Universität des Saarlandes
Prof. Dr. Dietrich Klakow
Lehrstuhl für Signalverarbeitung
NNIA Winter Term 2019/2020

# Exercise Sheet 5

Philip Georgis [s8phgeor], Pauline Sander [s8pasand], Vilem Zouhar [vizo00001]

*(Solutions)*

**Deadline: 15.12.2020, 23:59**

## Instructions

Submit the jupyter notebook with the solution for exercise 5.2 b) in an archive along with the latex file.

## Exercises

### Exercise 5.1 - Computing Jacobian and Hessian (1 + 1 = 2 points)

Let $f(x, y) = 3x^2y + 4x^3y^4 - 7x^9y^4$. Compute Jacobian and Hessian matrices of $f$.

*Solution 5.1*

$$\mathbf{J}_f = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} 6xy + 12x^2y^4 - 63x^8y^4 \\ 3x^2 + 16x^3y^3 - 28x^9y^3 \end{bmatrix}^T$$

$$\mathbf{H}_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 6y + 24xy^4 - 504x^7y^4 & 6x + 48x^2y^3 - 252x^8y^3 \\ 6x + 48x^2y^3 - 252x^8y^3 & 48x^3y^2 - 84x^9y^2 \end{bmatrix}$$

### Exercise 5.2 - Taylor Series and Newton's Method (1 + 2 + 1 + 1 = 5 points)

a) Derive the first 5 terms of the Taylor series about $x_0 = 0$ for $f(x) = cos(x)$, and write the series in sigma notation (e.g. as an infinite sum).

b) In python, apply Newton's method to find the nearest critical point of

$$f(x, y) = x^2 - y^2 + 4 - 3xy$$
from the initial point $x_0 = -0.3, y_0 = 0.3$.

After each iteration, check the value of the first derivative, i.e. Jacobian: if Jacobian is 0, then we reached the critical point.
Plot the original function for x and y in range from -0.5 to 0.5 with step size of 0.01, along with the initial point and the points computed after each iteration. Use method

.surface_plot() with parameter $alpha = 0.3$ for plotting the function and .scatter() for plotting the points.
What kind of problem of function minimization task is illustrated with this example?

c) How is Newton's Method related to gradient descent?

d) In which case is it impossible to apply Newton's method? Hint: look at the multidimensional generalization of the formula.

*Solution 5.2*

a) The Taylor series is defined as $\sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$. If $f(x) = cos(x)$ and $x_0 = 0$, the first five elements are

$$\begin{aligned} cos(x) &\approx cos(0) + -sin(0)x + \frac{1}{2!} - cos(0)x^2 + \frac{1}{3!}sin(0)x^3 + \frac{1}{4!}cos(0)x^4 \\ &= 1 + \frac{1}{2!} \cdot x^2 + \frac{1}{4!} \cdot x^4 \end{aligned} \tag{1}$$

or, written as a sum: $\sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!}$

b) In this example we end up on a saddle point, not at the minimum. We reach the critical point after only one iteration, because the function is quadratic and the parabola can be fitted onto the quadratic slope.

c) Both Newton's method and gradient descent find critical points of graphs using the first order derivative of the function. In contrast to gradient descent, Newton's method also takes the second order derivative into account and cannot distinguish between maxima and minima (it only works if the Hessian is positive definite). It therefore depends on the starting point. Newton's method can be faster than gradient descent (e.g. it found the critical point in b) after only one iteration).

d) Newtons method is computationally much more expensive because the inverse of the Hessian Metrix has to be computed every time the matrix is updated. Only minimazition problems with few parameters can be solved using Newton's method.

## Exercise 5.3 - Activation Functions (1.5 + 1 + 0.5 = 3 points)

a) Three of the most commonly-used activation functions are the sigmoid function, hyperbolic tangent, and ReLU. The equations for these are given below. Compute the first derivative of each function. Note that your final derivative for tanh should not be written in terms of other hyperbolic functions, though you may use these in your calculation. Hint: ReLU is not differentiable at x = 0. For the purposes of your derivative, you may define its derivative piecewise, ignoring this point.

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

$$ReLU(x) = \begin{cases} 0 & x < 0 \\ x & x > 0 \end{cases}$$

b) Using an online resource like Wolfram Alpha or Desmos, graph each function along with its derivative. Discuss the differences you observe. What are the advantages and disadvantages of each? In particular, think about how the range of the function and the amplitude of the derivative would affect a network.

c) Which activation function would be most appropriate for a classification problem when there are only two classes? Would adding more classes change your choice? Why or why not?

*Solution 5.3*

**a)**

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

$$tahn'(x) = \Big(\frac{e^{2x}-1}{e^{2x}+1}\cdot\frac{e^{-x}}{e^{-x}}\Big)'$$

$$= \Big(\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}\Big)' = \frac{(e^{x}-e^{-x})'\cdot(e^{x}+e^{-x}) - (e^{x}-e^{-x})\cdot(e^{x}+e^{-x})'}{(e^{x}+e^{-x})^2}$$

$$= \frac{(e^{x}+e^{-x})\cdot(e^{x}+e^{-x}) - (e^{x}+e^{-x})\cdot(e^{x}-e^{-x})}{(e^{x}-e^{-x})^2} = \frac{(e^{x}+e^{-x})^2 - (e^{x}-e^{-x})^2}{(e^{x}+e^{-x})^2}$$

$$= 1 - \frac{(e^{x}-e^{-x})^2}{(e^{x}+e^{-x})^2} \qquad \big(= 1 - tanh^2(x)\big)$$

$$ReLU'(x) = \begin{cases} 0 & x < 0 \\ 1 & x > 0 \end{cases}$$

**b)** All functions (sigmoid: Figure 1, tanh: Figure 2, ReLU: Figure 3) are non-decreasing. Positive properties of sigmoid and tanh are that their derivatives are continuous and that they are bounded. Both sigmoid and ReLU are non-negative, though this is not an unfixable property of tanh, as it can be moved upward and squashed to also span $(0, 1)$. In fact tanh is just manipulated sigmoid. Tanh and sigmoid require quite demanding computations (can be approximated by precomputed lookup tables), while ReLU is only an *if* statement.

Sigmoid is more stretched along the x axis and therefore in comparison the gradient of tanh is more concentrated around the origin. This is also adjustable by just multiplying the argument by a parameter: $\alpha x$.

Because of the almost constant derivative of ReLU, it could be difficult to navigate the function landscape with higher order approximation.

**c)** For binary classification, sigmoid or squashed tanh would make the most sense, because they would directly be able to output the probability of the first class and the probability of the other class as $1 - p$ (property of them being bounded).

For predicting multiple classes, both sigmoid and tanh would give the same argmax. This would not be the case for ReLU in case all of the feeding values would be $< 0$. For multiclass prediction something like softmax (based on sigmoid) would be the most useful.

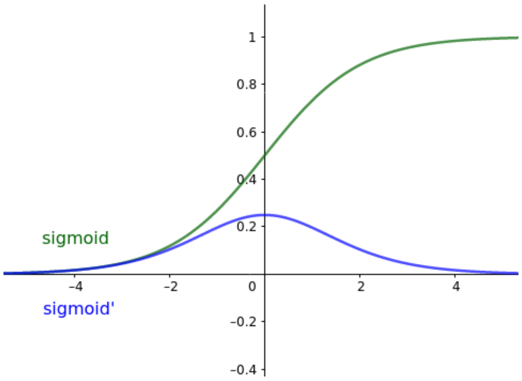

Figure 1: sigmoid function and its derivative, x:y ratio 5:1

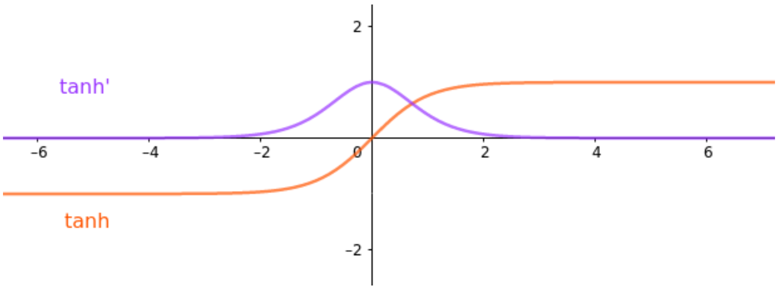

Figure 2: tanh function and its derivative, x:y ratio 1:1

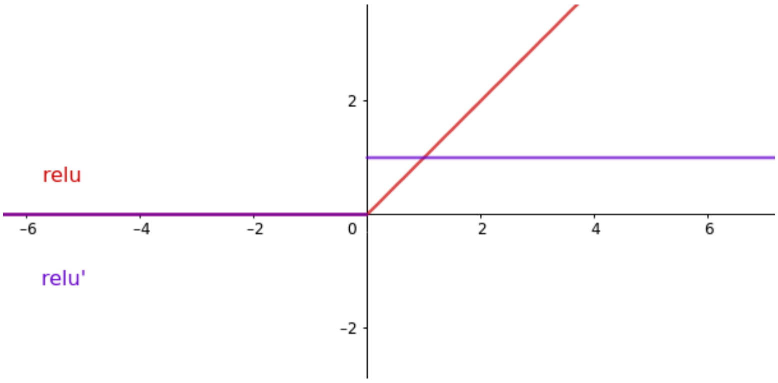

Figure 3: ReLU function and its derivative, x:y ratio 1:1

# Submission instructions

The following instructions are mandatory. If you are not following them, tutors can decide to not correct your exercise.

- You have to submit the solutions of this assignment sheet as a team of 2-3 students.
- Hand in a **single** PDF file with your solutions.
- Make sure to write the student teams ID and the name of each member of your team on your submission.
- Your assignment solution must be uploaded by only **one** of your team members to the course website.
- If you have any trouble with the submission, contact your tutor **before** the deadline.