entropy

MDPI

Article

Conditional Deep Gaussian Processes: Multi-Fidelity Kernel Learning

Chi-Ken Lu [1,*] **and Patrick Shafto** [1,2]

[1] Mathematics and Computer Science, Rutgers University, Newark, NJ 07102, USA; patrick.shafto@rutgers.edu
[2] School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540, USA
* Correspondence: cl1178@rutgers.edu

Abstract: Deep Gaussian Processes (DGPs) were proposed as an expressive Bayesian model capable of a mathematically grounded estimation of uncertainty. The expressivity of DPGs results from not only the compositional character but the distribution propagation within the hierarchy. Recently, it was pointed out that the hierarchical structure of DGP well suited modeling the multi-fidelity regression, in which one is provided sparse observations with high precision and plenty of low fidelity observations. We propose the conditional DGP model in which the latent GPs are directly supported by the fixed lower fidelity data. Then the moment matching method is applied to approximate the marginal prior of conditional DGP with a GP. The obtained effective kernels are implicit functions of the lower-fidelity data, manifesting the expressivity contributed by distribution propagation within the hierarchy. The hyperparameters are learned via optimizing the approximate marginal likelihood. Experiments with synthetic and high dimensional data show comparable performance against other multi-fidelity regression methods, variational inference, and multi-output GP. We conclude that, with the low fidelity data and the hierarchical DGP structure, the effective kernel encodes the inductive bias for true function allowing the compositional freedom.

Keywords: multi-fidelity regression; Deep Gaussian Process; approximate inference; moment matching; kernel composition; neural network.

Citation: Lu, C.-K.; Shafto, P. Conditional Deep Gaussian Processes: Multi-Fidelity Kernel Learning. *Entropy* **2021**, *23*, 1545. https://doi.org/10.3390/e23111545

Academic Editors: Eric Nalisnick and Dustin Tran

Received: 1 October 2021
Accepted: 18 November 2021
Published: 20 November 2021

1. Introduction

Multi-fidelity regression refers to a category of learning tasks in which a set of sparse data is given to infer the underlying function but a larger amount of less precise or noisy observations is also provided. Multifidelity tasks frequently occur in various fields of science because precise measurement is often costly while approximate measurements are more affordable (see [1,2] for example). The assumption that the more precise function is a function of the less precise one [1,3] is shared in some hierarchical learning algorithms (e.g., one-shot learning in [4], meta learning [5], and continual learning [6]). Thus, one can view the plentiful low fidelity data as a source of prior knowledge so the function can be efficiently learned with sparse data.

In Gaussian Process (GP) regression [7] domain experts can encode their knowledge into the combinations of covariance functions [8,9], building an expressive learning model. However, construction of an appropriate kernel becomes less clear when building a prior for the precise function in the context of multi-fidelity regression because the uncertainty, both epistemic and aleatoric, in the low fidelity function prior learned by the plentiful data should be taken into account. It is desirable to fuse the low fidelity data to an *effective* kernel as a prior, taking advantage of marginal likelihood being able to avoid overfitting, and then perform the GP regression as if only the sparse precise observations are given.

Deep Gaussian Process (DGP) [10] and the similar models [11,12] are expressive models with a hierarchical composition of GPs. As pointed out in [3], a hierarchical structure is particularly suitable for fusing data of different fidelity into one learning

model. Although full Bayesian inference is promising in obtaining expressiveness while avoiding overfitting, exact inference is not tractable and approximate solutions such as the variational approach [13–15] are employed. Ironically, the major difficulty in inference comes from marginalization of the latent GPs in Bayesian learning, which, on the flip side, is also why overfitting can be prevented.

We propose a conditional DGP model in which the intermediate GPs are supported by the lower fidelity data. We also define the corresponding marginal prior distribution which is obtained by marginalizing all GPs except the exposed one. For some families of kernel compositions, we previously developed an analytical method in calculating exact covariance in the marginal prior [16]. As such, the method is applied here so the marginal prior is approximated as a GP prior with an effective kernel. The high fidelity data are then connected to the exposed GP, and the hyperparameters throughout the hierarchy are optimized via the marginal likelihood. Our model, therefore, captures the expressiveness embedded in hierarchical composition, retains the Bayesian character hinted in the marginal prior, but loses the non-Gaussian aspect of DGP. From the analytical expressions, one can partially understand the propagation of uncertainty in latent GPs as it is responsible for the non-stationary aspect of effective kernels. Moreover, the compositional freedom, i.e., different compositions may result in the same target function, in a DGP model [17,18] can be shown to be intact in our approach.

The paper is organized as follows. In Section 2, we review the literature of multi-fidelity regression model and deep kernels. A background of GP, DGP, and the moment matching method is introduced in Section 3. The conditional DGP model defined as a marginal prior and the exact covariance associated with two families of kernel compositions are discussed in Section 4. The method of hyperparameter learning is given in Section 5, and the simulation of synthetic and high dimensional multi-fidelity regression in a variety of situations are presented in Section 6. A brief discussion followed by the conclusion appear in Sections 7 and 8, respectively.

2. Related Work

Assuming autoregressive relations between data of different fidelity, Kennedy and O'Hagan [1] proposed the AR1 model for multi-fidelity regression tasks. Le Gratiet and Garnier [19] improved computational efficiency with a recursive multi-fidelity model. Deep-MF [20] mapped the input space to the latent space and followed the work in Kennedy and O'Hagan [1]. NARGP [21] stacked a sequence of GPs in which the posterior mean about the low-fidelity function is passed to the input of the next GP while the associated uncertainty is not. GPAR [22] uses a similar conditional structure between functions of interest. MF-DGP in [3] exploited the DGP structure for the multi-fidelity regression tasks and used the approximate variational inference in [13]. Multi-output GPs [23,24] regard the observations from different data sets as realization of vector-valued function; [25] modeled the multi-output GP using general relation between multiple target functions and multiple hidden functions. Alignment learning [26,27] is an application of warped GP [11,12] to time series data. We model the multi-fidelity regression as a kernel learning, effectively taking the space of functions representing the low fidelity data into account.

As for general studies of deep and non-stationary kernels, Williams [28] and Cho and Saul [29] used the basis of error functions and Heaviside polynomial functions to obtain the arc-sine and arc-cosine kernel functions, respectively, of neural networks. Duvenaud et al. [30] employed the analogy between neural network and GP, and constructed the deep kernel for DGP. Dunlop et al. [31] analyzed variety of non-stationary kernel compositions in DGP, and Shen et al. [32] provided an insight from Wigner transformation of general two-input functions. Wilson et al. [33] proposed the general recipe for constructing the deep kernel with neural networks. Daniely et al. [34] computed the deep kernel from the perspective of two correlated random variables. Mairal et al. [35] and Van der Wilk et al. [36] studied the deep kernels in the convolutional models. The mo-

ment matching method [16] allows obtaining the effective kernel encoding the uncertainty in learning the lower fidelity function.

3. Background

3.1. Gaussian Process and Deep Gaussian Process

Gaussian Process (GP) [7] is a popular Bayesian learning model in which the prior over a continuous function is modeled as a Gaussian. Denoted by $f \sim \mathcal{GP}(\mu, k)$, the collection of any finite function values $f(\mathbf{x}_{1:N})$ with $\mathbf{x} \in \mathbb{R}^d$ has the mean $\mathbb{E}[f(\mathbf{x}_i)] = \mu(\mathbf{x}_i)$ and covariance $\mathbb{E}\{[f(\mathbf{x}_i) - \mu(\mathbf{x}_i)][f(\mathbf{x}_j) - \mu(\mathbf{x}_j)]\} = k(\mathbf{x}_i, \mathbf{x}_j)$. Thus, a continuous and deterministic mean function $\mu(\cdot)$ and a positive definite kernel function $k(\cdot, \cdot)$ fully specify the stochastic process. It is common to consider the zero-mean case and write down the prior distribution, $p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(0, K)$ with covariance matrix K. In the setting of a regression task with input and output of data $\{\mathbf{X}, \mathbf{y}\}$, the hyperparameters in the mean and kernel functions can be learned by optimizing the marginal likelihood, $p(\mathbf{y}|\mathbf{X}) = \int d\mathbf{f} p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X})$.

Deep Gaussian Process (DGP) was proposed in [10] as a hierarchical composition of GPs for superior expressivity. From a generative view, the distribution over the composite function $f(\mathbf{x}) = f_L \circ f_{L-1} \circ \cdots \circ f_1(\mathbf{x})$ is a serial product of Gaussian conditional distribution,

$$p(\mathbf{F}_L, \mathbf{F}_{L-1}, \cdots, \mathbf{F}_1|\mathbf{x}) = p(\mathbf{F}_L|\mathbf{F}_{L-1}) p(\mathbf{F}_{L-1}|\mathbf{F}_{L-2}) \cdots p(\mathbf{F}_1|\mathbf{X}), \tag{1}$$

in which the capital bolded face symbol $\mathbf{F}_i$ stands for a multi-output GP in ith layer and the independent components have $p(\mathbf{f}_i|\mathbf{F}_{i-1}) = \mathcal{N}(0, K(\mathbf{F}_{i-1}.\mathbf{F}_{i-1}))$. The above is the general DGP, and the width in each layer is denoted by $H_i := |\mathbf{F}_i|$. In such notation, the zeroth layer represents the collection of inputs $\mathbf{X}$. Here, we shall consider the DGP with $L = 2$ and $H_2 = H_1 = 1$ and the three-layer counterpart.

The intractability of exact inference is a result of the fact that the random variables $\mathbf{F}_i$ for $L > i > 0$ appear in the covariance matrix K. In a full Bayesian inference, the random variables are marginalized in order to estimate the evidence $p(\mathbf{y}|\mathbf{X})$ associated with the data.

3.2. Multi-Fidelity Deep Gaussian Process

The multi-fidelity model in [1] considered the regression task for a data set consisting of observations measured with both high and low precision. For simplicity, the more precise observations are denoted by $\{\mathbf{X}, \mathbf{y}\}$ and those with low precision by $\{\mathbf{X}_1, \mathbf{y}_1\}$. The main assumption made in [1] is that the less precise observations shall come from a function $f_1(\mathbf{x})$ modeled by a GP with zero mean and kernel k, while the more precise ones come from the combination $f(\mathbf{x}) = \alpha f_1(\mathbf{x}) + h(\mathbf{x})$. With the residual function h being a GP with kernel k_h, one can jointly model the two subsets with the covariance within precise observations $\mathbb{E}[f(\mathbf{x}_i) f(\mathbf{x}_j)] = \alpha^2 k_{ij} + k_{h_{ij}}$, within the less precise ones $\mathbb{E}[f_1(\mathbf{x}_i) f_1(\mathbf{x}_j)] = k_{ij}$, and the mutual covariance $\mathbb{E}[f(\mathbf{x}_i) f_1(\mathbf{x}_j)] = \alpha k_{ij}$. k_{ij} refers to the covariance between the two inputs at $\mathbf{x}_i$ and $\mathbf{x}_j$.

The work in [3] generalized the above the linear relationship between the more and less precise functions to a nonlinear one, i.e., $f(\mathbf{x}) = f_2(f_1(\mathbf{x})) + \text{noise}$. The hierarchical structure in DGP is suitable for nonlinear modeling. The variational inference scheme [13] was employed to evaluate the evidence lower bounds (ELBOs) for the data with all levels of precision, and the sum over all ELBOs is the objective for learning the hyperparameters and inducing points.

3.3. Covariance in Marginal Prior of DGP

The variational inference, e.g., [13], starts with connecting the joint distribution $p(f_1, f_2|\mathbf{X})$ with data $\mathbf{y}$, followed by applying the Jensen's inequality along with an approx-

imate posterior in evaluating the ELBO. In contrast, we proposed in [16] that the marginal prior for the DGP,

$$p(\mathbf{f}|\mathbf{X}) = \int d\mathbf{f}_1 p(\mathbf{f}_2|\mathbf{f}_1) p(\mathbf{f}_1|\mathbf{X}) \,, \tag{2}$$

with the bolded face symbols representing the set of function values, $f(\cdot) = f_2(f_1(\cdot))$, $f_2(\cdot)$, and $f_1(\cdot)$, can be approximated as a GP, i.e., $q(\mathbf{f}|\mathbf{X}) = \mathcal{N}(0, K_{\text{eff}})$ in the zero-mean case. The matching of covariance in p and q leads to the closed form of effective covariance function for certain families of kernel compositions. The SE[SC] composition, i.e., the squared exponential and squared cosine kernels being used in the GPs for $f_2|f_1$ and f_1, respectively, is an example. With the intractable marginalization over the intermediate f_1 being taken care of in the moment matching approximation, one can evaluate the approximate marginal likelihood for the data set $\{\mathbf{X}, \mathbf{y}\}$,

$$p(\mathbf{y}|\mathbf{X}) \approx \int d\mathbf{f} p(\mathbf{y}|\mathbf{f}) q(\mathbf{f}|\mathbf{X}) \,. \tag{3}$$

In the following, we shall develop along the line of [16] the approximate inference for the multi-fidelity data consisting of precise observations $\{\mathbf{X}, \mathbf{y}\}$ and less precise observations $\{\mathbf{X}_{1:L-1}, \mathbf{y}_{1:L-1}\}$ with the L-layer width-1 DGP models. The effective kernels k_{eff} shall encode the knowledge built on these less precise data, which allows modeling the precise function even with a sparse data set.

4. Conditional DGP and Multi-Fidelity Kernel Learning

In the simplest case, we are given two subsets of data, $\{\mathbf{X}, \mathbf{y}\}$ with high precision and $\{\mathbf{X}_1, \mathbf{y}_1\}$ with low precision. We can start with defining the conditional DGP in terms of the marginal prior,

$$p(\mathbf{f}|\mathbf{X}, \mathbf{X}_1, \mathbf{y}_1) = \int d\mathbf{f}_1 p(\mathbf{f}_2|\mathbf{f}_1) p(\mathbf{f}_1|\mathbf{X}, \mathbf{X}_1, \mathbf{y}_1) \,, \tag{4}$$

where the Gaussian distribution $p(\mathbf{f}_1|\mathbf{X}, \mathbf{X}_1, \mathbf{y}_1) = \mathcal{N}(f_1(\mathbf{x}_{1:N})|\mathbf{m}, \Sigma)$ has the conditional mean in the vector form,

$$\mathbf{m} = K_{\mathbf{X},\mathbf{X}_1} K_{\mathbf{X}_1,\mathbf{X}_1}^{-1} \mathbf{y}_1 \,, \tag{5}$$

and the conditional covariance in the matrix form,

$$\Sigma = K_{\mathbf{X},\mathbf{X}} - K_{\mathbf{X},\mathbf{X}_1} K_{\mathbf{X}_1,\mathbf{X}_1}^{-1} K_{\mathbf{X}_1,\mathbf{X}} \,. \tag{6}$$

The matrix $K_{\mathbf{X},\mathbf{X}_1}$ registers the covariance among the inputs in $\mathbf{X}$ and $\mathbf{X}_1$, and likewise for $K_{\mathbf{X},\mathbf{X}}$ and $K_{\mathbf{X}_1,\mathbf{X}_1}$. Thus, the set of function values $f_1(\mathbf{x}_{1:N})$ associated with the N inputs in $\mathbf{X}$ are supported by the low fidelity data.

The data $\{\mathbf{X}, \mathbf{y}\}$ with high precision are then associated with the function $f(\mathbf{x}) = f_2(f_1(\mathbf{x}))$. Following the previous discussion, we may write down the true evidence for the precise data conditioned on the less precise ones shown below,

$$p(\mathbf{y}|\mathbf{X}, \mathbf{X}_1, \mathbf{y}_1) = \int d\mathbf{f} p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \mathbf{X}_1, \mathbf{y}_1) = \int d\mathbf{f}_1 d\mathbf{f}_2 p(\mathbf{y}|\mathbf{f}_2) p(\mathbf{f}_2|\mathbf{f}_1) p(\mathbf{f}_1|\mathbf{X}, \mathbf{X}_1, \mathbf{y}_1) \,. \tag{7}$$

To proceed with the moment matching approximation of the true evidence in Equation (7), one needs to find the effective kernel in the approximate distribution $q(\mathbf{f}|\mathbf{X}, \mathbf{X}_1, \mathbf{y}_1) = \mathcal{N}(0, K_{\text{eff}})$ and replace the true distribution in Equation (4) with the approximate distribution,

$$p(\mathbf{y}|\mathbf{X}, \mathbf{X}_1, \mathbf{y}_1) \approx \int d\mathbf{f} p(\mathbf{y}|\mathbf{f}) q(\mathbf{f}|\mathbf{X}, \mathbf{X}_1, \mathbf{y}_1) = \mathcal{N}(\mathbf{y}|0, K_{\text{eff}} + \sigma_n^2 I_N) \,. \tag{8}$$

Therefore, the learning in the conditional DGP includes the hyperparameters in the exposed GP, $f_2|f_1$, and those in the intermediate GP, f_1. Standard gradient descent is applied to above approximate marginal likelihood. One can customize the kernel K_{eff} in the

GPy [37] framework and implement the gradient components $\partial K_{\text{eff}}/\partial\theta$ with $\theta \in \{\sigma_{1,2}, \ell_{1,2}\}$ in the optimization.

4.1. Analytic Effective Kernels

Here, we consider the conditional DGP with two-layer and width-1 hierarchy, focusing on the cases where the exposed GP for $f_2|f_1$ in Equation (4) uses the squared exponential (SE) kernel or the squared cosine (SC) kernel. We also follow the notation in [16] so that the composition denoted by $k_2[k_1]$ represents that k_2 is the kernel used in the exposed GP and k_1 used in the intermediate GP. For example, SE[SC] means k_2 is SE while k_1 is SC. Following [16], the exact covariance in the marginal prior Equation (4) is calculated,

$$\mathbb{E}_{\mathbf{f}}[f(\mathbf{x}_i)f(\mathbf{x}_j)] := \mathbb{E}_{\mathbf{f}_1}[\mathbb{E}_{\mathbf{f}_2|\mathbf{f}_1}[f_2(f_1(\mathbf{x}_i))f_2(f_1(\mathbf{x}_j))]] = \int d\mathbf{f}_1\, k_2(f_1(\mathbf{x}_i), f_1(\mathbf{x}_j))p(\mathbf{f}_1|\mathbf{X}, \mathbf{X}_1, \mathbf{y}_1)\,. \quad (9)$$

Thus, when the exposed GP has the kernel k_2 in the exponential family, the above integral is tractable and the analytic k_{eff} can be implemented as a customized kernel. The following two lemmas from [16] are useful for the present case with a nonzero conditional mean and a conditional covariance in f_1.

Lemma 1. *(Lemma 2 in [16]) For a vector of Gaussian random variables $g_{1:n} \sim \mathcal{N}(\mathbf{m}, \mathbf{C})$, the expectation of exponential quadratic form $\exp[-\frac{1}{2}Q(g_1, g_2, \cdots, g_n)]$ with $Q(g_{1:n}) = \sum_{i,j} A_{ij} g_i g_j \geq 0$ has the following closed form,*

$$\mathbb{E}[e^{-\frac{1}{2}Q(g_{1:n})}] = \frac{\exp\left[-\frac{1}{2}\mathbf{m}^T(I_n + \mathbf{AC})^{-1}\mathbf{Am}\right]}{\sqrt{|I_n + \mathbf{CA}|}}\,. \quad (10)$$

The n-dimensional matrix $\mathbf{A}$ appearing in the quadratic form Q is symmetric.

Lemma 2. *(Lemma 3 in [16]) With the same Gaussian vector $\mathbf{g}$ in Lemma 1, the expectation value of the exponential inner product $\exp[\mathbf{a}^t\mathbf{g}]$ between $\mathbf{g}$ and a constant vector $\mathbf{a}$ reads,*

$$\mathbb{E}[e^{\mathbf{a}^t\mathbf{g}}] = \exp\{\mathbf{a}^t\mathbf{m} + \frac{1}{2}Tr[\mathbf{Caa}^t]\}\,, \quad (11)$$

where the transpose operation on column vector is denoted by the superscript.

We shall emphasize that our previous work [16] did not discuss the cases when the intermediate GP for f_1 is conditioned on the low precision data $\{\mathbf{X}_1, \mathbf{y}_1\}$. Thus, the conditional mean and the non-stationary conditional covariance were not considered in [16].

Lemma 3. *The covariance in the marginal prior with a SE $k_2(x, y) = \sigma_2^2 \exp[-(x-y)^2/2\ell_2^2]$ in the exposed GP can be calculated analytically. With the Gaussian conditional distribution, $p(\mathbf{f}_1|\mathbf{X}, \mathbf{X}_1, \mathbf{y}_1)$, supported by the low fidelity data, the effective kernel reads,*

$$k_{\text{eff}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sigma_2^2}{\sqrt{1 + \delta_{ij}^2/\ell_2^2}} \exp\left[-\frac{(m_i - m_j)^2}{2(\ell_2^2 + \delta_{ij}^2)}\right], \quad (12)$$

where $m_{i,j} := \mathbb{E}[f_1(\mathbf{x}_{i,j})|\mathbf{X}_1, \mathbf{y}_1]$ being the conditional mean of f_1. The positive parameter $\delta_{ij}^2 := c_{ii} + c_{jj} - 2c_{ij}$ is defined with the conditional covariance $c_{ij} := \text{cov}[f_1(\mathbf{x}_i), f_1(\mathbf{x}_j)|\mathbf{X}_1, \mathbf{y}_1]$. δ_{ij}^2 and and the the length scale ℓ_2 in k_2 dictates how the uncertainty in $f_1(\mathbf{x})$ affects the function composition.

Proof. For SE[ ] composition, one can represent the kernel $k_2 = \exp\{-[f_1(\mathbf{x}_i) - f_1(\mathbf{x}_j)]^2/2\}$ as an exponential quadratic form $\exp[-\frac{Q}{2}]$ with $Q = \mathbf{f}_1^t\mathbf{A}\mathbf{f}_1$ with $\mathbf{A} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$. $\ell_2 = 1$ is set for ease of notation. Now $\mathbf{f}_1$ is a bivariate Gaussian variable with mean $\mathbf{m}$ and

covariance matrix $\mathbf{C}$. To calculate the expectation value in Equation (12), we need to compute the following 2-by-2 matrix and one can show $[I_2 + \mathbf{AC}]^{-1}$ can be reduced to

$$\frac{1}{1 + c_{ii} + c_{jj} - 2c_{ij}} \begin{pmatrix} 1 + c_{jj} - c_{ij} & c_{jj} - c_{ij} \\ c_{ii} - c_{ij} & 1 + c_{ii} - c_{ij} \end{pmatrix}. \tag{13}$$

The seemingly complicated matrix in fact is reducible as one can show that $(I_2 + \mathbf{AC})^{-1}\mathbf{A} = \mathbf{A}/(1 + \delta_{ij}^2)$, which leads to the exponential term in the kernel. With the determinant $|I_2 + \mathbf{CA}| = (1 + \delta_{ij}^2)$ and restoring back the length scale ℓ_2, the kernel in Equation (12) is reproduced. □

A few observations are in order. First, we can rewrite $\delta_{ij}^2 = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} c_{ii} & c_{ij} \\ c_{ji} & c_{jj} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, which guarantees the positiveness of δ_{ij}^2 as the two-by-two sub-block of covariance matrix is positive-definite. Secondly, there are deterministic and probabilistic aspects of the kernel in Equation (12). When the the length scale ℓ_2 is very large, the term δ^2 encoding the uncertainty in f_1 becomes irrelevant and the kernel is approximately a SE kernel with the input transformed via the conditonal mean Equation (5), which is reminiscent of the deep kernel proposed in [33] where GP is stacked on the output of a DNN. The kernel used in [21] similarly considers the conditional mean in f_1 as a deterministic transformation while the uncertainty is ignored. On the other hand, when δ^2 and ℓ_2^2 are comparable, it means that the (epistemic) uncertainty in f_1 shaped by the supports $\mathbf{y}_1$ is relevant. The effective kernel then represents the covariance in the ensemble of GPs, each of which receives the inputs transformed by one f_1 sampled from the intermediate GP. Thirdly, we shall stress that the appearance of δ^2 is a signature of marginalization over the latent function in deep probabilistic models. Similar square distance also appeared in [30] where the effectively deep kernel was derived from a recursive inner product in the space defined by neural network feature functions.

In the following lemma, we consider the composition where the kernel in outer layer is squared cosine, $k_h(x, y) = (\sigma_2^2/2)\{1 + \cos[(x - y)/\ell_2]\}$, which is a special case of spectrum mixture kernel [38].

Lemma 4. *The covariance in f of the marginal prior with SC kernel used in the exposed GP is given below,*

$$k_{\text{eff}}(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}_p[f_i f_j] = \frac{\sigma_2^2}{2}\left[1 + \cos(m_i - m_j)\exp(-\frac{\delta_{ij}^2}{2\ell_2^2})\right], \tag{14}$$

where δ_{ij}^2 has been defined in the previous lemma.

The form of product of cosine and exponential kernels is similar with the deep spectral mixture kernel [33]. In our case the cosine function has the warped input $m(x_i) - m(x_j)$, but the exponential function has the input $c(x_i, x_i) + c(x_j, x_j) - 2c(x_i, x_j)$ due to the conditional covariance in the intermediate GP.

4.2. Samples from the Marginal Prior

Now we study the samples from the approximate marginal prior with the effective kernel in Equation (12). We shall vary the low fidelity data $\mathbf{X}_1, \mathbf{y}_1$ to see how they affect the inductive bias for the target function. See the illustration in Figure 1. The top row displays the low-fidelity functions $f_1|\mathbf{X}_1, \mathbf{y}_1$, which are obtained by a standard GP regression. Here, the low-fidelity data are noiseless observations of three simple functions (linear in the left, hyper tangent in middle, and sine in right). The conditional covariance and condition mean are then fed into the effective kernel in Equation (12), and so we can sample functions from the prior carrying the effective kernel. The samples are displayed in the second row.

In such cases, it can be seen that $f_1|\mathbf{X}_1, \mathbf{y}_1$ is nearly a deterministic function (top row) given the sufficient amount of noiseless observations in $\{\mathbf{X}_1, \mathbf{y}_1\}$. In fact, the left panel in the second row is equivalent to the samples from a SE kernel as f_1 is the identity function. Moving to the second column, one can see the effect of nonlinear warping generates additional kinks in the target functions. The case on the third column with periodic warping results in periodic patterns to the sampled functions.

Next, we shall see the effect of uncertainty in $f_1|\mathbf{X}_1, \mathbf{y}_1$ (third row) on the sampled functions (bottom row). The increased uncertainty (shown by shadow region) in f_1 generates the weak and high frequency signal in the target function due to the non-stationary δ^2 in Equation (12). We stress that these weak signals are not white noise. The noise in the low fidelity data even reverses the sign of sampled functions, i.e., comparing the second against the bottom rows in the third column. Consequently, the expressivity of the effective kernel gets a contribution from the uncertainty in learning the low fidelity functions.

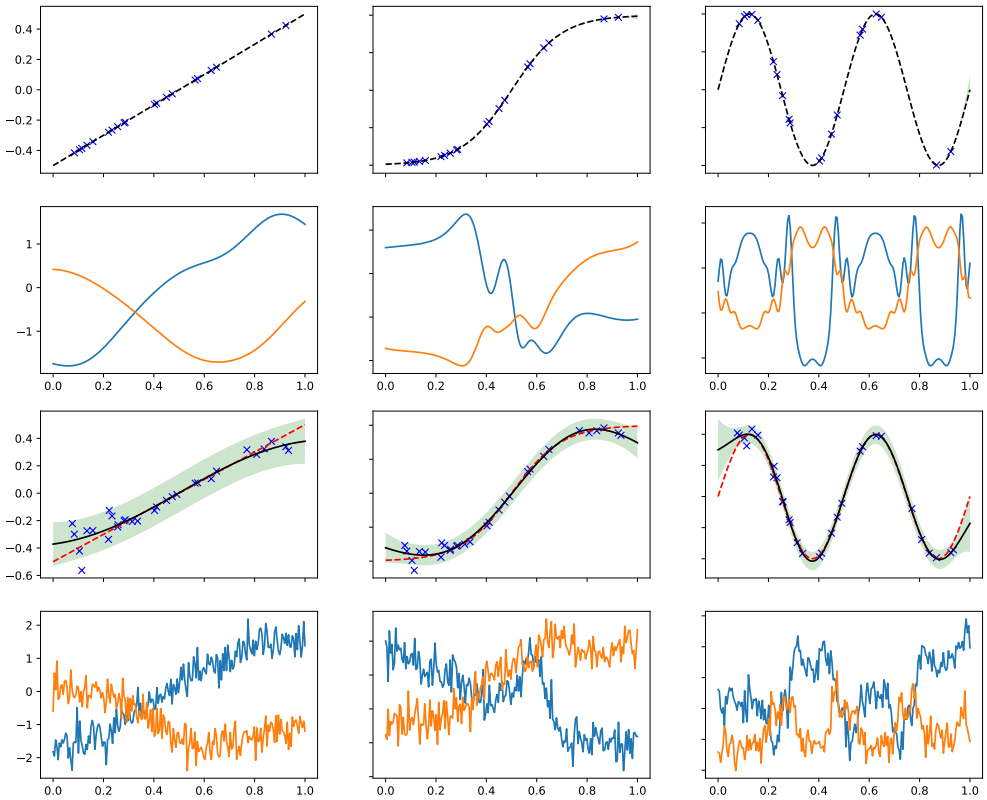


Figure 1. Sampling random functions from the approximate marginal prior $q(\mathbf{f})$ which carries the effective kernel in Equation (12). The low fidelity data $\mathbf{X}_1, \mathbf{y}_1$, marked by the cross symbols, and the low fidelity function $f_1|\mathbf{X}_1, \mathbf{y}_1$ and the uncertainty are shown in the top (noiseless) and the third (noisy) rows. Top row: the uncertainty in $\mathbf{X}_1, \mathbf{y}_1$ is negligible so f_1 is nearly a deterministic function, so the effective kernels are basically kernels with warped input. The corresponding samples from q are shown in the second row. Third row: the noise in $\mathbf{X}_1, \mathbf{y}_1$ generates the samples in bottom row which carry additional high-frequency signals due to the non-stationary δ^2 in Equation (12).

5. Method

Since we approximate the marginal prior for the conditional DGP with a GP, the corresponding approximate marginal likelihood should be the objective for jointly learning the hyperparameters including those in the exposed GP and the intermediate GPs. From the analytical expression for the effective kernel, e.g., Equation (12), the gradient components include the explicit derivatives $\partial K_{\text{eff}}/\partial\sigma_2$ and $\partial K_{\text{eff}}/\partial\ell_2$ as well as those implicit derivatives $\partial K_{\text{eff}}/\partial\sigma_1$ and $\partial K_{\text{eff}}/\partial\ell_1$ which can be computed via chain rule.

With the data consisting of observations of different fidelity, an alternative method can learn the hyperparameters associated with each layer of the hierarchy sequentially. See Algorithm 1 for details. The low fidelity data are fed into the first GP regression model for inferring f_1 and the hyperparameters ℓ_1 and σ_1. The conditional mean and conditional covariance in $f_1|\mathbf{X}_1, \mathbf{y}_1$ are then sent to the effective kernel. The second GP using the effective kernel is to infer the high fidelity function f with the marginal likelihood for the high fidelity data being the objective. Optimization of the second model results in the hyperparameters ℓ_2 and σ_2 in the second layer. Learning in the three-layer hierarchy can be generalized from the two-layer hierarchy. In the Appendix, a comparison of regression using the two methods is shown.

Algorithm 1 A learning algorithm for conditional DGP multi-fidelity regression

Input: two sources of data, low-fidelity data $(\mathbf{X}_1, \mathbf{y}_1)$ and high-fidelity data $(\mathbf{X}, \mathbf{y})$, kernel k_1 for function f_1, and the test input $\mathbf{x}_*$.
1. k_1 = Kernel(var=σ_1^2, lengthscale=ℓ_1) {Initialize the kernel for inferring g}
2. model$_1$ = Regression(kernel=k_1, data=$\mathbf{X}_1$ and $\mathbf{y}_1$) {Initialize regression model for f_1}
3. model$_1$.optimize()
4. $\mathbf{m}$, $\mathbf{C}$ = model$_1$.predict(input = $\mathbf{X}, \mathbf{x}_*$, full-cov=true) {Output pred. mean and post cov. of f_1}
5. k_{eff} = EffectiveKernel(var=σ_2^2, lengthscale=ℓ_2, $\mathbf{m}$, $\mathbf{C}$) {Initialize the effective kernel in Equation (12) for SE[ ] and Equation (14) for SC[ ].}
6. model$_2$ = Regression(kernel=k_{eff}, data = $\mathbf{X}, \mathbf{y}$) {Initialize regression model for f}
7. model$_2$.optimize()
8. μ_*,σ_*^2 = model$_2$.predict(input=$\mathbf{x}_*$)

Output: Optimal hyper-parameters $\{\sigma_{1,2}^2, \ell_{1,2}\}$ and predictive mean μ_* and variance σ_* at $\mathbf{x}_*$.

6. Results

In this section, we shall present the results of multi-fidelity regression given low fidelity data $\mathbf{X}_1, \mathbf{y}_1$ and high fidelity $\mathbf{X}, \mathbf{y}$ and use the 2-layer conditional DGP model. The cases where there are three levels of fidelity can be generalized with the 3-layer counterpart. The toy demonstrations in Section 6.1 focus on data sets in which the target function is a composite, $f(x) = f_2(f_1(x))$. The low fidelity data are observations of $f_1(x)$ while the high fidelity are those of $f(x)$. The aspect of compositional freedom is discussed in Section 6.2, and the same target function shall be inferred with the same high fidelity data but the low fidelity data now result from a variety of functions. In Section 6.3, we switch to the case where the low fidelity data are also observations of the target function f but with large noise. In Section 6.4, we compare our model with the work in [3] on the data set with high dimensional inputs.

6.1. Synthetic Two-Fidelity Function Regression

The first example in Figure 2 consists of 10 random observations of the target function $f(x) = (x - \sqrt{2})f_1^2(x)$ (red dashed line) along with 30 observations of the low fidelity function $f_1(x) = \sin 8\pi x$ (not shown). The 30 observations of f_1 with a period 0.25 in the range of $[0, 1]$ is more than sufficient to reconstruct f_1 with high confidence. In contrast, the 10 observations of f alone (shown in red dots) are difficult to reconstruct f if a GP with SE kernel is used. The above figures demonstrate the results from a set of multi-source nonparametric regression methods which incorporate the learning of f_1 into the target regression of f. Our result, the SE[SE] [panel (f)] kernel, and NARGP [panel (c)] successfully capture the periodic pattern inherited from the low fidelity function f_1, but the target function is fully covered in the confidence region in our prediction only. On the other hand, in the input space away from the target observations, AR1 [panel (a)] and MF-DGP [panel (e)] manages to only capture part of the oscillation. Predictions in LCM

[panel(b)] and DEEP-MF [panel (d)] are reasonable near the target observations but fail to capture the oscillation away from these observations.

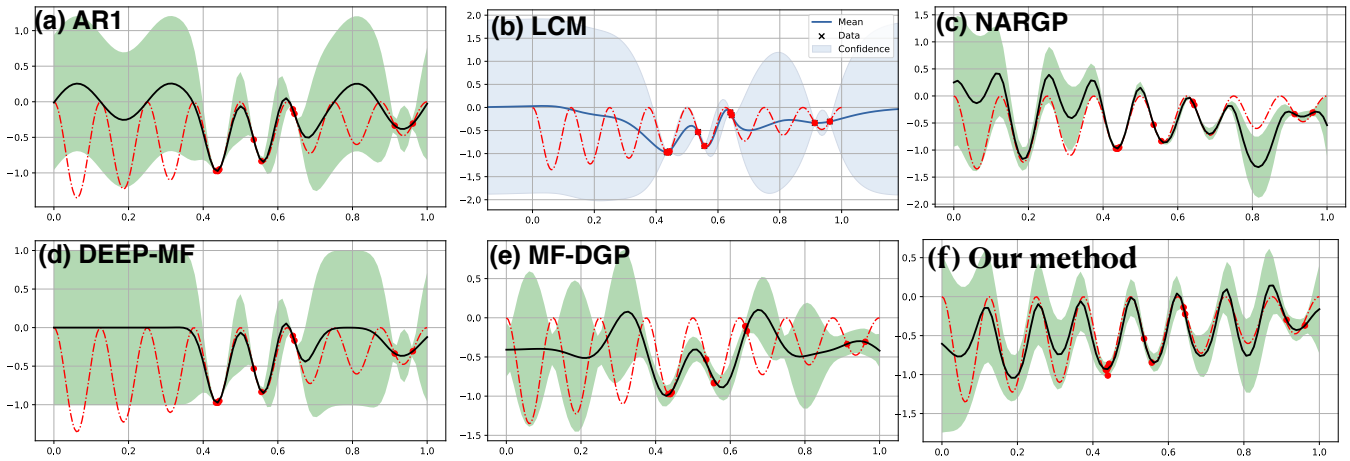


Figure 2. Multi-fidelity regression with 30 observations (not shown) of low fidelity $f_1(x) = \sin 8\pi x$ and 10 observations (red dots) from the target function, $f(x) = (x - \sqrt{2})f_1^2(x)$ (shown in red dashed line). Only the target prediction (solid dark) and associated uncertainty (shaded) are shown. Top row: (**a**) AR1, (**b**) LCM, (**c**) NARGP. Bottom row: (**d**) DEEP-MF, (**e**) MF-DGP, (**f**) our model with SE[SE] kernel.

Figure 3 demonstrates another example of multi-fidelity regression on the nonlinear composite function. The low fidelity function is also periodic, $f_1 = \cos 15x$, and the target is exponential function, $f = x \exp[f_1(2x - 2)] - 1$. The 15 observations of f (red dashed line) are marked by the red dots. The exponential nature in the mapping $f_1 \mapsto f$ might make the reconstruction more challenging than the previous case, which may lead to less satisfying results from LCM [panel (b)]. NARGP [panel (c)] and MF-DGP [panel (e)] have similar predictions which mismatch some of the observations, but the target function is mostly covered by the uncertainty estimation. Our model with SE[SE] kernel [panel (f)], on the other hand, has predictions consistent with all the target observations, and the target function is fully covered by the uncertainty region. Qualitatively similar results are also obtained from AR1 [panel (a)] and DEEP-MF [panel (d)].

6.2. Compositional Freedom and Varying Low-Fidelity Data

Given the good learning results in the previous subsection, one may wonder the effects of having a different low fidelity data set on inferring the high fidelity function. Here, we consider the same high fidelity data from the target function in Figure 2, but the low fidelity data are observations of $f_1(x) = x$, $f_1(x) = \tanh x$, $f_1(x) = \sin 4\pi x$, and $f_1(x) = \sin 8\pi x$. Figure 4 displays the results. Plots in the top rows represent $f_1|\mathbf{X}_1, \mathbf{y}_1$, while the bottom rows show the inferred target function given the high fidelity data (red dots). It can be seen in the left most column in panel (a) that the linear f_1 is not a probable low fidelity function as the true target function (red dashed line) in the bottom is outside the predictive confidence. Similarly in the second plot in (a), f_1 being a hyper tangent function is not probable to account for the true target function. In the end, f_1 being a periodic function is more likely to account for the true target function than the first two cases, but the right most plot with $f_1(x) = \sin 8\pi x$ leads to the predictive mean very close to the true target function.

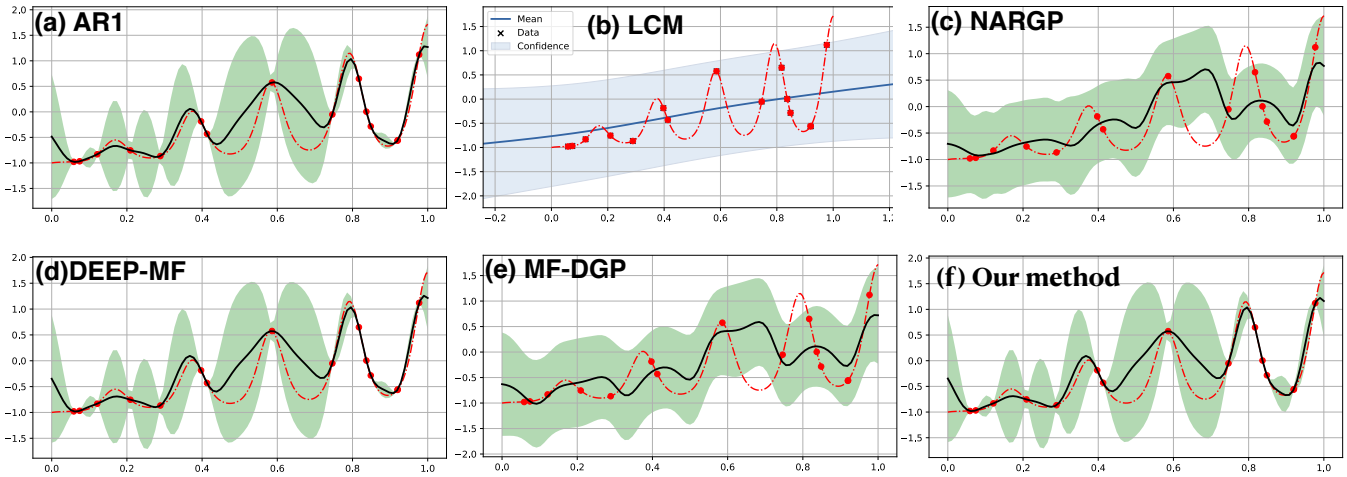
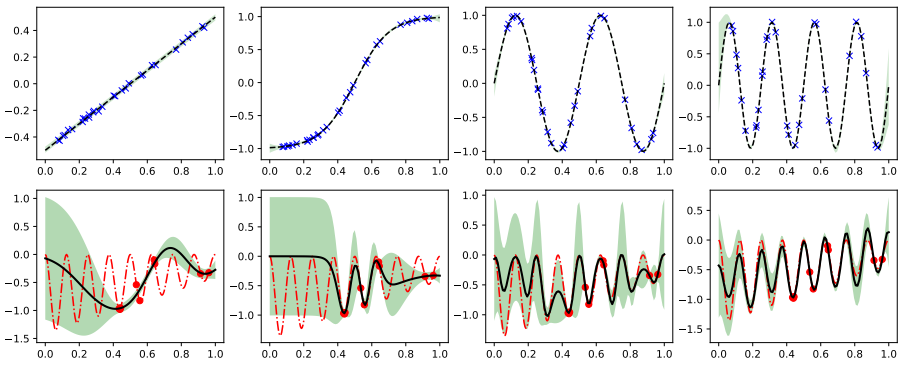
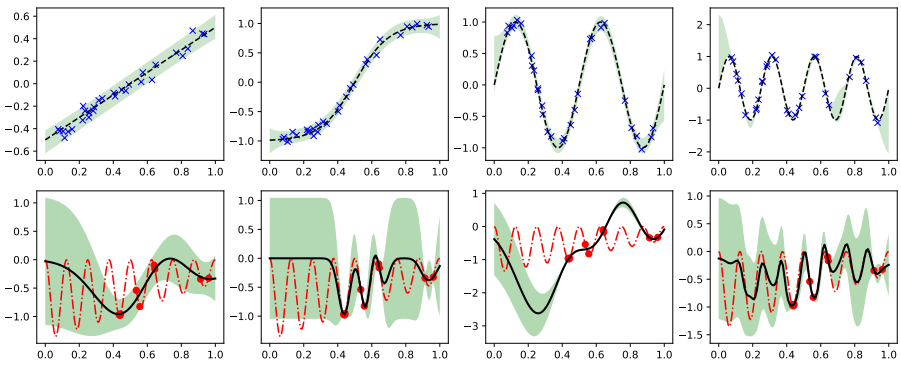


Figure 3. Multi-fidelity regression on the low-level true function, $h(x) = \cos 15x$, with 30 observations and high-level one, $f(x) = x \exp[h(2x - 0.2)] - 1$, with 15 observations. Top row: (**a**) AR1, (**b**) LCM, and (**c**) NARGP. Bottom row: (**d**) DEEP-MF, (**e**) MF-DGP, and (**f**) Our method with SE[SE] kernel.



(**a**) Inference with noiseless low-fidelity data from 4 functions



(**b**) Inference with noisy low-fidelity data from 4 functions

Figure 4. Demonstration of compositional freedom and effects of uncertainty in low fidelity function f_1 on the target function inference. Given the same high fidelity observations of target function, four different sets of observations of $f_1(x) = x$, $f_1(x) = \tanh x$, $f_1(x) = \sin 4\pi x$, and $f_1(x) = \sin 8\pi x$ are employed as low fidelity data in inferring the target function. In panel (**a**), the low fidelity data are noiseless observations of the four functions. The true target function is partially outside the model confidence for the first two cases. In panel (**b**), the low fidelity data are noisy observations of the same four functions. Now the first three cases result in the inferred function outside the model confidence. The effect of uncertainty in low fidelity is most dramatic when comparing the third subplots in (**a**,**b**).

Next, the low fidelity data become the noisy observations of the same four functions. As shown in panel (b), the increased variance in $f_1|\mathbf{X}_1, \mathbf{y}_1$ also results in raising the variance in f, especially comparing the first two cases in (a) against those in (b). A dramatic difference can be found in comparing the third plot in (a) against that in (b). In (b), the presence of noise in the low fidelity data slightly raises the uncertainty in f_1, but the ensuing inference in f generates the prediction which fails to contain most of the true target function within its model confidence. Thus, the likelihood that $f_1(x) = \sin 4\pi x$ is the probable low fidelity function is greatly reduced by the noise in the observation. Lastly, the noise in observing $f_1(x) = \sin 8\pi x$ as the low fidelity data does not significantly change the inferred target function.

Therefore, the inductive bias associated with the target function is indeed controllable by the intermediate function distribution conditioned on the lower fidelity data. The observation motivates the DGP learning from the common single-fidelity regression data with the intermediate GPs conditioned on some optimizable hyperdata [39]. These hyperdata constrain the space of intermediate function, and the uncertainty therein contribute to the expressiveness of the model.

6.3. Denoising Regression

Here we continue considering the inference of the same target function in $f(x) = (x - \sqrt{2}) \sin^2 8\pi x$, but now the low fidelity data set becomes the noisy observations of the target function. See Figure 5 for illustration. Now we have 15 observations of f with noise level of 0.001 (red dots) as high fidelity data and 30 observations of the same function with nosie level of 0.1 (dark cross symbol) as the low fidelity data. Next, we follow the same procedure in inferring f_1 with the low fidelity, and then use the conditional mean and covariance in constructing the effective kernel for inferring the target function f with the high fidelity data. Unlike the previous cases, the relation between f and f_1 here is not clear. However, the structure of DGP can be viewed as the intermediate GP emitting infinitely many samples of f_1 into the exposed GP. Qualitatively, one can expect that the actual prediction for f is the average over the GP models with different warping f_1. Consequently, we may expect the variance in predicting f is reduced.

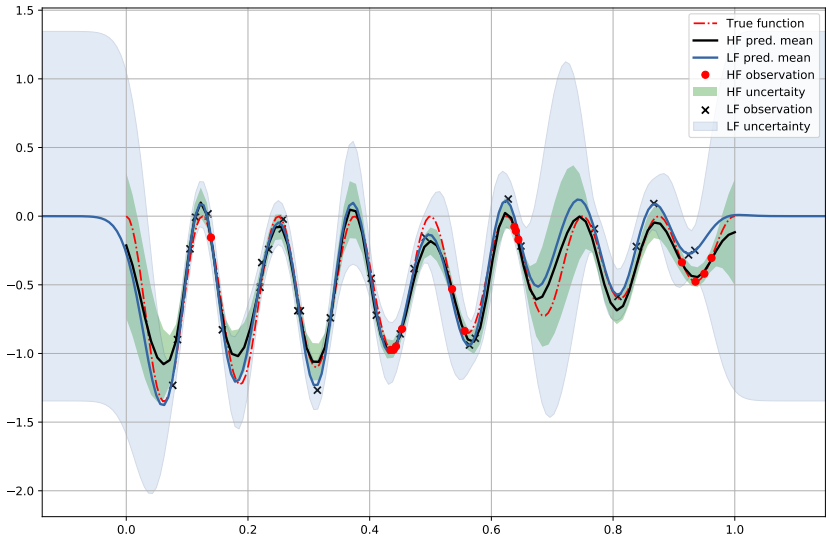


Figure 5. Denoising regression with 30 high-noise and 15 low-noise observations from the target function $y = (x - \sqrt{2}) \sin^2 8\pi x$ (red dashed line). The uncertainty is reduced in the GP learning with the SE[SE] kernels.

Indeed, as shown in Figure 5, the predictive variance using a GP with the low fidelity (high noise) observations only is marked by the light-blue region around the predictive mean (light-blue solid line). When the statistical information in $f_1|\mathbf{X}_1, \mathbf{y}_1$ is transferred

to the effective kernel, the new prediction and model confidence possess much tighter uncertainty (marked by the light-green shaded region) around the improved predictive mean (dark solid line) even in the region away from the low-noise observations.

6.4. Multi-Fidelity Data with High Dimensional Input

The work in [3] along with their public code in emukit [40] assembles a set of multi-fidelity regression data sets in which the input $\mathbf{x}$ is of high dimension. Here we demonstrate the simulation results on these data (see [3] for details). The simulation is performed using the effective kernels with compositions: SE[SE] and SC[SE] for the Borehole (two-fidelity) regression data set, SE[SE[SE]]] and SC[SC[SE]] for Branin (three-fidelity) regression data set. The data are obtained from deploying the modules in [40]. Algorithm 1 is followed to obtain the results here. The performance of generalization is measured in terms of mean negative log likelihood (MNLL). Table 1 displays the results using the same random seed from MF-DGP and our methods. We also include the simulation of standard GP regression with the high fidelity data only. It is seen that the knowledge about the low fidelity function is significant for predicting high-level simulation (comparing with vanilla GP) and that the informative kernels have better performance in these cases.

Table 1. MNLL results of multi-fidelity regression.

	MFDGP	**SE[ ]**	**SC[ ]**	**GP+$\mathcal{D}_f$**
Borehole	−1.87	**2.08**	**−2.08**	0.56
Branin	−2.7	−2.52	**−2.93**	5180

7. Discussion

In this paper, we propose a novel kernel learning which is able to fuse data of low fidelity into a prior for high fidelity function. Our approach addresses two limitations of prior research on GPs: the need to choose or design kernel [8,9] and the lack of explicit dependence on the observations in the prediction (in Student-t process [41] the latter is possible). We resolve limitations associated with reliance on designing kernels, introducing new data-dependent kernels together with effective approximate inference. Our results show that the method is effective, and we prove that our moment-matching approximation retains some multi-scale, multi-frequency, and non-stationary correlations that are characteristic of deep kernels, e.g., [33]. The compositional freedom [18] pertaining to hierarchical learning is also manifested in our approach.

8. Conclusions

Central to the allure of Bayesian methods, including Gaussian Processes, is the ability to calibrate model uncertainty through marginalization over hidden variables. The power and promise of DGP is in allowing rich composition of functions while maintaining the Bayesian character of inference over unobserved functions. Modeling the multi-fidelity data with the hierarchical DGP is able to exploit its expressive power and to consider the effects of uncertainty propagation. Whereas most approaches are based on variational approximations for inference and Monte Carlo sampling in the prediction stage, our approach uses a moment-based approximation in which the marginal prior of DGP is analytically approximated with a GP. For both, the full implications of these approximations are unknown. Continued research is required to understand the full strengths and limitations of each approach.

Author Contributions: Conceptualization, C.-K.L. and P.S.; methodology, C.-K.L.; software, C.-K.L.; validation, C.-K.L. and P.S.; formal analysis, C.-K.L.; investigation, C.-K.L.; resources, C.-K.L.; data curation, C.-K.L.; writing—original draft preparation, C.-K.L.; writing—review and editing, C.-K.L. and P.S.; visualization, C.-K.L.; supervision, P.S.; project administration, C.-K.L.; funding acquisition, P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Air Force Research Laboratory and DARPA under agreement number FA8750-17-2-0146.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to thank the helpful correspondences with the authors of [22].

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DGP	Deep Gaussian Process
SE	Squared Exponential
SC	Squared Cosine

Appendix A

Figure A1 shows the two results of multi-fidelity regressions with the same data. The left panel is obtained with jointly learning the hyperparameters in all layers with the standard gradient descent on the approximate marginal likelihood, while the right panel is from learning the hyperparameters sequentially with the Algorithm 1. It is noted that the right panel yields higher log of marginal likelihood.

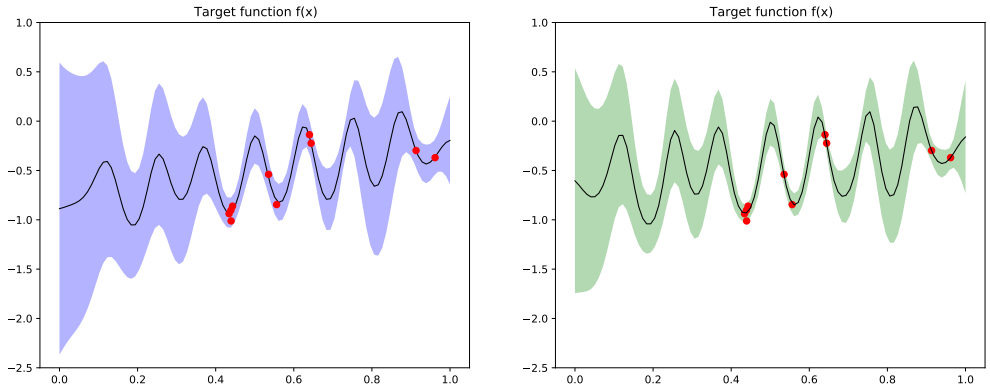


Figure A1. Comparison between the joint learning (**left**) and the sequential learning with Algorithm 1 (**right**). The same 10 training data are shown by the red dots. The joint learning algorithm results in a log marginal likelihood 1.65 while the alternative one 2.64. The hyperparameters are $\{\sigma_{1,2} = (3.3, 1.24), \ell_{1,2} = (0.12, 1.40)\}$ (**left**) and $\{\sigma_{1,2} = (1.22, 1.65), \ell_{1,2} = (0.08, 0.98)\}$ (**right**).

References

1. Kennedy, M.C.; O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **2000**, *87*, 1–13. [CrossRef]
2. Wang, Z.; Xing, W.; Kirby, R.; Zhe, S. Multi-Fidelity High-Order Gaussian Processes for Physical Simulation. In Proceedings of the International Conference on Artificial Intelligence and Statistics, San Diego, CA, USA, 13 April 2021; pp. 847–855.
3. Cutajar, K.; Pullin, M.; Damianou, A.; Lawrence, N.; González, J. Deep Gaussian processes for multi-fidelity modeling. *arXiv* **2019**, arXiv:1903.07320
4. Salakhutdinov, R.; Tenenbaum, J.; Torralba, A. One-shot learning with a hierarchical nonparametric bayesian model. In Proceedings of the ICML Workshop on Unsupervised and Transfer Learning, Bellevue, WA, USA, 2 July 2012; pp. 195–206.
5. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6 August 2017; pp. 1126–1135.

6. Titsias, M.K.; Schwarz, J.; Matthews, A.G.d.G.; Pascanu, R.; Teh, Y.W. Functional Regularisation for Continual Learning with Gaussian Processes. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
7. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Process for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
8. Duvenaud, D.; Lloyd, J.; Grosse, R.; Tenenbaum, J.; Zoubin, G. Structure discovery in nonparametric regression through compositional kernel search. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013; pp. 1166–1174.
9. Sun, S.; Zhang, G.; Wang, C.; Zeng, W.; Li, J.; Grosse, R. Differentiable compositional kernel learning for Gaussian processes. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4828–4837.
10. Damianou, A.; Lawrence, N. Deep Gaussian processes. In Proceedings of the Artificial Intelligence and Statistics, Scottsdale, AZ, USA, 29 April–1 May 2013; pp. 207–215.
11. Snelson, E.; Ghahramani, Z.; Rasmussen, C.E. Warped Gaussian processes. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 13–18 December 2004; pp. 337–344.
12. Lázaro-Gredilla, M. Bayesian warped Gaussian processes. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1619–1627.
13. Salimbeni, H.; Deisenroth, M. Doubly stochastic variational inference for deep Gaussian processes. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
14. Salimbeni, H.; Dutordoir, V.; Hensman, J.; Deisenroth, M.P. Deep Gaussian Processes with Importance-Weighted Variational Inference. *arXiv* **2019**, arXiv:1905.05435
15. Yu, H.; Chen, Y.; Low, B.K.H.; Jaillet, P.; Dai, Z. Implicit Posterior Variational Inference for Deep Gaussian Processes. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 14502–14513.
16. Lu, C.K.; Yang, S.C.H.; Hao, X.; Shafto, P. Interpretable deep Gaussian processes with moments. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Virtual, 26–28 August 2020; pp. 613–623.
17. Havasi, M.; Hernández-Lobato, J.M.; Murillo-Fuentes, J.J. Inference in deep Gaussian processes using stochastic gradient Hamiltonian Monte Carlo. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 7506–7516.
18. Ustyuzhaninov, I.; Kazlauskaite, I.; Kaiser, M.; Bodin, E.; Campbell, N.; Ek, C.H. Compositional uncertainty in deep Gaussian processes. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, Toronto, ON, Canada, 3 August 2020; pp. 480–489.
19. Le Gratiet, L.; Garnier, J. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *Int. J. Uncertain. Quantif.* **2014**, *4*, 365–386. [CrossRef]
20. Raissi, M.; Karniadakis, G. Deep multi-fidelity Gaussian processes. *arXiv* **2016**, arXiv:1604.07484
21. Perdikaris, P.; Raissi, M.; Damianou, A.; Lawrence, N.; Karniadakis, G.E. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proc. R. Soc. Math. Phys. Eng. Sci.* **2017**, *473*, 20160751. [CrossRef] [PubMed]
22. Requeima, J.; Tebbutt, W.; Bruinsma, W.; Turner, R.E. The Gaussian process autoregressive regression model (gpar). In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, Naha, Okinawa, Japan, 16–18 April 2019; pp. 1860–1869.
23. Alvarez, M.A.; Rosasco, L.; Lawrence, N.D. Kernels for vector-valued functions: A review. *arXiv* **2011**, arXiv:1106.6251.
24. Parra, G.; Tobar, F. Spectral mixture kernels for multi-output Gaussian processes. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6684–6693.
25. Bruinsma, W.; Perim, E.; Tebbutt, W.; Hosking, S.; Solin, A.; Turner, R. Scalable Exact Inference in Multi-Output Gaussian Processes. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1190–1201.
26. Kaiser, M.; Otte, C.; Runkler, T.; Ek, C.H. Bayesian alignments of warped multi-output Gaussian processes. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, Canada, 3–8 December 2018; pp. 6995–7004.
27. Kazlauskaite, I.; Ek, C.H.; Campbell, N. Gaussian Process Latent Variable Alignment Learning. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, Naha, Okinawa, Japan, 16–18 April 2019; pp. 748–757.
28. Williams, C.K. Computing with infinite networks. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 2–4 December1997; pp. 295–301.
29. Cho, Y.; Saul, L.K. Kernel methods for deep learning. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–8 December 2009; pp.342–350.
30. Duvenaud, D.; Rippel, O.; Adams, R.; Ghahramani, Z. Avoiding pathologies in very deep networks. In Proceedings of the Artificial Intelligence and Statistics, Reykjavik, Iceland, 22–25 April 2014; pp. 202–210.
31. Dunlop, M.M.; Girolami, M.A.; Stuart, A.M.; Teckentrup, A.L. How deep are deep Gaussian processes? *J. Mach. Learn. Res.* **2018**, *19*, 2100–2145.
32. Shen, Z.; Heinonen, M.; Kaski, S. Learning spectrograms with convolutional spectral kernels. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Virtual, 26–28 August 2020; pp. 3826–3836.
33. Wilson, A.G.; Hu, Z.; Salakhutdinov, R.; Xing, E.P. Deep kernel learning. In Proceedings of the Artificial Intelligence and Statistics, Cadiz, Spain, 9–11 May 2016; pp. 370–378.

34. Daniely, A.; Frostig, R.; Singer, Y. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. In Proceedings of the NIPS: Centre Convencions Internacional Barcelona, Barcelona, Spain, 5–10 December 2016.
35. Mairal, J.; Koniusz, P.; Harchaoui, Z.; Schmid, C. Convolutional kernel networks. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 8–13 December 2014; pp. 2627–2635.
36. Van der Wilk, M.; Rasmussen, C.E.; Hensman, J. Convolutional Gaussian processes. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 2849–2858.
37. GPy. GPy: A Gaussian Process Framework in Python. 2012. Available online: http://github.com/SheffieldML/GPy (accessed on 1 October 2021).
38. Wilson, A.; Adams, R. Gaussian process kernels for pattern discovery and extrapolation. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1067–1075.
39. Lu, C.K.; Shafto, P. Conditional Deep Gaussian Processes: Empirical Bayes Hyperdata Learning. *Entropy* **2021**, *23*, 1387. [CrossRef]
40. Paleyes, A.; Pullin, M.; Mahsereci, M.; Lawrence, N.; González, J. Emulation of physical processes with Emukit. In Proceedings of the Second Workshop on Machine Learning and the Physical Sciences, Vancouver, BC, Canada, 8–14 December 2019.
41. Shah, A.; Wilson, A.; Ghahramani, Z. Student-t processes as alternatives to Gaussian processes. In Proceedings of the Artificial Intelligence and Statistics, Reykjavic, Iceland, 22–25 April 2014; pp. 877–885.