Deep learning - Session 2

- Jonas Freiburghaus
- Vincent Moulin

03.03.2021

Exercice 2

a) Compute the derivative of the sigmoid function

$$\sigma = \frac{1}{1+e^{-z}}$$

The derivative is:

$$\sigma' = \frac{\partial}{\partial z}\frac{1}{1+e^{-z}} = \frac{\partial}{\partial z}(1+e^{-z})^{-1} = -(1+e^{-z})^{-2}\frac{\partial}{\partial z}(1+e^{-z}) = \frac{e^{-z}}{(1+e^{-z})^2}$$

b) Show that the derivative fullfills the equation

$$\sigma'(z) = \sigma(z)(1-\sigma(z))$$

Proof:

$$\sigma'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{(1+e^{-z})}\frac{e^{-z}}{(1+e^{-z})} = \sigma(z)\frac{e^{-z}+1-1}{(1+e^{-z})}$$

$$= \sigma(z)\left(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}}\right) = \sigma(z)(1-\sigma(z))$$

c) Compute the first and second derivative of the function

$$\zeta(z) = -log(\sigma(-z))$$

First derivative:

$$\zeta'(z) = -\frac{\partial}{\partial \sigma}log(\sigma(-z))\frac{\partial \sigma}{\partial z} = \frac{1}{\sigma(-z)}\sigma(-z)(1-\sigma(-z))$$

$$= 1-\sigma(-z) = \frac{1}{1} - \frac{1}{1+e^{z}} = \frac{1+e^{z}}{1+e^{z}} - \frac{1}{1+e^{z}} = \frac{e^{z}}{1+e^{z}} = \frac{1}{1+e^{-z}} = \sigma(z)$$

Second derivative:

$$\zeta''(z) = \sigma'(z) = \sigma(z)(1-\sigma(z))$$

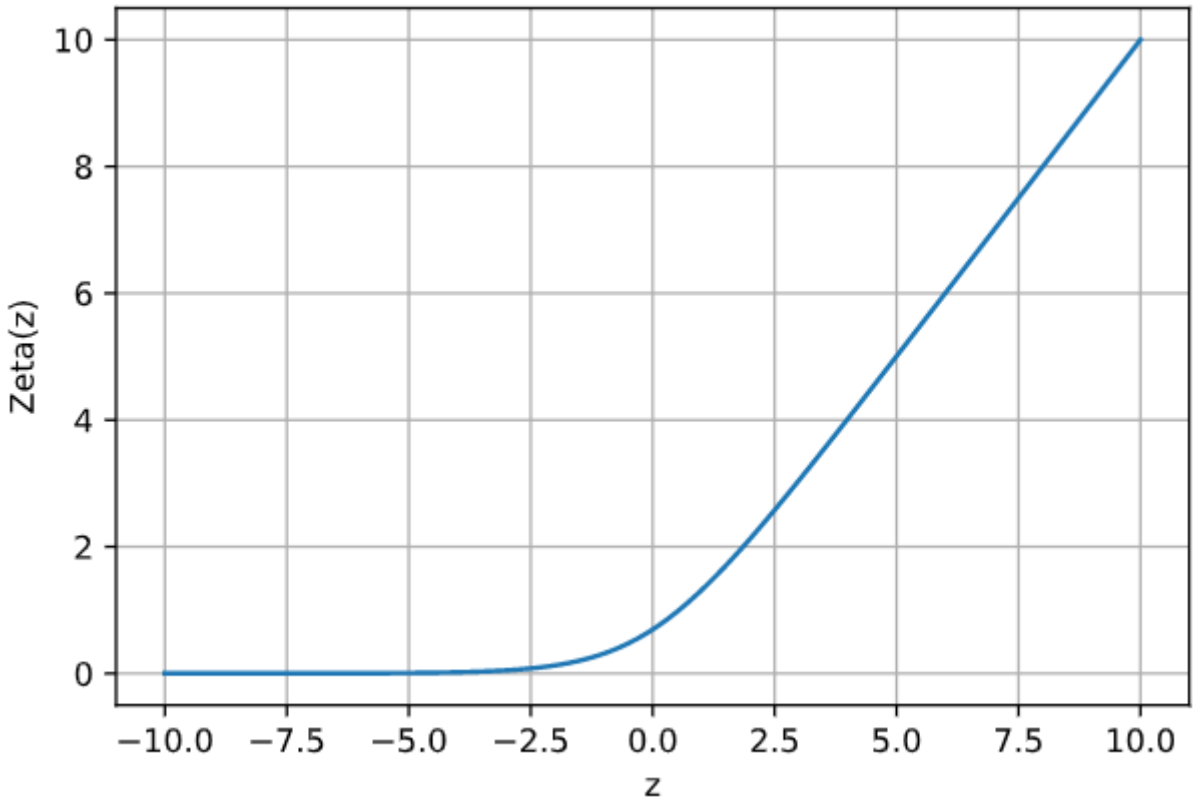
Compute the asymptots as z goes to +- infinity :

$$\lim_{z\to+\infty} -log\left(\frac{1}{1+e^{z}}\right) = -\lim_{z\to+0} log(z) = +\infty$$

$$\lim_{z \to -\infty} -log\left(\frac{1}{1+e^{z}}\right) = -\lim_{z \to +1} log(z) = 0$$

By using hôpital's rule

$$\lim_{z \to +\infty} \frac{\zeta(z)}{z} = \lim_{z \to +\infty} \frac{\zeta'(z)}{z'} = \lim_{z \to +\infty} \frac{\sigma(z)}{1} = \frac{1}{1} = 1$$



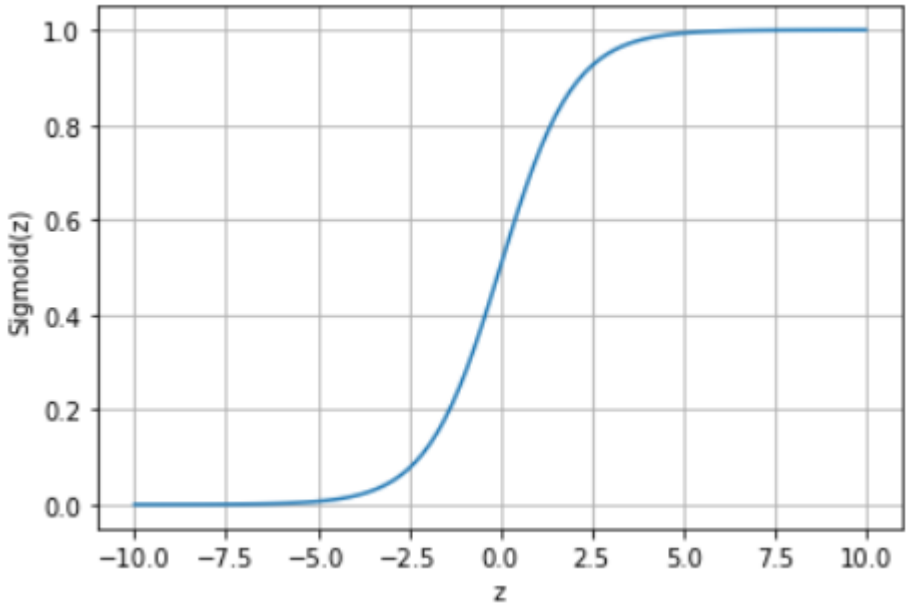
d) Implement the sigmoid function in a Jupyter Notebook

```
import matplotlib.pyplot as plt
import numpy as np

def sigmoid(z):
    return 1. / (1 + np.exp(-z))

x = np.linspace(-10, 10, 100)
y = sigmoid(x)

plt.plot(x, y)
plt.grid(True)
plt.xlabel("z")
plt.ylabel("Sigmoid(z)")
plt.show()
```



f) Show that the function is non-convex

$$c_1(x) = (\sigma(x) - 1)^2$$

First derivative

$$c_1'(x) = \frac{\partial}{\partial \sigma}(\sigma(x) - 1)^2 \frac{\partial \sigma}{\partial x} = 2(\sigma(x) - 1)\sigma'(x)$$

$$2(\sigma(x) - 1)\sigma(x)(1 - \sigma(x)) = -2\sigma(x)c_1(x)$$

The first derivative could be equal to 0 when:

- sigma(x) = 0 -> never the case as sigma(x) in]0; 1[
- sigma(x) - 1 = 0 -> sigma(x) = 1 -> never the case as sigma(x) in]0; 1[

So it is never equal to 0.

Second derivative

$$c_1''(x) = -2\left(\frac{\partial c_1}{\partial x}\sigma(x) + \frac{\partial \sigma}{\partial x}c_1(x)\right)$$

$$= -2\left(-2\sigma^2(x)c_1(x) + \sigma(x)(1 - \sigma(x))c_1(x)\right)$$

$$= -2c_1(x)\left(\sigma(x) - 3\sigma^2(x)\right)$$

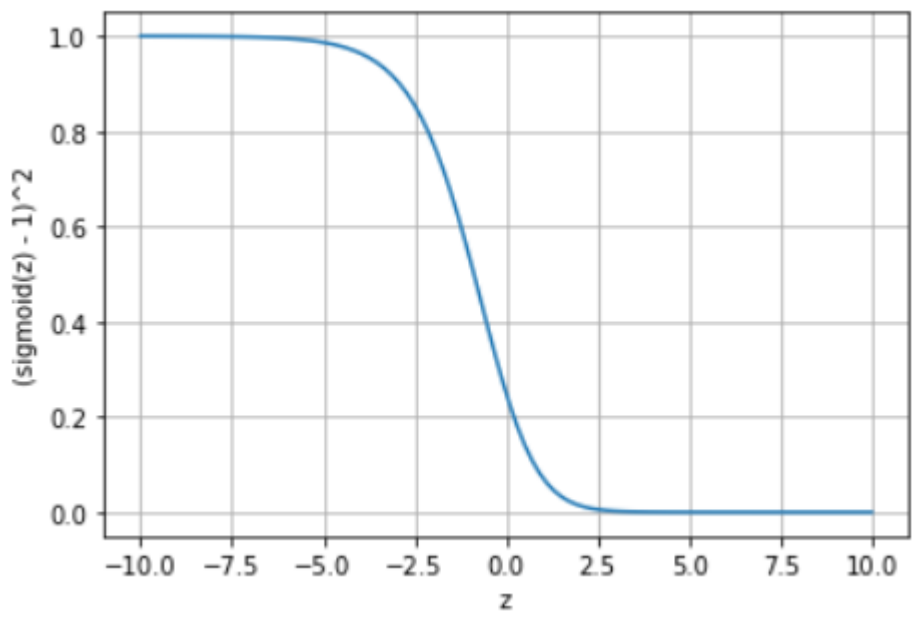
$$= -2(\sigma(x) - 1)^2\sigma(x)(1 - 3\sigma(x))$$

The second derivative could be equal 0 when :

- sigma(x) = 0 -> never the case as sigma(x) in]0; 1[
- sigma(x) - 1 = 0 -> sigma(x) = 1 -> never the case as sigma(x) in]0; 1[
- 1 - 3sigma(x) = 0 -> sigma(x) = 1/3

$$\frac{1}{1 + e^{-x}} = \frac{1}{3}$$

So when x = -log(2). Consequently on the left of x = -log(2) it is either > 0 or < 0 and the opposite on the right of x = -log(2). The function is non-convex.



Visually, if we were to draw a line between z = 5 and z = -5. The line would pass through the function. Affirming our previous conclusion.

Optimising with gradient descent may become difficult when the function is non-convex as we can fall in minimum local or we can face to a saddle point.

g) Compute the first and second derivative of the function

$$c_2(x) = -(y\log(\sigma(\omega x)) + (1-y)log(1-\sigma(\omega x)))$$

First derivative:

Let's rewrite:

$$z = \omega x$$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Such that

$$c_2(\sigma(z(x,w)),y)$$

By the chain rule :

$$\frac{\partial c_2}{\partial w} = \frac{\partial c_2}{\partial \sigma}\frac{\partial \sigma}{\partial z}\frac{\partial z}{\partial w}$$

$$\frac{\partial z}{\partial w} = -\left(y\frac{1}{\sigma} + (1-y)\frac{1}{1-\sigma}\right) = -\left(\frac{y}{\sigma} - \frac{1-y}{1-\sigma}\right) = \frac{\sigma - y}{\sigma(1-\sigma)}$$

$$\frac{\partial \sigma}{\partial z} = \sigma(z)(1-\sigma(z))$$

$$\frac{\partial z}{\partial w} = x$$

We obtain:

$$\frac{\partial c_2}{\partial w} = \frac{\sigma - y}{\sigma(1-\sigma)}\sigma(z)(1-\sigma(z))x = (\sigma(\omega x) - y)x$$

When y = 0

$$c_2'(\sigma(z(x,w)),0) = \sigma(\omega x)x$$

When y = 1

$$c_2'(\sigma(z(x,w)),1) = (\sigma(\omega x) - 1)x$$

Setting the derivative to 0:

$$c_2'(\sigma(z(x,w)),y) = 0$$

Which is true when :

- sigmoid(wx) = y, (prediction = ground truth)

Second derivative:

$$\frac{\partial^2 c_2}{\partial \omega} = x(\sigma(wx)(1-\sigma(wx))x = x^2\sigma'(\omega x)$$

The logistic loss is a convex function with respect to w as :

$$x^2\sigma'(\omega x) > 0, \forall x \in \real, \forall w \in \real$$

Exercice 3

Computation of the gradients :

Logistic loss

Remark: Missing summation for the whole training set for simplification purposes and not the full gradient

$$c_2(x) = -(y\log(\sigma(\omega x)) + (1-y)log(1-\sigma(\omega x)))$$

First derivative:

Let's rewrite:

$$z = \omega x$$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Such that

$$c_2(\sigma(z(x,w)),y)$$

By the chain rule :

$$\frac{\partial c_2}{\partial w} = \frac{\partial c_2}{\partial \sigma}\frac{\partial \sigma}{\partial z}\frac{\partial z}{\partial w}$$

$$\frac{\partial z}{\partial w} = -\left(y\frac{1}{\sigma} + (1-y)\frac{1}{1-\sigma}\right) = -\left(\frac{y}{\sigma} - \frac{1-y}{1-\sigma}\right) = \frac{\sigma - y}{\sigma(1-\sigma)}$$

$$\frac{\partial \sigma}{\partial z} = \sigma(z)(1-\sigma(z))$$

$$\frac{\partial z}{\partial w} = x$$

$$\frac{\partial c_2}{\partial w} = \frac{\sigma - y}{\sigma(1-\sigma)}\sigma(z)(1-\sigma(z))x = (\sigma(\omega x) - y)x$$

MSE with sigmoid activation function

$$\text{MSE} = J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$$

Which we can write as :

$$J(h(z(\theta, X)), Y)$$

Consequently the chain rule is :

$$\frac{\partial J}{\partial \theta_i} = \frac{\partial J}{\partial h}\frac{\partial h}{\partial z}\frac{\partial z}{\partial \theta_i}$$

Applying the chain rule :

$$\frac{\partial J(\theta)}{\partial \theta_i} = \frac{1}{2m}\sum_{i=1}^{m}\frac{\partial}{\partial h_\theta}(h_\theta(z_i) - y_i)^2\frac{\partial h_\theta(z_i)}{\partial z_i}\frac{\partial z_i}{\partial x_i}$$

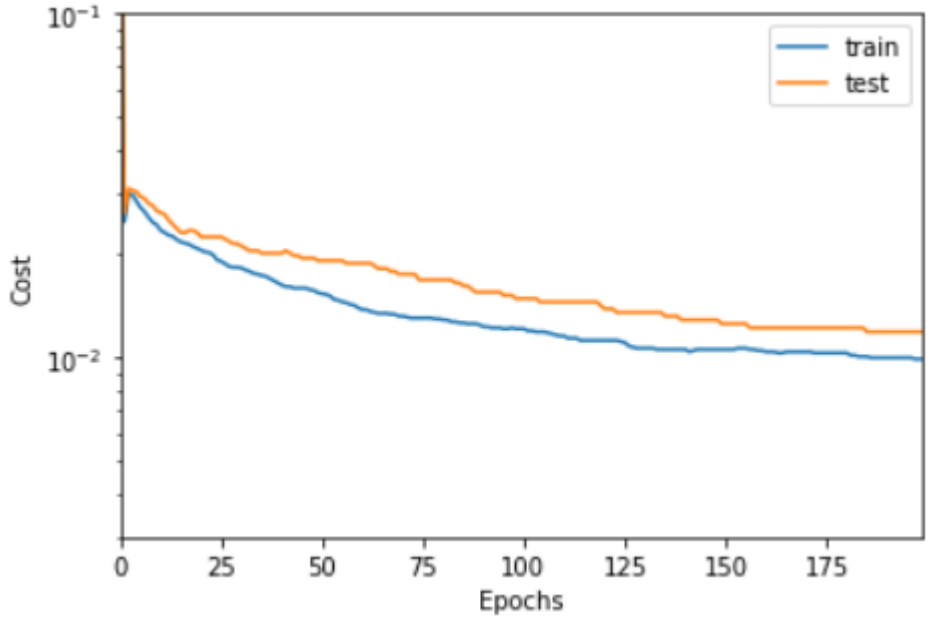
$$\frac{\partial J(\theta)}{\partial \theta_i} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(z_i) - y_i)(h_\theta(z_i)(1-h_\theta(z_i)))x_i$$

D) Analyse the dependency of the final error rate on the number of epochs. What is the goal of the learning and how many epochs make sense ? (Choose here the learning rate α = 0.5.)

The goal of the learning is to find a minima of the loss function. Ideally we would like to find the global minima which we should be able to find when the loss function is convex with respect to the model's weights. However most of the time we may find a local minima that could be good enough.

The number of epochs we should pick is dependent on the task. We should pick a number of epochs that brings the generalization error as low as possible.

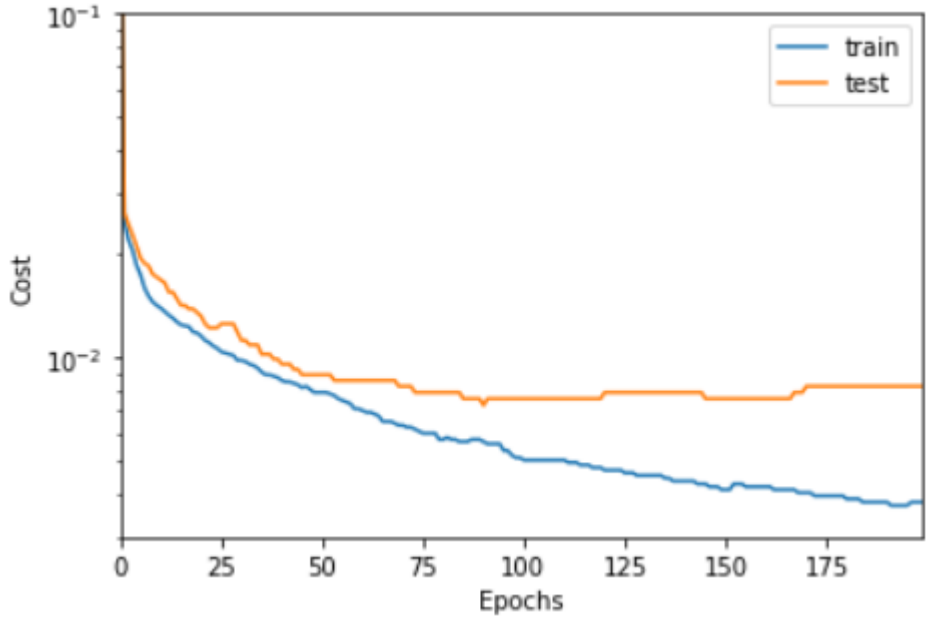
For a learning rate of 0.5, after 200 epochs it seems like the error could still decrease a little bit.



Both the curves for the training set and test set follow each other quite well. Meaning it does neither overfit nor underfit.

E) Analyse the dependency on the learning rate by trying different values

For a learning rate = 10 :



We observe that the model seems to have converged at about 100 epochs.

If the learning rate is too big there is a risk of divergence. The gradients could bounce and start climbing up the loss function.

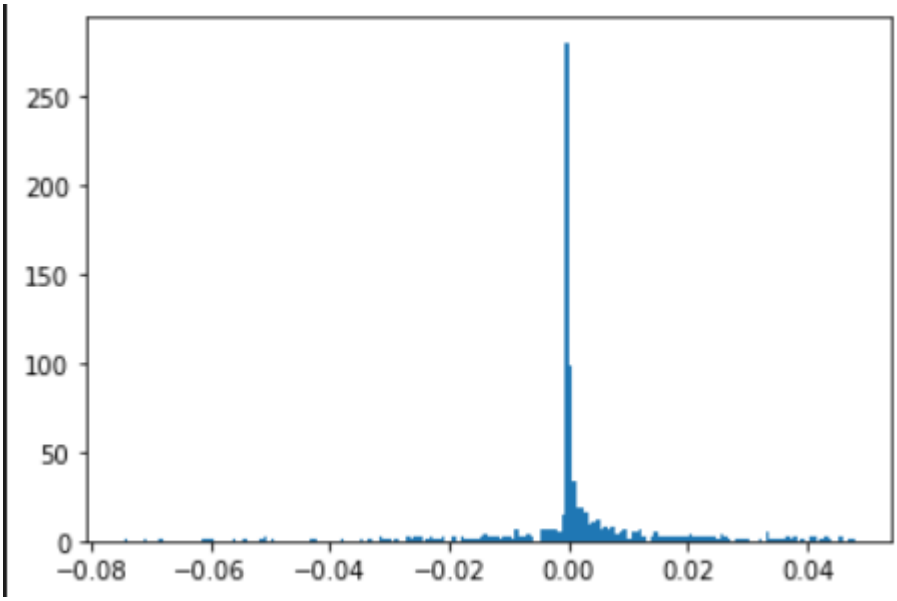
If the learning rate is too small the training will take too much time and consequently waste resources.

Reasonable number of epochs for the following learning rates:

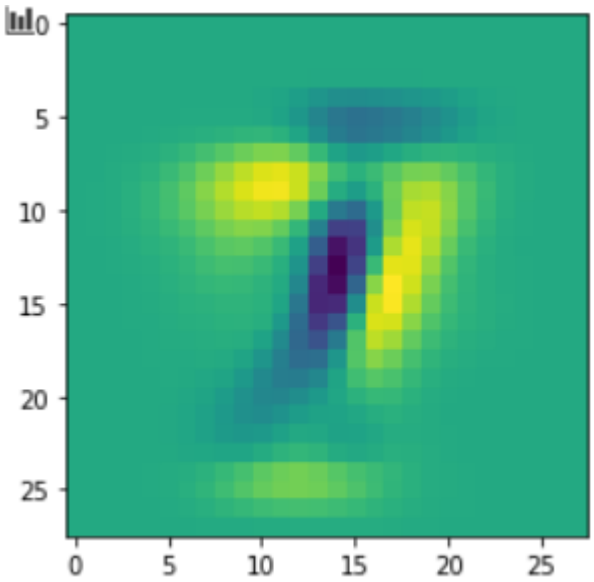
- 0.01 : not a reasonable learning rate
- 0.05 : 500 epochs

- 0.1 : 500 epochs
- 0.5 : 200 epochs
- 1.0 : 200 epochs
- 1.5 : 200 epochs
- 2.0 : 200 epcohs
- 5.0 : 150 epochs
- 10.0 : 100 epochs

F) Plot a histogram of the weights finally obtained from learing. A strong peak at zero remains. Why ?



The many zeroes comes from the fact that most pixels are part of the black background.



When looking at this image we can see which pixels would activate the perceptron for each class.

Some of these numbers are even difficult to classify by a human. Some of the 1 looks like 7 and vice versa.