The Angle Degeneracy Phenomenon in Deep Neural Networks: Analysis and Relation to Training Dynamics

by
Cameron Jakub

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master's of Science
in
Mathematics & Statistics
(Collaborative Specialization in Artificial Intelligence)

Guelph, Ontario, Canada

ABSTRACT

THE ANGLE DEGENERACY PHENOMENON IN DEEP NEURAL NETWORKS: ANALYSIS AND RELATION TO TRAINING DYNAMICS

Cameron Jakub
University of Guelph, 2023

Advisor:
Dr. Mihai Nica

Deep neural networks have proven to be powerful functions with many applications, but the theoretical behaviour of these functions is not fully understood. One such behaviour is the large depth degeneracy phenomenon, where inputs tend to become highly correlated as they travel deeper into a randomly initialized network. This can make the network effectively incapable of distinguishing between inputs, which has negative impacts on training performance. Through combinatorial expansions, we develop precise formulas to predict the expected value and variance of the angle between inputs at any layer of the initialized network. We provide a detailed analysis of how quickly the angle tends toward zero in a finite width setting, which proves to be qualitatively different than studying the problem in the infinite width limit. We validate our theoretical results through comparison to empirical simulations, and run experiments to explore how network degeneracy can impact training dynamics.

Dedication

This thesis is dedicated to my parents Mike and Donna. Love you guys! ◡̈

Acknowledgements

First and foremost, I would like to offer my deepest thanks to my advisor, Dr. Mihai Nica, for all of the time and effort he put into supporting me. Thank you for being an excellent teacher, providing valuable insight and mentorship, and creating a fun and comfortable learning environment for me. I've grown so much as a student and a researcher from your guidance. I would also like to thank Dr. Graham Taylor for all the work he put into being on my advisory committee and for providing me with valuable feedback. I would also like to express my gratitude to Dr. Rajesh Pereira for his work on the examination committee, being a fantastic teacher throughout my undergrad, and an excellent supervisor on the multiple research projects we worked on.

Thank you to Dr. Matt Demers and Dr. Kim Levere for their exceptional teaching, putting their trust in me to teach course content, and for always being a friend to chat with. I am particularly thankful for Dr. Steve Gismondi, whose infectious enthusiasm for math was the catalyst for me to pursue a math degree in the first place. Along with the aforementioned faculty, I also want to extend thanks to the following professors who were especially impactful during my time at Guelph due to their teaching style and mentorship: Dr. Hermann Eberl, Dr. Ayesha Ali, and Dr. Herb Kunze.

My degree would not have been such a positive experience without my amazing support network. Thank you to my parents, Mike and Donna, and my sisters Taryn, Kristen, and Kaela for always being there for me. I also want to thank some of my closest friends Jake, Luke, Matt, Dave, Grace, Emma, and Braydon. Thanks for the laughs, support, and generally making this degree a whole lot of fun. Further, thank you to Kate and Arya for being kind, supportive friends, and providing me with a welcoming place to stay during the final months of my degree.

Last, I want to extend a big thank you to Eat Thai in downtown Guelph for consistently making the best fried rice and pad thai in Guelph, keeping me fueled throughout my degree.

Contents

Abstract ii

Dedication iii

Acknowledgements iv

List of Tables vii

List of Figures viii

1 Introduction 1
- 1.1 The Large Depth Degeneracy Phenomenon in Neural Networks 1
- 1.2 Outline 3
- 1.3 Main Results for the Angle Process 4
- 1.4 Introduction to the J Functions 9

2 ReLU Neural Networks on Initialization 12
- 2.1 Expected Value Calculation 14
- 2.2 Variance Calculation 16

3 Network Degeneracy as an Indicator of Training Performance 17
- 3.1 Comparison to Infinite Width Networks 18

4 Explicit Formula for Mixed-Moment J Functions 20
- 4.1 Statement of Main Results and Outline of Method 20
- 4.2 Gaussian Integration-by-Parts Formulas 23
- 4.3 Recursive Formulas for $J_{a,b}(\theta)$ - Proof of Proposition 1 26
- 4.4 Solving the Recurrence to get an Explicit Formula for $J_{a,b}(\theta)$ - Proof of Theorem 2 27

5 Conclusion And Further Work 33

Bibliography 34

A Appendix A **38**
A.1 Expected Value Approximation 38
A.2 Variance Approximation 39
A.3 Covariance Approximation 40
A.4 Third and Fourth Moment Bound Lemma 40
A.5 Expected Value Calculations 43
A.5.1 Calculation of $\mathbf{E}\left[R^{\ell+1}\right]$ 43
A.5.2 Calculation of $\mathbf{E}\left[R^{\ell+1}\sin^2(\theta^{\ell+1})\right]$ 44
A.6 Variance and Covariance Calculations 44
A.6.1 Calculation of $\mathbf{Var}\left[R^{\ell+1}\right]$ 46
A.6.2 Calculation of $\mathbf{Var}\left[R^{\ell+1}\sin^2(\theta^{\ell+1})\right]$ 46
A.6.3 Calculation of $\mathbf{Cov}\left(R^{\ell+1}\sin^2(\theta^{\ell+1}), R^{\ell+1}\right)$ 47
A.7 Derivation of Useful Identities - Equations 2.1, 2.2 48
A.8 Cauchy-Binet and Determinant of the Gram Matrix - Equation 2.3 51
A.9 Infinite Width Update Rule 51

B Appendix B **53**
B.1 Derivation of Lower-Order J Functions - Proof of Proposition 2 53
B.2 Proof of Explicit Formulas for $J_{n,0}$ and $J_{n,1}$ 54
B.3 Bijection between Paths in Graphs of J Functions and the Bessel Number Graphs P,Q 57

C Appendix C **60**
C.1 P and Q numbers 60
C.2 Recursions for the P and Q numbers - Proof of Lemma 3 60

D Appendix D **62**

List of Tables

1.1 Low Order J Functions 11

2.1 Notation for Fully Connected ReLU Networks 13

A.1 Third Moment Bound Lemma: Irreducible Patterns 43
A.2 Variance of $R^{\ell+1}$ Calculation 47
A.3 Variance of $R^{\ell+1}\sin^2(\theta^{\ell+1})$ Calculation 48
A.4 $\mathbf{Cov}\left(R^{\ell+1}\sin^2(\theta^{\ell+1}), R^{\ell+1}\right)$ Calculation 49

D.1 Network Architecture Summary: Networks 1-20 62
D.2 Network Architecture Summary: Networks 21-45 63
D.3 List of Hidden Layer Widths: Networks 1-25 64
D.4 List of Hidden Layer Widths: Networks 26-45 65

List of Figures

1.1 Infinite Width Comparison and Monte Carlo Simulations 5
1.2 Mean and Variance Plots 8

3.1 Comparison of Network Degeneracy to Training Performance 18
3.2 Finite Versus Infinite Width Angle Prediction Comparison 19

4.1 Comparison of J and J^* Graphs 29
4.2 Graphs of the P, Q, and J Recursions 30

B.1 Bijection between Paths in J^* and P Graphs 58
B.2 Bijection between Paths in J^* and Q Graphs 59

Introduction

1.1 The Large Depth Degeneracy Phenomenon in Neural Networks

Creating *deep* neural networks by stacking many layers has achieved exceptional performance in many applications and contributed to the recent explosion of these methods. It has been shown that the "power" of a network's ability to approximate functions comes from the *depth* of the network, rather than the width. In fact, Poole et al. [21] and Eldan and Shamir [9] have shown that depth can *exponentially* improve the expressibility of a network. Specifically, Eldan and Shamir [9] proved that there exists a function which cannot be approximated by any two-layer feed forward neural network, but simply adding a layer to create a 3-layer network allows one to approximate the function.

These findings may suggest that creating deeper networks is advantageous. However, this is not always the case. As networks become deeper, they also are more susceptible to becoming *degenerate*. This concept of degeneracy can be observed in multiple ways. One sense in which a network can be considered degenerate is the concept of vanishing and exploding gradients [10]. Hanin [10] studied this phenomenon in feed-forward ReLU networks, and found that certain network architectures can cause the gradients of the network to vary wildly on initialization. The stability of the gradients depends on the sum of the inverse layer widths, where a larger sum corresponds to less stable gradients. Therefore, networks with more layers can be more susceptible to unstable gradients. The vanishing and exploding gradient problem poses a challenge to network training, and is an example of how deeper networks do not necessarily correspond to better prediction.

Another sense in which feed forward neural networks can become degenerate is that as inputs travel through a randomly initialized network, they tend to become more and more correlated (i.e. the angle between inputs tends toward 0 as the number of layers tends toward

infinity). Therefore, if an initialized network has too many layers, it may send all inputs to effectively the same output, making the network incapable of differentiating between any two inputs fed into it. This is the type of degeneracy we study in this thesis, which has been observed by many authors from different angles [2, 3, 8, 12, 18, 20, 23]. For example, Schoenholz et al. [23] found that there is a maximum depth for which information can propagate in random neural networks. They reason that when information is able to properly propagate through the networks, the networks are able to train precisely. However, there exists a maximum depth for which networks can be properly trained. Similarly, Hayou et al. [12] observed that an improper choice of activation can cause loss of information during the forward pass of training. Avelin and Karlsson [2] noticed the degeneracy as a "cut off" behaviour, for which after a certain depth, networks begin to behave very differently. Nachum et al. [20] even observed this phenomenon in convolutional neural networks, finding that the level of degeneracy was dependent on the type of input being fed into the networks.

Previous works have developed strategies to combat degeneracy in deep feed forward neural networks. For ReLU networks, Li et al. [18] demonstrated how *shaping* the ReLU activation function can prevent the angle between inputs from becoming trivially small as they travel deeper into the network. They shape the ReLU function by tweaking the slopes of the piecewise linear segments. Letting s_- represent the slope for negative inputs, and s_+ represent the slope for positive inputs, they define the "leaky" ReLU activation function $\varphi^* : \mathbb{R} \to \mathbb{R}$ as $\varphi^*(x) = s_+ \max(x, 0) + s_- \min(x, 0)$. Rather than shaping the activation function, other authors have shown that modifying the architecture of the network itself can preserve the variation between inputs. ResNets [14] have shown that strategically introducing skip connections (connections between non-adjacent layers) can allow very deep networks to train properly. Similarly, Srivastava et al. [24] introduced Highway Networks, where "highway layers" in the architecture can behave as a blend between a typical feed-forward layer and a layer which simply passes inputs through to the next layer, which reduces information loss over many layers. Martens et al. [19] introduced the method of "deep kernel shaping" to prevent deep networks from becoming degenerate. They suggest a strategy which involves function transformations on the activation function, precise parameter initialization, and alterations to the architecture of the network itself.

Further, previous studies have provided analyses for how the angle between inputs evolves in *infinite width* networks (i.e. a network studied in the limit that all layer widths tend towards infinity) [11, 12, 22, 23]. Studying the angle evolution in the infinite width limit suggests that the angle between inputs goes toward 0 polynomially fast. The infinite width

angle prediction uses the law of large numbers and thereby disregards any random fluctuations in $\theta^{\ell+1}$ given θ^{ℓ}. These random fluctuations, though small, can accumulate over many layers leading to inaccurate predictions for finite width networks.

The depth degeneracy phenomenon in ReLU networks has been observed in the past, and the angle evolution in neural networks has been studied in the infinite width case, but there has not yet been a study which provides a thorough analysis of the angle evolution for finite width networks with the ordinary, unshaped ReLU function. This thesis provides a detailed analysis of this problem, and provides a comparison of our finite width angle prediction to the infinite width prediction developed in previous studies.

1.2 Outline

There are two main theoretical contributions of this thesis. The first is to prove Theorem 1, which describes the angle evolution between inputs in terms of a collection of mixed-moment "J" functions, denoted $J_{a,b}$, for $a, b \in \mathbb{N}$. The second theoretical contribution is to separately derive an explicit formula for any of the mixed moments $J_{a,b}$, given in Theorem 2. With these theoretical results, we also run simulations to study the relationship between the predicted level of degeneracy in a network, and the network's training performance on multiple datasets.

Section 1.3 covers the main results for the angle process analysis in deep ReLU networks. Given the angle between inputs at layer ℓ, Theorem 1 introduces accurate formulas to predict the mean and variance of the angle at layer $\ell + 1$. Theorem 1 is simplified into a convenient finite width update rule for the angle in Approximation 1. We also compare our finite width update rule to the infinite width rule, given in Approximation 3. Section 1.4 provides an introduction to the J functions, which are essential for studying the angle evolution in a finite width setting. A more detailed analysis of the J functions is given in Chapter 4.

Chapter 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 $\mathbf{E}[\ln(\sin^2(\theta^{\ell+1}))]$ given θ^{ℓ}, which leads to the finite width update rule for θ^{ℓ} as in equation (1.1), while Section 2.2 outlines our approximation for $\mathbf{Var}[\ln(\sin^2(\theta^{\ell+1}))]$.

Chapter 3 uses the results from Chapter 2 to explore how network degeneracy can affect training performance. We run simulations which compare the predicted final angle between inputs to the accuracy of classification on the MNIST [7], Fashion-MNIST [25], and CIFAR-10 [17] datasets. In Section 3.1, we demonstrate the advantages of using our finite-width

prediction rule over the infinite width prediction rule.

In Chapter 4, we cover the derivation of the explicit formula for the J functions. We state the main results of this section in Section 4.1, and cover the mathematical tools needed to solve the expectations using Gaussian integration by parts in Section 4.2. We develop the formula for $J_{a,b}$ by first finding a recursive formula in Section 4.3, 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 4.4.

1.3 Main Results for the Angle Process

In this thesis, 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 $\mathbb{R}^{n_\ell}$.

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

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" [12, 23], where the variance of each layer is order one as layer width increases. See Table 2.1 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 1.1 for an illustration). This indicates that the hidden layer representation of *any* two inputs in a deep neural network becomes closer and closer to co-linear as depth increases.

In this thesis, 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.

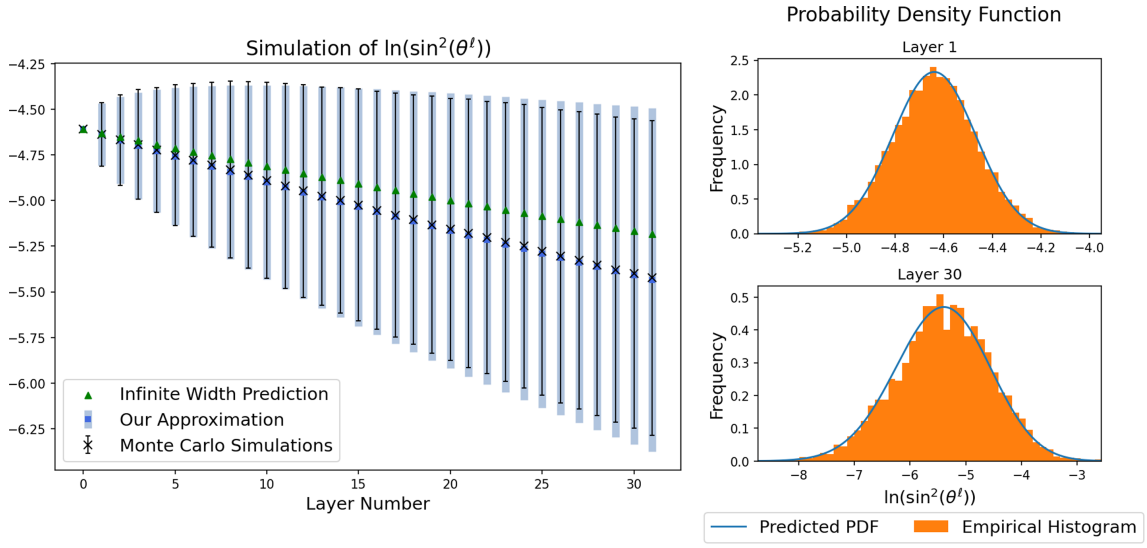


Figure 1.1: We feed 2 inputs with initial angle $\theta^0 = 0.1$ into 5000 Monte Carlo samples of independently initialized networks with network width $n_\ell = 256$ for all layers. Left: Using the Monte Carlo samples, we plot the empirical mean and standard deviation of $\ln(\sin^2(\theta^\ell))$ at each layer. We compare this to both the infinite width update rule and our prediction using Approximation 1 for the mean of $\ln(\sin^2(\theta^\ell))$. Our prediction for the standard deviation in each layer using Approximation 2 is also plotted as the shaded area. 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 (1.10) at Layer 1 (top) and Layer 30 (bottom). 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 **link**.

Approximation 1 (Finite Width Update Rule)**.** *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*

$$\ln \sin^2(\theta^{\ell+1}) \approx \ln \sin^2(\theta^\ell) - \frac{2}{3\pi}\theta^\ell - \rho(n_\ell), \tag{1.1}$$

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

$$\rho(n) := \ln\left(\frac{n+5}{n-1}\right) - \frac{10n}{(n+5)^2} + \frac{6n}{(n-1)^2} = \frac{2}{n} + \mathcal{O}\left(n^{-2}\right). \tag{1.2}$$

Figure 1.1 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 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.

Comparison to Infinite Width Networks

Approximation 1 predicts that $\theta^\ell \to 0$ *exponentially fast* in ℓ due to the term $\rho(n)$; 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 $n_\ell \to \infty$ limit predicts that $\theta^\ell \to 0$ like ℓ^{-1}, which is qualitatively very different! The prediction of this rate was first demonstrated under the name "edge of chaos" [12, 23] and again in Hanin [11], Roberts et al. [22]. 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 $1 - \cos(x) \approx \frac{1}{2}x^2$ as $x \to 0$. The update rule for $\cos(\theta_\ell)$ in the infinite width limit is given in Approximation 3. A derivation for this update rule in our notation is provided in Appendix A.9.

The fact that θ^ℓ decays polynomially fast in infinite width networks compared to exponentially fast in finite width networks means that information is preserved better layer-by-layer in infinite width networks. This highlights an interesting advantage of infinite width networks: They are less susceptible to the depth degeneracy phenomenon in the sense that infinite width architectures can be made deeper than finite width ones before becoming "degenerate".

We can also derive the infinite width prediction from our result by replacing $\rho(n)$ with 0 in the update rule (1.1). Exponentiating both sides and using $\sin(\theta) \approx \theta, e^\theta \approx 1 + \theta$ for $\theta \ll 1$, Approximation 1 becomes $(\theta^{\ell+1})^2 \approx \left(\theta^\ell\right)^2(1 - \frac{2}{3\pi}\theta^\ell)$, which is equivalent to the result of Proposition C.1 of Hanin [11] and is also a corollary of Lemma 1 of Hayou et al. [12]. In those papers, the rule was derived directly from the infinite width update rule for $\cos(\theta)$, and those results are equivalent to the fact that $\theta^\ell \approx \ell^{-1}$ 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.

More Detailed Results for the Mean and Variance

Approximation 1 is derived 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, the mean and variance of $\ln\sin^2(\theta^{\ell+1})$ obey the following limit as the layer width $n_\ell \to \infty$*

$$\mathbf{E}[\ln\sin^2(\theta^{\ell+1})] = \mu(\theta^\ell, n_\ell) + \mathcal{O}(n_\ell^{-2}), \quad \mathbf{Var}[\ln\sin^2(\theta^{\ell+1})] = \sigma^2(\theta^\ell, n_\ell) + \mathcal{O}(n_\ell^{-2}), \tag{1.3}$$

$$\begin{aligned}\mu(\theta, n) :=& \ln\left(\frac{(n-1)(1-4J_{1,1}^2)}{4J_{2,2}-1+n}\right) + \frac{4(J_{2,2}+1)}{n\left(\frac{4J_{2,2}-1}{n}+1\right)^2} \\ &- \frac{4\left(8J_{1,1}^2J_{2,2} - 8J_{1,1}^4 + 4J_{1,1}^2 - 8J_{1,1}J_{3,1} + J_{2,2} + 1\right)}{n\left(1-\frac{1}{n}\right)^2\left(1-4J_{1,1}^2\right)^2},\end{aligned} \tag{1.4}$$

$$\begin{aligned}\sigma^2(\theta, n) :=& \frac{8n(J_{2,2}+1)}{(4J_{2,2}-1+n)^2} + \frac{8n(8J_{1,1}^2J_{2,2} - 8J_{1,1}^4 + 4J_{1,1}^2 - 8J_{1,1}J_{3,1} + J_{2,2} + 1)}{(n-1)^2(1-4J_{1,1}^2)^2} \\ &- \frac{16n(2J_{1,1}^2 - 4J_{1,1}J_{3,1} + J_{2,2} + 1)}{(4J_{2,2}-1+n)(n-1)(1-4J_{1,1}^2)},\end{aligned} \tag{1.5}$$

where $\mathbf{E}, \mathbf{Var}$ denote the conditional mean and variance of quantities in layer $\ell+1$ given the value of θ^ℓ in the previous layer and $J_{a,b} := J_{a,b}(\theta^\ell)$ are the joint moments of correlated Gaussians passed through the ReLU function $\varphi(x) = \max\{x, 0\}$, namely

$$J_{a,b}(\theta) := \mathbf{E}_{G,\hat{G}}[\varphi^a(G)\varphi^b(\hat{G})], \tag{1.6}$$

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 4. 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 2. 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 1 (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*

$n_\ell \to \infty$

$$\mathbf{E}[\ln(\sin^2(\theta^{\ell+1}))] = \mu(\theta^\ell, n_\ell) + \mathcal{O}(n_\ell^{-2}), \quad \mathbf{Var}[\ln(\sin^2(\theta^{\ell+1}))] = \sigma^2(\theta^\ell, n_\ell) + \mathcal{O}(n_\ell^{-2}), \tag{1.7}$$

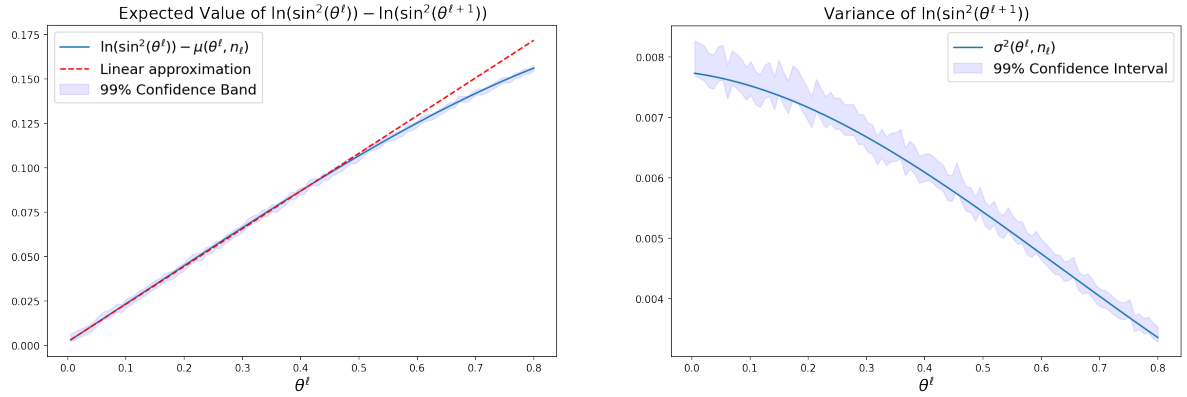
$$\mu(\theta, n) = \ln(\sin^2\theta) - \frac{2}{3\pi}\theta - \rho(n) - \frac{8\theta}{15\pi n} - \left(\frac{2}{9\pi^2} - \frac{68}{45\pi^2 n}\right)\theta^2 + \mathcal{O}(\theta^3), \tag{1.8}$$

$$\sigma^2(\theta, n) = \frac{8}{n} - \frac{64}{15\pi}\frac{\theta}{n} - \left(8 + \frac{296}{45\pi}\right)\frac{\theta^2}{n} + \mathcal{O}\left(\theta^3\right), \tag{1.9}$$

where $\rho(n)$ *is as defined in* (1.2).

To derive Approximation 1 from Theorem 1, we simply keep only the first few terms of the series expansion (1.8), 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 (1.9)). 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 averages over n pairs of independent Gaussian variables (see (2.1-2.3)). Thus, by central-limit-theorem type arguments, one would expect the following approximation by Gaussian laws which also accounts for the variability of $\ln(\sin^2(\theta^\ell))$ using our calculated value for the variance.



(a) Mean as a function of θ

(b) Variance as a function of θ

Figure 1.2: 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 10000 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) $\mathbf{Var}\left[\ln(\sin^2(\theta^{\ell+1}))\right]$.

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

$$\ln(\sin^2(\theta^{\ell+1})) \overset{d}{\approx} \mathcal{N}(\mu(\theta^\ell, n_\ell), \sigma^2(\theta^\ell, n_\ell)), \tag{1.10}$$

where μ, σ^2 are as in Theorem 1.

We find that the normal approximation (1.10) matches simulated finite neural networks remarkably well; see Monte Carlo simulations from real networks in Figure 1.1. 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 Chapter 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.

1.4 Introduction to the J Functions

In 2009, Cho and Saul [5] introduced the p-th moment for correlated ReLU-Gaussians, which they denoted with the letter J,

$$J_p(\theta) := 2\pi \mathbf{E}\left[\varphi^p(G)\varphi^p(\hat{G})\right], \tag{1.11}$$

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 (Infinite Width Update Rule)**.** *In the limit that the width of each layer tends to infinity, the* infinite width approximation *for the angle $\theta^{\ell+1}$ given θ^ℓ is*

$$\cos\left(\theta^{\ell+1}\right) = \frac{J_1(\theta^\ell)}{\pi} = \frac{\sin(\theta^\ell) + (\pi - \theta^\ell)\cos(\theta^\ell)}{\pi}. \tag{1.12}$$

The formula for J_1 is the $p = 1$ case of a remarkable explicit formula for J_p derived by Cho and Saul [5] 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. 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 ([18]). The infinite width limit uses the law of large numbers to obtain (1.12), thereby discarding random fluctuations. For very deep networks, microscopic fluctuations (on the order of $\mathcal{O}(n_\ell^{-1})$) 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 1.1; very deep networks are far from the infinite width limit in this case. See Figure 1.1 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 thesis. While the mean only depends on the p-th moment functions J_p from (1.11), these fluctuations depend on the *mixed* moments, which we denote by $J_{a,b}$ for $a, b \in \mathbb{N}$ as follows[1]

$$J_{a,b}(\theta) := \mathbf{E}\left[\varphi^a(G)\varphi^b(\hat{G})\right], \tag{1.13}$$

with $G, \hat{G}$ again as in (1.11) 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 Chapter 4, 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

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

$J_{a,b}$ and the Bessel numbers ([4]), 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 1.1, and the general explicit formula is presented in Theorem 2.

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

Table 1.1: Table of formulas for the first few J functions. The normalizing constant appearing in all entries, either $c_0 = 2\pi$, $c_1 = 2\sqrt{2\pi}$ depending on the parity of $a+b$, has been omitted, i.e. the table shows the value of $c_{(a+b \bmod 2)}J_{a,b}(\theta)$. (Note that when $a = 0$, the appropriate convention of $0^0 = 0$ is needed, see (4.1) for details). These generalize $J_p(\theta)$ of (1.11) 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 4.4.

ReLU Neural Networks on Initialization

In this chapter, 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 2.1. Note that the factor of $\sqrt{2/n_\ell}$ in our definition implements the so called He initialization [13], which ensures that $\mathbf{E}[\|z^\ell\|^2] = \|x\|^2$ for all layers ℓ. This initialization is known to be the "critical" initialization for taking large limits of the network [12, 22]. 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 are transformed from layer to layer.

The starting point for our calculation is to notice that because the weights are Gaussian, the values of $\varphi_\alpha^{\ell+1}, \varphi_\beta^{\ell+1}$ are jointly Gaussian given the vectors of $\varphi_\alpha^\ell, \varphi_\beta^\ell$. In fact, it turns out that by properties of Gaussian random variables, one only needs to know the values of the scalars $\|\varphi_\alpha^\ell\|, \|\varphi_\beta^\ell\|$ and θ^ℓ to understand the full distribution of $\varphi_\alpha^{\ell+1}, \varphi_\beta^{\ell+1}$. (see Appendix A.7 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.7 and A.8.

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 $\cos(\theta^\ell) := \frac{\langle \varphi^\ell_\alpha, \varphi^\ell_\beta \rangle}{\|\varphi^\ell_\alpha\| \|\varphi^\ell_\beta\|}$
$R^{\ell+1} \in \mathbb{R}$	Shorthand for the ratio $R^{\ell+1} := \frac{\|\varphi^{\ell+1}_\alpha\|^2 \|\varphi^{\ell+1}_\beta\|^2}{\|\varphi^\ell_\alpha\|^2 \|\varphi^\ell_\beta\|^2}$

Table 2.1: Definition and notation used for fully connected ReLU neural networks.

$$\|\varphi^{\ell+1}_\alpha\|^2 = \frac{\|\varphi^\ell_\alpha\|^2}{n_\ell} \sum_{i=1}^{n_\ell} 2\varphi^2(G_i), \qquad \|\varphi^{\ell+1}_\beta\|^2 = \frac{\|\varphi^\ell_\beta\|^2}{n_\ell} \sum_{i=1}^{n_\ell} 2\varphi^2(\hat{G}_i), \tag{2.1}$$

$$\langle \varphi^{\ell+1}_\alpha, \varphi^{\ell+1}_\beta \rangle = \frac{\|\varphi^\ell_\alpha\| \|\varphi^\ell_\beta\|}{n_\ell} \sum_{i=1}^{n_\ell} 2\varphi(G_i)\varphi(\hat{G}_i), \tag{2.2}$$

$$R^{\ell+1} \sin^2(\theta^{\ell+1}) = \frac{2}{n_\ell^2} \sum_{i,j=1}^{n_\ell} \Big(\varphi(G_i)\varphi(\hat{G}_j) - \varphi(G_j)\varphi(\hat{G}_i) \Big)^2, \tag{2.3}$$

where $R^{\ell+1} := \frac{\|\varphi^{\ell+1}_\alpha\|^2 \|\varphi^{\ell+1}_\beta\|^2}{\|\varphi^\ell_\alpha\|^2 \|\varphi^\ell_\beta\|^2}$, and $G_i, \hat{G}_i$ 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 (2.3) is derived using the *determinant of the Gram matrix* for vectors $\varphi^{\ell+1}_\alpha,\ \varphi^{\ell+1}_\beta$ (full derivation given in Appendix A.8).

Combining the equations in (2.1) gives us a useful identity for the ratio $R^{\ell+1}$, namely:

$$R^{\ell+1} = \frac{4}{n_\ell^2} \sum_{i,j=1}^{n_\ell} \varphi^2(G_i)\varphi^2(\hat{G}_j). \tag{2.4}$$

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 $\ln\sin^2(\theta^\ell) \in (-\infty, \infty)$ works well, whereas Gaussian approximations for θ^ℓ or $\cos(\theta^\ell) \in [-1, 1]$ are less accurate.) We first derive a formula for $\mathbf{E}\left[\ln(\sin^2(\theta^{\ell+1}))\right]$.

2.1 Expected Value Calculation

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, using properties of the logarithm, we rewrite this expectation as the difference

$$\mathbf{E}\left[\ln(\sin^2(\theta^{\ell+1}))\right] = \mathbf{E}\left[\ln\left(R^{\ell+1}\sin^2(\theta^{\ell+1})\right)\right] - \mathbf{E}\left[\ln\left(R^{\ell+1}\right)\right]. \tag{2.5}$$

The two random variables $R^{\ell+1}$ and $R^{\ell+1}\sin^2(\theta^{\ell+1})$ in (2.5) both have interpretations in terms of sums of Gaussians as in (2.3) and (2.4) 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):

$$\ln(X) = \ln(\mathbf{E}[X]) + \frac{X - \mathbf{E}[X]}{\mathbf{E}[X]} - \frac{(X - \mathbf{E}[X])^2}{2\mathbf{E}[X]^2} + \epsilon_2\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right), \tag{2.6}$$

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 (2.5),

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 (2.1 - 2.3). 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 (2.1 - 2.3). The key quantities we calculate are

$$\mathbf{E}\left[R^{\ell+1}\right] = \frac{4J_{2,2}-1}{n_\ell} + 1, \tag{2.7}$$

$$\mathbf{Var}\left[R^{\ell+1}\right] = \frac{4}{n_\ell}(J_{2,2}+1) + \frac{16}{n_\ell^2}\left(2J_{4,2} - \frac{5}{2}J_{2,2} + J_{2,2}^2 + \frac{5}{8}\right) + \mathcal{O}\left(n_\ell^{-3}\right), \tag{2.8}$$

$$\mathbf{E}\left[R^{\ell+1}\sin^2(\theta^{\ell+1})\right] = \frac{(n_\ell-1)(1-4J_{1,1}^2)}{n_\ell}, \tag{2.9}$$

$$\mathbf{Var}\left[R^{\ell+1}\sin^2(\theta^{\ell+1})\right] = \frac{8\left(-8J_{1,1}^4 + 8J_{1,1}^2J_{2,2} + 4J_{1,1}^2 - 8J_{1,1}J_{3,1} + J_{2,2} + 1\right)}{n_\ell} + \mathcal{O}\left(n_\ell^{-2}\right), \tag{2.10}$$

where $J_{a,b} = J_{a,b}(\theta^\ell)$. These formulas are calculated in Appendices A.5 and A.6 by a combinatorial expansion using the representations from (2.1-2.3). 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 (2.8, 2.10).

2.2 Variance Calculation

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:

$$\mathbf{Var}[\ln(\sin^2(\theta^{\ell+1}))] = \mathbf{Var}\left[\ln\left(R^{\ell+1}\sin^2(\theta^{\ell+1})\right) - \ln\left(R^{\ell+1}\right)\right] \tag{2.11}$$
$$= \mathbf{Var}\left[\ln\left(R^{\ell+1}\sin^2(\theta^{\ell+1})\right)\right] + \mathbf{Var}\left[\ln\left(R^{\ell+1}\right)\right] - 2\,\mathbf{Cov}\left(\ln\left(R^{\ell+1}\sin^2(\theta^{\ell+1})\right), \ln\left(R^{\ell+1}\right)\right).$$

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 (2.1 - 2.3) in our calculations. Appendices A.2 and 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 (2.1 - 2.3). 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 (2.1-2.3). (Full calculation given in Appendix A.6). 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 (1.5). 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 1.1.

Network Degeneracy as an Indicator of Training Performance

This chapter uses the theoretical results derived thus far as an input into experiments that investigate how the level of degeneracy can influence training. 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 Angle prediction between inputs for a feed-forward ReLU network with depth L and layer widths n_ℓ, $1 \leq \ell \leq L$. The function $\mu(\theta, n)$ is given in Theorem 1.

$\theta^0 =$ angle between inputs
for $\ell = 0, \ldots, L-1$ **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 $= \theta^L$

Algorithm 1 predicts the angle at the final layer on initialization based solely on the network architecture $n_1, n_2, \ldots n_L$. If all inputs into an initialized network tend to be highly correlated by the final layer, this could make it difficult for the network to distinguish the differences between inputs and therefore harder to train. Figure 3.1 demonstrates how networks which exhibit this type of degeneracy empirically tend to perform worse *after training*, and seem to train less consistently than networks which can better distinguish between inputs on initialization.

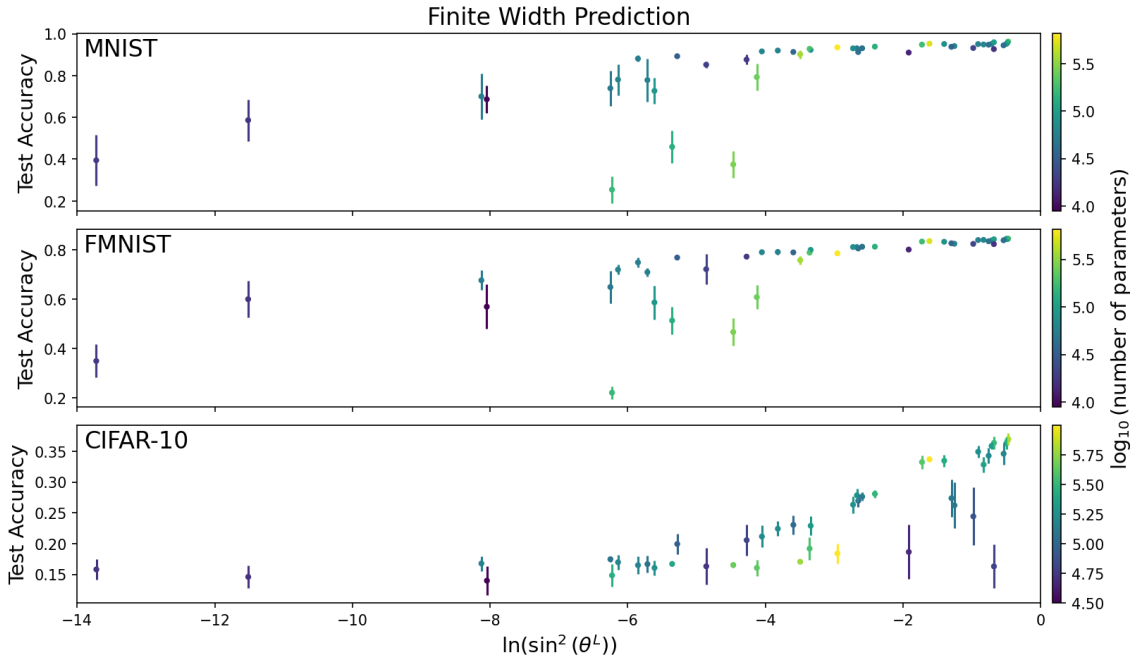


Figure 3.1: We compare 45 different network architectures trained on the MNIST [7], Fashion-MNIST [25], and CIFAR-10 [17] datasets 10 times each. Using the architecture of the network and Algorithm 1, we predict the angle between 2 orthogonal inputs at the final output layer of the network on initialization. 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. 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 [1]. See Appendix D for details on all of the network architectures used. The code which produced this figure can be found at the following **link**.

When Algorithm 1 predicts that the network architecture forces inputs to become highly correlated on initialization, this serves a warning that the network may train poorly. 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.

3.1 Comparison to Infinite Width Networks

The angle degeneracy phenomenon has been studied in previous works for networks in the limit of infinite width [11, 12, 22, 23]. The infinite width case uses the law of large numbers

and thereby disregards any random fluctuations in $\theta^{\ell+1}$ given θ^{ℓ}. These random fluctuations, though small, can accumulate over many layers leading to inaccurate predictions for finite width networks (see Figure 1.1). The infinite width update rule is given below in Approximation 3.

Another issue with using the infinite width prediction to study finite width networks is that all networks with the same depth are treated exactly the same, since it does not take into account the width of each layer. Both the depth of the network and the width of each layer affect how the angle between inputs propagates layer-by-layer through the network. Figure 3.2-Left illustrates how our method yields different angle predictions for different architectures with the same depth, while the infinite width method does not. Figure 3.2-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.

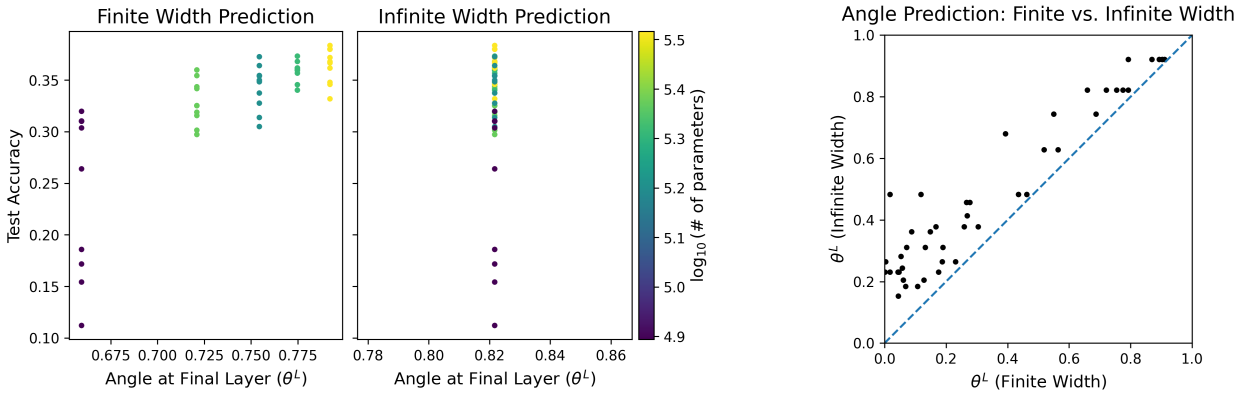


Figure 3.2: Left: Comparison of the finite and infinite width predictions for 5 network architectures with a depth of $L = 3$ trained 10 times each on the CIFAR-10 dataset [17]. The infinite width predicts the same final angle for all networks, since it only depends on network depth. Right: Using the same 45 network architectures as in Figure 3.1, 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.

Explicit Formula for Mixed-Moment J Functions

In this chapter, we develop a combinatorial method that allows us 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 (1.13) 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

$$J_{a,b}(\theta) = \mathbf{E}[G^a(G\cos\theta + W\sin\theta)^b \, 1\{G > 0\} \, 1\{G\cos\theta + W\sin\theta > 0\}], \tag{4.1}$$

where $a, b \in \mathbb{N} \cup \{0\}$ and $1\{A\}$ is the indicator function for condition A. 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 (4.1) generalizes (1.13) to the case $a = 0$. We also note that $J_{a,b}(\theta) = J_{b,a}(\theta)$ for all $a, b \in \mathbb{N} \cup \{0\}$.

4.1 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 4.2 where this is carried out. Then, by use of the generalized Gaussian integration by parts formula given in Section 4.2, 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:

$$J_{a,0}(\theta) = (a-1)J_{a-2,0}(\theta) + \frac{\sin^{a-1}\theta\cos\theta}{c_{a \bmod 2}}(a-2)!!, \tag{4.2}$$

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:*

$$J_{a,b}(\theta) = (a-1)J_{a-2,b}(\theta) + b\cos\theta J_{a-1,b-1}(\theta). \tag{4.3}$$

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*

$$J_{0,0}(\theta) = \frac{\pi-\theta}{2\pi}, \qquad J_{1,0}(\theta) = \frac{1+\cos\theta}{2\sqrt{2\pi}}, \qquad J_{1,1}(\theta) = \frac{\sin\theta + (\pi-\theta)\cos\theta}{2\pi}. \tag{4.4}$$

See Appendix B.1 for a derivation of these quantities. Note that Cho and Saul [5] 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, we are also 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.

Proof See Appendix B.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 we define below.

Definition 1 (P and Q Numbers)**.** *The numbers $P(a,b)$ and $Q(a,b)$ are defined as follows,*

$$P(a,b) = \begin{cases} \frac{a!}{b!\left(\frac{a-b}{2}\right)!\, 2^{\frac{a-b}{2}}}, & a \geq b,\ a \equiv b \pmod 2 \\ 0, & \textit{otherwise} \end{cases}, \tag{4.5}$$

$$Q(a,b) = \begin{cases} \frac{\left(\frac{a+b}{2}\right)!}{b!} 2^{\frac{b-a}{2}} \sum_{i=0}^{\frac{a-b}{2}} \binom{a+1}{i}, & a \geq b,\ a \equiv b \pmod 2 \\ 0, & \textit{otherwise} \end{cases}. \tag{4.6}$$

$P(a,b)$ represents a family of numbers known as the Bessel numbers of the second kind [4]. The Bessel numbers are the coefficients of the Bessel polynomials [6], which arise naturally in studies of the classical wave equation in a spherical coordinate system [15]. The Q numbers are closely related to the Bessel numbers [16], and both the P and Q numbers follow a similar recursion pattern to that of $J_{a,b}$ (see Lemma 3). Using these, we can express $J_{a,b}$ as follows:

Theorem 2 (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 $J_{1,n}$*

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

where $(b)_k = b(b-1)\cdots(b-k+1)$ is the falling factorial with k terms.

Remark 1. *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 $J_{0,n}$ and $J_{1,1}$. This substitution yields the following formula.*

For notational convenience, we will let $\delta := b - a$,

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

4.2 Gaussian Integration-by-Parts Formulas

This section covers two important formulas that together give us the tools for computing the expectations that appear in $J_{a,b}$.

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 with $\lim_{g \to \infty} f(g) e^{\frac{-g^2}{2}} = 0$. Then,*

$$\mathbf{E}[Gf(G)] = \mathbf{E}[f'(G)]. \tag{4.7}$$

Proof (of Fact 1). Applying integration by parts, we have

$$\begin{aligned} \mathbf{E}[Gf(G)] &= \int_{-\infty}^{\infty} g f(g) \frac{e^{\frac{-g^2}{2}}}{\sqrt{2\pi}} dg \\ &= \left[f(g) \left(\frac{-e^{\frac{-g^2}{2}}}{\sqrt{2\pi}} \right) \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \left(\frac{-e^{\frac{-g^2}{2}}}{\sqrt{2\pi}} \right) f'(g) dg \\ &= 0 + \int_{-\infty}^{\infty} f'(g) \frac{e^{\frac{-g^2}{2}}}{\sqrt{2\pi}} dg \\ &= \mathbf{E}[f'(G)] \end{aligned}$$

□

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.

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*

$$\mathbf{E}[1'\{G > a\}f(G)] = f(a)\frac{e^{\frac{-a^2}{2}}}{\sqrt{2\pi}}. \tag{4.8}$$

Remark 2. *The purpose of assigning a value to the expectation* (4.8) *is to allow one to compute "honest" expectations of the form* (4.7) *when $f(x)$ involves $1\{x > 0\}$; see Lemma 1 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 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 $\lim_{\epsilon\to 0} \mathbf{E}[1'_\epsilon\{G > a\}f(G)]$.

Proof (of Fact 2). Applying integration by parts, we formally have

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

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

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

where we have used the hypothesis on f once again. □

Corollary 2. *For two independent Gaussian variables G, W, and $f : \mathbb{R}^2 \to \mathbb{R}$ such that*

$\lim_{g\to\infty} \mathbf{E}[f(g,W)]e^{\frac{-g^2}{2}} = 0$, *we have that*

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

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 1 (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 (4.7) with $f(x) = 1\{x > 0\}$ and then applying (4.8) with $f(x) \equiv 1$,

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

Now, to see the induction, we apply (4.7) with $f(x) = x^{k-1}1\{x > 0\}$, $k \geq 2$. Due to the product rule, there are two terms in the derivative,

$$\begin{aligned} \mathbf{E}[\varphi(G)^k] &= \mathbf{E}[G \cdot G^{k-1}1\{G > 0\}] \\ &= (k-1)\mathbf{E}[G^{k-2}1\{G > 0\}] + \mathbf{E}[G^{k-1}1'\{G > 0\}] \\ &= (k-1)\mathbf{E}[\varphi(G)^{k-2}] + 0, \end{aligned} \tag{4.9}$$

where we have recognized that the second term is 0 by application of (4.8) 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. □

4.3 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 (4.7) to $f(x) = x^{a-1}1\{x > 0\}1\{x\cos\theta + W\sin\theta > 0\}$ to evaluate the expected value over G first. When applying the product rule there are three terms:

$$\begin{aligned} J_{a,0} =& \mathbf{E}[G \cdot G^{a-1}1\{G > 0\}1\{G\cos\theta + W\sin\theta > 0\}] \\ =& (a-1)\mathbf{E}[G^{a-2}1\{G > 0\}1\{G\cos\theta + W\sin\theta > 0\}] \\ &+ \mathbf{E}[G^{a-1}1\{G > 0\}1'\{G\cos\theta + W\sin\theta > 0\}]\cos\theta \\ &+ \mathbf{E}[G^{a-1}1'\{G > 0\}1\{G\cos\theta + W\sin\theta > 0\}]. \end{aligned} \tag{4.10}$$

The first term is simply $(a-1)J_{a-2,0}$. The last two terms can now be evaluated with the help of (4.8). The last term of (4.10) is (4.8) 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 (4.10), we introduce a change of variables to express $G\cos\theta + W\sin\theta$ in terms of two other independent Gaussian variables $Z, W \sim \mathcal{N}(0,1)$

$$\begin{aligned} Z &= G\cos\theta + W\sin\theta, \qquad & G &= Z\cos\theta + Y\sin\theta, \\ Y &= G\sin\theta - W\cos\theta, \qquad & W &= Z\sin\theta - Y\cos\theta, \end{aligned} \tag{4.11}$$

where Y, Z iid $\mathcal{N}(0,1)$. Under this change of variables, $J_{a,0},\ a \geq 2$ is setup to apply (4.8) 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 1 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 (4.7) 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 four terms in the product rule derivative. Fortunately in this case, the last two terms are simply zero by application of (4.8) 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. □

4.4 Solving the Recurrence to get an Explicit Formula for $J_{a,b}(\theta)$ - Proof of Theorem 2

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

$$J_{a,b} = (a-1)J_{a-2,b} + b\cos\theta J_{a-1,b-1}, \quad a \geq 2, b \geq 1, \tag{4.12}$$

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 2 (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 (4.12) *for* $J_{a,b}$ *can be expressed in the graph as a sum over incoming edges,*

$$J_{a,b} = \sum_{e:(a',b')\to(a,b)} w_e^J J_{a',b'}, \tag{4.13}$$

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 4.1.*

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 coefficient in front of each source is simply the weighed sum over all paths from the source to the node, namely*

$$J_{a,b} = \sum_{source\ vertices\ v} W^J_{v\to(a,b)} J_v = \sum_{n\geq 0} W^J_{(0,n)\to(a,b)} J_{0,n} + \sum_{n\geq 0} W^J_{(1,n)\to(a,b)} J_{1,n}, \tag{4.14}$$

$$W^J_{(a',b')\to(a,b)} := \sum_{\pi:(a',b')\to(a,b)} \prod_{e\in\pi} w_e^J, \tag{4.15}$$

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 (4.14), to prove Theorem 2, we have only to calculate the weighted sum of paths $W^J_{(0,n)\to(a,b)}$ 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 1.

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* (4.15)*,*

$$W^J_{(0,n)\to(a,b)} = (b)_{b-n}(\cos\theta)^{b-n}(P(a,b-n) - Q(a-1,b-n-1)), \tag{4.16}$$

$$W^J_{(1,n)\to(a,b)} = (b)_{b-n}(\cos\theta)^{b-n} Q(a-1,b-n). \tag{4.17}$$

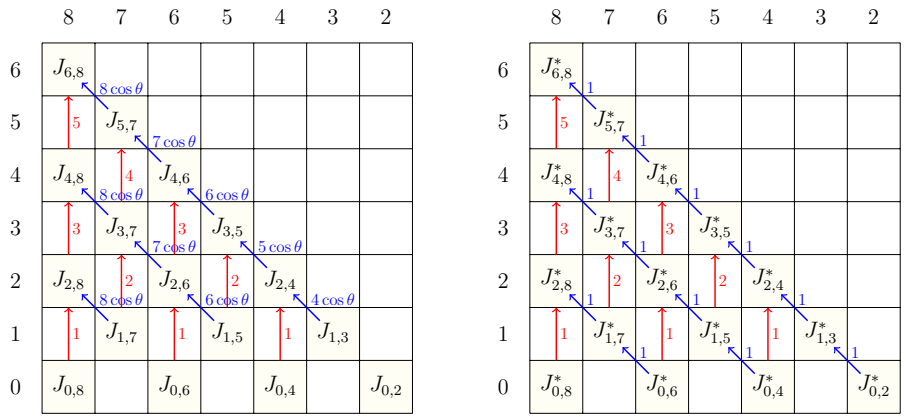
To prove this Proposition 4, we first create a simpler recursion, J^*, which we solve first and then slightly modify the solution to get the solution for J.

Lemma 2 (Weighted sums of paths for J^*)**.** *Let* $J^*_{a,b}$ *be defined to be the recursion:*

$$J^*_{a,b} := (a-1)J^*_{a-2,b} + 1J^*_{a-1,b-1} \text{ for } 2\leq a\leq b, \tag{4.18}$$

$$J^*_{1,b} := 0 + 1J^*_{0,b-1} \text{ for } 1\leq b. \tag{4.19}$$

Thinking of this recursion as a graph as in Definition 2 (see Figure 4.1 for an illustration),



(a) Directed graph associated with J (b) Directed graph associated with J^*

Figure 4.1: The graph associated with the recursions for J in (4.12) (left) and J^* in (4.19) (right). The graph is defined so that the recursion is given by a sum of incoming edges as in (4.13). The edges are color coded red and blue to match the coefficients in the recursion.

we have that the sum of weighted paths W^{J^*} defined analogously to those in (4.15)*, are given by P and Q numbers, namely*

$$W^{J^*}_{(0,n)\to(a,b)} = P(a, b-n), \quad W^{J^*}_{(1,n)\to(a,b)} = Q(a-1, b-n). \tag{4.20}$$

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

Lemma 3. *(Recursion for P and Q numbers) The P numbers, defined in Definition 1, satisfy $P(0,0) = 1$, $P(n,n) = P(n-1, n-1)$ for $n \geq 1$, and the recursion*

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

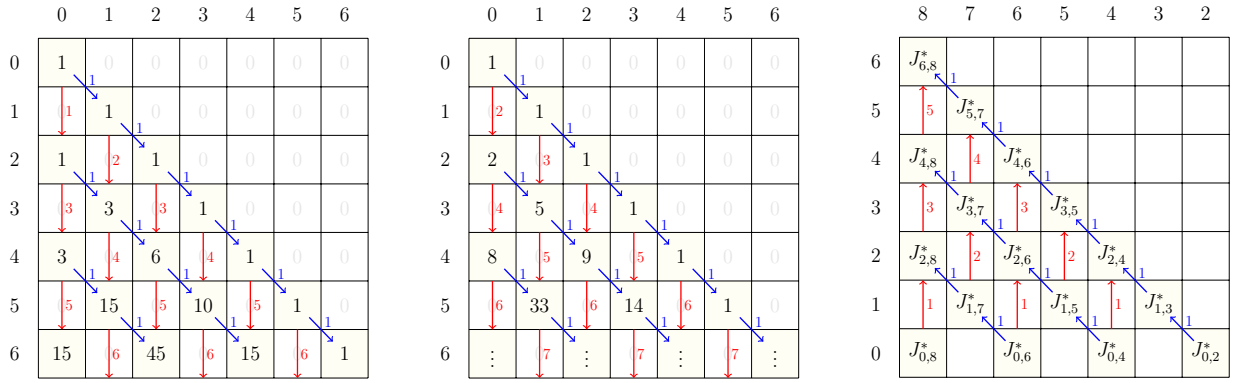
under the convention that $P(a,-1) = 0$. The Q numbers satisfy the same recursion as the P numbers, with a coefficient of a rather than $(a-1)$.

The proof of Lemma 3 is an easy consequence of known results from [16] and is deferred to Appendix C.2.

Proof (of Lemma 2). Using the same idea of recursions expressed as graphs as in Definition 2, the recursion from Lemma 3 means that P and Q can be expressed as weighted directed graphs. These are displayed in Figure 4.2. Since the P and Q graphs have only one single unit valued source vertex at $(0,0)$, (4.14) shows that the P and Q numbers are actually

themselves equal to sums over weighted paths in their respective graphs

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



(a) P (b) Q (c) $J^*_{6,8}$

Figure 4.2: 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.

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,

$$W^{J^*}_{(0,n)\to(a,b)} = W^{P}_{(0,0)\to(a,b-n)}, \quad W^{J^*}_{(1,n)\to(a,b)} = W^{Q}_{(0,0)\to(a-1,b-n)}. \tag{4.21}$$

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 4.2, 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. □

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 2. There are two differences between the formula for $J_{a,b}$ compared to $J^*_{a,b}$, which can both be seen in Figure 4.1. 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, $(b)_{b'-b}(\cos\theta)^{b'-b}$. (Recall that $(b)_k = b(b-1)\cdots(b-k+1)$ is the falling factorial with k terms.) This argument shows that the weighted sum of paths in J and J^* are related by

$$W^{J}_{(a,b)\to(a',b')} = (b)_{b'-b}(\cos\theta)^{b'-b} W^{J^*}_{(a,b)\to(a',b')}. \tag{4.22}$$

By the result of Lemma 2, 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 (4.22) 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

$$W^{J^*}_{(0,n)\to(a,b)} = W^{J^*}_{(1,n+1)\to(a,b)} + W^{J^*}_{(2,n)\to(a,b)}, \tag{4.23}$$

$$\implies W^{J^*}_{(2,n)\to(a,b)} = W^{J^*}_{(0,n)\to(a,b)} - W^{J^*}_{(1,n+1)\to(a,b)} \tag{4.24}$$

$$= P(a, b-n) - Q(a-1, b-n-1), \tag{4.25}$$

by the result of Lemma 2. By applying now (4.22) 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 2). The formula is immediate from (4.14), 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. □

Conclusion And Further Work

This thesis provides a detailed analysis of the angle process in deep feed forward ReLU networks on initialization. Our derivation of an explicit formula for the mixed-moment J functions allows for this analysis to be repeated with more accurate approximations for the mean and variance of $\ln(\sin^2(\theta^\ell))$ by including higher-order mixed J functions in the calculations.

We believe the methods proposed here are flexible enough to be modified to apply to non-linearities other than the ReLU. It would be interesting to repeat our analysis for other activations, such as $\tanh(x)$ to see how the angle evolution changes compared to the ReLU. The angle evolution could also be studied for ReLU networks with architectures beyond fully-connected networks. Adding skip connections or highway layers can improve information flow to deep layers of a network. Performing a similar analysis as ours on ResNets [14] or Highway Networks [24] could help us better understand how inputs behave as they travel through these modified architectures.

As mentioned previously, one method to combat network degeneracy is activation function shaping, where the activation function is tweaked in such a way to preserve the angle between inputs. It would be interesting to repeat our analysis for a more generalized ReLU function, such as the "leaky" ReLU studied in Li et al. [18], where they alter the slopes of the ReLU function on the negative and positive domains. Developing a more general update rule which depends on the modified slopes of the modified ReLU function would allow us to conduct a detailed analysis on how shaping the ReLU prevents the network from sending inputs to highly correlated outputs.

The experiments in Chapter 3 and would be an interesting starting point for more detailed experiments and/or theoretical explanations about training. We saw that network architectures which were predicted to have highly correlated outputs tended to perform worse and train more inconsistently than networks which better preserved the angle between inputs. This observation combined with the simplicity and efficiency of Algorithm 1 suggests our

prediction method may lend itself well to applications in neural architecture search.

Bibliography

[1] 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. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. `https://www.tensorflow.org/` Software available from tensorflow.org.

[2] Benny Avelin and Anders Karlsson. 2022. Deep Limits and a Cut-Off Phenomenon for Neural Networks. *Journal of Machine Learning Research* 23, 191 (2022), 1–29. `http://jmlr.org/papers/v23/21-0431.html`

[3] Sam Buchanan, Dar Gilboa, and John Wright. 2021. Deep Networks and the Multiple Manifold Problem. In *International Conference on Learning Representations*. `https://openreview.net/forum?id=O-6Pm_d_Q-`

[4] Gi-Sang Cheon, Ji-Hwan Jung, and Louis W. Shapiro. 2013. Generalized Bessel numbers and some combinatorial settings. *Discrete Mathematics* 313, 20 (2013), 2127–2138.

[5] Youngmin Cho and Lawrence Saul. 2009. Kernel Methods for Deep Learning. In *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), Vol. 22. Curran Associates, Inc. `https://proceedings.neurips.cc/paper/2009/file/5751ec3e9a4feab575962e78e006250d-Paper.pdf`

[6] Ji Young Choi and Jonathan D.H. Smith. 2003. On the unimodality and combinatorics of Bessel numbers. *Discrete Mathematics* 264, 1 (2003), 45–53. `https://doi.org/10.1016/S0012-365X(02)00549-6` The 2000 Com2MaC Conference on Association Schemes, Codes and Designs.

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

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

[9] Ronen Eldan and Ohad Shamir. 2015. The Power of Depth for Feedforward Neural Networks. In *Annual Conference Computational Learning Theory*.

[10] Boris Hanin. 2018. Which Neural Net Architectures Give Rise to Exploding and Vanishing Gradients?. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/13f9896df61279c928f19721878fac41-Paper.pdf

[11] Boris Hanin. 2023. Random Fully Connected Neural Networks as Perturbatively Solvable Hierarchies. arXiv:2204.01058 [math.PR] https://arxiv.org/abs/2204.01058

[12] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. 2019. On the Impact of the Activation Function on Deep Neural Networks Training. In *International Conference on Machine Learning*.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1026–1034. https://doi.org/10.1109/ICCV.2015.123

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE International Conference on Computer Vision and Pattern Recognition*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[15] Harry L Krall and Orrin Frink. 1949. A new class of orthogonal polynomials: The Bessel polynomials. *Trans. Amer. Math. Soc.* 65, 1 (1949), 100–115.

[16] Alexander Kreinin. 2016. Integer Sequences Connected to the Laplace Continued Fraction and Ramanujan's Identity. *Journal of Integer Sequences* 19 (06 2016), 1–12.

[17] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. (2009), 32–33. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[18] Mufan Bill Li, Mihai Nica, and Daniel M. Roy. 2022. The Neural Covariance SDE: Shaped Infinite Depth-and-Width Networks at Initialization. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=WG3vmsteqR_

[19] James Martens, Andy Ballard, Guillaume Desjardins, Grzegorz Swirszcz, Valentin Dalibard, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. 2021. Rapid training of deep neural networks without skip connections or normalization layers using Deep Kernel Shaping. *CoRR* (2021). arXiv:2110.01765 https://arxiv.org/abs/2110.01765

[20] Ido Nachum, Jan Hazla, Michael Gastpar, and Anatoly Khina. 2022. A Johnson-Lindenstrauss Framework for Randomly Initialized CNNs. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YX0lrvdPQc

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

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

[23] Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. 2017. Deep Information Propagation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1W1UN9gg

[24] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training Very Deep Networks. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/215a71a12769b056c3c32e7299f1c5ed-Paper.pdf

[25] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. http://arxiv.org/abs/1708.07747

Appendix A

A.1 Expected Value Approximation

Lemma 4. *Both the random variables $X = R^{\ell+1}$ and $X = R^{\ell+1}\sin^2(\theta^\ell)$ satisfy*

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

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

$$\ln(X) = \ln\left(\mathbf{E}[X]\left(\frac{\mathbf{E}[X] + (X - \mathbf{E}[X])}{\mathbf{E}[X]}\right)\right) = \ln(\mathbf{E}[X]) + \ln\left(1 + \frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right). \tag{A.2}$$

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 (2.3, 2.4), both $X = R^{\ell+1}$ and $X = R^{\ell+1}\sin^2(\theta^{\ell+1})$ can be expressed

as averages of the form

$$X = \frac{1}{n_\ell^2} \sum_{i,j}^{n_\ell} f(G_i, \hat{G}_j). \tag{A.3}$$

From the bound on the 3rd moment in Lemma 7, 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 5. *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 (A.2), and using the first term of the Taylor series approximation for $\ln(1+x) = x + \epsilon_1(x)$ now, we have that

$$\ln(X) = \ln(\mathbf{E}[X]) + \frac{X - \mathbf{E}[X]}{\mathbf{E}[X]} + \epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right). \tag{A.4}$$

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 $\mathbf{Var}[\ln(X)]$.

$$\begin{aligned}\mathbf{Var}[\ln(X)] &= \mathbf{Var}\left[\ln(\mathbf{E}[X]) + \frac{X - \mathbf{E}[X]}{\mathbf{E}[X]} + \epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right)\right] \\ &= \mathbf{Var}\left[\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right] + \mathbf{Var}\left[\epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right)\right] + 2\,\mathbf{Cov}\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}, \epsilon_1\left(\frac{X - \mathbf{E}[X]}{\mathbf{E}[X]}\right)\right).\end{aligned}$$

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 (A.3). 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 7. 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 6. *Both the random variables $X = R^{\ell+1}$ and $Y = R^{\ell+1}\sin^2(\theta^{\ell+1})$ 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 (A.4) 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 8. □

A.4 Third and Fourth Moment Bound Lemma

Lemma 7. *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, $\mathbf{E}[f(G_i, \hat{G}_j)^4] < \infty$. Then, the third and fourth central moment of A satisfy*

$$\mathbf{E}[(A - \mathbf{E}[A])^3] = \mathcal{O}(n^{-2}), \quad \mathbf{E}[(A - \mathbf{E}[A])^4] = \mathcal{O}(n^{-2}). \tag{A.5}$$

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:

$$A - \mathbf{E}[A] = \frac{1}{n^2} \sum_{i,j}^{n} \Big(f(G_i, \hat{G}_j) - \mathbf{E}[f(G_i, \hat{G}_j)] \Big)$$

$$\implies \mathbf{E}\left[(A - \mathbf{E}[A])^3\right] = \frac{1}{n^6} \sum_{\substack{i_1,i_2,i_3 \\ j_1,j_2,j_3}}^{n} \mathbf{E}\left[\prod_{k=1}^{3} \Big(f(G_{i_k}, \hat{G}_{j_k}) - \mathbf{E}[f(G_{i_k}, \hat{G}_{j_k})] \Big) \right]. \tag{A.6}$$

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

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

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

$$32(n)_4 + 68(n)_3 + 28(n)_2 + 1(n)_1. \tag{A.7}$$

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 A.1. 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 (Here, $(n)_k$ denotes the falling factorial with k terms).

Since the number of irreducible configurations is $\mathcal{O}(n^4)$, the normalization by n^6 in (A.6) 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 $\left|\bigcup_{k=1}^4 \{i_k, j_k\}\right| \geq 7$ 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 3. *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 8. *Let $G_i, \hat{G}_i$, $1 \leq i \leq n$ 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, $\mathbf{E}[f_1(G_i, \hat{G}_j)^4]$, $\mathbf{E}[f_2(G_i, \hat{G}_j)^4] < \infty$. Then,*

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

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

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

By the same argument as in Lemma 7, 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 arguments. □

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

Table A.1: 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 (A.6).

A.5 Expected Value Calculations

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

A.5.1 Calculation of $\mathbf{E}\left[R^{\ell+1}\right]$

First, we apply the identity as in (2.4):

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

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:

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

A.5.2 Calculation of $\mathbf{E}\left[\boldsymbol{R^{\ell+1}\sin^2(\theta^{\ell+1})}\right]$

Applying the identity (2.3), 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.6 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

$$\frac{1}{n_\ell^4}\sum_{\substack{i_1,j_1\\ i_2,j_2}}\left(\mathbf{E}\left[f_1(G_{i_1}, \hat{G}_{j_1})f_2(G_{i_2}, \hat{G}_{j_2})\right] - \mathbf{E}\left[f_1(G_{i_1}, \hat{G}_{j_1})\right]\mathbf{E}\left[f_2(G_{i_2}, \hat{G}_{j_2})\right]\right), \tag{A.8}$$

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 (A.8) 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 (A.8). We use the following formal notion of a pattern to organize this idea of repeated versus distinct indices.

Definition 3. *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 (A.8), many patterns have *no* contribution and can be ignored. We formalize this in the following definition.

Definition 4. *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 (A.8) completely vanishes! Therefore, to evaluate (A.8), 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 A.2.) 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 (A.8). This strategy is precisely how we evaluate each variance/covariance in this section.

A.6.1 Calculation of Var $\left[R^{\ell+1}\right]$

First, applying the identity in (2.4), 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 (A.8), 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 A.2. Combining all these cases and simplifying based on powers of $\frac{1}{n_\ell}$, we arrive at the following expression for $\mathbf{Var}\left[R^{\ell+1}\right]$:

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

A.6.2 Calculation of Var $\left[R^{\ell+1}\sin^2(\theta^{\ell+1})\right]$

Applying identity (2.3), we can express $\mathbf{Var}[R^{\ell+1}\sin^2(\theta^{\ell+1})]$ as

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

Note that we can express $\mathbf{Var}[R^{\ell+1}\sin^2(\theta^{\ell+1})]$ as in (A.8) 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 A.3. Combining all these cases and simplifying based on powers of $\frac{1}{n_\ell}$, we arrive at

$\mathbf{Var}[R^{\ell+1}]$ 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 A.2: $\mathbf{Var}[R^{\ell+1}]$ calculated in the form of (A.8) 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.

the following expression:

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

A.6.3 Calculation of $\mathbf{Cov}\left(\boldsymbol{R^{\ell+1}\sin^2(\theta^{\ell+1})}, \boldsymbol{R^{\ell+1}}\right)$

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

Applying known identities (2.3, 2.4) derived in Appendices A.7 and A.8, we can express this

$\mathbf{Var}[R^{\ell+1}\sin^2(\theta^{\ell+1})]$ 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 A.3: $\mathbf{Var}[R^{\ell+1}\sin^2(\theta^{\ell+1})]$ calculated in the form of (A.8) 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.

in the form of (A.8), 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 A.4 shows the calculation of all the irreducible patterns. Collecting all cases and simplifying based on powers of $\frac{1}{n_\ell}$ gives:

$$\mathbf{Cov}\left(R^{\ell+1}\sin^2(\theta^{\ell+1}), R^{\ell+1}\right) = \frac{1}{n_\ell}\left(16J_{1,1}^2 - 32J_{1,1}J_{3,1} + 8J_{2,2} + 8\right)$$
$$+\frac{1}{n_\ell^2}\left(32J_{1,1}^2J_{2,2} - 40J_{1,1}^2 + 96J_{1,1}J_{3,1} - 32J_{1,1}J_{3,3} + 16J_{2,2}^2 - 32J_{2,2} - 32J_{3,1}^2 + 16J_{4,2} + 10\right)$$
$$+\frac{1}{n_\ell^3}\left(24J_{1,1}^2 - 32J_{1,1}^2J_{2,2} - 64J_{1,1}J_{3,1} + 32J_{1,1}J_{3,3} - 16J_{2,2}^2 + 24J_{2,2} + 32J_{3,1}^2 - 16J_{4,2} - 18\right).$$

A.7 Derivation of Useful Identities - Equations 2.1, 2.2

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

$\mathbf{Cov}(R^{\ell+1}, R^{\ell+1}\sin^2(\theta^{\ell+1}))$ 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 A.4: $\mathbf{Cov}\left(R^{\ell+1}\sin^2(\theta^{\ell+1}), R^{\ell+1}\right)$ calculated in the form of (A.8) 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.

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 of $\varphi(z^\ell(x))$ with the i-th row of the weight matrix $W^{\ell+1}$:

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

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 $K_{i,j}^\ell$, for $i \in \{\alpha, \beta\}, j \in \{\alpha, \beta\}$ are actually *averages* of entries in the previous layer

$$K_{i,j}^\ell := \frac{2}{n_\ell}\langle\varphi_i^\ell, \varphi_j^\ell\rangle = \frac{1}{n_\ell}\sum_{k=1}^{n_\ell} 2\varphi(z_k^\ell(x_i))\varphi(z_k^\ell(x_j)).$$

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

$$z_i^{\ell+1}(x_\alpha) = \sqrt{\frac{2}{n_\ell}}\|\varphi_\alpha^\ell\| G_i, \quad z_i^{\ell+1}(x_\beta) = \sqrt{\frac{2}{n_\ell}}\|\varphi_\beta^\ell\|\hat{G}_i, \tag{A.9}$$

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

$$\varphi(z_i^{\ell+1}(x_\alpha)) = \sqrt{\frac{2}{n_\ell}}\|\varphi_\alpha^\ell\|\varphi(G_i), \quad \varphi(z_i^{\ell+1}(x_\beta)) = \sqrt{\frac{2}{n_\ell}}\|\varphi_\beta^\ell\|\varphi(\hat{G}_i). \tag{A.10}$$

Taking the norm/inner product of the vector now yields (2.1-2.3) as desired.

A.8 Cauchy-Binet and Determinant of the Gram Matrix - Equation 2.3

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

$$\det\begin{bmatrix} \|\varphi_\alpha^{\ell+1}\|^2 & \langle\varphi_\alpha^{\ell+1},\varphi_\beta^{\ell+1}\rangle \\ \langle\varphi_\alpha^{\ell+1},\varphi_\beta^{\ell+1}\rangle & \|\varphi_\beta^{\ell+1}\|^2 \end{bmatrix} = \sum_{1\le i<j\le n_\ell}\left(\varphi_{i;\alpha}^{\ell+1}\varphi_{j;\beta}^{\ell+1} - \varphi_{j;\alpha}^{\ell+1}\varphi_{i;\beta}^{\ell+1}\right)^2. \tag{A.11}$$

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 (A.11), and dividing both sides by $\|\varphi_\alpha^\ell\|^2\|\varphi_\beta^\ell\|^2$, we get our desired result.

A.9 Infinite Width Update Rule

Lemma 9. *Let $f(x)$ be a feed forward neural network as defined in 2.1. Conditional on the value of θ^ℓ in layer ℓ, the angle θ^ℓ between inputs at layer ℓ of f follows the following deterministic update rule in the limit $n_\ell \to \infty$.*

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

Remark 4. *Note that a more general proof of this result appears in prior work [11] which allows one to take the layer sizes $n_1, n_2, \ldots, n_\ell \to \infty$ in any order, rather than one layer at a time as we prove here.*

Proof. We begin with the identity (2.2), 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 (2.1) 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 (4.11), 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 (4.8) to evaluate the integrals involving $1'\{Z > 0\}$ and $\mathbf{E}[Y1\{Y > 0\}] = (\sqrt{2\pi})^{-1}$ from Lemma 1. We now have $J'_{0,0}(\theta) = -\frac{1}{2\pi}$ with initial condition given by $J_{0,0}(0) = \frac{1}{2}$. Solving this differential equation gives the desired result. □

Proof of Formula for $J_{1,0}$. Here we use the Gaussian integration-by-parts strategy (4.7 -4.8).

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

By using the change of variables as in (4.11) on the second term, we arrive at

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

□

Proof of Formula for $J_{1,1}$:

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

□

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.

Lemma 10. *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

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

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

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,

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

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

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

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. □

Lemma 11. *Let $J^{rec}_{n,1}$ be the recursively defined formula, and let $J^{exp}_{n,1}$ be the explicitly defined*

formula for $J_{n,1}$, which is given by

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

$$J_{n,1}^{rec} = (n-1)J_{n-2,1} + \cos\theta J_{n-1,0}, \quad J_{0,1}^{rec} := J_{0,1}, \quad J_{1,1}^{rec} := J_{1,1}.$$

Then, $J_{n,1}^{rec} = J_{n,1}^{exp}$ for all $n \geq 0$.

Proof. Let S_n, $n \in \mathbb{N}$, $n \geq 2$ be the statement $J_{n,1}^{rec} = J_{n,1}^{exp}$. We show that the base case S_2 is true.

$$J_{2,1}^{rec} = (2-1)J_{0,1} + \cos\theta J_{1,0} = (1+\cos\theta)J_{1,0} = \frac{(1+\cos\theta)^2}{2\sqrt{2\pi}},$$

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

Under the assumption that S_n is true, we show that S_{n+1} is also true.

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

□

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 4.2c. 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 4.2a, 4.2b. 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 B.1. 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 (see Figure B.2 for an illustration). This bijection establishes the equality of the weighted paths claim in (4.21).

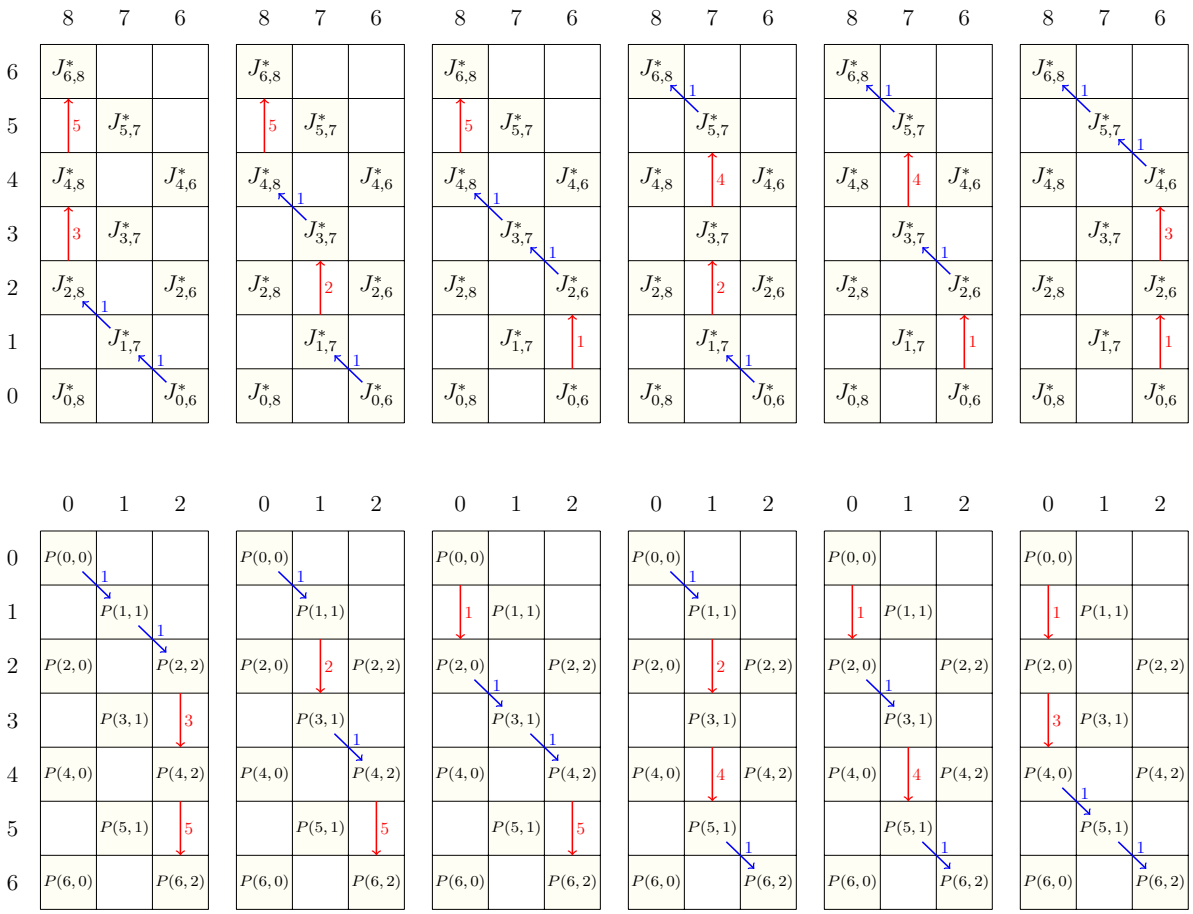


Figure B.1: Top: All paths $\pi \in \Gamma_{J^*}(6, 8, 0, 6)$. Bottom: All paths $\Lambda(\pi) \in \Gamma_P(6, 8, 0, 6)$. The paths are lined up so that for each path π in the top row, $\Lambda(\pi)$ appears in the bottom row.

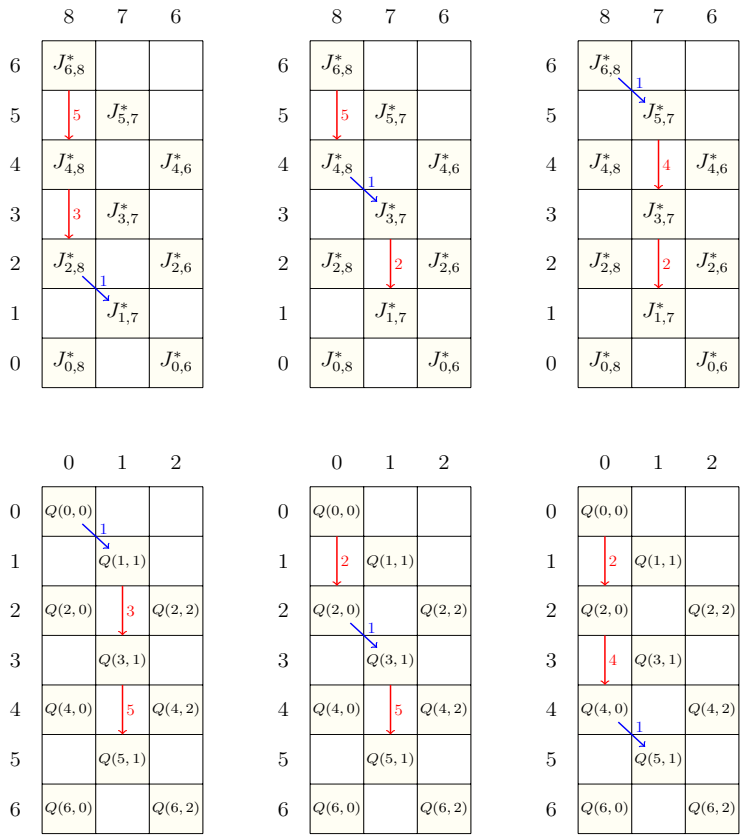


Figure B.2: Top: All paths $\pi \in \Gamma_{J^*}(6, 8, 1, 7)$. Bottom: All paths $\Lambda(\pi) \in \Gamma_Q(6, 8, 1, 7)$. The paths are lined up so that for each path π in the top row, $\Lambda(\pi)$ appears in the bottom row.

Appendix C

C.1 P and Q numbers

The P and Q numbers were introduced in [16] and can also be thought of as infinite matrices with elements in the a^{th} row and b^{th} column given by $P(a,b)$ and $Q(a,b)$, respectively.

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 3 & 0 & 1 & 0 & 0 & 0 & \cdots \\ 3 & 0 & 6 & 0 & 1 & 0 & 0 & \cdots \\ 0 & 15 & 0 & 10 & 0 & 1 & 0 & \cdots \\ 15 & 0 & 45 & 0 & 15 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array} \qquad Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 2 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 5 & 0 & 1 & 0 & 0 & 0 & \cdots \\ 8 & 0 & 9 & 0 & 1 & 0 & 0 & \cdots \\ 0 & 33 & 0 & 14 & 0 & 1 & 0 & \cdots \\ 48 & 0 & 87 & 0 & 20 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array} \tag{C.1}$$

C.2 Recursions for the P and Q numbers - Proof of Lemma 3

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

Theorem 3 (Kreinin [16])**.** *The elements of the matrices P and Q satisfy*

$$b \cdot P(a,b) = a \cdot P(a-1,b-1), \qquad a \geq 1, 1 \leq b \leq a, \tag{C.2}$$

$$P(a+1,b) = P(a,b-1) + (b+1) \cdot P(a,b+1), \qquad a \geq 0, 1 \leq b \leq a, \tag{C.3}$$

$$Q(a,b) = P(a,b) + (b+1) \cdot Q(a-1,b+1), \qquad a \geq 1, 1 \leq b \leq a, \tag{C.4}$$

$$Q(a,b) = a \cdot Q(a-2,b) + Q(a-1,b-1), \qquad a \geq 2, 1 \leq b \leq a. \tag{C.5}$$

Of Lemma 3. Equation (C.3) 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 (C.2) 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 (C.5), 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 3. □

Appendix D

This section details the architectures of the 45 different neural networks used to produce Figure 3.1.

#	Depth	Avg. Width	# Parameters		Avg. Test Accuracy ± Standard Deviation		
			(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 D.1: Summary of the architectures of the first 20 neural networks used in Figure 3.1, 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 D.2.

#	Depth	Avg. Width	# Parameters		Average Score ± Standard Deviation		
			(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 D.2: Continuation of Table D.1 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 D.3: Ordered list of hidden layer widths for the first 25 networks used in Figure 3.1. This table is continued in Table D.4.

#	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, 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, 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, 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, 20, 20, 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
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
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
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
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

Table D.4: Continuation of Table D.3 for networks 26 through 45.