Nonlinear Optimisation and Learning (202300028)

Exam December 21, 2023, 8:45 - 10:45 (two hours).

No additional materials may be used during this exam (no notes, calculators, etc.). In your proofs, you may use definitions and all theorems, lemmas, corollaries, propositions and conclusions drawn in the text of the reader with a reference like "We know that..." (make clear that you use a result from the reader). You may not use results from exercises. You may take this exam after handing in your work.

This exam has 6 exercises.

1. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(\mathbf{x}) = 100\,(x_2 - x_1^2)^2 + (1 - x_1)^2\,.$$

   (a) Compute the gradient vector and the Hessian matrix of f.

   (b) Find the critical point(s) of f, and investigate the nature of each critical point.

   (c) Is f a convex function? Motivate your answer.

   (d) Does f have a global minimiser? Motivate your answer.

2. Argue that in the Gradient Descent Method for minimising f, if the line search problem is solved exactly in each step, we have $\nabla f(\mathbf{x}^{(k+1)})\mathbf{d}^{(k)} = 0$. Here $\mathbf{d}^{(k)}$ denotes the direction of the k-th step of the Gradient Descent Method.

3. We consider the nonlinear least square problem where we want to fit the function

$$g(x_1, x_2, t) = x_1 t \cos(x_2 t)$$

   to the data points $(t_1, y_1) = (1, -\frac{1}{2})$ and $(t_2, y_2) = (\frac{1}{2}, 16)$.

   (a) Write down the function that we want to minimise.

   (b) Compute the direction of the first step of the Gauss-Newton Method starting with $(x_1^{(0)}, x_2^{(0)}) = (1, \pi)$.

4. Pick a suitable series of step sizes $\eta^{(k)}$ (argue why your pick is suitable) and apply one step of the subgradient method to

$$f(\mathbf{x}) = |x_1| + |x_2|$$

   starting at $(x_1^{(0)}, x_2^{(0)}) = (0, -\frac{1}{\sqrt{2}})$.

5. Figure 1 shows a perceptron.

   a What are the parameters of this network? For each layer, write down the weights as a matrix of the appropriate size and the biases as a vector of the appropriate size.

   b Assume that each layer, including the output layer, has a step function as activation function. Suppose x_1 represents whether there is a party at the Vestingbar, x_2 represents whether the officer internal affairs is present in the Vestingbar, and the output represents whether you are going to the Vestingbar. Is this a good model for you not going to the Vestingbar when there is a party in the Vestingbar but the officer internal affairs is not there? Briefly explain your answer.

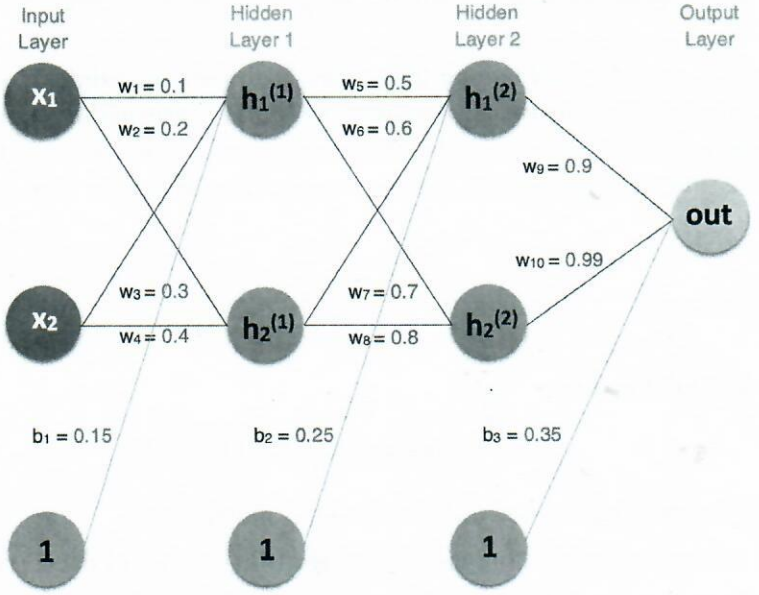

Figure 1: A neural network

6. Consider the function f_θ given by

$$f_\theta(x) = w_2\,\mathrm{ReLU}(w_1 x + b_1) + b_2\,,$$

where $\theta = (w_1, w_2, b_1, b_2)$. We want to use gradient descent to fit the parameters. Our dataset consists of the single data point $(x, y) = (1, -1)$.

(a) Write down the MSE loss function for this problem.

(b) Sketch the computational graph for the loss function.

(c) Compute the gradient of the loss function w.r.t. the parameters using backpropagation through the computational graph. Give the answer in terms of arbitrary parameter values for (w_1, w_2, b_1, b_2). Your answer may contain ReLU and Step.

You can use that

$$\mathrm{ReLU}(x) = \max(0, x)\,.$$

and the derivative of ReLU is

$$\frac{d\,\mathrm{ReLU}}{dx}(x) = \mathrm{Step}(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0. \end{cases}.$$

Points: 90 + 10 = 100

1.	(a)	:	5 pt.	2.		:	7 pt.	5.	(a)	:	4 pt.
	(b)	:	5 pt.	3.	(a)	:	5 pt.		(b)	:	6 pt.
	(c)	:	8 pt.		(b)	:	10 pt.	6.	(a)	:	5 pt.
	(d)	:	5 pt.	4.		:	10 pt.		(b)	:	10 pt.
									(c)	:	10 pt.