Foundations of Deep Learning **Fall 2024**

Homework 1: Basics, Approximation theory and Complexity

Lecturer: Aurelien Lucchi

The points of the best-two-out-of-three homeworks, including this one, will be contributed to the final score. The points of each problem in this exercise sheet are equally weighted. Period: 19 September 2024 18:00 - 24 October 2024 23:55 (Bern time).

Problem 1 (Neural Networks) (5 Points):
Consider a shallow neural network (NN) structure as follows:

- The input is from $\mathbb{R}^d$, and the output is real-valued;
- the first layer consists of m neurons with a continuously differentiable activation function $\sigma : \mathbb{R} \to \mathbb{R}$; [1]
- the second layer is just an affine transformation from $\mathbb{R}^m$ to $\mathbb{R}$;

In other words, let $\mathbf{x} \in \mathbb{R}^d$ be an input, $m \in \mathbb{N}$ the number of (hidden) neurons, $\sigma : \mathbb{R} \to \mathbb{R}$, $\mathbf{W} \in \mathbb{R}^{m\times d}$ the weight matrix and $\mathbf{b} \in \mathbb{R}^m$ the bias in the first layer, $\mathbf{a} \in \mathbb{R}^m$ the weight vector in the second layer, we have the neural network $f : \mathbb{R}^d \to \mathbb{R}$ defined by:

$$f(\mathbf{x}; \mathbf{W}, \mathbf{b}, \mathbf{a}) = \sum_{r=1}^{m} a_r \sigma(\mathbf{W}_{r,\bullet}\mathbf{x} + b_r) = \mathbf{a}^\top \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{1}$$

for all $x \in [0, 1]$.

a) Compute the derivative $\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}; \mathbf{W}, \mathbf{b}, \mathbf{a})$ of the neural network.

b) Compute the partial derivative $\frac{\partial}{\partial \mathbf{W}} f(\mathbf{x}; \mathbf{W}, \mathbf{b}, \mathbf{a})$ of the neural network.

c) Compute the partial derivative $\frac{\partial}{\partial \mathbf{b}} f(\mathbf{x}; \mathbf{W}, \mathbf{b}, \mathbf{a})$ of the neural network.

d) Compute the partial derivative $\frac{\partial}{\partial \mathbf{a}} f(\mathbf{x}; \mathbf{W}, \mathbf{b}, \mathbf{a})$ of the neural network.

Problem 2 (Deep ReLU network and parity function) (10 Points):
Consider the parity function $\chi : \{0, 1\}^n \to \{0, 1\}$, which maps the to 0 or 1 depending on whether there is an even or uneven number of 1s. For instance for $n = 5$, $\chi([0, 1, 0, 1, 1]^\top) = 1$ and $\chi([1, 0, 0, 1, 0]) = 0$.
In this problem we want to construct a deep network with ReLU activations that exactly expresses the parity function with $O(n)$ hidden units.

a) Consider first the case of $n = 2$. Construct a ReLU network which takes inputs $\mathbf{x} \in \{[0, 0], [0, 1], [1, 0], [1, 1]\}$ and gives the correct parity $y \in \{0, 1\}$. What are the weight matrices? Can you extend your network to the case of $n = 3$ and $n = 4$?

b) Now consider the general case of some arbitrary n. Show that the deep ReLU network only consists of $O(n)$ hidden units.

Problem 3 (Complexity) (10 Points):
For an integer $m \in \mathbb{N}$ and the ReLU-activation function $\sigma : \mathbb{R} \to \mathbb{R}$, denote the function class of the shallow NN with the hidden σ-layer of width m:

$$\mathcal{F}_\sigma^m = \{f(\cdot; \mathbf{w}, \mathbf{b}, \mathbf{a}) : [0, 1] \to \mathbb{R} \mid \mathbf{w}, \mathbf{b}, \mathbf{a} \in \mathbb{R}^m\} \tag{2}$$

a) What is the minimum number m such that the function class $\mathcal{F}_\sigma^m$ can interpolate the following dataset? i.e. there exist some $\mathbf{w}, \mathbf{b}, \mathbf{a} \in \mathbb{R}^m$ such that $f(x_i; \mathbf{w}, \mathbf{b}, \mathbf{a}) = y_i,\ \forall i = 1, \ldots, n$. Give an explicit $f \in \mathcal{F}_\sigma^m$ for each interpolation.

   i) $\{(x_i, y_i)\}_{i=1}^3 = \{(0, 0), (\frac{1}{2}, 0), (1, \frac{1}{2})\}$.

   ii) $\{(x_i, y_i)\}_{i=1}^3 = \{(0, 0), (\frac{1}{2}, \frac{1}{2}), (1, \frac{3}{4})\}$

[1] With abuse of notation, when we write $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ as a multivariate function, we mean the entry-wise evaluation of σ.

b) Let $\mathcal{PL}^m \subset C[0,1]$ be the function class consisting of all piecewise-linear continuous functions $g : [0,1] \to \mathbb{R}$ with $g(0) = 0$ and with less than or equal to m affine pieces. More precisely, we have $f \in \mathcal{PL}^m$ if and only if there exists a partition of the interval $[0,1]$, $0 = z_1 < z_2 < \ldots < z_{m-1} < z_m = 1$, such that f can be expressed as $f(x) = f_i(x) := m_i x + b_i$ for $x \in [z_{i-1}, z_i]$, for some constants $m_i, b_i \in R$, for all $i = 1, \ldots, m$ and additionally $f_i(z_i) = f_{i+1}(z_i)$ for $i = 1, \ldots, m-1$.

i) Show that $\mathcal{PL}^1 \subset \mathcal{F}^1_\sigma$.

ii) For $m > 1$, let $g \in \mathcal{PL}^m$ be a piecewise-linear continuous function with $g(0) = 0$ and with exactly m affine pieces: say we have $0 = x_0 < x_1 < x_2 < ... < x_{m-1} < x_m = 1$ and g is linear on each interval $[x_{r-1}, x_r]$, for $r = 1, ..., m$. Show that there exists a function $\hat{g} \in \mathcal{PL}^{m-1}$ with at most $m-1$ affine pieces such that $g = \hat{g}$ on $[x_0, x_{m-1}]$.

iii) Show that, by induction on m, $\mathcal{PL}^m \subset \mathcal{F}^m_\sigma$.

Problem 4 (Fundamentals of Unconstrained Optimization) (5 Points):

a) Define the Dixon-Price function $R : \mathbb{R}^2 \to \mathbb{R}$ to be

$$R(x_1, x_2) := (x_1 - 1)^2 + 2(2x_2^2 - x_1)^2$$

Find all critical points by calculating its gradient $\nabla R(x_1, x_2)$ and show that $\mathbf{x}_* = \left(1, \sqrt{\frac{1}{2}}\right)^\top$ is the global minimizer of this function. Show that the Hessian matrix $\nabla^2 R(x_1, x_2)$ at the global minimizer is indeed positive definite.

b) Show that the function $f(x_1, x_2) = 8x_1 + 12x_2 + x_1^2 - 2x_2^2$ has only one stationary point, and that it is neither a maximum or minimum, but a saddle point. Sketch the contour lines of f.