On the Impact of the Activation Function on Deep Neural Networks Training

Soufiane Hayou

University of Oxford

soufiane.hayou@stats.ox.ac.uk

Overview

1. Neural Networks as Gaussian Processes
   - Limit of large networks
2. Information Propagation
   - Depth Scales
   - Edge of Chaos
   - Impact of smoothness
3. Experiments

Random Neural Networks

Consider a fully connected feed-forward neural network of depth L, widths $(N_l)_{1\leq l\leq L}$, weights $W_{ij}^l \overset{iid}{\sim} \mathcal{N}(0, \frac{\sigma_w^2}{N_{l-1}})$ and bias $B_i^l \overset{iid}{\sim} \mathcal{N}(0, \sigma_b^2)$
For some input $a \in \mathbb{R}^d$, the propagation of this input through the network is given by

$$y_i^1(a) = \sum_{j=1}^{d} W_{ij}^1 a_j + B_i^1$$

$$y_i^l(a) = \sum_{j=1}^{N_{l-1}} W_{ij}^l \phi(y_j^{l-1}(a)) + B_i^l, \quad \text{for } l \geq 2.$$

Limit of infinite width

1. When N_{l-1} is large, $y_i^l(a)$ are iid centred Gaussian variables. By induction, this is true for all l.

Limit of infinite width

1. When N_{l-1} is large, $y_i^l(a)$ are iid centred Gaussian variables. By induction, this is true for all l.

2. Stronger result : when $N_l = +\infty$ for all l (recursively), $y_i^l(.)$ are independent (across i) centred Gaussian processes.
(first proposed by Neal [1995] in the single layer case and has been recently extended to the multiple layer case by Lee et al. [2018] and Matthews et al. [2018])

Information Propagation

For two inputs a, b, let $q^l(a)$ be the variance of $y_1^l(a)$ and c_{ab}^l the correlation of $y_1^l(a)$ and $y_1^l(b)$.

1. Variance propagation : $q^l = F(q^{l-1})$
   where $F(x) = \sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(\sqrt{x}Z)^2]\ , \quad Z \sim \mathcal{N}(0,1))$

2. Correlation propagation : $c^{l+1} = f_l(c^l)$
   where $f_l(x) = \frac{\sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(\sqrt{q_a^l}Z_1)\phi(\sqrt{q_b^l}(xZ_1 + \sqrt{1-x^2}Z_2))}{\sqrt{q_a^l}\sqrt{q_b^l}}$

Depth scales

Schoenholz et al. [2017] established the existence of $c \in [0, 1]$ such that $|c^l_{ab} - c| \sim e^{-l/\epsilon_c}$ where $\epsilon_c = -\log(\chi)$ and $\chi = \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q}Z)]$ (q is the limiting variance). The equation $\chi_1 = 1$ corresponds to an infinite depth scale of the correlation. It is called the edge of chaos as it separates two phases :

Depth scales

Schoenholz et al. [2017] established the existence of $c \in [0, 1]$ such that $|c^l_{ab} - c| \sim e^{-l/\epsilon_c}$ where $\epsilon_c = -\log(\chi)$ and $\chi = \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q}Z)]$ (q is the limiting variance). The equation $\chi_1 = 1$ corresponds to an infinite depth scale of the correlation. It is called the edge of chaos as it separates two phases :

- Ordered phase where $\chi_1 < 1$ ($c = 1$): the correlation converges (exponentially) to 1. In this case, two different inputs will have the same output.

Depth scales

Schoenholz et al. [2017] established the existence of $c \in [0, 1]$ such that $|c^l_{ab} - c| \sim e^{-l/\epsilon_c}$ where $\epsilon_c = -\log(\chi)$ and $\chi = \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q}Z)]$ (q is the limiting variance). The equation $\chi_1 = 1$ corresponds to an infinite depth scale of the correlation. It is called the edge of chaos as it separates two phases :

- Ordered phase where $\chi_1 < 1$ ($c = 1$): the correlation converges (exponentially) to 1. In this case, two different inputs will have the same output.
- Chaotic phase where $\chi_1 > 1$ ($c < 1$): the correlation converges (exponentially) to some value $c < 1$. In this case, very close inputs will have very different outputs (the output function is discontinuous everywhere).

Ordered phase

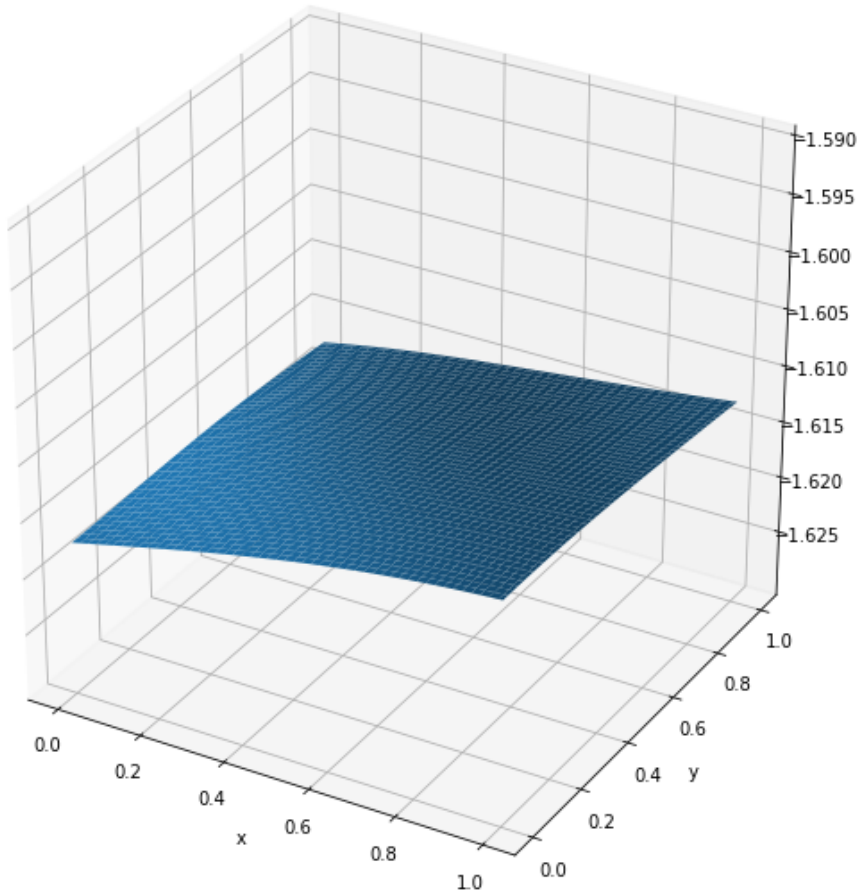


Figure: Output of a 300x20 Tanh network with $(\sigma_b, \sigma_w) = (1, 1)$(Ordered phase)

Chaotic phase

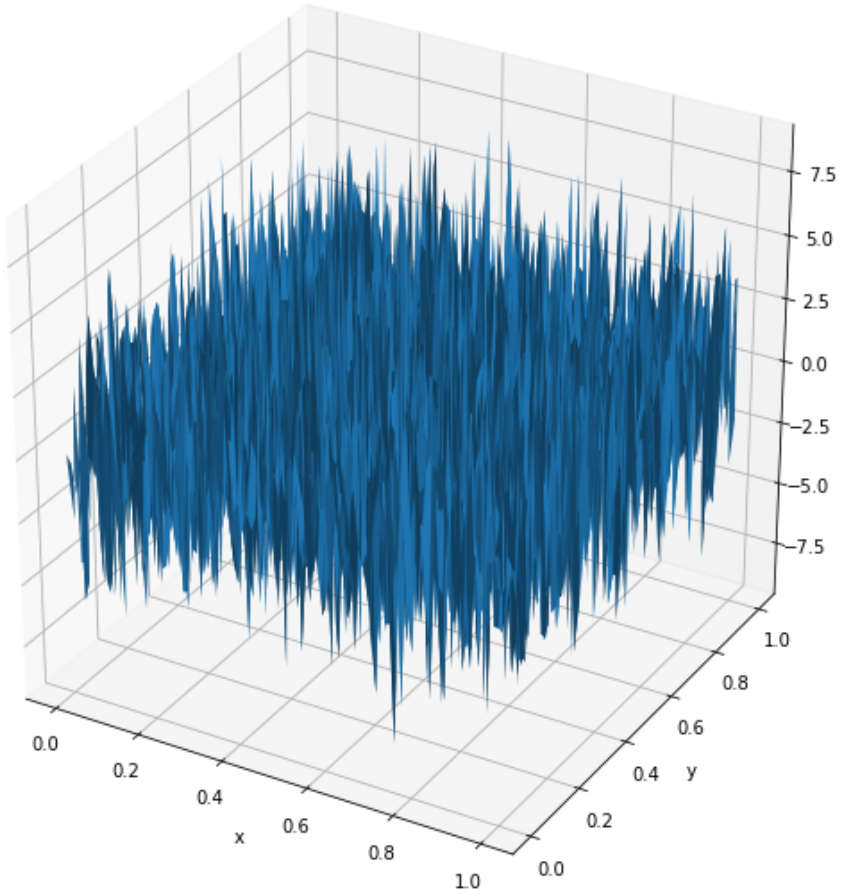


Figure: A draw of the output of a 300x20 Tanh network with $(\sigma_b, \sigma_w) = (0.3, 2)$ (chaotic phase)

Edge of Chaos

Definition

For $(\sigma_b, \sigma_w) \in D_{\phi,var}$, let q be the limiting variance. The Edge of Chaos, hereafter EOC, is the set of values of (σ_b, σ_w) satisfying $\chi_1 = \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q}Z)^2] = 1$.

- Having $\chi_1 = 1$ is linked to an infinite depth scale $\rightarrow$ Sub-exponential convergence rate for the correlation
- For ReLU, the $\mathrm{EOC} = \{(0, \sqrt{2})\}$. This coincides with the recommendation of He et al. [2015]

.

Edge of Chaos for ReLU

Proposition 1 : EOC acts as Residual connections

Consider a ReLU network with parameters $(\sigma_b^2, \sigma_w^2) = (0, 2) \in EOC$ and let c_{ab}^l be the corresponding correlation. Consider also a ReLU network with simple residual connections given by

$$\overline{y}_i^l(a) = \overline{y}_i^{l-1}(a) + \sum_{j=1}^{N_{l-1}} \overline{W}_{ij}^l \phi(\overline{y}_j^{l-1}(a)) + \overline{B}_i^l$$

where $\overline{W}_{ij}^l \overset{iid}{\sim} \mathcal{N}(0, \frac{\overline{\sigma}_w^2}{N_{l-1}})$ and $\overline{B}_i^l \overset{iid}{\sim} \mathcal{N}(0, \overline{\sigma}_b^2)$. Let $\overline{c}_{ab}^l$ be the corresponding correlation. Then, by taking $\overline{\sigma}_w > 0$ and $\overline{\sigma}_b = 0$, there exists a constant $\gamma > 0$ such that

$$1 - c_{ab}^l \sim \gamma(1 - \overline{c}_{ab}^l) \sim \frac{9\pi^2}{2l^2} \quad \text{as} \quad l \to \infty.$$

Impact of Smoothness

Class $\mathcal{A}$

Let $\phi \in \mathcal{D}_g^2$. We say that ϕ is in $\mathcal{A}$ if there exists $n \geq 1$, a partition $(S_i)_{1 \leq i \leq n}$ of $\mathbb{R}$ and $g_1, g_2, ..., g_n \in \mathcal{C}_g^2$ such that $\phi^{(2)} = \sum_{i=1}^{n} 1_{S_i} g_i$.

Proposition 3 : Convergence rate for smooth Activation functions

Let $\phi \in \mathcal{A}$ such that ϕ non-linear (i.e. $\phi^{(2)}$ is non-identically zero). Then, on the EOC, we have $1 - c^l \sim \frac{\beta_q}{l}$ where $\beta_q = \frac{2\mathbb{E}[\phi'(\sqrt{q}Z)^2]}{q\mathbb{E}[\phi''(\sqrt{q}Z)^2]}$.

- Example : Tanh, Swish, ELU (with $\alpha = 1$) ...
- The non-smoothness of ReLU-like Activations makes the convergence rate worse on the EOC

Impact of Smoothness

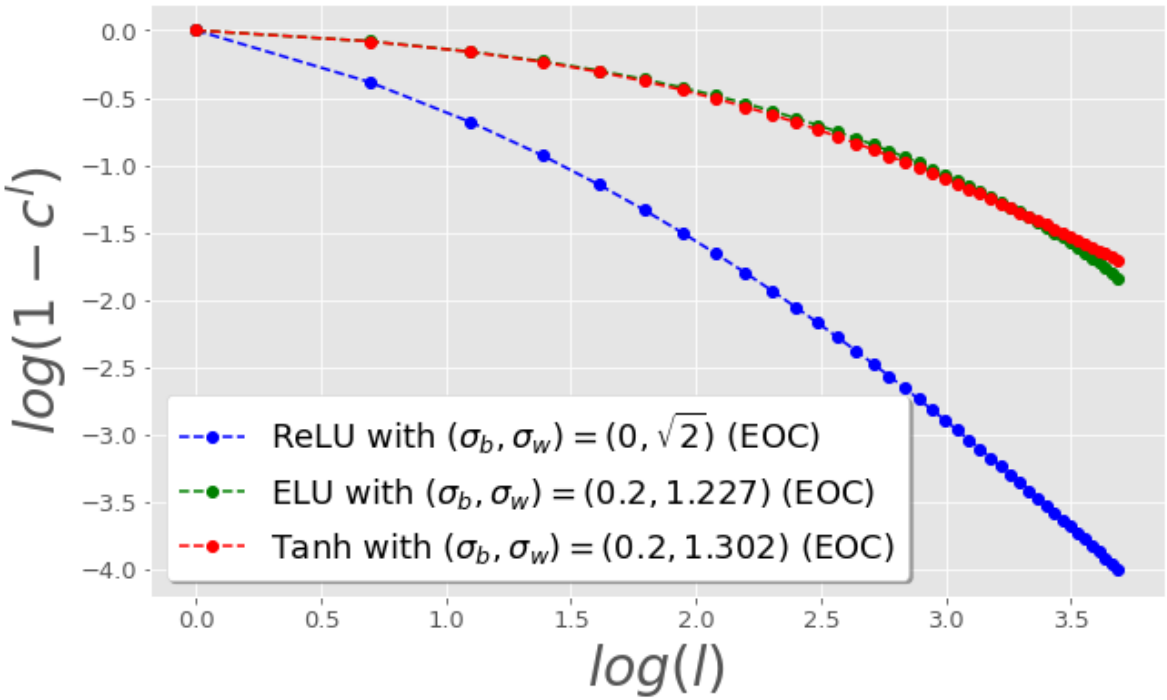
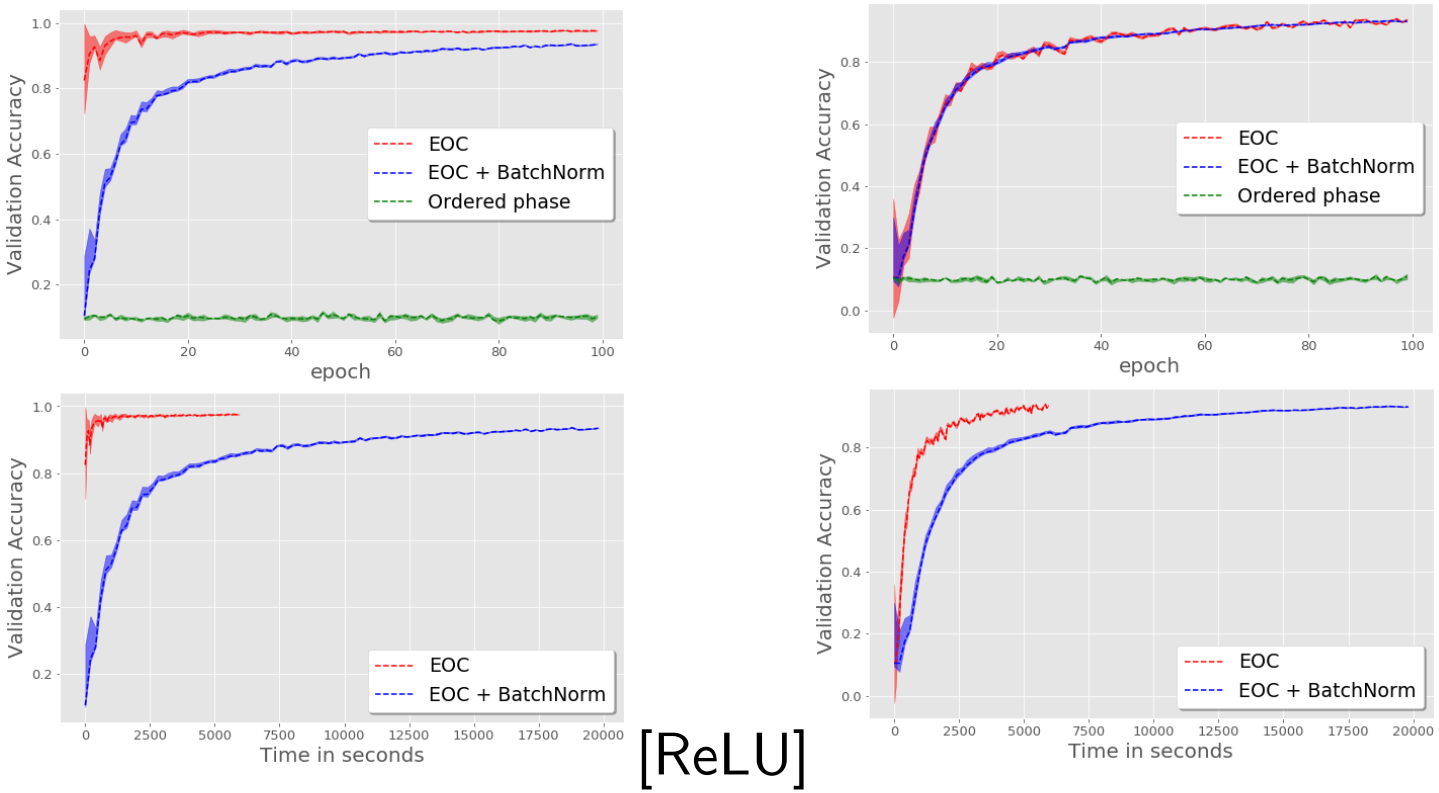


Figure: Impact of the smoothness of the activation function on the convergence of the correlation on the EOC. The convergence rate for ReLU is $\mathcal{O}(1/\ell^2)$ and $\mathcal{O}(1/\ell)$ for ELU and Tanh.

Experiments : Impact of Initializtion on the EOC



[ELU] [ReLU]

Figure: 100 epochs of the training curve (test accuracy) for different activation functions for depth 200 and width 300 using SGD. The red curves correspond to the EOC, the green ones correspond to an ordered phase, and the blue curves corresponds to an Initialization on the EOC plus a Batch Normalization after each layer. Upper figures show the test accuracies with respect to the epochs while lower figures show the accuracies with respect to time.

Experiments : Impact of Initiliaztion on the EOC

Table: Test accuracies for width 300 and depth 200 with different activations on MNIST and CIFAR10 after 100 epochs using SGD

MNIST	EOC	EOC + BN	Ord Phase
ReLU	**93.57± 0.18**	93.11± 0.21	10.09± 0.61
ELU	**97.62± 0.21**	93.41± 0.3	10.14± 0.51
Tanh	**97.20± 0.3**	10.74± 0.1	10.02± 0.13
S-Softplus	10.32± 0.41	9.92± 0.12	10.09± 0.53

CIFAR10	EOC	EOC + BN	Ord Phase
ReLU	**36.55± 1.15**	35.91± 1.52	9.91± 0.93
ELU	**45.76± 0.91**	44.12± 0.93	10.11± 0.65
Tanh	**44.11± 1.02**	10.15± 0.85	9.82± 0.88
S-Softplus	10.13± 0.11	9.81± 0.63	10.05± 0.71

Experiments : Impact of Smoothness

Table: Test accuracies for width 300 and depth 200 with different activations on MNIST and CIFAR10 using SGD

MNIST	Epoch 10	Epoch 50	Epoch 100
ReLU	66.76± 1.95	88.62± 0.61	93.57± 0.18
ELU	**96.09± 1.55**	**97.21± 0.31**	**97.62± 0.21**
Tanh	89.75± 1.01	96.51± 0.51	97.20± 0.3

CIFAR10	Epoch 10	Epoch 50	Epoch 100
ReLU	26.46± 1.68	33.74± 1.21	36.55± 1.15
ELU	**35.95± 1.83**	**45.55± 0.91**	**47.76± 0.91**
Tanh	34.12± 1.23	43.47± 1.12	44.11± 1.02

References

R.M. Neal. Bayesian learning for neural networks. *Springer Science & Business Media*, 118, 1995.

J. Lee, Y. Bahri, R. Novak, S.S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as gaussian processes. *6th International Conference on Learning Representations*, 2018.

A.G. Matthews, J. Hron, M. Rowland, R.E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. *6th International Conference on Learning Representations*, 2018.

S.S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep information propagation. *5th International Conference on Learning Representations*, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *ICCV*, 2015.