Width and Depth Limits Commute in Residual Networks

Soufiane Hayou [1] **Greg Yang** [2]

Abstract

We show that taking the width and depth to infinity in a deep neural network with skip connections, when branches are scaled by $1/\sqrt{depth}$, result in the same covariance structure no matter how that limit is taken. This explains why the standard infinite-width-then-depth approach provides practical insights even for networks with depth of the same order as width. We also demonstrate that the pre-activations, in this case, have Gaussian distributions which has direct applications in Bayesian deep learning. We conduct extensive simulations that show an excellent match with our theoretical findings.

1. Introduction

In recent years, deep neural networks have achieved remarkable success in a variety of tasks, such as image classification and natural language processing. However, the behavior of these networks in the limit of large depth and large width is still not fully understood.

The success of large language and vision models have recently amplified an existing trend of research on neural network limits. Two main limits are the large-width and the large-depth limits. While the former by itself is now relatively well understood (Neal, 1995; Schoenholz et al., 2017; Lee et al., 2018; Hayou, Doucet, et al., 2019a; Yang, 2020a), the latter and the interaction between the two have not been studied as much. In particular, a basic question is: do these two limits commute? Recent literature suggests that, at initialization, in certain kinds of multi-layer perceptrons (MLPs) or residual neural networks (resnets), the depth and width limits do not commute; this would imply that in practice, such kinds of networks would behave quite differently depending on whether width is much larger than depth or the other way around.

However, in this paper, we show: to the contrary, at initialization, for a resnet with branches scaled the natural way so as to avoid blowing up the output,[1] the width and depth limits *do commute*. This 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 (Hayou, Clerico, et al., 2021).

In addition to the significance of the results, the mathematical novelty of this paper is the proof technique: we take the depth limit first (fixing width), then take the width limit, in contrast to the typical prior work which takes the limits in the opposite order. In the process, we prove a concentration of measure result for a kind of McKean-Vlasov process (Mean-Field games). Our results provide new insights into the behavior of deep neural networks and we discuss implications for the design and analysis of these networks.

The proofs of the theoretical results are provided in the appendix and referenced after each result. Empirical evaluations support our theoretical findings.

2. Related Work

The theoretical analysis of randomly initialized neural networks with an infinite number of parameters has yielded a wealth 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 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 in this area, though it's not exhaustive. A more comprehensive literature review is provided in Appendix A.

2.1. 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

[1] Department of Mathematics, National University of Singapore [2] Microsoft Research AI. Correspondence to: Soufiane Hayou <hayou@nus.edu.sg>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] This contrasts with M. B. Li et al., 2022 whose non-commute result requires the branches to be large enough to blow up the network output in the case of standard resnet.

and Schoenholz, 2017; Hayou, Doucet, et al., 2019a), and the selection of activation functions (Hayou, Doucet, et al., 2019a; Martens et al., 2021; Wolinski et al., 2022; Zhang et al., 2022), 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; Matthews et al., 2018; Hron et al., 2020; Yang, 2020a) 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, Clerico, et al., 2021). [2]

The Neural Tangent Kernel (NTK) is another interesting area of research where the infinite-width limit proves useful. In this limit, the NTK converges to a deterministic kernel, given appropriate parameterization. This limiting kernel is fixed at initialization and remains constant throughout the training process. The optimization and generalization characteristics of the NTK have been the subject of extensive study in the literature (see e.g. Arora et al., 2019; Liu et al., 2022).

2.2. 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 research in this field can be categorized into three groups based on the approach and criteria used to consider the infinite-depth limit in relation to the width.

Infinite-width-then-depth limit. In this case, the width of the neural network is taken to infinity first, followed by the depth. This is the infinite-depth limit of infinite-width neural networks. 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 (Poole et al., 2016; Schoenholz et al., 2017; Yang and Schoenholz, 2017)), evaluating the impact of the activation function (Hayou, Doucet, et al., 2019a; Martens et al., 2021), and studying the behavior of the Neural Tangent Kernel (NTK) (Hayou, Doucet, et al., 2020; Xiao et al., 2020).

The joint infinite-width-and-depth limit. In this case, the ratio of depth to width is fixed, and the width and depth are jointly taken to infinity. There are only a limited number of works that have investigated the joint width-depth limit. In (M. 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 the depth, in which case the distribution of the network output is asymptotically normal ((Schoenholz et al., 2017; Hayou, Doucet, et al., 2019a)). Additionally, in (M. B. Li et al., 2022), the authors examined the covariance kernel 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 (Hanin and Nica, 2019; Noci et al., 2021; Zavatone-Veth et al., 2021; Hanin, 2022).

Infinite-depth limit of finite-width neural networks. In the previous limits, the width of the neural network was extended to infinity, either independently or in conjunction with the depth. However, it is natural to inquire about the behavior of networks in which the width is fixed, while the depth is increased towards infinity. In Peluchetti et al., 2020, it was shown that for a particular ResNet architecture, the pre-activations converge weakly to a diffusion process in the infinite-depth limit, which follows from existing results in stochastic calculus on the convergence of Euler-Maruyama discretization schemes to continuous Stochastic Differential Equations. More recent work by Hayou, 2022 evaluated 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.

In this work, we are particularly interested in the case where both the width and depth are taken to infinity.

3. Setup and Definitions

When analyzing the asymptotic behavior of randomly initialized neural networks, various notions of probabilistic convergence are employed, depending on the context. These notions are typically well-established definitions in probability theory. In this study, we particularly focus on two forms of convergence:

- Convergence in distribution (weak convergence): we show that the pre-activations converge weakly to a Gaussian distribution in the limit $\min(n, L) \to \infty$. We use the Wasserstein metric to quantify the convergence rate for the weak convergence.
- Convergence in L_2 (strong convergence): we show that the neural covariance[3] converges to a deterministic limit that is characterized by a differential flow q_t as $\min(n, L)$ approaches infinity.

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

[3] The neural covariance is a (linear) measure of similarity between the pre-activations for different inputs. We define this quantity in Section 4.

Definition 1 (Weak convergence)**.** *Let $d \geq 1$. We say that a sequence of $\mathbb{R}^d$-valued random variables $(X_k)_{k\geq 1}$ converges weakly to a random variable Z if the cumulative distribution function of X_k converges point-wise to that of Z.*

There are various metrics that can be utilized to measure the weak convergence rate. One commonly used metric is the Wasserstein metric.

Definition 2 (Wasserstein distance $\mathcal{W}_1$)**.** *Let μ and ν be two probability measures on $\mathbb{R}^d$. The Wasserstein distance between μ and ν is defined by*

$$\begin{aligned}\mathcal{W}_1 &= \sup_{f \in \text{Lip}_1} \left| \int f(x)(d\mu - d\nu) \right| \\ &= \sup_{f \in \text{Lip}_1} |\mathbb{E}_\mu f - \mathbb{E}_\nu f|,\end{aligned}$$

where Lip_1 *is the set of Lipschitz continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$ with a Lipschitz constant ≤ 1.*

In this work, we define *strong* convergence to be the L_2 convergence as described in the following definition.

Definition 3 (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 $\|X\|_{L_2} = \left(\mathbb{E}[\|X\|^2]\right)^{1/2}$.*

Both of these forms of convergence are valuable when analyzing the behavior of neural networks with an infinite number of parameters. They facilitate the understanding of the network's asymptotic behavior which enables predictions about the finite-but-large width-and-depth regimes.

4. Warmup: Depth and Width Generally Do Not Commute

In this section, we present corollaries of previously established results that demonstrate that depth and width typically do not commute. The width and depth of the network are denoted by n and L, respectively, and the input dimension is denoted by d. Let $d, n, L \geq 1$, and consider a simple MLP architecture given by the following:

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

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 $W_l^{ij} \sim \mathcal{N}(0, \frac{2}{n})$,[4] $W_{in}^{ij} \sim \mathcal{N}(0, \frac{1}{d})$. 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. While the activation function is only defined for real numbers (1-dimensional), we will 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 $\{Y_l, l = 0, \dots, L\}$ as *pre-activations* and the vectors $\{\phi(Y_l), l = 0, \dots, L\}$ as *post-activations*.

4.1. Distribution of the pre-activations in the limit $n, L \to \infty$

It is well-established that in fixed-depth neural networks of any type, as the width n approaches infinity, the pre-activations exhibit Gaussian behavior. This phenomenon was initially demonstrated for single-layer perceptrons by (Neal, 1995), and has since been extended to include multiple-layer perceptrons (MLPs) and general neural architectures (Yang, 2020a). This behavior can be roughly attributed to the Central Limit Theorem (CLT) (although a formal proof require careful application of CLT for exchangeable random variables in the MLP case, as detailed in Matthews et al., 2018, or Law of Large Numbers and Gaussian conditioning trick in the general case (Yang, 2019b)). A question that the reader may have in this context is: *Why is the Gaussian distribution of significance?* One of the key implications of the Gaussian behavior of infinite-width neural networks is their equivalence to Gaussian processes. By utilizing existing methods of Gaussian process regression, this equivalence facilitates the application of exact Bayesian inference to infinite-width neural networks, referred to as the neural network Gaussian process (NNGP, Lee et al., 2018). The Gaussian behavior also provides an interesting framework to study signal propagation in deep neural networks; since a Gaussian distribution is fully characterized by its mean and covariance structure, understanding these quantities is sufficient to capture what happens inside the network at initialization.

When the depth L is also taken to infinity, different behaviors may emerge. Specifically, in the case of the MLP architecture (1), if a fixed layer index $l < L$ is considered and the behavior of Y_l is examined as n and L approach infinity, Y_l will exhibit the same limiting behavior as in the case of $n \to \infty$ and the depth is fixed. Some simple intuitive calculations indicate that it is only meaningful to study the limiting behavior of layers where the layer index is proportional to the depth L (and not proportional to L^α for any $\alpha < 1$).[5] In this case, the quantity of interest is $Y_{\lfloor tL \rfloor}$ for

[4] 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.

[5] Indeed, the $(\frac{1}{n})$-scaled Gram matrix of $\{Y_l(a) : a \in \mathbb{R}^d\}$ fluctuates with size $\tilde{\Theta}(1/\sqrt{n})$ around its $n \to \infty$ limit for any fixed l. This fluctuation is asymptotically independent across every layer, so the accumulated fluctuation at layer $l = L^\alpha$ is

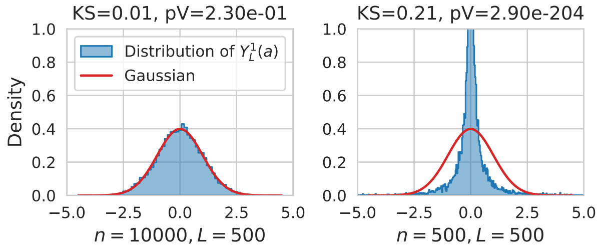


Figure 1: Histogram of $Y_L^1(a)$ for an MLP Eq. (1) with $(n, L) \in \{(10000, 500), (500, 500)\}$, $d = 30$, and $a = \sqrt{d}\frac{u}{\|u\|}$ and $u \in \mathbb{R}^d$ has all coordinates randomly sampled from the unifrom distribution $\mathcal{U}([0, 1])$. The histogram is based on $N = 10^4$ simulations. The red dashed line represents the theoretical distribution (Gaussian) predicted in Proposition 1. We also perform a Kolmogorov-Smirnov normality test and report the KS statistic and the p-value.

some $t \in [0, 1]$. Varying t between 0 and 1 encompasses all layer indices, even in the infinite-depth limit.

Let us now state some corollaries of existing results. The following is a trivial result from existing literature (see e.g. Matthews et al., 2018) that characterizes the distribution of the pre-activations in the limit $n \to \infty$ then $L \to \infty$.

Proposition 1 (Infinite-width-then-depth)**.** *Consider the MLP architecture given by Eq.* (1) *and let $a \in \mathbb{R}^d$ such that $a \neq 0$. Then, in the limit "$n \to \infty$, then $L \to \infty$", $Y_L^1(a)$*[6] *converges weakly to a Gaussian distribution.*

When the width and depth of a neural network both tend towards infinity, the limiting behavior can vary depending on the relative rates at which the width and depth increase. Specifically, if the width and depth both approach infinity while the ratio of width to depth remains constant, the distribution of the pre-activations in the last layer is not Gaussian. This is a corollary of a more general result established by (M. Li et al., 2021) (the case when $\alpha = 0$) under certain conditions and assumptions, which was also verified through empirical evidence. We omit here the rigorous statement of the result and only illustrate this behaviour with simulations.

Empirical evidence supports the existence of this difference in the limiting behavior of the distribution. As shown in Fig. 1, the distribution of $Y_L^1(a)$ is observed to be (nearly) Gaussian when the width is significantly greater than the depth, as evidenced by a small KS statistic. However, when the width is of the same magnitude as the depth, the distribution exhibits heavy tails. This can be seen by comparing the distribution for the settings $(n, L) \in (10000, 500), (500, 500)$.

4.2. Neural covariance/correlation

In the literature on signal propagation, there is a significant interest in understanding the covariance/correlation structure of neural networks. Specifically, researchers have sought to understand the covariance of the pre-activation vectors $Y_{\lfloor tL \rfloor}(a)$ and $Y_{\lfloor tL \rfloor}(b)$ (often called the neural covariance) for two different inputs $a, b \in \mathbb{R}^d$. A natural question in this context is: *Why do we study the covariance structure?*

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* $L \to \infty$" (Poole et al., 2016; Schoenholz et al., 2017; Hayou, Doucet, et al., 2019a; Yang and Salman, 2019). 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[7]. 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, resulting in a correlation kernel that does not converge to 1 (Hayou, Clerico, et al., 2021). This analysis was carried in the limit "$n \to \infty$, then, $L \to \infty$". 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 one in this limit.

Proposition 2 (Correlation, (Hayou, Doucet, et al., 2019a; M. B. Li et al., 2022))**.** *Consider the MLP architecture given by Eq.* (1) *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)), ren-

$\tilde{\Theta}(L^{\alpha/2}/\sqrt{n})$. This is $\tilde{\Theta}(1)$ iff $\alpha = 1$.

[6] $Y_L^1(a)$ refers to the first neuron in the last layer.

[7] 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.

[8] Note that weak convergence to a constant implies also convergence in probability.

dering 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.

Proposition 3 (Correlation SDE, Corollary of Thm 3.2 in M. B. Li et al., 2022)**.** *Consider the MLP architecture given by Eq. (1) 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 behaviour of the covariance structure. It is easy to check that even with shaped ReLU as in Proposition 3, taking the width to infinity first, then depth, the result is a deterministic covariance structure. The main takeaway from this section is the following:

Summary. *With MLPs (Eq. (1)), the width and depth limits do not commute in the sense that the behaviour of the distribution of the pre-activations and the covariance structure might differ depending on how the limit is taken.*

With the background information provided above, we are now able to present our findings. In contrast to MLPs, our next section demonstrates that the limits of width and depth for ResNet architectures commute.

5. Main results: Width and Depth Commute in ResNets

We use the same notation as in the MLP case. Let $d, n, L \geq 1$, and consider the following ResNet architecture of width n and depth L

$$
\begin{aligned}
&Y_0(a) = W_{in} a, \quad a \in \mathbb{R}^d \\
&Y_l(a) = Y_{l-1}(a) + \frac{1}{\sqrt{L}} W_l \phi(Y_{l-1}(a)), \; l \in [1:L],
\end{aligned}
\tag{2}
$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the ReLU activation function. We assume that the weights are randomly initialized with *iid* Gaussian variables $W_l^{ij} \sim \mathcal{N}(0, \frac{1}{n})$, $W_{in}^{ij} \sim \mathcal{N}(0, \frac{1}{d})$. 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. The $1/\sqrt{L}$ scaling in Eq. (2)is not chosen arbitrarily. It has been demonstrated that this specific scaling serves to stabilize the norm of Y_l and the gradient norms in the asymptotic limit of large depth (e.g. Hayou, Clerico, et al., 2021; Hayou, 2022; Marion et al., 2022).[10]

5.1. Distribution of the pre-activations in the limit $n, L \to \infty$

It turns out that for the ResNet architecture given by (2), the limiting distribution of the pre-activations $Y_{\lfloor tL \rfloor}$ is a zero-mean Gaussian distribution, with an analytic variance term, regardless of how the depth L and width n approach infinity, as long as $\min(n, L) \to \infty$. This is demonstrated in the following result, where an upper bound on the Wasserstein distance between the distribution of the neuron $Y^1_{\lfloor tL \rfloor}$ (the first coordinate of the pre-activations $Y_{\lfloor tL \rfloor}$)[11] and that of a zero-mean Gaussian random variable is provided.

Theorem 1 (Convergence of the pre-activations)**.** *Let $a \in \mathbb{R}^d$ such that $a \neq 0$. For $t \in [0,1]$, the random variable $(Y_{\lfloor tL \rfloor}(a))_{L \geq 1}$ converges weakly to a Gaussian random variable with law $\mathcal{N}(0, v(t,a))$ in the limit of $\min(n, L) \to \infty$, where $v(t, a) = d^{-1}\|a\|^2 \exp(t/2)$. Moreover, we have the following convergence rate*

$$
\sup_{t \in [0,1]} \mathcal{W}_1(\mu^t_{n,L}(a), \mu^t_{\infty,\infty}(a)) \leq C \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{L}} \right)
$$

where $\mu^t_{n,L}(a)$ is the distribution of $Y^1_{\lfloor tL \rfloor}(a)$, $\mu^t_{\infty,\infty}(a)$ is the distribution $\mathcal{N}(0, v(t,a))$, and C is a constant that depends only on $\|a\|$ and d.

Moreover, for two different $i, j \in [n]$, the neurons $Y^i_{\lfloor tL \rfloor}(a)$ and $Y^j_{\lfloor tL \rfloor}(a)$ become independent in the limit $\min(n, L) \to \infty$.

The proof of Theorem 1 is provided in Appendix D. It relies on two technical results: 1) Width-uniform convergence rate of the finite-width neural networks to an infinite-depth SDE.[12] 2) A new result on the convergence of particles to a mean field process. Both results are new. More details are provided in the Appendix.

Theorem 1 suggests that the distribution of the pre-activations becomes similar to a Gaussian distribution as $\min(n, L) \to \infty$ regardless of how n and L go to infinity. Note that the limiting distribution is the same as the one reported in (Hayou, 2022) where the author considered

[9] In M. B. Li et al., 2022, 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.

[10] A scaling of the form $L^{-\alpha}$ where $\alpha < 1/2$ yields exploding pre-activations, while a more aggressive scaling where $\alpha > 1/2$ yields trivial limiting covariance (identity covariance).

[11] Notice that the coordinate of the pre-activations are identically distributed (but not necessarily independent).

[12] By width-uniform, we refer to bounds with constants that do not depend on the width n.

the limit "$n \to \infty$, *then* $L \to \infty$". Our result generalizes these findings and establishes the universality of the Gaussian behaviour as long as $n \to \infty$ and $L \to \infty$. We validate these theoretical predictions in Section 6. An important consequence of the Gaussian behaviour is that the residual network can be seen as a Gaussian process in this limit with a well-specified kernel function (see next section). Leveraging this result to perform Bayesian inference with infinite-width-and-depth networks can be an interesting direction for future work.

5.2. Neural covariance

Unlike the covariance structure in MLPs which exhibits different limiting behaviors depending on how the width and depth limits are taken, we show in the next result that for the ResNet architecture given by (2), the neural covariance converges strongly to a deterministic kernel, which is given by the solution of a differential flow, in the limit $\min(n, L) \to \infty$ regardless of the relative rate at which n and L tend to infinity.

Theorem 2 (Neural covariance)**.** *Let $a, b \in \mathbb{R}^d$ such that $a, b \neq 0$ and $a \neq b$. Define the neural covariance kernel $\hat{q}_t(a,b) = \frac{\langle Y_{\lfloor tL \rfloor}(a), Y_{\lfloor tL \rfloor}(b)\rangle}{n}$. Then, we have the following*

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

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

$$\begin{cases} \frac{dq_t(a,b)}{dt} &= \frac{1}{2}\frac{f(c_t(a,b))}{c_t(a,b)} q_t(a,b), \\ c_t(a,b) &= \frac{q_t(a,b)}{\sqrt{q_t(a,a)}\sqrt{q_t(b,b)}}, \\ q_0(a,b) &= \frac{\langle a,b\rangle}{d}, \end{cases} \tag{3}$$

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

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

The proof of Theorem 2 is provided in Appendix E. The result of Theorem 2 unifies previous approaches to understanding the covariance structure in large width and depth ResNets. Perhaps the most important consequence of our result is that it implies that all previous results that considered the limit $n \to \infty$, then $L \to \infty$, in order to understand the covariance structure in ResNets still hold for ResNets where the depth is of the same order as the width and both are large. This is specific for ResNet and does not hold for instance for MLPs where the joint-limit yields different asymptotic behaviors (see Section 4). Notice that the limiting covariance kernel q_t is the same kernel found in (Hayou, Clerico, et al., 2021) in the limit $n \to \infty$, then $L \to \infty$.[13] It is also worth noting that constant C can be chosen independent of $\|a\|$ and $\|b\|$ provided that the inputs belong to a compact set that does not contain 0. The result of Theorem 2 can also be expressed in terms of the correlation. We demonstrate this in the next theorem.

Theorem 3 (Neural correlation)**.** *Under the same conditions of Theorem 2, we have the following*

$$\sup_{t\in[0,1]} \|\hat{c}_t(a,b) - c_t(a,b)\|_{L_2} \leq C'\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{L}}\right)$$

where C' is a constant that depends only on $\|a\|$, $\|b\|$, and d, and $\hat{c}_t(a,b) = \frac{\langle Y_{\lfloor tL \rfloor}(a), Y_{\lfloor tL \rfloor}(b)\rangle}{\|Y_{\lfloor tL \rfloor}(a)\|\|Y_{\lfloor tL \rfloor}(b)\|}$ is the neural correlation kernel, and $c_t(a,b)$ is defined in Theorem 2.

The proof of Theorem 3 relies on using a concentration inequality to control the inverse variance term, and conclude by using the bound in Theorem 2. We refer the reader to the Appendix for more details.

The differential flow satisfied by the kernel function q_t can actually be simplified and expressed as an ordinary differential equation (ODE). We show this in the next lemma.

Lemma 1. *Let $z = (a,b) \in \mathbb{R}^d \times \mathbb{R}^d$. The function q_t in Theorem 2 is the solution of the following ODE:*

$$\frac{dq_t(z)}{dt} = \frac{\exp(t/2)}{2}\xi(z) f\left(\xi(z)^{-1}\exp(-t/2)q_t(z)\right),$$

where $\xi(z) = \frac{\|a\|\,\|b\|}{d}$, and f is defined in Theorem 2.

Proof. The proof is straightforward by noticing that $f(1) = 1$. With this we get $\frac{dq_t(a,a)}{dt} = \frac{1}{2}q_t(a,a)$ which yields $q_t(a,a) = q_0(a,a)\exp(t/2) = d^{-1}\|a\|^2\exp(t/2)$. The same holds for b, which concludes the proof. □

Lemma 1 will prove useful in the experiments section when we will have to approximate the solution q_t using ODE solvers.

6. Experiments and Practical Implications

In this section, we validate our theoretical results with extensive simulations on large width and depth residual neural networks of the form Eq. (2).

6.1. Gaussian behavior and independence of neurons

Theorem 1 predicts that in the large depth and width limit, the neurons (pre-activations) converge weakly to a Gaussian distribution. To empirically validate this finding, we

[13] In Hayou, Clerico, et al., 2021, the authors showed that the kernel q_t is universal, meaning the network output is rich enough that we can approximate any continuous function on a compact set with features from this kernel.

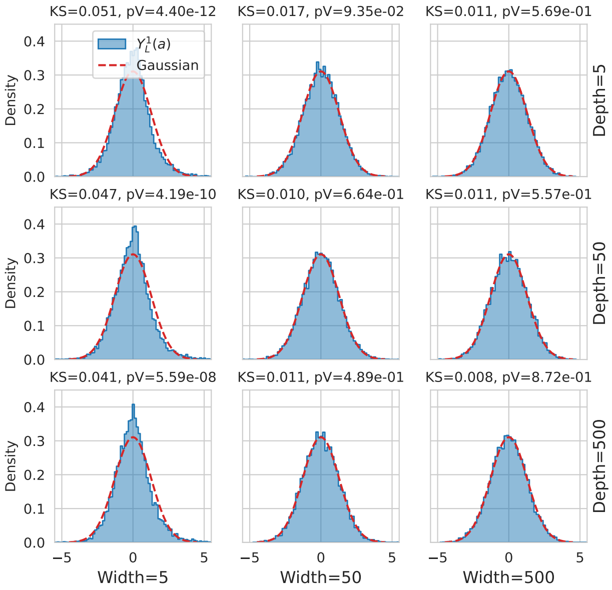


Figure 2: Histogram of $Y_L^1(a)$ for ResNet Eq. (2) with $n, L \in \{5, 50, 500\}$, $d = 30$, and $a = \sqrt{d}\frac{u}{\|u\|}$ and $u \in \mathbb{R}^d$ has all coordinates randomly sampled from the unfirom distribution $\mathcal{U}([0, 1])$. The histogram is based on $N = 10^4$ simulations. The red dashed line represents the theoretical distribution (Gaussian) predicted in Theorem 1. We also peform a Kolmogorov-Smirnov normality test and report the KS statistic and the p-value.

show in Fig. 2 the histograms of the first neuron in the last layer ($t = 1$ in Theorem 1) for a randomly chosen input a and $n, L \in \{5, 50, 500\}$. We also perform a Kolmogorov-Smirnov normality test and report the statistic (KS) and the p-value. As can be seen in Fig. 2, the histograms appear to fit the theoretical Gaussian distribution more closely as width and depth increase. Additionally, the KS statistic decreases as the width and depth increase. For smaller widths, the p-values are extremely small indicating a non-Gaussian behavior. This is expected as the Gaussian behavior arises primarily due to the average behavior when the width increases. The depth also plays a role in the goodness of fit, as can be seen for the pair $(n, L) = (500, 50)$ and $(n, L) = (500, 500)$ where the latter shows a better fit in terms of the KS statistic which measures the distance between the empirical cumulative distribution function and the theoretical one. Notice also the contrast with the previously reported case of MLP (Fig. 1) where the the distribution of the neurons in the last layer is heavy-tailed.

Another theoretical prediction of Theorem 1 is the independence of the neurons $(Y^i_{\lfloor tL \rfloor})_{1 \leq i \leq L}$. To validate this prediction, we show in figure Fig. 3 the pair-wise joint distributions of 3 randomly chosen neurons in the last layer ($t = 1$). We also perform a kernel density estimation (KDE) using the Gaussian kernel and illustrate the result on top of the histograms. The joint distributions show an excellent match with an isotropic 2-dimensional Gaussian distribution which indicates independence of the neurons.

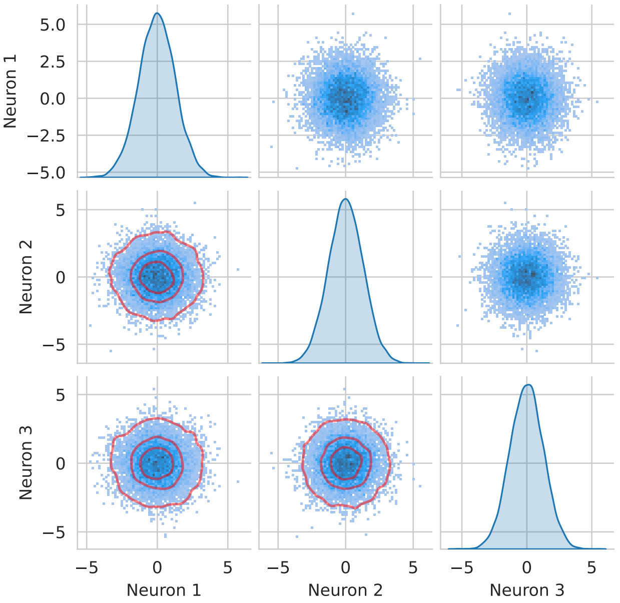
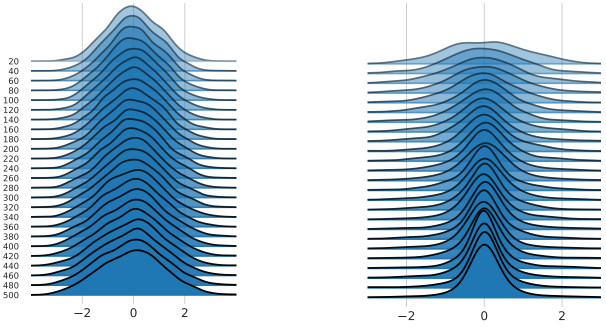


Figure 3: Joint distributions of $(Y_L^i(a), Y_L^j(a))$ for ResNet Eq. (2) with $n, L = 500$, $d = 30$, $i, j \in \{i_1, i_2, i_3\}$ where i_1, i_2, i_3 are randomly sampled from $\{1, \dots, n\}$, and $a = \sqrt{d}\frac{u}{\|u\|}$ and $u \in \mathbb{R}^d$ has all coordinates randomly sampled from the uniform distribution $\mathcal{U}([0, 1])$. The histograms are based on $N = 10^4$ simulations. The red curves represent an isotropic two-dimensional Gaussian distribution (i.e. independent coordinates).

In Fig. 4, we investigate the distribution of the first neuron in each layer in a ResNet/MLP of width $n = 500$ and depth $L = 500$. For the ResNet architecture, the distribution is relatively similar across layers which is expected since Theorem 1 predicts a Gaussian limit with a standard deviation that differs only by a factor of $e^{1/4} \approx 1.28$ between the first and the last layers. In MLPs, the distribution varies across layers with the neurons in the last layers displaying heavy-tailed shapes, which agrees with Fig. 1.

6.2. Convergence of neural covariance

Theorem 2 predicts that the covariance $\hat{q}_t(a, b)$ for two inputs a, b converges in L_2 norm to q_t in the limit $\min(n, L) \to \infty$. In Fig. 5, we compare the empirical covariance $\hat{q}_t$ with the theoretical prediction q_t for $(n, L) \in \{5, 50, 500, 5000\}$. The empirical L_2 error is also reported. As the width increases, we observe a good match with the



(a) ResNet

(b) MLP

Figure 4: Densities (approximated by Kernel Density Estimation) of the first neuron $Y_l^1(a)$ for $l \in \{20k, k = 1, \dots, 25\}$ for a ResNet Eq. (2) and an MLP Eq. (1) with $(n, L) = (500, 500)$. The input a is randomly sampled and normalized in the same way as in Fig. 2.

theory. The role of the depth is less visually noticeable, but for instance, with width $n = 5000$, we can see that the L_2 error is smaller with depth $L = 5000$ as compared to depth $L = 5$ (see Section 6.3 for a more in-depth discussion of the role of width and depth). The theoretical prediction q_t is approximated with a PDE solver (RK45 method, Fehlberg, 1968) for $t \in [0, 1]$ with a discretization step $\Delta t =$1e-4.

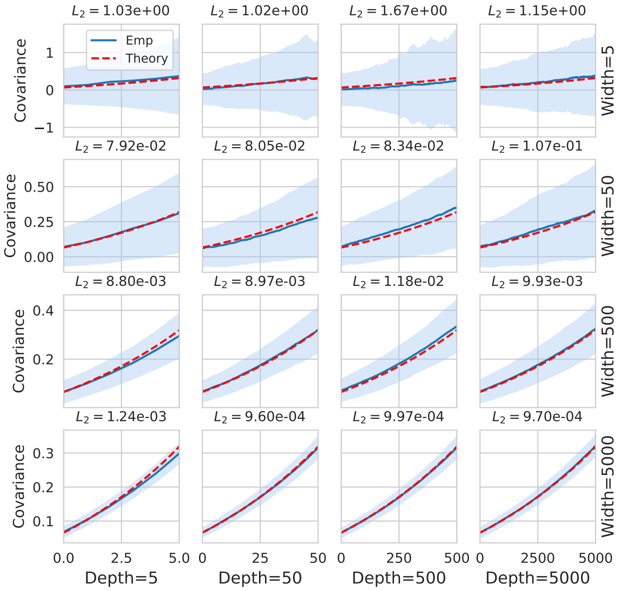


Figure 5: The blue curve represents the average covariance $\hat{q}_t(a, b)$ for ResNet Eq. (2) with $n, L \in \{5, 50, 500, 5000\}$, $d = 30$, and a and b are sampled following the same rule as in Fig. 2. The average is calculated based on $N = 100$ simulations. The shaded blue area represents 1 standard 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 is reported as well.

6.3. Role of width and depth

From Fig. 2 and Fig. 5, it appears that the role of the width is more important than that of the depth in the convergence to the limiting values. In this section, we provide an intuitive explanation as to why that happens. First of all, recall that in both figures, the impact of depth is less noticeable but reflected in some measures (KS statistic in Fig. 2, and L_2 error in Fig. 5). The bounds in Theorem 1 and Theorem 2 are of the form $C\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{L}}\right)$ for some constant C. This bound is sufficient to conclude on the convergence rate but it is not optimal in terms of the constants. We conjecture that a 'better' bound of the form $\frac{C_1}{\sqrt{n}} + \frac{C_2}{\sqrt{L}}$ can be obtained where the constant C_2 is much smaller than C_1, which would explain why the depth has less impact on the bound. To give the reader an intuition of why this should be the case, let us look at the case where the width is much larger than the depth, for instance $n = 500$ and $L \in \{5, 50\}$ (see Fig. 2). Since $n \gg L$, then we are essentially in the regime where the n goes to infinity first. In this case, the impact of depth is limited to how far the finite-depth variance is from infinite-depth one $v(t, a)$ (see Theorem 1). For an input satisfying $\|a\|^2 = d$, simple calculations yield that the infinite-width finite-depth L variance of the neurons in the last layer is given by $\sigma_L = (1 + \frac{1}{2L})^L$.[14] For $L = 5$, $\sigma_5 \approx 1.61$ and for $L = 50$, we have $\sigma_{50} \approx 1.644$. This is very close to the infinite-depth variance given by $v(1, a) = e^{1/2} \approx 1.648$. Hence, even for small depths, the finite-depth variance is close to the infinite-depth variance. Similar analysis can be carried for the covariance as well.

7. Conclusion and Limitations

In this paper, we have shown that, at initialization, in the most natural scaling of branches, the large-depth and large-width limits of a residual neural network (resnet) commute. We used a novel proof technique and proved a concentration of measure result for a kind of McKean-Vlasov process. Our results justify the calculations in prior works analyzing deep and wide neural networks that take the width limit first then depth. However, our technique cannot say anything about what happens when the network starts training. Potentially, different behaviors can occur depending on how the learning rate is chosen as a function

[14] See e.g. Hayou, Clerico, et al., 2021.

of width and depth. Because of the correlations between weights induced by training, such an analysis would likely require far more mathematical machinery than presented here, e.g., Tensor Programs (Yang, 2019a; Yang, 2019b; Yang, 2020b; Yang and E. Hu, 2021; Yang and Littwin, 2021; Yang, E. J. Hu, et al., 2022).

References

Fehlberg, E. (1968). “Classical Fifth-, Sixth-, Seventh-, and Eighth-Order Runge-Kutta Formulas with Stepsize Control”. *NASA Technical Report*.

Ingersoll, J. E. (1987). *Theory of Financial Decision Making*.

Kloeden, P. and E. Platen (1995). *Numerical Solution of Stochastic Differential Equations*. Springer Berlin, Heidelberg, pp. 342–343.

Neal, R. (1995). *Bayesian Learning for Neural Networks*. Vol. 118. Springer Science & Business Media.

Øksendal, B. (2003). *Stochastic Differential Equations*.

Jourdain, B., S. Meleard, and W. Woyczynski (Aug. 2007). “Nonlinear SDEs driven by Lévy processes and related PDEs”. *Latin American journal of probability and mathematical statistics* 4.

Xuerong, M. (2008). *Stochastic differential equations and applications*. Vol. 2. Woodhead Publishing.

Poole, B., S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli (2016). “Exponential expressivity in deep neural networks through transient chaos”. *30th Conference on Neural Information Processing Systems*.

Schoenholz, S., J. Gilmer, S. Ganguli, and J. Sohl-Dickstein (2017). “Deep Information Propagation”. In: *International Conference on Learning Representations*.

Yang, G. and S. Schoenholz (2017). “Mean field residual networks: On the edge of chaos”. In: *Advances in neural information processing systems*, pp. 7103–7114.

Lee, J., Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, and J. Sohl-Dickstein (2018). “Deep Neural Networks as Gaussian Processes”. In: *International Conference on Learning Representations*.

Matthews, A., J. Hron, M. Rowland, R. Turner, and Z. Ghahramani (2018). “Gaussian Process Behaviour in Wide Deep Neural Networks”. In: *International Conference on Learning Representations*.

Tankov, P. and N. Touzi (2018). *CALCUL STOCHASTIQUE ET FINANCE*.

Arora, S., S. Du, W. Hu, Z. Li, and R. Wang (2019). “Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 322–332.

Hanin, B. (2019). “Universal Function Approximation by Deep Neural Nets with Bounded Width and ReLU Activations”. *Mathematics* 7.10.

Hanin, B. and M. Nica (2019). “Products of Many Large Random Matrices and Gradients in Deep Neural Networks”. *Communications in Mathematical Physics* 376.1, pp. 287–322.

Hayou, S., A. Doucet, and J. Rousseau (2019a). “On the Impact of the Activation Function on Deep Neural Networks Training”. In: *International Conference on Machine Learning*.

Hayou, S., A. Doucet, and J. Rousseau (2019b). “Training dynamics of deep networks using stochastic gradient descent via neural tangent kernel”.

Vladimirova, M., J. Verbeek, P. Mesejo, and J. Arbel (Sept. 2019). “Understanding Priors in Bayesian Neural Networks at the Unit Level”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 6458–6467.

Yang, G. (2019a). “Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation”. *arXiv preprint arXiv:1902.04760*.

– (2019b). “Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are Gaussian processes”. *arXiv preprint arXiv:1910.12478*.

Yang, G. and H. Salman (2019). “A fine-grained spectral perspective on neural networks”. *arXiv preprint arXiv:1907.10599*.

Hanin, B. and M. Nica (2020). “Finite Depth and Width Corrections to the Neural Tangent Kernel”. In: *International Conference on Learning Representations*.

Hayou, S., A. Doucet, and J. Rousseau (2020). “Mean-field Behaviour of Neural Tangent Kernel for Deep Neural Networks”. *arXiv preprint arXiv:1905.13654*.

He, B., B. Lakshminarayanan, and Y. W. Teh (2020). “Bayesian Deep Ensembles via the Neural Tangent Kernel”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 1010–1022.

Hron, J., Y. Bahri, J. Sohl-Dickstein, and R. Novak (2020). “Infinite attention: NNGP and NTK for deep attention networks”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 4376–4386.

Peluchetti, S. and S. Favaro (2020). “Infinitely deep neural networks as diffusion processes”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 1126–1136.

Xiao, L., J. Pennington, and S. Schoenholz (2020). “Disentangling Trainability and Generalization in Deep Neural Networks”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 10462–10472.

Yang, G. (2020a). “Tensor Programs III: Neural Matrix Laws”. *arXiv preprint arXiv:2009.10685*.

Yang, G. (Aug. 2020b). “Tensor Programs II: Neural Tangent Kernel for Any Architecture”. *arXiv:2006.14548 [cond-mat, stat]*. arXiv: 2006.14548.

Hayou, S., J. Ton, A. Doucet, and Y. Teh (2021). “Robust Pruning at Initialization”. In: *International Conference on Learning Representations*.

Hayou, S. and F. Ayed (2021). “Regularization in ResNet with Stochastic Depth”. *Proceedings of Thirty-fifth Neural Information Processing Systems (NeurIPS)*.

Hayou, S., E. Clerico, B. He, G. Deligiannidis, A. Doucet, and J. Rousseau (13–15 Apr 2021). “Stable ResNet”. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Banerjee and K. Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 1324–1332.

Li, M., M. Nica, and D. Roy (2021). “The future is log-Gaussian: ResNets and their infinite-depth-and-width limit at initialization”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., pp. 7852–7864.

Martens, J., A. Ballard, G. Desjardins, G. Swirszcz, V. Dalibard, J. Sohl-Dickstein, and S. S. Schoenholz (2021). “Rapid training of deep neural networks without skip connections or normalization layers using Deep Kernel Shaping”. *arXiv, preprint 2110.01765*.

Noci, L., G. Bachmann, K. Roth, S. Nowozin, and T. Hofmann (2021). “Precise characterization of the prior predictive distribution of deep ReLU networks”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan.

Yang, G. and E. Hu (2021). “Tensor Programs IV: Feature Learning in Infinite-Width Neural Networks”. *ICML 2021*.

Yang, G. and E. Littwin (May 2021). “Tensor Programs IIb: Architectural Universality of Neural Tangent Kernel Training Dynamics”. *arXiv:2105.03703 [cs, math]*. arXiv: 2105.03703 version: 1.

Zavatone-Veth, J. and C. Pehlevan (2021). “Exact marginal prior distributions of finite Bayesian neural networks”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., pp. 3364–3375.

Hanin, B. (2022). “Correlation Functions in Random Fully Connected Neural Networks at Finite Width”.

Hayou, S. (2022). “On the infinite-depth limit of finite-width neural networks”. *Transactions on Machine Learning Research*.

Hayou, S., A. Doucet, and J. Rousseau (2022). “The Curse of Depth in Kernel Regime”. In: *Proceedings on ”I (Still) Can’t Believe It’s Not Better!” at NeurIPS 2021 Workshops*. Ed. by M. F. Pradier, A. Schein, S. Hyland, F. J. R. Ruiz, and J. Z. Forde. Vol. 163. Proceedings of Machine Learning Research. PMLR, pp. 41–47.

Jacot, A. (2022). *Theory of Deep Learning: Neural Tangent Kernel and Beyond*.

Jacot, A., F. Gabriel, F. Ged, and C. Hongler (2022). “Freeze and Chaos: NTK views on DNN Normalization, Checkerboard and Boundary Artifacts”. In: *Proceedings of Mathematical and Scientific Machine Learning*. Ed. by B. Dong, Q. Li, L. Wang, and Z.-Q. J. Xu. Vol. 190. Proceedings of Machine Learning Research. PMLR, pp. 257–270.

Li, M. B., M. Nica, and D. M. Roy (2022). “The Neural Covariance SDE: Shaped Infinite Depth-and-Width Networks at Initialization”. *arXiv*.

Liu, F., H. Yang, S. Hayou, and Q. Li (2022). “Connecting Optimization and Generalization via Gradient Flow Path Length”.

Lou, Y., C. E. Mingard, and S. Hayou (2022). “Feature Learning and Signal Propagation in Deep Neural Networks”. In: *Proceedings of the 39th International Conference on Machine Learning*, pp. 14248–14282.

Marion, P., A. Fermanian, G. Biau, and J.-P. Vert (2022). “Scaling ResNets in the Large-depth Regime”. *arXiv*.

Seleznova, M. and G. Kutyniok (2022). “Analyzing Finite Neural Networks: Can We Trust Neural Tangent Kernel Theory?” In: *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*. Ed. by J. Bruna, J. Hesthaven, and L. Zdeborova. Vol. 145. Proceedings of Machine Learning Research. PMLR, pp. 868–895.

Wolinski, P. and J. Arbel (2022). “Gaussian Pre-Activations in Neural Networks: Myth or Reality?”

Yang, G., E. J. Hu, I. Babuschkin, S. Sidor, X. Liu, D. Farhi, N. Ryder, J. Pachocki, W. Chen, and J. Gao (Mar. 2022). “Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer”. *arXiv:2203.03466 [cond-mat]*. arXiv: 2203.03466.

Yang, G., M. Santacroce, and E. J. Hu (2022). “Efficient Computation of Deep Nonlinear Infinite-Width Neural Networks that Learn Features”. In: *International Conference on Learning Representations*.

Zhang, G., A. Botev, and J. Martens (2022). “Deep Learning without Shortcuts: Shaping the Kernel with Tailored Rectifiers”. In: *International Conference on Learning Representations*.

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, Clerico, 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, Doucet, et al., 2019a, 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 behaviour*: Multiple papers (e.g. Neal, 1995; Lee et al., 2018; Matthews et al., 2018; Hron et al., 2020; Yang, 2020a) 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 behaviour 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, Clerico, et al., 2021, where the authors showed that with proper scaling, the infinite-depth (weak) limit is a Gaussian process with a universal kernel[15].

- *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 Arora et al., 2019; Hayou, Doucet, et al., 2019b; Liu et al., 2022; Seleznova et al., 2022. The large depth asymptotics of the NTK have been studied in (Hayou, Doucet, et al., 2020; Xiao et al., 2020; Hayou, Doucet, et al., 2022; Jacot et al., 2022). 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, 2019a; Yang, 2019b; Yang, 2020a; Yang and E. Hu, 2021).

- *Others*: the theory of infinite-width neural networks have also been utilized for network pruning (Hayou, Ton, et al., 2021), regularization (Vladimirova et al., 2019; Hayou and Ayed, 2021), feature learning (Lou et al., 2022), and ensembling methods (He et al., 2020).

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

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

scheme (Poole et al., 2016; Schoenholz et al., 2017), investigating the impact of the activation function (Hayou, Doucet, et al., 2019a), and analyzing the behavior of the Neural Tangent Kernel (NTK) (Hayou, Doucet, et al., 2020; Xiao et al., 2020).

The joint infinite-width-and-depth limit. In this case, the depth-to-width ratio is fixed[16], 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 (M. 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, Doucet, et al., 2019a)). Furthermore, in (M. B. Li et al., 2022), 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, Doucet, et al., 2020). In (Hanin, 2019; Hanin, 2022), 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 (Noci et al., 2021; Zavatone-Veth et al., 2021).

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 et al., 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, 2022 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.

B. Review of Stochastic Calculus

In this section, we present the mathematical framework for the study of stochastic differential equations (SDEs). We consider a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}t)t \geq 0)$, where Ω is the sample space, $\mathcal{F}$ is the sigma-algebra of events, $\mathbb{P}$ is the probability measure, and $(\mathcal{F}t)t \geq 0$ is the natural filtration of a standard n-dimensional Brownian motion B. This framework allows us to study the evolution of a stochastic process X over time, by considering the events that are measurable up to a given time t. Specifically, we focus on the class of Itô processes, which are defined through a specific type of stochastic differential equation.

B.1. Existence and uniqueness

Definition 4 (Itô diffusion process)**.** *A stochastic process $(X_t)_{t\in[0,T]}$ valued in $\mathbb{R}^n$ is called an Itô diffusion process if it can be expressed as*

$$X_t = X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dB_s,$$

where B is a n-dimensional Brownian motion and $\sigma_t \in \mathbb{R}^{n\times n}, \mu \in \mathbb{R}^n$ are predictable processes satisfying $\int_0^T (\|\mu_s\|_2 + \|\sigma_s \sigma_s^\top\|_2) ds < \infty$ almost surely.

The following result gives conditions under which a strong solution of a given SDE exists, and is unique.

Theorem 4 (Thm 3.1 and Lemma 3.2 in Xuerong, 2008)**.** *Let $n \geq 1$, and consider the following SDE*

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dB_t, \quad X_0 \in L_2,$$

[16] Other works consider the case when the depth-to-width ratio converge to a constant instead of being fixed.

where B is a m-dimensional Brownian process for some $m \geq 1$, and $\mu : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are measurable functions satisfying

1. *There exists a constant $K > 0$ such that for all $t \geq 0$, $x, x' \in \mathbb{R}^n$*

$$\|\mu(t,x) - \mu(t,x')\| + \|\sigma(t,x) - \sigma(t,x')\| \leq K\|x - x'\|.$$

2. *There exists a constant $K' > 0$ such that for all $t \geq 0$, $x \in \mathbb{R}^n$*

$$\|\mu(t,x)\| + \|\sigma(t,x)\| \leq K'(1 + \|x\|).$$

Then, for all $T \geq 0$, there exists a unique strong solution of the SDE above, and it satisfies the following

$$\mathbb{E} \sup_{0\leq t\leq T} \|X_t\|^2 \leq C(1 + \mathbb{E}\|X_0\|^2),$$

where C is a constant that depends only on K, K', and T.

B.2. Itô 's lemma

The following result, known as Itô 's lemma, is a classic result in stochastic calculus. We state a version of this result from Tankov et al., 2018. Other versions and extensions exist in the literature (e.g. Ingersoll (1987), Kloeden et al. (1995), and Øksendal (2003)).

Lemma 2 (Itô 's lemma, Thm 6.7 in Tankov et al., 2018)**.** *Let X_t be an Itô diffusion process (Definition 4) of the form*

$$dX_t = \mu_t dt + \sigma_t dB_t, t \in [0,T], X_0 \sim \nu$$

where ν is some given distribution. Let $g : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^{1,2}([0,T], \mathbb{R}^n)$ (i.e. $\mathcal{C}^1$ in the first variable t and $\mathcal{C}^2$ in the second variable x). Then, with probability 1, we have that

$$f(t,X_t) = f(0,X_0) + \int_0^t \nabla_x f(s,X_s) \cdot dX_s + \int_0^t \left(\partial_t f(s,X_s) + \frac{1}{2}\mathrm{Tr}\left[\sigma_s^\top \nabla_x^2 f(s,X_s)\sigma_s\right]\right) ds,$$

where $\nabla_x f$ and $\nabla_x^2 f$ refer to the gradient and the Hessian, respectively. This can also be expressed as an SDE

$$df(t,X_t) = \nabla_x f(t,X_t) \cdot dX_t + \left(\partial_t f(t,X_t) + \frac{1}{2}\mathrm{Tr}\left[\sigma_t^\top \nabla_x^2 f(t,X_t)\sigma_t\right]\right) dt.$$

B.3. Convergence of Euler's scheme to the SDE solution

The following result gives a convergence rate of the Euler discretization scheme to the solution of the SDE.

Theorem 5 (Corollary of Thm 7.3 in Xuerong, 2008)**.** *Let $d \geq 1$ and consider the $\mathbb{R}^d$-valued ito process X (Definition 4) given by*

$$X_t = X_0 + \int_0^t \mu(s,X_s)ds + \int_0^t \sigma(s,X_s)dB_s,$$

where B is a m-dimensional Brownian motion for some $m \geq 1$, X_0 satisfies $\mathbb{E}\|X_0\|^2 < \infty$, and $\mu : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$ are measurable functions satisfying the following conditions:

1. *There exists a constant $K > 0$ such that for all $t \in \mathbb{R}, x, x' \in \mathbb{R}^d$,*

$$\|\mu(t,x) - \mu(t,x')\|^2 + \|\sigma(t,x) - \sigma(t,x')\|^2 \leq \bar{K}\|x - x'\|^2.$$

2. *There exists a constant $K' > 0$ such that for all $t \in \mathbb{R}, x \in \mathbb{R}^d$*

$$\|\mu(t,x)\|^2 + \|\sigma(t,x)\|^2 \leq K(1 + \|x\|^2).$$

Let $\delta \in (0,1)$ such that $\delta^{-1} \in \mathbb{N}$ (integer), and consider the times $t_k = k\delta$ for $k \in \{1, \dots, \delta^{-1}\}$. Consider the Euler discretization scheme given by

$$\bar{X}^i_{k+1} = \bar{X}^i_k + \mu^i(t_k, \bar{X}^k_n)\delta + \sum_{j=1}^{m} \sigma^{i,j}(t_k, \bar{X}^k_n)\Delta B^j_k, \quad \bar{X}^i_0 = X^i_0,$$

where $\bar{X}^i, \mu^i, \sigma^{i,j}$ denote the coordinates of these vectors for $i \in [d], j \in [m]$, and $\Delta B^j_k = B^j_{k+1} - B^j_k \sim \mathcal{N}(0, \delta)$. Then, we have that

$$\mathbb{E} \sup_{t \in [0,1]} \|X_t - \bar{X}_{\lfloor t\delta^{-1} \rfloor}\|^2 \leq C\,\delta,$$

where $C = 80K\bar{K}(1 + (1 + 3\mathbb{E}\|X_0\|^2)\exp(6K))\exp(20\bar{K})$.

Proof. The proof is straightforward by taking $T = 1$ and $t_0 = 0$ in Thm 7.3 in Xuerong, 2008. □

Using this result, we prove the following width-uniform convergence result for infinite-depth, which is crucial to our results.

Theorem 6 (Width-uniform convergence)**.** *Assume that the activation function ϕ is Lipschitz on $\mathbb{R}$ with Lipschitz constant $\zeta > 0$ and that $\phi(0) = 0$, and let $a \in \mathbb{R}^d$ be a non-zero vector. Consider the process X_t the solution of the following SDE*

$$dX_t = \frac{1}{\sqrt{n}}\|\phi(X_t)\|dB_t, \quad X_0 = W_{in}a, \tag{4}$$

where $(B_t)_{t\geq 0}$ is a Brownian motion (Wiener process), and let $\bar{X}$ be its Euler scheme as in Theorem 5. Then, we have the following width-uniform bound on the discretization error:

$$\sup_{n \geq 1} n^{-1}\,\mathbb{E} \sup_{t \in [0,1]} \|X_t - \bar{X}_{\lfloor t\delta^{-1} \rfloor}\|^2 \leq C'\,\delta,$$

where $C' = 80\zeta^4(1 + (1 + 3d^{-1}\|a\|^2)\exp(6\zeta^2))\exp(20\zeta^2)$.

Proof. The key observation in this proof is that the constant C in Theorem 5 scales linearly with width. Indeed, in this case, the volatility term is given by $\sigma(x) = \frac{1}{\sqrt{n}}\|\phi(x)\|I_n$, which satisfies the linear growth condition

$$\|\sigma(x)\| = \frac{1}{\sqrt{n}}\|\phi(x)\|\|I_n\| = \|\phi(x)\| \leq \zeta\|x\|,$$

where we have used the fact that $\|I_n\| = \sqrt{\mathrm{Tr}(I_n I_n^\top)} = \sqrt{n}$[17]. Moreover, for any $x, x' \in \mathbb{R}^n$, we have that

$$\|\sigma(x) - \sigma(x')\| \leq \left| \frac{1}{\sqrt{n}}\|\phi(x)\| - \frac{1}{\sqrt{n}}\|\phi(x)\| \right| \|I_n\| \leq \zeta\|x - x'\|,$$

Hence, in this case we can set $\bar{K} = K = \zeta^2$. We conclude by observing that $\mathbb{E}\|X_0\|^2 = nd^{-1}\|a\|^2$ and using Theorem 5. □

The result of Theorem 6 can be generalized to the case of multiple inputs as we show in the next result. We omit the proof here as this result is not necessary for the proofs of the main results.

Theorem 7. *Let $a_1, a_2, \dots, a_k \in \mathbb{R}^d$ be non-zero inputs, and assume that the activation function ϕ is Lipschitz on $\mathbb{R}$ and that $\phi(0) = 0$. Consider the process X^k_t, the solution of the following SDE*

$$d\boldsymbol{X}^k_t = \frac{1}{\sqrt{n}}\Sigma(\boldsymbol{X}^k_t)^{1/2}d\boldsymbol{B}_t, \quad \boldsymbol{X}^k_0 = ((W_{in}a_1)^\top, \dots, (W_{in}a_k)^\top)^\top, \tag{5}$$

[17] In (almost) all the results on the existence, uniqueness, and Euler schemes in stochastic calculus, the default matrix norm is the Frobenius norm.

where $(\boldsymbol{B}_t)_{t\geq 0}$ is an kn-dimensional Brownian motion (Wiener process), independent from W_{in}, and $\Sigma(\boldsymbol{X}_t^k)$ is the covariance matrix given by

$$\Sigma(\boldsymbol{X}_t^k) = \left[\begin{array}{c|c|c|c} \alpha_{1,1}I_n & \alpha_{1,2}I_n & \dots & \alpha_{1,k}I_n \\ \hline \alpha_{2,1}I_n & \alpha_{2,2}I_n & \dots & \alpha_{2,k}I_n \\ \hline \vdots & \vdots & \vdots & \vdots \\ \alpha_{k,1}I_n & \dots & \dots & \alpha_{k,k}I_n \end{array}\right],$$

where $\alpha_{i,j} = \langle \phi(\boldsymbol{X}_t^{k,i}), \phi(\boldsymbol{X}_t^{k,j})\rangle$, with $(X_t^{k,1\top}, \dots, X_t^{k,k\top})^\top \overset{def}{=} \boldsymbol{X}_t^k$.

Let $\bar{\boldsymbol{X}}^k$ be its Euler scheme as in Theorem 5. Then, we have the following width-uniform bound on the discretization error:

$$\sup_{n\geq 1}(kn)^{-1}\,\mathbb{E}\sup_{t\in[0,1]}\|\boldsymbol{X}_t^k - \bar{\boldsymbol{X}}^k_{\lfloor t\delta^{-1}\rfloor}\|^2 \leq C'\,\delta,$$

where $C' = 80\zeta^4(1+(1+3d^{-1}\|a\|^2)\exp(6\zeta^2))\exp(20\zeta^2)$.

B.4. Convergence of Particles to the solution of Mckean-Vlasov process

The next result gives sufficient conditions for the system of particles to converge to its mean-field limit, known as the Mckean-Vlasov process.

Theorem 8 (Uniform Mckean-Vlasov process)**.** *Let $d \geq 1$ and consider the $\mathbb{R}^d$-valued ito process X (Definition 4) given by*

$$dX_t = \sigma(\nu_t^n)dB_t, \quad X_0 = W_{in}a,$$

where B is a d-dimensional Brownian motion, $W_{in}^{ij} \sim \mathcal{N}(0, 1/d)$, $a \in \mathbb{R}^d$ and $a \neq 0$, $\nu_t^n \overset{def}{=} \frac{1}{d}\sum_{i=1}^d \delta_{\{X_t^i\}}$ is the empirical distribution of the coordinates of X_t, and σ is real-valued given by $\sigma(\nu) = \left(\int \phi(y)^2 d\nu(y)\right)^{1/2}$ for any distribution ν, where ϕ is the ReLU activation function. Then, for all $T \in \mathbb{R}^+$, we have that

$$\sup_{i\in[n]}\mathbb{E}\left(\sup_{t\leq T}|X_t^i - \tilde{X}_t^i|^2\right) = \mathcal{O}(n^{-1}),$$

where $\tilde{X}^i$ is the solution of the following Mckean-Vlasov equation

$$d\tilde{X}_t^i = \sigma(\nu_t^i)dB_t^i = \frac{\|a\|}{\sqrt{2d}}\exp(t/4)dB_t^i, \quad \tilde{X}_0^i = X_0^i,$$

where ν_t^i is the distribution of $\tilde{X}^i$. The constant in the $\mathcal{O}$ depends only on T and the norm of a.

Proof. The first part of the proof is similar to that of Theorem 3 in (Jourdain et al., 2007). In the second part, we use a concentration argument to control the deviations of the volatility term which allow us to conclude.

Let $\tilde{\upsilon}_t^n$ denote the empirical distribution of the independent processes $\tilde{X}_t^i, i \in [n]$ defined in the statement of the theorem. Let $t \in [0,1]$. Following (Jourdain et al., 2007), for some $i \in [n]$, using Doob's inequality, there exists a universal constant $C > 0$ such that

$$\begin{aligned}\mathbb{E}\left(\sup_{s\leq t}|X_s^i - \tilde{X}_s^i|^2\right) &\leq C\int_0^t \mathbb{E}|\sigma(\nu_s^n) - \sigma(\nu_s)|^2 ds \\ &\leq C\int_0^t \mathbb{E}|\sigma(\nu_s^n) - \sigma(\tilde{\nu}_s^n)|^2 ds + C\int_0^t \mathbb{E}|\sigma(\tilde{\nu}_s^n) - \sigma(\nu_s)|^2 ds.\end{aligned}$$

For the first term, we have that

$$\begin{aligned}\int_0^t \mathbb{E}|\sigma(\nu_s^n) - \sigma(\tilde{\nu}_s^n)|^2 ds &= \int_0^t \mathbb{E}\left|\frac{1}{\sqrt{n}}\|\phi(X_s)\| - \frac{1}{\sqrt{n}}\|\phi(\tilde{X}_s)\|\right|^2 ds\\ &\leq \frac{1}{n}\int_0^t \mathbb{E}\|\phi(X_s) - \phi(\tilde{X}_s)\|^2 ds\\ &\leq \int_0^t \mathbb{E}\left(\sup_{r\leq s}|X_r^i - \tilde{X}_r^i|^2\right) ds,\end{aligned}$$

where we have used the exchangeability of the couples $(X_t^i, \tilde{X}_t^i)$ (across i) and the Lipschitz property of ζ. Therefore, using Gronwall's lemma, there exists a constant $C' > 0$ (independent of i) such that

$$\mathbb{E}\left(\sup_{s\leq t}|X_s^i - \tilde{X}_s^i|^2\right) \leq C'\int_0^t \mathbb{E}|\sigma(\tilde{\nu}_s^n) - \sigma(\nu_s)|^2 ds.$$

Since the bound is uniform in i, we then have

$$\sup_{i\in[n]}\mathbb{E}\left(\sup_{s\leq t}|X_s^i - \tilde{X}_s^i|^2\right) \leq C'\int_0^t \mathbb{E}|\sigma(\tilde{\nu}_s^n) - \sigma(\nu_s)|^2 ds.$$

Thus, it suffices to show that the right hand side is of order n^{-1} to conclude. Let us first show that the volatility of the process $\tilde{X}_t^i$ is given by $\sigma(\nu_t^i) = \frac{\|a\|}{\sqrt{2d}}\exp(t/4)$. We have that $d\tilde{X}_t^i = \sigma(\nu_t^i)dB_t^i$. A simple application of Itô 's lemma (Lemma 2) yields

$$d\mathbb{E}(\tilde{X}_t^i)^2 = \frac{1}{2}\mathbb{E}(\tilde{X}_t^i)^2 dt,$$

where we have used the fact that with ReLU $\mathbb{E}(\phi(\tilde{X}_t^i)^2) = \frac{1}{2}\mathbb{E}(\tilde{X}_t^i)^2$. Therefore, we obtain $\mathbb{E}(\tilde{X}_t^i)^2 = \mathbb{E}(\tilde{X}_0^i)^2\exp(t/2) = \frac{\|a\|^2}{d}\exp(t/2)$. Thus, the volatility term is given by stated formula. Notice that $\hat{X}_t^i$ has a normal distribution in this case.

We now use Hoeffding's inequality for random variables with sub-exponential growth to control the deviations of $\sigma(\tilde{\nu}_s^n)^2$. We have

$$\begin{aligned}\mathbb{P}\left(\sigma(\tilde{\nu}_s^n)^2 \leq \frac{\|a\|^2}{4d}\right) &\leq \mathbb{P}\left(\sigma(\tilde{\nu}_s^n)^2 \leq \sigma(\nu_s)^2/2\right)\\ &= \mathbb{P}\left(\sigma(\tilde{\nu}_s^n)^2 - \sigma(\nu_s)^2 \leq -\sigma(\nu_s)^2/2\right)\\ &\leq 2\exp(-nc),\end{aligned}$$

where $c > 0$ is a constant that depends only on the moments of $\phi(\tilde{X}_t^i)$ which can be upper-bounded uniformly for $t \in [0, T]$. Define the event $\mathcal{H}_n = \{\sigma(\tilde{\nu}_s^n)^2 \leq \frac{\|a\|^2}{4d}\}$ and let $\bar{\mathcal{H}}_n$ denote its complementary event. This yields for all $s \in [0, T]$

$$\begin{aligned}\mathbb{E}|\sigma(\tilde{\nu}_s^n) - \sigma(\nu_s)|^2 &= \mathbb{E}\,\mathbb{1}_{\mathcal{H}_n}|\sigma(\tilde{\nu}_s^n) - \sigma(\nu_s)|^2 + \mathbb{E}\mathbb{1}_{\bar{\mathcal{H}}_n}|\sigma(\tilde{\nu}_s^n) - \sigma(\nu_s)|^2\\ &\leq \frac{2\|a\|^2}{d}\exp\left(-n\,c + \frac{s}{2}\right) + \left(\frac{d}{4}\right)^{1/4}\mathbb{E}|\sigma(\tilde{\nu}_s^n)^2 - \sigma(\nu_s)^2|^2\\ &\leq \frac{2\|a\|^2}{d}\exp\left(-n\,c + \frac{s}{2}\right) + \left(\frac{\sqrt{d}}{2\|a\|}\right)\frac{\mathbb{E}\phi(\tilde{X}_s^1)^4}{n},\end{aligned}$$

where we have use the fact that $|\sqrt{z} - \sqrt{z'}| \leq \frac{1}{2\sqrt{z_0}}|z - z'|$ for $z, z' \geq z_0 > 0$. Since $\tilde{X}_s^1$ is a zero-mean Gaussian with a variance that depends only on s, we can therefore conclude that there exists C'' independent of n and $i \in [n]$ such that

$$\sup_{i\in[n]}\mathbb{E}\left(\sup_{s\leq t}|X_s^i - \tilde{X}_s^i|^2\right) \leq C''\,n^{-1},$$

which concludes the proof. □

B.5. Other results from probability and stochastic calculus

The next trivial lemma has been opportunely used in M. Li et al., 2021 to derive the limiting distribution of the network output (multi-layer perceptron) in the joint infinite width-depth limit. This simple result will also prove useful in our case of the finite-width-infinite-depth limit.

Lemma 3. *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. □

C. Some technical results for the proofs

Proposition 4. *Assume that the activation function ϕ is Lipschitz on $\mathbb{R}$ and let $a \in \mathbb{R}^d$ with $a \neq 0$. Then, in the limit $L \to \infty$, the process $X^L_t(a) = Y_{\lfloor tL\rfloor}(a)$, $t \in [0,1]$, converges in distribution to the solution of the following SDE*

$$dX_t(a) = \frac{1}{\sqrt{n}}\|\phi(X_t(a))\|dB_t, \quad X_0(a) = W_{in}a, \tag{6}$$

where $(B_t)_{t\geq 0}$ is a Brownian motion (Wiener process). Moreover, we have that

$$\sup_{n\geq 1}\sup_{1\leq t\leq 1} \mathcal{W}_1(\mu^t_{n,L}, \mu^t_{n,\infty}) \leq CL^{-1/2},$$

where $\mu^t_{n,L}(a)$ is the distribution of $Y^i_{\lfloor tL\rfloor}(a)$, $\mu^t_{n,\infty}(a)$ is the distribution $X^i_t(a)$ (for any i since the coordinates are identically distributed), and C is a constant that depends only on d and $\|a\|$.

Proof. The proof is based on Theorem 6 in the appendix. It remains to express Eq. (2) in the required form and make sure all the conditions are satisfied for the result to hold. To alleviate the notation, we denote $Y_l := Y_l(a)$. Using Lemma 3, we can write Eq. (2) as

$$Y_l = Y_{l-1} + \frac{1}{\sqrt{L}}\sigma(Y_{l-1})\zeta^L_{l-1},$$

where $\sigma(y) \overset{def}{=} \frac{1}{\sqrt{n}}\|\phi(y)\|$ for all $y \in \mathbb{R}^n$ and ζ^L_l are *iid* random Gaussian vectors with distribution $\mathcal{N}(0, I)$. This is equal in distribution to the Euler scheme of SDE Eq. (6). Since σ trivially inherits the Lipschitz or local Lipschitz properties of ϕ, we conclude for the convergence using Theorem 6.

Now let Ψ be 1-Lipschitz. We have that

$$|\mathbb{E}\Psi(Y_{\lfloor tL\rfloor}) - \mathbb{E}\Psi(X_t)| \leq \mathbb{E}\|\bar{X}_{\lfloor tL\rfloor} - X_t\| \leq CL^{-1/2}.$$

where $\bar{X}$ is the Euler scheme as in Theorem 6, and where we have used the fact that $Y_{\lfloor tL\rfloor}$ and $\bar{X}_{\lfloor tL\rfloor}$ have the same distribution, coupled with the Cauchy-Schwartz inequality. Since C depends only on d and $\|a\|$, the conclusion is straightforward. □

Proposition 5. *Assume that the activation function ϕ is Lipschitz on $\mathbb{R}$ and let $a, b \in \mathbb{R}^d$ with $a, b \neq 0$ and $a \neq b$. Then, there exists two n-dimensional Brownian motions $B_t(a)$ and $B_t(b)$ and a discretized Euler scheme $(\bar{X}(a))$ and $(\bar{X}(b))$ such that for any $t \in [0,1]$, the processes $(Y_{\lfloor tL \rfloor}(a), Y_{\lfloor tL \rfloor}(b))$ have the same distribution as $(\bar{X}_{\lfloor tL \rfloor}(a), \bar{X}_{\lfloor tL \rfloor}(b))$ and $\bar{X}_{\lfloor tL \rfloor}(a)$ and $\bar{X}_{\lfloor tL \rfloor}(b)$ converge (in L_2) to the solutions of the following SDEs*

$$\begin{aligned} dX_t(a) &= \frac{1}{\sqrt{n}} \|\phi(X_t(a))\| dB_t(a), \quad X_0(a) = W_{in} a, \\ dX_t(b) &= \frac{1}{\sqrt{n}} \|\phi(X_t(b))\| dB_t(b), \quad X_0(b) = W_{in} b, \end{aligned} \tag{7}$$

Moreover, we have that

$$\lim_{n \to \infty} \mathbb{E}\left[\frac{\langle X_t(a), X_t(b) \rangle}{n}\right] = q_t(a,b),$$

where $q_t(a,b)$ is the solution of the following Ordinary Differential Equation

$$\begin{aligned} \frac{dq_t(a,b)}{dt} &= \frac{1}{2} \frac{f(c_t(a,b))}{c_t(a,b)} q_t(a,b), \\ c_t(a,b) &= \frac{q_t(a,b)}{\sqrt{q_t(a,a)}\sqrt{q_t(b,b)}}, \\ q_0(a,b) &= \frac{\langle a, b \rangle}{d}, \end{aligned} \tag{8}$$

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

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

Proof. The proof is similar to that of Proposition 4. The only difference lies the definition of the Gaussian vector ζ_l^L. In this case, for $x \in \{a, b\}$, we have

$$Y_l(x) = Y_{l-1}(x) + \frac{1}{\sqrt{L}} \frac{1}{\sqrt{n}} \zeta_{l-1}^L(Y_{l-1}(x)),$$

where $\zeta_{l-1}^L(Y_{l-1}(x)) \overset{def}{=} \sqrt{n} W_l \phi(Y_{l-1}(x))$. It is straightforward that we can write $\frac{1}{\sqrt{L}} \zeta_{l-1}^L(Y_{l-1}(x))$ as a Brownian increment $\Delta B_l(x) = L^{-1/2} \zeta_{l-1}^L(Y_{l-1}(x))$. Defining the Euler schemes $\bar{X}(a), \bar{X}(b)$ with the Brownian motions $(B_t(x))_{x \in \{a,b\}}$ yields that the concatenated vector $(Y_{\lfloor tL \rfloor}(a), Y_{\lfloor tL \rfloor}(b))$ has the same distribution as $(\bar{X}_{\lfloor tL \rfloor}(a), \bar{X}_{\lfloor tL \rfloor}(b))$. In particular, this implies that

$$\mathbb{E}\left[\frac{\langle \bar{X}_{\lfloor tL \rfloor}(a), \bar{X}_{\lfloor tL \rfloor}(b) \rangle}{n}\right] = \mathbb{E}\left[\frac{\langle Y_{\lfloor tL \rfloor}(a), Y_{\lfloor tL \rfloor}(b) \rangle}{n}\right].$$

Now using Theorem 6, we know that for $x \in \{a, b\}$,

$$\sup_{n \geq 1} n^{-1} \mathbb{E} \sup_{t \in [0,1]} \|X_t(x) - \bar{X}_{\lfloor tL \rfloor}(x)\|^2 \leq C' \, \delta,$$

where C' depends only on the $\|x\|$ and d^{-1}. From this, and by observing that the L_2 norm of $X_t(x)$ and $\bar{X}_{\lfloor tL \rfloor}(x)$ are upperbounded (see Theorem 4), it is straightforward that

$$\left| \mathbb{E}\left[\frac{\langle X_t(a), X_t(b) \rangle}{n}\right] - \mathbb{E}\left[\frac{\langle Y_{\lfloor tL \rfloor}(a), Y_{\lfloor tL \rfloor}(b) \rangle}{n}\right] \right| \leq C L^{-1/2},$$

where C is a constant that depends only on $\|a\|$, $\|b\|$, and d. To conclude, we will take the width to infinity first then take the depth to infinity. Taking $n \to \infty$, then depth to ∞ (standard result, see Lemma 5 in Hayou, Clerico, et al., 2021) yields

$$\lim_{L \to \infty} \lim_{n \to \infty} \mathbb{E}\left[\frac{\langle Y_{\lfloor tL \rfloor}(a), Y_{\lfloor tL \rfloor}(b) \rangle}{n}\right] = q_t(a,b),$$

which concludes the proof.

□

D. Proof of Theorem 1

Theorem 1 [Width/Depth uniform convergence of the pre-activations]
Let $a \in \mathbb{R}^d$ such that $a \neq 0$. For $t \in [0,1]$ and $i \in [n]$ fixed, the random variable $(Y_{\lfloor tL \rfloor}(a))_{L \geq 1}$ converges weakly to a Gaussian random variable with law $\mathcal{N}(0, v(t,a))$ in the limit of $\min(n, L) \to \infty$, where $v(t,a) = d^{-1}\|a\|^2 \exp(t)$. Moreover, we have the following convergence rate

$$\sup_{t\in[0,1]} \mathcal{W}_1(\mu^t_{n,L}(a), \mu^t_{\infty,\infty}(a)) \leq C\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{L}}\right)$$

where $\mu^t_{n,L}(a)$ is the distribution of $Y^1_{\lfloor tL \rfloor}(a)$, $\mu^t_{\infty,\infty}(a)$ is the distribution $\mathcal{N}(0, v(t,a))$, and C is constant that depends only on $\|a\|$ and d.

Proof. The proof relies on a careful manipulation of the order of the depth and width limits. Unlike existing literature on the infinite-width-then-depth networks, we found that is much easier to control the convergence rate by looking at what happens when the L diverges first, then control over n. This uses two main ingredients:

- A new width-uniform convergence rate of the Euler discretization scheme of the infinite-depth SDE. We prove this in Theorem 6.
- A new particle convergence result to a McKean-Vlasov process (Mean-Field limit). We prove this result in Theorem 8.

Let $a \in \mathbb{R}^d$ with $a \neq 0$.

Part 1: Width-uniform infinite-depth limit. Let $n \geq 1$ be fixed for now, and let us look at what happens in the infinite depth limit. Using Proposition 4, we know that $Y^1_{\lfloor tL \rfloor}(a)$ converges in distribution to $X^1_t(a)$ with a width-uniform rate in terms of the Wasserstein distance

$$\sup_{1\leq t\leq 1} \mathcal{W}_1(\mu^t_{n,L}, \mu^t_{n,\infty}) \leq CL^{-1/2},$$

where C depends only on d and $\|a\|$.

Part 2: Taking the width to infinity. The rest of the proof rely on a new technical result that we prove in Theorem 8. The intuition is that the coordinates $(X^i(a)_t)_{1\leq i\leq n}$ can be seen as interacting particles of some underlying mean-field process. This is known as Mckean-Vlasov process. Using Theorem 8 with $T = 1$, we obtain that

$$\sup_{i\in[n]} \mathbb{E}\left(\sup_{0\leq t\leq 1} |X^i_t(a) - \tilde{X}^i_t(a)|^2\right) \leq C'n^{-1},$$

where $\tilde{X}^i_t(a)$ is the solution of the SDE

$$d\tilde{X}^i_t(a) = \frac{\|a\|}{\sqrt{2d}} \exp(t/4) dB^i_t, \tilde{X}^i_0(a) = X^i_0(a).$$

This is a special SDE since all the marginal distributions are zero-mean Gaussians (sum of Brownian increments) with variance $\mathbb{E}\tilde{X}^i_t(a)^2 = \frac{\|a\|^2}{d}\exp(t/2)$.

In particular, $X^i(a)_t$ converges weakly to $\tilde{X}^i_t(a)$ in the limit of infinite width n. Combining the bound in Part 1 with the Mckean-Vlasov bound above, we obtain

$$\sup_{t\in[0,1]} \mathcal{W}_1(\mu^t_{n,L}(a), \mu^t_{\infty,\infty}(a)) \leq C\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{L}}\right)$$

for some constant that depends only on d and $\|a\|$, and where $\mu^t_{\infty,\infty}(a)$ is the distribution of $\tilde{X}^i_t(a) \sim \mathcal{N}(0, d^{-1}\|a\|^2 \exp(t/2))$.

□

E. Proof of Theorem 2

Theorem 9 (Neural Covariance)**.** *Let $a, b \in \mathbb{R}^d$ such that $a, b \neq 0$ and $a \neq b$. Then, we have the following*

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

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

$$\begin{aligned} \frac{dq_t(a,b)}{dt} &= \frac{1}{2}\frac{f(c_t(a,b))}{c_t(a,b)} q_t(a,b), \\ c_t(a,b) &= \frac{q_t(a,b)}{\sqrt{q_t(a,a)}\sqrt{q_t(b,b)}}, \\ q_0(a,b) &= \frac{\langle a, b\rangle}{d}, \end{aligned} \tag{9}$$

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

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

Proof. Let $a, b \in \mathbb{R}^d$ and q_t be as in the statement of the theorem. Let $X_t(a)$ and $X_t(b)$ be the infinite-depth limits as in Proposition 5, and let $\bar{X}(a), \bar{X}(b)$ be the corresponding Euler schemes. Using the fact that $(Y_{\lfloor tL \rfloor}(a), Y_{\lfloor tL \rfloor}(b))$ has the same law as $(\bar{X}_{\lfloor tL \rfloor}(a), \bar{X}_{\lfloor tL \rfloor}(b))$, we trivially have

$$\mathbb{E}\left| \frac{\langle Y_{\lfloor tL \rfloor}(a), Y_{\lfloor tL \rfloor}(b)\rangle}{n} - q_t(a,b) \right|^2 = \mathbb{E}\left| \frac{\langle \bar{X}_{\lfloor tL \rfloor}(a), \bar{X}_{\lfloor tL \rfloor}(b)\rangle}{n} - q_t(a,b) \right|^2.$$

We have the following upperbound

$$\begin{aligned} \left\| \frac{\langle \bar{X}_{\lfloor tL \rfloor}(a), \bar{X}_{\lfloor tL \rfloor}(b)\rangle}{n} - q_t(a,b) \right\|_{L_2} \leq & \left\| \frac{\langle \bar{X}_{\lfloor tL \rfloor}(a), \bar{X}_{\lfloor tL \rfloor}(b)\rangle}{n} - \frac{\langle X_t(a), X_t(b)\rangle}{n} \right\|_{L_2} \\ &+ \left\| \frac{\langle X_t(a), X_t(b)\rangle}{n} - \frac{\langle \tilde{X}_t(a), \tilde{X}_t(b)\rangle}{n} \right\|_{L_2} \\ &+ \left\| \frac{\langle \tilde{X}_t(a), \tilde{X}_t(b)\rangle}{n} - q_t(a,b) \right\|_{L_2}, \end{aligned} \tag{10}$$

where $\tilde{X}(a), \tilde{X}(b)$ are the infinite-width limits of the processes $X(a), X(b)$ as in Theorem 8. Let us deal with each term in this bound.

- First term: from Theorem 6 and standard upperbounds on the second moments (Theorem 4), we have that

$$\left\| \frac{\langle \bar{X}_{\lfloor tL \rfloor}(a), \bar{X}_{\lfloor tL \rfloor}(b)\rangle}{n} - \frac{\langle X_t(a), X_t(b)\rangle}{n} \right\|_{L_2} \leq C_1 L^{-1/2},$$

  where C_1 is a constant that depends only on $\|a\|$, $\|b\|$, and d.

- Second term: from Theorem 8, there exists a constant C_2 such that

$$\left\| \frac{\langle X_t(a), X_t(b)\rangle}{n} - \frac{\langle \tilde{X}_t(a), \tilde{X}_t(b)\rangle}{n} \right\|_{L_2} \leq C_2 n^{-1/2},$$

  where C_2 depends only on $\|a\|$, $\|b\|$, and d.

- Third term: from Proposition 5, we know that $\lim_{n\to\infty} \mathbb{E}\left[\frac{\langle X_t(a), X_t(b)\rangle}{n}\right] = q_t(a,b)$. Using the bound above on the second term, we obtain that $\lim_{n\to\infty} \mathbb{E}\left[\frac{\langle \tilde{X}_t(a), \tilde{X}_t(b)\rangle}{n}\right] = q_t(a,b)$. Now the key observation is that
$$\frac{\langle \tilde{X}_t(a), \tilde{X}_t(b)\rangle}{n} = \frac{1}{n}\sum_{i=1}^{n} \tilde{X}_t^i(a)\tilde{X}_t^i(b),$$
and the terms in the sum above are iid with mean $q_t(a,b)$. Therefore,
$$\left\|\frac{\langle \tilde{X}_t(a), \tilde{X}_t(b)\rangle}{n} - q_t(a,b)\right\|_{L_2} = \left(\mathbb{E}\left|\frac{\langle \tilde{X}_t(a), \tilde{X}_t(b)\rangle}{n} - q_t(a,b)\right|^2\right)^{1/2} \leq (\mathbb{E}(\tilde{X}_t^1(a)\tilde{X}_t^1(b))^2)^{1/2}\, n^{-1/2}.$$
Observe that $\mathbb{E}(\tilde{X}_t^1(a)\tilde{X}_t^1(b))^2$ can be bounded with a constant C_3 depends only on $\|a\|$, $\|b\|$, and d.

We conclude by combining the three bounds above. □

F. Proof of Theorem 3

Theorem 3. [Neural correlation]
Under the same conditions of Theorem 2, we have the following

$$\sup_{t\in[0,1]} \|\hat{c}_t(a,b) - c_t(a,b)\|_{L_2} \leq C'\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{L}}\right)$$

where C' is a constant that depends only on $\|a\|$, $\|b\|$, and d, and $\hat{c}_t(a,b) = \frac{\langle Y_{\lfloor tL\rfloor}(a), Y_{\lfloor tL\rfloor}(b)\rangle}{\|Y_{\lfloor tL\rfloor}(a)\|\|Y_{\lfloor tL\rfloor}(b)\|}$ is the neural correlation kernel, and $c_t(a,b)$ is defined in Theorem 2.

Proof. Let a and b be as stated in the theorem. We have the following

$$\|\hat{c}_t(a,b) - c_t(a,b)\|_{L_2} \leq \left\|\frac{\hat{q}_t(a,b)}{\sqrt{\hat{q}_t(a,a)\hat{q}_t(b,b)}} - \frac{q_t(a,b)}{\sqrt{\hat{q}_t(a,a)\hat{q}_t(b,b)}}\right\|_{L_2} + \left\|\frac{q_t(a,b)}{\sqrt{\hat{q}_t(a,a)\hat{q}_t(b,b)}} - \frac{q_t(a,b)}{\sqrt{q_t(a,a)q_t(b,b)}}\right\|_{L_2}.$$

Using Markov's inequality along with Theorem 2, it is straightforward that there exists a constant C_1 that depends only on $\|a\|$, $\|b\|$, and d such that
$$\mathbb{P}\left(\hat{q}_t(a,a) < \frac{q_t(a,a)}{2}\right) \leq C_1 \min(n,L)^{-1},$$
and
$$\mathbb{P}\left(\hat{q}_t(b,b) < \frac{q_t(b,b)}{2}\right) \leq C_1 \min(n,L)^{-1}.$$

Let A denote the event $\{\hat{q}_t(a,a) \geq \frac{q_t(a,a)}{2}\} \cup \{\hat{q}_t(b,b) \geq \frac{q_t(b,b)}{2}\}$. With this, we obtain the following upperbound

$$\begin{aligned}\left\|\frac{\hat{q}_t(a,b)}{\sqrt{\hat{q}_t(a,a)\hat{q}_t(b,b)}} - \frac{q_t(a,b)}{\sqrt{\hat{q}_t(a,a)\hat{q}_t(b,b)}}\right\|_{L_2} &\leq \left\|\mathbb{1}_A\left(\frac{\hat{q}_t(a,b)}{\sqrt{\hat{q}_t(a,a)\hat{q}_t(b,b)}} - \frac{q_t(a,b)}{\sqrt{\hat{q}_t(a,a)\hat{q}_t(b,b)}}\right)\right\|_{L_2} \\ &+ \left\|\mathbb{1}_{A^c}\left(\frac{\hat{q}_t(a,b)}{\sqrt{\hat{q}_t(a,a)\hat{q}_t(b,b)}} - \frac{q_t(a,b)}{\sqrt{\hat{q}_t(a,a)\hat{q}_t(b,b)}}\right)\right\|_{L_2} \\ &\leq \frac{2}{\sqrt{q_t(a,a)q_t(b,b)}}\|\hat{q}_t(a,b) - q_t(a,b)\|_{L_2} + C_2\mathbb{P}(A^c)^{1/2},\end{aligned}$$

where A^c denote the complementary event of A and C_2 is a constant that depends only on $\|a\|$, $\|b\|$, and d. From Theorem 2 and the Markov inequality bound, we can upperbound this term by a term of order $\min(n,L)^{-1/2}$ with a constant that depends only on $\|a\|$, $\|b\|$ and d. A similar treatment of the second term yields the desired result. □

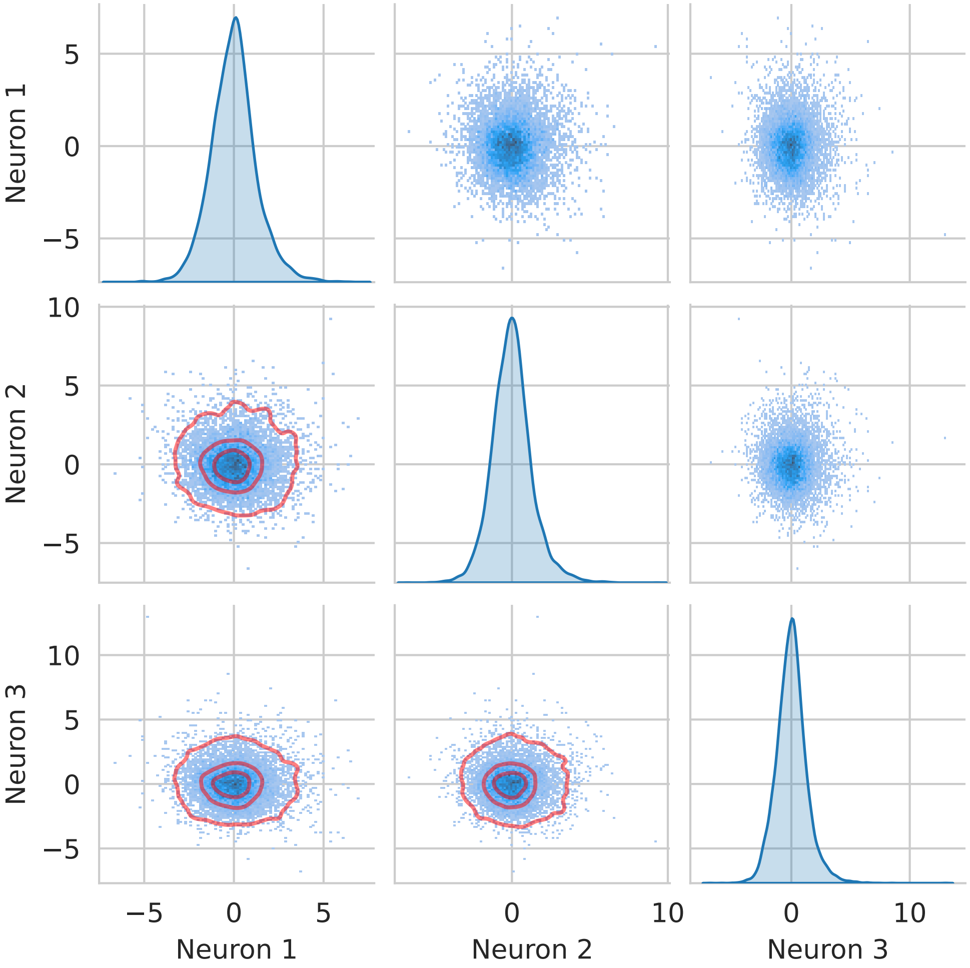


Figure 6: Same plot as Fig. 3 with $n = L = 5$

G. Additional experiments

G.1. Pairplots

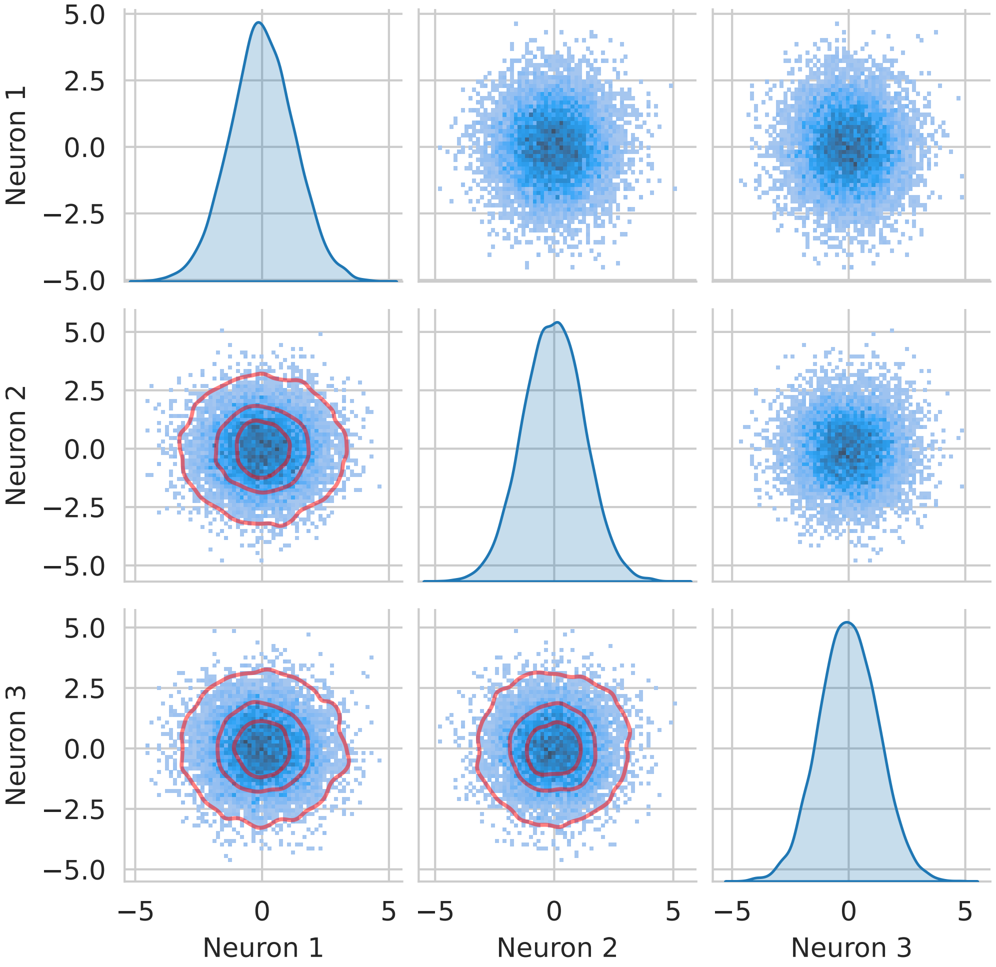


Figure 7: Same plot as Fig. 3 with $n = 100$ and $L = 5$

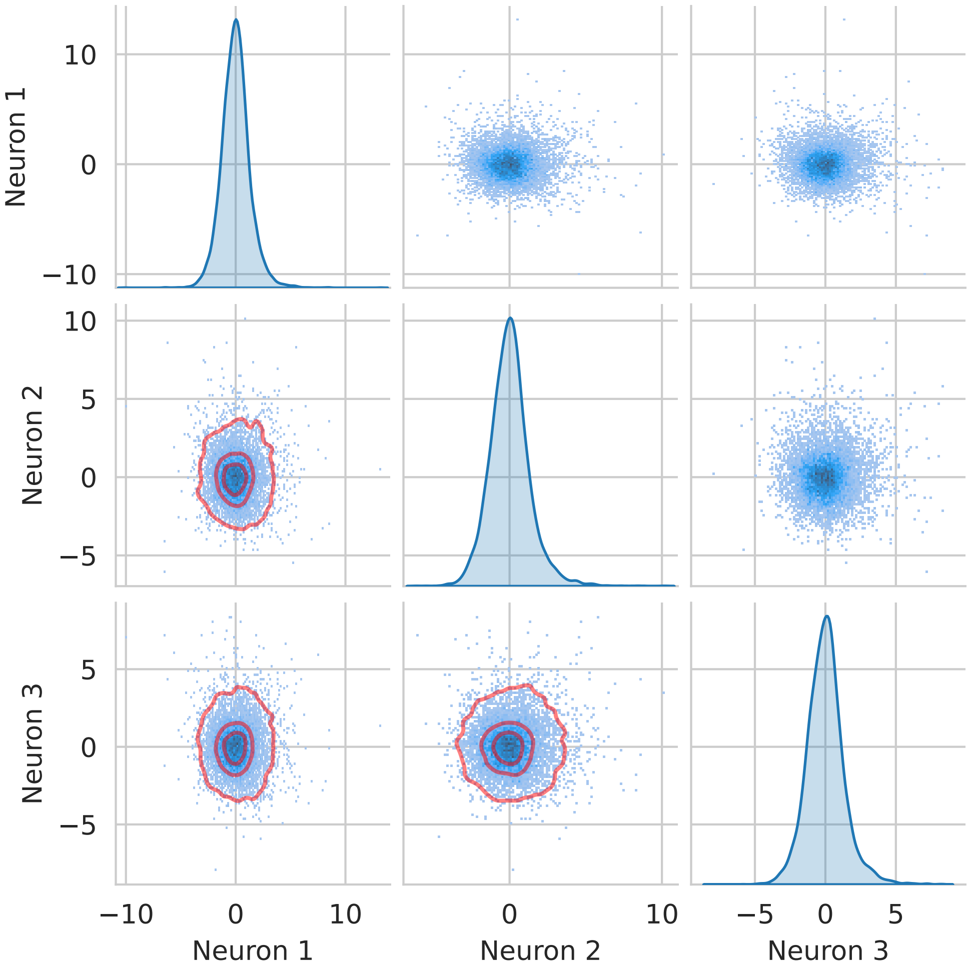


Figure 8: Same plot as Fig. 3 with $n = 5$ and $L = 100$