Interpretable Deep Gaussian Processes with Moments

Chi-Ken Lu **Scott Cheng-Hsin Yang** **Xiaoran Hao** **Patrick Shafto**
Department of Mathematics & Computer Science, Rutgers University Newark, NJ 07102

Abstract

Deep Gaussian Processes (DGPs) combine the expressiveness of Deep Neural Networks (DNNs) with quantified uncertainty of Gaussian Processes (GPs). Expressive power and intractable inference both result from the non-Gaussian distribution over composition functions. We propose interpretable DGP based on approximating DGP as a GP by calculating the exact moments, which additionally identify the heavy-tailed nature of some DGP distributions. Consequently, our approach admits interpretation as both NNs with specified activation functions and as a variational approximation to DGP. We identify the expressivity parameter of DGP and find non-local and non-stationary correlation from DGP composition. We provide general recipes for deriving the effective kernels for DGP of two, three, or infinitely many layers, composed of homogeneous or heterogeneous kernels. Results illustrate the expressiveness of our effective kernels through samples from the prior and inference on simulated and real data and demonstrate advantages of interpretability by analysis of analytic forms, and draw relations and equivalences across kernels.

1 INTRODUCTION

The success of deep learning models is generally perceived as stemming from greater expressivity which results in powerful generalization (Goodfellow et al., 2016). However, deep learning models are viewed as black boxes because their complexity arises from the enormous number of parameters and possible choices for different structures and activation units. Understanding these models remains an open and challenging problem (Zhang et al., 2016; Belkin et al., 2018; Mei et al., 2018; Advani et al., 2013; Liao and Couillet, 2018). Williams (1997) demonstrated that the characteristics of single-layer neural networks (NNs) can be understood from its effective kernel showing non-stationary and non-local correlation. It is thus appealing to create more interpretable methods through the correspondence between deep learning and kernel-based methods which have the advantage of an explicit mathematical formalization (Schölkopf et al., 1998; Rahimi and Recht, 2008; Cho and Saul, 2009; Mairal et al., 2014; Wilson et al., 2014, 2016; Van der Wilk et al., 2017; Sun et al., 2018).

Quantifying uncertainty of inferences is also a critical issue for deep learning models (Wang and Manning, 2013; Gal and Ghahramani, 2016). Gaussian processes (GPs) (Rasmussen and Williams, 2006), the infinitely wide limit of single-layer neural network with random weight parameters (Neal, 2012), are attractive alternative models capable of quantifying uncertainty through Bayesian inference. Moreover, GPs can be composed into DGPs (Damianou and Lawrence, 2013), with inference via variational approximations (Titsias and Lawrence, 2010; Bui et al., 2016; Cutajar et al., 2017; Salimbeni et al., 2019; Salimbeni and Deisenroth, 2017) and sampling-based method (Havasi et al., 2018), to achieve expressiveness that is comparable to DNNs. However, as for DNNs, with this expressivity comes difficulty in interpretability. It is desirable to leverage interpretability of explicit mathematical formalizations associated with kernel-based methods, while preserving expressivity and uncertainty quantification.

A brief review on GP and intractability of DGP is given in Sec. 2. The multi-modal posterior distribution is shown to result from latent function symmetry in the nested probabilistic model. We introduce calculation of the second and fourth moments in Sec. 3 and 4, respectively. With the second moment, we approximate DGP as GP in Sec. 5 where the validity and the heavy-tailed nature of DGP are discussed by

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

examining the fourth moments. This can be regarded as a new variational approximation to Bayesian inference in DGPs that effectively represents any depth as single layer. In Sec. 6, with these analytic effective covariance functions, we interpret DGP as generation of non-local, non-stationary, and multi-scale correlation through depth of homogeneous or heterogeneous function composition, which leads to the expressiveness of DGP. Sec. 7 demonstrates the expressivness of DGP with optimized marginal likelihood for three-layer DGP against the expressivity parameter defined in DNN model.

1.1 Related Works

DGP (Damianou and Lawrence, 2013) and the earlier Warped GP (Snelson et al., 2004; Lázaro-Gredilla, 2012) were proposed to generalize the idea of GP to serve as a prior distribution over composition function in Bayesian learning framework. The need for approximate inference presents challenges, however. Using variational approach (Titsias, 2009; Titsias and Lawrence, 2010; Salimbeni and Deisenroth, 2017) and expectation propagation (Minka, 2001; Bui et al., 2016) DGPs have demonstrated superior expressivity over GPs. Our contribution is the development of Bayesian DGPs with analytic forms for compositional homogeneous and heterogeneous kernels as well as identification of heavy-tailed distribution associated with some DGP models. Because activation functions of DNNs correspond with kernels in DGPs, our results shall also help interpret the heterogeneous networks using different activation units.

Williams (1997) was first to derive analytic kernels of one-layer neural networks, using sigmoidal and Gaussian activation functions. Later, Cho and Saul (2009) extended analytic results to polynomial activation functions and further made extension to deep models. Duvenaud et al. (2014) studied pathologies in DGPs, primarily focusing on squared exponential kernels and sampling functions from the prior; Dunlop et al. (2018) made connection between the effective depth and ergodicity. Daniely et al. (2016) extended these results for homogeneous deep networks by proposing the dual activation for the Hermite, step, and exponential functions. Poole et al. (2016) studied the DNN models and found expressivity parameter indicating chaotic phase and ordered phase from the correlation map. Our approach differs by focusing on DGPs for Bayesian inference, and allowing heterogeneous kernels.

Lee et al. (2017) extended the correspondence between NNs and GPs, established by Neal (2012), to DNNs and GPs, via the central limit theorem, focusing on homogeneous kernels.

2 GP and DGP Overview

A Gaussian process (GP) is a prior distribution $p(f)$ over continuous function f of which any subset of points $\{f_i\}$ follows the joint multivariate Gaussian distribution specified by mean function $\mu(\mathbf{x}_i)$ and covariance matrix $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. The properties of function, such as smoothness, are encoded in the covariance function $k(\mathbf{x}_i, \mathbf{x}_j) : R^D \times R^D \to R$. For example, functions that are sampled from the squared exponential (SE) kernel are infinitely differentiable, while the non-stationary neural network kernel (Williams, 1997) may generate functions which resemble step functions. Under the Bayesian framework, the posterior distribution $p(f|\mathcal{D}, \theta)$ is used to make predictions. The hyperparameters, denoted by θ, are determined by the maximizing the marginal distribution $p(\mathcal{D}|\theta)$. The benefit of Bayesian learning of θ is that over fitting is naturally avoided by the two competing terms, data-fit and complexity penalty, in the marginal likelihood.

Deep GP serves as a prior distribution over both *homogeneous* and *heterogeneous* compositions of functions, where the resultant function f from L layers of warping is given by

$$\mathbf{f} \sim \mathcal{GP}(\mu^L, k^{(L)}(\mathbf{h}^{L-1}, \mathbf{h}^{L-1})) \tag{1}$$

with the hidden functions $\mathbf{h}^{1:L}$

$$\mathbf{h}^m \sim \mathcal{GP}(\mu^m, k^{(m)}(\mathbf{h}^{m-1}, \mathbf{h}^{m-1}))\,. \tag{2}$$

The zeroth layer variable $\mathbf{h}^0$ represents the input $\mathbf{x} \in R^D$. The prior mean functions $\mu^{1:L}$ and covariance function $k^{(1:L)}$ in the latent layers are predetermined. Homogeneous deep GPs compose functions with the same kernel, $k^{(i)} = k^{(j)}$, $\forall i, j$, whereas heterogeneous deep GPs compose functions with different kernels $\exists i, j$ s.t. $k^{(i)} \neq k^{(j)}$. For a finite set of $\{f_i\}_{i=1}^N$ associated with the input $\{\mathbf{x}_i\}_{i=1}^N$, the joint distribution can be expressed as

$$p(\mathbf{f}, \mathbf{h}^{1:L}|\mathbf{X}) = p(\mathbf{f}|\mathbf{h}^L)p(\mathbf{h}^L|\mathbf{h}^{L-1}) \dots p(\mathbf{h}^1|\mathbf{X})\,, \tag{3}$$

in which each distribution in the product on right hand side is Gaussian. The DGP is defined as the marginal distribution,

$$p(\mathbf{f}|\mathbf{X}) = \int p(\mathbf{f}, \mathbf{h}^{1:L}|\mathbf{X}) d\mathbf{h}^{1:L} \tag{4}$$

over the composition function. However, the above marginal distribution is not Gaussian as the latent variables $\mathbf{h}^{1:L}$ appear in the inverse of covariance matrix.

Although the marginalization over the latent variable $\mathbf{h}$'s in Eq. (4) is intractable, we have the following observations regarding the joint and posterior distributions of DGP, which are consistent with the sampling-based DGP in (Havasi et al., 2018).

Lemma 1 *The joint distribution in Eq. (3) is invariant under the transformation* $\mathbf{h}^i \to -\mathbf{h}^i$, $1 \le i \le L$, *if the prior mean* $\mu^i = 0$.

The proof proceeds by noting that the latent variable $\mathbf{h}^i$ appears only in the two distributions, $p(\mathbf{h}^i|\mathbf{h}^{i-1})$ and $p(\mathbf{h}^{i+1}|\mathbf{h}^i)$ in Eq. (3). In the former distribution, the quadratic $\mathbf{h}^i$ in the exponent of Gaussian distribution permits the sign changing. In the latter distribution, the covariance function $k^{(i+1)}$ is also invariant because the distance $|h_a^i - h_b^i|$, $1 \le a, b \le N$, and the distance to origin, $|h_a^i|$ is intact by the sign change. Therefore the joint distribution is left invariant.

Remark 1 *If a set of function values associated with each latent layer,* $\{\overline{\mathbf{h}}^1, \overline{\mathbf{h}}^2, \cdots, \overline{\mathbf{h}}^L\}$, *is found to correspond to the maximum of the posterior distribution* $p(\mathbf{h}^{1:L}|\mathcal{D}) \propto \int p(\mathbf{f}, \mathbf{h}^{1:L}) p(\mathcal{D}|\mathbf{f}) d\mathbf{f}$, *with zero prior mean in each latent layer, then the rest of* 2^L *sets,* $\{\pm\overline{\mathbf{h}}^1, \cdots, \pm\overline{\mathbf{h}}^L\}$, *are also maximum of posterior distribution. Therefore, such DGP posterior distribution is multi-modal.*

In the following, we shall consider the family of few-layer DGP. Since DGP allows heterogeneous composition, it is convenient to label DGP by the covariance functions. For example, SE[SC] stands for the DGP that $L = 2$ and the covariance functions $k^{(2)}$ and $k^{(1)}$, respectively, are squared exponential and squared cosine functions in the output and input layers. The key finding in this paper is, despite the fact that marginalization in DGP inference is intractable, that the calculation of moments is tractable and analytic for some families of DGP.

To illustrate, we first consider the two-layer DGP, $L = 2$, and the prior mean at output layer is zero, $\mu^{(2)} = 0$. The first moment of the marginal distribution in Eq. (4) is obtained as $\mathbb{E}[f_i] = \int f_i p(\mathbf{f}, \mathbf{h}|\mathrm{X}) d\mathbf{f} d\mathbf{h}$, and it follows from the zero-mean assumption that this first moment is zero. Likewise, the third moment as well as other odd order moments is zero. Moreover, the second moment is $\mathbb{E}[f_i f_j]$, which can be shown to be the expected value of output layer covariance function with respect to the mutivariate normal distribution associated with the input layer GP,

$$\mathbb{E}[f_i f_j] = \int d\mathbf{h}\; k^{(2)}(h_i, h_j) \mathcal{N}(\mathbf{h}|\mu^{(1)}, k^{(1)}) \,. \tag{5}$$

As will be shown later, the second moment is useful when the DGP is approximated by GP where the effective covariance function is just the second moment. Additionally, the fourth moment is

$$\mathbb{E}[f_i f_j f_m f_l] = \sum \int d\mathbf{h}\; k_{ij}^{(2)} k_{ml}^{(2)} \mathcal{N}(\mathbf{h}|\mu^{(1)}, k^{(1)}) \tag{6}$$

where, according to the Isserlis' theorem (Isserlis, 1918), the sum is over all distinct ways of partitioning i, j, m, l into pairs ij and ml. $k_{ij}^{(2)} = k^{(2)}(h_i, h_j)$ is abbreviated for convenience. Although the marginal distribution for DGP in Eq. (4) is intractable, the fourth moments are able to reveal additional information about the true distribution and shed light on the validity of approximation scheme in this paper.

3 Second Moment of DGP

We first consider second moments of the DGP distributions, SE[·] and SC[·], where the covariance function $k^{(1)}$ in the input layer is arbitrary but $k^{(2)}$ in the output layer is in the exponential family with zero prior mean. The analytic calculation of second moment is possible due to the marginal property of Gaussian which leads to the bivariate distribution,

$$\mathcal{N}(h_i, h_j|\mathbf{v}, \mathbb{K}) = \int d\mathbf{h}_{\setminus ij}\; \mathcal{N}(\mathbf{h}|\mu, k), \tag{7}$$

where $d\mathbf{h}_{\setminus ij}$ denotes all but the integration with respect to h_i and h_j. $\mathbf{v}$ and $\mathbb{K}$ denotes the mean and covariance matrix in the remaining block related to h_i and h_j. Explicitly,

$$\mathbb{K} = \begin{pmatrix} k_{ii} & k_{ij} \\ k_{ij} & k_{jj} \end{pmatrix} , \tag{8}$$

and the matrix elements take the value $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. As such, following the expression in Eq. (5), the second moments of SE[·] is obtained as

$$\sigma^2\, \mathbb{E}_{h_i h_j}\left[e^{-\frac{(h_i - h_j)^2}{2\ell^2}} \right] , \tag{9}$$

with the pair of random variables h's sampled from the bivariate normal distribution in Eq. (7). Similarly, the second moment of SC[·] is given by

$$\frac{\sigma^2}{2}\, \mathbb{E}_{h_i h_j}\left(1 + e^{i\frac{h_i - h_j}{\ell}} \right) . \tag{10}$$

The signal magnitude σ_2 and length scale ℓ_2 are the hyperparameters in the output layer GP.

3.1 Squared Exponential DGP: SE[·] & NuN[·]

The following lemma is useful when calculating the expected value of random exponential quadratic function.

Lemma 2 *The expected value of the random exponential quadratic form* $Q(\mathbf{x}) = \exp(-\frac{\mathbf{x}^t J \mathbf{x}}{2})$ *with respect to multivariate normal distribution* $\mathcal{N}(\mathbf{x}|\mathbf{v}, \mathbb{K})$ *is given by*

$$\mathbb{E}[Q(\mathbf{x})] = \frac{\exp(-\frac{1}{2}\mathbf{v}^t \mathbb{A} \mathbf{v})}{\sqrt{|I + \mathbb{K}J|}} , \tag{11}$$

where J denotes the symmetric matrix of dimension $|\mathbf{x}|$ and the matrix $\mathbb{A} = \mathbb{K}^{-1}[I - (I + \mathbb{K}J)^{-1}]$.

For SE[·] with prior mean being zero in input and output layers, the calculation Eq. (9) for covariance function in above representation with $J = \frac{1}{\ell^2}\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$ leads to the second moment,

$$\sigma^2 \left(1 + \frac{k_{ii} + k_{jj} - 2k_{ij}}{\ell^2}\right)^{-\frac{1}{2}} . \tag{12}$$

Moreover, we may consider the covariance function, $k(x, y) = \sigma^2 \exp[-(\alpha x^2 - 2\beta xy + \alpha y^2)/2]$ with $\alpha > \beta > 0$, as a more general Gaussian kernel in the output layer. This kernel in fact corresponds to the neural network with Gaussian activation function [Eq. (13) in (Williams, 1997)]. For convenience, we label such DGP as NuN[·]. It should be noted that NuN[·] is identical to SE[·] as $\alpha = \beta = 1/\ell^2$. Writing the matrix $J = \left(\begin{smallmatrix} \alpha & -\beta \\ -\beta & \alpha \end{smallmatrix}\right)$, the resultant second moment reads,

$$\sigma^2 \left[1 + \alpha(k_{ii} + k_{jj}) - 2\beta k_{ij} + (\alpha^2 - \beta^2)|\mathbb{K}|\right]^{-\frac{1}{2}} . \tag{13}$$

Note that the determinant term $|\mathbb{K}| = k_{ii}k_{jj} - k_{ij}^2$ does not appear in the stationary SE[·] case. In addition, the nonstationariness, i.e. the covariance function value depending on the input $\mathbf{x}_i$ and $\mathbf{x}_j$ as the distance between pair vanishes, is intact as long as $\alpha \neq \beta$. Otherwise, Eq. (13) shall be constant as $k_{ij} \to k_{ii}$ and k_{jj}.

3.2 Squared Cosine DGP: SC[·]

To capture periodic patterns in data, we may be interested in squared cosine (SC) kernel, $k(x, y) = \sigma^2 \cos^2 \frac{x-y}{2\ell}$. The expectation of general exponential linear function is given by the following lemma.

Lemma 3 *The expected value of the random exponential linear function $L(\mathbf{x}) = \exp(\mathbf{u} \cdot \mathbf{x})$ with respect to the bivariate nomral distribution $\mathcal{N}(\mathbf{x}|\mathbf{v}, \mathbb{K})$ is given by*

$$\mathbb{E}[L(\mathbf{x})] = \exp(\mathbf{u} \cdot \mathbf{v} + \frac{\mathbf{u}^t \mathbb{K} \mathbf{u}}{2}) . \tag{14}$$

Using $\mathbf{u} = \frac{1}{\ell}\left(\begin{smallmatrix} i \\ -i \end{smallmatrix}\right)$, the second moment of SC[·] with zero prior mean in input layer, Eq. (10) becomes

$$\frac{\sigma^2}{2}\left[1 + \exp\frac{2k_{ij} - k_{ii} - k_{jj}}{2\ell^2}\right] . \tag{15}$$

3.3 Three-layer DGP: SE[SE[SE]]

For DGP with $L > 2$, the present approach for obtaining the second moment is still valid but exact closed form seems unlikely. To be able to shed light on the effect of depth on the second moment, we consider the special case for three-layer DGP, SE[SE[SE]], where SE kernel is employed in all three layers and all prior means are set to zero. Following from the second moment of SE[SE], Eq. (12), we can show that the second moment for three-layer SE[SE[·]] is,

$$\mathbb{E}_{h_i h_j}\left\{\sigma_3^2[1 + 2\frac{\sigma_2^2}{\ell_3^2}(1 - c)]^{-\frac{1}{2}}\right\} \tag{16}$$

with $c = \exp[-(h_i - h_j)^2/2\ell_2^2]$ and the h's are sampled from the bivariate normal distribution $\mathcal{N}(0, k)$. We may proceed by approximating the random second moment inside the braces in Eq. (16) as

$$\begin{cases} \sigma_3^2 & |h_i - h_j| \leq \ell_2 \\ \dfrac{\sigma_3^2}{\sqrt{1 + 2\frac{\sigma_2^2}{\ell_3^2}}} & \text{otherwise} . \end{cases}$$

As such, we obtain the following approximate second moment for three-layer DGP,

$$\frac{\sigma_3^2}{\sqrt{1 + 2\frac{\sigma_2^2}{\ell_3^2}}}[1 - \text{erf}(v)] + \sigma_3^2 \text{ erf}(v) \tag{17}$$

where erf() represents the error function and the parameter $v = \frac{\ell_2}{\sqrt{2(k_{ii} - k_{ij})}}$.

4 Fourth Moment of SE[·]

The fourth moment is useful in revealing some important properties, e.g. heavy-tailedness, of a distribution. For univariate distribution, the measure of excess kurtosis is linked to the Gaussianity. For example, the heavy-tailed Student t-distribution has positive excess kurtosis whereas the normal distribution has zero. Here, we quantify the non-Gaussianity of DGP in Eq. (4) by investigating the fourth moments of the special case SE[·].

Lemma 4 *The product of two exponential quadratic functions has the expected value with respect to the multivariate distribution $\mathcal{N}(h_i, h_j, h_m, h_l|0, K)$,*

$$\mathbb{E}[e^{-(h_i - h_j)^2 - (h_m - h_l)^2}] = [G_{ij}G_{ml} - V]^{-\frac{1}{2}} \tag{18}$$

where $\frac{1}{\sqrt{G_{ij}}} = \mathbb{E}[e^{-(h_i - h_j)^2}]$ corresponds to the second moment for SE[·]. The term $V = (k_{im} + k_{jl} - k_{il} - k_{jm})^2$, with the k's denoting the matrix elements of off-diagonal block of covariance matrix K.

The proof follows the generalization of Lemma 2 to the case of four random variables. The determinant $|I_4 + KJ|$ is a bit tedious but can be done with the help of the equality, $|\left(\begin{smallmatrix} A & C \\ D & B \end{smallmatrix}\right)| = |A||B||I_2 - A^{-1}CB^{-1}D|$.

5 Approximating DGP with GP

In previous sections, we have shown the second and fourth moments of some DGP models. Despite the fact that DGP distribution in Eq. (4) is not tractable and non-Gaussian, these moments are exact and analytic (except the case of SE[SE[SE]]). Here, we propose to approximate the true DPG distribution with GP. Namely, the variational distribution,

$$q = \mathcal{N}(\mathbf{f}|0, k_{\text{eff}}) \tag{19}$$

is a multivariate normal distribution which minimizes the KL divergence, $\text{KL}(p||q)$ (this usage is not standard as the conventional KL is between true and approximate posterior distributions). Due to the Gaussianity of q, the second moments for true distribution p in Eq. (4) have the exact correspondence,

$$K_{\text{eff}} = \mathbb{E}_q[\mathbf{f}\mathbf{f}^t] = \mathbb{E}_p[\mathbf{f}\mathbf{f}^t]\,. \tag{20}$$

Therefore, the second moments turn out to be the effective covariance function $k_{\text{eff}} : R^D \times R^D \to R$.

Considering the case of SE[SE], the approximate distribution q acquires the effective covariance function,

$$k_{\text{eff}}(\mathbf{x}, \mathbf{y}) = \frac{\sigma_2^2}{\sqrt{1 + 2\frac{\sigma_1^2}{\ell_2^2}[1 - \exp(-||\mathbf{x} - \mathbf{y}||^2/2\ell_1^2)]}}\,, \tag{21}$$

where the σ's are signal magnitude and ℓ's the length scale in the respective layers. In addition, the heterogeneous composition SE[Lin] with linear kernel $k(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ in input layer gives rise to the rational quadratic kernel of order $\frac{1}{2}$. The composition can also include the non-stationary kernel such as the neural network kernel (Williams, 1997) in the first layer.

Likewise, the composition SC[·] results in many novel and interesting kernels as well. Composition SC[SE] generates an exponential Gaussian kernel. Surprisingly, the periodic kernel (MacKay, 1998) is identical to the composition SC[SC], and SC[Lin] turns out to be the addition of SE and constant kernels. See Supplementary Material for effective covariance function of SC[NuN].

For the two-layer composition SE[·], it is clear that the first layer can be regarded as a deterministic mapping when the signal magnitude σ_1 is set to zero. Thus, with $\mu^{(1)} \neq 0$, this two-layer DGP is equivalent to $\mathcal{GP}(0, k(\mu^{(1)}(\cdot), \mu^{(1)}(\cdot)))$. The nature of DGP distribution becomes transparent by inspecting the fourth moment. The fourth moment of the approximate distribution for SE[·],

$$\mathbb{E}_q[f_i^2 f_j^2] = \sigma_2^4\{1 + [1 + (k_{ii} + k_{jj} - 2k_{ij})/\ell_2^2]^{-1}\}\,, \tag{22}$$

whereas the true distribution,

$$\mathbb{E}_p[f_i^2 f_j^2] = \sigma_2^4\{1 + [1 + 2(k_{ii} + k_{jj} - 2k_{ij})/\ell_2^2]^{-1/2}\}\,. \tag{23}$$

If we treat the terms inside the bracket as $1 + x$ with $x > 0$, following the fact that $1/\sqrt{1 + 2x} \geq 1/(1 + x)$ for $x \geq 0$, then Eq. (23) is larger than or equal to Eq. (22). Nevertheless, for large value of ℓ_2^2, Eq. (22) and (23) are identical up to $O(\ell_2^{-2})$ in the expansion. Therefore, the distributions q for SE[SE] is a good approximation to p in the regime $\sigma_1/\ell_2 \ll 1$. More generally, the following remark states the fourth moments from p and q and the nature of distribution of SE[·].

Remark 2 *The fourth moments of SE[·] associated with the true distribution $p(\mathbf{f}|\mathbf{X})$ in Eq. (4) and the approximate distribution $q(\mathbf{f}|\mathbf{X})$ in Eq. (19) have the relation,*

$$\mathbb{E}_p[f_i f_j f_m f_l] \geq \mathbb{E}_q[f_i f_j f_m f_l]\,, \tag{24}$$

provided that their second moments match $\mathbb{E}_p[f_i f_j] = \mathbb{E}_q[f_i f_j]$.

The proof follows from that the mutivariate normal distribution has $\mathbb{E}_q[f_i f_j f_m f_l] = \sum \mathbb{E}_q[f_a f_b]\,\mathbb{E}_q[f_c f_d]$ with the summation over all three distinct ways of partitioning the quartet $\{i, j, m, l\}$ into two doublets (Isserlis, 1918). As for p, the partition is the same but each term is instead given by Eq. (18), which is greater than the product or the second moments. The particular fourth moment considered in Eq. (22) and (23) is reproduced by setting the random variables $f_i = f_m$ and $f_j = f_l$ here.

An important implication of above remark is that the true distribution for the zero-mean SE[·] DGP (SC[·] as well) is heavy-tailed non-Gaussian distribution. The above analysis is analogous to the comparison between GP and Student-t process (TP) (Shah et al., 2014). A TP with same covariance function as a GP can be shown to possess larger fourth moment than the GP (Press, 2005), which may account for the more extreme behavior in TP.

5.1 Recursive Composition

The above composition for two-layer DGP can be generalized to the cases where $L > 2$ in a recursive manner. For *homogeneous* composition, e.g. SE[SE[SE]]], one can treat SE[SE] inside the bracket as a GP with zero mean and effective covariance function given in Eq. (21). Thus, the new effective covariance function for the three-layer DGP is obtained by plugging this kernel back into Eq. (12). This trick can be applied

recursively so in general the following relation holds,

$$k_{\text{eff}}^{(L+1)}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sigma_{L+1}^2}{\sqrt{1 + 2(\ell_{L+1}^{-2})[\sigma_L^2 - k_{\text{eff}}^{(L)}(\mathbf{x}_i, \mathbf{x}_j)]}}\,, \tag{25}$$

which relates the current covariance function $k^{(L)}$ to the covariance $k^{(L+1)}$ after another layer of composition. Similar approach can be taken for obtaining other homogeneous composition kernels SC[SC[SC[...]]] (Eq. 15) and the non-stationary Gaussian kernels (Eq. 13).

For *heterogeneous* composition, e.g. SE[SC[NuN]], one can obtain the effective covariance function in similar manner, which leads to the expression for the effective covariance function,

$$\frac{\sigma_3^2}{\sqrt{1 + 2\frac{\sigma_2^2}{\ell_3^2}[1 - e^{(2k_{ij} - k_{ii} - k_{jj})/2\ell_2^2}]}}, \tag{26}$$

where the NuN covariance function k can be the general Gaussian kernel or the arcsine kernel [Eq. (11) in (Williams, 1997)].

5.2 Characteristics of Effective Kernel

Here we consider the distribution over derivatives of composite functions sampled from approximate DGP in order to make connection with the expressivity parameter defined in NN model (Poole et al., 2016). By definition, any two values f_1 and f_2 from a function $f(x)$ must follow the joint Gaussian distribution $\mathcal{N}(f_1, f_2|0, \Sigma)$ with the two-by-two covariance matrix Σ. In order to shed light on the expressiveness of our effective kernels, we calculate the following expectation,

$$\mathbb{E}[(f_1 - f_2)^2] = \int df_1 df_2 (f_1 - f_2)^2 \mathcal{N}(f_1, f_2|0, \Sigma), \tag{27}$$

which shall asymptotically approach $\mathscr{E}\{[f'(x)]^2\}(x_1 - x_2)^2$ as their the inputs x_1 and x_2 are close to each other. Consequently, we obtain the expected value of squared derivative,

$$\mathbb{E}\{[f'(x)]^2\} = 2 \lim_{x_1 \to x_2} \frac{\sigma_2^2 - k_{\text{eff}}(x_1, x_2)}{(x_1 - x_2)^2}\,. \tag{28}$$

When applying to the effective kernel of SE[SE] in Eq. (12), one may show the normalized and squared average derivative has

$$\frac{\mathbb{E}\{[f'(x)]^2\}}{\sigma_2^2} = \frac{\sigma_1^2}{\ell_2^2 \ell_1^2} = \frac{\chi}{\ell_1^2}\,, \tag{29}$$

where we follow Poole et al. (2016) and define the parameter important for the analysis of expressivity,

$$\chi := \frac{\partial(k_{\text{eff}}/\sigma_2^2)}{\partial(k/\sigma_1^2)}|_{k=\sigma_1^2}\,, \tag{30}$$

Notation	Effective kernel
SE[SE]	$a[1 + b\ G(\Delta x^2, c)]^{-\frac{1}{2}}$
SC[SE]	$a\{1 + \exp[-bG(\Delta x^2, c)]\}$
SE[SC]	$a[1 + b \sin^2(\Delta x/c)]^{-\frac{1}{2}}$
SC[SC]	$a\{1 + \exp[-b \sin^2(\Delta x/c)]\}$
SE[Lin]	$\text{RQ}\frac{1}{2}$
SC[Lin]	SE + Const.
SE[Lin+SE]	$a\{1 + bG(\Delta x^2, c) + d\Delta x^2\}^{-\frac{1}{2}}$
NuN[SE]	$a[1 + f + bG(\Delta x^2, c/2) + dG(\Delta x^2, c)]^{-\frac{1}{2}}$

Table 1: List of compositional kernels by stacking two GPs with respective kernels denoted by SE (squared exponential), SC (squared cosine), Lin (linear), and NN (neural network with Gaussian activation function). The function symbol $G(x^2, c) = 1 - \exp(-x^2/c)$ appears with SE composition. The sequence can be understood from the notation, for example, SE[SC] means input data is directed to the first GP with SC kernels, followed by sending its output to the second GP with SE kernel which produces the final output. The hyperparameters are represented by the symbols a, b, c, d, and f, all of which are positive.

which was suggested to signal the phase transition between chaotic and ordered phases in neural network (Sompolinsky et al., 1988; Poole et al., 2016). With the chain rule for derivative, we can also show that the effective SE[SE[SE]] has the normalized and squared average derivative χ/ℓ_1^2 with $\chi = \sigma_2^2\sigma_1^2/\ell_3^2\ell_2^2$. Similar construction is applicable to squared cosine DGP. This characteristic of the derivative of our effective kernels is consistent with (Duvenaud et al., 2014).

6 Interpretation of Effectively DGP

Our approximation based on GPs yields benefits in interpretability for both DGPs and DNNs. Prior results establish that DNNs can be viewed as GPs (Williams, 1997; Cho and Saul, 2009; Lee et al., 2017). In both the DGP and DNNs paradigms, people tend toward homogeneous kernels / activation functions, in part because there are not effective tools for predicting how compositions will behave a priori. In Table 1, we list a few representative analytic kernels from stacking two GPs. Our analysis allows iterative generation of heterogeneous kernels such as SE[SC[NuN]] and homogeneous one SE[SE[SE[...]]] of any depth. Moreover, because our approach sits between these two paradigms, we can view our compositions as arising from activation functions (Daniely et al., 2016), kernels (Schölkopf et al., 1998; Duvenaud et al., 2013), or compositions of both.

We may leverage interpretability to analyze differences in expressivity of SE[SE] and SE. First, SE[SE] ap-

proaches RQ$\frac{1}{2}$ in small Δx limit. Because the rational quadratic kernel is known to arise from summing over infinitely many SE kernels with Gamma distribution over the inverse of length scale (Rasmussen and Williams, 2006), the effective SE[SE] kernel also possesses the multi-length scale feature. Second, the function composition $y = f(h(x))$ leads to the long-range correlation, i.e. when $\Delta x \to \infty$, SE[SE] kernel approaches some nonzero constant while SE becomes vanishingly small. This can be seen by noting that the first-layer outputs $h(x_1)$ and $h(x_2)$ are nearly independent if x_1 and x_2 are far apart. Nevertheless, the distance $h(x_1) - h(x_2)$, generally falling within $[0, \sigma_1]$ with high probability, fed into the second layer is greatly reduced if the signal magnitude σ_1 associated with h is small, which leads to finite correlation between $f(h(x_1))$ and $f(h(x_2))$.

The composition using non-stationary kernel, i.e. NuN[$\cdot$], is also of interest. Take NuN[SE] for example, the determinant $|\mathbb{K}|$ in Eq. (13) generate a squared term, k_{ij}^2, and as such the combination $e^{-||\mathbf{x}_i-\mathbf{x}_j||^2/\ell^2} + e^{-||\mathbf{x}_i-\mathbf{x}_j||^2/2\ell^2}$ appears in the effective covariance function.

6.1 SE vs. SE[SE]

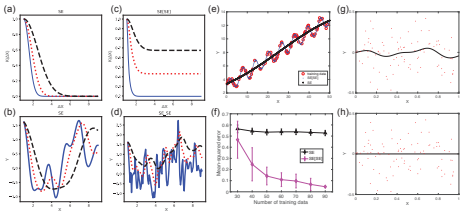

Figure 1: (a) Three different SE kernels and (b) functions sampled from these kernels. (c) Three different SE[SE] kernels and (d) functions sampled from these kernels. For (a) and (c), the kernel black dash, red dotted line, and blue solid line are 0.8, 0.5, and 0.2 at $\Delta x = 1$, respectively. All functions in (b) and (d) are generated using the same noise vector. (e)-(f) Regression on multiscale data with SE and SE[SE] kernels. (g)-(h) Regression with SE (g) and SE[SE] (h) kernels on pure noise data sampled from normal distribution with $\sigma = 0.2$. Red dots are data points, and black line is the predictive mean.

Figure 1(a)–(d) shows functions generated from SE and SE[SE]. The functions generated by SE[SE] seem to possess two length scales in variations: the rapidly varying blue line in Figure 1(d) seems to have slowly varying underlying trend that is similar to a shifted version of the black dash line. This is broadly consistent with the discussion in Section 6 of the additional expressivity of multi-length scale behavior in deeper structure.

To further demonstrate the expressivity of SE[SE] kernel over the SE kernel, we use them to fit a dataset with two length scales.[1] Figure 1(e) shows that the smooth SE kernel does not fit the small variation, while the SE[SE] kernel is able to capture these fast variation on top of the slowly varying one. Figure 1(f) shows that the SE[SE] kernel is better than SE kernel at prediction not only in the range of larger amount of data, but also in the limited data regime. We also remark in passing that the prediction accuracy for SE[SE] kernel is comparable with the RQ kernel, which is known to be multi-length scale.

Given the ability of the SE[SE] kernel to capture fast variation, one might have the concern that it will fit to noise. To explore this possibility, we trained both SE and SE[SE] kernels on pure noise data. In Figure 1(g) and (h), we show the prediction mean (black) from training on 90 noise data points sampled from the normal distribution with $\sigma = 0.2$. For this particular noise data, the SE kernel in (g) generates a non-zero prediction mean, while the SE[SE] kernel in (f) nicely predicts the underlying zero function.

We also apply the SE and SE[SE] kernels to two UCI regression data sets, the House Price dataset and the Abalone dataset. We investigated the test error as a function of the fraction of training number. For the House Price dataset, the SE[SE] kernel obtains lower test error than the SE kernel does when the fraction of training data is larger than 30% (see Supplementary Material Figure 1(a)). For the Abalone dataset, the two kernels have similar test error (see Supplementary Material Figure 1(b)).

6.2 Deep SE Compositions

We examine two different methods for obtaining the effective covariance function in SE[SE[SE]]. The straightforward way deals with the expectation of second moment from output to input layers. With approximation, Eq. (17) is obtained. On the other hand, the recursive approach treats the first two layers SE[SE] inside the bracket as a GP first, then take the three-layer DGP as another two-layer DGP, which results in Eq. (25). Fig. 2 (a) and (b), respectively, demonstrates the effective kernel functions from Eq. (17) and Eq. (25) for different hyperparameters. The prominent difference is the presence of plateau for small Δx in Fig. 2(a). For smaller value of σ_2/ℓ_3 and σ_1/ℓ_2, the two appraoches have closer results (green solid and dashed curves). The functions sampled from these kernels are shown in panel (c) and (d), respectively. With larger value of $\chi = (\frac{\sigma_2\sigma_1}{\ell_3\ell^2})^2$, the sampled

[1] The data is from MATLAB FITRGP example.

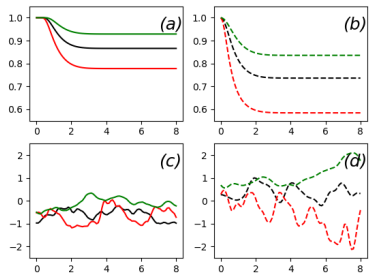

Figure 2: Sampling functions from effective kernels for SE[SE[SE]]. Panels (a) refers to the kernel using the approximate second moment in Eq. (17) while panel (b) uses the recursive relation in Eq. (25). The hyperparameters: green ($\sigma_2/\ell_3 = 0.8$, $\sigma_1/\ell_2 = 0.8$), black ($\sigma_2/\ell_3 = 1$, $\sigma_1/\ell_2 = 1$), and red ($\sigma_2/\ell_3 = 1.4$, $\sigma_1/\ell_2 = 1.2$). The common parameters: $\sigma_3 = \ell_1 = 1$. Panel (c) and (d) shows the functions sampled from panel (a) and (b), respectively.

functions in both panels seem to possess variation in shorter length scale. However, the sampled functions in panel (c) are more restricted in the range $[-\sigma_3, \sigma_3]$, but they are much less smooth than the functions obtained using kernel in (b).

7 Hyperparameter Optimization and Expressivity

From above analysis, it is clear that χ characterizes both the decay of kernel function [Fig. 2 (a) and (b)] and the degree associated with variation in sampled functions. Considering the recursive kernel of depth L, $\chi \propto (\frac{\sigma}{\ell})^{L-1}$ increases exponentially with the L if the ratio $\frac{\sigma_l}{\ell_{l+1}}$ is kept constant in the model. In (Poole et al., 2016), the parameter defined as $\chi_1 := \partial c_{12}^l / \partial c_{12}^{l-1}$ characterizes the the mapping of manifold in DNNs and serves as expressivity parameter which also signals the transition between chaotic and ordered phases in a mean-field theory for neural network (Sompolinsky et al., 1988).

As a demonstration that DGP with larger χ is more expressive, we generate data from the non-periodic kernels in Table 1. The hyperparameters of the data-generating kernel are chosen randomly, and we use SE[SE[SE]] kernel to fit the data by maximum marginal likelihood. We run the optimization 20 times with 20 random initialization where the hyperparameters are sampled uniformly from (-10,10) in log space. Figure 3 shows the final log of marginal likelihood versus the value of $\log \chi$ after optimization. It can be seen that a dramatic increase of marginal likelihood is

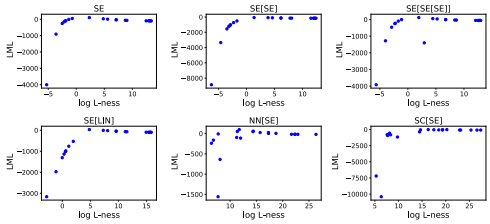

Figure 3: Log marginal likelihood (LML) for three-layer DGP SE[SE[SE]] [Eq. (25)] fitting the data generated by the kernel labeled on top of panel. The hyperparameters in generating kernels are chosen randomly. The LML shows a dramatic increase when the value of $\log \chi$, defined in Eq. (30), is close to zero in the top row of data. This agrees with Poole et al. (2016) who used χ from correlation maps to quantify the transition into the chaotic phase in which DNNs are much more expressive than shallow networks

observed near $\chi = O(1)$ in the top row of panels, while the transition is shifted in the bottom row. Therefore, the three-layer DGP have superior expressivity in the chaotic regime.

8 CONCLUSIONS

We have presented interpretable deep Gaussian Processes that combine increased expressiveness associated with deep NNs with uncertainty quantification of GPs. Marginalization over latent function variable is not tractable but calculation of moments is. Our approach is based on approximating deep GPs as GPs, which enables one to analytically integrate yielding effectively deep, single layer kernels. The heavy-tailed nature of some deep Gaussian process distribution is revealed by the fourth moment. We have provided a recipe for constructing effective kernels for cases including homogeneous and heterogeneous kernels (equivalently, activation functions), derived a variety of such kernels, analyzed their behavior, and confirmed behavior by prior and posterior predictive simulation. Simpler than alternative approaches to variational inference, our approach yields strong benefits in interpretability while retaining remarkable expressivity.

Acknowledgement

This material is based on research sponsored by the Air Force Research Laboratory and DARPA under agreement number FA8750-17-2-0146. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

Advani, M., Lahiri, S., and Ganguli, S. (2013). Statistical mechanics of complex neural systems and high dimensional data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03014.

Belkin, M., Ma, S., and Mandal, S. (2018). To understand deep learning we need to understand kernel learning. *arXiv preprint arXiv:1802.01396*.

Bui, T., Hernández-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*, pages 1472–1481.

Cho, Y. and Saul, L. K. (2009). Kernel methods for deep learning. In *Advances in neural information processing systems*, pages 342–350.

Cutajar, K., Bonilla, E. V., Michiardi, P., and Filippone, M. (2017). Random feature expansions for deep gaussian processes. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 884–893. JMLR. org.

Damianou, A. and Lawrence, N. (2013). Deep gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215.

Daniely, A., Frostig, R., and Singer, Y. (2016). Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *NIPS*.

Dunlop, M. M., Girolami, M. A., Stuart, A. M., and Teckentrup, A. L. (2018). How deep are deep gaussian processes? *The Journal of Machine Learning Research*, 19(1):2100–2145.

Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. (2013). Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922*.

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

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Havasi, M., Hernández-Lobato, J. M., and Murillo-Fuentes, J. J. (2018). Inference in deep gaussian processes using stochastic gradient hamiltonian monte carlo. In *Advances in Neural Information Processing Systems*, pages 7506–7516.

Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139.

Lázaro-Gredilla, M. (2012). Bayesian warped gaussian processes. In *Advances in Neural Information Processing Systems*, pages 1619–1627.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. (2017). Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*.

Liao, Z. and Couillet, R. (2018). The dynamics of learning: a random matrix approach. *arXiv preprint arXiv:1805.11917*.

MacKay, D. J. (1998). Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166.

Mairal, J., Koniusz, P., Harchaoui, Z., and Schmid, C. (2014). Convolutional kernel networks. In *Advances in neural information processing systems*, pages 2627–2635.

Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671.

Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. (2016). Exponential expressivity in deep neural networks through transient chaos. In *NIPS*.

Press, S. J. (2005). *Applied multivariate analysis: using Bayesian and frequentist methods of inference*. Courier Corporation.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Process for Machine Learning*. MIT press, Cambridge, MA.

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

Salimbeni, H., Dutordoir, V., Hensman, J., and Deisenroth, M. P. (2019). Deep gaussian processes with importance-weighted variational inference. *arXiv preprint arXiv:1905.05435*.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319.

Shah, A., Wilson, A., and Ghahramani, Z. (2014). Student-t processes as alternatives to gaussian processes. In *Artificial intelligence and statistics*, pages 877–885.

Snelson, E., Ghahramani, Z., and Rasmussen, C. E. (2004). Warped gaussian processes. In *Advances in neural information processing systems*, pages 337–344.

Sompolinsky, H., Crisanti, A., and Sommers, H.-J. (1988). Chaos in random neural networks. *Physical review letters*, 61(3):259.

Sun, S., Zhang, G., Wang, C., Zeng, W., Li, J., and Grosse, R. (2018). Differentiable compositional kernel learning for gaussian processes. *arXiv preprint arXiv:1806.04326*.

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574.

Titsias, M. and Lawrence, N. (2010). Bayesian gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851.

Van der Wilk, M., Rasmussen, C. E., and Hensman, J. (2017). Convolutional gaussian processes. In *Advances in Neural Information Processing Systems*, pages 2849–2858.

Wang, S. and Manning, C. (2013). Fast dropout training. In *international conference on machine learning*, pages 118–126.

Williams, C. K. (1997). Computing with infinite networks. In *Advances in neural information processing systems*, pages 295–301.

Wilson, A. G., Gilboa, E., Nehorai, A., and Cunningham, J. P. (2014). Fast kernel learning for multidimensional pattern extrapolation. In *Advances in Neural Information Processing Systems*, pages 3626–3634.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016). Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.