Initializing a Neural Network on the Edge of Chaos

Pierre Wolinski

University of Oxford/Inria Grenoble

06/10/2020

Framework

Neural network with layers $l \in \{1, \cdots, L\}$ of sizes n_l. Each neuron performs the following computation:

$$z_i^l = \sum_{j=1}^{n_{l-1}} W_{ij}^l y_j^l + b_i^l, \qquad y_i^{l+1} = \phi(z_i^l),$$

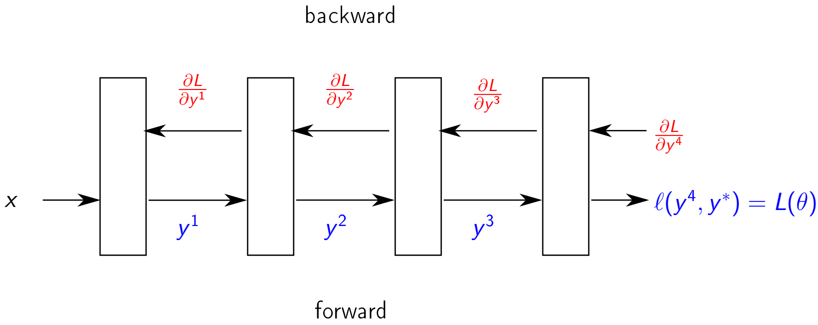
where:

- ϕ: activation function;
- z_i^l, y_i^l: pre-activation, activation of the preceding layer;
- W_{ij}^l, b_i^l: weight, bias.

Backpropagation of the gradient:

$$\frac{\partial L}{\partial W_{ij}^l} = \delta_i^l \phi(z_j^{l-1}) \qquad \delta_i^l = \frac{\partial L}{\partial y_i^l} = \phi'(z_i^l) \sum_j \delta_j^{l+1} W_{ji}^{l+1}$$

Forward and backward



Issues with the initialization of weights and biases

We want to:

- propagate information: avoid activations explosion/collapse;
- backpropagate information: avoid gradients explosion/collapse.

We must avoid these problems at initialization.

$\Rightarrow$ find a good initialization distribution for W_{ij}^l and b_i^l.

Idea: variance preservation

Initialization rule for a layer l. We assume that:

- the inputs $(y_i^0)_i$ and $(y_i^l)_i$ are random and i.i.d.;
- the gradients $\left(\frac{\partial L}{\partial y_i^l}\right)_i$ are i.i.d and independent from the inputs.

Ideas:

- preserve variance of the pre-activations during propagation;
- preserve variance of the gradients during backpropagation.

Idea: variance preservation

Initialization rule for a layer l. We assume that:

- the inputs $(y_i^0)_i$ and $(y_i^l)_i$ are random and i.i.d.;
- the gradients $\left(\frac{\partial L}{\partial y_i^l}\right)_i$ are i.i.d and independent from the inputs.

Ideas:

- preserve variance of the pre-activations during propagation;
- preserve variance of the gradients during backpropagation.

Therefore, we have, for all layers l:

$$\mathrm{Var}(y_\cdot^l) = \mathrm{Var}(y_\cdot^{l+1}) \quad \Rightarrow \quad \mathrm{Var}(\phi(z_\cdot^l)) \approx n_{l-1}\mathrm{Var}(W_{\cdot\cdot}^l) = 1$$

$$\mathrm{Var}\left(\frac{\partial L}{y_\cdot^{l+1}}\right) = \mathrm{Var}\left(\frac{\partial L}{y_\cdot^l}\right) \quad \Rightarrow \quad \mathrm{Var}(\phi'(z_\cdot^l)W_{\cdot\cdot}^l) \approx n_l\mathrm{Var}(W_{\cdot\cdot}^l) = 1$$

Results

Glorot initialization:

$$W^l_{..} \sim \mathcal{U}\left(-\frac{\sqrt{6}}{\sqrt{n_l + n_{l-1}}}, \frac{\sqrt{6}}{\sqrt{n_l + n_{l-1}}}\right)$$
$$\mathrm{Var}(W^l_{..}) = \frac{2}{n_l + n_{l-1}}.$$

Arithmetic compromise between forward and backward variance preservation.

Results

Glorot initialization:

$$W^l_{..} \sim \mathcal{U}\left(-\frac{\sqrt{6}}{\sqrt{n_l + n_{l-1}}}, \frac{\sqrt{6}}{\sqrt{n_l + n_{l-1}}}\right)$$

$$\mathrm{Var}(W^l_{..}) = \frac{2}{n_l + n_{l-1}}.$$

Arithmetic compromise between forward and backward variance preservation.

He initialization:

$$W^l_{..} \sim \mathcal{N}\left(0, \left(\frac{\sqrt{2}}{\sqrt{n_{l-1}}}\right)^2\right).$$

Factor $\sqrt{2}$: variance preservation when $\phi = \mathrm{ReLU}$.
Preserves the variance during the forward pass.
Backward pass: no variance collapse/explosion, even for arbitrarily deep NNs.

References

- *Understanding the difficulty of training deep feedforward neural networks*, Glorot and Bengio, 2010;
- *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, He et al., 2015.

Framework

Study:

- propagation of deterministic inputs;
- correlation between two data points across a neural network.

Given two inputs input a and b:

- $(x_{i;a})_i$: input vector of a data point a;
- $(z^l_{i;a})_i$: the pre-activation at layer l, given a data point a.
- $q^l_{aa} = \mathbb{E}[z^2_{i;a}]$: expected variance of a pre-activation at layer l;
- $q^l_{ab} = \mathbb{E}[z_{i;a} z_{i;b}]$: expected covariance between the pre-activations of two inputs a and b at layer l.

Goal: study the sequences $(q^l_{aa})_l$ and $(q^l_{ab})_l$. Fixed point? Convergence? ...

Recurrence Relations

Assumption: the pre-activations are Gaussian (i.e. the layers are very wide).
Initialization: $W_{ij}^l \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{n_{l-1}}\right)$ and $b_i^l \sim \mathcal{N}(0, \sigma_b^2)$

Recurrence relations:

$$q_{aa}^l = \sigma_w^2 \int \phi^2 \left(\sqrt{q_{aa}^{l-1}}z\right) \mathcal{D}z + \sigma_b^2$$

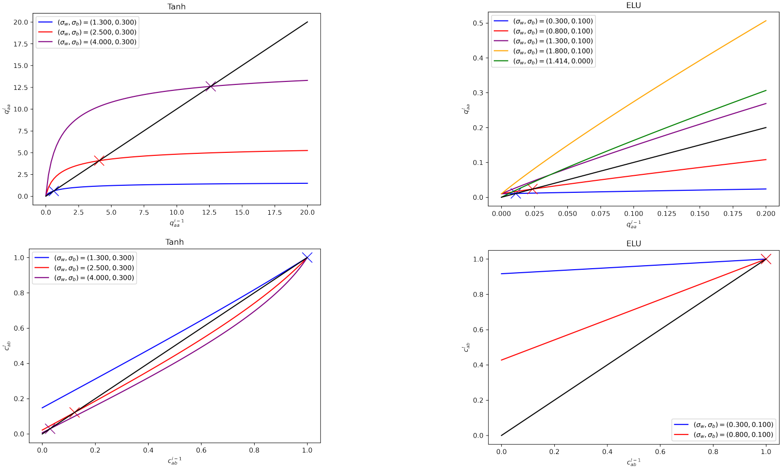
$$q_{ab}^l = \sigma_w^2 \int \phi(u_1)\phi(u_2)\mathcal{D}z_1\mathcal{D}z_2 + \sigma_b^2,$$

where $\mathcal{D}$ is the Gaussian distribution $\mathcal{N}(0,1)$, $u_1 = \sqrt{q_{aa}^{l-1}}z_1$,
$u_2 = \sqrt{q_{bb}^{l-1}}\left(c_{ab}^{l-1}z_1 + \sqrt{1-(c_{ab}^{l-1})^2}z_2\right)$.

Correlation: $c_{ab}^l = \frac{q_{ab}^l}{\sqrt{q_{aa}^l q_{bb}^l}}$.

Fixed points

Convergence towards fixed points for several activation functions ϕ.



Meaning of the fixed points

$\sqrt{q^*}$ is the limit of the norms of the pre-activations $(\|z_a^l\|_2)_l$ as $l \to \infty$.

Interpretation of the convergence of $(c_{ab}^l)_l$ towards c^*:

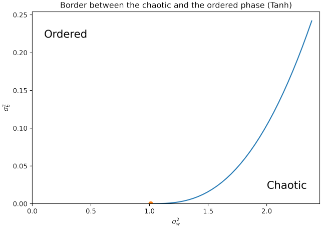
- $c^* = 1$: two inputs a and b, even far from each other, tend to be fully correlated
  $\Rightarrow$ order;
- $c^* \in [0, 1[$: two close inputs a and b, tend to decorrelate
  $\Rightarrow$ chaos.

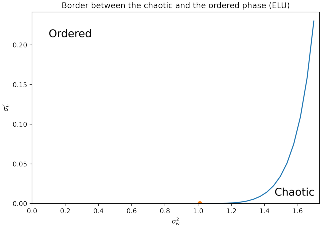
Discriminator χ_1

$c^* = 1$ is always a fixed point. Stability determined by $\chi_1 = \left.\frac{\partial c^l_{ab}}{\partial c^{l-1}_{ab}}\right|_{c=1}$.

Cases:

- $\chi_1 < 1 \Rightarrow c^* = 1$ is stable $\Rightarrow$ order;
- $\chi_1 = 1 \Rightarrow c^* = 1$ is "astable" $\Rightarrow$ transition order/chaos;
- $\chi_1 > 1 \Rightarrow c^* = 1$ is unstable, the limit is $< 1 \Rightarrow$ chaos;





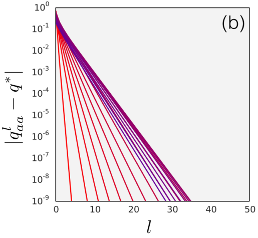
Exponential convergence of q^l_{aa} and c^l_{ab}

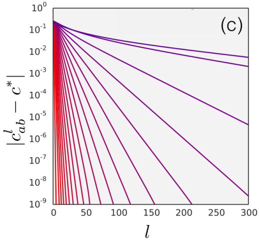
Across how many layers does information propagates/backpropagates?

Speed of convergence of q^l_{aa} and c^l_{ab} with respect to l.

Goal: the convergence of $(c^l_{ab})_l$ should be as slow as possible.
Preserve initial correlation.

Figures: red: $\sigma_w^2 = 0.01$; purple: $\sigma_w^2 = 1.7$.





Depth of Information Propagation

Guess: Exponential convergence.

- $|q^l_{aa} - q^*| \sim \exp\left(-\frac{l}{\xi_q}\right)$;
- $|c^l_{ab} - c^*| \sim \exp\left(-\frac{l}{\xi_c}\right)$.

Asymptotic recurrence relations give:

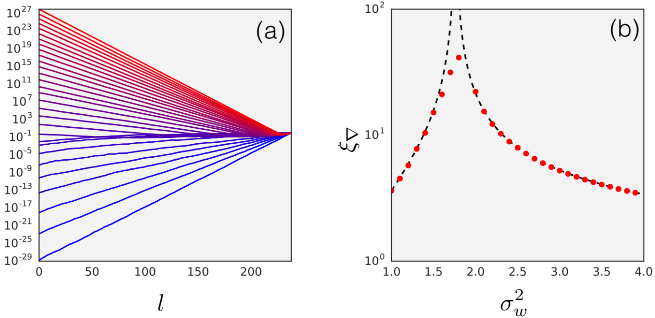
- $\xi_q^{-1} = -\log\left[\chi_1 + \sigma_w^2 \int \phi''(\sqrt{q^*}z)\phi(\sqrt{q^*}z)\mathcal{D}z\right]$;
- $\xi_c^{-1} = \log\left[\sigma_w^2 \int \phi'(u_1^*)\phi'(u_2^*)\mathcal{D}z_1\mathcal{D}z_2\right]$

Case $\chi_1 = 1$: $\xi_c = \infty$.
$(c^l_{ab})_l$ converges slower than exponentially towards 1.

About the gradients

Layers necessary to achieve convergence:

- expected gradient norm: $\xi_\nabla^{-1} = -\log \chi_1$;
- expected gradient correlation: $\xi_c^{-1} = \log \left[\sigma_w^2 \int \phi'(u_1^*)\phi'(u_2^*)\mathcal{D}z_1\mathcal{D}z_2\right]$.



Neural Tangent Kernels (NTK)

Dynamics of the training of a neural network:

$$\mathrm{d}f_t(\mathcal{X}) = -\frac{1}{N} K^L_{\theta_t}(\mathcal{X}, \mathcal{X}) \nabla_z \ell(f_t(\mathcal{X}), \mathcal{Y}) \, \mathrm{d}t,$$

where f_t is the NN, $(\mathcal{X}, \mathcal{Y})$ is the training set, θ_t is the vector of parameters, $\ell(z, y)$ is the loss, and $K^L_{\theta_t}$ is the "NTK".

Neural Tangent Kernels (NTK)

Dynamics of the training of a neural network:

$$\mathrm{d}f_t(\mathcal{X}) = -\frac{1}{N} K^L_{\theta_t}(\mathcal{X}, \mathcal{X}) \nabla_z \ell(f_t(\mathcal{X}), \mathcal{Y})\, \mathrm{d}t,$$

where f_t is the NN, $(\mathcal{X}, \mathcal{Y})$ is the training set, θ_t is the vector of parameters, $\ell(z, y)$ is the loss, and $K^L_{\theta_t}$ is the "NTK".

In the infinite-width limit:

$$\mathrm{d}f_t(\mathcal{X}) = -\frac{1}{N} K^L(\mathcal{X}, \mathcal{X}) \nabla_z \ell(f_t(\mathcal{X}), \mathcal{Y}) \mathrm{d}t,$$

where K^L is the limit kernel as the widths tend to infinity.
K^L *depends of the initialization distribution of weights and biases.*

Neural Tangent Kernels (NTK)

Dynamics of the training of a neural network:

$$\mathrm{d}f_t(\mathcal{X}) = -\frac{1}{N} K^L_{\theta_t}(\mathcal{X}, \mathcal{X}) \nabla_z \ell(f_t(\mathcal{X}), \mathcal{Y}) \, \mathrm{d}t,$$

where f_t is the NN, $(\mathcal{X}, \mathcal{Y})$ is the training set, θ_t is the vector of parameters, $\ell(z, y)$ is the loss, and $K^L_{\theta_t}$ is the "NTK".

In the infinite-width limit:

$$\mathrm{d}f_t(\mathcal{X}) = -\frac{1}{N} K^L(\mathcal{X}, \mathcal{X}) \nabla_z \ell(f_t(\mathcal{X}), \mathcal{Y}) \mathrm{d}t,$$

where K^L is the limit kernel as the widths tend to infinity.
K^L depends of the initialization distribution of weights and biases.

Example with $\ell(z, y) = \frac{1}{2} \|z, y\|_2^2$:

$$\mathrm{d}f_t(\mathcal{X}) = -\frac{1}{N} K^L(\mathcal{X}, \mathcal{X})(f_t(\mathcal{X}) - \mathcal{Y}) \mathrm{d}t.$$

NTK with EOC Initialization

Behavior of K^L as $L \to \infty$?

Results:

- if K^L is singular, then there exists $C > 0$ such that, for all t: $\|f_t(\mathcal{X}) - \mathcal{Y}\| \geq C$.
  $\Rightarrow$ untrainable NN;
- if the initialization parameters (σ_w, σ_b) lie in the ordered or chaotic phase, then K^L tends towards a constant kernel $K^\infty(x, x') = \lambda$, which is singular
  $\Rightarrow$ untrainable;
- if the initialization parameters (σ_w, σ_b) lie on the EOC, then K^∞ is invertible
  $\Rightarrow$ trainable.

Conclusion: EOC initialization is useful to make an arbitrarily deep NN trainable.

References

- EOC: *Exponential expressivity in deep neural networks through transient chaos*, Poole et al., 2016;
- EOC and layer scales: *Deep information propagation*, Schoenholz et al., 2016;
- NTK: *Neural tangent kernel: Convergence and generalization in neural networks*, Jacot et al., 2018;
- NTK and EOC: *Mean-field Behaviour of Neural Tangent Kernel for Deep Neural Networks*, Hayou et al., 2019.

Conclusion

- importance of the Edge of Chaos to study the training dynamics;
- question: validity of the infinite-width approximations (EOC and NTK);
- question: Gaussian behavior of the pre-activations (EOC);
- "philosophical" question: after Glorot/He and EOC approaches, is there another refinement of the initialization procedure?

Conclusion

- importance of the Edge of Chaos to study the training dynamics;
- question: validity of the infinite-width approximations (EOC and NTK);
- question: Gaussian behavior of the pre-activations (EOC);
- "philosophical" question: after Glorot/He and EOC approaches, is there another refinement of the initialization procedure?

Thank you for your attention!