November 24, 2025

Faculty of Computer Science, Heinrich-Heine-University Düsseldorf

Deep Learning

Winter term 25/26 – Exercise Sheet 7

Submission Deadline: Monday, December 1, 2025, 2:00 PM

1. **Newton's Method (4P)**
   The Newton's method is an optimization scheme based on using a second-order Taylor series expansion to approximate the objective function (that is the empirical risk $R_{emp}(\theta)$ in our case) near some current point (θ_0 in our case), ignoring derivatives of higher order:

$$R_{emp}(\theta) \approx R_{emp}(\theta_0) + (\theta - \theta_0)^T \nabla_\theta R_{emp}(\theta_0) + \frac{1}{2}(\theta - \theta_0)^T H_\theta (\theta - \theta_0) \ ,$$

   where H_θ is the Hessian of R_{emp} wrt θ evaluated at θ_0.

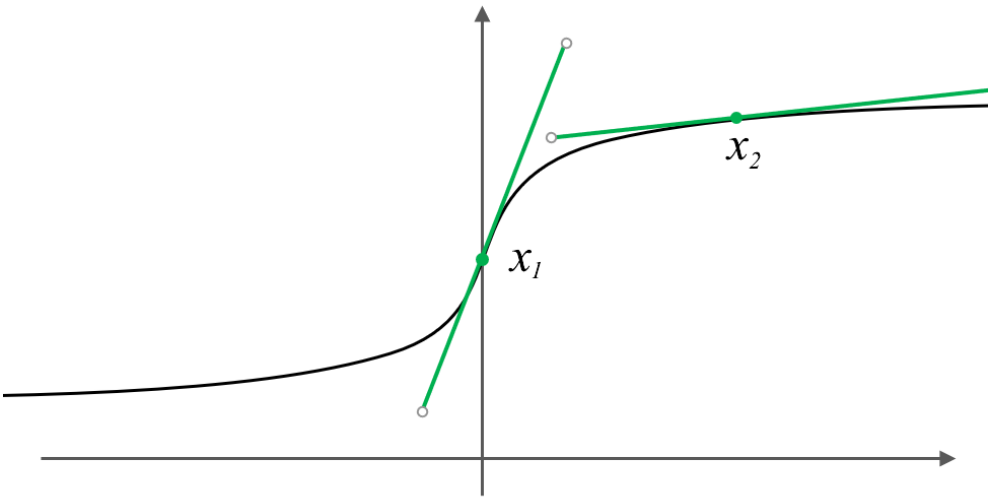
a) (1P) Solve for the critical point (corresponding to the minimum) of this function to obtain the Newton parameter update rule.

b) (2P) For a (locally) quadratic function (with positive definite H_θ) Newton's method jumps directly to the minimum. To demonstrate this apply Newton's method to the simple function $f(x) = (x-1)(x-3)$ and the current point $x_0 = 4$.

c) (1P) What makes it difficult to apply Newton's method and other second order methods to the optimization of neural networks?

2. **Vanishing Gradients (6P)**

a) (2P) Assume the following image to be the graph of a sigmoid function. Explain how the problem of vanishing gradients is exemplified by the slope of the graph in the points x_1 and x_2.



b) (4P) Assume we have a neural network with one input neuron, one hidden neuron and one output neuron, with the loss function given by the squared loss (the squared loss is normalized by 1/2, so its derivative does not have additional constant). The weights of the connections are $\theta_1 = 4$ and $\theta_2 = 1.5$, respectively, and the biases set to zero. Consider the training example $(x, y) = (1.5, 1)$. Calculate the forward pass (using a computer/calculator and giving results up to the fourth decimal position) and then calculate the gradient of the loss function

i) (2P) given that the activation functions in the hidden and output layer are sigmoids,

ii) (2P) given that the activation functions in the hidden and output layer are ReLUs.

Interpret the results in the light of part a). How does using ReLUs resolve the problem of vanishing gradients? Can you explain geometrically, why the same problem occurs when using the hyperbolic tangent activation function?