Controlling the exponential growth of variance and correlation

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

Prof. Jared Tanner
Mathematical Institute
University of Oxford

Oxford Mathematics

University of Oxford
Mathematical Institute

Xavier Glorot and Yoshua Bengio (2010) considered random initialized networks and the associated variance of pre-activation values and the gradients as they pass from layer to layer.

"Our objective here is to understand why standard gradient descent from random initialization is doing so poorly with deep neural networks.... we study how activations and gradients vary across layers and during training, with the idea that training may be more difficult when the singular values of the Jacobian associated with each layer are far from 1."

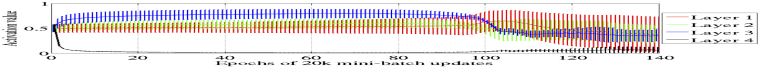


Figure 2: Mean and standard deviation (vertical bars) of the activation values (output of the sigmoid) during supervised learning, for the different hidden layers of a deep architecture. The top hidden layer quickly saturates at 0 (slowing down all learning), but then slowly desaturates around epoch 100.

`http://proceedings.mlr.press/v9/glorot10a.html`

Xavier weight initialization (Glorot et al.' 10)

Variance normalization; precursor to Pennington with $\sigma_b = 0$

Glorot and Bengio noted that (for symmetric $\phi(\cdot)$ such as $tanh(\cdot)$) the hidden layer values $h_{i+1} = \phi_i\left(W^{(i)}h_i + b^{(i)}\right)$ are approximately Gaussian with a variance depending on the variance of the weight matrices $W^{(i)}$.

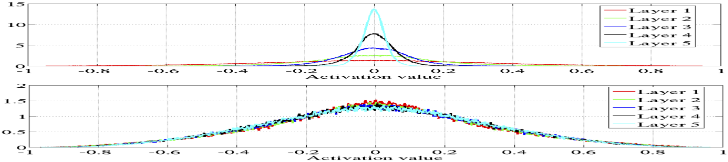


Figure 6: Activation values normalized histograms with hyperbolic tangent activation, with standard (top) vs normalized initialization (bottom). Top: 0-peak increases for higher layers.

In particular, if σ_w is selected appropriately ($\sigma_w^2 = 1/3n$) then the variance of h_i is approximately constant through layers (see the "bottom" plot), and if σ_w is to small then the variance of h_i converges towards zero with depth (see the "top" plot).

`http://proceedings.mlr.press/v9/glorot10a.html`

Xavier weight initialization (Glorot et al.' 10)

Variance normalization; precursor to Pennington with $\sigma_b = 0$

Similarly, the gradient of a loss function used for training showed similar Gaussian behaviour depending on σ_w.

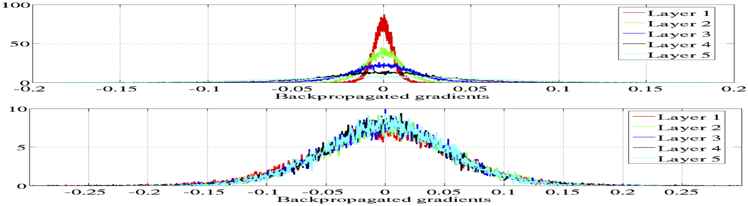


Figure 7: Back-propagated gradients normalized histograms with hyperbolic tangent activation, with standard (top) vs normalized (bottom) initialization. Top: 0-peak decreases for higher layers.

The suggested "Xavier" initialization $\sigma_w^2 = 1/3n$ follows from balancing the variance of h_i and the gradient to be constant through depth.

`http://proceedings.mlr.press/v9/glorot10a.html`

Xavier initialization (Glorot et al.' 10)

Impact on training rate

Proper initialization is essential for training:

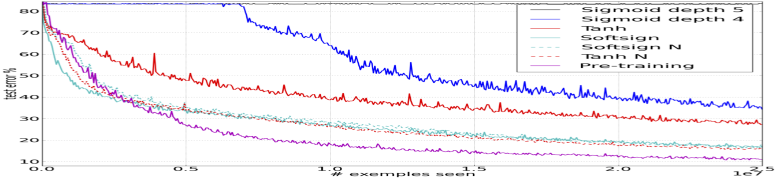


Figure 11: Test error during online training on the **Shapeset-3** $\times$ *2 dataset, for various activation functions and initialization schemes (ordered from top to bottom in decreasing final error). N after the activation function name indicates the use of normalized initialization.*

Five layer sigmoid fails to train while 4 layer sigmoid trains after substantial stagnation. Tanh Normalized trains substantially faster than Tanh without this initialization.

`http://proceedings.mlr.press/v9/glorot10a.html`

Let $f_{NN}(x)$ denote a random Gaussian DNN

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

which takes as input the vector x, and is parameterised by random weight matrices $W^{(\ell)}$ and bias vectors $b^{(\ell)}$ with entries sampled iid from the Gaussian normal distributions $\mathcal{N}(0, \sigma_w^2)$ and $\mathcal{N}(0, \sigma_b^2)$. Define the ℓ^2 length of the pre-activation hidden layer $h^{(\ell)} \in \mathbb{R}^{n_\ell}$ output as:

$$q^{\ell} = n_\ell^{-1} \left\| h^{(\ell)} \right\|_{\ell^2}^2 := \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \left(h^{(\ell)}(i) \right)^2 .$$

which is the sample mean of the random entries $\left(h^{(\ell)}(i)\right)^2$, each of which are identically distributed

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

Norm of hidden layer output dependence on prior layer

The norm of the hidden layer output $q^\ell = n_\ell^{-1} \left\| h^{(\ell)} \right\|_{\ell^2}^2$ has an expected value over the random draws of $W^{(\ell)}$ and $b^{(\ell)}$ which satisfies

$$\mathcal{E}(q^\ell) = \mathcal{E}\left(\left(h^{(\ell)}(i)\right)^2\right)$$

which as $h^{(\ell)} = W^{(\ell)}\phi(h^{(\ell-1)}) + b^{(\ell)}$ is

$$\begin{aligned}\mathcal{E}(q^\ell) &= \mathcal{E}\left(\left(W_i^{(\ell)}\phi\left(h^{(\ell-1)}\right)\right)^2\right) + \mathcal{E}\left(\left(b_i^{(\ell)}\right)^2\right) \\ &= \sigma_w^2 n_{\ell-1}^{-1} \sum_{i=1}^{n_{\ell-1}} \phi\left(h_i^{(\ell-1)}\right)^2 + \sigma_b^2\end{aligned}$$

where $W_i^{(\ell)}$ denotes the i^{th} row of $W^{(\ell)}$.
`https://arxiv.org/pdf/1606.05340.pdf`
For a more rigorous proof see the solutions to Assignment Sheet 2.

Approximating $h_i^{(\ell-1)}$ as Gaussian with variance $q^{(\ell-1)}$:

$$n_{\ell-1}^{-1}\sum_{i=1}^{n_{\ell-1}}\phi\left(h_i^{(\ell-1)}\right)^2=\mathcal{E}\left(\phi\left(h^{(\ell-1)}\right)^2\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\phi\left(\sqrt{q^{(\ell-1)}}z\right)^2 e^{-z^2/2}dz$$

which gives a recursive map of $q^{\ell}=n_{\ell}^{-1}\left\|h^{(\ell)}\right\|_{\ell^2}^2$ between layers

$$q^{(\ell)}=\sigma_b^2+\sigma_w^2\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\phi\left(\sqrt{q^{(\ell-1)}}z\right)^2 e^{-z^2/2}dz=:\mathcal{V}(q^{(\ell-1)}|\sigma_w,\sigma_b,\phi(\cdot)).$$

Note that the integral is larger for $\phi(x)=|x|$ than ReLU, which are larger than $\phi(x)=tanh(x)$, indicating smaller σ_w,σ_b needed to ensure $q^{(\ell)}$ has a finite nonzero limit q^*.

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

Example of DNN recursion fixed points (Poole et al. 16')

Dependence on σ_w, σ_b, $\phi(\cdot)$ for $\phi(\cdot) = tanh(\cdot)$.

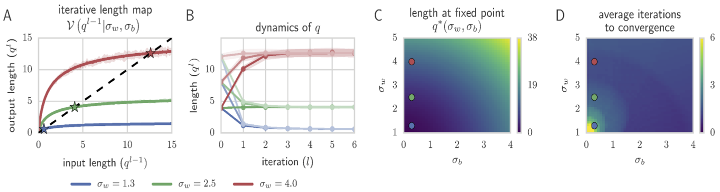


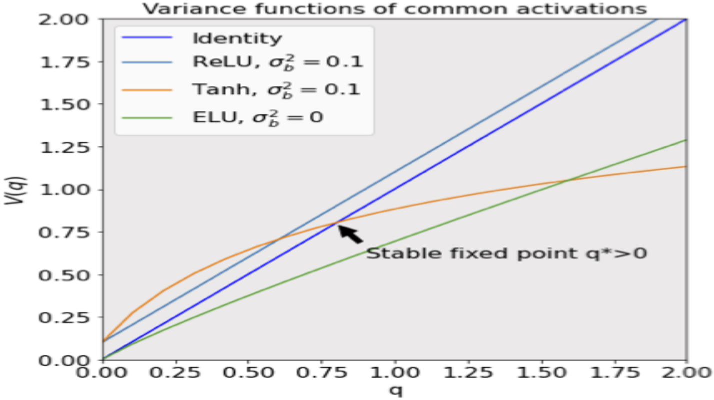
Figure 1: Dynamics of the squared length q^l for a sigmoidal network ($\phi(h) = \tanh(h)$) with 1000 hidden units. (A) The iterative length map in (3) for 3 different σ_w at $\sigma_b = 0.3$. Theoretical predictions (solid lines) match well with individual network simulations (dots). Stars reflect fixed points q^* of the map. (B) The iterative dynamics of the length map yields rapid convergence of q^l to its fixed point q^*, independent of initial condition (lines=theory; dots=simulation). (C) q^* as a function of σ_w and σ_b. (D) Number of iterations required to achieve $\leq 1\%$ fractional deviation off the fixed point. The (σ_b, σ_w) pairs in (A,B) are marked with color matched circles in (C,D).

Note that the fixed points here are all stable.

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

DNN recursion fixed points (Murray et al. 21')

Examples of different fixed points or divergence



ReLU $q^{(l)}$ is unbounded, ELU $q^{(l)} \to 0$, sigmoidal $q^{(l)} \to q^* > 0$.

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

A single input x has hidden pre-activation output converging to a fixed expected length.
Consider the map governing the angle between the hidden layer pre-activations of two distinct inputs $x^{(0,a)}$ and $x^{(0,b)}$:

$$q_{ab}^{(\ell)} = n_\ell^{-1} \sum_{i=1}^{n_l} h_i^{(\ell)}(x^{(0,a)}) h_i^{(\ell)}(x^{(0,b)}).$$

Similar to the analysis before, use the relation
$h^{(\ell)} = W^{(\ell)} \phi(h^{(\ell-1)}) + b^{(\ell)}$ to show the relation between layers.
`https://arxiv.org/pdf/1606.05340.pdf`

Replacing the average in the sum with the expected value gives

$$q_{ab}^{(\ell)} = n_\ell^{-1} \sum_{i=1}^{n_l} h_i^{(\ell)}(x^{(0,a)}) h_i^{(\ell)}(x^{(0,b)})$$

$$= \sigma_b^2 + \sigma_w^2 \mathcal{E} \left(\phi(h^{(\ell-1)}(x^{(0,a)})) \phi(h^{(\ell-1)}(x^{(0,b)})) \right)$$

where as before, $h_i^{(\ell-1)}$ are well modelled as being Gaussian with expected length $q^{(\ell-1)}$ which converge to fixed points q^*.
`https://arxiv.org/pdf/1606.05340.pdf`

Defining the angle between the hidden layers as $c^{(\ell)} = q_{12}^{(\ell)}/q^*$ and writing the expectation as integrals we have

$$c^{(\ell)} = \mathcal{C}\left(c^{(\ell-1)}|\sigma_w, \sigma_b, \phi(\cdot)\right) := \sigma_b^2 + \sigma_w^2 \int Dz_1 Dz_2 \phi(u_1)\phi(u_2)$$

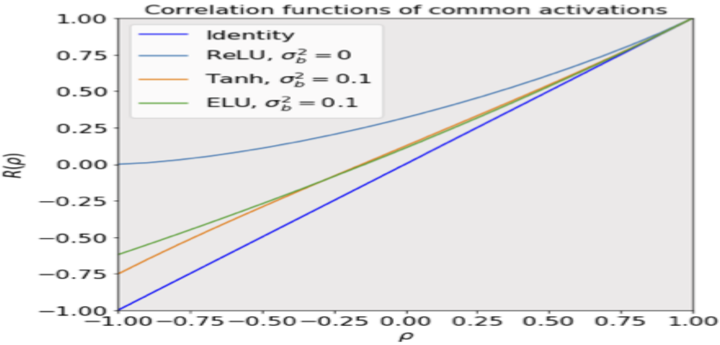
where the double integral is with respect to the measure $Dz = (2\pi)^{-1/2}e^{-z^2/2}dz$ where $u_1 = \sqrt{q^*}z_1$ and $u_2 = \sqrt{q^*}[c^{(\ell-1)}z_1 + \sqrt{1-(c^{(\ell-1)})^2}z_2]$ are a change of variables for the integrals. Normalizing by $q^{(*)}$ we have the correlation map

$$\rho^{(l+1)} = c^{(l+1)} := R(\rho^{(l)}; \sigma_w, \sigma_b, \phi(\cdot)).$$

By construction $R(1) = 1$ as $c^{(l)} \to q^*$ as the two inputs converge to one another. `https://arxiv.org/pdf/1606.05340.pdf`

DNN correlation fixed points (Murray et al. 21')

Examples of correlation functions for different $\phi(\cdot)$



ReLU requires $\sigma_b = 0$, ELU and sigmoidal have σ_w selected so that $R'(1) = 1$. `https://arxiv.org/abs/2105.07741`

By definition $R(1) = 1$. For $\sigma_b^2 > 0$ the correlation map satisfies $R(0) > 0$, orthogonal $x^{(0,a)}$ and $x^{(0,b)}$ become increasingly correlated with depth.

Of particular note is the slope of $R(\cdot)$ at $\rho = 1$

$$\chi := \frac{\partial R(\rho)}{\partial \rho}|_{\rho=1} = \frac{\partial \rho^{(\ell)}}{\partial \rho^{(\ell-1)}}|_{\rho=1} = \sigma_w^2 \int Dz[\phi'(\sqrt{q^*}z)^2].$$

Stability of the fixed point at $\rho = 1$ is determined by χ:

- $\chi < 1$: $\rho = 1$is locally stable and points which are sufficiently correlated all converge, with depth, to the same point.
- $\chi > 1$: $\rho = 1$ is unstable and nearby points become uncorrelated with depth.
- Preferable to choose $\chi = 1$ if possible for $(\sigma_w, \sigma_b, \phi(\cdot))$.

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

Example of DNN correlation fixed points (Poole et al. 16')

Dependence on σ_w, σ_b, $\phi(\cdot)$ for $\phi(\cdot) = tanh(\cdot)$.

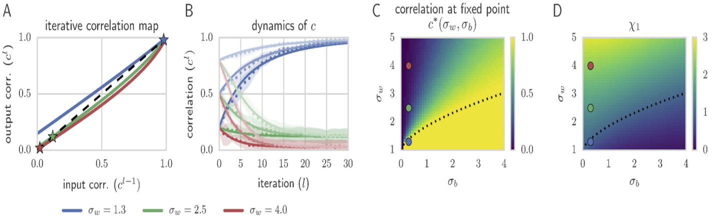


Figure 2: Dynamics of correlations, c_{12}^l, in a sigmoidal network with $\phi(h) = \tanh(h)$. (A) The $\mathcal{C}$-map in (6) for the same σ_w and $\sigma_b = 0.3$ as in Fig. 1A. (B) The $\mathcal{C}$-map dynamics, derived from both theory, through (6) (solid lines) and numerical simulations of (1) with $N_l = 1000$ (dots) (C) Fixed points c^* of the $\mathcal{C}$-map. (D) The slope of the $\mathcal{C}$-map at 1, χ_1, partitions the space (black dotted line at $\chi_1 = 1$) into chaotic ($\chi_1 > 1, c^* < 1$) and ordered ($\chi_1 < 1, c^* = 1$) regions.

Note three respective stable fixed points, determined in part by χ.

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

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 activation $\phi(\cdot)$ and $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$.

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

DNN Jacobian

Gradient back-propagation: role of DNN Jacobian

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

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

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

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

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

UNIVERSITY OF OXFORD
Mathematical Institute

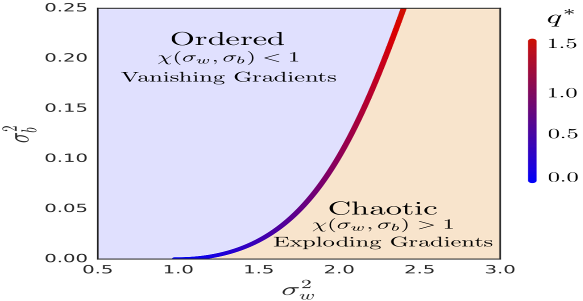


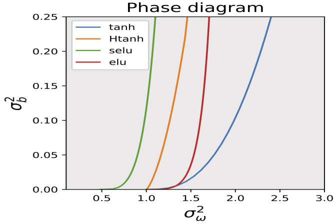
Figure 1: Order-chaos transition when $\phi(h) = \tanh(h)$. The critical line $\chi = 1$ determines the boundary between the two phases. In the chaotic regime $\chi > 1$ and gradients explode while in the ordered regime $\chi < 1$ and we expect gradients to vanish. The value of q^* along this line is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

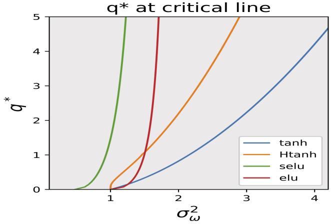
https://arxiv.org/pdf/1802.09979.pdf

Network variance control through depth

Edge of chaos curves for other nonlinear activations (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.

DNN random initialisation summary, so far

Dependence on $(\sigma_w, \sigma_b, \phi(\cdot))$.

- The hidden layers converge to fixed expected length.
- All inputs converge to either one another or prescribed correlation, independent of the class the data is in, typically happening at a rate which is exponential with depth.
- The rate with which these phenomenon occur, and values which they take, are determined by the choice of $(\sigma_w, \sigma_b, \phi(\cdot))$.
- Very DNNs can be especially hard to train for activations with unfavourable initialisations; e.g. ReLU with $\chi = 1$ requires $(\sigma_w, \sigma_b) = (\sqrt{2}, 0)$.
- The gradient of the network also have exponential depth dependence proportional to χ^L through the expected singular value of the Jacobian, making $\chi = 1$ essential for training.

- 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`