The Neural Covariance SDE: Shaped Infinite Depth-and-Width Networks at Initialization

Anonymous Author(s)
Affiliation
Address
email

Abstract

The logit outputs of a feedforward neural network at initialization are conditionally Gaussian, given a random covariance matrix defined by the penultimate layer. In this work, we study the distribution of this random matrix. Recent work has shown that shaping the activation function as network depth grows large is necessary for this covariance matrix to be non-degenerate. However, the current infinite-width-style understanding of this shaping method is unsatisfactory for large depth: infinite-width analyses ignore the microscopic fluctuations from layer to layer, but these fluctuations accumulate over many layers.

To overcome this shortcoming, we study the random covariance matrix in the shaped infinite-depth-and-width limit. We identify the precise scaling of the activation function necessary to arrive at a non-trivial limit, and show that the random covariance matrix is governed by a stochastic differential equation (SDE) that we call the Neural Covariance SDE. Using simulations, we show that the SDE closely matches the distribution of the random covariance matrix of finite networks. Additionally, we recover an if-and-only-if condition for exploding and vanishing norms of large shaped networks based on the activation function.

1 Introduction

Of the many milestones in deep learning theory, the precise characterization of the infinite-width limit of neural networks at initialization as a Gaussian process with a non-random covariance matrix [1, 2] was a turning point. The so-called Neural Network Gaussian process (NNGP) theory laid the mathematical foundation to study various limiting training dynamics under gradient descent [3–12]. The Neural Tangent Kernel (NTK) limit formed the foundation for a rush of theoretical work, including advances in our understanding of generalization for wide networks [13–15]. Besides the NTK limit, the infinite-width mean-field limit was developed, showing benefits for feature learning and hyperparameter tuning [16–22].

Fundamentally, the infinite-width paradigm derives results from the assumption that the depth of the network is held *fixed* while the widths of all layers grow to infinity. Unfortunately, this assumption can be problematic for modeling real-world networks, as the microscopic fluctuations from layer to layer are neglected in this limit (see Figure 1). In particular, infinite-width predictions are shown to be poor approximations of real network behaviour unless the depth is much less than the width [23, 24].

Impressive achievements of deep networks with billions of parameters crystallize the importance of understanding extremely large, deep neural networks (DNNs). An alternative to the infinite-width paradigm is the infinite-depth-and-width paradigm. In this setting, both the network depth d and the width n of each layer are simultaneously scaled to infinity, while their relative ratio d/n remains fixed [23, 25–29]. Recent work also explores using d/n as an effective perturbation parameter [30–32]

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

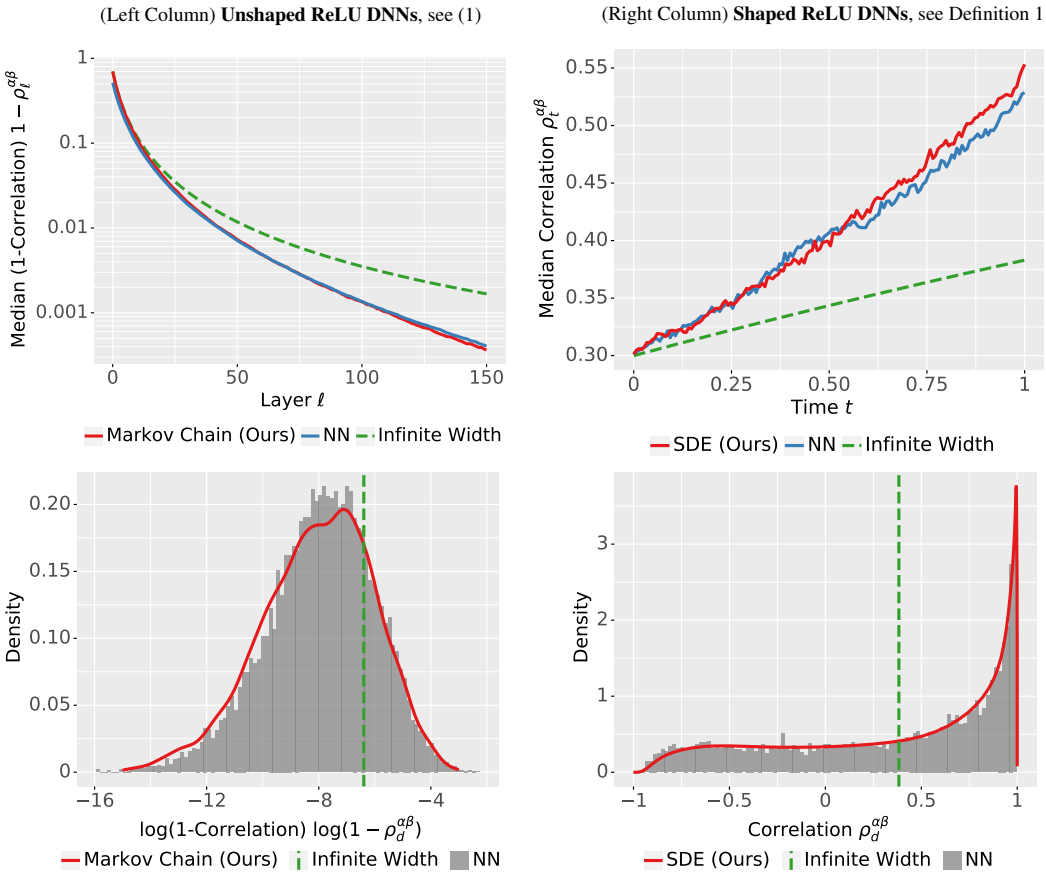

Figure 1: Simulations of correlation $\rho_\ell^{\alpha\beta} = \langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle / |\varphi_\ell^\alpha||\varphi_\ell^\beta|$ between post-activation vectors in ReLU networks, comparing *finite NNs* vs. our theoretical predictions vs. infinite-width paradigm. **Left Column:** $\rho_\ell^{\alpha\beta}$ vs. our Markov chain (10) vs. infinite-width update $\rho_{\ell+1} = cK_1(\rho_\ell)$ (see (10) and note the log scale and $1-\rho$ here). **Right Column:** $\rho_{\lfloor tn \rfloor}^{\alpha\beta}$ vs. our Neural Covariance SDE vs. ODE $\rho_t = \nu(\rho_t)$ as in Theorem 3. **Top Row:** *Median* ρ as a function of layer. **Bottom Row:** *Full distribution* at final layer $\ell = d$. **Simulation details:** $n = d = 150$, $\rho_0 = 0.3$, 2^{13} samples for each. In right column: $c_+ = 0, c_- = -1$, DE step size $1\mathrm{e}{-2}$. Density curves are kernel density estimation.

or to study concentration bounds in terms of d/n [5, 33]. This limit has the distinct advantage of being incredibly accurate at predicting the output distribution for finite size networks at initialization [27] — a significant improvement over the NNGP theory. Considering the mathematical success of the NNGP techniques, the infinite-depth-and-width limit hints at the possibility of developing an accurate theory for training and generalization.

An immediate issue of the infinite-depth limit is that this limit predicts that network output becomes degenerate as depth increases: on initialization the network becomes a constant function sending all inputs to the same (random) output [34, 35, 32]. While degenerate outputs are not necessarily an issue in theory, it poses a more serious problem in practice: degenerate correlations imply a "sharp" input–output Jacobian, and therefore exploding gradients [36, 25]. Intuitively, the output is not very sensitive to changes in the input, hence the gradient must be very large in the earlier layers.

Recently [37, 38] proposed a solution by modifying the activation function to reduce to the effect of degeneracy. The authors designed several criteria including reducing estimated output correlation, and numerically optimized the shape of activation functions for improved training results. However, their deterministic estimation of output correlation using the infinite-width limit leads to a poor approximation of real networks, as the additional randomness has both non-zero mean and heavy skew (see Figure 1 right column). Furthermore, numerically searching for the activation shape obscures the picture on how shaping should depend on the network depth and width.

Table 1: Notation

Notation	Description	Notation	Description
$n_{\text{in}} \in \mathbb{N}$	Input dimension	$n_{\text{out}} \in \mathbb{N}$	Output dimension
$n \in \mathbb{N}$	Hidden layer width	$d \in \mathbb{N}$	Number of hidden layers (depth)
$\varphi(\cdot)$	Base activation	$\varphi_s(\cdot)$	Shaped activation
$x^\alpha \in \mathbb{R}^{n_{\text{in}}}$	Input for $1 \leq \alpha \leq m$	$W_0 \in \mathbb{R}^{n_{\text{in}} \times n}$	Weight matrix at layer 0
$z^\alpha_{\text{out}} \in \mathbb{R}^{n_{\text{out}}}$	Network output	$W_{\text{out}} \in \mathbb{R}^{n \times n_{\text{out}}}$	Weight matrix at final layer
$z^\alpha_\ell \in \mathbb{R}^n$	Neurons (pre-activation) for layer $1 \leq \ell \leq d$	$W_\ell \in \mathbb{R}^{n \times n}$	Weight matrix at layer $1 \leq \ell \leq d$ **All weights initialized iid** $\sim \mathcal{N}(0,1)$
$\varphi^\alpha_\ell \in \mathbb{R}^n$	Neurons (post-activation) for layer $1 \leq \ell \leq d$	$c \in \mathbb{R}$	Normalizing constant $c := \left(\mathbb{E}\,\varphi(g)^2\right)^{-1}$ for $g \sim \mathcal{N}(0,1)$

In this paper, we seek to address these problems by providing a precise theory of shaped infinite-depth-and-width networks, extending both the NNGP theories and the activation shaping techniques. In particular, we prescribe an exact scaling of the activation function shape as a function of network width n that leads to a non-trivial limit. By keeping track of microscopic $O(n^{-1/2})$ random fluctuations in each layer of the network, we show that the cumulative effect is described by a stochastic differential equation (SDE) in the limit. In contrast to existing infinite-width theory, we are able to characterize the random distribution of the output covariance and correlation, which matches matches closely to simulations of real networks. In a similar spirit to how the NNGP theory laid the foundation for studying training and generalization in the infinite-width limit, we also see this work as building the mathematical tools for an infinite-depth-and-width theory of training and generalization.

1.1 Contributions

Similar to the NNGP approach, we use the fact that the output is Gaussian conditional on the penultimate layer. However, unlike in the infinite-width paradigm, the covariance matrix is no longer deterministic in the infinite-depth-and-width limit. Our focus in this paper is to study this random covariance matrix. Our main contributions are as follows:

1. We introduce the tool of stochastic $\sqrt{n}$-expansions and convergence to SDEs for analyzing the distribution of covariances in DNNs.
2. For *unshaped* ReLU-like activations, we show that the norm of each layer evolves according to geometric Brownian motion and correlations evolve according a discrete Markov process. See left column of Figure 1 and Section 2.
3. For both ReLU-like and a large class of smooth activation functions, we derive the Neural Covariance SDE characterizing the distribution of the shaped infinite-depth-and-width limit. See right column of Figure 1 and Section 3.
4. We show our prescribed shape scaling is exact, as other rates of scaling leads to either degenerate or linear network limits. See Proposition 4 and Proposition 10.
5. For smooth activations, we derive an if-and-only-if condition for exploding/vanishing norms based on properties of the activation function. See Proposition 7 and Section 4.
6. We provide simulations to verify theoretical predictions and help interpret properties of real DNNs. See Figures 1 and 2 and supplemental simulations in Appendix E.

2 Limits for Unshaped ReLU-Like Activations

Using the **notation in Table 1**, the output of a fully connected feedforward network with d hidden layers of width n on input x^α is defined by vectors of **pre-activations** z^α_ℓ and **post-activations** φ^α_ℓ:

$$z^\alpha_1 := \frac{1}{\sqrt{n_{\text{in}}}} W_0 x^\alpha, \quad \varphi^\alpha_\ell := \varphi(z^\alpha_\ell), \quad z^\alpha_{\ell+1} := \sqrt{\frac{c}{n}} W_\ell \varphi^\alpha_\ell, \quad z^\alpha_{\text{out}} := \sqrt{\frac{c}{n}} W_{\text{out}} \varphi^\alpha_d \,. \tag{1}$$

Note that factors of $\sqrt{cn^{-1}}$ are equivalent to intializing according to the so-called He initialization [39]. We use Greek indices $\alpha, \beta, \ldots$ to denote multiple different inputs. Note that while our results are all stated for fixed width n in each layer, they can be generalized to layer width n_ℓ in the limit where all $n_\ell \to \infty$ with $\sum_{\ell=1}^d {n_\ell}^{-1}$ replacing the role of the depth-to-width ratio d/n [25].

In this section, we analyze **ReLU-like** activations by which we mean activations which are linear on the negative and positive numbers given respectively by two slopes s_+ and s_-:

$$\varphi(x) \coloneqq s_+ \max(x,0) + s_- \min(x,0) = s_+ \varphi_{\text{ReLU}}(x) - s_- \varphi_{\text{ReLU}}(-x)\,. \tag{2}$$

These are precisely the **positive homogeneous** functions: $\varphi(ax) = |a|\,\varphi(x)\,\forall x, a \in \mathbb{R}$.

2.1 SDE Limits of Markov chains

We briefly review the main type of SDE convergence principle used in our main results (see Proposition 18 for a more precise version). Let X_t, $t \in \mathbb{R}^+$, be a continuous time diffusion process obeying an SDE with drift b and variance σ^2 as given in (3). Suppose that for each $n \in \mathbb{N}$, Y_ℓ^n is a discrete time Markov chain $\ell \in \mathbb{N}$ whose increments obey (3) in terms of the same functions b, σ^2:

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t\,, \qquad Y_{\ell+1}^n - Y_\ell^n = b(Y_\ell)\frac{1}{n} + \sigma(Y_\ell)\frac{\xi_\ell}{\sqrt{n}} + O(n^{-3/2}), \tag{3}$$

where ξ_ℓ are independent variables with $\mathbb{E}(\xi) = 0, \mathbf{Var}(\xi) = 1$. With this setup, under technical conditions described precisely in Appendix A, we have convergence of Y_ℓ at $\ell = \lfloor tn \rfloor$ to X_t, namely: $Y_{\lfloor tn \rfloor}^n \to X_t$ as $n \to \infty$ in the Skorohod topology. In our applications, n is always the width (i.e., number of neurons in each layer) which may appear implicitly and ℓ is always the layer number.

2.2 A Simple SDE: Geometric Brownian Motion Describes $|\varphi_\ell^\alpha|^2$

To motivate our approach of SDE limits, we illustrate the method using the example of, the squared norm of the ℓ-th layer, $|\varphi_\ell^\alpha|^2$, where we recall $\varphi_\ell^\alpha = \varphi(z_\ell^\alpha)$. For a single fixed input x^α and a ReLU-like activation φ, the norm of the post-activation neurons $|\varphi_\ell^\alpha|^2$ forms a Markov chain in the layer number ℓ. We use the fact that a Gaussian random matrix applied to any unit vector gives a Gaussian vector of iid $\mathcal{N}(0,1)$ entries. Hence, in each layer, we can define the Gaussian vector g^α as follows, and use (1) with the positive homogeneity of φ to write the Markov chain update rule:

$$g^\alpha \coloneqq W_\ell \frac{\varphi_\ell^\alpha}{|\varphi_\ell^\alpha|} \overset{d}{=} \mathcal{N}(0, I_n), \;\; \left|\varphi_{\ell+1}^\alpha\right|^2 = |\varphi_\ell^\alpha|^2 \frac{1}{n}\sum_{i=1}^n c\varphi(g_i^\alpha)^2\,. \tag{4}$$

At this point, the infinite-width approach applies the law of large numbers (LLN) to conclude $\lim_{n\to\infty} \left|\varphi_{\ell+1}^\alpha\right|^2 = |\varphi_\ell^\alpha|^2\,\mathbb{E}[c\varphi^2(g)] = |\varphi_\ell^\alpha|^2 \cdot 1$ a.s. by definition of c. However, the LLN cannot be applied when depth d is diverging with n, as the cumulative effect of the fluctuations over d layers does not vanish! Instead, we keep track of the $O(1/\sqrt{n})$ fluctuations in each layer by introducing the zero mean finite variance random variable $R_\ell^{\alpha\alpha} \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(c\varphi(g_i^\alpha)^2 - 1\right)$. This allows us to rewrite this Markov chain update rule as

$$\left|\varphi_{\ell+1}^\alpha\right|^2 = |\varphi_\ell^\alpha|^2 \left(1 + \frac{1}{\sqrt{n}} R_\ell^{\alpha\alpha}\right)\,, \tag{5}$$

which allows us to see that the Markov chain $Y_\ell^n = \frac{c}{n}|\varphi_\ell^\alpha|^2$ is now in the form of (3) with $b(Y) \equiv 0, \sigma^2(Y) = \mathbf{Var}(R_\ell^{\alpha\alpha})Y^2 = \mathbf{Var}(c\varphi(g)^2)Y^2$. Consequently, we have that the squared norm Markov chain converges to a geometric Brownian motion $dX_t = \sigma X_t dB_t$, or more precisely

$$\lim_{n\to\infty} \frac{c}{n}\left|\varphi_{\lfloor tn \rfloor}^\alpha\right|^2 = X_t \overset{d}{=} e^{\mathcal{N}(-\frac{\sigma^2}{2}t, \sigma^2 t)}\,, \tag{6}$$

where the convergence is in the Skorohod topology (see Appendix A). When φ is the ReLU function ($s_+ = 1, s_- = 0$), we have $c = 2$ and $\sigma^2 = 5$, which recovers known results in [25, 27–29]. We remark again this simple Markov chain example illustrates the main technique we use in later sections to establish SDE convergence for shaped networks in Section 3.

2.3 Non-SDE Markov Chains: the Gram Matrix $\left\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \right\rangle$ and Correlation $\rho_\ell^{\alpha\beta} = \frac{\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle}{|\varphi_\ell^\alpha|\,|\varphi_\ell^\beta|}$

We can generalize Section 2.2 to a collection of m inputs $\{x^\alpha\}_{\alpha=1}^m$ by looking at the entire Gram matrix $[\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle]_{\alpha,\beta=1}^m$, where we again recall $\varphi_\ell^\alpha = \varphi(z_\ell^\alpha)$. The Gram matrix is of particular interest

because the neurons in any layer are conditionally Gaussian **when conditioned on the previous layer**, with covariance matrix proportional to the Gram matrix:

$$[z_{\ell+1}^{\alpha}]_{\alpha=1}^{m} \stackrel{d}{=} \mathcal{N}\left(0, \frac{c}{n}[\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle]_{\alpha,\beta=1}^{m} \otimes I_n\right), [z_{\text{out}}^{\alpha}]_{\alpha=1}^{m} \stackrel{d}{=} \mathcal{N}\left(0, \frac{c}{n}[\langle \varphi_d^\alpha, \varphi_d^\beta \rangle]_{\alpha,\beta=1}^{m} \otimes I_{n_{\text{out}}}\right), \tag{7}$$

where $\otimes$ denotes the Kronecker product (here indicating conditionally independent entries in each vector). If we define g^α as in (4), we see that the g^α are all marginally $\mathcal{N}(0, I_n)$ and now with correlations between inputs $\rho_\ell^{\alpha\beta} \coloneqq \mathbb{E}[g_i^\alpha g_i^\beta] = \langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle / |\varphi_\ell^\alpha||\varphi_\ell^\beta| = \cos(\theta_\ell^{\alpha\beta})$, where $\theta_\ell^{\alpha\beta}$ is the angle between φ_ℓ^α and φ_ℓ^β. Similar to (4), we can write the update rule for the α, β-entry of the Gram matrix:

$$\left\langle \varphi_{\ell+1}^\alpha, \varphi_{\ell+1}^\beta \right\rangle = |\varphi_\ell^\alpha||\varphi_\ell^\beta| \frac{1}{n_\ell} \sum_{i=1}^{n} c\varphi(g_i^\alpha)\varphi(g_i^\beta) \tag{8}$$

Just as we did in (5), we can define $R_\ell^{\alpha\beta} \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} c\varphi(g_i^\alpha)\varphi(g_i^\beta) - \mathbb{E}[c\varphi(g_i^\alpha)\varphi(g_i^\beta)]$ and write

$$\left\langle \varphi_{\ell+1}^\alpha, \varphi_{\ell+1}^\beta \right\rangle = |\varphi_\ell^\alpha||\varphi_\ell^\beta| \left(\mathbb{E}\left[c\varphi(g_i^\alpha)\varphi(g_i^\beta)\right] + \frac{1}{\sqrt{n}} R_\ell^{\alpha\beta} \right) \tag{9}$$

where $R_\ell^{\alpha\beta}$ are mean zero with covariance $\mathbf{Cov}[R_\ell^{\alpha\beta}, R_\ell^{\gamma\delta}] = \mathbf{Cov}[c\varphi(g^\alpha)\varphi(g^\beta), c\varphi(g^\gamma)\varphi(g^\delta)]$. (By the Central Limit Theorem, $R_\ell^{\alpha\beta}$ will be approximately Gaussian for large n.)

Unlike the simple single-data-point case from Section 2.2, we do not have convergence to a continuous SDE because the differences $\left\langle \varphi_{\ell+1}^\alpha, \varphi_{\ell+1}^\beta \right\rangle - \left\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \right\rangle \nrightarrow 0$ as $n \to \infty$. Instead, (9) is a discrete recursion update with additive noise of the form $Y_{\ell+1}^n = f(Y_\ell^n) + \frac{1}{\sqrt{n}}\xi$ for some update function f, which does not vanish as $n \to \infty$. In Section 3, we will see that by shaping the activation function we can arrange so that the update function rescales to an approximate identity $f(x) = x + \frac{1}{n}\nu(x)$ (see also Lemma 20), recasting this update into the form of the Markov chain in (3), which will then lead to convergence to an SDE.

One quantity that can be isolated from the Gram matrix is the correlation between any pair of inputs. For such a pair, x^α, x^β, this is written in terms of the entries of the 2×2 Gram matrix: $\rho_\ell^{\alpha\beta} = \left\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \right\rangle / |\varphi_\ell^\alpha||\varphi_\ell^\beta|$. After some manipulation, one can see that $\rho_\ell^{\alpha\beta}$ on its own evolves as a Markov process in ℓ and we obtain the following approximate Markov chain evolution $\rho_\ell^{\alpha\beta}$ (see Appendix B.2 and (77) for details):

$$\rho_{\ell+1}^{\alpha\beta} = cK_1(\rho_\ell^{\alpha\beta}) + \frac{1}{n}\mu_{\text{ReLU}}(\rho_\ell^{\alpha\beta}) + \frac{\xi_\ell}{\sqrt{n}}\sigma_{\text{ReLU}}(\rho_\ell^{\alpha\beta}) + O(n^{-3/2}), \tag{10}$$

where $K_1(\rho) \coloneqq \mathbb{E}\left[\varphi(g)\varphi(g\rho + w\sqrt{1-\rho^2})\right]$ for g, w iid $\mathcal{N}(0,1)$ random variables, and $\mu_{\text{ReLU}}, \sigma_{\text{ReLU}}$ can be found through a series expansions in $1/\sqrt{n}$ for $\rho_{\ell+1}^{\alpha\beta}$ from (9) (see Appendix D for explicit formulas). For the ReLU case, $c = 2$ and $cK_1(\rho) = 1 + (\sqrt{1-\rho^2} - \rho\arccos(\rho))/\pi$ was first calculated in [40].

The infinite-width method of understanding (10) is to discard the terms that vanish as $n \to \infty$ and consider only the update $\rho_{\ell+1}^{\alpha\beta} = cK_1(\rho_\ell^{\alpha\beta})$. (Indeed, deriving this update is much simpler than the calculations needed to determine μ_{ReLU} and σ_{ReLU}.) Analysis of this deterministic equation leads to the prediction that $\rho_\ell^{\alpha\beta} = 1 - O(\ell^{-2})$ for $\ell \gg 1$ (see (4.8) in [32]). Using recent results about the classical logistic map [41], we derive in Appendix D a bound on this recursion, which empirically seems to be quite sharp.

In fact, simulations show that the infinite-width prediction $\rho_{\ell+1}^{\alpha\beta} = cK_1(\rho_\ell^{\alpha\beta})$ does not match real networks when $d/n \gg 0$ (see left column of Figure 1). In this case, the microscopic $O(n^{-1})$ and $O(n^{-1/2})$ terms in (10) accumulate to macroscopic differences! Empirically, we see their net effect is that $\rho_\ell^{\alpha\beta} \to 1$ **faster** than the infinite-width prediction. Heuristically speaking, the reason for this discrepancy is due to $\sigma_{\text{ReLU}}(\rho) \to 0$ as $\rho \to 1$: this means that the randomness can push $\rho_\ell^{\alpha\beta}$ closer to 1, but becomes "trapped" when $\rho_\ell^{\alpha\beta}$ is close to 1 because σ_{ReLU} is so small here. In the next section, we will also see that shaping the activation induces a limiting correlation SDE.

3 Neural Covariance SDEs: Shaped Infinite-Depth-and-Width Limit

In this section, we follow the ideas of [37, 38] to **reshape** the activation function φ. Reshaping means to replace the base activation function φ in (1) with φ_s that depends on width n. Specifically, we will choose φ_s to depend on n such that in the limit as $n \to \infty$, we have that φ is approximately an identity function, $\varphi_s \to \text{Id}$. Recalling from (7) that the output is conditionally Gaussian with covariance determined by the Gram matrix $[\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle]_{\alpha,\beta=1}^m$, therefore it is sufficient to characterize this random covariance matrix. All the proofs for results in this section can be found in Appendix B.

3.1 Neural Covariance SDE for Shaped ReLU-Like Activations

Definition 1. *We shape the ReLU-like activation $\varphi_s(x) \coloneqq s_+ \max(x, 0) + s_- \min(x, 0)$, by setting the slopes to depend on n according to $s_\pm \coloneqq 1 + \frac{c_\pm}{\sqrt{n}}$ for some given constants $c_+, c_- \in \mathbb{R}$.*

We will show that with shaping of Definition 1, one gets non-trivial SDEs that describe the covariance (Theorem 2) and correlations (Theorem 3) of the network. The precise scaling is shown to be the critical scaling for a non-trivial limit in Proposition 4.

Remark. Note that in the statement of our theorems, we abuse notation and use the same letter to denote the pre-limit Markov chain and the limiting SDE. For example, in Theorem 2 we use V_ℓ for the covariance at layer ℓ and V_t to denote the limiting SDE at time t.

Theorem 2 (Covariance SDE, ReLU)**.** *Let $V_\ell^{\alpha\beta} \coloneqq \frac{c}{n}\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle$, and define $V_\ell \coloneqq [V_\ell^{\alpha\beta}]_{1\le\alpha\le\beta=m}$ to be the upper triangular entries thought of as a vector in $\mathbb{R}^{m(m+1)/2}$ Then, with $s_\pm = 1 + \frac{c_\pm}{\sqrt{n}}$ as in Definition 1, in the limit as $n \to \infty$, $\frac{d}{n} \to T$, the interpolated process $V_{\lfloor tn \rfloor}$ converges in distribution to the solution of the following SDE in the Skorohod topology of $D_{\mathbb{R}_+,\mathbb{R}^{m(m+1)/2}}$*

$$dV_t = b(V_t)\,dt + \Sigma(V_t)^{1/2}\,dB_t\,, \quad V_0 = \left[\frac{1}{n_{in}}\langle x^\alpha, x^\beta \rangle\right]_{1\le\alpha\le\beta\le m}\,, \tag{11}$$

where we denote $\nu(\rho) \coloneqq \frac{(c_+-c_-)^2}{2\pi}\left(\sqrt{1-\rho^2} - \rho\arccos\rho\right)$, $\rho_t^{\alpha\beta} \coloneqq \frac{V_t^{\alpha\beta}}{\sqrt{V_t^{\alpha\alpha}V_t^{\beta\beta}}}$ and write

$$b(V_t) = \left[\nu\left(\rho_t^{\alpha\beta}\right)\sqrt{V_t^{\alpha\alpha}V_t^{\beta\beta}}\right]_{1\le\alpha\le\beta\le m}\,, \quad \Sigma(V_t) = \left[V_t^{\alpha\gamma}V_t^{\beta\delta} + V_t^{\alpha\delta}V_t^{\beta\gamma}\right]_{\alpha\le\beta,\gamma\le\delta}\,. \tag{12}$$

Furthermore, the output distribution can be described conditional on V_T evaluated at final time T

$$[z_{out}^\alpha]_{\alpha=1}^m \,|V_T \overset{d}{=} \mathcal{N}\left(0, [V_T^{\alpha\beta}]_{\alpha,\beta=1}^m\right)\,. \tag{13}$$

Here we remark that $\nu(1) = 0$, and therefore the drift component of diagonal entries ($V_t^{\alpha\alpha}$) are zero, as they are geometric Brownian motion. For example, for two points, $m = 2$ we have the SDE:

$$d\begin{bmatrix} V_t^{\alpha\alpha} \\ V_t^{\beta\beta} \\ V_t^{\alpha\beta} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ b(V_t)_{\alpha\beta} \end{bmatrix} dt + \begin{bmatrix} 2(V_t^{\alpha\alpha})^2 & 2(V_t^{\alpha\beta})^2 & 2V_t^{\alpha\alpha}V_t^{\alpha\beta} \\ 2(V_t^{\alpha\beta})^2 & 2(V_t^{\beta\beta})^2 & 2V_t^{\beta\beta}V_t^{\alpha\beta} \\ 2V_t^{\alpha\alpha}V_t^{\alpha\beta} & 2V_t^{\beta\beta}V_t^{\alpha\beta} & V_t^{\alpha\alpha}V_t^{\beta\beta} + (V_t^{\alpha\beta})^2 \end{bmatrix}^{1/2} dB_t\,. \tag{14}$$

However, we emphasize that the m-point joint output distribution is *not* characterized by the marginal for each of the pairs, as the V_t at any fixed time t is *not* Gaussian. In particular, we observe the diffusion matrix entry corresponding to $V_t^{\alpha\beta}, V_t^{\gamma\delta}$ involves other processes $V_t^{\alpha\gamma}, V_t^{\beta\delta}, V_t^{\alpha\delta}, V_t^{\beta\gamma}$! That being said, it is still instructive to study the marginal for a pair of data points. More specifically, it turns out in the generalized ReLU case, we can derive the marginal SDE for the correlation process.

Theorem 3 (Correlation SDE, ReLU)**.** *Let $\rho_\ell^{\alpha\beta} \coloneqq \frac{\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle}{|\varphi_\ell^\alpha|\,|\varphi_\ell^\beta|}$, where $\varphi_\ell^\alpha \coloneqq \varphi_s(z_\ell^\alpha)$. In the limit as $n \to \infty$ and $s_\pm = 1 + \frac{c_\pm}{\sqrt{n}}$, the interpolated process $\rho_{\lfloor tn \rfloor}^{\alpha\beta}$ converges in distribution to the solution of the following SDE in the Skorohod topology of $D_{\mathbb{R}_+,\mathbb{R}}$*

$$d\rho_t^{\alpha\beta} = \left[\nu(\rho_t^{\alpha\beta}) + \mu(\rho_t^{\alpha\beta})\right]\,dt + \sigma(\rho_t^{\alpha\beta})\,dB_t\,, \quad \rho_0^{\alpha\beta} = \frac{\langle x^\alpha, x^\beta \rangle}{|x^\alpha|\,|x^\beta|}\,, \tag{15}$$

where

$$\nu(\rho) = \frac{(c_+ - c_-)^2}{2\pi}\left[\sqrt{1-\rho^2} - \arccos(\rho)\rho\right]\,, \quad \mu(\rho) = -\frac{1}{2}\rho(1-\rho^2)\,, \quad \sigma(\rho) = 1 - \rho^2\,. \tag{16}$$

To help interpret the SDE, we observe that μ and σ are entirely independent of the activation function. In other words, these terms will be present in this limit even for linear networks. At the same time, ν describes the influence of the shaped activation function in this limit. [38] has derived a related ordinary differential equation (ODE) of $\partial_t \rho_t = \nu(\rho_t)$ in the sequential limit of $n \to \infty$ then $d \to \infty$. Here we also note that $\nu(\rho)$ is closely related to the J_1 function derived in [40]. See Appendix D for the joint version of the correlation SDE, and Appendix E for empirical evidence of convergence.

One immediate consequence of the correlation SDE is that we can show the $n^{-1/2}$ scaling in Definition 1 is the only case where the limit is neither degenerate nor a linear network.

Proposition 4 (Critical Exponent, ReLU)**.** *Let $\rho_\ell^{\alpha\beta} := \frac{\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle}{|\varphi_\ell^\alpha|\,|\varphi_\ell^\beta|}$, where $\varphi_\ell^\alpha := \varphi_s(z_\ell^\alpha)$. Consider the limit $n \to \infty$ and $s_\pm = 1 + \frac{c_\pm}{n^p}$ for some $p \geq 0$. Then depending on the value of p, the interpolated process $\rho_{\lfloor tn \rfloor}^{\alpha\beta}$ converges in distribution w.r.t. the Skorohod topology of $D_{\mathbb{R}_+,\mathbb{R}}$ to*

(i) the degenerate limit: $\rho_t^{\alpha\beta} = 1$ for all $t > 0$, if $0 \leq p < \frac{1}{2}$, and $c_+ \neq c_-$,

(ii) the critical limit: the SDE from Theorem 3, if $p = \frac{1}{2}$,

(iii) the linear network limit: if $p > \frac{1}{2}$, the following SDE, with μ, σ as defined in (16)*,*

$$d\rho_t^{\alpha\beta} = \mu(\rho_t^{\alpha\beta})\, dt + \sigma(\rho_t^{\alpha\beta})\, dB_t\,, \quad \rho_0^{\alpha\beta} = \frac{\langle x^\alpha, x^\beta \rangle}{|x^\alpha|\,|x^\beta|}\,, \tag{17}$$

Here we remark that the unshaped network case ($p = 0$) is contained by the above in case (i). At the same time, we observe that case (iii) is equivalent to the correlation SDE in Theorem 3 except with $\nu = 0$. In particular, we observe this limit is also reached when $c_+ = c_-$, which implies $\varphi_s(x) = s_+ x$ is linear, which is why we call this the linear network limit. Furthermore, without much additional work, the same argument also implies the joint covariance SDE also loses the drift component, i.e., $dV_t = \Sigma(V_t)^{1/2}\, dB_t$.

3.2 Neural Covariance SDE for Shaped Smooth Activations

In this section, we consider smooth activation functions satisfying the following assumption. All the proofs for results in this section can be found in Appendix C.

Assumption 5. *$\varphi \in C^4(\mathbb{R})$, $\varphi(0) = 0, \varphi'(0) = 1$, and $|\varphi^{(4)}(x)| \leq C(1 + |x|^p)$ for some $C, p > 0$.*

We note that for any non-constant function $\sigma \in C^1(\mathbb{R})$ and $x_0 \in \mathbb{R}$ such that $\sigma'(x_0) \neq 0$, we can always define $\varphi(x) := \frac{\sigma(x+x_0) - \sigma(x_0)}{\sigma'(x_0)}$ such that it satisfies $\varphi(0) = 0, \varphi'(0) = 1$. The choice of x_0 will be discussed further in Section 4. The fourth derivative growth condition is used to control the Taylor remainder term in expectation, but any control over the remainder will suffice.

Definition 6. *Following the ideas of [37], we consider the following shaping of a smooth base activation function φ. For some constant $a > 0$, we set $\varphi_s(x) := s\varphi\left(\frac{x}{s}\right)$ with $s = a\sqrt{n}$, and observe that in the limit $n \to \infty$ we will similarly achieve that $\varphi_s \to Id$ as desired.*

In this regime, we can similarly characterize the joint output distribution, however the limiting SDEs are not always well behaved. In particular, they can have finite time explosions as described by the Feller test for explosions [42, Theorem 5.29]. The Neural Covariance SDE we derive for smooth activation functions becomes equivalent to the SDE in Proposition 7. The parameter b is determined by the activation function φ and controls whether or not finite time explosions happen (see (20)) !

Proposition 7 (Finite Time Explosion)**.** *Let $X_t \in \mathbb{R}$ be a solution to the following SDE*

$$dX_t = bX_t(X_t - 1)\, dt + \sqrt{2} X_t\, dB_t\,, \quad X_0 = x_0\,, \tag{18}$$

for some $b \in \mathbb{R}, x_0 > 0$. Let $\tau^* = \sup_{M>0} \inf\{t : |X_t| > M\}$ be the first time X_t exits all *compact intervals. If $\tau^* < \infty$, we say X_t has a finite time explosion. For this equation, $\mathbf{E}[\tau^*] < \infty$ if and only if $b > 0$.*

Technically speaking, the main culprit behind finite time explosions is the non-Lipschitzness of the drift coefficient. This issue requires us to weaken the sense of convergence in this section; the

ordinary convergence in the Skorohod topology is in general not true when the diffusion has finite time explosions. A weakened type of convergence is the best we can hope for. To this goal, we introduce the following definition.

Definition 8. *We say a sequence of processes X^n* ***converge locally*** *to X in the Skorohod topology if for any $r > 0$, we define the following stopping times*

$$\tau^n \coloneqq \{t \geq 0 : |X_t^n| \geq r\} \ , \quad \tau \coloneqq \{t \geq 0 : |X_t| \geq r\} \ , \tag{19}$$

and we have that $X^n_{t\wedge\tau^n}$ converge to $X_{t\wedge\tau}$ in the Skorohod topology.

This weakened sense of convergence essentially constrains the processes X^n, X in a bounded set by adding an absorbing boundary condition. Not only do these stopping time rules out explosions, the drift coefficient is now also Lipschitz on a compact set. With this notition of convergence, we can now state a precise version of the covariance SDE result for general smooth activation functions.

Theorem 9 (Covariance SDE, Smooth)**.** *Let φ satisfy Assumption 5, $V_\ell^{\alpha\beta} \coloneqq \frac{c}{n}\langle\varphi_\ell^\alpha, \varphi_\ell^\beta\rangle$ where $\varphi_\ell^\alpha = \varphi_s(z_\ell^\alpha)$, and define $V_\ell \coloneqq [V_\ell^{\alpha\beta}]_{1\leq\alpha\leq\beta=m}$ to be the upper triangular part thought of as a vector in $\mathbb{R}^{m(m+1)/2}$ Then, with $s = a\sqrt{n}$ as in Definition 6, in the limit as $n \to \infty$, $\frac{d}{n} \to T$, the interpolated process $V_{\lfloor tn\rfloor}$ converges locally in distribution to the solution of the following SDE in the Skorohod topology of $D_{\mathbb{R}_+,\mathbb{R}^{m(m+1)/2}}$*

$$dV_t = b(V_t)\,dt + \Sigma(V_t)^{1/2}\,dB_t \ , \quad V_0 = \left[\frac{1}{n_{in}}\langle x^\alpha, x^\beta\rangle\right]_{1\leq\alpha\leq\beta\leq m} \ , \tag{20}$$

where $\Sigma(V_t)$ is the same as Theorem 2 and

$$b^{\alpha\beta}(V_t) = \frac{\varphi''(0)^2}{4a^2}\left(V_t^{\alpha\alpha}V_t^{\beta\beta} + V_t^{\alpha\beta}(2V_t^{\alpha\beta} - 3)\right) + \frac{\varphi'''(0)}{2a^2}V_t^{\alpha\beta}(V_t^{\alpha\alpha} + V_t^{\beta\beta} - 2)\,. \tag{21}$$

Furthermore, if V_T is finite, then the output distribution can be described conditional on V_T as

$$[z_{out}^\alpha]_{\alpha=1}^m \,|V_T \overset{d}{=} \mathcal{N}\left(0, [V_T^{\alpha\beta}]_{\alpha,\beta=1}^m\right) \ , \tag{22}$$

and otherwise the distribution of $[z_{out}^\alpha]_{\alpha=1}^m$ is undefined.

We also have a similar critical scaling result for general smooth activations.

Proposition 10 (Critical Exponent, Smooth)**.** *Let φ satisfy Assumption 5, $V_\ell^{\alpha\beta} \coloneqq \frac{c}{n}\langle\varphi_\ell^\alpha, \varphi_\ell^\beta\rangle$ where $\varphi_\ell^\alpha = \varphi_s(z_\ell^\alpha)$ with $s = an^p$ for some $p \geq 0$, and define $V_\ell \coloneqq [V_\ell^{\alpha\beta}]_{1\leq\alpha\leq\beta=m}$ to be the upper triangular component of the covariance matrix. Then in the limit as $n \to \infty$, $\frac{d}{n} \to T$, the interpolated process $V_{\lfloor tn\rfloor}^{\alpha\beta}$ converges locally in distribution w.r.t. the Skorohod topology of $D_{\mathbb{R}_+,\mathbb{R}^{m(m+1)/2}}$ to V, which depending on the value of p is*

(i) the degenerate limit: if $0 < p < \frac{1}{2}$

$$\begin{cases} V_t^{\alpha\alpha} = 0 \text{ or } \infty, & \text{if } \frac{3}{4}\varphi''(0)^2 + \varphi'''(0) > 0\,, \\ V_t^{\alpha\beta} = 1\,, & \text{if } \frac{3}{4}\varphi''(0)^2 + \varphi'''(0) \leq 0 \text{ and } \varphi''(0) \neq 0\,, \end{cases} \tag{23}$$

for all $t > 0$ and $1 \leq \alpha \leq \beta \leq m$,

(ii) the critical limit: the solution of the SDE from Theorem 9, if $p = \frac{1}{2}$,

(iii) the linear network limit: the stopped solution to the SDE $dV_t = \Sigma(V_t)\,dB_t$ with coefficient Σ defined in Theorem 3, if $p > \frac{1}{2}$.

Here we observe that case (i) also implies a degenerate correlation $\rho_t^{\alpha\beta}$ similar to the ReLU case Proposition 4. At the same time, the linear network limit in case (iii) also has the same covariance SDE as Proposition 4.

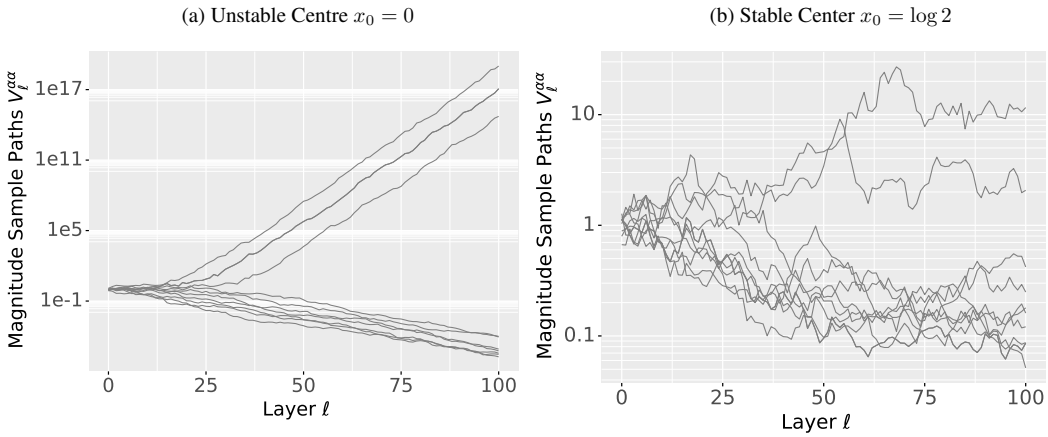

Figure 2: Simulation of 10 shaped softplus networks as in Example 12 with $n = d = 100, a = 1, V_0^{\alpha\alpha} = \frac{1}{n_{\text{in}}}|x^\alpha|^2 = 1$ centred at two different values. "Stable" here means the Neural Covariance SDE is guaranteed not to have finite time explosions; unstable networks can explode on initialization!

4 Consequences of the Neural Covariance SDE and Future Directions

So far, we have derived the Neural Covariance SDE. Analysis of this SDE reveals important behaviour of the network on initialization. Here we lay out one concrete example and some future directions.

Exploding and Vanishing Norms. Recall the shaping parameter a from Definition 6. Let V_t be the solution to the SDE in (20). We can write down the marginal SDE for $V_t^{\alpha\alpha}$ as

$$dV_t^{\alpha\alpha} = \left(\frac{\varphi''(0)^2}{4} + \varphi'''(0)\right)\frac{V_t^{\alpha\alpha}}{a^2}(V_t^{\alpha\alpha} - 1)\,dt + \sqrt{2}V_t^{\alpha\alpha}\,dB_t\,, \tag{24}$$

which implies by Proposition 7 that V_t has a finite time explosion **if and only if** $\frac{\varphi''(0)^2}{4} + \varphi'''(0) > 0$. This criterion can be used to help choose which activation function to use; below are two examples.

Example 11 (Sigmoid and $\tanh$ at $x_0 = 0$)**.** *We start with the sigmoid activation $\sigma(x) = \frac{1}{1+e^{-x}}$, then we can define $\varphi(x) := 4\sigma(x) - 2$ to satisfy Assumption 5, which leads to $\varphi''(0) = 0, \varphi'''(0) = -\frac{1}{2}$, and therefore leads to a stable network. It turns out $\varphi(x) := \tanh(x)$ already satisfies Assumption 5, which leads to $\varphi''(0) = 0, \varphi'''(0) = -2$, and therefore is also stable.*

More generally, if σ behaves like a cumulative distribution function for a symmetric unimodal density, we will have that $\varphi''(0) = 0$ and $\varphi'''(0) < 0$ as desired.

Example 12 (Soft Plus at General $x_0 \in \mathbb{R}$)**.** *Let us consider $x_0 \in \mathbb{R}$ and $\sigma(x) = \log(1 + e^{x+x_0})$, which implies $\varphi(x) := (1 + e^{-x_0})\log\frac{1+e^{x+x_0}}{1+e^{x_0}}$ satisfies Assumption 5. This gives us $\varphi''(0) = \frac{1}{1+e^{x_0}}, \varphi'''(0) = \frac{1-e^{\beta x_0}}{(1+e^{x_0})^2}$, and therefore $\frac{\varphi''(0)^2}{4} + \varphi'''(0) = \frac{1}{(1+e^{\beta x_0})^2}\left(\frac{5}{4} - e^{x_0}\right)$. In other words, the shaped network is stable if and only if $x_0 \geq \log\frac{5}{4}$ (see Figure 2).*

Future Directions. By providing a precise scaling limit in the shaped infinite-depth-and-width regime, we can use these covariance SDEs to study more precise properties of DNNs at initialization. Firstly, practitioners using the shaping techniques of [37, 38] can now simulate the covariance SDEs at a low computational cost to significantly improve estimates of the output correlation (see Figure 1 and additional simulations in Appendix E). At the same time, the techniques developed here and the covariance SDEs also serve as a mathematical foundation for studying training and generalization, in the same way that NNGP theory did in the infinite-width regime.

References

[1] R. M. Neal. *Bayesian learning for neural networks*. Vol. 118. Springer Science & Business Media, 1995.

[2] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. "Deep Neural Networks as Gaussian Processes." *Int. Conf. Learning Representations (ICLR)*. 2018.

[3] A. Jacot, F. Gabriel, and C. Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." *Advances in Information Processing Systems (NeurIPS)*. 2018. arXiv: 1806.07572.

[4] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. "Gradient descent finds global minima of deep neural networks." *Int. Conf. Machine Learning (ICML)*. PMLR. 2019, pp. 1675–1685.

[5] Z. Allen-Zhu, Y. Li, and Z. Song. "A convergence theory for deep learning via over-parameterization." *Int. Conf. Machine Learning (ICML)*. PMLR. 2019, pp. 242–252.

[6] D. Zou, Y. Cao, D. Zhou, and Q. Gu. "Gradient descent optimizes over-parameterized deep ReLU networks." In: *Machine Learning* 109.3 (2020), pp. 467–492.

[7] L. Chizat, E. Oyallon, and F. Bach. "On Lazy Training in Differentiable Programming." In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 2937–2947.

[8] J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. *Wide neural networks of any depth evolve as linear models under gradient descent*. 2019. arXiv: 1902.06720.

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

[10] G. Yang. *Tensor programs ii: Neural tangent kernel for any architecture*. 2020. arXiv: 2006.14548.

[11] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang. "On exact computation with an infinitely wide neural net." *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2019, pp. 8141–8150.

[12] Z. Chen, Y. Cao, D. Zou, and Q. Gu. "How Much Over-parameterization Is Sufficient to Learn Deep Re{LU} Networks?" *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=fgd7we_uZa6.

[13] Z. Ji and M. Telgarsky. "Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks." In: *arXiv preprint arXiv:1909.12292* (2019).

[14] J. Ba, M. Erdogdu, T. Suzuki, D. Wu, and T. Zhang. "Generalization of two-layer neural networks: An asymptotic viewpoint." *International conference on learning representations*. 2019.

[15] P. L. Bartlett, A. Montanari, and A. Rakhlin. "Deep learning: a statistical viewpoint." In: *Acta numerica* 30 (2021), pp. 87–201.

[16] G. M. Rotskoff and E. Vanden-Eijnden. *Trainability and Accuracy of Neural Networks: An Interacting Particle System Approach*. 2018. arXiv: 1805.00915.

[17] L. Chizat and F. Bach. *On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport*. 2018. arXiv: 1805.09545.

[18] J. Sirignano and K. Spiliopoulos. *Mean Field Analysis of Neural Networks: A Law of Large Numbers*. 2018. arXiv: 1805.01053.

[19] S. Mei, A. Montanari, and P.-M. Nguyen. "A mean field view of the landscape of two-layer neural networks." In: *Proceedings of the National Academy of Sciences* 115.33 (2018), E7665–E7671. ISSN: 0027-8424. DOI: 10.1073/pnas.1806579115. eprint: https://www.pnas.org/content/115/33/E7665.full.pdf.

[20] G. Yang and E. J. Hu. "Feature Learning in Infinite-Width Neural Networks." *Int. Conf. Machine Learning (ICML)*. 2021. arXiv: 2011.14522.

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

[22] J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang. "High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation." In: *arXiv preprint arXiv:2205.01445* (2022).

[23] B. Hanin and M. Nica. "Finite Depth and Width Corrections to the Neural Tangent Kernel." *Int. Conf. Learning Representations (ICLR)*. 2019.

[24] M. Seleznova and G. Kutyniok. *Analyzing Finite Neural Networks: Can We Trust Neural Tangent Kernel Theory?* 2020. arXiv: 2012.04477.

[25] B. Hanin and M. Nica. "Products of many large random matrices and gradients in deep neural networks." In: *Communications in Mathematical Physics* (2019), pp. 1–36.

[26] Z. Hu and H. Huang. "On the Random Conjugate Kernel and Neural Tangent Kernel." *International Conference on Machine Learning*. PMLR. 2021, pp. 4359–4368.

[27] M. Li, M. Nica, and D. Roy. "The future is log-Gaussian: ResNets and their infinite-depth-and-width limit at initialization." In: *Advances in Neural Information Processing Systems* 34 (2021).

[28] J. Zavatone-Veth and C. Pehlevan. "Exact marginal prior distributions of finite Bayesian neural networks." In: *Advances in Neural Information Processing Systems* 34 (2021).

[29] L. Noci, G. Bachmann, K. Roth, S. Nowozin, and T. Hofmann. "Precise characterization of the prior predictive distribution of deep ReLU networks." In: *Advances in Neural Information Processing Systems* 34 (2021).

[30] D. A. Roberts, S. Yaida, and B. Hanin. *The principles of deep learning theory*. Cambridge University Press, 2022.

[31] J. Zavatone-Veth, A. Canatar, B. Ruben, and C. Pehlevan. "Asymptotics of representation learning in finite Bayesian neural networks." In: *Advances in Neural Information Processing Systems* 34 (2021).

[32] B. Hanin. "Correlation Functions in Random Fully Connected Neural Networks at Finite Width." In: *arXiv preprint arXiv:2204.01058* (2022).

[33] S. Buchanan, D. Gilboa, and J. Wright. "Deep Networks and the Multiple Manifold Problem." *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=0-6Pm_d_Q-.

[34] G. Yang and S. S. Schoenholz. "Mean field residual networks: on the edge of chaos." *Advances in Neural Information Processing Systems*. 2017, pp. 2865–2873.

[35] S. Hayou, E. Clerico, B. He, G. Deligiannidis, A. Doucet, and J. Rousseau. "Stable ResNet." *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*. PMLR. 2021, pp. 1324–1332.

[36] B. Hanin and D. Rolnick. "How to Start Training: The Effect of Initialization and Architecture." In: *Advances in Neural Information Processing Systems* 31 (2018).

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

[38] G. Zhang, A. Botev, and J. Martens. "Deep Learning without Shortcuts: Shaping the Kernel with Tailored Rectifiers." In: *arXiv preprint arXiv:2203.08120* (2022).

[39] K. He, X. Zhang, S. Ren, and J. Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." *Proc. IEEE Int. Conf. Computer Vision*. 2015, pp. 1026–1034.

[40] Y. Cho and L. K. Saul. "Kernel methods for deep learning." *Advances in Neural Information Processing Systems (NeurIPS)*. 2009, pp. 342–350.

[41] T. Campbell and T. Broderick. "Automated scalable Bayesian inference via Hilbert coresets." In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 551–588.

[42] I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*. Vol. 113. Springer Science & Business Media, 2012.

[43] O. Kallenberg and O. Kallenberg. *Foundations of modern probability*. Vol. 2. Springer, 1997.

[44] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.

[45] D. W. Stroock and S. S. Varadhan. *Multidimensional diffusion processes*. Vol. 233. Springer Science & Business Media, 1997.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The contributions list directly references the theorems and sections corresponding to them.

(b) Did you describe the limitations of your work? [Yes] In our introduction, we discussed that our theory is only at initialization, and does not describe training dynamics and generalization.

(c) Did you discuss any potential negative societal impacts of your work? [N/A]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] The only assumption required is in Assumption 5.

(b) Did you include complete proofs of all theoretical results? [Yes] Please see the relevant sections in Appendices A to C.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All of our simulations (at initialization only) are contained the file *Correlation.ipynb*.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] Our experiments only required simulations at initialization, and therefore no training.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] We plotted the histogram and full density of the samples instead.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Our simulations were small enough that it did not require GPUs.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]