On the Effect of the Activation Function on the Distribution of Hidden Nodes in a Deep Network

Philip M. Long* and Hanie Sedghi*
Google Brain

Abstract

We analyze the joint probability distribution on the lengths of the vectors of hidden variables in different layers of a fully connected deep network, when the weights and biases are chosen randomly according to Gaussian distributions. We show that, if the activation function ϕ satisfies a minimal set of assumptions, satisfied by all activation functions that we know that are used in practice, then, as the width of the network gets large, the "length process" converges in probability to a length map that is determined as a simple function of the variances of the random weights and biases, and the activation function ϕ. We also show that this convergence may fail for ϕ that violate our assumptions. We show how to use this analysis to choose the variance of weight initialization, depending on the activation function, so that hidden variables maintain a consistent scale throughout the network.

Keywords: Initialization, theory, stability.

1 Introduction

The size of the weights of a deep network must be managed delicately. If they are too large, signals blow up as they travel through the network, leading to numerical problems, and if they are too small, the signals fade away. The practical state of the art in deep learning made a significant step forward due to schemes for initializing the weights that aimed in different ways at maintaining roughly the same scale for the hidden variables before and after a layer (LeCun et al., 1998; Glorot and Bengio, 2010). Later work (He et al., 2015; Poole et al., 2016; Daniely et al., 2016) took into account the effect of the non-linearities on the length dynamics of a deep network, informing initialization policies in a more refined way.

An influential theoretical analysis (Poole et al., 2016) considered whether signals tend to blow up or fade away as they propagate through a fully connected network with the same activation function ϕ at each hidden node. For a given input, they studied the probability distribution over the lengths of the vectors of

*Authors ordered alphabetically.

hidden variables, when the weights between nodes are chosen from a zero-mean Gaussian with variance σ_w^2/N, and where the biases are chosen from a zero-mean distribution with variance σ_b^2. They argued that, in a fully-connected network, as a width of the network approaches infinity, the (suitably normalized) lengths of the hidden layers approach a sequence of values, one for each layer, and characterized this length map as a function of ϕ, σ_w and σ_b. This analysis has since been widely used (Schoenholz et al., 2016; Yang and Schoenholz, 2017; Pennington et al., 2017; Lee et al., 2018; Xiao et al., 2018; Chen et al., 2018; Pennington et al., 2018; Hayou et al., 2018).

Poole et al. (2016) claimed that their analysis holds for arbitrary non-linearities ϕ. In contrast, we show that, for arbitrarily small positive σ_w, even if $\sigma_b = 0$, for $\phi(z) = 1/z$, the distribution of values of each of the hidden nodes in the second layer diverges as N gets large. For finite N, each node has a Cauchy distribution, which already has infinite variance, and as N gets large, the scale parameter of the Cauchy distribution gets larger, leading to divergence. We also show that the hidden variables in the second layer may not be independent, even for commonly used ϕ like the ReLU, contradicting a claim that is part of the analysis of (Poole et al., 2016).

These observations, together with the wide use of the length map from (Poole et al., 2016), motivate the search for a new analysis. This note provides such an analysis for activation functions ϕ that satisfy the following properties: (a) the restriction of ϕ to any finite interval is bounded; (b) as z gets large,[1] $|\phi(z)| \leq \exp(o(z^2))$, (c) ϕ is measurable. We refer to such ϕ as *permissible*. Note that conditions (a) and (c) both hold for any non-decreasing ϕ.

We show that, for all permissible ϕ and all σ_w and σ_b, as N gets large, the length process converges in probability to the length map described in (Poole et al., 2016).

Section 5 describes some simulation experiments verifying some of the findings of the paper, and illustrating the dependence among the values of the hidden nodes.

Section 6 describes one way to use our analysis to choose the variance of the weights depending on the activation function so that signals neither blow up nor vanish as computation flows through a wide and deep network.

Our analysis of the convergence of the length map borrows ideas from (Daniely et al., 2016), who studied the properties of the mapping from inputs to hidden representations resulting from random Gaussian initialization. Their theory applies in the case of activation functions with certain smoothness properties, and to a wide variety of architectures. Informally, they showed that, after random initialization, for wide networks, it is likely that the kernel associated with feature map computed by the network closely approximates a fixed kernel. Our analysis treats a wider variety of values of σ_w and σ_b, and uses weaker assumptions on ϕ. Motivated by Bayesian goals as in (Neal, 1996), Matthews et al. (2018) performed an analysis in a related setting, characterizing the distribution of kernels arising

[1]Here $o(z^2)$ denotes any function of z that grows strictly more slowly than z^2, such as $z^{2-\epsilon}$ for $\epsilon > 0$.

from a random initialization. Their analysis used a "linear envelope" condition on ϕ that is stronger than the assumption used here. Alternative but related uses of theory to guide the choice of weight variances may be found in (Schoenholz et al., 2016; Pennington et al., 2017). Hanin (2018) studied the effect of the widths of layers and the depth of a fully connected network on the size of the input-output Jacobian in the case of ReLU activations.

2 Preliminaries

2.1 Notation

For $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. If T is a $n \times m \times p$ tensor, then, for $i \in [n]$, let $T_{i,:,:}$ be the matrix A such that $A_{j,k} = T_{i,j,k}$, and define $T_{i,j,:}$, etc., analogously.

2.2 The finite case

Consider a deep fully connected width-N network with D layers. Let $W \in \mathbb{R}^{D \times N \times N}$. An activation function ϕ maps $\mathbb{R}$ to $\mathbb{R}$; we will also use ϕ to denote the function from $\mathbb{R}^N$ to $\mathbb{R}^N$ obtained by applying ϕ componentwise. Computation of the neural activity vectors $x_{0,:}, ..., x_{D,:} \in \mathbb{R}^N$ and preactivations $h_{1,:}, ..., h_{D,:} \in \mathbb{R}^N$ proceeds in the standard way as follows:

$$h_{\ell,:} = W_{\ell,:,:} x_{\ell-1,:} + b_{\ell,:} \quad x_{\ell,:} = \phi(h_{\ell,:}), \qquad \text{for } \ \ell = 1, \ldots, D.$$

We will study the process arising from fixing an arbitrary input $x_{0,:} \in \mathbb{R}^N$ and choosing the parameters independently at random: the entries of W are sampled from Gauss $\left(0, \frac{\sigma_w^2}{N}\right)$, and the entries of b from Gauss $(0, \sigma_b^2)$. For each $\ell \in \{0, ..., D\}$, define $q_\ell = \frac{1}{N} \sum_{i=1}^{N} h_{\ell,i}^2$.

Note that for all $\ell \geq 1$, all the components of $h_{\ell,:}$ and $x_{\ell,:}$ are identically distributed.

2.3 The wide-network limit

For the purpose of defining a limit, assume that, for a fixed, arbitrary function $\chi : \mathbb{N} \to \mathbb{R}$, for finite N, we have $x_{0,:} = (\chi(1), ..., \chi(N))$. We also assume that $\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{\infty} \chi(i)^2$ exists and is nonzero. For $\ell > 0$, let $\underline{x}_\ell$ be a random variable whose distribution is the limit of the distribution of $x_{\ell,1}$ as N goes to infinity, if this limit exists (in the sense of "convergence in distribution"). Define $\underline{h}_\ell$ and $\underline{q}_\ell$ similarly.

2.4 Total variation distance

If P and Q are probability distributions, then $d_{TV}(P, Q) = \sup_E P(E) - Q(E)$, and if p and q are their densities, $d_{TV}(P, Q) = \frac{1}{2} \int |p(x) - q(x)| \, dx$.

2.5 Permissible activation functions

Definition 1 *An activation function ϕ is* permissible *if,*

- *the restriction of ϕ to any finite interval is bounded*
- $|\phi(x)| = \exp(o(x^2))$ *as $|x|$ gets large.*[2]*, and*
- *ϕ is measurable.*

Conditions (b) and (c) ensure that a key integral can be computed. The proof of Lemma 1 is in Appendix A.

Lemma 1 *If ϕ is permissible, then, for all positive constants c, the function g defined by $g(x) = \phi(cx)^2 \exp(-x^2/2)$ is integrable.*

2.6 Length map

Next we recall the definition of a length map from (Poole et al., 2016); we will prove that the the length process converges to this length map. Define $\tilde{q}_1, ..., \tilde{q}_D$ and $\tilde{r}_0, ..., \tilde{r}_D$ recursively as follows. First $\tilde{r}_0 = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_{0,i}^2$. Then, for $\ell > 0$,

$$\tilde{q}_\ell = \sigma_w^2 \tilde{r}_{\ell-1} + \sigma_b^2$$

and

$$\tilde{r}_\ell = \mathbb{E}_{z\in \text{Gauss}(0,1)}[\phi(\sqrt{\tilde{q}_\ell} z)^2].$$

If ϕ is permissible, then, since $\phi(cz)^2 \exp(-z^2/2)$ is integrable for all c, we have that $\tilde{q}_0, ..., \tilde{q}_D, \tilde{r}_0, ..., \tilde{r}_D$ are well-defined finite real numbers.

3 Some surprising behaviors

In this section, we show that, for some activation functions, the probability distribution of hidden nodes can have some surprising properties.

3.1 Failure to converge

In this subsection, we will show that the probability distribution of the hidden variables may not converge. Our proof will refer to the Cauchy distribution.

Definition 2 *A distribution over the reals that, for $x_0 \in \mathbb{R}$ and $\gamma > 0$, has a density f given by $f(x) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$ is a* Cauchy distribution*, denoted by* Cauchy(x_0, γ)*. Cauchy$(0, 1)$ is the* standard Cauchy distribution*.*

Lemma 2 ((Hazewinkel, 2013)) *If $X_1, ..., X_n$ are i.i.d. random variables with a Cauchy distribution, then $\frac{1}{n}\sum_{i=1}^{n} X_i$ has the same distribution.*

[2] This condition may be expanded as follows, $\text{limsup}_{x\to\infty} \frac{\log|\phi(x)|}{x^2} = 0$ and $\text{limsup}_{x\to-\infty} \frac{\log|\phi(x)|}{x^2} = 0$.

Lemma 3 ((Lupton, 1993)) *If U and V are zero-mean normally distributed random variables with the same variance, then U/V has the standard Cauchy distribution.*

The following shows that there is a ϕ such that the limiting $\underline{h}_2$ is not defined. It contradicts a claim made on line 7 of Section A.1 of (Poole et al., 2016).

Proposition 4 *For any input function χ with range $\{-1, 1\}$, there is an activation function ϕ such that, for every $\sigma_w > 0$, if $\sigma_b = 0$, then (a) for finite N, $h_{2,1}$ has infinite variance, and (b) $h_{2,1}$ diverges as N goes to infinity.*

Proof: Consider ϕ defined by

$$\phi(y) = \begin{cases} 1/y & \text{if } y \neq 0 \\ 0 & \text{if } y = 0. \end{cases}$$

Fix a value of N and $\sigma_w > 0$, and take $\sigma_b = 0$. Each component of $h_{1,:}$ is a sum of zero-mean Gaussians with variance σ_w^2/N; thus, for all i, $h_{1,i} \sim \text{Gauss}(0, \sigma_w^2)$. Now, almost surely,

$$h_{2,1} = \sum_{j=1}^{N} W_{2,1,j}\phi(h_{1,j}) = \sum_{j=1}^{N} W_{2,1,j}/h_{1,j}.$$

By Lemma 3, for each j, $W_{2,1,j}/h_{1,j}$ has a Cauchy distribution, and since

$$(NW_{2,1,1}), ..., (NW_{2,1,N}) \sim \text{Gauss}(0, N\sigma_w^2),$$

recalling that $h_{1,1}, ..., h_{1,N} \sim \text{Gauss}(0, \sigma_w^2)$, we have that

$$NW_{2,1,1}/h_{1,1}, ..., NW_{2,1,N}/h_{1,N}$$

are i.i.d. Cauchy$(0, \sqrt{N})$. Applying Lemma 2,

$$h_{2,1} = \sum_{j=1}^{N} W_{2,1,j}\phi(h_{2,j}) = \frac{1}{N}\sum_{j=1}^{N} NW_{2,1,j}\phi(h_{1,j})$$

is also Cauchy$(0, \sqrt{N})$.

So, for all N, $h_{2,1}$ is Cauchy$(0, \sqrt{N})$. Suppose that $h_{2,1}$ converged in distribution to some distribution P. Since the cdf of P can have at most countably many discontinuities, we can cover the real line by a countable set of finite-length intervals $[a_1, b_1], [a_2, b_2], ...$ whose endpoints are points of continuity for P. Since Cauchy$(0, \sqrt{N})$ converges to P in distribution, for any i,

$$P([a_i, b_i]) \leq \lim_{N\to\infty} \frac{|b_i - a_i|}{\pi\sqrt{N}} = 0.$$

Thus, the probability assigned by P to the entire real line is 0, a contradiction. □

3.2 Independence

The following contradicts a claim made on line 8 of Section A.1 of (Poole et al., 2016).

Theorem 5 *If ϕ is either the ReLU or the Heaviside function, then, for every $\sigma_w > 0$, $\sigma_b \geq 0$, and $N \geq 2$, $(h_{2,1}, ..., h_{2,N})$ are not independent.*

Proof: We will show that $\mathbb{E}[h_{2,1}^2 h_{2,2}^2] \neq \mathbb{E}[h_{2,1}^2]\mathbb{E}[h_{2,2}^2]$, which will imply that $h_{2,1}$ and $h_{2,2}$ are not independent.

As mentioned earlier, because each component of $h_{1,:}$ is the dot product of $x_{0,:}$ with an independent row of $W_{1,:,:}$ plus an independent component of $b_{1,:}$, the components of $h_{1,:}$ are independent, and since $x_{1,:} = \phi(h_{1,:})$, this implies that the components of $x_{1,:}$ are independent. Since each row of $W_{1,:,:}$ and each component of the bias vector has the same distribution, $x_{1,:}$ is i.i.d.

We have

$$\begin{aligned}\mathbb{E}[h_{2,1}^2] &= \mathbb{E}\left[\left[\left(\sum_{i\in[N]} W_{2,1,i}x_{1,i}\right) + b_{2,1}\right]^2\right] \\ &= \sum_{(i,j)\in[N]^2} \mathbb{E}\left[W_{2,1,i}W_{2,1,j}x_{1,i}x_{1,j}\right] + 2\sum_{i\in[N]} \mathbb{E}\left[W_{2,1,i}x_{1,i}b_{2,1}\right] + \mathbb{E}\left[b_{2,1}^2\right].\end{aligned}$$

The components of $W_{2,:,:}$ and $x_{1,:}$, along with $b_{2,1}$, are mutually independent, so terms in the double sum with $i \neq j$ have zero expectation, and $\mathbb{E}[h_{2,1}^2] = \left(\sum_{i\in[N]} \mathbb{E}\left[W_{2,1,i}^2\right]\mathbb{E}\left[x_{1,i}^2\right]\right) + \mathbb{E}[b_{2,1}^2]$. For a random variable x with the same distribution as the components of $x_{1,:}$, this implies

$$\mathbb{E}[h_{2,1}^2] = \sigma_w^2\mathbb{E}\left[x^2\right] + \sigma_b^2. \tag{1}$$

Similarly,

$$
\begin{aligned}
&\mathbb{E}[h_{2,1}^2 h_{2,2}^2] \\
&= \mathbb{E}\left[\left[\sum_{i\in[N]} W_{2,1,i}x_{1,i} + b_{2,1}\right]^2 \left[\sum_{i\in[N]} W_{2,2,i}x_{1,i} + b_{2,2}\right]^2\right] \\
&= \sum_{(i,j,r,s)\in[N]^4} \mathbb{E}[W_{2,1,i}W_{2,1,j}W_{2,2,r}W_{2,2,s}x_{1,i}x_{1,j}x_{1,r}x_{1,s}] \\
&\quad + 2\sum_{(i,j,r)\in[N]^3} \mathbb{E}[W_{2,1,i}W_{2,1,j}W_{2,2,r}x_{1,i}x_{1,j}x_{1,r}b_{2,2}] \\
&\quad + 2\sum_{(i,r,s)\in[N]^3} \mathbb{E}[W_{2,1,i}W_{2,2,r}W_{2,2,s}x_{1,i}x_{1,r}x_{1,s}b_{2,1}] \\
&\quad + 4\sum_{(i,r)\in[N]^2} \mathbb{E}[W_{2,1,i}W_{2,2,r}x_{1,i}x_{1,r}b_{2,1}b_{2,2}] \\
&\quad + \sum_{(i,j)\in[N]^2} \mathbb{E}[W_{2,1,i}W_{2,1,j}x_{1,i}x_{1,j}b_{2,2}^2] + \sum_{(r,s)\in[N]^2} \mathbb{E}[W_{2,2,r}W_{2,2,s}x_{1,r}x_{1,s}b_{2,1}^2] \\
&\quad + 2\sum_{i\in[N]} \mathbb{E}[W_{2,1,i}x_{1,i}b_{2,1}b_{2,2}^2] + 2\sum_{r\in[N]} \mathbb{E}[W_{2,2,r}x_{1,r}b_{2,1}^2 b_{2,2}] \\
&\quad + \mathbb{E}[b_{2,1}^2 b_{2,2}^2] \\
&= \sum_{(i,r)\in[N]^2, i\neq r} \mathbb{E}[W_{2,1,i}^2 W_{2,2,r}^2]\mathbb{E}[x_{1,i}^2]\mathbb{E}[x_{1,r}^2] + \sum_{i\in[N]} \mathbb{E}[W_{2,1,i}^2 W_{2,2,i}^2]\mathbb{E}[x_{1,i}^4] \\
&\quad + \sum_{i\in[N]} \mathbb{E}[W_{2,1,i}^2]\mathbb{E}[x_{1,i}^2]\mathbb{E}[b_{2,2}^2] + \sum_{r\in[N]} \mathbb{E}[W_{2,2,r}^2]\mathbb{E}[x_{1,r}^2]\mathbb{E}[b_{2,1}^2] \\
&\quad + \mathbb{E}[b_{1,2}^2 b_{2,2}^2] \\
&= \frac{(N^2-N)\sigma_w^4\mathbb{E}[x^2]^2}{N^2} + \frac{N\sigma_w^4\mathbb{E}[x^4]}{N^2} + \frac{2N\sigma_w^2\mathbb{E}[x^2]\sigma_b^2}{N} + \sigma_b^4 \\
&= \sigma_w^4\mathbb{E}[x^2]^2 + \frac{\sigma_w^4(\mathbb{E}[x^4] - \mathbb{E}[x^2]^2)}{N} + 2\sigma_w^2\sigma_b^2\mathbb{E}[x^2] + \sigma_b^4.
\end{aligned}
$$

Putting this together with (1), we have

$$
\mathbb{E}[h_{2,1}^2 h_{2,2}^2] - \mathbb{E}[h_{2,1}^2]\mathbb{E}[h_{2,2}^2] = \frac{\sigma_w^4(\mathbb{E}[x^4] - \mathbb{E}[x^2]^2)}{N}. \tag{2}
$$

Now, we calculate the difference using (2) for the Heaviside and ReLU functions.

Heaviside. Suppose ϕ is Heaviside function, i.e. $\phi(z)$ is the indicator function for $z > 0$. In this case, since the components of $h_{1,:}$ are symmetric about 0, the distribution of $x_{1,:}$ is uniform over $\{0,1\}^N$. Thus $\mathbb{E}[x^4] = \mathbb{E}[x^2] = 1/2$, and so (2) gives $\mathbb{E}[h_{2,1}^2 h_{2,2}^2] - \mathbb{E}[h_{2,1}^2]\mathbb{E}[h_{2,2}^2] = \frac{3\sigma_w^4}{4N} \neq 0$.

ReLU. Next, we consider the case that ϕ is the ReLU. Recalling that, for all i, $h_{1,i} \sim \text{Gauss}(0, \sigma_w^2)$, we have $\mathbb{E}[x^2] = \frac{1}{\sqrt{2\pi\sigma_w^2}} \int_0^\infty z^2 \exp\left(\frac{-z^2}{2\sigma_w^2}\right) dz$. By symmetry

this is $\frac{1}{2}\mathbb{E}_{z\sim\text{Gauss}(0,\sigma_w^2)}[z^2] = \sigma_w^2/2$. Similarly, $\mathbb{E}[x^4] = \frac{1}{2}\mathbb{E}_{z\sim\text{Gauss}(0,\sigma_w^2)}[z^4] = \frac{3\sigma^4}{2}$. Plugging these into (2) we get that, in the case the ϕ is the ReLU, that

$$\mathbb{E}[h_{2,1}^2 h_{2,2}^2] - \mathbb{E}[h_{2,1}^2]\mathbb{E}[h_{2,2}^2] = \frac{\sigma_w^4\left((3/2)\sigma_w^4 - \sigma_w^4/4\right)}{N} = \frac{5\sigma_w^8}{4N} > 0,$$

completing the proof. □

Note that, informally, the degree of dependence between pairs of hidden nodes established in the proof of Theorem 5 approaches 0 as N gets large. On the other hand, the number of dependent pairs of hidden nodes is $\Omega(N^2)$.

3.3 Undefined length map

Here, we show, informally, that for ϕ at the boundary of the second condition in the definition of permissibility, the recursive formula defining the length map $\tilde{q}_\ell$ breaks down. Roughly, this condition cannot be relaxed.

Proposition 6 *For any $\alpha > 0$, if ϕ is defined by $\phi(x) = \exp(\alpha x^2)$, even if all components of all inputs are in $\{-1, 1\}$, there exists a σ_w, σ_b s.t. $\tilde{q}_\ell, \tilde{r}_\ell$ is undefined for all $\ell \geq 2$.*

Proof: Suppose $\sigma_w^2 + \sigma_b^2 = \frac{1}{4\alpha^2}$. Then $\tilde{q}_1 = \frac{1}{4\alpha^2}$, so that

$$\begin{aligned}\tilde{r}_1 &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\phi(\sqrt{\tilde{q}_1}z)\exp\left(-\frac{z^2}{2}\right)dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp(\alpha\sqrt{\tilde{q}_1}z^2)\exp\left(-\frac{z^2}{2}\right)dz \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp(z^2/2)\exp\left(-\frac{z^2}{2}\right)dz = \infty,\end{aligned}$$

and downsteam values of $\tilde{q}_\ell$ and $\tilde{r}_\ell$ are undefined. □

4 Convergence in probability

In this section that the length process $q_0, ..., q_D$ converges in probability to the length map $\tilde{q}_0, ..., \tilde{q}_D$ from (Poole et al., 2016).

Theorem 7 *For any permissible ϕ, $\sigma_w, \sigma_b \geq 0$, any depth D, and any $\epsilon, \delta > 0$, there is an N_0 such that, for all $N \geq N_0$, with probability $1 - \delta$, for all $\ell \in [D]$, we have $|q_\ell - \tilde{q}_\ell| \leq \epsilon$.*

The rest of this section is devoted to proving Theorem 7. Our proof will use the weak law of large numbers.

Lemma 8 ((Feller, 2008)) *For any random variable X with a finite expectation, and any $\epsilon, \delta > 0$, there is an N_0 such that, for all $N \geq N_0$, if $X_1, ..., X_N$ are i.i.d. with the same distribution as X, then*

$$\Pr\left(\left|\mathbb{E}[X] - \frac{1}{N}\sum_{i=1}^{N} X_i\right| > \epsilon\right) \leq \delta.$$

In order to divide our analysis into cases, we need the following lemma, whose proof is in Appendix B.

Lemma 9 *If ϕ is permissible and not zero a.e., for all $\sigma_w > 0$, for all ℓ, $\tilde{q}_\ell > 0$ and $\tilde{r}_\ell > 0$.*

We will also need a lemma that shows that small changes in σ lead to small changes in $\text{Gauss}(0, \sigma^2)$.

Lemma 10 (see (Klartag, 2007)) *There is an absolute constant C such that, for all $\sigma_1, \sigma_2 > 0$,*

$$d_{TV}(\text{Gauss}(0, \sigma_1^2), \text{Gauss}(0, \sigma_2^2)) \leq C \frac{|\sigma_1 - \sigma_2|}{\sigma_1}.$$

The following technical lemma, which shows that tail bounds hold uniformly over different choices of q, is proved in Appendix C.

Lemma 11 *If ϕ is permissible, for all $0 < r \leq s$, for all $\beta > 0$, there is an $a \geq 0$ such that, for all $q \in [r, s]$, $\int_a^\infty \phi(\sqrt{q}z)^2 \exp(-z^2/2)\ dz \leq \beta$ and $\int_{-\infty}^{-a} \phi(\sqrt{q}z)^2 \exp(-z^2/2)\ dz \leq \beta$.*

Armed with these lemmas, we are ready to prove Theorem 7.

First, if ϕ is zero a.e., or if $\sigma_w = 0$, Theorem 7 follows directly from Lemma 8, together with a union bound over the layers. Assume for the rest of the proof that $\phi(x)$ is non-zero on a set of positive measure, and that $\sigma_w > 0$, so that $\tilde{q}_\ell > 0$ and $\tilde{r}_\ell > 0$ for all ℓ.

For each $\ell \in \{0, ..., D\}$, define $r_\ell = \frac{1}{N} \sum_{i=1}^N x_{\ell,i}^2$.

Our proof of Theorem 7 is by induction. The inductive hypothesis is that, for any $\epsilon, \delta > 0$ there is an N_0 such that, if $N \geq N_0$, then, with probability $1 - \delta$, for all $\ell' \in \{1, ..., \ell\}$, $|q_{\ell'} - \tilde{q}_{\ell'}| \leq \epsilon$ and, for all $\ell' \in \{0, ..., \ell\}$, and $|r_{\ell'} - \tilde{r}_{\ell'}| \leq \epsilon$.

The base case, where $\ell = 0$, holds because $\tilde{r}_0$ is definied to be the limit of r_0 as N goes to infinity.

Now for the induction step; choose $\ell > 0$, $0 < \epsilon < \min\{\tilde{q}_\ell/4, \tilde{r}_\ell\}$ and $0 < \delta \leq 1/2$. (Note that these choices are without loss of generality.) Let $\epsilon' \in (0, \epsilon)$ take a value that will be described later, using quantities from the analysis. By the inductive hypothesis, whatever the value of ϵ', there is an N_0' such that, if $N \geq N_0'$, then, with probability $1 - \delta/2$, for all $\ell' \leq \ell - 1$, we have $|q_{\ell'} - \tilde{q}_{\ell'}| \leq \epsilon'$ and $|r_{\ell'} - \tilde{r}_{\ell'}| \leq \epsilon'$. Thus, to establish the inductive step, it suffices to show that, after conditioning on the random choices before the ℓth layer, if $|r_{\ell-1} - \tilde{r}_{\ell-1}| \leq \epsilon'$, there is an N_ℓ such that, if $N \geq N_\ell$, then with probability at least $1 - \delta/2$ with respect only to the random choices of $W_{\ell,:,:}$ and $b_{\ell,:}$, that $|q_\ell - \tilde{q}_\ell| \leq \epsilon$ and $|r_\ell - \tilde{r}_\ell| \leq \epsilon$. Given such an N_ℓ, the inductive step can be satisfied by letting N_0 be the maximum of N_0' and N_ℓ.

Let us do that. To simplify the notation, for the rest of the proof of the inductive step, let us condition on outcomes of the layers before layer ℓ; all expectations and probabilities will concern the randomness only in the ℓth layer. Let us further assume that $|r_{\ell-1} - \tilde{r}_{\ell-1}| \leq \epsilon'$.

Recall that $q_\ell = \frac{1}{N}\sum_{i=1}^N h_{\ell,i}^2$. Since the values of $h_{\ell-1,1}, ..., h_{\ell-1,N}$ have been fixed by conditioning, each component of $h_{\ell,i}$ is obtained by taking the dot-product of $x_{\ell-1,:} = \phi(h_{\ell-1,:})$ with $W_{\ell,i,:}$ and adding an independent $b_{\ell,i}$. Thus, conditioned on $h_{\ell-1,1}, ..., h_{\ell-1,N}$, we have that $h_{\ell,1}, ..., h_{\ell,N}$ are independent. Also, since $x_{\ell-1,:}$ is fixed by conditioning, each $h_{\ell,i}$ has an identical Gaussian distribution.

Since each component of W and b has zero mean, each $h_{\ell,i}$ has zero mean.

Choose an arbitrary $i \in [N]$. Since $x_{\ell-1,:}$ is fixed by conditioning and

$$W_{\ell,i,1}, ..., W_{\ell,i,N}$$

and $b_{\ell,i}$ are independent,

$$\mathbb{E}[q_\ell] = \mathbb{E}[h_{\ell,i}^2] = \sigma_b^2 + \frac{\sigma_w^2}{N}\sum_j x_{\ell-1,j}^2 = \sigma_b^2 + \sigma_w^2 r_{\ell-1} \stackrel{\text{def}}{=} \overline{q}_\ell. \tag{3}$$

We wish to emphasize the $\overline{q}_\ell$ is determined as a function of random outcomes before the ℓth layer, and thus a fixed, nonrandom quantity, regarding the randomization of the ℓth layer. By the inductive hypothesis, we have

$$|\mathbb{E}[q_\ell] - \tilde{q}_\ell| = |\mathbb{E}[h_{\ell,i}^2] - \tilde{q}_\ell| = |\overline{q}_\ell - \tilde{q}_\ell| = \sigma_w^2 |r_{\ell-1} - \tilde{r}_{\ell-1}| \le \epsilon' \sigma_w^2. \tag{4}$$

The key consequence of this might be paraphrased by saying that, to establish the portion of the inductive step regarding q_ℓ, it suffices for q_ℓ to be close to its mean. Now, we want to prove something similar for r_ℓ. We have

$$\mathbb{E}[r_\ell] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[x_{\ell,i}^2] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\phi(h_{\ell,i})^2] = \mathbb{E}[\phi(h_{\ell,1})^2],$$

since, recalling that we have conditioned on previous layers, $h_{\ell,1}, ..., h_{\ell,N}$ are i.i.d. Since $h_{\ell,i} \sim \text{Gauss}(0, \overline{q}_\ell)$, we have

$$\begin{aligned}
\mathbb{E}[r_\ell] &= \mathbb{E}_{z\sim\text{Gauss}(0,\overline{q}_\ell)}[\phi(z)^2] \\
&= \mathbb{E}_{z\sim\text{Gauss}(0,1)}[\phi(\sqrt{\overline{q}_\ell} z)^2] \\
&= \sqrt{\frac{1}{2\pi}} \int \phi(\sqrt{\overline{q}_\ell} z)^2 \exp(-z^2/2)\, dz
\end{aligned}$$

which gives

$$|\mathbb{E}[r_\ell] - \tilde{r}_\ell| \le \left|\mathbb{E}_{z\sim\text{Gauss}(0,\overline{q}_\ell)}[\phi(z)^2] - \mathbb{E}_{z\sim\text{Gauss}(0,\tilde{q}_\ell)}[\phi(z)^2]\right|.$$

Since $|\overline{q}_\ell - \tilde{q}_\ell| \le \epsilon' \sigma_w^2$ and we may choose ϵ' to ensure $\epsilon' \le \frac{\tilde{q}_\ell}{2\sigma_w^2}$, we have $\tilde{q}_\ell/2 \le \overline{q}_\ell \le 2\tilde{q}_\ell$.

For $\beta > 0$ and $\kappa \in (0, 1/2)$ to be named later, by Lemma 11, we can choose a such that, for all $q \in [\tilde{q}_\ell/2, 2\tilde{q}_\ell]$,

$$\int_{-\infty}^{-a} \phi(\sqrt{q} z)^2 \exp(-z^2/2)\, dz \le \beta/2 \quad \text{and} \quad \int_a^{\infty} \phi(\sqrt{q} z)^2 \exp(-z^2/2)\, dz \le \beta/2$$

and $\frac{1}{\sqrt{2\pi q}} \int_{-a}^{a} \exp\left(-\frac{z^2}{2q}\right)\, dz \geq 1-\kappa$. Choose such an a.

We claim that $\left|\int_{-a}^{a} \phi(\sqrt{q}z)^2 \exp(-z^2/2)\, dz - \int \phi(\sqrt{q}z)^2 \exp(-z^2/2)\, dz\right| \leq \beta$ for all $\tilde{q}_\ell/2 < q \leq 2\tilde{q}_\ell$. Choose such a q. We have

$$\begin{aligned}
&\left|\int_{-a}^{a} \phi(\sqrt{q}z)^2 \exp(-z^2/2)\, dz - \int \phi(\sqrt{q}z)^2 \exp(-z^2/2)\, dz\right| \\
&= \int_{-\infty}^{-a} \phi(\sqrt{q}z)^2 \exp(-z^2/2)\, dz + \int_{a}^{\infty} \phi(\sqrt{q}z)^2 \exp(-z^2/2)\, dz \\
&\leq 2\max\left\{\int_{-\infty}^{-a} \phi(\sqrt{q}z)^2 \exp(-z^2/2)\, dz, \int_{a}^{\infty} \phi(\sqrt{q}z)^2 \exp(-z^2/2)\, dz\right\} \\
&\leq \beta.
\end{aligned}$$

So now we are trying to bound

$$\left|\int_{-a}^{a} \phi(\sqrt{\overline{q}_\ell}z)^2 \exp(-z^2/2)\, dz - \int_{-a}^{a} \phi(\sqrt{\tilde{q}_\ell}z)^2 \exp(-z^2/2)\, dz\right|$$

using $\tilde{q}_\ell/2 \leq \overline{q}_\ell \leq 2\tilde{q}_\ell$.

Using changes of variables, we have

$$\begin{aligned}
&\left|\int_{-a}^{a} \phi(\sqrt{\overline{q}_\ell}z)^2 \exp(-z^2/2)\, dz - \int_{-a}^{a} \phi(\sqrt{\tilde{q}_\ell}z)^2 \exp(-z^2/2)\, dz\right| \\
&= \left|\frac{1}{\sqrt{\overline{q}_\ell}} \int_{-a\sqrt{\overline{q}_\ell}}^{a\sqrt{\overline{q}_\ell}} \phi(z)^2 \exp\left(-\frac{z^2}{2\overline{q}_\ell}\right)\, dz - \frac{1}{\sqrt{\tilde{q}_\ell}} \int_{-a\sqrt{\tilde{q}_\ell}}^{a\sqrt{\tilde{q}_\ell}} \phi(z)^2 \exp\left(-\frac{z^2}{2\tilde{q}_\ell}\right)\, dz\right|.
\end{aligned}$$

Since ϕ is permissible, ϕ^2 is bounded on $[-a\sqrt{2\tilde{q}_\ell}, a\sqrt{2\tilde{q}_\ell}]$. If P is the distribution obtained by conditioning Gauss$(0, \overline{q}_\ell)$ on $[-a\sqrt{\overline{q}_\ell}, a\sqrt{\overline{q}_\ell}]$, and $\tilde{P}$ by conditioning Gauss$(0, \tilde{q}_\ell)$ on $[-a\sqrt{\tilde{q}_\ell}, a\sqrt{\tilde{q}_\ell}]$, then if $M = \sqrt{2\pi}\sup_{z\in[-a\sqrt{2\tilde{q}_\ell}, a\sqrt{2\tilde{q}_\ell}]} \phi(z)^2$, since $\overline{q}_\ell \leq 2\tilde{q}_\ell$,

$$\begin{aligned}
&\left|\frac{1}{\sqrt{\overline{q}_\ell}} \int_{-a\sqrt{\overline{q}_\ell}}^{a\sqrt{\overline{q}_\ell}} \phi(z)^2 \exp(-\frac{z^2}{2\overline{q}_\ell})\, dz - \frac{1}{\sqrt{\tilde{q}_\ell}} \int_{-a\sqrt{\tilde{q}_\ell}}^{a\sqrt{\tilde{q}_\ell}} \phi(z)^2 \exp(-\frac{z^2}{2\tilde{q}_\ell})\, dz\right| \\
&\leq M d_{TV}(P, \tilde{P}).
\end{aligned}$$

But since, for $\kappa < 1/2$, conditioning on an event of probability at least $1-\kappa$ only changes a distribution by total variation distance at most 2κ, and therefore, applying Lemma 10 along with the fact that $|\overline{q}_\ell - \tilde{q}_\ell| \leq \epsilon'\sigma_w^2$, for the constant C

from Lemma 10, we get

$$\begin{aligned} d_{TV}(P, \tilde{P}) &\leq 4\kappa + d_{TV}(\text{Gauss}(0, \overline{q}_\ell), \text{Gauss}(0, \tilde{q}_\ell)) \\ &\leq 4\kappa + \frac{C|\sqrt{\overline{q}_\ell} - \sqrt{\tilde{q}_\ell}|}{\sqrt{\tilde{q}_\ell}} \\ &= 4\kappa + \frac{C|\overline{q}_\ell - \tilde{q}_\ell|}{|\sqrt{\overline{q}_\ell} + \sqrt{\tilde{q}_\ell}|\sqrt{\tilde{q}_\ell}} \\ &\leq 4\kappa + \frac{C\epsilon'\sigma_w^2}{\tilde{q}_\ell}. \end{aligned}$$

Tracing back, we have

$$\begin{aligned} &\left| \int_{-a}^{a} \phi(\sqrt{\overline{q}_\ell} z)^2 \exp(-z^2/2)\, dz - \int_{-a}^{a} \phi(\sqrt{\tilde{q}_\ell} z)^2 \exp(-z^2/2)\, dz \right| \\ &\leq M \left(4\kappa + \frac{C\epsilon'\sigma_w^2}{\tilde{q}_\ell} \right) \end{aligned}$$

which implies

$$\begin{aligned} |\mathbb{E}[r_\ell] - \tilde{r}_\ell| &\leq \left| \int \phi(\sqrt{\overline{q}_\ell} z)^2 \exp(-z^2/2)\, dz - \int \phi(\sqrt{\tilde{q}_\ell} z)^2 \exp(-z^2/2)\, dz \right| \\ &\leq M \left(4\kappa + \frac{C\epsilon'\sigma_w^2}{\tilde{q}_\ell} \right) + 2\beta. \end{aligned}$$

If $\kappa = \min\{\frac{\epsilon}{24M}, \frac{1}{3}\}$, $\beta = \frac{\epsilon}{12}$, and $\epsilon' = \min\left\{\frac{\epsilon}{2}, \frac{\epsilon}{2\sigma_w^2}, \frac{\tilde{q}_\ell}{2\sigma_w^2}, \frac{\tilde{q}_\ell \epsilon}{6CM\sigma_w^2}\right\}$ this implies $|\mathbb{E}[r_\ell] - \tilde{r}_\ell| \leq \epsilon/2$.

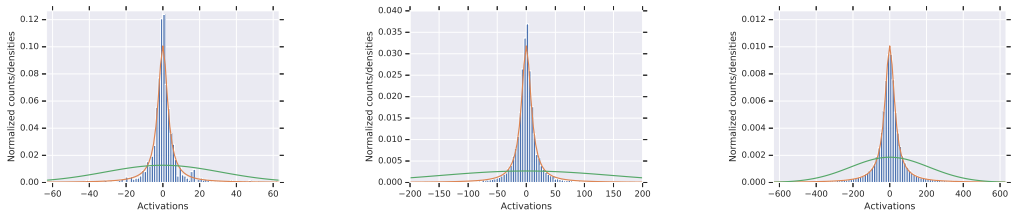
Recall that q_ℓ is an average of N identically distributed random variables with a mean between 0 and $2\tilde{q}_\ell$ (which is therefore finite) and r_ℓ is an average of N identically distributed random variables, each with mean between 0 and $\tilde{r}_\ell + \epsilon/2 \leq 2\tilde{r}_\ell$. Applying the weak law of large numbers (Lemma 8), there is an N_ℓ such that, if $N \geq N_\ell$, with probability at least $1 - \delta/2$, both $|q_\ell - \mathbb{E}[q_\ell]| \leq \epsilon/2$ and $|r_\ell - \mathbb{E}[r_\ell]| \leq \epsilon/2$ hold, which in turn implies $|q_\ell - \tilde{q}_\ell| \leq \epsilon$ and $|r_\ell - \tilde{r}_\ell| \leq \epsilon$, completing the proof of the inductive step, and therefore the proof of Theorem 7.

5 Experiments

Our first experiment fixed $x[0, :] = (1, ..., 1)$, $\sigma_w = 1$, $\sigma_b = 0$, $\phi(z) = 1/z$.

For each $N \in \{10, 100, 1000\}$, we (a) initialized the weights 100 times, (b) plotted the histograms of all of the values of $h[2, :]$, along with the Cauchy$(0, \sqrt{N})$ distribution from the proof of Proposition 4, and Gauss$(0, \sigma^2)$ for σ estimated from the data. Consistent with the theory, the Cauchy$(0, \sqrt{N})$ distribution fits the data well.

To illustrate the fact that the values in the second hidden layer are not independent, for $N = 1000$ and the parameters otherwise as in the other experiment,



(a) $N = 10$ (b) $N = 100$ (c) $N = 1000$

Figure 1: Histograms of $h[2, :]$, averaged over 100 random initializations, for $N \in \{10, 100, 1000\}$, along with Cauchy$(0, \sqrt{N})$ (shown in red) and Gauss$(0, \sigma^2)$ for σ estimated from the data (shown in green). When we average over multiple random initializations of the weights, the distribution of the activations matches the Cauchy distribution, and not the Gaussian.

we plotted histograms of the values seen in the second layer for nine random initializations of the weights in Figure 2. When some of the values in the first hidden layer have unusually small magnitude, then the values in the second hidden layer coordinately tend to be large. Note that this is consistent with Theorem 7 es-

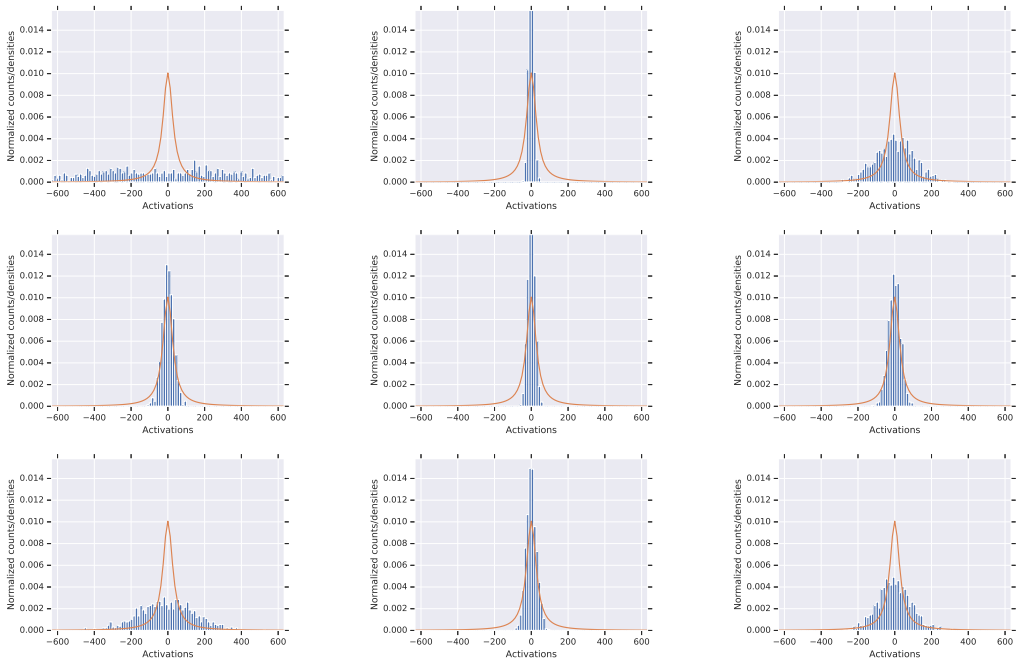


Figure 2: Histograms of $h[2, :]$ for nine random weight initializations. Plotting activations separately for different random initializations reveals the dependence among the activations in a layer.

tablishing convergence in probability for permissible ϕ, since the ϕ used in this experiment is not permissible.

activation	input variance	weight variance
Identity	1	1
ReLU	1/2	2
Heaviside	1/2	2
Exponential	e^2	$1/e^2$
Tanh	0.394	2.53

Table 1: Choices of input variance (r_0) and weight variances (σ_w^2) that theory suggests will promote an invariant that the preactivations maintain a constant scale as computation flows through the network.

6 Maintaining unit scale

In this section, we describe one use of our analysis to guide the design of initialization variances.

Our analysis shows that $q_1 \approx \sigma_w^2 r_0 + \sigma_b^2$, and

$$q_{\ell+1} \approx \sigma_w^2 \mathbb{E}_{z \in \text{Gauss}(0,1)}[\phi(\sqrt{q_\ell} z)^2] + \sigma_b^2.$$

If we achieve

$$1 \approx \sigma_w^2 r_0 + \sigma_b^2$$

and

$$1 = \sigma_w^2 \mathbb{E}_{z \in \text{Gauss}(0,1)}[\phi(z)^2] + \sigma_b^2,$$

this will promote $q_1 \approx 1, q_2 \approx 1, q_3 \approx 1$, and so on. Setting

$$\sigma_w^2 = 1/\mathbb{E}_{z \in \text{Gauss}(0,1)}[\phi(z)^2],\ r_0 = \mathbb{E}_{z \in \text{Gauss}(0,1)}[\phi(z)^2],\ \sigma_b^2 = 0$$

satisfies both. These values are collected from some common activation functions in Table 1.

7 Conclusion

We have given a rigorous analysis of the limiting value of the distribution of the lengths of the vectors of hidden nodes in a fully connected deep network, and described how to choose the variance of the weights at initialization using this analysis for various commonly used activation functions. Our analysis can be easily applied to other activation functions.

As in earlier work, our analysis concerned a limit in which the input grows along with the hidden layers. This simplifies the analysis, but it appears not to be difficult to remove this assumption (see (Matthews et al., 2018)).

After publication of some of this work in preliminary form (Long and Sedghi, 2019), elements of its analysis were used in (Novak et al., 2019).

Analysis of the length map in the case of ReLU activations was an important component of recent analyses of the convergence of deep network training (Zou

et al., 2018; Allen-Zhu et al., 2019). A non-asymptotic refinement of our analysis would be a step toward generalizing those results to more general activation functions.

Acknowledgements

We thank Ben Poole, Sam Schoenholz and Jascha Sohl-Dickstein for valuable conversations, and Jascha and anonymous reviewers for their helpful comments on earlier versions of this paper.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *ICML*, pages 242–252, 2019.

M. Chen, J. Pennington, and S. S. Schoenholz. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. *arXiv preprint arXiv:1806.05394*, 2018.

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

W. Feller. *An introduction to probability theory and its applications*. John Wiley & Sons, 2008.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*, pages 580–589, 2018.

S. Hayou, A. Doucet, and J. Rousseau. On the selection of initialization and activation function for deep neural networks. *arXiv preprint arXiv:1805.08266*, 2018.

M. Hazewinkel. Cauchy distribution. In *Encyclopaedia of Mathematics: Volume 6*. Springer Science & Business Media, 2013.

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

B. Klartag. A central limit theorem for convex sets. *Inventiones mathematicae*, 168(1):91–131, 2007.

Y. A. LeCun, L. Bottou, G. B. Orr, and K. Müller. Efficient backprop. In *Neural networks: Tricks of the trade*. Springer, 1998.

J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as gaussian processes. *ICLR*, 2018.

Philip M Long and Hanie Sedghi. On the effect of the activation function on the distribution of hidden nodes in a deep network. *arXiv preprint arXiv:1901.02104*, 2019.

R. Lupton. *Statistics in theory and practice*. Princeton University Press, 1993.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018. v2, 8/16/2018.

Radford M Neal. Bayesian learning for neural networks. 1996.

Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. *ICLR*, 2019.

J. Pennington, S. Schoenholz, and S. Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in neural information processing systems*, pages 4785–4795, 2017.

J. Pennington, S. S. Schoenholz, and S. Ganguli. The emergence of spectral universality in deep networks. *arXiv preprint arXiv:1802.09979*, 2018.

B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep information propagation. *arXiv preprint arXiv:1611.01232*, 2016.

L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. S. Schoenholz, and J. Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. *arXiv preprint arXiv:1806.05393*, 2018.

G. Yang and S. Schoenholz. Mean field residual networks: On the edge of chaos. In *Advances in neural information processing systems*, pages 7103–7114, 2017.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *CoRR*, abs/1811.08888, 2018.

A Proof of Lemma 1

Choose $c > 0$. Since $\text{limsup}_{x\to\infty} \frac{\log|\phi(x)|}{x^2} = 0$ and $\text{limsup}_{x\to-\infty} \frac{\log|\phi(x)|}{x^2} = 0$, we also have
$\text{limsup}_{x\to\infty} \frac{\log|\phi(cx)|}{x^2} = 0$ and $\text{limsup}_{x\to-\infty} \frac{\log|\phi(cx)|}{x^2} = 0$. Thus, there is an a such that, for all $x \notin [-a, a]$, $\log|\phi(cx)| \leq \frac{x^2}{8}$, which implies $\phi(cx)^2 \leq \exp\left(\frac{x^2}{4}\right)$. Since ϕ is permissible, it is bounded on $[-a, a]$. Thus, we have

$$\begin{aligned}
&\int \phi(cx)^2 \exp(-x^2/2)\, dx \\
&= \int_{-\infty}^{-a} \phi(cx)^2 \exp(-x^2/2) dx + \int_{-a}^{a} \phi(cx)^2 \exp(-x^2/2) dx \\
&\qquad + \int_{a}^{\infty} \phi(cx)^2 \exp(-x^2/2) dx \\
&\leq \int_{-\infty}^{-a} \exp(-x^2/4) dx + \left(\sup_{x\in[-a,a]} \phi(cx)^2\right) \int_{-a}^{a} \exp(-x^2/2) dx \\
&\qquad + \int_{a}^{\infty} \exp(-x^2/4) dx \\
&< \infty
\end{aligned}$$

completing the proof.

B Proof of Lemma 9

The proof is by induction. The base case holds since we have assumed that $\tilde{r}_0 > 0$.

To prove the inductive step, we need the following lemma.

Lemma 12 *If ϕ is not zero a.e., then, for all $c > 0$, $\mathbb{E}_{z\in\text{Gauss}(0,1)}(\phi(cz)^2) > 0$.*

Proof: If μ is the Lebesgue measure, since

$$\mu(\{x \in \mathbb{R} : \phi^2(cx) > 0\}) = \lim_{n\to\infty} \mu(\{x : \phi^2(cx) > 1/n\} \cap [-n, n]) > 0,$$

there exists n such that $\mu(\{x : \phi^2(cx) > 1/n\} \cap [-n, n]) > 0$. For such an n, we have

$$\mathbb{E}_{z\in\text{Gauss}(0,1)}(\phi(cz)^2) \geq \frac{1}{n} e^{-n^2/2} \mu(\{x : \phi^2(cx) > 1/n\} \cap [-n, n]) > 0.$$

□

Returning to the proof of Lemma 9, by the inductive hypothesis, $\tilde{r}_{\ell-1} > 0$, which, since $\sigma_w > 0$, implies $\tilde{q}_\ell > 0$. Applying Lemma 12 yields $\tilde{r}_\ell > 0$.

C Proof of Lemma 11

Since $\operatorname{limsup}_{x\to\infty} \frac{\log|\phi(x)|}{x^2} = 0$ there is an b such that, for all $x \geq b$, $\log|\phi(x)| \leq \frac{x^2}{8s}$, which implies $\phi(x)^2 \leq \exp\left(\frac{x^2}{4s}\right)$. Now, choose $q \in [r,s]$. For $a = b/\sqrt{r}$, we then have

$$\begin{aligned}
&\int_a^\infty \phi(\sqrt{q}x)^2 \exp(-x^2/2)\, dx \\
&= \frac{1}{\sqrt{q}} \int_{a\sqrt{q}}^\infty \phi(z)^2 \exp\left(-\frac{z^2}{2q}\right)\, dz \\
&\leq \frac{1}{\sqrt{q}} \int_{a\sqrt{q}}^\infty \exp\left(\frac{z^2}{4s}\right) \exp\left(-\frac{z^2}{2q}\right)\, dz \\
&\leq \frac{1}{\sqrt{q}} \int_{a\sqrt{q}}^\infty \exp\left(-\frac{z^2}{4q}\right)\, dz \\
&\leq \frac{1}{\sqrt{q}} \int_{b}^\infty \exp\left(-\frac{z^2}{4q}\right)\, dz.
\end{aligned}$$

By increasing b if necessary, we can ensure $\frac{1}{\sqrt{q}} \int_b^\infty \exp\left(-\frac{z^2}{4q}\right)\, dz \leq \beta$ which then gives
$\int_a^\infty \phi(\sqrt{q}x)^2 \exp(-x^2/2)\, dx \leq \beta$. A symmetric argument yields

$$\int_{-\infty}^a \phi(\sqrt{q}x)^2 \exp(-x^2/2)\, dx \leq \beta,$$

completing the proof.