Perceptrons and Multilayer Perceptrons

Be careful of the vanishing gradient !!

Vincent Barra
LIMOS, UMR 6158 CNRS, Université Clermont Auvergne

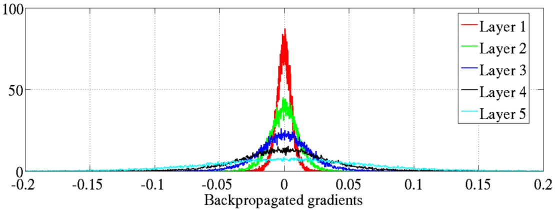
Be careful of the vanishing gradient !!

Problem statement

When L increases, small gradients tend to disappear

$\rightarrow$ Blocking gradient descent

$\rightarrow$ Limited capacity of learning



Backpropagated gradients normalized histograms (source: Glorot & Bengio, 2010)

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

Be careful of the vanishing gradient !!

Let consider a 3-layer MLP without bias

$$f_{\boldsymbol{w}}(x) = \sigma\left[w_3 \sigma\left(w_2\left(\sigma(w_1 x)\right)\right)\right]$$

To apply the chain rule, the MLP computes

$$u_1 = w_1 x, \quad u_2 = \sigma(u_1), \quad u_3 = w_2 u_2, \quad u_4 = \sigma(u_3), \quad u_5 = w_3 u_4$$

and thus

$$\begin{aligned}
\frac{df_{\boldsymbol{w}}(x)}{dw_1} &= \frac{\partial f_{\boldsymbol{w}}(x)}{\partial u_5} \frac{\partial u_5}{\partial u_4} \frac{\partial u_4}{\partial u_3} \frac{\partial u_3}{\partial u_2} \frac{\partial u_2}{\partial u_1} \frac{\partial u_1}{\partial w_1} \\
&= \frac{\partial \sigma(u_5)}{\partial u_5} w_3 \frac{\partial \sigma(u_3)}{\partial u_3} w_2 \frac{\partial \sigma(u_1)}{\partial u_1} x
\end{aligned}$$

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

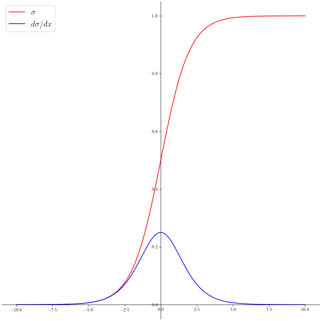
Be careful of the vanishing gradient !!

$$\frac{df_{\boldsymbol{w}}(x)}{dw_1} = \frac{d\sigma(u_5)}{du_5} w_3 \frac{d\sigma(u_3)}{du_3} w_2 \frac{d\sigma(u_1)}{du_1} x$$

But $0 \leq \frac{d\sigma(z)}{dz} = \sigma(z)\,(1 - \sigma(z)) \leq \frac{1}{4}$

So if $w_i \sim \mathcal{N}(0, \Sigma), \Sigma < 1$ then with high probability $w_i \leq 1$ and

$$0 \leq \frac{df_{\boldsymbol{w}}(x)}{dw_1} \leq \frac{1}{4^3} x$$



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

UNIVERSITÉ Clermont Auvergne

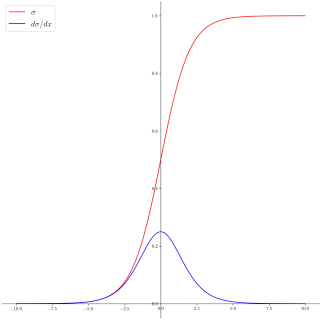
Be careful of the vanishing gradient !!

$$\frac{df_{\boldsymbol{w}}(x)}{dw_1} = \frac{d\sigma(u_5)}{du_5} w_3 \frac{d\sigma(u_3)}{du_3} w_2 \frac{d\sigma(u_1)}{du_1} x$$

But $0 \leq \frac{d\sigma(z)}{dz} = \sigma(z)\,(1-\sigma(z)) \leq \frac{1}{4}$

So if $w_i \sim \mathcal{N}(0, \Sigma), \Sigma < 1$ then with high probability $w_i \leq 1$ and

$$0 \leq \frac{df_{\boldsymbol{w}}(x)}{dw_1} \leq \frac{1}{4^3} x$$



→ $\frac{df_{\boldsymbol{w}}(x)}{dw_1} \rightarrow 0$ as L increases
→ True for almost all bounded activation functions (σ, $tanh$,...)
→ Proper initialization scheme needed for the w_i

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