Review of Lecture 7: random matrix theory view of loss function (Pennington et al. 17'[1])

Let $e_{i,\mu} = \hat{y}_{i,\mu} - y_{i,\mu}$ be the error in the i^{th} entry of the output for data entry indexed by μ, and $\theta = \{W^{(1)}, W^{(2)}\} \in \mathbb{R}^{2n^2}$ be the net parameters, then the hessian of the sum of squares loss function has entries

$$H_{\alpha,\beta} = \frac{\partial^2 \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta} =: H_0 + H_1$$

with H_0 positive semi-definite and H_1 indefinite.
Modelling H_0 and H_1 as Wishart and Wigner respectively, the additive spectra can be computed and fraction of non-negative eigenvalues determined along with point where $2n/m$ and $n^{-1}\mathcal{L}$ predict the loss function is convex.

[1] http://proceedings.mlr.press/v70/pennington17a.html

Outline for today

- Jacobian of the feed forward deep net, length propagation.
- Stability or exponential growth/shrinkage of length with depth; computation of the spectra through $\mathcal{S}$ Transform
- Role of nonlinear activations and length fixed point maps.
- Deep spectra and distributions of the activation derivatives
- Classes of universality in the spectra for diverse activation functions.

Jacobian of deep net

Consider a fully connected L layer deep net given by

$$h^{(\ell)} = \phi(\hat{h}^{(\ell)}) \quad \text{with} \quad \hat{h}^{(\ell)} = W^{(\ell)} h^{(\ell-1)} + b^{(\ell)}$$

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

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 $D_{ii}^{(\ell)} = \phi'(\hat{h}_i^{(\ell)})$.

We further consider the case of a random net, $W^{(\ell)}$ and $b^{(\ell)}$ drawn from specified distributions, and investigate how the ℓ^2 length of input vectors change as they are propagated through the net.

Length propagation (Poole et al. 16'[2])

Let $q^\ell = N^{-1}\|h^{(\ell)}\|_2^2$ be the average squared ℓ_2 length of the pre-activation $\hat{h}^{(\ell)} = W^{(\ell)}h^{(\ell-1)} + b^{(\ell)}$ at layer ℓ.

Treating the model of $W^{(\ell)}$ and $b^{(\ell)}$ being drawn from $\mathcal{N}(0,\sigma_w^2)$ and $\mathcal{N}(0,\sigma_b^2)$ respectively, we can express the evolution of the length as

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

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

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

[2] https://arxiv.org/pdf/1606.05340.pdf

Length propagation fixed point (Poole et al. 16'[3])

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 largest singular value of $J^T J$, that is $\|Ju\|_2^2 / \|u\|_2^2$ which is given by

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

[3] https://arxiv.org/pdf/1606.05340.pdf

Stability of pre-activation lengths (Pennington et al. 18'[4])

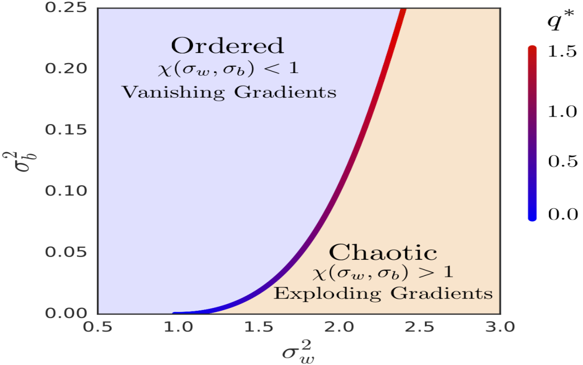


Figure 1: Order-chaos transition when $\phi(h) = \tanh(h)$. The critical line $\chi = 1$ determines the boundary between the two phases. In the chaotic regime $\chi > 1$ and gradients explode while in the ordered regime $\chi < 1$ and we expect gradients to vanish. The value of q^* along this line is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

[4] https://arxiv.org/pdf/1802.09979.pdf

Spectrum of the Jacobian pt. 1 (Pennington et al. 18'[5])

Recall the input-output 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 $D_{ii}^{(\ell)} = \phi'(h_i^{(\ell)})$.

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 $\mathcal{S}_{\rho_1\rho_2} = \mathcal{S}_{\rho_1}\mathcal{S}_{\rho_2}$.

[5] https://arxiv.org/pdf/1802.09979.pdf

Spectrum of the Jacobian pt. 2 (Pennington et al. 18'[6])

Recall the input-output 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 $D_{ii}^{(\ell)} = \phi'(\hat{h}_i^{(\ell)})$.
The $\mathcal{S}$ Transform of JJ^T 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 for which $\chi = 1$ has controlled growth through the layers..

[6] https://arxiv.org/pdf/1802.09979.pdf

Nonlinear activation stability (Pennington et al. 18'[7])

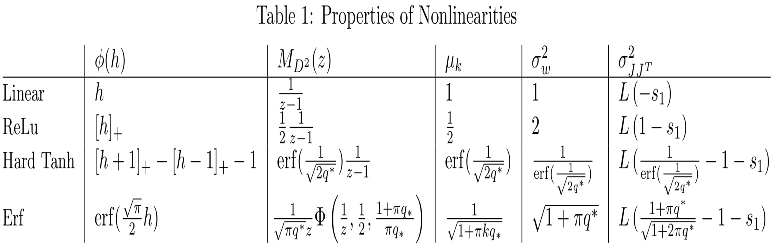


Table 1: Properties of Nonlinearities

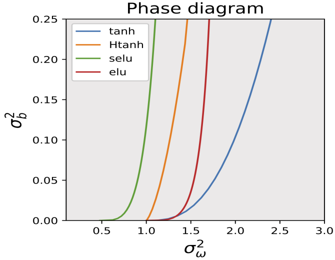
	$\phi(h)$	$M_{D^2}(z)$	μ_k	σ_w^2	$\sigma_{JJ^T}^2$
Linear	h	$\frac{1}{z-1}$	1	1	$L\left(-s_1\right)$
ReLu	$[h]_+$	$\frac{1}{2}\frac{1}{z-1}$	$\frac{1}{2}$	2	$L\left(1-s_1\right)$
Hard Tanh	$[h+1]_+ - [h-1]_+ - 1$	$\text{erf}(\frac{1}{\sqrt{2q^*}})\frac{1}{z-1}$	$\text{erf}(\frac{1}{\sqrt{2q^*}})$	$\frac{1}{\text{erf}(\frac{1}{\sqrt{2q^*}})}$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

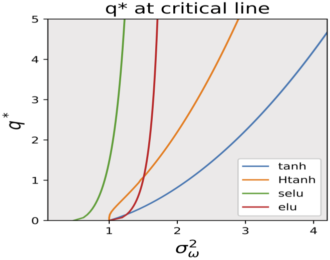
Where $M_{D^2}(z) = \int (2\pi)^{-1/2} \frac{\phi'\left(\sqrt{q^{(\star)}z}\right)^2}{z-\phi'\left(\sqrt{q^{(\star)}z}\right)^2} e^{-z^2/2} dz$ and for W Gaussian $s_1 = -1$ where as for W orthogonal $s_1 = 0$. Note that for all nonlinear activations for $\mu_1\sigma_w^2 = 1$, $\sigma_{JJ^T}^2$ grows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

[7] `https://arxiv.org/pdf/1802.09979.pdf`

Network variance control through depth (Abrol 19')

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, but adding batch-normalization can causes complete inability to train.

Deep spectra of Jacobian (Pennington et al. 18'[8])

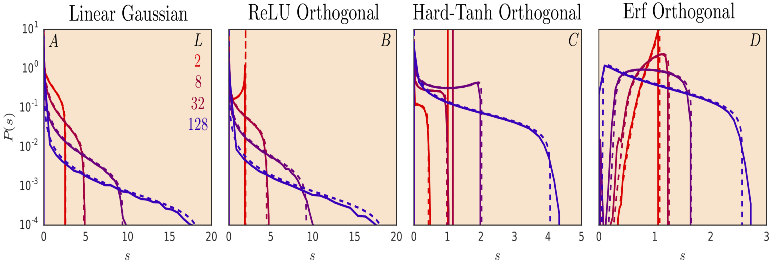


Figure 2: Examples of deep spectra at criticality for different nonlinearities at different depths. Singular values from empirical simulations of networks of width 1000 are shown with solid lines while theoretical predictions from the master equation and algorithm are overlaid with dashed lines. For 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.

Linear Gaussian and ReLU exhibit greater growth in the spectra with depth as compared to Hard Tahh and Erf.

[8] https://arxiv.org/pdf/1802.09979.pdf

Nonlinear activation stability (Pennington et al. 18'[9])

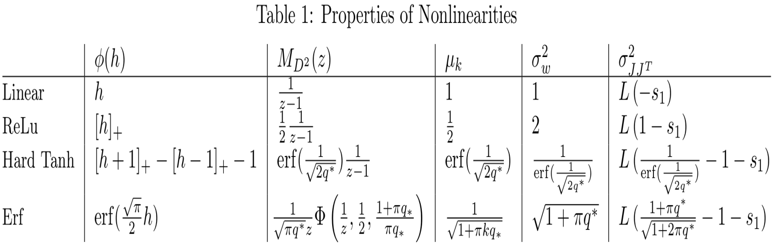


Table 1: Properties of Nonlinearities

	$\phi(h)$	$M_{D^2}(z)$	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	$\mathrm{erf}(\frac{\sqrt{\pi}}{2}h)$	$\frac{1}{\sqrt{\pi q^*}z}\Phi\left(\frac{1}{z}, \frac{1}{2}, \frac{1+\pi q_*}{\pi q_*}\right)$	$\frac{1}{\sqrt{1+\pi k q_*}}$	$\sqrt{1+\pi q^*}$	$L\left(\frac{1+\pi q^*}{\sqrt{1+2\pi q^*}} - 1 - s_1\right)$

Where $M_{D^2}(z) = \int (2\pi)^{-1/2} \frac{\phi'\left(\sqrt{q^{(\star)}z}\right)^2}{z-\phi'\left(\sqrt{q^{(\star)}z}\right)^2} e^{-z^2/2} dz$ and 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 and ReLU have $\sigma_{JJ^T}^2$ growing linearly with L (except linear orthogonal where $s_1 = 0$).
Hard Tanh and Erf: $q^*(L)$ can be selected such that $\sigma_{JJ^T}^2$ approaches a fixed value as $L \to \infty$.

[9] https://arxiv.org/pdf/1802.09979.pdf

Distribution of activations $\phi'(z)$ (Pennington et al. 18'[10])

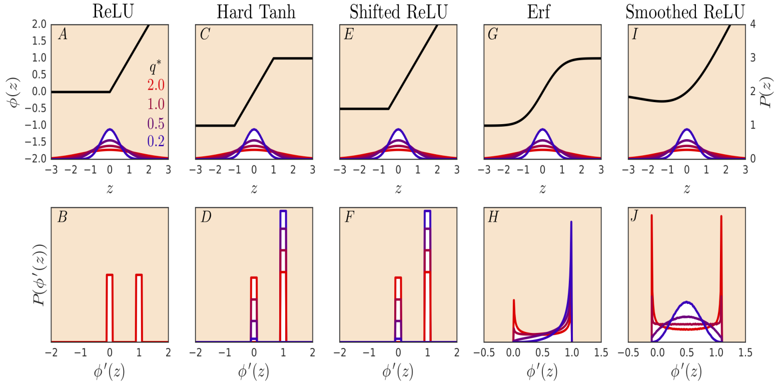


Figure 3: Distribution of $\phi'(h)$ for different nonlinearities. The top row shows the nonlinearity, $\phi(h)$, along with the Gaussian distribution of pre-activations h for four different choices of the variance, q^*. The bottom row gives the induced distribution of $\phi'(h)$. We see that for ReLU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

[10] `https://arxiv.org/pdf/1802.09979.pdf`

Distribution of Jacobian spectra (Pennington et al. 18'[11])

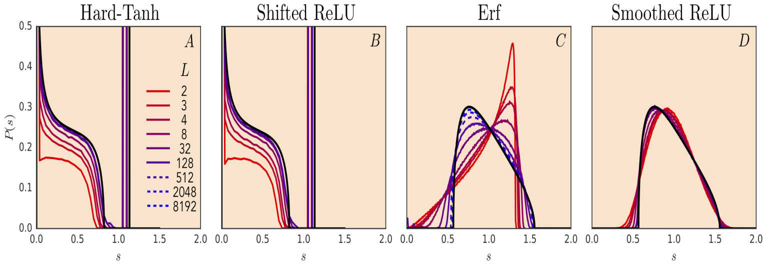


Figure 4: Two limiting universality classes of Jacobian spectra. Hard Tanh and Shifted ReLU fall into one class, characterized by Bernoulli-distributed $\phi'(h)^2$, while Erf and Smoothed ReLU fall into a second class, characterized by a smooth distribution for $\phi'(h)^2$. The black curves are theoretical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

[11] https://arxiv.org/pdf/1802.09979.pdf