Compositional uncertainty in deep Gaussian processes

Ivan Ustyuzhaninov*
University of Tübingen

Ieva Kazlauskaite*
University of Bath,
Electronic Arts

Markus Kaiser
Siemens AG,
TU Munich

Erik Bodin
University of Bristol

Neill D. F. Campbell[❦]
University of Bath,
Royal Society

Carl Henrik Ek[❦]
University of Bristol

Abstract

Gaussian processes (GPs) are nonparametric priors over functions. Fitting a GP implies computing a posterior distribution of functions consistent with the observed data. Similarly, deep Gaussian processes (DGPs) should allow us to compute a posterior distribution of compositions of multiple functions giving rise to the observations. However, exact Bayesian inference is intractable for DGPs, motivating the use of various approximations. We show that the application of simplifying mean-field assumptions across the hierarchy leads to the layers of a DGP collapsing to near-deterministic transformations. We argue that such an inference scheme is suboptimal, not taking advantage of the potential of the model to discover the compositional structure in the data. To address this issue, we examine alternative variational inference schemes allowing for dependencies across different layers and discuss their advantages and limitations.

1 INTRODUCTION

Hierarchical learning studies functions represented as compositions of other functions, $f = f_L \circ \ldots \circ f_1$. Such models provide a natural way to model data generated by a hierarchical process, as each f_ℓ represents a certain part of the hierarchy, and the prior assumptions on $\{f_\ell\}_{\ell=1}^L$ reflect the corresponding prior assumptions about the data generating process. DGPs (Damianou and Lawrence, 2013), which are compositions of GPs, allow us to impose explicit prior assumptions on $\{f_\ell\}$ by choosing the corresponding kernels. Since different compositions can fit the data equally well (see an illustration in Fig. 1), DGPs are inherently unidentifiable, and this lack of identifiability should be captured by an adequate Bayesian posterior, allowing us to quantify uncertainties pertaining to each f_ℓ. We refer to this uncertainty as *compositional uncertainty*. This uncertainty can be thought of as the epistemic uncertainty (Gal, 2016) describing how the layers of the hierarchy jointly compose the observed data.

While the DGP posterior captures compositional uncertainty, exact Bayesian inference in DGPs is intractable (Damianou and Lawrence, 2013). In this work we show that the typically used approximate inference schemes (*e.g.* Salimbeni and Deisenroth, 2017) impose strong simplifying assumptions, making intermediate DGP layers collapse to deterministic transformations.[1] This corresponds to representing a DGP as a single-layer GP with a transformed kernel (Dunlop et al., 2018), similar to GPs with kernels parametrised by a deterministic function (*e.g.* a neural network). Such behaviour might not be a problem in practice if the goal is to design a model that only provides a high marginal likelihood of the data, however, it does not make full use of the capacity of DGP as it fails to describe the uncertainty that stems from the potential decomposition in the hierarchy. Distributions over compositions, and the resulting compositional uncertainty, are important for applications, *e.g.* for temporal alignment of time series data (Kaiser et al., 2018; Kazlauskaite et al., 2019), in reinforcement learning (Jin et al., 2017) as well as for building more interpretable models where each layer in the hierarchy expresses a meaningful functional prior (Sun et al., 2019).

We address the issue of collapsing compositional uncertainty by proposing variational distributions and corre-

*, ❦ Equal contributions

[1] In cases where the data has high observational noise, the noise is explained by introducing an uncertainty in one or multiple layers of the composition. We focus on the case where the data is noiseless, thus the uncertainty in each of the layers arises only due to the ambiguity in the compositional structure.

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

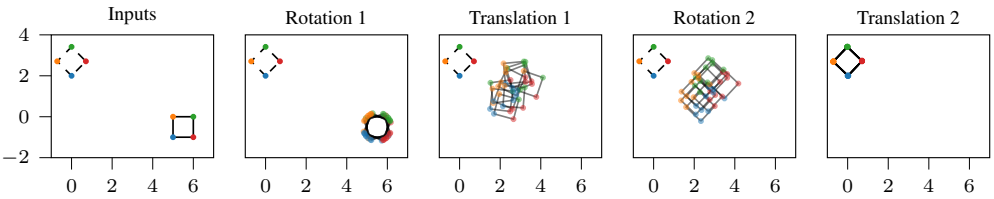

Figure 1: Compositional model (toy example): the transformation of the solid rectangle onto the dashed one is decomposed as $T_2 \circ R_2 \circ T_1 \circ R_1$ where R_i and T_i are rotations and translations. Different sampled realisations of these transformations are overlaid, showing the *compositional uncertainty*. Approximating R_i and T_i as independent transformations does not allow us to capture such uncertainty, collapsing to a single realisation of the composition.

sponding inference methods that explicitly model the dependencies between the layers, resulting in variational posteriors that capture compositional uncertainty. We highlight the limitations of existing approaches and lay the ground for future work in uncertainty quantification in DGPs. Our main contributions are:

- We demonstrate that variational distributions over the inducing points that are factorised *across layers* lead to a collapse of compositional uncertainty,
- We provide an intuitive as well as a quantitative argument for this behaviour by drawing a link between the work on mean-field variational inference for DGPs and the models of regression with noisy inputs (Girard et al., 2003);
- We propose modifications to the factorised variational distribution that incorporate the dependencies between the inducing points in different layers, and discuss the corresponding inference procedures,
- We use the proposed variational inference approaches to further illustrate how the correlations across the layers are necessary in order to argue about compositional uncertainty.

The remainder of the paper is structured as follows. We first provide a background to DGPs with an emphasis on approximate inference and discuss the method of (Salimbeni and Deisenroth, 2017) in detail, using it as the starting point for our argument on the collapse of compositional uncertainty, presented in Sec. 3. In Sec. 4 we propose variational distributions that aims to address the shortcomings of the layer-wise factorisation. In Sec. 5 we illustrate the behaviour of the proposed methods and discuss potential areas of applications.

2 BACKGROUND: MODELS OF DGPs

Previous work The hierarchical GP construction was originally motivated from the perspective of latent variable models (Lawrence, 2004) and was designed with a specific application in mind. In the early work on DGPs, Lawrence and Moore (2007) proposed a model that captured the hierarchical structure in the human skeleton, that allowed to produce interpretable generative models of human motion. However, most of the later work shifted the emphasis from uncovering specific interpretable hierarchical structures to employing a hierarchical construction to design models that are more flexible than a standard GP (in particular, by weakening the assumptions about a joint Gaussian structure in the observations). For example, Lázaro-Gredilla (2012) proposed a hierarchical (two-layer) GP model to allow for non-stationary observations. Damianou and Lawrence (2013) drew further parallels between DGPs and deep belief networks, and proposed a DGP construction beyond two layers for both supervised and unsupervised settings. Concurrently, the MAP estimation used in the early works (Lawrence and Moore, 2007) was replaced with variational inference schemes, initially proposed for the latent variable model (Titsias and Lawrence, 2010) and later adapted for the hierarchical DGP setting (Damianou and Lawrence, 2013).

However, the variational inference approach of Damianou and Lawrence (2013) was shown to be prohibitive for large data sets, motivating further research on inference schemes that scale to large data sets (Hensman et al., 2013; Hensman and Lawrence, 2014; Dai et al., 2016; Bui et al., 2016; Gal and Ghahramani, 2016; Hensman et al., 2017; Salimbeni and Deisenroth, 2017; Rudner and Sejdinovic, 2017; Cutajar, 2019). A different line of thought emerges from the work on inference using stochastic gradient Hamiltonian Monte Carlo (Havasi et al., 2018). The authors recognise the issue of compositional uncertainty, highlighting the fact that most of the existing (variational) approaches to inference are limited to estimating single modes of the posterior distributions in each layer of the hierarchy. As inference using MC is typically very costly, the authors note that it is beneficial to decouple the model in terms of the inducing points for the mean and the variance, which results in a highly non-convex optimization problem that requires careful parameterisation to improve the stability of convergence. Various issues with numerical stabil-

ity, poor convergence and underestimation of uncertainty have also been reported in the context of variational approximations (Hensman and Lawrence, 2014; Kaiser et al., 2018). Duvenaud et al. (2014) show a pathological behaviour of the concentration of density along a single dimension as the number of layers increases, and propose including direct links between the inputs and each individual layer.

Doubly stochastic variational inference (DSVI) Our work builds on the variational approximation scheme introduced by Salimbeni and Deisenroth (2017), thus here we provide a short recap of the main ideas from this work and introduce the notation that is used throughout the rest of this paper. Given a dataset[2] $\mathcal{D} = \{(x_j, y_j)\}_{j=1}^J$, with $x_j, y_j \in \mathbb{R}$, we model $y_j = (f_L \circ \ldots \circ f_1)(x_j)$, where $f_\ell \sim \mathcal{GP}(\mu_\ell(\cdot), k_\ell(\cdot,\cdot))$. We denote the inputs as $\mathbf{x} = (x_1, \ldots, x_J) \in \mathbb{R}^J$, and the evaluations of the intermediate layers at the entire vector of inputs $\mathbf{x}$ as $\mathbf{f}_\ell \sim (f_\ell \circ \ldots \circ f_1)(\mathbf{x})$ for $\ell = 2, \ldots, L$. The DGP joint distribution is

$$p(\mathbf{y}, \mathbf{f}_L, \ldots, \mathbf{f}_1 \mid \mathbf{x}) = p(\mathbf{y} \mid \mathbf{f}_L) \prod_{\ell=1}^{L} p(\mathbf{f}_\ell \mid \mathbf{f}_{\ell-1}), \quad (1)$$

where $p(\mathbf{f}_\ell \mid \mathbf{f}_{\ell-1}) \sim \mathcal{GP}(\mu_j(\mathbf{f}_{\ell-1}), k_j(\mathbf{f}_{\ell-1}, \mathbf{f}_{\ell-1}))$ is a GP prior for the ℓ-th layer, and we define $\mathbf{f}_0 = \mathbf{x}$. Integrating $\{\mathbf{f}_\ell\}$ from (1) to obtain a marginal likelihood is intractable, since that requires integrating a product of Gaussian factors, each of which contains $\mathbf{f}_\ell$ inside a non-linear kernel.

To overcome this limitation, variational inference is used to estimate the lower bound on (1). To this end, each DGP layer ℓ is augmented with M inducing locations $\mathbf{z}_{\ell-1} \in \mathbb{R}^M$ and inducing points $\mathbf{u}_\ell \in \mathbb{R}^M$, resulting in the following augmented joint distribution:

$$\begin{aligned} p(\mathbf{y}, \{\mathbf{f}_\ell\}, \{\mathbf{u}_\ell\} \mid \mathbf{x}, \{\mathbf{z}_\ell\}) &= p(\mathbf{y} \mid \mathbf{f}_L) \times \\ &\times \prod_{\ell=1}^{L} p(\mathbf{f}_\ell \mid \mathbf{f}_{\ell-1}, \mathbf{u}_\ell, \mathbf{z}_{\ell-1}) p(\mathbf{u}_\ell \mid \mathbf{z}_{\ell-1}), \end{aligned} \quad (2)$$

where $p(\mathbf{f}_\ell \mid \mathbf{f}_{\ell-1}, \mathbf{u}_\ell, \mathbf{z}_{\ell-1}) \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$ is a GP posterior at inputs $\mathbf{f}_{\ell-1}$ given values of $\mathbf{u}_\ell$ at $\mathbf{z}_{\ell-1}$. The specific form of $\boldsymbol{\mu}_\ell$ and $\boldsymbol{\Sigma}_\ell$ is as follows (note a slight abuse of notation: $\mu_\ell(\cdot)$ is a mean function, while $\boldsymbol{\mu}_\ell$ is a posterior mean):

$$\begin{aligned} \boldsymbol{\mu}_\ell &= \mu_\ell(\mathbf{f}_{\ell-1}) + \alpha_\ell(\mathbf{f}_{\ell-1})^T (\mathbf{u}_\ell - \mu_\ell(\mathbf{z}_{\ell-1})), \\ \boldsymbol{\Sigma}_\ell &= k_\ell(\mathbf{f}_{\ell-1}, \mathbf{f}_{\ell-1}) - \alpha_\ell(\mathbf{f}_{\ell-1})^T k_\ell(\mathbf{z}_{\ell-1}, \mathbf{z}_{\ell-1})\, \alpha_\ell(\mathbf{f}_{\ell-1}), \end{aligned}$$

where

$$\alpha_\ell(\mathbf{f}_{\ell-1}) = k_\ell(\mathbf{z}_{\ell-1}, \mathbf{z}_{\ell-1})^{-1} k_\ell(\mathbf{z}_{\ell-1}, \mathbf{f}_{\ell-1}). \quad (3)$$

[2]Throughout the paper we consider one-dimensional data but the general considerations also apply in many dimensions.

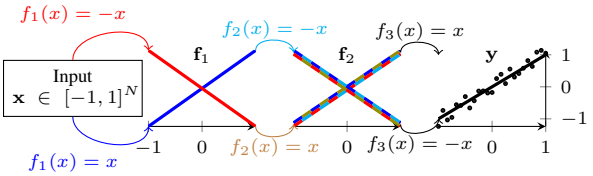

Figure 2: A toy example illustrating three layer compositions where each layer is either $f_\ell(x) = x$ or $f_\ell(x) = -x$. Multiple compositions map $\mathbf{x}$ to $\mathbf{y}$, this uncertainty is illustrated by showing the range of different values of $\mathbf{f}_1 = f_1(\mathbf{x})$ and $\mathbf{f}_2 = f_2(\mathbf{f}_1)$. If a variational distribution over $\{f_\ell\}$ is factorised, the posterior compositions collapse to a single realisation.

Introducing a factorised variational distribution over the inducing points

$$q(\{\mathbf{u}_\ell\}) = q(\mathbf{u}_1) \ldots q(\mathbf{u}_L), \quad q(\mathbf{u}_\ell) \sim \mathcal{N}(\mathbf{m}_\ell, \mathbf{S}_\ell) \quad (4)$$

the likelihood lower bound is as follows:

$$\begin{aligned} \mathcal{L}(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{f}_L)}[\log p(\mathbf{y} \mid \mathbf{f}_L)] - \\ - \sum_{\ell=1}^{L} \mathrm{KL}[q(\mathbf{u}_\ell) \,||\, p(\mathbf{u}_\ell \mid \mathbf{z}_{\ell-1})]. \end{aligned} \quad (5)$$

A key insight of Salimbeni and Deisenroth (2017) is that the expectation in (5) can be efficiently estimated by a Monte-Carlo estimator. This is possible by marginalising the inducing points $\{\mathbf{u}_\ell\}$ from the variational posterior, obtaining

$$\begin{aligned} q(\mathbf{f}_L, \ldots, \mathbf{f}_1) &= \prod_{\ell=1}^{L} \int p(\mathbf{f}_\ell \mid \mathbf{u}_\ell) q(\mathbf{u}_\ell)\, \mathrm{d}\mathbf{u}_\ell \\ &= q(\mathbf{f}_L \mid \mathbf{f}_{L-1}) \ldots q(\mathbf{f}_1 \mid \mathbf{x}), \end{aligned} \quad (6)$$

with $q(\mathbf{f}_\ell \mid \mathbf{f}_{\ell-1}) \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_\ell, \tilde{\boldsymbol{\Sigma}}_\ell)$, where

$$\tilde{\boldsymbol{\mu}}_\ell = \mu_\ell(\mathbf{f}_{\ell-1}) + \alpha_\ell(\mathbf{f}_{\ell-1})^T (\mathbf{m}_\ell - \mu_\ell(\mathbf{z}_{\ell-1})), \quad (7)$$

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_\ell = k_\ell(\mathbf{f}_{\ell-1}, \mathbf{f}_{\ell-1}) \\ - \alpha_\ell(\mathbf{f}_{\ell-1})^T (k_\ell(\mathbf{z}_{\ell-1}, \mathbf{z}_{\ell-1}) - \mathbf{S}_\ell)\, \alpha_\ell(\mathbf{f}_{\ell-1}). \end{aligned} \quad (8)$$

The bound in (5) can be estimated by sequentially sampling from $q(\mathbf{f}_\ell \mid \mathbf{f}_{\ell-1})$ using (7) and (8). The time complexity of this step is linear in the number of data points, since each marginal $[\mathbf{f}_\ell]_j$ can be drawn independently (we only need marginals of the final layer $\mathbf{f}_L$ in (5)).

3 MEAN-FIELD DGPs

In this section we argue that factorised variational distributions of inducing points, *e.g.* (4), imply that the layers in a DGP collapse to deterministic transformations.

3.1 INTUITION

If a DGP $f_L \circ \ldots \circ f_1$ maps fixed inputs $\mathbf{x}$ to fixed outputs $\mathbf{y}$, the functions $\{f_\ell\}$ must be dependent, because every realisation of this composition must map the *same* $\mathbf{x}$ to the *same* $\mathbf{y}$. This is illustrated in Fig. 2, which shows a composition of three layers, each of which could either be $f_\ell(x) = x$ or $f_\ell(x) = -x$. Depending on the choices of f_1 and f_2, the input is mapped by $f_2 \circ f_1$ to one of the two realisations of $\mathbf{f}_2$ (as shown by the colour code in the corresponding panel), and f_3 must be chosen in such a way that $\mathbf{f}_2$ is mapped to $\mathbf{y}$. Therefore, in this example, f_3 depends on the choice of f_1 and f_2. However, if $\{f_\ell\}$ were independent, then the only way to ensure that every realisation of the composition fits the data would be for each layer to implement a deterministic transformation (*i.e.* either $f_\ell(x) = x$ or $f_\ell(x) = -x$ such that there are zero or two instances of $f_\ell(x) = -x$). Another illustration of this idea is provided in Fig. 1, in which movement of a square is represented as a composition of correlated rotations and translations, allowing us to see a variety of possible movements. However, a model with independent transformations would converge to a single possible sequence of rotations and translations.

The same intuition holds for general DGPs. Analogously to choosing either x or $-x$ in Fig. 2, inducing locations $\mathbf{z}_\ell$ and points $\mathbf{u}_\ell$ define the transformation implemented by the corresponding layer through the predictive posterior $p(\mathbf{f}_\ell \mid \mathbf{f}_{\ell-1}, \mathbf{u}_\ell, \mathbf{z}_{\ell-1})$. Following a similar argument, the DGP layers collapse to deterministic transformations to ensure good data fits unless they are dependent to allow multiple different compositions to fit the data.

3.2 QUANTITATIVE ARGUMENT

Assume that the DGP layers $\{f_\ell\}$ are independent. Then the distribution of the outputs of layer $\ell - 1$ can be thought of as uncertain inputs[3] to the layer ℓ. Similarly to DGPs, the inference in such models is complicated by the need to propagate a distribution through a non-linear mapping. Such models have been studied in the context of GP regression (Girard et al., 2003; Mchutchon and Rasmussen, 2011; Bijl, 2018) and have also been discussed in relation to DGPs (Damianou, 2015), though not in the context of compositional uncertainty.

Assuming, for simplicity, that our dataset consists of a single point, *i.e.* $\mathcal{D} = \{(x, y)\}$, we can write $\mathbf{f}_\ell = (f_\ell \circ \ldots \circ f_1)(x) = f_\ell(\mathbf{f}_{\ell-1}) = f_\ell(\bar{\mathbf{f}}_{\ell-1} + \varepsilon_{\ell-1})$, with $\bar{\mathbf{f}}_{\ell-1}$ as the mean[4] of $\mathbf{f}_{\ell-1}$ and $\varepsilon_{\ell-1}$ as an appropriate zero-mean noise (not necessarily Gaussian, since marginals of DGP layers are not Gaussian in general (Damianou, 2015)), the variance of which we denote as $\boldsymbol{\sigma}^2_{\text{noise}} := \mathbb{V}\text{ar}\,[\varepsilon_{\ell-1}]$. We want to show that the variance of $\mathbf{f}_\ell$ increases with increasing variance of $\varepsilon_{\ell-1}$, which would imply that unless the layers collapse, *i.e.* $\varepsilon_{\ell-1} = 0$, there is finite variance in the final layer $\mathbf{f}_L$. That constitutes a poor fit to observations that contain low observational noise (noiseless in the limit), forcing the layers to collapse to deterministic transformations.

High observational noise might lead to the layers not collapsing despite being independent. However, such uncertainty is the observational noise spread across the layers, rather than compositional uncertainty due to multiple compositions explaining the data. To make our arguments clearer, we assume noiseless observations.

Linear approximation We can approximate $\mathbf{f}_\ell$ as

$$\mathbf{f}_\ell = f_\ell(\mathbf{f}_{\ell-1}) \approx f_\ell(\bar{\mathbf{f}}_{\ell-1}) + \varepsilon_{\ell-1}\, f'_\ell(\bar{\mathbf{f}}_{\ell-1}), \qquad (9)$$

where $f_\ell(\bar{\mathbf{f}}_{\ell-1}) \sim p(\mathbf{f}_\ell \mid \bar{\mathbf{f}}_{\ell-1}, \mathbf{u}_\ell, \mathbf{z}_{\ell-1}) = \mathcal{N}(\bar{\boldsymbol{\mu}}_\ell, \bar{\boldsymbol{\sigma}}^2_\ell)$, with $\bar{\boldsymbol{\mu}}_\ell$ and $\bar{\boldsymbol{\sigma}}^2_\ell$ given in (7) and (8). Note that both $\bar{\boldsymbol{\mu}}_\ell$ and $\bar{\boldsymbol{\sigma}}^2_\ell$ are functions of $\bar{\mathbf{f}}_{\ell-1}$, which we omit to not clutter the notation; the derivatives below are taken w.r.t. $\bar{\mathbf{f}}_{\ell-1}$.

The evaluation of a GP and its derivative are jointly distributed as follows (Rasmussen and Williams, 2005):

$$\begin{bmatrix} f_\ell(\bar{\mathbf{f}}_{\ell-1}) \\ f'_\ell(\bar{\mathbf{f}}_{\ell-1}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \bar{\boldsymbol{\mu}}_\ell \\ \bar{\boldsymbol{\mu}}'_\ell \end{bmatrix}, \begin{bmatrix} \boldsymbol{\sigma}^2_\ell & (\boldsymbol{\sigma}^2_\ell)' \\ (\boldsymbol{\sigma}^2_\ell)' & (\boldsymbol{\sigma}^2_\ell)'' \end{bmatrix} \right). \qquad (10)$$

Similarly to Mchutchon and Rasmussen (2011), we compute a linear transformation of (10) and obtain that

$$\begin{aligned} \mathbb{E}\,[\mathbf{f}_\ell \mid \varepsilon_{\ell-1}] &= \bar{\boldsymbol{\mu}}_\ell + \varepsilon_{\ell-1}\bar{\boldsymbol{\mu}}'_\ell, \\ \mathbb{V}\text{ar}\,[\mathbf{f}_\ell \mid \varepsilon_{\ell-1}] &= \boldsymbol{\sigma}^2_\ell - 2\varepsilon_{\ell-1}(\boldsymbol{\sigma}^2_\ell)' + \varepsilon^2_{\ell-1}(\boldsymbol{\sigma}^2_\ell)''. \end{aligned}$$

Using the law of total variance we have

$$\mathbb{V}\text{ar}\,[\mathbf{f}_\ell] = \mathbb{E}\,[\mathbb{V}\text{ar}\,[\mathbf{f}_\ell|\varepsilon_{\ell-1}]] + \mathbb{V}\text{ar}\,[\mathbb{E}\,[\mathbf{f}_\ell|\varepsilon_{\ell-1}]]\,,$$

where

$$\begin{aligned} &\mathbb{E}\,[\mathbb{V}\text{ar}\,[\mathbf{f}_\ell|\varepsilon_{\ell-1}]] = \boldsymbol{\sigma}^2_\ell + \boldsymbol{\sigma}^2_{\text{noise}}\,(\boldsymbol{\sigma}^2_\ell)'', \\ &\mathbb{V}\text{ar}\,[\mathbb{E}\,[\mathbf{f}_\ell|\varepsilon_{\ell-1}]] = \mathbb{V}\text{ar}\,[\bar{\boldsymbol{\mu}}_\ell + \varepsilon_{\ell-1}\,\bar{\boldsymbol{\mu}}'_\ell] = \boldsymbol{\sigma}^2_{\text{noise}} \cdot (\bar{\boldsymbol{\mu}}'_\ell)^2\,. \end{aligned}$$

Combining these results together we obtain

$$\mathbb{V}\text{ar}\,[\mathbf{f}_\ell] = \boldsymbol{\sigma}^2_\ell + \boldsymbol{\sigma}^2_{\text{noise}}\left[(\bar{\boldsymbol{\mu}}'_\ell)^2 + (\boldsymbol{\sigma}^2_\ell)''\right] + O(\varepsilon^2_{\ell-1}). \qquad (11)$$

The only term in (11) that can be negative is $(\boldsymbol{\sigma}^2_\ell)''$, potentially making the variance of the GP output at a noisy input smaller that the variance at a fixed input (i.e. $\boldsymbol{\sigma}^2_\ell$).

[3] Regression models that include input uncertainty can generally be formulated as: $\mathbf{y} = f(\mathbf{x} + \varepsilon_{\mathbf{x}})$, where $\mathbf{y}$ are observations, $\mathbf{x}$ are noise-free inputs and $\varepsilon_{\mathbf{x}}$ is zero-mean noise.

[4] We use bold notation for $\bar{\mathbf{f}}_{\ell-1}$, even though it refers to a scalar, to distinguish it from the notation we use for functions.

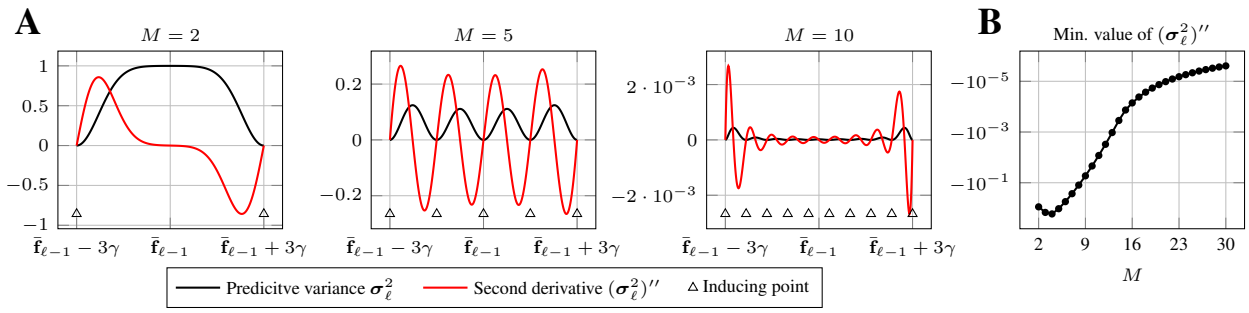

Figure 3: **A:** Predictive posterior variance $\boldsymbol{\sigma}_\ell^2$ and its second derivative $(\boldsymbol{\sigma}_\ell^2)''$ of ℓ-th layer in a $3\gamma_\ell$-neighbourhood Δ_ℓ of the noiseless input $\bar{\mathbf{f}}_{\ell-1}$ for different numbers of inducing points. **B:** Minimum value of $(\boldsymbol{\sigma}_\ell^2)''$ as a function of number of inducing points M.

Counterexample Such an example can indeed be constructed. Girard et al. (2003) study GPs with uncertain inputs and compute an exact expression for $\mathbb{V}\mathrm{ar}\,[\mathbf{f}_\ell]$ as a function of $\boldsymbol{\sigma}^2_{\text{noise}}$ assuming $\varepsilon_{\ell-1}$ is Gaussian. Assuming there is a single inducing point $\mathbf{u}_\ell = 0$, the derivative of $\mathbb{V}\mathrm{ar}\,[\mathbf{f}_\ell]$ is negative at 0 provided that $\bar{\mathbf{f}}_{\ell-1}$ is sufficiently far away from $\mathbf{u}_\ell$ (in comparison to the length scale; see the derivation in the Supplement). This means that the input noise might *reduce* the output variance. However, such an example relies on the inputs to the ℓ-th layer, $\bar{\mathbf{f}}_{\ell-1}$, appearing in the regions of the input space that are poorly covered by the inducing points. Consequently, this scenario only occurs if the inducing points are placed in a way that leads to a poor fit of the observed data.

Inducing points limit The counterexample above relies on a degenerate setting in which the inducing points are far from the observations. Here we consider a limiting case that corresponds to a more realistic situation of sufficiently many inducing points near $\mathbf{f}_\ell$ (the limit of arbitrary many inducing points is a conceptually desirable setting, complicated by the computational constraints). Specifically, in each layer we assume M linearly spaced inducing points $\mathbf{z}_\ell = \{z^1_{\ell-1}, \ldots, z^M_{\ell-1}\}$ in $\Delta_\ell := [\bar{\mathbf{f}}_{\ell-1} - 3\gamma_\ell, \bar{\mathbf{f}}_{\ell-1} + 3\gamma_\ell]$, where γ_ℓ is the kernel length scale in layer ℓ. This assumption means that the input to each layer is contained in an interval Δ_ℓ covered by the inducing points.

The behaviour of (11) under such an assumption is illustrated in Fig. 3. The minimum value of $(\boldsymbol{\sigma}_\ell^2)''$ approaches zero as M increases; this suggests that the input noise leads to increased predictive posterior variance apart from degenerate cases of inducing points not covering the input region corresponding to the observed inputs $\mathbf{x}$, and the predictive mean derivative $(\bar{\boldsymbol{\mu}}'_\ell)^2$ being sufficiently small (*i.e.* the function implemented by the ℓ-th layer being close to a constant one).

To summarise, we argue that under the assumption of $\mathbf{f}_\ell = (f_\ell \circ \ldots \circ f_1)(x)$ being contained in an interval covered by the inducing locations $\mathbf{z}_\ell$ for the next layer, the variance in $\mathbf{f}_\ell$ leads to increased variance in $\mathbf{f}_{\ell+1}$, and hence in $\mathbf{f}_L$. Therefore, for $\mathbf{f}_L$ to fit a noiseless observation y, the variance in intermediate layers has to be reduced, implying that the layers collapse to deterministic transformations.

4 BEYOND FACTORISED VARIATIONAL DISTRIBUTIONS

To further investigate the effect of the factorisation imposed by the mean-field variational inference, we propose two alternative variational inference schemes that allow for correlations between layers. By relaxing the mean-field assumption across layers, we aim to uncover a range of solutions that are consistent with the data and follow the prior belief about each of the individual layers. In Sec. 4.1 we present a natural generalisation of a factorised variational distribution (4) to capture the marginal dependencies between the layers. In Sec. 4.2 we present an alternative variational approximation that introduces dependencies between the layers by linking the inducing points and locations of the neighbouring layers.

4.1 JOINTLY GAUSSIAN INDUCING POINTS

A straightforward modification of the DSVI variational approximation (Sec. 2) allowing us to capture the dependencies between the layers is to introduce correlations between the inducing points by modelling them with a jointly Gaussian variational distribution:

$$q(\mathbf{u}_1, \ldots, \mathbf{u}_L) \sim \mathcal{N}(\mathbf{m}, \mathbf{S}), \tag{12}$$

with $\mathbf{m} \in \mathbb{R}^{LM}$, $\mathbf{S} \in \mathbb{R}^{LM \times LM}$. The variational posterior is then given by

$$q(\{\mathbf{f}_\ell\}, \{\mathbf{u}_\ell\}) = q(\mathbf{u}_1, \ldots, \mathbf{u}_L) \prod_{\ell=1}^{L} p(\mathbf{f}_\ell | \mathbf{f}_{\ell-1}, \mathbf{u}_\ell). \tag{13}$$

The corresponding likelihood lower bound has the same structure as (5) with the KL term, $\mathrm{KL}[q(\mathbf{u}_1, \ldots, \mathbf{u}_L) \,||\, p(\mathbf{u}_1) \ldots p(\mathbf{u}_L)]$, that can be computed in closed form (it involves two Gaussians). The expectation $\mathbb{E}_{q(\mathbf{f}_L)}[\log p(\mathbf{y} \mid \mathbf{f}_L)]$ is, however, harder to estimate in case of variational distribution (12). The integral (6) no longer factorises into a product of integrals, which means that we can no longer integrate $\{\mathbf{u}_\ell\}$ out from $q(\{\mathbf{f}_\ell\}, \{\mathbf{u}_\ell\})$ and draw samples from $q(\mathbf{f}_L)$ in the same way as in (Salimbeni and Deisenroth, 2017). We consider two approaches to address this issue.

Sampling $\{\mathbf{u}_\ell\}$ We start by noting that, conditioned on $\{\mathbf{u}_\ell\}$, we can draw samples from $q(\mathbf{f}_L)$ in the same way as in (Salimbeni and Deisenroth, 2017). Specifically, to estimate $\mathbb{E}_{q(\mathbf{f}_L)}[\log p(\mathbf{y} \mid \mathbf{f}_L)]$, we

1. Draw S samples $\{(\mathbf{u}_1^s, \ldots, \mathbf{u}_L^s)\}_{s=1}^S \overset{iid}{\sim} q(\mathbf{u}_1, \ldots, \mathbf{u}_L)$,
2. For each sample $(\mathbf{u}_1^s, \ldots, \mathbf{u}_L^s)$, draw $\mathbf{f}_L^s \sim q(\mathbf{f}_L \mid \mathbf{u}_1^s, \ldots, \mathbf{u}_L^s)$ by recursively drawing from $p(\mathbf{f}_\ell \mid \mathbf{f}_{\ell-1}, \mathbf{u}_\ell^s)$, which are regular GP posterior distributions conditioned on $\mathbf{u}_\ell^s$,
3. Compute a Monte Carlo estimate $\mathbb{E}_{q(\mathbf{f}_L)}[\log p(\mathbf{y} \mid \mathbf{f}_L)] \approx \frac{1}{S} \sum_s \log p(\mathbf{y} \mid \mathbf{f}_L^s)$.

This approach is easy to implement and it can be applied in a variety of settings (*e.g.* when $q(\{\mathbf{u}_i\})$ is not Gaussian, as long as we can sample from it and reparametrise the gradients). However, that comes at the cost of introducing another sampling step, resulting in $\mathbb{E}_{q(\mathbf{f}_L)}[\log p(\mathbf{y} \mid \mathbf{f}_L)]$ being estimated by two nested Monte-Carlo estimators, implying an increased overall variance of the estimator and the need to carefully choose the appropriate number of samples (Rainforth et al., 2019). Estimating the variance implied by the nested MC estimator offers a direction for future work. Moreover, drawing coherent samples from $q(\mathbf{u}_1, \ldots, \mathbf{u}_L)$ has computational complexity of $O(L^3M^3)$ leading to an overall complexity of $O(L^3M^3 + LNM^2)$ per estimation.

Analytic marginalisation To address statistical and computational limitations of the above method, we propose another approach consisting of analytically integrating $\{\mathbf{u}_\ell\}$ from (13). To do so we assume that $q(\{\mathbf{u}_\ell\})$ admits a chain-like factorisation, namely

$$q(\{\mathbf{u}_\ell\}) = q(\mathbf{u}_L \mid \mathbf{u}_{L-1}) \ldots q(\mathbf{u}_2 \mid \mathbf{u}_1) q(\mathbf{u}_1). \quad (14)$$

The precision matrix across all layers, $\Lambda = \mathbf{S}^{-1} \in \mathbb{R}^{LM \times LM}$, encodes the conditional independence assumptions, and (14) implies that such matrix is block-tridiagonal (Fig. 4). The advantage of this assumption is that the number of parameters in the unconstrained $\mathbf{S}$ scales quadratically with the number of layers, while (14) implies a linear growth.

Assuming that the variational distribution (12) satisfies the factorisation (14), we analytically marginalise $\{\mathbf{u}_\ell\}$ from the variational posterior (13), obtaining

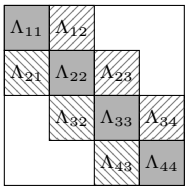

Figure 4: Precision matrix Λ induced by (14).

$$\int q(\{\mathbf{f}_\ell\}, \{\mathbf{u}_\ell\}) \, \mathrm{d}\{\mathbf{u}_\ell\} = \prod_{\ell=1}^{L} p(\mathbf{f}_\ell \mid \mathbf{f}_{\ell-1}, \ldots \mathbf{f}_1, \mathbf{x}), \quad (15)$$

where $p(\mathbf{f}_\ell \mid \mathbf{f}_{\ell-1}, \ldots \mathbf{f}_\ell, \mathbf{x}) \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_\ell, \tilde{\boldsymbol{\Sigma}}_\ell)$ with the mean and the covariance are defined recursively as follows:

$$\tilde{\boldsymbol{\mu}}_1 = \mu_1(\mathbf{x}) + \alpha_1(\mathbf{x})^T (\mathbf{m}_1 - \mu_1(\mathbf{z}_0))$$
$$\tilde{\boldsymbol{\Sigma}}_1 = k_1(\mathbf{x}, \mathbf{x}) - \alpha_1(\mathbf{x})^T (k_1(\mathbf{z}_0, \mathbf{z}_0) - \mathbf{S}_{11}) \, \alpha_1(\mathbf{x})$$

and $\alpha_1(\mathbf{x})$ is defined in (3). For $i > 1$, $\tilde{\mu}_i$ and $\tilde{\Sigma}_i$ are recursively defined as

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_\ell = \mu_\ell(\mathbf{f}_\ell) + \alpha_\ell(\mathbf{f}_{\ell-1})^T (\mathbf{m}_\ell + \mathbf{S}_{\ell,\ell-1} \, \alpha_{\ell-1}(\mathbf{f}_{\ell-1}) \times \\ \times \tilde{\boldsymbol{\Sigma}}_{\ell-1}^{-1} (\mathbf{f}_{\ell-1} - \tilde{\boldsymbol{\mu}}_{\ell-1} - \alpha_{\ell-1}(\mathbf{x})^T \times \\ \times (\mathbf{m}_{\ell-1} - \mu_{\ell-1}(\mathbf{z}_{\ell-2})) - \mu_\ell(\mathbf{z}_{\ell-1})), \end{aligned} \quad (16)$$

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_\ell = k_\ell(\mathbf{f}_{\ell-1}, \mathbf{f}_{\ell-1}) - \alpha_\ell(\mathbf{f}_{\ell-1})^T (k_\ell(\mathbf{z}_{\ell-1}, \mathbf{z}_{\ell-1}) - \\ - \mathbf{S}_{\ell\ell} + \mathbf{S}_{\ell,\ell-1} \, \alpha_{\ell-1}(\mathbf{f}_{\ell-1}) \tilde{\boldsymbol{\Sigma}}_{\ell-1}^{-1} \times \\ \times \alpha_{\ell-1}(\mathbf{f}_{\ell-1})^T \mathbf{S}_{\ell-1,\ell}) \alpha_\ell(\mathbf{f}_{\ell-1}), \end{aligned} \quad (17)$$

where $\mathbf{S}_{ij} = \mathrm{cov}(\mathbf{u}_i, \mathbf{u}_j)$.

The derivation is provided in the Supplement. Using these results, $\mathbb{E}_{q(\mathbf{f}_L)}[\log p(\mathbf{y} \mid \mathbf{f}_L)]$ can be estimated analogously to DSVI by recursively sampling $\mathbf{f}_i$ using (15).

4.2 INDUCING POINTS AS INDUCING LOCATIONS

In this section we discuss an alternative variational approximation, that connects the inducing points and the inducing locations of the neighbouring layers. Instead of directly modelling the inducing points in every layer, we only consider the inducing inputs $\mathbf{z}$ in the first layer and variational distributions over $\{\mathbf{f}_\ell^{\mathbf{z}} \sim (f_\ell \circ \ldots \circ f_1)(\mathbf{z})\}$. The advantage of such an approach is that unlike the variational distributions of inducing points, the factorisation of a variational distribution over $\{\mathbf{f}_i^{\mathbf{z}}\}$ does not imply that the variational posterior collapses to a single realisation of a composition fitting the data. In such a setting, $\mathbf{f}_{\ell-1}^{\mathbf{z}}$ and $\mathbf{f}_\ell^{\mathbf{z}}$ can be thought of as inducing pairs of the ℓ-th layer, meaning that the inducing points of a previous layer are the inducing locations of the next one.

Intuition Let us revisit the illustration given in Fig. 2. Assuming for this example that $\mathbf{z} = \mathbf{x}$, we independently sample values of $\mathbf{f}_1$ and $\mathbf{f}_2$ from $q(\mathbf{f}_1)q(\mathbf{f}_2)$ (*i.e.* one of the two types of coloured lines in panels $\mathbf{f}_1$ and $\mathbf{f}_2$). Given such a sample, we can deduce the functions f_1, f_2, f_3. For example, the colour of $\mathbf{f}_1$ denotes the choice of f_1, the second colour of $\mathbf{f}_2$ (the first colour is that of f_1) corresponds to f_2, and f_3 is chosen to map $\mathbf{f}_2$ to the observations. Thus each sample from $q(\mathbf{f}_1)q(\mathbf{f}_2)$ corresponds to a composition mapping $\mathbf{x}$ to $\mathbf{y}$ (different samples correspond to different compositions). This is in contrast to sampling from the factorised distribution of the inducing points (which directly parametrise each $\{f_i\}$). In such case, some compositions (*e.g.* $f_1(x) = -x, f_2(x) = f_3(x) = x$) do not fit the data, making the variational posterior collapse, as argued in Sec. 3.

Inducing inputs We introduce inducing inputs $\mathbf{z} \in \mathbb{R}^M$ (with $M < N$) in the input space and denote the evaluations of intermediate layers at $\mathbf{z}$ as $\mathbf{f}_\ell^{\mathbf{z}} \sim (f_\ell \circ \ldots \circ f_1)(\mathbf{z})$. The augmented DGP joint distribution is

$$\begin{aligned} &p(\mathbf{y}, \mathbf{f}_L, \ldots, \mathbf{f}_1, \mathbf{f}_L^{\mathbf{z}}, \ldots, \mathbf{f}_1^{\mathbf{z}} \mid \mathbf{x}, \mathbf{z}) = \\ &= p(\mathbf{y} \mid \mathbf{f}_L) \prod_{\ell=1}^{L} p(\mathbf{f}_\ell \mid \mathbf{f}_\ell^{\mathbf{z}}, \mathbf{f}_{\ell-1}, \mathbf{f}_{\ell-1}^{\mathbf{z}}) p(\mathbf{f}_\ell^{\mathbf{z}} \mid \mathbf{f}_{\ell-1}^{\mathbf{z}}), \end{aligned} \tag{18}$$

where $p(\mathbf{f}_\ell^{\mathbf{z}} \mid \mathbf{f}_{\ell-1}^{\mathbf{z}}) \sim \mathcal{N}(\mu_\ell(\mathbf{f}_{\ell-1}^{\mathbf{z}}), k_\ell(\mathbf{f}_{\ell-1}^{\mathbf{z}}, \mathbf{f}_{\ell-1}^{\mathbf{z}}))$ is an ℓ-th layer GP prior, and $p(\mathbf{f}_\ell \mid \mathbf{f}_\ell^{\mathbf{z}}, \mathbf{f}_{\ell-1}, \mathbf{f}_{\ell-1}^{\mathbf{z}})$ is an ℓ-th layer GP posterior at inputs $\mathbf{f}_{\ell-1}$ given $\mathbf{f}_\ell^{\mathbf{z}}$ and $\mathbf{f}_{\ell-1}^{\mathbf{z}}$ in ℓ-th and $(\ell-1)$-th layers respectively.

Variational lower bound We introduce the following variational distribution

$$q(\{\mathbf{f}_\ell\}, \{\mathbf{f}_\ell^{\mathbf{z}}\}) = \prod_{\ell=1}^{L} p(\mathbf{f}_\ell \mid \mathbf{f}_\ell^{\mathbf{z}}, \mathbf{f}_{\ell-1}, \mathbf{f}_{\ell-1}^{\mathbf{z}}) q(\mathbf{f}_\ell^{\mathbf{z}}), \tag{19}$$

where $q(\mathbf{f}_\ell^{\mathbf{z}}) \sim \mathcal{N}(\mathbf{m}_\ell, \mathbf{S}_\ell)$. The corresponding likelihood lower bound is as follows

$$\begin{aligned} &\mathcal{L}(\mathbf{y}) \geq \mathbb{E}_q \left[\log \frac{p(\mathbf{y}, \{\mathbf{f}_\ell\}, \{\mathbf{f}_\ell^{\mathbf{z}}\})}{q(\{\mathbf{f}_\ell\}, \{\mathbf{f}_\ell^{\mathbf{z}}\})} \right] = \\ &= \mathbb{E}_{q(\mathbf{f}_L)}[\log p(\mathbf{y} \mid \mathbf{f}_L)] - \qquad (20) \\ &\quad - \sum_{\ell=1}^{L} \mathbb{E}_{q(\mathbf{f}_\ell^{\mathbf{z}}) q(\mathbf{f}_{\ell-1}^{\mathbf{z}})} \left[\log \frac{q(\mathbf{f}_\ell^{\mathbf{z}})}{p(\mathbf{f}_\ell^{\mathbf{z}} \mid \mathbf{f}_{\ell-1}^{\mathbf{z}})} \right]. \qquad (21) \end{aligned}$$

Estimating (20) We are estimating an expectation over the marginal $q(\mathbf{f}_L) \sim (f_L \circ \ldots \circ f_1)(\mathbf{x})$, which can be computed by marginalising the intermediate layers in the joint variational posterior (19):

$$\begin{aligned} q(\mathbf{f}_L) &= \int q(\{\mathbf{f}_\ell\}, \{\mathbf{f}_\ell^{\mathbf{z}}\}) \, \mathrm{d}\{\mathbf{f}_\ell\}_{\ell=1}^{L-1} \, \mathrm{d}\{\mathbf{f}_i^{\mathbf{z}}\}_{\ell=1}^{L} \\ &= \int p(\mathbf{f}_L \mid \mathbf{f}_L^{\mathbf{z}}, \mathbf{f}_{L-1}, \mathbf{f}_{L-1}^{\mathbf{z}}) q(\mathbf{f}_L^{\mathbf{z}}) \, \mathrm{d}\mathbf{f}_L^{\mathbf{z}} \times \\ &\quad \times \prod_{\ell=1}^{L-1} p(\mathbf{f}_\ell \mid \mathbf{f}_\ell^{\mathbf{z}}, \mathbf{f}_{\ell-1}, \mathbf{f}_{\ell-1}^{\mathbf{z}}) q(\mathbf{f}_\ell^{\mathbf{z}}) \, \mathrm{d}\mathbf{f}_\ell \, \mathrm{d}\mathbf{f}_\ell^{\mathbf{z}}. \end{aligned} \tag{22}$$

The integrals in (22) are generally intractable since they require integrating the kernel matrices, thus we estimate them by sampling. Overall, the procedure is as follows:

1. Draw S samples $\{(\mathbf{f}_1^{\mathbf{z},s}, \ldots, \mathbf{f}_L^{\mathbf{z},s})\}_{s=1}^{S} \overset{iid}{\sim} q(\mathbf{f}_1^{\mathbf{z}}) \cdot \ldots \cdot q(\mathbf{f}_L^{\mathbf{z}})$,
2. Use the samples of $\{\mathbf{f}_\ell^{\mathbf{z}}\}$ to sequentially draw samples of intermediate layers $\mathbf{f}_\ell^s \sim p(\mathbf{f}_\ell \mid \mathbf{f}_\ell^{\mathbf{z},s}, \mathbf{f}_{\ell-1}^s, \mathbf{f}_{\ell-1}^{\mathbf{z},s})$ from a GP posterior given $\mathbf{f}_\ell^{\mathbf{z},s}$ and $\mathbf{f}_{\ell-1}^{\mathbf{z},s}$,
3. Use $\{\mathbf{f}_L^s\}_{s=1}^s$, the samples from $q(\mathbf{f}_L)$, to estimate the expectation in (20): $\mathbb{E}_{q(\mathbf{f}_L)}[\log p(\mathbf{y} \mid \mathbf{f}_L)] \approx \frac{1}{S} \sum_{s=1}^{S} \log p(\mathbf{y} \mid \mathbf{f}_L^s)$.

Estimating (21) We write the summands in (21) as

$$\begin{aligned} &\mathbb{E}_{q(\mathbf{f}_\ell^{\mathbf{z}}) q(\mathbf{f}_{\ell-1}^{\mathbf{z}})} \left[\log \frac{q(\mathbf{f}_\ell^{\mathbf{z}})}{p(\mathbf{f}_\ell^{\mathbf{z}} \mid \mathbf{f}_{\ell-1}^{\mathbf{z}})} \right] = \\ &\quad = \mathbb{E}_{q(\mathbf{f}_{\ell-1}^{\mathbf{z}})} \mathrm{KL}[q(\mathbf{f}_\ell^{\mathbf{z}}) \,||\, p(\mathbf{f}_\ell^{\mathbf{z}} \mid \mathbf{f}_{\ell-1}^{\mathbf{z}})]. \end{aligned} \tag{23}$$

KL divergence between the two Gaussians $q(\mathbf{f}_\ell^{\mathbf{z}})$ and $p(\mathbf{f}_\ell^{\mathbf{z}} \mid \mathbf{f}_{\ell-1}^{\mathbf{z}})$ is a function of $\mathbf{f}_{\ell-1}^{\mathbf{z}}$ and can be computed analytically for a given value of $\mathbf{f}_{\ell-1}^{\mathbf{z}}$. Therefore, to estimate it, we use the draws from $\mathbf{f}_{\ell-1}^{\mathbf{z}}$ (which are computed for the estimate of (20) as well): for every such draw $\mathbf{f}_{\ell-1}^{\mathbf{z},s}$, we analytically compute the KL divergence $\mathrm{KL}[q(\mathbf{f}_\ell^{\mathbf{z}}) \,||\, p(\mathbf{f}_\ell^{\mathbf{z}} \mid \mathbf{f}_{\ell-1}^{\mathbf{z},s})]$, and then average these values to obtain a Monte-Carlo estimate of the expectation in (23).

Learning and predictions We maximise the likelihood lower bound (20-21) w.r.t. the variational parameters $\{\mathbf{m}_\ell\}$ and $\{\mathbf{S}_\ell\}$. The gradients can be obtained using a reparametrisation trick (Kingma and Welling, 2014). Given a test input $\mathbf{x}^*$, we can draw the DGP outputs $\mathbf{f}_L^* \sim (f_L \circ \ldots \circ f_1)(\mathbf{x}^*)$ by drawing from $q(\mathbf{f}_l^*)$ using the procedure for estimating (22) described above. We substitute $\mathbf{x}^*$ instead of $\mathbf{x}$ replacing $\mathbf{f}_\ell$ with $\mathbf{f}_\ell^*$ in (22), while the rest of the procedure remains the same.

Time complexity The time complexity of estimating (20) is $O(LNM^3)$. Sampling from $q(\mathbf{f}_i^{\mathbf{z}})$ is $O(M^3)$, while, as discussed in (Salimbeni and Deisenroth, 2017), sampling from $p(\mathbf{f}_i | \mathbf{f}_i^{\mathbf{z}}, \mathbf{f}_{i-1}, \mathbf{f}_{i-1}^{\mathbf{z}})$ can be performed separately for each element of $\mathbf{f}_i$ only requiring drawing from univariate Gaussians, which scales linearly with the

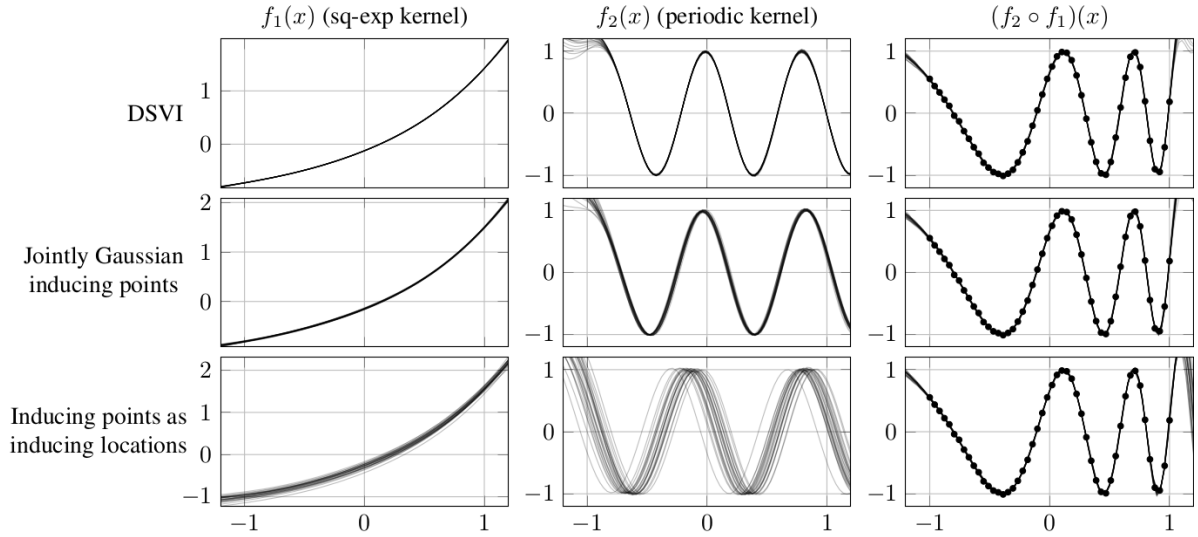

Figure 5: 25 random samples from 2-layer DGPs with squared-exponential and periodic kernels fitted to the observations in the third column (black dots) using DSVI as well as variational distributions discussed in Sec. 4. The first and second columns show samples from each of the two layers, while the third one shows samples from the entire composition (all such samples fit the data despite the variance in f_1 and f_2 because the two layers are dependent).

	ELBO	$\mathbb{V}\mathrm{ar}\left[f_1(0)\right]$	$\mathbb{V}\mathrm{ar}\left[f_2(0)\right]$
DSVI	13.43 ± 8.03	$1.99 \cdot 10^{-6} \pm 1.76 \cdot 10^{-7}$	$1.11 \cdot 10^{-4} \pm 1.35 \cdot 10^{-5}$
Jointly Gaussian	23.15 ± 6.80	$4.23 \cdot 10^{-5} \pm 3.17 \cdot 10^{-6}$	$3.33 \cdot 10^{-4} \pm 2.12 \cdot 10^{-5}$
Inducing points as inducing inputs	36.31 ± 3.55	$2.22 \cdot 10^{-3} \pm 2.73 \cdot 10^{-4}$	$4.98 \cdot 10^{-2} \pm 7.78 \cdot 10^{-3}$

Table 1: Evaluations of the DGPs fitted on a dataset in Fig. 5. First column shows lower bounds on marginal likelihood $p(\mathbf{y})$; the second and third ones show marginal variances of both layers at $x = 0$. The numbers are the means as well as standard deviations across 10 trials.

number of layers and training inputs. The estimate of (21) does not add additional complexity since we use the samples from $q(\mathbf{f}_i^{\mathbf{z}})$ drawn for estimating (20), while analytic computation of the KL divergence between $q(\mathbf{f}_\ell^{\mathbf{z}})$ and $p(\mathbf{f}_\ell^{\mathbf{z}} \mid \mathbf{f}_{\ell-1}^{\mathbf{z}})$ is $O(M^3)$ since it requires inversions of covariance matrices. Therefore, the overall complexity of estimating the lower bound is $O(LNM^3)$.

5 NUMERICAL SIMULATIONS

Compositional uncertainty As illustrated in Fig. 5 (first row) as well as in Table 1, the intermediate layers in a DGP with a factorised variational distribution over the inducing points collapse to nearly deterministic transformations in the range of the observed data ($[-1, 1]$). Meanwhile, the models with correlated inducing points (second and third rows) capture more uncertainty, with the approach proposed in Sec. 4.2 allowing us to capture more uncertainty than jointly Gaussian inducing points. Additional examples are provided in the Supplement.

Likelihood lower bounds In Table 1 we provide the variational lower bounds of the marginal likelihood[5], $p(\mathbf{y})$. We see that including the dependencies between the layers to the variational distribution leads to higher likelihood bounds, suggesting that factorised variational distributions are suboptimal for DGP inference.

6 APPLICATIONS

As compositions of functions, DGPs provide a natural way to represent data that is known to have a compositional structure and thus they may be used in applications to learn a more informative representation of the data.

Non-stationary time series Consider a sequence $\mathbf{y} \in \mathbb{R}^N$ that is observed at fixed time inputs $\mathbf{x} \in \mathbb{R}^N$. The

[5]The baseline estimate of the true marginal likelihood could be obtained by fitting the DGP using HMC (Havasi et al., 2018), however, we found the existing implementation of this scheme to be very unstable (as also noted by the authors) and the estimation of marginal likelihood from posterior samples to have high variance, hence we do not report such values.

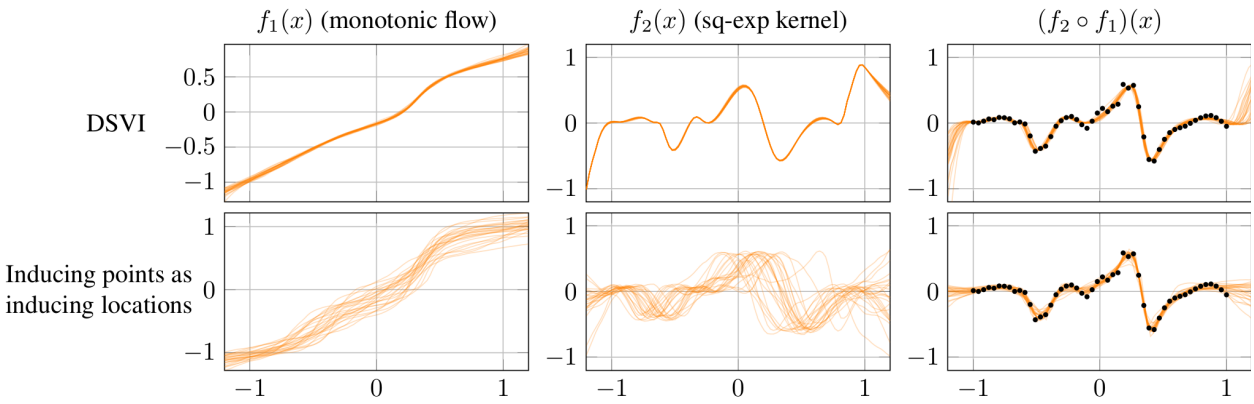

Figure 6: Compositional model of heartbeats data, comparing results without (top) and with correlations across layers.

observed sequence is assumed to be generated by temporally warping the inputs $\mathbf{x}$ as follows:

$$\mathbf{y} = f(g(\mathbf{x})) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{24}$$

where $g(\cdot)$ is the temporal warping, $f(\cdot)$ is the latent function that encodes the structure of the observed sequence. The model in (24) generates non-stationary sequences, which are convenient to model with a composition of a monotonic transformation of the inputs $\mathbf{x}$ and a GP with a stationary kernel. The previous work on such models treats the temporal warping $g(\cdot)$ as a deterministic transformation (Snoek et al., 2014; Kazlauskaite et al., 2019), disregarding the fact that many different compositions may explain the observed data equally well.

To illustrate this, we consider a recording of a heartbeat (Bentley et al., 2011), and fit a two layer DGP with monotonic flow (Ustyuzhaninov et al., 2020) in the first layer. Here the prior on the warping functions $g(\cdot)$ dictates that while an identity warp is preferred, other smooth warps are plausible. The latent functions $f(\cdot)$ are modelled using a GP with a stationary squared exponential kernel. Fig. 6 shows how introducing correlations between the layers allows us to uncover a wide range of possible solutions that follow the above-defined priors and are consistent with the data. Meanwhile, the model with the same prior assumptions that uses a mean-field approximation collapses to a near-deterministic transformation, concentrating the probability mass in both layers on one of the many possible solutions. An application to sequence alignment is provided in the Supplement.

7 DISCUSSION

We have discussed the issue of compositional uncertainty in the context of DGPs. This is in contrast to much of the existing work on DGPs (as well as other Bayesian deep learning approaches (Gal, 2016)) that primarily focuses on predictive uncertainty. We argued that the uncertainty about the function implemented by each individual layer in the hierarchy provides a more informative model of the data. The inference in DGP models is typically performed using variational approximations that factorise across the layers of the hierarchy. While computationally convenient, such a factorisation implies that the distributions of the intermediate layers collapse to deterministic transformations. Such behaviour diminishes some of the other benefits offered by a compositional model of GPs, such as a systematic way to impose informative functional priors over each of the layers in the hierarchy and a way to uncover distributions over each layer.

To gain further insight into the issue of compositional uncertainty, we proposed two alternatives to the factorised variational distributions of inducing points that include some correlations between the layers. Contrary to the factorised distributions in DSVI, the proposed variational distributions uncover a range of possible solutions, reinforcing the argument that mean-field approximations are prohibitive when it comes to capturing compositional uncertainty. These consideration pose many open questions, ranging from technical considerations of more efficient ways to introduce correlations across layers and ways to represent variational distributions that are multimodal (Lawrence, 2000), to broader questions about the structures captured by each layer of the hierarchy, and the applications that may benefit from the more accurate estimates of compositional uncertainty.

Acknowledgments

This work has been supported by EPSRC CDE (EP/L016540/1), CAMERA (EP/M023281/1), EPSRC DTP, Hans Werthén Fund at The Royal Swedish Academy of Engineering Sciences, German Federal Ministry of Education and Research (project 01 IS 18049 A) and the Royal Society.

References

Bentley, P., Nordehn, G., Coimbra, M., and Mannor, S. (2011). Pascal Classifying Heart Sounds Challenge.

Bijl, H. (2018). LQG and Gaussian process techniques: For fixed-structure wind turbine control. *PhD thesis, Delft University of Technology*.

Bui, T., Hernandez-Lobato, D., Hernandez-Lobato, J., Li, Y., and Turner, R. (2016). Deep Gaussian processes for regression using approximate expectation propagation. In *International Conference on Machine Learning*.

Cutajar, K. (2019). *Broadening the scope of Gaussian processes for large-scale learning*. PhD thesis, Thesis.

Dai, Z., Damianou, A., González, J., and Lawrence, N. D. (2016). Variational auto-encoded deep Gaussian processes. In *International Conference on Learning Representations*.

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

Damianou, A. and Lawrence, N. (2013). Deep Gaussian processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Dunlop, M. M., Girolami, M. A., Stuart, A. M., and Teckentrup, A. L. (2018). How deep are deep Gaussian processes? *Journal of Machine Learning Research*.

Duvenaud, D., Rippel, O., Adams, R. P., and Ghahramani, Z. (2014). Avoiding pathologies in very deep networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Gal, Y. (2016). *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*.

Girard, A., Rasmussen, C. E., Candela, J. Q., and Murray-Smith, R. (2003). Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. In *Neural Information Processing Systems*.

Havasi, M., Hernández-Lobato, J. M., and Murillo-Fuentes, J. J. (2018). Inference in deep Gaussian processes using stochastic gradient Hamiltonian Monte Carlo. In *Neural Information Processing Systems*.

Hensman, J., Durrande, N., and Solin, A. (2017). Variational fourier features for Gaussian processes. *Journal of Machine Learning Research (JMLR)*, 18(1).

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

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

Jin, M., Damianou, A., Abbeel, P., and Spanos, C. (2017). Inverse reinforcement learning via deep Gaussian process. *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Kaiser, M., Otte, C., Runkler, T., and Ek, C. H. (2018). Bayesian alignments of warped multi-output Gaussian processes. In *Neural Information Processing Systems*.

Kazlauskaite, I., Ek, C. H., and Campbell, N. (2019). Gaussian process latent variable alignment learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Lawrence, N. D. (2000). *Variational Inference in Probabilistic Models*. PhD thesis, Cambridge University.

Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. *Neural Information Processing Systems*.

Lawrence, N. D. and Moore, A. J. (2007). Hierarchical Gaussian process latent variable models. In *International Conference on Machine Learning*.

Lázaro-Gredilla, M. (2012). Bayesian warped Gaussian processes. In *Neural Information Processing Systems*.

Mchutchon, A. and Rasmussen, C. E. (2011). Gaussian process training with input noise. In *Neural Information Processing Systems*.

Rainforth, T., Cornish, R., Yang, H., Warrington, A., and Wood, F. (2019). On nesting Monte Carlo estimators. *Proceedings of Machine Learning Research*, 80.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. MIT Press.

Rudner, T. G. J. and Sejdinovic, D. (2017). Inter-domain deep Gaussian processes.

Salimbeni, H. and Deisenroth, M. (2017). Doubly stochastic variational inference for deep Gaussian processes. In *Neural Information Processing Systems*.

Snoek, J., Swersky, K., Zemel, R., and Adams, R. (2014). Input warping for Bayesian optimization of non-stationary functions. In *International Conference on Machine Learning*.

Sun, S., Zhang, G., Shi, J., and Grosse, R. (2019). Functional variational Bayesian neural networks. In *International Conference on Learning Representations*.

Titsias, M. and Lawrence, N. (2010). Bayesian Gaussian process latent variable model. *Journal of Machine Learning Research (JMLR)*, 9.

Ustyuzhaninov, I., Kazlauskaite, I., Ek, C. H., and Campbell, N. D. F. (2020). Monotonic Gaussian process flow. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.