The Emergence of Spectral Universality in Deep Networks

Jeffrey Pennington
Google Brain

Samuel S. Schoenholz
Google Brain

Surya Ganguli
Google Brain
Applied Physics, Stanford University

Abstract

Recent work has shown that tight concentration of the *entire* spectrum of singular values of a deep network's input-output Jacobian around one at initialization can speed up learning by orders of magnitude. Therefore, to guide important design choices, it is important to build a full theoretical understanding of the spectra of Jacobians at initialization. To this end, we leverage powerful tools from free probability theory to provide a detailed analytic understanding of how a deep network's Jacobian spectrum depends on various hyperparameters including the nonlinearity, the weight and bias distributions, and the depth. For a variety of nonlinearities, our work reveals the emergence of new universal limiting spectral distributions that remain concentrated around one even as the depth goes to infinity.

1 INTRODUCTION

A well-conditioned initialization is essential for successfully training neural networks. Seminal initial work focused on random weight initializations ensuring that the second moment of the spectrum of singular values of the network Jacobian from input to output remained one, thereby preventing exponential explosion or vanishing of gradients [1]. However, recent work has shown that even among different random initializations sharing this property, those whose *entire* spectrum tightly concentrates around one can often yield faster learning by orders of magnitude. For example, deep linear networks with orthogonal initializations, for which the entire spectrum is exactly one, can achieve depth-independent learning speeds, while the corresponding Gaussian initializations cannot [2].

Recently, it was shown [3] that a similarly well-conditioned Jacobian could be constructed for deep non-linear networks using a combination of orthogonal weights and tanh nonlinearities. The result of this improved conditioning was an orders-of-magnitude speedup in learning for tanh networks. However, the same study also proved that a well-conditioned Jacobian could not be achieved with Rectified Linear units (ReLUs). Together these results explained why, historically, in some cases orthogonal weight initialization had been found to improve training efficiency only slightly [4].

These empirical results connecting the conditioning of the Jacobian to a dramatic speedup in learning raise an important theoretical question. Namely, how does the entire shape of this spectrum depend on a network's nonlinearity, weight and bias distribution, and depth? Here we provide a detailed analytic answer by using powerful tools from free probability theory. Our answer provides theoretical guidance on how to choose these different network ingredients so as to achieve tight concentration of deep Jacobian spectra even at very large depths. Along the way, we find several surprises, and we summarize our results in the discussion.

2 PRELIMINARIES

2.1 Problem Setup

Consider an L-layer feed-forward neural network of width N with synaptic weight matrices $\mathbf{W}^l \in \mathbb{R}^{N\times N}$, bias vectors $\mathbf{b}^l$, pre-activations $\mathbf{h}^l$, and post-activations $\mathbf{x}^l$, with $l = 1, \ldots, L$. The forward-propagation dynamics are given by,

$$\mathbf{x}^l = \phi(\mathbf{h}^l)\,, \quad \mathbf{h}^l = \mathbf{W}^l\mathbf{x}^{l-1} + \mathbf{b}^l\,, \tag{1}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a pointwise nonlinearity and the input is $\mathbf{x}^0 \in \mathbb{R}^N$. Now consider the input-output

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

Jacobian $\mathbf{J} \in \mathbb{R}^{N \times N}$ given by

$$\mathbf{J} = \frac{\partial \mathbf{x}^L}{\partial \mathbf{x}^0} = \prod_{l=1}^{L} \mathbf{D}^l \mathbf{W}^l. \tag{2}$$

Here $\mathbf{D}^l$ is a diagonal matrix with entries $D^l_{ij} = \phi'(h^l_i)\,\delta_{ij}$, where δ_{ij} is the Kronecker delta function. The input-output Jacobian $\mathbf{J}$ is closely related to the backpropagation operator mapping output errors to weight matrices at a given layer, in the sense that if the former is well-conditioned, then the latter tends to be well-conditioned for all weight layers. We are therefore interested in understanding the entire singular value spectrum of $\mathbf{J}$ for deep networks with randomly initialized weights and biases.

In particular, we will take the biases $\mathbf{b}^l_i$ to be drawn i.i.d. from a zero-mean Gaussian with standard deviation σ_b. For the weights, we will consider two random matrix ensembles: (1) random *Gaussian* weights in which each W^l_{ij} is drawn i.i.d from a Gaussian with variance σ_w^2/N, and (2) random *orthogonal* weights, drawn from a uniform distribution over scaled orthogonal matrices obeying $(\mathbf{W}^l)^T\mathbf{W}^l = \sigma_w^2\,\mathbf{I}$.

2.2 Review of Signal Propagation

The random matrices $\mathbf{D}^l$ in (2) depend on the empirical distribution of pre-activations h^l_i for $i = 1, \ldots, N$ entering the nonlinearity ϕ in (1). The propagation of this empirical distribution through different layers l was studied in [5, 6]. In those works, it was shown that in the large N limit this empirical distribution converges to a Gaussian with zero mean and variance q^l, where q^l obeys a recursion relation induced by the dynamics in (1):

$$q^l = \sigma_w^2 \int \mathcal{D}h\, \phi\left(\sqrt{q^{l-1}}h\right)^2 + \sigma_b^2\,, \tag{3}$$

with initial condition $q^1 = \frac{\sigma_w^2}{N}\sum_{i=1}^{N}(x^0_i)^2 + \sigma_b^2$, and $\mathcal{D}h = \frac{dh}{\sqrt{2\pi}}\exp\left(-\frac{h^2}{2}\right)$ denoting the standard normal measure. This recursion has a fixed point obeying,

$$q^* = \sigma_w^2 \int \mathcal{D}h\, \phi\left(\sqrt{q^*}h\right)^2 + \sigma_b^2\,. \tag{4}$$

If the input $\mathbf{x}^0$ is chosen so that $q^1 = q^*$, then the dynamics start at the fixed point and the distribution of $\mathbf{D}^l$ is independent of l. Moreover, even if $q^1 \neq q^*$, a few layers is often sufficient to approximately converge to the fixed point (see [5, 6]). As such, when L is large, it is often a good approximation to assume that $q^l = q^*$ for all depths l when computing the spectrum of $\mathbf{J}$.

Another important quantity governing signal propagation through deep networks [5] is

$$\begin{aligned}\chi &= \frac{1}{N}\left\langle \mathrm{Tr}\,(\mathbf{DW})^T\mathbf{DW}\right\rangle \\ &= \sigma_w^2 \int \mathcal{D}h\left[\phi'\left(\sqrt{q^*}h\right)\right]^2,\end{aligned} \tag{5}$$

where ϕ' is the derivative of ϕ. Here χ is second moment of the distribution of squared singular values of the matrix $\mathbf{DW}$, when the pre-activations are at their fixed point distribution with variance q^*. As shown in [5, 6], $\chi(\sigma_w, \sigma_b)$ separates the (σ_w, σ_b) plane into two regions: (a) when $\chi > 1$, forward signal propagation expands and folds space in a chaotic manner and back-propagated gradients exponentially explode; and (b) when $\chi < 1$, forward signal propagation contracts space in an ordered manner and back-propagated gradients exponentially vanish. Thus the constraint $\chi(\sigma_w, \sigma_b) = 1$ determines a critical line in the (σ_w, σ_b) plane separating the ordered and chaotic regimes. Moreover, the second moment of the distribution of squared singular values of $\mathbf{J}$ was shown simply to be χ^L in [5, 6]. Fig. 1 shows an example of an order-chaos transition for the tanh nonlinearity.

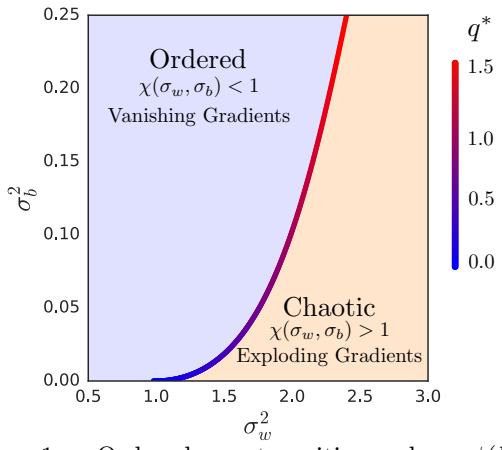

Figure 1: Order-chaos transition when $\phi(h) = \tanh(h)$. The critical line $\chi = 1$ determines the boundary between the two phases. In the chaotic regime $\chi > 1$ and gradients explode while in the ordered regime $\chi < 1$ and we expect gradients to vanish. The value of q^* along this line is shown as a heatmap.

2.3 Review of Free Probability

The previous section revealed that the mean squared singular value of $\mathbf{J}$ is χ^L. Indeed when $\chi \ll 1$ or $\chi \gg 1$ the vanishing or explosion of gradients, respectively, dominates the learning dynamics and provide a compelling case for choosing an initialization that is critical with $\chi = 1$. We would like to investigate the question of whether or not all cases where $\chi = 1$ are the same and, in particular, to obtain more detailed

information about entire the singular value distribution of $\mathbf{J}$ when $\chi = 1$. Since (2) consists of a product of random matrices, free probability becomes relevant as a powerful tool to compute the spectrum of $\mathbf{J}$, as we now review. See [7] for a pedagogical introduction, and [3, 8] for prior work applying free probability to deep learning.

In general, given a random matrix $\mathbf{X}$, its limiting spectral density is defined as

$$\rho_X(\lambda) \equiv \left\langle \frac{1}{N} \sum_{i=1}^{N} \delta(\lambda - \lambda_i) \right\rangle_X , \tag{6}$$

where $\langle \cdot \rangle_X$ denotes an average w.r.t to the distribution over the random matrix $\mathbf{X}$. The *Stieltjes transform* of ρ_X is defined as,

$$G_X(z) \equiv \int_{\mathbb{R}} \frac{\rho_X(t)}{z - t} dt , \qquad z \in \mathbb{C} \setminus \mathbb{R} , \tag{7}$$

which can be inverted using,

$$\rho_X(\lambda) = -\frac{1}{\pi} \lim_{\epsilon \to 0^+} \operatorname{Im} G_X(\lambda + i\epsilon) . \tag{8}$$

G_X is related to the moment generating function M_X,

$$M_X(z) \equiv z G_X(z) - 1 = \sum_{k=1}^{\infty} \frac{m_k}{z^k} , \tag{9}$$

where m_k is the kth moment of the distribution ρ_X,

$$m_k = \int d\lambda \; \rho_X(\lambda) \lambda^k = \frac{1}{N} \langle \operatorname{tr} \mathbf{X}^k \rangle_X . \tag{10}$$

In turn, we denote the functional inverse of M_X by M_X^{-1}, which by definition satisfies $M_X(M_X^{-1}(z)) = M_X^{-1}(M_X(z)) = z$. Finally, the *S-transform* [9, 10] is defined as,

$$S_X(z) = \frac{1 + z}{z M_X^{-1}(z)} . \tag{11}$$

The utility of the S-transform arises from its behavior under multiplication. Specifically, if $\mathbf{A}$ and $\mathbf{B}$ are two freely independent random matrices, then the S-transform of the product random matrix ensemble $\mathbf{AB}$ is simply the product of their S-transforms,

$$S_{AB}(z) = S_A(z) S_B(z) . \tag{12}$$

3 MASTER EQUATION FOR SPECTRAL DENSITY

3.1 S-transform for Jacobians

We can now write down an implicit expression of the spectral density of $\mathbf{J}\mathbf{J}^T$, which is also the distribution of the square of the singular values of $\mathbf{J}$. In particular, in the supplementary material (SM) Sec. 1, we combine (12) with the facts that the S-transform depends only on traces of moments through (9), and that these traces are invariant under cyclic permutations, to derive a simple expression for the S-transform of $\mathbf{J}\mathbf{J}^T$,

$$S_{JJ^T} = \prod_{l=1}^{L} S_{(D^l)^2} S_{(W^l)^T W^l} = S_{D^2}^L S_{W^T W}^L . \tag{13}$$

Here the lack of dependence on the layer index l on the RHS is valid if the input $\mathbf{x}^0$ is such that $q^1 = q^*$.

Thus, given expressions for the S-transforms associated with the nonlinearity, S_{D^2}, and the weights, $S_{W^T W}^L$, one can compute the S-transform of the input-output Jacobian S_{JJ^T} at any network depth L through (13). Then from S_{JJ^T}, one can invert the sequence (7), (9), and (11) to obtain $\rho_{JJ^T}(\lambda)$.

3.2 An Efficient Master Equation

The previous section provides a naive method for computing the spectrum $\rho_{JJ^T}(\lambda)$, through a complex sequence of calculations. One must start from $\rho_{W^T W}(\lambda)$ and $\rho_{D^2}(\lambda)$, compute their respective Stieltjes transforms, moment generating functions, inverse moment generating functions, and S-transforms, take the product in (13), and then invert this sequence of steps to finally arrive at $\rho_{JJ^T}(\lambda)$. Here we provide a much simpler "master" equation for extracting information about $\rho_{JJ^T}(\lambda)$ and its moments directly from knowledge of the moment generating function of the nonlinearity, $M_D^2(z)$, and the S-transform of the weights, $S_{W^T W}(z)$. As we shall see, these latter two functions are the simplest functions to work with for arbitrary nonlinearities.

To derive the master equation, we insert (11), for $\mathbf{X} = \mathbf{D}^2$, into (13), and perform some algebraic manipulations (see SM Sec. 3 for details) to obtain implicit functional equations for $M_{JJ^T}(z)$ and $G(z)$,

$$M_{JJ^T}(z) = M_{D^2}\left(z^{\frac{1}{L}} F\big(M_{JJ^T}(z)\big) \right) , \tag{14}$$

$$z G(z) - 1 = M_{D^2}\left(z^{\frac{1}{L}} F\big(z G(z) - 1\big) \right) , \tag{15}$$

where,

$$F(x) = S_{W^T W}(x) \left(\frac{1 + x}{x} \right)^{1 - \frac{1}{L}} . \tag{16}$$

In principle, a solution to eq. (15) allows us to compute the entire spectrum of $\boldsymbol{J}\boldsymbol{J}^T$. In practice, when an exact solution in terms of elementary functions is lacking, it is still possible to extract robust numerical solutions, as we describe in the next subsection.

Table 1: Properties of Nonlinearities

	$\phi(h)$	$M_{D^2}(z)$	μ_k	σ_w^2	$\sigma_{JJ^T}^2$
Linear	h	$\frac{1}{z-1}$	1	1	$L\left(-s_1\right)$
ReLu	$[h]_+$	$\frac{1}{2}\frac{1}{z-1}$	$\frac{1}{2}$	2	$L\left(1-s_1\right)$
Hard Tanh	$[h+1]_+ - [h-1]_+ - 1$	$\mathrm{erf}(\frac{1}{\sqrt{2q^*}})\frac{1}{z-1}$	$\mathrm{erf}(\frac{1}{\sqrt{2q^*}})$	$\frac{1}{\mathrm{erf}(\frac{1}{\sqrt{2q^*}})}$	$L\left(\frac{1}{\mathrm{erf}(\frac{1}{\sqrt{2q^*}})}-1-s_1\right)$
Erf	$\mathrm{erf}(\frac{\sqrt{\pi}}{2}h)$	$\frac{1}{\sqrt{\pi q^*}z}\Phi\left(\frac{1}{z},\frac{1}{2},\frac{1+\pi q_*}{\pi q_*}\right)$	$\frac{1}{\sqrt{1+\pi k q_*}}$	$\sqrt{1+\pi q^*}$	$L\left(\frac{1+\pi q^*}{\sqrt{1+2\pi q^*}}-1-s_1\right)$

3.3 Numerical Extraction of Spectra

Here we describe how to solve (15) numerically. The difficulty is that (15) implicitly defines $G(z)$ through an equation of the form $\mathcal{F}(G,z)=0$. Notice that, for any given z, this equation may have multiple roots in G. The correct branch can be chosen by requiring that $z\to\infty$, $G(z)\sim 1/z$ [11]. Therefore, one point on the correct branch can be found by taking $|z|$ large, and finding the solution to $\mathcal{F}(G,z)=0$ that is closest to $G=1/z$. Recall that to obtain the density $\rho_{JJ^T}(\lambda)$ through the inversion formula ((8)), we need to extract the behavior of $G(z)$ near the real axis at a point $z=\lambda+i\epsilon$ where $\rho_{JJ^T}(\lambda)$ has support. So, practically speaking, for each λ we can walk along the imaginary direction obeying $\mathrm{Re}(z)=\lambda$ from large imaginary values to small, and repeatedly solve $\mathcal{F}(G,z)=0$, always choosing the root that is closest to the previous root.

A potential pitfall arises if we approach a point z where $\mathcal{F}(G,z)=0$ has a double root in G, which could cause us to leave the correct branch of roots and then traverse an incorrect branch. However, points in the complex two dimensional plane $(G,z)\in\mathbb{C}^2$ where $\mathcal{F}$ has a double root in G are expected to be a set of measure 0, and in practice they do not seem to be a concern. Algorithm 1 summarizes our heuristic for computing $\rho(\lambda)$ for each λ of interest.

Algorithm 1 Root finding procedure

1. Choose to take $2N$ steps of size $b>1$
2. Initialize $z_0=\lambda+ib^N$ and $G_0=1/z_0$
3. For k in $1\dots 2N$:
   $z_k\leftarrow\lambda+ib^{N-k}$
   $G_k\leftarrow$ Root of (15) nearest to $G_{k-1}(z_k)$
4. Return $-\frac{1}{\pi}\mathrm{Im}\,G_{2N}\approx\rho(\lambda)$

In the following sections, we demonstrate through many examples a precise numerical match between the outcome of Algorithm 1 and direct simulations of various random neural networks, thereby justifying not only (15), but also the efficacy our algorithm.

3.4 Moments of Deep Spectra

In addition to numerically extracting the spectrum of $\mathbf{J}\mathbf{J}^T$, we can also calculate its moments m_k encoded in the function

$$M_{JJ^T}(z)\equiv\sum_{k=1}^{\infty}\frac{m_k}{z^k}\,. \tag{17}$$

These moments in turn can be computed in terms of the series expansions of S_{W^TW} and M_{D^2}, which we define as

$$S_{W^TW}(z)\equiv\ \sigma_w^{-2}\left(1+\textstyle\sum_{k=1}^{\infty}s_kz^k\right) \tag{18}$$

$$M_{D^2}(z)\equiv\ \textstyle\sum_{k=1}^{\infty}\frac{\mu_k}{z^k}\,, \tag{19}$$

where the moments μ_k of $\mathbf{D}^2$ are given by,

$$\mu_k=\int\mathcal{D}h\;\phi'(\sqrt{q^*}h)^{2k}\,. \tag{20}$$

Substituting these expansions into (14), we obtain equations for the unknown moments m_k in terms of the known moments μ_k and s_k. We can solve for the low-order moments by expanding (14) in powers of z^{-1}. By equating the coefficients of z^{-1} and z^{-2}, we find equations for m_1 and m_2 whose solution yields (see SM Sec. 3),

$$\begin{aligned}m_1&=(\sigma_w^2\mu_1)^L\\ m_2&=(\sigma_w^2\mu_1)^{2L}\,L\left(\frac{\mu_2}{\mu_1^2}+\frac{1}{L}-1-s_1\right)\,.\end{aligned} \tag{21}$$

Note the combination $\sigma_w^2\mu_1$ is none other than χ defined in (5), and so (21) recovers the result that the mean squared singular value m_1 of $\mathbf{J}$ either exponentially explodes or vanishes unless $\chi(\sigma_w,\sigma_b)=1$ on a critical boundary between order and chaos. However, *even* on this critical boundary where the mean m_1 of the spectrum of $\mathbf{J}\mathbf{J}^T$ is one for any depth L, the variance

$$\sigma_{JJ^T}^2=m_2-m_1^2=L\left(\frac{\mu_2}{\mu_1^2}-1-s_1\right) \tag{22}$$

grows linearly with depth L for generic values of μ_1, μ_2 and s_1. Thus $\mathbf{J}$ can be highly ill-conditioned at large depths L for generic choices of nonlinearities and weights, even when σ_w and σ_b are tuned to criticality.

4 SPECIAL CASES OF DEEP SPECTRA

Exploiting the master equation (14) requires information about $M_{D^2}(z)$, and $S_{WW^T}(z)$. We first provide this information and then use it to look at special cases of deep networks.

4.1 Transforms of Nonlinearities

First, for any nonlinearity $\phi(h)$, we have, through (7) and (9),

$$M_{D^2}(z) = \int \mathcal{D}h \frac{\phi'(\sqrt{q^*}h)^2}{z - \phi'(\sqrt{q^*}h)^2} \,. \tag{23}$$

The integral over the Gaussian measure $\mathcal{D}h$ reflects a sum over all the activations h_i^l in a layer l, since in the large N limit the empirical distribution of activations converges to a Gaussian with standard deviation $\sqrt{q^*}$. Moreover, an activation h_i^l feels a squared slope $\phi'(h_i^l)^2$, which appears as an eigenvalue of the diagonal matrix $(\mathbf{D}^l)^2$. Thus $M_{D^2}(z)$ naturally involves an integral over a function of $\phi'(\cdot)^2$ against a Gaussian.

Table 1 provides the moment generating function and moments of $\mathbf{D}^2$ for several nonlinearities. Detailed derivations of the results in Table 1, which follow from performing the integral in (23), can be found in the SM Sec. 3. In the Erf case, Φ is a special function known as the Lerch transcendent, which can be defined by its moments μ_k.

4.2 Transforms of Weights

Table 2: Transforms of weights

Random Matrix $\mathbf{W}$	$S_{W^TW}(z)$	s_1
Scaled Orthogonal	σ_w^{-2}	0
Scaled Gaussian	$\sigma_w^{-2}(1+z)^{-1}$	-1

The S-transforms of the weights can be obtain through the sequence of equations (7), (9), and (11), starting with $\rho_{W^TW}(\lambda) = \delta(\lambda - 1)$ for an orthogonal random matrix $\mathbf{W}$, and $\rho_{W^TW}(\lambda) = (2\pi)^{-1}\sqrt{4-\lambda}$ for $\lambda \in [0, 4]$, for a Gaussian random matrix $\mathbf{W}$ with variance $\frac{1}{N}$ (see SM Sec. 5). Furthermore, by scaling $\mathbf{W} \to \sigma_w \mathbf{W}$, the S-transform scales as $S_{W^TW} \to \sigma_w^{-2} S_{W^TW}$, yielding the S-transforms and first moments in Table 2.

4.3 Exact Properties of Deep Spectra

Now for different randomly initialized deep networks, we insert the appropriate expressions in Tables 1 and 2 into our master equations (14) and (15) to obtain information about the spectrum of $\mathbf{JJ}^T$, including its entire shape, through Algorithm 1, and its variance $\sigma^2_{JJ^T}$ through (21) and (22). We always work at criticality, so that in (5), $\chi = \sigma_w^2 \mu_1 = 1$. The resulting condition for σ_w^2 at criticality and the value of $\sigma^2_{JJ^T}$ are shown in Table 1 for different nonlinearities, both for orthogonal ($s_1 = 0$) and Gaussian ($s_1 = -1$) weights.

4.3.1 Linear Networks

For linear networks, the fixed point equation (4) reduces to $q^* = \sigma_w^2 q^* + \sigma_b^2$, and $(\sigma_w, \sigma_b) = (1, 0)$ is the only critical point. Moreover, linear Gaussian networks behave very differently from orthogonal ones. The latter are well conditioned, with $\sigma^2_{JJ^T} = 0$ because the product of orthogonal matrices is orthogonal and so $\rho_{JJ^T}(\lambda) = \delta(\lambda - 1)$ for all L. However, $\sigma^2_{JJ^T} = L$ for Gaussian weights. This radically different behavior of the spectrum of $\mathbf{JJ}^T$ is shown in Fig. 2A.

4.3.2 ReLU Networks

For ReLU networks, the fixed point equation (4) reduces to $q^* = \frac{1}{2}\sigma_w^2 q^* + \sigma_b^2$, and $(\sigma_w, \sigma_b) = (\sqrt{2}, 0)$ is the only critical point. Unlike the linear case, $\sigma^2_{JJ^T}$ becomes L for orthogonal and $2L$ for Gaussian weights. In essence, the ReLU nonlinearity destroys the qualitative scaling advantage that linear networks possess for orthogonal weights versus Gaussian. The qualitative similarity of spectra for ReLU Orthogonal and linear Gaussian is shown in Fig. 2AB.

4.3.3 Hard Tanh and Erf Networks

For Hard Tanh and Erf Networks, the criticality condition $\sigma_w^2 = \mu_1^{-1}$ does not determine a unique value of σ_w^2 because μ_1, the mean squared slope $\phi'(h)^2$, now depends on the variance q^* of the distribution of preactivations h. Since q^* itself is a function of σ_w and σ_b through (4), these networks enjoy an entire critical curve in the (σ_w, σ_b) plane, similar to that shown in Fig. 1. As q^* decreases monotonically towards zero, the corresponding point on this curve approaches the point $(\sigma_w, \sigma_b) = (1, 0)$.

Moreover, Table 1 shows that $\sigma^2_{JJ^T} = L(\mathcal{F}(q^*) - 1 - s_1)$ with $\lim_{q^* \to 0} \mathcal{F}(q^*) = 1$. This implies that for Gaussian weights ($s_1 = -1$), no matter how small one makes σ_w, $\sigma^2_{JJ^T} \propto L$. However, for orthogonal weights ($s_1 = 0$), for any fixed L, one can reduce σ_w and therefore q^*, so as to make $\sigma^2_{JJ^T}$ arbitrarily small. Thus Hard Tanh and Erf nonlinearities rescue the scaling advantage that orthogonal weights possess over Gaussian, which was present in linear networks, but destroyed in ReLU networks. Examples of the well-conditioned nature of orthogonal Hard Tanh and Erf networks compared to orthogonal ReLu networks are shown in Fig. 2.

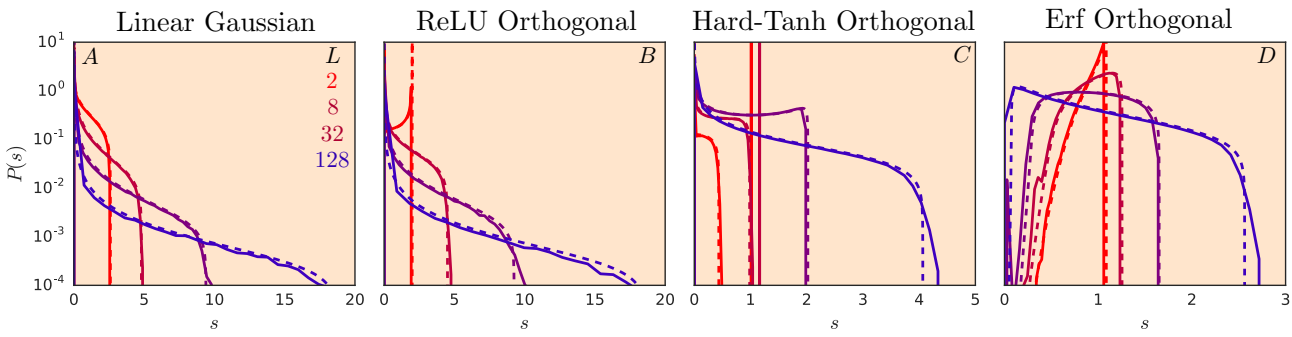

Figure 2: Examples of deep spectra at criticality for different nonlinearities at different depths. Singular values from empirical simulations of networks of width 1000 are shown with solid lines while theoretical predictions from the master equation and algorithm are overlaid with dashed lines. For each panel, the weight variance σ_w^2 is held constant as the depth increases. Notice that linear Gaussian and orthogonal ReLU have similarly-shaped distributions, especially for large depths, where poor conditioning and many large singular values are observed. Erf and Hard Tanh are better conditioned, but at 128 layers we begin to observe some spread in the distributions.

5 UNIVERSALITY IN DEEP SPECTRA

Table 1 shows that for orthogonal Erf and Hard Tanh networks (but not ReLU networks), since $\sigma^2_{JJ^T} = L(\mathcal{F}(q^*) - 1)$ with $\lim_{q^* \to 0} \mathcal{F}(q^*) = 1$, one can always choose q^* to vary inversely with L so as to achieve a desired L-independent constant variance $\sigma^2_{JJ^T} \equiv \sigma_0^2$. To achieve this scaling, $q^*(L)$ should satisfy the equation $\mathcal{F}(q^*(L)) = 1 + \frac{\sigma_0^2}{L}$, which implies $\sigma_w \to 1$ and $q^* \to 0$ as $L \to \infty$.

Remarkably, in this double scaling limit, not only does the variance of the spectrum of $\mathbf{J}\mathbf{J}^T$ remain constant at the fixed value σ_0^2, but the entire *shape* of the distribution converges to a *universal limiting distribution* as $L \to \infty$. There is more than one possible limiting distribution, but its form depends on ϕ only through the distribution of $\phi'(h)^2$ as $q^* \to 0$ via the expression for $M_{D^2}(z)$ in (23). Therefore, many qualitatively different activation functions may in fact be members of the same *universality class*. We identify two universality classes that correspond to many common activation functions: the *Bernoulli* universality class and the *smooth* universality class, named based on the distribution of $\phi'(h)^2$ as $q^* \to 0$.

The Bernoulli universality class contains many piecewise linear activation functions, such as Hard Tanh (Fig. 3C) and a version of ReLU shifted so as to be linear at the origin, which for concreteness we define as $\phi(x) = [x + \frac{1}{2}]_+ - \frac{1}{2}$ (Fig. 3E). While these functions look quite different, their derivatives are both Bernoulli-distributed (Fig. 3DF) and the limiting spectra of their corresponding Jacobians are the same (Fig. 4AB).

The smooth universality class contains many smooth activation functions, such as Erf (Fig. 3G) and a smoothed version of ReLU that we take to be the sigmoid-weighted linear unit (SiLU) [12, 13] (Fig. 3I). In this case, not only do the activation functions themselves look different, but so too do their derivatives (Fig. 3HJ). Nevertheless, in the double scaling limit, the limiting spectra of their corresponding Jacobians are the same (Fig. 4CD). The rate of convergence to the limiting distribution is different, because the moments μ_k differ substantially for non-zero q^*.

Unlike the smoothed and shifted versions of ReLU, the vanilla ReLU activation (Fig. 3AB) behaves entirely differently and has no limiting distribution because the μ_k are independent of q^* and therefore it is impossible to attain an L-independent constant variance $\sigma^2_{JJ^T} \equiv \sigma_0^2$ in this case.

To understand the mechanism behind the emergence of spectral universality, we now examine orthogonal networks whose activation functions have squared derivatives obeying a Bernoulli distribution and show that they all share a *universal* limiting distribution as $L \to \infty$. To this end, we suppose that,

$$M_{D^2} = p(q^*) \frac{1}{z - 1}, \tag{24}$$

for some function $p(q^*)$ that measures the probability of the nonlinearity having slope one as a function of q^*. We will assume that $p(q^*) \to 1$ as $q^* \to 0$. The relevant ratio of moments and the weight variance σ_w^2 are given as,

$$\frac{\mu_2}{\mu_1^2} = \frac{1}{\mu_1} = \sigma_w^2 = \frac{1}{p(q^*)}. \tag{25}$$

From (22), we have,

$$\sigma^2_{JJ^T} = \sigma_0^2 = L\left(\frac{1}{p(q^*)} - 1\right) \Rightarrow p(q^*) = 1 + \frac{\sigma_0^2}{L}. \tag{26}$$

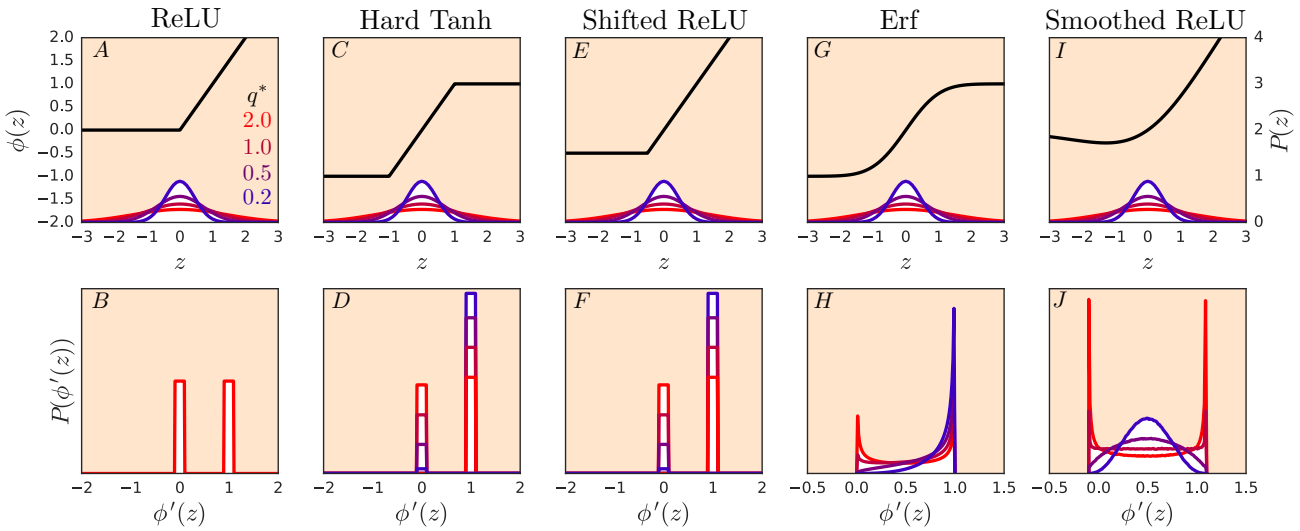

Figure 3: Distribution of $\phi'(h)$ for different nonlinearities. The top row shows the nonlinearity, $\phi(h)$, along with the Gaussian distribution of pre-activations h for four different choices of the variance, q^*. The bottom row gives the induced distribution of $\phi'(h)$. We see that for ReLU the distribution is independent of q^*. This implies that there is no stable limiting distribution for the spectrum of $\mathbf{J}\mathbf{J}^T$. By contrast for the other nonlinearities the distribution is a relatively strong function of q^*.

Notice that a solution $q^*(L)$ to (22) will exist for large L since we are assuming $p(q^*) \to 1$ as $q^* \to 0$. Substituting this solution in (24) and (25) gives for large L,

$$M_{D^2} = \frac{L}{L+\sigma_0^2}\frac{1}{z-1} \quad \text{and} \quad \mu_1 = \frac{L}{L+\sigma_0^2}\,. \tag{27}$$

Using these expressions and (11), we find that the S-transform obeys,

$$S_{JJ^T}^{\text{Bernoulli}} = (\mu_1 \frac{1+z}{z M_{D^2}^{-1}})^L = \left(1 + \frac{z\sigma_0^2}{L(1+z)}\right)^{-L}. \tag{28}$$

The large depth limit gives,

$$S_{JJ^T}^{\text{Bernoulli}} = e^{-\frac{z\sigma_0^2}{(1+z)}}\,. \tag{29}$$

Using (9) and (11) to solve for $G(z)$ gives,

$$G(z) = \frac{1}{z}\frac{\sigma_0^2}{\sigma_0^2 + W(\frac{-\sigma_0^2}{z})}\,, \tag{30}$$

where W denotes the principal branch of the Lambert-W function [14] and solves the transcendental equation,

$$W(x)e^{W(x)} = x\,. \tag{31}$$

The spectral density can be extracted from (30) easily using (8). The results are shown in black lines in Fig. 4AB. Both Hard Tanh and Shifted ReLU have Bernoulli-distributed $\phi'(h)^2$ and, despite being qualitatively different activation functions, have the same limiting spectral distributions. It is evident that the empirical spectral densities converge to this universal limiting distribution as the depth increases.

Next we build some additional understanding of the spectral density implied by (30). Because the spectral density is proportional to the imaginary part of $G(z)$, we expect the locations of the spectral edges to be related to branch points of $G(z)$, or more generally to poles in its derivative. Using the relation,

$$W'(x) = \frac{1}{x + e^{W(x)}}\,, \tag{32}$$

we can inspect the derivative of $G(z)$. It may be expressed as,

$$G'(z) = -\frac{\sigma_0^2\big(\sigma_0^2 + W(\frac{-\sigma_0^2}{z})(\sigma_0^2 + W(\frac{-\sigma_0^2}{z}))\big)}{z^2\big(1 + W(\frac{-\sigma_0^2}{z})\big)\big(\sigma_0^2 + W(\frac{-\sigma_0^2}{z})\big)^2}\,. \tag{33}$$

By inspection, we find that $G'(z)$ has double poles at,

$$z = \lambda_0 = 0, \quad z = \lambda_2 = e^{\sigma_0^2}\,, \tag{34}$$

which are locations where the spectral density diverges, i.e. there are delta function peaks at λ_0 and λ_2. Note that there is only a pole at λ_2 if $\sigma_0 \leq 1$. There is also a single pole at,

$$\lambda_1 = \sigma_0^2 e\,, \tag{35}$$

which defines the right spectral edge, i.e. the maximum value of the bulk of the density.

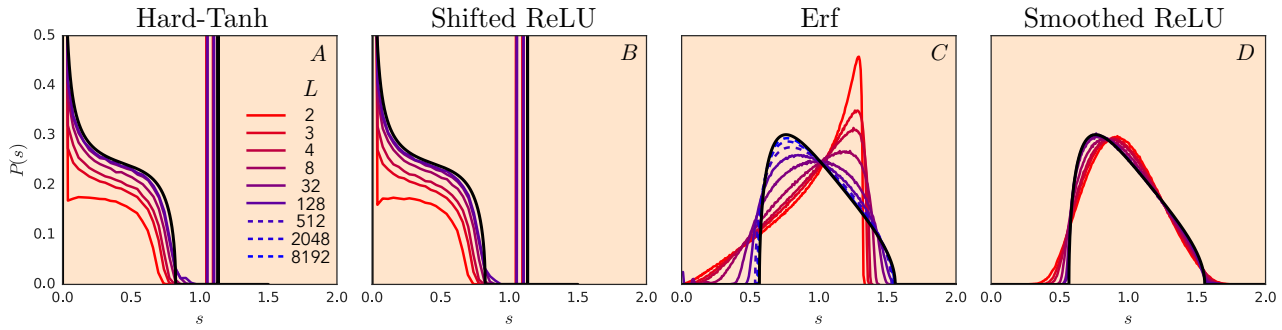

Figure 4: Two limiting universality classes of Jacobian spectra. Hard Tanh and Shifted ReLU fall into one class, characterized by Bernoulli-distributed $\phi'(h)^2$, while Erf and Smoothed ReLU fall into a second class, characterized by a smooth distribution for $\phi'(h)^2$. The black curves are theoretical predictions for the limiting distributions with variance $\sigma_0^2 = 1/4$. The colored lines are emprical spectra of finite-depth width-1000 orthogonal neural networks. The empirical spectra converge to the limiting distributions in all cases. The rate of convergence is similar for Hard-Tanh and Shifted ReLU, whereas it is significantly different for Erf and Smoothed Relu, which converge to the same limiting distribution along distinct trajectories. In all cases, the solid colored lines go from shallow $L = 2$ networks (red) to deep networks (purple). In all cases but Erf the deepest networks have $L = 128$. For Erf, the dashed lines show solutions to (15) for very large depth up to $L = 8192$.

The above observations regarding λ_0, λ_1, and λ_2 are evident in Fig. 4AB. Noting that in the figure, $\sigma_0 = 1/2$, we predict that the bulk of the density to have its right edge located at $s = \sqrt{\lambda_1} = \sqrt{e}/2 \approx 0.82$ and that there should be a delta function peak at $s = \sqrt{\lambda_2} = e^{1/8} \approx 1.13$, both of which are reflected in the figure.

A similar analysis can be carried out for activation functions for which the distribution of $\phi'(h)^2$ is smooth and concentrates around one as $q^* \to 0$. The analysis for Erf is presented in the SM. We find that,

$$S_{JJ^T}^{\text{Smooth}} = e^{-z\sigma_0^2}, \tag{36}$$

and that $G(z)$ can be expressed in terms of a generalized Lambert-W function [15]. The locations of the spectral edges are given by $s_\pm = e^{-\frac{1}{4}\sigma_\pm^2}\sqrt{1+\frac{1}{2}\sigma_\mp^2}$, where,

$$\sigma_\pm^2 = \sigma_0\left(\sigma_0 \pm \sqrt{\sigma_0^2+4}\right). \tag{37}$$

For $\sigma_0 = 1/2$, these results give $s_- \approx 0.57$ and $s_+ = 1.56$, which is in excellent agreement with the behavior observed in Fig. 4CD. Overall, Fig. 4 provides strong evidence supporting our predictions that orthogonal Hard Tanh and shifted ReLU networks have the Bernoulli limit distribution, while orthogonal Erf and smoothed Relu networks have the smooth limit distribution.

Finally, we derived these universal limits assuming orthogonal weights. In the SM we show that orthogonality is in fact necessary for the existence of a stable limiting distribution for the spectrum of $\mathbf{J}\mathbf{J}^T$. No other random matrix ensemble can yield a stable distribution for *any* choice of nonlinearity with $\phi'(0) = 1$. Essentially, any spread in the singular values of $\mathbf{W}$ grows in an unbounded way with depth and cannot be nonlinearly damped.

6 DISCUSSION

In summary, motivated by a lack of theoretical clarity on when and why different weight initializations and nonlinearities combine to yield well-conditioned spectra that speed up deep learning, we developed a calculational framework based on free probability to provide, with unprecedented detail, analytic information about the *entire* Jacobian spectrum of deep networks with *arbitrary* nonlinearities. Our results provide a principled framework for the initialization of weights and the choice of nonlinearities in order to produce well-conditioned Jacobians and fast learning. Intriguingly, we find novel universality classes of deep spectra that remain well-conditioned as the depth goes to infinity, as well as theoretical conditions for their existence. Our results lend additional support to the surprising conclusions revealed in [3], namely that using either Gaussian initializations or ReLU nonlinearities precludes the possibility of obtaining stable spectral distributions for very deep networks. Beyond the sigmoidal units advocated in [3], our results suggest that a wide variety of nonlinearities, including shifted and smoothed variants of ReLU, can achieve dynamical isometry, provided the weights are orthogonal. Interesting future work could involve the discovery of new universality classes of well-conditioned deep spectra for more diverse nonlinearities than considered here.

References

[1] Xavier Glorot and Yoshua 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.

[2] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *International Conference on Learning Representations*, abs/1312.6120, 2014.

[3] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in neural information processing systems*, pages 4788–4798, 2017.

[4] Dmytro Mishkin and Jiri Matas. All you need is a good init. *CoRR*, abs/1511.06422, 2015.

[5] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 2016.

[6] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep Information Propagation. *ArXiv e-prints*, November 2016.

[7] James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

[8] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, pages 2798–2806, 2017.

[9] Roland Speicher. Multiplicative functions on the lattice of non-crossing partitions and free convolution. *Mathematische Annalen*, 298(1):611–628, 1994.

[10] Dan V Voiculescu, Ken J Dykema, and Alexandru Nica. *Free random variables*. Number 1. American Mathematical Soc., 1992.

[11] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Society Providence, RI, 2012.

[12] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 2018.

[13] P. Ramachandran, B. Zoph, and Q. V. Le. Searching for Activation Functions. *ArXiv e-prints*.

[14] Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the lambertw function. *Advances in Computational mathematics*, 5(1):329–359, 1996.

[15] István Mező and Árpád Baricz. On the generalization of the lambert function. *Transactions of the American Mathematical Society*, 369(11):7917–7934, 2017.