Journal of Machine Learning Research 25 (2024) 1-41 Submitted 9/23; Revised 4/24; Published 8/24

Commutative Scaling of Width and Depth in Deep Neural Networks

Soufiane Hayou HAYOU@BERKELEY.EDU
Simons Institute
UC Berkeley

Editor: Daniel Roy

Abstract

In this paper, we study the commutativity of infinite width and depth limits in deep neural networks. Our aim is to understand the behavior of neural functions (functions that depend on a neural network model) as width and depth go to infinity (in some sense), and eventually identify settings under which commutativity holds, i.e. the neural function tends to the same limit no matter how width and depth limits are taken. In this paper, we formally introduce and define the commutativity framework, and discuss its implications on neural network design and scaling. We study commutativity for the neural covariance kernel which reflects how network layers separate data. Our findings extend previous results established in Hayou and Yang (2023) by showing that taking the width and depth to infinity in a deep neural network with skip connections, when branches are suitably scaled to avoid exploding behavior, result in the same covariance structure no matter how that limit is taken. This has a number of theoretical and practical implications that we discuss in the paper. The proof techniques in this paper are new and rely on tools that are more accessible to readers who are not familiar with stochastic calculus (used in the proofs of Hayou and Yang (2023)).

1. Introduction

The success of large language and vision models have recently amplified an existing trend of research on large size neural network. There are generally two ways to increase the size of a neural network model: increasing the width, for instance the number of neurons in hidden layers in a fully-connected network, the number of channels in a convolutional network, or the number of attention heads in a transformer architecture; and increasing the depth of the network, i.e. the number of layers. A suitable appraoch to understand the behavior of large neural networks is by analyzing some pre-defined quantity as the width and/or depth tend to infinity. While the width limit by itself is now relatively well understood in different contexts (Neal, 1995; Schoenholz et al., 2017; Lee et al., 2018; Yang, 2021b; Hayou et al., 2019), the depth limit and the interaction between the two have not been studied as much. In particular, given some pre-defined quantity of interest that depends on the network model, a basic question is: *do these two limits commute?* (in the sense that the behavior of the quantity of interest as width and depth go to infinity does not change

©2024 Soufiane Hayou.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/23-1163.html.

depending on the order of which these limits are taken). One statistical quantity of interest is the *neural covariance* kernel which reflects how layers in a neural network model separate input data. In this context, recent literature suggests that, at initialization, in certain kinds of multi-layer perceptrons (MLPs) or residual neural networks (ResNets) with scaled main branch, the depth and width limits generally do *not* commute (Li et al., 2023; Noci et al., 2023); this would imply that in practice, such networks would behave quite differently depending on whether width is much larger than depth or the other way around. However, in the case of ResNets with suitably scaled residual blocks, Hayou and Yang (2023) showed that, to the contrary, at initialization, for a ResNet with blocks scaled the natural way so as to avoid blowing up the output, the width and depth limits *do commute*. An interesting practical implication of this result is that it justifies prior calculations that take the width limit first, then depth, to understand the behavior of deep residual networks, such as prior works in the signal propagation literature (Schoenholz et al., 2017; Yang and Schoenholz, 2017; Hayou et al., 2021).

In this work, we introduce and formalize the framework of commutativity of the width and depth limits and generalize (and improve) existing results on the covariance from Hayou and Yang (2023) for arbitrary sequences of scaling factors; these sequences are used to scale the residual blocks so as to avoid exploding behavior as depth grows. We discuss the theoretical and practical implications of commutativity by addressing the natural question; *why should care about commutativity at all?* (see Section 3).

In addition to the significance of the results and the new framework, the mathematical novelty of this paper lies in the proof techniques: in contrast to Hayou and Yang (2023) where the depth limit is taken first (fixing the width), followed by the width limit, we first take the width to infinity this time, which is a more conventional approach in the theory of signal propagation in deep networks. As such, the proof techniques in this paper can be seen as 'orthogonal' to the machinery developed in Hayou and Yang (2023), and are more accessible to readers who are not familiar with stochastic calculus. Our results provide new insights into the behavior of deep neural networks with general depth scaling factors and we discuss implications for the design and analysis of these networks.

All the proofs are deferred to the appendix and referenced after each result. Empirical evaluations are provided to illustrate the theoretical results.

2. Related Work

The theoretical analysis of randomly initialized neural networks with an infinite number of parameters has yielded a stream of interesting results, both theoretical and practical. A majority of this research has concentrated on examining the scenario in which the width of the network is taken to infinity while the depth is considered fixed. However, in recent years, there has been a growing interest in exploring the large depth limit of these networks. In this overview, we present a summary of existing results on this topic, though it is not exhaustive. A more comprehensive literature review is provided in Appendix A.

Infinite-Width Limit: The study of the infinite-width limit of neural network architectures has been a topic of significant research interest, yielding various theoretical and algorithmic innovations. These include initialization methods, such as the Edge of Chaos (Poole et al., 2016; Schoenholz et al., 2017; Yang and Schoenholz, 2017; Hayou et al., 2019), and the

selection of activation functions (Hayou et al., 2019; Martens et al., 2021; Zhang et al., 2022; Wolinski and Arbel, 2023), which have been shown to have practical benefits. In the realm of Bayesian analysis, the infinite-width limit presents an intriguing framework for Bayesian deep learning, as it is characterized by a Gaussian process prior. Several studies (e.g. Neal (1995); Lee et al. (2018); Yang (2021b); Matthews et al. (2018); Hron et al. (2020)) have investigated the weak limit of neural networks as the width increases towards infinity, and have demonstrated that the network's output converges to a distribution modeled by a Gaussian process. Bayesian inference utilizing this "neural" Gaussian process has been explored in Lee et al. (2018); Hayou et al. (2021). [1]

Infinite-Depth Limit: The infinite-depth limit of neural networks with random initialization is a less explored area compared to the study of the infinite-width limit. Existing results can be categorized depending on how the two limits are taken. For instance, in the case of sequential limits, the width of the neural network is taken to infinity first, followed by the depth. This limit has been extensively utilized to explore various aspects of neural networks, such as examining the neural covariance, deriving the Edge of Chaos initialization scheme (cited in (Schoenholz et al., 2017; Poole et al., 2016; Yang and Schoenholz, 2017)), evaluating the impact of the activation function (Hayou et al., 2019; Martens et al., 2021), and studying the behavior of the Neural Tangent Kernel (NTK) (Hayou et al., 2022; Xiao et al., 2020). Another interesting limit is the proportional limit where the ratio of depth to width is fixed, and both are jointly taken to infinity. In Li et al. (2021), the authors showed that for a particular type of residual neural networks (ResNets), the network output exhibits a (scaled) log-normal behavior in this limit, which differs from the sequential limit in which the width is first taken to infinity followed by depth, in which case the distribution of the network output is asymptotically normal (Schoenholz et al., 2017; Hayou et al., 2019). Additionally, in Li et al. (2023), the authors examined the neural covariance of a multi-layer perceptron (MLP) in the joint limit and proved that it weakly converges to the solution of a Stochastic Differential Equation (SDE). Other works have investigated this limit and found similar results (Noci et al., 2021; Zavatone-Veth and Pehlevan, 2021; Hanin and Nica, 2019; Hanin, 2022; Noci et al., 2023). A third interesting approach is the general limit $\min\{n, L\} \to \infty$, where width and depth can to infinity in any order. To the best of our knowledge, this limit was only studied in Hayou and Yang (2023) where the authors showed convergence of the neural covariance in this limit for suitably scaled ResNet.

3. Setup and Definitions: Commutativity and Neural Functions

When analyzing the asymptotic behavior of randomly initialized neural networks, various notions of probabilistic convergence are employed, depending on the context. In this work, we particularly focus on strong convergence, defined to be the L_2 convergence as described in the following definition.

Definition 1 (Strong convergence)**.** *Let $d \geq 1$. We say that a sequence of $\mathbb{R}^d$-valued random variables $(X_k)_{k\geq 1}$ converges in L_2 (or strongly) to a continuous random variable Z if $\lim_{k\to\infty} \|X_k - Z\|_{L_2} = 0$, where the L_2 is defined by $\|X\|_{L_2} = \left(\mathbb{E}[\|X\|^2]\right)^{1/2}$.*

1. It is worth mentioning that kernel methods such as NNGP and NTK significantly underperform properly tuned finite-width network trained using SGD, see Yang et al. (2022).

With this notion of strong convergence, we are now ready to introduce the commutativity framework for general neural network models.

Notation. Throughout the paper, the width and depth of a neural network model are denoted by n and L, respectively, and the input dimension is denoted by d. We write $[N] := \{1, 2, \dots, N\}$ for any $N \geq 1$.

Let us now consider a general neural network model of width $n \geq 1$ and depth $L \geq 1$, given by

$$\begin{cases} Y_0(a) = W_{in}a, \quad a \in \mathbb{R}^d \\ Y_l(a) = \mathcal{F}_l(W_l, Y_{l-1}(a)),\ l \in [L], Y_l(a) \in \mathbb{R}^n, \end{cases} \tag{1}$$

where $\mathcal{F}_l$ is a mapping that defines the nature of the l^{th} layer and $W_{in} \in \mathbb{R}^{n\times d}, W_l \in \mathbb{R}^{n\times n}$ are model weights. For the sake of simplification, we omit the dependence of Y_l on n and L in the notation. We refer to the vectors $\{Y_l, l = 0, \dots, L\}$ as *pre-activations*. Let $\theta_{n,L} = (W_{in}, W_1, \dots, W_l)$ be the model weights and assume that $\theta^0_{n,L} \sim \mu^0_{n,L}$, where $\theta^0_{n,L}$ are the weights at initialization and μ^0 is a distribution that (naturally) depends on network width n and depth L. The distribution $\mu^0_{n,L}$. Let us now define the notion of neural functions.

Definition 2 (Neural Function)**.** *Given a general neural network model (Eq.* (1)*) of width n and depth L, a set of network inputs $\mathbf{a} = (a_1, a_2, \dots, a_k) \in (\mathbb{R}^d)^k$, a neural function T is any function of the form $T(n, L, \mathbf{a}) = \mathcal{G}(\theta^0_{n,L}, \mathbf{a})$, where $\mathcal{G}$ is a general mapping with output in $\mathbb{R}$.*[2]

Note that (almost) any quantity in the training process of neural networks can be represented as a neural function. This remark was first observed in the series of Tensor Programs (Yang and Hu, 2022) where the result of any neural computation can be seen as a random quantity where the randomness is inherited from the initialization weights. The training dataset is considered deterministic in this case and consists of a sequence of inputs $(a_1, a_2, \dots, a_k)$. Other neural functions that cannot be expressed with Tensor Programs include the generalization error for instance. In this paper, we think of neural functions as proxy functions that track some behavior of the network as we scale width and depth with the goal of providing insights on scaling strategies (see below for a specific choice of the neural function).

With this definition of neural functions, we now formalize the notion of commutativity of the width and depth limits.

Definition 3 (Commutativity)**.** *Given a neural function T,*[3] *we say that T satisfies universality for the width and depth limits if for any set of inputs $\mathbf{a} = (a_1, a_2, \dots, a_k)$, $T(n, L, \mathbf{a})$ converges in L_2 in the limit $\min\{n, L\} \to \infty$.*

2. This definition of neural functions can be extended to general mappings $\mathcal{G}$ with outputs in $\mathbb{R}^p$ for some $p \geq 1$. This is not required in this paper since we will be focusing on neural covariance kernel which has output in $\mathbb{R}$.
3. Note that by definition, a neural function is associated with a network model. When we consider a neural function T, the underlying model is assumed to be fixed.

We can define a weak notion of commutativity where only sequential limits are considered, i.e. n or L limits are taken in a sequential order.

Definition 4 (Weak Commutativity)**.** *Given neural function T, we say that T satisfies commutativity for the width and depth limits if for any set of inputs $\mathbf{a} = (a_1, a_2, \ldots, a_k)$, both $\lim_{L\to\infty}\lim_{n\to\infty} T(n, L, \mathbf{a})$ and $\lim_{n\to\infty}\lim_{L\to\infty} T(n, L, \mathbf{a})$ exist in L_2 and are equal.*

Weak commutativity is trivially implied by commutativity. Intuitively, weak commutativity only deals with the 'extreme' scenarios $L \gg n \gg 1$ and $n \gg L \gg 1$ and does not consider the cases where for instance $L \approx n \gg 1$.

Implications of Commutativity. Naturally, one might ask why we should care about commutativity at first. Commutativity of width and depth limits in neural networks holds significant importance for several compelling reasons:

1. *Unification of Width and Depth Scaling:* when we aim to scale a neural network for improved performance, we often encounter scenarios where we must decide whether to increase the network's width or depth. Each of these choices generally lead to different design considerations, including variations in initialization schemes, activation functions, and learning rates. However, commutativity of the width and depth limits for some neural function T ensures that regardless of how we scale the network—whether by increasing width before depth, growing both width and depth proportionally, or taking width to infinity before depth—the resulting limiting behavior remains consistent. This means that once an effective scaling strategy is identified for a specific scenario with large width and depth, it remains a viable choice as long as both width and depth are large, simplifying the scaling process.

2. *Robust Scaling:* as a result of commutativity, scaling the width and depth becomes robust to extreme changes in neural functions. This allows some flexibility in the scaling procedure; in practice, one might want to increase width significantly while fixing depth, or the opposite, while preserving desirable properties captured by the neural function.

3. *Transfer of Insights:* commutativity facilitates the transfer of insights from simplified theoretical settings to practical applications. When dealing with neural networks of large width and depth, it can be challenging to analyze their behavior directly. However, commutativity allows us to explore different limits, such as taking width to infinity first and then depth or vice versa, to gain a better understanding of the network's behavior.

4. *Commutativity is Achievable in Practice:* we show that by introducing a simple scaling factor in front of the residual block in ResNets, commutativity holds for the neural covariance function at initialization (defined below). This neural function is used as a measure of how network layer separate input data, and led to many interesting practical methods (initialization schemes, neural network Gaussian process, choice of the activation function etc.) (Lee et al., 2018; Schoenholz et al., 2017; Hayou et al., 2019). An in-depth discussion on this topic is provided below.

Neural Covariance. In this paper, we focus on neural functions given by the covariance/correlation functions *at initialization*. Given two inputs $a, b \in \mathbb{R}^d \backslash \{0\}$,[4] the neural covariance and correlation kernels at layer l are given by

$$\begin{cases} q_{l,n}(a,b) = \frac{\langle Y_l(a), Y_l(b)\rangle}{n} \\ c_{l,n}(a,b) = \frac{\langle Y_l(a), Y_l(b)\rangle}{\|Y_l(a)\|\|Y_l(b)\|}, \end{cases}$$

where the correlation is only defined when $\|Y_l(a)\|, \|Y_l(b)\| \neq 0$.

Note that in general, if commutativity holds for the covariance kernel, then it holds for the neural correlation kernel, and vice-versa. This is true as long as pre-activations norms $\|Y_l(a)\|$ are non-zero with high probability, which is generally satisfied, see Lemma 5 for a rigorous proof of this result. Hereafter, we will interchangeably discuss commutativity for neural covariance and correlation, while stating the theoretical results only for neural covariance. The results on the convergence of neural covariance are stated for two inputs a, b, but they can be readily generalized to the case of multiple inputs $a_1, a_2, \ldots, a_k \in \mathbb{R}^d$, where we can define the neural covariance matrix at layer l by

$$\mathbf{q}_{l,n}(a_1, a_2, \ldots, a_k) = \begin{pmatrix} q_{l,n}(a_1,a_1) & \ldots & q_{l,n}(a_1,a_k) \\ \vdots & \ddots & \vdots \\ q_{l,n}(a_k,a_1) & \ldots & q_{l,n}(a_k,a_k) \end{pmatrix}.$$

Why Neural Covariance/Correlation? In the literature on signal propagation, there is a significant interest in understanding the covariance/correlation between the pre-activation vectors $Y_{\lfloor tL \rfloor}(a)$ and $Y_{\lfloor tL \rfloor}(b)$ for two different inputs $a, b \in \mathbb{R}^d$. A natural question in this context is: *Why should we care about this covariance function?*

It is well-established that even for properly initialized multi-layer perceptrons (MLPs), the network outputs $Y_L(a)$ and $Y_L(b)$ become perfectly correlated (correlation=1) in the limit of "$n \to \infty$, *then* $L \to \infty$" (Schoenholz et al., 2017; Poole et al., 2016; Hayou et al., 2019; Yang and Salman, 2020). This can lead to unstable behavior of the gradients and make the model untrainable as the depth increases and also results in the inputs being non-separable by the network[5]. To address this issue, several techniques involving targeted modifications of the activation function have been proposed (Martens et al., 2021; Zhang et al., 2022). In the case of ResNets, the correlation still converges to 1, but at a polynomial rate (Yang and Schoenholz, 2017). A solution to this problem has been proposed by introducing well-chosen scaling factors in the residual branches, preventing the correlation kernel from converging to 1. This analysis was carried in the limit "$n \to \infty$, then, $L \to \infty$" in Hayou et al. (2021), and recently extended in Hayou and Yang (2023) to the case where

4. Here, we assume that the inputs are non-zero, other all the pre-activations Y_l are zero, and the correlation is undefined in this case. All the results in this paper are trivial if $a = 0$ or $b = 0$. We will therefore always assume that $a, b \neq 0$.
5. To see this, assume that the inputs are normalized. In this case, the correlation between the pre-activations of the last layer for two different inputs converges to 1. This implies that as the depth grows, the network output becomes similar for all inputs, and the network no longer separates the data. This is problematic for the first step of gradient descent as it implies that the information from the data is (almost) unused in the first gradient update.

"$\min(n, L) \to \infty$", showing that commutativity holds in this case. Most of these works have provided empirical evidence showing an association between favorable characteristics of the neural covariance/correlation and good trainability properties of deep networks.[6]

4. Existing Results

In this section, we present corollaries of existing results showing different scenarios where commutativity is satisfied or not for the neural covariance. The aim of this section is show that commutativity depends on the neural architecture.

4.1 Non-Commutativity in MLPs

Let $d, n, L \geq 1$, and consider a simple MLP architecture given by the following:

$$\begin{cases} Y_0(a) = W_{in}a, \quad a \in \mathbb{R}^d \\ Y_l(a) = W_l\phi(Y_{l-1}(a)),\ l \in [L], \end{cases} \tag{2}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the ReLU activation function, $W_{in} \in \mathbb{R}^{n\times d}$, and $W_l \in \mathbb{R}^{n\times n}$ is the weight matrix in the l^{th} layer. We assume that the weights are randomly initialized with *iid* Gaussian variables $W_l^{ij} \sim \mathcal{N}(0, \frac{2}{n})$,[7] $W_{in}^{ij} \sim \mathcal{N}(0, \frac{1}{d})$. While the activation function is only defined for real numbers (1-dimensional), we abuse the notation and write $\phi(z) = (\phi(z^1), \dots, \phi(z^k))$ for any k-dimensional vector $z = (z^1, \dots, z^k) \in \mathbb{R}^k$ for any $k \geq 1$. We refer to the vectors $\{\phi(Y_l), l = 0, \dots, L\}$ as *post-activations*.

In the case of the joint limit $n, L \to \infty$ with n/L fixed, it has been shown that the covariance/correlation between $Y_{\lfloor tL \rfloor}(a)$ and $Y_{\lfloor tL \rfloor}(b)$ becomes similar to that of a Markov chain that incorporates random terms. However, the correlation still converges to 1 in this limit.

Proposition 1 (Correlation, (Hayou et al., 2019; Li et al., 2023)). *Consider the MLP architecture given by Eq.* (2) *and let* $a, b \in \mathbb{R}^d$ *such that* $a, b \neq 0$. *Then, in the limit "*$n \to \infty$*, then* $L \to \infty$*" or the the joint limit "*$n, L \to \infty$*,* L/n *fixed", the correlation* $\frac{\langle Y_L(a), Y_L(b)\rangle}{\|Y_L(a)\|\|Y_L(b)\|}$ *converges*[8] *weakly to 1.*

The convergence of the correlation to 1 in the infinite depth limit of a neural network poses a significant issue, as it indicates that the network loses all of the covariance structure from the inputs as the depth increases. This results in degenerate gradients (see e.g. Schoenholz et al. (2017)), rendering the network untrainable. To address this problem in MLPs, various studies have proposed the use of depth-dependent shaped ReLU activations, which prevent the correlation from converging to 1 and exhibit stochastic differential equation (SDE) behavior. As a result, the correlation of the last layer does not converge to a deterministic value in this case.

6. By favorable characteristics of the neural covariance, we refer for instance to non-degeneracy as $L \to \infty$ as reported in (Hayou et al., 2021).
7. This is the standard He initialization which coincides with the Edge of Chaos initialization (Schoenholz et al., 2017). This is the only choice of the variance that guarantees stability in both the large-width and the large-depth limits.
8. Note that weak convergence to a constant implies also convergence in probability.

Proposition 2 (Correlation SDE, Corollary of Thm 3.2 in Li et al. (2023)). *Consider the MLP architecture given by Eq.* (2) *with the following activation function* $\phi_L(z) = z + \frac{1}{\sqrt{L}}\phi(z)$ *(a modified ReLU). Let* $a, b \in \mathbb{R}^d$ *such that* $a, b \neq 0$. *Then, in the joint limit "*$n, L \to \infty$, L/n *fixed", the correlation* $\frac{\langle Y_L(a), Y_L(b)\rangle}{\|Y_L(a)\|\|Y_L(b)\|}$ *converges weakly to a nondeterministic random variable.*[9]

The joint limit, therefore, yields non-deterministic behavior of the covariance structure. It is easy to check that even with shaped ReLU as in Proposition 2, taking the width to infinity first, then depth, the result is a deterministic covariance structure. The main takeaway from this section is the following:

Corollary 1. *With MLPs (Eq.* (2)*), the width and depth limits do not commute for the neural covariance/correlation.*

4.2 Commutativity with Scaled Residual Networks

Using the same notation as in the MLP case, consider the following ResNet architecture of width n and depth L

$$\begin{aligned} Y_0(a) &= W_{in}a, \quad a \in \mathbb{R}^d \\ Y_l(a) &= Y_{l-1}(a) + \frac{1}{\sqrt{L}} W_l \phi(Y_{l-1}(a)),\ l \in [1:L], \end{aligned} \tag{3}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the ReLU activation function. Assume that the weights are randomly initialized with *iid* Gaussian variables $W_l^{ij} \sim \mathcal{N}(0, \frac{1}{n})$, $W_{in}^{ij} \sim \mathcal{N}(0, \frac{1}{d})$. If we consider the set of scaling factors of the form $L^{-\gamma}$ for $\gamma > 0$, then the choice of $\gamma = 1/2$ is the smallest value of γ such that the network output do not explode in the infinite-depth limit (see Lemma 1). Therefore, in some sense, this scaling is the 'optimal' amongst uniform scalings (meaning all residual branches are scaled with the same factor) for two reasons: it stabilizes the network as depth increases, and it does not result in trivial behavior (see discussion after Proposition 3).

With the ResNet architecture Eq. (3), we have the following result for the covariance kernel, which establishes commutativity in this case.

Proposition 3 (Thm 2 in Hayou and Yang (2023)). *Let* $a, b \in \mathbb{R}^d$ *such that* $a, b \neq 0$ *and* $a \neq b$. *Then, we have the following*

$$\sup_{t\in[0,1]} \left\| q_{\lfloor tL \rfloor, n}(a, b) - q_t(a, b) \right\|_{L_2} \leq C \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{L}} \right)$$

9. In Li et al. (2023), the authors show that the correlation of $\frac{\langle \phi_L(Y_L(a)), \phi_L(Y_L(b))\rangle}{\sqrt{\|\phi_L(Y_L(a))\|}\sqrt{\|\phi_L(Y_L(b))\|}}$ converges to a random variable in the joint limit. Since ϕ_L converges to the identity function in this limit, simple calculations show that the correlation between the pre-activations $\frac{\langle Y_L(a), Y_L(b)\rangle}{\|Y_L(a)\|\|Y_L(b)\|}$ is also random in this limit.

where C is a constant that depends only on $\|a\|$, $\|b\|$, and d, and $q_t(a,b)$ is the solution of the following differential flow

$$\begin{cases} \frac{dq_t(a,b)}{dt} &= \frac{1}{2}\frac{f(c_t(a,b))}{c_t(a,b)} q_t(a,b), \\ c_t(a,b) &= \frac{q_t(a,b)}{\sqrt{q_t(a,a)}\sqrt{q_t(b,b)}}, \\ q_0(a,b) &= \frac{\langle a,b\rangle}{d}, \end{cases} \tag{4}$$

where the function $f : [-1,1] \to [-1,1]$ is given by

$$f(z) = \frac{1}{\pi}(z \arcsin(z) + \sqrt{1-z^2}) + \frac{1}{2}z.$$

This result suggests that commutativity for the neural covariance depends on the architecture, and holds in this particular case. More importantly, with this residual architecture, taking the width and depth limits to infinity yield a non-trivial limit of the neural covariance given by the function q_t. In (Hayou et al., 2021), it was shown that q_t is a universal kernel, meaning that, it is not only non-trivial, but one can approximate any sufficiently smooth function on some compact set with features from the kernel q_t. This has a number of implications, especially in the context of neural network Gaussian processes. We invite the reader to check (Hayou et al., 2021) for a more in-depth discussion. Another recent result showed that trivial behavior can be avoided by scaling the main branch of the ResNet. The neural covariance converges weakly to a random variable in the proportional limit, which implies that such scaling breaks commutativity.

Proposition 4 (Corrollary of Thm 3.2 in Noci et al. (2023))**.** *Conider a ResNet where the hidden layers are of the form $Y_l(a) = \beta Y_{l-1}(a) + \sqrt{1-\beta^2} W_l \phi_L(\tilde{W}_l Y_{l-1}(a))$, where $\beta \in (0,1)$ is a constant, W_l and $\tilde{W}_l$ are weight matrices initialized as $\mathcal{N}(0, n^{-1})$, and ϕ_L is the shaped ReLU (defined in Proposition 2). Then, the width and depth limits for the covariance kernel do not commute in this case.*

Scaling the main branch of the residual network results in a similar behavior to the of the MLP case. Intuitively, with the factor β, the direct contribution of any layer to the main branch decreases exponentially with depth, hence simulating the 'compositional' nature of MLPs. Note that the use of shaped ReLU is essential with this scaling in order to avoid degeneracy problems; with ReLU, the correlation converges to 1 in the infinite-depth limit. In the same work, the authors show a similar result for Transformers which is a more modern residual architecture.

With the background information provided above, we are now able to present our findings. In the next section, we demonstrates commutativity of the width and depth limits for a general class of ResNet architectures, extending the results of Hayou and Yang (2023).

5. Main Results: Commutativity under General Scaling

In this section, we present our main results regarding commutativity of the width and depth limits under general scaling rules. All the proofs are deferred to the Appendix. We first define the *sequence of scaling factors*, a notion that will be frequently used in the paper.

Definition 5 (Sequence of Scaling Factors)**.** *A sequence of scaling factors is an infinite triangular array of non-negative real numbers. It has the form* $\alpha = (\alpha_{l,L})_{l\in\{1,\dots,L\},L\geq 1}$.

Visually, one can think of α as an infinite object of the form

$$\alpha = \begin{cases} \alpha_{1,1} & & & \\ \alpha_{1,2} & \alpha_{2,2} & & \\ \vdots & \vdots & \ddots & \\ \alpha_{1,L} & \dots & \alpha_{L,L} & \\ \vdots & \dots & \dots & \ddots \end{cases}$$

The use of such notation will come handy when we scale up the depth of a neural network. Such sequences will are used to define a scaling strategy as network depth grows.

Setup. Recall the previously introduced notation, the width and depth of the network are denoted by n and L, respectively, and the input dimension is denoted by d. Let $n, L, d \geq 1$, and consider the following neural network model with skip connections

$$\begin{cases} Y_0(a) = W_{in}a, \quad a \in \mathbb{R}^d, \\ Y_l(a) = Y_{l-1}(a) + \alpha_{l,L} W_l \, \phi(Y_{l-1}(a)), \quad l \in [L], \end{cases} \tag{5}$$

where ϕ is the ReLU activation function, $W_{in} \in \mathbb{R}^{n\times d}$ is the input layer weight matrix, and $W_l \in \mathbb{R}^{n\times n}$ is the weight matrix in the l^{th} layer. We assume that the weights are randomly initialized as $W_d^{ij} \sim \mathcal{N}(0, 1/d)$, and $W_l^{ij} \sim \mathcal{N}(0, 1/n)$ for $l \in [L]$, $a \neq 0$ is an arbitrary input in $\mathbb{R}^d$, $\alpha = (\alpha_{l,L})_{L\geq 1, l\in[L]}$ is a sequence of scaling factors. For the sake of simplification, we only consider networks with no bias, and we omit the dependence of Y_l on n and L in the notation. For a vector $Z \in \mathbb{R}^k$, we write $Z = (Z^1, Z^2, \dots, Z^k) \in \mathbb{R}^k$ to denote its entries. Hereafter, we consider two inputs $a, b \in \mathbb{R}^d$ satisfying $a, b \neq 0$ and $\langle a, b\rangle$.[10]

As depth increases, the pre-activations might grow arbitrarily large, depending on the choice of the sequence α. The next result fully characterizes sequences that guarantee stability in terms of the L_2 norm.

Lemma 1. *For all* $L \geq 1, l \in [L], i \in [n]$

$$\mathbb{E}\left[Y_l^i(a)^2\right] = \frac{\|a\|^2}{d} \prod_{k=1}^{l} \left(1 + \frac{\alpha_{k,L}^2}{2}\right).$$

As a result, $\sup_{l\in[L],L\geq 1,i\in[n]} \mathbb{E}\left[Y_l^i(a)^2\right]$ *is bounded iff* $\sup_{L\geq 1} \sum_{l=1}^{L} \alpha_{l,L}^2 < \infty$.

Proof. Simple calculations yield

$$\mathbb{E}\left[Y_l^i(a)^2\right] = \mathbb{E}\left[Y_{l-1}^i(a)^2\right] + \alpha_{l,L}^2 \mathbb{E}\left[\phi(Y_l^i(a))^2\right].$$

To conclude, it suffices to see that $Y_l^i(a)^2$ is a symmetric random variable, and therefore $\mathbb{E}\left[\phi(Y_l^i(a))^2\right] = \frac{1}{2}\mathbb{E}\left[Y_l^i(a)^2\right]$. □

10. These conditions on a, b are generally satisfied in practical scenarios. From a theoretical standpoint, we added these conditions in order to avoid dealing with division by 0 etc. These cases are trivial and can be easily incorporated in the main results. However, we believe this is an unnecessary complication that does not add any value to the results.

The result of Lemma 1 is independent from the width n. Hence, a necessary and sufficient condition so that the pre-activations do not explode with depth (in L_2 norm), for any width n, is to have $\sup_{L\geq 1}\sum_{l=1}^{L}\alpha_{l,L}^2 < \infty$. We say that such sequences of scaling factors are stable.

Definition 6 (Stable Sequence of Scaling Factors)**.** *Let α be a sequence of scaling factors. We say that α is stable if it satisfies $\sup_{L\geq 1}\sum_{l=1}^{L}\alpha_{l,L}^2 < \infty$. We denote the space of stable sequences of scaling factors by $\mathcal{S}$. For $\alpha \in \mathcal{S}$, we define the $\mathcal{S}$-norm of α by $\|\alpha\|_S = \sqrt{\sup_{L\geq 1}\sum_{l=1}^{L}\alpha_{l,L}^2}$.*[11]

Stable Sequences of Scaling Factors have appeared Hayou et al. (2021). In that work, the sequential limit 'infinite-width, then infinite-depth' was considered, and such sequences were proven to stabilize the gradients as well, and yield other favorable network properties regarding the neural covariance kernel and the neural tangent kernel.

In the next two (sub)sections, we show that unlike in MLPs or residual networks with scaled main branch where the neural covariance/correlation exhibits different limiting behaviors depending on how the width and depth limits are taken, under general conditions on the sequence α, for the ResNet architecture given by Eq. (5), the neural covariance converges strongly to a deterministic kernel, which depends on the choice of the sequence α, in the limit $\min(n, L) \to \infty$ regardless of the relative rate at which n and L tend to infinity. We show different examples and recover and strengthen previous results as special cases.

5.1 Sequence of Scaling Factors as Convergent Series

In this section, we consider sequences α that "converge" to a series in a specific way. We show that in this case, the neural covariance kernel converges to the same limiting kernel with a specific convergence rate as long as $\min(n, L)$ goes to infinity, hence inducing commutativity.

Theorem 1. *Let $\alpha \in \mathcal{S}$. Assume that there exists a sequence $\zeta = (\zeta_i)_{i\geq 1} \in \ell_2(\mathbb{N})$ such that $\sum_{l=1}^{L}|\alpha_{l,L}^2 - \zeta_l^2| \to 0$ as $L \to \infty$. Then, we have that for all $t \in (0, 1]$*

$$\|q_{\lfloor tL\rfloor,n}(a,b) - q_\infty^\zeta(a,b)\|_{L_2} \leq C\left(n^{-1/2} + \sum_{l=1}^{L}|\alpha_{l,L}^2 - \zeta_l^2| + \sum_{l\geq L}\zeta_l^2\right),$$

where C is a constant that depends only on $t, \|a\|, \|b\|, d, \|\zeta\|_S$, and $q_\infty^\zeta(a,b) = \lim_{L\to\infty} q_L^\zeta(a,b)$ and q_L^ζ are given by the recursive formula

$$\begin{cases} q_L^\zeta(a,b) = q_{L-1}^\zeta(a,b) + \frac{1}{2}\zeta_L^2 \frac{f(c_{L-1}(a,b))}{c_{L-1}(a,b)} q_{L-1}(a,b), \quad L \geq 1 \\ c_L(a,b) = \frac{q_L(a,b)}{\sqrt{q_L(a,a)q_L(b,b)}}, \\ q_0^\zeta(a,b) = \frac{\langle a,b\rangle}{d}, \end{cases}$$

where $f : [-1, 1] \to [-1, 1]$ is given by

$$f(z) = \pi^{-1}(z\arcsin z + \sqrt{1-z^2}) + \frac{1}{2}z.$$

11. If we allow negative values for $\alpha_{l,L}$, then we can show that the space $\mathcal{S}$, endowed with the inner product $\langle\alpha,\beta\rangle_S = \sup_{L\geq 1}\sum_{l=1}^{L}\alpha_{l,L}\beta_{l,L}$, is a complete space (Banach space). We omit these technicalities in this paper.

Theorem 1 shows that the neural covariance kernel converges to the same limiting kernel no matter how the width and depth limits are taken. In the proof, provided in Appendix D, we first show the existence of the limit of q_L, then proceed to bound the difference with the neural covariance kernel. The convergence rate depends on he properties of the series ζ that approximates α as depth grows. Notice that the limiting kernel q_∞^ζ does not depend on $t \in (0, 1]$. This is because the entries of ζ do not depend on depth L.

Examples. The conditions of Theorem 1 are satisfied by many sequences α. Examples include:

- "Decreasing" scaling: assume that $\alpha_{l,L} = \zeta_l$ for all $L \geq 1, l \in [L]$, where $\zeta \in \ell_2(\mathbb{N})$. We call this scaling decreasing because $\lim_{l\to\infty} \zeta_l = 0$. This choice of scaling factors trivially satisfies the conditions of Theorem 1 and the convergence rate is given by $\mathcal{O}(n^{-1} + \sum_{l\geq L} \zeta_l^2)$. An examples of such scaling was studied in (Hayou et al., 2021) and empirical results (performance of trained networks) were reported with $\zeta = \left((l \log(l+1)^2)^{-1/2}\right)_{l\geq 1}$.

- "Aggressive" Uniform scaling: assume that $\alpha_{l,L} = L^{-\gamma}$ for some constant $\gamma > 1/2$. This scaling is called uniform because all the residual branches have the same scaling factor. This sequence of scaling factors satisfies the conditions of Theorem 1 with $\zeta = 0_{\ell_2(\mathbb{N})}$. The convergence rate is given by $\mathcal{O}(n^{-1} + L^{-(2\gamma-1)})$, and the limiting kernel is trivial and given by $q_\infty^\zeta = q_0^\zeta$, hence the wording 'aggressive' since this scaling removes all contributions of the hidden layers in the limiting kernel. Note that this case covers the Neural ODE limit with scaling factors $\alpha_{l,L} = L^{-1}$. In the next section, we will see that another kind of uniform scaling(non-aggressive) that yield non-trivial limits.

5.2 Normalized Sequences of Scaling Factors

In this section, we discuss another type of sequences of scaling factors. We know from Hayou and Yang (2023) that with $\alpha_{l,L} = L^{-1/2}$, the limiting kernel is given by the solution of an ODE. In this section, we generalize this result by considering all sequences α that satisfy the condition $\sum_{l=1}^L \alpha_{l,L}^2 = 1$ for all $L \geq 1$. Let us first give a formal definition of such sequences.

Definition 7 (Normalized Sequence of Scaling Factors)**.** *Let α be a sequence of scaling factors. We say that α is normalized if it satisfies $\sum_{l=1}^L \alpha_{l,L}^2 = 1$ for all $L \geq 1$. The space of normalized sequences of scaling factors is denoted by $\mathcal{S}_1$.*

It is trivial that $\mathcal{S}_1 \subset \mathcal{S}$, and for all $\alpha \in \mathcal{S}_1, \|\alpha\|_{\mathcal{S}} = 1$ (hence the subscript in $\mathcal{S}_1$). The next result establishes commutativity of the infinite width and depth limit for normalized sequences.

Theorem 2 (Universal Limits for Normalized Sequences)**.** *Consider a sequence of scaling factors $\alpha \in \mathcal{S}_1$. Let $h_L = \max_{1\leq l\leq L} \alpha_{l,L}^2$ and assume that $Lh_L^2 = o(1)$. Then, we have that*

$$\sup_{t\in(0,1]} \|q_{\lfloor tL\rfloor,n}(a,b) - q_{t_L}(a,b)\|_{L_2} \leq C\left(n^{-1/2} + h_L + Lh_L^2\right),$$

where C depends only on $\|a\|, \|b\|, d$, $t_L = \sum_{k=1}^{\lfloor tL \rfloor} \alpha_{k,L}^2$, and q_t is given by the solution of the following differential flow

$$\begin{cases} \frac{dq_t(a,b)}{dt} &= \frac{1}{2} \frac{f(c_t(a,b))}{c_t(a,b)} q_t(a,b), \\ c_t(a,b) &= \frac{q_t(a,b)}{\sqrt{q_t(a,a)}\sqrt{q_t(b,b)}}, \\ q_0(a,b) &= \frac{\langle a,b \rangle}{d}, \end{cases} \tag{6}$$

where the function $f : [-1,1] \to [-1,1]$ is given by

$$f(z) = \frac{1}{\pi}(z \arcsin(z) + \sqrt{1-z^2}) + \frac{1}{2}z.$$

Moreover, assume that there exists a function $\lambda : [0,1] \to [0,1]$ such that the sequence α satisfies $\sup_{t\in[0,1]} \left| \sum_{k=1}^{\lfloor tL \rfloor} \alpha_{k,L}^2 - \lambda(t) \right| \leq r_L$ and $\lim_{L\to\infty} r_L = 0$. Then, we have

$$\sup_{t\in(0,1]} \|q_{\lfloor tL \rfloor, n}(a,b) - q_{\lambda(t)}(a,b)\|_{L_2} \leq C' \left(n^{-1/2} + h_L + L h_L^2 + r_L \right),$$

where C' depends only on $\|a\|, \|b\|, d$.

Theorem 2 generalizes previous results from Hayou and Yang (2023) to arbitrary normalized sequences. Using this theorem, we recover those results by choosing $\alpha_{l,L} = L^{-1/2}$ and verifying the conditions in the theorem. In particular, with the new proof techniques developed in this paper, we obtain a stronger convergence rate for depth.

Corollary 2 (Normalized Uniform Scaling)**.** *Assume that $\alpha_{l,L} = L^{-1/2}$ for all $L \geq 1$ and $l \in [L]$. Then, the results of Theorem 2 are satisfied with $\lambda(t) = t$, $r_L = L^{-1}$, and $h_L = L^{-1}$. As a result, we have that*

$$\sup_{t\in(0,1]} \|q_{\lfloor tL \rfloor, n}(a,b) - q_t(a,b)\|_{L_2} \leq C \left(n^{-1/2} + L^{-1} \right),$$

where C depends only on $\|a\|, \|b\|, d$, and q_t is defined in Theorem 2.

Proof. With $\alpha_{l,L} = L^{-1/2}$, we trivially have $h_L = L^{-1}$ and $L h_L^2 = L^{-1}$. Moreover, given $t \in (0,1]$ we have that $\sum_{k=1}^{\lfloor tL \rfloor} \alpha_{l,L}^2 = \frac{\lfloor tL \rfloor}{L}$, and therefore $\left| \sum_{k=1}^{\lfloor tL \rfloor} \alpha_{l,L}^2 - t \right| \leq L^{-1}$. □

Improved Depth Rate. In this paper, we obtain a depth rate of L^{-1} in contrast to the $L^{-1/2}$ convergence rate reported in Hayou and Yang (2023). The reason lies in the differences of the proof techniques used to derive the results. The proof techniques in both results are essentially 'orthogonal' in the following sense: in Hayou and Yang (2023), the proofs rely on taking the depth to infinity first, while controlling the effect of width at the same time. With this approach, the best depth rate one can obtain is $L^{-1/2}$ which is induced by the Euler disctretization error (note that with $\alpha_{l,L} = L^{-1/2}$, the ResNet behaves as the solution of a Stochastic Differential Equation (SDE) in the infinite depth limit when the width is fixed). However, in the present work, we first take the width to infinity while controlling the depth. By doing this, all the randomness in the covariance is removed as $n \to \infty$, regardless

of the L. As a result, by taking depth to infinity, we deal with deterministic dynamical systems instead of stochastic ones (the SDE case), in which case the Euler disctretization error is of order L^{-1}. We refer the reader to Section 7 for more details about the proof techniques.

Remark. The normalized uniform scaling is optimal in terms of the depth-related error in Theorem 2. More precisely, the depth-related error is given by the term $\mathcal{R}_L(\alpha) = h_L + Lh_L^2 + r_L$ up to constant C. A natural question is to ask what properties should the sequence of scaling factors satisfy in order to minimize this error. Given a fixed depth L, this problem can be formulated as a constrained minimization problem

$$\min_{\alpha \in \mathcal{S}_1} \mathcal{R}_L(\alpha) = h_L + Lh_L^2 + r_L, \tag{7}$$

where the constraint is given by the fact that $\alpha \in \mathcal{S}_1$.

Lemma 2. *The normalized uniform scaling given by $\alpha_{l,L} = L^{-1/2}$ is a solution to problem (7).*

To explain the intuition behind the result of Lemma 2, we first need to understand what each term in $\mathcal{R}_L$ represents. The first term h_L is well-known in numerical methods and represents the Euler discretization (global) error. The second term Lh_L^2 is a bound on the error between the Euler scheme of the ODE satisfied by q_t and the actual neural covariance kernel from the finite depth network. The last term r_L is induced by the behavior of the scaling sequence as L grows. If we consider just the sum of the first two terms, uniform scaling balances the two terms which should intuitively minimize that sum. It also happens that for this choice of scaling r_L is of the same order as $h_L + Lh_L^2$.

Note that the uniform scaling $\alpha_{l,L} = L^{-1/2}$ was used in the parametrization of the initialization weights of the GPT2 model (Radford et al., 2019).

6. Experiments

In this section, we validate our theoretical results with simulations on large width and depth residual neural networks of the form Eq. (5) with different choices of the sequence α.

6.1 Convergence of the Neural Covariance

Theorem 2 and Theorem 1 predict that the covariance $q_{l,n}(a, b)$ for two inputs a, b converges in L_2 norm in the limit $\min(n, L) \to \infty$. We empirically investigate this convergence in the the case of uniform scaling.

Uniform Scaling $\alpha_{l,L} = L^{-1/2}$. In Fig. 1, we compare the empirical covariance $q_{l,n}$ with the theoretical prediction q_t from Theorem 2 for $n \in \{2^3, 2^8, 2^{14}\}$ and $L \in \{2^1, 2^3, 2^8\}$. We chose maximum depth to be much smaller than maximum width to take into account the difference in the width and depth convergence rates: $n^{-1/2}$ versus L^{-1} in this case. The empirical L_2 error between $q_{L,n}$ and q_1 (from Theorem 2) is also reported. As the width increases, we observe an excellent match with the theory. The role of the depth is less noticeable, but for instance, with width $n = 2^14$, we can see that the L_2 error is smaller with depth $L = 256$ as compared to depth $L = 2$. The theoretical prediction q_t

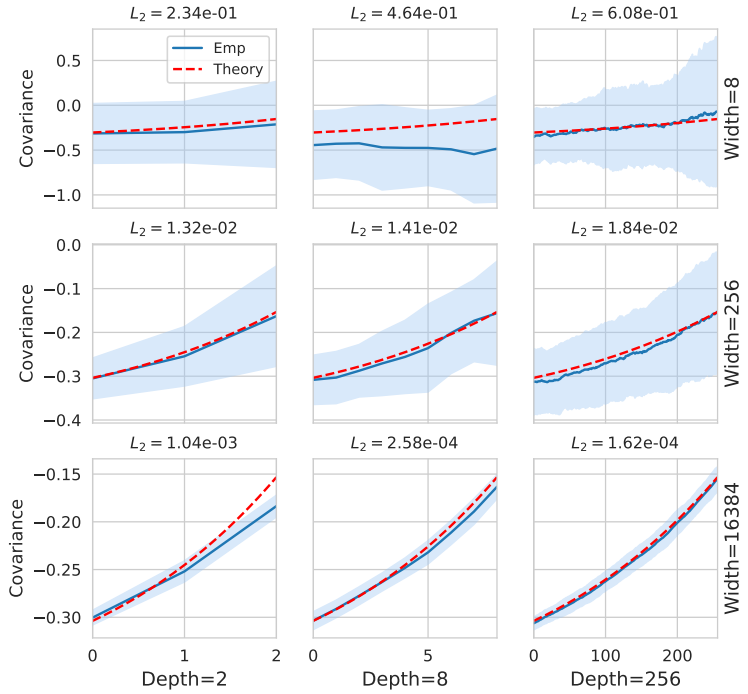

Figure 1: The blue curve represents the average covariance $q_{l,n}(a, b)$ for ResNet Eq. (5) with $n \in \{2^3, 2^8, 2^14\}$, $L \in \{2^1, 2^3, 2^8\}$, $d = 30$, and a and b are sampled randomly from $\mathcal{N}(0, I_d)$ and normalized to have $\|a\| = \|b\| = 1$. The average is calculated based on $N = 100$ simulations. The shaded blue area represents 1 standard deviation of the observations. The red dashed line represents the theoretical covariance $q_t(a, b)$ predicted in Theorem 2. The empirical L_2 error for $t = 1$ is reported.

is approximated with a PDE solver (RK45 method, Fehlberg (1968)) for $t \in [0, 1]$ with a discretization step Δt =1e-6.

6.2 Comparison with Other Architectures

In Fig. 2, we show the evolution of the distribution of $q_{L,n}$ for three different architectures with $L = n$. With our choice of scaling factors $\alpha_{l,L}$, the distribution concentrates around the deterministic limit given by the solution of the ODE described in Theorem 2. For MLP with shaped ReLU, and the Shaped ResNet (Proposition 4, the main branch is scaled with $\beta = 1/2$), we observe that the neural covariance remains random as width (and depth) grows. The sequential infinite-width-then-depth is illustrated in blue, and shows that with our choice of scaling factors, the covariance concentrates around this sequential limit even when $n = L \to \infty$. In contrast, with shaped MLP/ResNet, the two limits (sequential vs proportional) exhibit different behaviors, confirming that commutativity does not hold in these two cases.

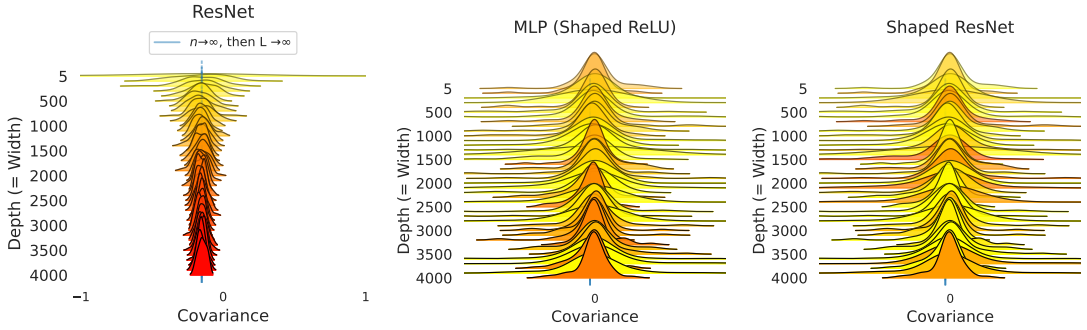

Figure 2: The distribution of the $q_{L,n}$ with $L = n$ for varying $n \in [5, 4000]$. **(Left)** ResNet described in Eq. (5) with $\alpha_{l,L} = L^{-1/2}$. **(Center)** MLP described in Eq. (2) with shaped ReLU ϕ_L Li et al. (2023). **(Right)** Shaped ResNet with $\beta = 1/2$ (Proposition 4). The vertical blue line represents the limit sequential limit $\lim_{L\to\infty} \lim_{n\to\infty} q_{L,n}$. The inputs a, b are sampled following the same procedure in Fig. 1.

6.3 Improved Depth Rate

In Hayou and Yang (2023), commutativity was established with the choice of scaling factors $\alpha_{l,L} = L^{-1/2}$. The reported convergence rate (as n and L go to infinity) of the neural covariance is of the form $\mathcal{O}(n^{-1/2} + L^{-1/2})$. In this paper, we established an improved convergence rate of order $\mathcal{O}(n^{-1/2} + L^{-1})$, which suggests that convergence is more sensitive to the width than to depth. To validate this result, we conduct the following experiment: we let n grow to infinity while fixing L and obtain the infinite-width neural covariance $q_{L,\infty}$ (infinite-width covariance for the last layer). We then measure $\Delta_L = |q_{L,\infty}(a, b) - q_{t=1}(a, b)|$ where q_t is given in Theorem 2 and (a, b) are sampled randomly following the procedure in Fig. 1. We observe a perfect match of the L^{-1} convergence rate (where the intercept was adjusted so that all the lines start from the same initial value).

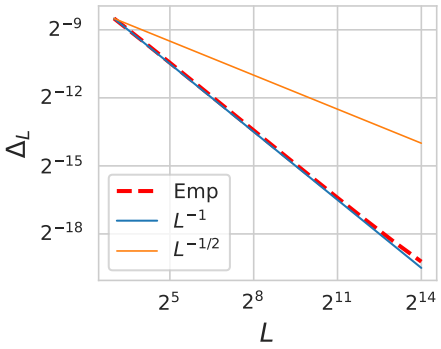

Figure 3: The curve of $\Delta_L = |q_{L,\infty}(a, b) - q_{t=1}(a, b)|$ for $L \in \{2^k, k = 3, \dots, 14\}$.

7. Outline of Proof Techniques

In Hayou and Yang (2023), it was shown that the neural covariance satisfies commutativity with the specific scaling $\alpha_{l,L} = L^{-1/2}$. The main technical novelty in that work is taking the depth L to infinity first, while controlling the dependence of the constants on the width n. Given a fixed width n, taking depth L to infinity results in an SDE behavior (Stochastic Differential Equation), and the main tools to study such convergence are numerical methods for SDEs (Euler discretization scheme). In this case, it is possible to obtain infinite-depth

strong convergence where the constants do not depend on width n. Commutativity then follows by studying the infinite-width limit of these SDEs. This involves the use of tools from mean-field stochastic calculus (namely McKean-Vlasov processes).

In the present work however, we take an orthogonal approach where the width is taken to infinity first, and the constants are well chosen so that they do not depend on depth, followed by infinite-depth which concludes the proof. The main innovation in the proofs is related to the introduction fot *the auxiliary process* $\tilde{Y}$: given a residual network of the form Eq. (5), we introduce an auxiliary process $\tilde{Y}_l$ that shares some properties with the original neural process Y_l. We bound the difference between Y_l and $\tilde{Y}_l$ using Gronwall's type of techniques, and show that the constants in this bound can be chosen to be independent of depth, hence providing a depth-uniform bound for the infinite-width limit. More importantly, the auxiliary process has iid Gaussian entries, which make it easier to study the covariance kernel related to $\tilde{Y}_l$, and allow us to conclude on commutativity. Some technical results involve the use of concentration inequalities to treat low probability events such as $\phi(Y_l) = 0$ when n is large.

8. Conclusion and Limitations

In this paper, we have shown that, at initialization, under general assumptions on the sequence of scaling factors, the large-depth and large-width limits of a residual neural network (resnet) commute for the neural covariance. We used novel proof techniques. Our results generalize and strengthen previous works on commutativity. While ReLU was specifically considered due to the closed-form expression for the covariance/correlation kernel, the proofs can be extended to general Lipschitz continuous activation functions. The Lipschitz continuity is needed to bound the term $f(c_{l-1}^{\alpha}) - f(c_{l-1}^{\beta})$ in the proof of Lemma 8, where f is the correlation kernel. However, there will be some technical differences in the proofs. For instance, with Tanh activation function, the event $\|\phi(Y_{l-1})\| = 0$ becomes a zero probability event, eliminating the need (and it is mathematically wrong) for the nested conditional expectation trick used in the proof of Theorem 4 (bounding T_1 in page 30). Extending the results to general layers is more challenging. Intuitively, for L_2 Lipschitz-continuous residual blocks (i.e. neural networks $Y_l = Y_{l-1} + \alpha_{l,L}\mathcal{F}_l(Y_{l-1}, W_l)$, where $\mathbb{E}\|\mathcal{F}_l(Z, W_l) - \mathcal{F}_l(Z', W_l)\|^2 \leq \kappa \times \mathbb{E}\|Z - Z'\|^2$), the results should in-principle hold, although additional assumptions might be needed. Unfortunately, this is not satisfied with quadratic-type layers like self-attention in Transformers architecture, in which case a more delicate analysis might be needed to prove commutativity.

Note also that our results are restricted to the neural covariance function and do not extend to other neural functions (Definition 2). More importantly, it is unclear what happens during training, and potentially, different behaviors can occur depending on how the learning rate is chosen as a function of width and depth.

One might also ask whether commutativity is needed in the current context of Large Language Models, where most architectures are in the regime $n \gg L \gg 1$ (e.g. $n \sim 1000, L \sim 50$) and that this regime can be fairly described by the sequential limit '$n \to \infty$, then $L \to \infty$'. While this might be true to some extent, note that convergence of neural functions can happen at different width and depth rates (e.g. $n^{-1/2}$ and L^{-1} in the case of

neural covariance), which implies that small changes in depth (or width) could completely change the behavior of the neural function. We leave this question for future work.

Acknowledgments

SH is supported by the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and 814639. We would like to thank the anonymous reviewers for suggesting multiple modifications that improved the paper.

References

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 2019.

John Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, Ltd, 2003. ISBN 9780470868270. doi: https://doi.org/10.1002/0470868279.fmatter. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/0470868279.fmatter`.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL `https://proceedings.neurips.cc/paper_files/paper/2009/file/5751ec3e9a4feab575962e78e006250d-Paper.pdf`.

E. Fehlberg. Classical fifth-, sixth-, seventh-, and eighth-order runge-kutta formulas with stepsize control. *NASA Technical Report*, 1968.

Boris Hanin. Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics*, 7(10), 2019.

Boris Hanin. Correlation functions in random fully connected neural networks at finite width, 2022.

Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2019.

Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations*, 2020.

Soufiane Hayou. On the infinite-depth limit of finite-width neural networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=RbLsYz1Az9`.

Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato,

and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12700–12723. PMLR, 2023. URL https://proceedings.mlr.press/v202/hayou23a.html.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2672–2680. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hayou19a.html.

Soufiane Hayou, Eugenio Clerico, Bobby He, George Deligiannidis, Arnaud Doucet, and Judith Rousseau. Stable resnet. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1324–1332. PMLR, 13–15 Apr 2021.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. Mean-field behaviour of neural tangent kernel for deep neural networks, 2022.

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: NNGP and NTK for deep attention networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4376–4386. PMLR, 2020.

Arthur Jacot. *Theory of Deep Learning: Neural Tangent Kernel and Beyond*. 2022. URL https://infoscience.epfl.ch/record/295831/files/EPFL_TH9825.pdf.

Arthur Jacot, Franck Gabriel, Francois Ged, and Clement Hongler. Freeze and chaos: Ntk views on dnn normalization, checkerboard and boundary artifacts. In Bin Dong, Qianxiao Li, Lei Wang, and Zhi-Qin John Xu, editors, *Proceedings of Mathematical and Scientific Machine Learning*, volume 190 of *Proceedings of Machine Learning Research*, pages 257–270. PMLR, 2022.

J. Lee, Y. Bahri, R. Novak, S.S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

Mufan Li, Mihai Nica, and Dan Roy. The future is log-gaussian: Resnets and their infinite-depth-and-width limit at initialization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7852–7864. Curran Associates, Inc., 2021.

Mufan Bill Li, Mihai Nica, and Daniel M. Roy. The neural covariance sde: Shaped infinite depth-and-width networks at initialization, 2023.

Fusheng Liu, Haizhao Yang, Soufiane Hayou, and Qianxiao Li. From optimization dynamics to generalization bounds via łojasiewicz gradient inequality, 2022.

James Martens, Andy Ballard, Guillaume Desjardins, Grzegorz Swirszcz, Valentin Dalibard, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping, 2021.

A.G. Matthews, J. Hron, M. Rowland, R.E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

R.M. Neal. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 1995.

Lorenzo Noci, Gregor Bachmann, Kevin Roth, Sebastian Nowozin, and Thomas Hofmann. Precise characterization of the prior predictive distribution of deep reLU networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=DTA7Bgrai-Q.

Lorenzo Noci, Chuning Li, Mufan Bill Li, Bobby He, Thomas Hofmann, Chris Maddison, and Daniel M. Roy. The shaped transformer: Attention models in the infinite depth-and-width limit, 2023.

Stefano Peluchetti and Stefano Favaro. Infinitely deep neural networks as diffusion processes. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1126–1136. PMLR, 2020.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/148510031349642de5ca0c544f31b2ef-Paper.pdf.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In *OpenAI blog (2019)*, 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

S.S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

Mariia Seleznova and Gitta Kutyniok. Analyzing finite neural networks: Can we trust neural tangent kernel theory? In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova, editors, *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 868–895. PMLR, 2022.

Pierre Wolinski and Julyan Arbel. Gaussian pre-activations in neural networks: Myth or reality?, 2023.

Lechao Xiao, Jeffrey Pennington, and Samuel Schoenholz. Disentangling trainability and generalization in deep neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10462–10472. PMLR, 2020.

Ge Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/81c650caac28cdefce4de5ddc18befa0-Paper.pdf.

Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation, 2020.

Greg Yang. Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are gaussian processes, 2021a.

Greg Yang. Tensor programs iii: Neural matrix laws, 2021b.

Greg Yang and Edward J. Hu. Feature learning in infinite-width neural networks, 2022.

Greg Yang and Hadi Salman. A fine-grained spectral perspective on neural networks, 2020.

Greg Yang, Michael Santacroce, and Edward J Hu. Efficient computation of deep nonlinear infinite-width neural networks that learn features. In *International Conference on Learning Representations*, 2022.

Jacob Zavatone-Veth and Cengiz Pehlevan. Exact marginal prior distributions of finite bayesian neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3364–3375. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/1baff70e2669e8376347efd3a874a341-Paper.pdf.

Guodong Zhang, Aleksandar Botev, and James Martens. Deep learning without shortcuts: Shaping the kernel with tailored rectifiers. In *International Conference on Learning Representations*, 2022.

Appendix

Table of Contents

Appendix A. A More Comprehensive Literature Review

Theoretical analysis of randomly initialized neural networks with an infinite number of parameters has yielded a wealth of interesting results, both theoretical and practical. Most of the research in this area has focused on the case where the depth of the network is fixed and the width is taken to infinity. However, in recent years, motivated by empirical observations, there has been an increased interest in studying the large depth limit of these networks. We provide here a non-exhaustive summary of existing results of these limits.

A.1 Infinite-Width Limit

The infinite-width limit of neural network architectures has been extensively studied in the literature and has led to many interesting theoretical and algorithmic innovations. We summarize these results below.

- *Initialization schemes*: the infinite-width limit of different neural architectures has been extensively studied in the literature. In particular, for multi-layer perceptrons (MLP), a new initialization scheme that stabilizes forward and backward propagation (in the infinite-width limit) was derived in Poole et al. (2016); Schoenholz et al. (2017). This initialization scheme is known as the Edge of Chaos, and empirical results show that it significantly improves performance. In Yang and Schoenholz (2017); Hayou et al. (2021), the authors derived similar results for the ResNet architecture, and showed that this architecture is *placed* by-default on the Edge of Chaos for any choice of the variances of the initialization weights (Gaussian weights). In Hayou et al. (2019), the authors showed that an MLP that is initialized on the Edge of Chaos exhibits similar properties to ResNets, which might partially explain the benefits of the Edge of Chaos initialization.

- *Gaussian process behavior*: Multiple papers (e.g. Neal (1995); Lee et al. (2018); Yang (2021b); Matthews et al. (2018); Hron et al. (2020)) studied the weak limit of neural networks when the width goes to infinity. The results show that a randomly initialized neural network (with Gaussian weights) has a similar behavior to that of a Gaussian process, for a wide range of neural architectures, and under mild conditions on the activation function. In Lee et al. (2018), the authors leveraged this result and introduced the neural network Gaussian process (NNGP), which is a Gaussian process model with a neural kernel that depends on the architecture and the activation function. Bayesian regression with the NNGP showed that NNGP surprisingly achieves performance close to the one achieved by an SGD-trained finite-width neural network.

  The large depth limit of this Gaussian process was studied in Hayou et al. (2021), where the authors showed that with proper scaling, the infinite-depth (weak) limit is a Gaussian process with a universal kernel[12].

- *Neural Tangent Kernel (NTK)*: the infinite-width limit of the NTK is the so-called NTK regime or Lazy-training regime. This topic has been extensively studied in the literature. The optimization and generalization properties (and some other aspects) of the NTK have been studied in Liu et al. (2022); Arora et al. (2019); Seleznova and Kutyniok (2022).

12. A kernel is called universal when any continuous function on some compact set can be approximated arbitrarily well with kernel features.

The large depth asymptotics of the NTK have been studied in Jacot et al. (2022); Xiao et al. (2020). We refer the reader to Jacot (2022) for a comprehensive discussion on the NTK.

- *Tensor programs*: It is worth mentioning that a series of works called *Tensor Programs* studied the dynamics of infinite-width limit of finite-depth general neural networks both at initialization and at finite training step t with gradient descent (Yang, 2021a,b, 2020; Yang and Hu, 2022).

A.2 Infinite-Depth Limit

Infinite-width-then-infinite-depth limit. In this case, the width of the neural network is taken to infinity first, followed by the depth. This is known as the infinite-depth limit of infinite-width neural networks. This limit has been widely used to study various aspects of neural networks, such as analyzing neural correlations and deriving the Edge of Chaos initialization scheme (Schoenholz et al., 2017; Poole et al., 2016), investigating the impact of the activation function (Hayou et al., 2019), and analyzing the behavior of the Neural Tangent Kernel (NTK) (Hayou et al., 2022; Xiao et al., 2020).

The joint infinite-width-and-depth limit. In this case, the depth-to-width ratio is fixed[13], the width and depth are jointly taken to infinity. There are a limited number of studies that have examined the joint width-depth limit. For example, in Li et al. (2021), the authors demonstrated that for a specific form of residual neural networks (ResNets), the network output exhibits a (scaled) log-normal behavior in this joint limit, which is distinct from the sequential limit where the width is taken to infinity first followed by the depth, in which case the distribution of the network output is asymptotically normal (Schoenholz et al., 2017; Hayou et al., 2019). Furthermore, in Li et al. (2023), the authors studied the covariance kernel of a multi-layer perceptron (MLP) in the joint limit and found that it weakly converges to the solution of a Stochastic Differential Equation (SDE). In Hanin and Nica (2020), it was shown that in the joint limit case, the Neural Tangent Kernel (NTK) of an MLP remains random when the width and depth jointly go to infinity, which is different from the deterministic limit of the NTK when the width is taken to infinity before depth (Hayou et al., 2022). In Hanin (2022, 2019), the authors explored the impact of the depth-to-width ratio on the correlation kernel and the gradient norms in the case of an MLP architecture and found that this ratio can be interpreted as an effective network depth. Similar results have been discussed in Zavatone-Veth and Pehlevan (2021); Noci et al. (2021, 2023).

Infinite-depth limit of finite-width neural networks. In both previous limits, the width of the neural network is taken to infinity, either in isolation or jointly with the depth. However, it is natural to question the behavior of networks where the width is fixed and the depth is taken to infinity. For example, in Hanin (2019), it was shown that neural networks with bounded width are still universal approximators, motivating the examination of finite-width large depth neural networks. The limiting distribution of the network output at initialization in this scenario has been investigated in the literature. In Peluchetti

13. Other works consider the case when the depth-to-width ratio converges to a constant instead of being fixed.

and Favaro (2020), it was demonstrated that for a specific ResNet architecture, the pre-activations converge weakly to a diffusion process in the infinite-depth limit. This a simple corollary of existing results in stochastic calculus on the convergence of Euler-Maruyama disctretization schemes to continuous Stochastic Differential Equations. Other recent work by Hayou (2023) examined the impact of the activation function on the distribution of the pre-activation, and characterized the distribution of the post-activation norms in this limit.

General limit $\min\{n, L\} \to \infty$**.** This limit is understudied, and to the best of our knowledge, it has been only studied in Hayou and Yang (2023).

Appendix B. Proof Outline

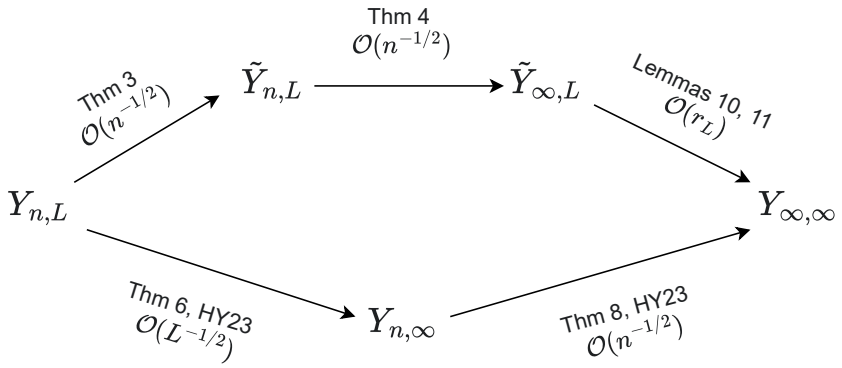

Figure 4: Proof outline and comparison with Hayou and Yang (2023).

The structure of the proofs is depicted in Fig. 4. In Hayou and Yang (2023), the proof relies on taking the infinite-depth limit first, followed by the infinite-width limit. In this paper, we first construct an auxiliary process $\tilde{Y}$ that remains within $\mathcal{O}(n^{-1/2})$ distance from Y, then take the infinite-width limit first, followed by the infinite-depth limit. We thank the anonymous reviewer for suggesting this figure in their review.

Appendix C. Depth-Uniform Infinite Width Limit: The Auxiliary Process

In this section, we aim to understand the infinite-width behavior of the pre-activations Y_l as a function of depth L. We will show that there exists a process $\tilde{Y}_l(.) : \mathbb{R}^d \to \mathbb{R}^n$ such that for any $a \in \mathbb{R}^d$, the entries $(\tilde{Y}_l^i(a))_{i\in[n]}$ are iid Gaussian random variables, and

$$n^{-1}\mathbb{E}\|Y_l(a) - \tilde{Y}_l(a)\|^2 \leq C\sum_{i=1}^{l} \alpha_{l,L}^2,$$

where $C > 0$ is a constant that depends only on the input a, and which can be made independent of $a$if the input is chosen in a compact set. A straightforward result is that if the sequence αsatisfies $\sup_{L\geq 1}\sum_{l=1}^{L}\alpha_{l,L}^2 \leq M$ for some constant M, then the convergence

rate of the neural processes Y_l to $\tilde{Y}_l$ can be upperbounded by a quantity that does not depend on depth.

C.1 Constructing $\tilde{Y}_l$

We can write the forward propagation as follows

$$Y_l(a) = Y_{l-1}(a) + \alpha_{l,L}\frac{1}{\sqrt{n}}\|\phi(Y_{l-1}(a))\|G_l(a)$$

$$G_l(a) = \begin{cases} \sqrt{n}W_l\frac{\phi(Y_{l-1}(a))}{\|\phi(Y_{l-1}(a))\|} & \text{if} \quad \|\phi(Y_{l-1}(a))\| \neq 0, \\ \sqrt{n}W_l e & \text{otherwise,} \end{cases}$$

where $e = n^{-1/2}(1, \dots, 1)^\top \in \mathbb{R}^n$ (the choice of e here is arbitrary and does not impact the identity above). As such, the vector $G_l(a)$ consists of iid standard Gaussian variables as a result of Lemma 13. Moreover, for any $l \neq l'$, the processes G_l and $G_{l'}$ are independent.

Using this auxiliary process G_l, we define the process $\tilde{Y}_l$ as follows

$$\tilde{Y}_l(a) = \tilde{Y}_{l-1}(a) + \alpha_{l,L}\left(\mathbb{E}[\phi(\tilde{Y}^1_{l-1}(a))^2]\right)^{1/2} G_l(a)$$

The *volatility* term $\mathbb{E}[\phi(\tilde{Y}^1_{l-1}(a))^2]$ in the definition of the process $\tilde{Y}_l$ can be expressed analytically. We state this result in the next lemma.

Lemma 3. *For all $l \in [L]$,*

$$q_l(a) = \mathbb{E}[(Y^1_l)^2] = \mathbb{E}[(\tilde{Y^1_l})^2] = \frac{\|a\|^2}{d}\prod_{k=1}^{l}\left(1 + \frac{\alpha^2_{k,L}}{2}\right).$$

As a result, we also have

$$\mathbb{E}\left[\phi(Y^1_l(a))^2\right] = \mathbb{E}\left[\phi(\tilde{Y}^1_l(a))^2\right] = \frac{\|a\|^2}{2d}\prod_{k=1}^{l}\left(1 + \frac{\alpha^2_{k,L}}{2}\right).$$

.

Proof. Simple calculations yield.

$$\begin{aligned}\mathbb{E}[(Y^1_l)^2] &= \mathbb{E}[(Y^1_{l-1})^2] + \alpha^2_{l,L}\mathbb{E}[\phi(Y^1_{l-1})^2] \\ &= \left(1 + \frac{\alpha^2_{l,L}}{2}\right)\mathbb{E}[(Y^1_{l-1})^2].\end{aligned}$$

Knowing that $\mathbb{E}[(Y^1_0)^2] = \frac{\|a\|^2}{d}$, we obtain $\mathbb{E}[(Y^1_l)^2] = \frac{\|a\|^2}{d}\prod_{k=1}^{l}\left(1 + \frac{\alpha^2_{l,L}}{2}\right)$. Similar calculations hold for $\mathbb{E}[(\tilde{Y^1_l})^2]$.

As a result of Lemma 3, we can express the process $\tilde{Y}_l$ by substituting the volatility term with its analytical expression. This allows us to conclude that $\tilde{Y}_l$ has iid Gaussian weights with a analytical expression of the variance. □

Lemma 4. *The process $\tilde{Y}_l$ satisfies the following*

$$\tilde{Y}_l(a) = \tilde{Y}_{l-1}(a) + \alpha_{l,L}\frac{\|a\|}{\sqrt{2d}}\prod_{k=1}^{l}\left(1+\frac{\alpha_{k,L}^2}{2}\right)^{1/2} G_l(a).$$

As a result, the entries of $\tilde{Y}_l(a)$ are iid centered Gaussian random variables with variance $Var(\tilde{Y}_l^1(a)) = \frac{\|a\|^2}{d}\prod_{k=1}^{l}\left(1+\frac{\alpha_{l,L}^2}{2}\right)$.

Note that while the the entries of $\tilde{Y}_l(a)$ are Gaussian, the process $\tilde{Y_l}(.)$ is not necessarily a Gaussian process.

C.2 Convergence Rate

In this section, we will analyze the convergence properties of different quantities as width goes to infinity.

Theorem 3. *(Depth-Uniform strong convergence rate) Let α be a stable sequence of scaling factors. Then, there exists a constant $C > 0$ that depends only on $\|a\|, d, \|\alpha\|_S$ such that*

$$\sup_{L\geq 1}\sup_{l\in[L]}\mathbb{E}\|Y_l(a) - \tilde{Y}_l(a)\|^2 \leq C.$$

As a result, we have that

$$\sup_{L\geq 1}\sup_{l\in[L]}\sup_{i\in[n]}\mathbb{E}\|Y_l^i(a) - \tilde{Y}_l^i(a)\|^2 \leq Cn^{-1}.$$

Proof. Let $a \in \mathbb{R}^d$. To alleviate the notation, we write $Y_l := Y_l(a)$ and $\tilde{Y}_l := \tilde{Y}_l(a)$. We would like to obtain recursive bounds on $\mathbb{E}\|Y_l - \tilde{Y}_l\|^2$ which would allow us to conclude. The proof technique follows Gronwall's style inequalities. We have the following

$$\mathbb{E}\|Y_l - \tilde{Y}_l\|^2 = \mathbb{E}\|Y_{l-1} - \tilde{Y}_{l-1}\|^2 + n\alpha_{l,L}^2\underbrace{\mathbb{E}\left(\frac{1}{\sqrt{n}}\|\phi(Y_{l-1})\| - \left(\mathbb{E}[\phi(\tilde{Y}_{l-1}^1)^2]\right)^{1/2}\right)^2}_{T}.$$

We bound the term T as follows

$$\begin{aligned}\mathbb{E}\left(\frac{1}{\sqrt{n}}\|\phi(Y_{l-1})\| - \left(\mathbb{E}[\phi(\tilde{Y}_{l-1}^1)^2]\right)^{1/2}\right)^2 &\leq 2\mathbb{E}\left(\frac{1}{\sqrt{n}}\|\phi(Y_{l-1})\| - \frac{1}{\sqrt{n}}\|\phi(\tilde{Y}_{l-1})\|\right)^2 \\ &\quad + 2\mathbb{E}\left(\frac{1}{\sqrt{n}}\|\phi(\tilde{Y}_{l-1})\| - \left(\mathbb{E}[\phi(\tilde{Y}_{l-1}^1)^2]\right)\right)^2 \\ &\leq \frac{2}{n}\mathbb{E}\|Y_{l-1} - \tilde{Y}_{l-1}\|^2 + 2\mathbb{E}\left(\frac{1}{\sqrt{n}}\|\phi(\tilde{Y}_{l-1})\| - \left(\mathbb{E}[\phi(\tilde{Y}_{l-1}^1)^2]\right)\right)^2\end{aligned}$$

where we have used the fact that ϕ is 1-Lipschitz.

Using the fact that the entries of $\tilde{Y}_{l-1}$ are iid, and that $q_l(a) \in \left[\frac{\|a\|^2}{d}, \frac{\|a\|^2}{d} e^{\frac{1}{2}\|\alpha\|_S^2}\right]$, standard concentration inequalities (Hoeffding's inequality) ensure that with probability at least $1 - e^{-nc}$ (where c is a constant that depends only on $\|a\|, \|\alpha\|_S, d$), we have that $\frac{1}{n}\|\phi(\tilde{Y}_{l-1})\|^2 > \frac{1}{2}\mathbb{E}[\phi(\tilde{Y}_{l-1}^1)^2] = \frac{1}{4}q_{l-1}(a)$.

Using this results combined with the fact that $|\sqrt{x_1} - \sqrt{x_2}| \leq \frac{1}{2\sqrt{x_0}}|x_1 - x_2|$ for all $x_1, x_2 > x_0 > 0$, we obtain that

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{\sqrt{n}}\|\phi(\tilde{Y}_{l-1})\| - \left(\mathbb{E}[\phi(\tilde{Y}_{l-1}^1)^2]\right)\right)^2 &\leq 2e^{-nc}q_{l-1}(a) + \frac{2}{q_{l-1}(a)}\mathbb{E}\left(\frac{1}{n}\|\phi(\tilde{Y}_{l-1})\|^2 - \mathbb{E}[\phi(\tilde{Y}_{l-1}^1)^2]\right)^2 \\
&\leq 2e^{-nc}q_{l-1}(a) + \frac{2}{nq_{l-1}(a)}\mathbb{E}[\phi(\tilde{Y}_{l-1}^1)^4] \\
&\leq 2e^{-nc}q_{l-1}(a) + \frac{2q_{l-1}(a)}{n}\mathbb{E}[\phi(Z)^4]
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$. As a result, there exists a constant $C_1 > 0$ that depends only on $\|a\|, d$, and $\|\alpha\|_S$, such that

$\mathbb{E}\left(\frac{1}{\sqrt{n}}\|\phi(\tilde{Y}_{l-1})\| - \left(\mathbb{E}[\phi(\tilde{Y}_{l-1}^1)^2]\right)\right)^2 \leq C_1 n^{-1}$. Hence, denoting $\Delta_l = n^{-1}\mathbb{E}\|Y_l - \tilde{Y}_l\|^2$, we have that

$$\Delta_l \leq (1 + 2\alpha_{l,L}^2)\Delta_{l-1} + 2C\alpha_{l,L}^2 n^{-1}.$$

Given that $\Delta_0 = 0$, we obtain

$$\Delta_l \leq n^{-1} \times 2C_1 \sum_{i=1}^{l} \alpha_{i,L}^2 \prod_{k=i+1}^{l} (1 + \alpha_{k,L}^2) \leq 2C_1\|\alpha\|_S^2 e^{\|\alpha\|_S^2} n^{-1},$$

which concludes the proof. □

As a result of this theorem, we have the following result (a useful lemma for subsequent proofs).

Lemma 5. *Let $a \in \mathbb{R}^d, \zeta \in [0, (8d)^{-1/2}\|a\|)$. For $L \geq 1$ and $l \in [L]$, define the event*

$$\mathcal{H}_a^l = \{\|\phi(Y_l(a))\| > \zeta n^{1/2}\} \cap \{\|\phi(\tilde{Y}_l(a))\| > \zeta n^{1/2}\}.$$

Then, we have that $\mathbb{P}(\mathcal{H}_a^l) \geq 1 - Cn^{-1}$, where C is a constant that depends only on $\|a\|, d$, and $\|\alpha\|_S$.

Proof. We have that $\mathcal{H}_a^l = E_a \cap \tilde{E}_a$, where $E_a = \{\|\phi(Y_l(a))\| > \zeta n^{1/2}\}$, and $\tilde{E}_a = \{\|\phi(\tilde{Y}_l(a))\| > \zeta n^{1/2}\}$. For some event A, let A^c denote its complimentary event. Using the fact that the entries of $\tilde{Y}_l(a)$ are iid zero-mean Gaussians, we have that

$$\begin{aligned}
\mathbb{P}(\tilde{E}_a^c) = \mathbb{P}(\|\phi(\tilde{Y}_l(a))\| \leq \zeta n^{1/2}) &= \mathbb{P}(\frac{1}{n}\|\phi(\tilde{Y}_l(a))\|^2 \leq \zeta^2) \\
&\leq \mathbb{P}(\frac{1}{n}\|\phi(\tilde{Y}_l(a))\|^2 \leq \frac{1}{4}q_l(a)) \\
&\leq e^{-nC_1}
\end{aligned}$$

where C_1 is a constant that depends only on $\|a\|, d, \|\alpha\|_S, \zeta$, where we have used the same techniques as in the proof of Theorem 3 (Hoeffding's inequality).

Now let $\kappa = \left(\frac{q_l(a)}{8} n\right)^{1/2}$. We have that

$$\mathbb{P}(E_a^c) = \mathbb{P}(\|\phi(Y_l(a))\| \leq \zeta n^{1/2}) \leq \mathbb{P}(\|\phi(\tilde{Y}_l(a))\| \leq \kappa + \zeta n^{1/2}) + \mathbb{P}(\|\phi(Y_l(a)) - \phi(\tilde{Y}_l(a))\| > \kappa).$$

Therefore we obtain

$$\mathbb{P}(\|\phi(\tilde{Y}_l(a))\| \leq \kappa + \zeta n^{1/2}) \leq \mathbb{P}(\frac{1}{n}\|\phi(\tilde{Y}_l(a))\| \leq \frac{1}{4} q_l(a)) \leq e^{-nC_1}.$$

Using Theorem 3, Markov's inequality, and the fact that ϕ is 1-Lipschitz, we have that

$$\begin{aligned}\mathbb{P}(\|\phi(Y_l(a)) - \phi(\tilde{Y}_l(a))\| > \kappa) &\leq \kappa^{-2}\mathbb{E}\|\phi(Y_l(a)) - \phi(\tilde{Y}_l(a))\|^2 \\ &\leq \frac{4K}{q_l(a)} n^{-1} \leq \frac{4Kd}{\|a\|^2} n^{-1}.\end{aligned}$$

Combining both bounds, there exists a constant C_2 that depends only on $\|a\|, d, \|\alpha\|_S, \zeta$, such that $\mathbb{P}(E_a^c) \leq C_2 n^{-1}$. □

C.3 Infinite-Width Limits of the Neural Covariance

The auxiliary process $\tilde{Y}_l$ was introduced for two reasons:

1. The distance between $\tilde{Y}_l$ and Y_l as n grows can be upperbounded so that the constants do not depend on depth.
2. It is easier to study the covariance kernel of the $\tilde{Y}_l$ instead of that of Y_l as n and l go to infinity.

We dealt with (1) in the previous section, now we deal with (2).

Define the covariance kernel of the auxiliary process

$$\tilde{q}_{l,n}(a, b) = n^{-1}\langle \tilde{Y}_l(a), \tilde{Y}_l(b)\rangle.$$

This covariance kernel satisfies the following recursion

$$\begin{aligned}\tilde{q}_{l,n}(a, b) = \tilde{q}_{l-1,n}(a, b) &+ \alpha_{l,L}^2 n^{-1}(1/2q_{l-1}(a))^{1/2}(1/2q_{l-1}(b))^{1/2}\langle G_l(a), G_l(b)\rangle \\ &+ \alpha_{l,L} n^{-1}((1/2q_{l-1}(b))^{1/2}\langle \tilde{Y}_{l-1}(a), G_l(b)\rangle + (1/2q_{l-1}(a))^{1/2}\langle \tilde{Y}_{l-1}(b), G_l(a)\rangle)\end{aligned}$$

In the following, we will show that in the infinite-width limit, the kernel $\tilde{q}_{l,n}$ converges to a kernel $\tilde{q}_{l,\infty}$ that satisfies the following recursion

$$\tilde{q}_{l,\infty}(a, b) = \tilde{q}_{l-1,\infty}(a, b) + \alpha_{l,L}^2 (1/2q_{l-1}(a))^{1/2}(1/2q_{l-1}(b))^{1/2} f(c_{l-1}(a, b)),$$

where $f(c) := 2\mathbb{E}[\phi(Z_1)\phi(cZ_1 + \sqrt{1-c^2}Z_2)]$ with $Z_1, Z_2 \sim \mathcal{N}(0, 1)$, and $c_{l-1,\infty}(a, b) := \frac{\tilde{q}_{l-1,\infty}(a,b)}{\tilde{q}_{l-1,\infty}(a,a)^{1/2}\tilde{q}_{l-1,\infty}(b,b)^{1/2}}$ (the infinite-width correlation kernel).

Remark. Observe that $\tilde{q}_{l-1,\infty}(a,a) = q_{l-1}(a)$. (proof is straightforward by induction).

Now we derive non-asymptotic convergence rates for the covariance kernel $\tilde{q}_{l,n}$ in the infinite-width limit. Similar to the analysis in the previous section, define the L_2 error between the kernels by $\tilde{\Delta}_{l,n} := \mathbb{E}\,|\tilde{q}_{l,n}(a,b) - \tilde{q}_{l,\infty}(a,b)|^2$. Simple calculations yield

$$\begin{aligned}
\tilde{\Delta}_{l,n} = &\tilde{\Delta}_{l-1,n} + \mathbb{E}(\alpha_{l,L}^2 (1/2q_{l-1}(a))^{1/2}(1/2q_{l-1}(b))^{1/2}(n^{-1}\langle G_l(a), G_l(b)\rangle - f(c_{l-1}(a,b))) \\
&+ \alpha_{l,L} n^{-1}((1/2q_{l-1}(b))^{1/2}\langle \tilde{Y}_{l-1}(a), G_l(b)\rangle + (1/2q_{l-1}(a))^{1/2}\langle \tilde{Y}_{l-1}(b), G_l(a)\rangle))^2 \\
\leq &\tilde{\Delta}_{l-1,n} + \underbrace{\frac{1}{2}\alpha_{l,L}^4 q_{l-1}(a)q_{l-1}(b)\mathbb{E}\left(n^{-1}\langle G_l(a), G_l(b)\rangle - f(c_{l-1}(a,b))\right)^2}_{T_1} \\
&+ \underbrace{2\alpha_{l,L}^2 n^{-2}\left(q_{l-1}(b)\mathbb{E}\langle \tilde{Y}_{l-1}(a), G_l(b)\rangle^2 + q_{l-1}(a)\mathbb{E}\langle \tilde{Y}_{l-1}(b), G_l(a)\rangle^2\right)}_{T_2}
\end{aligned}$$

We will deal with the diffferent terms separately.

Bounding T_2: We have that

$$q_{l-1}(b)\,\mathbb{E}\langle \tilde{Y}_{l-1}(a), G_l(b)\rangle^2 = q_{l-1}(b)\,\mathbb{E}\|\tilde{Y}_{l-1}(a)\|^2 = nq_{l-1}(b)q_{l-1}(a) \leq \frac{\|a\|^2\|b\|^2}{d^2}e^{\|\alpha\|_S^2}$$

As a result, we obtain

$$\begin{aligned}
T_2 &= 2\alpha_{l,L}^2 n^{-2}\left(q_{l-1}(b)\mathbb{E}\langle \tilde{Y}_{l-1}(a), G_l(b)\rangle^2 + q_{l-1}(a)\mathbb{E}\langle \tilde{Y}_{l-1}(b), G_l(a)\rangle^2\right) \\
&\leq 2\frac{\|a\|^2\|b\|^2}{d^2}e^{\|\alpha\|_S^2}\alpha_{l,L}^2 n^{-1}.
\end{aligned}$$

Bounding T_1: Define the events $\mathcal{H}_a^l = \{\|\phi(Y_l(a))\| \neq 0\} \cap \{\|\phi(\tilde{Y}_l(a))\| \neq 0\}$ and $\mathcal{H}_b^l = \{\|\phi(Y_l(b))\| \neq 0\} \cap \{\|\phi(\tilde{Y}_l(b))\| \neq 0\}$. We will condition on the event $\mathcal{H}_a^{l-1} \cap \mathcal{H}_b^{l-1}$ to avoid dividing by zero. This allows us to control a conditional expectation in the following manner

$$\begin{aligned}
\mathbb{E}\left(n^{-1}\langle G_l(a), G_l(b)\rangle - f(c_{l-1}(a,b))\right)^2 \leq\, &C_1 n^{-1} \\
&+ \mathbb{E}\left[\left(n^{-1}\langle G_l(a), G_l(b)\rangle - f(c_{l-1}(a,b))\right)^2 \mid \mathcal{H}_a^{l-1} \cap \mathcal{H}_b^{l-1}\right],
\end{aligned}$$

where C_1 is a constant that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$ (using Lemma 5 with $\zeta = 0$). To alleviate the notation, we denote $\mathbb{E}_l[.] = \mathbb{E}[. \mid \mathcal{H}_a^{l-1} \cap \mathcal{H}_b^{l-1}]$. We therefore have

$$\mathbb{E}_l\left(n^{-1}\langle G_l(a),G_l(b)\rangle - f(c_{l-1}(a,b))\right)^2 \leq$$
$$3\underbrace{\mathbb{E}_l\left(n^{-1}\langle G_l(a),G_l(b)\rangle - \mathbb{E}_l\frac{\langle\phi(Y_{l-1}(a)),\phi(Y_{l-1}(b))\rangle}{\|\phi(Y_{l-1}(a))\|\|\phi(Y_{l-1}(b))\|}\right)^2}_{T_{11}}$$
$$+3\underbrace{\left(\mathbb{E}_l\frac{\langle\phi(Y_{l-1}(a)),\phi(Y_{l-1}(b))\rangle}{\|\phi(Y_{l-1}(a))\|\|\phi(Y_{l-1}(b))\|} - \mathbb{E}_l\frac{\langle\phi(\tilde{Y}_{l-1}(a)),\phi(\tilde{Y}_{l-1}(b))\rangle}{\|\phi(\tilde{Y}_{l-1}(a))\|\|\phi(\tilde{Y}_{l-1}(b))\|}\right)^2}_{T_{12}}$$
$$+3\underbrace{\left(\mathbb{E}_l\frac{\langle\phi(\tilde{Y}_{l-1}(a)),\phi(\tilde{Y}_{l-1}(b))\rangle}{\|\phi(\tilde{Y}_{l-1}(a))\|\|\phi(\tilde{Y}_{l-1}(b))\|} - f(c_{l-1}(a,b))\right)^2}_{T_{13}}$$

We deal with each one of the terms T_{11}, T_{12}, T_{13} separately.

- The first term T_{11} satisfies

$$\begin{aligned}T_{11} &= \mathbb{E}_l\left(n^{-1}\langle G_l(a),G_l(b)\rangle - \mathbb{E}_l\frac{\langle\phi(Y_{l-1}(a)),\phi(Y_{l-1}(b))\rangle}{\|\phi(Y_{l-1}(a))\|\|\phi(Y_{l-1}(b))\|}\right)^2 \\ &= n^{-1}\mathrm{Var}_l\left((w^\top\frac{\phi(Y_{l-1}(a))}{\|\phi(Y_{l-1}(a))\|})\times(w^\top\frac{\phi(Y_{l-1}(b))}{\|\phi(Y_{l-1}(b))\|})\right) \\ &\leq n^{-1}\mathbb{E}_l\left((w^\top\frac{\phi(Y_{l-1}(a))}{\|\phi(Y_{l-1}(a))\|})\times(w^\top\frac{\phi(Y_{l-1}(b))}{\|\phi(Y_{l-1}(b))\|})\right)^2,\end{aligned}$$

  where $w\sim\mathcal{N}(0,I)$. We bound this quantity using the following lemma.

  Lemma 6. *We have that* $\mathbb{E}_l\left((w^\top\frac{\phi(Y_{l-1}(a))}{\|\phi(Y_{l-1}(a))\|})\times(w^\top\frac{\phi(Y_{l-1}(b))}{\|\phi(Y_{l-1}(b))\|})\right)^2 \leq 3$, *for* $w\sim\mathcal{N}(0,I)$.

  Proof. Let $u(a) := \frac{\phi(Y_{l-1}(a))}{\|\phi(Y_{l-1}(a))\|}$ and the same for b. Expanding the term inside the expectation yields $\left((w^\top u(a))\times(w^\top u(b))\right)^2 = \sum_{i=1} w_i^4 u_i(a)^2u_i(b)^2 + \sum_{i\neq j} w_i^2w_j^2u_i(a)^2u_j(b)^2 + \zeta$, where ζ consists of terms with at least one weight having an odd exponent. For such terms, the expectation is null, and we obtain

$$\begin{aligned}\mathbb{E}_l\left((w^\top u(a))\times(w^\top u(b))\right)^2 &= \mathbb{E}_l\left(3\sum_{i=1}u_i(a)^2u_i(b)^2 + \sum_{i\neq j}u_i(a)^2u_j(b)^2\right) \\ &= \mathbb{E}_l\left(2\sum_{i=1}u_i(a)^2u_i(b)^2 + 1\right) \leq 3.\end{aligned}$$

  □

Using this Lemma, we obtain

$$T_{11} = \mathbb{E}_l \left(n^{-1} \langle G_l(a), G_l(b) \rangle - \mathbb{E} \frac{\langle \phi(Y_{l-1}(a)), \phi(Y_{l-1}(b)) \rangle}{\|\phi(Y_{l-1}(a))\| \|\phi(Y_{l-1}(b))\|} \right)^2 \leq 3n^{-1}.$$

- Now let us deal with the second term T_{12}. We use the uniform bound we obtained in Theorem 3. We have that

$$\begin{aligned}
&\left(\mathbb{E}_l \frac{\langle \phi(Y_{l-1}(a)), \phi(Y_{l-1}(b)) \rangle}{\|\phi(Y_{l-1}(a))\| \|\phi(Y_{l-1}(b))\|} - \mathbb{E}_l \frac{\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\|\phi(\tilde{Y}_{l-1}(a))\| \|\phi(\tilde{Y}_{l-1}(b))\|} \right)^2 \\
&\leq 2 \left(\mathbb{E}_l \frac{\langle \phi(Y_{l-1}(a)), \phi(Y_{l-1}(b)) \rangle}{\|\phi(Y_{l-1}(a))\| \|\phi(Y_{l-1}(b))\|} - \mathbb{E}_l \frac{\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\|\phi(Y_{l-1}(a))\| \|\phi(Y_{l-1}(b))\|} \right)^2 \\
&+ 2 \left(\mathbb{E}_l \frac{\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\|\phi(Y_{l-1}(a))\| \|\phi(Y_{l-1}(b))\|} - \mathbb{E}_l \frac{\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\|\phi(\tilde{Y}_{l-1}(a))\| \|\phi(\tilde{Y}_{l-1}(b))\|} \right)^2
\end{aligned} \tag{8}$$

Using Lemma 5 with $\zeta = 12^{-1/2} d^{-1/2} \min\{\|a\|, \|b\|\}$, then with probability at least $1 - C_2 n^{-1}$, where C_2 is a constant that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$, we have that $\|\phi(Y_{l-1}(a))\| \geq \frac{\|a\|}{2\sqrt{3d}} n^{1/2}$ and $\|\phi(Y_{l-1}(b))\| \geq \frac{\|b\|}{2\sqrt{3d}} n^{1/2}$. Therefore,

$$\begin{aligned}
&\left(\mathbb{E}_l \frac{\langle \phi(Y_{l-1}(a)), \phi(Y_{l-1}(b)) \rangle}{\|\phi(Y_{l-1}(a))\| \|\phi(Y_{l-1}(b))\|} - \mathbb{E}_l \frac{\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\|\phi(Y_{l-1}(a))\| \|\phi(Y_{l-1}(b))\|} \right)^2 \\
&\leq \frac{144 d^2}{\|a\|^2 \|b\|^2} n^{-2} \mathbb{E} \left(\langle \phi(Y_{l-1}(a)), \phi(Y_{l-1}(b)) \rangle - \langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle \right)^2 \\
&\leq \frac{288 d^2}{\|a\|^2 \|b\|^2} n^{-2} \big[\mathbb{E} \left(\langle \phi(Y_{l-1}(a)), \phi(Y_{l-1}(b)) - \phi(\tilde{Y}_{l-1}(b)) \rangle \right)^2 \\
&+ \mathbb{E} \left(\langle \phi(Y_{l-1}(a)) - \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle \right)^2 \big] \\
&\leq C_3 n^{-1},
\end{aligned}$$

where C_3 depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$ and where we have used Jensen's inequality (first line), the Lipschitz property of ReLU, and the bounds on $\mathbb{E}\|Y_{l-1} - \tilde{Y}_{l-1}\|^2$ from Theorem 3. Using the same techniques for the second term in the RHS of Eq. (8), we obtain a similar bound, and we finally get

$$T_{12} = \left(\mathbb{E}_l \frac{\langle \phi(Y_{l-1}(a)), \phi(Y_{l-1}(b)) \rangle}{\|\phi(Y_{l-1}(a))\| \|\phi(Y_{l-1}(b))\|} - \mathbb{E}_l \frac{\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\|\phi(\tilde{Y}_{l-1}(a))\| \|\phi(\tilde{Y}_{l-1}(b))\|} \right)^2 \leq C_4 n^{-1}$$

where C_4 depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$.

- It remains to bound the third term T_{13} to conclude. We have that

$$
\begin{aligned}
T_{13} &= \left(\mathbb{E}_l \frac{\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\|\phi(\tilde{Y}_{l-1}(a))\| \|\phi(\tilde{Y}_{l-1}(b))\|} - f(c_{l-1}(a,b)) \right)^2 \\
&\leq 3 \underbrace{\left(\mathbb{E}_l \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{n^{-1/2}\|\phi(\tilde{Y}_{l-1}(a))\| n^{-1/2}\|\phi(\tilde{Y}_{l-1}(b))\|} - \mathbb{E}_l \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\sqrt{\frac{1}{2}q_{l-1}(a)}\sqrt{\frac{1}{2}q_{l-1}(b)}} \right)^2}_{T_{131}} \\
&\quad + 3 \underbrace{\left(\mathbb{E}_l \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\sqrt{\frac{1}{2}q_{l-1}(a)}\sqrt{\frac{1}{2}q_{l-1}(b)}} - \mathbb{E} \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\sqrt{\frac{1}{2}q_{l-1}(a)}\sqrt{\frac{1}{2}q_{l-1}(b)}} \right)^2}_{T_{132}} \\
&\quad + 3 \underbrace{\left(\mathbb{E} \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\sqrt{\frac{1}{2}q_{l-1}(a)}\sqrt{\frac{1}{2}q_{l-1}(b)}} - f(c_{l-1}(a,b)) \right)^2}_{T_{133}}.
\end{aligned}
$$

  Simple calculations yield

$$
T_{131} = \left(\mathbb{E} \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{n^{-1/2}\|\phi(\tilde{Y}_{l-1}(a))\| n^{-1/2}\|\phi(\tilde{Y}_{l-1}(b))\|} - \mathbb{E} \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\sqrt{\frac{1}{2}q_{l-1}(a)}\sqrt{\frac{1}{2}q_{l-1}(b)}} \right)^2 \leq C_5 n^{-1},
$$

  where C_5 depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$.

  For the second term T_{132}, we have that

$$
\begin{aligned}
T_{132} &= \left(\mathbb{E}_l \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\sqrt{\frac{1}{2}q_{l-1}(a)}\sqrt{\frac{1}{2}q_{l-1}(b)}} - \mathbb{E} \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\sqrt{\frac{1}{2}q_{l-1}(a)}\sqrt{\frac{1}{2}q_{l-1}(b)}} \right)^2 \\
&\leq (1 - \mathbb{P}(\mathcal{H}_a^l \cap \mathcal{H}_b^l))^2 \left(\mathbb{E}_l \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\sqrt{\frac{1}{2}q_{l-1}(a)}\sqrt{\frac{1}{2}q_{l-1}(b)}} - \mathbb{E}_l^c \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\sqrt{\frac{1}{2}q_{l-1}(a)}\sqrt{\frac{1}{2}q_{l-1}(b)}} \right)^2 \\
&\leq C_6 n^{-2},
\end{aligned}
$$

  where we write $\mathbb{E}_l^c[.] = \mathbb{E}[. \mid (\mathcal{H}_a^l \cap \mathcal{H}_b^l)^c]$, and where C_6 is a constant that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$.

  Now let us deal with the last term T_{133}. Observe that $\mathbb{E} \frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b)) \rangle}{\sqrt{\frac{1}{2}q_{l-1}(a)}\sqrt{\frac{1}{2}q_{l-1}(b)}} = f(\tilde{c}_{l-1}(a,b))$ where $\tilde{c}_{l-1}(a,b) := \frac{\mathbb{E}_l \tilde{Y}_{l-1}^1(a)\tilde{Y}_{l-1}^1(b)}{\sqrt{q_{l-1}(a)}\sqrt{q_{l-1}(b)}}$. Using the Lipschitz property of f,

we obtain

$$\begin{aligned}
T_{133} &= \left(\mathbb{E}\frac{n^{-1}\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b))\rangle}{\sqrt{\frac{1}{2}q_{l-1}(a)}\sqrt{\frac{1}{2}q_{l-1}(b)}} - f(c_{l-1}(a,b)) \right)^2 \\
&\leq (\tilde{c}_{l-1}(a,b) - c_{l-1}(a,b))^2 \\
&= (q_{l-1}(a)q_{l-1}(b))^{-1} \left(\mathbb{E}\tilde{Y}^1_{l-1}(a)\tilde{Y}^1_{l-1}(b) - \tilde{q}_{l-1,\infty}(a,b) \right)^2 \\
&\leq 2(q_{l-1}(a)q_{l-1}(b))^{-1} \left(\mathbb{E}(\tilde{Y}^1_{l-1}(a)\tilde{Y}^1_{l-1}(b) - \tilde{q}_{l-1,n}(a,b))^2 + \mathbb{E}(\tilde{q}_{l-1,n}(a,b) - \tilde{q}_{l-1,\infty}(a,b))^2 \right) \\
&\leq C_7(n^{-1} + \tilde{\Delta}_{l-1,n})
\end{aligned}$$

where we have used the bounds on q_{l-1}, and where C_7 is a constant that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$. As a result, we have that

$$T_{13} = \left(\mathbb{E}\frac{\langle \phi(\tilde{Y}_{l-1}(a)), \phi(\tilde{Y}_{l-1}(b))\rangle}{\|\phi(\tilde{Y}_{l-1}(a))\|\|\phi(\tilde{Y}_{l-1}(b))\|} - f(c_{l-1}(a,b)) \right)^2 \leq C_8(n^{-1} + \tilde{\Delta}_{l-1,n}),$$

where C_8 is a constant that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$.

Combining all the results we obtain the following upperbound on $\tilde{\Delta}_{l,n}$

$$\tilde{\Delta}_{l,n} \leq (1 + C_9\alpha^4_{l,L})\tilde{\Delta}_{l-1,n} + C_{10}\alpha^2_{l,L}n^{-1}$$

where C_9, C_{10} are constants that depend only on $\|a\|, \|b\|, d, \|\alpha\|_S$. Using the fact that $\tilde{\Delta}_{0,n} = 0$, and that $\sum_{l=1}^{L} \alpha^4_{l,L} \leq \|\alpha\|^2_S$, we obtain that

$$\sup_{L\geq 1} \sup_{l\in[L]} \tilde{\Delta}_{l,n} \leq C_{11}n^{-1},$$

where C_{11} depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$.

We state this result formally in the next theorem.

Theorem 4. *Let α be a stable sequence of scaling factors. There exists a constant $C > 0$ that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$, such that $\sup_{L\geq 1} \sup_{l\in[L]} \mathbb{E}\,|\tilde{q}_{l,n}(a,b) - \tilde{q}_{l,\infty}(a,b)|^2 \leq Cn^{-1}$.*

Combining the results of Theorem 4 and Theorem 3, we obtain the following result.

Theorem 5. *Let α be a stable sequence of scaling factors. There exists a constant $C > 0$ that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$, such that $\sup_{L\geq 1} \sup_{l\in[L]} \mathbb{E}\,|q_{l,n}(a,b) - \tilde{q}_{l,\infty}(a,b)|^2 \leq Cn^{-1}$.*

We will see in the next section that the result of Theorem 5 is the cornerstone of commutativity; indeed, it suffices to study the infinite-depth limit of $\tilde{q}_{l,\infty}$ to obtain commutativity with explicit convergence rates for width and depth.

Appendix D. Infinite-Depth Limits

In this section, we study the infinite depth limit of the covariance kernel for different choices of the sequence α. We begin by proving a general commutativity result.

D.1 Infinite Depth Convergence of the Neural Covariance

Now that we have depth-uniform bounds for the kernels, we can look at what happens to the infinite-width kernels when depth increases.

Theorem 6 (General Theorem)**.** *Let $\alpha \in \mathcal{S}$. Assume that the kernel $\tilde{q}_{\lfloor tL \rfloor,\infty}(a,b)$ converges to some limiting kernel $q_t^{\infty}(a,b)$ in the limit $L \to \infty$ with some rate r_L. Then we have that*

$$\sup_{t\in[0,1]} \|q_{\lfloor tL \rfloor,n}(a,b) - q_t^{\infty}(a,b)\|_{L_2} \leq C\left(n^{-1/2} + r_L\right),$$

where C is a constant that depends only on $\|a\|.\|b\|, d, \|\alpha\|_S$.

Proof. We have that

$$\begin{aligned}\|q_{\lfloor tL \rfloor,n}(a,b) - q_t^{\infty}(a,b)\|_{L_2} &\leq \|q_{\lfloor tL \rfloor,n}(a,b) - \tilde{q}_{\lfloor tL \rfloor,\infty}(a,b)\|_{L_2} \\ &\quad + \|\tilde{q}_{\lfloor tL \rfloor,\infty}(a,b) - q_t^{\infty}(a,b)\|_{L_2} \\ &\leq C_1 n^{-1/2} + r_L,\end{aligned}$$

where we have used Theorem 5 to obtain the constant C_1 which depends only on $\|a\|.\|b\|, d, \|\alpha\|_S$. We conclude the proof by taking $C = \max(C_1, 1)$.

□

The result of Theorem 6 requires that the kernel $\tilde{q}_{\lfloor tL \rfloor,\infty}(a,b)$ converges in the infinite-depth limit with some rate r_L. In the following sections, we refine our analysis and study two scenarios where such convergence occurs.

D.2 Sequence of Scaling factors as 'Quasi-Convergent' Series.

Recall that for $L \geq 1, l \in [L]$

$$\tilde{q}_{l,\infty}(a,b) = \tilde{q}_{l-1,\infty}(a,b) + \alpha_{l,L}^2 (1/2q_{l-1}(a))^{1/2}(1/2q_{l-1}(b))^{1/2} f(c_{l-1}(a,b)),$$

where $c_{l-1,\infty}(a,b) := \frac{\tilde{q}_{l-1,\infty}(a,b)}{\tilde{q}_{l-1,\infty}(a,a)^{1/2}\tilde{q}_{l-1,\infty}(b,b)^{1/2}}$ is the infinite-width correlation kernel.

Given a sequence of scaling factors α, define $Q_l^{\alpha}(a,b) = \tilde{q}_{l,\infty}(a,b)$ with the scaling factors being $\alpha_{l,L}$.

To analyze the infinite-depth behavior of the kernel Q_l^{α}, it is crucial to understand the sensitivity of Q_l^{α} to α. The first result characterizes the sensitivity of the variance to a change in α.

Lemma 7. *Consider two stable sequences of scaling factors $\alpha, \beta \in \mathcal{S}$. Then, for all $L \geq 1, l \in \{1, \ldots, L\}$, we have that*

$$\sup_{l\in\{1,\ldots,L\}} |Q_l^{\alpha}(a,a) - Q_l^{\beta}(a,a)| \leq \frac{\|a\|^2}{2d} e^{\sup\{\|\alpha\|_S^2, \|\beta\|_S^2\}} \sum_{l=1}^{L} |\alpha_{l,L}^2 - \beta_{l,L}^2|$$

Proof. To alleviate the notation, we write $Q_l^\alpha = Q_l^\alpha(a,a)$. We have that

$$\begin{aligned}|Q_l^\alpha - Q_l^\beta| &\leq |Q_{l-1}^\alpha - Q_{l-1}^\beta| + \frac{1}{2}Q_{l-1}^\alpha|\alpha_{l,L}^2 - \beta_{l,L}^2| + \frac{1}{2}\beta_{l,L}^2|Q_{l-1}^\alpha - Q_{l-1}^\beta| \\ &\leq (1 + \frac{1}{2}\beta_{l,L}^2)|Q_{l-1}^\alpha - Q_{l-1}^\beta| + \frac{\|a\|^2}{2d}e^{\|\alpha\|_S^2/2}|\alpha_{l,L}^2 - \beta_{l,L}^2|\end{aligned}$$

The result follows immediately by induction.

□

Next, we prove a similar result for the kernel (not just the variance terms).

Lemma 8. *Consider two sequences of scaling factors $\alpha, \beta \in \mathcal{S}$. Then, for all $L \geq 1, l \in [L]$, we have that*

$$\sup_{l\in\{1,\dots,L\}} |Q_l^\alpha(a,b) - Q_l^\beta(a,b)| \leq C\sum_{l=1}^{L}|\alpha_{l,L}^2 - \beta_{l,L}^2|,$$

where C is a constant that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S, \|\beta\|_S$.

Proof. Let $I_l(\alpha) = (1/2Q_{l-1}^\alpha(a,a))^{1/2}(1/2Q_{l-1}^\alpha(b,b))^{1/2}$. We also use the notation c_l^α to denote the previously defined correlation kernel c_l. We have that

$$\begin{aligned}|Q_l^\alpha(a,b) - Q_l^\beta(a,b)| \leq |Q_{l-1}^\alpha(a,b) - Q_{l-1}^\beta(a,b)| + |\alpha_{l,L}^2 - \beta_{l,L}^2|I_l(\alpha) \\ + \beta_{l,L}^2|I_l(\alpha)f(c_{l-1}^\alpha(a,b)) - I_l(\beta)f(c_{l-1}^\beta(a,b))|\end{aligned} \tag{9}$$

Using Lemma 3, we have that $I_l(\alpha) \leq \frac{\|a\|\|b\|}{2d}e^{\|\alpha\|_S^2/2}$ and $I_l(\beta) \leq \frac{\|a\|\|b\|}{2d}e^{\|\beta\|_S^2/2}$. Moreover, using Lemma 7, we have that

$$\begin{aligned}|I_l(\alpha) - I_l(\beta)| &\leq (1/2Q_{l-1}^\alpha(a,a))^{1/2}|(1/2Q_{l-1}^\alpha(b,b))^{1/2} - 1/2Q_{l-1}^\beta(b,b))^{1/2}| \\ &\quad + (1/2Q_{l-1}^\beta(b,b))^{1/2}|(1/2Q_{l-1}^\alpha(a,a))^{1/2} - 1/2Q_{l-1}^\beta(a,a))^{1/2}| \\ &\leq C_1\sum_{l=1}^{L}|\alpha_{l,L}^2 - \beta_{l,L}^2|,\end{aligned}$$

where C_1 is a constant that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$, $\|\beta\|_S$, and where we have used the fact that $|\sqrt{x_1} - \sqrt{x_2}| \leq \frac{1}{2\sqrt{x_0}}|x_1 - x_2|$ for $x_1, x_2 > x_0 > 0$.

For the third term in the RHS of Eq. (9), we have that

$$\begin{aligned}|I_l(\alpha)f(c_{l-1}^\alpha(a,b)) - I_l(\beta)f(c_{l-1}^\beta(a,b))| &\leq |I_l(\alpha) - I_l(\beta)| + |f(c_{l-1}^\alpha(a,b)) - f(c_{l-1}^\beta(a,b))|I_l(\beta) \\ &\leq C_1\sum_{l=1}^{L}|\alpha_{l,L}^2 - \beta_{l,L}^2| + |c_{l-1}^\alpha(a,b) - c_{l-1}^\beta(a,b)|I_l(\beta).\end{aligned} \tag{10}$$

The second term on the RHS in the Eq. (10) can be upperbounded using the following

$$\begin{aligned}|c^{\alpha}_{l-1}(a,b) - c^{\beta}_{l-1}(a,b)| &= \left|(2I_{l-1}(\alpha))^{-1}Q^{\alpha}_{l-1}(a,b) - (2I_{l-1}(\beta))^{-1}Q^{\beta}_{l-1}(a,b)\right| \\ &\leq (2I_{l-1}(\alpha))^{-1}\left|Q^{\alpha}_{l-1}(a,b) - Q^{\beta}_{l-1}(a,b)\right| \\ &+ \left|(2I_{l-1}(\alpha))^{-1} - (2I_{l-1}(\beta))^{-1}\right| |Q^{\beta}_{l-1}(a,b)| \\ &\leq (2I_{l-1}(\alpha))^{-1}\left|Q^{\alpha}_{l-1}(a,b) - Q^{\beta}_{l-1}(a,b)\right| \\ &+ \frac{1}{2}(I_{l-1}(\alpha)I_{l-1}(\beta))^{-1}\,|I_{l-1}(\alpha) - I_{l-1}(\beta)|\,|Q^{\beta}_{l-1}(a,b)|\end{aligned}$$

Using the fact that $Q^{\beta}_{l-1}(a,b) \leq 2I_{l-1}(\beta)$, and (from Lemma 3) that $I_{l-1}(\alpha)^{-1}, I_{l-1}(\beta)^{-1} \leq 2d\,\|a\|^{-1}\|b\|^{-1}$, we conclude that there exist constants $\tilde{C}_1, C_2$ that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$, $\|\beta\|_S$, such that

$$|I_l(\alpha)f(c^{\alpha}_{l-1}(a,b)) - I_l(\beta)f(c^{\beta}_{l-1}(a,b))| \leq \tilde{C}_1 \sum_{l=1}^{L} |\alpha^2_{l,L} - \beta^2_{l,L}| + C_2|Q^{\alpha}_{l-1}(a,b) - Q^{\beta}_{l-1}(a,b)|.$$

As a result, we obtain

$$\begin{aligned}|Q^{\alpha}_{l}(a,b) - Q^{\beta}_{l}(a,b)| &\leq (1 + C_3\beta^2_{l,L})|Q^{\alpha}_{l-1}(a,b) - Q^{\beta}_{l-1}(a,b)| \\ &+ C_4|\alpha^2_{l,L} - \beta^2_{l,L}| + C_5\beta^2_{l,L}\sum_{l=1}^{L}|\alpha^2_{l,L} - \beta^2_{l,L}|,\end{aligned}$$

where C_3, C_4, C_5 are constants that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$, $\|\beta\|_S$.

Using the fact that $Q^{\alpha}_0 = Q^{\beta}_0$, there exists a constant C that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$, $\|\beta\|_S$, such that

$$|Q^{\alpha}_{l}(a,b) - Q^{\beta}_{l}(a,b)| \leq C\sum_{l=1}^{L}|\alpha^2_{l,L} - \beta^2_{l,L}|.$$

□

Let us now prove convergence in the case where α is the truncation (at level L) of a convergent series. The convergence is straightforward in this case and similar results have appeared in (Hayou et al., 2021).

Lemma 9. *Let $\alpha \in \mathcal{S}$ such that there exists a sequence $\zeta = (\zeta_i)_{i\geq 1} \in \ell_2(\mathbb{N})$ such that $\alpha_{l,L} = \zeta_l$ for all $L \geq 1, l \in [L]$. Then, there exists a limiting kernel Q^{α}_{∞} such that*

$$|Q^{\alpha}_{L}(a,b) - Q^{\alpha}_{\infty}(a,b)| \leq C\sum_{l\geq L}\zeta^2_l,$$

where C is a constant that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S = \sqrt{\sum_{l=1}^{\infty}\zeta^2_l}$.

Proof. Let $L' \geq L \geq 1$. It is straightforward that there exists a constant $C_1 > 0$ that depends only on $\|a\|, \|b\|, d, \|\alpha\|_S$ such that

$$|Q_L^\alpha(a,b) - Q_{L'}^\alpha(a,b)| \leq C_1 \sum_{L<l\leq L'} \zeta_l^2,$$

which shows that $(Q_L^\alpha)_{L\geq 1}$ is a Cauchy sequence and therefore it converges to some limit Q_∞^α. Taking L' to infinity provide the convergence rate.

□

Note that we only show the existence of the limit (and the convergence rate) in Lemma 9. Some properties of the limiting kernel in this case were studied in Hayou et al. (2021), these include continuity, universality etc.

Combining the results of Lemma 9 and Lemma 8, we conclude the following.

Lemma 10. *Let $\alpha \in \mathcal{S}$. Assume that there exists a sequence $\zeta = (\zeta_i)_{i\geq 1} \in \ell_2(\mathbb{N})$ such that $\sum_{l=1}^L |\alpha_{l,L}^2 - \zeta_l^2| \to 0$ as $L \to \infty$. Then, for all $t \in (0,1]$, we have*

$$|Q_{\lfloor tL\rfloor}^\alpha(a,b) - Q_\infty^\zeta(a,b)| \to 0,$$

where the convergence rate is given by $r_L = \Theta(\sum_{l=1}^{\lfloor tL\rfloor} |\alpha_{l,L}^2 - \zeta_l^2| + \sum_{l\geq \lfloor tL\rfloor} \zeta_l^2)$.

Combining Theorem 6 and Lemma 10, we conclude for Theorem 1.

D.3 Convergence with "Normalized" Sequences

For the specific choice of $\alpha_{l,L} = L^{-1/2}$, we know from (Hayou and Yang, 2023) that the covariance kernel $\tilde{q}_{\lfloor tL\rfloor,\infty}(a,b)$ converges to the solution of the following ODE

$$\frac{dq_t(a,b)}{dt} = \frac{e^{t/2}}{2}\zeta(a,b) f(\zeta(a,b)^{-1}e^{-t/2}q_t(a,b)) = F(t, q_t(a,b)), \tag{11}$$

where $\zeta(a,b) = d^{-1}\|a\|\|b\|$. The Euler scheme of Eq. (11) is given by

$$q_l^E(a,b) = q_{l-1}^E(a,b) + \alpha_{l,L}^2 F(t_{l-1}, q_{l-1}^E(a,b)), \quad q_0^E(a,b) = q_0(a,b),$$

where $t_l = \sum_{i=1}^l \alpha_{l,L}^2$.

For an ODE of the form $\dot{z}(t) = F(t, z(t))$, we call F the ODE functional. It is well known that under some conditions on this functional, the discretization error of the Euler scheme with steps $\delta_1, \dots, \delta_L$ is of order $\mathcal{O}(\max_{i\in[L]} \delta_i)$.

Theorem 7 (Corollary of Thm 212A in (Butcher, 2003))**.** *Consider an ODE of the form $\dot{z}(t) = F(t, z(t)), t \in [0,1]$, and consider the Euler discretization scheme with steps $\delta_1, \dots, \delta_L$ given by $z_l^E = z_{l-1}^E + \delta_l F(t_{l-1}, z_{l-1}^E)$ with the initial condition $z_0^E = z_0$, where $t_l = \sum_{i=1}^l \delta_i$. Assume that the following hold*

- *There exists a constant* $L > 0$ *such that* $|F(t,z) - F(t,z')| \leq |z - z'|$ *for all* $t \in [0,1], z, z' \in \mathbb{R}$.
- $M = \sup_{t\in[0,1]} \left| \frac{d^2 z(t)}{dt^2} \right| < \infty$.

Then, we have

$$\sup_{l\in[L]} |z_l^E - z_{t_l}| \leq C \max_{i\in[L]} \delta_i,$$

where C *depends only on* M *and* L.

In the next result, we use this result to show convergence rate of the Euler scheme presented above.

Lemma 11. *Consider a normalized sequence of scaling fatcors* α *and let* $h_L = \max_{1\leq l\leq L} \alpha_{l,L}^2$. *We have*

$$\sup_{1\leq l\leq L} |q_l^E(a,b) - q_{t_l}(a,b)| \leq C\, h_L,$$

where C *is a constant that depends only on* $\|a\|, \|b\|, d$.

Proof. Let us verify the conditions of Theorem 7 one by one.

- Lipschitz property: from Lemma 14, we know that the function f is Lipschitz. Therefore, we have $|F(t,z) - F(t,z')| \leq \frac{1}{2} e^{t/2} |\zeta(a,b)| |e^{-t/2} \zeta(a,b)^{-1} (z - z')| = |z - z'|$.
- For $t \in [0,1]$, we have

$$\begin{aligned}\frac{d^2 q_t(a,b)}{dt^2} =& \frac{t e^{t/2}}{4} \zeta(a,b) f(\zeta(a,b)^{-1} e^{-t/2} q_t(a,b)) + \frac{e^{t/2}}{2} \zeta(a,b) (-\zeta(a,b)^{-1} \frac{t e^{t/2}}{2} q_t(a,b) \\ &+ \zeta(a,b)^{-1} e^{-t/2} \frac{dq_t(a,b)}{dt}) f'(\zeta(a,b)^{-1} e^{-t/2} q_t(a,b)),\end{aligned}$$

  Replacing $\frac{dq_t(a,b)}{dt}$ by its valye, it is straightforward that $M = \sup_{t\in[0,1]} \left| \frac{d^2 q_t(a,b)}{dt^2} \right|$ is finite and depends only on $\|a\|, \|b\|, d$.

This concludes the proof.

□

Now it remains to bound the difference between $q_l^E(a,b)$ and $\tilde{q}_{l,\infty}(a,b)$, the covariance kernel of the auxiliary process. We deal with this in the next result.

Lemma 12. *Consider a normalized sequence of scaling factors* α. *Let* $h_L = \max_{1\leq l\leq L} \alpha_{l,L}^2$ *and assume that* $L h_L^2 = o(1)$. *Then, we have*

$$\sup_{1\leq l\leq L} |\tilde{q}_{l,\infty}(a,b) - q_l^E(a,b)| \leq C L h_L^2,$$

where C *is a constant that depends only on* $\|a\|, \|b\|, d$.

Proof. Assume that $L \times h_L^2 = o(1)$. We write $\zeta = \zeta(a,b)$ to simplify the notation. We have

$$\begin{aligned}|\tilde{q}_{l,\infty}(a,b) - q_l^E(a,b)| &\leq |\tilde{q}_{l-1,\infty}(a,b) - q_{l-1}^E(a,b)| \\ &+ \frac{1}{2}\alpha_{l,L}^2\zeta f(\zeta^{-1}\prod_{i=1}^{l-1}(1+\frac{\alpha_{l,L}^2}{2})^{-1}\tilde{q}_{l-1,\infty}(a,b))|\prod_{i=1}^{l-1}(1+\frac{\alpha_{l,L}^2}{2}) - e^{\frac{1}{2}\sum_{i=1}^{l-1}\alpha_{l,L}^2}| \\ &+ \frac{1}{2}\alpha_{l,L}^2\zeta e^{\frac{1}{2}\sum_{i=1}^{l-1}\alpha_{l,L}^2}|f(\zeta^{-1}\prod_{i=1}^{l-1}(1+\frac{\alpha_{l,L}^2}{2})^{-1}\tilde{q}_{l-1,\infty}(a,b)) - f(\zeta^{-1}e^{-\frac{1}{2}\sum_{i=1}^{l-1}\alpha_{l,L}^2}q_{l-1}^E(a,b))|\end{aligned} \tag{12}$$

Notice that

$$\begin{aligned}\prod_{i=1}^{l-1}(1+\frac{\alpha_{l,L}^2}{2}) &= \prod_{i=1}^{l-1}(e^{\frac{1}{2}\alpha_{l,L}^2} + \mathcal{O}(\alpha_{l,L}^4)) = \prod_{i=1}^{l-1}e^{\frac{1}{2}\alpha_{l,L}^2}(1+\mathcal{O}(\alpha_{l,L}^4)) \\ &= e^{\frac{1}{2}\sum_{i=1}^{l-1}\alpha_{l,L}^2}(1+\mathcal{O}(h_L^2))^{l-1} = e^{\frac{1}{2}\sum_{i=1}^{l-1}\alpha_{l,L}^2} + \mathcal{O}(Lh_L^2),\end{aligned}$$

where the constant in "$\mathcal{O}$" is universal. As a result, there exists a constant C_1 that depends only on $\|a\|, \|b\|, d$, such that the second term in the RHS of Eq. (12) is smaller than $C \times Lh_L^2$.

We also have

$$\prod_{i=1}^{l-1}(1+\frac{\alpha_{l,L}^2}{2})^{-1} = \prod_{i=1}^{l-1}e^{\frac{1}{2}\alpha_{l,L}^2} + \mathcal{O}(Lh_L^2),$$

where the constant in "$\mathcal{O}$" is universal. Using the Lipschitz property of f (Lemma 14), we obtain that

$$\begin{aligned}|f(\zeta^{-1}\prod_{i=1}^{l-1}(1+\frac{\alpha_{l,L}^2}{2})^{-1}\tilde{q}_{l-1,\infty}(a,b)) - f(\zeta^{-1}e^{-\frac{1}{2}\sum_{i=1}^{l-1}\alpha_{l,L}^2}q_{l-1}^E(a,b))| \\ \leq \zeta^{-1}|\prod_{i=1}^{l-1}(1+\frac{\alpha_{l,L}^2}{2})^{-1} - e^{-\frac{1}{2}\sum_{i=1}^{l-1}\alpha_{l,L}^2}|\tilde{q}_{l-1,\infty}(a,b) \\ + \zeta^{-1}e^{-\frac{1}{2}\sum_{i=1}^{l-1}\alpha_{l,L}^2}|\tilde{q}_{l-1,\infty}(a,b) - q_{l-1}^E(a,b)|,\end{aligned}$$

This yield,

$$|\tilde{q}_{l,\infty}(a,b) - q_l^E(a,b)| \leq (1+C_2\alpha_{l,L}^2)|\tilde{q}_{l-1,\infty}(a,b) - q_{l-1}^E(a,b)| + C_3Lh_L^2,$$

where C_2, C_3 are constants that depend only on $\|a\|, \|b\|, d$. An induction argument allows us to conclude.

□

Combining the results of Lemma 11 and Lemma 12, we obtain the following result.

Theorem 8. *Consider a sequence of scaling factors α such that $\sum_{l=1}^{L}\alpha_{l,L}^2 = 1$. Let $h_L = \max_{1\leq l\leq L}\alpha_{l,L}^2$ and assume that $Lh_L^2 = o(1)$. Then, we have that*

$$\sup_{1\leq l\leq L}|\tilde{q}_{l,\infty}(a,b) - q_{t_l}(a,b)| \leq C(h_L + Lh_L^2)$$

By combining the results of Theorem 6 and Theorem 8, we obtain the first part of Theorem 2. It remains to show the second part of the theorem when $\sup_{t\in[0,1]} \left|\sum_{k=1}^{\lfloor tL\rfloor} \alpha_{k,L}^2 - \lambda(t)\right| \leq r_L$ and $\lim_{L\to\infty} r_L = 0$. Assume that this condition holds. From the ODE Eq. (11), it is straightforward that $|q_t(a,b) - q_{t'}(a,b)| \leq C_1|t-t'|$ holds for all $t, t' \in [0,1]$ for some constant $C_1 > 0$ that depends only on $\zeta(a,b)$.
Let $t \in [0,1]$ and $t_L = \sum_{k=1}^{\lfloor tL\rfloor} \alpha_{k,L}^2$. As a result of the inequality above, we have $|q_{t_L}(a,b) - q_{\lambda(t)}| \leq C_1|t_L - \lambda(t)| \leq C_1 r_L$. We conclude using the first part of the theorem and the triangular inequality.

Appendix E. Other Technical Results

E.1 Lemma for the Auxiliary process

We use the next lemma to prove that the Auxiliary process has *iid* coordinates. This is a trivial result, but we include the proof for better readability.

Lemma 13. *Let $W \in \mathbb{R}^{n\times n}$ be a matrix of standard Gaussian random variables $W_{ij} \sim \mathcal{N}(0,1)$. Let $v \in \mathbb{R}^n$ be a random vector independent from W and satisfies $\|v\|_2 = 1$. Then, $Wv \sim \mathcal{N}(0, I)$.*

Proof. The proof follows a simple characteristic function argument. Indeed, by conditioning on v, we observe that $Wv \sim \mathcal{N}(0, I)$. Let $u \in \mathbb{R}^n$, we have that

$$\begin{aligned}\mathbb{E}_{W,v}[e^{i\langle u, Wv\rangle}] &= \mathbb{E}_v[\mathbb{E}_W[e^{i\langle u, Wv\rangle}|v]] \\ &= \mathbb{E}_v[e^{-\frac{\|u\|^2}{2}}] \\ &= e^{-\frac{\|u\|^2}{2}}.\end{aligned}$$

This concludes the proof as the latter is the characteristic function of a random Gaussian vector with Identity covariance matrix. □

E.2 Lemma for the (correlation) function f

Lemma 14 (Function f)**.** *Let $f : [-1,1] \to [-1,1]$ be the function defined by $f(c) := 2\mathbb{E}[\phi(Z_1)\phi(cZ_1 + \sqrt{1-c^2}Z_2)]$. Then, we have*

$$f(c) = \frac{1}{\pi}(c \arcsin c + \sqrt{1-c^2}) + \frac{1}{2}c.$$

Thus, f is 1-Lipschitz.

Proof. The closed-form expression of f has appeared in a series of papers under different forms (Cho and Saul, 2009; Hayou et al., 2019, 2021). Here, we only show the Lipschitz property, which is straightforward. From the closed-form expression of f, we obtain

$$f'(c) = \pi^{-1} \arcsin(c) + \frac{1}{2},$$

which shows that $|f'| \leq 1$ and concludes the proof. □