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

Mufan (Bill) Li
University of Toronto,
Vector Institute

Mihai Nica
University of Guelph,
Vector Institute

Daniel M. Roy
University of Toronto,
Vector Institute

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 [16–19], where the different parameterization demonstrates benefits for feature learning and hyperparameter tuning [20–22].

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 networks 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–

Correspondence: mufan.li@mail.utoronto.ca; nicam@uoguelph.ca; daniel.roy@utoronto.ca.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

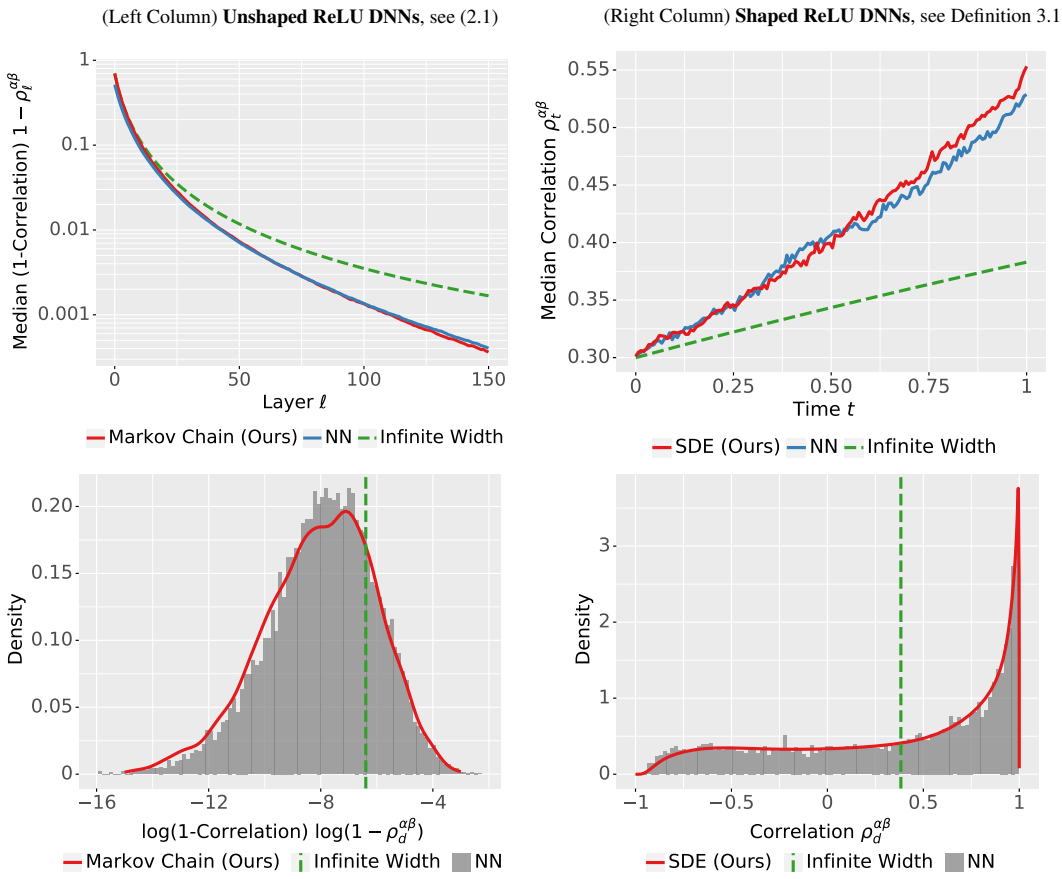

Figure 1: Simulations of correlation $\rho_\ell^{\alpha\beta} = \frac{\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle}{|\varphi_\ell^\alpha||\varphi_\ell^\beta|}$ between post-activation vectors in ReLU networks, comparing *finite NNs* vs. our theoretical predictions vs. infinite-width paradigm. **Left Column:** $\rho_\ell^{\alpha\beta}$ vs. our Markov chain (2.10) vs. infinite-width update $\rho_{\ell+1} = cK_1(\rho_\ell)$ (see (2.10) and note the log scale and $1-\rho$ here). **Right Column:** $\rho_{\lfloor tn \rfloor}^{\alpha\beta}$ vs. our Neural Covariance SDE vs. ODE $d\rho_t = \nu(\rho_t)\,dt$ (see Theorem 3.3). **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 1e$-$2. Densities from kernel density estimation.

33] or to study concentration bounds in terms of d/n [5, 34]. 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. Furthermore, it has also been shown that there is feature learning in this limit [23], in contrast to the linear regime of infinite-width limits [8]. 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 [35, 36, 33]. 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 [37, 25]. Intuitively, the output is not very sensitive to changes in the input, hence the gradient must be very large in the earlier layers.

A promising new attack on this problem is to modify the activation function ("shaping") to reduce to the effect of degeneracy [38, 39]. In this prior work, extensive experiments show that shaping the activation significantly improves training speed *without the need for normalization layers*. This method has been proven effective for problems as large as standard ResNets on ImageNet data. The authors designed several criteria including reducing estimated output correlation, and numerically

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 $1 \le \alpha \le m$	$W_0 \in \mathbb{R}^{n_{\text{in}} \times n}$	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 $1 \le \ell \le d$	$W_\ell \in \mathbb{R}^{n \times n}$	Weight matrix at layer $1 \le \ell \le d$ **All weights initialized iid** $\sim \mathcal{N}(0,1)$
$\varphi^\alpha_\ell \in \mathbb{R}^n$	Neurons (post-activation) for layer $1 \le \ell \le d$	$c \in \mathbb{R}$	Normalizing constant $c := \left(\mathbb{E}\,\varphi(g)^2\right)^{-1}$ for $g \sim \mathcal{N}(0,1)$

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.

In this paper, we 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 nonlinear 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, which 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 to 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 3.4 and Proposition 3.10.
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 3.7 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 F.

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:

$$z^\alpha_1 := \frac{1}{\sqrt{n_{\text{in}}}} W_0 x^\alpha, \quad \varphi^\alpha_\ell := \varphi(z^\alpha_\ell), \quad z^\alpha_{\ell+1} := \sqrt{\frac{c}{n}} W_\ell \varphi^\alpha_\ell, \quad z^\alpha_{\text{out}} := \sqrt{\frac{c}{n}} W_{\text{out}} \varphi^\alpha_d \,. \tag{2.1}$$

Note that factors of $\sqrt{cn^{-1}}$ are equivalent to intializing according to the so-called He initialization [40]. 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_-:

$$\varphi(x) \coloneqq s_+ \max(x, 0) + s_- \min(x, 0) = s_+ \varphi_{\text{ReLU}}(x) - s_- \varphi_{\text{ReLU}}(-x)\,. \tag{2.2}$$

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 A.6 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 (2.3). Suppose that for each $n \in \mathbb{N}$, Y_ℓ^n is a discrete time Markov chain $\ell \in \mathbb{N}$ whose increments obey (2.3) in terms of the same functions b, σ^2:

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t\,, \qquad Y_{\ell+1}^n - Y_\ell^n = b(Y_\ell^n)\frac{1}{n} + \sigma(Y_\ell^n)\frac{\xi_\ell}{\sqrt{n}} + O(n^{-3/2}), \tag{2.3}$$

where ξ_ℓ are independent variables with $\mathbb{E}(\xi_\ell) = 0, \mathbf{Var}(\xi_\ell) = 1$. 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, or more precisely: with $X_t^n \coloneqq Y_{\lfloor tn \rfloor}^n$ we have $X^n \to X$ 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 matrix with iid Gaussian entries 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 (2.1) with the positive homogeneity of φ to write the Markov chain update rule:

$$\left|\varphi_{\ell+1}^\alpha\right|^2 = |\varphi_\ell^\alpha|^2 \frac{1}{n}\sum_{i=1}^{n} c\varphi(g_i^\alpha)^2, \quad \text{where } g^\alpha \coloneqq W_\ell \frac{\varphi_\ell^\alpha}{|\varphi_\ell^\alpha|} \overset{d}{=} \mathcal{N}(0, I_n)\,. \tag{2.4}$$

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

$$\left|\varphi_{\ell+1}^\alpha\right|^2 = |\varphi_\ell^\alpha|^2\left(1 + \frac{1}{\sqrt{n}}R_\ell^{\alpha\alpha}\right), \tag{2.5}$$

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 (2.3) with $Y_0^n = \frac{1}{n_{\text{in}}}|x^\alpha|^2$, $b(Y) \equiv 0, \sigma^2(Y) = \mathbf{Var}(R_\ell^{\alpha\alpha})Y^2 = \mathbf{Var}(c\varphi(g)^2)Y^2$. 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

$$\lim_{n\to\infty}\frac{c}{n}\left|\varphi_{\lfloor tn \rfloor}^\alpha\right|^2 = X_t \overset{d}{=} e^{\mathcal{N}(-\frac{\sigma^2}{2}t, \sigma^2 t)}\,, \tag{2.6}$$

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 $\rho_\ell^{\alpha\beta}$

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 $\varphi_\ell^\alpha = \varphi(z_\ell^\alpha)$. We note that the convergence of Markov chains to SDEs in (2.3) can be generalized to $Y_\ell^n \in \mathbb{R}^N$ by considering $\mathbf{Cov}(\xi_\ell) = I_N$, $b : \mathbb{R}^N \to \mathbb{R}^N$, and $\sigma : \mathbb{R}^N \to \mathbb{R}^{N\times N}$. 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:

$$\begin{aligned} [z_{\ell+1}^\alpha]_{\alpha=1}^m \big|\, \mathcal{F}_\ell &\overset{d}{=} \mathcal{N}\left(0, \frac{c}{n}[\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle]_{\alpha,\beta=1}^m \otimes I_n\right), \\ [z_{\text{out}}^\alpha]_{\alpha=1}^m \big|\, \mathcal{F}_d &\overset{d}{=} \mathcal{N}\left(0, \frac{c}{n}[\langle \varphi_d^\alpha, \varphi_d^\beta \rangle]_{\alpha,\beta=1}^m \otimes I_{n_{\text{out}}}\right), \end{aligned} \tag{2.7}$$

where $\mathcal{F}_\ell$ denotes the sigma-algebra generated by the ℓ-th layer $[z_\ell^\alpha]_{\alpha=1}^m$, and $\otimes$ denotes the Kronecker product (here indicating conditionally independent entries in each vector). With this property in mind, we will introduce $\mathbb{E}_\ell[\,\cdot\,] \coloneqq \mathbb{E}[\,\cdot\,|\mathcal{F}_\ell]$ to denote the conditional expectation, and $\mathbf{Var}_\ell(\,\cdot\,), \mathbf{Cov}_\ell(\,\cdot\,)$ similarly to denote the conditional variance and covariance. If we define g^α as in (2.4), we see that the g^α are all marginally $\mathcal{N}(0, I_n)$. Similar to (2.4), we can write the update rule for the α, β-entry of the Gram matrix:

$$\langle \varphi_{\ell+1}^\alpha, \varphi_{\ell+1}^\beta \rangle = |\varphi_\ell^\alpha||\varphi_\ell^\beta| \frac{1}{n} \sum_{i=1}^n c\varphi(g_i^\alpha)\varphi(g_i^\beta), \tag{2.8}$$

Just as we did in (2.5), we can define $R_\ell^{\alpha\beta} \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^n c\varphi(g_i^\alpha)\varphi(g_i^\beta) - \mathbb{E}_\ell[c\varphi(g_i^\alpha)\varphi(g_i^\beta)]$ and write

$$\langle \varphi_{\ell+1}^\alpha, \varphi_{\ell+1}^\beta \rangle = |\varphi_\ell^\alpha||\varphi_\ell^\beta| \left(\mathbb{E}_\ell\left[c\varphi(g_i^\alpha)\varphi(g_i^\beta)\right] + \frac{1}{\sqrt{n}} R_\ell^{\alpha\beta} \right), \tag{2.9}$$

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

However, unlike the simple single-data-point case from Section 2.2, **we do not have convergence to a continuous time SDE.** This is because the differences $\langle \varphi_{\ell+1}^\alpha, \varphi_{\ell+1}^\beta \rangle - \langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle \nrightarrow 0$ as $n \to \infty$. Instead, (2.9) 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 function f, and consequently $Y_{\ell+1}^n - Y_\ell^n$ does not vanish as $n \to \infty$.

For a clarifying example, we can consider the one-dimensional Markov chain of hidden layer correlations. More precisely, we can define $\rho_\ell^{\alpha\beta} = \langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle / |\varphi_\ell^\alpha||\varphi_\ell^\beta|$, which we observe can be extracted from the entries of the Gram matrix. In fact, we can write down an approximate recursion update for $\rho_\ell^{\alpha\beta}$ (see Appendix B and Proposition B.8 for details):

$$\rho_{\ell+1}^{\alpha\beta} \approx cK_1(\rho_\ell^{\alpha\beta}) + \frac{1}{n}\mu_{\text{ReLU}}(\rho_\ell^{\alpha\beta}) + \frac{\xi_\ell}{\sqrt{n}}\sigma_{\text{ReLU}}(\rho_\ell^{\alpha\beta}), \quad \rho_0^{\alpha\beta} = \frac{\langle x^\alpha, x^\beta \rangle}{n_{\text{in}}}, \tag{2.10}$$

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 ξ_ℓ are iid $N(0,1)$. For the ReLU case, $c = 2$ and $cK_1(\rho) = (\sqrt{1-\rho^2} + \rho \arccos(-\rho))/\pi$ was first calculated in [41]. In fact, we can observe that as $n \to \infty$, $\rho_{\lfloor tn \rfloor}^{\alpha\beta}$ converges to the fixed point of $cK_1(\cdot)$ at $\rho = 1$ for all $t > 0$. **We note this limiting behaviour cannot be described by an SDE**, as the solution must jump from the initial condition to the fixed point at $t = 0$.

Despite not having an SDE limit, we observe that the approximate Markov chain (2.10) already provides a much better approximation to finite size networks compared to the infinite-width theory (see left column of Figure 1). This is because the infinite-width approach discards the terms in (2.10) that vanish as $n \to \infty$ and consider only the update $\rho_{\ell+1}^{\alpha\beta} = cK_1(\rho_\ell^{\alpha\beta})$. 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 [33] and a new bound in Appendix E).

Furthermore, we observe that in this case, the microscopic $O(n^{-1})$ and $O(n^{-1/2})$ terms in (2.10) accumulate to macroscopic differences! For the examples in Figure 1, we see their net effect is that

$\rho_\ell^{\alpha\beta} \to 1$ faster than the infinite-width prediction. Heuristically, 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 see that we are just one step away from achieving limiting SDEs.

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

In this section, we follow the ideas of [38, 39] to *reshape* the activation function φ. Reshaping means to replace the base activation function φ in (2.1) with φ_s that depends on width n. We will also replace the normalizing constant $c = \left(\mathbb{E}\,\varphi_s(g)^2\right)^{-1}$ for $g \sim \mathcal{N}(0,1)$. 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 (2.7) 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 we recover a complete characterization by describing the random covariance matrix.

3.1 Neural Covariance SDE for Shaped ReLU-Like Activations

Definition 3.1. *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 also set $c = \left(\mathbb{E}\,\varphi_s(g)^2\right)^{-1}$ for $g \sim \mathcal{N}(0,1)$.*

We will show that with shaping of Definition 3.1, one gets non-trivial SDEs that describe the covariance (Theorem 3.2) and correlations (Theorem 3.3) of the network. The precise scaling is shown to be the critical scaling for a non-trivial limit in Proposition 3.4. All proofs for results in this section appear in Appendix C.

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 3.2 we use V_ℓ for the covariance at layer ℓ and V_t to denote the limiting SDE at time t.

Theorem 3.2 (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 3.1, in the limit as $n \to \infty, \frac{d}{n} \to T$, the interpolated process $V_{\lfloor tn \rfloor}$ converges in distribution in the Skorohod topology of $D_{\mathbb{R}_+, \mathbb{R}^{m(m+1)/2}}$ to the solution of the SDE*

$$dV_t = b(V_t)\,dt + \Sigma(V_t)^{1/2}\,dB_t\,, \quad V_0 = \left[\frac{1}{n_{in}}\langle x^\alpha, x^\beta \rangle\right]_{1 \le \alpha \le \beta \le m}\,, \tag{3.1}$$

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

$$b(V_t) = \left[\nu\left(\rho_t^{\alpha\beta}\right)\sqrt{V_t^{\alpha\alpha} V_t^{\beta\beta}}\right]_{1 \le \alpha \le \beta \le m}, \quad \textit{and} \quad \Sigma(V_t) = \left[V_t^{\alpha\gamma} V_t^{\beta\delta} + V_t^{\alpha\delta} V_t^{\beta\gamma}\right]_{\alpha \le \beta, \gamma \le \delta}. \tag{3.2}$$

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

$$[z_{out}^\alpha]_{\alpha=1}^m \,|V_T \overset{d}{=} \mathcal{N}\left(0, [V_T^{\alpha\beta}]_{\alpha,\beta=1}^m\right). \tag{3.3}$$

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. However, we emphasize that the m-point joint output distribution is *not* characterized by the marginal for each of the pairs, as the output z_{out}^α is *not* Gaussian. In particular, we observe the diffusion matrix entry corresponding to $V_t^{\alpha\beta}, V_t^{\gamma\delta}$ involves other processes $V_t^{\alpha\gamma}, V_t^{\beta\delta}, V_t^{\alpha\delta}, V_t^{\beta\gamma}$! *This implies that the Neural Covariance SDE limit cannot be described by a kernel, unlike stacking random features or NNGP.*

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 3.3 (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* $D_{\mathbb{R}_+,\mathbb{R}}$

$$d\rho_t^{\alpha\beta} = \Big[\nu(\rho_t^{\alpha\beta}) + \mu(\rho_t^{\alpha\beta})\Big]\, dt + \sigma(\rho_t^{\alpha\beta})\, dB_t\,, \quad \rho_0^{\alpha\beta} = \frac{\langle x^\alpha, x^\beta \rangle}{|x^\alpha|\,|x^\beta|}\,, \tag{3.4}$$

where

$$\nu(\rho) = \frac{(c_+ - c_-)^2}{2\pi}\left[\sqrt{1-\rho^2} - \arccos(\rho)\rho\right], \quad \mu(\rho) = -\frac{1}{2}\rho(1-\rho^2)\,, \quad \sigma(\rho) = 1 - \rho^2\,. \tag{3.5}$$

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. [39] has derived a related ordinary differential equation (ODE) of $d\rho_t = \nu(\rho_t)\, dt$ in the sequential limit of $n \to \infty$ then $d \to \infty$, where the activation is shaped depending on depth. Here we also note that $\nu(\rho)$ is closely related to the J_1 function derived in [41]. See Appendix C.3 for the m-point joint version of the correlation SDE, and Appendix F for an empirical measure of convergence in the Kolmogorov–Smirnov distance.

One immediate consequence of the correlation SDE is that we can show the $n^{-1/2}$ scaling in Definition 3.1 is the only case where the limit is neither degenerate nor a linear network.

Proposition 3.4 (Critical Exponent, ReLU)**.** *Let* $\rho_\ell^{\alpha\beta} \coloneqq \frac{\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle}{|\varphi_\ell^\alpha|\,|\varphi_\ell^\beta|}$*, where* $\varphi_\ell^\alpha \coloneqq \varphi_s(z_\ell^\alpha)$*. 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 3.3, if $p = \frac{1}{2}$*,*

(iii) the linear network limit: if $p > \frac{1}{2}$ *, the following SDE, with* μ, σ *as defined in* (3.5)*,*

$$d\rho_t^{\alpha\beta} = \mu(\rho_t^{\alpha\beta})\, dt + \sigma(\rho_t^{\alpha\beta})\, dB_t\,, \quad \rho_0^{\alpha\beta} = \frac{\langle x^\alpha, x^\beta \rangle}{|x^\alpha|\,|x^\beta|}\,. \tag{3.6}$$

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 3.3 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 the reason 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 and derive a similar covariance SDE. All the proofs for results in this section can be found in Appendix D.

Assumption 3.5. $\varphi \in C^4(\mathbb{R})$*,* $\varphi(0) = 0, \varphi'(0) = 1$*, 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 $\varphi(x) \coloneqq \frac{\sigma(x+x_0) - \sigma(x_0)}{\sigma'(x_0)}$ 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.

Following the ideas of [38], we consider the following shaping of a smooth activation function φ.

Definition 3.6. *For some constant* $a > 0$*, we set* $\varphi_s(x) \coloneqq s\varphi\left(\frac{x}{s}\right)$ *with* $s = a\sqrt{n}$*, and* $c = \left(\mathbb{E}\,\varphi_s(g)^2\right)^{-1}$ *for* $g \sim \mathcal{N}(0,1)$*.*

Observe that in the limit $n \to \infty$, we will achieve that $\varphi_s \to \text{Id}$ as desired. We also observe that the shaping factor s outside the activation cancels out with the next layer's $\frac{1}{s}$ factor, therefore it is equivalent shape the entire network. More precisely, if we view z_{out} as an input-output map $f : \mathbb{R}^{n_{\text{in}}} \to \mathbb{R}^{n_{\text{out}}}$ of an unshaped network, then shaping the smooth activation functions is equivalent to the modification $sf\left(\frac{x}{s}\right)$.[1]

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 5.5.29]. Here the SDE in Proposition 3.7 is exactly the $V_t^{\alpha\alpha}$ marginal of the Neural Covariance SDE, with the parameter b determined by the activation function φ and controls whether or not finite time explosions happen (see (4.1)).

Proposition 3.7 (Finite Time Explosion)**.** *Let $X_t \in \mathbb{R}_+$ be a solution to the following SDE*

$$dX_t = bX_t(X_t - 1)\, dt + \sqrt{2}X_t\, dB_t\,, \quad X_0 = x_0 > 0\,, b \in \mathbb{R}\,. \tag{3.7}$$

Let $\tau^* = \sup_{M>0} \inf\{t : X_t \geq M \text{ or } X_t \leq M^{-1}\}$ be the explosion time, and we say X_t has a finite time explosion if $\tau^* < \infty$. For this equation, $\mathbb{P}[\tau^* = \infty] = 1$ if and only if $b \leq 0$.

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 3.8. *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*

$$\tau^n \coloneqq \{t \geq 0 : |X_t^n| \geq r\}\,, \quad \tau \coloneqq \{t \geq 0 : |X_t| \geq r\}\,, \tag{3.8}$$

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 times rule out explosions, the drift coefficient is now also Lipschitz on a compact set. With this notion of convergence, we can now state a precise Neural Covariance SDE result for general smooth activation functions.

Theorem 3.9 (Covariance SDE, Smooth)**.** *Let φ satisfy Assumption 3.5, $V_\ell^{\alpha\beta} \coloneqq \frac{c}{n}\langle\varphi_\ell^\alpha, \varphi_\ell^\beta\rangle$ 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 entries thought of as a vector in $\mathbb{R}^{m(m+1)/2}$. Then, with $s = a\sqrt{n}$ as in Definition 3.6, 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 $D_{\mathbb{R}_+,\mathbb{R}^{m(m+1)/2}}$*

$$dV_t = b(V_t)\, dt + \Sigma(V_t)^{1/2}\, dB_t\,, \quad V_0 = \left[\frac{1}{n_{in}}\langle x^\alpha, x^\beta\rangle\right]_{1\leq\alpha\leq\beta\leq m}\,, \tag{3.9}$$

where $\Sigma(V_t)$ is the same as Theorem 3.2 and

$$b^{\alpha\beta}(V_t) = \frac{\varphi''(0)^2}{4a^2}\left(V_t^{\alpha\alpha}V_t^{\beta\beta} + V_t^{\alpha\beta}(2V_t^{\alpha\beta} - 3)\right) + \frac{\varphi'''(0)}{2a^2}V_t^{\alpha\beta}(V_t^{\alpha\alpha} + V_t^{\beta\beta} - 2)\,. \tag{3.10}$$

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

$$[z_{out}^\alpha]_{\alpha=1}^m\,|V_T \overset{d}{=} \mathcal{N}\left(0, [V_T^{\alpha\beta}]_{\alpha,\beta=1}^m\right)\,, \tag{3.11}$$

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 3.10 (Critical Exponent, Smooth)**.** *Let φ satisfy Assumption 3.5, $V_\ell^{\alpha\beta} \coloneqq \frac{c}{n}\langle\varphi_\ell^\alpha, \varphi_\ell^\beta\rangle$ where $\varphi_\ell^\alpha = \varphi_s(z_\ell^\alpha)$ with $s = an^p$ for some $p > 0$, and define $V_\ell \coloneqq [V_\ell^{\alpha\beta}]_{1\leq\alpha\leq\beta=m}$ to be the upper triangular entries thought of as a vector. Then in the limit as $n \to \infty$, $\frac{d}{n} \to T$, the interpolated process $V_{\lfloor tn\rfloor}$ 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*

[1] We want to thank Boris Hanin for observing this equivalent parameterization.

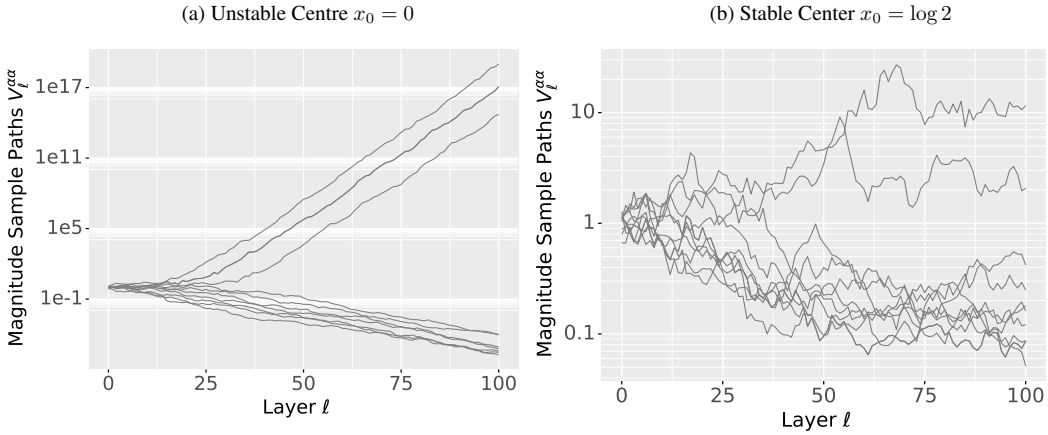

Figure 2: Simulation of 10 shaped softplus networks as in Example 4.2 with $n = d = 100, a = 1, V_0^{\alpha\alpha} = \frac{1}{n_{\text{in}}}|x^\alpha|^2 = 1$ centred at two different values. "Stable" here means the Neural Covariance SDE is guaranteed not to have finite time explosions; unstable networks can explode on initialization!

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

$$\begin{cases} V_t^{\alpha\alpha} = 0 \text{ or } \infty, & \text{if } \frac{3}{4}\varphi''(0)^2 + \varphi'''(0) > 0 \text{ and } V_0^{\alpha\alpha} \neq 0\,, \\ V_t^{\alpha\beta} = \text{const.}\,, & \text{if } \frac{3}{4}\varphi''(0)^2 + \varphi'''(0) \leq 0\,, \end{cases} \tag{3.12}$$

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

(ii) the critical limit: the solution of the SDE from Theorem 3.9, 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 3.3, if* $p > \frac{1}{2}$.

Here we observe that in case (i) when $\frac{3}{4}\varphi''(0)^2 + \varphi'''(0) \leq 0$, we also have a constant (in time) correlation $\rho_t^{\alpha\beta}$ similar to the ReLU case in Proposition 3.4, however in this case $\rho_t^{\alpha\beta}$ is not necessarily equal to 1. At the same time, the linear network limit in case (iii) also has the same covariance SDE as Proposition 3.4.

4 Consequences, Discussion, 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 provide some discussion and future directions.

Exploding and Vanishing Norms. Here we consider the behaviour of shaping smooth activation functions, as it is done in the experiments of [38]. While the authors here avoided exploding and vanishing norms by numerically optimizing shaping parameters, we can actually describe the precise behaviour a priori with the Neural Covariance SDE. Recall the shaping parameter a from Definition 3.6. Let V_t be the solution to the SDE in (3.9). We can write down the marginal SDE for $V_t^{\alpha\alpha}$ as

$$dV_t^{\alpha\alpha} = \left(\frac{3}{4}\varphi''(0)^2 + \varphi'''(0)\right)\frac{V_t^{\alpha\alpha}}{a^2}(V_t^{\alpha\alpha} - 1)\, dt + \sqrt{2}V_t^{\alpha\alpha}\, dB_t\,, \tag{4.1}$$

which implies by Proposition 3.7 that V_t has a finite time explosion (with non-zero probability) **if and only if** $\frac{3}{4}\varphi''(0)^2 + \varphi'''(0) > 0$. This criterion can be used to help choose how activation functions should be centered for shaping; below are two examples.

Example 4.1 (Sigmoid and $\tanh$ at $x_0 = 0$)**.** *We start with the sigmoid activation* $\sigma(x) = \frac{1}{1+e^{-x}}$*, then we can define* $\varphi(x) \coloneqq 4\sigma(x) - 2$ *to satisfy Assumption 3.5, which leads to* $\varphi''(0) = 0, \varphi'''(0) = -\frac{1}{2}$*, and therefore leads to a stable network. It turns out* $\varphi(x) \coloneqq \tanh(x)$ *already satisfies Assumption 3.5, 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 4.2 (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 3.5. This gives us $\varphi''(0) = \frac{1}{1+e^{x_0}}, \varphi'''(0) = \frac{1-e^{x_0}}{(1+e^{x_0})^2}$, and therefore $\frac{3}{4}\varphi''(0)^2 + \varphi'''(0) = \frac{1}{(1+e^{x_0})^2}\left(\frac{5}{4} - e^{x_0}\right)$. In other words, the shaped network is stable if and only if $x_0 \geq \log \frac{5}{4}$ (see Figure 2). We note that the authors of [38] numerically found a shift of $x_0 \approx 0.41$, which is in the stable regime of $x_0 \geq \log \frac{5}{4} \approx 0.097$.*

Relationship to Edge of Chaos. The finite time explosion example above resembles the Edge of Chaos (EOC) analysis of gradient stability [43, 35, 44, 45], where the weight and bias variance at initialization determines a stability criterion. However, we note that the EOC regime is sufficiently different that the results are not directly comparable. More precisely, the EOC analysis is in the sequential limit of infinite-width and then infinite-depth, which also leaves the activation function unchanged. Under very weak assumptions, the variance (diagonal of V_t) will not explode in this regime; instead, the gradient can explode due to the covariance (off diagonals). On the other hand, our finite explosion result is in the joint limit of depth and width, where the variance (diagonal of V_t) can explode instead.

Posterior Inference. Similar to the NNGP setting, we can use the Neural Covariance SDE to generate a prior over functions $f : \mathbb{R}^{n_{\text{in}}} \to \mathbb{R}^{n_{\text{out}}}$. Consequently, an interesting future direction would be to study the posterior distribution, i.e. the output z_{out}^{m+1} conditioned on x^{m+1} and a training dataset $(x^\alpha, z_{\text{out}}^\alpha)_{\alpha=1}^m$. However, to our best knowledge, it is not straightforward to explicitly compute or sample from the conditional distributions for this SDE structure. It would be desirable to extend existing approaches in the perturbative regime [30, 31] to our setting.

Extension to Other Architectures. The key step to deriving the covariance SDE is the conditional Gaussian distribution in (2.7), which directly leads to a Markov chain. It follows immediately that ResNets [46] admit a similar conditional structure. With a bit more work for convolutional networks, we can obtain $z_{\ell+1}^\alpha | \mathcal{F}_\ell \sim \mathcal{N}(0, \mathcal{A}(V_\ell) \otimes I_n)$ where $\mathcal{A}$ is an affine transformation and V_ℓ is the previous layer's Gram matrix [47]. We note that recurrent networks will not lead to a Markov chain or SDE limit, as the weight matrix is reused from layer to layer.

Simulating SDEs. Both the Markov chains and SDEs predict neural networks at initialization very well (see Figure 1), but the SDE is significantly faster to simulate. In particular, we can view the Markov chain as an approximate Euler discretization of the SDE, but with a very small step size n^{-1}. In contrast, to simulate the SDE we should only need a step size that is small on the scale of depth-to-width ratio $T = d/n$, which is *independent of width* n. Therefore, practitioners using the shaping techniques of [38, 39] 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 F).

Analytical Tractability of SDEs. Besides numerical tractability, the SDEs are also far more tractable to analyze. For example, in the one input case, we arrive at geometric Brownian motion (2.6), which is known to have a log-normal distribution at fixed times. Similarly, our finite time explosions hinge on the fact we identified an SDE limit. In the same way that NNGP theory played a major role in the infinite-width regime, the Neural Covariance SDEs and the techniques developed here also serve as a mathematical foundation for studying training and generalization.

Acknowledgement

We would like to thank Sinho Chewi, James Foster, Boris Hanin, Cameron Jakub, Jeffrey Negrea, Nuri Mert Vural, Guodong Zhang, Matthew S. Zhang, and Yuchong Zhang for helpful discussions and draft feedback. We would like to thank Sam Buchanan and Soufiane Hayou for pointing out a gap in the proof of Proposition B.8. ML is supported by Ontario Graduate Scholarship and the Vector Institute. MN is supported by an NSERC Discovery Grant. DMR is supported in part by Canada CIFAR AI Chair funding through the Vector Institute, an NSERC Discovery Grant, Ontario Early Researcher Award, a stipend provided by the Charles Simonyi Endowment, and a New Frontiers in Research Exploration Grant.

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] S. Yaida. "Non-Gaussian processes and neural networks at finite widths." *Mathematical and Scientific Machine Learning*. PMLR. 2020, pp. 165–192.

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

[32] 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).

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

[34] 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=O-6Pm_d_Q-.

[35] 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.

[36] 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.

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

[38] 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).

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

[40] 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.

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

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

[43] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. "Deep information propagation." In: *arXiv preprint arXiv:1611.01232* (2016).

[44] S. Hayou, A. Doucet, and J. Rousseau. "On the impact of the activation function on deep neural networks training." *International conference on machine learning*. PMLR. 2019, pp. 2672–2680.

[45] M. Murray, V. Abrol, and J. Tanner. "Activation function design for deep networks: linearity and effective initialisation." In: *Applied and Computational Harmonic Analysis* 59 (2022), pp. 117–154.

[46] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[47] R. Novak, L. Xiao, J. Lee, Y. Bahri, G. Yang, J. Hron, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein. "Bayesian deep convolutional networks with many channels are gaussian processes." In: *arXiv preprint arXiv:1810.05148* (2018).

[48] O. Kallenberg. *Foundations of Modern Probability*. Probability theory and stochastic modelling. Springer, 2021. ISBN: 9783030618728.

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

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

[51] A. Meurer, C. P. Smith, M. Paprocki, O. ertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel,. Rouka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, and A. Scopatz. "SymPy: symbolic computing in Python." In: *PeerJ Computer Science* 3 (Jan. 2017), e103. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.103. URL: https://doi.org/10.7717/peerj-cs.103.

[52] 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.

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 3.5.
   (b) Did you include complete proofs of all theoretical results? [Yes] Please see the relevant sections in Appendices A, C and D.
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]

A Background on Markov Chain Convergence to SDEs

In this section we briefly review the background and technical results required to characterize the convergence of a Markov chain to an SDE. Majority of the content in this section are based on [48–50].

To start we first introduce the Skorohod J_1-topology [48, Appendix 5]. Let S be a complete separable metric space, and $D_{\mathbb{R}_+,S}$ be the space of càdlàg functions (right continuous with left limits) from $\mathbb{R}_+ \to S$. Here we write $x_n \xrightarrow{ul} x$ to denote locally uniform convergence (i.e., uniform on compact subsets of $\mathbb{R}_+$). We also consider bijections λ on $\mathbb{R}_+$ so that λ is strictly increasing with $\lambda_0 = 0$. We can now define *Skorohod convergence* $x_n \xrightarrow{s} x$ on $D_{\mathbb{R}_+,S}$ if there exists a sequence of bijections λ_n satisfying the above conditions and

$$\lambda_n \xrightarrow{ul} \mathrm{Id}\,, \quad x_n \circ \lambda_n \xrightarrow{ul} x\,. \tag{A.1}$$

The most important result is that $D_{\mathbb{R}_+,S}$ equipped with the above sense of convergence is indeed a well behaved probability space, which we state below.

Theorem A.1 (Theorem A5.3, [48])**.** *For any separable complete metric space S, there exists a topology $\mathcal{T}$ on $D_{\mathbb{R}_+,S}$ such that*

(i) $\mathcal{T}$ induces the Skorohod convergence $x_n \xrightarrow{s} x$,

(ii) $D_{\mathbb{R}_+,S}$ is Polish (separable completely metrizable topological space) under $\mathcal{T}$,

(iii) $\mathcal{T}$ generates the Borel σ-field generated by the evaluation maps π_t, $t \geq 0$, where $\pi_t(x) = x_t$.

We also need to define Feller semi-groups. To start we let S be a locally compact separable metric space and $C_0 := C_0(S)$ be the space of continuous functions that vanishes at infinity, and we equip C_0 with the sup norm to make it a Banach space. $T : C_0 \to C_0$ is a *positive contraction operator* if for all $0 \leq f \leq 1$ we have $0 \leq Tf \leq 1$. A semi-group of such operators (T_t) on C_0 is called a *Feller semi-group* if it additionally satisfies

$$\begin{aligned} &T_t C_0 \subset C_0\,, \quad t \geq 0\,, \\ &T_t f(x) \to x \text{ as } t \to 0\,, \quad f \in C_0, x \in S\,. \end{aligned} \tag{A.2}$$

Let $\mathcal{D} \subset C_0$ and $A : \mathcal{D} \to C_0$, and we say that $(A, \mathcal{D})$ is a *generator* of (T_t) if $\mathcal{D}$ is the maximal set such that for all $f \in \mathcal{D}$, we have that

$$\lim_{t \to 0} \frac{T_t f - f}{t} = Af\,. \tag{A.3}$$

An operator A with domain $\mathcal{D}$ on a Banach space B is said to be *closed*, if its graph $G = \{(f, Af) | f \in \mathcal{D}\}$ is a closed subset of $B \times B$. If the closure of G is the graph of an operator

$\bar{A}$, we say $\bar{A}$ is the *closure* of A. Finally, we will define a linear subspace $D \subset \mathcal{D}$ as a **core** of A if the closure of $A|_D$ is A. If $(A, \mathcal{D})$ is a generator of a Feller semigroup, every dense invariant subspace $D \subset D$ is a core of A [48, Proposition 17.9]. In particular, we will work with the core C_0^∞ of smooth functions vanishing at infinity.

We will state a sufficient condition required for an semi-group to be Feller based on its generator.

Theorem A.2 (Section 8, Theorem 2.5, [49]). *Let $a^{ij} \in C^2(\mathbb{R}^d)$ with $\partial_k \partial_\ell a^{ij}$ be bounded for all $i, j, k, \ell \in [d]$. Further let $b : \mathbb{R}^d \to \mathbb{R}^d$ be Lipschitz. Then the generator defined by*

$$Af = \frac{1}{2} \sum_{i,j=1}^{d} a^{ij} \partial_i \partial_j f + \sum_{i=1}^{d} b^i \partial_i f \,, \tag{A.4}$$

generates a Feller semi-group on C_0.

We will next state a set of equivalent criterion for convergence of Feller processes.

Theorem A.3 (Theorem 17.25, [48]). *Let $X, X^1, X^2, X^3, \cdots$ be Feller processes in S with semi-groups $(T_t), (T_{n,t})$ and generators $(A, \mathcal{D}), (A_n, \mathcal{D}_n)$, respectively, and fix a core D for A. Then these conditions are equivalent:*

(i) for any $f \in D$, there exists some $f_n \in \mathcal{D}_n$ with $f_n \to f$ and $A_n f_n \to Af$,

(ii) $T_{n,t} \to T_t$ strongly for each $t > 0$,

(iii) $T_{n,t} f \to T_t f$ for every $f \in C_0$, uniformly for bounded $t > 0$,

(iv) $X_0^n \xrightarrow{d} X_0$ in $S \Rightarrow X^n \xrightarrow{d} X$ in the Skohorod topology of $D_{\mathbb{R}_+, S}$.

Once again, we note that it is common to choose the core $D = C_0^\infty$, and that checking condition (i) is sufficient for convergence in the Skorohod topology. This is translated to the Markov chain setting by the next theorem.

Theorem A.4 (Theorem 17.28, [48]). *Let $Y^1, Y^2, Y^3, \cdots$ be discrete time Markov chains in S with transition operators $U_1, U_2, U_3, \cdots$, and let X be a Feller process with semi-group (T_t) and generator A. Fix a core D for A, and let $0 < h_n \to 0$. Then conditions $(i) - (iv)$ of Theorem A.3 remain equivalent for the operators and processes*

$$A_n = h_n^{-1}(U_n - I) \,, \quad T_{n,t} = U_n^{\lfloor t/h_n \rfloor} \,, \quad X_t^n = Y_{\lfloor t/h_n \rfloor}^n \,. \tag{A.5}$$

It remains to check that the generators A_n converges to A with respect to the core $D = C_0^\infty$, and we will use a criterion from [50]. Here we will first let $\Pi_n(x, dy)$ be the Markov transition kernel of Y^n, and define

$$\begin{aligned} a_n^{ij}(x) &= \frac{1}{h_n} \int_{|y-x| \le 1} (y_i - x_i)(y_j - x_j) \, \Pi_n(x, dy) \,, \\ b_n^i(x) &= \frac{1}{h_n} \int_{|y-x| \le 1} (y_i - x_i) \, \Pi_n(x, dy) \,, \\ \Delta_n^\epsilon(x) &= \frac{1}{h_n} \Pi_n(x, \mathbb{R}^d \setminus B(x, \epsilon)) \,. \end{aligned} \tag{A.6}$$

Lemma A.5 (Lemma 11.2.1, [50]). *The following two conditions are equivalent:*

(i) For any $R > 0, \epsilon > 0$ we have that

$$\lim_{n \to \infty} \sup_{|x| \le R} \|a_n(x) - a(x)\|_{op} + |b_n(x) - b(x)| + \Delta_n^\epsilon(x) = 0 \,, \tag{A.7}$$

(ii) For each $f \in C_0^\infty(\mathbb{R}^d)$, we have that

$$\frac{1}{h_n} A_n f \to Af \,, \tag{A.8}$$

uniformly on compact sets of $\mathbb{R}^d$, where A is defined as (A.4).

Finally, we summarize the above results in a user friendly form for our applications.

Proposition A.6 (Convergence of Markov Chains to SDE)**.** *Let Y^n be a discrete time Markov chain on $\mathbb{R}^N$ defined by the following update for $p, \delta > 0$*

$$Y^n_{\ell+1} = Y^n_\ell + \frac{\widehat{b}_n(Y^n_\ell, \omega^n_\ell)}{n^{2p}} + \frac{\sigma_n(Y^n_\ell)}{n^p}\xi^n_\ell + O(n^{-2p-\delta}), \tag{A.9}$$

where $\xi^n_\ell \in \mathbb{R}^N$ are iid random variables with zero mean, identity covariance, and moments uniformly bounded in n. Furthermore, ω^n_ℓ are also iid random variables such that $\mathbb{E}[\widehat{b}_n(Y^n_\ell, \omega^n_\ell)|Y^n_\ell = y] = b_n(y)$ and $\widehat{b}_n(y, \omega^n_\ell)$ has uniformly bounded moments in n. Finally, σ_n is a deterministic function, and the remainder terms in $O(n^{-2p-\delta})$ have uniformly bounded moments in n.

Suppose b_n, σ_n are uniformly Lipschitz functions in n and converges to b, σ uniformly on compact sets, then in the limit as $n \to \infty$, the process $X^n_t = Y^n_{\lfloor tn^{2p} \rfloor}$ converges in distribution to the solution of the following SDE in the Skorohod topology of $D_{\mathbb{R}_+,\mathbb{R}^N}$

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t, \quad X_0 = \lim_{n\to\infty} Y^n_0. \tag{A.10}$$

Suppose otherwise b_n, σ_n are only locally Lipschitz (but still uniform in n), then X^n converges locally to X in the same topology (see Definition 3.8). More precisely, for any fixed $r > 0$, we consider the stopping times

$$\tau^n \coloneqq \inf\{t \geq 0 : |X^n_t| \geq r\}, \quad \tau \coloneqq \inf\{t \geq 0 : |X_t| \geq r\}, \tag{A.11}$$

then the stopped process $X^n_{t\wedge\tau^n}$ converges in distribution to the stopped solution $X_{t\wedge\tau}$ of the above SDE in the same topology.

Proof. We will essentially check the criterion of Theorem A.4 directly for the metric space $S = \mathbb{R}^N$ if b, σ is globally Lipschitz, and $S = B(0, r)$ otherwise. In both of these cases, b, σ are Lipschitz on S, therefore the limiting process (either X_t or $X_{t\wedge\tau}$) is Feller in S by Theorem A.2.

In the equivalent criteria of Theorem A.3, we will use the implication of (i) $\Rightarrow$ (iv) to get convergence of X^n to X in the Skorohod topology of $D_{\mathbb{R}_+,S}$. More precisely, it is sufficient to choose $h_n = \frac{1}{n^{2p}}$ as the natural time scale, and check $\frac{1}{h_n}A_n f \to Af$ for any $f \in C^\infty_0$. Given Lemma A.5, it is sufficient to check the convergence of the coefficients and Δ^ϵ_n.

We start with $\Delta^\epsilon_n(x)$. Given that the randomness in the Markov chain have bounded moments (uniform in n), then by a Markov inequality we have that for any $q > 0$

$$\Pi_n(x, \mathbb{R}^d \setminus B(x,\epsilon)) = \mathbb{P}\left[\left|\frac{\widehat{b}(x,\omega^n_\ell)}{n^{2p}} + \frac{\sigma}{n^p}\xi^n_\ell + O(n^{-2p-\delta})\right|^{2q} \geq \epsilon^{2q}\right] \leq O\left(\epsilon^{-2q}n^{-2pq}\right), \tag{A.12}$$

therefore choosing $q > 1$ we have $\sup_{|x|\leq R}\Delta^\epsilon_n(x) = O(n^{-2p(q-1)}) \to 0$ for any fixed ϵ.

We can rewrite $b_n(x)$ as

$$b_n(x) = n^{2p}\mathbb{E}[Y^n_{\ell+1} - Y^n_\ell | Y^n_\ell = x] + O(n^{-\delta}) \to b(x), \tag{A.13}$$

since ξ^n_ℓ has zero mean and the remainder terms have bounded moments (uniform in n), which also gives the desired convergence of $\sup_{x\leq|R|}|b_n(x) - b(x)| \to 0$.

Similarly we can rewrite $a_n(x)$ as

$$a_n(x) = n^{2p}\mathbb{E}[(Y^n_{\ell+1} - Y^n_\ell)(Y^n_{\ell+1} - Y^n_\ell)^\top | Y^n_\ell = x] + O(n^{-2\delta} + n^{-2p}) \to \sigma(x)\sigma(x)^\top, \tag{A.14}$$

where we note the drift's randomness contributes the higher order n^{-2p} term and therefore also vanishes in the limit. This implies $\sup_{x\leq|R|}\|a_n(x) - a(x)\|_{op} \to 0$, which gives us the desired result.

□

B Unshaped ReLU Markov Chain

In this section, we will derive the Markov chain update (2.10) with explicit coefficients. For the rest of this section, we will adopt the following notation. Let $\varphi(x) \coloneqq \max(x, 0)$ be the ReLU activation function. Let $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ be the density of a standard Gaussian, and let $F(x) = \int_{-\infty}^{x} f(t)\, dt$ be the cumulative distribution function (CDF).

Lemma B.1 (Gaussian Integration-by-Parts with Indicator Function)**.** *For $g \sim \mathcal{N}(0,1)$ and h is weakly differentiable, we have that*

$$\mathbb{E}\, g \mathbb{1}_{\{g>-a\}} h(g) = h(-a) f(a) + \mathbb{E}\, \mathbb{1}_{\{g>-a\}} h'(g)\,, \tag{B.1}$$

where f is the standard Gaussian density.

Proof. We start by writing the expectation as an integral

$$\mathbb{E}\, g \mathbb{1}_{\{g>-a\}} h(g) = \int_{-a}^{\infty} x h(x) f(x)\, dx\,. \tag{B.2}$$

Here by observing that $f'(x) = -xf(x)$, we can use integration by parts for $u = h(x), dv = xf(x)\, dx$ to get $du = h'(x)\, dx, v = -f(x)$, and therefore

$$\int_{-a}^{\infty} x h(x) f(x)\, dx = [-h(x) f(x)]_{-a}^{\infty} + \int_{-a}^{\infty} h'(x) f(x)\, dx = h(-a) f(-a) + \mathbb{E}\, \mathbb{1}_{\{g>-a\}} h'(g)\,. \tag{B.3}$$

Finally we recover the desired result using symmetry of $f(-a) = f(a)$.

□

We will note the special case of $a = 0$ to get

$$\mathbb{E}\, g \mathbb{1}_{\{g>0\}} h(g) = \frac{h(0)}{\sqrt{2\pi}} + \mathbb{E}\, \mathbb{1}_{\{g>0\}} h'(g)\,. \tag{B.4}$$

Lemma B.2 (Gaussian Density Substitution)**.** *Let $g \sim \mathcal{N}(0,1), \rho \in [0,1], q = \sqrt{1-\rho^2}$, then we have that*

$$\mathbb{E}\, h(g) f\left(\frac{\rho g + a}{q}\right) = q f(a) \mathbb{E}\, h(qg - \rho a)\,. \tag{B.5}$$

Proof. We will again write the expectation as an integral

$$\mathbb{E}\, h(g) f\left(\frac{\rho g + a}{q}\right) = \int h(x) f\left(\frac{\rho x + a}{q}\right) f(x)\, dx\,. \tag{B.6}$$

Here observe that

$$f\left(\frac{\rho x + a}{q}\right) f(x) = \frac{1}{2\pi} \exp\left[-\frac{(\rho x + a)^2}{2q^2} - \frac{x^2}{2}\right] = \frac{1}{2\pi} \exp\left[-\frac{\rho^2 x^2 + a^2 + 2a\rho x + q^2 x^2}{2q^2}\right], \tag{B.7}$$

at this point, we can complete the square to write

$$\rho^2 x^2 + a^2 + 2a\rho x + q^2 x^2 = (x + a\rho)^2 - a^2\rho^2 + a = (x + a\rho)^2 - a^2 q^2\,. \tag{B.8}$$

This implies that we have

$$f\left(\frac{\rho x + a}{q}\right) f(x) = \frac{1}{2\pi} \exp\left[-\frac{(x + a\rho)^2}{2q^2} - \frac{a^2}{2}\right] = f\left(\frac{x + a\rho}{q}\right) f(a)\,. \tag{B.9}$$

Finally, we can use the substitution $y = \frac{x+a\rho}{q}, dy = \frac{1}{q} dx$ to get

$$\int h(x) f\left(\frac{\rho x + a}{q}\right) f(x)\, dx = \int h(qy - \rho a) f(y) f(a) q\, dy = q f(a) \mathbb{E}\, h(qg - \rho a)\,, \tag{B.10}$$

which is the desired result.

□

We will start by calculating simpler quantities.

Lemma B.3 (Moments)**.** *Let $g \sim \mathcal{N}(0,1)$, then*

$$\mathbb{E}\,\varphi(g) = \frac{1}{\sqrt{2\pi}}\,, \quad \mathbb{E}\,\varphi(g)^2 = \frac{1}{2}\,, \quad \mathbb{E}\,\varphi(g)^4 = \frac{3}{2}\,. \tag{B.11}$$

Proof. For the second and fourth moments, we simply observe that g^2 is symmetric and φ is exactly half of of the integral. For the first integral we will use Gaussian integration-by-parts with $h(g) = 1$ to get

$$\mathbb{E}\,\varphi(g) = \mathbb{E}\,g\mathbb{1}_{\{g>0\}} = \frac{1}{\sqrt{2\pi}}\,, \tag{B.12}$$

which is the desire result.

□

We will also recall the following result from [41]

Lemma B.4 ($\bar{J}_0, \bar{J}_1, \bar{J}_2$)**.** *Let $\rho \in [0,1], q = \sqrt{1-\rho^2}$ and let $\rho, w \sim \mathcal{N}(0,1)$ be independent. Then we have that*

$$\begin{aligned}
\bar{J}_0(\rho) &= \mathbb{E}\,\mathbb{1}_{\{g>0\}}\mathbb{1}_{\{\rho g+qw>0\}} = \frac{\arccos(-\rho)}{2\pi}\,,\\
\bar{J}_1(\rho) &= \mathbb{E}\,\varphi(g)\varphi(\rho g+qw) = \frac{q+\rho\arccos(-\rho)}{2\pi}\,,\\
\bar{J}_2(\rho) &= \mathbb{E}\,\varphi(g)^2\varphi(\rho g+qw)^2 = \frac{3\rho q+\arccos(-\rho)(1+2\rho^2)}{2\pi}\,.
\end{aligned} \tag{B.13}$$

We will need to compute the following quantity.

Lemma B.5 ($\bar{J}_{3,1}$)**.** *Let $\rho \in [0,1], q = \sqrt{1-\rho^2}$ and let $\rho, w \sim \mathcal{N}(0,1)$ be independent. Then we have that*

$$\bar{J}_{3,1}(\rho) = \mathbb{E}\,\varphi(g)^3\varphi(\rho g+qw) = \frac{q(2+\rho^2)+3\arccos(-\rho)\rho}{2\pi}\,. \tag{B.14}$$

Proof. We start by using Gaussian integration-by-parts with $h(g) = \mathbb{E}_g\, g^2\varphi(\rho g+qw)$ where we use $\mathbb{E}_g[\,\cdot\,] \coloneqq \mathbb{E}\,[\,\cdot\,|g]$ to denote conditional expectation

$$\begin{aligned}
\mathbb{E}\,\varphi(g)^3\varphi(\rho g+qw) &= \mathbb{E}\,g\mathbb{1}_{\{g>0\}}h(g)\\
&= \mathbb{E}\,\mathbb{1}_{\{g>0\}}\left[2g\mathbb{E}_g\,\varphi(\rho g+qw) + g^2\mathbb{E}_g\rho\mathbb{1}_{\{\rho g+qw>0\}}\right]\\
&= 2\bar{J}_1(\rho) + \rho\mathbb{E}\,g\mathbb{1}_{\{g>0\}}\mathbb{E}_g g\mathbb{1}_{\{\rho q+qw>0\}}\,.
\end{aligned} \tag{B.15}$$

Here we observe that $\mathbb{E}_g g\mathbb{1}_{\rho q+qw>0} = gF(\rho g/q)$, and we can again set this to the new $h(g)$ and use integration-by-parts to write

$$\mathbb{E}\,g\mathbb{1}_{\{g>0\}}\mathbb{E}_g g\mathbb{1}_{\{\rho q+qw>0\}} = \mathbb{E}\,\mathbb{1}_{\{g>0\}}\mathbb{1}_{\{\rho g+qw>0\}} + \mathbb{E}\,\mathbb{1}_{\{g>0\}}\frac{\rho g}{q}f\left(\frac{\rho g}{q}\right)\,. \tag{B.16}$$

At this point we can use the substitution formula from Lemma B.2 to write

$$\mathbb{E}\,\mathbb{1}_{\{g>0\}}\frac{\rho g}{q}f\left(\frac{\rho g}{q}\right) = \frac{\rho}{q}f(0)\mathbb{E}\varphi(qg) = \frac{\rho q}{2\pi}\,. \tag{B.17}$$

Putting this together, we have

$$\bar{J}_{3,1}(\rho) = 2\bar{J}_1(\rho) + \rho\bar{J}_0(\rho) + \frac{\rho^2 q}{2\pi}\,, \tag{B.18}$$

which is the desired result after simplifying.

□

We will now recall the ReLU-like activations for $s = (s_+, s_-) \in \mathbb{R}^2$

$$\varphi_s(x) := s_+ \max(x, 0) + s_- \min(x, 0) = s_+\varphi(x) - s_-\varphi(-x)\,, \tag{B.19}$$

where $\varphi(x) := \max(x, 0)$ is the usual ReLU activation.

We will compute several basic moments first.

Lemma B.6 (Moments, c, M_2)**.** *Let $g \sim N(0, 1)$, we have that*

$$\mathbb{E}\,\varphi_s(g) = \frac{s_+ - s_-}{\sqrt{2\pi}}\,, \quad \mathbb{E}\,\varphi_s(g)^2 = \frac{s_+^2 + s_-^2}{2}\,, \quad \mathbb{E}\,\varphi_s(g)^4 = \frac{3}{2}(s_+^4 + s_-^4)\,. \tag{B.20}$$

Furthermore, this implies the normalizing constant is $c = \frac{2}{s_+^2+s_-^2}$ and

$$M_2 \coloneqq \mathbb{E}\,[c\varphi_s(g)^2 - 1]^2 = 6\frac{s_+^4 + s_-^4}{(s_+^2 + s_-^2)^2} - 1\,. \tag{B.21}$$

Proof. To start we first recall the Gaussian integration by parts calculation

$$\mathbb{E}\,\varphi(g) = f(0) = \frac{1}{\sqrt{2\pi}}\,, \tag{B.22}$$

then the first moment follows immediately from rewriting in terms of φ

$$\mathbb{E}\,\varphi_s(g) = s_+\mathbb{E}\,\varphi(g) - s_-\mathbb{E}\,\varphi(-g) = \frac{s_+ - s_-}{\sqrt{2\pi}}\,. \tag{B.23}$$

For the second moment, we will also rewrite in terms of φ

$$\mathbb{E}\,\varphi_s(g)^2 = \mathbb{E}\,s_+^2\varphi(g)^2 + s_-^2\varphi(-g)^2 - 2s_+s_-\varphi(g)\varphi(-g) = (s_+^2 + s_-^2)\mathbb{E}\,\varphi(g)^2\,, \tag{B.24}$$

where we used that $\varphi(g)\varphi(-g) = 0$ almost surely and $g \overset{d}{=} -g$, and the desire result follows from Gaussian integration by parts

$$\mathbb{E}\,\varphi(g)^2 = 0f(0) + \mathbb{E}\,\mathbb{1}_{\{g>0\}} = \frac{1}{2}\,. \tag{B.25}$$

For the fourth moment, we will similarly observe that all mixed moments $\varphi(g)^p\varphi(-g)^r = 0$ almost surely whenever $p, r > 0$, which allows us to write

$$\mathbb{E}\,\varphi_s(g)^4 = \mathbb{E}\,s_+^4\varphi(g)^4 + s_-^4\varphi(-g)^4 = (s_+^4 + s_-^4)\mathbb{E}\,\varphi(g)^4\,, \tag{B.26}$$

and the desire result follows from the Gaussian integration by parts calculation

$$\mathbb{E}\,\varphi(g)^4 = 0^3 f(0) + \mathbb{E}\,3g^2\mathbb{1}_{\{g>0\}} = 3(0^3 f(0) + \mathbb{E}\,\mathbb{1}_{\{g>0\}}) = \frac{3}{2}\,. \tag{B.27}$$

□

We will also convert the $\bar{J}_{k,\ell}$ formulas to $K_{k,\ell}$ formulas, i.e. the following quantities

$$K_{p,r}(\rho) := \mathbb{E}\,\varphi_s(g)^p\varphi_s(\hat{g})^r\,, \tag{B.28}$$

where $g, w \sim N(0, 1)$ and we define $\hat{g} = \rho g + qw$ with $q = \sqrt{1 - \rho^2}$. We will also use the short hand notation to write $\bar{J}_p := \bar{J}_{p,p}, K_p := K_{p,p}$.

Lemma B.7 ($K_1, K_2, K_{3,1}$)**.** *Let $\rho \in [-1, 1]$, $q = \sqrt{1 - \rho^2}$, $g, w \sim \mathcal{N}(0, 1)$, and $\hat{g} = \rho g + qw$. Then we have the following formulas*

$$\begin{aligned}
K_1(\rho) &= (s_+^2 + s_-^2)\bar{J}_1(\rho) - 2s_+s_-\bar{J}_1(-\rho)\,,\\
K_2(\rho) &= (s_+^4 + s_-^4)\bar{J}_2(\rho) + 2s_+^2s_-^2\bar{J}_2(-\rho)\,,\\
K_{3,1}(\rho) &= (s_+^4 + s_-^4)\bar{J}_{3,1}(\rho) - s_+s_-(s_+^2 + s_-^2)\bar{J}_{3,1}(-\rho)\,.
\end{aligned} \tag{B.29}$$

Proof. Before we start, we will make several observations. Using the fact that $(g, w) \stackrel{d}{=} (\pm g, \pm w)$, we have the following equality in distribution relations

$$\begin{aligned}(g, \rho g + qw) &\stackrel{d}{=} (-g, -\rho g - qw) = (-g, -\hat{g})\,, \\ (g, -\hat{g}) &\stackrel{d}{=} (g, -\rho g + qw) \stackrel{d}{=} (-g, \rho g + qw) = (-g, \hat{g})\,.\end{aligned} \tag{B.30}$$

In particular, we note that the two Gaussian random variable $(g, -\hat{g})$ have correlation $-\rho$.

This allows us to simplify K_1

$$\begin{aligned}K_1(\rho) &= \mathbb{E}\,\varphi_s(g)\varphi_s(\hat{g}) \\ &= \mathbb{E}\,s_+^2\varphi(g)\varphi(\hat{g}) + s_-^2\varphi(-g)\varphi(-\hat{g}) - s_+s_-\varphi(g)\varphi(-\hat{g}) - s_+s_-\varphi(-g)\varphi(\hat{g}) \\ &= (s_+^2 + s_-^2)\bar{J}_1(\rho) - 2s_+s_-\bar{J}_1(-\rho)\,,\end{aligned} \tag{B.31}$$

which is the desired result.

With K_2, we will additionally make use of the fact that $\varphi(g)\varphi(-g) = 0$ almost surely to write

$$\begin{aligned}K_2(\rho) &= \mathbb{E}\,(s_+^2\varphi(g)^2 + s_-^2\varphi(-g)^2)(s_+^2\varphi(\hat{g})^2 + s_-^2\varphi(-\hat{g})^2) \\ &= \mathbb{E}\,s_+^4\varphi(g)^2\varphi(\hat{g})^2 + s_-^4\varphi(-g)^2\varphi(-\hat{g})^2 + s_+^2s_-^2\varphi(g)^2\varphi(-\hat{g})^2 + s_+^2s_-^2\varphi(-g)^2\varphi(\hat{g})^2 \\ &= (s_+^4 + s_-^4)\bar{J}_2(\rho) + 2s_+^2s_-^2\bar{J}_2(-\rho)\,.\end{aligned} \tag{B.32}$$

$K_{3,1}$ follows from a similar calculation

$$\begin{aligned}K_{3,1}(\rho) &= \mathbb{E}\,(s_+^3\varphi(g)^3 - s_-^3\varphi(-g)^3)(s_+\varphi(\hat{g}) - s_-\varphi(-\hat{g})) \\ &= \mathbb{E}\,s_+^4\varphi(g)^3\varphi(\hat{g})s_-^4\varphi(-g)^3\varphi(-\hat{g}) - s_+^3s_-\varphi(g)^3\varphi(-\hat{g}) - s_+s_-^3\varphi(-g)^3\varphi(\hat{g}) \\ &= (s_+^4 + s_-^4)\bar{J}_{3,1}(\rho) - s_+s_-(s_+^2 + s_-^2)\bar{J}_{3,1}(-\rho)\,.\end{aligned} \tag{B.33}$$

□

Finally, we to get to state the desired formulas for the approximate Markov chain. Here we will make introduce several definitions first. In the event that $|\varphi_\ell^\alpha| = 0$ or $|\varphi_\ell^\beta| = 0$, the formula $\rho_\ell^{\alpha\beta} \coloneqq \frac{\langle\varphi_\ell^\alpha, \varphi_\ell^\beta\rangle}{|\varphi_\ell^\alpha|\,|\varphi_\ell^\alpha|}$ is undefined. We will remedy this by introducing an additional point $\mathbf{e}$ in the state space $\mathbb{R} \cup \{\mathbf{e}\}$, and set $\rho_\ell^{\alpha\beta} = \mathbf{e}$ in this event. We note that once $\rho_\ell^{\alpha\beta} = \mathbf{e}$, then the next step $\rho_{\ell+1}^{\alpha\beta} = \mathbf{e}$ as well since either $z_{\ell+1}^\alpha, z_{\ell+1}^\beta = 0$. For all $x \in \mathbb{R}$ we will define the distance $|x - \mathbf{e}| = \infty$. Consequently, $\mathbb{R} \cup \{\mathbf{e}\}$ is a Polish space (complete separable metric space), and therefore it's a well behaved probability space (e.g. admits conditional densities). For a random variable X, we write $X = O(n^p)$ if all moments of $n^{-p}X$ are bounded by a constant independent of n.

We will also define the bounded Lipschitz function norm as

$$\|h\|_{BL} \coloneqq \|h\|_\infty + \sup_{x \neq y} \frac{|h(x) - h(y)|}{|x - y|}\,, \tag{B.34}$$

which induces the bounded Lipschitz distance for probability measures

$$d_{BL}(\mu, \nu) \coloneqq \sup_{\|h\|_{BL} \leq 1} \int h\,d\mu - \int h\,d\nu\,. \tag{B.35}$$

Proposition B.8 (Unshaped ReLU Correlation)**.** *Let $\rho_\ell^{\alpha\beta} \coloneqq \frac{\langle\varphi_\ell^\alpha, \varphi_\ell^\beta\rangle}{|\varphi_\ell^\alpha|\,|\varphi_\ell^\alpha|}$ when defined, and $\mathbf{e}$ when either $|\varphi_\ell^\alpha|\,, |\varphi_\ell^\alpha| = 0$. Let us also define the approximate Markov chain*

$$p_{\ell+1} = cK_1(p_\ell) + \frac{\mu_{ReLU}(p_\ell)}{n} + \sigma_{ReLU}(p_\ell)\frac{z_\ell}{\sqrt{n}}\,, \tag{B.36}$$

where z_ℓ are iid $\mathcal{N}(0, 1)$ and

$$\begin{aligned}\mu_{ReLU}(\rho_\ell^{\alpha\beta}) &= \frac{c}{4}\left[K_1(c^2K_2 + 3M_2 + 3) - 4cK_{3,1}\right]\,, \\ \sigma_{ReLU}^2(\rho_\ell^{\alpha\beta}) &= \frac{c^2}{2}\left[K_1^2(c^2K_2 + M_2 + 1) - 4cK_1K_{3,1} + 2K_2\right]\,,\end{aligned} \tag{B.37}$$

where we write $K_\cdot = K_\cdot(\rho_\ell^{\alpha\beta})$*, and the formulas for* $K_1, K_2, K_{3,1}, c, M_2$ *are calculated in Lemma B.6 and Lemma B.7.*

Let $\Pi(x, dy), P(x, dy)$ *be the Markov transition kernels of* $\rho_\ell^{\alpha\beta}$ *and* p_ℓ *respectively, then*

$$d_{BL}(\Pi(x,\cdot), P(x,\cdot)) = O(n^{-1}), \quad \text{for all } x \in \mathbb{R} \cup \{\mathbf{e}\}. \tag{B.38}$$

Remark B.9. The infinite-width ($n \to \infty$) approximation of the Markov chain corresponds to the update $q_{\ell+1} = cK_1(q_\ell)$, and this is an $O(n^{-1/2})$ approximation to the chain $\{\rho_\ell^{\alpha\beta}\}$. On the other hand, the $\{p_\ell\}$ chain we propose is an *improved approximation* up to the zero mean terms up to $O(n^{-1/2})$, and the expected value of non-zero mean terms up to $O(n^{-1})$. In the SDE limit of Proposition A.6, these are exactly the terms that do not vanish, which leads us to speculate that this approximation is sufficiently close when studying the infinite-depth-and-width limit.

We will also note that $O(n^{-1})$ error in the result arise from replacing the $O(n^{-1/2})$ with a Gaussian due to Berry–Esseen, and the $O(n^{-1})$ term with its expectation, as these are the dominant error terms in the approximation.

Proof. We start by defining the notations

$$g_\ell^\alpha \coloneqq W_\ell \frac{\varphi_\ell^\alpha}{|\varphi_\ell^\alpha|}, \quad R_\ell^{\alpha\beta} \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^n c\varphi_s(g_{\ell,i}^\alpha)\varphi_s(g_{\ell,i}^\beta) - cK_1(\rho_\ell^{\alpha\beta}), \tag{B.39}$$

and using positive homogeneity we can write $\varphi_s(\sqrt{\frac{c}{n}} W_\ell \varphi_\ell^\alpha) = \sqrt{\frac{c}{n}} |\varphi_\ell^\alpha| \varphi_s(g_\ell^\alpha)$, which gives us

$$\langle \varphi_{\ell+1}^\alpha, \varphi_{\ell+1}^\beta \rangle = |\varphi_\ell^\alpha|\,|\varphi_\ell^\beta| \frac{c}{n} \sum_{i=1}^n \varphi_s(g_{\ell,i}^\alpha)\varphi_s(g_{\ell,i}^\beta) = |\varphi_\ell^\alpha|\,|\varphi_\ell^\beta| \left(cK_1(\rho_\ell^{\alpha\beta}) + \frac{1}{\sqrt{n}} R_\ell^{\alpha\beta} \right). \tag{B.40}$$

Now consider the same case for $R_\ell^{\alpha\alpha}$ and $R_\ell^{\beta\beta}$ with $K_1(1) = c^{-1}$, we also get

$$\rho_{\ell+1}^{\alpha\beta} = \begin{cases} \frac{\langle \varphi_{\ell+1}^\alpha, \varphi_{\ell+1}^\beta \rangle}{|\varphi_{\ell+1}^\alpha|\,|\varphi_{\ell+1}^\alpha|} = \frac{cK_1(\rho_\ell^{\alpha\beta}) + \frac{1}{\sqrt{n}} R_\ell^{\alpha\beta}}{\sqrt{(1+\frac{1}{\sqrt{n}} R_\ell^{\alpha\alpha})(1+\frac{1}{\sqrt{n}} R_\ell^{\beta\beta})}}, & \text{if } |\varphi_{\ell+1}^\alpha|, |\varphi_{\ell+1}^\beta| > 0, \\ \mathbf{e}, & \text{otherwise.} \end{cases} \tag{B.41}$$

We observe that whenever $|\varphi_\alpha^\ell| > 0$, we have that $1 + \frac{1}{\sqrt{n}} R_\ell^{\alpha\alpha} = \frac{|\varphi_{\ell+1}^\alpha|^2}{|\varphi_\ell^\alpha|^2} \geq 0$. Therefore the event $E \coloneqq \{R_\ell^{\alpha\alpha}, R_\ell^{\beta\beta} \leq -\sqrt{n}\}$ is the same as $\{\rho_{\ell+1}^{\alpha\beta} = \mathbf{e}\}$, which is equivalent to when $z_{\ell+1}^\alpha$ or $z_{\ell+1}^\beta$ has only non-positive entries. When conditioned on the previous layer, all the entries are independent, this event has probability $\Pi(x, \{\mathbf{e}\}) = O(2^{-n})$. We will see later that modifying this Markov chain to remove this event will incur only a "minor cost" of $O(2^{-n})$.

Let us fix any realization of $R^{\alpha\alpha}, R^{\beta\beta}, R^{\alpha\beta}$ outside of the event E (i.e. by viewing it as a map $R^{\alpha\alpha} : \Omega \to \mathbb{R}$ from the probability space for some fixed $\omega \in \Omega$), we can compute the Taylor expansion with respect to $1/\sqrt{n}$ about 0 (Taylor expansion done using SymPy [51] Python package)

$$\begin{aligned} \rho_{\ell+1}^{\alpha\beta} = cK_1(\rho_\ell^{\alpha\beta}) &+ \frac{1}{\sqrt{n}} \left[R_\ell^{\alpha\beta} - \frac{cK_1(\rho_\ell^{\alpha\beta})}{2} (R_\ell^{\alpha\alpha} + R_\ell^{\beta\beta}) \right] \\ &+ \frac{1}{n} \left[\frac{cK_1(\rho_\ell^{\alpha\beta})}{8} (3(R_\ell^{\alpha\alpha} + R_\ell^{\beta\beta})^2 - 4R_\ell^{\alpha\alpha} R_\ell^{\beta\beta}) - \frac{1}{2} R_\ell^{\alpha\beta} (R_\ell^{\alpha\alpha} + R_\ell^{\beta\beta}) \right] + O(n^{-3/2}), \end{aligned} \tag{B.42}$$

where we recall the $X = O(n^{-3/2})$ notation denotes a random variable (the Taylor remainder term) where all moments of $n^{3/2}X$ are bounded by a constant independent of n.

We can simplify these terms further by computing the mean and variance of the expansion (without conditioning on E^c). More specifically, each of the $R_\ell^{\alpha\alpha}, R_\ell^{\beta\beta}, R_\ell^{\alpha\beta}$ have zero mean and covariance

$$\mathbf{Cov}_\ell \left(\begin{bmatrix} R_\ell^{\alpha\alpha} \\ R_\ell^{\beta\beta} \\ R_\ell^{\alpha\beta} \end{bmatrix} \right) = \begin{bmatrix} M_2 & c^2K_2 - 1 & c^2K_{3,1} - cK_1 \\ c^2K_2 - 1 & M_2 & c^2K_{3,1} - cK_1 \\ c^2K_{3,1} - cK_1 & c^2K_{3,1} - cK_1 & c^2(K_2 - K_1^2) \end{bmatrix}, \tag{B.43}$$

where we recall $\mathbf{Cov}_\ell$ is the conditional covariance given the sigma-algebra $\mathcal{F}_\ell$ generated by the ℓ-th layer $[z_\ell^\alpha]_{\alpha=1}^m$. We can now recover the desired result by calculating the drift and variance coefficients using SymPy [51] again

$$\begin{aligned}
\sigma^2_{\text{ReLU}}(\rho_\ell^{\alpha\beta}) &:= \mathbb{E}_\ell\left[R_\ell^{\alpha\beta} - \frac{cK_1}{2}(R_\ell^{\alpha\alpha} + R_\ell^{\beta\beta})\right]^2 \\
&= \frac{c^2}{2}\left[K_1^2(c^2K_2 + M_2 + 1) - 4cK_1K_{3,1} + 2K_2\right] , \\
\mu_{\text{ReLU}}(\rho_\ell^{\alpha\beta}) &:= \mathbb{E}_\ell\left[\frac{cK_1}{8}(3(R_\ell^{\alpha\alpha} + R_\ell^{\beta\beta})^2 - 4R_\ell^{\alpha\alpha}R_\ell^{\beta\beta}) - \frac{1}{2}R_\ell^{\alpha\beta}(R_\ell^{\alpha\alpha} + R_\ell^{\beta\beta})\right] \\
&= \frac{c}{4}\left[K_1(c^2K_2 + 3M_2 + 3) - 4cK_{3,1}\right] ,
\end{aligned} \tag{B.44}$$

where we recall $\mathbb{E}_\ell[\,\cdot\,] = \mathbb{E}[\,\cdot\,|\mathcal{F}_\ell]$ is the conditional expectation given the sigma-algebra $\mathcal{F}_\ell$ generated by the ℓ-th layer $[z_\ell^\alpha]_{\alpha=1}^m$.

This allows us to write (considering the well defined case)

$$\rho_{\ell+1}^{\alpha\beta} = cK_1(\rho_\ell^{\alpha\beta}) + \frac{\sigma_{\text{ReLU}}(\rho_\ell^{\alpha\beta})}{\sqrt{n}}\xi_\ell + \frac{\mu_{\text{ReLU}}(\rho_\ell^{\alpha\beta}) + \eta(\rho_\ell^{\alpha\beta})}{n} + O(n^{-3/2}) , \tag{B.45}$$

where ξ_ℓ is has zero mean and unit variance (when not conditioned on $\rho_{\ell+1}^{\alpha\beta} = \mathbf{e}$), and $\eta(\rho_\ell^{\alpha\beta})$ has zero mean. Observe that there are three differences between $\{\rho_\ell^{\alpha\beta}\}$ and the approximate chain $\{p_\ell\}$:

1. $\rho_{\ell+1}^{\alpha\beta} = \mathbf{e}$ with probability $O(2^{-n})$,
2. ξ_ℓ is replaced by $z_\ell \sim \mathcal{N}(0,1)$,
3. $\eta(\rho_\ell^{\alpha\beta})$ and the higher order $O(n^{-3/2})$ terms in the Taylor expansion are removed.

To complete the proof, we will need to control these differences in terms of the bounded Lipschitz distance on the Markov transition kernels. To this goal, we let h be such that $\|h\|_{BL} \le 1$, hence it must be both bounded by 1 and at worst 1-Lipschitz. We will first condition on E^c to write the Taylor expansion, and then "uncondition" to recover the original distribution, both at a cost of an $O(2^{-n})$ error term. More precisely, we will write

$$\begin{aligned}
&\mathbb{E}_\ell\, h(\rho_{\ell+1}^{\alpha\beta}) \\
&= \mathbb{E}_\ell[h(\rho_{\ell+1}^{\alpha\beta})|E^c]\,\mathbb{P}_\ell(E^c) + \mathbb{E}_\ell[h(\mathbf{e})|E]\,\mathbb{P}_\ell(E) \\
&= \mathbb{E}_\ell[h(\rho_{\ell+1}^{\alpha\beta})|E^c]\,\mathbb{P}_\ell(E^c) + O(1)O(2^{-n}) \\
&= \mathbb{E}_\ell\left[h\left(cK_1(\rho_\ell^{\alpha\beta}) + \frac{\sigma_{\text{ReLU}}(\rho_\ell^{\alpha\beta})}{\sqrt{n}}\xi_\ell + \frac{\mu_{\text{ReLU}}(\rho_\ell^{\alpha\beta}) + \eta(\rho_\ell^{\alpha\beta})}{n} + O(n^{-3/2})\right)\middle| E^c\right]\,\mathbb{P}_\ell(E^c) + O(2^{-n}) ,
\end{aligned} \tag{B.46}$$

where we recall $\mathbb{E}_\ell[\,\cdot\,] = \mathbb{E}[\,\cdot\,|\mathcal{F}_\ell]$, and we define $\mathbb{P}_\ell(E) \coloneqq \mathbb{E}_\ell \mathbb{1}_E$.

At this point we observe that we can now "uncondition" the Taylor expansion by essentially doing the same trick, or more precisely observe that

$$\mathbb{E}_\ell\left[h\left(cK_1(\rho_\ell^{\alpha\beta}) + \frac{\sigma_{\text{ReLU}}(\rho_\ell^{\alpha\beta})}{\sqrt{n}}\xi_\ell + \frac{\mu_{\text{ReLU}}(\rho_\ell^{\alpha\beta}) + \eta(\rho_\ell^{\alpha\beta})}{n} + O(n^{-3/2})\right)\middle| E\right]\,\mathbb{P}_\ell(E) = O(2^{-n}) , \tag{B.47}$$

therefore we can write

$$\begin{aligned}
&\mathbb{E}_\ell\, h(\rho_{\ell+1}^{\alpha\beta}) \\
&= \mathbb{E}_\ell\left[h(\cdots)|\,E^c\right]\,\mathbb{P}_\ell(E^c) + \mathbb{E}_\ell\left[h(\cdots)|\,E\right]\,\mathbb{P}_\ell(E) + O(2^{-n}) \\
&= \mathbb{E}_\ell\, h\left(cK_1(\rho_\ell^{\alpha\beta}) + \frac{\sigma_{\text{ReLU}}(\rho_\ell^{\alpha\beta})}{\sqrt{n}}\xi_\ell + \frac{\mu_{\text{ReLU}}(\rho_\ell^{\alpha\beta}) + \eta(\rho_\ell^{\alpha\beta})}{n} + O(n^{-3/2})\right) + O(2^{-n}) .
\end{aligned} \tag{B.48}$$

Since h is 1-Lipschitz, we have that $h(x+y) \le h(x) + |y|$, and therefore we can write

$$\begin{aligned}\mathbb{E}_\ell\, h(\rho_{\ell+1}^{\alpha\beta}) \le \mathbb{E}_\ell\, h\left(cK_1(\rho_\ell^{\alpha\beta}) + \frac{\sigma_{\text{ReLU}}(\rho_\ell^{\alpha\beta})}{\sqrt{n}} z_\ell + \frac{\mu_{\text{ReLU}}(\rho_\ell^{\alpha\beta})}{n}\right) \\ + \mathbb{E}_\ell\, \frac{\sigma_{\text{ReLU}}(\rho_\ell^{\alpha\beta})}{\sqrt{n}} |\xi_\ell - z_\ell| + \mathbb{E}_\ell\, \frac{|\eta(\rho_\ell^{\alpha\beta})|}{n} + O(n^{-3/2} + 2^{-n})\,.\end{aligned} \tag{B.49}$$

Observe that the first term is exactly the transition kernel of p_ℓ applied to h, i.e. $\mathbb{E}_\ell\, h(p_{\ell+1}) = \int h(y)\, P(p_\ell, dy)$, which means it's sufficient to control the leftover terms at order $O(n^{-1})$ for a chosen coupling of ξ_ℓ and z_ℓ. Since clearly $\mathbb{E}_\ell\, \eta(\rho_\ell^{\alpha\beta}) = O(1)$ as it does not depend on n, we just need to show $\mathbb{E}_\ell\, |\xi_\ell - z_\ell| = O(n^{-1/2})$. Observe that by definition, we have

$$\xi_\ell = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\sigma_{\text{ReLU}}(\rho_\ell^{\alpha\beta})} \left[c\varphi_s(g_{\ell,i}^\alpha)\varphi_s(g_{\ell,i}^\beta) - cK_1(\rho_\ell^{\alpha\beta}) - \frac{cK_1(\rho_\ell^{\alpha\beta})}{2}\left(c\varphi_s(g_{\ell,i}^\alpha)^2 + c\varphi_s(g_{\ell,i}^\beta)^2 - 2\right)\right], \tag{B.50}$$

where the terms of the sum are iid with zero mean and unit variance (since each neuron is independent conditioned on the previous layer). Therefore, we can invoke a standard L^1 Berry–Esseen bound, e.g. Theorem 4.2 of [**chen2011normal**]. In this case, we let F be the CDF of ξ_ℓ and G be the CDF of z_ℓ, and by duality of L^1 (equation 4.6 of [**chen2011normal**]) we have that

$$\inf \mathbb{E}\, |\xi_\ell - z_\ell| = \|F - G\|_{L^1} \le O(n^{-1/2})\,, \tag{B.51}$$

where the $\inf$ is over all couplings of ξ_ℓ, z_ℓ.

Finally since the above results do not depend on the choice of the test function h, so we have that

$$d_{BL}(\Pi(x,\cdot), P(x,\cdot)) = \sup_{\|h\|_{BL}\le 1} \mathbb{E}_\ell \left(h(\rho_{\ell+1}^{\alpha\beta}) - h(p_{\ell+1})\right) \le O(n^{-1})\,, \tag{B.52}$$

which is the desired result.

□

C Proofs for ReLU Shaping Results

In this section, we first recall the ReLU-like activation function for $s = (s_+, s_-) \in \mathbb{R}^2$ defined as

$$\varphi_s(x) \coloneqq s_+ \max(x, 0) + s_- \min(x, 0) = s_+\varphi(x) - s_-\varphi(-x)\,, \tag{C.1}$$

where $\varphi(x) \coloneqq \max(x, 0)$ is the usual ReLU activation.

We will also recall the definitions

$$\bar{J}_{p,r}(\rho) \coloneqq \mathbb{E}\, \varphi(g)^p \varphi(\hat{g})^r\,, \quad K_{p,r}(\rho) \coloneqq \mathbb{E}\, \varphi_s(g)^p \varphi_s(\hat{g})^r\,, \tag{C.2}$$

where g, w are iid $\mathcal{N}(0,1)$ and we define $\hat{g} = \rho g + qw$ with $q = \sqrt{1-\rho^2}$. We will also use the short hand notation to write $\bar{J}_p \coloneqq \bar{J}_{p,p}, K_p \coloneqq K_{p,p}$.

Here we recall from [41]

$$\bar{J}_1(\rho) = \frac{\sqrt{1-\rho^2} + (\pi - \arccos\rho)\rho}{2\pi}\,. \tag{C.3}$$

We will also recall from Lemma B.6 the following moment calculations

$$\begin{aligned} c^{-1} = \mathbb{E}\, \varphi_s(g)^2 = \frac{s_+^2 + s_-^2}{2}\,, \\ K_1(\rho) = \mathbb{E}\, \varphi_s(g)\, \varphi_s(\hat{g}) = (s_+^2 + s_-^2)\bar{J}_1(\rho) - 2s_+s_-\bar{J}_1(-\rho)\,. \end{aligned} \tag{C.4}$$

In the shaped case, we will calculate a Taylor expansion for the function $cK_1(\rho)$.

Lemma C.1 (Shaping Correlation Function Expansion)**.** *Let* $s_\pm = 1 + \frac{c_\pm}{\sqrt{n}}$*, then*

$$cK_1(\rho) = \rho + \frac{\nu(\rho)}{n} + O(n^{-3/2}), \tag{C.5}$$

where $\nu(\rho) = \frac{(c_+ - c_-)^2}{2\pi}\left(\sqrt{1-\rho^2} + \rho \arccos \rho\right)$.

Proof. We start by consider plugging in the formula from (C.4) to get

$$\begin{aligned} cK_1(\rho) &= \frac{2}{s_+^2 + s_-^2}\left((s_+^2 + s_-^2)\bar{J}_1(\rho) - 2s_+ s_- \bar{J}_1(-\rho)\right) \\ &= \frac{2}{s_+^2 + s_-^2}\frac{1}{2\pi}\left((s_+^2 + s_-^2)\left(\sqrt{1-\rho^2} + (\pi - \arccos\rho)\rho\right) - 2s_+s_-\left(\sqrt{1-\rho^2} - (\pi - \arccos(-\rho))\rho\right)\right) \\ &= \frac{2}{s_+^2 + s_-^2}\frac{1}{2\pi}\left((s_+^2 + s_-^2)\left(\sqrt{1-\rho^2} + (\pi - \arccos\rho)\rho\right) - 2s_+s_-\left(\sqrt{1-\rho^2} - (\arccos\rho)\rho\right)\right). \end{aligned} \tag{C.6}$$

where we used the fact that $\arccos(-\rho) = \pi - \arccos(\rho)$.

After substituting $s_\pm = 1 + \frac{c_\pm}{\sqrt{n}}$, we can use SymPy [51] to Taylor expand with respect to the variable $x = n^{-1/2}$ about $x_0 = 0$ and get

$$\begin{aligned} cK_1(\rho) = &\frac{\rho \arccos(\rho)}{\pi} + \frac{\rho(\pi - \arccos(\rho))}{\pi} \\ &+ \left(n^{-1/2}\right)^2 \left(\frac{-\rho c_+^2 \arccos(\rho) + 2\rho c_+ c_- \arccos(\rho) - \rho c_-^2 \arccos(\rho)}{2\pi} \right. \\ &\qquad \left. + \frac{c_+^2\sqrt{1-\rho^2} - 2c_+c_-\sqrt{1-\rho^2} + c_-^2\sqrt{1-\rho^2}}{2\pi}\right) \\ &+ O\left(\left(n^{-1/2}\right)^3\right), \end{aligned} \tag{C.7}$$

where we used the simplify function on the coefficients to reduce the size of the expression.

We can further simplify to get

$$cK_1(\rho) = \rho + \frac{1}{n}\frac{(c_+ - c_-)^2}{2\pi}\left(\sqrt{1-\rho^2} - \rho\arccos\rho\right) + O(n^{-3/2}), \tag{C.8}$$

which is the desired result.

□

We will also need an approximation result for fourth moments.

Lemma C.2 (Fourth Moment Approximation)**.** *Let* $g^\alpha, g^\beta, g^\gamma, g^\delta \in \mathbb{R}$ *be jointly Gaussian such that*

$$\begin{bmatrix} g^\alpha \\ g^\beta \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \rho^{\alpha\beta} \\ \rho^{\alpha\beta} & 1 \end{bmatrix}\right), \tag{C.9}$$

and similarly for other pairs of $\alpha, \beta, \gamma, \delta$*. Then*

$$\mathbb{E}\,\varphi_s(g^\alpha)\varphi_s(g^\beta)\varphi_s(g^\gamma)\varphi_s(g^\delta) = \mathbb{E}\,g^\alpha g^\beta g^\gamma g^\delta + O(n^{-1/2}) = \rho^{\alpha\beta}\rho^{\gamma\delta} + \rho^{\alpha\gamma}\rho^{\beta\delta} + \rho^{\alpha\delta}\rho^{\beta\gamma} + O(n^{-1/2}), \tag{C.10}$$

where the constant in the $O(\cdot)$ *notation is universal.*

Proof. We start by writing

$$\varphi_s(x) = x + \frac{1}{\sqrt{n}}\left(c_+\varphi(x) - c_-\varphi(-x)\right), \tag{C.11}$$

and this allows us to write

$$\mathbb{E}\,\varphi_s(g^\alpha)\varphi_s(g^\beta)\varphi_s(g^\gamma)\varphi_s(g^\delta) = \mathbb{E}\,g^\alpha g^\beta g^\gamma g^\delta + O(n^{-1/2}). \tag{C.12}$$

Then by Isserlis' Theorem, we can write

$$\mathbb{E}\, g^\alpha g^\beta g^\gamma g^\delta = \mathbb{E}\, g^\alpha g^\beta \mathbb{E}\, g^\gamma g^\delta + \mathbb{E}\, g^\alpha g^\gamma \mathbb{E}\, g^\beta g^\delta + \mathbb{E}\, g^\alpha g^\delta \mathbb{E}\, g^\beta g^\gamma\,, \tag{C.13}$$

which gives us the desired result.

□

We will also calculate a useful covariance.

Lemma C.3 (Covariance of $R^{\alpha\beta}$)**.** *Let $g^\alpha, g^\beta, g^\gamma, g^\delta \in \mathbb{R}^n$ be jointly Gaussian vectors such that*

$$\begin{bmatrix} g^\alpha \\ g^\beta \end{bmatrix} \sim \mathcal{N}\left(0\,, \begin{bmatrix} 1 & \rho^{\alpha\beta} \\ \rho^{\alpha\beta} & 1 \end{bmatrix} \otimes I_n\right)\,, \tag{C.14}$$

and similarly for other pairs of $\alpha, \beta, \gamma, \delta$. If we also define

$$R^{\alpha\beta} \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[c\varphi_s(g_i^\alpha)\varphi_s(g_i^\beta) - cK_1(\rho^{\alpha\beta})\right]\,, \tag{C.15}$$

then we have the following covariance formula:

$$\mathbb{E}\, R^{\alpha\beta} R^{\gamma\delta} = \rho^{\alpha\gamma}\rho^{\beta\delta} + \rho^{\alpha\delta}\rho^{\beta\gamma} + O(n^{-1/2})\,. \tag{C.16}$$

Proof. We first observe that since each entry of the sum in $R^{\alpha\beta}$ are iid and zero mean, it is sufficient to just compute the covariance a single term. In other words

$$\mathbb{E}\, R^{\alpha\beta} R^{\gamma\delta} = \mathbb{E}\, c^2 \left(\varphi_s(g_i^\alpha)\varphi_s(g_i^\beta) - K_1(\rho^{\alpha\beta})\right)\left(\varphi_s(g_i^\gamma)\varphi_s(g_i^\delta) - K_1(\rho^{\gamma\delta})\right)\,. \tag{C.17}$$

Since $c = 1 + O(n^{-1/2})$ and $K_1(\rho) = \rho + O(n^{-1})$ from Lemma C.1, we can further write this as

$$\mathbb{E}\, R^{\alpha\beta} R^{\gamma\delta} = \mathbb{E}\left(\varphi_s(g_i^\alpha)\varphi_s(g_i^\beta) - \rho^{\alpha\beta}\right)\left(\varphi_s(g_i^\gamma)\varphi_s(g_i^\delta) - \rho^{\gamma\delta}\right) + O(n^{-1/2})\,, \tag{C.18}$$

and we can use the fourth moment approximation Lemma C.2 to get

$$\begin{aligned} \mathbb{E}\, R^{\alpha\beta} R^{\gamma\delta} &= \rho^{\alpha\beta}\rho^{\gamma\delta} + \rho^{\alpha\gamma}\rho^{\beta\delta} + \rho^{\alpha\delta}\rho^{\beta\gamma} - \rho^{\alpha\beta}\rho^{\gamma\delta} - \rho^{\alpha\beta}\rho^{\gamma\delta} + \rho^{\alpha\beta}\rho^{\gamma\delta} + O(n^{-1/2}) \\ &= \rho^{\alpha\gamma}\rho^{\beta\delta} + \rho^{\alpha\delta}\rho^{\beta\gamma} + O(n^{-1/2})\,, \end{aligned} \tag{C.19}$$

which is the desired result.

□

C.1 Proof of Theorem 3.2 (Covariance SDE, ReLU)

We start by restating the theorem.

Theorem C.4 (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 3.1, 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 $D_{\mathbb{R}_+, \mathbb{R}^{m(m+1)/2}}$*

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

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

$$b(V_t) = \left[\nu\left(\rho_t^{\alpha\beta}\right)\sqrt{V_t^{\alpha\alpha}V_t^{\beta\beta}}\right]_{1\le\alpha\le\beta\le m}\,, \quad \Sigma(V_t) = \left[V_t^{\alpha\gamma}V_t^{\beta\delta} + V_t^{\alpha\delta}V_t^{\beta\gamma}\right]_{\alpha\le\beta,\gamma\le\delta}\,. \tag{C.21}$$

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

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

Proof. We start by recalling the definitions

$$V_{\ell+1}^{\alpha\beta} \coloneqq \frac{c}{n}\langle \varphi_{\ell+1}^{\alpha}, \varphi_{\ell+1}^{\beta}\rangle = \frac{c}{n}\left\langle \varphi_s\left(\sqrt{\frac{c}{n}}W_\ell \varphi_\ell^\alpha\right), \varphi_s\left(\sqrt{\frac{c}{n}}W_\ell \varphi_\ell^\beta\right)\right\rangle . \tag{C.23}$$

At this point, we can define

$$g_\ell^\alpha \coloneqq W_\ell \frac{\varphi_\ell^\alpha}{|\varphi_\ell^\alpha|}, \tag{C.24}$$

and observe that

$$\begin{bmatrix} g_\ell^\alpha \\ g_\ell^\beta \end{bmatrix} \Bigg| \mathcal{F}_\ell \overset{d}{=} N\left(0, \begin{bmatrix} 1 & \rho_\ell^{\alpha\beta} \\ \rho_\ell^{\alpha\beta} & 1 \end{bmatrix} \otimes I_n\right), \tag{C.25}$$

where $\mathcal{F}_\ell$ is the sigma-algebra generated by $[z_\ell^\alpha]_{\alpha=1}^m$, $\rho_\ell^{\alpha\beta} \coloneqq \frac{\langle \varphi_\ell^\alpha, \varphi_\ell^\beta\rangle}{|\varphi_\ell^\alpha|\,|\varphi_\ell^\beta|}$, and $\otimes$ denotes the Kronecker product. Then we can use positive homogeneity (i.e. $\varphi_s(cx) = |c|\varphi_s(x)$) to write

$$\begin{aligned} V_{\ell+1}^{\alpha\beta} &= \frac{c}{n}|\varphi_\ell^\alpha|\,|\varphi_\ell^\beta|\,\frac{c}{n}\langle \varphi_s(g_\ell^\alpha), \varphi_s(g_\ell^\beta)\rangle \\ &= \sqrt{V_\ell^{\alpha\alpha}V_\ell^{\beta\beta}}\left(cK_1(\rho_\ell^{\alpha\beta}) + \frac{1}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[c\varphi_s(g_{\ell,i}^\alpha)\varphi_s(g_{\ell,i}^\beta) - cK_1(\rho_\ell^{\alpha\beta})\right]\right) \\ &=: \sqrt{V_\ell^{\alpha\alpha}V_\ell^{\beta\beta}}\left(cK_1(\rho_\ell^{\alpha\beta}) + \frac{1}{\sqrt{n}}R_\ell^{\alpha\beta}\right), \end{aligned} \tag{C.26}$$

where we defined $R_\ell^{\alpha\beta} \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^n\left[c\varphi_s(g_{\ell,i}^\alpha)\varphi_s(g_{\ell,i}^\beta) - cK_1(\rho_\ell^{\alpha\beta})\right]$.

Next we use the expansion of $cK_1(\rho_\ell^{\alpha\beta})$ from Lemma C.1 to write

$$\begin{aligned} V_{\ell+1}^{\alpha\beta} &= \sqrt{V_\ell^{\alpha\alpha}V_\ell^{\beta\beta}}\left(\rho_\ell^{\alpha\beta} + \frac{\nu(\rho_\ell^{\alpha\beta})}{n} + \frac{1}{\sqrt{n}}R_\ell^{\alpha\beta}\right) + O(n^{-3/2}) \\ &= V_\ell^{\alpha\beta} + \frac{1}{n}\nu(\rho_\ell^{\alpha\beta})\sqrt{V_\ell^{\alpha\alpha}V_\ell^{\beta\beta}} + \frac{1}{\sqrt{n}}\sqrt{V_\ell^{\alpha\alpha}V_\ell^{\beta\beta}}R_\ell^{\alpha\beta} + O(n^{-3/2}), \end{aligned} \tag{C.27}$$

which essentially recovers the Markov chain form we want from Proposition A.6, where the drift is

$$b(V) = \nu\left(\rho_\ell^{\alpha\beta}\right)\sqrt{V_\ell^{\alpha\alpha}V_\ell^{\beta\beta}}, \tag{C.28}$$

as desired.

It remains to simply compute the covariance conditioned on previous layer. To this end, we will use Lemma C.3 to write

$$\begin{aligned} \Sigma(V_\ell)_{\alpha\beta,\gamma\delta} &= \mathbb{E}_\ell\left[\sqrt{V_\ell^{\alpha\alpha}V_\ell^{\beta\beta}}R_\ell^{\alpha\beta}\sqrt{V_\ell^{\gamma\gamma}V_\ell^{\delta\delta}}R_\ell^{\gamma\delta}\right] \\ &= \sqrt{V_\ell^{\alpha\alpha}V_\ell^{\beta\beta}V_\ell^{\gamma\gamma}V_\ell^{\delta\delta}}\left(\rho_\ell^{\alpha\gamma}\rho_\ell^{\beta\delta} + \rho_\ell^{\alpha\delta}\rho_\ell^{\beta\gamma} + O(n^{-1/2})\right) \\ &= V_\ell^{\alpha\gamma}V_\ell^{\beta\delta} + V_\ell^{\alpha\delta}V_\ell^{\beta\gamma} + O(n^{-1/2}), \end{aligned} \tag{C.29}$$

where we recall $\mathbb{E}_\ell[\,\cdot\,] = \mathbb{E}[\,\cdot\,|\mathcal{F}_\ell]$ is the conditional expectation given the sigma-algebra generated by $\{z_\ell^\alpha\}_{\alpha=1}^m$. By setting $\sigma = \Sigma^{1/2}$, we then recover the desired SDE via Proposition A.6 on the Markov chain of $V_\ell^{\alpha\beta}$.

□

C.2 Proof of Theorem 3.3 (Correlation SDE, ReLU)

We start by restating the theorem.

Theorem C.5 (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* $\rho^{\alpha\beta}_{\lfloor tn \rfloor}$ *converges in distribution to the solution of the following SDE in the Skorohod topology of* $D_{\mathbb{R}_+,\mathbb{R}}$

$$d\rho_t^{\alpha\beta} = \left[\nu(\rho_t^{\alpha\beta}) + \mu(\rho_t^{\alpha\beta})\right] dt + \sigma(\rho_t^{\alpha\beta})\, dB_t\,, \quad \rho_0^{\alpha\beta} = \frac{\langle x^\alpha, x^\beta \rangle}{|x^\alpha|\,|x^\beta|}\,, \tag{C.30}$$

where

$$\nu(\rho) = \frac{(c_+ - c_-)^2}{2\pi}\left[\sqrt{1-\rho^2} - \arccos(\rho)\rho\right], \quad \mu(\rho) = -\frac{1}{2}\rho(1-\rho^2)\,, \quad \sigma(\rho) = 1 - \rho^2\,. \tag{C.31}$$

Proof. While it is possible to obtain this result as a consequence of Theorem 3.2 via Itô's Lemma, we will show an alternative derivation by extending the steps of Proposition B.8, where we can directly compute the Taylor expansion in the event $E \coloneqq \{|\varphi_{\ell+1}^\alpha|, |\varphi_{\ell+1}^\beta| > 0\}$

$$\rho_{\ell+1}^{\alpha\beta} = \frac{\langle \varphi_{\ell+1}^\alpha, \varphi_{\ell+1}^\beta \rangle}{|\varphi_{\ell+1}^\alpha|\,|\varphi_{\ell+1}^\beta|} = cK_1(\rho_\ell^{\alpha\beta}) + \frac{\widetilde{\mu}(\rho_\ell^{\alpha\beta})}{n} + \sigma(\rho_\ell^{\alpha\beta})\frac{\xi_\ell}{\sqrt{n}} + O(n^{-3/2})\,, \tag{C.32}$$

where (unconditioned on E) ξ_ℓ are iid with mean zero variance one and

$$\begin{aligned} \mu(\rho_\ell^{\alpha\beta}) &\coloneqq \mathbb{E}_\ell\, \widetilde{\mu}(\rho_\ell^{\alpha\beta}) = \frac{c}{4}\left[K_1(c^2K_2 + 3M_2 + 3) - 4cK_{3,1}\right], \\ \sigma^2(\rho_\ell^{\alpha\beta}) &\coloneqq \frac{c^2}{2}\left[K_1^2(c^2K_2 + M_2 + 1) - 4cK_1K_{3,1} + 2K_2\right], \end{aligned} \tag{C.33}$$

where we replaced $\mu_{\text{ReLU}}, \sigma_{\text{ReLU}}$ with μ, σ as we will be shaping the activation function, and we recall $\mathbb{E}_\ell[\,\cdot\,] = \mathbb{E}[\,\cdot\,|\mathcal{F}_\ell]$ is the conditional expectation given the sigma-algebra generated by $\{z_\ell^\alpha\}_{\alpha=1}^m$.

We note that the undefined event E occurs only when $z_{\ell+1}^\alpha$ or $z_{\ell+1}^\beta$ has all negative entries, which occurs with probability $O(2^{-n})$. Since all the terms of interest have finite moments, we can proceed by removing this event E in a similar fashion as Proposition B.8.

Using the expansion of $cK_1(\rho)$ from Lemma C.1, we can now write

$$\rho_{\ell+1}^{\alpha\beta} = \rho_\ell^{\alpha\beta} + \frac{\nu(\rho_\ell^{\alpha\beta}) + \widetilde{\mu}(\rho_\ell^{\alpha\beta})}{n} + \sigma(\rho_\ell^{\alpha\beta})\frac{\xi_\ell}{\sqrt{n}} + O(n^{-3/2})\,. \tag{C.34}$$

Furthermore, we also have that by Lemma C.1 and Lemma C.2

$$\begin{aligned} K_1 &= \rho_\ell^{\alpha\beta} + O(n^{-1})\,, \quad K_2 = 2(\rho_\ell^{\alpha\beta})^2 + 1 + O(n^{-1/2})\,, \\ K_{3,1} &= 3\rho_\ell^{\alpha\beta} + O(n^{-1/2})\,, \quad M_2 = 2 + O(n^{-1/2})\,, \end{aligned} \tag{C.35}$$

which gives us the desired formula of

$$\mu(\rho) = -\frac{1}{2}\rho(1-\rho^2)\,, \quad \sigma(\rho) = 1 - \rho^2\,. \tag{C.36}$$

Finally, we can recover the desired SDE via Proposition A.6.

□

C.3 Joint Correlation SDE

In this section, we will extend Theorem 3.3 to a general joint process over all the possible pairs of correlations.

Theorem C.6 (Joint Correlation SDE)**.** *Let* $\rho_\ell^{\alpha\beta} \coloneqq \frac{\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle}{|\varphi_\ell^\alpha|\,|\varphi_\ell^\beta|}$*, and define* $\rho_\ell \coloneqq [\rho_\ell^{\alpha\beta}]_{1\le\alpha\le\beta=m}$ *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 3.1, in the limit as $n \to \infty, \frac{d}{n} \to T$, the interpolated process $\rho_{\lfloor tn \rfloor}$ converges in distribution to the solution of the following SDE in the Skorohod topology of $D_{\mathbb{R}_+,\mathbb{R}^{m(m+1)/2}}$

$$d\rho_t = b(\rho_t)\,dt + \Sigma(\rho_t)^{1/2}\,dB_t\,, \quad \rho_0 = \left[\frac{\langle x^\alpha, x^\beta\rangle}{|x^\alpha|\,|x^\beta|}\right]_{1\le\alpha\le\beta\le m}\,, \tag{C.37}$$

where the coefficients are defined by

$$\begin{aligned} b(\rho_t) &= \left[\nu(\rho_t^{\alpha\beta}) + \mu(\rho_t^{\alpha\beta})\right]_{1\le\alpha\le\beta\le m}\,, \\ \Sigma(\rho_t) &= \Big[\rho^{\alpha\gamma}\rho^{\beta\delta} + \rho^{\alpha\delta}\rho^{\beta\gamma} + \frac{1}{2}\rho^{\alpha\beta}\rho^{\gamma\delta}\left((\rho^{\alpha\gamma})^2 + (\rho^{\beta\gamma})^2 + (\rho^{\alpha\delta})^2 + (\rho^{\beta\delta})^2\right) \\ &\qquad - \rho^{\alpha\beta}\left(\rho^{\alpha\gamma}\rho^{\alpha\delta} + \rho^{\beta\gamma}\rho^{\beta\delta}\right) - \rho^{\gamma\delta}\left(\rho^{\alpha\gamma}\rho^{\beta\gamma} + \rho^{\alpha\delta}\rho^{\beta\delta}\right)\Big]_{\alpha\le\beta,\gamma\le\delta}\,, \end{aligned} \tag{C.38}$$

with ν, μ defined as in Theorem 3.3.

Proof. It's sufficient to just compute the covariance matrix Σ for the random terms of the Markov chain (B.42), which reduces down to

$$\Sigma(\rho_\ell)_{\alpha\beta,\gamma\delta} = \mathbb{E}_\ell\left(R_\ell^{\alpha\beta} - \frac{c}{2}K_1^{\alpha\beta}(R_\ell^{\alpha\alpha} + R_\ell^{\beta\beta})\right)\left(R_\ell^{\gamma\delta} - \frac{c}{2}K_1^{\gamma\delta}(R_\ell^{\gamma\gamma} + R_\ell^{\delta\delta})\right)\,, \tag{C.39}$$

where we recall $\mathbb{E}_\ell[\,\cdot\,] = \mathbb{E}[\,\cdot\,|\mathcal{F}_\ell]$ is the conditional expectation given the sigma-algebra generated by $\{z_\ell^\alpha\}_{\alpha=1}^m$, and we write $K_1^{\alpha\beta} := K_1(\rho_\ell^{\alpha\beta})$.

Using Lemma C.1 and Lemma C.3, we can calculate this explicitly as

$$\begin{aligned} \Sigma(\rho_\ell)_{\alpha\beta,\gamma\delta} &= \mathbb{E}_\ell\, R_{\alpha\beta}R_{\gamma\delta} + \frac{c^2}{4}K_1^{\alpha\beta}K_1^{\gamma\delta}\mathbb{E}_\ell\,(R_{\alpha\alpha} + R_{\beta\beta})(R_{\gamma\gamma} + R_{\delta\delta}) \\ &\quad - \frac{c}{2}K^{\alpha\beta}\mathbb{E}_\ell\, R_{\gamma\delta}(R_{\alpha\alpha} + R_{\beta\beta}) - \frac{c}{2}K^{\gamma\delta}\mathbb{E}_\ell\, R_{\alpha\beta}(R_{\gamma\gamma} + R_{\delta\delta}) \\ &= \rho^{\alpha\gamma}\rho^{\beta\delta} + \rho^{\alpha\delta}\rho^{\beta\gamma} + \frac{1}{2}\rho^{\alpha\beta}\rho^{\gamma\delta}\left((\rho^{\alpha\gamma})^2 + (\rho^{\beta\gamma})^2 + (\rho^{\alpha\delta})^2 + (\rho^{\beta\delta})^2\right) \\ &\quad - \rho^{\alpha\beta}\left(\rho^{\alpha\gamma}\rho^{\alpha\delta} + \rho^{\beta\gamma}\rho^{\beta\delta}\right) - \rho^{\gamma\delta}\left(\rho^{\alpha\gamma}\rho^{\beta\gamma} + \rho^{\alpha\delta}\rho^{\beta\delta}\right) + O(n^{-1/2})\,, \end{aligned} \tag{C.40}$$

which is the desired result.

□

C.4 Proof for Proposition 3.4 (Critical Exponent, ReLU)

We start by restating the proposition.

Proposition C.7 (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 $p \ge 0$. 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 $0 \le p < \frac{1}{2}$, and $c_+ \ne c_-$,

(ii) the critical limit: the SDE from Theorem 3.3, if $p = \frac{1}{2}$,

(iii) the linear network limit: if $p > \frac{1}{2}$, the following SDE, with μ, σ as defined in (3.5)*,*

$$d\rho_t^{\alpha\beta} = \mu(\rho_t^{\alpha\beta})\,dt + \sigma(\rho_t^{\alpha\beta})\,dB_t\,, \quad \rho_0^{\alpha\beta} = \frac{\langle x^\alpha, x^\beta\rangle}{|x^\alpha|\,|x^\beta|}\,, \tag{C.41}$$

Proof. Case (ii) follows from Theorem 3.3, therefore it is sufficient to only consider cases (i) and (iii). In the case that $p = 0$, we can recover the following recursion in the limit as $n \to \infty$

$$\rho_{\ell+1}^{\alpha\beta} = cK_1(\rho_\ell^{\alpha\beta})\,, \tag{C.42}$$

which matches the infinite-width limit, and it is known that $\rho_\ell^{\alpha\beta} \to 1$ as $\ell \to 1$ (see also Appendix E for an upper bound).

Next we will recall the result of Lemma C.1 and observe that we can simply replace $\sqrt{n}$ with n^p to recover the expansion

$$cK_1(\rho) = \rho + \frac{\nu(\rho)}{n^{2p}} + O(n^{-3p})\,. \tag{C.43}$$

This gives us the following Markov chain from the proof of Theorem 3.3

$$\rho_{\ell+1}^{\alpha\beta} = \rho_\ell^{\alpha\beta} + \frac{\nu(\rho_\ell^{\alpha\beta})}{n^{2p}} + \frac{\mu(\rho_\ell^{\alpha\beta})}{n} + \sigma(\rho_\ell^{\alpha\beta})\frac{\xi_\ell}{\sqrt{n}} + O(n^{-3p} + n^{-3/2})\,. \tag{C.44}$$

In the case that $0 < p < 1/2$, we can consider the time step size $h_n = n^{-2p}$ instead of n^{-1} and apply Proposition A.6, where we recover the ODE

$$\partial_s \hat{\rho}_s^{\alpha\beta} = \nu(\hat{\rho}_s^{\alpha\beta})\,, \tag{C.45}$$

but on the time scale of $\hat{\rho}_s^{\alpha\beta,n} = \hat{\rho}^{\alpha\beta}_{\lfloor sn^{2p} \rfloor}$. Converting it back to the time scale of $\rho_t^{\alpha\beta,n} = \rho^{\alpha\beta}_{\lfloor tn \rfloor}$ implies that we have

$$\rho_t^{\alpha\beta} = \hat{\rho}_\infty^{\alpha\beta}\,, \quad \text{for all } t > 0\,. \tag{C.46}$$

And since $\nu(\rho) > 0$ for all $\rho < 1$ and that $\nu(\rho) = C(1-\rho)^{3/2} + O((1-\rho)^{5/2})$ as $\rho \to 1$, we have that $\hat{\rho}_\infty^{\alpha\beta} = 1$ as desired.

In the case $p > \frac{1}{2}$, we have that since ν is deterministic, we observe the drift term used in Proposition A.6 in the limit as $n \to \infty$ is

$$b_n(\rho) = \nu(\rho)n^{1-2p} + \mu(\rho) \to b(\rho) = \mu(\rho)\,, \tag{C.47}$$

which would simply recover the desired SDE with drift μ only.

□

D Proofs for Smooth Shaping Results

In this section, we consider smooth activation functions φ satisfying Assumption 3.5, that is $\varphi \in C^4, \varphi(0) = 0, \varphi'(0) = 1$, and that $|\varphi^{(4)}(x)| \le C(1 + |x|^p)$ for some $C, p > 0$. We recall the shaping we consider for activations of this type is via the following definition for $s > 0$

$$\varphi_s(x) \coloneqq s\varphi\left(\frac{x}{s}\right)\,, \tag{D.1}$$

so that $\lim_{s\to\infty} \varphi_s(x) = x$.

Before we start, we will calculate the behaviour of the normalizing constant c up an error order of s^3.

Lemma D.1. *Let φ_s be defined as above with φ satisfying Assumption 3.5. Then if $g \sim N(0,1)$, we have that*

$$c = 1 + \frac{1}{s^2}\left(\frac{3}{4}\varphi''(0)^2 + \varphi'''(0)\right) + O(s^{-3})\,. \tag{D.2}$$

Proof. We will first Taylor expand $\varphi_s(g)$ about $g = 0$

$$\varphi_s(g) = 0 + g + \frac{\varphi''(0)}{2s}g^2 + \frac{\varphi'''(0)}{6s^2}g^3 + O(s^{-3})\,, \tag{D.3}$$

where we note by Assumption 3.5 the remainder term is at most polynomial in g.

Therefore the second moment satisfies

$$\begin{aligned}\mathbb{E}\,\varphi_s(g)^2 &= \mathbb{E}\,g^2 + \frac{\varphi''(0)}{s}g^3 + \frac{1}{s^2}\left(\frac{1}{4}\varphi''(0)^2 + \frac{2}{6}\varphi'''(0)\right)g^4 + O(s^{-3}) \\ &= 1 + \frac{1}{s^2}\left(\frac{3}{4}\varphi''(0)^2 + \varphi'''(0)\right) + O(s^{-3})\,,\end{aligned} \tag{D.4}$$

where $O(s^{-3})$ is bounded due to Gaussians have all bounded moments.

Therefore, for $s > 0$ sufficiently small, we have the following expansion

$$c = \frac{1}{\mathbb{E}\,\varphi_s(g)^2} = \frac{1}{1-(-bs^{-2}+O(s^{-3}))} = 1 - \frac{b}{s^2} + O(s^{-3})\,, \tag{D.5}$$

where $b = \frac{3}{4}\varphi''(0)^2 + \varphi'''(0)$, which is the desired result.

□

D.1 Proof of Theorem 3.9 (Covariance SDE, Smooth)

We start by restating the theorem.

Theorem D.2 (Covariance SDE, Smooth)**.** *Let φ satisfy Assumption 3.5, $V_\ell^{\alpha\beta} \coloneqq \frac{c}{n}\langle \varphi_\ell^\alpha, \varphi_\ell^\beta\rangle$ where $\varphi_\ell^\alpha = \varphi_s(z_\ell^\alpha)$, and define $V_\ell \coloneqq [V_\ell^{\alpha\beta}]_{1\le\alpha\le\beta=m}$ to be the upper triangular entries thought of as a vector in $\mathbb{R}^{m(m+1)/2}$. Then, with $s = a\sqrt{n}$ as in Definition 3.6, 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 $D_{\mathbb{R}_+,\mathbb{R}^{m(m+1)/2}}$*

$$dV_t = b(V_t)\,dt + \Sigma(V_t)^{1/2}\,dB_t\,, \quad V_0 = \left[\frac{1}{n_{in}}\langle x^\alpha, x^\beta\rangle\right]_{1\le\alpha\le\beta\le m}\,, \tag{D.6}$$

where $\Sigma(V_t)$ is the same as Theorem 3.2 and

$$b^{\alpha\beta}(V_t) = \frac{\varphi''(0)^2}{4a^2}\left(V_t^{\alpha\alpha}V_t^{\beta\beta} + V_t^{\alpha\beta}(2V_t^{\alpha\beta}-3)\right) + \frac{\varphi'''(0)}{2a^2}V_t^{\alpha\beta}(V_t^{\alpha\alpha}+V_t^{\beta\beta}-2)\,. \tag{D.7}$$

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

$$[z_{out}^\alpha]_{\alpha=1}^m\,|V_T \overset{d}{=} \mathcal{N}\left(0, [V_T^{\alpha\beta}]_{\alpha,\beta=1}^m\right)\,, \tag{D.8}$$

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

Proof. We start by defining $g_\ell^\alpha \coloneqq W_\ell \frac{\varphi_\ell^\alpha}{|\varphi_\ell^\alpha|}$, and observe that

$$\begin{bmatrix} g_\ell^\alpha \\ g_\ell^\beta \end{bmatrix}\Bigg|\,\mathcal{F}_\ell \overset{d}{=} \mathcal{N}\left(0\,, \begin{bmatrix} 1 & \rho_\ell^{\alpha\beta} \\ \rho_\ell^{\alpha\beta} & 1 \end{bmatrix} \otimes I_n\right)\,, \tag{D.9}$$

where $\mathcal{F}_\ell$ is the sigma-algebra generated by the ℓ-th layer $[z_\ell^\alpha]_{\alpha=1}^m$, $\rho_\ell^{\alpha\beta} \coloneqq \frac{\langle\varphi_\ell^\alpha,\varphi_\ell^\beta\rangle}{|\varphi_\ell^\alpha|\,|\varphi_\ell^\beta|}$, and $\otimes$ denotes the Kronecker product. We can then write the Taylor expansion for φ_s about 0 as

$$\begin{aligned}
\varphi_{\ell+1,i}^\alpha &= \varphi_s\left(\sqrt{\frac{c}{n}}|\varphi_\ell^\alpha|g_{\ell,i}^\alpha\right) \\
&= \varphi_s(0) + \varphi_s'(0)\sqrt{\frac{c}{n}}|\varphi_\ell^\alpha|g_{\ell,i}^\alpha + \frac{\varphi_s''(0)}{2}\left(\sqrt{\frac{c}{n}}|\varphi_\ell^\alpha|g_{\ell,i}^\alpha\right)^2 + \frac{\varphi_s'''(0)}{6}\left(\sqrt{\frac{c}{n}}|\varphi_\ell^\alpha|g_{\ell,i}^\alpha\right)^3 \\
&\quad + R_3\left(\sqrt{\frac{c}{n}}|\varphi_\ell^\alpha|g_{\ell,i}^\alpha\right)\,,
\end{aligned} \tag{D.10}$$

where $R_3(\cdot)$ is the Taylor remainder term, which has polynomial growth by Assumption 3.5.

By using the fact that $\varphi(0) = 0, \varphi'(0) = 1$ and observing that the derivatives of φ_s satisfies $\varphi_s^{(k)}(0) = \frac{\varphi^{(k)}(0)}{s^{k-1}}$, we can further write

$$\varphi_{\ell+1,i}^\alpha = \sqrt{\frac{c}{n}}|\varphi_\ell^\alpha|g_{\ell,i}^\alpha + \frac{\varphi''(0)}{2s}\left(\sqrt{\frac{c}{n}}|\varphi_\ell^\alpha|g_{\ell,i}^\alpha\right)^2 + \frac{\varphi'''(0)}{6s^2}\left(\sqrt{\frac{c}{n}}|\varphi_\ell^\alpha|g_{\ell,i}^\alpha\right)^3 + O(s^{-3})\,, \tag{D.11}$$

where the remainder term is at most polynomial in $g_{\ell,i}^\alpha$.

Then we can compute the inner product with the same expansion as

$$\begin{aligned}
&\frac{c}{n}\langle\varphi_{\ell+1}^{\alpha},\varphi_{\ell+1}^{\beta}\rangle \\
&= \frac{c}{n}\sum_{i=1}^{n}\left(\sqrt{\frac{c}{n}}|\varphi_{\ell}^{\alpha}|g_{\ell,i}^{\alpha}+\frac{\varphi''(0)}{2s}\frac{c}{n}|\varphi_{\ell}^{\alpha}|^2(g_{\ell,i}^{\alpha})^2+\frac{\varphi'''(0)}{6s^2}\left(\frac{c}{n}|\varphi_{\ell}^{\alpha}|^2\right)^{3/2}(g_{\ell,i}^{\alpha})^3+O(s^{-3})\right) \\
&\qquad\left(\sqrt{\frac{c}{n}}|\varphi_{\ell}^{\beta}|g_{\ell,i}^{\beta}+\frac{\varphi''(0)}{2s}\frac{c}{n}|\varphi_{\ell}^{\beta}|^2(g_{\ell,i}^{\beta})^2+\frac{\varphi'''(0)}{6s^2}\left(\frac{c}{n}|\varphi_{\ell}^{\beta}|^2\right)^{3/2}(g_{\ell,i}^{\beta})^3+O(s^{-3})\right),
\end{aligned} \tag{D.12}$$

and we will proceed by analyzing the product terms separately. We start with the terms of order $O(s^0)$ first, which are

$$\begin{aligned}
\frac{c}{n}\sum_{i=1}^{n}\frac{c}{n}|\varphi_{\ell}^{\alpha}||\varphi_{\ell}^{\beta}|g_{\ell,i}^{\alpha}g_{\ell,i}^{\beta} &= \frac{c}{n}|\varphi_{\ell}^{\alpha}||\varphi_{\ell}^{\beta}|c\left(\rho_{\ell}^{\alpha\beta}+\frac{1}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_{\ell,i}^{\alpha}g_{\ell,i}^{\beta}-\rho_{\ell}^{\alpha\beta}\right) \\
&= \sqrt{V_{\ell}^{\alpha\alpha}V_{\ell}^{\beta\beta}}c\left(\rho_{\ell}^{\alpha\beta}+\frac{1}{\sqrt{n}}R_{\ell}^{\alpha\beta}\right) \\
&= cV_{\ell}^{\alpha\beta}+c\sqrt{V_{\ell}^{\alpha\alpha}V_{\ell}^{\beta\beta}}\frac{R_{\ell}^{\alpha\beta}}{\sqrt{n}},
\end{aligned} \tag{D.13}$$

where we used the definitions $V_{\ell}^{\alpha\beta}\coloneqq\frac{c}{n}\langle\varphi_{\ell}^{\alpha},\varphi_{\ell}^{\beta}\rangle$ and $R_{\ell}^{\alpha\beta}\coloneqq\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_{\ell,i}^{\alpha}g_{\ell,i}^{\beta}-\rho_{\ell}^{\alpha\beta}$.

For the first order terms, i.e., terms of order $O(s^{-1})$, we have the terms

$$\begin{aligned}
&\frac{c}{n}\sum_{i=1}^{n}\sqrt{\frac{c}{n}}|\varphi_{\ell}^{\alpha}|g_{\ell,i}^{\alpha}\frac{\varphi''(0)}{2s}\frac{c}{n}|\varphi_{\ell}^{\beta}|^2(g_{\ell,i}^{\beta})^2+\sqrt{\frac{c}{n}}|\varphi_{\ell}^{\beta}|g_{\ell,i}^{\beta}\frac{\varphi''(0)}{2s}\frac{c}{n}|\varphi_{\ell}^{\alpha}|^2(g_{\ell,i}^{\alpha})^2 \\
&= \frac{\varphi''(0)}{2s}\sqrt{V_{\ell}^{\alpha\alpha}V_{\ell}^{\beta\beta}}\frac{c}{n}\sum_{i=1}^{n}g_{\ell,i}^{\alpha}g_{\ell,i}^{\beta}\left(\sqrt{V_{\ell}^{\alpha\alpha}}g_{\ell,i}^{\alpha}+\sqrt{V_{\ell}^{\beta\beta}}g_{\ell,i}^{\beta}\right) \\
&= \frac{\varphi''(0)}{2s}\sqrt{V_{\ell}^{\alpha\alpha}V_{\ell}^{\beta\beta}}\frac{c}{\sqrt{n}}\widehat{R}_{\ell}^{\alpha\beta},
\end{aligned} \tag{D.14}$$

where we define $\widehat{R}_{\ell}^{\alpha\beta}\coloneqq\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_{\ell,i}^{\alpha}g_{\ell,i}^{\beta}\left(\sqrt{V_{\ell}^{\alpha\alpha}}g_{\ell,i}^{\alpha}+\sqrt{V_{\ell}^{\beta\beta}}g_{\ell,i}^{\beta}\right)$ and observe this random variable has zero mean and a finite variance. Therefore in view of Proposition A.6, this term cannot contribute to the drift due to having zero mean, nor can this term contribute to the diffusion term due to $s=a\sqrt{n}$ leading to the term being order $\frac{1}{n}$. In other words, the effect of this term will vanish in the limit as $n\to\infty$.

We then turn our attention to the second order terms, i.e., terms of order $O(s^{-2})$

$$\begin{aligned}
&\frac{c}{n}\sum_{i=1}^{n}\frac{\varphi''(0)^2}{4s^2}\frac{c}{n}|\varphi_{\ell}^{\alpha}|^2\frac{c}{n}|\varphi_{\ell}^{\beta}|^2(g_{\ell,i}^{\alpha})^2(g_{\ell,i}^{\beta})^2 \\
&+\frac{\varphi'''(0)}{6s^2}\left(\sqrt{\frac{c}{n}}|\varphi_{\ell}^{\alpha}|\left(\sqrt{\frac{c}{n}}|\varphi_{\ell}^{\beta}|\right)^3g_{\ell,i}^{\alpha}(g_{\ell,i}^{\beta})^3+\left(\sqrt{\frac{c}{n}}|\varphi_{\ell}^{\alpha}|\right)^3\sqrt{\frac{c}{n}}|\varphi_{\ell}^{\beta}|(g_{\ell,i}^{\alpha})^3g_{\ell,i}^{\beta}\right).
\end{aligned} \tag{D.15}$$

Since this term is order $s^{-2}=\frac{1}{a^2n}$, it can only contribute to the drift term, and in view of Proposition A.6, we only need to compute its mean. To this goal, we will simply invoke Isserlis' Theorem and calculate

$$\mathbb{E}_{\ell}\,(g_{\ell,i}^{\alpha})^2(g_{\ell,i}^{\beta})^2=1+2(\rho_{\ell}^{\alpha\beta})^2,\quad \mathbb{E}_{\ell}\,g_{\ell,i}^{\alpha}(g_{\ell,i}^{\beta})^3=\mathbb{E}_{\ell}\,(g_{\ell,i}^{\alpha})^3g_{\ell,i}^{\beta}=3\rho_{\ell}^{\alpha\beta}, \tag{D.16}$$

where we recall $\mathbb{E}_{\ell}[\,\cdot\,]=\mathbb{E}[\,\cdot\,|\mathcal{F}_{\ell}]$ is the conditional expectation given the sigma-algebra generated by $\{z_{\ell}^{\alpha}\}_{\alpha=1}^{m}$. This allows us to compute the conditional expectation $\mathbb{E}_{\ell}$ for the terms of order s^{-2} as

$$\begin{aligned}
&c\left[\frac{\varphi''(0)^2}{4s^2}\frac{c}{n}|\varphi_{\alpha}^{\ell}|^2\frac{c}{n}|\varphi_{\beta}^{\ell}|^2(1+2(\rho_{\ell}^{\alpha\beta})^2)+\frac{\varphi'''(0)}{6s^2}3\rho^{\alpha\beta}\sqrt{\frac{c}{n}}|\varphi_{\alpha}^{\ell}|\sqrt{\frac{c}{n}}|\varphi_{\beta}^{\ell}|\left(\frac{c}{n}|\varphi_{\alpha}^{\ell}|^2+\frac{c}{n}|\varphi_{\beta}^{\ell}|^2\right)\right] \\
&=c\left[\frac{\varphi''(0)^2}{4s^2}(V_{\ell}^{\alpha\alpha}V_{\ell}^{\beta\beta}2(V_{\ell}^{\alpha\beta})^2)+\frac{\varphi'''(0)}{2s^2}V_{\ell}^{\alpha\beta}(V_{\ell}^{\alpha\alpha}+V_{\ell}^{\beta\beta})\right],
\end{aligned} \tag{D.17}$$

Putting these terms together with the fact that $c = 1 - \frac{b}{s^2} + O(s^{-3})$ with $b = \frac{3}{4}\varphi''(0)^2 + \varphi'''(0)$, we can write the update rule for $V_\ell^{\alpha\beta}$ as

$$V_{\ell+1}^{\alpha\beta} = V_\ell^{\alpha\beta} + \frac{1}{n}\left[\frac{\varphi''(0)^2}{4a^2}\left(V_\ell^{\alpha\alpha}V_\ell^{\beta\beta} + V_\ell^{\alpha\beta}(2V_\ell^{\alpha\beta} - 3)\right) + \frac{\varphi'''(0)}{2a^2}V_\ell^{\alpha\beta}(V_\ell^{\alpha\alpha} + V_\ell^{\beta\beta} - 2)\right] + c\sqrt{V_\ell^{\alpha\alpha}V_\ell^{\beta\beta}}\frac{R_\ell^{\alpha\beta}}{\sqrt{n}} + O(n^{-3/2})\,. \tag{D.18}$$

At this point, we have fully recovered the drift term, and we observe the covariance structure is the same as Lemma C.3 in the limit as $n \to \infty$. Therefore we can invoke Proposition A.6 to recover the desired SDE.

□

D.2 Proof of Proposition 3.10 (Critical Exponent, Smooth)

We will restate and prove the proposition.

Proposition D.3 (Critical Exponent, Smooth)**.** *Let φ satisfy Assumption 3.5, $V_\ell^{\alpha\beta} \coloneqq \frac{c}{n}\langle\varphi_\ell^\alpha, \varphi_\ell^\beta\rangle$ where $\varphi_\ell^\alpha = \varphi_s(z_\ell^\alpha)$ with $s = an^p$ for some $p > 0$, 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. Then in the limit as $n \to \infty, \frac{d}{n} \to T$, the interpolated process $V_{\lfloor tn\rfloor}$ 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 $0 < p < \frac{1}{2}$

$$\begin{cases} V_t^{\alpha\alpha} = 0 \text{ or } \infty, & \text{if } \frac{3}{4}\varphi''(0)^2 + \varphi'''(0) > 0 \text{ and } V_0^{\alpha\alpha} \neq 0\,, \\ V_t^{\alpha\beta} = \text{const.}\,, & \text{if } \frac{3}{4}\varphi''(0)^2 + \varphi'''(0) \le 0\,, \end{cases} \tag{D.19}$$

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

(ii) the critical limit: the solution of the SDE from Theorem 3.9, 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 3.3, if $p > \frac{1}{2}$.

Proof. Similar to the proof of Theorem 3.9, we will borrow the same notation and write down the Markov chain update and consider the time scale depending on the value of p. In case (i) where $0 < p < \frac{1}{2}$, we will consider the time scale $h_n = \frac{1}{s^2} = \frac{1}{a^2n^{2p}}$ and observe that based on the Taylor expansion of φ_s about 0, we can write

$$\begin{aligned} V_{\ell+1}^{\alpha\alpha} &= \frac{c}{n}\sum_{i=1}^n\left(\sqrt{V_\ell^{\alpha\alpha}}g_{\ell,i}^\alpha + \frac{\varphi''(0)}{2s}V_\ell^{\alpha\alpha}(g_{\ell,i}^\alpha)^2 + \frac{\varphi'''(0)}{6s^2}(V_\ell^{\alpha\alpha})^{3/2}(g_{\ell,i}^\alpha)^3 + O(s^{-3})\right)^2 \\ &= cV_\ell^{\alpha\alpha}\frac{1}{n}\sum_{i=1}^n(g_{\ell,i}^\alpha)^2 + (V_\ell^{\alpha\alpha})^{3/2}\frac{\varphi''(0)}{2s}\frac{c}{n}\sum_{i=1}^n 2(g_{\ell,i}^\alpha)^3 \\ &\quad + (V_\ell^{\alpha\alpha})^2\left(\frac{\varphi'''(0)}{3s^2} + \frac{\varphi''(0)^2}{4s^2}\right)\frac{c}{n}\sum_{i=1}^n(g_{\ell,i}^\alpha)^4 + O(s^{-3}) \\ &= cV_\ell^{\alpha\alpha} + cV_\ell^{\alpha\alpha}\frac{1}{\sqrt{n}}R_\ell^{\alpha\alpha} + c(V_\ell^{\alpha\alpha})^{3/2}\frac{\varphi''(0)}{s}\frac{1}{\sqrt{n}}\widehat{R}_\ell^{\alpha\alpha} + c(V_\ell^{\alpha\alpha})^2\left(\frac{\varphi'''(0)}{s^2} + \frac{3\varphi''(0)^2}{4s^2}\right) \\ &\quad + c(V_\ell^{\alpha\alpha})^2\left(\frac{\varphi'''(0)}{3s^2} + \frac{\varphi''(0)^2}{4s^2}\right)\frac{1}{\sqrt{n}}\widetilde{R}_\ell^{\alpha\alpha} + O(s^{-3})\,, \end{aligned} \tag{D.20}$$

where we define $R_\ell^{\alpha\alpha} \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^n(g_{\ell,i}^\alpha)^2 - 1$, $\widehat{R}_\ell^{\alpha\alpha} \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^n(g_{\ell,i}^\alpha)^3$, $\widetilde{R}_\ell^{\alpha\alpha} \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^n(g_{\ell,i}^\alpha)^4 - 3$ and observe they all have zero mean and finite variance.

In view of the time scale s^{-2} for Proposition A.6, it is then only important to keep track of the expected value of the s^{-2} terms and the covariance of the s^{-1} terms. However, since there is no terms on the order of s^{-1}, we essentially have

$$V_{\ell+1}^{\alpha\alpha} = V_\ell^{\alpha\alpha} + \frac{1}{s^2}\left(\varphi'''(0) + \frac{3}{4}\varphi''(0)^2\right) V_\ell^{\alpha\alpha}(V_\ell^{\alpha\alpha} - 1) + O(s^{-3} + n^{-1})\,, \tag{D.21}$$

where we used the fact that $c = 1 - \frac{b}{s^2} + O(s^{-3})$ for $b = \varphi'''(0) + \frac{3}{4}\varphi''(0)^2$ from Lemma D.1.

Hence, we have that $U_t^{\alpha\alpha,n} \coloneqq V_{\lfloor ts^2 \rfloor}^{\alpha\alpha}$ converging to the ODE via Proposition A.6

$$\partial_t U_t^{\alpha\alpha} = b U_t^{\alpha\alpha}(U_t^{\alpha\alpha} - 1)\,, \tag{D.22}$$

where we observe if $b > 0$ this ODE is "mean avoiding" as it will drift towards 0 or ∞. And since the V_t time scale is on the order of $\frac{1}{n}$, for all $t > 0$ we have that

$$V_t^{\alpha\alpha} = U_\infty^{\alpha\alpha}\,, \tag{D.23}$$

therefore if $b > 0$ we have that $V_t^{\alpha\alpha} = 0$ or ∞ as desired in the first case of (i). When $b = 0$ we observe that $V_t^{\alpha\alpha} = V_0^{\alpha\alpha}$ since the time derivative is zero. Furthermore if $b < 0$ we also have that $V_t^{\alpha\alpha} = 1$ in the second case of (i).

When $b \leq 0$, we can also write down the ODE for $U_t^{\alpha\beta}$ using a similar argument and keeping only the s^{-2} terms. More precisely, we can modify (D.18) to get

$$\begin{aligned} V_{\ell+1}^{\alpha\beta} = V_\ell^{\alpha\beta} + \frac{1}{s^2}\left[\frac{\varphi''(0)^2}{4}\left(V_\ell^{\alpha\alpha}V_\ell^{\beta\beta} + V_\ell^{\alpha\beta}(2V_\ell^{\alpha\beta} - 3)\right) + \frac{\varphi'''(0)}{2}V_\ell^{\alpha\beta}(V_\ell^{\alpha\alpha} + V_\ell^{\beta\beta} - 2)\right] \\ + c\sqrt{V_\ell^{\alpha\alpha}V_\ell^{\beta\beta}}\frac{R_\ell^{\alpha\beta}}{\sqrt{n}} + O(n^{-3/2})\,, \end{aligned} \tag{D.24}$$

which leads to the following ODE

$$\partial_t U_t^{\alpha\beta} = \frac{\varphi''(0)^2}{4}\left(U_t^{\alpha\alpha}U_t^{\beta\beta} + U_t^{\alpha\beta}\left(2U_t^{\alpha\beta} - 3\right)\right) + \frac{\varphi'''(0)}{2}U_t^{\alpha\beta}\left(U_t^{\alpha\alpha} + U_t^{\beta\beta} - 2\right)\,. \tag{D.25}$$

Since $U_t^{\alpha\alpha}, U_t^{\beta\beta}$ converge to constants as $t \to \infty$, $|U_t^{\alpha\beta}| \leq \sqrt{U_t^{\alpha\alpha}U_t^{\beta\beta}}$ by definition and Cauchy–Schwarz inequality, and that $U_t^{\alpha\beta}$ satisfies a first order ODE (so it cannot have a periodic solution), we must also have that $\lim_{t\to\infty} U_t^{\alpha\beta} = \text{const}$. This completes the proof for case (i).

Case (ii) follows directly from Theorem 3.9, therefore we can then consider case (iii) with the same Taylor expansion, however this time on the time scale of n^{-1} instead. We will again follow Proposition A.6 to only track the mean of the order n^{-1} term and the variance of the $n^{-1/2}$ term. Since $p > \frac{1}{2}$, the only term that remains is the diffusion on the order of $n^{-1/2}$

$$V_{\ell+1}^{\alpha\beta} = V_\ell^{\alpha\beta} + V_\ell^{\alpha\beta}\frac{1}{\sqrt{n}}R_\ell^{\alpha\beta}\,, \tag{D.26}$$

which gives us the desired SDE from calculating the covariance from Theorem 3.9.

□

D.3 Proof of Proposition 3.7 (Finite Time Explosion Criterion)

We will start by recalling several definitions from [42, Section 5.5]. Firstly, we consider the one dimensional Itô diffusion on $I \coloneqq (0, \infty)$

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t\,, \tag{D.27}$$

where the drift and diffusion coefficients satisfy the following conditions

$$\begin{aligned} &\sigma^2(x) > 0\,, \forall x \in I\,, \\ &\forall x \in I\,, \exists \epsilon > 0 : \int_{x-\epsilon}^{x+\epsilon} \frac{|b(y)|}{\sigma^2(y)}\,dy < \infty\,. \end{aligned} \tag{D.28}$$

We will also define the following functions for some fixed $c \in I$

$$\begin{aligned} p(x) &\coloneqq \int_c^x \exp\left(-\int_c^\xi \frac{2b(z)}{\sigma^2(z)} dz\right) d\xi\,, \\ m(dx) &\coloneqq \frac{2\,dx}{p'(x)\sigma^2(x)}\,, \\ v(x) &\coloneqq \int_c^x p'(x) \int_c^y \frac{2\,dz}{p'(z)\sigma^2(z)}\,dy = \int_c^x p(x) - p(y)\,m(dx)\,. \end{aligned} \tag{D.29}$$

We will also define the following sequence of stopping times for $M > 0$

$$\tau_M \coloneqq \inf\left\{t \geq 0 : X_t \geq M \text{ or } X_t \leq M^{-1}\right\}\,, \tag{D.30}$$

and let $\tau^* \coloneqq \sup_{M>0} \tau_M$. Now we will state the main results we need for finite time explosions.

Lemma D.4 ([42, Problem 5.5.27])**.** *We have the following implications*

$$\begin{aligned} \lim_{x\to 0} p(x) = -\infty \implies \lim_{x\to 0} v(x) = \infty\,, \\ \lim_{x\to \infty} p(x) = \infty \implies \lim_{x\to \infty} v(x) = \infty\,. \end{aligned} \tag{D.31}$$

Theorem D.5 (Feller's Test for Explosions [42, Theorem 5.5.29])**.** *Assume the conditions in (D.28) are satisfied. Then $\mathbb{P}[\tau^* = \infty] = 1$ if and only if*

$$\lim_{x\to 0} v(x) = \lim_{x\to\infty} v(x) = \infty\,. \tag{D.32}$$

We will begin our derivations for the SDE (D.27).

Lemma D.6 (Geometric Brownian Motion, the $b = 0$ Case)**.** *Let X_t be a solution to the following SDE*

$$dX_t = \sqrt{2} X_t\,dB_t\,, \quad X_0 = x_0 > 0\,, \tag{D.33}$$

then we have that $\tau^* = \infty$ a.s.

Proof. Here we observe that

$$p'(x) = \exp(0) = 1 \implies p(x) = x - 1\,. \tag{D.34}$$

Then we have that

$$m(dx) = \frac{2\,dx}{p'(x)\sigma^2(x)} = \frac{dx}{x^2}\,, \tag{D.35}$$

which implies

$$v(x) = (x-1)\int_1^x \frac{dy}{y^2} - \int_1^x \frac{y-1}{y^2} dy = x - \log x - 1\,, \tag{D.36}$$

and therefore

$$\lim_{x\to 0} v(x) = \lim_{x\to\infty} v(x) = \infty\,. \tag{D.37}$$

By Feller's test for explosions Theorem D.5, we have the desired result.

□

Proposition D.7 (Calculate $p(x), m(dx)$ and the $b \leq -1$ Case)**.** *Suppose X_t is a solution of the following equation*

$$dX_t = bX_t(X_t - 1)\,dt + \sqrt{2} X_t\,dB_t\,, \quad X_0 = x_0 > 0\,, \tag{D.38}$$

then for all $b \neq 0$ we have that

$$p(x) = e^b \int_1^x e^{-by} y^b dy\,, \quad m(dx) = \frac{dx}{e^b e^{-bx} x^{b+2}}\,. \tag{D.39}$$

This implies that

$$\lim_{x\to 0} p(x) = \begin{cases} -\infty\,, & b \leq -1\,, \\ \textit{finite}\,, & b > -1\,, \end{cases} \quad \lim_{x\to\infty} p(x) = \begin{cases} \infty\,, & b \leq 0\,, \\ \textit{finite}\,, & b > 0\,. \end{cases} \tag{D.40}$$

In particular, when $b \leq -1$, we have that $\lim_{x\to 0} v(x) = \lim_{x\to\infty} v(x) = \infty$.

Proof. We start by writing

$$p'(x) = \exp\left(-\int_1^x \frac{2b(y)}{\sigma^2(y)}dy\right) = \exp\left(-b(x-1) + b\log x\right) = e^b e^{-bx} x^b \,. \tag{D.41}$$

Then we can also calculate the integral via a substitution of $y = bx$ to get the desired result.

At this time, we observe that when $b > 0$

$$\begin{aligned} p(x) &= \int_1^x p'(y)\,dy \\ &= e^b \int_1^x e^{-by} y^b \,dy \\ &= e^b b^{-b-1} \int_b^{bx} e^{-z} z^b \,dz \\ &= e^b b^{-b-1} \left(\gamma(b+1, bx) - \gamma(b+1, b)\right) \,, \end{aligned} \tag{D.42}$$

where γ is the lower incomplete gamma function, and therefore finite for all values of x including the limits $x \to 0, \infty$.

The $b = 0$ case follows from Lemma D.6. Finally when $b < 0$ we can write

$$p(x) = e^{-|b|} \int_1^x \frac{e^{|b|y}}{y^{|b|}}\,dy \,, \tag{D.43}$$

which clearly diverges to ∞ as $x \to \infty$.

On the other hand, we can observe as that as $x \to 0$, we have that $y \in [0,1]$ and therefore $1 \le e^{|b|y} \le e^{|b|}$. This implies we only need to consider the integral $-\int_x^1 y^{-|b|}\,dy$, which diverges to $-\infty$ if and only if $|b| \ge 1$. In other words we have

$$\lim_{x\to 0} p(x) = \begin{cases} -\infty \,, & b \le -1 \,, \\ \text{finite} \,, & b > -1 \,. \end{cases} \tag{D.44}$$

The limits on $v(x)$ follows from Lemma D.4.

□

Proposition D.8 (The $b > -1$ Case)**.** *Suppose X_t is a solution of the following equation*

$$dX_t = bX_t(X_t - 1)\,dt + \sqrt{2}X_t\,dB_t \,, \quad X_0 = x_0 > 0 \,, \tag{D.45}$$

then when $b > -1$, we have that

$$\lim_{x\to 0} v(x) = \infty \,, \quad \lim_{x\to\infty} v(x) = \begin{cases} \infty \,, & b \in (-1, 0] \,, \\ < \infty \,, & b > 0 \,. \end{cases} \tag{D.46}$$

Proof. We will start by calculating the following integral using the exponential series expansion

$$\begin{aligned} \int_1^y \frac{2\,dz}{p'(z)\sigma^2(z)} &= e^{-b} \int_1^y e^{bz} z^{-(b+2)}\,dz \\ &= e^{-b} \int_1^y \sum_{k\ge 0} \frac{(bz)^k}{k!} z^{-(b+2)}\,dz \\ &= e^{-b} \sum_{k\ge 0, k\ne b+1} \frac{b^k}{k!} \frac{y^{k-b-1} - 1}{k-b-1} + \frac{b^{b+1}}{(b+1)!}\log(y)\mathbb{1}_{\{k=b+1\}} \,. \end{aligned} \tag{D.47}$$

Now we can compute $v(x)$

$$\begin{aligned} v(x) &= \int_1^x e^{-by} y^b \left(\sum_{k\ge 0, k\ne b+1} \frac{b^k}{k!} \frac{y^{k-b-1} - 1}{k-b-1} + \frac{b^{b+1}}{(b+1)!}\log(y)\mathbb{1}_{\{k=b+1\}} \right) dy \\ &= \sum_{k\ge 0, k\ne b+1} \frac{b^k}{k!(k-b-1)} \int_1^x e^{-by}(y^{k-1} - y^b)dy + \frac{b^{b+1}}{(b+1)!}\mathbb{1}_{\{k=b+1\}} \int_1^x e^{-by} y^b \log y\,dy \,. \end{aligned} \tag{D.48}$$

We first consider the case when $x \to 0$, in which case we have $e^{-|b|} \leq e^{-by} \leq e^{|b|}$ and therefore will not affect convergence or divergence, so we can safely ignore the factor e^{-by} and write (for $k > 0, x \to 0$)

$$\int_1^x e^{-by}(y^{k-1} - y^b)\, dy \approx \left[\frac{y^k}{k} - \frac{y^{b+1}}{b+1}\right]_1^x = \frac{x^k - 1}{k} - \frac{x^{b+1} - 1}{b+1} \to \frac{-1}{k} + \frac{1}{b+1}\,. \tag{D.49}$$

Since the exponential series $\sum_{k>0} \frac{b^k}{k!} = e^b - 1$ converges, and we have terms strictly smaller than the exponential series, we have convergence of these terms when $k > 0$. We now return to handle a couple of edge case terms, firstly when $k = 0$

$$\frac{1}{-b-1}\int_1^x e^{-by} y^{-1}\, dy \approx \frac{-\log x}{|1+b|} \to \infty\,, \quad \text{as } x \to \infty\,, \tag{D.50}$$

which is a desired behaviour. Secondly we consider when $k = b + 1$

$$\int_1^x y^b \log y\, dy = \frac{x^{b+1}\left[(b+1)\log x - 1\right] + 1}{(b+1)^2} \to (b+1)^{-2}\,, \quad \text{as } x \to \infty\,, \tag{D.51}$$

from which we can conclude $\lim_{x\to 0} v(x) = \infty$.

Next we consider the case when $x \to \infty$. Firstly, since we already have that $p(x) \to \infty$ when $b \leq 0$, therefore Lemma D.4 implies $v(x) \to \infty$. Therefore we only need to consider when $b > 0$.

Since $b > 0$ we will have that e^{-bx} will dominate, and therefore we can safely ignore all the edge case terms and consider the series

$$v(x) \approx \sum_{k>0, k\neq b+1} \frac{b^k}{k!(k-b-1)} \int_1^x e^{-by}(y^{k-1} - y^b)\, dy\,. \tag{D.52}$$

Observe that as $x \to \infty$ we actually recover the gamma integral in the terms i.e.

$$\int_1^\infty e^{-by}(y^{k-1} - y^b)\, dy = -b^{-k}\Gamma(k) + b^{-b-1}\Gamma(b+1)\,, \tag{D.53}$$

where we observe the second term is independent of k, and therefore the series converges due to comparison with the exponential Taylor series. This implies we only need to focus on the first term, which is

$$v(x) \approx \sum_{k>0, k\neq b+1} \frac{b^k}{k!(k-b-1)}(-b^{-k})(k-1)! = \sum_{k>0, k\neq b+1} \frac{-1}{k(k-b-1)} < \infty\,, \tag{D.54}$$

where the series converges since it's a sum of k^{-2} type. This allows us to conclude that $\lim_{x\to\infty} v(x) < \infty$ as desired.

□

We can now prove the desired result of Proposition 3.7, which we restate below.

Proposition D.9 (Finite Time Explosion)**.** *Let $X_t \in \mathbb{R}_+$ be a solution to the following SDE*

$$dX_t = bX_t(X_t - 1)\, dt + \sqrt{2}X_t\, dB_t\,, \quad X_0 = x_0 > 0\,, b \in \mathbb{R}\,. \tag{D.55}$$

Let $\tau^* = \sup_{M>0} \inf\{t : X_t \geq M \text{ or } X_t \leq M^{-1}\}$ be the explosion time, and we say X_t has a finite time explosion if $\tau^* < \infty$. For this equation, $\mathbb{P}[\tau^* = \infty] = 1$ if and only if $b \leq 0$.

Proof. Putting the results of Proposition D.7 and Proposition D.8 together, we have the following table

	$\lim_{x\to 0} v(x)$	$\lim_{x\to\infty} v(x)$	$\lim_{x\to 0} p(x)$	$\lim_{x\to\infty} p(x)$
$b \leq -1$	∞	∞	$-\infty$	∞
$-1 < b \leq 0$	∞	∞	finite	∞
$b > 0$	∞	finite	finite	finite

Therefore, invoking Feller's test for explosions from Theorem D.5, we have that $\mathbb{P}[\tau^* = \infty]$ if and only if $b \leq 0$.

□

E Lower Bound for the Recursion $\rho_{\ell+1} = cK_1(\rho_\ell)$

In this section, we consider a Taylor expansion of $cK_1(\rho)$ around $\rho \to 1$ from the left hand side to get

$$\rho_{\ell+1} = cK_1(\rho_\ell) = \rho_\ell + \frac{2\sqrt{2}}{3\pi}(1-\rho_\ell)^{3/2} + O((1-\rho_\ell)^{5/2}), \tag{E.1}$$

which we can rewrite using $r_\ell = 1 - \rho_\ell$ as

$$r_{\ell+1} = r_\ell - \frac{2\sqrt{2}}{3\pi} r_\ell^{3/2} + O(r_\ell^{3/2}). \tag{E.2}$$

We will compute an upper bound on r_ℓ inspired by the following result.

Lemma E.1 (Lemma A.6, [52])**.** *The logistic recursion*

$$x_{n+1} \leq \alpha x_n(1-x_n), \tag{E.3}$$

for $x_0, \alpha \in [0,1]$ *satisfies*

$$x_n \leq \frac{x_0}{\alpha^{-n} + x_0 n}. \tag{E.4}$$

We will extend the above Lemma to a slightly modified update as well.

Lemma E.2. *Suppose the recursive map satisfies*

$$x_{n+1} \leq x_n(1 - x_n^{1/2}), \tag{E.5}$$

for $x_0 \in [0,1]$*, then we also have that*

$$x_n \leq \frac{x_0}{\left(1 + \frac{1}{3} n x_0^{1/2}\right)^2}. \tag{E.6}$$

Proof. We will start the induction proof at $n = 1$

$$x_1 \leq x_0(1 - x_0^{1/2}) \leq \frac{x_0}{1 + x_0^{1/2}}. \tag{E.7}$$

When $x_0 \leq 9$ we have that

$$1 + x_0^{1/2} \geq 1 + \frac{1}{9}x_0 + \frac{2}{3}x_0^{1/2} = \left(1 + \frac{1}{3}x_0^{1/2}\right)^2, \tag{E.8}$$

and hence

$$x_1 \leq \frac{x_0}{\left(1 + \frac{1}{3}x_0^{1/2}\right)}, \tag{E.9}$$

which proves the case for $n = 1$.

Then we assume the inequality holds for x_n, we will similarly write

$$x_{n+1} \leq x_n(1 - x_n^{1/2}) \leq \frac{x_n}{1 + x_n^{1/2}}, \tag{E.10}$$

and plugging in the inequality for x_n we get

$$x_{n+1} \leq \frac{\frac{x_0}{\left(1+\frac{1}{3}nx_0^{1/2}\right)^2}}{1 + \frac{x_0^{1/2}}{1+\frac{1}{3}nx_0^{1/2}}} = \frac{x_0}{\left(1 + \left(\frac{n}{3} + 1\right)x_0^{1/2}\right)\left(1 + \left(\frac{n}{3}\right)x_0^{1/2}\right)}, \tag{E.11}$$

To complete the proof it's sufficient to show

$$\left(1 + \left(\frac{n}{3} + 1\right)x_0^{1/2}\right)\left(1 + \left(\frac{n}{3}\right)x_0^{1/2}\right) \geq \left(1 + \left(\frac{n+1}{3}\right)x_0^{1/2}\right)^2, \tag{E.12}$$

which is equivalent to

$$\left(\frac{n}{3}+1\right)\frac{n}{3}x_0+\left(\frac{2n}{3}+1\right)x_0^{1/2}\geq\frac{(n+1)^2}{9}x_0+\frac{2(n+1)}{3}x_0^{1/2}\,. \tag{E.13}$$

Since $\frac{2n}{3}+1\geq\frac{2(n+1)}{3}$, we only need to compare the first coefficient, which is

$$\frac{n^2+3n}{9}\geq\frac{n^2+2n+1}{9}\,, \tag{E.14}$$

and this is equivalent to $n\geq 1$, and therefore satisfied by the induction. This completes the proof.

□

At the same time, we also conjecture the following bound.

Conjecture E.3. *Suppose the recursive map satisfies*

$$x_{n+1}=x_n(1-x_n^{1/2})\,, \tag{E.15}$$

for $x_0\in[0,1]$*, then*

$$x_n\approx\frac{x_0}{\left(1+\frac{1}{2}nx_0^{1/2}\right)^2}\,. \tag{E.16}$$

Sketch of Conjecture. Suppose we want to establish the approximation of

$$x_n\leq\frac{x_0}{\left(1+bnx_0^{1/2}\right)^2}\,. \tag{E.17}$$

Then for the initial induction $n=1$ step, we only need

$$1+x_0^{1/2}\geq(1+bx_0^{1/2})^2\,, \tag{E.18}$$

which is equivalent to

$$x_0\leq\frac{(1-2b)^2}{b^4}\,. \tag{E.19}$$

Using WolframAlpha (probably through the quartic formula), we find the desired solution for $b\in(0,1/2)$ is

$$b=\frac{\sqrt{1+\sqrt{x_0}}-1}{\sqrt{x_0}}\,. \tag{E.20}$$

This function $b(x_0)$ is a strictly decreasing function on $[0,1]$, and it satisfies $b(0)=\frac{1}{2}, b(1)=\sqrt{2}-1$. This implies that whenever x_0 is small, we can choose b closer to $\frac{1}{2}$ in the $n=1$ step of the induction.

Similarly, for the induction step, it's sufficient to show

$$\left(1+(nb+1)x_0^{1/2}\right)\left(1+nbx_0^{1/2}\right)\geq\left(1+(n+1)bx_0^{1/2}\right)^2\,, \tag{E.21}$$

which is equivalent to

$$nbx_0^{1/2}+1\geq(2n+1)b^2x_0^{1/2}+2b\,. \tag{E.22}$$

Again, since we are always choosing $b\leq 1/2$, therefore we have $1\geq 2b$, and we will only need to focus on the first coefficient. To this end we rewrite the first term as

$$b(n(1-2b)-b)x_0^{1/2}=b(1-2b)\left(n-\frac{b}{1-2b}\right)x_0^{1/2}\,. \tag{E.23}$$

This implies we require $n\geq\frac{b}{1-2b}$, which increases as we choose b closer to $1/2$. However, if the induction starts the step $\lceil\frac{b}{1-2b}\rceil$, then this is not a problem, which leads to our conjecture.

□

Using the above results, we can have a similar approximation for r_ℓ given the infinite-width update

$$r_{\ell+1} = r_\ell - \frac{2\sqrt{2}}{3\pi} r_\ell^{3/2}, \tag{E.24}$$

which we can rewrite using $\hat{r}_\ell := \left(\frac{2\sqrt{2}}{3\pi}\right)^2 r_\ell$ to get

$$\hat{r}_{\ell+1} = \hat{r}_\ell (1 - \hat{r}_\ell^{1/2}) \,. \tag{E.25}$$

This allows us to consider the upper bound

$$\hat{r}_\ell \leq \frac{\hat{r}_0}{\left(1 + \frac{1}{3}\ell\, \hat{r}_0^{1/2}\right)^2}, \tag{E.26}$$

or equivalently

$$r_\ell \leq \frac{r_0}{\left(1 + \frac{2\sqrt{2}}{9\pi}\ell\, r_0^{1/2}\right)^2} \,. \tag{E.27}$$

Similarly, the conjecture leads to the following approximation

$$r_\ell \approx \frac{r_0}{\left(1 + \frac{\sqrt{2}}{3\pi}\ell\, r_0^{1/2}\right)^2} \,. \tag{E.28}$$

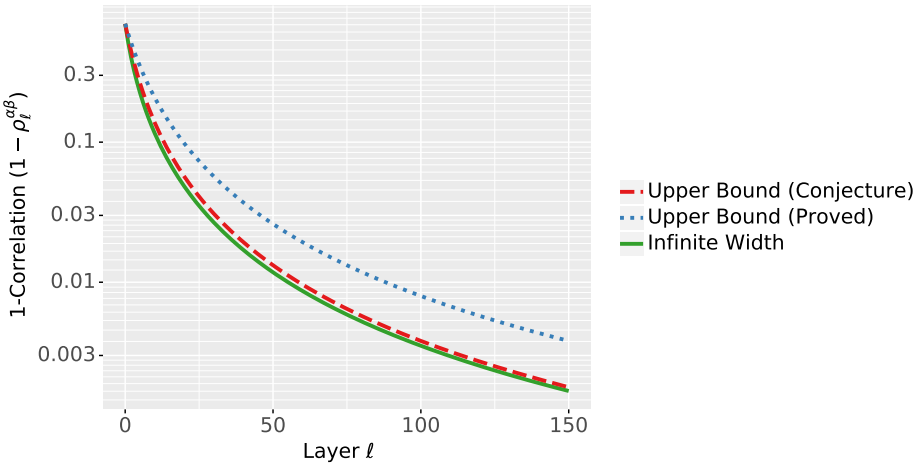

Figure 3: Plot of the convergence of correlation $\rho_\ell^{\alpha\beta}$ to 1 for a ReLU network, and the lower bounds (E.27) and (E.28). Computed with $d = n = 150, \rho_0^{\alpha\beta} = 0.3$, using the usual ReLU activation i.e. $\varphi_s(x) = \max(x, 0)$.

F Additional Simulations and Discussions

In this section, we have additional simulations plotting the densities of $\rho_d^{\alpha\beta}$ and $V_d^{\alpha\beta}$ for shaped ReLU-like, sigmoid, and softplus networks. In particular, the density of $V_d^{\alpha\beta}$ for ReLU-like networks can be found in Figure 4, the densities for sigmoid in Figure 5, and the densities for softplus in Figure 6.

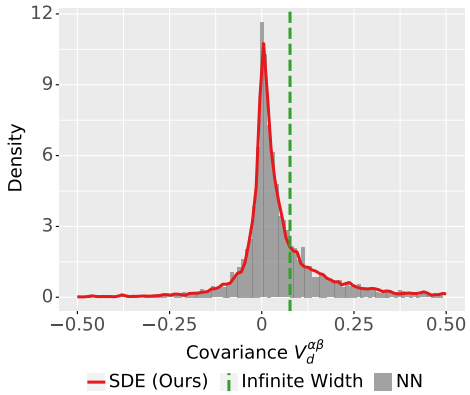

Figure 4: Empirical distribution of the covariance $V_d^{\alpha\beta}$ for a ReLU-like network, SDE sample density computed via kernel density estimation. Infinite width prediction simulated from the ODE $\partial_t \rho_t^{\alpha\beta} = \nu(\rho_t^{\alpha\beta})$, and we note $V_t^{\alpha\alpha} = V_0^{\alpha\alpha}$ in the infinite width limit. Simulated with $n = d = 150, c_+ = 0, c_- = -1, \rho_0^{\alpha\beta} = 0.3$, SDE and ODE step size 10^{-2}, and 2^{13} samples.

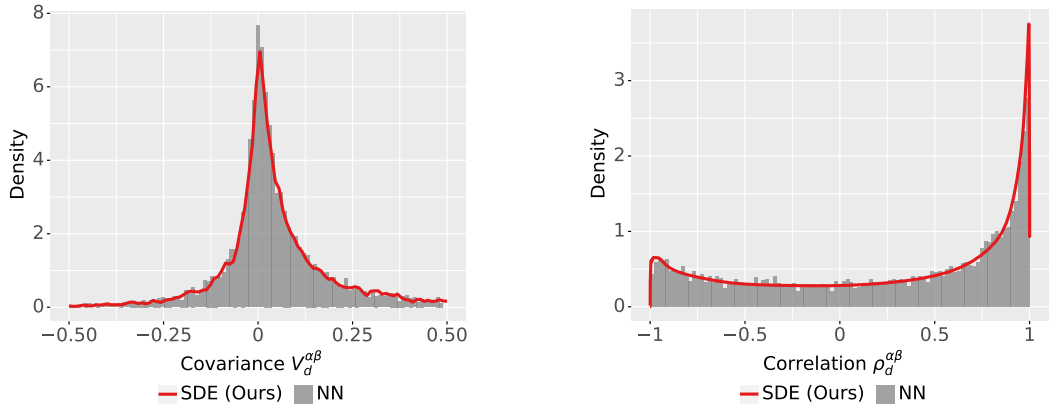

Figure 5: Empirical distribution of the covariance $V_d^{\alpha\beta}$ and correlation $\rho_d^{\alpha\beta}$ for a shaped sigmoid network, SDE sample density computed via kernel density estimation. Simulated with $n = d = 150, a = 1, \rho_0^{\alpha\beta} = 0.3$, SDE step size 10^{-2}, and 2^{13} samples.

F.1 Convergence in Kolmogorov–Smirnov Distance

From Figure 7, we can show that our results (Theorem 3.3) converges at a rate of $n^{-1/2}$ in terms of the KS-distance.

F.2 Tuning Shape and Depth-to-Width Ratio

Since the existing shaping methods [38, 39] estimates the output correlation based on the infinite-width limit, we can easily improve the shape tuning based on the covariance SDEs. In particular, we consider the example of ReLU-like activations with correlation described by the SDE (3.4). By simulating both the SDE and the infinite-width limit ODE, we arrive at the results in Figure 8.

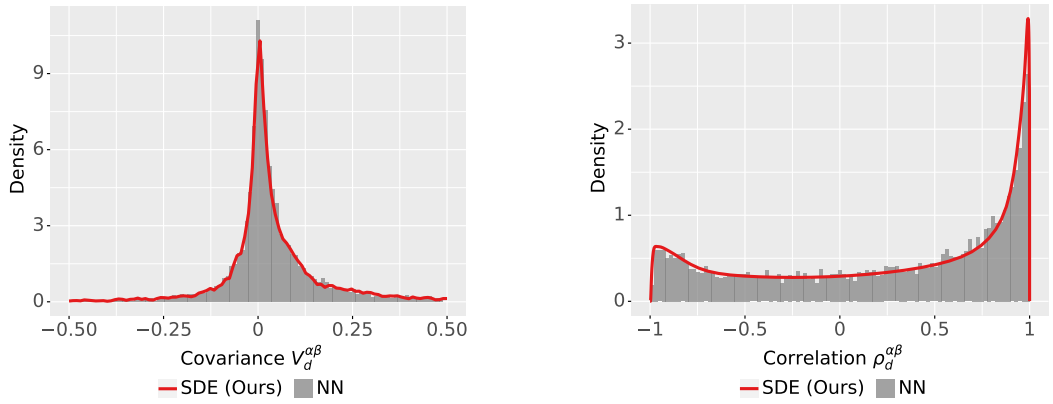

Figure 6: Empirical distribution of the covariance $V_d^{\alpha\beta}$ and correlation $\rho_d^{\alpha\beta}$ for a shaped softplus network (centered at $x_0 = \log 2$), SDE sample density computed via kernel density estimation. Simulated with $n = d = 150, a = 1, \rho_0^{\alpha\beta} = 0.3$, SDE step size 10^{-2}, and 2^{13} samples.

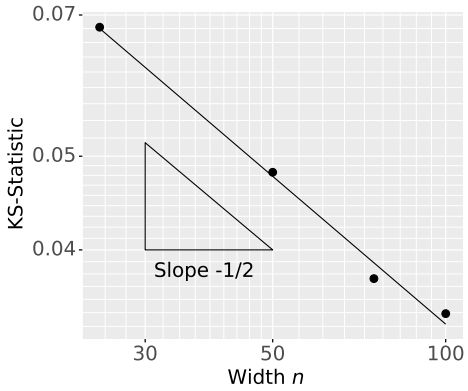

Figure 7: The Kolmogorov–Smirnov statistic (sup norm of the difference between two empirical CDFs) for the empirical samples of the correlation SDE (3.4) and from a neural network at initialization. Simulated with $c_+ = 0, c_- = -1, \rho_0^{\alpha\beta} = 0.3, \frac{d}{n} = T = 1$, SDE step size 10^{-2}, and 2^{13} samples.

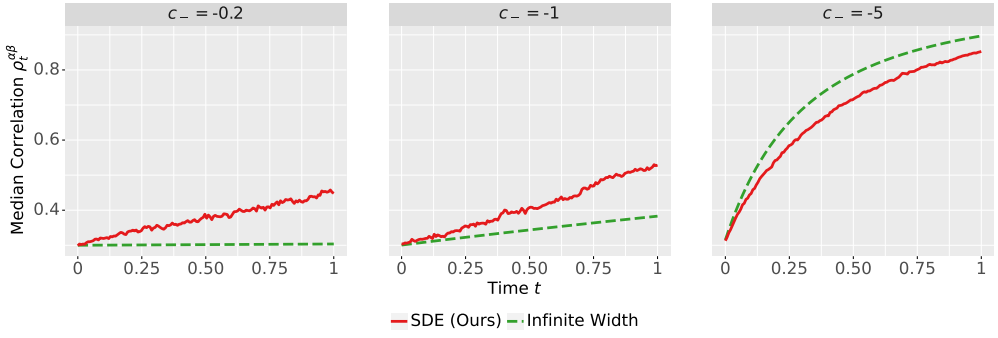

Figure 8: ReLU Correlation SDE (3.4) and ODE simulated with $c_+ = 0, \rho_0^{\alpha\beta} = 0.3$ varying c_- values, 2^{12} samples, and step size 10^{-2}. infinite-width is from ODE $\partial_t \rho_t^{\alpha\beta} = \nu(\rho_t^{\alpha\beta})$ with ν.

We observe that simply by increasing c_- towards zero does not automatically reduce effects on the correlation when time t (the depth-to-width ratio) is large. In other words, even a linear network will observe an increase in correlation when depth is large enough. Therefore *shaping the activation alone is insufficient*, but we also need to account for the depth-to-width ratio.

We also remark that Figure 8 only plotted the median for simplicity, but if we recall the density plots from Figure 1, correlation is heavily skewed and concentrated near 1. More precisely, while the median correlation is approximately 0.55, roughly 20% of the samples are larger than 0.9. In other words, one in five random initializations will lead to a correlation worse than 0.9! As a consequence, practitioners implementing the shaping methods of [38, 39] should consider simulating the correlation SDE to account for the heavy skew.