Neural Network Gaussian Processes by Increasing Depth

Shao-Qun Zhang†, Fei Wang‡, Senior Member, IEEE, and Feng-Lei Fan†, Member, IEEE

Abstract—Recent years have witnessed an increasing interest in the correspondence between infinitely wide networks and Gaussian processes. Despite the effectiveness and elegance of the current neural network Gaussian process theory, to the best of our knowledge, all the neural network Gaussian processes (NNGPs) are essentially induced by increasing width. However, in the era of deep learning, what concerns us more regarding a neural network is its depth as well as how depth impacts the behaviors of a network. Inspired by a width-depth symmetry consideration, we use a shortcut network to show that increasing the depth of a neural network can also give rise to a Gaussian process, which is a valuable addition to the existing theory and contributes to revealing the true picture of deep learning. Beyond the proposed Gaussian process by depth, we theoretically characterize its uniform tightness property and the smallest eigenvalue of the Gaussian process kernel. These characterizations can not only enhance our understanding of the proposed depth-induced Gaussian process but also pave the way for future applications. Lastly, we examine the performance of the proposed Gaussian process by regression experiments on two benchmark datasets.

Index Terms—Deep neural networks, generalized central limit theorem, neural network Gaussian processes (NNGPs), smallest eigenvalue, uniform tightness, weak dependence.

I. INTRODUCTION

Currently, kernel methods and deep neural networks are two of the most remarkable machine learning methodologies. Recent years have witnessed lots of works on their connection. Lee et al. [1] pointed out that randomly initializing parameters of an infinitely wide network gives rise to a Gaussian process, which is referred to as neural network Gaussian processes (NNGPs). Due to the attraction of this idea, the studies of NNGP have been scaled into more types of networks, such as attention-based models [2] and recurrent networks [3].

A Gaussian process is a classical nonparametric model. The equivalence between an infinitely wide fully connected network and a Gaussian process has been established in [1] and [4]. Given a fully connected multilayer network whose parameters are independent identically distributed (i.i.d.) randomly initialized, the output of each neuron is an aggregation of neurons in the preceding layer whose outputs are also i.i.d. When the network width goes infinitely large, according to the Central Limit Theorem [5], the output of each neuron conforms to the Gaussian distribution. As a result, the output function expressed by the network is essentially a Gaussian process. The correspondence between neural networks and Gaussian processes allows the exact Bayesian inference using the neural network [1].

Despite the achievements of the current NNGP theory, it has an important limitation that is not addressed satisfactorily. So far, the neural network Gaussian process is essentially induced by increasing width, regardless of how many layers are stacked in a network. But in the era of deep learning, what concerns us more regarding deep learning is its depth and how the depth affects the behaviors of a neural network, since the depth is the major element accounting for the power of deep learning. Although that the current NNGP theory is beautiful and elegant in its form, unfortunately, it cannot accommodate our concern adequately. Therefore, it is highly necessary to expand the scope of the existing theory to include the depth issue. Specifically, our natural curiosity is what is going to happen if we have an infinitely deep but finitely wide network. Can we derive an NNGP by increasing depth rather than width, which contributes to understanding the true picture of deep learning? If this question is positively answered, we are able to reconcile the successes of deep networks and the elegance of the NNGP theory. What’s more, as a valuable addition, the depth-induced NNGP greatly enlarges the scope of the existing NNGP theory, which is posited to open lots of doors for research and translation opportunities in this area.

The above idea is well-motivated based on a width-depth symmetry consideration. Previously, Lu et al. [6] and Hornik et al. [7] have respectively proved that the width-bounded and depth-bounded neural networks are universal approximators. Fan et al. [8] suggested that a wide network and a deep network can be converted to each other with a negligible error by De Morgan’s law. Since somehow there exists a symmetry between width and depth, deepening a neural network in certain conditions can likely lead to an NNGP as well. Along this direction, we investigate the feasibility of inducing an NNGP by depth (NNGP(d)), with a network of a shortcut topology in Fig. 1. The characteristic of this topology is that outputs of intermediate layers with a gap of $h$ are aggregated in the final layer, yielding the network output. Such a shortcut topology has been successfully applied to medical imaging [9] and computer vision [10] as a backbone structure.

An NNGP by width (NNGP(w)) is accomplished by summing the i.i.d. output terms of infinitely many neurons and applying Central Limit Theorem. In contrast, for the topology in Fig. 1, as the depth increases, the outputs of increasingly many neurons are aggregated together. We constrain the random weights and biases such that those summed neurons turn weakly dependent by the virtue of their separation. Consequently, when going infinitely deep,
the network is also a function drawn from a Gaussian process according to the generalized Central Limit Theorem under weak dependences [11]. Beyond the proposed NNGP, we theoretically prove that NNGP is uniformly tight and provide a tight bound of the smallest eigenvalue of the concerned NNGP kernel. From the former, one can determine the properties of NNGP such as the functional limit and continuity, while the non-trivial lower and upper bounds mirror the characteristics of the derived kernel, which constitutes a cornerstone for its optimization and generalization properties.

Main Contributions: In this manuscript, we establish the NNGP by increasing depth, in contrast to the present mainstream NNPs that are induced by width. Our work substantially enlarges the scope of the existing elegant NNGP theory, making a stride toward understanding the true picture of deep learning. Furthermore, we investigate the essential properties of the proposed NNGP and its associated kernel, which lays a solid foundation for future research and applications. Lastly, we implement an NNGP kernel and apply it for regression experiments on benchmark datasets.

II. PRELIMINARIES

Let \([N] = \{1, 2, \ldots, N\}\) be the set for an integer \(N > 0\). Given a function \(g(n)\), we denote by \(h_1(n) = \Theta(g(n))\) if there exist positive constants \(c_1, c_2,\) and \(n_0\) such that \(c_1 g(n) \leq h_1(n) \leq c_2 g(n)\) for every \(n \geq n_0\); \(h_2(n) = O(g(n))\) if there exist positive constants \(c\) and \(n_0\) such that \(h_2(n) \leq c g(n)\) for every \(n \geq n_0\); \(h_3(n) = \Omega(g(n))\) if there exist positive constants \(c\) and \(n_0\) such that \(h_3(n) \geq c g(n)\) for every \(n \geq n_0\). Let \([\mathbf{W}]\) denote the matrix norm for the matrix \(\mathbf{W} \in \mathbb{R}^{n_m \times n_m} \). Throughout this brief, we employ the maximum spectral norm \(\|\mathbf{W}\| = \max_k |\lambda_k|, \) for \(k \in \{\min[m, n]\}\) as the matrix norm [12], where \(\lambda_k\) denotes the \(k\)th singular value of the matrix \(\mathbf{W}\). Let \(|\cdot|\) be the number of elements, e.g., \(|\mathbf{W}| = nm\). Finally, we provide several definitions for the characterization of inputs and parameters.

Definition 1: A data distribution \(P\) is said to be well-scaled, if the following conditions hold for \(x \in \mathbb{R}^d\):

1) \(\int x \, dP(x) = 0\).
2) \(\int |x| \, dP(x) = \Theta(\sqrt{d})\).
3) \(\int |x|^{3/2} \, dP(x) = \Theta(d)\).

Definition 2: A function \(\sigma : \mathbb{R} \rightarrow \mathbb{R}\) is said to be well-scaled, if \(\sigma\) is first-order differentiable, and its derivative is bounded by a certain constant \(C_\sigma\). Specially, the commonly used activation functions like ReLU, tanh, and sigmoid are well-scaled (Please see Table I).

Definition 3: A matrix \(\mathbf{V}\) is said to be well-scaled for a well-scaled activation function \(\sigma\), in short \(\mathbf{V} \in \text{SP}(\sigma)\), if the inequality \(C_\sigma \|\mathbf{V}\| < 1\) holds.

III. MAIN RESULTS

In this section, we formally present the neural network Gaussian process NNGP, led by an infinitely deep but finitely wide neural network with i.i.d. weight parameters. We also derive the uniform tightness for NNGP with the increased depth and the bound estimation of its associated kernel’s smallest eigenvalue. These two valuable characterizations serve as the solid cornerstones for NNGP.

A. Neural Network Gaussian Process With Increasing Depth

Consider an \(L\)-layer neural network whose topology is illustrated as Fig. 1, the feed-forward propagation follows:

\[
\begin{align*}
\mathbf{z}_0 &= \mathbf{x} \\
\mathbf{z}_l &= \sigma (\mathbf{W}_l \mathbf{z}_{l-1} + \mathbf{b}_l) \quad (1)
\end{align*}
\]

where \(\mathbf{W}_l\) and \(\mathbf{b}_l\) are the weight matrix and bias vector of the \(l\)th layer, respectively, and \(\sigma\) is the activation function. Invoking shortcut connections, the final output of this network is a mean of \(x \in \mathbb{N}^+\) previous layers with an equal separation \(h \in \mathbb{N}^+\) and \(l_1 \in [L]\)

\[
f(x; \theta) = \frac{1}{\sqrt{|\mathbf{M}_L|}} \sum_{k=0}^{K} 1^{l_1+k-h} \mathbf{z}_{l_1+k} \tag{2}
\]

where the matrix \(1^{l_1+k-h} \in \{1\}^{n_{l_1+k-h}}\) indicates the unit shortcut connection between \(\mathbf{z}_{l_1+k}\) and the final layer, and \(M_L\) denotes the summed number of concerned hidden neurons

\[
M_L = \sum_{k=0}^{K} n_k \quad \text{with} \quad n_k = |1^{l_1+k-h}|_v.
\]

Let \(\theta = \text{concat}(\bigcup_{l=1}^{L} \text{vec} (\mathbf{W}_l))\) be the concatenation of all vectorized weight matrices and \(n = |\theta|\). Regarding the neural network \(f : \mathbb{R}^d \rightarrow \mathbb{R}^v\), we present the first main theorem as follows:

Theorem 1: The infinitely deep neural network, defined by (1) and (2), is equivalent to a Gaussian process NNGP, if \(\sigma\) is well-scaled and the augmented parameter matrix of each layer is stable-pertinent for \(\sigma\), that is, \((\mathbf{W}_l, \mathbf{b}_l) \in \text{SP}(\sigma)\), for \(\forall l \in [L]\).

Theorem 1 states that our proposed neural network converges to a Gaussian process as \(L \rightarrow \infty\). Given a dataset \(D = \{(x_i, y_i)\}_{i=0}^{m}\), the limit output variables of this network belongs to a multivariate Gaussian distribution \(N(0, \mathbf{K}_D, \nu)\) whose mean equals to 0 and covariance matrix is an \(N \times N\) matrix, the \((i,j)\)-entry of which is defined as

\[
\mathbf{K}(x_i, x_j) = \mathbb{E}[f(x_i; \theta), f(x_j; \theta)], \quad \text{for} \quad x_i, x_j \in D. \tag{3}
\]

The key idea of proving Theorem 1 is to show that our proposed neural network converges to a Gaussian process as depth increases according to the generalized Central Limit Theorem with weakly dependent variables instead of random ones. To implement this idea, we constrain the weights and biases to enable that random variables of two hidden layers with a sufficient separation degenerate to weak dependence, i.e., mixing processes. By aggregating the weak dependent variables to the final layer via shortcut connections, the output of the proposed network converges to a Gaussian process as the depth grows to infinity. The key steps are formally stated by Lemmas 1 and 2 as follows.

Lemma 1: Provided a well-scaled \(\sigma\) and stable-pertinent parameter matrices, the concerned neural network comprises a stochastic sequence of weakly dependent variables as the depth goes to infinity.

Proof: Let \(\mathcal{H}_t^\sigma\) denote the distribution of the random variable sequence \(\{Z^t, Z^{t+1}, \ldots, Z^t\}\), where \(0 \leq t < \tau\), and \(Z^t = (Z^t, \ldots, Z^t)\) indicates the vector of random variables before the timestamp \(t\). We define a coefficient [13] as

\[
\beta(\sigma) = \sup_{\tau} \mathbb{E}_{Z^t} \left[ \|\mathcal{H}_{t+\tau}^\sigma(\cdot | Z^t) - \mathcal{H}_{t+\tau}^\sigma(\cdot)\|_\mu \right]
\]

where \(\mathcal{H}(\cdot|\cdot)\) stands for a conditional probability distribution, and \(\mu\) denotes a probability measure, or equally the \(\sigma\)-algebra of

| Activations | Well-Posedness |
|-------------|----------------|
| ReLU        | \(|\sigma'(x)| \leq 1\) |
| tanh        | \(|\sigma'(x)| = |1 - \sigma'(x)| \leq 1\) |
| sigmoid     | \(|\sigma'(x)| = |\sigma(x)(1 - \sigma(x))| \leq 1/4\) |
Therefore, the sequence \( \{ Z_t \} \) converges in distribution to an exponential distribution with rate \( \beta \), which completes the proof. □

Theorem 2: Suppose that: 1) a random variable sequence \( \{ Z_t \} \) is weakly dependent, satisfying \( \beta \)-mixing with an exponential convergence rate and 2) for \( s \in \mathbb{N} \), we have

\[
\mathbb{E}[Z_s] = 0 \quad \text{and} \quad \mathbb{E}[(Z_s)^2] < \infty.
\]

Let \( \Lambda_t = Z_1^2 + Z_2^2 + \cdots + Z_t^2 \), then we have

\[
\mu \overset{d}{=} \lim_{t \to \infty} \mathbb{E}[\Lambda_t] = 0 \quad \text{and} \quad v^2 \overset{d}{=} \lim_{t \to \infty} \mathbb{E}(\Lambda_t^2)/t < \infty.
\]

Further, the limit variable \( \Lambda_t/(v \sqrt{t}) \) converges in distribution to \( \mathcal{N}(0, 1) \) as \( t \to \infty \), provided \( v \neq 0 \).

Lemma 2 is a variant of the generalized Central Limit Theorem under weak dependence. The proof idea can be summarized as follows. From [15], it’s observed that an \( \alpha \)-mixing sequence with an exponential convergence rate can be covered by the \( \alpha \)-mixing one with \( O(1/t) \). Thus, the conditions of Lemma 2 satisfy the preconditions of the generalized Central Limit Theorem under weak dependence [11, Theorem 27.5]. This lemma also has alternative proofs according to the encyclopedic treatment of limit theorems under mixing conditions. Interested readers can refer to [16] for more details.

1) Finishing the Proof of Theorem 1: Let \( z^i \) denote the output variables of the \( i \)-th layer, which satisfies that \( z^i+1 = \sigma(W^iz^i + b^i+1) \) and \( z^i = x \). Because the weights and biases are taken to be i.i.d., the sequence \( \{ z^i \} \) leads to a stochastic process, and the post-activations in the same layer, such as \( z^i \) and \( z^j \), are independent for \( i \neq j \). Given an integer \( h \in \mathbb{N} \), we select a sub-sequence of \( \{ z^i \} \) as follows:

\[
\exists z^i_h = \{ z^i+1_h, z^{i+2}_h, \ldots, z^{i+k}_h, \ldots \}
\]

for \( i \in [L] \) and \( k \in \mathbb{N} \), which satisfies \( l_i + kh \leq L \). From Lemma 1, the sequence \( z^i_h \) leads to a weakly dependent stochastic process. Aggregating this sub-sequence with \( k \) shortcut connections to the output layer, the output of the concerned neural network converges to a Gaussian process as \( k \to \infty \) as well as \( L \to \infty \), from Lemma 2.

2) Discussions: To the best of our knowledge, our proposed NNGP\((d)\) is the first NNGP induced by increasing depth. Currently, there is no rigorous definition for width and depth. The way we claim depth just aligns with the conventional usage of the width and depth for a neural network, in which the depth is understood as the maximum number of neurons among all possible routes from the input to the output, and the width is the maximum number of neurons in a layer. As illustrated in Fig. 2, if examined in an unraveled view, our network is a simultaneously wide and deep network due to the layer reuse in different routes. However, we argue that this will not affect our claim because not every layer has an infinite width in the unraveled view, which is different from the key character of NNGP\((d)\). What’s more, the conventional usage is more acceptable relative to the unraveled view; otherwise, it is against common sense because the ResNet is also a wide network in the unraveled view.

The existence of the proposed NNGP\((d)\) kernel relies heavily on the generalized Central Limit Theorem, which holds on three conditions as mentioned in Lemma 2: 1) the random variable sequence is weakly dependent; 2) the random variable maintains a finite mathematical variance; and 3) the input data are drawn from a compact set. According to these conditions, we make two remarks. First, as shown in Lemma 1, \( h \) provides a separation of the network depth to ensure that the layers at both ends of the separation interval are weakly dependent. Therefore, \( h \) is not necessarily an equal separation. Second, our proof doesn’t prescribe the distribution of the input data, as long as the input data are drawn from a compact set.

B. Uniform Tightness of NNGP\((d)\)

In this subsection, we delineate the asymptotic behavior of NNGP\((d)\) as the depth goes to infinity. Here, we assume that the weights and biases are i.i.d. sampled from \( \mathcal{N}(0, \sigma^2) \). Per the conditions of Theorem 1, we have the following theorem.

Theorem 2: For any \( l_i \in [L] \), the stochastic process, described in Lemma 1, is uniformly tight in \( \mathcal{C}(\mathbb{R}^d, \mathbb{R}) \).

Theorem 2 reveals that the stochastic process contained by our network (illustrated in Fig. 1) is uniformly tight, which is an intrinsic characteristic of NNGP\((d)\). Based on Theorem 2, one can obtain not only the functional limit and continuity properties of NNGP\((d)\) in analogy to the results of NNGP\((0)\) [17]. Similarly, we start the proof of Theorem 2 with some useful lemmas.

Lemma 3: Let \( \{ Z^1, Z^2, \ldots, Z^r \} \) denote a sequence of random variables in \( \mathcal{C}(\mathbb{R}^d, \mathbb{R}) \). This stochastic process is uniformly tight in \( \mathcal{C}(\mathbb{R}^d, \mathbb{R}) \), if (1) \( x = 0 \) is a uniformly tight point of \( Z^r(x) \) \( (s \in [r]) \) in \( \mathcal{C}(\mathbb{R}^d, \mathbb{R}) \); (2) for any \( x, x' \in \mathbb{R}^d \) and \( s \in [r] \), there exist \( \alpha, \beta, C > 0 \), such that \( \mathbb{E}[\| Z^r(x) - Z^r(x') \|^\beta] \leq C\| x - x' \|^\alpha \).
Lemma 3 is the core guidance for proving Theorem 2. This lemma can be straightforwardly derived from Kolmogorov Continuity Theorem [18], provided the Polish space \((\mathbb{R}, |\cdot|)\).

**Lemma 4:** Based on the notations of Lemma 3, \(x = 0\) is a uniformly tight point of \(Z'() (s \in [1])\) in \(C(\mathbb{R}^d, \mathbb{R})\).

**Proof:** It suffices to prove that \(1 \cdot x = 0\) is a tight point of \(Z'() (s \in [1])\) in \(C(\mathbb{R}^d, \mathbb{R})\) and the statistic \((Z'(0) + \ldots + Z'(0))/s\) converges in distribution as \(s \to \infty\). Note that: (1) is self-evident since every probability measure in \((\mathbb{R}, |\cdot|)\) is tight [19] and (2) has been proven by Theorem 1. Therefore, we finish the proof of this lemma.

**Remark:** Notice that the convergence in distribution \((\Rightarrow)\) from Lemmas 2 and 4 paves the way for the convergence of expectations. Specifically, provided a linear and bounded functional \(F : C(\mathbb{R}^d; \mathbb{R}^r) \to \mathbb{R}\) as \(L \to \infty\) and a function \(f\) which satisfies that \(f(x; \theta) \Rightarrow f^*\), then we have \(F(f(x; \theta)) \Rightarrow F(f^*)\) and \(\mathbb{E}[F(f(x; \theta))] \Rightarrow \mathbb{E}[F(f^*)]\) according to General Transformation Theorem [20, Theorem 2.3] and Uniform Integrability [21], respectively. These results may serve as solid bases for development and applications of NNPP\(_d\) in the future.

**Lemma 5:** Based on the notations of Lemma 3, for any \(x, x' \in \mathbb{R}^d\) and \(s \in [1]\), there exist \(a, \alpha, C > 0\), such that

\[
\mathbb{E}\left[\sup_i |Z'_i(x) - Z'_i(x')|^\alpha\right] \leq C||x - x'||^{\alpha + d}.
\]

The proof of Lemma 5 can be accessed from Appendix A. Further, Theorem 2 can be completely proved by invoking Lemmas 4 and 5 into Lemma 3.

**C. Tight Bound for the Smallest Eigenvalue**

In this subsection, we provide a tight bound for the smallest eigenvalue of the NNPP\(_d\) kernel. For the NNPP\(_d\) with ReLU activation, we have the following theorem.

**Theorem 3:** Suppose that \(x_1, \ldots, x_N\) are i.i.d. sampled from \(P_X = N(0, \eta_i^2)\), and \(P_X\) is a well-scaled distribution, then for an integer \(r \geq 2\), with probability \(1 - \delta > 0\), we have \(\lambda_{\min}(K_{P, D}) = \Theta(d)\), where

\[
\delta \leq Ne^{-\Omega(d)} + N^2e^{-\Omega(d)N^{1/2(-\alpha/3)}}.
\]

Theorem 3 provides a tight bound for the smallest eigenvalue of the NNPP\(_d\) kernel. This nontrivial estimation mirrors the characteristics of this kernel, and usually be used as a key assumption for optimization and generalization.

The key idea of proving Theorem 3 is based on the following inequalities about the smallest eigenvalue of real-valued symmetric square matrices. Given two symmetric matrices \(P, Q \in \mathbb{R}^{n \times n}\), it's observed that

\[
\lambda_{\min}(PQ) \geq \lambda_{\min}(P) \min_{i \in [n]} Q(i, i),
\]

\[
\lambda_{\min}(P + Q) \geq \lambda_{\min}(P) + \lambda_{\min}(Q).
\]

From (2) to (3), we can unfold \(K(x, x_j)\) as a sum of covariance of the sequence of random variables \(Z_i^k \sim \mathcal{N}(0, \eta_i^2)\). Thus, we can bound \(\lambda_{\min}(K_{P, D})\) by \(\text{Cov}(Z_i^k, Z_j^k)\) via a chain of feedforward compositions in (1). For conciseness, we put the proof of Theorem 3 into Appendix B.

**IV. EXPERIMENTS**

Generally, the depth can endow a network with a more powerful representation ability than the width. However, it is unclear whether or not the superiority of depth can sustain in the setting of NNPP, as all parameters are random rather than trained. In other words, it is unclear whether our established NNPP\(_d\) is more expressive than NNPP\(_{wo}\). To answer this question, in this section, we apply the NNPP\(_d\) kernel into the generic regression task and then compare its performance on the Fashion-MNIST (FMNIST) and CIFAR10 datasets with that of NNPP\(_{wo}\).

**A. NNPP\(_d\) Regression**

Provided the dataset \(D = \{(x_i, y_i)\}_{i=1}^n\), where \(x_i \in \mathbb{R}^{d_x}\) is the input, and \(y_i \in \mathbb{R}\) is the corresponding label, our goal is to predict \(y^*\) for the test sample \(x^*\). From Theorem 1, \(x^*\) and \(x^t\) belong to a multivariate Gaussian process \(\mathcal{N}(0, K^*)\), whose mean equals to 0, and covariance matrix has the following form:

\[
K^* = \begin{bmatrix}
K_{D, D} & K_{x^*, D} \\
K_{x^*, D} & K_{x^*, x^*}
\end{bmatrix}
\]

where \(K_{D, D}\) is an \(N \times N\) matrix computed by (3), and the \(i\)th element of \(K_{x^*, D} \in \mathbb{R}^{1 \times N}\) is \(K(x^*, x_i)\) for \(x_i \in D\). It's observed that (5) provides a division paradigm corresponding to the training set and test sample, respectively. Thus, we have \((\cdot | D, x^*) \in \mathcal{N}(\mu^*, K^*)\) with

\[
\mu^* = K_{x^*, D}(K_{D, D} + \eta_1^2I_n)^{-1}y^*
\]

\[
K^* = K_{x^*, x^*} - K_{x^*, D}(K_{D, D} + \eta_1^2I_n)^{-1}K_{x^*, D}^T
\]

where \(y^*\) and \(\eta_1\) denote the label vector. When the observations are corrupted by the Gaussian additive noise of \(\mathcal{N}(0, \eta^2)\), (6) becomes

\[
\mu^* = K_{x^*, D}(K_{D, D} + \eta_1^2I_n)^{-1}y^* + \mathcal{N}(0, \eta^2)
\]

\[
K^* = K_{x^*, x^*} - K_{x^*, D}(K_{D, D} + \eta_1^2I_n + \eta_2^2I_n)^{-1}K_{x^*, D}^T
\]

where \(I_n\) is the \(n \times n\) identity matrix. For numerical implementation, we calculate the kernels as, for \(x_1, x_2 \in D\),

\[
K(x_1, x_2) = \mathbb{E}([g(x_1; \theta), g(x_2; \theta)]^T)
\]

where \(g(\cdot; \theta)\) indicates the deep network or wide network.

**B. Experimental Setups**

We conduct regression experiments on FMNIST and CIFAR10 datasets. We respectively sample 1k, 2k, and 3k data from the training sets to construct two kernels and then test the performance of kernels on the test sets. Here, we employ a one-hidden-layer wide network to compute the NNPP\(_{wo}\) kernel, whereas the width of the deep network is set to the number of classes which is the smallest possible width for prediction tasks. For a fair comparison, the depth of NNPP\(_d\) and the width of NNPP\(_{wo}\) are equally set to 200 (\(h = 1\)). For classification tasks, the class labels are encoded into an opposite regression formation, where incorrect classes are \(-0.1\) and the correct class is 0.9 [1]. For two networks, we employ tanh as the activation function. Following the setting of NNPP\(_{wo}\) [1], all weights are initialized with a Gaussian distribution of the mean 0 and the variance of 0.3/\(n_l\) for normalization in each layer, where \(n_l\) is the number of neurons in the \(i\)th layer. The initialization is repeated 200 times to compute the empirical statistics of the NNPP\(_d\) and NNPP\(_{wo}\) based on (8). We also run each experiment 5 times for counting the mean and variance of accuracy. All experiments are conducted on Intel Core-i7-6500U.

**C. Results**

Table II lists the performance of the regression experimental results using NNPP\(_d\) and NNPP\(_{wo}\) kernels. It is observed that the test accuracy of NNPP\(_d\) and NNPP\(_{wo}\) kernels are comparable to each other, which implies that NNPP\(_d\) and NNPP\(_{wo}\) kernels are similar to each other in representation ability. The reason may be
that both NNGP\textsuperscript{(d)} and NNGP\textsuperscript{(w)} kernels are not stacked kernels. Their difference is mainly the aggregation of independent or weakly dependent variables. Thus, their ability should be similar [1].

Next, we use the angular plot to investigate how the separation \( h \) affects the representation ability of the NNGP\textsuperscript{(d)} kernel. The angle is computed according to

\[
\alpha = \arccos \left( \frac{K(x_1, x_2)}{\sqrt{K(x_1, x_1) \cdot K(x_2, x_2)}} \right)
\]

and the angular plot manifests the relationship between kernel values and angles. If an angular plot comes near zero, the kernel cannot well recognize the difference between samples. Otherwise, the kernel is regarded to have a better discriminative ability. We set the network depth to 200 × \( h \) so that the NNGP\textsuperscript{(d)} kernel is empirically computed by aggregating \( k = 200 \) shortcut connections with a separation of \( h \) between neighboring shortcut connections. Fig. 3 illustrates the angularities of NNGP\textsuperscript{(d)} kernels with \( h = 1, 3 \) for FMNIST-1k training data. It is observed that the angular plot of the kernel with \( h = 3 \) is compressed to be closer to zero relative to that of the kernel with \( h = 1 \), which implies that a smaller separation \( h \) may induce a powerful NNGP\textsuperscript{(d)} kernel.

To have a better understanding of the proposed NNGP\textsuperscript{(d)} kernel, we explore the impacts of the separation \( h \), the number of samples, the parameter variance, and the network size on it, as well as the computation time of the kernel in Appendix C. We have shared all our code in link1 and link2.

V. RELATED WORK

A. Deep Learning and Kernel Methods

There have been great efforts on correspondence between deep neural networks and Gaussian processes. Neal [4] presented the seminal work by showing that a one-hidden-layer network of infinite width turns into a Gaussian process. Cho and Saul [22] linked the multi-layer networks using rectified polynomial activation with compositional Gaussian kernels. Lee et al. [1] showed that the infinitely wide fully-connected neural networks with commonly-used activation functions can converge to Gaussian processes. Recently, the NNGP has been scaled to many types of networks including Bayesian networks [23], deep networks with convolution [24], and recurrent networks [3]. Furthermore, Wang et al. [25] wrote an inclusive review for studies on connecting neural networks and kernel learning. Despite great progress, all existing works about NNGP still rely on increasing width to induce the Gaussian processes, yet we go into the depth paradigm and offer an NNGP by increasing depth, which not only complements the existing theory to a good degree but also enhances our understanding to the true picture of “deep” learning.

B. Developments of NNGPs

Recent years have witnessed a growing interest in NNGPs. NNGPs can provide a quantitative characterization of how likely certain outcomes are if some aspects of the system are not exactly known. In the experiments of [1], an explicit estimate in the form of variance prediction is given to each test sample. Besides, Pang et al. [26] showed that the NNGP is good at handling data with noise and is superior to discretizing differential operators in solving some linear or nonlinear partially differential equations. Park et al. [27] employed the NNGP kernel in the performance measurement of network architectures for the purpose of speeding up the neural architecture search. Dutordoir et al. [28] presented the translation insensitive convolutional kernel by relaxing the translation invariance of deep convolutional Gaussian processes. Lu et al. [29] proposed an interpretable NNGP by approximating an NNGP with its low-order moments.

VI. CONCLUSION AND PROSPECTS

In this brief, we have presented the first depth-induced NNGP (NNGP\textsuperscript{(d)}) based on a width-depth symmetry consideration. Next, we have characterized the basic properties of the proposed NNGP\textsuperscript{(d)} kernel by proving its uniform tightness and estimating its smallest eigenvalue, respectively. Such results serve as a solid base for the understanding and application of the derived NNGP, such as the generalization and optimization properties and Bayesian inference with the NNGP\textsuperscript{(d)}. Lastly, we have conducted regression experiments on image classification and showed that our proposed NNGP\textsuperscript{(d)} kernel can achieve a performance comparable to the NNGP\textsuperscript{(w)} kernel. Future efforts can be put into scaling the proposed NNGP\textsuperscript{(d)} kernel into more applications.

REFERENCES

[1] J. Lee, Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, “Deep neural networks as Gaussian processes,” in Proc. 6th Int. Conf. Learn. Represent., 2018, pp. 1–19.
[2] J. Hron, Y. Bahri, J. Sohl-Dickstein, and R. Novak, “Infinite attention: NNGP and NTK for deep attention networks,” in Proc. 37th Int. Conf. Mach. Learn., 2020, pp. 4376–4386.
[3] G. Yang, “Wide feedforward or recurrent neural networks of any architecture are Gaussian processes,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 9947–9960.
[4] M. R. Neal, Prioris for Infinite Networks. Cham, Switzerland: Springer, 1996, pp. 29–53.
[5] H. Fischer, A History of the Central Limit Theorem: From Classical to Modern Probability Theory. Cham, Switzerland: Springer, 2010.
[6] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, “The expressive power of neural networks: A view from the width,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 6232–6240.
[7] M. Stinchcombe and H. White, “Multilayer feedforward networks are universal approximators,” Neural Netw., vol. 2, no. 5, pp. 359–366, Jun. 1988.
[8] F.-L. Fan, R. Lai, and G. Wang, “Quasi-equivalence of width and depth of neural networks,” 2020, arXiv:2002.02515.Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
[9] C. You et al., “CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE),” IEEE Trans. Med. Imag., vol. 39, no. 1, pp. 188–203, Jan. 2020.
[10] F. Fan et al., “On a sparse shortcut topology of artificial neural networks,” 2018, arXiv:1811.09003.
[11] P. Billingsley, Probability and Measure. Hoboken, NJ, USA: Wiley, 1995.
[12] C. Meyer, Matrix Analysis and Applied Linear Algebra, vol. 71. Philadelphia, PA, USA: SIAM, 2000.
[13] S.-Q. Zhang and Z.-H. Zhou, “Harmonic recurrent process for time series forecasting,” in Proc. 24th Eur. Conf. Artif. Intell., 2020, pp. 1714–1721.
[14] H. Joe, Multivariate Models and Multivariate Dependence Concepts. Boca Raton, FL, USA: CRC Press, 1997.
[15] P. Doukhan, Mixing: Properties and Examples. Cham, Switzerland: Springer, 2012.
[16] R. Bradley, Introduction to Strong Mixing Conditions. Los Angeles, CA, USA: Kendrick Press, 2007.
[17] D. Bracale, S. Favaro, S. Fortini, and S. Peluchetti, “Large-width functional asymptotics for deep Gaussian neural networks,” in Proc. 8th Int. Conf. Learn. Represent., 2020. [Online]. Available: https://openreview.net/pdf?id=OaW6dYOYB7d
[18] D. Stroock and S. Varadhan, Multidimensional Diffusion Processes, vol. 233. Cham, Switzerland: Springer, 1997.
[19] S.-Q. Zhang and Z.-H. Zhou, “ARISE: ApeRIodic SEmi-parametric process for efficient markets without periodogram and gaussianity assumptions,” 2021, arXiv:2111.06222.
[20] A. W. V. D. Vaart, Asymptotic Statistics. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[21] P. Billingsley, Convergence of Probability Measures. Hoboken, NJ, USA: Wiley, 2013.
[22] Y. Cho and L. Saul, “Kernel methods for deep learning,” in Proc. Adv. Neural Inf. Process. Syst., 2009, pp. 342–350.
[23] R. Novak et al., “Bayesian deep convolutional networks with many channels are Gaussian processes,” in Proc. 6th Int. Conf. Learn. Represent., 2018. [Online]. Available: https://openreview.net/pdf?id=B1g30j0qF7
[24] A. Garriga-Alonso, L. Aitchison, and CE. Rasmussen, “Deep convolutional networks as shallow Gaussian processes,” in Proc. 7th Int. Conf. Learn. Represent., 2019. [Online]. Available: https://openreview.net/pdf?id=Bklfsi0cKm
[25] T. Wang, L. Zhang, and W. Hu, “Bridging deep and multiple kernel learning: A review,” Inf. Fusion, vol. 67, pp. 3–13, Mar. 2021.
[26] G. Pang, L. Yang, and G. E. Karniadakis, “Neural-net-induced Gaussian process regression for function approximation and PDE solution,” J. Comput. Phys., vol. 384, pp. 270–288, May 2019.
[27] D. S. Park, J. Lee, D. Peng, Y. Cao, and J. Sohl-Dickstein, “Towards NNGP-guided neural architecture search,” 2020, arXiv:2011.06006.
[28] V. Dutordoir, M. Wilk, A. Artemev, and J. Hensman, “Bayesian image classification with deep convolutional Gaussian processes,” in Proc. 23rd Int. Conf. Artif. Intell. Statist., 2020, pp. 1529–1539.
[29] C.-K. Lu, S. Yang, X. Hao, and P. Shafto, “Interpretable deep Gaussian processes with moments,” in Proc. 23rd Int. Conf. Artif. Intell. Statist., 2020, pp. 613–623.