Journal of Machine Learning Research 21 (2020) 1-66 Submitted 1/20; Revised 7/20; Published 9/20

Wide Neural Networks with Bottlenecks are Deep Gaussian Processes

Devanshu Agrawal DAGRAWA2@VOLS.UTK.EDU
The Bredesen Center
University of Tennessee
Knoxville, TN 37996-3394, USA

Theodore Papamarkou PAPAMARKOUT@ORNL.GOV
Computational Sciences and Engineering Division
Oak Ridge National Lab
Oak Ridge, TN 37830-8050, USA

Jacob Hinkle HINKLEJD@ORNL.GOV
Computational Sciences and Engineering Division
Oak Ridge National Lab
Oak Ridge, TN 37830-8050, USA

Editor: Mohammad Emtiyaz Khan

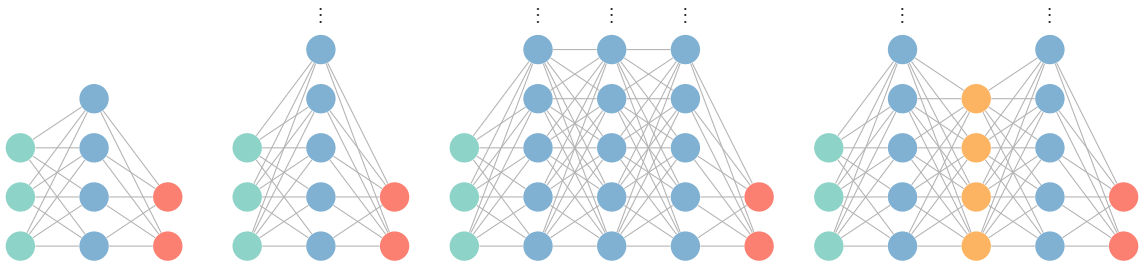
Abstract

There has recently been much work on the "wide limit" of neural networks, where Bayesian neural networks (BNNs) are shown to converge to a Gaussian process (GP) as all hidden layers are sent to infinite width. However, these results do not apply to architectures that require one or more of the hidden layers to remain narrow. In this paper, we consider the wide limit of BNNs where some hidden layers, called "bottlenecks", are held at finite width. The result is a composition of GPs that we term a "bottleneck neural network Gaussian process" (bottleneck NNGP). Although intuitive, the subtlety of the proof is in showing that the wide limit of a composition of networks is in fact the composition of the limiting GPs. We also analyze theoretically a single-bottleneck NNGP, finding that the bottleneck induces dependence between the outputs of a multi-output network that persists through extreme post-bottleneck depths, and prevents the kernel of the network from losing discriminative power at extreme post-bottleneck depths.

Keywords: Bayesian neural networks, deep learning, Gaussian processes, kernels, phase transitions

1. Introduction

Deep neural networks have found great empirical success, achieving state-of-the-art performance on a variety of tasks such as those in computer vision and natural language understanding (Krizhevsky et al., 2012; Antipov et al., 2015; Liang et al., 2017). There is considerable interest in understanding the theoretical aspects of deep neural networks both to establish guarantees on the behavior of these models on certain classes of problems as well as to guide architecture design and optimization. One avenue of pursuit in this endeavor leads to the study of Bayesian neural networks (BNNs), where the parameters



(a) MLP (b) NNGP (c) Deep NNGP (d) Bottleneck NNGP

Figure 1: Depiction of various NNGP architectures with three predictors (green nodes) and two response variables (red nodes). Blue nodes indicate hidden layers, with ellipses indicating layers that increase in width toward infinity. In our bottleneck NNGP model, one or more bottlenecks (finite-width hidden layers with orange nodes) are surrounded by infinite-width hidden layers. In the historical development of NNGP architectures, MLPs (sub-figure (a)) have been succeeded by shallow NNGPs (sub-figure (b), see Neal (1996)), which in turn have been succeeded by deep NNGPs (sub-figure (c), see Lee et al., 2017; Matthews et al., 2018). In our paper, we propose bottleneck NNGPs (sub-figure (d)).

of the network are random variables following some probability distributions. BNNs thus bring the formalism and machinery of probability theory to bear on neural networks.

It is a foundational result that a BNN converges to a Gaussian process (GP) in the "wide limit"—i.e., as the widths of all hidden layers are sent to infinity while the prior distributions on weights are sharpened accordingly (Neal, 1996). The resulting GP is called a "neural network Gaussian process" (NNGP). Although NNGP limits have been derived from various BNN architectures, they cannot be obtained from architectures requiring some hidden layers to remain narrow, such as certain autoencoders. It seems intuitive that the wide limit of a BNN with some hidden layers restricted to finite-width "bottlenecks" is a composition of NNGPs, but until now this claim has not been proven. Such a composition of GPs is called a "deep Gaussian process" (DGP) in the literature (Damianou and Lawrence, 2013). Although DGPs were inspired by the compositional structure of deep neural networks, their connection to BNNs has not been established formally.

In this paper, we give a formal proof of the convergence of BNNs with bottleneck layers to a DGP in the wide limit, where the DGP is a composition of NNGPs. In doing so, we unify the two major approaches to making GPs "deeper"—NNGPs and DGPs, thus allowing NNGPs to be examined in the DGP framework. We will refer to the limiting DGP as a "bottleneck NNGP". Even though the result is intuitive, the proof is nontrivial as it requires us to formally justify that the limit of a composition of BNNs equals the composition of the limiting NNGPs.

In the bottleneck NNGP limit, we consider a sequence of BNNs all having the same architecture except that some hidden layers are growing to infinite width (Fig. 1). We call the hidden layers held to finite widths "bottleneck layers" or simply "bottlenecks", and we call each network in the sequence a "bottleneck BNN". We use the term "component" to refer to any subnetwork that is either (1) between the input layer and the first bottleneck

layer, (2) between two bottleneck layers with no bottlenecks in between, or (3) between the last bottleneck layer and the output layer; each BNN is thus a composition of components, and each component maintains constant input and output dimensions with only its hidden layers growing in width over the sequence of networks.

We know that each sequence of corresponding components converges to an NNGP in the wide limit. It is therefore intuitive to expect that the sequence of bottleneck BNNs (each BNN being a composition of components) converges to the composition of NNGPs—i.e., a bottleneck NNGP. However, this fact is not immediate, and care must be taken to verify that the limit procedure can be exchanged with the composition of components. In particular, we find that this exchangeability holds if each post-bottleneck component converges to an NNGP with a sufficient amount of uniformity with respect to its inputs.

We demonstrate the utility of bottleneck NNGPs and their link to no-bottleneck NNGPs empirically, showing that restricting a hidden layer of an NNGP to a bottleneck can boost its model likelihood on three example datasets[1].

We also characterize the effect of a bottleneck layer theoretically by analyzing an example multi-output single-bottleneck NNGP with rectified linear unit (ReLU) activation. We find that the bottleneck induces dependence between distinct response variables and derive a closed-form expression for the correlation between the squares (i.e., quadratic correlation) of the response variables. We show that in the deep post-bottleneck limit (infinitely many infinite-width hidden layers after the bottleneck), the quadratic correlation tends to 0 when the network is in the "disordered phase"—so that response variables decouple—but remains a nontrivial function of the inputs in the "ordered phase"—so that information about the inputs can be recovered. We identify the prior variance of the network weights as the order parameter responsible for the phase transition.

Similarly, in the deep post-bottleneck limit, we obtain a closed-form expression for the quadratic correlation of outputs of a single response variable given two inputs. We find that the quadratic correlation is 100% in the "disordered phase"—so that the network has lost all discriminative power at infinite depth—but is surprisingly a nontrivial function of the inputs in the "ordered phase". This behavior in the ordered phase stands in stark contrast to that of no-bottleneck NNGPs and indicates that bottleneck layers are essential for a very deep network to maintain discriminative power.

2. Preliminaries

In this section, we review prior work on DGPs and NNGPs to contextualize and motivate the bottleneck NNGP model. We also review the main theorem of Matthews et al., 2018 in Sec. 2.3, introducing notation that will be essential to stating our main result in Sec. 3.1.

2.1 Deep Gaussian processes

Compositions of GPs are known as deep Gaussian processes (DGPs) in the literature and were originally motivated by the success of deep neural networks and the hope to obtain similar success on small data sets where Bayesian methods generally shine (Damianou and Lawrence, 2013). DGPs have indeed been shown to outperform shallow GPs on a vari-

1. Code for our simulations and experiments is available at `https://code.ornl.gov/d0a/bottleneck_nngp`.

ety of regression and classification tasks (Damianou and Lawrence, 2013; Salimbeni and Deisenroth, 2017). Damianou and Lawrence (2013) implement DGPs in a sparse variational inducing points framework based on the work of Titsias (2009) in order to simplify the composition of GPs to a set of separate but coupled GPs, but their implementation is restricted to small data sets with only a few hundred entries. Much of the DGP literature has therefore been dedicated to developing more efficient and scalable implementations (Hensman and Lawrence, 2014; Dai et al., 2015; Bui et al., 2016; Wang et al., 2016; Salimbeni and Deisenroth, 2017). Salimbeni and Deisenroth (2017) in particular show that DGPs can be put into a stochastic variational framework as in Hensman et al. (2013), allowing the models to be applied to much larger data sets.

DGPs offer the additional advantage that they can capture correlation between multiple outputs (Wang et al., 2016). In contrast, the outputs of a shallow multi-output GP are by default independent, which can limit predictive performance for multi-output problems. Although methods have been proposed to model correlation in a shallow multi-output GP, such as through linear mixing of latent outputs (Bonilla et al., 2008; Alvarez and Lawrence, 2009), DGPs capture correlation naturally through shared feature representations in the latent "bottleneck space"—similar to the approach taken with multi-task deep neural networks (Ruder, 2017). DGPs have therefore been applied to problems that can benefit from modeling correlation between multiple outputs, such as multi-task regression (Alaa and van der Schaar, 2017) and tasks involving partially observed multivariate outputs—i.e., missing values (Wang et al., 2016).

The mechanism by which the outputs of a DGP are made dependent predates the DGP model itself, as it was first introduced in the context of the Bayesian Gaussian process latent variable model (GP-LVM), where a Gaussian prior is placed on the latent inputs of a GP (Lawrence, 2004; Titsias and Lawrence, 2010); for regression, each input is concatenated with such a latent random variable before it is fed into the GP (Dutordoir et al., 2018). The outputs then become dependent through their dependence on the common set of latent random variable inputs, which are analogous to the bottleneck activations of a DGP. However, it still remains quantitatively unclear how the introduction of a bottleneck layer—or in the case of an NNGP, the restriction of a layer to finite width—induces correlation between multiple outputs under the prior and how this translates to correlation under the posterior. Moreover, although this mechanism is well-established for the GP-LVM and DGP models, it is conspicuously absent in the NNGP literature and thus its implications for NNGPs are not fully understood.

The DGP prior has been studied by Lu et al. (2019), who show that for a single-bottleneck DGP with a single response variable, the prior has heavy tails, in contrast to shallow GPs. Their calculation of the prior kurtosis is similar to that of the quadratic correlation between distinct response variables of a multi-output DGP, but this connection is not discussed. Moreover, they only consider a bottleneck of width one and primarily focus on stationary kernels that do not arise from NNGP limits with common activation functions.

There is considerable interest in understanding the "deep limit" of DGPs—i.e. when arbitrarily many GPs are composed together. Duvenaud et al. (2014) and Dunlop et al. (2018) show that DGPs with a certain class of kernels have trivial, pathological, or convergent deep limits, meaning that increasing the depth of a DGP beyond some point is either detrimental

to performance or diminishingly beneficial. However, they do not consider NNGP kernels and thus do not analyze deep limits of architectures with both bottlenecks and infinitely many infinite-width hidden layers.

Although DGPs were inspired by deep neural networks, there is little literature concretely establishing their connection. Duvenaud et al. (2014) discuss the connection between DGPs and neural networks at a high level to motivate studying the deep limit of DGPs with radial basis function (RBF) kernels, but the implication for neural networks is not treated formally. Gal and Ghahramani (2015) consider a DGP where the kernel of each GP layer is an integral as in Williams (1997). They show that a Monte Carlo estimation of the kernels leads to a BNN approximation of the DGP, where the width of a hidden layer corresponds to the size of the Monte Carlo sample. However, they do not formally verify convergence of a BNN to a DGP in the limit of infinite width. Moreover, their bottleneck layers have no activation function and are not scaled to allow an NNGP to be recovered as the bottlenecks are sent to infinite width.

2.2 Wide neural networks as GPs

A foundational result in the study of BNNs came when Neal (1996) showed that a BNN with one hidden layer converges to a GP in the "wide limit"—i.e., as the number of hidden neurons is sent to infinity. Shortly after, Williams (1997) derived analytic expressions for the kernel of the GPs corresponding to neural networks with sigmoidal and Gaussian hidden units. These works connected neural networks to the world of Bayesian nonparametrics and kernel methods and thus offered a new perspective to interrogate and probe the behavior of neural networks. In particular, while training neural networks is challenging since it requires the optimization of highly non-convex objective functions, GPs are nonparametric models that admit exact Bayesian inference, where the predictive posterior distribution can be written in closed form (Rasmussen and Williams, 2006).

Since the works of Neal (1996); Williams (1997), new insights into BNNs have steadily emerged. Cho and Saul (2009) interpreted a BNN as a feature embedding map and derived the equations for the propagation of a kernel through the layers of a deep neural network with rectified polynomial unit activations. Subsequent works built upon these findings to elucidate key theoretical aspects of neural networks including expressivity (Poole et al., 2016), generalization power (Hazan and Jaakkola, 2015), initialization (Daniely et al., 2016), and trainability (Schoenholz et al., 2016). More recently, the original result by Neal (1996) has been extended to deep architectures by showing that a deep BNN converges to a GP as the widths of all hidden layers are simultaneously sent to infinity (Lee et al., 2017; Matthews et al., 2018). We refer to GPs that arise from such a limit as "neural network Gaussian processes" (NNGPs). As the work of Matthews et al., 2018 illustrates, this extension is nontrivial; the proof by Neal (1996) relies on the Central Limit Theorem, but the assumption of independent and identically distributed (IID) random variables necessary for the Central Limit Theorem does not hold for deep architectures. The proof by Matthews et al., 2018 for deep architectures is instead based on a more exotic central limit theorem as given in Blum et al. (1958).

Since the extension of the NNGP limit to deep architectures, there have been a number of works establishing and analyzing analogous wide-limit results for more modern neural

network architectures that are used in practice today. These include convolutional neural networks (Garriga-Alonso et al., 2019; Novak et al., 2019), weight-tied autoencoders (Li and Nguyen, 2019), and most generally any network that can be represented as a "tensor program"—including recurrent neural networks and attention networks among others (Yang, 2019). Alongside these works, new insights into the trainability and generalization power of neural networks have continued to emerge, based on the tractable learning dynamics of neural networks in the wide limit (Jacot et al., 2018; Lee et al., 2019; Arora et al., 2019).

One application of the NNGP limit that is of particular note is that it can make the analysis of the "deep limit"—i.e., as the number of hidden layers is sent to infinity—tractable (Poole et al., 2016; Schoenholz et al., 2016; Yang and Schoenholz, 2017; Lee et al., 2017). Poole et al. (2016) and Schoenholz et al. (2016) show that the correlation between two inputs transformed through an NNGP with sigmoidal activation function has a fixed point at 100% in the deep limit that transitions from stable to unstable (i.e., ordered to chaotic) when the variances of the Gaussian weights and biases cross a certain phase boundary; the network is shown to be highly expressive in the chaotic phase and optimally trainable near the phase boundary. In contrast, for networks with rectified linear unit (ReLU) activations, the correlation between transformed inputs has a stable fixed point at 100% regardless of the weight variance, implying that an NNGP with ReLU activation has no discriminative power at infinite depth (Lee et al., 2017).

The works described above all consider BNNs in the wide limit, and thus the results and insights therein do not apply to neural network architectures that require one or more finite width or "bottleneck" layers. One of the most important classes of neural networks that frequently require bottleneck layers is that of autoencoders (Hinton and Salakhutdinov, 2006; Kingma and Welling, 2014). Another example is neural networks with a word embedding layer, which is currently key to the successful application of neural networks to natural language understanding (Mikolov et al., 2013). Both word embedding layers and many autoencoder models aim to find dense feature representations and therefore depend on low-dimensional spaces. Even for neural network architectures that are not directly meant for dense representation learning, it has still been argued and demonstrated that bottleneck layers perform data compression and therefore help to boost generalization power (Tishby and Zaslavsky, 2015). Particularly for fully-connected architectures, which is what we consider in this work, it has been shown that the insertion of linear bottleneck layers between two linear ReLU layers boosts predictive performance by reducing sparsity and improving gradient flow (Lin et al., 2015). This prompts the question: How can insights based on very wide BNNs be generalized to networks in which one or more hidden layers are held fixed to a finite width? The first step in addressing this question is to understand what happens if we let all but finitely many hidden layers of a BNN grow to infinite width. We call these finite-width hidden layers "bottleneck layers". It is already established that the component networks between consecutive bottleneck layers converge to GPs, and thus we intuitively expect a BNN with bottleneck layers to converge in the wide limit to a composition of GPs. In Sec. 3.1, we formally verify that this is the case.

2.3 The no-bottleneck NNGP limit

The bottleneck NNGP limit is a generalization of the (no-bottleneck) NNGP result proved by Matthews et al., 2018. Moreover, one component of our proof is verifying that BNNs converge in distribution uniformly on compact sets, and our approach to proving this closely follows the proof of Matthews et al., 2018. In this section, we state the NNGP limit result by Matthews et al., 2018, which also allows us to introduce key concepts and notation along the way.

We consider a traditional fully-connected network mapping $\mathbb{R}^M$ to $\mathbb{R}^L$ with D hidden layers and nonlinearity ϕ. Let H_μ be the width of the μ-th hidden layer. The propagation of an input x through the network is then governed by a recursion with initial step

$$f_i^{(1)}(x) = b_i^{(1)} + \sum_{j=1}^{M} w_{ij}^{(1)} x_j, \tag{1}$$

and for $\mu \in \{1, \ldots, D\}$,

$$g_i^{(\mu)}(x) = \phi[f_i^{(\mu)}(x)], \tag{2}$$

$$f_i^{(\mu+1)}(x) = b_i^{(\mu+1)} + \sum_{j=1}^{H_\mu} w_{ij}^{(\mu+1)} g_j^{(\mu)}(x). \tag{3}$$

In Eq. (2), i ranges from 1 to H_μ. In Eq. (3), i ranges from 1 to $H_{\mu+1}$ for $\mu = 1, \ldots, D-1$, and from 1 to L for $\mu = D$. We refer to $f^{(\mu)}(x)$ and the $g^{(\mu)}(x)$ as the preactivations into and activations out of the μ-th hidden layer, respectively. The top-most preactivations $f^{(D+1)}(x)$ are the outputs of the network.

We require mild assumptions on the nonlinearity ϕ for our main theorem to hold; these are the same assumptions made by Matthews et al., 2018, namely the linear envelope condition.

Definition 1 (Linear envelope condition) *A nonlinearity $\phi : \mathbb{R} \mapsto \mathbb{R}$ is said to satisfy the linear envelope condition if it is continuous and there exist positive constants C and M such that*

$$|\phi(x)| < C + M|x| \text{ for all } x \in \mathbb{R}.$$

Many popular activation functions such as tanh, ReLU, and leaky ReLU satisfy the linear envelope condition, and thus our result is quite general with regards to the choice of nonlinearity.

We now turn the above network into a random network by placing IID normal distributions on the weights $w^{(\mu)}$ and biases $b^{(\mu)}$ of the network:

$$w_{ij}^{(\mu)} \sim \mathcal{N}\left(0, \frac{v_w^{(\mu)}}{H_{\mu-1}}\right), \tag{4}$$

$$b_i^{(\mu)} \sim \mathcal{N}(0, v_b^{(\mu)}), \tag{5}$$

where we set $H_0 = 1$ for the purpose of defining these distributions. The variance of the weights after the first layer are scaled inversely to the preceding hidden layer width so that

the Central Limit Theorem can be applied to the convergence of BNNs to a GP. With a slight abuse of terminology, we will call the constants $v_w^{(\mu)}$ and $v_b^{(\mu)}$ "weight and bias variance hyperparameters" even though $v_w^{(\mu)}$ is not the actual variance of the weights.

The output $f^{(D+1)}(x)$ is now a random vector of dimension L for each input x, and we therefore understand a BNN as an instance of a stochastic process. We give a formal definition next, after we introduce some notation. If Ω is a probability space and E is a measurable space, then an E-valued stochastic process F with index set X is a function $F : X \times \Omega \mapsto E$ such that $F(x, \cdot)$ is a measurable function for each $x \in X$. By the notation $F(x)$, we refer to the random variable $F(x) : \Omega \mapsto E$ defined by $F(x)(\omega) = F(x, \omega)$.

Definition 2 (Bayesian neural network) *A Bayesian neural network (BNN) F mapping $\mathbb{R}^M$ to $\mathbb{R}^L$ with D hidden layers of widths H_μ, $\mu \in \{1, \ldots, D\}$, and nonlinearity $\phi : \mathbb{R} \mapsto \mathbb{R}$ is a stochastic process $F : \mathbb{R}^M \times \Omega \mapsto \mathbb{R}^L$ defined such that $F(x) = f^{(D+1)}(x)$, where $f^{(D+1)}$ is the neural network output defined through the recursion of preactivations and activations of Eqs.* (2)-(3).

Matthews et al., 2018 prove the following theorem concerning the convergence of BNNs with no bottleneck layers.

Theorem 3 (NNGP theorem, Matthews et al., 2018) *Let $\{F[n]\}_{n=1}^\infty$ be a sequence of BNNs mapping $\mathbb{R}^M$ to $\mathbb{R}^L$ each with D hidden layers of widths $H_\mu[n]$, $\mu \in \{1, \ldots, D\}$, and nonlinearity $\phi : \mathbb{R} \mapsto \mathbb{R}$ that satisfies the linear envelope condition. If $H_\mu[n]$ is strictly increasing in n for each μ, then $\{F[n]\}_{n=1}^\infty$ restricted to a countable index set $\mathcal{X} \subset \mathbb{R}^M$ converges in distribution to the Gaussian process $\mathcal{GP}(0, K)$, where K is a kernel defined recursively by Eqs.* (6)-(7).

Note that we use the suffix $[n]$ instead of a subscript to index a sequence of stochastic processes. Convergence in distribution is defined in the measurable space $((\mathbb{R}^L)^\infty, \mathcal{A})$ of $\mathbb{R}^L$-valued sequences; details are provided in Appendix A. The limiting GP in Thm. 3 is called a neural network Gaussian process (NNGP). If $f_i^{(\mu)}$ is the limiting NNGP of $\{f_i^{(\mu)}[n]\}_{n=1}^\infty$, then the NNGP kernel is defined through a recursion with initial step

$$K_{ij}^{(1)}(x_1, x_2) = \mathrm{E}[f_i^{(1)}(x_1) f_j^{(1)}(x_2)] = \delta_{ij}(v_b^{(1)} + v_w^{(1)} x_1 \cdot x_2), \tag{6}$$

and for $\mu \in \{1, \ldots, D\}$,

$$\begin{aligned} K_{ij}^{(\mu+1)}(x_1, x_2) &= \mathrm{E}[f_i^{(\mu+1)}(x_1) f_j^{(\mu+1)}(x_2)] \\ &= \delta_{ij} \left(v_b^{(\mu+1)} + v_w^{(\mu+1)} \, \mathrm{E}_{z_1, z_2 \sim \mathcal{N}(0, C^{(\mu)})}[\phi(z_1)\phi(z_2)] \right), \end{aligned} \tag{7}$$

where $C^{(\mu)}$ is the 2×2 matrix with entries $c_{ab}^{(\mu)} = K_{11}^{(\mu)}(x_a, x_b)$; here we could have used $K_{ii}^{(\mu)}$ in place of $K_{11}^{(\mu)}$ for any i since the NNGP preactivations $f_i^{(\mu)}(x)$ into the μ-th hidden layer are IID over i. The (countably infinite) kernel matrix $K^{(\mu)}(\mathcal{X}, \mathcal{X})$ is therefore block-diagonal with the (i, j)-th block $K_{ij}^{(\mu)}(\mathcal{X}, \mathcal{X})$.

The kernel $K : \mathbb{R}^M \times \mathbb{R}^M \mapsto \mathbb{R}^{L \times L}$ in Thm. 3 is given by $K = K^{(D+1)}$. Observe that the L outputs of the BNNs converge to IID GPs so that all correlations between the outputs of the networks are lost in the infinite width limit. We will see that bottleneck layers help to preserve some correlations between outputs.

3. The bottleneck NNGP theorem

In this section, we state and prove the bottleneck NNGP theorem, we show that a single-bottleneck NNGP approximates a no-bottleneck NNGP as the bottleneck width is increased, and we explore the effect of depth and width on bottleneck NNGPs using three example datasets.

3.1 Statement of main theorem

We now state our main theorem, which is a direct generalization of Thm. 3 to compositions of BNNs. Given two stochastic processes $F^{(1)} : X \times \Omega^{(1)} \mapsto Y$ and $F^{(2)} : Y \times \Omega^{(2)} \mapsto Z$, we define the composition $F^{(2)} \circ F^{(1)}$ as the stochastic process $F^{(2)} \circ F^{(1)} : X \times (\Omega^{(1)} \times \Omega^{(2)}) \mapsto Z$ with

$$(F^{(2)} \circ F^{(1)})(x, (\omega_1, \omega_2)) = F^{(2)}(F^{(1)}(x, \omega_1), \omega_2).$$

Theorem 4 (Bottleneck NNGP theorem) *Let $\{B_d \in \mathbb{N}\}_{d=0}^{D}$ for $D \in \mathbb{N}$ with $B_0 = M$ and $B_D = L$. For each $d \in \{1, \ldots, D\}$, let $\{F^{(d)}[n]\}_{n=1}^{\infty}$ be a sequence of BNNs mapping $\mathbb{R}^{B_{d-1}}$ to $\mathbb{R}^{B_d}$ with D_d hidden layers of widths $H_\mu^{(d)}[n]$, $\mu \in \{1, \ldots, D_d\}$, and nonlinearity ϕ that satisfies the linear envelope condition. If $H_\mu^{(d)}[n]$ is strictly increasing in n for each $d \in \{1, \ldots, D\}$ and $\mu \in \{1, \ldots, D_d\}$, then the sequence of bottleneck random neural networks $\{F^{(D)}[n] \circ \cdots \circ F^{(1)[n]}\}_{n=1}^{\infty}$ restricted to a countable index set $\mathcal{X} \in \mathbb{R}^M$ converges in distribution in $((\mathbb{R}^L)^\infty, \mathcal{A})$ to $F^{(D)} \circ \cdots \circ F^{(1)}$, where $F^{(d)}$ is the limiting NNGP of $\{F^{(d)}[n]\}_{n=1}^{\infty}$.*

Remark 5 (Nonlinear bottleneck) *Theorem 4 as stated above assumes no nonlinearity on the bottleneck layers. However, the theorem also holds when we replace $F^{(d)}[n]$ and $F^{(d)}$ with $F^{(d)}[n] \circ \left(\frac{1}{\sqrt{B_{d-1}}}\phi\right)$ and $F^{(d)} \circ \left(\frac{1}{\sqrt{B_{d-1}}}\phi\right)$ respectively for $d \in \{2, \ldots, D\}$, that is when we scale the weights after each bottleneck layer by layer width in the same way as all other weights after a hidden layer and and we place nonlinearities ϕ on the bottleneck layers. The proof is nearly identical to the proof of Thm. 4 (see Remark 28 in Appendix B.2 for details).*

Remark 6 (Discontinuous nonlinearity) *Theorem 4 holds even if the nonlinearity $\phi : \mathbb{R} \mapsto \mathbb{R}$ is continuous only almost everywhere, as long as ϕ is continuous at 0 or $v_b > 0$. Each of these two conditions ensures that the Continuous Mapping Theorem is still applicable in Lemmas 31-33 (see Remark 29 in Appendix B.2 for details). This extends the class of allowable nonlinearities to include such prominent examples as the Heaviside step function used in the first perceptron model (Rosenblat, 1958).*

Remark 7 (Converse of the main theorem) *The converse of Thm. 4—that every DGP is the bottleneck NNGP limit of a BNN with IID priors and a nonlinearity satisfying the linear envelope condition—does not hold. A simple counterexample is a no-bottleneck single-hidden-layer NNGP with the rectified polynomial unit activation $\phi(x) = \max(0, x)^n$ for $n \geq 2$ (Cho and Saul, 2009); this is a GP that can only result from a wide limit if the linear envelope condition is violated. A more trivial counterexample is any GP with a linear kernel $k(x_1, x_2) = x_1^\top G x_2$ where G is a symmetric positive-semidefinite matrix not proportional to the identity matrix. By Eq. (6), the NNGP kernel depends on its inputs only through their dot product and is thus invariant under rotations. The kernel k with metric G can therefore only arise from a wide limit if the "IID priors" condition is violated.*

Here we consider a sequence of BNNs with $D-1$ bottleneck layers of widths $B_1, \ldots, B_{D-1}$. As all hidden layers except the bottleneck layers tend to infinite width, each component network converges to an NNGP by Thm. 3, but it is less obvious that the composition of components tends to the composition of the limiting NNGPs. Our proof depends on several original lemmas (Lemmas 24-27 in Appendix B.2). However, Lemma 27 is a simple generalization of Lemma 12 in Matthews et al., 2018, and its proof therefore runs in parallel to that in Matthews et al., 2018. The complete proof of the main theorem as well as proofs for all supporting lemmas can be found in Appendix B; we recommend readers to start at the introduction of Appendix B, where we provide a detailed sketch of the proof and discuss the high-level function of each lemma. Definitions and properties of the various modes of convergence of stochastic processes pertinent to the proof are discussed in Appendix A.

3.2 Correspondence to the no-bottleneck NNGP

We expect that in the limit as bottlenecks are sent to infinite width, the bottleneck NNGP converges to the (no-bottleneck) NNGP with the same number of hidden layers. The next theorem gives this result for the case of a bottleneck NNGP with one bottleneck layer.

Theorem 8 (Wide bottleneck correspondence) *Let $\{F^{(H)}\}_{H=1}^{\infty}$ be a sequence of single-bottleneck NNGPs mapping $\mathbb{R}^M$ to $\mathbb{R}^L$ with D_1 hidden layers and D_2 hidden layers before and after the bottleneck of width H and with nonlinearity $\phi : \mathbb{R} \mapsto \mathbb{R}$ that satisfies the linear envelope condition. Suppose the nonlinearity ϕ is also applied to the bottleneck and that its activations are scaled by $\frac{1}{\sqrt{H}}$, in accordance with Remark 5. Suppose also that IID Gaussian noise $\mathcal{N}(0, v_n)$ is added to the networks for any $v_n > 0$. Then*

(a) $\{F^{(H)}\}_{H=1}^{\infty}$ restricted to a countable index set $\mathcal{X} \subset \mathbb{R}^M$ converges in distribution in $((\mathbb{R}^L)^{\infty}, \mathcal{A})$ to an NNGP F with $D_1 + D_2 + 1$ hidden layers and Gaussian noise $\mathcal{N}(0, v_n)$.

(b) For every finite set of inputs $X \subset \mathbb{R}^M$, the sequence of probability density functions (PDFs) of $\{F^{(H)}\}_{H=1}^{\infty}$ converges pointwise to the PDF of the NNGP F.

Remark 9 *Statement (a) of Thm. 8 holds even if there is no additive Gaussian noise ($v_n = 0$); the proof uses the technique in the proof of Lemma 24, where the function $X \to \Pr(F(X) \in U)$ is shown to be continuous for an NNGP F and a continuity set U.*

The proof of Thm. 8 is given in Appendix C. It is based on the observation that the activations in the bottleneck layer of a single-bottleneck NNGP are IID. Since the post-bottleneck NNGP depends on these activations through their Gram matrix only and since the activations are inversely scaled by the bottleneck width, then the post-bottleneck NNGP is a function of the sample covariance of bottleneck activations, which converges to the pre-bottleneck NNGP kernel in the limit of infinite bottleneck width by the Law of Large Numbers. Extending this result to the case of multiple bottlenecks is left for future work.

3.3 Experiments

Statement (b) of Thm. 8 implies that the marginal log-likelihood (MLL) of a single-bottleneck NNGP architecture given data (X, Y) and fixed variance hyperparameters v_n, v_b, v_w converges to the MLL of the corresponding NNGP as the bottleneck is sent to infinite width.

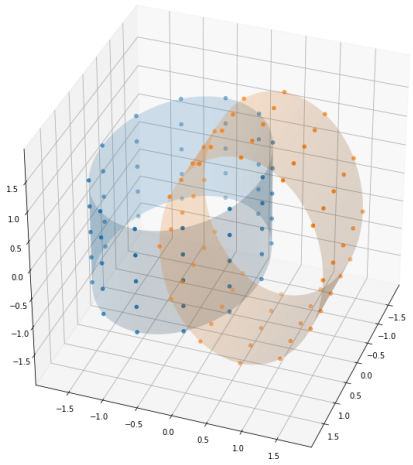


Figure 2: Visualization of the simulated Rings dataset. We take 60 regularly spaced points from each ring. We assign a binary label to each point based on the ring to which it belongs.

The MLL of the bottleneck NNGP is just the logarithm of the PDF given in Eq. (41) evaluated at the dataset. This formally validates the intuition that a bottleneck NNGP with a sufficiently wide bottleneck is a similar model to a no-bottleneck NNGP. However, the utility of bottleneck NNGPs with narrower bottlenecks as measured by MLL is less clear. We investigate this question on a simulated dataset that we call Rings and on two publicly available datasets—Fisher's Iris data set (Anderson, 1935; Fisher, 1936) and the US Census Boston housing prices dataset (Harrison and Rubinfeld, 1978).

The Rings dataset consists of 120 points lying on two interlocked cylindrical bands or "rings" orthogonal to one another and passing through one another's centers (Fig. 2). We took a regular 12×5 lattice of 60 points in $[0, 2\pi) \times \left[-\frac{1}{2}, \frac{1}{2}\right]$ and mapped it onto one of the rings in $\mathbb{R}^3$ by

$$(\theta, z) \mapsto (\cos\theta, \sin\theta, z).$$

To generate the second ring, we rotated a copy of the first by 90° in the xz-plane and translated it by 1 along the y-axis. We assigned a label of 0 to the 60 points on the first ring and a label of 1 to the 60 points on the second. The Rings dataset is therefore a non-linearly separable binary classification problem where the dimension of the data manifold is less than that of the linear span of the data points.

We also considered the Iris and Boston House-Prices datasets. Like Rings, Iris is a classification problem, but we one-hot encoded its labels to include a multivariate dataset with strongly correlated labels. For simplicity, following Lee et al. (2017), we implemented all three problems as regression tasks. We standardized both the input and target sets of all three datasets to help place the three problems on similar scales, as well as to set a reasonable scale for the variance hyperparameters v_b and v_w.

We used a single-bottleneck NNGP with one infinitely wide hidden layer before the bottleneck, bottleneck width H, and post-bottleneck depth D (i.e., D infinitely wide hidden layers after the bottleneck) for various values of H and D. We equipped all hidden neurons

with the normalized ReLU activation

$$\phi(x) = \sqrt{2}\max(0, x). \tag{8}$$

The propagation of the NNGP kernel through the hidden layers given in Eq. (7) admits a closed form for the normalized ReLU activation and is given by (Cho and Saul, 2009):

$$K_{ij}^{(\mu+1)}(x_1, x_2) = \delta_{ij}\left[v_b + v_w \cdot \frac{1}{\pi}\sqrt{K_{11}^{(\mu)}(x_1, x_1)K_{11}^{(\mu)}(x_2, x_2)}J_1(\theta^{(\mu)})\right], \tag{9}$$

$$\theta^{(\mu)} = \cos^{-1}\left[\frac{K_{11}^{(\mu)}(x_1, x_2)}{\sqrt{K_{11}^{(\mu)}(x_1, x_1)K_{11}^{(\mu)}(x_2, x_2)}}\right],$$

where the function J_1 is defined as

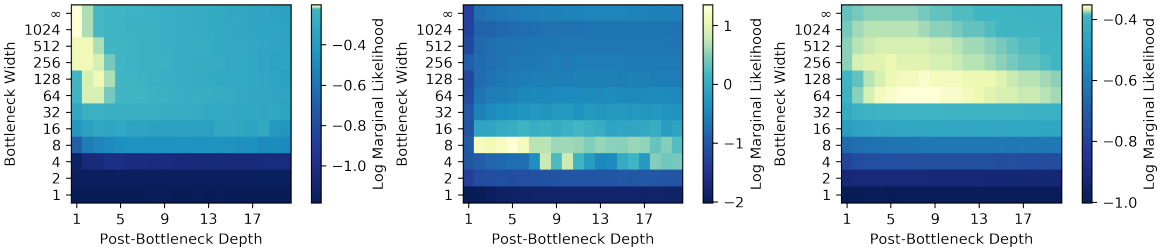
$$J_1(\theta) = \sin\theta + (\pi - \theta)\cos\theta.$$

Our goal is to find the bottleneck NNGP architecture (i.e., combination of bottleneck width H and post-bottleneck depth D) with the greatest likelihood given a dataset (X, Y) of N observations. We calculated the MLL of each bottleneck NNGP architecture (H, D) using

$$\text{MLL}(H, D; X, Y) = \log p(Y; X, H, D, v_{b*}(H, D), v_{w*}(H, D), v_{n*}(H, D)), \tag{10}$$

where on the right-hand side, p is the PDF in Eq. (41) of the data outputs given the data inputs and network architecture, and where the variance hyperparameters are set to their maximum likelihood estimates v_{b*}, v_{w*} and v_{n*}. We found the optimal variance hyperparameters iteratively through gradient descent. During the forward pass through the network in each iteration, we estimated the integral in Eq. (41) by drawing 100 IID Monte Carlo (MC) samples—each an $N \times H$ matrix with IID columns—from the pre-bottleneck NNGP. We did so using the local reparameterization trick (Kingma et al., 2015), so that each sample is a transformation of a draw from the $(N \times H)$-dimensional standard normal distribution. We used the Adam optimizer (Kingma and Ba, 2014) to take advantage of the gradient noise generated by MC sampling during optimization; we set the initial learning rate to 0.1. In order to ensure that the noise observed in the learning curves was due only to MC sampling and not due to a large learning rate, we decayed the learning rate as follows: After the backward pass of each iteration, we re-evaluated the MLL using the same draw from the $(N \times H)$-dimensional standard normal distribution for the MC samples; if the new MLL was less than the value obtained from the initial forward pass of the iteration, then we multiplied the learning rate by 0.9. We iterated the optimization procedure until convergence of the MLL learning curves; once complete, we evaluated Eq. (10) once more—this time with 1000 MC samples—to obtain the final MLL estimate for each network architecture.

On all three datasets, the maximum MLL is attained at a finite bottleneck width and post-bottleneck depth ($H_* = 1024$ and $D_* = 1$ for Rings; $H_* = 8$ and $D_* = 5$ for Iris; $H_* = 64$ and $D_* = 7$ for Boston), thus demonstrating the utility of bottleneck layers in NNGP models (Fig. 3). On Rings and Boston, we also observe that the optimal post-bottleneck depth conditional on a bottleneck width roughly decreases as the bottleneck width increases.



(a) Rings (b) Iris (c) Boston

Figure 3: Marginal log-likelihoods (MLL) of three datasets normalized by number of observations (data points) under a bottleneck NNGP for various bottleneck widths and post-bottleneck depths. Infinite bottleneck width corresponds to the limiting no-bottleneck NNGP. On all three datasets, the maximum MLL is attained at some finite bottleneck width and post-bottleneck depth.

Since the no-bottleneck NNGP kernel with ReLU activation is known to degenerate to a constant kernel with no discriminative power (Lee et al., 2017), it makes sense that a deeper network may require a narrower bottleneck to help information propagate through the network. Although not conclusive from the figures alone, this observation at least warrants further investigation. In Sec. 4, we do just that and find that when the variance hyperparameters are fixed, then the bottleneck width and post-bottleneck depth are indeed intimately related.

4. Bottleneck layers induce dependence

In Sec. 3.3, we showed empirically that the model likelihood of an NNGP is improved if one of the hidden layers is restricted to a finite-width bottleneck and speculated that the optimal post-bottleneck depth may increase as the bottleneck is narrowed. In this section, we investigate possible mechanisms underlying these observed trends in model performance. We start by showing that although linear correlations between distinct response variables or "output neurons" of an NNGP remain zero even in the presence of bottlenecks, the corresponding *quadratic* correlations are often non-zero. We also analyze the behavior of this quadratic correlation in the deep post-bottleneck limit—i.e., as the post-bottleneck depth is sent to infinity. This deep limit is distinct from the ones typically considered for DGPs, where the number of GP components is sent to infinity, and for NNGPs, where the depth of a single GP component is sent to infinity. Proposition 14 provides a striking result, which implies that bottleneck layers help a network retain discriminative power even at extreme post-bottleneck depths.

Note that in this section, we primarily consider the ReLU activation defined in Eq. (8) as it is by far the most common nonlinearity used in deep learning today. In Sec. 4.5, we briefly consider other nonlinearities and contrast their deep limit behaviors with that of the ReLU activation, thereby highlighting the peculiarities of the ReLU activation.

4.1 An exact formula for quadratic correlation

The outputs of a multi-output GP prior are IID, and it follows that the outputs of the corresponding posterior remain independent (though not necessarily identically distributed). This is a limitation of GPs for multi-task learning applications, since information cannot be shared across tasks. One method to solve this problem was proposed by Bonilla et al. (2008) who introduce a coupling matrix hyperparameter through which distinct output neurons or tasks can interact. This method, if applied to finite-width neural networks, would be superfluous since tasks could share information through a common set of features learned in the earlier layers of the network.

The key to correlating tasks in neural networks is clearly not depth alone, since the outputs of a (no-bottleneck) NNGP—however deep—are independent. Rather, following from the DGP framework as discussed in Sec. 2.1, the outputs of an NNGP become correlated if bottleneck layers are introduced, so that bottleneck NNGPs support multi-task learning out of the box. Correlation arising from finite-width bottleneck layers is exactly the type of behavior we expect in neural network architectures such as word embedding layers and many kinds of autoencoders, where the bottleneck width forces dense feature representations (i.e., feature representations that capture correlation) to be learned. The correlation structure that is induced in an NNGP prior through bottleneck layers is, however, subtle; distinct outputs of a bottleneck NNGP prior remain linearly uncorrelated (i.e., have zero covariance) but can be quadratically correlated (i.e., the squares of the outputs have non-zero covariance).

The expression for the quadratic correlation of outputs in a single-bottleneck NNGP prior can be obtained in closed form. Consider a bottleneck NNGP $F : \mathbb{R}^M \times \Omega \mapsto \mathbb{R}^2$ with one bottleneck layer of width H, any number of infinitely wide hidden layers before the bottleneck, and $D-1$ infinitely wide hidden layers after the bottleneck. Suppose all hidden neurons (including in the bottleneck) are equipped with the normalized ReLU activation defined in Eq. (8). We also scale the bottleneck activations by $\frac{1}{H}$ in accordance with Remark 5. Let v_b and v_w be the bias variance and weight variance hyperparameters of both the pre-bottleneck and post-bottleneck components of the bottleneck NNGP defined as in Eqs. (4)-(5). Suppose the bottleneck NNGP is fed two inputs $x_1, x_2 \in \mathbb{R}^M$. Then the preactivations into the H neurons in the bottleneck layer are IID with common 2D normal distribution $\mathcal{N}(0, C)$ for some covariance matrix

$$C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}.$$

As is commonly done with DGPs (Damianou and Lawrence, 2013; Salimbeni and Deisenroth, 2017), we assume that IID Gaussian noise $\mathcal{N}(0, v_n)$ (with v_n arbitrarily small) is added to the preactivations of the bottleneck layer; We include the variance of the noise in the covariance matrix C, so that C is invertible. We also add IID Gaussian noise $\mathcal{N}(0, v_n)$ to each of the two outputs of the bottleneck NNGP.

It is easy to verify that the two outputs $F_1(x_a)$ and $F_2(x_b)$ (for any $a, b \in \{1, 2\}$) are linearly uncorrelated; this immediately follows from the conditional independence of outputs given the bottleneck activations. The relationship between the squares of the outputs is, however, less trivial. Let $Q^{\times}$ denote the matrix of correlations between the squares of the

two outputs of F (the superscript $\times$ is used to emphasize that the correlation is between distinct output neurons).

Proposition 10 (Quadratic correlation between outputs) *Consider the single-bottleneck NNGP F with normalized ReLU activation defined above. Then*

$$\begin{aligned}\mathrm{Cov}[F_1(x_a)^2, F_2(x_b)^2] &= \frac{v_w^{2D}}{H}\,\mathrm{Cov}_{(z,z')\sim\mathcal{N}(0,K)}[\phi(z)^2, \phi(z')^2] \\ &= \frac{v_w^{2D} c_{aa} c_{bb}}{H}\left(\frac{2}{\pi}J_2(\beta) - 1\right),\end{aligned} \tag{11}$$

where $\beta = \frac{c_{ab}}{\sqrt{c_{aa}c_{bb}}}$ and the function J_2 is defined as (Cho and Saul, 2009):

$$J_2(\beta) = 3\sin\beta\cos\beta + (\pi - \beta)(1 + 2\cos^2\beta).$$

The corresponding correlation is

$$\begin{aligned}q_{ab}^{\times} &= \frac{\mathrm{Cov}[F_1(x_a)^2, F_2(x_b)^2]}{\sqrt{\mathrm{V}[F_1(x_a)^2]\,\mathrm{V}[F_2(x_b)^2]}} \\ &= \frac{\left(\frac{2}{\pi}J_2(\beta) - 1\right)}{\sqrt{\left[15 + 2H\left(\frac{r_D}{c_{aa}} + 1\right)^2\right]\left[15 + 2H\left(\frac{r_D}{c_{bb}} + 1\right)^2\right]}},\end{aligned} \tag{12}$$

where

$$r_D = \begin{cases} v_n + Dv_b & \text{if } v_w = 1 \\ \frac{v_n}{v_w^D} + \frac{v_b}{1 - v_w}\left(\frac{1}{v_w^D} - 1\right) & \text{otherwise.}\end{cases} \tag{13}$$

Remark 11 (Quadratic correlation for stationary kernels) *In Prop. 10, if we instead consider a single-bottleneck DGP F where the post-bottleneck GP has a stationary kernel (such as the RBF kernel), then the quadratic correlation between outputs is $q_{ab}^{\times} = 0$. The non-stationarity of the NNGP kernel is therefore key to capturing some amount of correlation.*

The proof of Prop. 10 as well as the proofs of all other propositions in Sec. 4 are given in Appendix D. The significance of Prop. 10 is two-fold. First, quadratic correlation under the prior—although subtle—may translate to stronger dependence (such as linear correlation) under the posterior. Indeed, as stated in Remark 11, a DGP with RBF kernel captures less correlation under its prior than does a bottleneck NNGP with ReLU activation, and yet the former has been shown to be useful in modeling dependence in practice (Alaa and van der Schaar, 2017; Wang et al., 2016). This suggests that a bottleneck NNGP with ReLU activation may be just as useful in modeling dependence. Second, the ability of the network to capture quadratic correlation is closely linked to its ability to operate effectively at extreme depths; we discuss this in more detail in Sec. 4.4.

The pre-bottleneck component of the bottleneck NNGP is a map that sends input vectors in $\mathbb{R}^M$ to normally distributed real-valued random variables (preactivations of bottleneck

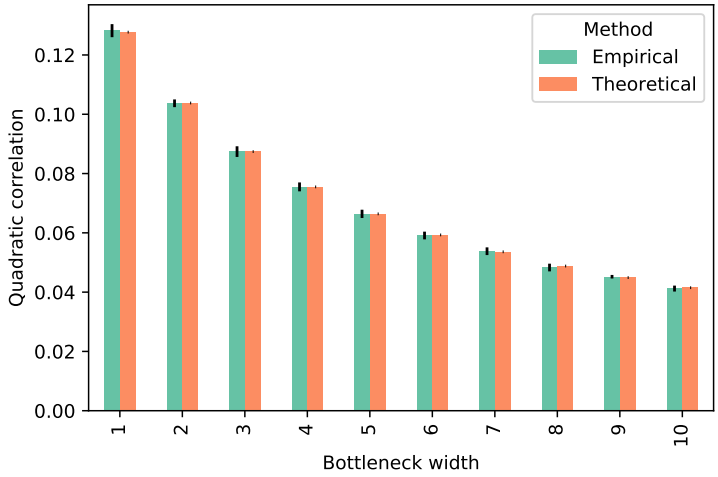


Figure 4: Empirical estimates with standard errors of the quadratic correlation of the outputs of an example bottleneck NNGP with one bottleneck surrounded by two infinite hidden layers, compared to theory (Eq. (12)) for various widths of the bottleneck layer. The theoretical values are all within one standard error of the empirical mean value.

neurons). By understanding covariance as an inner product on the space of finite-variance random variables, we can see that the covariance matrix C is the Gram matrix of bottleneck preactivations, and the angle β appearing in Prop. 10 is the angle between the preactivations of two inputs at one bottleneck neuron; we will call β the bottleneck angle.

The quadratic correlation (Eq. (12)) varies with the bottleneck width H roughly as $\frac{1}{H}$ and thus vanishes in the limit of infinite bottleneck width, recovering the independence of outputs of an NNGP with no bottlenecks. We empirically verified Eq. (12) for a bottleneck NNGP prior with one hidden layer before the bottleneck, one hidden layer after the bottleneck ($D = 2$), and with variance parameters $v_b = v_w = 1$ and $v_n = 10^{-4}$. We fed the example bottleneck NNGP two inputs $x_1 = (1, 0)$ and $x_2 = (0, 1)$. Then for each bottleneck width $H \in \{1, \ldots, 10\}$, we generated 10^6 IID samples of $(F_1(x_1), F_2(x_2))$ and used them to estimate $q_{12}^{\times}$. We repeated this simulation 10 times, and we report the mean estimate of $q_{12}^{\times}$ along with its standard deviation for each bottleneck width (Fig. 4). The empirical quadratic correlation estimates are very close to the theoretical values predicted by Eq. (12), with standard deviations all on the order of 10^{-3}.

We performed additional simulations to understand how multiple bottleneck layers affect the correlation of outputs, as we found this to be intractable theoretically. We still consider a bottleneck NNGP with $v_b = v_w = 1$ and $v_n = 10^{-4}$ that is fed two 2D inputs $x_1 = (1, 0)$ and $x_2 = (0, 1)$, but now we suppose the bottleneck NNGP has 11 hidden layers (including all bottleneck layers). We chose 11 hidden layers since it allows us to restrict zero to three hidden layers to bottlenecks such that the bottlenecks are equally spaced in depth. For each of the zero to three bottleneck layers, we ran the experiment described above for a single bottleneck and estimated the quadratic correlation $q_{12}^{\times}$ along with its standard deviation over ten runs for various bottleneck widths (all bottleneck layers have the same

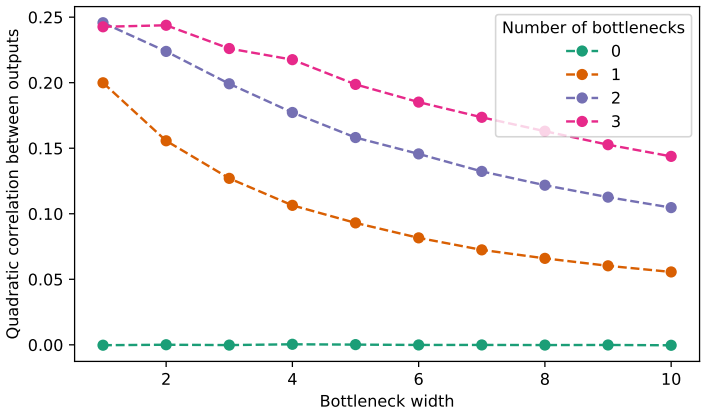


Figure 5: Empirical quadratic correlation of the outputs of an example bottleneck NNGP with 11 hidden layers, some of which are restricted to regularly spaced bottlenecks. Quadratic correlation increases with more numerous and narrower bottlenecks.

width). The quadratic correlation tends to zero with increasing bottleneck width regardless of the number of bottleneck layers, as we expect (Fig. 5). We additionally observe that for bottleneck widths $H \geq 2$, the quadratic correlation increases with the number of bottlenecks even if the overall depth of the bottleneck NNGP remains the same, suggesting that adding more bottlenecks has a similar effect to further narrowing existing bottlenecks.

4.2 Quadratic correlation as a function of depth

The inverse dependence of the quadratic correlation on bottleneck width is intuitive since we know that the outputs of an NNGP are independent in the absence of bottlenecks. There is also an interesting and less obvious dependence of the quadratic correlation on the post-bottleneck depth D (where there are $D-1$ post-bottleneck hidden layers of infinite width) as well as on the angle β between the random bottleneck preactivations of the inputs x_a $(a = 1, 2)$ in the bottleneck layer, as captured by the covariance matrix C. We denote the quadratic correlation by $q_{ab}^{\times(D)}$ to make explicit its dependence on the post-bottleneck depth D and will sometimes write $q_{ab}^{\times(D)}(\beta)$ to further clarify its dependence on the bottleneck angle β. By Prop. 10, it is easy to verify that $q_{ab}^{\times(D)}(\beta)$ is strictly decreasing in β on $[0, \pi]$ with $\beta < \frac{\pi}{2}$ giving positive correlation, $\beta = \frac{\pi}{2}$ giving zero correlation, and $\beta > \frac{\pi}{2}$ giving negative correlation at all depths D. The quadratic correlation between outputs therefore encodes the correlation of inputs in the bottleneck layer. By Eq. (13), r_D is strictly increasing in D regardless of the values of $v_b, v_w > 0$. It follows that the absolute quadratic correlation $|q_{ab}^{\times(D)}(\beta)|$ strictly decreases with D for $\beta \neq 0$ (and remains 0 otherwise). A final property of the quadratic correlation is its range of possible values, which easily follows from the ranges of J_2 and r_D:

$$-\frac{1}{17} < q_{ab}^{\times(D)}(\beta) < \frac{5}{17}. \tag{14}$$

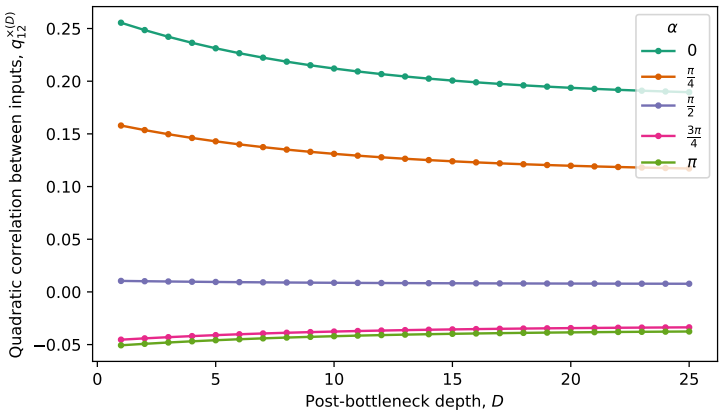


Figure 6: Theoretical quadratic correlation between outputs of a single-bottleneck NNGP over the post-bottleneck depth D for various angles α between the inputs. These quadratic correlations are asymptotically non-zero whenever $v_w > 1$ and the bottleneck preactivations (as random variables) of the inputs are not orthogonal (bottleneck orthogonality occurs at $\alpha \approx 0.526\pi$).

These properties are apparent in the plot of $q_{ab}^{\times(D)}(\beta)$ over D for an example single-bottleneck NNGP ($H = 2$, $v_b = 0.09$, $v_w = 1.1$) with no pre-bottleneck hidden layers that is fed two inputs $x_1 = (1, 0)$ and $x_2 = (\cos\alpha, \sin\alpha)$ for various values of the input angle α (Fig. 6); since there are no pre-bottleneck hidden layers, the input angle α and bottleneck angle β are related through the equation

$$\cos\beta = \frac{v_b + v_w \cos\alpha}{v_b + v_w}.$$

Orthogonal bottleneck preactivations ($\beta = \frac{\pi}{2}$) and thus 0 quadratic correlation in the example bottleneck NNGP are then achieved at an input angle $\alpha \approx 0.526\pi$. Observe in general that if $v_b > 0$, then $\beta < \alpha$ and the range of β is strictly smaller than $[0, \pi]$, indicating that bias units promote positive quadratic correlation.

The behavior of r_D (Eq. (13)) in the limit of infinite post-bottleneck depth ($D \to \infty$) is easily analyzed. We have

$$r_\infty = \lim_{D\to\infty} r_D = \begin{cases} \frac{v_b}{v_w - 1} & \text{if } v_w > 1 \\ \infty & \text{otherwise.} \end{cases} \tag{15}$$

This lets us determine what happens to the quadratic correlation of outputs as the number of post-bottleneck hidden layers grows to infinity.

Proposition 12 (Infinite-depth quadratic correlation between outputs) *Consider the single-bottleneck NNGP F with normalized ReLU activation from Prop. 10, and suppose we send the post-bottleneck depth to infinity.*

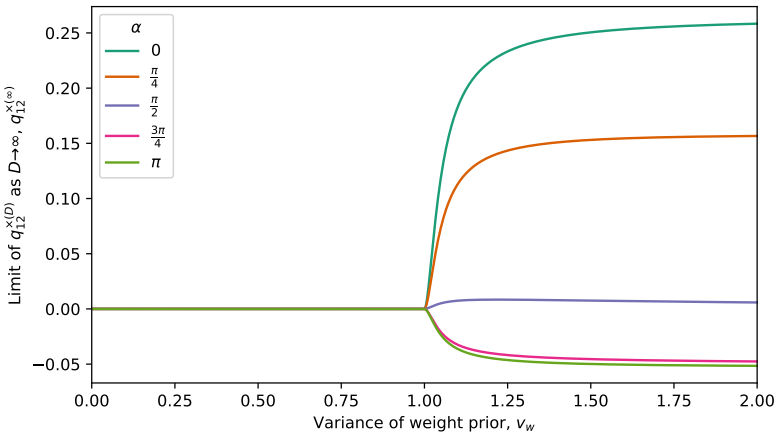


Figure 7: The infinite-depth quadratic correlation between outputs as a function of the weight-variance hyperparameter v_w for an example bottleneck NNGP for various angles α between the inputs.

(a) The infinite-depth quadratic correlation matrix $Q^{\times(\infty)}$ *has* (a,b)*-th element* $q_{ab}^{\times(\infty)} = \lim_{D\to\infty} q_{ab}^{\times(D)}$ *given by*

$$q_{ab}^{\times(\infty)} = \begin{cases} \dfrac{\left(\frac{2}{\pi}J_2(\beta)-1\right)}{\prod\limits_{c=c_{aa},c_{bb}} \sqrt{15+2H\left(\dfrac{v_b}{(v_w-1)c}+1\right)^2}} & \text{if } v_w > 1 \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

(b) Let G *be the Gram matrix of inputs with entries* $g_{ab} = x_a^\top x_b$*. If* $v_w > 1$*, then the mapping* $G \mapsto Q^{\times(\infty)}$ *is invertible.*

We visualize the infinite-depth quadratic correlation by plotting Eq. (16) as a function of v_w for a single-bottleneck NNGP ($H = 2$, $v_b = 0.09$) with no pre-bottleneck hidden layers that is fed two inputs $x_1 = (1, 0)$ and $x_2 = (\cos\alpha, \sin\alpha)$ for various values of the input angle α (Fig. 7). Note that the covariance matrix C and hence the bottleneck angle β are themselves functions of v_w, which is why we plot the quadratic correlation for fixed values of α instead of β. The infinite-depth quadratic correlation exhibits interesting behavior around $v_w = 1$; it is continuous but not differentiable at $v_w = 1$. The bottleneck NNGP therefore undergoes a phase transition at $v_w = 1$. The quadratic correlation tends to 0 in the $v_w \leq 1$ regime—meaning that the outputs of a bottleneck NNGP decouple (up through second order correlations) at infinite depth—while quadratic correlation is maintained through infinitely many, infinitely wide hidden layers in the regime $v_w > 1$, though even then the limiting correlation is weak (Eq. (14)).

A phase transition at $v_w = 1$ has already been noted in the literature in the behavior of no-bottleneck NNGP models at infinite depth (Schoenholz et al., 2016; Poole et al., 2016; Lee et al., 2017). Specifically, the kernel of an NNGP with normalized ReLU activation

degenerates to a constant kernel at infinite depth with a value of either $\frac{v_b}{1-v_w}$ if $v_w < 1$ and ∞ otherwise. Bottleneck layers help to reveal a richer structure of this phase transition, as we explain next. Drawing an analogy to the classical Ising model in statistical mechanics (Baxter, 2016), the hyperparameter v_w operates as an inverse temperature with a critical value at $v_w = 1$. The quadratic correlation is then an order parameter analogous to magnetization whose derivative contains a discontinuity at the phase boundary $v_w = 1$. In the $v_w < 1$ phase, the infinite-depth quadratic correlation is 0 regardless of the bottleneck angle β; information about the inputs into the bottleneck NNGP is therefore lost, analogous to a disordered system at large scale. However, as v_w crosses the phase boundary from below, the system undergoes a symmetry breaking with the infinite-depth quadratic correlation taking a distinct value for each bottleneck angle β as well as for each input angle α. This lets us recover information about the inputs from the infinite-depth quadratic correlation, indicating that bottlenecks help information propagate to extreme depths (Prop. 12 (b)).

The symmetry breaking discussed above is not apparent in the phase transition of degenerate NNGP kernels noted in the literature; i.e., all information about the inputs are lost in an infinite-depth (no-bottleneck) NNGP in either phase. We see, however, that the restriction of just one hidden layer to a bottleneck is sufficient to break this symmetry. Moreover, the symmetry can be recovered by sending the bottleneck width to infinity (Thm. 8); the bottleneck is therefore analogous to an external magnetic field in the Ising model.

4.3 A divergent depth scale

Schoenholz et al. (2016) show that the characteristic depth scale on which the kernel of a (no-bottleneck) NNGP degenerates exponentially to its deep limit diverges at a phase boundary in (v_b, v_w)-space, and they use this result to argue that values of (v_b, v_w) near criticality or "at the edge of chaos" optimize trainability by maximizing the depth to which information can penetrate in an NNGP. We show that the depth scale on which the quadratic correlation $Q^{\times(D)}$ converges to its limit also diverges at the phase boundary $v_w = 1$.

Proposition 13 (Characteristic depth scale) *Consider the single-bottleneck NNGP F with normalized ReLU activation from Prop. 10, and suppose it is fed two inputs x_1 and x_2 with $\|x_1\| = \|x_2\|$ and bottleneck angle $\beta \neq \frac{\pi}{2}$. Then for all positive $v_w \neq 1$, the quadratic correlation $q_{ab}^{\times(D)}$ is asymptotically exponential in D, meaning that the limit*

$$L = \lim_{D\to\infty} \frac{q_{ab}^{\times(D)} - q_{ab}^{\times(\infty)}}{e^{-\frac{D}{\lambda}}} \tag{17}$$

exists and is finite and non-zero for some $\lambda > 0$, which we find to be

$$\lambda = \begin{cases} \ln\left(\frac{1}{v_w^2}\right)^{-1} & \text{if } v_w < 1 \\ \ln(v_w)^{-1} & \text{if } v_w > 1. \end{cases} \tag{18}$$

In the case $v_w = 1$, the limit L is infinite for all finite $\lambda > 0$.

Proposition 13 excludes the case of orthogonal bottleneck preactivations ($\beta = \frac{\pi}{2}$) since it leads to trivial asymptotic behavior ($q_{ab}^{\times(D)} = 0$ for all D). The quantity λ given in

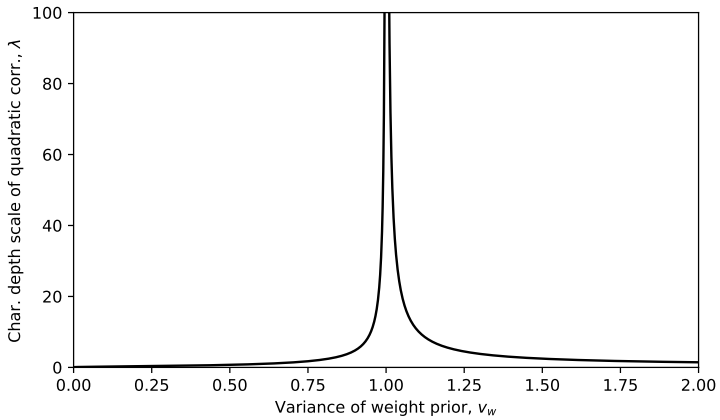


Figure 8: Characteristic depth scale of the convergence of the quadratic correlation between outputs to its infinite-depth limit as a function of the weight-variance hyperparameter v_w.

Eq. (18) is called the characteristic depth scale and is proportional to the "half-life" of quadratic correlation at large depth. The divergence of the depth scale at $v_w = 1$ gives us another perspective on the phase transition in bottleneck NNGP models (Fig. 8). Based on this, we expect optimal values of the v_w hyperparameter to be greater than but close to 1; although $v_w = 1$ gives the smallest decay rate (i.e., largest depth scale) to the infinite-depth quadratic correlation, this limiting value is 0. Larger values of v_w admit non-zero infinite-depth quadratic correlations, but values that are too large lead to fast decay rates (i.e., small depth scales). We hypothesize that this tension between large depth scales (near $v_w = 1$) and non-zero quadratic correlations ($v_w \gg 1$) is the main driving force determining the optimal value of v_w in the $v_w > 1$ phase.

4.4 Non-degenerate kernels at extreme depths

The non-trivial dependence of the infinite-depth quadratic correlation $q_{ab}^{\times(\infty)}(\beta)$ on the bottleneck angle β has remarkable implications for the kernel or covariance matrix $K^{(\infty)}$ of individual output neurons of a bottleneck NNGP at infinite depth. Consider again the single-bottleneck NNGP F described at the beginning of Sec. 4.1, and suppose it is fed two linearly independent inputs x_1 and x_2. Assume that the infinite-depth kernel $K^{(\infty)}$ is degenerate so that the associated correlation matrix $\hat{K}^{(\infty)}$ is a matrix of ones; this is indeed the case when there are no bottlenecks, where the correlations in $\hat{K}$ tend to 100% even when all elements of the kernel K grow to infinity at infinite depth. Then the outputs $F_i(x_1)$ and $F_i(x_2)$ of a single output neuron i at infinite depth are linearly dependent and are in fact equal. It follows that the elements $q_{ab}^{\times(\infty)}$ of the quadratic correlation matrix $Q^{\times(\infty)}$ between distinct output neurons are all equal, but this is impossible in the $v_w > 1$ phase since $q_{12}^{\times(\infty)}(\beta)$ is a one-one function of β (Fig. 7) and the diagonal elements assume $\beta = 0$ while the off-diagonal elements assume $\beta > 0$ (since x_1 and x_2 are linearly independent). We therefore learn that the kernel of an NNGP does not degenerate to a constant in the $v_w > 1$ phase if at least one hidden layer has finite width.

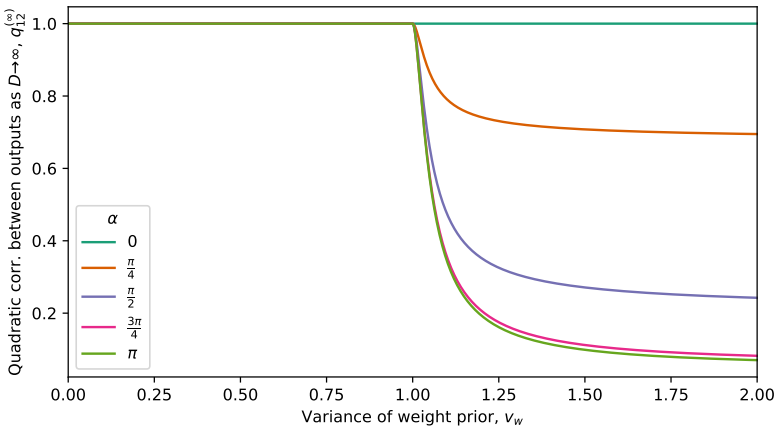


Figure 9: The infinite-depth quadratic correlation of a single output as a function of the v_w hyperparameter for an example bottleneck NNGP.

Unfortunately, the kernel—and thus the associated correlation matrix—of each output neuron of a bottleneck NNGP do not admit closed forms, even in the infinite depth limit. The quadratic correlation matrix $Q^{(D)}$ (without a superscript $\times$) for individual outputs is intractable as well, but its infinite depth limit does admit an elegant closed form.

Proposition 14 (Infinite-depth quadratic correlation for single output) *Consider the single-bottleneck NNGP F with normalized ReLU activation from Prop. 10, and suppose we send the post-bottleneck depth to infinity.*

(a) The infinite-depth quadratic correlation matrix $Q^{(\infty)}$ of a single output neuron has (a,b)-th element $q_{ab}^{(\infty)} = \lim_{D\to\infty} q_{ab}^{(D)}$ given by

$$q_{ab}^{(\infty)} = \begin{cases} 3q_{ab}^{\times(\infty)} + \dfrac{\left(\frac{v_b}{(v_w-1)c_{aa}}+1\right)\left(\frac{v_b}{(v_w-1)c_{bb}}+1\right)}{\prod_{c=c_{aa},c_{bb}} \sqrt{\frac{15}{2H} + \left(\frac{v_b}{(v_w-1)c}+1\right)^2}} & \text{if } v_w > 1 \\ 1 & \text{otherwise.} \end{cases} \tag{19}$$

(b) Let G be the Gram matrix of inputs with entries $g_{ab} = x_a^\top x_b$. If $v_w > 1$, then the mapping $G \mapsto (Q^{(\infty)}, \mathrm{diag}(G))$ is invertible.

The infinite-depth single-output quadratic correlation $Q^{(\infty)}$ carries many of the same properties as the infinite-depth between-output quadratic correlation $Q^{\times(\infty)}$. Recalling that $q_{ab}^{\times(\infty)} \to 0$ as $H \to \infty$, it is easy to verify that $q_{ab}^{(\infty)} \to 1$. The single-output quadratic correlation also exhibits the same symmetry breaking at the phase boundary $v_w = 1$ as the between-output quadratic correlation; this is evident in the example plot of $q_{12}^{(\infty)}(\beta)$ as a function of v_w, using the same setup as for Fig. 7 (see Fig. 9).

In the phase $v_w < 1$, the single-output quadratic correlation is 100%, suggesting degeneracy at infinite depth. In contrast, in the phase $v_w > 1$—where symmetry breaks—$q_{ab}^{(\infty)}$ becomes a strictly increasing function of the input angle α, allowing us to recover some information about the inputs (Prop. 14 (b)). In particular, Prop. 14 (b) implies that in the ordered phase $v_w > 1$, if the norms of the inputs are known (if the inputs are constrained to a sphere, for example), then the input angle can be recovered from the single-output quadratic correlation even at infinite depth. This stands in stark contrast to no-bottleneck NNGP models, where all information is lost at infinite depth regardless of the phase, and it suggests that bottleneck layers are vital for the trainability of very deep models.

4.5 Other nonlinearities

The behavior described in Prop. 14 does not extend to other common nonlinearities and is thus all the more striking. Two notable examples are the sigmoidal nonlinearity $\phi(x) = \tanh x$ and the sinusoidal nonlinearity

$$\phi(x) = \cos x + \sin x. \tag{20}$$

The significance of the latter is that the corresponding NNGP has an RBF kernel; the random features literature (Rahimi and Recht, 2008; Cutajar et al., 2017) hints at this connection but does not discuss it in the context of NNGPs or neural network nonlinearities. We make this connection more precise in the next proposition and subsequent remark (see Appendix D.1 for the proofs of both Props. 15, 17).

Proposition 15 (RBF-NNGP kernel recursion) *Consider an NNGP mapping $\mathbb{R}^M$ to $\mathbb{R}$ with D hidden layers and the sinusoidal nonlinearity in Eq.* (20)*.*

(a) The NNGP kernel $K^{(\mu)} : \mathbb{R}^M \times \mathbb{R}^M \mapsto \mathbb{R}$ for the μ-th hidden layer is given recursively as

$$K^{(1)}(x, x') = v_b + v_w x^\top x' \tag{21}$$

$$K^{(\mu+1)}(x, x') = v_b + v_w e^{-\frac{1}{2}\left[K^{(\mu)}(x,x) + K^{(\mu)}(x',x') - 2K^{(\mu)}(x,x')\right]}, \tag{22}$$

for $\mu = 1, \ldots, D$, where $K^{(D+1)}$ is the kernel of the output layer.

(b) In the deep limit $D \to \infty$, the NNGP kernel converges pointwise to

$$K^{(\infty)}(x, x') = v_*(v_b, v_w) \begin{cases} 1, & \text{if } x = x' \\ c_*(v_b, v_w), & \text{otherwise,} \end{cases} \tag{23}$$

where

$$v_*(v_b, v_w) = v_b + v_w \tag{24}$$

$$c_*(v_b, v_w) = \begin{cases} 1, & \text{if } v_w < 1 \\ c', & \text{if } v_w > 1. \end{cases} \tag{25}$$

Remark 16 (RBF-DGP as a bottleneck NNGP) *Since the sinusoidal nonlinearity in Eq.* (20) *clearly satisfies the linear envelope condition, then we can apply Thm. 4 to a BNN with the sinusoidal nonlinearity and obtain a bottleneck NNGP limit. If no bottlenecks are imposed, then the wide limit of such a BNN is the NNGP described in Prop. 15. In particular, using Eqs.* (21)-(22)*, the single-hidden-layer NNGP with sinusoidal nonlinearity has kernel*

$$\begin{aligned} K^{(2)}(x,x') &= v_b + v_w e^{-\frac{1}{2}[K^{(1)}(x,x)+K^{(1)}(x',x')-2K^{(1)}(x,x')]} \\ &= v_b + v_w e^{-\frac{1}{2}[v_b+v_w\|x\|^2+v_b+v_w\|x'\|^2-2(v_b+v_w x\cdot x')]} \\ &= v_b + v_w e^{-\frac{v_w}{2}\|x-x'\|^2}, \end{aligned}$$

which we recognize as the RBF kernel. More generally, if the BNN with sinusoidal nonlinearity has an odd number of hidden layers D and the μ-th hidden layers are restricted to bottlenecks for $\mu = 2, 4, 6, \ldots, D-1$, then the bottleneck NNGP limit is a DGP with $\frac{D+1}{2}$ GP components each with RBF kernel (and sinusoidal nonlinearities applied to the bottleneck layers).

The deep limit of an NNGP with sinusoidal nonlinearity is described in Prop. 15 (b). It is identical to the behavior of an NNGP with sigmoidal nonlinearity as described by Poole et al. (2016), except that $v_*(v_b, v_w)$ and $c_*(v_b, v_w)$ take different forms and the phase boundary has a different location. In one phase ($v_w < 1$ for the sinusoidal nonlinearity), all inputs tend to 100% correlation as depth is increased without bound—similar to the ReLU activation. However, in the other phase ($v_w > 1$ for the sinusoidal nonlinearity)—the "chaotic phase"—the infinite-depth correlation for distinct inputs is a constant less than 100%; thus, unlike the ReLU activation, the sigmoidal and sinusoidal nonlinearities allow distinct inputs to remain distinct through infinite depth, although all information about the distance between distinct inputs is lost. Moreover, the introduction of a bottleneck does not remove this degeneracy—in sharp contrast to Prop. 14. We substantiate this with the next proposition.

Proposition 17 (Deep post-bottleneck limit for sigmoidal and sinusoidal nonlinearities) *Consider a single-bottleneck NNGP F mapping $\mathbb{R}^M$ to $\mathbb{R}$ with either sigmoidal or sinusoidal nonlinearity $\phi : \mathbb{R} \mapsto \mathbb{R}$. Suppose IID Gaussian noise $\mathcal{N}(0, v_n)$ is added to the outputs for any $v_n > 0$. Given two distinct inputs $x_1, x_2 \in \mathbb{R}^M$, let $p^{(D)} : \mathbb{R}^2 \mapsto [0, \infty)$ be the PDF of $\{F(x_1), F(x_2)\}$, where D is the post-bottleneck depth. Then in the deep post-bottleneck limit $D \to \infty$, the PDF converges pointwise to*

$$p^{(\infty)}(y) = \mathcal{N}(y; 0, K^{(\infty)}(X, X) + v_n I), \tag{26}$$

where $K^{(\infty)}(X, X)$ is a 2×2 matrix with entries $K^{(\infty)}(x_a, x_b)$ with $K^{(\infty)}$ given in Eq. (23)*.*

Note that Eq. (26) is independent of bottleneck width; in fact, it is exactly the PDF of a no-bottleneck NNGP, indicating that the bottleneck has no effect at all on the deep post-bottleneck limit given either the sigmoidal or sinusoidal nonlinearity. In contrast to Prop. 14 (b) for the ReLU activation, the only information about the inputs that can be recovered from the deep post-bottleneck limit is whether the inputs are distinct. This

suggests a fundamental difference between the ReLU activation and the sigmoidal nonlinearity and between the NNGP kernel with ReLU activation and the RBF kernel—namely, that the ReLU activation allows a network to operate at very large depths as long as one bottleneck is present. Classifying activation functions based on their deep limit and deep post-bottleneck limit behaviors could help to better understand which activation functions are useful in practice and is a topic for future work.

4.6 Other deep limits

Besides the deep post-bottleneck limit discussed in Secs. 4.2-4.4, there are two other limits we could have considered: 1) The deep pre-bottleneck limit where a single-bottleneck NNGP has infinitely many infinitely wide hidden layers before its bottleneck but only finitely many after, and 2) the doubly deep bottleneck limit where the single-bottleneck NNGP has infinitely many infinitely wide hidden layers both before and after its bottleneck. The quadratic correlation between the activations of distinct output neurons can be analyzed in both of these limits by replacing the pre-bottleneck NNGP covariance matrix C in Prop. (10) with a sequence of such covariance matrices indexed by the pre-bottleneck depth. Since the NNGP kernel with ReLU activation tends to a constant in the deep limit, then the bottleneck angle β tends to 0 in both the deep pre-bottleneck limit and doubly deep bottleneck limit. Although the quadratic correlation remains nontrivial in both limits and even exhibits a first-order phase transition at $v_w = 1$ in the doubly deep bottleneck limit, these results are uninteresting as all information about the original inputs into the network is lost due to vanishing bottleneck angle.

The quadratic correlation between the activations of a single output neuron at two different inputs can also be analyzed in both the deep pre-bottleneck and doubly deep bottleneck limits; calculations proceed similarly as in the proof of Prop. 14. Again, since the bottleneck angle tends to 0, we find that the quadratic correlation tends to 100% in both limits at every post-bottleneck depth; the bottleneck NNGP kernel therefore degenerates, unlike in the deep post-bottleneck limit. We conclude that out of the three possible deep limits, only the deep post-bottleneck limit is interesting since it is the only one that admits a phase ($v_w > 1$) in which information about network inputs is preserved.

5. Conclusion

Our main theorem, Thm. 4, generalizes the result of Matthews et al., 2018 concerning deep neural networks whose hidden layer widths are increased without bound to a setting in which some intermediate hidden layers, called bottlenecks, are fixed to a finite width. From a theoretical perspective, this result connects the NNGP literature with that of DGPs, as the resulting probability model is in fact a DGP consisting of a composition of NNGP components. Additionally, we have explored the effect of these bottleneck layers on the resulting probability model from a practical perspective, showing that model likelihood peaks at a finite bottleneck width and is superior to that of no-bottleneck NNGPs.

Surprisingly, in contrast to no-bottleneck NNGP models, the behavior of a bottleneck NNGP with ReLU activation at extreme post-bottleneck depths is not always degenerate (Props. 12, 14); in particular, the input Gram matrix can be fully recovered from the between-output quadratic correlation matrix of the bottleneck NNGP at infinite

post-bottleneck depth, and the input angle can be recovered even from the single-output quadratic correlation matrix at infinite depth if the input norms are known. Bottleneck layers are therefore fundamental as they allow networks to "go deeper". However, this non-degeneracy in the deep limit manifests only when the network weight prior is weaker than a standard normal. We have just begun to explore the dependence of the deep post-bottleneck limit on the prior weight variance, showing that convergence to the limit is asymptotically exponential in depth and that the characteristic depth scale diverges at a critical value of the weight prior variance.

So far, we have not directly connected bottleneck NNGPs with BNNs, aside from studying the limits of BNNs in the wide regime. However, an interesting special case of bottleneck NNGPs emerges when *every* hidden layer is fixed at finite width. The result is a BNN, but our result lends an interesting perspective that suggests that one might approach BNN inference with a non-parametric DGP-based approach. Indeed, in follow-up work we intend to explore the practicality of such a method. Moreover, we postulate that our main theorem can easily be extended to convolutional architectures (by introducing a second index for each hidden layer), and thus we plan to explore the implications of our work for convolutional BNNs.

Finally, in this work we did not consider the implications of the bottleneck NNGP for the learning dynamics of Gaussian-initialized deterministic neural networks (DNNs). It is now a celebrated result that the evolution over training time of a Gaussian-initialized DNN is described in the wide limit by an exactly solvable linear ODE, where the time evolution operator is termed the neural tangent kernel (NTK) and is related to the NNGP kernel (Jacot et al., 2018). As part of additional follow-up work, we plan to investigate the "bottleneck NTK", where the wide limit is relaxed to allow for some bottleneck layers. The result is a system of coupled ODEs that is more challenging to analyze but carries potential for a more refined description of the evolution of finite-width DNNs.

Acknowledgments

This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DEAC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725. Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US Department of Energy under contract DE-AC05-00OR22725. This research used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paidup, irrevocable, world-wide license

Appendix A. Convergence of stochastic processes

Here we define three notions of convergence that are important for formally stating and proving the bottleneck NNGP limit. The first notion of convergence is the one that Matthews et al., 2018 employ to prove the (no-bottleneck) NNGP limit. They consider a sequence of BNNs mapping $\mathbb{R}^M$ to $\mathbb{R}^L$ restricted to a countable set of inputs $\mathcal{X} \subset \mathbb{R}^M$. Each BNN is then equivalent to a random sequence of L-vectors, i.e. a random variable taking values in $(\mathbb{R}^L)^\infty$. Following Billingsley (1999), Matthews et al., 2018 equip the sequence space $(\mathbb{R}^L)^\infty$ with the metric

$$\rho(s,t) = \sum_{k=1}^{\infty} 2^{-k} \min(1, \|s_k - t_k\|),$$

where $s, t \in (\mathbb{R}^L)^\infty$. The metric ρ induces the Euclidean topology on $(\mathbb{R}^L)^\infty$, and we endow $(\mathbb{R}^L)^\infty$ with the associated Borel algebra $\mathcal{A}$, giving us a measurable space of sequences. Demonstrating convergence in $((\mathbb{R}^L)^\infty, \mathcal{A})$ can prove challenging, but Matthews et al., 2018 simplify the task by invoking the following theorem.

Theorem 18 (Billingsley (1999)) *A sequence of stochastic processes $\{F[n] : \mathcal{X} \times \Omega \mapsto \mathbb{R}^L\}_{n=1}^\infty$ with countable index set $\mathcal{X}$ converges in distribution to a stochastic process $F : \mathcal{X} \times \Omega \mapsto \mathbb{R}^L$ on the measurable space $((\mathbb{R}^L)^\infty, \mathcal{A})$ if and only if every sequence of finite-dimensional marginals $\{(F(x_1)[n], \ldots, F(x_T)[n])\}_{n=1}^\infty$ converges in distribution to the corresponding limiting marginal $(F(x_1), \ldots, F(x_T))$.*

Theorem 18 effectively reduces the task of proving the convergence of a sequence of random sequences to that of a sequence of random vectors. When looking at the convergence of T-dimensional marginal distributions, it is convenient to introduce the following notation.

Definition 19 (Batch stochastic process) *Let $F : \mathbb{R}^M \times \Omega \mapsto \mathbb{R}^L$ be a stochastic process. Then the batch stochastic process of size $T \in \mathbb{N}$ associated with F is the stochastic process $\tilde{F} : (\mathbb{R}^M)^T \times \Omega \mapsto (\mathbb{R}^L)^T$ defined by*

$$\tilde{F}(\{x_t\}_{t=1}^T, \omega) = \{F(x_t, \omega)\}_{t=1}^T.$$

By working with batch stochastic processes, we can think of T inputs as constituting a single input. Thus, to show $\{F[n]\}_{n=1}^\infty \to F$, it is enough to show $\{\tilde{F}(x)[n]\}_{n=1}^\infty \to \tilde{F}(x)$ for each input x for every batch size T.

Our proof of the bottleneck NNGP limit takes the approach of showing that each component of a BNN after the first bottleneck layer converges to an NNGP in the wide limit with some uniformity. We specify the appropriate notion of uniform convergence next.

Definition 20 (Uniform convergence in distribution (Sweeting, 1980)) *A sequence of stochastic processes $\{F[n] : X \times \Omega \mapsto \mathbb{R}^L\}_{n=1}^\infty$ is said to converge in distribution to*

$\{F : X \times \Omega \mapsto \mathbb{R}^L\}$ *uniformly on* X *if for every continuity set* $U \subseteq \mathbb{R}^L$ *of* F*, i.e., a set satisfying*

$$\Pr(F(x) \in \partial U) = 0 \text{ for all } x \in X,$$

we have the limit

$$\lim_{n\to\infty} \Pr(F(x)[n] \in U) = \Pr(F(x) \in U) \text{ uniformly for all } x \in X.$$

We denote this by $F[n] \xrightarrow{UD} F$.

Note that uniform convergence in distribution is distinct from and is not a stronger version of convergence in distribution in $((\mathbb{R}^L)^\infty, \mathcal{A})$ since the former notion concerns only the singly-indexed marginals of a stochastic process while the latter deals with the joint distribution of all elements of a stochastic process. However, the uniform convergence in distribution of the batch stochastic processes $\{\tilde{F}[n]\}_{n=1}^\infty$ is stronger than the convergence in distribution of the original stochastic processes $\{F[n]\}_{n=1}^\infty$. Proving uniform convergence in distribution can be challenging, but fortunately there is another closely related notion of convergence, which we define next.

Definition 21 (Continuous convergence in distribution (Sweeting, 1980)) *Let* X *be a topological space. A sequence of stochastic processes* $\{F[n] : X \times \Omega \mapsto \mathbb{R}^L\}_{n=1}^\infty$ *is said to converge in distribution to* $\{F : X \times \Omega \mapsto \mathbb{R}^L\}$ *continuously on* X *if for every* $x \in X$ *and sequence* $\{x_n \in X\}_{n=1}^\infty$ *converging to* x*, the sequence of random variables* $\{F(x_n)[n]\}_{n=1}^\infty$ *converges in distribution to* $F(x)$*. We denote this by* $F[n] \xrightarrow{CD} F$.

Uniform and continuous convergence in distribution are related through the following proposition.

Proposition 22 (Saikkonen (1993)) *Let* X *be a topological space. Let* $\{F[n] : X \times \Omega \mapsto \mathbb{R}^L\}_{n=1}^\infty$ *be a sequence of stochastic processes and* $\{F : X \times \Omega \mapsto \mathbb{R}^L\}$ *a stochastic process. Then the following are equivalent statements:*

(a) $F[n] \xrightarrow{CD} F$.

(b) $F[n] \xrightarrow{UD} F$ *on every compact subset of* X *and* $x \mapsto \Pr(F(x) \in U)$ *is a continuous function for every continuity set* U *of* F.

We should mention that Sweeting (1980) and Saikkonen (1993) do not state Defs. 20-21 and Prop. 22 directly but instead define and work with the equivalent notions of uniform and continuous weak convergence of probability measures. The proof of Prop. 22 is a simple application of the equivalence of uniform and continuous convergence of real-valued functions on compact sets.

Appendix B. The bottleneck NNGP theorem

Here we list the proof of our main theorem (Thm. 4) along with all supporting lemmas. We give a sketch of the proof next, highlighting the role of each lemma and its position in the general proof strategy.

The first step is to apply Thm. 18 so that it is sufficient to prove convergence of BNNs to a bottleneck NNGP restricted to an arbitrary finite set of inputs. Since each component $F^{(d)}[n]$ in Thm. 4 is being evaluated at T inputs, then it is convenient to utilize the concept of a batch stochastic process of size T (Def. 19). By working with batch BNNs, we can think of T inputs as constituting a single "batch" input. This reduces our task to proving the convergence of batch BNNs to a batch bottleneck NNGP given a single arbitrary input.

The next step is to find sufficient conditions under which the distributional limit of an element-wise composition of a sequence of stochastic processes with a sequence of random variables equals the composition of the limiting stochastic process with the limiting random variable. This trick can then be iterated via induction to prove that the limit of compositions is the composition of limits for stochastic processes. Lemma 24 provides such sufficient conditions, which include a notion of uniform convergence in distribution (Def. 20).

Proving Thm. 4 now comes down to verifying the conditions of Lemma 24. Condition (a) is given to us by Thm. 3 for single-bottleneck networks and holds by induction for multi-bottleneck networks. Condition (b) is also immediate in the single-bottleneck case, although it is less obvious for multi-bottleneck architectures. We verify condition (b) directly in the proof of Thm. 4 but take aid from Lemma 25. Lemma 26 establishes condition (c), which amounts to showing that the NNGP kernel is a continuous function. Condition (d) is the trickiest to verify; it states that the outer sequence of stochastic processes (that is composed with an inner sequence of random variables) must converge in distribution uniformly (Def. 20) on compact sets, meaning that the rate of convergence in distribution should be independent of the input to the stochastic processes. Condition (d) is verified with the help of Lemma 27.

Lemma 27 is a direct generalization of Lemma 12 in Matthews et al., 2018; the latter states that given a fixed finite batch of inputs, BNNs with no bottlenecks converge in distribution to an NNGP in the wide limit. In Lemma 27, we strengthen the mode of convergence to continuous convergence in distribution (Def. 21). More specifically, Lemma 27 states that a BNN converges in distribution to an NNGP even if we replace the fixed batch of inputs with a convergent sequence of input batches. Continuous convergence in distribution is in fact equivalent to uniform convergence in distribution on compact sets (Prop. 22), thus granting condition (d).

The proof of Lemma 27 runs in parallel to the proof of Lemma 12 in Matthews et al., 2018. It depends on several lemmas (Appendices B.3-B.4) that are all simple extensions of (or help to extend) the lemmas in Matthews et al., 2018; at each step, we simply replace the fixed batch of inputs in Matthews et al., 2018 with a convergent sequence of input batches and verify that convergence in distribution still holds. Only a few key modifications are made to the lemmas establishing uniform integrability (Appendix B.4).

B.1 Notation

We start with some notation. Let $\{F[n] : \mathbb{R}^M \times \Omega \mapsto \mathbb{R}^L\}_{n=1}^{\infty}$ be a sequence of BNNs each with D hidden layers of widths $H_\mu[n]$, $\mu \in \{1, \ldots, D\}$, and nonlinearity $\phi : \mathbb{R} \mapsto \mathbb{R}$ that satisfies the linear envelope condition. Let $F : \mathbb{R}^M \times \Omega \mapsto \mathbb{R}^L$ be the limiting NNGP of the sequence of BNNs. Let $f_i^{(\mu)}(x)[n]$ (resp. $f_i^{(\mu)}(x)$) and $g_i^{(\mu)}(x)[n]$ (resp. $g_i^{(\mu)}(x)$) be the preactivation and activation of the i-th neuron in the μ-th hidden layer of $F[n]$ (resp. F).

For each $n \in \mathbb{N}$, let $X[n] = \{x_t[n] \in \mathbb{R}^M\}_{t=1}^T$ be a batch of T inputs, and suppose the sequence of batches $\{X[n]\}_{n=1}^{\infty}$ converges to some finite $X = \{x_t\}_{t=1}^T$. Let $\alpha \in \mathbb{R}^{T \times |\mathbb{N}|}$ be a countably infinite block vector whose blocks are indexed by $\mathbb{N}$ and where each block has T elements. Let α have finite support $\{1, \ldots, T\} \times I$, where I is a finite subset of $\mathbb{N}$; i.e., only finitely many blocks indexed by I are permitted to have non-zero elements. Let α_{ti} denote the t-th element in the i-th block. For each $\mu \in \{1, \ldots, D+1\}$, define the preactivation projections of a BNN and its limiting NNGP as

$$f^{(\mu)}(X[n], \alpha)[n] = \sum_{t=1}^{T} \sum_{i \in I} \alpha_{ti} f_i^{(\mu)}(x_t[n])[n],$$

$$f^{(\mu)}(X, \alpha) = \sum_{t=1}^{T} \sum_{i \in I} \alpha_{ti} f_i^{(\mu)}(x_t).$$

Let $k^{(\mu)} : \mathbb{R}^M \times \mathbb{R}^M \mapsto \mathbb{R}$ be the NNGP kernel of $f_i^{(\mu)}$. The kernel $k^{(\mu)}$ relates to the block kernel $K^{(\mu)}$ (Eqs. (6)-(7)) through the equation

$$K^{(\mu)}(X, X) = \begin{cases} k^{(\mu)}(X, X) \otimes \mathbf{I}_\infty & \text{for } \mu \in \{1, \ldots, D\} \\ k^{(\mu)}(X, X) \otimes \mathbf{I}_L & \text{for } \mu = D+1, \end{cases}$$

where $\mathbf{I}_L$ is the $L \times L$ identity matrix, $\mathbf{I}_\infty$ is the countably infinite identity matrix, and $\otimes$ denotes the Kronecker product. We also define

$$L_{ij}^{(\mu)}(x_1, x_2) = \operatorname{Cov}[g_i^{(\mu)}(x_1), g_j^{(\mu)}(x_2)] = \delta_{ij} \operatorname{Cov}[g_1^{(\mu)}(x_1), g_1^{(\mu)}(x_2)],$$

which satisfies the relation

$$K_{ij}^{(\mu+1)}(x_1, x_2) = v_b \delta_{ij} + v_w L_{ij}^{(\mu)}(x_1, x_2).$$

We let $L^{(\mu)}(X, X)$ denote a block matrix where $L_{ij}^{(\mu)}(X, X)$ is the (i, j)-th block. The block matrices $K^{(\mu)}(X, X)$ and $L^{(\mu)}(X, X)$ for $\mu \in \{1, \ldots, D\}$ have infinitely many blocks since the NNGP has infinitely wide hidden layers. However, given a block vector $\alpha \in \mathbb{R}^{T \times |\mathbb{N}|}$ of finite support $\{1, \ldots, T\} \times I$, the quadratic forms $\alpha^\top K^{(\mu)}(X, X)\alpha$ and $\alpha^\top L^{(\mu)}(X, X)\alpha$ are still finite sums:

$$\begin{aligned} \alpha^\top K^{(\mu)}(X, X)\alpha &= \sum_{t,u=1}^{T} \sum_{i,j \in I} \alpha_{ti} \alpha_{uj} K^{(\mu)}(X, X)_{T(i-1)+t, T(j-1)+u} \\ &= \mathrm{V}[\sum_{t=1}^{T} \sum_{i \in I} \alpha_{ti} f_i^{(\mu)}(x_t)] \\ &= \mathrm{V}[f^{(\mu)}(X, \alpha)], \end{aligned}$$

We define a quadratic form for $L^{(\mu)}(X, X)$ similarly.

Next we define the quantities that are at the heart of the proof of Lemma 27. This definition is similar to Definition 7 in Matthews et al., 2018. We discuss the purpose of this definition in more detail in Appendix B.3 in the context of the Central Limit Theorem.

Definition 23 (Projections and summands) *For each $\mu \in \{2, \ldots, D+1\}$ and for each $n \in \mathbb{N}$ and $j \in \{1, \ldots, n\}$, define the summands*

$$\gamma_j^{(\mu)}(X[n], \alpha)[n] = \sqrt{H_{\mu-1}[n]} \sum_{t=1}^{T} \sum_{i \in I} \alpha_{ti} w_{ij}^{(\mu)} g_j^{(\mu-1)}(x_t[n])[n], \tag{27}$$

and the projections

$$\begin{aligned} S^{(\mu)}(X[n], \alpha)[n] &= \sum_{t=1}^{T} \sum_{i \in I} \alpha_{ti} \left(f_i^{(\mu)}(x_t[n])[n] - b_i^{(\mu)} \right) \\ &= \frac{1}{\sqrt{H_{\mu-1}[n]}} \sum_{j=1}^{H_{\mu-1}[n]} \gamma_j^{(\mu)}[n]. \end{aligned} \tag{28}$$

Finally, for $\mu \in \{1, \ldots, D+1\}$, define the variances

$$\sigma^2_{(\mu)}(X[n], \alpha)[n] = \mathrm{V}[\gamma_j^{(\mu)}(X[n], \alpha)[n]] \tag{29}$$

$$\begin{aligned} \sigma^2_{(\mu)}(X, \alpha) &= v_w \alpha^\top L^{(\mu-1)}(X, X) \alpha \\ &= \alpha^\top \left[K^{(\mu)}(X, X) - v_b \mathbf{I} \otimes \mathbf{1}_{T \times T} \right] \alpha. \end{aligned} \tag{30}$$

B.2 Main lemmas and theorem

This section contains the proof of the main theorem (Thm. 4) as well as all original lemmas supporting it. See the proof sketch in Sec. 3.1 for an overview and guide to the logical flow of the lemmas. We start with a lemma that gives sufficient conditions under which the distributional limit of a sequence of compositions of stochastic processes and random variable indices equals the composition of limits.

Lemma 24 (Limit of stochastic process compositions) *Let $\{X[n]\}_{n=1}^{\infty}$ be a sequence of random vectors and X a random vector of dimension B. Let $\{F[n] : \mathbb{R}^B \times \Omega \mapsto \mathbb{R}^L\}_{n=1}^{\infty}$ be a sequence of stochastic processes and $F : \mathbb{R}^B \times \Omega \mapsto \mathbb{R}^L$ a stochastic process with $F(x) \sim \mathcal{N}(0, \Sigma(x))$. If*

(a) $X[n]$ converges in distribution to X, denoted $X[n] \xrightarrow{D} X$,

(b) $\{\mathrm{E}\{|X[n]|^2\}\}_{n=1}^{\infty}$ is eventually bounded,

(c) $\Sigma : \mathbb{R}^B \mapsto \mathbb{R}^{L \times L}$ is a continuous function, and

(d) $F[n] \xrightarrow{UD} F$ on every compact ball in $\mathbb{R}^B$ centered at 0,

then the sequence of random variables $F(X[n])[n] \xrightarrow{D} F(X)$.

Proof We first prove the claim for the case that $F[n]$ and F are real-valued stochastic processes ($L = 1$) and $\Sigma(x) > 0$ for all $x \in \mathbb{R}^B$. For this case, we will use the notation $\sigma^2(x)$ in place of $\Sigma(x)$ to emphasize that $\Sigma(x)$ is a scalar.

Let c be a continuity point of $F(X)$. We want to show that

$$\lim_{n\to\infty} \Pr(F(X[n])[n] < c) = \Pr(F(X) < c).$$

Let $\varepsilon > 0$. We have

$$\begin{aligned}
&|\Pr(F(X[n])[n] < c) - \Pr(F(X) < c)| \\
&= |\Pr(F(X[n])[n] < c) - \Pr(F(X[n]) < c) + \Pr(F(X[n]) < c) - \Pr(F(X) < c)| \\
&\leq |\Pr(F(X[n])[n] < c) - \Pr(F(X[n]) < c)| + |\Pr(F(X[n]) < c) - \Pr(F(X) < c)|.
\end{aligned} \tag{31}$$

We will show that both terms on the right-hand side of Inequality (31) tend to 0.

We start with the second term. Let $\mu[n]$ and μ be the probability distributions associated with $X[n]$ and X, respectively. Then the second term becomes

$$|\Pr(F(X[n]) < c\} - \Pr(F(X) < c\}| = \Big| \int_{\mathbb{R}^B} \Pr(F(x) < c)\, \mathrm{d}\mu(x)[n] - \int_{\mathbb{R}^B} \Pr(F(x) < c)\, \mathrm{d}\mu(x) \Big|.$$

Since $F(x) \sim \mathcal{N}(0, \sigma^2(x))$ with $\sigma^2(x) > 0$, then we have

$$\Pr(F(x) < c) = \Phi\left(\frac{c}{\sigma(x)}\right),$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution. Since σ is continuous, then the map $x \mapsto \Pr(F(x) < c)$ is immediately seen to be continuous as well. Moreover, $\Pr(F(x) < c)$ is clearly bounded. Since $X[n] \xrightarrow{D} X$, then $\{\mu[n]\}_{n=1}^{\infty}$ converges weakly to μ, so that

$$\lim_{n\to\infty} \int_{\mathbb{R}^B} \Pr(F(x) < c)\, \mathrm{d}\mu(x)[n] = \int_{\mathbb{R}^B} \Pr(F(x) < c)\, \mathrm{d}\mu(x).$$

Therefore, there exists an integer N_2 such that

$$|\Pr(F(X[n]) < c) - \Pr(F(X) < c)| < \frac{\varepsilon}{2} \text{ for all } n > N_2. \tag{32}$$

We next bound the first term on the right-hand side of Inequality (31). We have

$$\begin{aligned}
&|\Pr(F(X[n])[n] < c) - \Pr(F(X[n]) < c)| \\
&= \Big| \int_{\mathbb{R}^B} \Pr(F(x)[n] < c)\, \mathrm{d}\mu(x)[n] - \int_{\mathbb{R}^B} \Pr(F(x) < c)\, \mathrm{d}\mu(x)[n] \Big| \\
&\leq \int_{\mathbb{R}^B} |\Pr(F(x)[n] < c) - \Pr(F(x) < c)|\, \mathrm{d}\mu(x)[n].
\end{aligned} \tag{33}$$

We will bound the integrand. Since $\{\mathrm{E}\{|X[n]|^2\}\}_{n=1}^{\infty}$ is eventually bounded, then there exists $V > 0$ and an integer N_V such that

$$\mathrm{E}\{|X[n]|^2\} < V \text{ for all } n > N_V.$$

Define

$$R_\varepsilon = \sqrt{\max\left(0, \frac{2}{\varepsilon}(1+V) - 1\right)}.$$

R_ε is defined such that $\|x\| > R_\varepsilon$ implies

$$\frac{\varepsilon}{2} \cdot \frac{1 + \|x\|^2}{1+V} > 1.$$

We therefore have

$$|\Pr(F[n](x) < c) - \Pr(F(x) < c)| \leq 1 < \frac{\varepsilon}{2} \cdot \frac{1 + \|x\|^2}{1+V} \text{ for all } x \mid \|x\| > R_\varepsilon. \tag{34}$$

Since $F(x)$ follows a normal distribution, then c is trivially a continuity point of $F(x)$ for every $x \in \mathbb{R}^B$. Since $\{F[n]\}_{n=1}^{\infty}$ converges in distribution to F uniformly on every zero-centered compact ball, then there exists an integer $N_1 > N_V$ such that

$$|\Pr(F(x)[n] < c) - \Pr(F(x) < c)| < \frac{\varepsilon}{2} \cdot \frac{1}{1+V} \text{ for all } n > N_1 \text{ and } \|x\| \leq R_\varepsilon.$$

Since $\|x\|^2 \geq 0$, then we have the weaker bound

$$|\Pr(F(x)[n] < c) - \Pr(F(x) < c)| < \frac{\varepsilon}{2} \cdot \frac{1 + \|x\|^2}{1+V} \text{ for all } n > N_1 \text{ and } \|x\| \leq R_\varepsilon.$$

Combining this with Eq. (34) gives

$$|\Pr(F(x)[n] < c) - \Pr(F(x) < c)| < \frac{\varepsilon}{2} \cdot \frac{1 + \|x\|^2}{1+V} \text{ for all } n > N_1 \text{ and } x \in \mathbb{R}^B.$$

Using this bound in Inequality (33), we get

$$\begin{aligned} |\Pr(F(X[n])[n] < c) - \Pr(F(X[n]) < c)| &\leq \int_{\mathbb{R}^B} \frac{\varepsilon}{2} \cdot \frac{1 + \|x\|^2}{1+V} \, \mathrm{d}\mu_n(x) \\ &= \frac{\varepsilon}{2} \cdot \frac{1 + \mathrm{E}\{|X[n]|^2\}}{1+V} \\ &\leq \frac{\varepsilon}{2} \cdot \frac{1+V}{1+V} \\ &= \frac{\varepsilon}{2} \text{ for all } n > N_1. \end{aligned} \tag{35}$$

Let $N = \max(N_1, N_2)$. Combining Inequalities (31), (32), and (35), we obtain the bound

$$|\Pr(F(X[n])[n] < c) - \Pr(F(X) < c)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \text{ for all } n > N,$$

implying $F(X[n])[n] \xrightarrow{D} F(X)$.

Now consider the more general case where the processes $F[n]$ and F take values in $\mathbb{R}^L$ for $L \geq 1$ and where the kernel k of F is not necessarily strictly positive definite. Consider any $\alpha \in \mathbb{R}^L$, and define the processes

$$\hat{F}(x)[n] = Z + \alpha^\top F(x)[n],$$
$$\hat{F}(x) = Z + \alpha^\top F(x),$$

where $Z \sim \mathcal{N}(0,1)$ is independent of $F[n]$ and F. $\hat{F}[n]$ and $\hat{F}$ are real-valued stochastic processes and $\hat{F}(x)$ is normally distributed with variance

$$\hat{\sigma}^2(x) = 1 + \alpha^\top \Sigma(x)\alpha > 0.$$

By the case already proven above, $\hat{F}(X[n])[n] \xrightarrow{D} \hat{F}(X)$. Since the addition of an independent normally distributed random variable Z preserves convergence in distribution, then $\alpha^\top F(X[n])[n] \xrightarrow{D} \alpha F(x)$. Since this holds for any vector α, then by the Cramér-Wold Device (Billingsley, 1995), we obtain the conclusion $F(X[n])[n] \xrightarrow{D} F(X)$. ■

Lemma 24 can be applied inductively to show that a sequence of compositions of stochastic processes converges in distribution to the composition of limit processes. The next lemma verifies condition (b) of Lemma 24.

Lemma 25 (Uniformly bounded neural network variances) *Let $\{F[n]\}_{n=1}^\infty$ be a sequence of BNNs mapping $\mathbb{R}^M$ to $\mathbb{R}^L$ with D hidden layers of widths $H_\mu[n]$, $\mu \in \{1, \ldots, D\}$, and nonlinearity ϕ^μ on the μ-th hidden layer satisfying the linear envelope condition. Then for every $x \in \mathbb{R}^M$, the sequence of second moments $\mathrm{E}[\|F(x)[n]\|^2]\}_{n=1}^\infty$ is uniformly bounded by $A + B\|x\|^2$ for some constants $A, B > 0$.*

Proof Let $x \in \mathbb{R}^M$. The claimed uniform bound on $\mathrm{E}[\|F(x)[n]\|^2]$ holds if we can uniformly bound $\{\mathrm{E}[\|F_i(x)[n]\|^2]\}_{n=1}^\infty$ for each i. Since $F_i(x)[n] = f_i^{(D+1)}(x)[n]$, then we need to establish $\mathrm{E}[f_i^{(D+1)}(x)[n]^2] \leq A + B\|x\|^2$ for sufficiently large n. We proceed by induction on μ. In the case $\mu = 1$, we have

$$\mathrm{E}[f_i^{(1)}(x)[n]] = v_b + v_w\|x\|^2.$$

Taking $A = v_b$ and $B = v_w$, this is clearly bounded by $A + B\|x\|^2$ independently of n. By exchangeability, this same bound holds for all i.

Now suppose for some μ that the claimed uniform bound holds. We then need to establish the bound $\mathrm{E}[f_i^{(\mu+1)}(x)[n]] \leq A + B\|x\|^2$ for some $A, B > 0$. We have

$$\begin{aligned}
\mathrm{E}[f_i^{(\mu+1)}(x)[n]^2] &= v_b + \frac{v_w}{H_\mu[n]} \sum_{j=1}^{H_\mu[n]} \mathrm{E}[g_j^{(\mu)}(x)[n]^2] \\
&= v_b + v_w \, \mathrm{E}[g_i^{(\mu)}(x)[n]^2] \\
&= v_b + v_w \int_{-\infty}^{\infty} \phi^{(\mu)}(z)^2 \, \mathrm{d}\mu_n(z),
\end{aligned}$$

where μ_n is the probability distribution of $f_i^{(\mu)}(x)[n]$. By the linear envelope condition,

$$\begin{aligned}
\mathrm{E}[f_i^{(\mu+1)}(x)[n]^2] &\le v_b + v_w \int_{-\infty}^{\infty} (C + M|z|)^2 \,\mathrm{d}\mu_n(z) \\
&\le v_b + v_w \int_{-\infty}^{\infty} 2(C^2 + M^2|z|^2) \,\mathrm{d}\mu_n(z) \\
&= v_b + 2v_w(C^2 + M^2 \,\mathrm{E}[f_i^{(\mu)}(x)[n]^2]) \\
&\le v_b + 2v_w(C^2 + M^2(A + B\|x\|^2)),
\end{aligned}$$

which is clearly bounded by an expression of the form $A' + B'\|x\|^2$ independently of n for some $A', B' > 0$. The claim then follows by induction. ■

The next lemma verifies condition (c) of Lemma 24, which amounts to showing that the NNGP kernel is continuous.

Lemma 26 (Continuity of batch NNGP kernel) *Let $F : \mathbb{R}^M \times \Omega \mapsto \mathbb{R}^L$ be an NNGP with D hidden layers and nonlinearity ϕ that satisfies the linear envelope condition. Then the associated batch NNGP $\tilde{F} : (\mathbb{R}^M)^T \times \Omega \mapsto (\mathbb{R}^L)^T$ of size T has marginal $\tilde{F}(X) \sim \mathcal{N}(0, \Sigma(X))$, where the batch NNGP kernel $\Sigma : (\mathbb{R}^M)^T \mapsto \mathbb{R}^{LT \times LT}$ given by $\Sigma(X) = K(X, X)$ is a continuous function.*

Proof All we need to show is that Σ is a continuous function. Since $\Sigma(X) = K(X, X) = k(X, X) \otimes I_L$, then it is sufficient to show that the NNGP kernel $k : \mathbb{R}^M \times \mathbb{R}^M \mapsto \mathbb{R}$ is continuous. We do so inductively by showing that $k^{(\mu)}$ is continuous for $\mu \in \{1, \ldots, D+1\}$, where $k = k^{(D+1)}$.

For $\mu = 1$, $k^{(1)}(x, x') = v_b + v_w x \cdot x'$ is clearly continuous. Now suppose for some $\mu \in \{1, \ldots, D\}$, that $k^{(\mu)}$ is continuous. We then need to show that $k^{(\mu+1)}$ is continuous. Let $\{X_n = (x_n, x_n') \in \mathbb{R}^M \times \mathbb{R}^M\}_{n=1}^{\infty}$ be a convergent sequence of pairs of inputs such that $X_n \to X = (x, x')$. Since $k^{(\mu)}$ is continuous, then $k^{(\mu)}(X_n, X_n) \to k^{(\mu)}(X, X)$. Let $s_i(X_n)$ (resp. $s_i(X)$) denote the i-th column of the symmetric positive semidefinite square root $S(X_n)$ (resp. $S(X)$) of $k^{(\mu)}(X_n, X_n)$ (resp. $k^{(\mu)}(X, X)$). Then by the work of Cho and Saul (2009), the kernel recursion in Eq. (7) can be expressed as

$$k^{(\mu+1)}(x_n, x_n') = v_b + v_w \frac{1}{2\pi} \int_{\mathbb{R}^2} \phi(w^\top s_1(X_n)) \phi(w^\top s_2(X_n)) \, e^{-\frac{\|w\|^2}{2}} \,\mathrm{d}w,$$

and $k^{(\mu+1)}(x, x')$ is given similarly. To ensure continuity, we will show that $k^{(\mu+1)}(x_n, x_n') \to k^{(\mu+1)}(x, x')$. We do so by verifying the conditions of the Dominated Convergence Theorem.

First, by the linear envelope condition, there exist positive constants C and M such that

$$\begin{aligned}
\phi(w^\top s_1(X_n))\phi(w^\top s_2(X_n)) &\leq [C + Mw^\top s_1(X_n)][C + Mw^\top s_2(X_n)] \\
&= C^2 + CM\sum_{i=1}^{2} w^\top s_i(X_n) + M^2 w^\top s_1(X_n) w^\top s_2(X_n) \\
&\leq C^2 + CM\sum_{i=1}^{2} w^\top s_i(X_n) + \frac{M^2}{2}\sum_{i=1}^{2}(w^\top s_i(X_n))^2 \\
&= \frac{1}{2}\left[(C + Mw^\top s_1(X_n))^2 + (C + Mw^\top s_2(X_n))^2\right] \\
&\leq [C^2 + M^2(w^\top s_1(X_n))^2] + [C^2 + M^2(w^\top s_2(X_n))^2] \\
&= 2C^2 + M^2\|S^\top(X_n)w\|^2 \\
&\leq 2C^2 + M^2\|S^\top(X_n)\|_2^2\|w\|^2 \\
&\leq 2C^2 + M^2\|S^\top(X_n)\|_F^2\|w\|^2 \\
&= 2C^2 + M^2\|S(X_n)\|_F^2\|w\|^2,
\end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Since the matrix square root operation is continuous under the Frobenius norm, then $\|S(X_n)\|_F$ is bounded by some $B > 0$. We therefore have

$$\phi(w^\top s_1(X_n))\phi(w^\top s_2(X_n)) \leq 2C^2 + M^2B^2\|w\|^2,$$

hence

$$\frac{1}{2\pi}\phi(w^\top s_1(X_n))\phi(w^\top s_2(X_n))e^{-\frac{\|w\|^2}{2}} \leq \frac{1}{2\pi}(2C^2 + M^2B^2\|w\|^2)e^{-\frac{\|w\|^2}{2}},$$

where the bound on the right-hand side is clearly integrable over $w \in \mathbb{R}^2$. Moreover, since the matrix square root operation is continuous and ϕ is continuous, then

$$\frac{1}{2\pi}\phi(w^\top s_1(X_n))\phi(w^\top s_2(X_n))e^{-\frac{\|w\|^2}{2}} \to \frac{1}{2\pi}\phi(w^\top s_1(X))\phi(w^\top s_2(X))e^{-\frac{\|w\|^2}{2}} \text{ pointwise in } w. \tag{36}$$

Therefore, by the Dominated Convergence Theorem, $k^{(\mu+1)}(x_n, x_n') \to k^{(\mu+1)}(x, x')$ so that $k^{(\mu+1)}$ is continuous. The continuity of the kernel k then follows by induction. ■

The next lemma will help to verify condition (d) of Lemma 24. It is a generalization of Lemma 12 in Matthews et al., 2018 and depends on several additional lemmas (Appendices B.3-B.4) similar to those in Matthews et al., 2018.

Lemma 27 (Continuous convergence in distribution of batch BNNs) *Consider a sequence $\{F[n]\}_{n=1}^{\infty}$ of BNNs mapping $\mathbb{R}^M$ to $\mathbb{R}^L$ with D hidden layers of widths $H_\mu[n]$, $\mu \in \{1, \ldots, D\}$, and nonlinearity ϕ that satisfies the linear envelope condition. Let F be the NNGP limit of the BNNs as given by Thm. 3. Then for any $T \in \mathbb{N}$, the corresponding sequence of batch BNNs $\{\tilde{F}[n]\}_{n=1}^{\infty}$ converges in distribution to the batch NNGP $\tilde{F}$ continuously.*

Proof For each $n \in \mathbb{N}$, let $X[n] = \{x_t[n]\}_{t=1}^T \in (\mathbb{R}^M)^T$ be a batch of inputs such that the sequence of batches $\{X[n]\}_{n=1}^\infty$ converges to some finite $X \in (\mathbb{R}^M)^T$. We need to show that $\{\tilde{F}[n]\}_{n=1}^\infty$ converges in distribution to $\tilde{F}$ continuously, i.e. that the sequence of random variables $\{\tilde{F}(X[n])[n]\}_{n=1}^\infty$ converges in distribution to $\tilde{F}(X)$ and thus

$$\{f^{(D+1)}(x_t[n])[n]\}_{t=1}^T \xrightarrow{D} \{f^{(D+1)}(x_t)\}_{t=1}^T. \tag{37}$$

We will do so by establishing $\{f^{(\mu)}(x_t[n])[n]\}_{t=1}^T \xrightarrow{D} \{f^{(\mu)}(x_t)\}_{t=1}^T$ inductively for every $\mu \in \{1, \ldots, D+1\}$.

For the case $\mu = 1$, let $\alpha \in \mathbb{R}^{T \times |\mathbb{N}|}$ with finite support $\{1, \ldots, T\} \times I$.

By definition (Eq. (1)), it is straightforward to verify that

$$f^{(1)}(X[n], \alpha)[n] \sim \mathcal{N}(0, \alpha^\top K^{(1)}(X[n], X[n])\alpha)$$
$$f^{(1)}(X, \alpha) \sim \mathcal{N}(0, \alpha^\top K^{(1)}(X, X)\alpha).$$

Let c be a continuity point of $f^{(1)}(X, \alpha)$ so that $c \neq 0$ if $\alpha^\top K^{(1)}(X, X)\alpha = 0$. Extend the CDF Φ of the standard normal distribution by setting $\Phi(-\infty) = 0$ and $\Phi(\infty) = 1$. Then the map

$$z \mapsto \Phi\left(\frac{c}{\sqrt{z}}\right) \tag{38}$$

is continuous on $(0, \infty)$ and is right-continuous at $z = 0$ if $c \neq 0$. Now since the kernel is a continuous function (Lemma 26) and since $X[n] \to X$, then $\alpha^\top K^{(1)}(X[n], X[n])\alpha \to \alpha^\top K^{(1)}(X, X)\alpha$. Moreover, since we just established that the map given by Eq. (38) is continuous, then it follows that

$$\Phi\left(\frac{c}{\sqrt{\alpha^\top K^{(1)}(X[n], X[n])\alpha}}\right) \to \Phi\left(\frac{c}{\sqrt{\alpha^\top K^{(1)}(X, X)\alpha}}\right) \text{ as } n \to \infty,$$

and hence $f^{(1)}(X[n], \alpha)[n] \xrightarrow{D} f^{(1)}(X, \alpha)$. By the Cramér-Wold Device, we deduce that $f_I^{(1)}(X[n])[n] \xrightarrow{D} f_I^{(1)}(X)$.

Now suppose that $f_I^{(\mu)}(X[n])[n] \xrightarrow{D} f_I^{(\mu)}(X)$ for every finite subset $I \subseteq \mathbb{N}$ and for some $\mu \in \{1, \ldots, D\}$. We then want to show that this same convergence holds for $\mu + 1$. Let $\alpha \in \mathbb{R}^{T \times |\mathbb{N}|}$ with finite support $\{1, \ldots, T\} \times I$. We view α as a block vector where α_{ti} is the t-th element in the i-th block. By Lemmas 31-33, the sequence of summands $\{\gamma_j^{(\mu+1)}(X[n], \alpha)\}_{j=1}^{H_\mu[n]}$ for $n \in \mathbb{N}$ satisfies the conditions of Thm. 30; condition 1 is immediate since the weights $w_{i1}^{(\mu+1)}$ and $w_{j2}^{(\mu+1)}$ are independent and have mean 0. Theorem 30 then tells us that the projections $S^{(\mu+1)}(X[n], \alpha)[n] \xrightarrow{D} \mathcal{N}(0, \sigma^2(X, \alpha))$, where the limiting variance is given by $\sigma^2_{(\mu+1)}(X, \alpha)$ (Eq. (30)). By the Cramér-Wold Device, this implies

$$f_I^{(\mu+1)}(X[n])[n] - b_I^{(\mu+1)} \otimes \mathbf{1}_T \xrightarrow{D} \mathcal{N}(0, v_w L_{II}^{(\mu)}(X, X)),$$

which in turn implies

$$F_I^{(\mu+1)}(X[n])[n] \xrightarrow{D} F_I^{(\mu+1)}(X) \sim \mathcal{N}(0, K_{II}^{(\mu+1)}(X, X)).$$

Equation (37) then follows by induction, thus establishing continuous distributional convergence. ■

Next is the proof of the bottleneck NNGP theorem, which is the main theorem of our paper.

Proof [Proof of Thm. 4] We proceed by induction on $d \in \{1, \ldots, D\}$. The case $d = 1$ is given to us by Thm. 3 (i.e., no hidden bottlenecks). Now suppose the claim holds for some $d \in \{1, \ldots, D-1\}$. We will prove the claim for the case $d+1$.

Let $X = \{x_t\}_{t=1}^T$ be a finite subset of $\mathcal{X}$. Define the random variables

$$Z[n] = \{(F^{(d)}[n] \circ \cdots \circ F^{(1)}[n])(x_t)\}_{t=1}^T,$$
$$Z = \{(F^{(d)} \circ \cdots \circ F^{(1)})(x_t)\}_{t=1}^T.$$

Let $\tilde{F}^{(d+1)}[n]$ (resp. $\tilde{F}^{(d+1)}$) be the batch BNN (resp. batch NNGP) corresponding to $F^{(d+1)}[n]$ (resp. $F^{(d+1)}$), and observe that

$$\tilde{F}^{(d+1)}(Z[n])[n] = \{(F^{(d+1)}[n] \circ \cdots \circ F^{(1)}[n])(x_t)\}_{t=1}^T,$$
$$\tilde{F}^{(d+1)}(Z) = \{(F^{(d+1)} \circ \cdots \circ F^{(1)})(x_t)\}_{t=1}^T.$$

We proceed to establish the four conditions of Lemma 24 in order to prove

$$\tilde{F}^{(d+1)}(Z[n])[n] \xrightarrow{D} \tilde{F}^{(d+1)}(Z). \tag{39}$$

By the inductive hypothesis, $F^{(d)}[n] \circ \cdots \circ F^{(1)}[n] \xrightarrow{D} F^{(d)} \circ \cdots \circ F^{(1)}$ in $((\mathbb{R}^L)^\infty, \mathcal{A})$ and thus in particular $Z[n] \xrightarrow{D} Z$, establishing condition (a). Observe that

$$\mathrm{E}[\|Z[n]\|^2] = \sum_{t=1}^T \mathrm{E}[\|(f^{(d)}[n] \circ \cdots \circ F^{(1)}[n])(x_t)\|^2].$$

Since a composition of BNNs is still a BNN (with some hidden layers having linear activation), then we can apply Lemma 25 to each expectation in the sum to get the bound

$$\mathrm{E}[\|Z[n]\|^2] \leq \sum_{t=1}^T (A_t + B_t \|x_t\|^2),$$

for some constants $A_t, B_t > 0$. In other words, the sequence of second moments of $\{Z[n]\}_{n=1}^\infty$ is bounded, establishing condition (b). Lemma 26 gives us condition (c). Finally, Lemma 27 tells us that $\tilde{F}^{(d+1)}[n] \xrightarrow{CD} \tilde{F}^{(d+1)}$. By Prop. 22, we immediately have $\tilde{F}^{(d+1)}[n] \xrightarrow{UD} \tilde{F}^{(d+1)}$ on every compact subset of $\mathbb{R}^{T \times B_d}$, establishing condition (d).

Having verified its four conditions, Lemma 24 implies Eq. (39) and hence

$$\{(F^{(d+1)}[n] \circ \cdots \circ F^{(1)}[n])(x_t)\}_{t=1}^T \xrightarrow{D} \{(F^{(d+1)} \circ \cdots \circ F^{(1)})(x_t)\}_{t=1}^T.$$

Since this holds for any T inputs in $\mathcal{X}$, then by Thm. 18 the desired convergence in $((\mathbb{R}^L)^\infty, \mathcal{A})$ follows. ■

Remark 28 (Nonlinear bottleneck) *Theorem 4 holds even if we replace* $F^{(d)}[n]$ *and* $F^{(d)}$ *with* $F^{(d)}[n] \circ \left(\frac{1}{\sqrt{B_{d-1}}}\phi\right)$ *and* $F^{(d)} \circ \left(\frac{1}{\sqrt{B_{d-1}}}\phi\right)$ *respectively for* $d \in \{2, \ldots, D\}$*. The proof is nearly identical, making the necessary replacements where appropriate. The only additional step needed is to verify condition (c) of Lemma 24 for* $\tilde{F}^{(d+1)}[n] \circ \left(\frac{1}{\sqrt{B_d}}\phi\right)$ *in the inductive step; by Lemma 24,* $\tilde{F}^{(d+1)}[n] \xrightarrow{UD} \tilde{F}^{(d+1)}$ *and hence* $\tilde{F}^{(d+1)}[n] \xrightarrow{CD} \tilde{F}^{(d+1)}$ *by Prop. 22. Now since* $x \mapsto \frac{1}{\sqrt{B_d}}\phi(x)$ *is (sequentially) continuous, then* $\tilde{F}^{(d+1)}\left(\frac{1}{\sqrt{B_d}}\phi(x_n)\right) \to \tilde{F}^{(d+1)}\left(\frac{1}{\sqrt{B_d}}\phi(x)\right)$ *whenever* $x_n \to x$*. By Prop. 22,* $\tilde{F}^{(d+1)}[n] \circ \left(\frac{1}{\sqrt{B_d}}\phi\right) \xrightarrow{UD} \tilde{F}^{(d+1)} \circ \left(\frac{1}{\sqrt{B_d}}\phi\right)$*, establishing condition (c).*

Remark 29 (Discontinuous nonlinearity) *Theorem 4 holds even if the nonlinearity* $\phi : \mathbb{R} \mapsto \mathbb{R}$ *is continuous only almost everywhere (AE), as long as* ϕ *is continuous at* 0 *or* $v_b > 0$*. If* ϕ *is continuous AE, then the pointwise convergence in Eq.* (36) *holds AE, which is still sufficient for the Dominated Convergence Theorem. Moreover, the Continuous Mapping Theorem used in Lemmas 31-33 is still applicable as long as the set of discontinuities of* ϕ *has measure* 0 *with respect to the distribution of the NNGP preactivation* $f_i^{(\mu)}(x)$*. If* $v_b = 0$*, then it becomes possible for the distribution of* $f_i^{(\mu)}(x)$ *to degenerate to a delta distribution concentrated at* 0*; if* ϕ *is also discontinuous at* 0*, then its set of discontinuities will have measure* 1 *with respect to the delta distribution, hence the requirement that* $v_b > 0$ *if* ϕ *is discontinuous at* 0*.*

B.3 Verifying the conditions of the CLT for exchangeable processes

The results in this section serve to support the proof of Lemma 27. Since Lemma 27 is similar to Lemma 12 in Matthews et al., 2018, then the results in this section are also similar to results in Matthews et al., 2018. The approach to proving Lemma 27 is to show that in the (no-bottleneck) NNGP limit, if the preactivations into one hidden layer converge in distribution continuously to a GP, then so do the preactivations into the next hidden layer. This is done using a special central limit theorem. The challenge is that the preactivations into any hidden layer after the first hidden layer are independent only in the wide limit. Moreover, the distribution of each preactivation changes as the preceding hidden layer grows in width. The following is a central limit theorem adapted specifically for this case; it is a restatement of Lemma 10 in Matthews et al., 2018, which is in turn an adaptation of a central limit theorem for exchangeable processes by Blum et al. (1958).

Theorem 30 (CLT for sequences of exchangeable sequences (Matthews et al., 2018)) *For each positive integer* n*, let* $\{X_i[n]\}_{i=1}^{\infty}$ *be an exchangeable sequence of random variables with mean* 0*, variance* $\sigma^2[n]$*, and finite absolute third moment. Suppose also that the variances converge to the limit* $\lim_{n\to\infty} \sigma^2[n] = \sigma^2$*. If*

(a) $E\{X_1[n]X_2[n]\} = 0$*,*

(b) $\lim_{n\to\infty} E\{X_1[n]^2 X_2[n]^2\} = \sigma^4$*, and*

(c) $E\{|X_1[n]|^3\} = o(\sqrt{n})$

then for any strictly increasing sequence H, the sequence of standardized partial sums $\{S[n]\}_{n=1}^{\infty}$ with

$$S[n] = \frac{1}{\sqrt{H[n]}} \sum_{i=1}^{H[n]} X_i[n]$$

converges in distribution to $\mathcal{N}(0, \sigma^2)$, where $\mathcal{N}(0, 0)$ is interpreted as the constant 0.

We will apply Thm. 30 to the summands $\gamma_j^{(\mu)}(X[n], \alpha)[n]$ (Eq. (27)) to show that the projection $S^{(\mu)}(X[n], \alpha)[n]$ (Eq. (28)) converges to a GP. This requires us to verify the conditions of Thm. 30. We verify the existence of the limit $\lim_{n\to\infty} \sigma^2[n] = \sigma^2$ first. The following lemma is analogous to Lemma 11 in Matthews et al., 2018. The main difference is that the batch input X is replaced with a convergent sequence of input batches $\{X[n]\}_{n=1}^{\infty}$. We maintain the notation introduced in Sec. B.1.

Lemma 31 *Suppose that $f_I^{(\mu)}(X[n])[n] \xrightarrow{D} f_I^{(\mu)}(X)$ for some $\mu \in \{1, \ldots, D\}$, and for every finite set $I \subset \mathbb{N}$. Then*

$$\lim_{n\to\infty} \sigma_{\mu+1}^2(X[n], \alpha)[n] = \sigma_{\mu+1}^2(X, \alpha),$$

where these variances are defined in Eqs. (29)-(30)*.*

Proof It is clear that $\mathrm{E}[\gamma_j^{(\mu+1)}(X[n], \alpha)[n]] = 0$ since the weights $w_{ij}^{(\mu+1)}$ have 0 mean. We therefore have

$$\begin{aligned}
\sigma_{\mu+1}^2(X[n], \alpha)[n] &= \mathrm{E}[\gamma_1^{(\mu+1)}(X[n], \alpha)[n]^2] \\
&= \mathrm{E}\left[\left(\sqrt{H_\mu[n]} \sum_{t=1}^{T} \sum_{i\in I} \alpha_{ti} w_{i1}^{(\mu)} g_1^{(\mu)}(x_t[n])[n]\right)^2\right] \\
&= H_\mu[n] \sum_{t,u=1}^{T} \sum_{i,j\in I} \alpha_{ti}\alpha_{uj} \,\mathrm{E}[w_{i1}^{(\mu)} w_{j1}^{(\mu)}] \,\mathrm{E}[g_1^{(\mu)}(x_t[n])[n] g_1^{(\mu)}(x_u[n])[n]] \\
&= v_w \sum_{t,u=1}^{T} \sum_{i,j\in I} \alpha_{ti}\alpha_{uj}\delta_{ij} \,\mathrm{E}[g_1^{(\mu)}(x_t[n])[n] g_1^{(\mu)}(x_u[n])[n]].
\end{aligned}$$

Theorem 3.5 in Billingsley (1999) tells us that a limit can be moved inside an expectation operator if the sequence inside the expectation converges in distribution and is uniformly integrable. Since the preactivations $f_1^{(\mu)}(X[n])[n]$ converge in distribution and since the nonlinearity ϕ and multiplication mapping $\mathbb{R}^2$ to $\mathbb{R}$ are continuous functions, then the Continuous Mapping Theorem implies that the products of activations in the above expectations also converge in distribution. Uniform integrability holds by Cor. 39. We therefore have

the limit

$$\begin{aligned}\lim_{n\to\infty}\sigma^2_{\mu+1}(X[n],\alpha)[n] &= v_w\sum_{t,u=1}^{T}\sum_{i,j\in I}\alpha_{ti}\alpha_{uj}\delta_{ij}\,\mathrm{E}[g_1^{(\mu)}(x_t)g_1^{(\mu)}(x_u)]\\ &= v_w\sum_{t,u=1}^{T}\sum_{i,j\in I}\alpha_{ti}\alpha_{uj}L_{11}^{(\mu)}(x_t,x_u)\\ &= v_w\alpha^\top L^{(\mu)}(X,X)\alpha\\ &= \sigma^2_{(\mu+1)}(X,\alpha),\end{aligned}$$

completing the proof. ■

Condition (a) of Thm. 30 is easily verified directly in the proof of Lemma 27. We thus move to condition (b). The following lemma is analogous to Lemma 15 in Matthews et al., 2018.

Lemma 32 *Suppose that* $f_I^{(\mu)}(X[n])[n] \xrightarrow{D} f_I^{(\mu)}(X)$ *for some* $\mu\in\{1,\ldots,D\}$. *and for every finite set* $I\subset\mathbb{N}$. *Then*

$$\lim_{n\to\infty}\mathrm{E}[\gamma_1^{(\mu+1)}(X[n],\alpha)[n]^2\gamma_2^{(\mu+1)}(X[n],\alpha)[n]^2] = \sigma_\mu^4(X,\alpha).$$

Proof We proceed in direct analogy to the proof of Lemma 31. We have

$$\begin{aligned}&\mathrm{E}[\gamma_1^{(\mu+1)}(X[n],\alpha)[n]^2\gamma_2^{(\mu+1)}(X[n],\alpha)[n]^2]\\ &= \mathrm{E}\left[\left(\sqrt{H_\mu[n]}\sum_{r=1}^{T}\sum_{i\in I}w_{i1}^{(\mu)}g_1^{(\mu)}(x_r[n])[n]\right)^2\left(\sqrt{H_\mu[n]}\sum_{t=1}^{T}\sum_{k\in I}w_{k2}^{(\mu)}g_2^{(\mu)}(x_t[n])[n]\right)^2\right]\\ &= H_\mu^2[n]\sum_{r,s,t,u=1}^{T}\sum_{i,j,k,\ell\in I}\Big(\alpha_{ri}\alpha_{sj}\alpha_{tk}\alpha_{u\ell}\cdot\mathrm{E}[w_{i1}^{(\mu)}w_{j1}^{(\mu)}]\cdot\mathrm{E}[w_{k2}^{(\mu)}w_{\ell2}^{(\mu)}]\\ &\cdot\mathrm{E}[g_1^{(\mu)}(x_r[n])[n]g_1^{(\mu)}(x_s[n])[n]g_2^{(\mu)}(x_t[n])[n]g_2^{(\mu)}(x_u[n])[n]]\Big)\\ &= v_w^2\sum_{r,s,t,u=1}^{T}\sum_{i,j,k,\ell\in I}\Big(\alpha_{ri}\alpha_{sj}\alpha_{tk}\alpha_{u\ell}\delta_{ij}\delta_{k\ell}\\ &\cdot\mathrm{E}[g_1^{(\mu)}(x_r[n])[n]g_1^{(\mu)}(x_s[n])[n]g_2^{(\mu)}(x_t[n])[n]g_2^{(\mu)}(x_u[n])[n]]\Big).\end{aligned}$$

Since the preactivations $f_I^{(\mu)}(X[n])[n]$ converge in distribution for $I=\{1,2\}$, and since the nonlinearity ϕ and multiplication from $\mathbb{R}^4$ to $\mathbb{R}$ are continuous functions, then the Continuous Mapping Theorem implies the four-way products of activations in each expectation above converge in distribution as well. Corollary 38 also tells us that the set of these four-way products of activations is uniformly integrable. By Theorem 3.5 in Billingsley (1999),

we have the limit

$$\begin{aligned}&\lim_{n\to\infty} \mathrm{E}[\gamma_1^{(\mu+1)}(X[n],\alpha)[n]^2\gamma_2^{(\mu+1)}(X[n],\alpha)[n]^2]\\&= v_w^2 \sum_{r,s,t,u=1}^{T} \sum_{i,j,k,\ell\in I} \alpha_{ri}\alpha_{sj}\alpha_{tk}\alpha_{u\ell}\delta_{ij}\delta_{k\ell}\,\mathrm{E}[g_1^{(\mu)}(x_r)g_1^{(\mu)}(x_s)g_2^{(\mu)}(x_t,\alpha)g_2^{(\mu)}(x_u,\alpha)].\end{aligned} \tag{40}$$

Since parallel activations in a layer decorrelate in an NNGP, then we have

$$\begin{aligned}&\lim_{n\to\infty} \mathrm{E}[\gamma_1^{(\mu+1)}(X[n],\alpha)[n]^2\gamma_2^{(\mu+1)}(X[n],\alpha)[n]^2]\\&= v_w^2 \sum_{r,s,t,u=1}^{T} \sum_{i,j,k,\ell\in I} \alpha_{ri}\alpha_{sj}\alpha_{tk}\alpha_{u\ell}\delta_{ij}\delta_{k\ell}\,\mathrm{E}[g_1^{(\mu)}(x_r)g_1^{(\mu)}(x_s)]\,\mathrm{E}[g_2^{(\mu)}(x_t)g_2^{(\mu)}(x_u)]\\&= \left(v_w \sum_{r,s=1}^{T} \sum_{i,j\in I} \alpha_{ri}\alpha_{sj}\delta_{ij}\,\mathrm{E}[g_1^{(\mu)}(x_r)g_1^{(\mu)}(x_s)]\right)\\&\cdot \left(v_w \sum_{t,u=1}^{T} \sum_{k,\ell\in I} \alpha_{tk}\alpha_{u\ell}\delta_{k\ell}\,\mathrm{E}[g_2^{(\mu)}(x_t)g_2^{(\mu)}(x_u)]\right)\\&= v_w\alpha^\top L^{(\mu)}(X,X)\alpha v_w\alpha^\top L^{(\mu)}(X,X)\alpha\\&= \sigma^4_{(\mu+1)}(X,\alpha),\end{aligned}$$

completing the proof. ■

Finally, we verify condition (c) of Thm. 30. The following lemma is analogous to Lemma 16 in Matthews et al., 2018.

Lemma 33 *Suppose that* $f_I^{(\mu)}(X[n])[n] \xrightarrow{D} f_I^{(\mu)}(X)$ *for some* $\mu \in \{1,\dots,D\}$*, and for every finite set* $I \subset \mathbb{N}$*. Then*

$$\mathrm{E}[|\gamma_1^{(\mu+1)}(X[n],\alpha)[n]|^3] = o(\sqrt{n}).$$

Proof We will prove the stronger result that the third absolute moment of $\gamma_1^{(\mu+1)}(X[n],\alpha)[n]$ is bounded over n. By Hölder's inequality,

$$\mathrm{E}[|\gamma_1^{(\mu+1)}(X[n],\alpha)[n]|^3] \le \mathrm{E}[\gamma_1^{(\mu+1)}(X[n],\alpha)[n]^4]^{\frac{3}{4}}.$$

Thus, to bound the left side independently of n, it is sufficient to do the same for the fourth moment of $\gamma_1^{(\mu+1)}(X[n],\alpha)[n]$. Observe that

$$\mathrm{E}[\gamma_1^{(\mu+1)}(X[n],\alpha)[n]^4]^{\frac{3}{4}} = \mathrm{E}[\gamma_1^{(\mu+1)}(X[n],\alpha)[n]^2\gamma_1^{(\mu+1)}(X[n],\alpha)[n]^2],$$

where the right-hand side is similar to the quantity discussed in Lemma 32. Therefore, calculations proceed in direct analogy to the proof of Lemma 32 up to and including Eq. (40).

Thus, we have

$$\lim_{n\to\infty} \mathrm{E}[\gamma_1^{(\mu+1)}(X[n],\alpha)[n]^4]^{\frac{3}{4}}$$
$$= v_w^2 \sum_{r,s,t,u=1}^{T} \sum_{i,j,k,\ell\in I} \alpha_{ri}\alpha_{sj}\alpha_{tk}\alpha_{u\ell}\delta_{ij}\delta_{k\ell}\,\mathrm{E}[g_1^{(\mu)}(x_r,\alpha)g_1^{(\mu)}(x_s,\alpha)g_1^{(\mu)}(x_t,\alpha)g_1^{(\mu)}(x_u,\alpha)].$$

The right-hand side can be shown to be finite by applying Lemma 35 to bound the expectation of the four-way product by a product of eighth moments, applying the linear envelope property to obtain bounds in terms of preactivations, and finally noting that the eighth moment of a normal distribution is finite; this gives us the desired bound on the fourth and hence third absolute moment. ■

B.4 Establishing uniform integrability

The results in this section serve to support the proofs in Appendix B.3. As in Appendix B.3, the results in this appendix are stronger versions of results appearing in Matthews et al., 2018. The key results in this section are Lemma 35 and Cors. 38 and 39 and are the only ones referenced outside of this section.

Lemma 34 *Let X be a random variable. Then $E[X^4] \leq E[X^8]^{\frac{1}{2}}$.*

Proof By Hölder's Inequality, we have

$$E[X^4] = E[X^4 \cdot 1] \leq E[(X^4)^2]^{\frac{1}{2}} E[1^2]^{\frac{1}{2}} = E[X^8]^{\frac{1}{2}}.$$

■

The following lemma is a stronger version of Lemma 18 in Matthews et al., 2018. Matthews et al., 2018 proves that the expectation $E\left[\prod_{i=1}^4 |X_i|^{p_i}\right]$ is uniformly bounded by a polynomial in the eighth moments $E[X_i^8] < \infty$ for $i \in \{1,2,3,4\}$ without specifying the polynomial. Lemma 35 below provides the explicit bound $\prod_{i=1}^4 E[X_i^8]^{\frac{p_i}{8}}$, which is a polynomial in the eighth moments. This bound is important when proving uniform convergence with respect to the inputs of a random neural network, since the coefficients and exponents in the bound are independent of the network's input.

Lemma 35 *Let X_i be random variables with $E[X_i^8] < \infty$ for $i \in \{1,2,3,4\}$. Then for any choice of $p_i \in \{0,1,2\}$ it holds that*

$$E\left[\prod_{i=1}^4 |X_i|^{p_i}\right] \leq \prod_{i=1}^4 E[X_i^8]^{\frac{p_i}{8}}.$$

Proof Using Hölder's inequality twice, we have

$$\begin{aligned}
E[|X_1|^{p_1}|X_2|^{p_2}|X_3|^{p_3}|X_4|^{p_4}] &\leq E[(|X_1|^{p_1}|X_2|^{p_2})^2]^{\frac{1}{2}} E[(|X_3|^{p_3}|X_4|^{p_4})^2]^{\frac{1}{2}} \\
&= E[X_1^{2p_1} X_2^{2p_2}]^{\frac{1}{2}} E[X_3^{2p_3} X_4^{2p_4}]^{\frac{1}{2}} \\
&\leq \left(E[(X_1^{2p_1})^2]^{\frac{1}{2}} E[(X_2^{2p_2})^2]^{\frac{1}{2}}\right)^{\frac{1}{2}} \left(E[(X_3^{2p_3})^2]^{\frac{1}{2}} E[(X_4^{2p_4})^2]^{\frac{1}{2}}\right)^{\frac{1}{2}} \\
&= E[X_1^{4p_1}]^{\frac{1}{4}} E[X_2^{4p_2}]^{\frac{1}{4}} E[X_3^{4p_3}]^{\frac{1}{4}} E[X_4^{4p_4}]^{\frac{1}{4}} \\
&= \prod_{i=1}^{4} E[X_i^{4p_i}]^{\frac{1}{4}}.
\end{aligned}$$

If $p_i = 0$, then $E[X_i^{4p_i}]^{\frac{1}{4}} = E[1]^{\frac{1}{4}} = 1$, which can be written as $E[X_i^8]^0$. If $p_i = 1$, then by Lemma 34, $E[X_i^{4p_i}]^{\frac{1}{4}} \leq E[X_i^8]^{\frac{1}{8}}$. If $p_i = 2$, then we simply have $E[X_i^{4p_i}]^{\frac{1}{4}} = E[X_i^8]^{\frac{2}{8}}$. We therefore see that for any $p_i \in \{0, 1, 2\}$, $E[X_i^{4p_i}]^{\frac{1}{4}} = E[X_i^8]^{\frac{p_i}{8}}$. Substituting this into the above product yields the desired bound. ■

The following lemma extends Lemma 20 in Matthews et al., 2018 to stochastic processes in the sense that the input into the BNN is now a variable. We can achieve a uniform bound if we assume that the input space is compact.

Lemma 36 *Let $\mathcal{X} \subset \mathbb{R}^M$ be a compact input space. Then for each $\mu \in \{1, \ldots, D+1\}$, the eighth moments of the normally distributed random variables $f_i^{(\mu)}(x)[n]$ defined by equation (1) are uniformly bounded over all $i \in \{1, \ldots, h_\mu(n)\}$, $n \in \mathbb{N}$ and $x \in \mathcal{X}$.*

Proof We proceed by induction on μ. The case $\mu = 1$ is trivial; the random variables $f_i^{(1)}(x)[n]$ are IID over i and follow the normal distribution $\mathcal{N}(0, v_b^{(1)} + v_w^{(1)}\|x\|^2)$. The eighth moments are therefore

$$E[f_i^{(1)}(x)[n]^8] = 105(v_b^{(1)} + v_w^{(1)}\|x\|^2)^4.$$

Clearly the eighth moment is independent of i and n. Moreover, since $\mathcal{X}$ is compact, then $\sup_{x \in \mathcal{X}} E[f_i^{(1)}(x)[n]^8] < \infty$. The eighth moments are therefore uniformly bounded over i, n and x.

Now assume that the eighth moments of $f_i^{(\mu)}(x)[n]$ are uniformly bounded over i, n and x for all $\mu \in \{1, \ldots, t-1\}$ and for some $t \in \{2, \ldots, D+1\}$. We wish to prove that the eighth moments of $f_i^{(t)}(x)[n]$ are uniformly bounded over i, n and x. Using the inequality $|u(x)+v(x)|^p \leq 2^{p-1}(|u(x)|^p + |v(x)|^p)$ for elements u and v of the L^p space for $p \geq 1$, which follows from the convexity of $h(x) := x^p$ for $p > 1$, the bound

$$E[f_i^{(t)}(x)[n]^8] \leq 2^7 E\left[(b_i^{(t)})^8 + \left(\sum_{j=1}^{h_{t-1}(n)} w_{ij}^{(t)} g_j^{(t)}(x)[n]\right)^8\right]$$

is first established. The term $E[(b_i^{(t)})^8]$ is bounded since the biases are normally distributed. Moreover, the biases are IID over i and are independent of n. Therefore, to achieve the

desired uniform bound, we only need to show that the term

$$S_i(x)[n] := E\left[\left(\sum_{j=1}^{h_{t-1}(n)} w_{ij}^{(t)} g_j^{(t-1)}(x)[n]\right)^8\right]$$

is uniformly bounded over i, n and x. By Lemma 20 in Matthews et al., 2018,

$$S_i(x)[n] \leq \frac{1}{h_{t-1}(n)^4} E\left[\left(\sum_{i=1}^{h_{t-1}(n)} (c^2 + 2cm|f_i^{(t-1)}(x)[n]| + m^2|f_i^{(t-1)}(x)[n]|^2)\right)^4\right],$$

where $c, m > 0$ are constants from the linear envelope property of the activation function. Letting $a = \max\{c^2, 2cm, m^2\}$ and multiplying out the quantity in the above expectation, we have

$$\begin{aligned}
S_i(x)[n] &\leq \frac{a^4}{h_{t-1}(n)^4} E\left[\sum_{i,j,k,\ell=1}^{h_{t-1}(n)} \sum_{p,q,r,s=0}^{2} |f_i^{(t-1)}(x)[n]|^p \cdot |f_j^{(t-1)}(x)[n]|^q \right. \\
&\quad \left. \cdot |f_k^{(t-1)}(x)[n]|^r \cdot |f_\ell^{(t-1)}(x)[n]|^s\right] \\
&= \frac{a^4}{h_{t-1}(n)^4} \sum_{i,j,k,\ell=1}^{h_{t-1}(n)} \sum_{p,q,r,s=0}^{2} E\left[|f_i^{(t-1)}(x)[n]|^p \cdot |f_j^{(t-1)}(x)[n]|^q \right. \\
&\quad \left. \cdot |f_k^{(t-1)}(x)[n]|^r \cdot |f_\ell^{(t-1)}(x)[n]|^s\right].
\end{aligned}$$

Using Lemma 35 and the fact that the moments of $f_i^{(t-1)}(x)[n]$ are independent of i by exchangeability, we have

$$\begin{aligned}
S_i(x)[n] &\leq \frac{a^4}{h_{t-1}(n)^4} \sum_{i,j,k,\ell=1}^{h_{t-1}(n)} \sum_{p,q,r,s=0}^{2} \left(E[f_i^{(t-1)}(x)[n]^8]^{\frac{p}{8}} \cdot E[f_j^{(t-1)}(x)[n]^8]^{\frac{q}{8}} \right. \\
&\quad \left. \cdot E[f_k^{(t-1)}(x)[n]^8]^{\frac{r}{8}} \cdot E[f_\ell^{(t-1)}(x)[n]^8]^{\frac{s}{8}}\right) \\
&= \frac{a^4}{h_{t-1}(n)^4} \sum_{i,j,k,\ell=1}^{h_{t-1}(n)} \sum_{p,q,r,s=0}^{2} E[f_1^{(t-1)}(x)[n]^8]^{\frac{p+q+r+s}{8}} \\
&= \frac{a^4}{h_{t-1}(n)^4} \cdot h_{t-1}(n)^4 \sum_{p,q,r,s=0}^{2} E[f_1^{(t-1)}(x)[n]^8]^{\frac{p+q+r+s}{8}} \\
&= a^4 \sum_{p,q,r,s=0}^{2} E[f_1^{(t-1)}(x)[n]^8]^{\frac{p+q+r+s}{8}} \\
&= a^4 \sum_{j=1}^{81} E[f_1^{(t-1)}(x)[n]^8]^{m_j},
\end{aligned}$$

where each m_j is a rational number between 0 and 1. Define the function

$$\psi(z) = a^4 \sum_{j=1}^{81} z^{m_j},$$

and note that a and the m_j are independent of the hidden width index n, the hidden neuron index i, and the input x. Moreover, ψ is increasing on the interval $(0, \infty)$. Since we assumed as our inductive hypothesis that $E[f_i^{(t-1)}(x)[n]^8] < \infty$, then it follows that

$$\sup_{i,n,x} S_i(x)[n] \leq \psi(\sup_{i,n,x} E[f_1^{(t-1)}(x)[n]^8]) < \infty,$$

implying that $E[f_i^{(t)}(x)[n]^8] < \infty$ uniformly over i n, and x, thereby completing the proof. ■

The following lemma extends Lemma 21 in Matthews et al., 2018 to stochastic processes in the same sense as Lemma 36 above.

Lemma 37 *Let $\mathcal{X} \subset \mathbb{R}^M$ be a compact input space. Then for any $\mu \in \{1, \ldots, D+1\}$ and indices $i, j, k, \ell \in \mathbb{N}$, the set of random variables*

$$S := \{g_i^{(\mu)}(x_1)[n] g_j^{(\mu)}(x_2)[n] g_k^{(\mu)}(x_3)[n] g_\ell^{(\mu)}(x_4)[n] : n \in \mathbb{N} \text{ and } x_1, x_2, x_3, x_4 \in \mathcal{X}\}$$

is uniformly integrable.

Proof By the de la Vallée-Poussin Theorem (Meyer, 1966, p.19, Theorem T22), S is uniformly integrable if

$$\sup_{n,x_1,x_2,x_3,x_4} E[|g_i^{(\mu)}(x_1)[n] g_j^{(\mu)}(x_2)[n] g_k^{(\mu)}(x_3)[n] g_\ell^{(\mu)}(x_4)[n]|^{1+\varepsilon}] < \infty \text{ for some } \varepsilon > 0.$$

We consider $\varepsilon = 1$. By Lemma 35,

$$\begin{aligned}
&E[|g_i^{(\mu)}(x_1)[n] g_j^{(\mu)}(x_2)[n] g_k^{(\mu)}(x_3)[n] g_\ell^{(\mu)}(x_4)[n]|^2] \\
&\leq E[g_i^{(\mu)}(x_1)[n]^8]^{\frac{1}{4}} E[g_j^{(\mu)}(x_2)[n]^8]^{\frac{1}{4}} E[g_k^{(\mu)}(x_3)[n]^8]^{\frac{1}{4}} E[g_\ell^{(\mu)}(x_4)[n]^8]^{\frac{1}{4}} \\
&= \prod_{q=1}^{4} E[g_1^{(\mu)}(x_q)[n]^8]^{\frac{1}{4}},
\end{aligned}$$

where we obtained the last line by exchangeability over the indices i, j, k, ℓ. We therefore have

$$\begin{aligned}
\sup_{n,x_1,x_2,x_3,x_4} E[|g_i^{(\mu)}(x_1)[n] g_j^{(\mu)}(x_2)[n] g_k^{(\mu)}(x_3)[n] g_\ell^{(\mu)}(x_4)[n]|^2] &\leq \prod_{q=1}^{4} \sup_{n,x_q} E[g_1^{(\mu)}(x_q)[n]^8]^{\frac{1}{4}} \\
&= \prod_{q=1}^{4} \sup_{n,x} E[g_1^{(\mu)}(x)[n]^8]^{\frac{1}{4}} \\
&= \sup_{n,x} E[g_1^{(\mu)}(x)[n]^8].
\end{aligned}$$

It thus suffices to show that the supremum in the last line is finite. By the linear envelope property of the activation function,

$$E[g_1^{(\mu)}(x)[n]^8] \leq 2^7 \left(c^8 + m^8 E[f_1^{(\mu)}(x)[n]^8] \right).$$

By Lemma 36, the right-hand side is uniformly bounded over all $n \in \mathbb{N}$ and $x \in \mathcal{X}$, completing the proof. ■

Cors. 38 and 39 of Lemma 37, below, are used in Lemmas 32 and 31, respectively.

Corollary 38 *Let $\{x_q[n] \in \mathbb{R}^M\}_{n=1}^{\infty}$ for $q \in \{1, \ldots, 4\}$ be four convergent sequences with finite limits. Then for any $\mu \in \{1, \ldots, D+1\}$ and indices $i, j, k, \ell \in \mathbb{N}$, the set of random variables*

$$S = \{g_i^{(\mu)}(x_1[n])[n] g_j^{(\mu)}(x_2[n])[n] g_k^{(\mu)}(x_3[n])[n] g_\ell^{(\mu)}(x_4[n])[n] : n \in \mathbb{N}\}$$

is uniformly integrable.

Proof Since the sequences $\{x_q[n]\}_{n=1}^{\infty}$ for $q \in \{1, \ldots, 4\}$, converge to finite limits, then there exists a compact set $\mathcal{X} \subset \mathbb{R}^M$ that contains $x_q[n]$ for all n and q. By (the proof of) Lemma 37, we have

$$\sup_{n, x_1, x_2, x_3, x_4} E[|g_i^{(\mu)}(x_1)[n] g_j^{(\mu)}(x_2)[n] g_k^{(\mu)}(x_3)[n] g_\ell^{(\mu)}(x_4)[n]|^2] < \infty,$$

where the x_q are elements of the compact set $\mathcal{X}$. It then holds in particular that

$$\sup_{n} E[|g_i^{(\mu)}(x_1[n])[n] g_j^{(\mu)}(x_2[n])[n] g_k^{(\mu)}(x_3[n])[n] g_\ell^{(\mu)}(x_4[n])[n]|^2] < \infty.$$

Uniform integrability then follows by the de la Vallee-Poussin Theorem. ■

Corollary 39 *Let $\{x_q[n] \in \mathbb{R}^M\}_{n=1}^{\infty}$ for $q \in \{1, 2\}$, be two convergent sequences with finite limits. Then for any $\mu \in \{1, \ldots, D+1\}$ and indices $i, j \in \mathbb{N}$, the set of random variables*

$$S = \{g_i^{(\mu)}(x_1[n])[n] g_j^{(\mu)}(x_2[n])[n] : n \in \mathbb{N}\}$$

is uniformly integrable.

Proof By the de la Vallée-Poussin Theorem, S is uniformly integrable if

$$\sup_{n, x_1, x_2, x_3, x_4} E[|g_i^{(\mu)}(x_1)[n] g_j^{(\mu)}(x_2)[n]|^{1+\varepsilon}] < \infty \text{ for some } \varepsilon > 0.$$

We consider $\varepsilon = 3$. We have

$$\begin{aligned} &\sup_{n, x_1, x_2, x_3, x_4} E[|g_i^{(\mu)}(x_1)[n] g_j^{(\mu)}(x_2)[n]|^4] \\ &= \sup_{n, x_1, x_2, x_3, x_4} E[g_i^{(\mu)}(x_1)[n]^2 g_i^{(\mu)}(x_1)[n]^2 g_j^{(\mu)}(x_2)[n]^2 g_j^{(\mu)}(x_2)[n]^2], \end{aligned}$$

which is finite by (the proof of) Cor. 38. The claim then follows. ■

Appendix C. Correspondence to the no-bottleneck NNGP

The following is our proof of the Wide Bottleneck Correspondence Theorem for the case of a single-bottleneck NNGP.

Proof [Proof of Thm. 8] First we prove statement (b). We will do so for $L = 1$; the case $L > 1$ proceeds similarly. Let $X = \{x_t\}_{t=1}^T$ be a finite set of inputs. Let $p_H : \mathbb{R}^T \mapsto \mathbb{R}$ and $p : \mathbb{R}^T \mapsto \mathbb{R}$ be the PDFs of $F^{(H)}$ and F, respectively. Let $k^{(D_1)} : \mathbb{R}^M \times \mathbb{R}^M \mapsto \mathbb{R}$ and $k^{(D_2)} : \mathbb{R}^H \times \mathbb{R}^H \mapsto \mathbb{R}$ be the NNGP kernels of the pre-bottleneck and post-bottleneck components (with respective depths D_1 and D_2) of $F^{(H)}$; note that the kernels are independent of H. Then the PDF p_H is given by

$$p_H(y) = \int_{(\mathbb{R}^T)^H} \left(\mathcal{N}\left(y; 0, k^{(D_2)}\left(\frac{1}{\sqrt{H}}\phi(\{h_i\}_{i=1}^H), \frac{1}{\sqrt{H}}\phi(\{h_j\}_{j=1}^H)\right) + v_n \mathbf{I}_T\right) \cdot \prod_{k=1}^H \mathcal{N}(h_k; 0, k^{(D_1)}(X, X)) \right) \mathrm{d}h_k, \tag{41}$$

where $h_i \in \mathbb{R}^T$ is the vector of preactivations into the i-th hidden neuron in the bottleneck, and where we use the notation $\mathcal{N}(z; \mu, \Sigma)$ to mean the normal PDF in the variable z with mean μ and covariance Σ. Observing that the NNGP kernel in Eqs. (6)-(7) depends on its inputs only through their Gram matrix and writing the kernel of the first layer explicitly, we can define a function $\tilde{k}^{(D_2)}$ on the space of symmetric positive semidefinite matrices such that

$$\tilde{k}^{(D_2)}(v_b + v_w A) = k^{(D_2)}(B, B), \quad A = BB^\top.$$

Defining the random $T \times T$ matrix

$$Z_H = \frac{1}{H}\sum_{i=1}^H \phi(h_i)\phi(h_i)^\top, \quad h_i \sim \mathcal{N}(0, k^{(D_1)}(X, X)) \text{ IID}, \tag{42}$$

and letting μ_H denote the probability measure associated with Z_H, the PDF in Eq. (41) can be written as

$$p_H(y) = \int_{\mathbb{R}^{T\times T}} \mathcal{N}(y; 0, \tilde{k}^{(D_2)}(v_b + v_w z) + v_n \mathbf{I}_T)\, \mathrm{d}\mu_H(z).$$

Here z is a dummy variable. Now since the h_i in Eq. (42) are IID, then so are the matrices $\phi(h_i)\phi(h_i)^\top$. Therefore, Z_H is an empirical average of H IID random matrices. By the Law of Large Numbers, we have

$$\{Z_H\}_{H=1}^\infty \xrightarrow{P} Z = \mathrm{E}_{h\sim\mathcal{N}(0,k^{(D_1)}(X,X))}[\phi(h)\phi(h)^\top],$$

where the convergence is in probability. In particular, $\{Z_H\}_{H=1}^\infty \xrightarrow{D} Z$ so that the sequence of measures $\{\mu_H\}_{H=1}^\infty$ weakly converges to the probability measure μ associated with Z. Note that μ is a delta distribution concentrated at Z. Furthermore, thanks to the Gaussian noise, the function $z \to \mathcal{N}(y; 0, \tilde{k}^{(D_2)}(z) + v_n \mathbf{I}_T)$ is bounded over $\mathbb{R}^{T\times T}$; it is continuous as well, as the matrix inversion and determination operations and the NNGP

kernel (Lemma 26) are all continuous. By the weak convergence of measures and the delta distribution μ, we have

$$\begin{aligned}
\lim_{H\to\infty} p_H(y) &= \lim_{H\to\infty} \int_{\mathbb{R}^{T\times T}} \mathcal{N}(y; 0, \tilde{k}^{(D_2)}(v_b + v_w z) + v_n \mathbf{I}_T)\, \mathrm{d}\mu_H(z) \\
&= \int_{\mathbb{R}^{T\times T}} \mathcal{N}(y; 0, \tilde{k}^{(D_2)}(v_b + v_w z) + v_n \mathbf{I}_T)\, \mathrm{d}\mu(z) \\
&= \mathcal{N}\left(y; 0, \tilde{k}^{(D_2)}\left(v_b + v_w \,\mathrm{E}_{h\sim\mathcal{N}(0,k^{(D_1)}(X,X)}[\phi(h)\phi(h)^\top]\right) + v_n \mathbf{I}_T\right) \\
&= \mathcal{N}\left(y; 0, \tilde{k}^{(D_2)}\left(k^{(D_1+1)}(X,X)\right) + v_n \mathbf{I}_T\right) \\
&= \mathcal{N}(y; 0, k^{(D_1+D_2+1)}(X,X) + v_n \mathbf{I}_T) \\
&= p(y),
\end{aligned}$$

which is the PDF of an NNGP with $D_1 + D_2 + 1$ hidden layers.

To prove statement (a) of the theorem, we first note that the pointwise convergence $p_H \to p$ ensures the convegence in distribution $F^{(H)}(X) \xrightarrow{D} F(X)$ according to Scheffé's Lemma. Since this holds for any finite set of inputs X and in particular any finite subset of a countable set $\mathcal{X} \subset \mathbb{R}^M$, then by Thm. 18, we have that $\{F^{(H)}\}_{H=1}^{\infty} \xrightarrow{D} F$ in $((\mathbb{R}^L)^\infty, \mathcal{A})$ for inputs restricted to $\mathcal{X}$ as claimed. ■

Appendix D. Bottleneck layers induce correlation

Recall the single-bottleneck NNGP F defined in Sec. 4.1. Each output $(F_i(x_1), F_i(x_2))$ conditional on the activations of the bottleneck layer follow the two-dimensional normal distribution $\mathcal{N}(0, K)$ where

$$K = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix}.$$

It can be shown that the diagonal entries of K are given by

$$\begin{aligned}
k_{aa} &= b_D + \frac{w_D}{H} \sum_{i=1}^{H} \phi(h_i^a)^2, \\
b_D &= v_n + v_b \sum_{d=0}^{D-1} v_w^d, \\
w_D &= v_w^D.
\end{aligned}$$

The expression for the off-diagonals k_{12}, k_{21} will not be important.

The proof of Prop. 10 regarding the quadratic correlation between bottleneck NNGP outputs follows.

Proof [Proof of Prop. 10] We have

$$\begin{aligned}
\mathrm{E}[F_1(x_a)^2] &= \int_{(\mathbb{R}^2)^H} \int_{(\mathbb{R}^2)^2} (y_1^a)^2 \mathcal{N}(y_1; 0, K)\mathcal{N}(y_2; 0, K) \prod_{m=1}^{H} \mathcal{N}(h_m; 0, C) \,\mathrm{d}y \,\mathrm{d}h \\
&= \int_{(\mathbb{R}^2)^H} k_{aa} \prod_{m=1}^{H} \mathcal{N}(h_m; 0, C) \,\mathrm{d}h \\
&= \int_{(\mathbb{R}^2)^H} \left(b_D + \frac{w_D}{H} \sum_{i=1}^{H} \phi(h_i^a)^2 \right) \prod_{m=1}^{H} \mathcal{N}(h_m; 0, C) \,\mathrm{d}h \\
&= b_D + \frac{w_D}{H} \sum_{i=1}^{H} \int_{(\mathbb{R}^2)^H} \phi(h_i^a)^2 \prod_{m=1}^{H} \mathcal{N}(h_m; 0, C) \,\mathrm{d}h \\
&= b_D + \frac{w_D}{H} \sum_{i=1}^{H} \int_{(\mathbb{R}^2)^H} \phi(h_i^a)^2 \mathcal{N}(h_i; 0, C) \,\mathrm{d}h_i \\
&= b_D + \frac{w_D}{H} \sum_{i=1}^{H} E_{z \sim \mathcal{N}(0, c_{aa})}[\phi(z)^2] \\
&= b_D + w_D E_{z \sim \mathcal{N}(0, c_{aa})}[\phi(z)^2].
\end{aligned} \tag{43}$$

We similarly have

$$\mathrm{E}[F_2(x_b)^2] = b_D + w_D E_{z \sim \mathcal{N}(0, c_{bb})}[\phi(z)^2]. \tag{44}$$

Combining Eqs. (43) and (44) yields

$$\begin{aligned}
&\mathrm{E}[F_1(x_a)^2] \,\mathrm{E}[F_2(x_b)^2] \\
&= b_D^2 + b_D w_D E_{z \sim \mathcal{N}(0, c_{aa})}[\phi(z)^2] + b_D w_D E_{z \sim \mathcal{N}(0, c_{bb})}[\phi(z)^2] \\
&+ E_{z \sim \mathcal{N}(0, c_{aa})}[\phi(z)^2] E_{z \sim \mathcal{N}(0, c_{bb})}[\phi(z)^2].
\end{aligned}$$

We also have

$$
\begin{aligned}
&\mathrm{E}[F_1(x_a)^2 F_2(x_b)^2] \\
&= \int_{(\mathbb{R}^2)^H} \int_{(\mathbb{R}^2)^2} \prod_{y=y_1^a, y_2^b} y^2 \mathcal{N}(y; 0, K) \prod_{m=1}^{H} \mathcal{N}(h_m; 0, C) \,\mathrm{d}y \,\mathrm{d}h \\
&= \int_{(\mathbb{R}^2)^H} k_{aa} k_{bb} \prod_{m=1}^{H} \mathcal{N}(h_m; 0, C) \,\mathrm{d}h \\
&= \int_{(\mathbb{R}^2)^H} \prod_{h=h_i^a, h_j^b} \left(b_D + \frac{w_D}{H} \sum_{j=1}^{H} \phi(h)^2 \right) \prod_{m=1}^{H} \mathcal{N}(h_m; 0, C) \,\mathrm{d}h \\
&= b_D^2 + b_D w_D E_{z\sim\mathcal{N}(0,c_{aa})}[\phi(z)^2] + b_D w_D E_{z\sim\mathcal{N}(0,c_{bb})}[\phi(z)^2] \\
&+ \frac{w_D^2}{H^2} \sum_{i,j=1}^{H} \int_{(\mathbb{R}^2)^2} \phi(h_i^a)^2 \phi(h_j^b)^2 \mathcal{N}(h_i; 0, C) \mathcal{N}(h_j; 0, C) \,\mathrm{d}h_i \,\mathrm{d}h_j \\
&= \mathrm{E}[F_1(x_a)^2] \,\mathrm{E}[F_2(x_b)^2] - E_{z\sim\mathcal{N}(0,c_{aa})}[\phi(z)^2] E_{z\sim\mathcal{N}(0,c_{bb})}[\phi(z)^2] \\
&+ \frac{w_D^2}{H^2} \sum_{i=1}^{H} \int_{\mathbb{R}^2} \phi(h_i^a)^2 \phi(h_i^b)^2 \mathcal{N}(h_i; 0, C) \,\mathrm{d}h_i \\
&+ \frac{w_D^2}{H^2} \sum_{i\neq j=1}^{H} \int_{(\mathbb{R}^2)^2} \phi(h_i^a)^2 \phi(h_j^b)^2 \mathcal{N}(h_i; 0, C) \mathcal{N}(h_j; 0, C) \,\mathrm{d}h_i \,\mathrm{d}h_j \\
&= \mathrm{E}[F_1(x_a)^2] \,\mathrm{E}[F_2(x_b)^2] - E_{z\sim\mathcal{N}(0,c_{aa})}[\phi(z)^2] E_{z\sim\mathcal{N}(0,c_{bb})}[\phi(z)^2] \\
&+ \frac{w_D^2}{H} E_{(z,z')\sim\mathcal{N}(0,C)}[\phi(z)^2 \phi(z')^2] \\
&+ w_D^2 \left(1 - \frac{1}{H}\right) E_{z\sim\mathcal{N}(0,c_{aa})}[\phi(z)^2] E_{z\sim\mathcal{N}(0,c_{bb})}[\phi(z)^2] \\
&= \mathrm{E}[F_1(x_a)^2] \,\mathrm{E}[F_2(x_b)^2] + \frac{w_D^2}{H} E_{(z,z')\sim\mathcal{N}(0,C)}[\phi(z)^2 \phi(z')^2] \\
&- \frac{w_D^2}{H} E_{z\sim\mathcal{N}(0,c_{aa})}[\phi(z)^2] E_{z\sim\mathcal{N}(0,c_{bb})}[\phi(z)^2] \\
&= \mathrm{E}[F_1(x_a)^2] \,\mathrm{E}[F_2(x_b)^2] + \frac{w_D^2}{H} \operatorname{Cov}_{(z,z')\sim\mathcal{N}(0,C)}[\phi(z)^2, \phi(z')^2].
\end{aligned}
$$

We therefore have

$$
\operatorname{Cov}[F_1(x_a)^2, F_2(x_b)^2] = \frac{w_D^2}{H} \operatorname{Cov}_{(z,z')\sim\mathcal{N}(0,C)}[\phi(z)^2, \phi(z')^2].
$$

On the right-hand side, we have the covariance of two rectified quadratic units with respect to the Gaussian measure $\mathcal{N}(0, C)$. By the work of Cho and Saul (2009) and by adjusting for differences in normalization, we have the expectations

$$
E_{z\sim\mathcal{N}(0,c_{aa})}[\phi(z)^2] = c_{aa}, \tag{45}
$$

$$
E_{(z,z')\sim\mathcal{N}(0,C)}[\phi(z)^2 \phi(z')^2] = \frac{2}{\pi} c_{aa} c_{bb} J_2(\beta), \tag{46}
$$

where $\beta = \cos^{-1}\left(\frac{c_{ab}}{\sqrt{c_{aa}c_{bb}}}\right)$ and $J_2(\beta) = 3\sin\beta\cos\beta + (\pi - \beta)(1 + 2\cos^2\beta)$. Using also the fact that $w_D = v_w^D$, we obtain

$$\begin{aligned}\operatorname{Cov}[F_1(x_a)^2, F_2(x_b)^2] &= \frac{(v_w^D)^2}{H}\left(\frac{2}{\pi}c_{aa}c_{bb}J_2(\beta) - c_{aa}c_{bb}\right) \\ &= \frac{v_w^{2D}c_{aa}c_{bb}}{H}\left(\frac{2}{\pi}J_2(\beta) - 1\right),\end{aligned}$$

establishing Eq. (11).

The corresponding correlation is defined as

$$q_{ab}^{\times} = \frac{\operatorname{Cov}[F_1(x_a)^2, F_2(x_b)^2]}{\sqrt{\operatorname{V}[F_1(x_a)^2]\operatorname{V}[F_2(x_b)^2]}}. \tag{47}$$

We already know the numerator on the right-hand side, but we need to calculate the variances in the denominator. Using the fact that $F_1(x_a)$ and $F_2(x_b)$ are identically (but not independently) distributed, we have

$$\begin{aligned}&\operatorname{V}[F_1(x_a)^2] \\ &= \operatorname{E}[F_1(x_a)^4] - \operatorname{E}[F_1(x_a)^2]^2 \\ &= \int_{(\mathbb{R}^2)^H}\int_{(\mathbb{R}^2)^2}(y_1^a)^4\prod_{i=1}^{2}\mathcal{N}(y_i;0,K)\prod_{m=1}^{H}\mathcal{N}(h_m;0,C)\,\mathrm{d}y\,\mathrm{d}h - \operatorname{E}[F_1(x_a)^2]^2 \\ &= 3\int_{(\mathbb{R}^2)^H}k_{aa}k_{aa}\prod_{m=1}^{H}\mathcal{N}(h_m;0,C)\,\mathrm{d}h - \operatorname{E}[F_1(x_a)^2]^2 \\ &= 3\operatorname{E}[F_1(x_a)^2, F_2(x_a)^2] - \operatorname{E}[F_1(x_a)^2]\operatorname{E}[F_2(x_a)^2] \\ &= 3\operatorname{Cov}[F_1(x_a)^2, F_2(x_a)^2] + 3\operatorname{E}[F_1(x_a)^2]\operatorname{E}[F_2(x_a)^2] - \operatorname{E}[F_1(x_a)^2]\operatorname{E}[F_2(x_a)^2] \\ &= 3\operatorname{Cov}[F_1(x_a)^2, F_2(x_a)^2] + 2\operatorname{E}[F_1(x_a)^2]\operatorname{E}[F_2(x_a)^2] \\ &= 3\operatorname{Cov}[F_1(x_a)^2, F_2(x_a)^2] + 2\operatorname{E}[F_1(x_a)^2]^2.\end{aligned}$$

By Eqs. (43), (45), and (11), we have

$$\begin{aligned}\operatorname{V}[F_1(x_a)^2] &= \frac{3w_D^2c_{aa}c_{aa}}{H}\left(\frac{2}{\pi}J_2(0) - 1\right) + 2\left(b_D + w_Dc_{aa}\right)^2 \\ &= \frac{3w_D^2(c_{aa})^2}{H}\left(\frac{2\cdot 3\pi}{\pi} - 1\right) + 2\left(b_D + w_Dc_{aa}\right)^2 \\ &= \frac{15w_D^2(c_{aa})^2}{H} + 2\left(b_D + w_Dc_{aa}\right)^2.\end{aligned} \tag{48}$$

We similarly have

$$\operatorname{V}[F_2(x_b)^2] = \frac{15w_D^2(c_{bb})^2}{H} + 2\left(b_D + w_Dc_{bb}\right)^2.$$

Substituting these variances into Eq. (47), we obtain

$$\begin{aligned} q_{ab}^{\times} &= \frac{\frac{w_D^2 c_{aa} c_{bb}}{H}\left(\frac{2}{\pi}J_2(\beta)-1\right)}{\prod\limits_{c=c_{aa},c_{bb}}\sqrt{\frac{15w_D^2c^2}{H}+2\left(b_D+w_Dc\right)^2}} \\ &= \frac{\left(\frac{2}{\pi}J_2(\beta)-1\right)}{\prod\limits_{c=c_{aa},c_{bb}}\sqrt{\frac{H}{w_Dc^2}\left(\frac{15w_D^2c^2}{H}+2\left(b_D+w_Dc\right)^2\right)}} \\ &= \frac{\left(\frac{2}{\pi}J_2(\beta)-1\right)}{\prod\limits_{c=c_{aa},c_{bb}}\sqrt{15+2H\left(\frac{r_D}{c}+1\right)^2}}, \end{aligned}$$

where

$$\begin{aligned} r_D &= \frac{b_D}{w_D} \\ &= \frac{v_n}{v_w^D}+\frac{v_b}{v_w^D}\sum_{d=0}^{D-1}v_w^d \\ &= \frac{v_n}{v_w^D}+v_b\sum_{d=1}^{D}\frac{1}{v_w^d} \\ &= \begin{cases} v_n+Dv_b & \text{if } v_w=1 \\ \frac{v_n}{v_w^D}+\frac{v_b}{1-v_w}\left(\frac{1}{v_w^D}-1\right) & \text{otherwise,} \end{cases} \end{aligned}$$

establishing Eqs. (12) and (13). ■

Proof [Proof of Prop. 12] For part (a), Eq. (16) follows by substitution of Eq. (15) into Eq. (12).

For part (b), the map $G \mapsto Q^{\times(\infty)}$ is a composition of the maps $G \mapsto C$ and $C \mapsto Q^{\times(\infty)}$, and thus it suffices to show that these two maps are invertible. The map $G \mapsto C$ sends the input Gram matrix to the NNGP kernel at the bottleneck layer. Inverting Eq. (9) for the case $i = j$, we obtain the recursion for the backward propagation of the NNGP kernel:

$$K_{ii}^{(\mu-1)}(x_1,x_2) = \frac{1}{v_w}\left(\prod_{a=1,2}\sqrt{K_{ii}^{(\mu)}(x_a,x_a)-v_b}\right)\cos J_1^{-1}\left(\pi\frac{K_{ii}^{(\mu)}(x_1,x_2)-v_b}{\prod_{a=1,2}\sqrt{K_{ii}^{(\mu)}(x_a,x_a)-v_b}}\right),$$

where we note J_1 is strictly decreasing on $[0, \pi]$. Applying this recursion to C d times (where d is the depth of the pre-bottleneck NNGP) gives $K_{ii}^{(1)}(x_1, x_2)$, and by solving Eq. (6) we obtain G. Thus, $G \mapsto C$ is invertible.

To show $C \mapsto Q^{\times(\infty)}$ is invertible, we inspect Eq. (16) for the case $v_w > 1$ in Prop. 12 and observe that $q_{aa}^{\times(\infty)}$ depends only on c_{aa} (the bottleneck angle β is 0 when the two inputs

are identical). We may then solve for c_{aa}. Substituting c_{aa} for $a \in \{1, 2\}$ into Eq. (16) and noting that J_2 is strictly decreasing, we may solve for the bottleneck angle β from $q_{12}^{\times(\infty)}$ and thus obtain c_{12}, recovering C. ■

Proof [Proof of Prop. 13] Since $\|x_1\| = \|x_2\|$, then $c_{11} = c_{22}$. Letting $c = c_{11} = c_{22}$, we have by Eq. (12) that

$$q_{ab}^{\times(D)} = \frac{\left(\frac{2}{\pi} J_2(\beta) - 1\right)}{15 + 2H\left(\frac{r_D}{c} + 1\right)^2}.$$

We will find a $\lambda > 0$ such that the limit L in Eq. (17) is finite and non-zero. Note that while evaluating the limit, we will drop (non-zero) constants of proportionality. Observing that both the numerator and denominator inside the limit L in Eq. (17) tend to 0 as $D \to \infty$ and using L'Hôpital's rule, we have

$$\begin{aligned}
L &= \lim_{D\to\infty} \frac{\frac{\mathrm{d}}{\mathrm{d}D} q_{ab}^{\times(D)} - \frac{\mathrm{d}}{\mathrm{d}D} q_{ab}^{\times(\infty)}}{\frac{\mathrm{d}}{\mathrm{d}D} e^{-\frac{D}{\lambda}}} \\
&= \lim_{D\to\infty} \frac{\frac{\mathrm{d}}{\mathrm{d}D} \frac{\left(\frac{2}{\pi} J_2(\beta) - 1\right)}{15 + 2H\left(\frac{r_D}{c} + 1\right)^2} - 0}{-\frac{1}{\lambda} e^{-\frac{D}{\lambda}}} \\
&\propto -\lim_{D\to\infty} e^{\frac{D}{\lambda}} \cdot \frac{\mathrm{d}}{\mathrm{d}D} \frac{1}{15 + 2H\left(\frac{r_D}{c} + 1\right)^2} \\
&= \lim_{D\to\infty} e^{\frac{D}{\lambda}} \cdot \frac{4H\left(\frac{r_D}{c} + 1\right) \frac{1}{c} \cdot \left(v_n + \frac{v_b}{1 - v_w}\right) \cdot \frac{1}{v_w^D} \ln\left(\frac{1}{v_w}\right)}{\left[15 + 2H\left(\frac{r_D}{c} + 1\right)^2\right]^2} \\
&\propto \lim_{D\to\infty} \frac{e^{\frac{D}{\lambda}}}{v_w^D} \cdot \frac{\left(\frac{r_D}{c} + 1\right)}{\left[15 + 2H\left(\frac{r_D}{c} + 1\right)^2\right]^2}.
\end{aligned} \tag{49}$$

In the case $v_w > 1$, r_D tends to a finite positive limit as $D \to \infty$, so that the second fraction in the limit in Eq. (49) tends to a finite positive limit as well. We therefore have

$$L \propto \lim_{D\to\infty} \frac{e^{\frac{D}{\lambda}}}{v_w^D},$$

which is finite and non-zero (and exists) if and only if $\frac{e^{\frac{1}{\lambda}}}{v_w} = 1$, implying $\lambda = \ln(v_w)^{-1}$.

In the case $v_w < 1$, $r_D \to \infty$ as $D \to \infty$, so that Eq. (49) simplifies to

$$\begin{aligned} L &\propto \lim_{D\to\infty} \frac{e^{\frac{D}{\lambda}}}{v_w^D} \cdot \frac{\left(\frac{r_D}{c}\right)}{\left[2H\left(\frac{r_D}{c}\right)^2\right]^2} \\ &\propto \lim_{D\to\infty} \frac{e^{\frac{D}{\lambda}}}{v_w^D} \cdot \frac{1}{r_D^3} \\ &= \lim_{D\to\infty} \frac{e^{\frac{D}{\lambda}}}{v_w^D} \cdot \frac{1}{\left[\frac{v_n}{v_w^D} + \frac{v_b}{1-v_w}\left(\frac{1}{v_w^D} - 1\right)\right]^3} \\ &= \lim_{D\to\infty} \frac{e^{\frac{D}{\lambda}}}{v_w^D} \cdot \frac{1}{\left[\left(v_n + \frac{v_b}{1-v_w}\right)\frac{1}{v_w^D} - \frac{v_b}{1-v_w}\right]^3}. \end{aligned} \tag{50}$$

Since $\frac{1}{v_w^D} \to \infty$ as $D \to \infty$, then

$$L \propto \lim_{D\to\infty} \frac{e^{\frac{D}{\lambda}}}{v_w^D} \cdot \frac{1}{\left[\left(v_n + \frac{v_b}{1-v_w}\right)\frac{1}{v_w^D}\right]^3} \propto \lim_{D\to\infty} v_w^{2D} e^{\frac{D}{\lambda}},$$

which is finite and non-zero (and exists) if and only if $v_w^2 e^{\frac{1}{\lambda}} = 1$, implying $\lambda = \ln\left(\frac{1}{v_w^2}\right)^{-1}$.

In the case $v_w = 1$, we again have $r_D \to \infty$ as $D \to \infty$ and thus still obtain Eq. (50). Substituting $v_w = 1$ and $r_D = v_n + v_b D$ into Eq. (50) gives

$$L \propto \lim_{D\to\infty} \frac{e^{\frac{D}{\lambda}}}{(v_n + v_b D)^3},$$

which is infinite for all finite $\lambda > 0$. ■

Proof [Proof of Prop. 14] For part (a), let

$$\mathrm{d}\mu(h) = \prod_{m=1}^{H} \mathcal{N}(h_m; 0, C)\,\mathrm{d}h.$$

on $(\mathbb{R}^2)^H$. Note that μ is a non-degenerate normal distribution. Then we have

$$\begin{aligned} \mathrm{Cov}[F_1(x_a)^2, F_1(x_b)^2] &= \mathrm{E}[F_1(x_a)^2 F_2(x_b)^2] - \mathrm{E}[F_1(x_a)^2]\,\mathrm{E}[F_1(x_b)^2] \\ &= \int_{(\mathbb{R}^2)^H} \int_{(\mathbb{R}^2)^2} (y^a)^2 (y^b)^2 \mathcal{N}(y; 0, K)\,\mathrm{d}y\,\mathrm{d}\mu(h) - \mathrm{E}[F_1(x_a)^2]\,\mathrm{E}[F_1(x_b)^2] \\ &= \int_{(\mathbb{R}^2)^H} (2k_{ab}^2 + k_{aa}k_{bb})\,\mathrm{d}\mu(h) - \mathrm{E}[F_1(x_a)^2]\,\mathrm{E}[F_1(x_b)^2] \\ &= 2\int_{(\mathbb{R}^2)^H} k_{ab}^2\,\mathrm{d}\mu(h) + \int_{(\mathbb{R}^2)^H} k_{aa}k_{bb}\,\mathrm{d}\mu(h) - \mathrm{E}[F_1(x_a)^2]\,\mathrm{E}[F_1(x_b)^2]. \end{aligned}$$

Noting that $F_1(x_b)$ and $F_2(x_b)$ are identically distributed and recalling the proof of Prop. 10, we have

$$\begin{aligned}\operatorname{Cov}[F_1(x_a)^2, F_1(x_b)^2] &= 2\int_{(\mathbb{R}^2)^H} k_{ab}^2 \,\mathrm{d}\mu(h) + \int_{(\mathbb{R}^2)^H} k_{aa}k_{bb} \,\mathrm{d}\mu(h) - \mathrm{E}[F_1(x_a)^2]\,\mathrm{E}[F_2(x_b)^2] \\ &= 2\int_{(\mathbb{R}^2)^H} k_{ab}^2 \,\mathrm{d}\mu(h) + \operatorname{Cov}[F_1(x_a)^2, F_2(x_b)^2].\end{aligned}$$

Using again the fact that $F_1(x_b)$ and $F_2(x_b)$ are identically distributed, we have the correlation

$$\begin{aligned}q_{ab}^{(D)} &= \frac{\operatorname{Cov}[F_1(x_a)^2, F_1(x_b)^2]}{\sqrt{\mathrm{V}[F_1(x_a)^2]\,\mathrm{V}[F_1(x_b)^2]}} \\ &= 2\int_{(\mathbb{R}^2)^H} \frac{k_{ab}^2}{\sqrt{\mathrm{V}[F_1(x_a)^2]\,\mathrm{V}[F_1(x_b)^2]}} \,\mathrm{d}\mu(h) + \frac{\operatorname{Cov}[F_1(x_a)^2, F_2(x_b)^2]}{\sqrt{\mathrm{V}[F_1(x_a)^2]\,\mathrm{V}[F_1(x_b)^2]}} \\ &= 2\int_{(\mathbb{R}^2)^H} \frac{k_{ab}^2}{\sqrt{\mathrm{V}[F_1(x_a)^2]\,\mathrm{V}[F_1(x_b)^2]}} \,\mathrm{d}\mu(h) + q_{ab}^{\times(D)}.\end{aligned}$$

To make the dependence on the post-bottleneck depth D more explicit, we write

$$q_{ab}^{(D)} = 2\int_{(\mathbb{R}^2)^H} \frac{(k_{ab}^{(D)})^2}{\sqrt{\mathrm{V}[f_1^{(D)}(x_a)^2]\,\mathrm{V}[f_1^{(D)}(x_b)^2]}} \,\mathrm{d}\mu(h) + q_{ab}^{\times(D)}.$$

By Eq. (48), this becomes

$$q_{ab}^{(D)} = 2\int_{(\mathbb{R}^2)^H} \frac{(k_{ab}^{(D)})^2}{w_D^2 c_{aa} c_{bb} \sqrt{\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{aa}} + 1\right)^2\right)\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{bb}} + 1\right)^2\right)}} \,\mathrm{d}\mu(h) + q_{ab}^{\times(D)}.$$

The correlation at infinite depth is then

$$q_{ab}^{(\infty)} = \lim_{D\to\infty} 2\int_{(\mathbb{R}^2)^H} \frac{(k_{ab}^{(D)})^2}{w_D^2 c_{aa} c_{bb} \sqrt{\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{aa}} + 1\right)^2\right)\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{bb}} + 1\right)^2\right)}} \,\mathrm{d}\mu(h) + q_{ab}^{\times(\infty)}.$$

The limit can be moved inside the integral. To justify this, observe that the integrand (as a function of $h \in (\mathbb{R}^2)^H$) can be expressed as

$$\begin{aligned}
I_D(h) &:= \frac{(k_{ab}^{(D)})^2}{w_D^2 c_{aa} c_{bb} \sqrt{\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{aa}} + 1\right)^2\right)\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{bb}} + 1\right)^2\right)}} \\
&\leq \frac{k_{aa}^{(D)} k_{bb}^{(D)}}{w_D^2 c_{aa} c_{bb} \sqrt{\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{aa}} + 1\right)^2\right)\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{bb}} + 1\right)^2\right)}} \\
&= \frac{\left(b_D + \frac{w_D}{H}\sum_i \phi(h_i^a)^2\right)\left(b_D + \frac{w_D}{H}\sum_i \phi(h_i^b)^2\right)}{w_D^2 c_{aa} c_{bb} \sqrt{\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{aa}} + 1\right)^2\right)\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{bb}} + 1\right)^2\right)}} \\
&\leq \frac{\left(b_D + \frac{w_D}{H}\sum_i \phi(\max(h))^2\right)\left(b_D + \frac{w_D}{H}\sum_i \phi(\max(h))^2\right)}{w_D^2 c_{aa} c_{bb} \sqrt{\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{aa}} + 1\right)^2\right)\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{bb}} + 1\right)^2\right)}} \\
&= \frac{\left(b_D + w_D \phi(\max(h))^2\right)^2}{w_D^2 c_{aa} c_{bb} \sqrt{\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{aa}} + 1\right)^2\right)\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{bb}} + 1\right)^2\right)}} \\
&= \frac{\left(r_D + \phi(\max(h))^2\right)^2}{c_{aa} c_{bb} \sqrt{\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{aa}} + 1\right)^2\right)\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{bb}} + 1\right)^2\right)}} \\
&\leq \frac{2r_D^2}{c_{aa} c_{bb} \sqrt{\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{aa}} + 1\right)^2\right)\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{bb}} + 1\right)^2\right)}} \\
&+ \frac{2\phi(\max(h))^4}{c_{aa} c_{bb} \sqrt{\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{aa}} + 1\right)^2\right)\left(\frac{15}{H} + 2\left(\frac{r_D}{c_{bb}} + 1\right)^2\right)}}.
\end{aligned}$$

Recall that $r_D \to \frac{v_b}{v_w - 1}$ if $v_w > 1$ and $r_D \to \infty$ otherwise. In either case, it is easy to verify that the first term and the denominator of the second term converge to non-negative numbers independent of D. Therefore, there exist positive constants A and B such that

$$I_D(h) < A + B\phi(\max(h))^4 \text{ for sufficiently large } D.$$

Note the right-hand side is integrable with respect to the non-degenerate Gaussian measure μ since it is a piecewise polynomial in h (with finitely many pieces). We can therefore use the Dominated Convergence Theorem. Recall also that the NNGP kernel (post-bottleneck) degenerates to a constant-element kernel corresponding to a correlation matrix of 1's given

any fixed input h from the bottleneck layer. Using the Dominated Convergence Theorem twice, we therefore have

$$
\begin{aligned}
q_{ab}^{(\infty)} &= 2\int_{(\mathbb{R}^2)^H} \lim_{D\to\infty} \frac{(k_{ab}^{(D)})^2}{w_D^2 c_{aa} c_{bb} \prod_{c=c_{aa},c_{bb}} \sqrt{\frac{15}{H} + 2\left(\frac{r_D}{c}+1\right)^2}} \,\mathrm{d}\mu(h) + q_{ab}^{\times(\infty)} \\
&= 2\int_{(\mathbb{R}^2)^H} \lim_{D\to\infty} \frac{(k_{ab}^{(D)})^2}{k_{aa}^{(D)} k_{bb}^{(D)}} \frac{k_{aa}^{(D)} k_{bb}^{(D)}}{w_D^2 c_{aa} c_{bb} \prod_{c=c_{aa},c_{bb}} \sqrt{\frac{15}{H} + 2\left(\frac{r_D}{c}+1\right)^2}} \,\mathrm{d}\mu(h) + q_{ab}^{\times(\infty)} \\
&= 2\int_{(\mathbb{R}^2)^H} \lim_{D\to\infty} \frac{k_{aa}^{(D)} k_{bb}^{(D)}}{w_D^2 c_{aa} c_{bb} \prod_{c=c_{aa},c_{bb}} \sqrt{\frac{15}{H} + 2\left(\frac{r_D}{c}+1\right)^2}} \,\mathrm{d}\mu(h) + q_{ab}^{\times(\infty)} \\
&= \lim_{D\to\infty} \frac{2}{w_D^2 c_{aa} c_{bb} \prod_{c=c_{aa},c_{bb}} \sqrt{\frac{15}{H} + 2\left(\frac{r_D}{c}+1\right)^2}} \int_{(\mathbb{R}^2)^H} k_{aa}^{(D)} k_{bb}^{(D)} \,\mathrm{d}\mu(h) + q_{ab}^{\times(\infty)} \\
&= \lim_{D\to\infty} \frac{2}{w_D^2 c_{aa} c_{bb} \prod_{c=c_{aa},c_{bb}} \sqrt{\frac{15}{H} + 2\left(\frac{r_D}{c}+1\right)^2}} \\
&\cdot \left((b_D + w_D c_{aa})(b_D + w_D c_{bb}) + \mathrm{Cov}[f_1^{(D)}(x_a)^2, f_2^{(D)}(x_b)^2]\right) + q_{ab}^{\times(\infty)} \\
&= \lim_{D\to\infty} \frac{2(b_D + w_D c_{aa})(b_D + w_D c_{bb})}{w_D^2 c_{aa} c_{bb} \prod_{c=c_{aa},c_{bb}} \sqrt{\frac{15}{H} + 2\left(\frac{r_D}{c}+1\right)^2}} + 3q_{ab}^{\times(\infty)} \\
&= \lim_{D\to\infty} \frac{\left(\frac{r_D}{c_{aa}}+1\right)\left(\frac{r_D}{c_{bb}}\right)}{\prod_{c=c_{aa},c_{bb}} \sqrt{\frac{15}{2H} + \left(\frac{r_D}{c}+1\right)^2}} + 3q_{ab}^{\times(\infty)}.
\end{aligned}
$$

Evaluating the limit by recalling the limit of r_D, we obtain Eq. (19) as desired.

For part (b), given $(Q^{(\infty)}, \mathrm{diag}(G))$, we will show that we recover G. We can obtain $\mathrm{diag}(C)$ from $\mathrm{diag}(G)$ by applying the NNGP kernel propagation defined in Eq. (9). Given $\mathrm{diag}(C)$, we can solve for $Q^{\times(\infty)}$ using Eq. (19) for the case $v_w > 1$ in Prop. 14. We can then obtain G by Prop. 12 (b). ■

D.1 Other nonlinearities

The following is the proof for the proposition linking the sinusoidal nonlinearity in Eq. (20) to the RBF kernel.

Proof [Proof of Prop. 15] First we prove part (a). The case $\mu = 0$ is trivial. So, consider $\mu \geq 1$. Define the function $F_\phi : \mathbb{R}^3 \mapsto \mathbb{R}$ by

$$F_\phi(a, b, c) = \int_{\mathbb{R}^2} \phi(z_1)\phi(z_2)\mathcal{N}\left(z; 0, \begin{bmatrix} a & b \\ b & c \end{bmatrix}\right) \mathrm{d}z, \tag{51}$$

where ϕ is the sinusoidal nonlinearity in Eq. (20) and

$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

The NNGP kernel recursion can then be written as

$$K^{(\mu+1)}(x, x') = v_b + v_w F_\phi[K^{(\mu)}(x, x), K^{(\mu)}(x, x'), K^{(\mu)}(x', x')].$$

All we need to show is that

$$F_\phi(a, b, c) = e^{-\frac{1}{2}(a+c-2b)}. \tag{52}$$

Let X be a 2×2 matrix with columns $x_1, x_2 \in \mathbb{R}^2$ such that

$$X^\top X = \begin{bmatrix} a & b \\ b & c \end{bmatrix}. \tag{53}$$

Such a matrix X exists since the matrix on the right side is symmetric positive semidefinite. Performing the change of variables $z = X^\top w$, Eq. (51) becomes

$$\begin{aligned} F_{phi}(a, b, c) &= \int_{\mathbb{R}^2} \phi(x_1^\top w)\phi(x_2^\top w)\mathcal{N}(w; 0, I)\, \mathrm{d}w \\ &= \frac{1}{2\pi} \int_{\mathbb{R}^2} \phi(w \cdot x_1)\phi(w \cdot x_2) e^{-\frac{1}{2}\|w\|^2}\, \mathrm{d}w. \end{aligned}$$

The product of activations in the integrand can be rewritten as

$$\begin{aligned} \phi(w \cdot x_1)\phi(w \cdot x_2) &= [\cos(w \cdot x_1) + \sin(w \cdot x_1)][\cos(w \cdot x_2) + \sin(w \cdot x_2)] \\ &= \cos(w \cdot x_1)\cos(w \cdot x_2) + \sin(w \cdot x_1)\sin(w \cdot x_2) \\ &+ \cos(w \cdot x_1)\sin(w \cdot x_2) + \sin(w \cdot x_1)\cos(w \cdot x_2) \\ &= \cos[w \cdot (x_1 - x_2)] + \sin[w \cdot (x_1 + x_2)]. \end{aligned}$$

We therefore have

$$\begin{aligned} F_\phi(a, b, c) &= \frac{1}{2\pi} \int_{\mathbb{R}^2} \cos[w \cdot (x_1 - x_2)] e^{-\frac{1}{2}\|w\|^2}\, \mathrm{d}w \\ &+ \frac{1}{2\pi} \int_{\mathbb{R}^2} \sin[w \cdot (x_1 + x_2)] e^{-\frac{1}{2}\|w\|^2}\, \mathrm{d}w. \end{aligned}$$

The integrand of the second integral on the right side is odd in w for all x_1 and x_2, and thus this integral is zero. We can therefore replace $x_1 + x_2$ with $x_1 - x_2$ in the second integral

and multiply the integral by the imaginary unit i without changing its value:

$$\begin{aligned}
F_\phi(a,b,c) &= \frac{1}{2\pi}\int_{\mathbb{R}^2} \cos[w\cdot(x_1-x_2)]e^{-\frac{1}{2}\|w\|^2}\,\mathrm{d}w \\
&+ i\frac{1}{2\pi}\int_{\mathbb{R}^2} \sin[w\cdot(x_1-x_2)]e^{-\frac{1}{2}\|w\|^2}\,\mathrm{d}w \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2} \left(\cos[w\cdot(x_1-x_2)] + i\sin[w\cdot(x_1-x_2)]\right)e^{-\frac{1}{2}\|w\|^2}\,\mathrm{d}w \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2} e^{iw\cdot(x_1-x_2)}e^{-\frac{1}{2}\|w\|^2}\,\mathrm{d}w \\
&= e^{-\frac{1}{2}\|x_1-x_2\|^2},
\end{aligned}$$

where the last line holds because the Gaussian is an eigenfunction of the Fourier transform. Using Eq. (53), this becomes

$$\begin{aligned}
F_\phi(a,b,c) &= e^{-\frac{1}{2}(\|x_1\|^2+\|x_2\|^2-2x_1\cdot x_2)} \\
&= e^{-\frac{1}{2}(a+c-2b)},
\end{aligned}$$

giving us Eq. (52) as desired.

For part (b), consider two distinct inputs $x, x' \in \mathbb{R}^M$ and define

$$v^{(\mu)} = K^{(\mu)}(x,x') \text{ and } c^{(\mu)} = \frac{K^{(\mu)}(x,x')}{\sqrt{K^{(\mu)}(x,x)K^{(\mu)}(x',x')}}.$$

All we need to show is that $(v^{(\mu)}, c^{(\mu)})$ has a globally attractive fixed point (v_*, c_*) of the form given in the statement of the proposition. The dynamics of $v^{(\mu)}$ is given by $v^{(\mu+1)} = f_v(v^{(\mu)})$ where $f_v : (0,\infty) \mapsto (0,\infty)$ is given by

$$\begin{aligned}
f_v(v) &= v_b + v_w F_\phi(v,v,v) \\
&= v_b + v_w.
\end{aligned}$$

It is thus trivial that the global fixed point of f_v is $v_*(v_b, v_w) = v_b + v_w$. The dynamics of $c^{(\mu)}$ is given by $c^{(\mu+1)} = f_c(c^{(\mu)})$ where $f_c : [-1,1] \mapsto [-1,1]$ is given by

$$f_c(c) = v_b + v_w F_\phi(v_*, v_* c, v_*),$$

where we approximate $v^{(\mu)}$ with v_*; this approximation becomes exact in the deep limit. Substituting in $v_* = v_b + v_w$, we get

$$f_c(c) = \frac{v_b}{v_b+v_w} + \frac{v_w}{v_b+v_w}e^{(v_b+v_w)(c-1)}.$$

We make use of the fact that if a function $f : \mathbb{R} \mapsto \mathbb{R}$ has strictly positive first and second derivatives, then f has either no fixed points, one fixed point that is neither stable nor unstable, or two fixed points $c_1 < c_2$ where c_1 is stable with basin of attraction $(-\infty, c_2)$ and c_2 is unstable. Observe that $f_c(1) = 1$ and $f_c'(1) = v_w$. If $v_w < 1$, then $c = 1$ is a stable fixed point and is thus globally attractive on $[-1,1]$, thus establishing $c_*(v_b, v_w) = 1$ for $v_w < 1$.

On the other hand, if $v_w > 1$, then $c = 1$ is an unstable fixed point. Since $f_c(0) > 0$ and $f_c(1) = 1$, then by the Intermediate Value Theorem, f_c has a stable fixed point $c' \in (0, 1)$ that is globally attractive on $[-1, 1)$. Given the initial condition $c^{(0)} = \frac{x^\top x'}{\|x\|\|x'\|} < 1$ (since $x \neq x'$) and that $c = 1$ is an unstable fixed point, then we must have $c_*(v_b, v_w) < 1$ for $v_w > 1$ and thus in particular $c_*(v_b, v_w) = c'$, concluding the proof. ■

Proof [Proof of Prop. 17] Let H be the bottleneck width, and let $h_1, h_2 \in \mathbb{R}^H$ be the bottleneck preactivations—i.e., the outputs of the pre-bottleneck NNGP component—given network inputs $x_1, x_2 \in \mathbb{R}^M$. Let μ be the (Gaussian) probability measure associated to (h_1, h_2). Let $z_1, z_2 \in \mathbb{R}^H$ be the corresponding bottleneck activations that are fed into the post-bottleneck NNGP component, where $z_a = \frac{1}{\sqrt{H}}\phi(h_a)$ for $a = 1, 2$. Finally, let $K^{(D)} : \mathbb{R}^H \times \mathbb{R}^H \mapsto \mathbb{R}$ be the kernel of the post-bottleneck NNGP component assuming post-bottleneck depth D. Then the PDF of the bottleneck NNGP outputs is given by

$$p^{(D)}(y) = \int_{(\mathbb{R}^H)^2} \mathcal{N}(y; 0, K^{(D)}(Z, Z) + v_n I)\, \mathrm{d}\mu(h_1, h_2),$$

where $K^{(D)}(Z, Z)$ is a 2×2 matrix with entries $K^{(D)}(z_a, z_b)$ for $a, b = 1, 2$. Using the fact that for any 2×2 positive semidefinite matrix A with eigenvalues $\lambda_1, \lambda_2 \geq 0$,

$$\det(A + v_n I) = (\lambda_1 + v_n)(\lambda_2 + v_n) \geq v_n^2,$$

we have the bound

$$\begin{aligned}
\mathcal{N}(y; 0, K^{(D)}(Z, Z) + v_n I) &\leq \frac{1}{2\pi} \det[K^{(D)}(Z, Z) + v_n I]^{-\frac{1}{2}} \\
&\leq \frac{1}{2\pi}(v_n^2)^{-\frac{1}{2}} \\
&= \frac{v_n}{2\pi}.
\end{aligned}$$

Since the bound is clearly an integrable function with respect to μ and since the bound holds for all D, then we may apply the Bounded Convergence Theorem. By the continuity of the matrix determinant and inversion operations, the PDF converges to

$$\begin{aligned}
p^{(\infty)}(y) &= \lim_{D \to \infty} p^{(D)}(y) \\
&= \int_{(\mathbb{R}^H)^2} \mathcal{N}(y; 0, \lim_{D \to \infty} K^{(D)}(Z, Z) + v_n I)\, \mathrm{d}\mu(h_1, h_2) \\
&= \int_{(\mathbb{R}^H)^2} \mathcal{N}(y; 0, K^{(\infty)}(Z, Z) + v_n I)\, \mathrm{d}\mu(h_1, h_2). \qquad (54)
\end{aligned}$$

Define the set

$$\begin{aligned}
S &= \left\{(h_1, h_2) \in (\mathbb{R}^H)^2 : z_1 = z_2\right\} \\
&= \left\{(h_1, h_2) \in (\mathbb{R}^H)^2 : \frac{1}{\sqrt{H}}\phi(h_1) = \frac{1}{\sqrt{H}}\phi(h_2)\right\} \\
&= \{(h_1, h_2) \in (\mathbb{R}^H)^2 : \phi(h_1) = \phi(h_2)\} \\
&= \{(h_1, h_2) \in (\mathbb{R}^H)^2 : \exists n \in \mathbb{Z}^H \mid h_1 = h_2 + 2n\pi\} \\
&= \bigcup_{n \in \mathbb{Z}^H} \{(h, h + 2n\pi) : h \in \mathbb{R}^H\}.
\end{aligned}$$

We see that S is a countable disjoint union of H-dimensional planes embedded in a $2h$-dimensional space. Since the network inputs x_1 and x_2 are distinct, then μ is a non-degenerate Gaussian distribution on $(\mathbb{R}^H)^2$ so that $\mu(S) = 0$. We can therefore remove S from the region of integration in Eq. (54), so that z_1 and z_2 are distinct inputs into $K^{(\infty)}$ under the integral. We thus evaluate the covariance of the integrand in Eq. (54) using Prop. 15 (b) and Eq. (23) and get

$$\begin{aligned}
p^{(\infty)}(y) &= \int_{(\mathbb{R}^H)^2 \setminus S} \mathcal{N}(y; 0, K^{(\infty)}(Z, Z) + v_n I)\, \mathrm{d}\mu(h_1, h_2) \\
&= \int_{(\mathbb{R}^H)^2 \setminus S} \mathcal{N}\left(y; 0, v_*(v_b, v_w) \begin{bmatrix} 1 & c_*(v_b, v_w) \\ c_*(v_b, v_w) & 1 \end{bmatrix} + v_n I\right) \mathrm{d}\mu(h_1, h_2) \\
&= \mathcal{N}\left(y; 0, v_*(v_b, v_w) \begin{bmatrix} 1 & c_*(v_b, v_w) \\ c_*(v_b, v_w) & 1 \end{bmatrix} + v_n I\right),
\end{aligned}$$

which according to Prop. 15 (b) is precisely the deep limit with no bottleneck given two distinct inputs, thus establishing Eq. (26). ■

References

Ahmed M Alaa and Mihaela van der Schaar. Deep multi-task Gaussian processes for survival analysis with competing risks. In *Advances in Neural Information Processing Systems*, pages 2326–2334. Curran Associates Inc., 2017.

Mauricio Alvarez and Neil D Lawrence. Sparse convolved Gaussian processes for multi-output regression. In *Advances in Neural Information Processing Systems*, pages 57–64, 2009.

Edgar Anderson. The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

Grigory Antipov, Sid-Ahmed Berrani, Natacha Ruchaud, and Jean-Luc Dugelay. Learned vs. hand-crafted features for pedestrian gender recognition. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1263–1266. ACM, 2015.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019.

Rodney J Baxter. *Exactly solved models in statistical mechanics*. Elsevier, 2016.

Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, third edition, 1995. ISBN 978-0-471-00710-4.

Patrick Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. ISBN 0-471-19745-9.

J. R. Blum, H. Chernoff, M. Rosenblatt, and H. Teicher. Central limit theorems for interchangeable processes. *Canadian Journal of Mathematics*, 10, 1958.

Edwin V Bonilla, Kian M Chai, and Christopher Williams. Multi-task Gaussian process prediction. In *Advances in neural information processing systems*, pages 153–160, 2008.

Thang Bui, Daniel Hernández-Lobato, Jose Hernandez-Lobato, Yingzhen Li, and Richard Turner. Deep Gaussian processes for regression using approximate expectation propagation. In *International conference on machine learning*, pages 1472–1481, 2016.

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

Kurt Cutajar, Edwin V Bonilla, Pietro Michiardi, and Maurizio Filippone. Random feature expansions for deep Gaussian processes. In *International Conference on Machine Learning*, pages 884–893, 2017.

Zhenwen Dai, Andreas Damianou, Javier González, and Neil Lawrence. Variational auto-encoded deep Gaussian processes. *arXiv preprint arXiv:1511.06455*, 2015.

Andreas Damianou and Neil Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pages 2253–2261, 2016.

Alexander G. de Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *International Conference on Learning Representations*, 2018. accepted as poster.

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

Vincent Dutordoir, Hugh Salimbeni, James Hensman, and Marc Deisenroth. Gaussian process conditional density estimation. In *Advances in neural information processing systems*, pages 2385–2395, 2018.

David Duvenaud, Oren Rippel, Ryan Adams, and Zoubin Ghahramani. Avoiding pathologies in very deep networks. In *Artificial Intelligence and Statistics*, pages 202–210, 2014.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *ArXiv preprint arXiv:1506.02142*, 2015.

Adria Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow Gaussian processes. In *International Conference on Learning Representations*, 2019.

David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

Tamir Hazan and Tommi Jaakkola. Steps toward deep kernel methods from infinite neural networks. *arXiv preprint arXiv:1508.05133*, 2015.

James Hensman and Neil D Lawrence. Nested variational compression in deep Gaussian processes. *arXiv preprint arXiv:1412.1370*, 2014.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

Diederik Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*, pages 2575–2583, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Neil D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems*, pages 329–336, 2004.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

Ping Li and Phan-Minh Nguyen. On random deep weight-tied autoencoders: Exact asymptotic analysis, phase transitions, and implications to training. In *International Conference on Learning Representations*, 2019.

Hong Liang, Xiao Sun, Yunlei Sun, and Yuan Gao. Text feature extraction based on deep learning: a review. *EURASIP journal on wireless communications and networking*, 2017 (1):211, 2017.

Zhouhan Lin, Roland Memisevic, and Kishore Konda. How far can we go without convolution: Improving fully-connected networks. *arXiv preprint arXiv:1511.02580*, 2015.

Chi-Ken Lu, Scott Cheng-Hsin Yang, Xiaoran Hao, and Patrick Shafto. Interpretable deep Gaussian processes with moments. *arXiv preprint arXiv:1905.10963*, 2019.

Paul André Meyer. *Probability and potentials*, volume 1318. Blaisdell Pub. Co., 1966.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Bayesian deep convolutional networks with many channels are Gaussian processes. In *International Conference on Learning Representations*, 2019.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

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

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006.

F Rosenblat. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Pentti Saikkonen. Continuous weak convergence and stochastic equicontinuity results for integrated processes with an application to the estimation of a regression model. *Econometric Theory*, 9(2):155–188, 1993.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *arXiv preprint arXiv:1611.01232*, 2016.

TJ Sweeting. Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, pages 1375–1381, 1980.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

Michalis Titsias and Neil D Lawrence. Bayesian Gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851, 2010.

Yali Wang, Marcus Brubaker, Brahim Chaib-Draa, and Raquel Urtasun. Sequential inference for deep Gaussian process. In *Artificial Intelligence and Statistics*, pages 694–703, 2016.

Christopher KI Williams. Computing with infinite networks. In *Advances in neural information processing systems*, pages 295–301, 1997.

Ge Yang and Samuel Schoenholz. Mean field residual networks: on the edge of chaos. In *Advances in Neural Information Processing Systems*, pages 7103–7114, 2017.

Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.