Anna Krasilnikova 2562668
Mirza Misbah Mubeen Baig 2571567
Shahzain Mehboob 2571564

Exercise 8.1 – *Computational Graph*

In the Exercise 6.3 part (b), you have applied a forward and backward pass to compute the derivative of a term represented by a computational graph.

In this exercise we want to do the same with a logistic regression classifier $f(x) = \sigma(x_1 w_1 + x_2 w_2 + b)$ *with the sigmoid function* $\sigma(x) = \frac{1}{1+e^{-x}}$

a) *Setup the computation graph of the classifier. Split the whole model term into simple operations (i.e. Add, subtract, multiply, divide and exponential) that form the nodes of the graph. For each operation, state its derivative.*

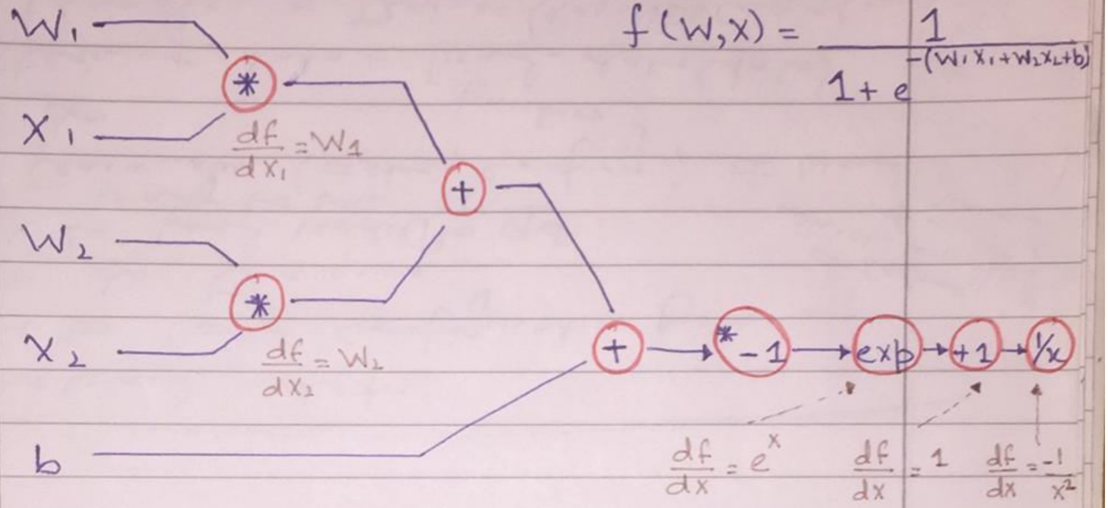
b) *Perform a forward and backward pass using the following values: x1 = 1, x2 = 0.5, w1 = 0.25, w2 = 0.3, b = 1. Keeping all other weights fix, how must w1 be adapted in order to flip the classification result of the given input? Justify your*

(For part (b) both forward and backward values are written in same graph)

EXERCISE 8.1 COMPUTATIONAL GRAPH.

(a) Solution:

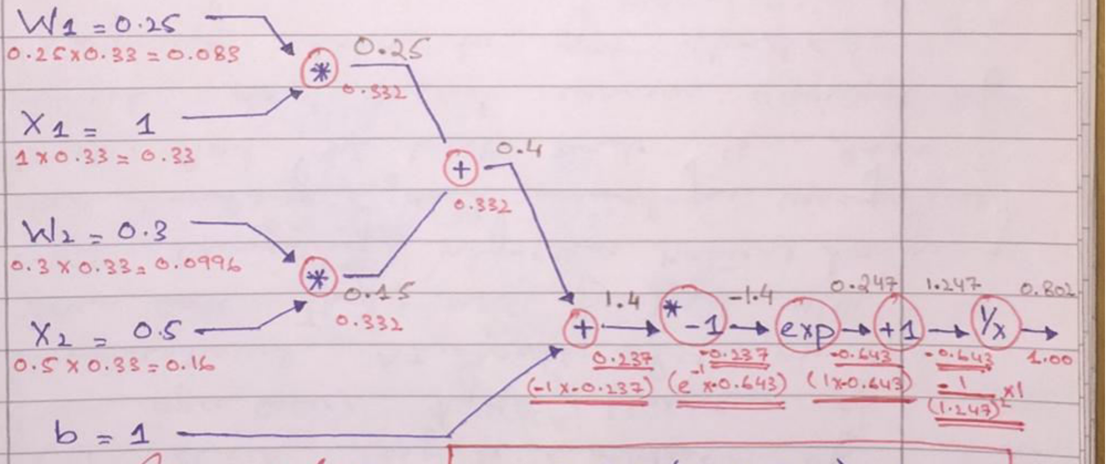
$$\begin{cases} f(x) = \sigma(X_1W_1 + X_2W_2 + b) \\ \sigma(x) = \dfrac{1}{1+e^{-x}} \end{cases}$$



$$f(W,x) = \frac{1}{1+e^{-(W_1X_1+W_2X_2+b)}}$$

(b) Solution: $x_1 = 1$, $x_2 = 0.5$, $w_1 = 0.25$, $w_2 = 0.3$, $b = 1$

$$f(1,0.5) = \frac{1}{1+e^{-(0.25\times1+0.3\times0.5+1)}} = \frac{1}{1+e^{-1.4}} = \underline{\underline{0.802}}$$



Sigmoid Gate $\boxed{\Rightarrow 0.802(1-0.802) = 0.159}$

Keeping other weights fix, $\frac{1}{1+e^{-(W_1+0.15+1)}} = 1-0.802$

$\Rightarrow 1+e^{-(W_1+1.15)} = 5.051$

$\Rightarrow \boxed{W_1 = -2.549}$

Anna Krasilnikova 2562668
Mirza Misbah Mubeen Baig 2571567
Shahzain Mehboob 2571564

Exercise 8.2 – ReLU and Tanh Functions

In Exercise 6.2. we calculated and sketched the derivative for sigmoid function, we will do the same for other activation functions to understand the difference between them and the benefit of each in Neural Networks:

a) *Calculate the derivative of the tanh and ReLU function by hand.*

Derivative of ReLU:

$$ReLU(x) = \max(0, x)$$

$$\frac{d}{dx}\max(0, x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

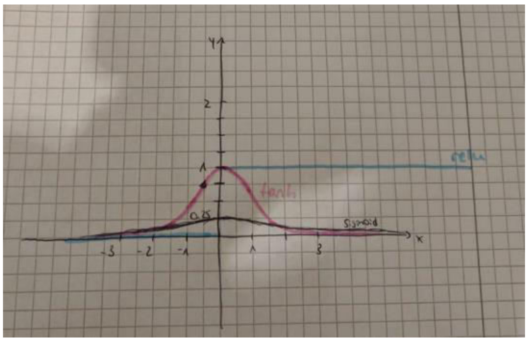
Derivative of tanh:

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

$$\Rightarrow \frac{d}{dz} tahn(z) = \frac{(1 + e^{-2z}) \cdot 0 - 2 \cdot e^{-2z} \cdot (-2)}{(1 + e^{-2z})^2} - 0$$

$$= \frac{4 \cdot e^{-2z}}{(1 + e^{-2z})^2}$$

b) *Sketch the gradient of the tanh, ReLU functions and also the sigmoid you sketched in the Exercise 6.2, all in the same graph (please indicate ticks on the axes). Now explain what do you observe from this graph (i.e what are the differences between activation functions' gradient, which one would be more suitable for back-propagation and which one would create more problems and why)?*



⟹ With ReLU the gradient has no relationship with x, which means that the changes made in back propagation are constant and do not depend on input x.

⇒ Sigmoid curve is very flat, so the significant changes cannot be made. This means network refuses to learn further.

Exercise 8.3. – *Lagrange Multiplier*

Lagrange Multiplier is a widely used method for optimizing functions under constraints. You can read more about it on https://en.wikipedia.org/wiki/Lagrange_multiplier.
Find critical points of the function $f(x, y) = x^3 + xy^2$ under the constraint $2x + y^2 = 2$ using Lagrange Multiplier. (No need to specify whether they are minimum or maximum).

We can start by representing the constraint as following:

$g(x, y) = 2x + y^2 - 2$ and $g(x, y) = 0$.

Then the Lagrangian function can be defined as:

$$L(x, y, \lambda) = x^3 + xy^2 + \lambda(2x + y^2 - 2)$$

Knowing this, we can find derivatives and calculate the gradient:

$$\nabla_{x,y,\lambda} L(x, y, \lambda) = \left[\frac{\partial L}{\partial x} \ \frac{\partial L}{\partial y} \ \frac{\partial L}{\partial \lambda}\right] = [3x^2 + y^2 + 2\lambda,\ \ 2xy + 2\lambda y,\ \ 2x + y^2 - 2]$$

Therefore:

$$\nabla_{x,y,\lambda} L(x,y,\lambda) = 0 \iff \begin{cases} 3x^2 + y^2 + 2\lambda = 0 \\ 2xy + 2\lambda y = 0 \\ 2x + y^2 - 2 = 0 \end{cases}$$

The second equation yields to $x = -\lambda$.

Then we can define $y^2 = 2 - 2x$, which yields to $3x^2 + 2 - 2x + 2\lambda = 0$

Knowing that $x = -\lambda$, we have:

$$3\lambda^2 + 2 + 2\lambda + 2\lambda = 0$$

$$3\lambda^2 + 4\lambda + 2 = 0$$

$$D = -8$$

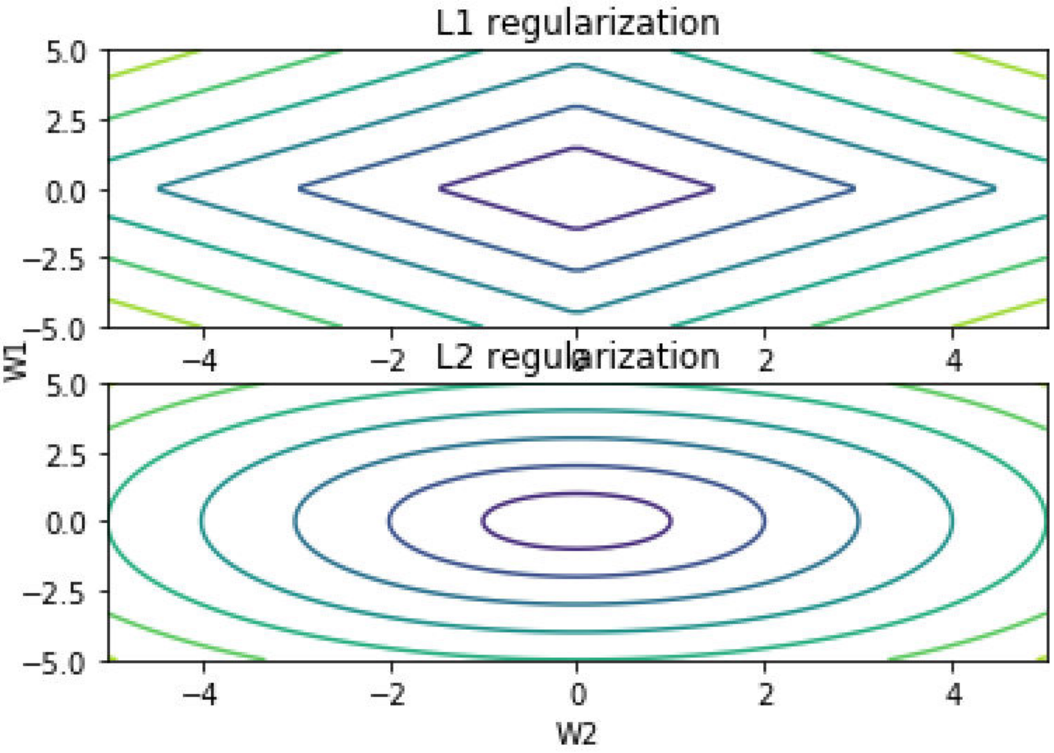
Since $D < 0$, there no real number solutions.

Exercise 8.4. – *Bridge and Lasso*

The p-Norm is defined as:

$$\|w\|_p = \left(\sum_{i=1}^{n} |\omega_i|^p\right)^{\frac{1}{p}}, where\ p \geq 1\ real\ number$$

a) Sketch a contour plot of L_p norm for $p \in \{1, 2\}$.



b) From the previous sketch, what are the main advantages and drawbacks of using lasso instead of ridge regression? (Hint: think about differentiability and the effect on weights).

According to the skech, we can see, that L1 will give sparser output to the weights (due to the fact, that some of them are set to 0). Essentially, it means that it has built the feature selection.

L1-norm, on the other hand, makes it more difficult to compute because of differentiation, since it doesn't have this analytic solution component as L1 has.

c) *In this lecture Ridge Regression was expressed as an unconstrained optimization problem:* $\min_w(\frac{1}{2}||Xw - y||_2^2 + \lambda||w||_2^2)$
Show that this is equivalent to the constrained optimization problem:$\min_w(\frac{1}{2}||Xw - y||_2^2)$*, constrained by* $||w||_2^2 \leq s$.
What is the relation between λ *and* s?

We can define the constraint function as:

$$g() = \sum \omega^2 - s$$

Then we can define the lambda-term of Lagrangian function as:

$$\lambda g = \lambda(\sum \omega^2 - s) = 0$$

Assuming that $\lambda \neq 0$ we have $\sum \omega^2 - s = 0$, which means that $\sum \omega^2 = s$.
Suppose the unconstrained problem solution to be $\sum \omega_{unconstrained}^2$.
Then we can assume that the solution for the constrained problem will contain two parts:
$\lambda_{constrained} = \lambda_{unconstrained}$

$$\sum \omega_{unconstrained}^2 = \sum \omega_{constrained}^2 = s$$

As the last step, we can derive the closed form of the ridge regression, which essentially defines the relation between s and λ:

$$s = \sum \omega^2 = \sum \left(\frac{\sum x_i y_i}{\sum x_i^2 + \lambda}\right)^2$$

According to this relation we can see that s and λ are inversely associated with each other. This concludes the prove.