# Neural Network Gaussian Processes by Increasing Depth

Shao-Qun Zhang[1], Fei Wang[2], *Senior Member, IEEE*, Feng-Lei Fan[3*], *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 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 data sets.

*Index Terms*—Deep neural networks, neural network Gaussian processes, generalized Central Limit Theorem, weak dependence, uniform tightness, smallest eigenvalue

## 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* (NNGP). 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 non-parametric model. The equivalence between an infinitely wide fully-connected network and a Gaussian process has been established in [1], [4]. Given a fully-connected multi-layer network whose parameters are 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

*Dr. F.-L. Fan (hitfanfenglei@gmail.com) is the corresponding author.

[1] Shao-Qun Zhang is with National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. Email: zhangsq@lamda.nju.edu.cn

[2] Dr. Fei Wang is with Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY 10065, USA.

[3] Feng-Lei Fan was with Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA. Now he is a postdoctoral associate in Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY 10065, USA.
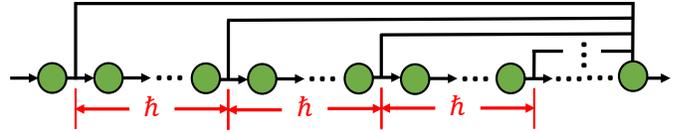


Fig. 1. A deep topology that can induce a neural network Gaussian process by increasing depth.

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 limit 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 can not 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 Figure 1. The characteristic of this topology is that outputs of intermediate layers with a gap of $\hbar$ 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 Figure 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 dependence [11]. Beyond the proposed NNGP$^{(d)}$, we theoretically prove that NNGP$^{(d)}$ is uniformly tight and provide a tight bound of the smallest eigenvalue of the concerned NNGP$^{(d)}$ kernel. From the former, one can determine the properties of NNGP$^{(d)}$ such as the functional limit and continuity, while the nontrivial 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 NNGPs that are induced by width. Our work substantially enlarges the scope of the existing elegant NNGP theory, making a stride towards 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$^{(d)}$ 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 $n \geq n_0$; $h_2(n) = \mathcal{O}(g(n))$ if there exist positive constants $c$ and $n_0$ such that $h_2(n) \leq cg(n)$ for every $n \geq n_0$; $h_3(n) = \Omega(g(n))$ if there exist positive constants $c$ and $n_0$ such that $h_3(n) \geq cg(n)$ for every $n \geq n_0$. Let $\|\mathbf{W}\|$ denote the matrix norm for the matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$. Throughout this paper, we employ the maximum spectral norm

$$\|\mathbf{W}\| \overset{\text{def}}{=} \max_k |\lambda_k|, \quad \text{for} \quad k \in [\min\{m, n\}],$$

as the matrix norm [12], where $\lambda_k$ denotes the $k$-th singular value of the matrix $\mathbf{W}$. Let $|\cdot|_\#$ denote 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 $\boldsymbol{x} \in \mathbb{R}^d$:*
1) $\int \boldsymbol{x} \, dP(\boldsymbol{x}) = 0$;
2) $\int \|\boldsymbol{x}\|_2 \, dP(\boldsymbol{x}) = \Theta(\sqrt{d})$;
3) $\int \|\boldsymbol{x}\|_2^2 \, dP(\boldsymbol{x}) = \Theta(d)$.

**Definition 2.** *A function $\sigma : \mathbb{R} \to \mathbb{R}$ is said to be **well-posed**, 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-posed (Please see Table I).*

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

TABLE I
WELL-POSEDNESS OF THE COMMONLY-USED ACTIVATION FUNCTIONS.

| Activations | Well-Posedness |
|---|---|
| ReLU | $\|\sigma'(\boldsymbol{x})\| \leq 1$ |
| tanh | $\|\sigma'(\boldsymbol{x})\| = \|1 - \sigma^2(\boldsymbol{x})\| \leq 1$ |
| sigmoid | $\|\sigma'(\boldsymbol{x})\| = \|\sigma(\boldsymbol{x})(1 - \sigma(\boldsymbol{x}))\| \leq 1/4$ |

## III. MAIN RESULTS

In this section, we formally present the neural network Gaussian process NNGP$^{(d)}$, led by an infinitely deep but finitely wide neural network with i.i.d. weight parameters. We also derive the uniform tightness for NNGP$^{(d)}$ 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$^{(d)}$.

### A. Neural Network Gaussian Process with Increasing Depth

Consider an $L$-layer neural network whose topology is illustrated as Figure 1, the feed-forward propagation follows

$$\begin{cases} \boldsymbol{z}^0 = \boldsymbol{x} \\ \boldsymbol{z}^l = \sigma(\mathbf{W}^l \boldsymbol{z}^{l-1} + \boldsymbol{b}^l) \, , \end{cases} \quad (1)$$

where $\mathbf{W}^l$ and $\boldsymbol{b}^l$ 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 $\kappa \in \mathbb{N}^+$ previous layers with an equal separation $\hbar \in \mathbb{N}^+$ and $l_1 \in [L]$

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{\sqrt{M_{\boldsymbol{z}}}} \sum_{\kappa=0}^{K} \mathbf{1}^{l_1 + \kappa\hbar} \boldsymbol{z}^{l_1 + \kappa\hbar} \, , \quad (2)$$

where the matrix $\mathbf{1}^{l_1 + \kappa\hbar} \in \{1\}^{n_o \times n_{l_1 + \kappa\hbar}}$ indicates the unit shortcut connection between $\boldsymbol{z}^{l_1 + \kappa\hbar}$ and the final layer, and $M_{\boldsymbol{z}}$ denotes the summed number of concerned hidden neurons

$$M_{\boldsymbol{z}} = \sum_{\kappa=0}^{K} n_\kappa \quad \text{with} \quad n_\kappa = |\boldsymbol{z}^{l_1 + \kappa\hbar}|_\# \, .$$

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

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

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

$$\mathbf{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathbb{E}[\langle f(\boldsymbol{x}_i; \boldsymbol{\theta}), f(\boldsymbol{x}_j; \boldsymbol{\theta}) \rangle], \quad \text{for} \quad \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D} \, . \quad (3)$$

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 weakly dependent variables to the final layer via shortcut connections, the output of the proposed network converges to a Gaussian process as the depth goes to infinity. The key steps are formally stated by Lemmas 1 and 2 as follows:

**Lemma 1.** *Provided a well-posed $\sigma$ 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 $\mathcal{H}_s^t$ denote the distribution of the random variable sequence $\{Z^s, Z^{s+1}, \ldots, Z^t\}$, where $0 \leq s < t$, and $\mathbf{Z}^{-t} = (Z^0, \ldots, Z^t)$ indicates the vector of random variables before the timestamp $t$. We define a coefficient [13] as

$$\beta(s) = \sup_t \mathbb{E}_{\mathbf{Z}^{-t}} \left[ \| \mathcal{H}_{t+s}^{+\infty}(\cdot \mid \mathbf{Z}^{-t}) - \mathcal{H}_{t+s}^{+\infty}(\cdot) \|_\mu \right] ,$$

where $\mathcal{H}(\cdot|\cdot)$ stands for a conditional probability distribution, and $\mu$ denotes a probability measure, or equally the $\sigma$-algebra of events $\mathcal{G}$ [14], which satisfies

$$\| P - Q \|_\mu = \sup_{z \in \mathcal{G}} |P(z) - Q(z)| ,$$

for two probability distributions $P$ and $Q$. According to Eq. (1), we have $Z^l = \sigma(\tilde{\mathbf{W}}^l \tilde{Z}^{l-1})$ for all $l \in [L]$, where $\tilde{\mathbf{W}}^l = (\mathbf{W}^l, \boldsymbol{b}^l)$ and $\tilde{Z}^{l-1} = (Z^{l-1}; 1)$. Given the well-posed $\sigma$ and stable-pertinent parameter matrices, *i.e.*, $\tilde{\mathbf{W}}^l \in SP(\sigma)$ for any $l \in [L]$, the followings hold

$$\frac{\partial Z^{l+s}}{\partial Z^l} \leq R^s \quad \text{and} \quad \mathbb{E}_{Z^l, \tilde{\mathbf{W}}} \left[ \frac{|Z^{l+s} Z^l|}{|Z^l|} \right] \leq R^s |Z^l| ,$$

where $C_\sigma \| \tilde{\mathbf{W}}^l \| \leq R < 1$ and $s \in \mathbb{N}^+$. This implies that (informally) the "dependence" between variables $Z^l$ and $Z^{l+s}$ goes to be weak as $s \to \infty$. From Sklar's theorem, we have

$$\mathcal{H}^{l+s}(\cdot) \wedge \mathcal{H}^l(\cdot) = C_l(s) \cdot \mathcal{H}^{l+s}(\cdot) \cdot \mathcal{H}^l(\cdot) ,$$

where $C_l(s) \in \Omega(R^s)$ is the corresponding Copula function. Further, it holds

$$\mathcal{H}_{l+s}^{+\infty}(\cdot \mid \mathbf{Z}^{-l}) - \mathcal{H}_{l+s}^{+\infty}(\cdot) = \sum_l C_l(s) \cdot \mathcal{H}_{l+s}^{+\infty}(\cdot) \cdot \mathcal{H}^l(\cdot) .$$

Since $C_l(s)$ is independent to the layer (*i.e., time*) index $l$, we assert that $\beta(s)$ is proportional to $C_l(s)$. Thus, we have

$$\beta(s) \to 0 \quad \text{as} \quad s \to +\infty .$$

Therefore, the sequence $\{Z^t\}$ led by Eq. (1) is $\beta$-mixing, or equally weakly dependent, which completes the proof. □

**Lemma 2.** *Suppose that (i) a random variable sequence $\{Z^s\}_{s=1}^t$ is weakly independent, satisfying $\beta$-mixing with an exponential convergence rate, (ii) for $\forall s \in [t]$, we have*

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

*Let $\Lambda_t = Z^1 + Z^2 + \cdots + Z^t$, then we have*

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

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

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 $\beta$-mixing sequence with an exponential convergence rate can be covered by the $\alpha$-mixing one with $\mathcal{O}(t^{-5})$. 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.

***Finishing the Proof of Theorem 1***. Let $\boldsymbol{z}^l$ denote the output variables of the $l$-th layer, which satisfies that $\boldsymbol{z}^{l+1} = \sigma(\mathbf{W}^{l+1} \boldsymbol{z}^l + \boldsymbol{b}^{l+1})$ and $\boldsymbol{z}^0 = \boldsymbol{x}$. Because the weights and biases are taken to be i.i.d., the sequence $\{\boldsymbol{z}^l\}$ ($l \in [L]$) leads to a stochastic process, and the post-activations in the same layer, such as $\boldsymbol{z}_i^l$ and $\boldsymbol{z}_j^l$ are independent for $i \neq j$. Given an integer $\hbar \in \mathbb{N}^+$, we select a sub-sequence of $\{\boldsymbol{z}^l\}$ as follows:

$$\mathcal{Z}_\hbar^{l_1} = \{\boldsymbol{z}^{l_1 + \hbar}, \boldsymbol{z}^{l_1 + 2\hbar}, \ldots, \boldsymbol{z}^{l_1 + \kappa\hbar}, \ldots\} ,$$

for $l_1 \in [L]$ and $\kappa \in \mathbb{N}^+$, which satisfies $l_1 + \kappa\hbar \leq L$. From Lemma 1, the sequence $\mathcal{Z}_\hbar^{l_1}$ leads to a weakly dependent stochastic process. Aggregating this sub-sequence with $\kappa$ shortcut connections to the output layer, the output of the concerned neural network converges to a Gaussian process as $\kappa \to \infty$ as well as $L \to \infty$, from Lemma 2. □
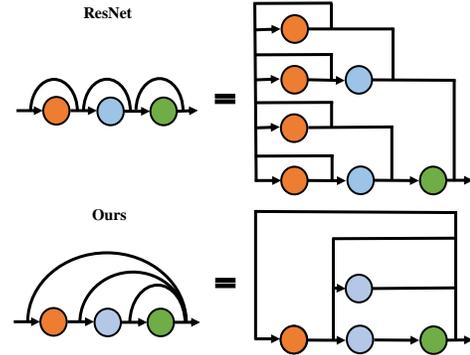


Fig. 2. Both ResNet and ours can be regarded as wide networks in the unraveled view.

**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 Figure 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 $\text{NNGP}^{(w)}$. 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 $\text{NNGP}^{(d)}$ kernel relies heavily on the generalized Central Limit Theorem, which holds on three conditions as mentioned in Lemma 2: i) The random variable sequence is weakly dependent; ii) the random variable maintains a finite mathematical variance; iii) the input data are drawn from a compact set. According to these conditions, we make two remarks. First, as shown in Lemma 1, $\hbar$ provides a separation of the network depth to ensure that the layers at both ends of the separation interval are weakly dependent. Therefore, $\hbar$ 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 $\text{NNGP}^{(d)}$

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

**Theorem 2.** *For any $l_1 \in [L]$, 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 Figure 1) is uniformly tight, which is an intrinsic characteristic of $\text{NNGP}^{(d)}$. Based on Theorem 2, one can obtain not only the functional limit and continuity properties of $\text{NNGP}^{(d)}$, in analogy to the results of $\text{NNGP}^{(w)}$ [17]. Similarly, we start the proof of Theorem 2 with some useful lemmas.

**Lemma 3.** *Let $\{Z^1, Z^2, \ldots, Z^t\}$ 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) $\boldsymbol{x} = \boldsymbol{0}$ is a uniformly tight point of $Z^s(\boldsymbol{x})$ ($s \in [t]$) in $\mathcal{C}(\mathbb{R}^d, \mathbb{R})$; (2) for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ and $s \in [t]$, there exist $\alpha, \beta, C > 0$, such that $\mathbb{E}\left[|Z^s(\boldsymbol{x}) - Z^s(\boldsymbol{x}')|^\alpha\right] \leq C\|\boldsymbol{x} - \boldsymbol{x}'\|^{\beta+d}$.*

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, $\boldsymbol{x} = \boldsymbol{0}$ is a uniformly tight point of $Z^s(\boldsymbol{x})$ ($s \in [t]$) in $\mathcal{C}(\mathbb{R}^d, \mathbb{R})$.*

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

**Remark.** Notice that the convergence in distribution ($\xrightarrow{\text{d}}$) from Lemmas 2 and 4 paves the way for the convergence of expectations. Specifically, provided a linear and bounded functional $\mathcal{F} : \mathcal{C}(\mathbb{R}^d; \mathbb{R}^{n^*}) \to \mathbb{R}$ as $L \to \infty$ and a function $f$ which satisfies that $f(\boldsymbol{x}; \boldsymbol{\theta}) \xrightarrow{\text{d}} f^*$, then we have $\mathcal{F}(f(\boldsymbol{x}; \boldsymbol{\theta})) \xrightarrow{\text{d}} \mathcal{F}(f^*)$ and $\mathbb{E}\left[\mathcal{F}(f(\boldsymbol{x}; \boldsymbol{\theta}))\right] \to \mathbb{E}\left[\mathcal{F}(f^*)\right]$ 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 $\text{NNGP}^{(d)}$ in the future.

**Lemma 5.** *Based on the notations of Lemma 3, for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ and $s \in [t]$, there exist $\alpha, \beta, C > 0$, such that*

$$\mathbb{E}\left[\sup_i \left|Z_i^s(\boldsymbol{x}) - Z_i^s(\boldsymbol{x}')\right|^\alpha\right] \leq C\|\boldsymbol{x} - \boldsymbol{x}'\|^{\beta+d} .$$

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 $\text{NNGP}^{(d)}$ kernel. For the $\text{NNGP}^{(d)}$ with ReLU activation, we have the following theorem:

**Theorem 3.** *Suppose that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are i.i.d. sampled from $P_X = \mathcal{N}(0, \eta^2)$, and $P_X$ is a well-scaled distribution, then for an integer $r \geq 2$, with probability $1 - \delta > 0$, we have $\lambda_{\min}(\mathbf{K}_{\mathcal{D}, \mathcal{D}}) = \Theta(d)$, where*

$$\delta \leq N e^{-\Omega(d)} + N^2 e^{-\Omega\left(dN^{-2/(r-0.5)}\right)} .$$

Theorem 3 provides a tight bound for the smallest eigenvalue of the $\text{NNGP}^{(d)}$ 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 $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{m \times m}$, it's observed that

$$\begin{cases} \lambda_{\min}(\mathbf{PQ}) \geq \lambda_{\min}(\mathbf{P}) \min_{i \in [m]} \mathbf{Q}(i, i) , \\ \lambda_{\min}(\mathbf{P} + \mathbf{Q}) \geq \lambda_{\min}(\mathbf{P}) + \lambda_{\min}(\mathbf{Q}) . \end{cases} \quad (4)$$

From Eqs. (2) and (3), we can unfold $\mathbf{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ as a sum of covariance of the sequence of random variables $\{\boldsymbol{z}^{l_1+\kappa\hbar}\}$. Thus, we can bound $\lambda_{\min}(\mathbf{K}_{\mathcal{D}, \mathcal{D}})$ by $\text{Cov}(\boldsymbol{z}^{l_1}, \boldsymbol{z}^{l_1})$ via a chain of feedforward compositions in Eq. (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 NNGP, as all parameters are random rather than trained. In other words, it is unclear whether our established $\text{NNGP}^{(d)}$ is more expressive than $\text{NNGP}^{(w)}$. To answer this question, in this section, we apply the $\text{NNGP}^{(d)}$ kernel into the generic regression task and then compare its performance on the Fashion-MNIST (FMNIST) and CIFAR10 data sets with that of $\text{NNGP}^{(w)}$.

**NNGP**$^{(d)}$ **regression.** Provided the data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^{d \times 1}$ is the input, and $y_i \in \mathbb{R}$ is the corresponding label, our goal is to predict $y^*$ for the test sample $\boldsymbol{x}^*$. From Theorem 1, $\boldsymbol{x}_i$ and $\boldsymbol{x}^*$ belong to a multivariate Gaussian process $\mathcal{N}(0, \mathbf{K}^*)$, whose mean equals to 0, and covariance matrix has the following form:

$$\mathbf{K}^* = \begin{bmatrix} \mathbf{K}_{\mathcal{D},\mathcal{D}} & \mathbf{K}_{\boldsymbol{x}^*,\mathcal{D}}^\top \\ \mathbf{K}_{\boldsymbol{x}^*,\mathcal{D}} & \mathbf{K}_{\boldsymbol{x}^*,\boldsymbol{x}^*} \end{bmatrix} , \tag{5}$$

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

$$\begin{cases} \mu^* = \mathbf{K}_{\boldsymbol{x}^*,\mathcal{D}} \mathbf{K}_{\mathcal{D},\mathcal{D}} \boldsymbol{y}^\top, \\ K^* = \mathbf{K}_{\boldsymbol{x}^*,\boldsymbol{x}^*} - \mathbf{K}_{\boldsymbol{x}^*,\mathcal{D}} \mathbf{K}_{\mathcal{D},\mathcal{D}}^{-1} \mathbf{K}_{\boldsymbol{x}^*,\mathcal{D}}^\top , \end{cases} \tag{6}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ denotes the label vector. When the observations are corrupted by the Gaussian additive noise of $\mathcal{N}(0, \eta^2)$, Eq. (6) becomes

$$\begin{cases} \mu^* = \mathbf{K}_{\boldsymbol{x}^*,\mathcal{D}} (\mathbf{K}_{\mathcal{D},\mathcal{D}} + \eta^2 \mathbf{I}_n) \boldsymbol{y}^\top , \\ K^* = \mathbf{K}_{\boldsymbol{x}^*,\boldsymbol{x}^*} - \mathbf{K}_{\boldsymbol{x}^*,\mathcal{D}} (\mathbf{K}_{\mathcal{D},\mathcal{D}} + \eta^2 \mathbf{I}_n)^{-1} \mathbf{K}_{\boldsymbol{x}^*,\mathcal{D}}^\top , \end{cases} \tag{7}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. For numerical implementation, we calculate the kernels as, for $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}$,

$$\mathbf{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathbb{E}[\langle g(\boldsymbol{x}_i; \boldsymbol{\theta}), g(\boldsymbol{x}_j; \boldsymbol{\theta}) \rangle] , \tag{8}$$

where $g(\cdot; \boldsymbol{\theta})$ indicates the deep network or wide network.

**Experimental setups.** We conduct regression experiments on FMNIST and CIFAR10 data sets. 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 NNGP$^{(w)}$ 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 NNGP$^{(d)}$ and the width of NNGP$^{(w)}$ are equally set to 200 ($\hbar = 1$). 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 NNGP$^{(w)}$ [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 $l$-th layer. The initialization is repeated 200 times to compute the empirical statistics of the NNGP$^{(d)}$ and NNGP$^{(w)}$ based on Eq. (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.

**Results.** Table II lists the performance of the regression experimental results using NNGP$^{(d)}$ and NNGP$^{(w)}$ kernels. It is observed that the test accuracy of NNGP$^{(d)}$ and NNGP$^{(w)}$ kernels are comparable to each other, which implies that NNGP$^{(d)}$ and NNGP$^{(w)}$ kernels are similar to each other in representation ability. The reason may be that both NNGP$^{(d)}$ and NNGP$^{(w)}$ kernels are not stacked kernels. Their difference

TABLE II
TEST ACCURACY OF REGRESSION EXPERIMENTS BASED ON NNGP$^{(d)}$ AND NNGP$^{(w)}$ KERNELS.

| Model | FMNIST | Test accuracy | CIFAR10 | Test accuracy |
|---|---|---|---|---|
| NNGP$^{(d)}$ | 1k | **0.345±0.016** | 1k | 0.166±0.018 |
| NNGP$^{(w)}$ | | 0.342±0.021 | | **0.187±0.018** |
| NNGP$^{(d)}$ | 2k | 0.352±0.019 | 2k | 0.178±0.007 |
| NNGP$^{(w)}$ | | **0.373±0.030** | | **0.188±0.012** |
| NNGP$^{(d)}$ | 3k | **0.372 ±0.024** | 3k | 0.182±0.005 |
| NNGP$^{(w)}$ | | 0.365±0.007 | | **0.185±0.019** |

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 $\hbar$ affects the representation ability of the NNGP$^{(d)}$ kernel. The angle is computed according to

$$\alpha = \arccos\left( \frac{\mathbf{K}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\sqrt{\mathbf{K}(\boldsymbol{x}_1, \boldsymbol{x}_1) \cdot \mathbf{K}(\boldsymbol{x}_2, \boldsymbol{x}_2)}} \right),$$

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 $200 \times \hbar$ so that the NNGP$^{(d)}$ kernel is empirically computed by aggregating $\kappa = 200$ shortcut connections with a separation of $\hbar$ between neighboring shortcut connections. Figure 3 illustrates the angularities of NNGP$^{(d)}$ kernels with $\hbar = 1, 3$ for FMNIST-1k training data. It is observed that the angular plot of the kernel with $\hbar = 3$ is compressed to be closer to zero relative to that of the kernel with $\hbar = 1$, which implies that a smaller separation $\hbar$ may induce a powerful NNGP$^{(d)}$ kernel.

To have a better understanding of the proposed NNGP$^{(d)}$ kernel, we explore the impacts of the separation $\hbar$, 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.
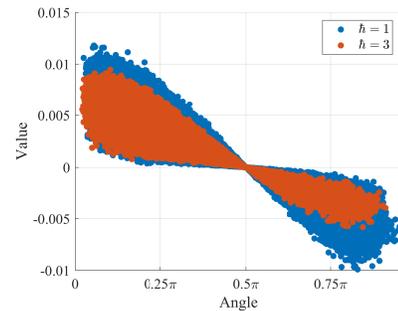


Fig. 3. Angularities of NNGP$^{(d)}$ kernels with various $\hbar$.

## V. RELATED WORK

**Deep Learning and Kernel Methods.** There have been great efforts on correspondence between deep neural networks and Gaussian processes. Neal *et al.* [4] presented the seminal work by showing that a one-hidden-layer network of infinite width turns into a Gaussian process. Cho *et al.* [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.

**Developments of NNGPs.** Recent years have witnessed a growing interest in neural network Gaussian processes. 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. CONCLUSIONS AND PROSPECTS

In this paper, we have presented the first depth-induced NNGP (NNGP$^{(d)}$) based on a width-depth symmetry consideration. Next, we have characterized the basic properties of the proposed NNGP$^{(d)}$ 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$^{(d)}$. Lastly, we have conducted regression experiments on image classification and showed that our proposed NNGP$^{(d)}$ kernel can achieve a performance comparable to the NNGP$^{(w)}$ kernel. Future efforts can be put into scaling the proposed NNGP$^{(d)}$ kernel into more applications.

## VII. ACKNOWLEDGMENTS

## REFERENCES

[1] J. Lee, Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as gaussian processes. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[2] Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: nngp and ntk for deep attention networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4376–4386, 2020.

[3] Greg Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. In *Advances in Neural Information Processing Systems 32*, pages 9947–9960, 2019.

[4] Radford M. Neal. *Priors for Infinite Networks*, pages 29–53. Springer, 1996.

[5] Hans Fischer. *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer Science & Business Media, 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 *Advances in Neural Information Processing Systems 30*, pages 6232–6240, 2017.

[7] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[8] Feng-Lei Fan, Rongjie Lai, and Ge Wang. Quasi-equivalence of width and depth of neural networks. *ArXiv e-prints*, 2020.

[9] Chenyu You, Guang Li, Yi Zhang, Xiaoliu Zhang, Hongming Shan, Mengzhou Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, et al. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE Transactions on Medical Imaging*, 39(1):188–203, 2019.

[10] Fenglei Fan, Dayang Wang, Hengtao Guo, Qikui Zhu, Pingkun Yan, Ge Wang, and Hengyong Yu. On a sparse shortcut topology of artificial neural networks. *ArXiv e-prints*, 2018.

[11] P. Billingsley. *Probability and Measure*. John Wiley & Sons, 1995.

[12] C. Meyer. *Matrix Analysis and Applied Linear Algebra*, volume 71. SIAM, 2000.

[13] S.-Q. Zhang and Z.-H. Zhou. Harmonic recurrent process for time series forecasting. In *Proceedings of the 24th European Conference on Artificial Intelligence*, pages 1714–1721, 2020.

[14] H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. CRC Press, 1997.

[15] P. Doukhan. *Mixing: Properties and Examples*. Springer Science & Business Media, 2012.

[16] R. Bradley. *Introduction to Strong Mixing Conditions*. Kendrick Press, 2007.

[17] D. Bracale, S. Favaro, S. Fortini, and S. Peluchetti. Large-width functional asymptotics for deep gaussian neural networks. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

[18] D. Stroock and S. Varadhan. *Multidimensional Diffusion Processes*, volume 233. Springer Science & Business Media, 1997.

[19] S.-Q. Zhang and Z.-H. Zhou. Arise: Aperiodic semi-parametric process for efficient markets without periodogram and gaussianity assumptions. *arXiv:2111.06222*, 2021.

[20] Aad W Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

[21] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013.

[22] Y. Cho and L. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems 22*, pages 342–350, 2009.

[23] R. Novak, L. Xiao, Y. Bahri, J. Lee, G. Yang, J. Hron, D.A. Abolafia, J. Pennington, and J. Sohl-dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. In *Proceedings of 6th International Conference on Learning Representations*, 2018.

[24] A. Garriga-Alonso, L. Aitchison, and CE. Rasmussen. Deep convolutional networks as shallow gaussian processes. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[25] T. Wang, L. Zhang, and W. Hu. Bridging deep and multiple kernel learning: a review. *Information Fusion*, 67:3–13, 2020.

[26] G. Pang, L. Yang, and G.E. Karniadakis. Neural-net-induced gaussian process regression for function approximation and pde solution. *Journal of Computational Physics*, 384:270–288, 2019.

[27] Daniel S Park, Jaehoon Lee, Daiyi Peng, Yuan Cao, and Jascha Sohl-Dickstein. Towards nngp-guided neural architecture search. *ArXiv e-prints*, 2020.

[28] V. Dutordoir, M. Wilk, A. Artemev, and J. Hensman. Bayesian image classification with deep convolutional gaussian processes. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 1529–1539, 2020.

[29] C.-K. Lu, S. Yang, X. Hao, and P. Shafto. Interpretable deep gaussian processes with moments. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 613–623, 2020.

# Appendices of 'Neural Network Gaussian Processes by Increasing Depth'

Shao-Qun Zhang[1], Fei Wang[2], *Senior Member, IEEE*, Feng-Lei Fan[3*], *Member, IEEE*

In Appendices A-B, we prove the uniform tightness of the depth-induced kernel and provide the tight bound of the smallest eigenvalue. In Appendix C, to have a better understanding of the proposed NNGP$^{(d)}$ kernel, we explore the impacts of the separation $\hbar$, the number of samples, the parameter variance, and the network size on it, as well as the computation time of the kernel.

## A. Uniform Tightness of NNGP$^{(d)}$

**Lemma 1** (Lemma 5 in the manuscript). *Based on the notations in the manuscript, for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ and $s \in [t]$, there exist $\alpha, \beta, C > 0$, such that*

$$\mathbb{E}\left[\sup_i \left|Z_i^s(\boldsymbol{x}) - Z_i^s(\boldsymbol{x}')\right|^\alpha\right] \leq C\|\boldsymbol{x} - \boldsymbol{x}'\|^{\beta+d}.$$

*Proof.* This proof follows mathematical induction. Before that, we show the following preliminary result. Let $\theta$ be one element of the augmented matrix $(\mathbf{W}^l, \boldsymbol{b}^l)$ at the $l$-th layer, then we can formulate its characteristic function as

$$\varphi(t) = \mathbb{E}\left[e^{\mathrm{i}\theta t}\right] = e^{-\eta^2 t^2/2} \quad \text{with} \quad \theta \sim \mathcal{N}(0, \eta^2),$$

where i denotes the imaginary unit with $\mathrm{i} = \sqrt{-1}$. Thus, the variance of hidden random variables at the $l^{th}$ layer becomes

$$\sigma_l^2 = \eta^2 \left[1 + \frac{1}{n_l}\sum_{i=1}^{n_l}\left|\varphi \circ Z_i^{l-1}\right|^2\right]. \tag{1}$$

Since the activation $\sigma$ is a well-posed function and $(\mathbf{W}^l, \boldsymbol{b}^l) \in SP(\sigma)$, we affirm that $\varphi$ is Lipschitz continuous (with Lipschitz constant $L_\varphi$).

Now we start the mathematical induction. When $s = 1$, for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ and $s \in [t]$, we have

$$\mathbb{E}\left[\sup_i \left|Z_i^1(\boldsymbol{x}) - Z_i^1(\boldsymbol{x}')\right|^\alpha\right] \leq C_{\eta,\theta,\alpha}\|\boldsymbol{x} - \boldsymbol{x}'\|^\alpha,$$

where $C_{\eta,\theta,\alpha} = \eta^\alpha \mathbb{E}[|\mathcal{N}(0,1)|^\alpha]$. Per mathematical induction, for $s \geq 1$, we have

$$\mathbb{E}\left[\sup_i \left|Z_i^s(\boldsymbol{x}) - Z_i^s(\boldsymbol{x}')\right|^\alpha\right] \leq C_{\eta,\theta,\alpha}\|\boldsymbol{x} - \boldsymbol{x}'\|^\alpha.$$

*Dr. F.-L. Fan (hitfanfenglei@gmail.com) is the corresponding author.

[1]Shao-Qun Zhang is with National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. Email: zhangsq@lamda.nju.edu.cn

[2]Dr. Fei Wang is with Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY 10065, USA.

[3]Feng-Lei Fan was with Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA. Now he is a postdoctoral associate in Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY 10065, USA.

Thus, one has

$$\mathbb{E}\left[\sup_i \left|Z_i^s(\boldsymbol{x}) - Z_i^s(\boldsymbol{x}')\right|^\alpha\right]$$
$$\leq (C_\sigma)^\alpha \,\mathbb{E}[|\mathcal{N}(0,1)|^\alpha]\,\left|Z_j^{s-1}(\boldsymbol{x}) - Z_j^{s-1}(\boldsymbol{x}')\right|^\alpha, \tag{2}$$

where

$$C_\sigma = \sigma_0^2(\boldsymbol{x}) - 2\Sigma_{\boldsymbol{x},\boldsymbol{x}'} + \sigma_0^2(\boldsymbol{x}')$$
$$= \frac{\eta^2}{n_{s-1}}\sum_{j=1}^{n_{s-1}}\left|\varphi \circ Z_j^{s-1}(\boldsymbol{x}) - \varphi \circ Z_j^{s-1}(\boldsymbol{x}')\right|^2 \quad \text{(from Eq. (1))}$$
$$\leq \frac{\eta^2 L_\varphi^2}{n_{s-1}}\sum_{j=1}^{n_{s-1}}\left|Z_j^{s-1}(\boldsymbol{x}) - Z_j^{s-1}(\boldsymbol{x}')\right|^2.$$

Thus, Eq. (2) becomes

$$\mathbb{E}\left[\sup_i \left|Z_i^s(\boldsymbol{x}) - Z_i^s(\boldsymbol{x}')\right|^\alpha\right] \leq C'_{\eta,\theta,\alpha}\left|Z_j^{s-1}(\boldsymbol{x}) - Z_j^{s-1}(\boldsymbol{x}')\right|^\alpha,$$

where

$$C'_{\eta,\theta,\alpha} = \frac{(\eta L_\varphi)^\alpha}{n_{s-1}}\sum_{j=1}^{n_{s-1}}\left|Z_j^{s-1}(\boldsymbol{x}) - Z_j^{s-1}(\boldsymbol{x}')\right|^\alpha \,\mathbb{E}[|\mathcal{N}(0,1)|^\alpha].$$

Iterating this argument, we obtain

$$\mathbb{E}\left[\sup_i \left|Z_i^s(\boldsymbol{x}) - Z_i^s(\boldsymbol{x}')\right|^\alpha\right] \leq C_{\eta,\theta,\alpha}\|\boldsymbol{x} - \boldsymbol{x}'\|^\alpha,$$

where

$$C_{\eta,\theta,\alpha} = \eta^{\alpha(s+1)} L_\varphi^{\alpha s} \,\mathbb{E}[|\mathcal{N}(0,1)|^\alpha]^{s+1}.$$

The above induction holds for any positive even $\alpha$. Let $\beta = \alpha - d > 0$, then this lemma is proved as desired. $\qquad\square$

## B. Tight Bound for the Smallest Eigenvalue

**Theorem 1** (Theorem 3 in the manuscript). *Suppose that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are i.i.d. sampled from $P_X = \mathcal{N}(0, \eta^2)$ and $P_X$ is a well-scaled distribution, then for an integer $r \geq 2$, with probability $1 - \delta > 0$, we have $\lambda_{\min}(\mathbf{K}_{\mathcal{D},\mathcal{D}}) = \Theta(d)$, where*

$$\delta \leq Ne^{-\Omega(d)} + N^2 e^{-\Omega\left(dN^{-2/(r-0.5)}\right)}.$$

We begin this proof with the following lemmas.

**Lemma 2.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz continuous function with constant $L$ and $P_X$ denote the Gaussian distribution $\mathcal{N}(0, \eta^2)$, then for $\forall\, \delta > 0$, there exists $c > 0$, s.t.*

$$\mathbb{P}\left(\left|f(\boldsymbol{x}) - \int f(\boldsymbol{x}')\,\mathrm{d}P_X(\boldsymbol{x}')\right| > \delta\right) \leq 2e^{\frac{-c\delta^2}{L^2}}. \tag{3}$$

Lemma 2 shows that the Gaussian distribution corresponding to our samples satisfies the log-Sobolev inequality (i.e.,

Eq. (3)) with some constants unrelated to dimension $d$. This result also holds for the uniform distributions on the sphere or unit hypercube [1].

**Lemma 3.** *Suppose that $x_1, \ldots, x_N$ are i.i.d. sampled from $\mathcal{N}(0, \eta^2)$, then with probability $1 - \delta > 0$, we have*

$$\|x_i\|_2 = \Theta(\sqrt{d}) \quad and \quad |\langle x_i, x_j \rangle|^r \leq dN^{-1/(r-0.5)},$$

*for $i \neq j$, where*

$$\delta \leq Ne^{-\Omega(d)} + N^2 e^{-\Omega\left(dN^{-2/(r-0.5)}\right)}.$$

*Proof.* From Definition 1 of the manuscript, we have

$$\int \|x\|_2^2 \, \mathrm{d}P_X(x) = \Theta(d).$$

Since $x_1, \ldots, x_n$ are i.i.d. sampled from $P_X = \mathcal{N}(0, \eta^2)$, for $\forall\, i \in [N]$, we have $\|x_i\|_2^2 = \Theta(d)$ with probability at least $1 - Ne^{\Omega(d)}$. Provided $x_i$, the single-sided inner product $\langle x_i, \cdot \rangle$ is Lipschitz continuous with the constant $L = \mathcal{O}(\sqrt{d})$. As such, from Lemma 2, for $\forall\, j \neq i$, we have

$$\mathbb{P}\left(|\langle x_i, x_j \rangle| > \delta^*\right) \leq 2e^{-\delta^2/L^2}.$$

Then, for $r \geq 2$, we have

$$\mathbb{P}\left(\max_{j \neq i} |\langle x_i, x_j \rangle|^r > \delta^*\right) \leq N^2 e^{-\Omega\left(\delta^{*2}\right)}.$$

We complete the proof by setting $\delta^* \leq dN^{-1/(r-0.5)}$. $\square$

***Proof of Theorem 1.*** We start this proof with some notations. Recall the empirical $\text{NNGP}^{(d)}$ kernel $\mathbf{K}_{\mathcal{D},\mathcal{D}}$. For convenience, we force $n^* = |z_1|_\# = |z_2|_\# = \cdots = |z_L|_\#$. We also abbreviate the covariance $\text{Cov}(z^{l_1+\kappa\hbar}, z^{l_1+\kappa\hbar})$ as $\mathbf{C}_{l_1+\kappa\hbar}$ and pick $l_1 = 1$ throughout this proof.

Unfolding the $\text{NNGP}^{(d)}$ kernel equation

$$\mathbf{K}(x_i, x_j) = \mathbb{E}[\langle f(x_i; \theta), f(x_j; \theta)\rangle], \quad \text{for} \quad x_i, x_j \in \mathcal{D},$$
(4)

we have

$$\mathbf{K}(x_i, x_j) = \frac{1}{M_z}\left[\sum_\kappa \varphi_\kappa + \sum_{\kappa_1 \neq \kappa_2} \phi_{\kappa_1, \kappa_2}\right],$$
(5)

where

$$\begin{cases} \varphi_\kappa = \mathbb{E}\left[\langle z^{l_1+\kappa\hbar}, z^{l_1+\kappa\hbar}\rangle\right], \\ \phi_{\kappa_1, \kappa_2} = \sum_{p,q} \mathbb{E}\left[z_p^{l_1+\kappa_1\hbar} z_q^{l_1+\kappa_2\hbar}\right], \quad \text{for} \quad \kappa_1 \neq \kappa_2, \end{cases}$$

in which the subscript $p$ indicates the $p$-th element of vector $z^{l_1+\kappa_1\hbar}$. From Theorem 1 of the manuscript, the sequence of random variables $\{z^{l_1}, z^{l_1+\hbar}, \ldots, z^{l_1+\kappa\hbar}\}$ is weakly dependent with $\beta(s) \to \infty$ as $s \to \infty$. Thus, $\phi_{\kappa_1, \kappa_2}$ is an infinitesimal with respect to $l_1 + |\kappa_2 - \kappa_1|\hbar$ when $\kappa_1 \neq \kappa_2$ and $\hbar$ is sufficiently large.

Invoking the following equations

$$\begin{cases} \lambda_{\min}(\mathbf{PQ}) \geq \lambda_{\min}(\mathbf{P}) \min_{i \in [m]} \mathbf{Q}(i,i), \\ \lambda_{\min}(\mathbf{P} + \mathbf{Q}) \geq \lambda_{\min}(\mathbf{P}) + \lambda_{\min}(\mathbf{Q}) \end{cases}$$
(6)

into Eq. (5), we have

$$\lambda_{\min}(\mathbf{K}_{\mathcal{D},\mathcal{D}}) \geq \sum_\kappa \lambda_{\min}\left(\mathbf{C}_{l_1+\kappa\hbar}\right),$$
(7)

$$\lambda_{\min}\left(\mathbf{C}_{l_1+\kappa\hbar}\right) \geq \lambda_{\min}\left(\mathbf{C}_{l_1+\kappa\hbar-1}\right), \quad \text{for} \quad \kappa \in \mathbb{N}.$$
(8)

Iterating Eq. (8) and then invoking it into Eq. (7), we have

$$\lambda_{\min}(\mathbf{K}_{\mathcal{D},\mathcal{D}}) \geq \sum_\kappa \lambda_{\min}\left(\mathbf{C}_1\right).$$
(9)

From the Hermite expansion [2] of ReLU function, we have

$$\mu_r(\sigma) = (-1)^{\frac{r-2}{2}} (r-3)!! / \sqrt{2\pi r!},$$
(10)

where $r \geq 2$ indicates the expansion order. Thus, we have

$$\begin{aligned}
\lambda_{\min}\left(\mathbf{C}_1\right) &= \lambda_{\min}\left(\sigma(\mathbf{W}^1 \mathbf{X})\sigma(\mathbf{W}^1 \mathbf{X})^\top\right) \\
&\geq \mu_r(\sigma)^2 \lambda_{\min}\left(\mathbf{X}^{(r)}\left(\mathbf{X}^{(r)}\right)^\top\right) \\
&\geq \mu_r(\sigma)^2 \left(\min_{i \in [N]} \|x_i\|_2^{2r} - (N-1)\max_{j \neq i}|\langle x_i, x_j\rangle|^r\right) \\
&\geq \mu_r(\sigma)^2 \Omega(d),
\end{aligned}$$
(11)

where the superscript $(r)$ denotes the $r$-th Khatri Rao power of the matrix $\mathbf{X}$, the first inequality follows from Eq. (10), the second one holds from Gershgorin Circle Theorem [3], and the third one follows from Lemma 3. Therefore, we can obtain the lower bound of the smallest eigenvalue by plugging Eq. (11) into Eq. (9).

On the other hand, it's observed from Lemma 1 of the manuscript that for $l \in [L]$,

$$\begin{cases} \|z_p^l\|_2^2 = \mathbb{E}_{\mathbf{W}_p^l}\left[\sigma(\mathbf{W}_p^l z^{l-1})^2\right] = \|z_q^l\|^2, \quad \text{for} \quad \forall q \neq p, \\ \|z^l\|_2^2 = \mathbb{E}_{\mathbf{W}^l}\left[\sigma(\mathbf{W}^l z^{l-1})^2\right] \leq \|z^l\|_2^2. \end{cases}$$
(12)

Thus, we have

$$\begin{aligned}
\lambda_{\min}(\mathbf{K}_{\mathcal{D},\mathcal{D}}) &\leq \frac{\text{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{N} = \frac{1}{N}\sum_i^N \mathbf{K}(x_i, x_i) \\
&\leq \frac{1}{N}\sum_i^N \frac{1}{M_z}\left[\sum_\kappa \varphi_\kappa + \sum_{\kappa_1 \neq \kappa_2} \phi_{\kappa_1, \kappa_2}\right] \\
&\leq \frac{1}{N}\sum_i^N \left(\frac{1}{\kappa}\sum_\kappa \max_{j \in [N]} \|x_j\|_2^2 + \Omega(d)\right) \\
&\leq \Theta(d),
\end{aligned}$$

where the second inequality follows from Eq. (5), the third one follows from Eq. (12), and the fourth one holds from Lemma 3. This completes the proof. $\square$

### C. Analysis Experiments

To have a better understanding of the proposed $\text{NNGP}^{(d)}$ kernel, here we explore the impacts of the separation $\hbar$, the number of samples, the parameter variance, and the network size on it, as well as the computation time. Now we introduce them one by one.

**The impact of $\hbar$.** We set the network depth to $200 \times \hbar$ so that the $\text{NNGP}^{(d)}$ kernel is empirically computed by aggregating $\kappa = 200$ shortcut connections with a separation of $\hbar$. For a comprehensive comparison, $\hbar$ is selected from

| $\hbar$ | Test accuracy |
|---|---|
| 1 | 0.329±0.027 |
| 2 | 0.198±0.027 |
| 3 | 0.179±0.022 |
| 4 | 0.145±0.014 |
| 5 | 0.146±0.017 |

$\{1, 2, 3, 4, 5\}$. Next, the NNGP$^{(d)}$ kernels are constructed with these networks and FMNIST-4k training data. All parameters are initialized with a mean of 0 and a variance of 0.5. Table I demonstrates the testing performance of so-built NNGP$^{(d)}$ kernels with respect to the FMNIST test data. As suggested by our angular plot analysis in the main body, the kernel with a larger $\hbar$ is compressed to be closer to zero relative to the kernel with a lower $\hbar$. Correspondingly, the kernel with a larger $\hbar$ should have lower discriminative ability. Table I shows that a larger $\hbar$ leads to an inferior test accuracy, which agrees with our analysis. We conclude that the separation $\hbar$ should be set to a smaller number to make a powerful NNGP$^{(d)}$ kernel.

**The impact of number of samples.** Here we investigate the impact of the number of training samples on the model's performance. We still conduct regression experiments on FM-NIST and CIFAR-10 data sets. Following the configurations in the main body, we respectively sample 1k, 2k, 3k, 4k, 5k data from the training sets to construct NNGP$^{(d)}$ and NNGP$^{(w)}$ kernels. Figure 1 shows the testing accuracy and its associated error bars of two kernels on FMNIST and CIFAR-10. Regarding the NNGP$^{(d)}$ kernel, its test accuracy culminates at 3k for both datasets. While for NNGP$^{(w)}$ kernel, the maximum accuracy is reached at 2k and 3k for FMNIST and CIFAR-10, respectively. We conclude that NNGP$^{(w)}$ and NNGP$^{(d)}$ kernels have a similar performance-sample behavior.
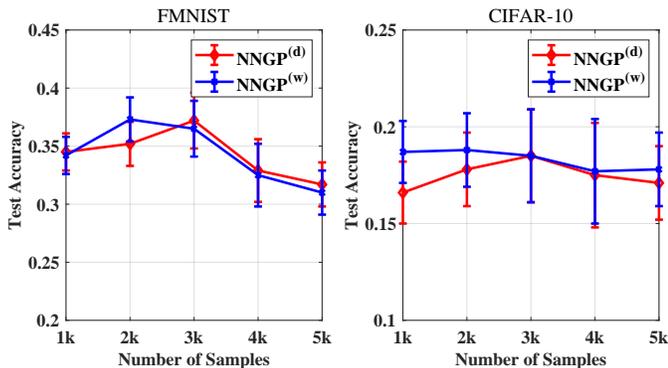


Fig. 1. The performance of the NNGP$^{(d)}$ and NNGP$^{(w)}$ kernels constructed by different number of samples on FMNIST and CIFAR-10.

**The impact of parameter variance and network size.** We construct a synthetic data set to investigate the impact of variance and network size on the model's performance. The task is to use the NNGP$^{(w)}$ and NNGP$^{(d)}$ kernels to fit a function: $f(x) = \sin(x)$ over $[0, \pi]$. A total of 200 data points

are evenly sampled from $[0, \pi]$, from which 100 points are randomly sampled for training and the rest for testing.

Similarly, we employ a one-hidden-layer wide network for computing the NNGP$^{(w)}$ kernel whose width is cast from $\{500, 1000, 2000\}$. In contrast, we use a deep network for the NNGP$^{(d)}$ kernel whose depth is cast from $\{100, 200, 500\}$. The width of the deep network is set to 30, and $\hbar = 1$. No label encoding is needed here because this is not a classification task. For two networks, we take $\tanh$ as the activation function. For the NNGP$^{(w)}$, all weights are initialized with a Gaussian distribution of mean 0 and variance of $\{0.3, 0.5, 0.8\}/n_l$, where $n_l$ is the number of neurons in the $l$-th layer. For NNGP$^{(d)}$, all weights are initialized with a Gaussian distribution of mean 0 and variance of $\{0.2, 0.3, 0.5\}$. The initialization is repeated 200 times to compute the empirical statistics of the NNGP$^{(d)}$ and NNGP$^{(w)}$. We run each model 10 times to count the mean and variance of accuracy. All experiments are conducted on an NVIDIA TITAN Xp GPU.



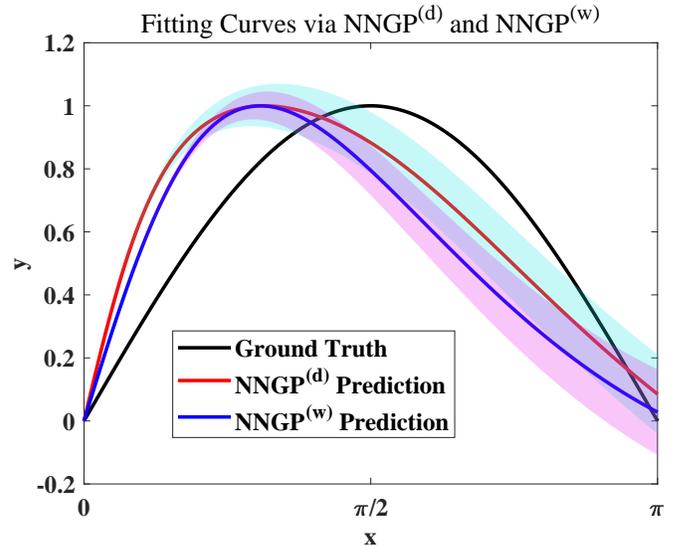Fig. 2. The fitting curves of the NNGP$^{(d)}$ and NNGP$^{(w)}$ kernels for a sine function over $[0, \pi]$.

Table II shows the performance of NNGP$^{(d)}$ and NNGP$^{(w)}$ kernels with respect to different variances and network sizes, from which we draw two highlights. The first is that with the same network size, the NNGP$^{(d)}$ kernel favors a lower variance, while the NNGP$^{(w)}$ kernel is on the contrary. The second one is that increasing the network size may not necessarily give rise to a lower MSE. In fact, it depends on the variance. For the NNGP$^{(d)}$ kernel, when $\sigma = 0.2$, increasing the network depth is beneficial, whereas when $\sigma = 0.3$, increasing the network depth hurts the performance. Figure 2 presents the fitting curves of the NNGP$^{(d)}$ ($\sigma = 0.2$, depth=200) and NNGP$^{(w)}$ ($\sigma = 0.5$, width=1000) kernels for $\sin(x)$, where the curve of NNGP$^{(w)}$ is more accurate in $[0, \pi/2]$ and the curve of NNGP$^{(d)}$ is more accurate in $[\pi/2, \pi]$.

**Computation time.** Here, we also compare the time spent on constructing the NNGP$^{(d)}$ and NNGP$^{(w)}$ kernels relative to different numbers of samples. We sample the data from FMNIST. The network size is 200 for both deep and wide

TABLE II
THE MSE SCORES OF NNGP$^{(d)}$ AND NNGP$^{(w)}$ KERNELS WITH RESPECT TO DIFFERENT VARIANCES AND NETWORK SIZES ON THE SYNTHETIC DATASET.

| NNGP$^{(w)}$ | width=500 | | | width=1000 | | | width=2000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\sigma=0.3$ | $\sigma=0.5$ | $\sigma=0.8$ | $\sigma=0.3$ | $\sigma=0.5$ | $\sigma=0.8$ | $\sigma=0.3$ | $\sigma=0.5$ | $\sigma=0.8$ |
| | 0.1240±0.0279 | 0.1051±0.0317 | 0.1143±0.0208 | 0.1350±0.0243 | 0.1092±0.0438 | 0.1075±0.0389 | 0.1255±0.0158 | 0.0920±0.0444 | 0.1006±0.0275 |
| NNGP$^{(d)}$ | depth=100 | | | depth=200 | | | depth=500 | | |
| | $\sigma=0.2$ | $\sigma=0.3$ | $\sigma=0.5$ | $\sigma=0.2$ | $\sigma=0.3$ | $\sigma=0.5$ | $\sigma=0.2$ | $\sigma=0.3$ | $\sigma=0.5$ |
| | 0.1437±0.0217 | 0.1808±0.0443 | 0.2184±0.0188 | 0.1310±0.0635 | 0.1926±0.0436 | 0.1742±0.0442 | 0.0917±0.0391 | 0.2056±0.0304 | 0.2384±0.1308 |

networks. $\hbar$ is 1 for the deep network. Previously, we repeat the initialization 200 times to compute a kernel. Here, the repetition time is set to 1 for convenience. The experiment is conducted on Intel Core-i7-6500U. As shown in Table III, generally, it is more expensive to construct the NNGP$^{(d)}$ kernel than the NNGP$^{(w)}$ kernel. However, the difference in computation time is no more than $2\times$, as the number of samples increases. This might be because we use the CPU which does not admit parallel acceleration.

TABLE III
THE COMPUTATION TIME IN CONSTRUCTING THE NNGP$^{(d)}$ AND NNGP$^{(w)}$ KERNELS.

| #Sample | NNGP$^{(d)}$ | NNGP$^{(w)}$ |
|---|---|---|
| 1k | 0.011s | 0.057s |
| 2k | 0.173s | 0.083s |
| 3k | 0.228s | 0.160s |
| 4k | 0.286s | 0.226s |
| 5k | 0.444s | 0.342s |

## REFERENCES

[1] Q. Nguyen, M. Mondelli, and G. Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8119–8129, 2021.

[2] S.-Q. Zhang and Z.-H. Zhou. Arise: Aperiodic semi-parametric process for efficient markets without periodogram and gaussianity assumptions. *arXiv:2111.06222*, 2021.

[3] Hector N Salas. Gershgorin's theorem for matrices of operators. *Linear Algebra and its Applications*, 291(1-3):15–36, 1999.