Nonlinear Optimisation and Learning

Exam December 18, 2025 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, lemmata, corollaries and propositions from 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 7 exercises.

$x^2 - 2xy + y^2 + y^2 - 2yz + z^2 + z^2 + 2z + 1$

1. Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined as $f(x, y, z) = (x - y)^2 + (y - z)^2 + (z + 1)^2$. Suppose we want to do two steps of Conjugate Gradient Descent with Exact Line Search for f, starting from the point $(0, 0, 0)$.

    (a) What is the gradient function of f? What is the gradient of f at $(0, 0, 0)$?

    (b) Apply the first step of CGD to f, starting from $(0, 0, 0)$. What is the descent direction $\boldsymbol{u}^{(0)}$ for this step? What is the point $\boldsymbol{x}^{(1)}$ reached after this step?

    (c) What is the gradient $\boldsymbol{g}$ of f at the point $\boldsymbol{x}^{(1)}$?

    (d) The second step of CGD uses a descent direction of the form $\boldsymbol{u} = -\boldsymbol{g} + \alpha \boldsymbol{u}^{(0)}$. Using the fact that $\boldsymbol{u}$ and $\boldsymbol{u}^{(0)}$ are conjugate directions, derive the correct value of α that we should use.

    (e) How many steps will Newton Descent need to find a local minimizer of f? Justify your answer.

2. Let $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = x^2 + y^3 - x^2 y$.

    (a) Determine the critical points of f on $\mathbb{R}^2$ (i.e. points for which gradient vanishes).

    (b) Which of these points are local minimizers and why?

    (c) Suppose we want to execute one step of Newton Descent starting at the point $(0, 0)$. Is that possible? If not, what is the typical method to handle such a problem in practice?

3. Consider the function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{x} + 42 x_1 \,.$$

    Prove that the Gradient Descent method with exact line minimisation finds the minimiser of f in at most n steps independent of the starting point $\mathbf{x}^{(0)}$.

4. Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ be two convex functions. Define $h : \mathbb{R}^n \to \mathbb{R}$ as

$$h(\boldsymbol{x}) = \max\{f(\boldsymbol{x}), g(\boldsymbol{x})\}$$

    Show that h is also convex.

5. Consider a neural network with widths $[2, 3, 2, 4]$ and ReLU as activation function for a regression task.
    (a) Visually represent this neural network both drawing styles.
    (b) Briefly explain the use-cases for both drawing styles.
    (c) Is this a deep neural network? Briefly explain why.

6. Hyperparameter optimization is an important aspect of finding the best neural network given a task.
    (a) What are hyperparameters?
    (b) We discussed 3 ways for doing hyperparameter optimization. Name them, and briefly describe how they work.

7. Consider the function $f_{\boldsymbol{\theta}}$ given by

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = w_1 x_1 + x_3 e^{w_2 x_2 + b}.$$

We want to use gradient descent to fit the parameters. Our dataset consists of a single data point $(\mathbf{x}, y) = ((-1, 1, 1), -8)$.

    (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. Your final answer should be expressed in terms of w_1, w_2, b only.

Points: 90 + 10 = 100

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