Journal of Machine Learning Research 25 (2024) 1-45 Submitted 3/23; Revised 7/24; Published 8/24

Depth Degeneracy in Neural Networks: Vanishing Angles in Fully Connected ReLU Networks on Initialization

Cameron Jakub cjakub@uoguelph.ca
Department of Mathematics & Statistics
University of Guelph

Mihai Nica nicam@uoguelph.ca
Department of Mathematics & Statistics
University of Guelph

Editor: Miguel Carreira-Perpinan

Abstract

Despite remarkable performance on a variety of tasks, many properties of deep neural networks are not yet theoretically understood. One such mystery is the depth degeneracy phenomenon: the deeper you make your network, the closer your network is to a constant function on initialization. In this paper, we examine the evolution of the angle between two inputs to a ReLU neural network as a function of the number of layers. By using combinatorial expansions, we find precise formulas for how fast this angle goes to zero as depth increases. These formulas capture microscopic fluctuations that are not visible in the popular framework of infinite width limits, and leads to qualitatively different predictions. We validate our theoretical results with Monte Carlo experiments and show that our results accurately approximate finite network behaviour. We also empirically investigate how the depth degeneracy phenomenon can negatively impact training of real networks. The formulas are given in terms of the mixed moments of correlated Gaussians passed through the ReLU function. We also find a surprising combinatorial connection between these mixed moments and the Bessel numbers that allows us to explicitly evaluate these moments.

Keywords: deep learning theory, infinite limits of neural networks, network initialization, Markov chains, combinatorics

1. Introduction

The idea of stacking many layers to make truly *deep* neural networks (DNNs) is what arguably led to the neural net revolution in the 2010s. Indeed, from a function-space point of view, it is known that depth exponentially improves expressibility (Poole et al., 2016; Eldan and Shamir, 2015). However, an important but less well known fact is that under standard initialization schemes, deep neural networks become more and more *degenerate* as depth gets larger and larger. One sense in which this happens is the phenomenon of vanishing and exploding gradients (Hanin, 2018). Another sense in which networks become degenerate is that a neural network gets closer and closer to a (random) constant function, i.e. the network sends all inputs to the same output and cannot distinguish input points. This phenomenon seems to have been discovered and analyzed from different points of view by several authors (Avelin and Karlsson, 2022; Dherin et al., 2022; Li et al., 2022; Hayou et al., 2019; Schoenholz et al., 2017; Nachum et al., 2022; Buchanan et al., 2021). Nachum et al. (2022) found for convolutional neural networks, the level of degeneracy was dependent on the type of input fed into the network.

One method already proposed to deal with the degeneracy phenomenon is the idea of activation function *shaping* (Martens et al., 2021). In particular, Li et al. (2022), showed that rescaling the non-linear activation function (i.e. using Leaky ReLUs with leakiness depending on network depth) can lead to a non-trivial angle between inputs. However, a detailed analysis of the evolution of the angle θ *without* any scaling (e.g. using an ordinary ReLU in all layers) remained an outstanding problem. This is the gap we fill in this article.

1.1 Main Results for the Angle Process θ_ℓ

In this paper, we examine the evolution of the *angle* θ_ℓ between two arbitrary inputs $x_\alpha, x_\beta \in \mathbb{R}^{n_{in}}$ after passing through ℓ layers of a fully connected ReLU network (a.k.a. a multi-layer perceptron) on initialization. The angle is defined in the usual way by the inner product between two vectors in $\mathbb{R}^{n_\ell}$

$$\cos(\theta_\ell) := \frac{\langle F^\ell(x_\alpha), F^\ell(x_\beta)\rangle}{\|F^\ell(x_\alpha)\|\|F^\ell(x_\beta)\|},$$

where n_ℓ is the width (i.e. number of neurons) of the ℓ-th layer and $F^\ell : \mathbb{R}^{n_{in}} \to \mathbb{R}^{n_\ell}$ is the (random) neural network function mapping input to the post-activation logits in layer ℓ on initialization. We assume here that the initialization is done with appropriately scaled independent Gaussian weights so that the network is on the "edge of chaos" (Hayou et al., 2019; Schoenholz et al., 2017), where the variance of each layer is order one as layer width increases. See Table 2 for our precise definition of the fully connected ReLU neural network.

With this setup, since the effect of each layer is independent of everything previous, θ_ℓ can be thought of as a Markov chain evolving as layer number ℓ increases. As expected by the aforementioned "large depth degeneracy" phenomenon, we observe that the angle concentrates $\theta_\ell \to 0$ as $\ell \to \infty$ (see Figure 1 for an illustration). This indicates that the hidden layer representation of *any* two inputs becomes closer to co-linear as depth increases.

We obtain a simple, yet remarkably accurate, approximation for the evolution of θ_ℓ as a function of ℓ that captures precisely how quickly this degeneracy happens for small angles θ_ℓ and large layer widths n_ℓ. In Section 1.2, we also empirically investigate how these predictions relate to network performance *after* training and show they may have applications in neural architecture search.

Approximation 1 *For small angles $\theta_\ell \ll 1$ and large layer width $n_\ell \gg 1$, the angle $\theta_{\ell+1}$ at layer $\ell+1$ is well approximated by*

$$\ln\sin^2(\theta_{\ell+1}) \approx \ln\sin^2(\theta_\ell) - \frac{2}{3\pi}\theta_\ell - \rho(n_\ell), \tag{1}$$

where $\rho(n_\ell)$ is a constant which depends on the width n_ℓ of layer ℓ, namely:

$$\rho(n) := \ln\left(\frac{n+5}{n-1}\right) - \frac{10n}{(n+5)^2} + \frac{6n}{(n-1)^2} = \frac{2}{n} + \mathcal{O}\left(n^{-2}\right). \tag{2}$$

Figure 1 illustrates how well this prediction matches Monte Carlo simulations of θ_ℓ sampled from real networks. Also illustrated is the *infinite width* prediction for θ_ℓ (discussed

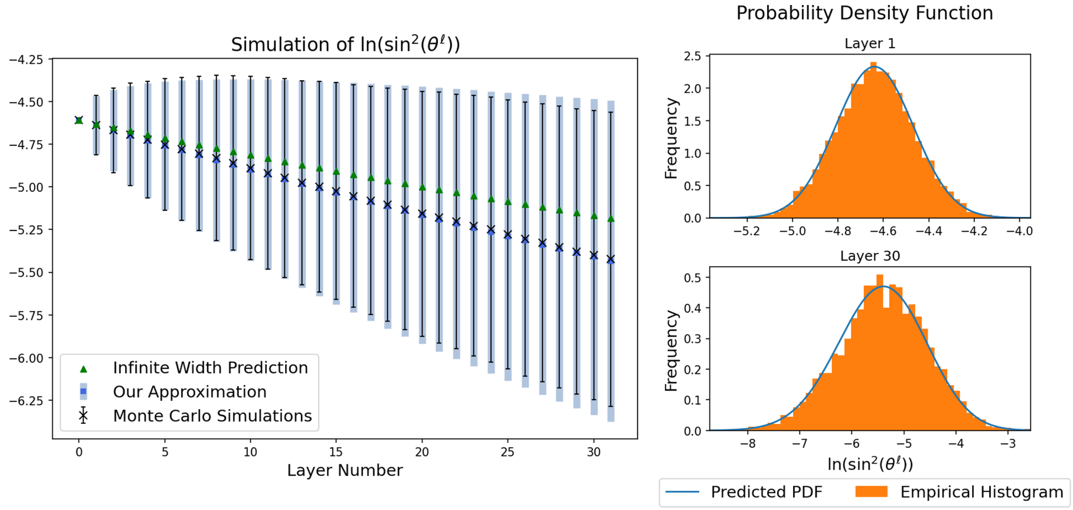


Figure 1: We feed 2 inputs with initial angle $\theta_0 = 0.1$ into 5000 Monte Carlo samples of independently initialized networks with network width $n_\ell = 256$ for all layers. *Left:* Using the Monte Carlo samples, we plot the empirical mean and standard deviation of $\ln(\sin^2(\theta_\ell))$ at each layer. We compare this to both the infinite width update rule and our prediction using Approximation 1 for the mean of $\ln(\sin^2(\theta_\ell))$ (Shown as the blue square). Our prediction for the standard deviation in each layer using Approximation 2 is also plotted as the shaded area. To compute this, we iterate Approximation 2 to estimate the PDF in each layer, and then compute the variance using the PDF. In contrast to our prediction, the infinite width rule predicts 0 variance in all layers. *Right:* We plot histograms of our simulations as well as our predicted probability density function using Approximation 2 from (10) at Layer 1 (top) and Layer 30 (bottom). The predicted PDF is computed numerically by iterating Approximation 2 over the 30 layers, using the PDF in layer ℓ to get the PDF in layer $\ell + 1$. The predicted and empirical distribution are statistically indistinguishable according to a Kolmogorov-Smirnov test, with p values $0.987 > 0.05$ (top) and $0.186 > 0.05$ (bottom). The code which produced this figure can be found at the following `GitHub link`.

in Appendix A.9) which is less accurate at predicting finite width network behaviour than our formula, due to the n_ℓ^{-1} effects that our formula captures in the term $\rho(n_\ell)$ but are not present in the infinite width formula. Approximation 1 is a simple corollary to the mathematically rigorous statement, Theorem 1, for the mean and variance of the random variable $\ln \sin^2(\theta_\ell)$. Approximation 1 is obtained by doing a Taylor series expansion around $\theta = 0$ and ignoring terms of size $\mathcal{O}(n^{-2})$ from the more precise Theorem 1 concerning the random variable $\ln \sin^2(\theta_\ell)$, which is why Approximation 1 only holds for $\theta \ll 1$ and $n \gg 1$. See also Corollary 2 for related expansions.

1.1.1 Theoretical Consequences and Comparison to Previous Work

Approximation 1 predicts that $\theta_\ell \to 0$ *exponentially fast* in ℓ due to $\rho(n)$; it predicts:

$$\theta_\ell \leq \exp\left(-\frac{1}{2}\sum_{i=1}^{\ell}\rho(n_i)\right) = \exp\left(-\sum_{i=1}^{\ell}\frac{1}{n_i} + \mathcal{O}(n_i^{-2})\right).$$

(Note that the exponential behaviour vanishes when $n_\ell \to \infty$ with ℓ fixed). In contrast to this prediction, an analysis using only expected values or equivalently working in the infinite-width $n \to \infty$ limit predicts that $\theta_\ell \to 0$ like ℓ^{-1}, which is qualitatively very different! This difference in the rate of degeneracy demonstrates significant difference between "real world" and infinite width networks. See also Figure 4 for comparisons of the infinite vs finite width predictions for some real architectures.

The prediction of the infinite width degeneracy was first demonstrated under the name "edge of chaos" (Hayou et al., 2019; Schoenholz et al., 2017) and again in Roberts et al. (2022); Hanin (2023). These earlier works studied the correlation $\cos(\theta_\ell)$ as a function of layer number, and showed showed that $1 - \cos(\theta_\ell) \to 0$ like ℓ^{-2}, which is equivalent to $\theta_\ell \to 0$ like ℓ^{-1} by Taylor series expansion $\cos(x) \approx \frac{1}{2}x^2$ as $x \to 0$. To unify the notation, we also present a derivation of the update rule for $\cos(\theta_\ell)$ in the infinite width limit in our notation in Appendix A.9.

We can also recover the infinite width prediction from our result by replacing $\rho(n)$ with 0 in Approximation 1 in the update rule (1). Exponentiating both sides and using $\sin(\theta) \approx \theta, e^\theta \approx 1 + \theta$ for $\theta \ll 1$, Approximation 1 becomes $(\theta_{\ell+1})^2 \approx \big(\theta^\ell)^2(1 - \frac{2}{3\pi}\theta_\ell\big)$, which is equivalent to the result of Proposition C.1 of Hanin (2023) and is also a corollary of Lemma 1 of Hayou et al. (2019). In those papers, the rule was derived directly from the infinite width update rule for $\cos(\theta)$, is equivalent since $\theta_\ell \approx \frac{1}{\ell}$ as $\ell \to \infty$.

One of the main limitations of the infinite width predictions is that they predict zero variance in the random variable θ_ℓ. In contrast to this, our methods allow us to also understand the variance of this random variable, as discussed below.

1.1.2 More detailed results for the mean and variance

Approximation 1 comes from a simplification of more precise formulas for the mean and variance of the random variable $\ln(\sin^2(\theta_\ell))$, which are stated in Theorem 1 below.

Theorem 1 (Formula for mean and variance in terms of J functions) *Conditionally on the angle θ_ℓ in layer ℓ (see Table 2 for a precise definition of all the notations), the mean and variance of* $\ln\sin^2(\theta_{\ell+1})$ *obey the following limit as the layer width $n_\ell \to \infty$*

$$\mathbf{E}[\ln\sin^2(\theta_{\ell+1})|\theta_\ell] = \mu(\theta_\ell, n_\ell) + \mathcal{O}(n_\ell^{-2}), \quad \mathbf{Var}[\ln\sin^2(\theta_{\ell+1})|\theta_\ell] = \sigma^2(\theta_\ell, n_\ell) + \mathcal{O}(n_\ell^{-2}), \tag{3}$$

$$\begin{aligned}\mu(\theta, n) :=& \ln\left(\frac{(n-1)(1-4J_{1,1}^2)}{4J_{2,2}-1+n}\right) + \frac{4(J_{2,2}+1)}{n\left(\frac{4J_{2,2}-1}{n}+1\right)^2} \\ &- \frac{4\left(8J_{1,1}^2J_{2,2} - 8J_{1,1}^4 + 4J_{1,1}^2 - 8J_{1,1}J_{3,1} + J_{2,2} + 1\right)}{n\left(1-\frac{1}{n}\right)^2\left(1-4J_{1,1}^2\right)^2},\end{aligned} \tag{4}$$

$$\begin{aligned}\sigma^2(\theta, n) :=& \frac{8n(J_{2,2}+1)}{(4J_{2,2}-1+n)^2} + \frac{8n(8J_{1,1}^2J_{2,2} - 8J_{1,1}^4 + 4J_{1,1}^2 - 8J_{1,1}J_{3,1} + J_{2,2} + 1)}{(n-1)^2(1-4J_{1,1}^2)^2} \\ &- \frac{16n(2J_{1,1}^2 - 4J_{1,1}J_{3,1} + J_{2,2} + 1)}{(4J_{2,2}-1+n)(n-1)(1-4J_{1,1}^2)},\end{aligned} \tag{5}$$

where $J_{a,b} := J_{a,b}(\theta_\ell)$ are the joint moments of correlated Gaussians passed through the ReLU function $\varphi(x) = \max\{x, 0\}$, namely

$$J_{a,b}(\theta) := \mathbf{E}_{G,\hat{G}}[\varphi^a(G)\varphi^b(\hat{G})], \tag{6}$$

where $G, \hat{G}$ are marginally $\mathcal{N}(0,1)$ random variables with correlation $\mathbf{E}[G\hat{G}] = \cos(\theta)$.

The joint moments $J_{a,b}(\theta)$ are discussed in detail in Section 3. A new combinatorial method of computing these moments is presented, which is used to give an explicit formula is given for these joint-moments, which is presented in Theorem 5. Using the explicit formula for $J_{a,b}$, the result of Theorem 1 can be used to obtain useful asymptotic formulas for μ and σ, as in the following corollary.

Corollary 2 (Small θ asymptotics for mean and variance) *Conditionally on the angle θ_ℓ in layer ℓ, the mean and variance of $\ln\sin^2(\theta_{\ell+1})$ obey the following limit as the layer width $n_\ell \to \infty$*

$$\mathbf{E}[\ln\sin^2(\theta_{\ell+1})] = \mu(\theta_\ell, n_\ell) + \mathcal{O}(n_\ell^{-2}), \quad \mathbf{Var}[\ln\sin^2(\theta_{\ell+1})] = \sigma^2(\theta_\ell, n_\ell) + \mathcal{O}(n_\ell^{-2}), \tag{7}$$

$$\mu(\theta, n) = \ln\sin^2\theta - \frac{2}{3\pi}\theta - \rho(n) - \frac{8\theta}{15\pi n} - \left(\frac{2}{9\pi^2} - \frac{68}{45\pi^2 n}\right)\theta^2 + \mathcal{O}(\theta^3), \tag{8}$$

$$\sigma^2(\theta, n) = \frac{8}{n} - \frac{64}{15\pi}\frac{\theta}{n} - \left(8 + \frac{296}{45\pi}\right)\frac{\theta^2}{n} + \mathcal{O}\left(\theta^3\right), \tag{9}$$

where $\rho(n)$ is as defined in (2).

To derive Approximation 1 from Theorem 1, we simply keep only the first few terms of the series expansion (8), and then also completely drop the variability, essentially approximating $\sigma^2(\theta_\ell, n) \approx 0$ (Note that in reality $\sigma^2(\theta, n) \approx 8/n$ from (9)). Therefore Approximation 1 is a greatly simplified consequence of Theorem 1.

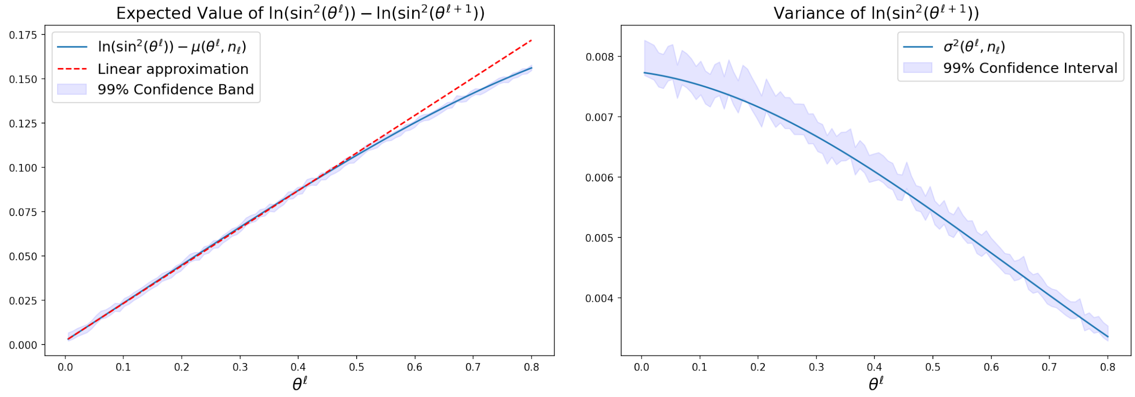
Moreover, our derivation shows that $\ln\sin^2(\theta_\ell)$ can be expressed in terms of sums over n pairs of independent Gaussian variables (see (14-16)). Thus, by central-limit-theorem type arguments, one would expect the following approximation by Gaussian laws which also accounts for the variability of $\ln\sin^2(\theta)$ using our calculated value for the variance.

Approximation 2 *Conditional on the value of θ_ℓ, the angle at layer $\ell + 1$ is well approximated by a Gaussian random variable*

$$\ln\sin^2(\theta_{\ell+1}) \overset{d}{\approx} \mathcal{N}(\mu(\theta_\ell, n_\ell), \sigma^2(\theta_\ell, n_\ell)), \tag{10}$$

where μ, σ^2 are as in Corollary 2. This approximation is understood in the sense that in the limit $n_\ell \to \infty$, we have

$$\frac{\ln\sin^2(\theta_{\ell+1}) - \mu(\theta_\ell, n_\ell)}{\sqrt{\sigma^2(\theta_\ell, n_\ell)}} \overset{d}{\Rightarrow} \mathcal{N}(0,1),$$



(a) Mean as a function of θ

(b) Variance as a function of θ

Figure 2: Plots comparing the functions $\mu(\theta, n)$ and $\sigma^2(\theta, n)$ to simulated neural networks. The linear approximation of μ, used to create Approximation 1 is also displayed. Confidence bands are constructed by randomly initializing 10,000 neural networks with layer width $n_\ell = 1024$, and a range of 100 initial angles $0.005 \leq \theta_\ell \leq 0.8$. We study $\theta_{\ell+1}$ and use the simulations to construct 99% confidence intervals for a) $\mathbf{E}\left[\ln(\sin^2(\theta_\ell)) - \ln(\sin^2(\theta_{\ell+1}))\right]$ and b) $\mathbf{Var}\left[\ln(\sin^2(\theta_{\ell+1}))\right]$.

We find that the normal approximation (10) matches simulated finite neural networks remarkably well; see Monte Carlo simulations from real networks in Figure 1. The big advantage of this approximation is that it very accurately captures the variance of $\ln(\sin^2(\theta_\ell))$, not just its mean. This variance grows as ℓ increases, so it is crucial for understanding behaviour of very deep networks.

The methods we use to obtain these approximations are quite flexible. For example, more accurate approximations can be obtained by incorporating higher moments $J_{a,b}(\theta)$ (see Section 2 for a discussion). We also believe that it should be possible to extend these methods to other non-linearities beyond ReLU and more complicated neural network architectures through the same basic principles we introduce here.

1.2 Practical Consequences: Depth Degeneracy Negatively Impacts Training

In this section, we empirically investigate how the theoretical prediction of large depth degeneracy phenomenon can lead to poor results *after training*. In other words, we show evidence that the depth degeneracy phenomenon (identified and studied only at *initialization*) can be used as a screening tool for neural architecture search to identify problematic neural architectures before they are trained. This could potentially add to the arsenal of existing tools used for neural architecture search (see e.g. Elsken et al. (2019))

We use the formula $\mu(\theta, n)$ developed in Theorem 1 to create a simple algorithm which accurately predicts the angle between inputs after travelling through the layers of an initialized network up to an error of size $\mathcal{O}(n_\ell^{-2})$ in layer ℓ.

Algorithm 1 predicts the angle at the final layer on initialization based solely on the network architecture $n_1, n_2, \ldots n_L$. Figure 3 demonstrates how networks which exhibit this type of degeneracy empirically tend to perform worse after training. When Algorithm 1

Algorithm 1 Angle prediction between inputs for a feed-forward ReLU network with depth L and layer widths n_ℓ, $1 \le \ell \le L$. The function $\mu(\theta, n)$ is given in Theorem 2.

```
$\theta^0$ = angle between inputs
for $\ell = 0, \dots, L-1$ do
    $x = \mu(\theta_\ell, n_\ell)$                    ▷ $x$ represents $\mathbf{E}[\ln(\sin^2(\theta_{\ell+1}))]$
    $\theta_{\ell+1} = \arcsin(e^{\frac{x}{2}})$
end for
Final angle = $\theta_L$
```

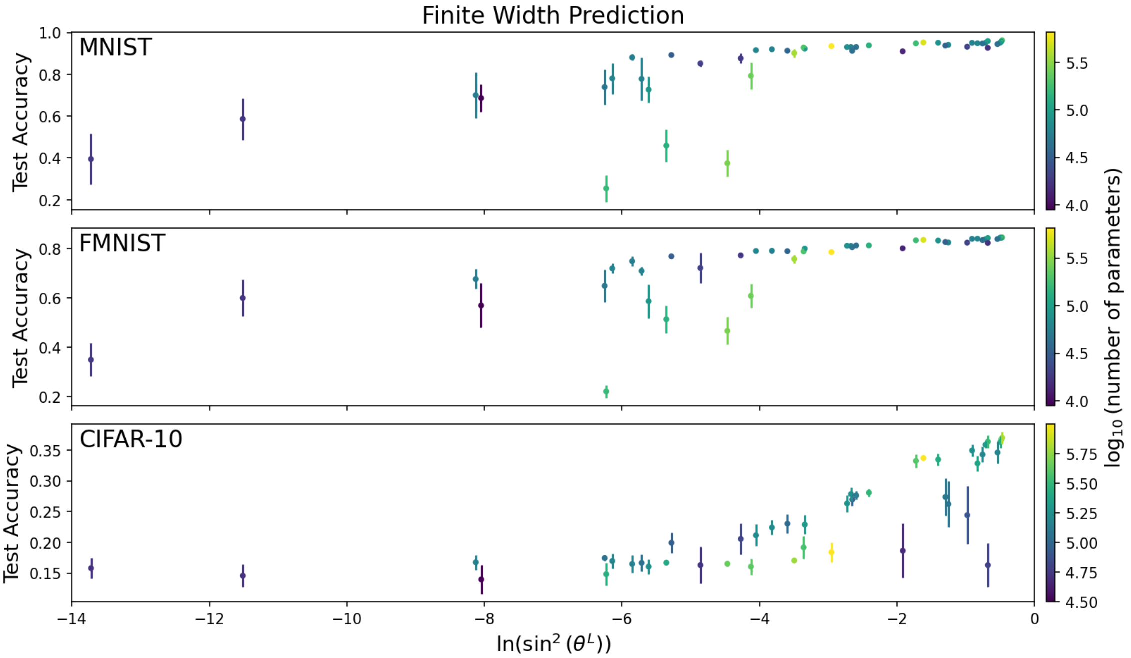


Figure 3: We compare 45 different network architectures trained on the MNIST Deng (2012), Fashion-MNIST Xiao et al. (2017), and CIFAR-10 Krizhevsky (2009) datasets 10 times each. Using the architecture of the network and Algorithm 1, we predict the angle between 2 orthogonal inputs at the final output layer of the network on initialization. We express the angle as $\ln(\sin^2(\theta_L))$, to follow the form used when developing the finite width approximations. The angle is plotted against the accuracy of each network on the test data after training, with error bars representing a 95% confidence interval across the 10 runs. We observe that small angle θ_L is related to lower test accuracy. All networks are trained using 1 epoch, batch size = 100, categorical cross-entropy loss, the ADAM optimizer, and default learning rate in the Keras module of TensorFlow Abadi et al. (2015). See Appendix D for details on all of the network architectures used. The code which produced this figure can be found at `GitHub link`.

predicts that θ is small at initialization, this serves a warning that the network may train poorly i.e. the test accuracy seems to be lower. Before going through the computationally expensive process of training many networks to assess their performance, this prediction could be used to quickly filter out network architectures that are unlikely to perform well due to excessive degeneracy.

1.2.1 Comparison to Infinite Width Update Rule

We compare the performance of Algorithm 1 (which takes into account the layer size n) to the infinite width update rule (given in Approximation 3). Since the infinite width prediction cannot account for the differences in layer widths, all networks with the same depth have the same prediction. In contrast, our method considers both the depth and width of each layer to predict how the angle propagates layer-by-layer through the network. Figure 4-Left illustrates how our method yields different angle predictions for different architectures with the same depth, while the infinite width method does not. Figure 4-Right shows the how the infinite width predictions differ from our "finite width" method which takes into account fluctuations of size $\mathcal{O}(n^{-1})$ in each layer.

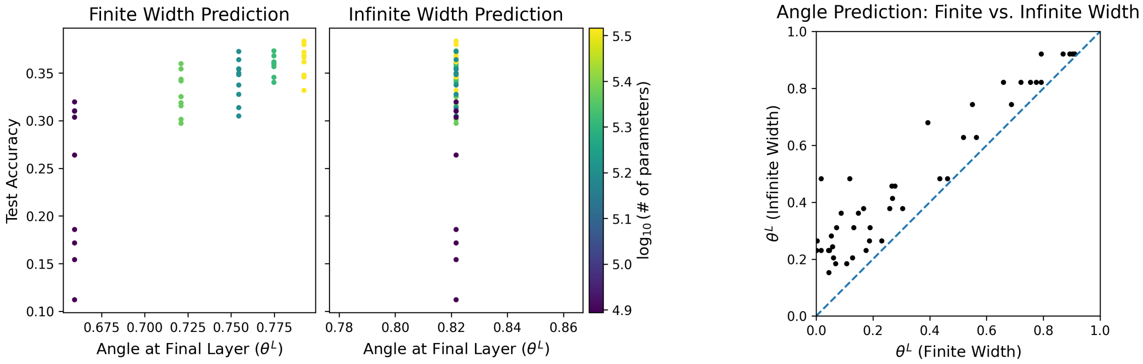


Figure 4: Left: Comparison of the finite and infinite width predictions for 5 network architectures with a depth of $L = 3$ trained 10 times each on the CIFAR-10 dataset Krizhevsky (2009). The infinite width predicts the same final angle for all networks, since it only depends on network depth. Right: Using the same 45 network architectures as in Figure 3, we plot a comparison of the predicted angle θ_L using Algorithm 1 (finite width) versus the infinite width prediction. We see that the infinite width prediction tends to underestimate the rate at which θ_ℓ tends towards 0.

1.3 J Functions and Infinite Width Limits

In 2009, Cho and Saul (2009) introduced the p-th moment for correlated ReLU-Gaussians, which they denoted with the letter J,

$$J_p(\theta) := 2\pi \mathbf{E}\left[\varphi^p(G)\varphi^p(\hat{G})\right], \tag{11}$$

where $p \in \mathbb{N}$, $\varphi(x) = \max\{x, 0\}$ is the ReLU function, and $G, \hat{G} \in \mathbb{R}$ are marginally two standard $\mathcal{N}(0,1)$ Gaussian random variables with correlation $\mathbf{Cov}(G, \hat{G}) = \cos(\theta)$. This quantity has found numerous applications for infinite width networks. One simple application of J_1 appears in the infinite width approximation for $\cos(\theta_\ell)$, where ℓ is fixed and we take the limit $n_1, n_2, \ldots n_\ell \to \infty$ (see Appendix A.9 for a detailed derivation):

Approximation 3 *The infinite-width approximation for the angle $\theta_{\ell+1}$ given θ_ℓ is*

$$\cos\left(\theta_{\ell+1}\right) = \frac{J_1(\theta_\ell)}{\pi} = \frac{\sin(\theta_\ell) + (\pi - \theta_\ell)\cos(\theta_\ell)}{\pi}. \tag{12}$$

The formula for J_1 is the $p = 1$ case of a remarkable explicit formula for J_p derived by Cho and Saul (2009) namely,

$$J_p(\theta) = (-1)^p (\sin\theta)^{2p+1} \left(\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\right)^p \left(\frac{\pi - \theta}{\sin\theta}\right).$$

This allows one to derive asymptotics of θ_ℓ in the infinite width limit, as in Section 1.1.1. However, there are several limitations to this approach. Most important is that the infinite width limit is not a good approximation when the network depth ℓ is comparable to the network width n (Li et al. (2022)). The infinite width limit uses the law of large numbers to obtain (12), thereby discarding random fluctuations. For very deep networks, microscopic fluctuations (on the order of $\mathcal{O}(1/n_\ell)$) from layer to layer can accumulate over ℓ layers to give macroscopic effects. This is why the infinite width predictions for θ_ℓ are not a good match to the simulations in Figure 1; very deep networks are far from the infinite width limit in this case. See Figure 1 where the infinite width predictions are compared to finite networks.

Instead, to analyze the evolution of the angle θ_ℓ more accurately, we need to do something more precise than the law of large numbers to capture the effect of these microscopic fluctuations. This is the approach we carry out in this paper. While the mean only depends on the p-th moment functions J_p from (11), these fluctuations depend on the *mixed* moments, which we denote by $J_{a,b}$ for $a, b \in \mathbb{N}$ as follows[1]

$$J_{a,b}(\theta) := \mathbf{E}\left[\varphi^a(G)\varphi^b(\hat{G})\right], \tag{13}$$

with $G, \hat{G}$ again as in (11) are marginally $\mathcal{N}(0,1)$ with correlation $\cos(\theta)$. In Section 2.1 we carry out a detailed asymptotic analysis to write the evolution of θ_ℓ in terms of the mixed moments $J_{a,b}$. In order to make useful predictions, one must also calculate a formula for $J_{a,b}(\theta)$. Unfortunately, the method that Cho-Saul originally proposed for this does *not* seem to work when $a \neq b$. This is because that method used contour integrals, and relied on using certain trig identities which do not hold when $a \neq b$. Instead, in Section 3, we introduce a new method, based on Gaussian integration by parts, to compute $J_{a,b}$ for general a, b via a recurrence relation. By serendipity[2], we find a remarkable combinatorial connection between $J_{a,b}$ and the Bessel numbers (Cheon et al. (2013)), which allows one to find an explicit (albeit complicated) formula for $J_{a,b}$ in terms of binomial coefficients. The formula for the first few functions are shown in Table 1, and the general explicit formula is presented in Theorem 5.

[1] Note that compared to Cho and Saul's definition for J_p, we omit the factor of 2π in our definition of $J_{a,b}$. The factor of 2π seems natural when $a+b$ is even (like the case $a = b = p$ that Cho-Saul considered), but when $a+b$ is odd a different factor of $2\sqrt{2\pi}$ appears! Therefore the factor of 2π would confuse things in the general case (see Table 1). The correct translation between Cho-Saul J_p and our $J_{a,b}$ is $J_p = 2\pi J_{p,p}$.

[2] This connection was first noticed by calculating the first few J functions, and then using the On-Line Encyclopedia of Integer Sequence to discover the connection to Bessel number (`https://oeis.org/A001498`).

$a \backslash b$	0	1	2	3
0	$\frac{\pi-\theta}{2\pi}$	$\frac{\cos\theta+1}{2\sqrt{2\pi}}$	$\frac{(\pi-\theta)+\sin\theta\cos\theta}{2\pi}$	$\frac{2(\cos\theta+1)+\sin^2\theta\cos\theta}{2\sqrt{2\pi}}$
1		$\frac{\sin\theta+(\pi-\theta)\cos\theta}{2\pi}$	$\frac{(\cos\theta+1)^2}{2\sqrt{2\pi}}$	$\frac{3(\pi-\theta)\cos\theta+\sin\theta\cos^2\theta+2\sin\theta}{2\pi}$
2			$\frac{(\pi-\theta)(2\cos^2\theta+1)+3\sin\theta\cos\theta}{2\pi}$	$\frac{3\cos\theta(\cos\theta+1)^2+2(\cos\theta+1)+\sin^2\theta\cos\theta}{2\sqrt{2\pi}}$
3				$\frac{(\pi-\theta)(6\cos^2\theta+9)\cos\theta}{2\pi}$ $\frac{+5\sin\theta\cos^2\theta+(6\cos^2\theta+4)\sin\theta}{2\pi}$

Table 1: Table of formulas for the first few J functions. Note that all entries have a denominator either $c_0 = 2\pi$ or $c_1 = 2\sqrt{2\pi}$ depending on the parity of $a+b$. These generalize $J_p(\theta)$ of (11) which appear on the diagonal of this table. Note that $J_{a,b} = J_{b,a}$ so only upper triangular entries are shown. An explicit formula for all $J_{a,b}$ is derived in Section 3.4.

1.4 Outline

The two main contributions of this paper are to prove Theorem 1 for the evolution of θ_ℓ in terms of the mixed moments $J_{a,b}$, and then separately derive an explicit formula, Theorem 5, for any of the mixed moments $J_{a,b}$. Combined, these allow for the explicit formula Corollary 2 and the useful simpler Approximations 1 and 2. See Figure 1 and Figure 2 for comparisons of these predictions to Monte-Carlo simulations. We also believe the methods proposed here are flexible enough to be modified to apply to non-linearities other than ReLU and to different neural network architectures beyond fully-connected networks in future work.

Section 2 contains the analysis of the angle process and predicted distribution of $\ln(\sin^2(\theta_\ell))$ in deep ReLU networks. Section 2.1 covers our approximation of $\mathbf{E}[\ln(\sin^2(\theta_{\ell+1}))]$ which leads to the rule for θ_ℓ as in equation (1), while Section 2.2 outlines our approximation for $\mathbf{Var}[\ln(\sin^2(\theta_{\ell+1}))]$.

In Section 3, we cover the derivation of the explicit formula for the J functions. We state the main results of this section in Section 3.1, and cover the mathematical tools needed to solve the expectations using Gaussian integration by parts in Section 3.2. We develop the formula for $J_{a,b}$ by first finding a recursive formula in Section 3.3, which reveals a connection between the J functions and the Bessel numbers. This recursion is studied to develop an explicit formula for $J_{a,b}$ in Section 3.4.

2. ReLU Neural Networks on Initialization

In this section, we analyze ReLU neural networks and show how the the mixed moments $J_{a,b}$ appear in evolution of the angle θ_ℓ on initialization. We define the notation we use for a fully connected ReLU neural network, along with other notations we will use in Table 2. Note that the factor of $\sqrt{2/n_\ell}$ in our definition is implementing the so called He initialization (He et al., 2015), which ensures that $\mathbf{E}[\|z^\ell\|^2] = \|x\|^2$ for all layers ℓ. This initialization

Symbol	Definition
$x \in \mathbb{R}^{n_{in}}$	Input (e.g. training example) in the input dimension $n_{in} \in \mathbb{N}$
$\ell \in \mathbb{N}$	Layer number. $\ell = 0$ is the input
$n_\ell \in \mathbb{N}$	Width of hidden layer ℓ (i.e. number of neurons in layer ℓ)
$W^\ell \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$	Weight matrix for layer ℓ. Initialized with iid standard Gaussian entries $W^\ell_{a,b} \sim \mathcal{N}(0,1)$
$\varphi : \mathbb{R}^n \to \mathbb{R}^n$	Entrywise ReLU activation function $\varphi(x)_i = \varphi(x_i) = \max\{x_i, 0\}$
$z^\ell(x) \in \mathbb{R}^{n_\ell}$	Pre-activation vector in the ℓ^{th} layer for input x (a.k.a logits of layer ℓ) $z^1(x) := W^1 x, \qquad z^{\ell+1}(x) := \sqrt{\frac{2}{n_\ell}} W^{\ell+1} \varphi(z^\ell(x)).$
$\varphi^\ell_\alpha, \varphi^\ell_\beta \in \mathbb{R}^{n_\ell}$	Post-activation vector on inputs x_α, x_β respectively $\varphi^\ell_\alpha := \varphi(z^\ell(x_\alpha)), \qquad \varphi^\ell_\beta := \varphi(z^\ell(x_\beta))$
$\theta_\ell \in [0, \pi]$	Angle between φ^ℓ_α and φ^ℓ_β defined by $\cos(\theta_\ell) := \frac{\langle \varphi^\ell_\alpha, \varphi^\ell_\beta \rangle}{\Vert\varphi^\ell_\alpha\Vert \Vert\varphi^\ell_\beta\Vert}$
$R_{\ell+1} \in \mathbb{R}$	Shorthand for the ratio $R_{\ell+1} := \frac{\Vert\varphi^{\ell+1}_\alpha\Vert^2 \Vert\varphi^{\ell+1}_\beta\Vert^2}{\Vert\varphi^\ell_\alpha\Vert^2 \Vert\varphi^\ell_\beta\Vert^2}$

Table 2: Definition and notation used for fully connected ReLU neural networks.

is known to be the "critical" initialization for taking large limits of the network (Roberts et al., 2022; Hayou et al., 2019). Given this neural network, we wish to study the evolution of 2 inputs x_α and x_β as they traverse through the layers of the network. Specifically, we wish to study how the angle θ between the inputs changes as the inputs move from layer to layer.

The starting point for our calculation is to notice that because the weights are Gaussian, the values of $\varphi^{\ell+1}_\alpha, \varphi^{\ell+1}_\beta$ are jointly Gaussian given the vectors of $\varphi^\ell_\alpha, \varphi^\ell_\beta$. In fact, it turns out that by properties of Gaussian random variables, one only needs to know the values of the scalars $\Vert\varphi^\ell_\alpha\Vert, \Vert\varphi^\ell_\beta\Vert$ and θ_ℓ to understand the full distribution of $\varphi^{\ell+1}_\alpha, \varphi^{\ell+1}_\beta$. (see Appendix A.5 for details) By using the positive homogeneity of the ReLU function $\varphi(\lambda x) = \lambda \varphi(x)$ for $\lambda > 0$, we can factor out the effect of the norm of each vector in layer ℓ. After some manipulations, these ideas lead us to the following identities that are the heart of our calculations; a full derivation of these quantities are provided in Appendix A.5 and A.6.

$$\|\varphi_\alpha^{\ell+1}\|^2 = \frac{\|\varphi_\alpha^\ell\|^2}{n_\ell}\sum_{i=1}^{n_\ell} 2\varphi^2(G_i), \qquad \|\varphi_\beta^{\ell+1}\|^2 = \frac{\|\varphi_\beta^\ell\|^2}{n_\ell}\sum_{i=1}^{n_\ell} 2\varphi^2(\hat{G}_i), \tag{14}$$

$$\langle \varphi_\alpha^{\ell+1}, \varphi_\beta^{\ell+1}\rangle = \frac{\|\varphi_\alpha^\ell\|\|\varphi_\beta^\ell\|}{n_\ell}\sum_{i=1}^{n_\ell} 2\varphi(G_i)\varphi(\hat{G}_i), \tag{15}$$

$$\frac{\|\varphi_\alpha^{\ell+1}\|^2\|\varphi_\beta^{\ell+1}\|^2}{\|\varphi_\alpha^\ell\|^2\|\varphi_\beta^\ell\|^2}\sin^2(\theta_{\ell+1}) = \frac{2}{n_\ell^2}\sum_{i,j=1}^{n_\ell}\Big(\varphi(G_i)\varphi(\hat{G}_j) - \varphi(G_j)\varphi(\hat{G}_i)\Big)^2, \tag{16}$$

where G_i, $\hat{G}_i$ are all marginally $\mathcal{N}(0,1)$, with correlation $\mathbf{Cov}(G_i, \hat{G}_i) = \cos(\theta_\ell)$ and independent for different indices i. The identity in (16) is derived using the *determinant of the Gram matrix* for vectors $\varphi_\alpha^{\ell+1}$, $\varphi_\beta^{\ell+1}$ (full derivation given in Appendix A.6). Combining the equations in (14) gives us a useful identity for the ratio $R_{\ell+1}$, namely:

$$R_{\ell+1} = \frac{4}{n_\ell^2}\sum_{i,j=1}^{n_\ell}\varphi^2(G_i)\varphi^2(\hat{G}_j). \tag{17}$$

Given some θ_ℓ, we wish to predict the behaviour of $\theta_{\ell+1}$. Rather than studying $\theta_{\ell+1}$ directly, we instead study the quantity $\ln(\sin^2(\theta_{\ell+1}))$. This allows us to use convenient approximations and identities for quantities we are interested in. And indeed, a post-hoc analysis shows that as $\theta \to 0$, the random variable $\ln\sin^2(\theta_{\ell+1})$ has a *non-zero constant* variance which depends only on n_ℓ. This is in contrast to θ_ℓ itself which has variance tending to *zero.* This is one reason why the Gaussian approximation for $\ln\sin^2(\theta_\ell) \in (-\infty, 0] \subset (-\infty, \infty)$ works well, whereas Gaussian approximations for θ_ℓ or $\cos(\theta_\ell) \in [-1, 1]$ are less accurate. This observation can equivalently be understood as the observation that the random fluctuations seem to be multiplicative, rather than additive, which is why taking the log makes them more amenable to calculation. We first derive a formula for $\mathbf{E}\left[\ln(\sin^2(\theta_{\ell+1}))\right]$.

2.1 Expected Value

In this section, we show how to compute the expected value of $\ln(\sin^2(\theta^\ell))$ in terms of the J functions as in Theorem 1. Firstly, we rewrite this expectation as the difference

$$\mathbf{E}\left[\ln(\sin^2(\theta_{\ell+1}))\right] = \mathbf{E}\left[\ln\left(R_{\ell+1}\sin^2(\theta_{\ell+1})\right)\right] - \mathbf{E}\left[\ln\left(R_{\ell+1}\right)\right]. \tag{18}$$

The two random variables $R_{\ell+1}$ and $R_{\ell+1}\sin^2(\theta_{\ell+1})$ in (18) both have interpretations in terms of sums of Gaussians as in (16) and (17) which makes it possible to calculate their moments in terms of the J functions. To enable our use of the moments here, we use the following approximation of $\ln(X)$ for a random variable X, which is based on the Taylor expansion for $\ln(1+x) = x - \frac{1}{2}x^2 + \ldots$ (a full derivation is given in Appendix A.1):

$$\ln(X) = \ln(\mathbf{E}[X]) + \frac{X - \mathbf{E}[X]}{\mathbf{E}[X]} - \frac{(X - \mathbf{E}[X])^2}{2\mathbf{E}[X]^2} + \epsilon_2\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right), \tag{19}$$

where $\epsilon_2(x)$ is the Taylor remainder term in $\ln(1+x) = x - \frac{x^2}{2} + \epsilon_2(x)$ and satisfies $\epsilon_2(x) = \mathcal{O}(x^3)$. Applying this approximation to the terms appearing on the right hand side of (18), and taking expected value of both sides, we obtain the estimates

$$\mathbf{E}\left[\ln\left(R_{\ell+1}\sin^2(\theta_{\ell+1})\right)\right] = \ln\left(\mathbf{E}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right]\right) - \frac{\mathbf{Var}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right]}{2\mathbf{E}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right]^2} + \mathcal{O}(n_\ell^{-2}),$$

$$\mathbf{E}\left[\ln\left(R_{\ell+1}\right)\right] = \ln\left(\mathbf{E}\left[R_{\ell+1}\right]\right) - \frac{\mathbf{Var}\left[R_{\ell+1}\right]}{2\mathbf{E}\left[R_{\ell+1}\right]^2} + \mathcal{O}(n_\ell^{-2}).$$

To control the error here, we have used here the fact that $R_{\ell+1}$ and $R_{\ell+1}\sin^2(\theta_{\ell+1})$ can be written as averages over random variables as in (14 - 16). This allows us to show the 3rd central moments for $R_{\ell+1}$ and $R_{\ell+1}\sin^2(\theta_{\ell+1})$ are $\mathcal{O}(n_\ell^{-2})$; see Appendix A.1 for details. This approximation is convenient because we are able to calculate the values on the right hand side of the equations in terms of the moments $J_{a,b}$ by expanding/taking expectations of the representations (14 - 16). The key quantities we calculate are

$$\mathbf{E}\left[R_{\ell+1}\right] = \frac{4J_{2,2}-1}{n_\ell} + 1, \tag{20}$$

$$\mathbf{Var}\left[R_{\ell+1}\right] = \frac{4}{n_\ell}(J_{2,2}+1) + \frac{16}{n_\ell^2}\left(2J_{4,2} - \frac{5}{2}J_{2,2} + J_{2,2}^2 + \frac{5}{8}\right) + \mathcal{O}\left(n_\ell^{-3}\right), \tag{21}$$

$$\mathbf{E}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right] = \frac{(n_\ell - 1)(1-4J_{1,1}^2)}{n_\ell}, \tag{22}$$

$$\mathbf{Var}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right] = \frac{8\left(-8J_{1,1}^4 + 8J_{1,1}^2J_{2,2} + 4J_{1,1}^2 - 8J_{1,1}J_{3,1} + J_{2,2} + 1\right)}{n_\ell} + \mathcal{O}\left(n_\ell^{-2}\right), \tag{23}$$

where $J_{a,b} = J_{a,b}(\theta_\ell)$. These formulas are calculated in Appendix A.7 and Appendix A.8 by a combinatorial expansion using the representations from (14-16). Combining these gives the result for $\mu(\theta, n)$ in Theorem 1. Note that to obtain a more accurate approximation, we would simply include more terms in the variance expressions in (21, 23).

2.2 Variance of $\ln(\sin^2(\theta_{\ell+1}))$

In this section, we show how to compute the variance of $\ln(\sin^2(\theta^\ell))$ in terms of the J functions as in Theorem 1. We can rewrite $\mathbf{Var}[\ln(\sin^2(\theta_{\ell+1}))]$ in the following way:

$$\mathbf{Var}[\ln(\sin^2(\theta_{\ell+1}))] = \mathbf{Var}\left[\ln\left(R_{\ell+1}\sin^2(\theta_{\ell+1})\right) - \ln\left(R_{\ell+1}\right)\right] \tag{24}$$

$$= \mathbf{Var}\left[\ln\left(R_{\ell+1}\sin^2(\theta_{\ell+1})\right)\right] + \mathbf{Var}\left[\ln\left(R_{\ell+1}\right)\right] - 2\mathbf{Cov}\left(\ln\left(R_{\ell+1}\sin^2(\theta_{\ell+1})\right), \ln\left(R_{\ell+1}\right)\right).$$

We have now expressed this in terms of $R^{\ell+1}$ and $R_{\ell+1}\sin^2(\theta_{\ell+1})$ which will allow us to use identities as in (14 - 16) in our calculations. Appendix A.2 and Appendix A.3 cover the method used to approximate the unknown variance and covariance terms above. Once again, we control the error term arising from moments in the error term of the Taylor series by using representation as sums (14 - 16). We have already calculated most of the quantities on the right hand side already in our calculation for $\mu(\theta, n)$. The only new term is

$$\mathbf{Cov}\left(R_{\ell+1}\sin^2(\theta_{\ell+1}), R_{\ell+1}\right) = \frac{1}{n_\ell}\left(16J_{1,1}^2 - 32J_{1,1}J_{3,1} + 8J_{2,2} + 8\right) + \mathcal{O}\left(n_\ell^{-2}\right).$$

This is again computed by a combinatorial expansion of the sums (14-16). (Full calculation given in Appendix A.8). We now have solved for all of the functions needed to perform our approximation of $\mathbf{Var}[\ln(\sin^2(\theta_{\ell+1}))]$. Putting it together, we end up with the expression for $\sigma^2(\theta, n)$ as in (5). We compare the predicted probability distribution of $\ln(\sin^2(\theta))$ using our formulas $\mu(\theta, n)$ and $\sigma^2(\theta, n)$ to empirical probability distributions in Figure 1.

3. Explicit Formula for the Mixed-Moment J Functions

In this section we develop a combinatorial method that allows us to compute exact formulas for the J functions. The method is to use Gaussian integration by parts to find a recurrence relationship between the moments $J_{a,b}$, and then solve it explicitly. We begin by generalizing the definition of $J_{a,b}$ from (13) to include $a = 0$ and/or $b = 0$ as follows. Let G, W be *independent* $\mathcal{N}(0,1)$ variables. Then, we define the functions $J_{a,b}(\theta)$ as

$$J_{a,b}(\theta) = \mathbf{E}[G^a (G\cos\theta + W\sin\theta)^b \, 1\{G > 0\} \, 1\{G\cos\theta + W\sin\theta > 0\}], \tag{25}$$

where $a, b \in \mathbb{N} \cup \{0\}$. Note that $G\cos\theta + W\sin\theta = \hat{G}$ is marginally $\mathcal{N}(0,1)$ and has correlation $\cos(\theta)$ with G, matching the original definition. The ReLU function satisfies the identity $\varphi(x)^a = x^a 1\{x > 0\}$ for $a \geq 1$, so (25) generalizes (13) to the case $a = 0$. We also note that $J_{a,b}(\theta) = J_{b,a}(\theta)$ for all $a, b \in \mathbb{N} \cup \{0\}$.

Remark 3 *Note that* (25) *can equivalently be written in terms of the correlation* $\rho = \cos\theta$ *and* $\sqrt{1-\rho^2} = \sin\theta$ *as follows.*

$$G^a(\rho G + \sqrt{1-\rho^2}W)^b 1\{G > 0\} 1\{\rho G + \sqrt{1-\rho^2}W\}$$

Some authors prefer to work with the correlation ρ rather than the angle θ. In this work we choose to work with the angle θ since our technique works well to directly see how quickly $\theta \to 0$.

3.1 Statement of Main Results and Outline of Method

By using the method of Gaussian integration by parts, we are able to derive recurrence relations for the $J_{a,b}$ functions. Since the definition of $J_{a,b}$ involves the indicator function $1\{G > 0\}$, we must make sense of what the derivative of this function means for the purposes of integration by parts; see Section 3.2 where this is carried out. Then, by use of the generalized Gaussian integration by parts formula, we obtain the following recurrence relations for $J_{a,b}$.

Proposition 1 (Recurrence relations for** $J_{a,b}$**) *For $a \geq 2$, the sequence $J_{a,0}$ satisfies the recurrence relation:*

$$J_{a,0}(\theta) = (a-1)J_{a-2,0}(\theta) + \frac{\sin^{a-1}\theta\cos\theta}{c_{a \bmod 2}}(a-2)!\,!\,, \tag{26}$$

where $c_0 = 2\pi$, $c_1 = 2\sqrt{2\pi}$. For $a \geq 2$, and $b \geq 1$, the collection $J_{a,b}$ satisfies the following two-index recurrence relation:

$$J_{a,b}(\theta) = (a-1)J_{a-2,b}(\theta) + b\cos\theta J_{a-1,b-1}(\theta). \tag{27}$$

The same integration by parts technique that yields the recurrence relation also makes it easy to evaluate the first few J functions. They are as follows:

Proposition 2 (Explicit Formula for $J_{0,0}, J_{1,0}, J_{1,1}$***)*** *$J_{0,0}$, $J_{1,0}$, and $J_{1,1}$ are given by*

$$J_{0,0}(\theta) = \frac{\pi - \theta}{2\pi}, \qquad J_{1,0}(\theta) = \frac{1 + \cos\theta}{2\sqrt{2\pi}}, \qquad J_{1,1}(\theta) = \frac{\sin\theta + (\pi - \theta)\cos\theta}{2\pi}. \tag{28}$$

See Appendix B.1 for a derivation of these quantities. Note that Cho and Saul (2009) have previously discovered the formulas for $J_{0,0}$ and $J_{1,1}$ by use of a completely different contour-integral based method.

The combination of Propositions 1 and 2 make it possible to practically calculate any value of $J_{a,b}$ when a, b are not too large. However, by serendipity, we are able to find remarkable explicit formulas for $J_{a,b}$, which we report below.

Proposition 3 (Explicit Formulas for $J_{a,0}(\theta)$***,*** $J_{a,1}(\theta)$***)*** *Let $a \geq 2$. Then, $J_{a,0}$ and $J_{a,1}$ are explicitly given by the following:*

$$J_{a,0}(\theta) = (a-1)!!\left(J_{a \bmod 2, 0} + \frac{\cos\theta}{c_{a \bmod 2}} \sum_{\substack{i \not\equiv a (\bmod 2) \\ 0 < i < a}} \frac{(i-1)!!}{i!!} \sin^i \theta \right),$$

where $c_0 = 2\pi$, $c_1 = 2\sqrt{2\pi}$. We can then use the explicit formula for $J_{a,0}$ in the formula for $J_{a,1}$:

$$J_{a,1}(\theta) = (a-1)!!\left(J_{a \bmod 2, 1} + \cos\theta \sum_{\substack{i \not\equiv a (\bmod 2) \\ 0 < i < a}} \frac{J_{i,0}(\theta)}{i!!} \right),$$

where an explicit formula for the first term (either $J_{1,0}$ or $J_{1,1}$ depending on the parity of a) is given in Proposition 2.

We can finally express $J_{a,b}$ as a linear combination of $J_{0,n}$ and $J_{1,n}$, as follows. (In light of the previous explicit formulas, this is an explicit formula for $J_{a,b}$.) It turns out that the coefficients are given in terms of two special numbers $P(a,b)$ and $Q(a,b)$ which are known as the Bessel numbers.

Definition 4 (Bessel numbers) *The numbers $P(a,b)$ and $Q(a,b)$ are defined as follows,*

$$P(a,b) = \begin{cases} \frac{a!}{b!\left(\frac{a-b}{2}\right)! 2^{\frac{a-b}{2}}}, & a \geq b,\ a \equiv b \pmod 2 \\ 0, & \textit{otherwise} \end{cases}, \tag{29}$$

$$Q(a,b) = \begin{cases} \frac{\left(\frac{a+b}{2}\right)!}{b!} 2^{\frac{b-a}{2}} \sum_{i=0}^{\frac{a-b}{2}} \binom{a+1}{i}, & a \geq b,\ a \equiv b \pmod 2 \\ 0, & \textit{otherwise} \end{cases}. \tag{30}$$

$P(a,b)$ represents a family of numbers known as the Bessel numbers of the second kind (Cheon et al. (2013)), and $Q(a,b)$ comes from a closely related family of numbers (Kreinin (2016)). Using these, we can express $J_{a,b}$ as follows.

Theorem 5 (Explicit Formula for $J_{a,b}(\theta)$) *Let $b \geq 2, a \geq 1,\ b \geq a$. Then, we have the following formula for $J_{a,b}(\theta)$ in terms of $J_{0,n}$ and $J_{1,n}$*

$$\begin{aligned} J_{a,b} = & \sum_{\substack{i \equiv 0 (\bmod 2) \\ 0 < i \leq a}} (b)_{a-i} (\cos\theta)^{a-i} \left(P(a, a-i) - Q(a-1, a-1-i)\right) J_{0,b-a+i} \\ & + \sum_{\substack{i \equiv 1 (\bmod 2) \\ 0 < i \leq a}} (b)_{a-i} (\cos\theta)^{a-i} Q(a-1, a-i) J_{1,b-a+i}. \end{aligned}$$

Remark 6 *Since $J_{1,n}$ is also given in terms of $J_{0,n}$, one may further simplify the formula for $J_{a,b}$ to be in terms of only $J_{0,n}$, $J_{0,0}$ and $J_{0,1}$. This substitution yields the following formula. For notational convenience, we will let $\delta := b - a$,*

$$\begin{aligned} J_{a,b} = & \sum_{\substack{i \equiv 0 (\bmod 2) \\ 0 < i \leq a}} (b)_{a-i} (\cos\theta)^{a-i} (P(a, a-i) - Q(a-1, a-1-i)) J_{0,\delta+i} \\ & + \sum_{\substack{i \equiv 1 (\bmod 2) \\ 0 < i \leq a}} (b)_{a-i} (\cos\theta)^{a-i} Q(a-1, a-i) (\delta + i - 1)!!\, J_{(\delta+1) \bmod 2, 1} \\ & + \cos\theta \sum_{\substack{i \equiv 1 (\bmod 2) \\ 0 < i \leq a}} \sum_{\substack{j \equiv \delta (\bmod 2) \\ 0 < j < \delta + i}} (b)_{a-i} (\cos\theta)^{a-i} Q(a-1, a-i) \frac{(\delta + i - 1)!!}{j!!} J_{0,j}. \end{aligned}$$

3.2 Gaussian Integration-by-Parts Formulas

In this section, we state two important formulas that together give us the tools for computing the expectations that appear in $J_{a,b}$, based on the well known Gaussian integration-by-parts trick; see for example Chapter 7.2 of the textbook (Vershynin, 2018).

Fact 1 (Gaussian Integration by Parts) *Let $G \sim \mathcal{N}(0,1)$ be a Gaussian variable and $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function. Then,*

$$\mathbf{E}[Gf(G)] = \mathbf{E}[f'(G)]. \tag{31}$$

Using this type of Gaussian integration by parts formula, we can generalize the expected value of Gaussians to derivatives of functions which are not necessarily differentiable. For example the indicator function $1\{x > a\}$ is not differentiable, but for the purposes of computing Gaussian expectation, we can use the following integration formula.

Remark 7 *An alternative way to view this kind of Gaussian expected value calculation is Wick's formula / Isserlis theorem. The Gaussian integration by parts trick can be thought of as the extension of those formulas from polynomials to arbitrary functions.*

Fact 2 (Gaussian expectations involving** $1'\{x > a\}$**) *Let G be a Gaussian variable and $a \in \mathbb{R}$. Let $f : \mathbb{R} \to \mathbb{R}$ such that $\lim_{g\to\infty} f(g)e^{\frac{-g^2}{2}} = 0$. Then, using the Gaussian integration by parts formula to assign a meaning to expectations involving the "derivative of the indicator function", $1'\{x > a\}$, we have*

$$\mathbf{E}[1'\{G > a\}f(G)] = f(a)\frac{e^{\frac{-a^2}{2}}}{\sqrt{2\pi}}. \tag{32}$$

Remark 8 *The purpose of assigning a value to the expectation (32) is to allow one to compute "honest" expectations of the form (31) when $f(x)$ involves $1\{x > 0\}$; see Lemma 9 for an illustrative example. The final result does* not *require interpreting "$1'\{x > a\}$"; this is only a useful intermediate step in the sequence of calculations leading to the final result.*

The formula can also be understood or proven in a number of different alternative ways. One is simply to say that $1'\{x > a\} = \delta\{x = a\}$ is a "Dirac delta function" at $x = a$. A more rigorous way would be to take any differentiable family of functions $1_\epsilon\{x > a\}$ which suitably converge to $1\{x > a\}$ as $\epsilon \to 0$ and then interpret the result as the limit of the expectation $\lim_{\epsilon\to 0} \mathbf{E}[1'_\epsilon\{G > a\}f(G)]$. Here is the argument that is used to obtain Fact 2: Applying integration by parts, we formally have

$$\begin{aligned}\mathbf{E}[1'\{G > a\}f(G)] &= \int\limits_{-\infty}^{\infty} 1'\{g > a\}f(g)\frac{e^{\frac{-g^2}{2}}}{\sqrt{2\pi}}dg \\ &= \left[1\{g > a\}f(g)\frac{e^{\frac{-g^2}{2}}}{\sqrt{2\pi}}\right]_{-\infty}^{\infty} - \int\limits_{-\infty}^{\infty} 1\{g > a\}\frac{d}{dg}\left(f(g)\frac{e^{\frac{-g^2}{2}}}{\sqrt{2\pi}}\right)dg.\end{aligned}$$

Note that the first term is 0 by the hypothesis $\lim\limits_{g\to\infty} f(g)e^{\frac{-g^2}{2}} = 0$, and we have then

$$\mathbf{E}[1'\{G > a\}f(G)] = -\int\limits_{a}^{\infty} \frac{d}{dg}\left(f(g)\frac{e^{\frac{-g^2}{2}}}{\sqrt{2\pi}}\right)dg = -\left[f(g)\frac{e^{\frac{-g^2}{2}}}{\sqrt{2\pi}}\right]_{a}^{\infty} = 0 + f(a)\frac{e^{\frac{-a^2}{2}}}{\sqrt{2\pi}},$$

where we have used the hypothesis on f once again.

The two facts about Gaussian integration by parts can be combined to create recurrence relations for expectations involving $1\{G > a\}$. A simple example is the following lemma, which we will also use later in our derivation. The proof strategy of this lemma is a microcosm of the proof strategy we use to compute $J_{a,b}$ in general, namely to use Gaussian integration by parts to derive a recurrence relation and initial condition, and then solve.

Lemma 9 (Moments of** $\varphi(G)$**) *For $k \geq 0$, we have*

$$\mathbf{E}[\varphi(G)^k] = \mathbf{E}[G^k 1\{G > 0\}] = \begin{cases} \frac{(k-1)!!}{2} & k \text{ is even} \\ \frac{(k-1)!!}{\sqrt{2\pi}} & k \text{ is odd} \end{cases} = \sqrt{2\pi}\frac{(k-1)!!}{c_{k-1 \bmod 2}},$$

where $c_0 = 2\pi$ and $c_1 = 2\sqrt{2\pi}$.

Proof *We prove this for even and odd k separately by induction on k. The base case for $k = 0$ is trivial since $(0-1)!! = 1$ is the empty product. The base case $k = 1$ follows by first applying* (31) *with $f(x) = 1\{x > 0\}$ and then applying* (32) *with $f(x) \equiv 1$,*

$$\mathbf{E}[\varphi(G)] = \mathbf{E}[G1\{G > 0\}] = \mathbf{E}[1'\{G > 0\}] = \sqrt{2\pi}^{-1}.$$

Now, to see the induction, we apply (31) *with $f(x) = x^{k-1}1\{x > 0\}$, $k \geq 2$. Due to the product rule, there are two terms in the derivative,*

$$\begin{aligned} \mathbf{E}[\varphi(G)^k] &= \mathbf{E}[G \cdot G^{k-1}1\{G > 0\}] \\ &= (k-1)\mathbf{E}[G^{k-2}1\{G > 0\}] + \mathbf{E}[G^{k-1}1'\{G > 0\}] \\ &= (k-1)\mathbf{E}[\varphi(G)^{k-2}] + 0, \end{aligned} \tag{33}$$

where we have recognized that the second term is 0 *by application of* (32) *with $f(x) = x^{k-1}$ which has $f(0) = 0$. The recurrence $\mathbf{E}[\varphi(G)^k] = (k-2)\mathbf{E}[\varphi(G)^{k-2}]$ along with initial condition leads to the stated result by induction.* ■

3.3 Recursive Formulas for $J_{a,b}(\theta)$ - Proof of Proposition 1

Proof *[Of Proposition 1] To find a recursive formula for $J_{a,0}, a \geq 2$, we apply the Gaussian integration by parts formula* (31) *to $f(x) = x^{a-1}1\{x > 0\}1\{\cos\theta x + W \sin\theta > 0\}$ to evaluate the expected value over G first. When applying product rule there are three terms*

$$\begin{aligned} J_{a,0} =&\mathbf{E}[G \cdot G^{a-1}1\{G > 0\}1\{G\cos\theta + W\sin\theta > 0\}] \\ =&(a-1)\mathbf{E}[G^{a-2}1\{G > 0\}1\{G\cos\theta + W\sin\theta > 0\}] \\ &+ \mathbf{E}[G^{a-1}1\{G > 0\}1'\{G\cos\theta + W\sin\theta > 0\}]\cos\theta \\ &+ \mathbf{E}[G^{a-1}1'\{G > 0\}1\{G\cos\theta + W\sin\theta > 0\}]. \end{aligned} \tag{34}$$

The first term is simply $(a-1)J_{a-2,0}$. The last two terms can now be evaluated with the help of (32)*. The last term of* (34) *is* (32) *with the function $f(x) = x^{a-1}1\{x\cos\theta + W\sin\theta > 0\}$ which has $f(0) = 0$ for $a \geq 2$. Therefore, this term simply vanishes.*

To evaluate the middle term of (34)*, we introduce a change of variables to express $G\cos\theta + W\sin\theta$ in terms of two other independent Gaussian variables $Z, W \sim \mathcal{N}(0, 1)$*

$$\begin{aligned} Z &= G\cos\theta + W\sin\theta, \qquad & G &= Z\cos\theta + Y\sin\theta, \\ Y &= G\sin\theta - W\cos\theta, \qquad & W &= Z\sin\theta - Y\cos\theta, \end{aligned} \tag{35}$$

where Y, Z iid $\mathcal{N}(0, 1)$. Under this change of variables, $J_{a,0}$, $a \geq 2$ is setup to apply (32) *with $f(x) = 1\{x\cos\theta + Y\sin\theta\}^{a-1}1\{x\cos\theta + Y\sin\theta > 0\}$*

$$\begin{aligned} J_{a,0} &= (a-1)J_{a-2,0} + \mathbf{E}[G^{a-1}1\{G > 0\}1'\{G\cos\theta + W\sin\theta > 0\}]\cos\theta \\ &= (a-1)J_{a-2,0} + \mathbf{E}[(Z\cos\theta + Y\sin\theta)^{a-1}1\{Z\cos\theta + Y\sin\theta > 0\}1'\{Z > 0\}]\cos\theta \\ &= (a-1)J_{a-2,0} + \mathbf{E}[(0 + Y\sin\theta)^{a-1}1\{0 + Y\sin\theta > 0\}]\frac{1}{\sqrt{2\pi}}\cos\theta \\ &= (a-1)J_{a-2,0} + \frac{\sin^{a-1}\theta\cos\theta}{c_{a \bmod 2}}(a-2)!!, \end{aligned}$$

where we have applied Lemma 9 to evaluate the last expectation.

A similar argument is used to find the recursive formula for $J_{a,b}$, $a \geq 2, b \geq 1$, by using (31) *with the function $f(x) = x^{a-1}(x\cos\theta + W\sin\theta)^b 1\{x > 0\}1\{x\cos\theta + W\sin\theta > 0\}$. There are 4 terms in the product rule derivative. Fortunately in this case, the last two terms are simply zero by application of* (32) *since the expressions vanish when $G = 0$, so we get*

$$\begin{aligned} J_{a,b} &= \mathbf{E}[G \cdot G^{a-1}(G\cos\theta + W\sin\theta)^b 1\{G>0\}1\{G\cos\theta + W\sin\theta > 0\}] \\ &= \mathbf{E}[(a-1)G^{a-2}(G\cos\theta + W\sin\theta)^b 1\{G>0\}1\{G\cos\theta + W\sin\theta > 0\}] \\ &\quad + \mathbf{E}[G^{a-1} b\cos\theta(G\cos\theta + W\sin\theta)^{b-1} 1\{G>0\}1\{G\cos\theta + W\sin\theta > 0\}] \\ &\quad + \mathbf{E}[G^{a-1}(G\cos\theta + W\sin\theta)^b 1'\{G>0\}1\{G\cos\theta + W\sin\theta > 0\}] \\ &\quad + \mathbf{E}[G^{a-1}(G\cos\theta + W\sin\theta)^b 1\{G>0\}1'\{G\cos\theta + W\sin\theta > 0\}\cos\theta] \\ &= (a-1)J_{a-2,b} + b\cos\theta J_{a-1,b-1} + 0 + 0, \end{aligned}$$

as desired. ■

3.4 Solving the Recurrence to get an Explicit Formula for $J_{a,b}(\theta)$ - Proof of Theorem 5

Solving the recurrence for the sequences $J_{a,0}$ and $J_{a,1}$ to get the claimed explicit formula for $J_{a,0}$ is a simple induction proof. We defer these to Appendix B.2. More difficult and interesting is the 2D array $J_{a,b}$. To solve the recurrence

$$J_{a,b} = (a-1)J_{a-2,b} + b\cos\theta J_{a-1,b-1}, \quad a \geq 2, b \geq 1, \tag{36}$$

we will apply the recursion repeatedly until $J_{a,b}$ can be expressed as a linear combination of $J_{0,n}$ and $J_{1,n}$ terms for which we already have an explicit formula developed. To determine the coefficients in front of $J_{0,n}$ and $J_{1,n}$, we take a combinatorial approach by thinking of the recurrence relation as a weighted directed graph as defined below.

Definition 10 (Viewing a recursion as a directed weighted graph) *We can view the recurrence relation for $J_{a,b}$ as a weighted directed graph on the vertex set $(a,b) \in \mathbb{Z}^2$ where vertices represent the values of $J_{a,b}$ and directed edges capture how values of $J_{a,b}$ are connected through the recurrence relation. To be precise, the graph edges and edge weights w_e are defined so that the recursion* (36) *for $J_{a,b}$ can be expressed in the graph as a sum over incoming edges,*

$$J_{a,b} = \sum_{e:(a',b')\to(a,b)} w_e^J J_{a',b'}, \tag{37}$$

where the sum is over the edges e with weight w_e^J incoming to the vertex (a,b). An example of the graph to calculate $J_{6,8}$ is illustrated in Figure 5.

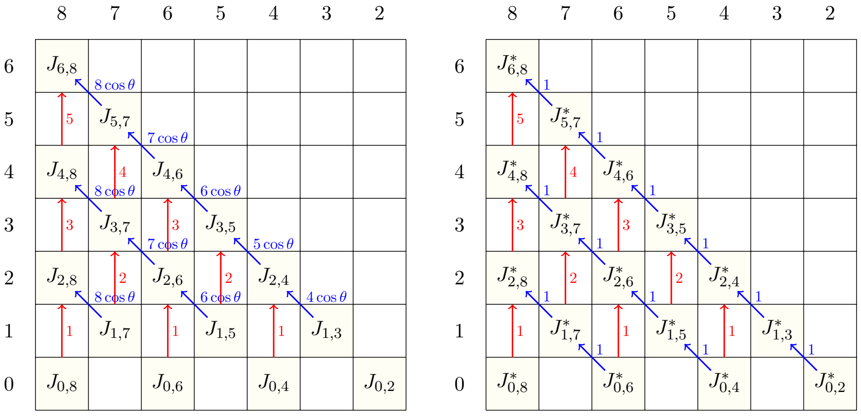
By repeatedly applying the recursion, $J_{a,b}$ can be expressed as a linear combination of the values at the source vertices *of the graph (i.e. those with no incoming edges). For the* *recurrence $J_{a,b}$, the source vertices are $J_{0,n}$ and $J_{1,n}$. The coefficients in front of each source is simply the weighed sum over all paths from the source to the node, namely*

$$J_{a,b} = \sum_{\text{source vertices } v} W^J_{v\to(a,b)} J_v = \sum_{n\geq 0} W^J_{(0,n)\to(a,b)} J_{0,n} + \sum_{n\geq 0} W^J_{(1,n)\to(a,b)} J_{1,n}, \tag{38}$$

$$W^J_{(a',b')\to(a,b)} := \sum_{\pi:(a',b')\to(a,b)} \prod_{e\in\pi} w^J_e, \tag{39}$$

where the sum is over all paths π from the vertex (a',b') to the vertex (a,b) in the J graph.

In light of (38)*, to prove Theorem 5, we have only to calculate the weighted sum of path $W^J_{(0,n)\to(a,b)}$ and $W^J_{(1,n)\to(a,b)}$. These weighted sums turn out to be given in terms of the P and Q numbers which were defined in Definition 4.*



(a) Directed graph associated with J

(b) Directed graph associated with J^*

Figure 5: The graph associated with the recursions for J in (36) (left) and J^* in (43) (right). The graph is defined so that the recursion is given by a sum of incoming edges as in (37). The edges are color coded red and blue to match the coefficients in the recursion.

Proposition 4 (Weighted sums of paths for J) *In the graph for J, we have the following formulas for the sum over weighted paths W^J defined in* (39)*,*

$$W^J_{(0,n)\to(a,b)} = (b)_{b-n}(\cos\theta)^{b-n}(P(a,b-n) - Q(a-1,b-n-1)), \tag{40}$$

$$W^J_{(1,n)\to(a,b)} = (b)_{b-n}(\cos\theta)^{b-n} Q(a-1,b-n). \tag{41}$$

To prove this Proposition 4, we first create a simpler recursion, J^*, which we solve first and then slightly modify the solution to get the solution for J.

Lemma 11 (Weighted sums of paths for J^*) *Let $J^*_{a,b}$ be defined to be the recursion:*

$$J^*_{a,b} := (a-1)J^*_{a-2,b} + 1J^*_{a-1,b-1} \text{ for } 2 \leq a \leq b, \tag{42}$$

$$J^*_{1,b} := 0 + 1J^*_{0,b-1} \text{ for } 1 \leq b. \tag{43}$$

Thinking of this recursion as a graph as in Definition 10 (see Figure 5 for an illustration), we have that the sum of weighted paths W^{J^*} defined analogously to those in (39)*, are given by P and Q numbers, namely*

$$W^{J^*}_{(0,n)\to(a,b)} = P(a, b-n), \quad W^{J^*}_{(1,n)\to(a,b)} = Q(a-1, b-n). \tag{44}$$

The connection between the J^* and the P, Q numbers is through the following recursion for the P, Q numbers.

Lemma 12 *(Recursion for P and Q numbers) The P numbers, defined in Definition 4, satisfy $P(0,0) = 1$, $P(n,n) = P(n-1,n-1)$ for $n \geq 1$, and the recursion*

$$P(a,b) = (a-1) \cdot P(a-2,b) + 1 \cdot P(a-1,b-1), \text{ for } a \geq 2,\ 0 \leq b \leq a-2,$$

under the convention that $P(a,-1) = 0$. The Q numbers satisfy the same recursion as the P numbers, with a coefficient of a rather than $(a-1)$.

The proof of Lemma 12 is an easy consequence of known results from Kreinin (2016) and is deferred to Appendix C.

Proof *[Of Lemma 11] Using the same idea of recursions expressed as graphs as in Definition 10, the recursion from Lemma 12 means that P and Q can be expressed as weighted directed graphs. These are displayed in Figure 6. Since the P and Q graphs have only one single unit valued source vertex at $(0,0)$,* (38) *shows that the P and Q numbers are actually themselves equal to sums over weighted paths in their respective graphs*

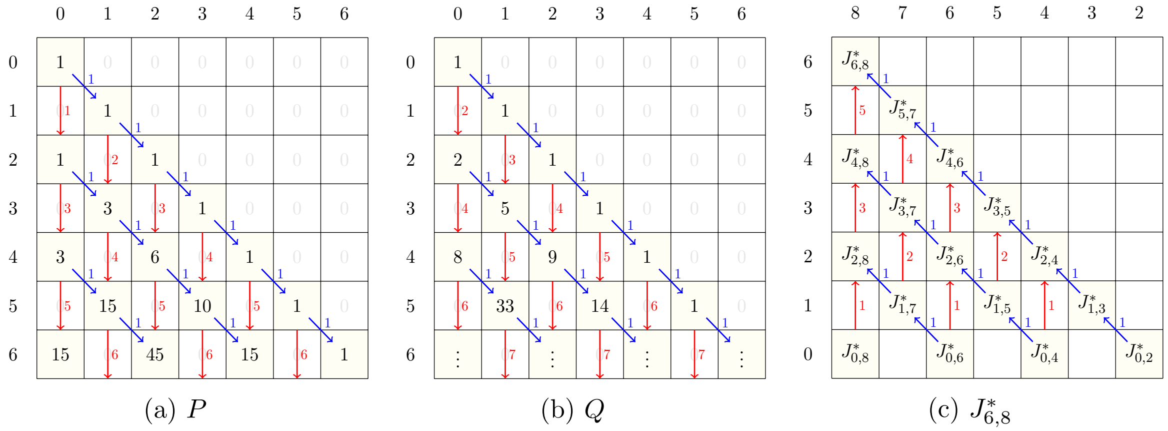
$$P(a,b) = W^{P}_{(0,0)\to(a,b)}, \quad Q(a,b) = W^{Q}_{(0,0)\to(a,b)}.$$

Therefore the statement of the lemma is that sum over weighted paths in the J^* graph are the same as other sums over weighted paths in the P graph/Q graphs,

$$W^{J^*}_{(0,n)\to(a,b)} = W^{P}_{(0,0)\to(a,b-n)}, \quad W^{J^*}_{(1,n)\to(a,b)} = W^{Q}_{(0,0)\to(a-1,b-n)}. \tag{45}$$

The fact that these are equal is demonstrated by establishing a simple bijection between weighted paths in the P graph/Q graph, and weighted paths in the J^* graph. For example, in Figure 6, there is a bijection between the weighted paths in the P graph which connect $P(0,0)$ to $P(6,2)$, to the paths which connect $J^*_{6,8}$ to $J^*_{0,6}$ in the J^* graph. The bijection is simply to flip *any path in the P-graph by rotating it by* 180° *to get a valid path in the J^*-graph. Moreover, the edge weights for J^* and P are precisely set up so that under this bijection, the paths will have the same set of weighted edges in the same order. A full, more detailed, explanation of this bijection is given in Appendix B.3. This argument shows that $W^{J^*}_{(0,n)\to(a,b)} = P(a,b-n)$ as desired.*

The P numbers do not apply for paths between $J^*_{1,n}$ and $J^*_{a,b}$ because we are starting one row higher so the first vertical upward edge is weight 2*. In this case, there is a bijection to the Q-graph after flipping the path. For any path which runs from a node in row 1 to the top left corner of the J^*-graph, we can find the same "flipped" path in the graph of the Q, running from the top left entry to the corresponding node in row $a-1$. (The bijection is explained in detail in Appendix B.3.) Hence $W^{J^*}_{(1,n)\to(a,b)} = Q(a-1,b-n)$ as desired.* ■



(a) P

(b) Q

(c) $J^*_{6,8}$

Figure 6: Graphs associated with the recursions for the P numbers (left), Q numbers (middle), and J^* (right). The weighted edges indicate the coefficients in the recursions for P, Q, J^* respectively. The diagrams are lined up so that the sum of weighted paths in from $J^*_{6,8}$ can be directly read from the P and Q entries in the same location. By reading from the bottom displayed row of P we see that the weighted sum over paths $W^{J^*}_{(0,n)\to(6,8)}$ are 15, 45, 15 and 1 for $n = 8, 6, 4$ and 2 respectively. (Since these are the source vertices, this shows that $J^*_{6,8} = 15J^*_{0,8} + 45J^*_{0,6} + 15J^*_{0,4} + 1J^*_{0,2}$.) From the bottom displayed row of Q we see that the values for sums of weighted paths from vertices $W^{J^*}_{(1,n)\to(6,8)}$ are 33, 14 and 1 for $n = 7, 5$ and 3 respectively.

Having solved for J^* in terms of P and Q, it remains to translate these into the weights for J to obtain Proposition 4.

Proof *[Of Proposition 4]*
The proof follows by relating the weighted sum of paths for J in terms of J^* and then applying the result of Lemma 11. There are two differences between the formula for $J_{a,b}$ compared to $J^*_{a,b}$, which can both be seen in Figure 5. We handle both differences as follows:
Difference #1: *J has a weight of $b\cos\theta$ on the blue diagonal edges $(a, b) \to (a + 1, b + 1)$ vs J^* has a weight of 1.*

This difference is handled by the following observation: any path from $(a, b) \to (a', b')$ in the graph goes through each column between b and b' exactly once. This means that the contribution of the edge weights from these edges do not depend on the details of which path was taken, only the starting and ending points. They always contribute the same factor, $(b)_{b'-b}(\cos\theta)^{b'-b}$. (here $(b)_k = b(b-1)\cdots(b-k+1)$ is the falling factorial with k terms). This argument shows that the weighted sum of paths in J and J^* are related by

$$W^{J}_{(a,b)\to(a',b')} = (b)_{b'-b}(\cos\theta)^{b'-b}W^{J^*}_{(a,b)\to(a',b')}. \tag{46}$$

By the result of Lemma 11, this shows that $W^{J}_{(1,n)\to(a,b)} = (b)_{b-n}(\cos\theta)^{b-n}Q(a-1, b-n)$ as desired. Equation (46) holds for all paths with starting point $a \geq 1$. When $a = 0$, there is one additional difference between J and J^* which is accounted for below.
Difference #2: *$J_{0,n}$ has* no *diagonal edge vs $J^*_{0,n}$ has a diagonal blue edge of weight 1.*

Because of this "missing edge", the only choice in J for paths starting from $(0,n)$ is to first go vertically up by 2 units to $(2,n)$. Hence $W^{J}_{(0,n)\to(a,b)} = W^{J}_{(2,n)\to(a,b)}$. To evaluate this, we use the decomposition of paths in J^* by what their first step is, either a diagonal blue step or a red vertical up step, to see that

$$\begin{aligned} W^{J^*}_{(0,n)\to(a,b)} &= W^{J^*}_{(1,n+1)\to(a,b)} + W^{J^*}_{(2,n)\to(a,b)}, && (47)\\ \implies W^{J^*}_{(2,n)\to(a,b)} &= W^{J^*}_{(0,n)\to(a,b)} - W^{J^*}_{(1,n+1)\to(a,b)} && (48)\\ &= P(a,b-n) - Q(a-1,b-n-1), && (49)\end{aligned}$$

by the result of Lemma 11. By applying now (46) to relate J and J^*, we obtain $W^{J}_{(0,n)\to(a,b)} = W^{J}_{(2,n)\to(a,b)} = (b)_{b-n}(\cos\theta)^{b-n}(P(a,b-n) - Q(a-1,b-n-1))$ as desired. ■

Proof *[Of Theorem 5] The formula is immediate from (38), which writes $J_{a,b}$ as a linear combination of $J_{0,n}$ and $J_{1,n}$, and Proposition 4 which gives the the coefficients.* ■

Appendix A.

A.1 Expected Value Approximation

Lemma 13 *Both the random variables $X = R^{\ell+1}$ and $X = R^{\ell+1}\sin^2(\theta_\ell)$ satisfy*

$$\mathbf{E}[\ln(X)] = \ln(\mathbf{E}[X]) - \frac{\mathbf{Var}[X]}{2\mathbf{E}[X]^2} + \mathcal{O}(n_\ell^{-2}). \tag{50}$$

Proof *First note that by the properties of the logarithm, we have*

$$\ln(X) = \ln\left(\mathbf{E}[X]\left(\frac{\mathbf{E}[X] + (X - \mathbf{E}[X])}{\mathbf{E}[X]}\right)\right) = \ln(\mathbf{E}[X]) + \ln\left(1 + \frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right). \tag{51}$$

We can now apply the Taylor series $\ln(1+x) = x - \frac{x^2}{2} + \epsilon_2(x)$, where $\epsilon_2(x)$ is the Taylor series remainder and satisfies $\epsilon_2(x) = \mathcal{O}(x^3)$. Hence

$$\ln(X) = \ln(\mathbf{E}[X]) + \frac{X - \mathbf{E}[X]}{\mathbf{E}[X]} - \frac{(X - \mathbf{E}[X])^2}{2\mathbf{E}[X]^2} + \epsilon_2\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right).$$

Note that $\mathbf{E}[X - \mathbf{E}[X]] = 0$, and $\mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{Var}[X]$. Thus, if we take the expected value of our above approximation, we get the following:

$$\mathbf{E}[\ln(X)] = \ln(\mathbf{E}[X]) - \frac{\mathbf{Var}[X]}{2\mathbf{E}[X]^2} + \mathbf{E}\left[\epsilon_2\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right)\right].$$

By using bounds on the Taylor series error term $\epsilon_2(x) = \mathcal{O}(x^3)$, one can obtain bounds for this last error term. By (16, 17), both $X = R_{\ell+1}$ and $X = R_{\ell+1}\sin^2(\theta_{\ell+1})$ can be expressed as averages of the form

$$X = \frac{1}{n_\ell^2}\sum_{i,j}^{n_\ell} f(G_i, \hat{G}_j). \tag{52}$$

From the bound on the 3rd moment in Lemma 16, it follows that $\mathbf{E}[\epsilon_2(X - \mathbf{E}[X])] = \mathcal{O}(n_\ell^{-2})$*, thus giving the desired result.* ■

A.2 Variance Approximation

Lemma 14 *Both the random variables* $X = R^{\ell+1}$ *and* $X = R^{\ell+1}\sin^2(\theta_\ell)$ *satisfy*

$$\mathbf{Var}[\ln(X)] = \frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2} + \mathcal{O}(n_\ell^{-2}).$$

Proof *Starting with* (51)*, and using the first term of the Taylor series approximation for* $\ln(1+x) = x + \epsilon_1(x)$ *now, we have that*

$$\ln(X) = \ln(\mathbf{E}[X]) + \frac{X - \mathbf{E}[X]}{\mathbf{E}[X]} + \epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right). \tag{53}$$

where $\epsilon_1(x)$ *is the Taylor error term and satisfies* $\epsilon_1(x) = \mathcal{O}(x^2)$*. Taking the variance of this, we arrive at an approximation of* $\mathbf{Var}[\ln(X)]$*.*

$$\mathbf{Var}[\ln(X)] = \mathbf{Var}\left[\ln(\mathbf{E}[X]) + \frac{X - \mathbf{E}[X]}{\mathbf{E}[X]} + \epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right)\right]$$
$$= \mathbf{Var}\left[\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right] + \mathbf{Var}\left[\epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right)\right] + \mathbf{Cov}\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}, \epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right)\right).$$

As with the expected value approximation, this approximation for variance is used twice, once for $X = R_{\ell+1}$*, and once for* $X = R_{\ell+1}\sin^2(\theta_{\ell+1})$ *(see Section 2.2), both of which can can be expressed as a sum as in (52). Since* $\epsilon_1(x) = \mathcal{O}(x^2)$*, we have that the terms with* $\epsilon_1(x)$ *are both* $\mathcal{O}(n_\ell^{-2})$ *from Lemma 16. Simplifying the first term,* $\mathbf{Var}\left[\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right] = \frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2}$ *gives the result of the Lemma.* ■

A.3 Covariance Approximation

Lemma 15 *Both the random variables* $X = R^{\ell+1}$ *and* $X = R^{\ell+1}\sin^2(\theta_\ell)$ *satisfy*

$$\mathbf{Cov}(\ln(X), \ln(Y)) = \frac{\mathbf{Cov}(X, Y)}{\mathbf{E}[X]\mathbf{E}[Y]} + \mathcal{O}(n_\ell^{-2}).$$

Proof *Using the approximation in (53) for* $\ln(X)$ *and* $\ln(Y)$*, we get the following expression for the covariance:*

$$\begin{aligned}
&\mathbf{Cov}(\ln(X), \ln(Y)) \\
&= \mathbf{Cov}\left(\ln(\mathbf{E}[X]) + \frac{X - \mathbf{E}[X]}{\mathbf{E}[X]} + \epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right), \ln(\mathbf{E}[Y]) + \frac{Y - \mathbf{E}[Y]}{\mathbf{E}[Y]} + \epsilon_1\left(\frac{Y - \mathbf{E}[Y]}{\mathbf{E}[Y]}\right)\right) \\
&= \mathbf{Cov}\left(\frac{X}{\mathbf{E}[X]} + \epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right), \frac{Y}{\mathbf{E}[Y]} + \epsilon_1\left(\frac{Y - \mathbf{E}[Y]}{\mathbf{E}[Y]}\right)\right) \\
&= \mathbf{Cov}\left(\frac{X}{\mathbf{E}[X]}, \frac{Y}{\mathbf{E}[Y]}\right) + \mathbf{Cov}\left(\frac{X}{\mathbf{E}[X]}, \epsilon_1\left(\frac{Y - \mathbf{E}[Y]}{\mathbf{E}[Y]}\right)\right) + \mathbf{Cov}\left(\epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right), \frac{Y}{\mathbf{E}[Y]}\right) \\
&\quad + \mathbf{Cov}\left(\epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right), \epsilon_1\left(\frac{Y - \mathbf{E}[Y]}{\mathbf{E}[Y]}\right)\right).
\end{aligned}$$

We get the desired result from the fact that that the error term $\epsilon_1(x)$ *satisfies* $\epsilon_1(x) = \mathcal{O}(x^2)$ *and from our result in Lemma 18.* ■

A.4 Third and Fourth Moment Bound Lemma

Lemma 16 *Let* $G_i, \hat{G}_i$*,* $1 \le i \le n$ *be marginally* $\mathcal{N}(0,1)$ *random variables with correlation* $\cos(\theta)$ *and independent for different indices i. Let* $A = \frac{1}{n^2}\sum_{i,j}^n f(G_i, \hat{G}_j)$ *be the average over all* n^2 *pairs of some function* $f : \mathbb{R}^2 \to \mathbb{R}$ *which has finite fourth moment,* $\mathbf{E}[f(G_i, \hat{G}_i)^4] < \infty$*. Then, the third and fourth central moment of A satisfy*

$$\mathbf{E}[(A - \mathbf{E}[A])^3] = \mathcal{O}(n^{-2}), \quad \mathbf{E}[(A - \mathbf{E}[A])^4] = \mathcal{O}(n^{-2}). \tag{54}$$

Proof *We begin by showing the third moment bound. First, we can express* $\mathbf{E}[(A - \mathbf{E}[A])^3]$ *as a sum in the following way:*

$$\begin{aligned}
A - \mathbf{E}[A] &= \frac{1}{n^2}\sum_{i,j}^n \left(f(G_i, \hat{G}_j) - \mathbf{E}[f(G_i, \hat{G}_j)]\right) \\
\implies \mathbf{E}\left[(A - \mathbf{E}[A])^3\right] &= \frac{1}{n^6}\sum_{\substack{i_1,i_2,i_3 \\ j_1,j_2,j_3}}^n \mathbf{E}\left[\prod_{k=1}^3 \left(f(G_{i_k}, \hat{G}_{j_k}) - \mathbf{E}[f(G_{i_k}, \hat{G}_{j_k})]\right)\right].
\end{aligned} \tag{55}$$

Note that many of these terms are mean zero. For example, for any configuration of the indices where there is no overlap between the indices (i_1, j_1) *and the other two index pairs* $(\{i_1, j_1\} \cap \{i_2, j_2, i_3, j_3\} = \varnothing)$*, we may use independence to observe that*

$$\begin{aligned}
&\mathbf{E}\left[\prod_{k=1}^3 \left(f(G_{i_k}, \hat{G}_{j_k}) - \mathbf{E}[f(G_{i_k}, \hat{G}_{j_k})]\right)\right] \\
=&\mathbf{E}\left[f(G_{i_1}, \hat{G}_{j_1}) - \mathbf{E}[f(G_{i_1}, \hat{G}_{j_1})]\right] \mathbf{E}\left[\prod_{k=2}^3 \left(f(G_{i_k}, \hat{G}_{j_k}) - \mathbf{E}[f(G_{i_k}, \hat{G}_{j_k})]\right)\right] = 0.
\end{aligned}$$

When this happens we say that (i_1, j_1) is a "reducible point". Similarly, (i_2, j_2) or (i_3, j_3) can be reducible if they have no overlap with the other two index pairs. To control $\mathbf{E}\left[(A - \mathbf{E}[A])^3\right]$, it will suffice to enumerate the number of indices $\{i_1, j_1, i_2, j_2, i_3, j_3\}$ so that all three points $(i_1, j_1), (i_2, j_2), (i_3, j_3)$ are not *reducible. We call these "irreducible configurations".*

We now observe that at least one of the points (i_1, j_1),(i_2, j_2) or (i_3, j_3) is reducible whenever the number of unique numbers is $\left|\bigcup_{k=1}^{3}\{i_k, j_k\}\right| \geq 5$. This is because, by the pigeonhole principle, if there are no repeated or only one repeated number between 6 indices, then at least one of the 3 pairs (i_1, j_1),(i_2, j_2) or (i_3, j_3) must consist of two unique numbers and therefore is a reducible point.

Since the irreducible configurations can only have at most 4 unique numbers, the number of irreducible configurations is $\mathcal{O}(n^4)$ as $n \to \infty$. In fact, a detailed enumeration of the number of configurations reveals that the number of irreducible configurations is precisely

$$32(n)_4 + 68(n)_3 + 28(n)_2 + 1(n)_1. \tag{56}$$

Here, $(n)_k = n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)$ denotes the falling factorial with k terms. The leading term is 32 because there are 32 possible "patterns" for how the indices can be arranged to be both irreducible and contain exactly 4 unique numbers $\left|\bigcup_{k=1}^{3}\{i_k, j_k\}\right| = 4$; these patterns are listed in Table 3. Each pattern contributes $(n)_4 = n(n-1)(n-2)(n-3)$ possible index configurations by filling in the 4 unique numbers in all the possible ways. Similarly, there are respectively 68, 28, and 1 pattern(s) for irreducible configurations with 3,2 and 1 unique number(s) in them which each contribute $(n)_3, (n)_2$ and $(n)_1$ configurations per pattern

Since the number of irreducible configurations is $\mathcal{O}(n^4)$, the normalization by n^6 in (55) shows that $\mathbf{E}[(A - \mathbf{E}[A])^3]$ is $\mathcal{O}(n^{-2})$ as desired for the third moment.

The argument for the 4th moment is similar. We write $\mathbf{E}[(A - \mathbf{E}[A])^4]$ as a sum over $i_1, j_1, i_2, j_2, i_3, j_3, i_4, j_4$ and again enumerate irreducible configurations. In this case, once again by the pigeonhole principle any configuration with 7 or more unique points $\left|\bigcup_{k=1}^{3}\{i_k, j_k\}\right| \geq 7$ will be reducible. Since there are at most 6 unique numbers, there will be $\mathcal{O}(n^6)$ irreducible configurations. A detailed enumeration of all the possible irreducible patterns and the number of unique elements in each yields that the number of irreducible configurations is precisely

$$48(n)_6 + 544(n)_5 + 1268(n)_4 + 844(n)_3 + 123(n)_2 + 1(n)_1.$$

The normalization factor of n^{-8} then shows that $\mathbf{E}[(A - \mathbf{E}[A])^4] = \mathcal{O}(n^{-2})$. ■

Remark 17 *A more detailed enumeration of the 4th moment actually shows that the dominant terms in the 4th moment correspond to the terms in the 2nd moment written twice, and asymptotically*

$$\mathbf{E}[(A - \mathbf{E}[A])^4] = 3\mathbf{E}[(A - \mathbf{E}[A])^2]^2 + \mathcal{O}(n^{-3}).$$

Here, 3 arises as the number of pair partitions of 4 items, and is related to the fact that $3 = \mathbf{E}[G^4]$.

Lemma 18 *Let $G_i, \hat{G}_i$, $1 \le i \le n$ be marginally $\mathcal{N}(0,1)$ random variables with correlation $\cos(\theta)$ and independent for different indices i. Let $A_1 = \frac{1}{n^2}\sum_{i,j}^n f_1(G_i, \hat{G}_j)$, and let $A_2 = \frac{1}{n^2}\sum_{i,j}^n f_2(G_i, \hat{G}_j)$, where $f_1, f_2 : \mathbb{R}^2 \to \mathbb{R}$ have finite fourth moments, $\mathbf{E}[f_1(G_i, \hat{G}_i)^4]$, $\mathbf{E}[f_2(G_i, \hat{G}_i)^4] < \infty$. Then,*

$$\mathbf{E}[(A_1 - \mathbf{E}[A_1])^2(A_2 - \mathbf{E}[A_2])] = \mathcal{O}\left(n^{-2}\right).$$

Proof *We can express $\mathbf{E}[(A_1 - \mathbf{E}[A_2])^2(A_2 - \mathbf{E}[A_2])]$ using sums as follows:*

$$\begin{aligned}&\mathbf{E}[(A_1 - \mathbf{E}[A_1])^2(A_2 - \mathbf{E}[A_2])] \\ &= \frac{1}{n^6}\sum_{\substack{i_1,i_2,i_3 \\ j_1,j_2,j_3}}^{n} \mathbf{E}\left[\prod_{k=1}^{2}\Big(f_1(G_{i_k}, \hat{G}_{j_k}) - \mathbf{E}[f_1(G_{i_k}, \hat{G}_{j_k})]\Big)\Big(f_2(G_{i_3}, \hat{G}_{j_3}) - \mathbf{E}[f_2(G_{i_3}, \hat{G}_{j_3})]\Big)\right].\end{aligned}$$

By the same argument as in Lemma 16, we can show that the number of nonzero terms in the above summation is $\mathcal{O}(n^4)$ as $n \to \infty$. Thus, we have that $\mathbf{E}[(A_1 - \mathbf{E}[A_1])^2(A_2 - \mathbf{E}[A_2])] = \mathcal{O}(n^{-2})$. We can also show that $\mathbf{E}[(A_1 - \mathbf{E}[A_1])^2(A_2 - \mathbf{E}[A_2])^2] = \mathcal{O}(n^{-2})$ by the same argument. ■

(i_1, j_1)		(i_2, j_2)		(i_3, j_3)	
$\{(a,b),(b,a)\}$	×	$\{(a,c),(c,a)\}$	×	$\{(a,d),(d,a)\}$	8 patterns
$\{(a,b),(b,a)\}$	×	$\{(a,c),(c,a)\}$	×	$\{(c,d),(d,c)\}$	8 patterns
$\{(a,b),(b,a)\}$	×	$\{(c,d),(d,c)\}$	×	$\{(a,c),(c,a)\}$	8 patterns
$\{(a,c),(c,a)\}$	×	$\{(a,b),(b,a)\}$	×	$\{(c,d),(c,b)\}$	8 patterns

Table 3: All 32 irreducible patterns using exactly 4 unique index values a, b, c, d. For example the pattern $(i_1, j_1), (i_2, j_2), (i_3, j_3) = (a, b), (a, c), (a, d)$ represents all configurations where $i_1 = i_2 = i_3$ and the j's are all unique and different from i. For each pattern, there are $(n)_4 = n(n-1)(n-2)(n-3)$ configurations by filling in a, b, c, d with unique numbers in $[n]$. These are the dominant terms in (55).

A.5 Derivation of Useful Identities - Equations (14, 15)

Let $G \in \mathbb{R}^n$ be a Gaussian vector with iid entries $G_i \sim \mathcal{N}(0,1)$. Then, by standard properties of Gaussians, the function $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(x) = \langle G, x\rangle$ is a Gaussian random variable. Further, $f(x) \sim \mathcal{N}(0, \|x\|^2)$ for all $x \in \mathbb{R}^n$, and for any two vectors $x_\alpha, x_\beta \in \mathbb{R}^n$, the joint distribution of $f(x_\alpha), f(x_\beta)$ is jointly Gaussian with

$$\begin{bmatrix} f(x_\alpha) \\ f(x_\beta) \end{bmatrix} \sim \mathcal{N}\left(0, \Sigma(x_\alpha, x_\beta)\right), \qquad \Sigma(x_\alpha, x_\beta) := \begin{bmatrix} \|x_\alpha\|^2 & \langle x_\alpha, x_\beta\rangle \\ \langle x_\alpha, x_\beta\rangle & \|x_\beta\|^2 \end{bmatrix},$$

where $\Sigma(x_\alpha, x_\beta)$ is sometimes called the 2×2 Gram matrix of the vectors x_α, x_β. In the setting of our fully connected neural network, any index $i \in [n_{\ell+1}]$ in the vector of $z^{\ell+1}$ is actually the inner product with the i-th row

$$z_i^{\ell+1}(x) = \sqrt{\frac{2}{n_\ell}} \langle W_{i,\cdot}^{\ell+1}, \varphi(z^\ell(x)) \rangle.$$

Note that each row $W_{i,\cdot}^{\ell+1}$ is a Gaussian vector, so the previous fact about Gaussians applies and we see that the entries of $z^{\ell+1}$ are conditionally Gaussian given the value of the previous layer. By the previous Gaussian fact, we have that $z_i^{\ell+1}(x_\alpha)$, $z_i^{\ell+1}(x_\beta)$ are jointly Gaussian with

$$\begin{bmatrix} z_i^{\ell+1}(x_\alpha) \\ z_i^{\ell+1}(x_\beta) \end{bmatrix} \sim \mathcal{N}\left(0, \frac{2}{n_\ell} \begin{bmatrix} \|\varphi_\alpha^\ell\|^2 & \langle \varphi_\alpha^\ell, \varphi_\beta^\ell \rangle \\ \langle \varphi_\alpha^\ell, \varphi_\beta^\ell \rangle & \|\varphi_\beta^\ell\|^2 \end{bmatrix}\right) =: \mathcal{N}\left(0, K^\ell\right),$$

where we use K^ℓ to denote the 2×2 covariance matrix. K^ℓ is precisely the 2×2 Gram matrix of the previous layer φ_α^ℓ, φ_β^ℓ scaled by $2/n_\ell$ and its entries $K_{\gamma\delta}^\ell$, for $\gamma \in \{\alpha, \beta\}, \delta \in \{\alpha, \beta\}$ are actually averages *of entries in the previous layer*

$$K_{\gamma,\delta}^\ell := \frac{2}{n_\ell} \langle \varphi_\gamma^\ell, \varphi_\delta^\ell \rangle = \frac{1}{n_\ell} \sum_{k=1}^{n_\ell} 2\varphi(z_k^\ell(x_i))\varphi(z_k^\ell(x_j)).$$

Moreover, in the weight matrix $W^{\ell+1}$, the i^{th} and j^{th} rows ($W_{i,\cdot}^{\ell+1}$ and $W_{j,\cdot}^{\ell+1}$, respectively) are independent. Therefore, all entries of $z^{\ell+1}$ are identically distributed and conditionally independent given $\varphi(z^\ell)$. From this fact, we can equivalently write the entries explicitly as

$$z_i^{\ell+1}(x_\alpha) = \sqrt{\frac{2}{n_\ell}} \|\varphi_\alpha^\ell\| G_i, \quad z_i^{\ell+1}(x_\beta) = \sqrt{\frac{2}{n_\ell}} \|\varphi_\beta^\ell\| \hat{G}_i, \tag{57}$$

where $G_i, \hat{G}_i$ are marginally $\mathcal{N}(0,1)$ variables with covariance $\mathbf{Cov}(G_i, \hat{G}_i) = \cos(\theta_\ell)$ and independent for different indices. This formulation precisely ensures that the covariance structure for the entries is exactly what is specified by the covariance kernel K^ℓ.

With this representation of $z_i^{\ell+1}(x_\alpha)$ and $z_i^{\ell+1}(x_\beta)$, we can apply $\varphi(\cdot)$ to each entry. By using the property of ReLU $\varphi(\lambda x) = \lambda \varphi(x)$ for $\lambda > 0$ to factor out the norms, we obtain

$$\varphi(z_i^{\ell+1}(x_\alpha)) = \sqrt{\frac{2}{n_\ell}} \|\varphi_\alpha^\ell\| \varphi(G_i), \quad \varphi(z_i^{\ell+1}(x_\beta)) = \sqrt{\frac{2}{n_\ell}} \|\varphi_\beta^\ell\| \varphi(\hat{G}_i). \tag{58}$$

Taking the norm/inner product of the vector now yields (14-16) as desired.

A.6 Cauchy-Binet and Determinant of the Gram Matrix - Equation (16)

To prove this identity, we begin with the fact that

$$\|\varphi_\alpha^{\ell+1}\|^2 \|\varphi_\beta^{\ell+1}\|^2 \sin^2(\theta_{\ell+1}) = \det \begin{bmatrix} \|\varphi_\alpha^{\ell+1}\|^2 & \langle \varphi_\alpha^{\ell+1}, \varphi_\beta^{\ell+1} \rangle \\ \langle \varphi_\alpha^{\ell+1}, \varphi_\beta^{\ell+1} \rangle & \|\varphi_\beta^{\ell+1}\|^2 \end{bmatrix}.$$

By the Cauchy-Binet identity, we can express the determinant as

$$\det\begin{bmatrix} \|\varphi_\alpha^{\ell+1}\|^2 & \langle\varphi_\alpha^{\ell+1},\varphi_\beta^{\ell+1}\rangle \\ \langle\varphi_\alpha^{\ell+1},\varphi_\beta^{\ell+1}\rangle & \|\varphi_\beta^{\ell+1}\|^2 \end{bmatrix} = \sum_{1\le i<j\le n_\ell}\left(\varphi_{i;\alpha}^{\ell+1}\varphi_{j;\beta}^{\ell+1}-\varphi_{j;\alpha}^{\ell+1}\varphi_{i;\beta}^{\ell+1}\right)^2. \tag{59}$$

Due to the fact that the summand is equal to 0 when $i=j$, we can equivalently take the sum over all indices $i,j\in[n_\ell]$ and halve the result. We can also express layer $\ell+1$ using the following conditioning on the previous layer

$$\varphi_{i;\alpha}^{\ell+1}=\sqrt{\frac{2}{n_\ell}}\|\varphi_\alpha^\ell\|\cdot\varphi(G_i),\quad \varphi_{i;\beta}^{\ell+1}=\sqrt{\frac{2}{n_\ell}}\|\varphi_\beta^\ell\|\cdot\varphi(\hat{G}_i).$$

Applying these facts to our expression in (59), and dividing both sides by $\|\varphi_\alpha^\ell\|^2\|\varphi_\beta^\ell\|^2$, we get our desired result.

A.7 Expected Value Calculations

In this section, we derive the formulas for $\mathbf{E}[R_{\ell+1}]$, $\mathbf{E}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right]$. We use $J_{a,b}$ to represent $J_{a,b}(\theta_\ell)$. Note that $\mathbf{E}[\varphi^2(G)]=\frac{1}{2}$, $\mathbf{E}[\varphi^4(G)]=\frac{3}{2}$.

Calculation of $\mathbf{E}[R_{\ell+1}]$:

First, we apply the identity as in (17)*:*

$$\mathbf{E}[R_{\ell+1}]=\left(\frac{2}{n_\ell}\right)^2\mathbf{E}\left[\sum_{i,j=1}^{n_\ell}\varphi^2(G_i)\varphi^2(\hat{G}_j)\right].$$

Whenever $i=j$, taking the expected value will give us a $J_{2,2}$ term. When $i\neq j$, the expected value of this term will be $\mathbf{E}[\varphi^2(G)]^2=\frac{1}{4}$. Since $i=j$ happens n_ℓ times, and therefore $i\neq j$ happens $n_\ell^2-n_\ell$ times, we arrive at the following expression:

$$\mathbf{E}[R_{\ell+1}]=\left(\frac{2}{n_\ell}\right)^2\left(n_\ell J_{2,2}+(n_\ell^2-n_\ell)\left(\frac{1}{4}\right)\right)=\frac{4J_{2,2}-1}{n_\ell}+1.$$

Calculation of $\mathbf{E}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right]$:

Applying the identity (16)*, we get*

$$\begin{aligned}\mathbf{E}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right]&=\frac{2}{n_\ell^2}\mathbf{E}\left[\sum_{i,j}^{n_\ell}\left(\varphi(G_i)\varphi(\hat{G}_j)-\varphi(G_j)\varphi(\hat{G}_i)\right)^2\right]\\&=\frac{2}{n_\ell^2}\mathbf{E}\left[\sum_{i,j}^{n_\ell}\left(\varphi^2(G_i)\varphi^2(\hat{G}_j)-2\varphi(G_i)\varphi(\hat{G}_i)\varphi(G_j)\varphi(\hat{G}_j)+\varphi^2(G_j)\varphi^2(\hat{G}_i)\right)\right].\end{aligned}$$

In the case where $i=j$, the expected value is equal to 0. Thus, we only need to consider the case where $i\neq j$, which happens $n_\ell^2-n_\ell$ times. When $i\neq j$, the expectation of

$\varphi(G_i)\varphi(\hat{G}_i)\varphi(G_j)\varphi(\hat{G}_j)$ is $J_{1,1}^2$, and the expectation of $\varphi^2(G_i)\varphi^2(\hat{G}_j)$ is $\frac{1}{4}$. All together, we have

$$\mathbf{E}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right] = \left(\frac{2}{n_\ell^2}\right)(n_\ell^2 - n_\ell)\left(\frac{1}{4} - 2J_{1,1}^2 + \frac{1}{4}\right) = \frac{(n_\ell - 1)(1 - 4J_{1,1}^2)}{n_\ell}.$$

A.8 Variance and Covariance Calculations

In this section, $\mathbf{Var}\left[R_{\ell+1}\right]$, $\mathbf{Var}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right]$, and $\mathbf{Cov}\left(R_{\ell+1}\sin^2(\theta_{\ell+1}), R_{\ell+1}\right)$ are evaluated. We use $J_{a,b}$ to represent $J_{a,b}(\theta_\ell)$. Note that $\mathbf{E}[\varphi^2(G)] = \frac{1}{2}$, $\mathbf{E}[\varphi^4(G)] = \frac{3}{2}$. We will see that there are simple functions $f_1, f_2 : \mathbb{R}^2 \to \mathbb{R}$ so that all of the variance and covariance calculations can be expressed as sums over i_1, j_1, i_2, j_2 of the form

$$\frac{1}{n_\ell^4}\sum_{\substack{i_1,j_1\\ i_2,j_2}}\left(\mathbf{E}\left[f_1(G_{i_1},\hat{G}_{j_1})f_2(G_{i_2},\hat{G}_{j_2})\right] - \mathbf{E}\left[f_1(G_{i_1},\hat{G}_{j_1})\right]\mathbf{E}\left[f_2(G_{i_2},\hat{G}_{j_2})\right]\right), \qquad (60)$$

where the sum goes over index configurations $(i_1, j_1), (i_2, j_2) \in [n_\ell]^4$. We will use this form to organize our calculations of the variance and covariance formulas. The strategy is to evaluate each term in the sum (60) individually.

Since the random variables $\{G_i, \hat{G}_i\}_{i=1}^n$ are exchangeable, the only thing that matters is the "pattern" of which of the indices i_1, j_1, i_2, j_2 are repeated versus which are distinct. For example, there will be n index configurations where $i_1 = j_1 = i_2 = j_2$ are all equal. All n of these give same contribution. There are $(n)_4 = n(n-1)(n-2)(n-3)$ configurations where i_1, j_1, i_2, j_2 are all distinct. Knowing which indices are repeated/distinct allows us to evaluate the corresponding term in (60). We use the following formal notion of a pattern to organize this idea of repeated versus distinct indices.

***Definition 19** A **pattern** for $(i_1, j_1), (i_2, j_2)$ is a subset of all possible index configurations $(i_1, j_1), (i_2, j_2) \in [n]^4$ represented by an assignment of each index to the letters a, b, c, d. Each letter a, b, c, d represents a choice of* unique *indices from $[n]$.*

For example, the pattern $(i_1, j_1), (i_2, j_2) = (a, a), (a, a)$ represents the set of all index configurations where all indices are equal and the pattern $(i_1, j_1), (i_2, j_2) = (a, b), (c, d)$ represents the set with all indices unique. The pattern $(i_1, j_1), (i_2, j_2) = (a, b), (a, c)$ represents all configurations where $i_1 = i_2$ and j_1, j_2 are unique and different from $i_1 = i_2$. For this pattern, there are $(n)_3 = n(n-1)(n-2)$ configurations by filling in a, b, c with unique numbers in $[n]$. More generally, for a pattern with k letters, there are $(n)_k$ configurations that fall into that pattern.

Fortunately, when enumerating (60), many patterns have no *contribution and can be ignored. We formalize this in the following definition.*

***Definition 20** We say that the configuration of indices $(i_1, j_1), (i_2, j_2)$ is **reducible** if $\{i_1, j_1\} \cap \{i_2, j_2\} = \varnothing$. Otherwise, the index configuration is called **irreducible**. A pattern is called reducible if all index configuration in that pattern are reducible.*

By the independence of the random variables $f_1(G_{i_1}, G_{j_1})$ and $f_2(G_{i_2}, G_{j_2})$, whenever $(i_1, j_1), (i_2, j_2)$ is reducible, we see that the corresponding term in (60) completely vanishes!

Therefore, to evaluate (60), we have only to understand the contribution of irreducible configurations. The irreducible configurations can be organized into irreducible patterns. For example, the pattern $(a,b),(c,c)$ is reducible (since formally $\{a,b\} \cap \{c\} = \varnothing$) and so any configuration from this pattern has no contribution in the expectation.

There are 11 irreducible patterns. (All these patterns are listed as part of Table 4.) The expected value of the terms for each pattern will give a contribution that is expressed in terms of the $J_{a,b}$ depending on the details of exactly which indices are repeated. Then by enumerating the number of configurations in each pattern, we can evaluate (60). This strategy is precisely how we evaluate each variance/covariance in this section.

Calculation of $\mathbf{Var}\left[R_{\ell+1}\right]$:

First, applying the identity in (17), we get

$$\begin{aligned}\mathbf{Var}\left[R_{\ell+1}\right] &= \left(\frac{2}{n_\ell}\right)^4 \mathbf{Var}\left[\sum_{i,j=1}^{n_\ell} \varphi^2(G_i)\varphi^2(\hat{G}_j)\right] \\ &= \frac{16}{n_\ell^4}\left(\mathbf{E}\left[\sum_{\substack{i_1,j_1\\ i_2,j_2}} \varphi^2(G_{i_1})\varphi^2(\hat{G}_{j_1})\varphi^2(G_{i_2})\varphi^2(\hat{G}_{j_2})\right] - \mathbf{E}\left[\sum_{i,j=1}^{n_\ell} \varphi^2(G_i)\varphi^2(\hat{G}_j)\right]^2\right).\end{aligned}$$

$\mathbf{Var}[R_{\ell+1}]$ follows the form of (60), with $f_1(G_i,\hat{G}_i) = f_2(G_i,\hat{G}_i) = \varphi^2(G_i)\varphi^2(\hat{G}_j)$. We then evaluate the contribution from each irreducible pattern in Table 4. Combining all these cases and simplifying based on powers of $\frac{1}{n_\ell}$, we arrive at the following expression for $\mathbf{Var}\left[R_{\ell+1}\right]$:

$$\frac{4}{n_\ell}(J_{2,2}+1) + \frac{16}{n_\ell^2}\left(2J_{4,2} - \frac{5}{2}J_{2,2} + J_{2,2}^2 + \frac{5}{8}\right) + \frac{16}{n_\ell^3}\left(J_{4,4} - 2J_{4,2} - 2J_{2,2}^2 + 2J_{2,2} - \frac{9}{8}\right).$$

Calculation of $\mathbf{Var}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right]$:

Applying identity (16), we can express $\mathbf{Var}[R_{\ell+1}\sin^2(\theta_{\ell+1})]$ as

$$\mathbf{Var}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right] = \frac{1}{4}\left(\frac{2}{n_\ell}\right)^4 \mathbf{Var}\left[\sum_{i,j=1}^{n_\ell}\Big(\varphi(G_i)\varphi(\hat{G}_j) - \varphi(G_j)\varphi(\hat{G}_i)\Big)^2\right].$$

Note that we can express $\mathbf{Var}[R_{\ell+1}\sin^2(\theta_{\ell+1})]$ as in (60) by letting $f_1(G_i,\hat{G}_j) = f_2(G_i,\hat{G}_j) = (\varphi(G_i)\varphi(\hat{G}_j) - \varphi(G_j)\varphi(\hat{G}_i))^2$. We then evaluate the contribution from each irreducible pattern in Table 5. Combining all these cases and simplifying based on powers of $\frac{1}{n_\ell}$, we arrive at the following expression:

$$\begin{aligned}\mathbf{Var}\left[R_{\ell+1}\sin^2(\theta_{\ell+1})\right] &= \frac{8}{n_\ell}\left(-8J_{1,1}^4 + 8J_{1,1}^2J_{2,2} + 4J_{1,1}^2 - 8J_{1,1}J_{3,1} + J_{2,2} + 1\right) \\ &+ \frac{2}{n_\ell^2}\left(80J_{1,1}^4 - 96J_{1,1}^2J_{2,2} - 40J_{1,1}^2 + 96J_{1,1}J_{3,1} + 24J_{2,2}^2 - 12J_{2,2} - 32J_{3,1}^2 + 5\right) \\ &+ \frac{2}{n_\ell^3}\left(-48J_{1,1}^4 + 64J_{1,1}^2J_{2,2} + 24J_{1,1}^2 - 64J_{1,1}J_{3,1} - 24J_{2,2}^2 + 8J_{2,2} + 32J_{3,1}^2 - 9\right).\end{aligned}$$

$\mathbf{Var}[R_{\ell+1}]$ Calculation

#	(i_1, j_1)	(i_2, j_2)	$\mathbf{E}[f_1(G_{i_1}, \hat{G}_{j_1})]$	$\mathbf{E}[f_2(G_{i_2}, \hat{G}_{j_2})]$	$\mathbf{E}[f_1(G_{i_1}, \hat{G}_{j_1}) f_2(G_{i_2}, \hat{G}_{j_2})]$
$(n)_1$	(a, a)	(a, a)	$J_{2,2}$	$J_{2,2}$	$J_{4,4}$
$(n)_2$	(a, b)	(a, b)	$\left(\frac{1}{2}\right)^2$	$\left(\frac{1}{2}\right)^2$	$\left(\frac{3}{2}\right)^2$
	(a, b)	(b, a)			$J_{2,2}^2$
	(a, a)	(a, b)	$J_{2,2}$	$\left(\frac{1}{2}\right)^2$	$\frac{1}{2} J_{4,2}$
	(a, a)	(b, a)			
	(a, b)	(a, a)	$\left(\frac{1}{2}\right)^2$	$J_{2,2}$	
	(b, a)	(a, a)			
$(n)_3$	(a, b)	(a, c)	$\left(\frac{1}{2}\right)^2$	$\left(\frac{1}{2}\right)^2$	$\frac{3}{2} \left(\frac{1}{2}\right)^2$
	(a, b)	(c, b)			
	(a, b)	(c, a)			$\left(\frac{1}{2}\right)^2 J_{2,2}$
	(a, b)	(b, c)			

Table 4: $\mathbf{Var}[R_{\ell+1}]$ calculated in the form of (60) with $f_1(G_i, \hat{G}_j) = f_2(G_i, \hat{G}_j) = \varphi^2(G_i)\varphi^2(\hat{G}_j)$. The contribution from all 11 possible *irreducible* patterns of the indices are shown.

Calculation of $\mathbf{Cov}\left(R_{\ell+1} \sin^2(\theta_{\ell+1}), R_{\ell+1}\right)$:

$$\mathbf{Cov}\left(R_{\ell+1} \sin^2(\theta_{\ell+1}), R_{\ell+1}\right) = \mathbf{E}\left[(R_{\ell+1})^2 \sin^2(\theta_{\ell+1})\right] - \mathbf{E}\left[R_{\ell+1} \sin^2(\theta_{\ell+1})\right] \mathbf{E}\left[R_{\ell+1}\right].$$

Applying known identities (16, 17) derived in Appendix A.5 and Appendix A.6, we can express this in the form of (60)*, where* $f_1(G_i, \hat{G}_j) = (\varphi(G_i)\varphi(\hat{G}_j) - \varphi(G_j)\varphi(\hat{G}_i))^2$*, and* $f_2(G_i, \hat{G}_j) = \varphi^2(G_i)\varphi^2(\hat{G}_j)$*. Table 6 shows the calculation of all the irreducible patterns. Collecting all cases and simplifying based on powers of* $\frac{1}{n_\ell}$ *gives:*

$$\mathbf{Cov}\left(R_{\ell+1} \sin^2(\theta_{\ell+1}), R_{\ell+1}\right) = \frac{1}{n_\ell}\left(16J_{1,1}^2 - 32J_{1,1}J_{3,1} + 8J_{2,2} + 8\right)$$

$$+\frac{1}{n_\ell^2}\left(32J_{1,1}^2 J_{2,2} - 40J_{1,1}^2 + 96J_{1,1}J_{3,1} - 32J_{1,1}J_{3,3} + 16J_{2,2}^2 - 32J_{2,2} - 32J_{3,1}^2 + 16J_{4,2} + 10\right)$$

$$+\frac{1}{n_\ell^3}\left(24J_{1,1}^2 - 32J_{1,1}^2 J_{2,2} - 64J_{1,1}J_{3,1} + 32J_{1,1}J_{3,3} - 16J_{2,2}^2 + 24J_{2,2} + 32J_{3,1}^2 - 16J_{4,2} - 18\right).$$

A.9 Infinite Width Update Rule

Lemma 21 *Let* $f(x)$ *be a feed forward neural network as defined in 2. Conditional on the value of* θ_ℓ *in layer* ℓ*, the angle* θ_ℓ *between inputs at layer* ℓ *of* f *follows the following*

$\mathbf{Var}[R_{\ell+1}\sin^2(\theta_{\ell+1})]$ Calculation

#	(i_1, j_1)	(i_2, j_2)	$\mathbf{E}[f(G_{i_1}, \hat{G}_{j_1})]$	$\mathbf{E}[f(G_{i_2}, \hat{G}_{j_2})]$	$\mathbf{E}[f(G_{i_1}, \hat{G}_{j_1})f(G_{i_2}, \hat{G}_{j_2})]$
$(n)_2$	(a, b)	(a, b)	$\frac{1}{2} - 2J_{1,1}^2$	$\frac{1}{2} - 2J_{1,1}^2$	$6J_{2,2}^2 - 8J_{3,1}^2 + \frac{9}{2}$
	(a, b)	(b, a)			
$(n)_3$	(a, b)	(a, c)			$4J_{2,2}J_{1,1}^2 - 4J_{3,1}J_{1,1} + \frac{1}{2}J_{2,2} + \frac{3}{4}$
	(a, b)	(c, a)			
	(a, b)	(c, b)			
	(a, b)	(b, c)			

Table 5: $\mathbf{Var}[R_{\ell+1}\sin^2(\theta_{\ell+1})]$ calculated in the form of (60) with $f_1(G_i, \hat{G}_j) = f_2(G_i, \hat{G}_j) = (\varphi(G_i)\varphi(\hat{G}_j) - \varphi(G_j)\varphi(\hat{G}_i))^2$. The *non-zero* contribution *irreducible* patterns of the indices are shown. Note that because $f_1(G_{i_1}, G_{j_1}) = 0$ when $i_1 = j_1$ and $f_2(G_{i_2}, G_{j_2}) = 0$ when $i_2 = j_2$, there are 5 irreducible patterns (of the possible 11) that have zero contribution and are not displayed in this table.

$\mathbf{Cov}(R_{\ell+1}, R_{\ell+1}\sin^2(\theta_{\ell+1}))$ Calculation

#	(i_1, j_1)	(i_2, j_2)	$\mathbf{E}[f_1(G_{i_1}, \hat{G}_{j_1})]$	$\mathbf{E}[f_2(G_{i_2}, \hat{G}_{j_2})]$	$\mathbf{E}[f_1(G_{i_1}, \hat{G}_{j_1})f_2(G_{i_2}, \hat{G}_{j_2})]$
$(n)_2$	(a, b)	(b, b)	$\frac{1}{2} - 2J_{1,1}^2$	$J_{2,2}$	$J_{4,2} - 2J_{1,1}J_{3,3}$
	(a, b)	(a, a)			
	(a, b)	(a, b)		$\left(\frac{1}{2}\right)^2$	$J_{2,2}^2 - 2J_{3,1}^2 + \left(\frac{3}{2}\right)^2$
	(a, b)	(b, a)			
$(n)_3$	(a, b)	(a, c)			
	(a, b)	(c, a)			
	(a, b)	(b, c)			
	(a, b)	(c, b)			

Table 6: $\mathbf{Cov}\left(R_{\ell+1}\sin^2(\theta_{\ell+1}), R_{\ell+1}\right)$ calculated in the form of (60) with $f_1(G_i, \hat{G}_j) = (\varphi(G_i)\varphi(\hat{G}_j) - \varphi(G_j)\varphi(\hat{G}_i))^2$, and $f_2(G_i, \hat{G}_j) = \varphi^2(G_i)\varphi^2(\hat{G}_j)$. The *non-zero* contribution from *irreducible* patterns of the indices are shown. Note that because $f_1(G_{i_1}, G_{j_1}) = 0$ when $i_1 = j_1$, there are 3 irreducible patterns (of the possible 11) that have zero contribution which are *not* displayed in this table.

deterministic update rule in the limit $n_\ell \to \infty$.

$$\cos(\theta_{\ell+1}) = 2J_{1,1}(\theta_\ell).$$

Remark 22 *Note that a more general proof of this result appears in prior work Hanin (2023) which allows one to take the layer sizes $n_1, n_2, \dots, n_\ell \to \infty$ in any order, rather than one layer at a time as we prove here.*

Proof *We begin with the identity (15), and use the inner product to introduce* $\cos(\theta_{\ell+1})$,

$$\frac{\|\varphi_\alpha^\ell\|\|\varphi_\beta^\ell\|}{n_\ell}\sum_{i=1}^{n_\ell} 2\varphi(G_i)\varphi(\hat{G}_i) = \langle \varphi_\alpha^{\ell+1}, \varphi_\beta^{\ell+1}\rangle = \|\varphi_\alpha^{\ell+1}\|\|\varphi_\beta^{\ell+1}\|\cos(\theta_{\ell+1}).$$

Applying the identities in (14) *to* $\|\varphi_\alpha^{\ell+1}\|$ *and* $\|\varphi_\beta^{\ell+1}\|$, *we get*

$$\begin{aligned}\frac{\|\varphi_\alpha^\ell\|\|\varphi_\beta^\ell\|}{n_\ell}\sum_{i=1}^{n_\ell} 2\varphi(G_i)\varphi(\hat{G}_i) &= \sqrt{\frac{\|\varphi_\alpha^\ell\|^2}{n_\ell}\sum_{i=1}^{n_\ell} 2\varphi^2(G_i)}\sqrt{\frac{\|\varphi_\beta^\ell\|^2}{n_\ell}\sum_{i=1}^{n_\ell} 2\varphi^2(\hat{G}_i)}\cos(\theta_{\ell+1}),\\ \implies \frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\varphi(G_i)\varphi(\hat{G}_i) &= \sqrt{\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\varphi^2(G_i)}\sqrt{\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\varphi^2(\hat{G}_i)}\cos(\theta_{\ell+1}).\end{aligned}$$

Now, in the limit $n_\ell \to \infty$ we have by application of the Law of Large Numbers,

$$\begin{aligned}\lim_{n_\ell\to\infty}\left(\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\varphi(G_i)\varphi(\hat{G}_i)\right) &= \lim_{n_\ell\to\infty}\left(\sqrt{\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\varphi^2(G_i)}\sqrt{\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\varphi^2(\hat{G}_i)}\cos(\theta_{\ell+1})\right)\\ \implies \mathbf{E}\left[\varphi(G_i)\varphi(\hat{G}_i)\right] &= \sqrt{\mathbf{E}\left[\varphi^2(G_i)\right]}\sqrt{\mathbf{E}[\varphi^2(\hat{G}_i)]}\cos(\theta_{\ell+1})\\ \implies J_{1,1}(\theta_\ell) &= \frac{1}{2}\cos(\theta_{\ell+1}),\end{aligned}$$

where we have used the definition of $J_{1,1}(\theta)$ and the fact that $\mathbf{E}[\varphi^2(G)] = \frac{1}{2}$. ■

Appendix B.

B.1 Derivation of Lower-Order J Functions - Proof of Proposition 2

Proof *[Of formula for $J_{0,0}$] We find a differential equation that $J_{0,0}$ satisfies and solve it to obtain the formula. First note the initial condition* $J_{0,0}(0) = \mathbf{E}[1\{G > 0\}] = \frac{1}{2}$. *To find $J_{0,0}'(\theta)$, we take the derivative inside the expectation and have by the chain rule that*

$$\begin{aligned}J_{0,0}'(\theta) = &\,\mathbf{E}[1\{G > 0\}1'\{G\cos\theta + W\sin\theta > 0\}G](-\sin\theta)\\ &+ \mathbf{E}[1\{G > 0\}1'\{G\cos\theta + W\sin\theta > 0\}W]\cos\theta.\end{aligned}$$

Applying the change of variables as in (35), *we have*

$$\begin{aligned}J_{0,0}'(\theta) &= \mathbf{E}[(Z\cos\theta + Y\sin\theta)1\{Z\cos\theta + Y\sin\theta > 0\}1'\{Z > 0\}](-\sin\theta)\\ &\quad + \mathbf{E}[(Z\sin\theta - Y\cos\theta)1\{Z\cos\theta + Y\sin\theta > 0\}1'\{Z > 0\}]\cos\theta\\ &= \mathbf{E}[(Y\sin\theta)1\{Y\sin\theta > 0\}]\frac{-\sin\theta}{\sqrt{2\pi}} + \mathbf{E}[(-Y\cos\theta)1\{Y\sin\theta > 0\}]\frac{\cos\theta}{\sqrt{2\pi}}\\ &= (-\sin^2\theta - \cos^2\theta)\mathbf{E}[Y1\{Y > 0\}]\frac{1}{\sqrt{2\pi}} = -\frac{1}{2\pi},\end{aligned}$$

where we have used (32) to evaluate the integrals involving $1'\{Z > 0\}$ *and* $\mathbf{E}[Y1\{Y > 0\}] = (\sqrt{2\pi})^{-1}$ *from Lemma 9. We now have* $J'_{0,0}(\theta) = -\frac{1}{2\pi}$ *with initial condition given by* $J_{0,0}(0) = 0$. *Solving this differential equation gives the desired result.* ■

Proof *[$J_{1,0}$ and $J_{1,1}$] Here we use the Gaussian integration-by-parts strategy (31 -32).*

Formula for $J_{1,0}$***:***

$$\begin{aligned}J_{1,0}(\theta) &= \mathbf{E}[G1\{G > 0\}\ 1\{G\cos\theta + W\sin\theta > 0\}]\\ &= \mathbf{E}\left[\frac{d}{dg}\left(\ 1\{G > 0\}\ 1\{G\cos\theta + W\sin\theta > 0\}\right)\right]\\ &= \mathbf{E}\left[1'\{G > 0\}\ 1\{G\cos\theta + W\sin\theta > 0\}\right] + \mathbf{E}\left[1\{G > 0\}\ 1'\{G\cos\theta + W\sin\theta > 0\}\right]\cos\theta.\end{aligned}$$

By using the change of variables as in (35) on the second term, we arrive at

$$\begin{aligned}J_{1,0}(\theta) &= \mathbf{E}[1\{W\sin\theta > 0\}]\frac{1}{\sqrt{2\pi}} + \cos\theta\mathbf{E}[1\{Z\cos\theta + Y\sin\theta > 0\}1'\{Z > 0\}]\\ &= \frac{1}{2}\frac{1}{\sqrt{2\pi}} + \cos\theta\mathbf{E}[1\{Y\sin\theta > 0\}]\frac{1}{\sqrt{2\pi}} = \frac{1}{2}\frac{1}{\sqrt{2\pi}} + \frac{\cos\theta}{2}\frac{1}{\sqrt{2\pi}} = \frac{1+\cos\theta}{2\sqrt{2\pi}}.\end{aligned}$$

Formula for $J_{1,1}$***:***

$$\begin{aligned}J_{1,1}(\theta) &= \mathbf{E}[G(G\cos\theta + W\sin\theta)1\{G > 0\}1\{G\cos\theta + W\sin\theta > 0\}]\\ &= \mathbf{E}[\cos\theta\ 1\{G > 0\}1\{G\cos\theta + W\sin\theta\}]\\ &+ \mathbf{E}[(G\cos\theta + W\sin\theta)1'\{G > 0\}1\{G\cos\theta + W\sin\theta > 0\}]\\ &+ \mathbf{E}[(G\cos\theta + W\sin\theta)1\{G > 0\}1'\{G\cos\theta + W\sin\theta > 0\}]\cos\theta\\ &= \cos\theta J_{0,0} + \mathbf{E}[W\sin\theta\ 1\{W\sin\theta > 0\}]\frac{1}{\sqrt{2\pi}} + \mathbf{E}[Z1\{Z\cos\theta + Y\sin\theta\}1'\{z > 0\}]\\ &= \cos\theta J_{0,0} + \sin\theta\mathbf{E}[\varphi(W)]\frac{1}{\sqrt{2\pi}} + 0 = \frac{\sin\theta + (\pi-\theta)\cos\theta}{2\pi}.\end{aligned}$$

■

B.2 Proof of Explicit Formulas for $J_{n,0}$ and $J_{n,1}$

Once the recursion is established, the formula for both $J_{n,0}$ *and* $J_{n,1}$ *is a simple proof by induction. We provide a detailed proof for* $J_{n,0}$ *here;* $J_{n,1}$ *is similar.*

Lemma 23 *Let* $J^{rec}_{n,0}$ *be the recursively defined formula, and let* $J^{exp}_{n,0}$ *be the explicitly defined formula for* $J_{n,0}$*, namely*

$$J^{rec}_{n,0} := (n-1)J^{rec}_{n-2,0} + \frac{\sin^{n-1}\theta\cos\theta}{c_{n \bmod 2}}(n-2)!!\,,\qquad J^{rec}_{1,0} := J_{1,0},\qquad J^{rec}_{0,0} := J_{0,0},$$

$$J^{exp}_{n,0} := (n-1)!!\left(J_{n \bmod 2,0} + \frac{\cos\theta}{c_{n \bmod 2}}\sum_{\substack{i\not\equiv n(\bmod 2)\\ 0<i<n}}\frac{(i-1)!!}{i!!}\sin^i\theta\right).$$

Then $J^{rec}_{n,0} = J^{exp}_{n,0}$ for all $n \geq 0$.

Proof *Let S_n, $n \in \mathbb{N}$, $n \geq 2$ be the statement $J^{rec}_{n,0} = J^{exp}_{n,0}$ and $J^{rec}_{n-1,0} = J^{exp}_{n-1,0}$. We prove S_n is true by induction. The base case S_2 is true because,*

$$\begin{aligned} J^{rec}_{2,0} &= (2-1)J_{0,0} + \frac{\sin\theta\cos\theta}{c_{2 \bmod 2}}(2-2)!! = J_{0,0} + \frac{\cos\theta\sin\theta}{2\pi}, \\ J^{exp}_{2,0} &= (2-1)!!\, J_{0,0} + \cos\theta \sum_{i=1}^{1} \frac{(2-1)!!}{(2i-1)!!}(2i-2)!!\, \frac{\sin^{2i-1}\theta}{2\pi} = J_{0,0} + \frac{\cos\theta\sin\theta}{2\pi}, \end{aligned}$$

and the fact that $J^{rec}_{1,0} = J^{exp}_{1,0}$ is trivial. Induction step: Assume S_n is true. To prove S_{n+1}, we have only to show that $J^{exp}_{n+1,0} = J^{rec}_{n+1,0}$. To do this, we separate the last term of the sum to get

$$\begin{aligned} J^{exp}_{n+1,0} = n!!\left(J_{(n+1) \bmod 2,0} + \frac{\cos\theta}{c_{(n+1) \bmod 2}} \sum_{\substack{i \not\equiv (n+1) (\bmod 2) \\ 0<i<n-1}} \frac{(i-1)!!}{i!!}\sin^i\theta \right) \\ + n!!\,\frac{\cos\theta}{c_{(n+1) \bmod 2}}\frac{(n-1)!!}{n!!}\sin^n\theta. \end{aligned}$$

Because the parity of $n+1$ and $n-1$ are the same, and using $n!! = n(n-2)!!$ we recognize the first term as $nJ^{exp}_{n-1,0}$. So after simplifying the last term, we remain with

$$J^{exp}_{n+1,0} = nJ^{exp}_{n-1,0} + \frac{\sin^n\theta\cos\theta}{c_{(n+1) \bmod 2}}(n-1)!! = J^{rec}_{n+1,0},$$

by the induction hypothesis. This completes the induction. ■

B.3 Bijection between Paths in Graphs of J Functions and the Bessel Number graphs P,Q

Let $G_{J^*} = (V_{J*}, E_{J*})$ be the graph of $J^*_{a,b}$ as in Figure 6c. Similarly, let $G_P = (V_P, E_P)$ and $G_Q = (V_Q, E_Q)$ be the graph of the P and Q matrices up to row a, respectively, as in Figures 6a, 6b. We define a map $\lambda : \mathbb{Z}^2 \times \mathbb{Z}^2 \to \mathbb{Z}^2$ as follows: Let $((i,j),(m,n)) \in \mathbb{Z}^2 \times \mathbb{Z}^2$, $0 \leq i \leq a$, $b-a+m \leq j \leq b$. Then define λ by

$$\lambda((i,j),(m,n)) := (i-m, j-n), \quad \lambda^{-1}((i,j),(m,n)) = (i+m, j+n).$$

The function λ can be used as a map between vertices of graph G_{J^*} to vertices of graph G_P or G_Q. Let $\pi = (v_1, v_2, ..., v_{k-1}, v_k)$ be a path in G_{J^*} from vertex $v_1 = (m,n)$ to vertex $v_k = (a,b)$, where $v_i \in \mathbb{Z}^2$, $1 \leq i \leq k$ is a vertex on the graph. λ extends to a map on paths, Λ, defined by

$$\begin{aligned} \Lambda((v_1, v_2, ..., v_{k-1}, v_k)) :=& (\lambda(v_1, v_1), \lambda(v_2, v_1), ..., \lambda(v_{k-1}, v_1), \lambda(v_k, v_1)), \\ \Lambda^{-1}((v_1, v_2, ..., v_{k-1}, v_k)) =& (\lambda^{-1}(v_1, v_1), \lambda^{-1}(v_2, v_1), ..., \lambda^{-1}(v_{k-1}, v_1), \lambda^{-1}(v_k, v_1)). \end{aligned}$$

Now, let $\Gamma_{J^*}(a,b,m,n)$ be the set of all paths in the graph of J^* from $J^*_{m,n}$ to $J^*_{a,b}$, and let $\Gamma_P(a,b,m,n)$ be the set of all paths in the graph of P from $P(0,0)$ to $P(a-m,b-n) = P(\lambda((a,b),(m,n)))$. For example, $\Gamma_{J^*}(6,8,0,4)$ is the set of all paths which run from $J^*_{6,8}$ to $J^*_{0,4}$, and $\Gamma_P(6,8,0,4)$ is the set of all paths which run from $P(0,0)$ to $P(6,4)$.

With these definitions, $\Lambda : \Gamma_{J^*}(a,b,0,n) \to \Gamma_P(a,b,0,n)$ is a bijection. An illustration of all paths $\pi \in \Pi(6,8,0,6)$ and the corresponding paths $\Lambda(\pi) \in \Gamma_P(6,8,0,6)$ is given in Figure 7. Similarly, if we let $\Gamma_Q(a,b,m,n)$ be the set of all paths from $Q(0,0)$ to $Q(a-m,b-n) = Q(\lambda((a,b),(m,n)))$ then $\Lambda : \Gamma_{J^*}(a,b,1,n) \to \Gamma_Q(a,b,1,n)$ is a bijection. This bijection establishes the equality of the weighted paths claim in (45).

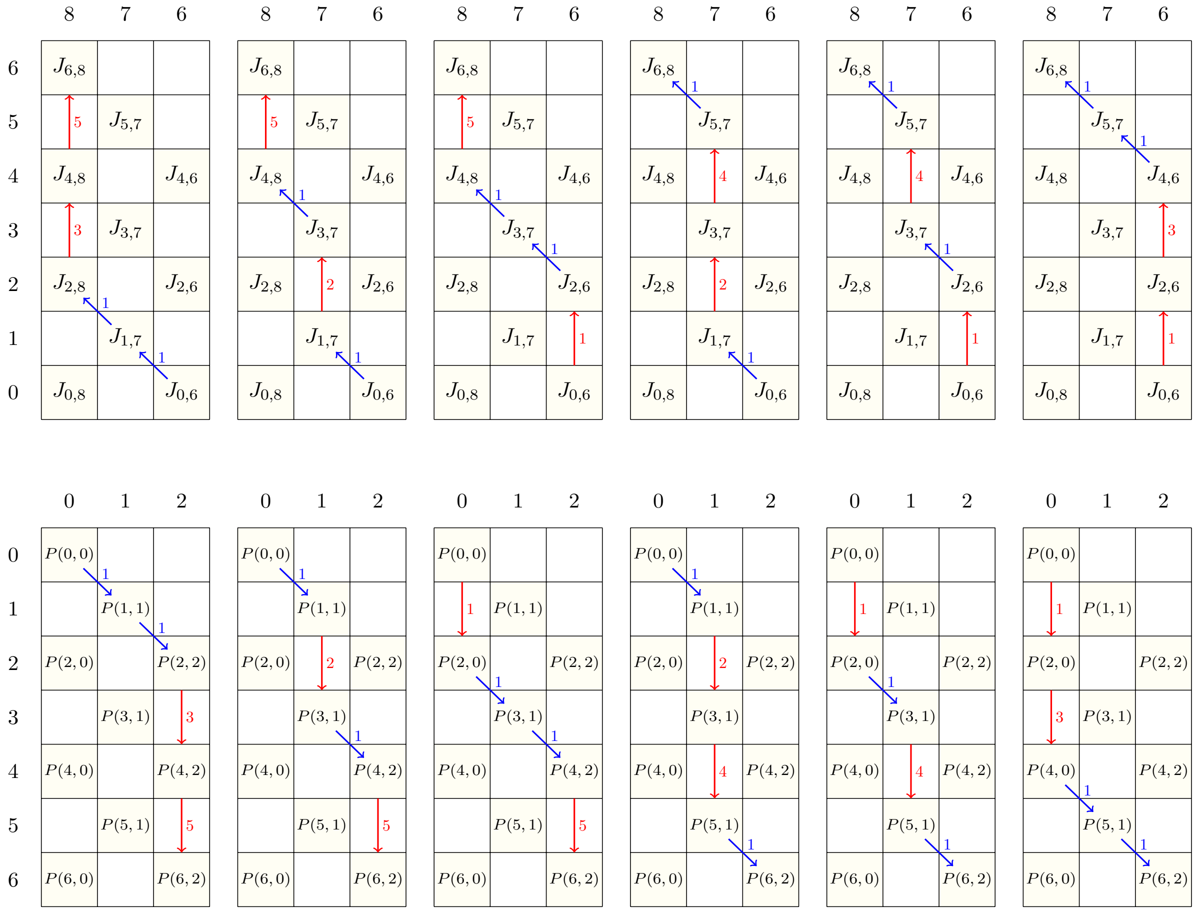


Figure 7: Top: All paths $\pi \in \Gamma_{J^*}(6,8,0,6)$. Bottom: All paths $\Lambda(\pi) \in \Gamma_P(6,8,0,6)$. The paths are lined up so that for each path π in the top row, $\Lambda(\pi)$ appears in the bottom row.

Appendix C. Recursions for the P and Q numbers - Proof of Lemma 12

Earlier work established the following properties of the P and Q numbers.

Theorem 24 (Kreinin (2016)) *The elements of the matrices P and Q satisfy*

$$\begin{aligned} b \cdot P(a,b) &= a \cdot P(a-1,b-1), && a \geq 1, 1 \leq b \leq a, && (61)\\ P(a+1,b) &= P(a,b-1) + (b+1) \cdot P(a,b+1), && a \geq 0, 1 \leq b \leq a, && (62)\\ Q(a,b) &= P(a,b) + (b+1) \cdot Q(a-1,b+1), && a \geq 1, 1 \leq b \leq a, && (63)\\ Q(a,b) &= a \cdot Q(a-2,b) + Q(a-1,b-1), && a \geq 2, 1 \leq b \leq a. && (64) \end{aligned}$$

Proof *[Of Lemma 12] Equation (62) tells us that $P(a,b) = P(a-1,b-1)+(b+1)\cdot P(a-1,b+1)$ for $a \geq 1, 1 \leq b \leq a-1$, while equation (61) tells us that $P(a-1,b+1) = \frac{(a-1)}{(b+1)} \cdot P(a-2,b)$ for $a \geq 2, 0 \leq b \leq a-2$. Putting these together, we get the following recurrence equation for $P(a,b)$:*

$$\begin{aligned} P(a,b) &= P(a-1,b-1) + (b+1)\left(\frac{(a-1)}{(b+1)} \cdot P(a-2,b)\right)\\ &= (a-1) \cdot P(a-2,b) + P(a-1,b-1), \end{aligned}$$

which holds for $a \geq 3, 1 \leq b \leq a-2$. Further, looking at equation (64), we see that the recursion for the Q numbers is very similar to that of the P numbers, but with a coefficient of a rather than $(a-1)$. This establishes Lemma 12. ■

Appendix D. Details of Network Architectures

This section details the architectures of the 45 different network architectures used to produce Figure 3.

#	Depth	Avg.	# Parameters		Avg. Test Accuracy ± Standard Deviation		
		Width	(F)MNIST	CIFAR	MNIST	FMNIST	CIFAR
1	2	50	58880	165790	0.924 ± 0.007	0.79 ± 0.02	0.211 ± 0.029
2	2	85	57350	135510	0.837 ± 0.051	0.709 ± 0.028	0.276 ± 0.011
3	2	200	19930	54250	0.878 ± 0.009	0.721 ± 0.098	0.163 ± 0.048
4	2	25	138300	201600	0.94 ± 0.004	0.812 ± 0.009	0.229 ± 0.025
5	2	125	31725	88925	0.89 ± 0.005	0.768 ± 0.013	0.199 ± 0.027
6	3	25	43990	114550	0.928 ± 0.008	0.812 ± 0.013	0.167 ± 0.022
7	3	50	62830	173280	0.916 ± 0.002	0.79 ± 0.012	0.224 ± 0.019
8	3	100	59700	96756	0.952 ± 0.004	0.839 ± 0.003	0.27 ± 0.016
9	3	67.67	87200	309900	0.924 ± 0.006	0.799 ± 0.011	0.281 ± 0.011
10	3	50	17310	189100	0.553 ± 0.181	0.599 ± 0.119	0.263 ± 0.022
11	4	30	369400	366150	0.877 ± 0.052	0.757 ± 0.026	0.192 ± 0.029
12	4	75	99400	105060	0.957 ± 0.003	0.842 ± 0.006	0.23 ± 0.025
13	5	21	74700	51630	0.931 ± 0.005	0.811 ± 0.009	0.146 ± 0.029
14	6	55	8840	976400	0.715 ± 0.088	0.569 ± 0.146	0.337 ± 0.008
15	6	87.5	169400	398200	0.949 ± 0.008	0.833 ± 0.007	0.332 ± 0.018
16	10	10	79020	180010	0.951 ± 0.003	0.832 ± 0.01	0.278 ± 0.018
17	10	100	64850	122050	0.939 ± 0.004	0.824 ± 0.008	0.262 ± 0.059
18	10	200	54170	262060	0.933 ± 0.005	0.81 ± 0.014	0.335 ± 0.016
19	10	17.5	49920	1002300	0.794 ± 0.052	0.648 ± 0.106	0.184 ± 0.026
20	11	34.55	518800	31720	0.955 ± 0.006	0.835 ± 0.011	0.14 ± 0.037

Table 7: Summary of the architectures of the first 20 neural networks used in Figure 3, as well as their performance on the test datasets. Note that the number of parameters differs between the (F)MNIST and CIFAR-10 datasets due to the fact that CIFAR-10 images are in colour requiring 3 colour channels, while the MNIST and FMNIST images are in grayscale. This table is continued in Table 8.

#	Depth	Avg.	# Parameters		Average Score ± Standard Deviation		
		Width	(F)MNIST	CIFAR	MNIST	FMNIST	CIFAR
21	11	35	21100	269195	0.93 ± 0.005	0.823 ± 0.007	0.363 ± 0.016
22	13	42	36420	328200	0.91 ± 0.008	0.789 ± 0.01	0.364 ± 0.016
23	15	30	41844	174100	0.92 ± 0.004	0.805 ± 0.011	0.349 ± 0.015
24	15	50	13860	235650	0.909 ± 0.005	0.8 ± 0.012	0.328 ± 0.02
25	15	75	16580	206848	0.927 ± 0.003	0.823 ± 0.007	0.359 ± 0.009
26	16	35	42200	159100	0.943 ± 0.004	0.838 ± 0.004	0.343 ± 0.021
27	16	22.5	198800	656400	0.963 ± 0.003	0.845 ± 0.01	0.37 ± 0.016
28	20	25	94900	323700	0.955 ± 0.002	0.843 ± 0.006	0.367 ± 0.006
29	20	50	60416	62340	0.951 ± 0.003	0.837 ± 0.005	0.163 ± 0.058
30	20	37.5	44700	156600	0.948 ± 0.003	0.834 ± 0.008	0.346 ± 0.028
31	23	31.30	194550	598200	0.927 ± 0.005	0.788 ± 0.008	0.17 ± 0.004
32	25	15	64050	48180	0.951 ± 0.002	0.84 ± 0.004	0.186 ± 0.071
33	25	75	55160	125880	0.899 ± 0.014	0.748 ± 0.033	0.274 ± 0.048
34	25	150	53760	64390	0.782 ± 0.077	0.676 ± 0.064	0.206 ± 0.041
35	28	35.71	74715	78300	0.953 ± 0.001	0.844 ± 0.001	0.244 ± 0.075
36	30	15	60860	152380	0.819 ± 0.08	0.719 ± 0.033	0.17 ± 0.02
37	30	30	18630	145280	0.862 ± 0.08	0.772 ± 0.017	0.168 ± 0.02
38	30	100	34360	146680	0.941 ± 0.003	0.826 ± 0.009	0.165 ± 0.022
39	30	26.67	659100	118560	0.932 ± 0.014	0.785 ± 0.011	0.175 ± 0.007
40	30	31.67	18435	52755	0.313 ± 0.131	0.349 ± 0.109	0.158 ± 0.026
41	35	40	86160	276600	0.753 ± 0.074	0.586 ± 0.11	0.148 ± 0.029
42	35	75	250800	450525	0.725 ± 0.163	0.608 ± 0.077	0.165 ± 0.007
43	40	50	137200	251600	0.522 ± 0.141	0.513 ± 0.089	0.167 ± 0.007
44	40	75	278925	422400	0.467 ± 0.123	0.466 ± 0.09	0.161 ± 0.022
45	50	50	162200	177680	0.242 ± 0.064	0.22 ± 0.042	0.161 ± 0.019

Table 8: Continuation of Table 7 for networks 21 through 45.

#	Hidden Layer Widths
1	50, 50
2	85, 85
3	200, 200
4	20, 30
5	100, 150
6	25, 25, 25
7	50, 50, 50
8	100, 100, 100
9	64, 75, 64
10	75, 50, 25
11	40, 40, 20, 20
12	50, 100, 100, 50
13	15, 15, 15, 30, 30
14	80, 70, 60, 50, 40, 30
15	25, 50, 75, 100, 125, 150
16	10, 10, 10, 10, 10, 10, 10, 10, 10, 10
17	100, 100, 100, 100, 100, 100, 100, 100, 100, 100
18	200, 200, 200, 200, 200, 200, 200, 200, 200, 200
19	20, 20, 20, 20, 20, 15, 15, 15, 15, 15
20	55, 30, 30, 30, 30, 30, 30, 30, 30, 30, 55
21	40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30
22	24, 27, 30, 33, 36, 39, 42, 45, 48, 51, 54, 57, 60
23	30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30
24	50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50
25	75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75

Table 9: Ordered list of hidden layer widths for the first 25 networks used in Figure 3. This table is continued in Table 10.

#	Hidden Layer Widths
26	50, 48, 46, 44, 42, 40, 38, 36, 34, 32, 30, 28, 26, 24, 22, 20
27	15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30
28	25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25
29	50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50
30	45, 45, 45, 45, 45, 40, 40, 40, 40, 40, 35, 35, 35, 35, 35, 30, 30, 30, 30, 30
31	40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20
32	15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15
33	75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75
34	150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 150
35	25, 25, 25, 25, 50, 50, 50, 50, 25, 25, 25, 25, 50, 50, 50, 50, 25, 25, 25, 25, 50, 50, 50, 50, 25, 25, 25, 25
36	15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15
37	30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30
38	100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100
39	40, 40, 40, 40, 40, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 40, 40, 40, 40, 40
40	40, 40, 40, 40, 40, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30
41	40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40
42	75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75
43	50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50
44	75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75
45	50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50

Table 10: Continuation of Table 9 for networks 26 through 45.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. *Software available from tensorflow.org.*

Benny Avelin and Anders Karlsson. Deep limits and a cut-off phenomenon for neural networks. Journal of Machine Learning Research, *23(191):1–29, 2022. URL* `http://jmlr.org/papers/v23/21-0431.html`.

Sam Buchanan, Dar Gilboa, and John Wright. Deep networks and the multiple manifold problem. In International Conference on Learning Representations, *2021. URL* `https://openreview.net/forum?id=O-6Pm_d_Q-`.

Gi-Sang Cheon, Ji-Hwan Jung, and Louis W. Shapiro. Generalized Bessel numbers and some combinatorial settings. Discrete Mathematics, *313(20):2127–2138, 2013. ISSN 0012-365X.*

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

Li Deng. The MNIST database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, *29(6):141–142, 2012.*

Benoit Dherin, Michael Munn, Mihaela Rosca, and David GT Barrett. Why neural networks find simple solutions: The many regularizers of geometric complexity. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, Advances in Neural Information Processing Systems, *2022. URL* `https://openreview.net/forum?id=-ZPeUAJlkEu`.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In Annual Conference Computational Learning Theory, *2015.*

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: a survey. J. Mach. Learn. Res., *20(1):1997–2017, jan 2019. ISSN 1532-4435.*

Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, *volume 31. Curran*

Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/13f9896df61279c928f19721878fac41-Paper.pdf`.

Boris Hanin. Random fully connected neural networks as perturbatively solvable hierarchies, 2023. URL `https://arxiv.org/abs/2204.01058`.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In International Conference on Machine Learning*, 2019.*

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.*

Alexander Kreinin. Integer sequences connected to the Laplace continued fraction and Ramanujan's identity. Journal of Integer Sequences*, 19:1–12, 06 2016.*

Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Mufan Bill Li, Mihai Nica, and Daniel M. Roy. The neural covariance SDE: Shaped infinite depth-and-width networks at initialization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, Advances in Neural Information Processing Systems*, 2022. URL* `https://openreview.net/forum?id=WG3vmsteqR_`.

James Martens, Andy Ballard, Guillaume Desjardins, Grzegorz Swirszcz, Valentin Dalibard, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping. CoRR*, 2021. URL* `https://arxiv.org/abs/2110.01765`.

Ido Nachum, Jan Hazla, Michael Gastpar, and Anatoly Khina. A Johnson-Lindenstrauss framework for randomly initialized CNNs. In International Conference on Learning Representations*, 2022. URL* `https://openreview.net/forum?id=YX0lrvdPQc`.

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

Daniel A. Roberts, Sho Yaida, and Boris Hanin. The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press, 2022. doi: 10.1017/9781009023405.*

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In International Conference on Learning Representations*, 2017. URL* `https://openreview.net/forum?id=H1W1UN9gg`.

Roman Vershynin. High-Dimensional Probability: An Introduction with Applications in Data Science. *Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN 978-1-108-41519-4.*

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017. URL `http://arxiv.org/abs/1708.07747`.