Invariance of Weight Distributions in Rectified MLPs

Russell Tsuchida [1] **Farbod Roosta-Khorasani** [2 3] **Marcus Gallagher** [1]

Abstract

An interesting approach to analyzing neural networks that has received renewed attention is to examine the equivalent kernel of the neural network. This is based on the fact that a fully connected feedforward network with one hidden layer, a certain weight distribution, an activation function, and an infinite number of neurons can be viewed as a mapping into a Hilbert space. We derive the equivalent kernels of MLPs with ReLU or Leaky ReLU activations for all rotationally-invariant weight distributions, generalizing a previous result that required Gaussian weight distributions. Additionally, the Central Limit Theorem is used to show that for certain activation functions, kernels corresponding to layers with weight distributions having 0 mean and finite absolute third moment are asymptotically universal, and are well approximated by the kernel corresponding to layers with spherical Gaussian weights. In deep networks, as depth increases the equivalent kernel approaches a pathological fixed point, which can be used to argue why training randomly initialized networks can be difficult. Our results also have implications for weight initialization.

1. Introduction

Neural networks have recently been applied to a number of diverse problems with impressive results (van den Oord et al., 2016; Silver et al., 2017; Berthelot et al., 2017). These breakthroughs largely appear to be driven by application rather than an understanding of the capabilities and training of neural networks. Recently, significant work has been done to increase understanding of neural networks (Choromanska et al., 2015; Haeffele & Vidal, 2015; Poole et al., 2016; Schoenholz et al., 2017; Zhang et al., 2016; Martin & Mahoney, 2017; Shwartz-Ziv & Tishby, 2017; Balduzzi et al., 2017; Raghu et al., 2017). However, there is still work to be done to bring theoretical understanding in line with the results seen in practice.

The connection between neural networks and kernel machines has long been studied (Neal, 1994). Much past work has been done to investigate the equivalent kernel of certain neural networks, either experimentally (Burgess, 1997), through sampling (Sinha & Duchi, 2016; Livni et al., 2017; Lee et al., 2017), or analytically by assuming some random distribution over the weight parameters in the network (Williams, 1997; Cho & Saul, 2009; Pandey & Dukkipati, 2014a;b; Daniely et al., 2016; Bach, 2017a). Surprisingly, in the latter approach, rarely have distributions other than the Gaussian distribution been analyzed. This is perhaps due to early influential work on Bayesian Networks (MacKay, 1992), which laid a strong mathematical foundation for a Bayesian approach to training networks. Another reason may be that some researchers may hold the intuitive (but *not necessarily principled*) view that the Central Limit Theorem (CLT) should somehow apply.

In this work, we investigate the equivalent kernels for networks with Rectified Linear Unit (ReLU), Leaky ReLU (LReLU) or other activation functions, one-hidden layer, and more general weight distributions. Our analysis carries over to deep networks. We investigate the consequences that weight initialization has on the equivalent kernel at the beginning of training. While initialization schemes that mitigate exploding/vanishing gradient problems (Hochreiter, 1991; Bengio et al., 1994; Hochreiter et al., 2001) for other activation functions and weight distribution combinations have been explored in earlier works (Glorot & Bengio, 2010; He et al., 2015), we discuss an initialization scheme for Muli-Layer Perceptrons (MLPs) with LReLUs and weights coming from distributions with 0 mean and finite absolute third moment. The derived kernels also allow us to analyze the loss of information as an input is propagated through the network, offering a complementary view to the shattered gradient problem (Balduzzi et al., 2017).

[1]School of ITEE, University of Queensland, Brisbane, Queensland, Australia [2]School of Mathematics and Physics, University of Queensland, Brisbane, Queensland, Australia [3]International Computer Science Institute, Berkeley, California, USA. Correspondence to: Russell Tsuchida <s.tsuchida@uq.edu.au>, Farbod Roosta-Khorasani <fred.roosta@uq.edu.au>, Marcus Gallagher <marcusg@uq.edu.au>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

2. Preliminaries

Consider a fully connected (FC) feedforward neural network with m inputs and a hidden layer with n neurons. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the activation function of all the neurons in the hidden layer. Further assume that the biases are 0, as is common when initializing neural network parameters. For any two inputs $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ propagated through the network, the dot product in the hidden layer is

$$\frac{1}{n}\boldsymbol{h}(\mathbf{x}) \cdot \boldsymbol{h}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} \sigma(\mathbf{w_i} \cdot \mathbf{x})\sigma(\mathbf{w_i} \cdot \mathbf{y}), \qquad (1)$$

where $\boldsymbol{h}(\cdot)$ denotes the n dimensional vector in the hidden layer and $\mathbf{w_i} \in \mathbb{R}^m$ is the weight vector into the i^{th} neuron. Assuming an infinite number of hidden neurons, the sum in (1) has an interpretation as an inner product in feature space, which corresponds to the kernel of a Hilbert space. We have

$$k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^m} \sigma(\mathbf{w} \cdot \mathbf{x})\sigma(\mathbf{w} \cdot \mathbf{y}) f(\mathbf{w})\, \mathbf{dw}, \qquad (2)$$

where $f(\mathbf{w})$ is the probability density function (PDF) for the identically distributed weight vector $\boldsymbol{W} = (W_1, ..., W_m)^T$ in the network. The connection of (2) to the kernels in kernel machines is well-known (Neal, 1994; Williams, 1997; Cho & Saul, 2009).

Probabilistic bounds for the error between (1) and (2) have been derived in special cases (Rahimi & Recht, 2008) when the kernel is shift-invariant. Two specific random feature mappings are considered: **(1)** Random Fourier features are taken for the σ in (1). Calculating the approximation error in this way requires being able to sample from the PDF defined by the Fourier transform of the target kernel. More explicitly, the weight distribution f is the Fourier transform of the target kernel and the n samples $\sigma(\mathbf{w_i} \cdot \mathbf{x})$ are replaced by some appropriate scale of $\cos(\mathbf{w_i} \cdot \mathbf{x})$. **(2)** A random bit string $\sigma(\mathbf{x_i})$ is associated to each input according to a grid with random pitch δ sampled from f imposed on the input space. This method requires having access to the second derivative of the target kernel to sample from the distribution f.

Other work (Bach, 2017b) has focused on the smallest error between a target function g in the reproducing kernel Hilbert space (RKHS) defined by (2) and an approximate function $\hat{g}$ expressible by the RKHS with the kernel (1). More explicitly, let $g(x) = \int_{\mathbb{R}^m} G(\mathbf{w})\sigma(\mathbf{w}, \mathbf{x}) f(\mathbf{w})\, d\mathbf{w}$ be the representation of g in the RKHS. The quantity $\left\|\hat{g} - g\right\| = \left\|\sum_{i=1}^{n} \alpha_i \sigma(\mathbf{w_i}, \cdot) - \int_{\mathbb{R}^m} G(\mathbf{w})\sigma(\mathbf{w}, \cdot) f(\mathbf{w})\, d\mathbf{w}\right\|$ (with some suitable norm) is studied for the best set of α_i and random $\mathbf{w_i}$ with an optimized distribution.

Yet another measure of kernel approximation error is investigated by Rudi & Rosasco (2017). Let $\hat{g}$ and g be the optimal solutions to the ridge regression problem of minimizing a regularized cost function C using the kernel (1) and the kernel (2) respectively. The number of datapoints n required to probabilistically bound $C(\hat{g}) - C(g)$ is found to be $O(\sqrt{n}\log n)$ under a suitable set of assumptions. This work notes the connection between kernel machines and one-layer Neural Networks with ReLU activations and Gaussian weights by citing Cho & Saul (2009). We extend this connection by considering other weight distributions and activation functions.

In this work our focus is on deriving expressions for the target kernel, not the approximation error. Additionally, we consider random mappings that have not been considered elsewhere. Our work is related to work by Poole et al. (2016) and Schoenholz et al. (2017). However, our results apply to the unbounded (L)ReLU activation function and more general weight distributions, and their work considers random biases as well as weights.

3. Equivalent Kernels for Infinite Width Hidden Layers

The kernel (2) has previously been evaluated for a number of choices of f and σ (Williams, 1997; Roux & Bengio, 2007; Cho & Saul, 2009; Pandey & Dukkipati, 2014a;b). In particular, the equivalent kernel for a one-hidden layer network with spherical Gaussian weights of variance $\mathbb{E}[W_i^2]$ and mean 0 is the Arc-Cosine Kernel (Cho & Saul, 2009)

$$k(\mathbf{x}, \mathbf{y}) = \frac{\mathbb{E}[W_i^2]\|\mathbf{x}\|\|\mathbf{y}\|}{2\pi}\big(\sin\theta_0 + (\pi - \theta_0)\cos\theta_0\big), \quad (3)$$

where $\theta_0 = \cos^{-1}\left(\frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}\right)$ is the angle between the inputs $\mathbf{x}$ and $\mathbf{y}$ and $\|\cdot\|$ denotes the ℓ^2 norm. Noticing that the Arc-Cosine Kernel $k(\mathbf{x}, \mathbf{y})$ depends on $\mathbf{x}$ and $\mathbf{y}$ only through their norms, with an abuse of notation we will henceforth set $k(\mathbf{x}, \mathbf{y}) \equiv k(\theta_0)$. Define the *normalized kernel* to be the cosine similarity between the signals in the hidden layer. The normalized Arc-Cosine Kernel is given by

$$\cos\theta_1 = \frac{k(\mathbf{x}, \mathbf{y})}{\sqrt{k(\mathbf{x}, \mathbf{x})}\sqrt{k(\mathbf{y}, \mathbf{y})}} = \frac{1}{\pi}\big(\sin\theta_0 + (\pi - \theta_0)\cos\theta_0\big),$$

where θ_1 is the angle between the signals in the first layer. Figure 1 shows a plot of the normalized Arc-Cosine Kernel. One might ask how the equivalent kernel changes for a different choice of weight distribution. We investigate the equivalent kernel for networks with (L)ReLU activations and general weight distributions in Section 3.1 and 3.2. The equivalent kernel can be composed and applied to deep networks. The kernel can also be used to choose good weights for initialization. These, as well as other implications for practical neural networks, are investigated in Section 5.

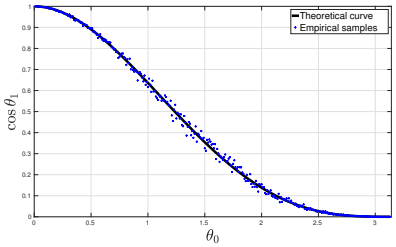


Figure 1. Normalized Arc-Cosine Kernel as a function of θ_0 for a single hidden layer network, Gaussian weights, and ReLU activations. Empirical samples from a network with 1000 inputs and 1000 hidden units are plotted alongside the theoretical curve. Samples are obtained by generating R from a QR decomposition of a random matrix, then setting $\mathbf{x} = R^T(1, 0, ..., 0)^T$ and $\mathbf{y} = R^T(\cos\theta, \sin\theta, 0, ..., 0)^T$.

3.1. Kernels for Rotationally-Invariant Weights

In this section we show that (3) holds more generally than for the case where f is Gaussian. Specifically, (3) holds when f is any rotationally invariant distribution. We do this by casting (2) as the solution to an ODE, and then solving the ODE. We then extend this result using the same technique to the case where σ is LReLU.

A rotationally-invariant PDF one with the property $f(\mathbf{w}) = f(R\mathbf{w}) = f(\|\mathbf{w}\|)$ for all $\mathbf{w}$ and orthogonal matrices R. Recall that the class of rotationally-invariant distributions (Bryc, 1995), as a subclass of elliptically contoured distributions (Johnson, 2013), includes the Gaussian distribution, the multivariate t-distribution, the symmetric multivariate Laplace distribution, and symmetric multivariate stable distributions.

Proposition 1. *Suppose we have a one-hidden layer feed-forward network with ReLU σ and random weights $\mathbf{W}$ with uncorrelated and identically distributed rows with rotationally-invariant PDF $f : \mathbb{R}^m \to \mathbb{R}$ and $\mathbb{E}[W_i^2] < \infty$. The equivalent kernel of the network is* (3)*.*

Proof. First, we require the following.

Proposition 2. *With the conditions in Proposition 1 and inputs $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ the equivalent kernel of the network is the solution to the Initial Value Problem (IVP)*

$$k''(\theta_0) + k(\theta_0) = F(\theta_0), \quad k'(\pi) = 0, \quad k(\pi) = 0, \quad (4)$$

where $\theta_0 \in (0, \pi)$ is the angle between the inputs $\mathbf{x}$ and $\mathbf{y}$. The derivatives are meant in the distributional sense; they are functionals applying to all test functions in $C_c^\infty(0, \pi)$. $F(\theta_0)$ is given by the $m-1$ dimensional integral

$$F(\theta_0) = \int_{\mathbb{R}^{m-1}} f\Big((s\sin\theta_0, -s\cos\theta_0, w_3, ..., w_m)^T\Big) \Theta(s)s^3\, ds\, dw_3\, dw_4 ... dw_m \|\mathbf{x}\|\|\mathbf{y}\| \sin\theta_0, \quad (5)$$

where Θ is the Heaviside step function.

The proof is given in Appendix A. The main idea is to rotate $\mathbf{w}$ (following Cho & Saul (2009)) so that

$$k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^m} \Theta(w_1)\Theta(w_1\cos\theta_0 + w_2\sin\theta_0) w_1 (w_1\cos\theta_0 + w_2\sin\theta_0) f(\mathbf{w})\, d\mathbf{w}\|\mathbf{x}\|\|\mathbf{y}\|.$$

Now differentiating twice with respect to θ_0 yields the second order ODE (4). The usefulness of the ODE in its current form is limited, since the forcing term $F(\theta_0)$ as in (5) is difficult to interpret. However, regardless of the underlying distribution on weights $\mathbf{w}$, as long as the PDF f in (5) corresponds to any rotationally-invariant distribution, the integral enjoys a much simpler representation.

Proposition 3. *With the conditions in Proposition 1, the forcing term $F(\theta_0)$ in the kernel ODE is given by $F(\theta_0) = K\sin\theta_0$, where*

$$K = \int_{\mathbb{R}^{m-1}} \Theta(s)s^3 f\big((s, 0, w_3, ..., w_m)^T\big)\, ds\, dw_3, ... dw_m \|\mathbf{x}\|\|\mathbf{y}\| < \infty,$$

and the solution to the distributional ODE (4) is the solution to the corresponding classical ODE.

The proof is given in Appendix B.

Note that in the representation $F(\theta_0) = K\sin\theta_0$ of the forcing term, the underlying distribution appears only as a constant K. For all rotationally-invariant distributions, the forcing term in (4) results in an equivalent kernel with the *same form*. We can combine Propositions 2 and 3 to find the equivalent kernel assuming rotationally-invariant weight distributions.

Due to the rotational invariance of f, $k(0) = \int_{\mathbb{R}^m} \Theta(w_1)w_1^2 f(R\mathbf{w})\, d\mathbf{w}\|\mathbf{x}\|\|\mathbf{y}\| = \frac{\|\mathbf{x}\|\|\mathbf{y}\|\mathbb{E}[W_i^2]}{2}$. The solution to the ODE in Proposition 2 using the forcing term from Proposition 3 is $k(\theta_0) = c_1\cos\theta_0 + c_2\sin\theta_0 - \frac{1}{2}K\theta_0\cos\theta_0$. Using the conditions from the IVP and $k(0)$, the values of c_1, c_2 and K give the required result. □

One can apply the same technique to the case of LReLU activations $\sigma(z) = \big(a + (1-a)\Theta(z)\big)z$, where a specifies the gradient of the activation for $z < 0$.

Proposition 4. *Consider the same situation as in Proposition 1 with the exception that the activations are LReLU. The integral (2) is then given by*

$$k(\mathbf{x}, \mathbf{y}) = \Big[\frac{(1-a)^2}{2\pi}\big(\sin\theta_0 + (\pi - \theta_0)\cos\theta_0\big) + a\cos\theta_0\Big] \mathbf{E}[W_i^2]\|\mathbf{x}\|\|\mathbf{y}\|, \quad (6)$$

where $a \in [0, 1)$ is the LReLU gradient parameter.

This is just a slightly more involved calculation than the ReLU case; we defer our proof to the supplementary material.

3.2. Asymptotic Kernels

In this section we approximate k for large m and more general weight PDFs. We invoke the CLT as $m \to \infty$, which requires a condition that we discuss briefly before presenting it formally. The dot product $\mathbf{w}\cdot\mathbf{x}$ can be seen as a linear combination of the weights, with the coefficients corresponding to the coordinates of $\mathbf{x}$. Roughly, such a linear combination will obey the CLT if many coefficients are non-zero. To let $m \to \infty$, we *construct* a sequence of inputs $\{\mathbf{x}^{(m)}\}_{m=2}^{\infty}$. This may appear unusual in the context of neural networks, since m is fixed and finite in practice. The sequence is used only for asymptotic analysis.

As an example if the dataset were CelebA (Liu et al., 2015) with 116412 inputs, one would have $\mathbf{x}^{(116412)}$. To generate an artificial sequence, one could down-sample the image to be of size 116411, 116410, and so on. At each point in the sequence, one could normalize the point so that its ℓ^2 norm is $\|\mathbf{x}^{(116412)}\|$. One could similarly up-sample the image.

Intuitively, if the up-sampled image does not just insert zeros, as m increases the we expect the ratio $\frac{|x_i^{(m)}|}{\|\mathbf{x}^{(m)}\|}$ to decrease because the denominator stays fixed and the numerator gets smaller. In our proof the application of CLT requires $\max_{i=1}^{m} \frac{|x_i^{(m)}|}{\|\mathbf{x}^{(m)}\|}$ to decrease faster than $m^{1/4}$. Hypothesis 5 states this condition precisely.

Hypothesis 5. *For $\mathbf{x}^{(m)}, \mathbf{y}^{(m)} \in \mathbb{R}^m$, define sequences of inputs $\{\mathbf{x}^{(m)}\}_{m=2}^{\infty}$ and $\{\mathbf{y}^{(m)}\}_{m=2}^{\infty}$ with fixed $\|\mathbf{x}^{(m)}\|=\|\mathbf{x}\|$, $\|\mathbf{y}^{(m)}\|=\|\mathbf{y}\|$, and $\theta_0=\cos^{-1}\frac{\mathbf{x}^{(m)}\cdot\mathbf{y}^{(m)}}{\|\mathbf{x}\|\|\mathbf{y}\|}$ for all m.*

Letting $x_i^{(m)}$ be the i^{th} coordinate of $\mathbf{x}^{(m)}$, assume that $\lim_{m\to\infty} m^{(1/4)} \max_{i=1}^{m} \frac{|x_i^{(m)}|}{\|\mathbf{x}\|}$ and $\lim_{m\to\infty} m^{(1/4)} \max_{i=1}^{m} \frac{|y_i^{(m)}|}{\|\mathbf{y}\|}$ are both 0.

Figures 2 and 5 empirically investigate Hypothesis 5 for two datasets, suggesting it makes reasonable assumptions on high dimensional data such as images and audio.

Theorem 6. *Consider an infinitely wide FC layer with almost everywhere continuous activation functions σ. Suppose the random weights $\mathbf{W}$ come from an IID distribution with PDF f_m such that $\mathbf{E}[W_i] = 0$ and $\mathbf{E}|W_i^3| < \infty$. Suppose that the conditions in Hypothesis 5 are satisfied. Then*

$$\sigma(\mathbf{W}^{(m)} \cdot \mathbf{x}^{(m)})\sigma(\mathbf{W}^{(m)} \cdot \mathbf{y}^{(m)}) \xrightarrow{D} \sigma(Z_1)\sigma(Z_2),$$

where $\xrightarrow{D}$ denotes convergence in distribution and $\mathbf{Z} = (Z_1, Z_2)^T$ is a Gaussian random vector with covariance matrix $\mathbb{E}[W_i^2]\begin{bmatrix} \|\mathbf{x}\|^2 & \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta_0 \\ \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta_0 & \|\mathbf{y}\|^2 \end{bmatrix}$ and $\mathbf{0}$ mean. Every $\mathbf{Z}^{(m)} = (\mathbf{W}^{(m)} \cdot \mathbf{x}^{(m)}, \mathbf{W}^{(m)} \cdot \mathbf{y}^{(m)})^T$ has the same mean and covariance matrix as $\mathbf{Z}$.

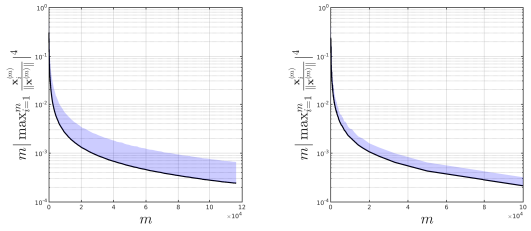


Figure 2. The solid line is an average over 1000 randomly sampled datapoints. The shaded region represents 1 standard deviation in the worst-case direction. Data is preprocessed so that each dimension is in the range $[0, 255]$. (Left) Aligned and cropped CelebA dataset (Liu et al., 2015), with true dimensionality $m = 116412$. The images are compressed using Bicubic Interpolation. (Right) CHiME3_embedded_et05_real live speech data from The 4th CHiME Speech Separation and Recognition Challenge (Vincent et al., 2017; Barker et al., 2017). Each clip is trimmed to a length of 6.25 seconds and the true sample rate is 16000 Hz, so the true dimensionality is $m = 100000$. Compression is achieved through subsampling by integer factors.

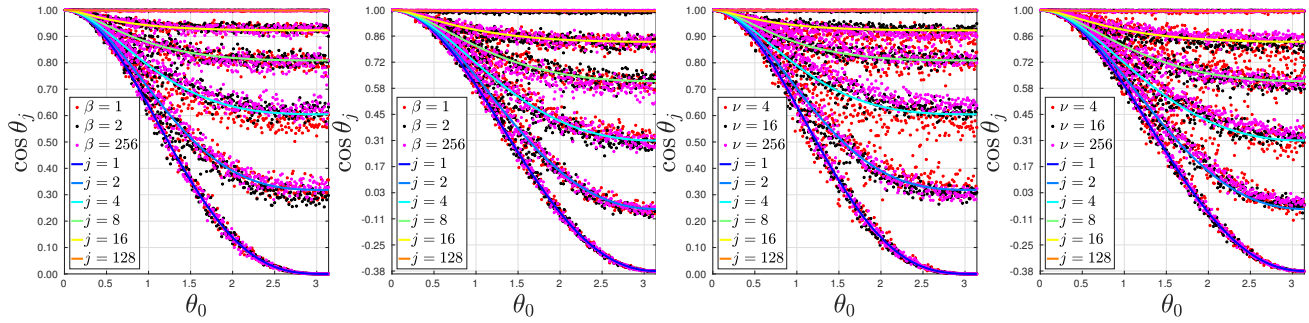
Convergence in distribution is a weak form of convergence, so we cannot expect in general that all kernels should converge asymptotically. For some special cases however, this is indeed possible to show. We first present the ReLU case.

Corollary 7. *Let m, $\mathbf{W}$, f_m, $\mathbf{E}[W_i]$ and $\mathbf{E}|W_i^3|$ be as defined in Theorem 6. Define the corresponding kernel to be $k_f^{(m)}(\mathbf{x}^{(m)}, \mathbf{y}^{(m)})$. Consider a second infinitely wide FC layer with m inputs. Suppose the random weights come from a spherical Gaussian with $\mathbb{E}[W_i] = 0$ and finite variance $\mathbb{E}[W_i^2]$ with PDF g_m. Define the corresponding kernel to be $k_g^{(m)}(\mathbf{x}^{(m)}, \mathbf{y}^{(m)})$. Suppose that the conditions in Hypothesis 5 are satisfied and the activation functions are $\sigma(z) = \Theta(z)z$. Then for all $s \geq 2$,*

$$\lim_{m\to\infty} k_f^{(m)}(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}) = k_g^{(s)}(\mathbf{x}^{(s)}, \mathbf{y}^{(s)}) = \mathbb{E}[\sigma(Z_1)\sigma(Z_2)],$$

where $\mathbf{Z}$ is as in Theorem 6. Explicitly, $k_f^{(m)}$ converges to (3).

The proof is given in Appendix D. This implies that the Arc-Cosine Kernel is well approximated by ReLU layers with weights from a wide class of distributions. Similar results hold for other σ including the LReLU and ELU (Clevert et al., 2016), as shown in the supplementary material.



(a) (b) (c) (d)

Figure 3. Theoretical normalized kernel for networks of increasing depth. Empirical samples from a network with between 1 and 128 hidden layers, 1000 hidden neurons in each layer, $m = 1000$ and weights coming from different symmetric distributions. The sampling process for each θ_0 is as described in Figure 1. The variance is chosen according to (8). (a) ReLU Activations, (7) distribution. (b) LReLU Activations with $a = 0.2$, (7) distribution. (c) ReLU Activations, t-distribution. (d) LReLU Activations with $a = 0.2$, t-distribution.

4. Empirical Verification of Results

We empirically verify our results using two families of weight distributions. First, consider the m-dimensional t-distribution

$$f(\mathbf{w}) = \frac{\Gamma[(\nu+m)/2]}{\Gamma(\nu/2)\nu^{m/2}\pi^{m/2}\sqrt{|\det(\Sigma)|}} \left[1 + \frac{1}{\nu}(\mathbf{w}^T\Sigma^{-1}\mathbf{w})\right]^{-(\nu+m)/2},$$

with degrees of freedom ν and identity shape matrix $\Sigma = I$. The multivariate t-distribution approaches the multivariate Gaussian as $\nu \to \infty$. Random variables drawn from the multivariate t-distribution are uncorrelated but not independent. This distribution is rotationally-invariant and satisfies the conditions in Propositions (1) and (4).

Second, consider the multivariate distribution

$$f(\mathbf{w}) = \prod_{i=1}^{m} \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-|w_i/\alpha|^\beta}, \tag{7}$$

which is not rotationally-invariant (except when $\beta = 2$, which coincides with a Gaussian distribution) but whose random variables are IID and satisfy the conditions in Theorem 6. As $\beta \to \infty$ this distribution converges pointwise to the uniform distribution on $[-\alpha, \alpha]$.

In Figure 3, we empirically verify Propositions 1 and 4. In the one hidden layer case, the samples follow the blue curve $j = 1$, regardless of the specific multivariate t weight distribution which varies with ν. We also observe that the universality of the equivalent kernel appears to hold for the distribution (7) regardless of the value of β, as predicted by theory. We discuss the relevance of the curves $j \neq 1$ in Section 5.

5. Implications for Practical Networks

5.1. Composed Kernels in Deep Networks

A recent advancement in understanding the difficulty in training deep neural networks is the identification of the shattered gradients problem (Balduzzi et al., 2017). Without skip connections, the gradients of deep networks approach white noise as they are backpropagated through the network, making them difficult to train.

A simple observation that complements this view is obtained through repeated composition of the normalized kernel. As $m \to \infty$, the angle between two inputs in the j^{th} layer of a LReLU network random weights with $\mathbb{E}[W] = 0$ and $\mathbb{E}|W^3| < \infty$ approaches $\cos\theta_j = \frac{1}{1+a^2}\left(\frac{(1-a)^2}{\pi}\left(\sin\theta_{j-1} + (\pi - \theta_{j-1})\cos\theta_{j-1}\right) + 2a\cos\theta_0\right)$.

A result similar to the following is hinted at by Lee et al. (2017), citing Schoenholz et al. (2017). Their analysis, which considers biases in addition to weights (Poole et al., 2016), yields insights on the trainability of random neural networks that our analysis cannot. However, their argument does not appear to provide a complete formal proof for the case when the activation functions are unbounded, e.g., ReLU. The degeneracy of the composed kernel with more general activation functions is also proved by Daniely (2016), with the assumption that the weights are Gaussian distributed.

Corollary 8. *The normalized kernel corresponding to LReLU activations converges to a fixed point at* $\theta^* = 0$.

Proof. Let $z = \cos\theta_{j-1}$ and define

$$T(z) = \frac{1}{1+a^2}\left(\frac{(1-a)^2}{\pi}\left(\sqrt{1-z^2} + (\pi - \cos^{-1} z)z\right) + 2az\right).$$

The magnitude of the derivative of T is $\left|1 - \left(\frac{1-a}{1+a}\right)^2 \frac{\cos^{-1} z}{\pi}\right|$

which is bounded above by 1 on $[-1, 1]$. Therefore, T is a contraction mapping. By Banach's fixed point theorem there exists a *unique* fixed point $z^* = \cos\theta^*$. Set $\theta^* = 0$ to verify that $\theta^* = 0$ is a solution, and θ^* is unique. □

Corollary 8 implies that for this deep network, the angle between any two signals at a deep layer approaches 0. No matter what the input is, the kernel "sees" the same thing after accounting for the scaling induced by the norm of the input. Hence, it becomes increasingly difficult to train deeper networks, as much of the information is lost and the outputs will depend merely on the norm of the inputs; the signals decorrelate as they propagate through the layers.

At first this may seem counter-intuitive. An appeal to intuition can be made by considering the corresponding linear network with deterministic and equal weight matrices in each layer, which amounts to the celebrated power iteration method. In this case, the repeated application of a matrix transformation A to a vector v converges to the dominant eigenvector (i.e. the eigenvector corresponding to the largest eigenvalue) of A.

Figure 3 shows that the theoretical normalized kernel for networks of increasing depth closely follows empirical samples from randomly initialized neural networks.

In addition to convergence of direction, by also requiring that $\|\mathbf{x}\| = \|\mathbf{y}\|$ it can be shown that after accounting for scaling, the magnitude of the signals converge as the signals propagate through the network. This is analogous to having the dominant eigenvalue equal to 1 in the power iteration method comparison.

Corollary 9. *The quantity* $\mathbb{E}\Big[\big(\sigma^{(j)}(\mathbf{x}) - \sigma^{(j)}(\mathbf{y})\big)^2\Big]/\mathbb{E}[\sigma^{(j)}(\mathbf{x})^2]$ *in a j-layer random (L)ReLU network of infinite width with random uncorrelated and identically distributed rotationally-invariant weights with $\|\mathbf{x}\|{=}\|\mathbf{y}\|$ approaches 0 as $j \to \infty$.*

Proof. Denote the output of one neuron in the j^{th} layer of a network $\sigma(\boldsymbol{W^{(1)}} \cdot \boldsymbol{\sigma}(...\boldsymbol{\sigma}(W^{(j)}\mathbf{x}))$ by $\sigma^{(j)}(\mathbf{x})$ and let k_j be the kernel for the j-layer network. Then

$$\begin{aligned}&\mathbb{E}\Big[\big(\sigma^{(j)}(\mathbf{x}) - \sigma^{(j)}(\mathbf{y})\big)^2\Big]/\mathbb{E}[\sigma^{(j)}(\mathbf{x})^2]\\ &\quad= \big(k_j(\mathbf{x},\mathbf{x}) - 2k_j(\mathbf{x},\mathbf{y}) + k_j(\mathbf{y},\mathbf{y})\big)/k_j(\mathbf{x},\mathbf{x}),\\ &\quad= 2 - 2\cos\theta_j\end{aligned}$$

which approaches 0 as $j \to \infty$. □

Contrary to the shattered gradients analysis, which applies to gradient based optimizers, our analysis relates to any optimizers that initialize weights from some distribution satisfying conditions in Proposition 4 or Corollary 7. Since information is lost during signal propagation, the network's output shares little information with the input. An optimizer that tries to relate inputs, outputs and weights through a suitable cost function will be "blind" to relationships between inputs and outputs.

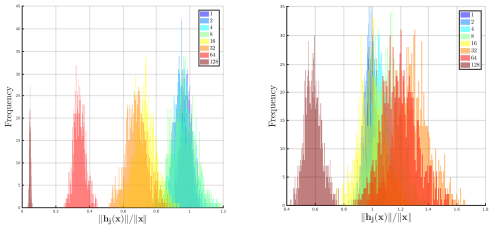


Figure 4. Histograms showing the ratio of the norm of signals in layer j to the norm of the input signals. Each histogram contains 1000 data points randomly sampled from a unit Gaussian distribution. The network tested has 1000 inputs, 1000 neurons in each layer, and LReLU activations with $a = 0.2$. The legend indicates the number of layers in the network. The weights are randomly initialized from a Gaussian distribution. (Left) Weights initialized according to the method of He et al. (2015). (Right) Weights initialized according to (8).

Our results can be used to argue against the utility of controversial Extreme Learning Machines (ELM) (Huang et al., 2004), which randomly initialize hidden layers from symmetric distributions and only learn the weights in the final layer. A single layer ELM can be replaced by kernel ridge regression using the equivalent kernel. Furthermore, a Multi-Layer ELM (Tang et al., 2016) with (L)ReLU activations utilizes a pathological kernel as shown in Figure 3. It should be noted that ELM bears resemblance to early works (Schmidt et al., 1992; Pao et al., 1994).

5.2. Initialization

Suppose we wish to approximately preserve the ℓ^2 norm from the input to hidden layer. By comparing (1) and (2), we approximately have $\|\mathbf{h}(\mathbf{x})\| \approx \sqrt{k(\mathbf{x},\mathbf{x})n}$. Letting $\theta_0 = 0$ in (6), we have $\|\mathbf{h}(\mathbf{x})\| = \|\mathbf{x}\|\sqrt{\frac{n\mathbb{E}[W_i^2](1+a^2)}{2}}$. Setting $\|\mathbf{h}(\mathbf{x})\| = \|\mathbf{x}\|$,

$$\sqrt{\mathbb{E}[W_i^2]} = \sqrt{\frac{2}{\big(1+a^2\big)n}}. \tag{8}$$

This applies whenever the conditions in Proposition 4 or Corollary 12 are satisfied. This agrees with the well-known case when the elements of $\mathbf{W}$ are IID (He et al., 2015) and $a = 0$. For small values of a, (8) is well approximated by the known result (He et al., 2015). For larger values of a, this approximation breaks down, as shown in Figure 4.

An alternative approach to weight initialization is the data-driven approach (Mishkin & Matas, 2016), which can be applied to more complicated network structures such as

convolutional and max-pooling layers commonly used in practice. As parameter distributions change during training, batch normalization inserts layers with learnable scaling and centering parameters at the cost of increased computation and complexity (Ioffe & Szegedy, 2015).

6. Conclusion

We have considered universal properties of MLPs with weights coming from a large class of distributions. We have theoretically and empirically shown that the equivalent kernel for networks with an infinite number of hidden ReLU neurons and all rotationally-invariant weight distributions is the Arc-Cosine Kernel. The CLT can be applied to approximate the kernel for high dimensional input data. When the activations are LReLUs, the equivalent kernel has a similar form. The kernel converges to a fixed point, showing that information is lost as signals propagate through the network.

One avenue for future work is to study the equivalent kernel for different activation functions, noting that some activations such as the ELU may not be expressible in a closed form (we do show in the supplementary material however, that the ELU does have an asymptotically universal kernel).

Since wide networks with centered weight distributions have approximately the same equivalent kernel, powerful trained deep and wide MLPs with (L)ReLU activations should have asymmetric, non-zero mean, non-IID parameter distributions. Future work may consider analyzing the equivalent kernels of trained networks and more complicated architectures. We should not expect that $k(\mathbf{x}, \mathbf{y})$ may be expressed neatly as $k(\theta_0)$ in these cases. This work is a crucial first step in identifying invariant properties in neural networks and sets a foundation from which we hope to expand in future.

A. Proof of Proposition 2

Proof. The kernel with weight PDF $f(\boldsymbol{\omega})$ and ReLU σ is

$$k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^m} \Theta(\boldsymbol{\omega} \cdot \mathbf{x})\Theta(\boldsymbol{\omega} \cdot \mathbf{y})(\boldsymbol{\omega} \cdot \mathbf{x})(\boldsymbol{\omega} \cdot \mathbf{y}) f(\boldsymbol{\omega})\, d\boldsymbol{\omega}.$$

Let θ_0 be the angle between $\mathbf{x}$ and $\mathbf{y}$. Define $\mathbf{u} = (\|\mathbf{x}\|, 0, ..., 0)^T$ and $\mathbf{v} = (\|\mathbf{y}\| \cos\theta_0, \|\mathbf{y}\| \sin\theta_0, 0, ..., 0)^T$ with $\mathbf{u}, \mathbf{v} \in \mathbb{R}^m$. Following Cho & Saul (2009), there exists some $m \times m$ rotation matrix R such that $\mathbf{x} = R\mathbf{u}$ and $\mathbf{y} = R\mathbf{v}$. We have

$$k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^m} \Theta(\boldsymbol{\omega} \cdot R\mathbf{u})\Theta(\boldsymbol{\omega} \cdot R\mathbf{v})(\boldsymbol{\omega} \cdot R\mathbf{u})(\boldsymbol{\omega} \cdot R\mathbf{v}) f(\boldsymbol{\omega})\, d\boldsymbol{\omega}.$$

Let $\boldsymbol{\omega} = R\mathbf{w}$ and note that the dot product is invariant under rotations and the determinant of the Jacobian of the transformation is 1 since R is orthogonal. We have

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= \int_{\mathbb{R}^m} \Theta(\mathbf{w} \cdot \mathbf{u})\Theta(\mathbf{w} \cdot \mathbf{v})(\mathbf{w} \cdot \mathbf{u})(\mathbf{w} \cdot \mathbf{v}) f(R\mathbf{w})\, d\mathbf{w}, \\ &= \int_{\mathbb{R}^m} \Theta(\|\mathbf{x}\| w_1)\Theta(\|\mathbf{y}\|(w_1 \cos\theta_0 + w_2 \sin\theta_0)) w_1(w_1 \cos\theta_0 + w_2 \sin\theta_0) f(\mathbf{w})\, d\mathbf{w} \|\mathbf{x}\| \|\mathbf{y}\|. \end{aligned} \tag{9}$$

One may view the integrand as a functional acting on test functions of θ_0. Denote the set of infinitely differentiable test functions on $(0, \pi)$ by $C_c^\infty(0, \pi)$. The linear functional acting over $C_c^\infty(0, \pi)$ is a Generalized Function and we may take distributional derivatives under the integral by Theorem 7.40 of Jones (1982). Differentiating twice,

$$\begin{aligned} k'' + k &= \int_{\mathbb{R}^m} \Theta(w_1) w_1(-w_1 \sin\theta_0 + w_2 \cos\theta_0)^2 \delta\big(w_1 \cos\theta_0 + w_2 \sin\theta_0\big) f(\mathbf{w})\, d\mathbf{w} \|\mathbf{x}\| \|\mathbf{y}\|, \\ &= \int_{\mathbb{R}^{m-1}} f\Big((s \sin\theta_0, -s \cos\theta_0, w_3, ..., w_m)^T\Big) \Theta(s) s^3\, ds\, dw_3\, dw_4 ... dw_m \|\mathbf{x}\| \|\mathbf{y}\| \sin\theta_0. \end{aligned}$$

The initial condition $k(\pi) = 0$ is obtained by putting $\theta_0 = \pi$ in (9) and noting that the resulting integrand contains a factor of $\Theta(w_1)\Theta(-w_1)w_1$ which is 0 everywhere. Similarly, the integrand of $k'(\pi)$ contains a factor of $\Theta(w_2)\Theta(-w_2)w_2$.

The ODE is meant in a distributional sense, that

$$\int_0^\pi \psi(\theta_0)\big(k''(\theta_0) + k(\theta_0) - F(\theta_0)\big)\, d\theta_0 = 0$$

$\forall \psi \in C_c \infty(0, \pi)$, where k is a distribution with a distributional second derivative k''. □

B. Proof of Proposition 3

Proof. Denote the marginal PDF of the first two coordinates of $\mathbf{W}$ by f_{12}. Due to the rotational invariance of f, $f(O\mathbf{x}) = f(\|\mathbf{x}\|) = f(\mathbf{x})$ for any orthogonal matrix O. So

$$\begin{aligned} F(\theta_0) &= \int_{\mathbb{R}^{m-1}} f\big((s \sin\theta_0, -s \cos\theta_0, w_3, ..., w_m)^T\big) \sin\theta_0 \Theta(s) s^3\, ds\, dw_3, ... dw_m \|\mathbf{x}\| \|\mathbf{y}\|, \\ &= \sin\theta_0 \int_{\mathbb{R}} \Theta(s) s^3 f_{12}\big((s, 0,)^T\big)\, ds \|\mathbf{x}\| \|\mathbf{y}\|, \\ &= K \sin\theta_0, \quad K \in (0, \infty]. \end{aligned}$$

It remains to check that $K < \infty$. F is integrable since

$$\int_{\mathbb{R}^2} \int_0^\pi \Theta(w_1) w_1(-w_1 \sin\theta_0 + w_2 \cos\theta_0)^2 \delta(w_1 \cos\theta_0 + w_2 \sin\theta_0) f_{12}(w_1, w_2) d\theta_0 dw_1 dw_2$$

$$= \int_{\mathbb{R}^2} \Theta(w_1) w_1 \left|(w_1^2+w_2^2)^{1/2}\right| f_{12}(w_1,w_2) dw_1 dw_2,$$
$$\leq \sqrt{\mathbb{E}\left[\Theta^2(W_1)W_1^2\right]}\sqrt{\mathbb{E}\left[W_1^2+W_2^2\right]} < \infty.$$

Therefore, F is finite almost everywhere. This is only true if $K < \infty$. $k'' = F - k$ must be a function, so the distributional and classical derivatives coincide. □

C. Proof of Theorem 6

Proof. There exist some orthonormal $\mathbf{R_1}, \mathbf{R_2} \in \mathbb{R}^m$ such that $\mathbf{y}^{(m)} = \|\mathbf{y}^{(m)}\|(\mathbf{R_1}\cos\theta_0 + \mathbf{R_2}\sin\theta_0)$ and $\mathbf{x}^{(m)} = \|\mathbf{x}^{(m)}\|\mathbf{R_1}$. We would like to examine the asymptotic distribution of $\sigma\big(\|\mathbf{y}^{(m)}\|\mathbf{W}^{(m)}\cdot\big(\mathbf{R_1}\cos\theta_0 + \mathbf{R_2}\sin\theta_0\big)\big)$ $\sigma\big(\|\mathbf{x}^{(m)}\|\mathbf{W}^{(m)}\cdot\mathbf{R_1}\big)$.

Let $U_1^{(m)} = \mathbf{W}\cdot\mathbf{R_1}\cos\theta_0 + \mathbf{W}\cdot\mathbf{R_2}\sin\theta_0$ and $U_2^{(m)} = -\mathbf{W}\cdot\mathbf{R_1}\sin\theta_0 + \mathbf{W}\cdot\mathbf{R_2}\cos\theta_0$. Note that $\mathbb{E}[U_1^{(m)2}] = \mathbb{E}[U_2^{(m)2}] = \mathbb{E}[W_i^2]$ and $\mathbb{E}[U_1^{(m)}] = \mathbb{E}[U_2^{(m)}] = 0$. Also note that $U_1^{(m)}$ and $U_2^{(m)}$ are uncorrelated since $\mathbb{E}[U_1^{(m)}U_2^{(m)}] = \mathbb{E}\Big[(\mathbf{W}\cdot\mathbf{R_1})(\mathbf{W}\cdot\mathbf{R_2})(\cos^2\theta_0 + \sin^2\theta_0) - \cos\theta_0\sin\theta_0\big((\mathbf{W}\cdot\mathbf{R_1})^2 - (\mathbf{W}\cdot\mathbf{R_2})^2\big)\Big] = 0$.

Let $M_k = \mathbb{E}\left|W_i^k\right|$, $\mathbf{U}^{(m)} = (U_1, U_2)^T$, I be the 2×2 identity matrix and $\mathbf{Q} \sim N\big(\mathbf{0}, M_2 I\big)$. Then for any convex set $S \in \mathbb{R}^2$ and some $C \in \mathbb{R}$, by the Berry-Esseen Theorem, $\left|\mathbb{P}[\mathbf{U}\in S] - \mathbb{P}[\mathbf{Q}\in S]\right|^2 \leq C\gamma^2$ where γ^2 is given by

$$\Big(\sum_{j=1}^m \mathbb{E}\Big\|M_2^{\frac{-1}{2}} W_i\, I \begin{pmatrix} R_{1j}\cos\theta_0 + R_{2j}\sin\theta_0 \\ -R_{1j}\sin\theta_0 + R_{2j}\cos\theta_0 \end{pmatrix}\Big\|^3\Big)^2,$$
$$= \Big(M_2^{\frac{-3}{2}} M_3 \sum_{j=1}^m \mathbb{E}\Big\| \begin{pmatrix} R_{1j}\cos\theta_0 + R_{2j}\sin\theta_0 \\ -R_{1j}\sin\theta_0 + R_{2j}\cos\theta_0 \end{pmatrix}\Big\|^3\Big)^2,$$
$$= \Big(M_2^{\frac{-3}{2}} M_3 \sum_{j=1}^m \left|R_{1j}^2 + R_{2j}^2\right|^{(3/2)}\Big)^2,$$
$$\leq M_2^{-3} M_3^2 m \sum_{j=1}^m \left|R_{1j}^2 + R_{2j}^2\right|^3,$$
$$= M_2^{-3} M_3^2 m \sum_{j=1}^m \left|R_{1j}^6 + 3R_{1j}^4 R_{2j}^2 + 3R_{1j}^2 R_{2j}^4 + R_{2j}^6\right|,$$
$$\leq M_2^{-3} M_3^2 m\Big(4\max_{k=1}^m R_{1k}^4 + 4\max_{k=1}^m R_{2k}^4\Big).$$

The last line is due to the fact that

$$\sum_{j=1}^m \left|R_{1j}^6 + 3R_{1j}^4 R_{2j}^2\right| \leq \max_{k=1}^m R_{1k}^4 \Big(\sum_{j=1}^m R_{1j}^2 + 3R_{2j}^2\Big).$$

Now $R_{1k} = \frac{x_k}{\|\mathbf{x}\|}$ and $R_{2k} = \frac{1}{\sin\theta_0}\Big(\frac{y_k}{\|\mathbf{y}\|} - \frac{x_k}{\|\mathbf{x}\|}\cos\theta_0\Big)$, so if $\theta_0 \neq 0, \pi$ by Hypothesis 5 $\mathbf{U}^{(m)}$ converges *in distribution* to the bivariate spherical Gaussian with variance $\mathbb{E}[W_i^2]$. Then the random vector $\mathbf{Z}^{(m)} = (Z_1^{(m)}, Z_2^{(m)})^T = \big(\|\mathbf{x}\|\mathbf{W}\cdot\mathbf{R_1}, \|\mathbf{y}\|(\mathbf{W}\cdot\mathbf{R_1}\cos\theta_0 + \mathbf{W}\cdot\mathbf{R_2}\sin\theta_0)\big)^T = \big(\|\mathbf{x}\|(U_1\cos\theta_0 - U_2\sin\theta_0), \|\mathbf{y}\|U_1\big)^T$ converges *in distribution* to the bivariate Gaussian random variable with covariance matrix $\mathbb{E}[W_i^2]\begin{bmatrix} \|\mathbf{x}\|^2 & \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta_0 \\ \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta_0 & \|\mathbf{y}\|^2 \end{bmatrix}$. Since σ is continuous almost everywhere, by the Continuous Mapping Theorem,

$$\sigma(\mathbf{W}^{(m)}\cdot\mathbf{x}^{(m)})\sigma(\mathbf{W}^{(m)}\cdot\mathbf{y}^{(m)}) \xrightarrow{D} \sigma(Z_1)\sigma(Z_2).$$

If $\theta_0 = 0$ or $\theta_0 = \pi$, we may treat $\mathbf{R_2}$ as $\mathbf{0}$ and the above still holds. □

D. Proof of Corollary 7

Proof. We have $\lim_{m\to\infty} k_f^{(m)}\big(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}\big) = \lim_{m\to\infty}\mathbb{E}\big[\sigma(Z_1^{(m)})\sigma(Z_2^{(m)})\big]$ and would like to bring the limit inside the expected value. By Theorem 6 and Theorem 25.12 of Billingsley (1995), it suffices to show that $\sigma(Z_1^{(m)})\sigma(Z_2^{(m)})$ is uniformly integrable. Define h to be the joint PDF of $\mathbf{Z}^{(m)}$. We have

$$\lim_{\alpha\to\infty}\int_{|\sigma(z_1)\sigma(z_2)|>\alpha} |\sigma(z_1)\sigma(z_2)| h(z_1,z_2)\, dz_1 dz_2$$
$$= \lim_{\alpha\to\infty}\int_{|\Theta(z_1)\Theta(z_2)z_1z_2|>\alpha} |\Theta(z_1)\Theta(z_2)z_1z_2| h(z_1,z_2)\, dz_1 dz_2,$$

but the integrand is 0 whenever $z_1 \leq 0$ *or* $z_2 \leq 0$. So

$$\int_{|\sigma(z_1)\sigma(z_2)|>\alpha} |\sigma(z_1)\sigma(z_2)| h(z_1,z_2)\, dz_1 dz_2$$
$$= \int_{\mathbb{R}^2} z_1 z_2 \Theta(z_1z_2 - \alpha)\Theta(z_1)\Theta(z_2) h(z_1,z_2)\, dz_1 dz_2.$$

We may raise the Heaviside functions to any power without changing the value of the integral. Squaring the Heaviside functions and applying Hölder's inequality, we have

$$\Big(\int_{\mathbb{R}^2} z_1 z_2 \Theta^2(z_1z_2-\alpha)\Theta^2(z_1)\Theta^2(z_2) h(z_1,z_2) dz_1 dz_2\Big)^2$$
$$\leq \mathbb{E}[z_1^2\Theta(z_1z_2-\alpha)\Theta(z_1)\Theta(z_2)]$$
$$\mathbb{E}[z_2^2\Theta(z_1z_2-\alpha)\Theta(z_1)\Theta(z_2)].$$

Examining the first of these factors,

$$\int_0^\infty \int_{\alpha/z_1}^\infty z_1^2 h(z_1,z_2)\, dz_2 dz_1,$$
$$= \int_0^\infty z_1^2 \int_{\alpha/z_1}^\infty h(z_1,z_2)\, dz_2 dz_1.$$

Now let $g_\alpha(z_1) = \int_{\alpha/z_1}^\infty h(z_1,z_2)\, dz_2$. $g_\alpha(z_1)z_1^2$ is monotonically pointwise non-increasing to 0 in α for all $z_1 > 0$ and $\int z_1^2 g_0(z_1) dz_1 \leq \mathbb{E}[Z_1^2] < \infty$. By the Monotone Convergence Theorem $\lim_{\alpha\to\infty}\mathbb{E}[z_1^2\Theta(z_1z_2-\alpha)\Theta(z_1)] = 0$. The second factor has the same limit, so the limit of the right hand side of Hölder's inequality is 0. □

Acknowledgements

We thank the anonymous reviewers for directing us toward relevant work and providing helpful recommendations regarding the presentation of the paper. Farbod Roosta-Khorasani gratefully acknowledges the support from the Australian Research Council through a Discovery Early Career Researcher Award (DE180100923). Russell Tsuchida's attendance at the conference was made possible by an ICML travel award.

References

Bach, F. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017a.

Bach, F. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017b.

Balduzzi, D., Frean, M., Leary, L., Lewis, J.P., Ma, K.W., and McWilliams, B. The shattered gradients problem: If resnets are the answer, then what is the question? In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 342–350, 2017.

Barker, J., Marxer, R., Vincent, E., and Watanabe, S. The third chime speech separation and recognition challenge: Analysis and outcomes. *Computer Speech and Language*, 46:605–626, 2017.

Bengio, Y., Simard, P., and Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Berthelot, D., Schumm, T., and Metz, L. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

Billingsley, P. *Probability and Measure*. Wiley-Interscience, 3rd edition, 1995. ISBN 0471007102.

Bryc, W. Rotation invariant distributions. In *The Normal Distribution*, pp. 51–69. Springer, 1995.

Burgess, A.N. Estimating equivalent kernels for neural networks: A data perturbation approach. In *Advances in Neural Information Processing Systems*, pp. 382–388, 1997.

Cho, Y. and Saul, L.K. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, pp. 342–350, 2009.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G.B., and LeCun, Y. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pp. 192–204, 2015.

Clevert, D., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations*, 2016.

Daniely, A., Frostig, R., and Singer, Y. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pp. 2253–2261, 2016.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

Haeffele, B.D. and Vidal, R. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Hochreiter, S. Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 91, 1991.

Hochreiter, S., Bengio, Y., and Frasconi, P. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In Kolen, J. and Kremer, S. (eds.), *Field Guide to Dynamical Recurrent Networks*. IEEE Press, 2001.

Huang, G., Zhu, Q., and Siew, C. Extreme learning machine: a new learning scheme of feedforward neural networks. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 2, pp. 985–990. IEEE, 2004.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

Johnson, M.E. *Multivariate statistical simulation: A guide to selecting and generating continuous multivariate distributions*. John Wiley & Sons, 2013.

Jones, D.S. *The Theory of Generalised Functions*, chapter 7, pp. 263. Cambridge University Press, 2nd edition, 1982.

Krizhevsky, Alex and Hinton, Geoffrey. Learning multiple layers of features from tiny images. 2009.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S.S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1611.01232*, 2017.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Livni, R., Carmon, D., and Globerson, A. Learning infinite layer networks without the kernel trick. In *International Conference on Machine Learning*, pp. 2198–2207, 2017.

MacKay, D.J.C. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

Martin, C.H. and Mahoney, M.W. Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. *arXiv preprint arXiv:1710.09553*, 2017.

Mishkin, D. and Matas, J. All you need is a good init. In *International Conference on Learning Representations*, 2016.

Neal, R.M. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1994.

Pandey, G. and Dukkipati, A. To go deep or wide in learning? In *Artificial Intelligence and Statistics*, pp. 724–732, 2014a.

Pandey, G. and Dukkipati, A. Learning by stretching deep networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1719–1727, 2014b.

Pao, Y., Park, G., and Sobajic, D.J. Learning and generalization characteristics of the random vector functional-link net. *Neurocomputing*, 6(2):163–180, 1994.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In *Advances In Neural Information Processing Systems*, pp. 3360–3368, 2016.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. On the expressive power of deep neural networks. In Precup, D. and Teh, Y.W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2847–2854, 2017.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

Roux, N. Le and Bengio, Y. Continuous neural networks. In *Artificial Intelligence and Statistics*, pp. 404–411, 2007.

Rudi, Alessandro and Rosasco, Lorenzo. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pp. 3218–3228, 2017.

Schmidt, W.F., Kraaijveld, M.A., and Duin, R.P.W. Feedforward neural networks with random weights. In *Pattern Recognition, 1992. Vol. II. Conference B: Pattern Recognition Methodology and Systems, Proceedings., 11th IAPR International Conference on*, pp. 1–4. IEEE, 1992.

Schoenholz, S.S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. Deep information propagation. In *International Conference on Learning Representations*, 2017.

Shwartz-Ziv, R. and Tishby, N. Opening the Black Box of Deep Neural Networks via Information. *arXiv preprint arXiv:1703.00810*, 2017.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, T., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Sinha, A. and Duchi, J.C. Learning kernels with random features. In *Advances in Neural Information Processing Systems*, pp. 1298–1306, 2016.

Tang, J., Deng, C., and Huang, G. Extreme learning machine for multilayer perceptron. *IEEE Transactions on Neural Networks and Learning Systems*, 27(4):809–821, 2016.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Vincent, E., Watanabe, S., Nugraha, A., Barker, J., and Marxer, R. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech and Language*, 46:535–557, 2017.

Williams, C.K.I. Computing with infinite networks. In *Advances in Neural Information Processing Systems*, pp. 295–301, 1997.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.