Mathematical
Institute

Random DNN: Jacobian and the exploding / vanishing gradient problem

Theories of Deep Learning: C6.5, Video 7

Prof. Jared Tanner
Mathematical Institute
University of Oxford

Oxford
Mathematics

University of Oxford
Mathematical Institute

Consider a fully connected L layer deep net given by

$$h^{(\ell)} = W^{(\ell)} z^{(\ell)} + b^{(\ell)}, \qquad z^{(\ell+1)} = \phi(h^{(\ell)}), \qquad \ell = 0, \dots, L-1,$$

for $\ell = 1, \dots, L$ with nonlinear activation $\phi(\cdot)$ and $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_\ell}$.

Its Jacobian is given by

$$J = \frac{\partial h^{(L)}}{\partial x^{(0)}} = \Pi_{\ell=1}^{L} D^{(\ell)} W^{(\ell)}$$

where $D^{(\ell)}$ is diagonal with entries $D_{ii}^{(\ell)} = \phi'(h_i^{(\ell)})$.
Which, amongst other things, can bound the local stability of the DNN: $\|H(x+\delta;\theta) - H(x;\theta)\| = \|J\delta + \mathcal{O}(\|\delta\|^2)\| \leq \|\delta\| \max \|J\|$.

UNIVERSITY OF OXFORD
Mathematical Institute

$$\mathcal{L}(\theta; X, Y) = (2m)^{-1} \sum_{\mu=1}^{m} \sum_{i=1}^{n_L} (H(x_\mu(i); \theta) - y_{i,\mu})^2$$

Letting $\delta_\ell := \frac{\partial \mathcal{L}}{\partial h^{(\ell)}}$ and as before $D^{(\ell)}$ the diagonal matrix with $D^{(\ell)}_{ii} = \phi'(h^{(\ell)}_i)$ we have

$$\delta_\ell = D^{\ell}(W^{(\ell)})^T \delta_{\ell+1} \quad \text{and} \quad \delta_L = D^{(L)} \text{grad}_{h^{(L)}} \mathcal{L}.$$

which gives the formula for computing the δ_ℓ for each layer as

$$\delta_\ell = \left(\Pi_{k=\ell}^{L-1} D^{(k)} (W^{(k)})^T \right) D^{(L)} \text{grad}_{h^{(L)}} \mathcal{L}.$$

and the resulting gradient $\text{grad}_\theta \mathcal{L}$ with entries as

$$\frac{\partial \mathcal{L}}{\partial W^{(\ell)}} = \delta_{\ell+1} \cdot h_\ell^T \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial b^{(\ell)}} = \delta_{\ell+1}$$

Spectrum of the Jacobian pt. 1(Pennington et al. 18')

How to compute the product of $D^{(\ell)}W^{(\ell)}$

Computing the spectrum of products of matrices, e.g. for $J = \frac{\partial h^{(L)}}{\partial x^{(0)}} = \Pi_{\ell=1}^{L} D^{(\ell)} W^{(\ell)}$ where $D_{ii}^{(\ell)} = \phi'(h_i^{(\ell)})$.

Stieltjes and $\mathcal{S}$ Transforms

For $z \in \mathbb{C}/\mathbb{R}$ the Stieltjes Transform, $G_\rho(z)$, of a probability distribution and its inverse are given by

$$G_\rho(z) = \int_{\mathbb{R}} \frac{\rho(t)}{z-t} dt \quad \text{and} \quad \rho(\lambda) = -\pi^{-1} \lim_{\epsilon \to 0_+} Imag(G_\rho(\lambda + i\epsilon)).$$

The Stieltjes Transform and moment generating function are related by $M_\rho(z) := zG_\rho(z) - 1 = \sum_{k=1}^{\infty} \frac{m_k}{z^k}$, and the $\mathcal{S}$ Transform is defined as $S_\rho(z) = \frac{1+z}{zM_\rho^{-1}(z)}$. The $\mathcal{S}$ Transform has the property that if ρ_1 and ρ_2 are freely independent then $\mathcal{S}_{\rho_1\rho_2} = \mathcal{S}_{\rho_1}\mathcal{S}_{\rho_2}$.

`https://arxiv.org/pdf/1802.09979.pdf`

The $\mathcal{S}$ Transform of JJ^T with $J = \frac{\partial h^{(L)}}{\partial x^{(0)}} = \Pi_{\ell=1}^{L} D^{(\ell)} W^{(\ell)}$ is then given by

$$\mathcal{S}_{JJ^T} = \mathcal{S}_{D^2}^{L} \mathcal{S}_{W^T W}^{L}.$$

This can be computed through the moments $M_{JJ^T}(z) = \sum_{k=1}^{\infty} \frac{m_k}{z^k}$, $M_{D^2}(z) = \sum_{k=1}^{\infty} \frac{\mu_k}{z^k}$, and $\mathcal{S}_{W^T W} = \sigma_w^{-2} \left(1 + \sum_{k=1}^{\infty} s_k z^k\right)$ where

$$\mu_k = \int (2\pi)^{-1/2} \phi' \left(\sqrt{q^{(\star)}} z\right)^{2k} e^{-z^2/2} dz.$$

In particular: $m_1 = (\sigma_w^2 \mu_1)^L$ and $\sigma_w^2 \mu_1 = \chi$ is the growth factor we observed before, requiring $\chi = 1$ to avoid rapid convergence of correlations to fixed points.

`https://arxiv.org/pdf/1802.09979.pdf`

Random DNN recursion map (Poole et al. 16')

Norm of hidden layer output dependence on prior layer

UNIVERSITY OF OXFORD
Mathematical Institute

Let $q^{(\ell)} = n_\ell^{-1} \|h^{(\ell)}\|_2^2$ be the average squared ℓ_2 length of the pre-activation $h^{(\ell)} = W^{(\ell)} z^{(\ell-1)} + b^{(\ell)}$ at layer ℓ, then with $W^{(\ell)}$ and $b^{(\ell)}$ being drawn from $\mathcal{N}(0, \sigma_w^2/n_\ell)$ and $\mathcal{N}(0, \sigma_b^2/n_\ell)$ respectively, we can express the evolution of the length as

$$q^{(\ell)} = \sigma_w^2 n_{\ell-1}^{-1} \|\phi(h^{(\ell-1)})\|_2^2 + \sigma_b^2.$$

Replacing the average squared length $n^{-1}\|\cdot\|_2^2$ for large n by the squared integral we could instead consider the propagation

$$q^{(\ell)} := \sigma_w^2 \int (2\pi)^{-1/2} \phi\left(\sqrt{q^{(\ell-1)}} z\right)^2 e^{-z^2/2} dz + \sigma_b^2.$$

`https://arxiv.org/pdf/1606.05340.pdf`

Random DNN recursion map (Poole et al. 16')

Norm of hidden layer output dependence on prior layer

UNIVERSITY OF OXFORD
Mathematical Institute

The average squared length $q^\ell = N^{-1}\|h^{(\ell)}\|_2^2$ of the pre-activation following the recursion

$$q^{(\ell)} := \sigma_w^2 \int (2\pi)^{-1/2} \phi\left(\sqrt{q^{(\ell-1)}} z\right)^2 e^{-z^2/2} dz + \sigma_b^2.$$

has a fixed point $q^\star = \sigma_w^2 \int (2\pi)^{-1/2} \phi\left(\sqrt{q^{(\star)}} z\right)^2 e^{-z^2/2} dz + \sigma_b^2$ whose stability governs the ability of the network to train. In fact, the growth of a perturbation is given by the expected mean singular value of $J^T J$ which is given by

$$\chi = \sigma_w^2 \int (2\pi)^{-1/2} \phi'\left(\sqrt{q^{(\star)}} z\right)^2 e^{-z^2/2} dz.$$

`https://arxiv.org/pdf/1606.05340.pdf`

Stability of pre-activation lengths (Pennington et al. 18')

The "Edge of Chaos Curve" for $\phi(\cdot) = tanh(\cdot)$.

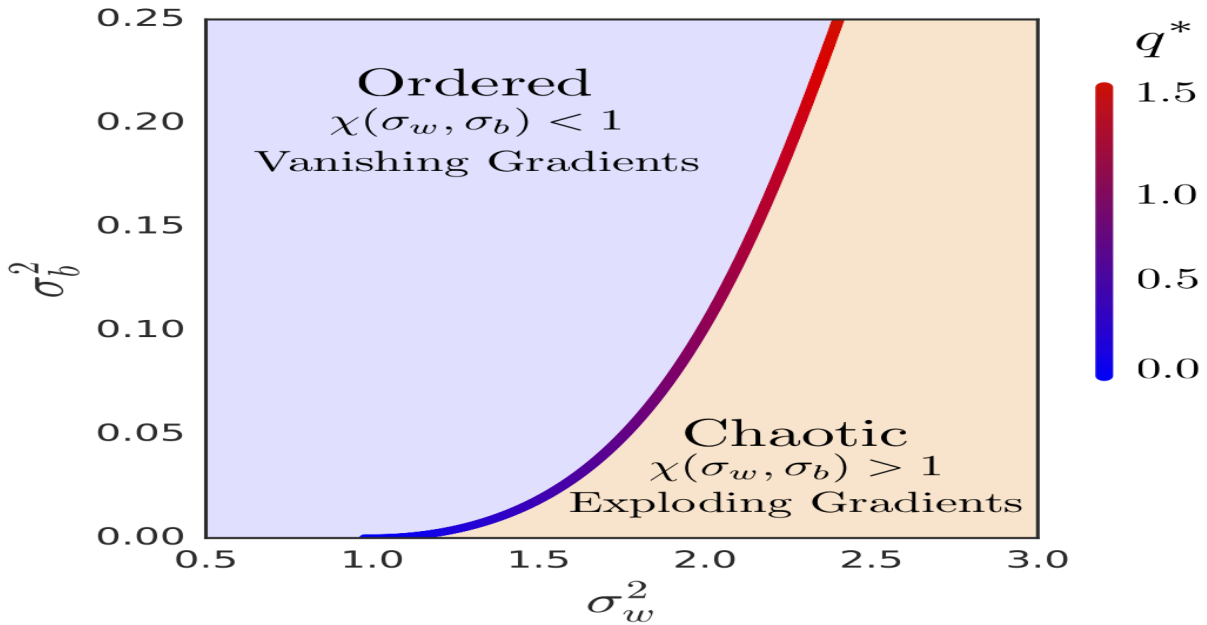


Figure 1: Order-chaos transition when $\phi(h) = \tanh(h)$. The critical line $\chi = 1$ determines the boundary between the two phases. In the chaotic regime $\chi > 1$ and gradients explode while in the ordered regime $\chi < 1$ and we expect gradients to vanish. The value of q^* along this line is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

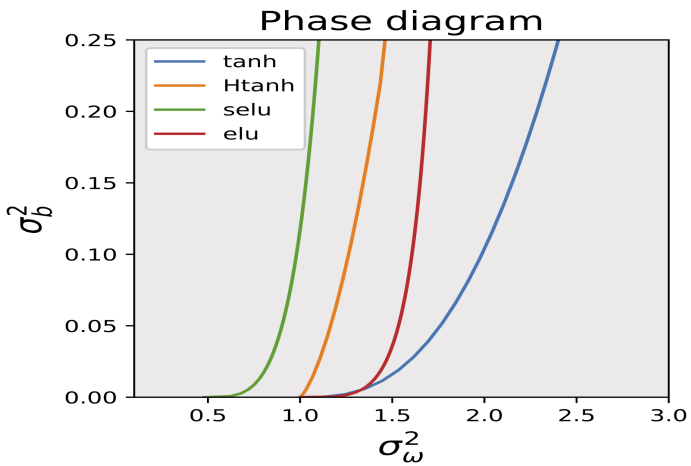
https://arxiv.org/pdf/1802.09979.pdf

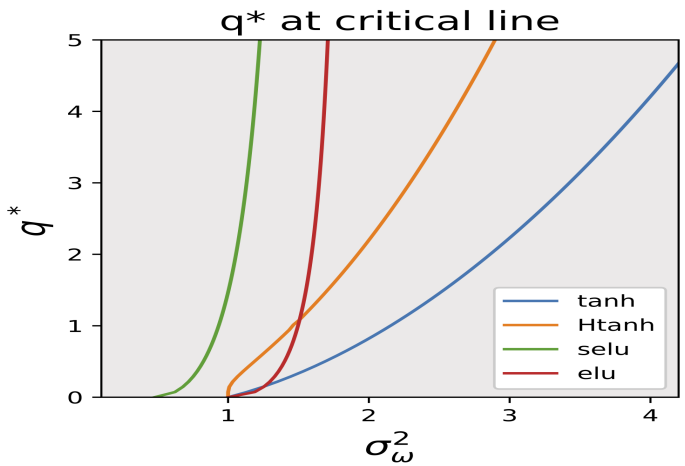
Network variance control through depth (Abrol 19')

Edge of chaos curves for other nonlinear activations

UNIVERSITY OF OXFORD
Mathematical Institute

The pre-activation output of networks converge to a zero-mean Gaussian distribution with variance, q^*, specified by the nonlinear activation, weight and bias variance, σ_w and σ_b respective. The distribution of the network input-output spectrum has a mean at layer d given by χ^d. Level curves of $\chi = 1$ overcome the exponential dependence on depth and allow training.





Initialisation on this curve allows training very deep networks.

Nonlinear activation stability (Pennington et al. 18')

Examples of moment generating functions

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	$L\,(-s_1)$
ReLu	$[h]_+$	$\frac{1}{2}\frac{1}{z-1}$	$\frac{1}{2}$	2	$L\,(1-s_1)$
Hard Tanh	$[h+1]_+ - [h-1]_+ - 1$	$\mathrm{erf}(\frac{1}{\sqrt{2q^*}})\frac{1}{z-1}$	$\mathrm{erf}(\frac{1}{\sqrt{2q^*}})$	$\frac{1}{\mathrm{erf}(\frac{1}{\sqrt{2q^*}})}$	$L\,(\frac{1}{\mathrm{erf}(\frac{1}{\sqrt{2q^*}})} - 1 - s_1)$
Erf	$\mathrm{erf}(\frac{\sqrt{\pi}}{2}h)$	$\frac{1}{\sqrt{\pi q^* z}}\Phi\left(\frac{1}{z}, \frac{1}{2}, \frac{1+\pi q_*}{\pi q_*}\right)$	$\frac{1}{\sqrt{1+\pi k q_*}}$	$\sqrt{1+\pi q^*}$	$L\,(\frac{1+\pi q^*}{\sqrt{1+2\pi q^*}} - 1 - s_1)$

Where $M_{D^2}(z) = \sum_{k=1}^{\infty} \frac{\mu_k}{z^k}$ with $\mu_k = \int (2\pi)^{-1/2} \phi'\left(\sqrt{q^{(\star)}} z\right)^{2k} e^{-z^2/2} dz$. For W Gaussian $s_1 = -1$ where as for W orthogonal $s_1 = 0$. Note that for all nonlinear activations for $\mu_1 \sigma_w^2 = 1$, $\sigma_{JJ^T}^2$ grows linearly with L.

Linear $\phi(\cdot)$: $q^* = \sigma_w^2 q^* + \sigma_b^2$, and fixed point $(\sigma_w, \sigma_b) = (1, 0)$.

ReLU $\phi(\cdot)$: $q^* = \frac{1}{2}\sigma_w^2 q^* + \sigma_b^2$, and fixed point $(\sigma_w, \sigma_b) = (\sqrt{2}, 0)$.

Hard Tanh and Erf have curves as fixed points $\chi(\sigma_w, \sigma_b)$.

`https://arxiv.org/pdf/1802.09979.pdf`

Distribution of activations $\phi'(z)$ (Pennington et al. 18')

$$\mu_k = \int (2\pi)^{-1/2} \phi' \left(\sqrt{q^{(\star)}} z \right)^{2k} e^{-z^2/2} dz.$$

University of Oxford
Mathematical Institute

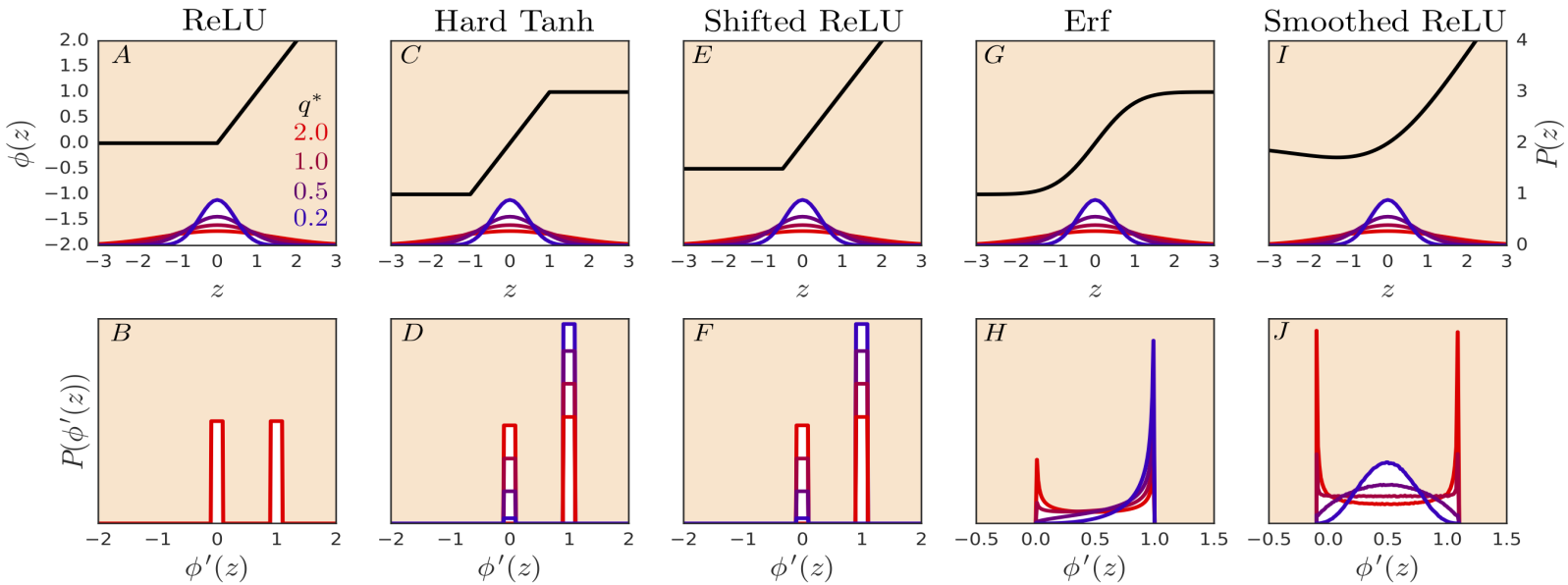


Figure 3: Distribution of $\phi'(h)$ for different nonlinearities. The top row shows the nonlinearity, $\phi(h)$, along with the Gaussian distribution of pre-activations h for four different choices of the variance, q^*. The bottom row gives the induced distribution of $\phi'(h)$. We see that for ReLU 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^*.

https://arxiv.org/pdf/1802.09979.pdf

Distribution of Jacobian spectra (Pennington et al. 18')

Observed universality of spectra based on $\phi(\cdot)$

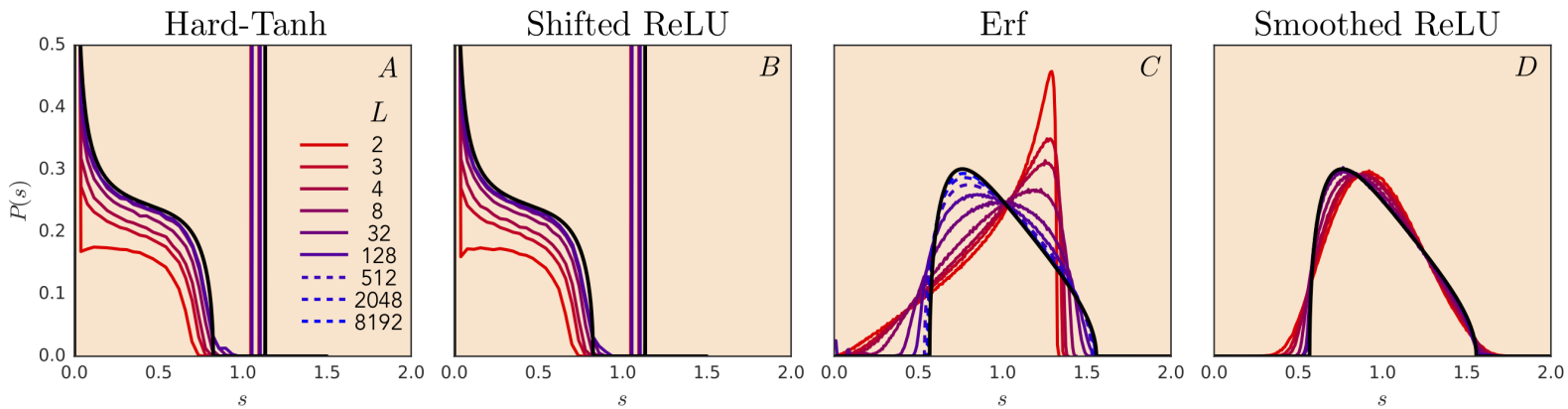


Figure 4: Two limiting universality classes of Jacobian spectra. Hard Tanh and Shifted ReLU fall into one class, characterized by Bernoulli-distributed $\phi'(h)^2$, while Erf and Smoothed ReLU fall into a second class, characterized by a smooth distribution for $\phi'(h)^2$. The black curves are theoretical 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 $L = 8192$.

`https://arxiv.org/pdf/1802.09979.pdf`

Summary of random DNN initialisation

Dependence between $\sigma_w, \sigma_b, \phi(\cdot)$

UNIVERSITY OF OXFORD
Mathematical Institute

- Poole et al. 16' showed pre-activation output is well modelled as Gaussian with variance q^* determined by $\sigma_w, \sigma_b, \phi(\cdot)$. Moreover, the correlation between two inputs follows a similar map with correlations converging to a fixed point, with the behaviour determined in part by χ where $\chi = 1$ avoids correlation to the same point, or nearby points diverging. `https://arxiv.org/pdf/1606.05340.pdf`
- Pennington et al 18' showed more generally how to compute the moments for the Jacobian spectra, where $\chi = 1$ is needed to avoid exponential growth or shrinkage with depth of gradients. `https://arxiv.org/pdf/1802.09979.pdf`