Mathematical
Institute

Controlling the variance of the Jacobian's spectrum and Rank-Collapse of Attention layers

Theories of Deep Learning: C6.5,
Lecture / Video 6

Prof. Jared Tanner
Mathematical Institute
University of Oxford

Oxford
Mathematics

UNIVERSITY OF OXFORD
Mathematical Institute

The DNN with weight matrices $W^{(\ell)}$ and bias vectors $b^{(\ell)}$ with Gaussian entries $\mathcal{N}(0, \sigma_w^2)$ and $\mathcal{N}(0, \sigma_b^2)$

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

has computable map $R(\rho)$ of how the correlation between two inputs evolve through the layers. The stability of a point and its perturbation is determined by
$\chi := \frac{\partial R(\rho)}{\partial \rho}|_{\rho=1} = \sigma_w^2 \int Dz[\phi'(\sqrt{q^*}z)^2]$.

- $\chi \leq 1$: locally stable and points which are sufficiently correlated all converge, with depth, to the same point.
- $\chi > 1$: small perturbations are unstable with nearby points become uncorrelated with depth.

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

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

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

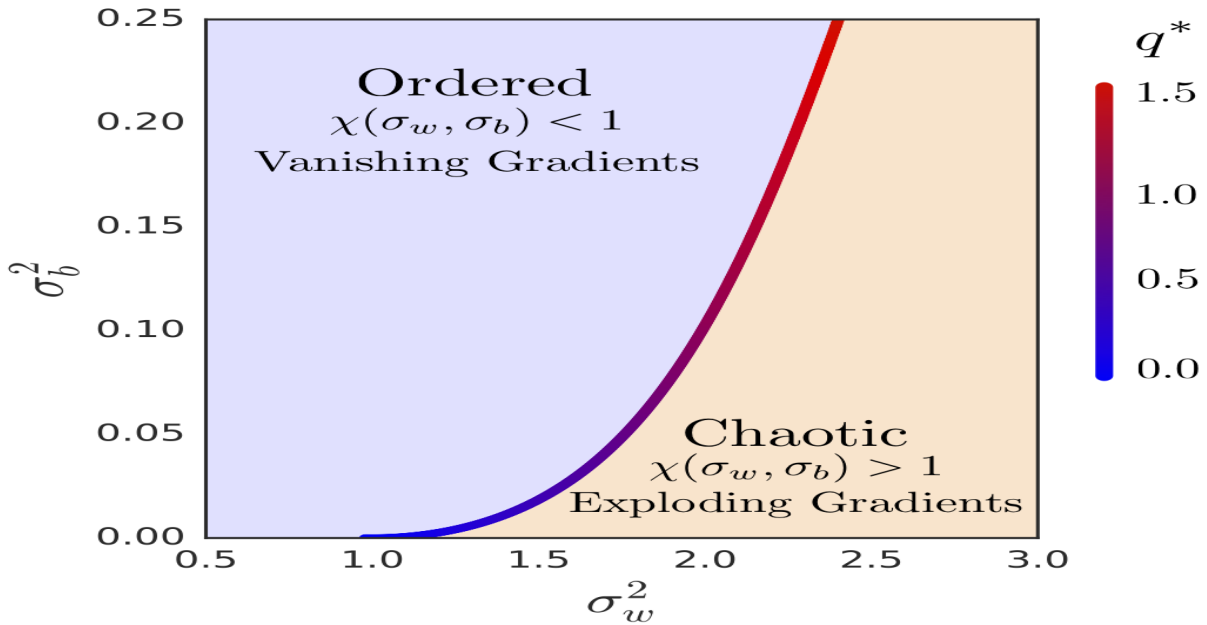


Figure 1: Order-chaos transition when $\phi(h) = \tanh(h)$. The critical line $\chi = 1$ determines the boundary between the two phases. In the chaotic regime $\chi > 1$ and gradients explode while in the ordered regime $\chi < 1$ and we expect gradients to vanish. The value of q^* along this line is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

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

The Jacobian of the feed forward net is given by

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

where $D^{(\ell)}$ is diagonal with entries $D_{ii}^{(\ell)} = \phi'(h_i^{(\ell)})$.
Moreover, for the sum of squares loss, the gradient is computed as

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

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

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

UNIVERSITY OF OXFORD
Mathematical Institute

In the infinite width limit, the average trace of $(DW)^T(DW)$ is the average of the singular values

$$\chi = N^{-1}\left\langle \mathrm{Tr}((\mathrm{DW})^\top \mathrm{DW})\right\rangle$$

The growth of a perturbation is given by the expected mean singular value of $J^T J$ from one layer to the next which is given by

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

Consider the spectrum of $J^T J$ more fully, in particular how it varies around its expected value.

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

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

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

Mathematical Institute

Computing the spectrum of products of matrices, e.g. for $J = \frac{\partial z^{(L)}}{\partial x^{(0)}} = \Pi_{\ell=0}^{L-1} D^{(\ell)} W^{(\ell)}$ 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 $J = \frac{\partial z^{(L)}}{\partial x^{(0)}} = \Pi_{\ell=0}^{L-1} D^{(\ell)} W^{(\ell)}$ 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 $M_{JJ^T}(z) = \sum_{k=1}^{\infty} \frac{m_k}{z^k}$, $M_{D^2}(z) = \sum_{k=1}^{\infty} \frac{\mu_k}{z^k}$, 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
$m_2 = (\sigma_w^2 \mu_1)^{2L} L(\mu_2^{-1} \mu_1^2 + L^{-1} - 1 - s_1)$.
Importantly, $\sigma_w^2 \mu_1 = \chi$ is the growth factor we observed with the edge of chaos, requiring $\chi = 1$ to avoid rapid convergence of correlations to fixed points.
`https://arxiv.org/pdf/1802.09979.pdf`

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	$\sigma^2_{JJ^T}$
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 $\mu_k = \int (2\pi)^{-1/2} \phi'\left(\sqrt{q^{(\star)}} z\right)^{2k} e^{-z^2/2} dz$.

Recall that $m_1 = \chi^L$ is the expected value of the spectrum of JJ^T; while the variance of the spectrum of JJ^T is given by $\sigma^2_{JJ^T} = m_2 - m_1^2 = L(\mu_2 \mu_1^{-2} - 1 - s_1)$, where for W Gaussian $s_1 = -1$ and for W orthogonal $s_1 = 0$.

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

ReLU $\phi(\cdot)$: $q^* = \frac{1}{2}\sigma_w^2 q^* + \sigma_b^2$, has 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

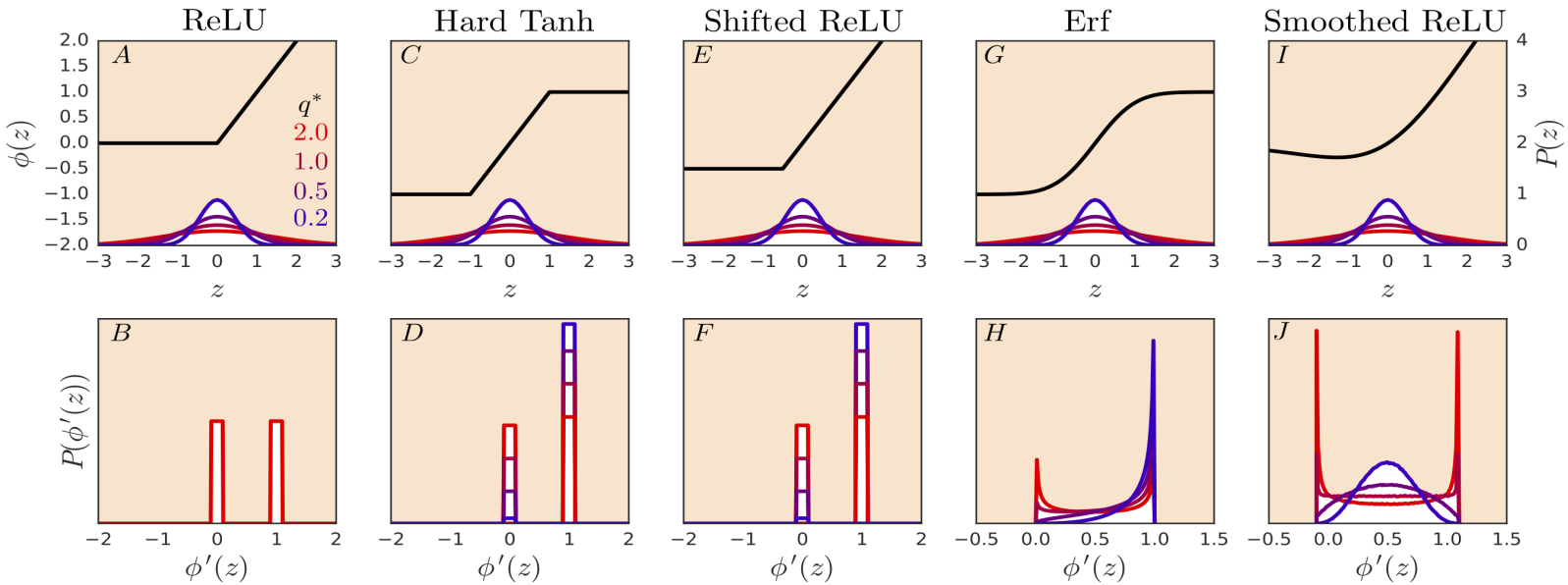


Figure 3: Distribution of $\phi'(h)$ for different nonlinearities. The top row shows the nonlinearity, $\phi(h)$, along with the Gaussian distribution of pre-activations h for four different choices of the variance, q^*. The bottom row gives the induced distribution of $\phi'(h)$. We see that for ReLU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

Definition (scaled-bounded activations)

We refer to the set of activation functions $\phi : \mathbb{R} \to \mathbb{R}$ which satisfy the following properties as scaled-bounded activations.

1. Continuous.
2. Odd, meaning that $\phi(z) = -\phi(-z)$ for all $z \in \mathbb{R}$.
3. Linear around the origin and bounded: in particular there exists $a, k \in \mathbb{R}_{>0}$ such that $\phi(z) = kz$ for all $z \in [-a, a]$ and $\phi(z) \leq ak$ for all $z \in \mathbb{R}$.
4. Twice differentiable at all points $z \in \mathbb{R} \backslash \mathcal{D}$, where $\mathcal{D} \subset \mathbb{R}$ is a finite set. Furthermore $|\phi'(z)| \leq k$ for all $z \in \mathbb{R} \backslash \mathcal{D}$.

`https://arxiv.org/abs/2105.07741`

Correlation map and variance convergence (Murray et al. 21')

Increasing the linear region drives the Jacobian spectra to 1

UNIVERSITY OF OXFORD
Mathematical Institute

Theorem (Murray 21')

Let ϕ be a scaled-bounded activation, $\sigma_b^2 > 0$, $\chi_1 := \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q^*}Z)^2] = 1$ where $q^* > 0$ is a fixed point of V_ϕ. Let inputs x satisfy $||\mathrm{x}||_2^2 = q^*$.
Then as $y := \sigma_b^2/a^2 \to 0$, both

$$\max_{\rho\in[0,1]} |R_{\phi,q^*}(\rho) - \rho|, \left|\mu_2/\mu_1^2 - 1\right| \to 0,$$

with rates available in Murray 21'.

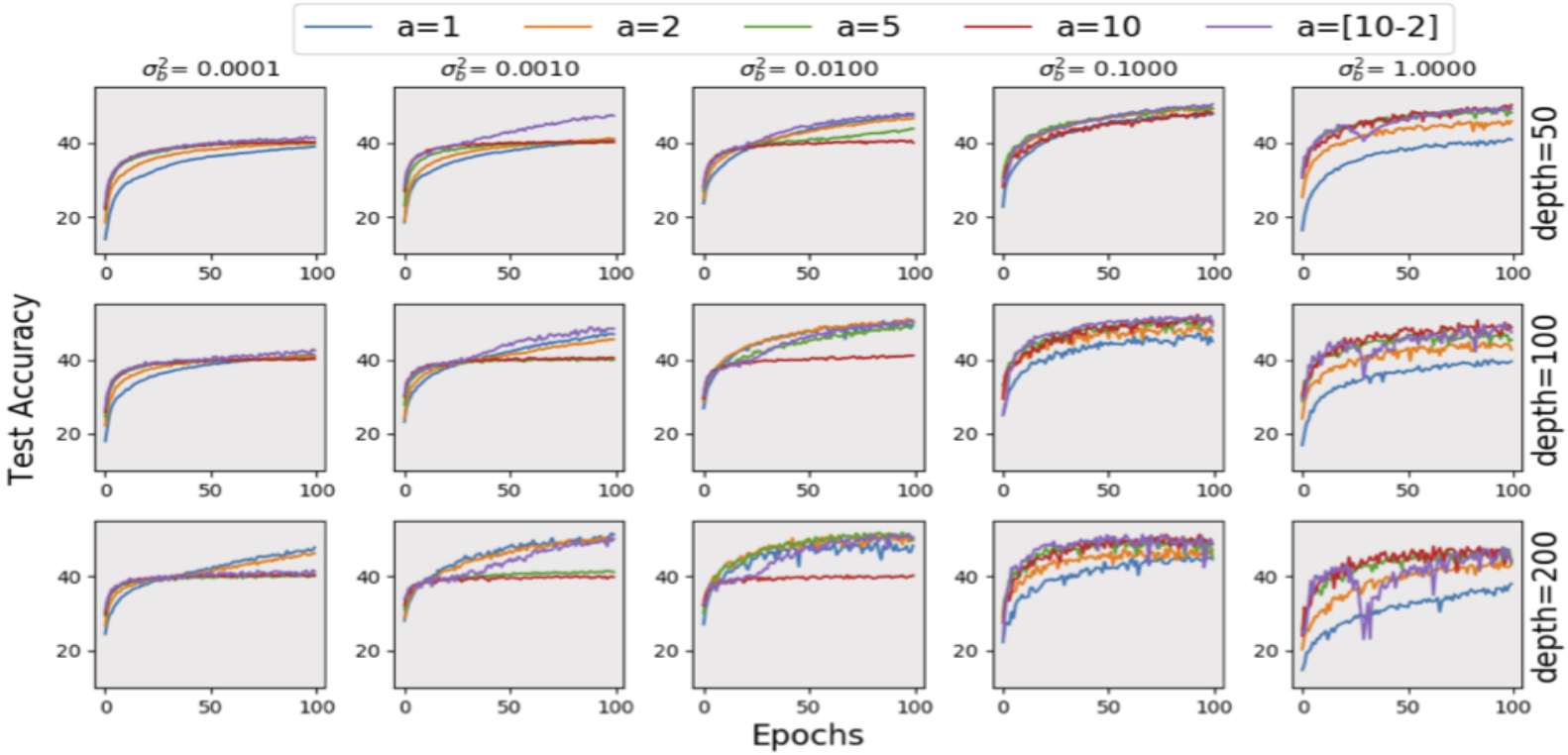
Note that this is independent of details of $\phi(\cdot)$ outside its linear region $[-a, a]$. Best performance is observed with $a \sim 3$, or preferably a decreasing from about 5 to 2 during training.
`https://arxiv.org/abs/2105.07741`

Training very DNNs with Shtanh (Murray et al. 21')

Improved accuracy with dynamic linearity decay

UNIVERSITY OF OXFORD
Mathematical Institute

Test accuracy of a trained very deep feed forward net on CIFAR-10.



(b) Shtanh with orthogonal initialisation

https://arxiv.org/abs/2105.07741

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

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

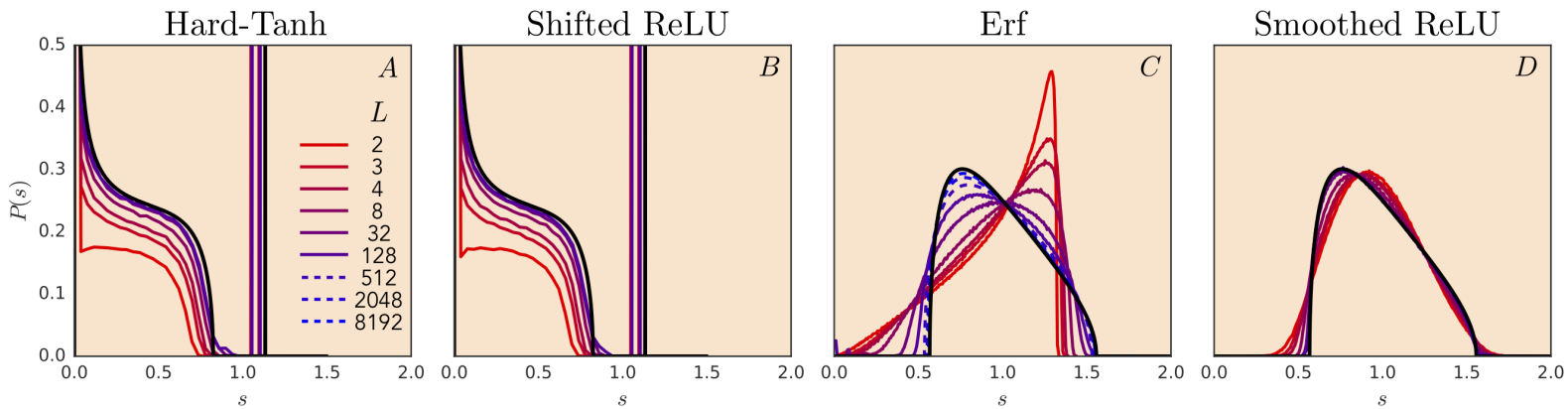


Figure 4: Two limiting universality classes of Jacobian spectra. Hard Tanh and Shifted ReLU fall into one class, characterized by Bernoulli-distributed $\phi'(h)^2$, while Erf and Smoothed ReLU fall into a second class, characterized by a smooth distribution for $\phi'(h)^2$. The black curves are theoretical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

From fully connected to CNN EoC (Pennington et al. 18')

Hidden layer covariance matrix becomes block diagonal for overlapping filters

UNIVERSITY OF OXFORD
Mathematical Institute

- In fully connected (MLP) networks the entries in a hidden layer are mean zero and have a diagonal, due to row independence in $W^{(\ell)}$, covariance matrix with diagonal entries $q^{(\ell)}$.
- In a CNN the hidden layers are again mean zero and have variance $q^{(\ell)}$. The covariance matrix is zero when considering different filters which are drawn randomly. Moreover, when a filter acts on disjoint locations the randomness from the prior layer induces a zero covariance. However, for output corresponding to a filter overlapping spatially the limit is $q^*\rho^*$ where $\rho^* = R(\rho^*)$ from the correlation map. The covariance matrix is then typically block diagonal with nonzero entries converging to q^* in depth.

`https://arxiv.org/pdf/1806.05393`

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`

Further associated reading 1 of 2

Related results

- Identifying natural depth scales of information propagation
  `https://arxiv.org/pdf/1611.01232.pdf`
- Further details on the role of activation functions
  `https://arxiv.org/pdf/1902.06853.pdf`
- Principles for selecting activation functions
  `https://arxiv.org/pdf/2105.07741.pdf`

Further associated reading 2 of 2

Convergence of representations at each layer of a neural network to a Gaussian Process & wider reading

- Early results on correlation of inputs (Chapter 2 in particular) `https://www.cs.toronto.edu/~radford/ftp/thesis.pdf`
- Rigorous treatment of Gaussian Process perspective, infinite width `https://arxiv.org/pdf/1711.00165.pdf`
- Rigorous treatment of Gaussian Process perspective, finite width `https://arxiv.org/pdf/1804.11271.pdf`
- Higher order terms and width proportional to depth scaling `https://arxiv.org/pdf/2106.10165.pdf`
- Specifics for random ReLU nets `https://arxiv.org/pdf/1801.03744.pdf` `https://arxiv.org/pdf/1803.01719.pdf`

Attention mechanisms: Rank-collapse and alternative nonlinear activations

Attention mechanism (Vaswani 17'), equations

Key and Query quadratic form to highlight relations

Mathematical Institute

Input is a matrix $X \in \mathbb{R}^{n \times d}$ where n is the "context length" and d is the "embedding dimension". They queries, keys, and values are then computed with matrix-products $Q^T = W_Q X^T$, $K^T = W_K X^T$, and $V^T = W_V X^T$ then the attention layer is

$$H = \text{softmax}\left(\frac{X W_Q W_K^T X^T}{n^{1/2}}\right) X W_V$$

where the softmax acts row-wise to give non-negative entries that sum to one.

$$\text{softmax}(x)_i = exp(x_i) / \sum_\ell exp(x_\ell)$$

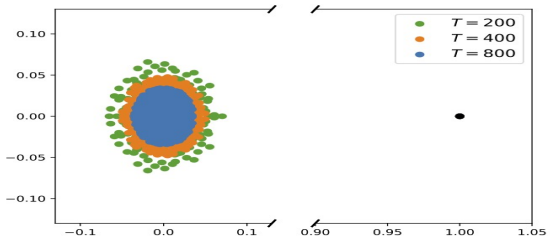
Generally Q and K have layer-norm applied to enforce fixed mean and variance. Intuitively the softmax helps highlight the rows in X that deserve "attention."

- The query and key matrices, $Q^T = W_Q X^T$ and $K^T = W_K X^T$, are used to measure alignment between the input X in their embedding space determined by W_Q and W_K. To measure angles one always layer-norm so that each row of Q and K have ℓ^2 length of 1. Their initialization largely ignored for this reason.
- The value matrix is very similar to fully connected networks, but note there isn't a nonlinear activation here.
- What is the impact of the softmax activation?

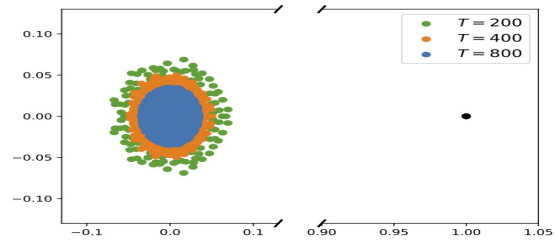
$$\text{softmax}\left(\frac{XW_Q W_K^T X^T}{n^{1/2}}\right) XW_v$$

Spectrum of the attention matrix (Nait Saada 24')

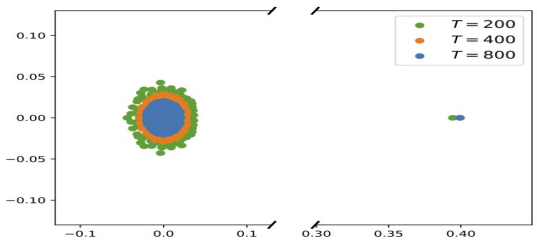
The spectral gap of random Markov matrices



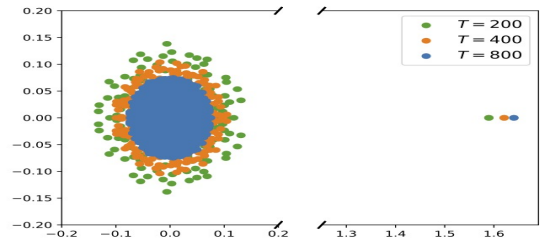
(c) Softmax attention



(d) i.i.d. Markov



(e) ReLU attention



(f) Sigmoid attention

Random matrices with a non-zero mean have an associated outlying eigenvalue. For Markov / softmax there is one eigenvalue at 1 and the remainder uniform in a disc with radius $\sim n^{-1/2}$. This causes rank-collapse with outputs increasingly one-dimensional.

`https://arxiv.org/pdf/2410.07799`

Rank-collapse due to the spectral gap (Nait Saada 24')

Excessive over smoothing in width and depth

- The single eigenvalue outlier causes an L layer need attention network to converge to a one dimensional output at the exponential rate $n^{-L/2}$.
- This is a form of excessive stability and slows initial training. It is reflected in a form of vanishing gradients analogous to $\chi = 1$
- Similarly, (Dong et al. 23) showed rank-collapse through depth with attention converging to the constant matrix $T^{-1}1_T 1_T^T$. (`https://arxiv.org/pdf/2103.03404`)

The resolution of rank-collapse is an active area of research `https://arxiv.org/pdf/2410.07799`

UNIVERSITY OF OXFORD
Mathematical Institute

- Centering (Ali et al. 23') subtracts the matrix of constant values T^{-1}, i.e. $T^{-1}1_T 1_T^T$, corresponding to the eigenvalue outlier `https://arxiv.org/pdf/2306.01610`
- Differential transformers (Ye et al. 24') compute two attention matrices with different key and query matrices and then subtract them. This has the same effect as centering, but doubles the number of parameters. `https://arxiv.org/pdf/2410.05258`
- Mind the gap (Nait Saada et al. 24') project out the direction 1_T from the columns of the value matrix V, mathematically equivalent to centering but more efficient. `https://arxiv.org/pdf/2410.07799`

Rather than remove the rank-collapse effect the softmax activation, Hron et al. 20', amongst others, instead consider not using softmax. They showed superior training accuracy by instead using linear attention where the softmax is omitted, referred to as linear attention, or ReLU entrywise instead of the softmax. `https://arxiv.org/pdf/2006.10540`

Similar linear attention is now widely discussed in the literature of Vision Transformers, see for instance `https://arxiv.org/pdf/2309.08586` by Wortsman et al. 23'.

Efficient training of LLMs is one of the dominant research questions being actively studied.