mathbf{K}_{\mathbf{f}^\ell\mathbf{f}^\ell} - \mathbf{K}_{\mathbf{f}^\ell\mathbf{u}^\ell}\mathbf{K}_{\mathbf{u}^\ell\mathbf{u}^\ell}(\mathbf{K}_{\mathbf{u}^\ell\mathbf{u}^\ell} - \boldsymbol{\Sigma}^{(\ell)})\mathbf{K}_{\mathbf{u}^\ell\mathbf{u}^\ell}\mathbf{K}_{\mathbf{u}^\ell\mathbf{f}^\ell}, \end{aligned} \tag{D.4}$$

with mean functions $\mathbf{m}_{\mathbf{f}^\ell} \overset{\text{def}}{=} m(\mathbf{f}^{(\ell-1)})$ and $\mathbf{m}_{\mathbf{u}^\ell} \overset{\text{def}}{=} m(\mathbf{Z}^{(\ell-1)})$ and inducing inputs $\mathbf{Z}^{(\ell-1)}$ for $\ell = 1, ..., L$.

The marginals within each layer thus only depend on the corresponding inputs, and so the nth marginal of the final layer of the DGP posterior predictive distribution can be expressed as

$$q(\mathbf{f}_n^{(L)}) = \int \prod_{\ell=1}^{L-1} q(\mathbf{f}_n^{(\ell)} \,|\boldsymbol{\mu}^{(\ell)}, \boldsymbol{\Sigma}^{(\ell)}; \mathbf{f}_n^{(\ell-1)}, \mathbf{Z}^{(\ell-1)}) \, \mathrm{d}\mathbf{f}_n^{(\ell)}, \tag{D.5}$$

where $\mathbf{f}_n^{(\ell)}$ is the nth row of $\mathbf{f}^{(\ell)}$. This quantity is easy to compute using the reparameterization trick, that allows for sampling from the nth instances of the variational posteriors across layers by defining

$$\hat{\mathbf{f}}_n^{(\ell)} = \widetilde{\mathbf{m}}\left(\hat{\mathbf{f}}_n^{(\ell-1)}\right) + \boldsymbol{\epsilon}_n^{(\ell)} \odot \sqrt{\widetilde{\mathbf{S}}\left(\hat{\mathbf{f}}_n^{(\ell-1)}, \hat{\mathbf{f}}_n^{(\ell-1)}\right)} \tag{D.6}$$

and sampling from $\boldsymbol{\epsilon}_n^{(\ell)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D^{(\ell)}})$ (Kingma & Welling, 2014; Salimbeni & Deisenroth, 2017).

E. ELBO Derivation

Starting from the log-likelihood,

$$\log p(\mathbf{y}) = \log \mathbb{E}_{q(\{\mathbf{f}^{(\ell)}, \mathbf{u}^{(\ell)}\}_{\ell=1}^{L})} \left(\frac{p(\mathbf{y}, \{\mathbf{f}^{(\ell)}, \mathbf{u}^{(\ell)}\}_{\ell=1}^{L})}{q(\{\mathbf{f}^{(\ell)}, \mathbf{u}^{(\ell)}\}_{\ell=1}^{L})} \right),$$

with variational posterior

$$\mathcal{Q} = q(\{\mathbf{f}^{\ell}, \mathbf{u}^{(\ell)}\}_{\ell=1}^{L}) = \prod_{\ell=1}^{L} p(\mathbf{f}^{(\ell)} \,|\, \mathbf{u}^{(\ell)}, \mathbf{f}^{(\ell-1)}) q(\mathbf{u}^{(\ell)})$$

and joint distribution

$$p(\mathbf{y}, \{\mathbf{f}^{(\ell)}, \mathbf{u}^{(\ell)}\}_{\ell=1}^{L}) = \prod_{n=1}^{N} p(\mathbf{y}_n \,|\, \mathbf{f}_n^{(L)}) \prod_{\ell=1}^{L} p(\mathbf{f}^{(\ell)} \,|\, \mathbf{u}^{(\ell)}; \mathbf{f}^{(\ell-1)}, \boldsymbol{\Omega}^{(\ell-1)}) p(\mathbf{u}^{(\ell)}; \boldsymbol{\Omega}^{(\ell-1)}),$$

Lower bounding it by applying Jensen's inequality, we get the evidence lower bound

$$\begin{aligned}\log p(\mathbf{y}) &= \log \mathbb{E}_{q(\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})}\left[\frac{p(\mathbf{y},\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})}{q(\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})}\right] \\ &\geq \mathbb{E}_{q(\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})}\left[\log\left(\frac{p(\mathbf{y},\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})}{q(\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})}\right)\right] = \mathcal{L}.\end{aligned}$$

Writing the expectation as an integral and substituting in the variational posterior and joint distribution, we get

$$\begin{aligned}\mathcal{L} &= \int\int q(\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})\log\left(\frac{p(\mathbf{y},\{\mathbf{f}^{\ell},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})}{q(\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})}\right)\mathrm{d}\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L} \\ &= \int\int q(\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})\log\left(\frac{\prod_{i=1}^{N}p\left(\mathbf{y}_i|\mathbf{f}_i^{L}\right)\prod_{l=1}^{L}p\left(\mathbf{f}^{\ell}|\mathbf{u}^{\ell};\mathbf{f}^{l-1},\mathbf{\Omega}^{l-1}\right)p\left(\mathbf{u}^{\ell};\mathbf{\Omega}^{l-1}\right)}{\prod_{l=1}^{L}p\left(\mathbf{f}^{\ell}|\mathbf{u}^{\ell};\mathbf{f}^{l-1};\mathbf{\Omega}^{l-1}\right)q\left(\mathbf{u}^{\ell}\right)}\right)\mathrm{d}\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L}.\end{aligned}$$

Cancelling out the identical terms in the logarithm and rewriting the resulting expression, we get

$$\begin{aligned}\mathcal{L} &= \iint q(\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})\log\left(\frac{\prod_{n=1}^{N}p\left(\mathbf{y}_n|\mathbf{f}_n^{L}\right)\prod_{l=1}^{L}p(\mathbf{u}^{\ell};\mathbf{\Omega}^{l-1})}{\prod_{l=1}^{L}q\left(\mathbf{u}^{\ell}\right)}\right)\mathrm{d}\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L} \\ &= \iint q(\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})\log\prod_{n=1}^{N}p(\mathbf{y}_n|\mathbf{f}_n^{L})\,d\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L} \\ &\quad + \iint q(\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})\log\left(\frac{\prod_{l=1}^{L}p\left(\mathbf{u}^{\ell};\mathbf{\Omega}^{(\ell-1)}\right)}{\prod_{l=1}^{L}q\left(\mathbf{u}^{\ell}\right)}\right)\mathrm{d}\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L} \\ &= \int q(\{\mathbf{f}^{(\ell)},\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})\log\prod_{n=1}^{N}p(\mathbf{y}_n|\mathbf{f}_n^{L})\,\mathrm{d}\{\mathbf{f}^{\ell}\}_{l=1}^{L} \\ &\quad + \iint q(\mathbf{u}^{(\ell)}\}_{\ell=1}^{L})\log\left(\frac{\prod_{l=1}^{L}p\left(\mathbf{u}^{\ell};\mathbf{\Omega}^{(\ell-1)}\right)}{\prod_{l=1}^{L}q\left(\mathbf{u}^{\ell}\right)}\right)\mathrm{d}\{\mathbf{u}^{(\ell)}\}_{\ell=1}^{L} \\ &= \sum_{n=1}^{N}\mathbb{E}_{q(\mathbf{f}_n^{(L)})}\left[\log p(\mathbf{y}_n\,|\,\mathbf{f}_n^{(L)})\right] - \sum_{\ell=1}^{L}\mathrm{KL}(q(\mathbf{u}^{(\ell)})\,||\,p(\mathbf{u}^{(\ell)})).\end{aligned}$$