Initialization

Vincent Barra
LIMOS, UMR 6158 CNRS, Université Clermont Auvergne

LABORATOIRE D'INFORMATIQUE,
DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

What we know

- Initialization should break symmetry
- the relative scale of weights is fundamental.
- and so is initialization.

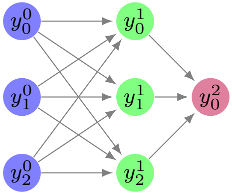
LABORATOIRE D'INFORMATIQUE,
DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

First bad idea

$$\forall i \ \boldsymbol{w}_i = 0$$

1 hidden layer MLP. $(\forall l \in \{1, 2\})(\forall i \in \{0, 2\})$

$$y_i^l = \sigma\left(\sum_{j=0}^{2} w_{ij}^l y_i^{l-1}\right)$$



Using backpropagation on error E

$$\frac{\partial E}{\partial w_{0j}^1} = \frac{\partial E}{\partial w_{1j}^1} = \frac{\partial E}{\partial w_{2j}^1}$$

Hidden neurons learn the same parameters
$\Rightarrow$ redundance.

Second bad idea

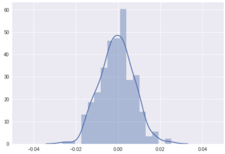
$\forall i\ \boldsymbol{w}_i = \delta$: Same result!

Second bad idea

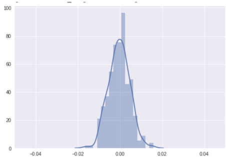
$\forall i \; \boldsymbol{w}_i = \delta$: Same result!

A better idea ?

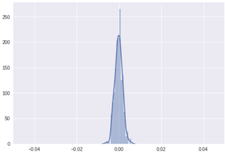
$$\forall i \; \boldsymbol{w}_i \rightsquigarrow 10^{-m}\mathcal{N}(0,1) \; m > 0$$



After 10 epochs



After 20 epochs



After 50 epochs

$\Rightarrow$ vanishing gradient (See dedicated slides).

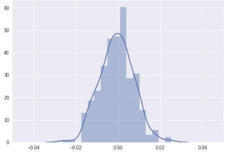
LABORATOIRE D'INFORMATIQUE,
DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

Second bad idea

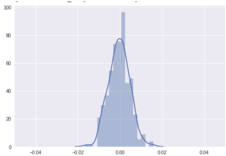
$\forall i \ \boldsymbol{w}_i = \delta$: Same result!

A better idea ?

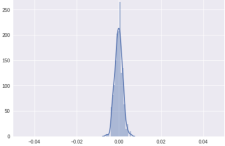
$$\forall i \ \boldsymbol{w}_i \rightsquigarrow 10^{-m}\mathcal{N}(0,1) \ m > 0$$



After 10 epochs



After 20 epochs



After 50 epochs

$\Rightarrow$ vanishing gradient (See dedicated slides).

A second better idea ?

$$\forall i \ \boldsymbol{w}_i \rightsquigarrow \ \textbf{high random values}$$

The scalar products of the post synaptic potentials saturate he activation functions σ. The derivatives tends towards 0.
$\Rightarrow$ vanishing gradient (See dedicated slides).

Xavier/Glorot initialization

MLP $\boldsymbol{w} \in \mathbb{R}^d i.i.d \rightsquigarrow \mathcal{N}(0,1)$ with input $\boldsymbol{x} \in \mathbb{R}^d$, $x_i \;\; i.i.d \rightsquigarrow \mathcal{N}(0,1)$.
Post synaptic potential of a neuron in the first hidden layer: $\boldsymbol{w}^T\boldsymbol{x}$.

$$\begin{aligned} Var(\boldsymbol{w}^T\boldsymbol{x}) &= \sum_{i=1}^{d} Var(w_i x_i) \\ &= \sum_{i=1}^{d} \left(\mathbb{E}(w_i)^2 Var(x_i) + \mathbb{E}(x_i)^2 Var(w_i) + Var(w_i)Var(x_i)\right) \end{aligned}$$

Since $\mathbb{E}(w_i) = \mathbb{E}(x_i) = 0$

$$Var(\boldsymbol{w}^T\boldsymbol{x}) = \sum_{i=1}^{d} Var(w_i)Var(x_i)$$

$w_i,\; x_i$ i.i.d $\Rightarrow \; Var(\boldsymbol{w}^T\boldsymbol{x}) = dVar(w_i)Var(x_i)$
More generally

$$Var(y^l) = (dVar(w_i))^l \, Var(x_i)$$

$\Rightarrow$ Each neuron van vary about d times the variation of its input.

Xavier/Glorot initialization

- if $dVar(w_i) > 1$ the gradient increases when going deeper in the network
- if $dVar(w_i) < 1$ vanishing gradient as l increases

$\Rightarrow$ impose $dVar(w_i) = 1$, thus $Var(w_i) = 1/d$ and

$$\boldsymbol{w}_{ij}^{l} \rightsquigarrow \frac{1}{\sqrt{m^{l-1}}}\mathcal{N}(0,1)$$

where m^{l-1} : number of neurons of layer $l-1$.
If $\sigma = ReLU$ one can multiply by $\frac{\sqrt{2}}{\sqrt{m^{l-1}}}$ to take into account the negarive part that does not participate to the variance.

LABORATOIRE D'INFORMATIQUE, DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

UNIVERSITÉ Clermont Auvergne