The Distribution of Elements of the Neural Tangent Kernel on Initialization in ReLU Networks

David Shustin[♡] **Evan Dogariu**[♠]
[♡]Department of Electrical Engineering, Princeton University
[♠]Department of Computer Science, Princeton University
{dshustin, edogariu}@princeton.edu

Abstract

In this work, we consider a practical and general setting: the ReLU neural network of arbitrary width and depth with no biases and output dimension 1, parameterized in the NTK regime. We study the distribition of entries of the Neural Tangent Kernel (NTK) upon reasonable initialization. We use a combinatorial approach to investigate the off-diagonal elements of the NTK, and find an expression for the expectation of all entries of the NTK in terms of the two-point correlator of the activation indicators for the same neuron on different datapoints. Using our results and well known methods from random matrix theory, we may place bounds on the initialized spectrum of the NTK.

We motivate and then explore a geometric approach to derive a closed form expectation for this two-point correlator in the first layer of a ReLU network. We also suggest research directions that might be explored by finding tighter bounds on this two-point correlator in deeper layers of the neural network.

Contents

1 Expectations of NTK Matrix Elements

We are interested firstly in solving for the expected NTK for an arbitrary dataset. Doing so will allow us to make claims about the expected training dynamics and trajectories of ReLU networks in a data-informed way. We state and prove the result in the next section.

1.1 Path Counting

Theorem 1.1. Consider the following ReLU neural network with no biases and output dimension 1, parameterized in the NTK regime. We have depth L and input dimension n_0, output dimension $n_{L+1} = 1$, and layer widths $n_1, ..., n_L \in \mathbb{N}$ arbitrary. For an input $x_\alpha \in \mathbb{R}^{n_0}$, we denote the preactivations of neuron i in layer l as $z_{i;\alpha}^{(l)}$, suppressing the dependence on x_α. The network is defined recursively by

$$z_{i;\alpha}^{(l+1)} = \begin{cases} \sum_{j=1}^{n_l} W_{ij}^{(l+1)} \sigma\left(z_{j;\alpha}^{(l)}\right) & l \geq 1 \\ \sum_{j=1}^{n_l} W_{ij}^{(l+1)} x_{j;\alpha} & l = 0 \end{cases}$$

with weights parameterized as $W_{ij}^{(l)} = \sqrt{\frac{2}{n_{l-1}}}\widehat{W}_{ij}^{(l+1)}$ and $\widehat{W}_{ij}^{(l+1)}$ drawn i.i.d from a distribution μ that is symmetric about 0 with mean 0, unit variance, and no atoms. We consider a dataset of m inputs $\{x_\alpha\}_{\alpha=1}^m$ that are distinct. Define the neural tangent kernel $K \in \mathbb{R}^{m\times m}$ to be the Gram matrix of the Jacobians for different inputs, with elements

$$K_{\alpha\beta} = \left(\vec{\nabla}_\theta z_\alpha^{(L+1)}\right)^T \left(\vec{\nabla}_\theta z_\beta^{(L+1)}\right) = \sum_{l=1}^{L+1}\sum_{i=1}^{n_l}\sum_{j=1}^{n_{l-1}} \frac{\partial z_\alpha^{(L+1)}}{\partial \widehat{W}_{ij}^{(l)}} \frac{\partial z_\beta^{(L+1)}}{\partial \widehat{W}_{ij}^{(l)}}$$

Here, $z_\alpha^{(L+1)}$ denotes the value of the single output neuron given input x_α. Lastly, define

$$\xi_{i;\alpha}^{(l)} = \mathbb{1}_{z_{i;\alpha}^{(l)} > 0}$$

to be an indicator random variable for if the i^{th} neuron in layer l activates when passed input x_α. Then, the elementwise expectation over initializations satisfies

$$\boxed{\mathbb{E}[K_{\alpha\beta}] = (x_\alpha \cdot x_\beta)\frac{L}{n_0}\left(2^{L+1}\right)\prod_{l=1}^{L}\mathbb{E}\left[\xi_{i;\alpha}^{(l)}\xi_{i;\beta}^{(l)}\right]}$$

Proof. We will approach this with the path counting framework, following up on a line of work presented in [1] [3] [2] [4] [5]. Formally, let $[n]$ denote $\{1, ..., n\}$ and let $\Gamma = [n_0] \times \ldots \times [n_{L+1}]$ be the set of all possible paths through the network. So, each element $\gamma \in \Gamma$ is a tuple of the form

$$\gamma = (\gamma(0), \ldots, \gamma(L+1))$$

where each $\gamma(l)$ denotes the index of the neuron in layer l that the path passes through. Denote

$$\Gamma_p = \{\gamma \in \Gamma \quad s.t. \quad \gamma(0) = p\}$$

to be the set of all paths through the network starting at input neuron p. Also, let $\widehat{W}_\gamma^{(l)} = \widehat{W}_{\gamma(l),\gamma(l-1)}^{(l)}$ denote the weight that the path γ passes through at layer l, and similarly let $\xi_{\gamma;\alpha}^{(l)}$ denote $\xi_{\gamma(l);\alpha}^{(l)}$. We start with the well-known path counting result that

$$z_\alpha^{(L+1)} = \sum_{p=1}^{n_0} x_{p;\alpha} \sum_{\gamma\in\Gamma_p} W_\gamma^{(L+1)} \prod_{l=1}^{L} W_\gamma^{(l)} \xi_\gamma^{(l)} = \sum_{p=1}^{n_0} x_{p;\alpha} \sum_{\gamma\in\Gamma_p} \sqrt{\frac{2}{n_L}}\widehat{W}_\gamma^{(L+1)} \prod_{l=1}^{L} \sqrt{\frac{2}{n_{l-1}}}\widehat{W}_\gamma^{(l)} \xi_\gamma^{(l)}$$

From here, we can compute the partials with respect to each weight simply, as

$$\frac{\partial z_\alpha^{(L+1)}}{\widehat{W}_{ij}^{(l)}} = \sum_{p=1}^{n_0} x_{p;\alpha} \sum_{\substack{\gamma\in\Gamma_p \\ \gamma(l)=i \\ \gamma(l-1)=j}} \frac{\sqrt{\frac{2}{n_L}}\widehat{W}_\gamma^{(L+1)} \prod_{l'=1}^{L} \sqrt{\frac{2}{n_{l'-1}}}\widehat{W}_\gamma^{(l')} \xi_{\gamma;\alpha}^{(l')}}{\widehat{W}_{ij}^{(l)}}$$

This is clear since only paths that pass through $\widehat{W}_{ij}^{(l)}$ will contribute to this derivative. Now, let

$$B_{ijl;\alpha\beta} = \frac{\partial z_\alpha^{(L+1)}}{\widehat{W}_{ij}^{(l)}} \frac{\partial z_\beta^{(L+1)}}{\widehat{W}_{ij}^{(l)}}$$

be a random variable denoting the contribution of weight $\widehat{W}_{ij}^{(l)}$ to the NTK[1]. We can simply write

$$B_{ijl;\alpha\beta} = \sum_{p_\alpha,p_\beta=1}^{n_0} x_{p_\alpha;\alpha}x_{p_\beta;\beta} \sum_{\substack{\gamma_\alpha\in\Gamma_{p_\alpha}\\ \gamma_\beta\in\Gamma_{p_\beta}\\ \gamma_\alpha(l)=\gamma_\beta(l)=i\\ \gamma_\alpha(l-1)=\gamma_\beta(l-1)=j}} \frac{\frac{2}{n_L}\widehat{W}_{\gamma_\alpha}^{(L+1)}\widehat{W}_{\gamma_\beta}^{(L+1)}\prod_{l'=1}^{L}\frac{2}{n_{l'-1}}\widehat{W}_{\gamma_\alpha}^{(l')}\widehat{W}_{\gamma_\beta}^{(l')}\xi_{\gamma_\alpha;\alpha}^{(l')}\xi_{\gamma_\beta;\beta}^{(l')}}{\left(\widehat{W}_{ij}^{(l)}\right)^2}$$

$$= \left(\prod_{l'=0}^{L}\frac{2}{n_l'}\right)\sum_{p_\alpha,p_\beta=1}^{n_0} x_{p_\alpha;\alpha}x_{p_\beta;\beta} \sum_{\substack{\gamma_\alpha\in\Gamma_{p_\alpha}\\ \gamma_\beta\in\Gamma_{p_\beta}\\ \gamma_\alpha(l)=\gamma_\beta(l)=i\\ \gamma_\alpha(l-1)=\gamma_\beta(l-1)=j}} \widehat{W}_{\gamma_\alpha}^{(L+1)}\widehat{W}_{\gamma_\beta}^{(L+1)}\xi_{\gamma_\alpha;\alpha}^{(l)}\xi_{\gamma_\beta;\beta}^{(l)}\prod_{\substack{l'=1\\ l'\neq l}}^{L}\widehat{W}_{\gamma_\alpha}^{(l')}\widehat{W}_{\gamma_\beta}^{(l')}\xi_{\gamma_\alpha;\alpha}^{(l')}\xi_{\gamma_\beta;\beta}^{(l')}$$

where for the second equality we simply used the fact that both paths γ_α and γ_β agree at layer l. We would like to consider the expectation of $B_{ijl;\alpha\beta}$ over random initializations. Note that each element is a product of many things, many of which are independent from each other. We have seen in Lecture that each random variable $\xi_{\gamma_\alpha;\alpha}^{(l)}$ is distributed as $Bernoulli(\frac{1}{2})$ and is independent from all other $\xi_{\gamma_\alpha;\alpha}^{(l')}, l' \neq l$. Furthermore, we have seen that each $\xi_{\gamma_\alpha;\alpha}^{(l)}$ is independent of any even function of the weights[2]. What this (in addition to the fact that each $\widehat{W}_{ij}^{(l)}$ is i.i.d.) means is that for each pair of paths $\gamma_\alpha, \gamma_\beta$, we have that the four random variables $\left(\widehat{W}_{\gamma_\alpha}^{(L+1)}\widehat{W}_{\gamma_\beta}^{(L+1)}\right), \left(\xi_{\gamma_\alpha;\alpha}^{(l)}\xi_{\gamma_\beta;\beta}^{(l)}\right), \left(\widehat{W}_{\gamma_\alpha}^{(l')}\widehat{W}_{\gamma_\beta}^{(l')}\right), \left(\xi_{\gamma_\alpha;\alpha}^{(l')}\xi_{\gamma_\beta;\beta}^{(l')}\right)$ are all pairwise independent for $l' \neq l$. So, we can use the rules of expectation to see that each element of the inner sum, in expectation, looks like

$$\mathbb{E}\left[\widehat{W}_{\gamma_\alpha}^{(L+1)}\widehat{W}_{\gamma_\beta}^{(L+1)}\right]\mathbb{E}\left[\xi_{\gamma_\alpha;\alpha}^{(l)}\xi_{\gamma_\beta;\beta}^{(l)}\right]\prod_{\substack{l'=1\\ l'\neq l}}^{L}\mathbb{E}\left[\widehat{W}_{\gamma_\alpha}^{(l')}\widehat{W}_{\gamma_\beta}^{(l')}\right]\mathbb{E}\left[\xi_{\gamma_\alpha;\alpha}^{(l')}\xi_{\gamma_\beta;\beta}^{(l')}\right]$$

From here, we note that if for any l' it is the case that γ_α and γ_β don't agree (meaning either $\gamma_\alpha(l'-1) \neq \gamma_\beta(l')$ or $\gamma_\alpha(l') \neq \gamma_\beta(l'-1)$), we get that $\mathbb{E}\left[\widehat{W}_{\gamma_\alpha}^{(l')}\widehat{W}_{\gamma_\beta}^{(l')}\right]$ separates because different weights are independent; since μ is 0 mean, the entire product becomes 0 and this term doesn't contribute to the sum. This means that the only nonzero elements of the sum must have $\gamma_\alpha = \gamma_\beta$ everywhere (they already must agree at l and we just found that they must agree for all $l' \neq l$). This also means that the only nonzero elements must have $p_\alpha = p_\beta$. This simplifies our expression for $\mathbb{E}[B_{ijl;\alpha\beta}]$ to

$$\mathbb{E}[B_{ijl;\alpha\beta}] = \left(\prod_{l'=0}^{L}\frac{2}{n_l'}\right)\sum_{p=1}^{n_0} x_{p;\alpha}x_{p;\beta} \sum_{\substack{\gamma\in\Gamma_p\\ \gamma(l)=i\\ \gamma(l-1)=j}} \mathbb{E}\left[\left(\widehat{W}_{\gamma}^{(L+1)}\right)^2\right]\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l)}\xi_{\gamma;\beta}^{(l)}\right]\prod_{\substack{l'=1\\ l'\neq l}}^{L}\mathbb{E}\left[\left(\widehat{W}_{\gamma}^{(l')}\right)^2\right]\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l')}\xi_{\gamma;\beta}^{(l')}\right]$$

[1] B for Boris :)

[2] To see these facts, it suffices to consider the symmetricity of the distribution μ from which the weights are drawn. We can invert any weight vector (since it is symmetric about 0 and each element is i.i.d.) to flip $\xi_{\gamma_\alpha;\alpha}^{(l)}$ without changing any other $\xi_{\gamma_\alpha;\alpha}^{(l')}$ or any even function of the weights. The distribution of $\xi_{\gamma_\alpha;\alpha}^{(l)}$ given the previous activations is the same as the overall distribution of $\xi_{\gamma_\alpha;\alpha}^{(l)}$, meaning they are independent. In general, the ability to flip the $\xi_{\gamma_\alpha;\alpha}^{(l)}$'s with equal probability without affecting another variable at all proves independence. We went over this line of reasoning in class, and a similar result is found in Proposition 2 of [3].

Since these weights are sampled from μ with second moment 1, this equals

$$\mathbb{E}[B_{ijl;\alpha\beta}] = \left(\prod_{l'=0}^{L}\frac{2}{n_l'}\right)\sum_{p=1}^{n_0}x_{p;\alpha}x_{p;\beta}\sum_{\substack{\gamma\in\Gamma_p\\ \gamma(l)=i\\ \gamma(l-1)=j}}\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l)}\xi_{\gamma;\beta}^{(l)}\right]\prod_{\substack{l'=1\\ l'\neq l}}^{L}\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l')}\xi_{\gamma;\beta}^{(l')}\right]$$
$$= \left(\prod_{l'=0}^{L}\frac{2}{n_l'}\right)\sum_{p=1}^{n_0}x_{p;\alpha}x_{p;\beta}\sum_{\substack{\gamma\in\Gamma_p\\ \gamma(l)=i\\ \gamma(l-1)=j}}\prod_{l'=1}^{L}\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l')}\xi_{\gamma;\beta}^{(l')}\right]$$

Another observation we can make is that the quantity $\prod_{l'=1}^{L}\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l')}\xi_{\gamma;\beta}^{(l')}\right]$ is independent of path γ due to symmetry of the weights: we can scramble the neurons in any order and, because the weights are i.i.d., the two-point correlator at each layer remains the same. This yields that

$$\mathbb{E}[B_{ijl;\alpha\beta}] = \left(\prod_{l'=0}^{L}\frac{2}{n_l'}\right)\left(\prod_{l'=1}^{L}\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l')}\xi_{\gamma;\beta}^{(l')}\right]\right)\sum_{p=1}^{n_0}x_{p;\alpha}x_{p;\beta}\sum_{\substack{\gamma\in\Gamma_p\\ \gamma(l)=i\\ \gamma(l-1)=j}}1$$
$$= \left(\prod_{l'=0}^{L}\frac{2}{n_l'}\right)\left(\prod_{l'=1}^{L}\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l')}\xi_{\gamma;\beta}^{(l')}\right]\right)\sum_{p=1}^{n_0}x_{p;\alpha}x_{p;\beta}\left|\{\gamma\in\Gamma_p \quad s.t. \quad \gamma(l)=i \text{ and } \gamma(l-1)=j\}\right|$$

The last observation to make is that there are $n_1\cdot n_2\cdot\ldots\cdot n_L$ distinct paths (recall $n_{L+1}=1$) in Γ_P; however the additional constraint that γ passes through $\widehat{W}_{ij}^{(l)}$ forces us to divide this by $n_{l-1}\cdot n_l$. So, the cardinality of the set in the above equation is $\frac{1}{n_{l-1}n_l}\prod_{l'=1}^{L}n_{l'}$, and therefore

$$\mathbb{E}[B_{ijl;\alpha\beta}] = \left(\prod_{l'=0}^{L}\frac{2}{n_l'}\right)\left(\prod_{l'=1}^{L}\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l')}\xi_{\gamma;\beta}^{(l')}\right]\right)\sum_{p=1}^{n_0}x_{p;\alpha}x_{p;\beta}\frac{1}{n_{l-1}n_l}\prod_{l'=1}^{L}n_{l'}$$
$$= \frac{2^{L+1}}{n_0 n_{l-1}n_l}\left(\prod_{l'=1}^{L}\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l')}\xi_{\gamma;\beta}^{(l')}\right]\right)\sum_{p=1}^{n_0}x_{p;\alpha}x_{p;\beta}$$
$$= \frac{2^{L+1}}{n_0 n_{l-1}n_l}x_\alpha\cdot x_\beta\prod_{l'=1}^{L}\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l')}\xi_{\gamma;\beta}^{(l')}\right]$$

Plugging this into our original form

$$K_{\alpha\beta} = \left(\vec{\nabla}_\theta z_\alpha^{(L+1)}\right)^T\left(\vec{\nabla}_\theta z_\beta^{(L+1)}\right) = \sum_{l=1}^{L+1}\sum_{i=1}^{n_l}\sum_{j=1}^{n_{l-1}}\frac{\partial z_\alpha^{(L+1)}}{\widehat{W}_{ij}^{(l)}}\frac{\partial z_\beta^{(L+1)}}{\widehat{W}_{ij}^{(l)}} = \sum_{l=1}^{L+1}\sum_{i=1}^{n_l}\sum_{j=1}^{n_{l-1}}B_{ijl;\alpha\beta}$$

yields that

$$\mathbb{E}\left[K_{\alpha\beta}\right] = \sum_{l=1}^{L+1}\frac{2^{L+1}}{n_0}x_\alpha\cdot x_\beta\prod_{l'=1}^{L}\mathbb{E}\left[\xi_{\gamma;\alpha}^{(l')}\xi_{\gamma;\beta}^{(l')}\right]$$
$$= \boxed{(x_\alpha\cdot x_\beta)\frac{L}{n_0}\left(2^{L+1}\right)\prod_{l=1}^{L}\mathbb{E}\left[\xi_{i;\alpha}^{(l)}\xi_{i;\beta}^{(l)}\right]}$$

as desired. □

Corollary 1.1. The expectation of the on-diagonal of the NTK of a ReLU network is

$$\boxed{\mathbb{E}\left[K_{\alpha\alpha}\right] = \frac{2L\|x_\alpha\|^2}{n_0}}$$

Proof. We immediately apply Theorem 1.1 with $\alpha = \beta$ to find that

$$\begin{aligned}\mathbb{E}[K_{\alpha\alpha}] &= (x_\alpha \cdot x_\alpha)\frac{L}{n_0}\left(2^{L+1}\right)\prod_{l=1}^{L}\mathbb{E}\left[\xi_{i;\alpha}^{(l)}\xi_{i;\alpha}^{(l)}\right] \\ &= \|x_\alpha\|^2\frac{L}{n_0}\left(2^{L+1}\right)\prod_{l=1}^{L}\mathbb{E}\left[\left(\xi_{i;\alpha}^{(l)}\right)^2\right]\end{aligned}$$

We note that in all cases, $\left(\xi_{i;\alpha}^{(l)}\right)^2 = \xi_{i;\alpha}^{(l)}$ due to the properties of indicator functions, so

$$\begin{aligned}\mathbb{E}[K_{\alpha\alpha}] &= \|x_\alpha\|^2\frac{L}{n_0}\left(2^{L+1}\right)\prod_{l=1}^{L}\mathbb{E}[\xi_{i;\alpha}^{(l)}] \\ &= \|x_\alpha\|^2\frac{L}{n_0}\left(2^{L+1}\right)\left(\frac{1}{2}\right)^L \\ &= \boxed{\frac{2L\|x_\alpha\|^2}{n_0}},\end{aligned}$$

which matches the result found in Proposition 3 of [2] up to the difference in architectures of the problems studied ([2] solves for network with a final layer parameterized by $W_{ij}^{L+1} = \sqrt{\frac{1}{n_L}}\widehat{W}_{ij}^{L+1}$ instead of our $W_{ij}^{L+1} = \sqrt{\frac{2}{n_L}}\widehat{W}_{ij}^{L+1}$, yielding an additional factor of 2 in our result). □

2 Towards the Two-Point Correlator $\mathbb{E}\left[\xi_{i;\alpha}^{(l)}\xi_{i;\beta}^{(l)}\right]$

The result of Theorem 1.1 shows that it is of interest to calculate the two-point correlator $\mathbb{E}\left[\xi_{i;\alpha}^{(l)}\xi_{i;\beta}^{(l)}\right]$ at different depths in the network. In the following sections, we explicitly compute this in the first layer via a geometric argument and provide bounds for later layers.

2.1 $\mathbb{E}\left[\xi_{i;\alpha}^{(l)}\xi_{i;\beta}^{(l)}\right]$ in Layer $l = 1$

Consider a neural network in the same setting as in the statement of Theorem 1.1 (ReLU network with output dimension 1, parameterized in the NTK regime with depth L and input dimension n_0, output dimension $n_{L+1} = 1$, and layer widths $n_1, \ldots, n_L \in \mathbb{N}$ arbitrary). However, we now set stricter constraints on the initialization. Specifically, we initialize the weights

$$\widehat{W}_{ij}^{(l)} \sim \mathcal{N}(0, 1)$$

This scheme amounts to the Kaiming initialization of our network in the NTK parameterization, a very reasonable and practical setting. Recall that vectors of componentwise i.i.d. random normal variables with mean 0 are

rotationally invariant. That is, if we constructed a random vector

$$\vec{X} = \left[X_i\right]_{i=0}^{N}$$

for N arbitrary, and $X_i \sim \mathcal{N}(0,1)$ i.i.d., then $\text{Rot}(\vec{X})$ has the same probability distribution as $\vec{X}$ for any rotational transformation Rot. It is a well-known corollary that $\vec{X}$ is equally likely to point in any direction. This rotational symmetry of our randomly initialized weights allows the following result.

Theorem 2.1. Consider the same setting as in Theorem 1.1, except that $\widehat{W}_{ij}^{(l)} \sim \mathcal{N}(0,1)$. Then the two-point correlator of the activation indicator for a single neuron in the first hidden layer when passed two data points x_α, x_β satisfies

$$\boxed{\mathbb{E}\left[\xi_{i;\alpha}^{(1)}\xi_{i;\beta}^{(1)}\right] = \begin{cases} \frac{1}{2} - J_n^{\frac{\pi}{2}-\theta_{\alpha\beta}}/A_n, & x_\alpha \cdot x_\beta \geq 0 \\ J_n^{\theta_{\alpha\beta}-\frac{\pi}{2}}/A_n, & x_\alpha \cdot x_\beta < 0 \end{cases}}$$

where

$$A_n = \frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}$$

is the surface area of an n-sphere,

$$\theta_{\alpha\beta} = \cos^{-1}\left(\frac{x_\alpha \cdot x_\beta}{\|x_\alpha\|\|x_\beta\|}\right)$$

is the angle between x_α and x_β, and

$$J_n^\theta = \frac{\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} \int_\theta^{\frac{\pi}{2}} \sin(\phi)^{n-2} I_{1-\frac{\tan^2(\theta)}{\tan^2(\phi)}}\left(\frac{n-2}{2}, \frac{1}{2}\right) \mathrm{d}\phi$$

where $I_x(a,b)$ is the regularized incomplete beta function

$$I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$$

Proof. First, we note that the expectation of an indicator is the probability that the indicator is one.

$$\mathbb{E}\left[\xi_{i;\alpha}^{(1)}\xi_{i;\beta}^{(1)}\right] = \mathbb{P}\left\{\xi_{i;\alpha}^{(1)}\xi_{i;\beta}^{(1)} = 1\right\}$$

This product equals one if and only if both indicators take the value of 1, yielding

$$\mathbb{E}\left[\xi_{i;\alpha}^{(1)}\xi_{i;\beta}^{(1)}\right] = \mathbb{P}\left\{\xi_{i;\alpha}^{(1)} = 1 \wedge \xi_{i;\beta}^{(1)} = 1\right\}$$

We now rewrite each indicator in terms of its equivalent event,

$$\begin{aligned}\mathbb{E}\left[\xi_{i;\alpha}^{(1)}\xi_{i;\beta}^{(1)}\right] &= \mathbb{P}\left\{z_{i;\alpha}^{(1)} > 0 \wedge z_{i;\beta}^{(1)} > 0\right\} \\ &= \mathbb{P}\left\{x_\alpha \cdot W_i^{(1)} > 0 \wedge x_\beta \cdot W_i^{(1)} > 0\right\}\end{aligned}$$

where $W_i^{(1)} \in \mathbb{R}^{n_0}$ is the vector of weights connecting each input node to the ith neuron in the first layer,

$$W_i^{(1)} = \left[W_{ij}^{(1)}\right]_{j=1}^{n_0}$$

We have that each $W_{ij}^{(1)}$ is i.i.d. normal with mean zero. We now consider the unit vector

$$V_i^{(1)} = \frac{W_i^{(1)}}{\|W_i^{(1)}\|}$$

$V_i^{(1)}$ has the same orientation as $W_i^{(1)}$ because it is defined as the corresponding unit vector. It is known that $W_i^{(1)}$ is oriented with a uniform distribution in every direction because its PDF is rotationally invariant, and also that $V_i^{(1)}$ has unit length. Denote the unit n_0-sphere by S^{n_0}. Then, we may say that $V_i^{(1)}$ is distributed uniformly over the surface of S^{n_0}. Therefore,

$$W_i^{(1)} \cdot x_\alpha > 0 \iff V_i^{(1)} \cdot x_\alpha > 0$$

$$\implies \mathbb{E}\left[\xi_{i;\alpha}^{(1)}\xi_{i;\beta}^{(1)}\right] = \mathbb{P}\left\{x_\alpha \cdot V_i^{(1)} > 0 \wedge x_\beta \cdot V_i^{(1)} > 0\right\} = \mathbb{P}\left\{V_i^{(1)} \in \mathcal{C}\right\}$$

where

$$\mathcal{C} = \{v \in S^{n_0} : v \cdot x_\alpha > 0 \wedge v \cdot x_\beta > 0\}$$

We can also describe $\mathcal{C}$ as the set of vectors on the unit sphere that makes an angle of less than $\frac{\pi}{2}$ with both x_α and x_β. This set is the intersection of the two n_0-sphere caps that are centered around x_α and x_β, with azimuthal angle $\frac{\pi}{2}$. Since $V_i^{(1)}$ has uniform distribution over S^{n_0}, we can say that its probability of being in $\mathcal{C} \subseteq S^{n_0}$ is proportional to the surface area of $\mathcal{C}$, and we can normalize this probability to find

$$\mathbb{E}\left[\xi_{i;\alpha}^{(1)}\xi_{i;\beta}^{(1)}\right] = \frac{\text{surface area of } \mathcal{C}}{\text{surface area of } S^{n_0}}$$

We denote the surface area of S^{n_0} by A_n.

There is a known, closed form expression for the surface area of $\mathcal{C}$, given below. This is derived from a general result for the surface area of a hypercap intersection [6]; we note that $\mathcal{C}$ is the intersection of two hemi-n_0-spheres to get

$$\text{surface area of } \mathcal{C} = \begin{cases} \frac{1}{2}A_n - J_n^{\frac{\pi}{2}-\theta_{\alpha\beta}}, & x_\alpha \cdot x_\beta \geq 0 \\ J_n^{\theta_{\alpha\beta}-\frac{\pi}{2}}, & x_\alpha \cdot x_\beta < 0 \end{cases},$$

where $\theta_{\alpha\beta}$ is the angle between x_α and x_β. Dividing by A_n, we recover the claim of this theorem. □

Remark. *This approach develops some nice geometric intuition for the correlation between a particular neuron's activations for different datapoints. We are looking for the surface area of a hypersphere on which a unit vector (the $V_i^{(l)}$) has a positive dot product (is aligned) with both datapoints. The further the datapoints are from each other, the smaller the overlap of this area, and the less likely the activations are to be correlated. This correlation passes through the neural network, but is warped by nonlinearities and rectifiers in a way that we will discuss in later sections. Also, if the data points are perfectly aligned, the intersection $\mathcal{C}$ is simply a single hemi-n_0-sphere, and so $\mathbb{E}\left[\xi_{i;\alpha}^{(1)}\xi_{i;\beta}^{(1)}\right] = \frac{1}{2}$. This matches Corollary 1.1.*

2.2 $\mathbb{E}\left[\xi_{i;\alpha}^{(l)}\xi_{i;\beta}^{(l)}\right]$ in Layers $l > 1$

For the later layers, we can apply a naive bound to get a closed form bound for the NTK elements. This is performed in the following theorem.

Theorem 2.2. There exists an upper bound for the expectation of each matrix element of the NTK. Each entry satisfies

$$\boxed{|\mathbb{E}\left[K_{\alpha\beta}\right]| \leq |x_\alpha \cdot x_\beta| \frac{4L}{n_0} \mathbb{E}\left[\xi_{i;\alpha}^{(1)} \xi_{i;\beta}^{(1)}\right]}$$

Proof. From Theorem 1.1, we have that

$$\mathbb{E}\left[K_{\alpha\beta}\right] = (x_\alpha \cdot x_\beta) \frac{L}{n_0} \left(2^{L+1}\right) \prod_{l=1}^{L} \mathbb{E}\left[\xi_{i;\alpha}^{(l)} \xi_{i;\beta}^{(l)}\right]$$

Consider the following bound of the two-point correlator.

$$\begin{aligned} \forall l, \quad \mathbb{E}\left[\xi_{i;\alpha}^{(l)} \xi_{i;\beta}^{(l)}\right] &= \mathbb{P}\left\{\xi_{i;\alpha}^{(l)} = 1 \wedge \xi_{i;\beta}^{(l)} = 1\right\} \\ &= \mathbb{P}\left\{\xi_{i;\alpha}^{(l)} = 1 | \xi_{i;\beta}^{(l)} = 1\right\} \mathbb{P}\left\{\xi_{i;\beta}^{(l)} = 1\right\} \\ &= \frac{1}{2} \mathbb{P}\left\{\xi_{i;\alpha}^{(l)} = 1 | \xi_{i;\beta}^{(l)} = 1\right\} \end{aligned}$$

But all probabilities $\mathbb{P}\{x\} \leq 1$, so

$$\forall l, \quad \mathbb{E}\left[\xi_{i;\alpha}^{(l)} \xi_{i;\beta}^{(l)}\right] \leq \frac{1}{2}$$

From Theorem 2.1, we have an expression for $\mathbb{E}\left[\xi_{i;\alpha}^{(1)} \xi_{i;\beta}^{(1)}\right]$. So if $L \geq 2$,

$$\begin{aligned} |\mathbb{E}[K_{\alpha\beta}]| &= |x_\alpha \cdot x_\beta| \frac{L}{n_0} \left(2^{L+1}\right) \prod_{l=1}^{L} \mathbb{E}\left[\xi_{i;\alpha}^{(l)} \xi_{i;\beta}^{(l)}\right] \\ &= |x_\alpha \cdot x_\beta| \frac{L}{n_0} \left(2^{L+1}\right) \mathbb{E}\left[\xi_{i;\alpha}^{(1)} \xi_{i;\beta}^{(1)}\right] \prod_{l=2}^{L} \mathbb{E}\left[\xi_{i;\alpha}^{(l)} \xi_{i;\beta}^{(l)}\right] \\ &\leq |x_\alpha \cdot x_\beta| \frac{L}{n_0} \left(2^{L+1}\right) \mathbb{E}\left[\xi_{i;\alpha}^{(1)} \xi_{i;\beta}^{(1)}\right] \cdot 2^{-(L-1)} \end{aligned}$$

We simplify and conclude that

$$\boxed{|\mathbb{E}\left[K_{\alpha\beta}\right]| \leq |x_\alpha \cdot x_\beta| \frac{4L}{n_0} \mathbb{E}\left[\xi_{i;\alpha}^{(1)} \xi_{i;\beta}^{(1)}\right]}$$

where we have a closed form expression for $\mathbb{E}\left[\xi_{i;\alpha}^{(1)} \xi_{i;\beta}^{(1)}\right]$ in terms of x_α and x_β. □

Remark. *There are certainly tighter bounds to be made on $\mathbb{E}\left[\xi_{i;\alpha}^{(l)} \xi_{i;\beta}^{(l)}\right]$ than the naive $\frac{1}{2}$. We conjecture that the ξ's for different inputs become more uncorrelated with increasing width and depth. Intuitively, the more complex the model is, the harder it is for two different datapoints to yield the same activation pattern, especially deeper in the network. At later layers, the randomness of the weights should complicate the activations enough that things appear independent; in this setting, we expect $\mathbb{E}\left[\xi_{i;\alpha}^{(l)} \xi_{i;\beta}^{(l)}\right]$ to approach $\frac{1}{4}$. A rigorous study of the asymptotics of the two-point correlator has eluded us so far during this research, but would tighten our bounds and allow stronger claims to be made in the regime of wide and deep ReLU nets.*

Note that the geometric intuition for when the same neuron is activated by two different inputs in terms of cap intersection in hyperspace still holds deeper in the network. Things get tricky because the two vectors that generate the caps in the first layer (i.e. the datapoints) are fixed, but in later layers are governed by random vectors

with rectified Gaussian distributions (i.e. the postactivations). In this case there is a nonuniform directional distribution of these two vectors, with the density functions being more concentrated along the borders of the positive quadrant as a result of the rectification. This introduces casework and messiness into the hemisphere intersection argument, but this geometric framework still could be explored in a way that would be informative. Understanding of the high dimensional geometry of nonlinearly transformed Gaussians is a generally useful area of study, and perhaps broader results could be attained through the two-point correlator explored in this paper.

3 Conclusion and Future Work

In this work, we have investigated the distribution of all elements of the NTK at initialization in a common practical architecture (ReLU network with Kaiming initialization). Building on the results of [2] about the distribution of the on-diagonal elements of the NTK, we derived a general form for the expectations of *all* elements of the NTK matrix. This result reveals the importance of studying the two-point correlator $\mathbb{E}\left[\xi^{(l)}_{\gamma;\alpha}\xi^{(l)}_{\gamma;\beta}\right]$ and how it evolves with increasing depth. We found a closed form expression for this correlator in the first layer (allowing us to write down a closed form expression for the full expected NTK of a 1-hidden-layer network) and developed a geometric approach and intuition for reasoning about its evolution in later layers.

There are many different directions for future work that we would like to (and likely will) pursue. We enumerate a few seemingly promising ones below.

1. Given the distributional understanding we have of the NTK on initialization (our results about the expectation and results about the variance of the diagonal from [2]), we can apply results from random matrix theory to investigate the concentration of the spectrum of the NTK on initialization. With the current understanding that the top eigenvalue of the initialized NTK sheds light on how the learning rate governs later stage training dynamics (see [7]), concentration-style bounds on the top eigenvalue of the NTK at initialization could have large practical and widespread use. A high level trajectory of a potential approach that we are currently attempting is described below.

   (a) Using our results, we can create concentration bounds on the top eigenvalue of the *expected NTK* using the upper bounds granted by the corollaries, matrix norms, and Chebyshev's Inequality.

   (b) With this, we can apply matrix concentration bounds such as the Matrix Chernoff Inequality in order to bound the tails of the distribution of the top eigenvalue of the NTK [8].

   (c) This would loosely grant us knowledge about a data-dependent way to understand some late stage training dynamics via our study of NTK initialization, which seems attractive.

   It would be constructive to investigate the variance of the off diagonal elements as well, as this would allow a more complete distributional description of the NTK on initialization. Pursuit of this result would look much like the work done in [2] and would require statistical handling of pairs/quadruples of shared paths.

2. It appears that it would be greatly useful to obtain a more complete description of the dynamics of the two-point correlator $\mathbb{E}\left[\xi^{(l)}_{\gamma;\alpha}\xi^{(l)}_{\gamma;\beta}\right]$ during forward propagation. This is a very rich area of study, as we could imagine tools from dynamical systems or statistical physics being used to complement our geometric perspective. This would grant more power to Theorem 1.1 and also any description of the off-diagonal variances, which would certainly include this correlator as well.

3. Taking a step back, we can view the product $\prod_l \mathbb{E}\left[\xi^{(l)}_{\gamma;\alpha}\xi^{(l)}_{\gamma;\beta}\right]$ that appears in the result of Theorem 1.1 from a higher-level viewpoint. Since the two-point correlators are independent between layers, this expression is equivalent to the probability that two inputs x_α, x_β result in precisely the same totally-on activation pattern

in the initialized network. A complete study of this distribution of activation patterns as a function of inputs could shed interesting light on the theory of the complexity/capacity of ReLU networks on initialization. Current theory focuses mainly on number of linear regions and number of different activation patterns one could see with decent probability (see [4], [5]), but an investigation into when two data points activate the network the same way, facilitated by $\prod_l \mathbb{E}\left[\xi^{(l)}_{\gamma;\alpha}\xi^{(l)}_{\gamma;\beta}\right]$, could have broad implications and use cases.

References

[1] Boris Hanin. *Which Neural Net Architectures Give Rise To Exploding and Vanishing Gradients?* 2018. DOI: 10.48550/ARXIV.1801.03744. URL: https://arxiv.org/abs/1801.03744.

[2] Boris Hanin and Mihai Nica. *Finite Depth and Width Corrections to the Neural Tangent Kernel*. 2019. DOI: 10.48550/ARXIV.1909.05989. URL: https://arxiv.org/abs/1909.05989.

[3] Boris Hanin and Mihai Nica. "Products of Many Large Random Matrices and Gradients in Deep Neural Networks". In: *Communications in Mathematical Physics* 376.1 (Dec. 2019), pp. 287–322. DOI: 10.1007/s00220-019-03624-z. URL: https://doi.org/10.1007%2Fs00220-019-03624-z.

[4] Boris Hanin and David Rolnick. *Complexity of Linear Regions in Deep Networks*. 2019. DOI: 10.48550/ARXIV.1901.09021. URL: https://arxiv.org/abs/1901.09021.

[5] Boris Hanin and David Rolnick. "Deep ReLU Networks Have Surprisingly Few Activation Patterns". In: (2019). DOI: 10.48550/ARXIV.1906.00904. URL: https://arxiv.org/abs/1906.00904.

[6] Yongjae Lee and Woo Chang Kim. *Concise Formulas for the Surface Area of the Intersection of Two Hyperspherical Caps*. Feb. 2014.

[7] Aitor Lewkowycz et al. *The large learning rate phase of deep learning: the catapult mechanism*. 2020. DOI: 10.48550/ARXIV.2003.02218. URL: https://arxiv.org/abs/2003.02218.

[8] Joel A. Tropp. *An Introduction to Matrix Concentration Inequalities*. 2015. DOI: 10.48550/ARXIV.1501.01571. URL: https://arxiv.org/abs/1501.01571.