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* DGP*s*.

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

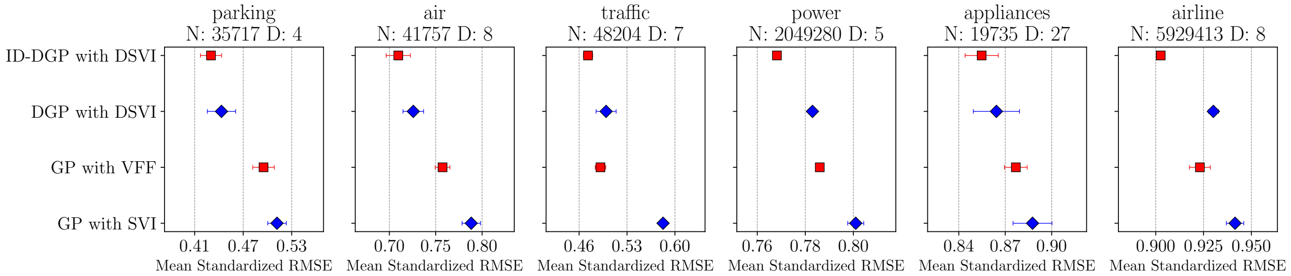

Figure 6: Average standardized root mean squared error (lower is better) and standard errors (over 10 random random seeds) on a set of real-world datasets exhibiting global structure. All models were trained with 50 inducing points.

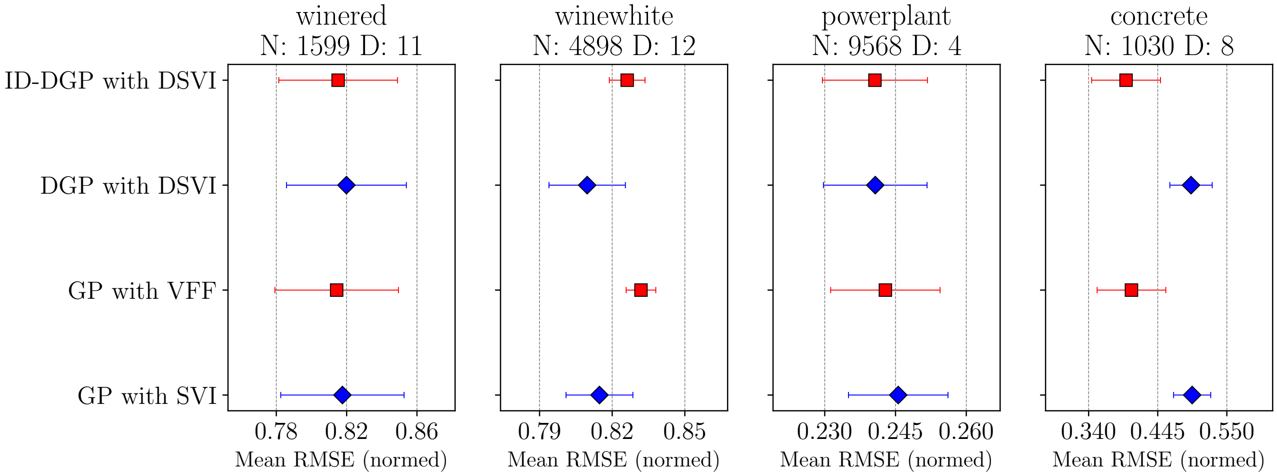

Figure 7: Average standardized root mean squared errors (lower is better) and standard errors (over 10 random seeds) on a set of small- and medium-scale regression problems. Each model was trained with 20 inducing points/inducing frequencies.

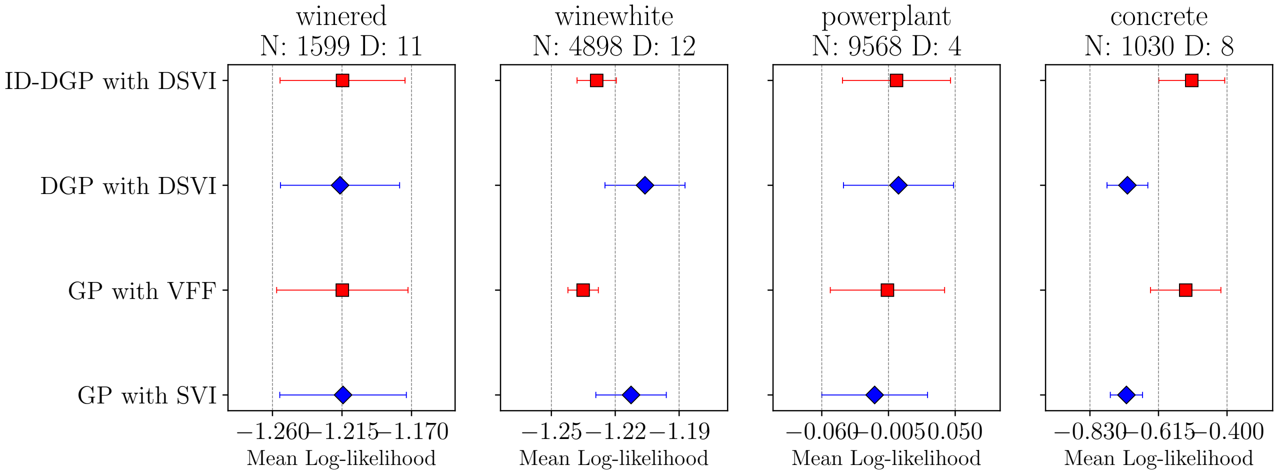

Figure 8: Average test log-likelihoods (higher is better) and standard errors (over 10 random seeds) on a set of small- and medium-scale regression problems. Each model was trained with 20 inducing points/inducing frequencies.

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.

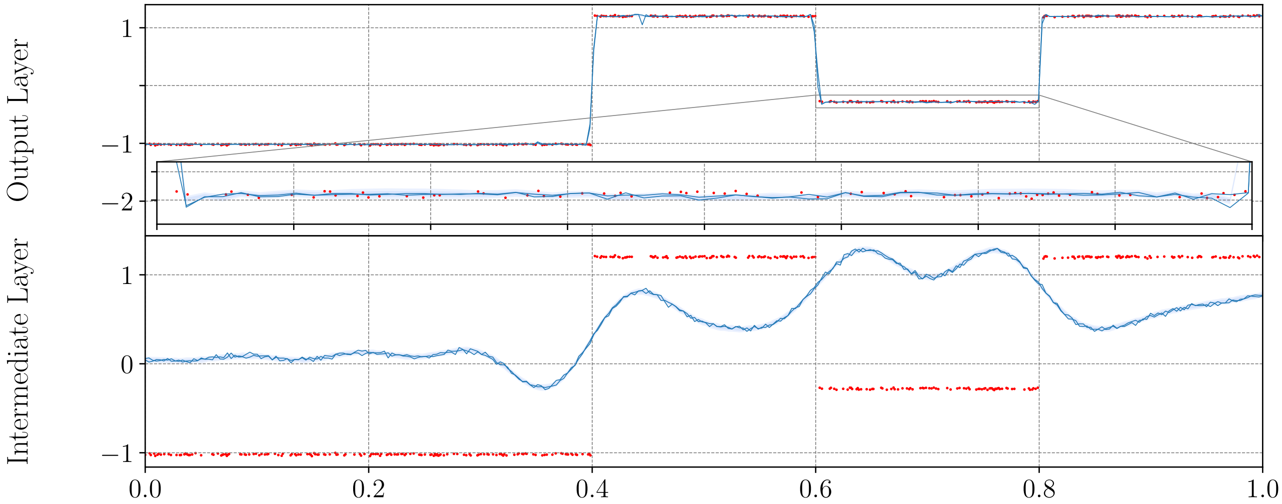

Figure 9: Inter-domain DGP with DSVI (two layers). Top: DGP posterior predictive distribution. Bottom: Marginal distribution at intermediate layer. The model is trained using 20 inducing frequencies. Training points are shown in red. Each shade of blue represents one standard deviation in the posterior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

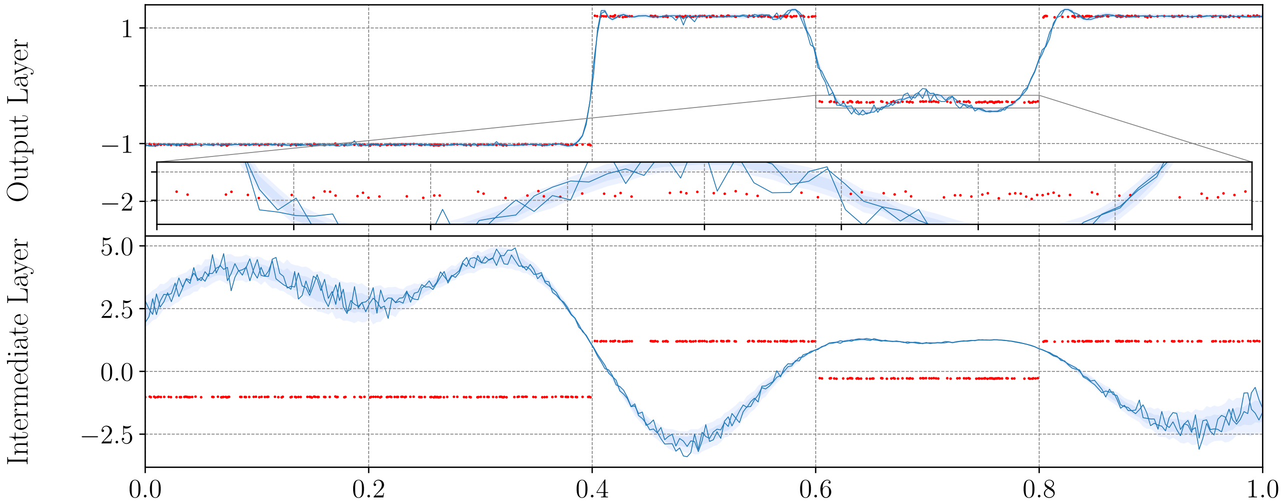

Figure 10: Conventional DGP with DSVI (two layers). Top: DGP posterior predictive distribution. Bottom: Marginal distribution at intermediate layer. The model is trained using 20 inducing frequencies. Training points are shown in red. Each shade of blue represents one standard deviation in the posterior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

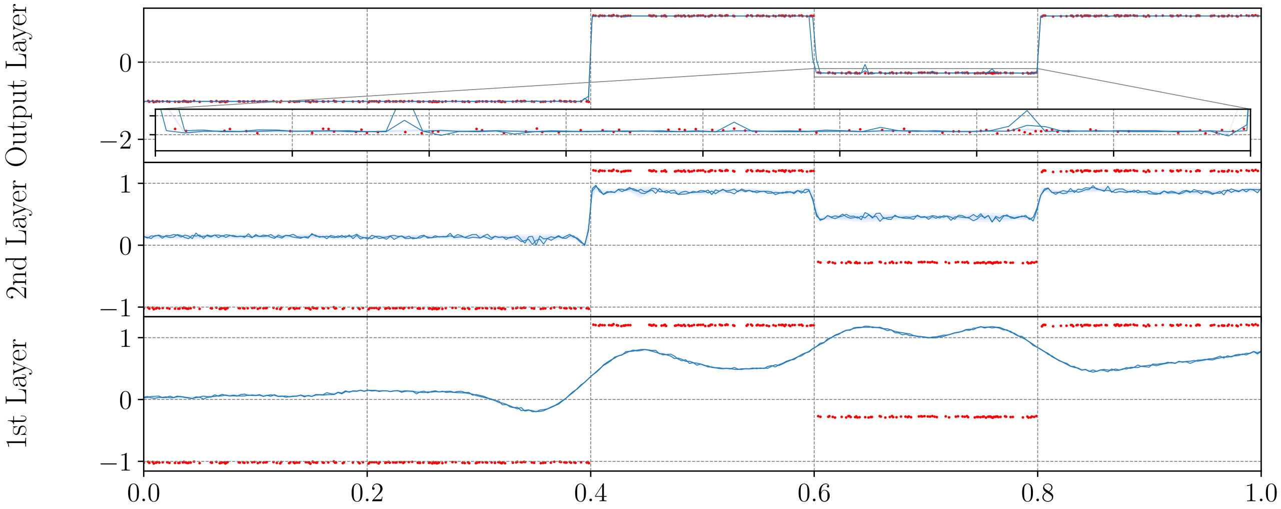

Figure 11: Inter-domain DGP with DSVI (three layers). Top: DGP posterior predictive distribution. Bottom: Marginal distribution at intermediate layer. The model is trained using 20 inducing frequencies. Training points are shown in red. Each shade of blue represents one standard deviation in the posterior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

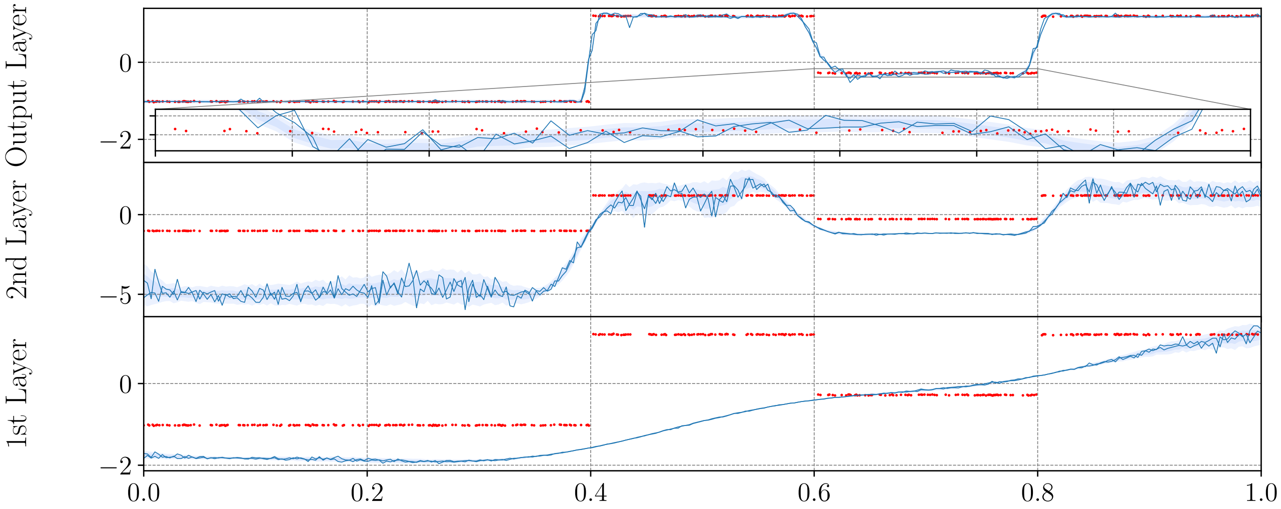

Figure 12: Conventional DGP with DSVI (three layers). Top: DGP posterior predictive distribution. Bottom: Marginal distribution at intermediate layer. The model is trained using 20 inducing frequencies. Training points are shown in red. Each shade of blue represents one standard deviation in the posterior 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 $\mathbf{K}_{\mathbf{uu}}$ and $\mathbf{k}_{\mathbf{u}}(\mathbf{x})$ 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 $\mathbf{k}_{\mathbf{u}}(x)$ and $\mathbf{K}_{\mathbf{uu}}$ 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,

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

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

$$\begin{aligned} \widetilde{\mathbf{m}}_{\mathbf{f}}^{(\ell)} &\overset{\text{def}}{=} \widetilde{\mathbf{m}}(\mathbf{f}^{(\ell)}) \\ &= \mathbf{m}_{\mathbf{f}^{\ell}} - \mathbf{K}_{\mathbf{f}^{\ell}\mathbf{u}^{\ell}} \mathbf{K}_{\mathbf{u}^{\ell}\mathbf{u}^{\ell}} (\boldsymbol{\mu}^{(\ell)} - \mathbf{m}_{\mathbf{u}^{\ell}}^{\phi}), && \text{(D.3)} \\ \widetilde{\mathbf{S}}_{\mathbf{f}}^{(\ell)} &\overset{\text{def}}{=} \widetilde{\mathbf{S}}(\mathbf{f}^{(\ell)}, \mathbf{f}^{(\ell)}) \\ &= \mathbf{K}_{\mathbf{f}^{\ell}\mathbf{f}^{\ell}} - \mathbf{K}_{\mathbf{f}^{\ell}\mathbf{u}^{\ell}} \mathbf{K}_{\mathbf{u}^{\ell}\mathbf{u}^{\ell}} (\mathbf{K}_{\mathbf{u}^{\ell}\mathbf{u}^{\ell}} - \boldsymbol{\Sigma}^{(\ell)}) \mathbf{K}_{\mathbf{u}^{\ell}\mathbf{u}^{\ell}} \mathbf{K}_{\mathbf{u}^{\ell}\mathbf{f}^{\ell}}, && \text{(D.4)} \end{aligned}$$

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

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

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

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