Assignment 2

Due March 21, beginning of class

INSTRUCTIONS. Apart from correctness, clarity and conciseness of your solutions are important. Please write your solutions legibly. You may collaborate in groups of two or three, but you must write your solution on your own and should list the name of the persons with whom you collaborated. If you need to, you are welcome to come to my office hours to discuss the problems; I may give you some hints. Solutions to a number of problems here can be found in books, research papers, or elsewhere on the web. Please do not refer to these to solve the problems. If you happen to have seen a solution please mention this in your work. If you think there is an error in some problem or need a clarification, please let us know.

Problem 1 (10 pts)**.** *Consider neural networks with a polynomial activation function of a fixed degree d. More precisely, the activation function is of the form $x \mapsto a_d x^d + a_{d-1} x^{d-1} + \ldots + a_0$ for constants $a_d, \ldots, a_0 \in \mathbb{R}$. Suppose that we have such a network with L hidden layers and with output dimension 1. Prove that the output computed by such a neural network is also a polynomial. What is the maximum possible degree of this polynomial in terms of L and d? For the special case when $d = 1$, such networks are also known as linear networks. Work out the degree of the neural network in this case separately in terms of L.*

Describe a classification problem for which linear networks cannot achieve zero training loss. By the description of a classification problem we mean describe the dataset $((x_1, y_1), \ldots, (x_n, y_n)$.

Problem 2 (15 pts)**.** *We saw formulas for the first and second derivatives of the cross-entropy loss in the logistic regression problem. For one hidden layer networks with ReLU activations and cross-entropy loss prove formula (1) in <* `https://arxiv.org/pdf/1808.01204.pdf` *> for the first derivative. (Please read the setting in the paper carefully—it's assumed there that the weights of the second layer (the one closer to the output) are fixed and the derivative is only w.r.t. the layer that's trained.)*

Prove or disprove: the optimization problem for one hidden layer networks with ReLU activations and cross-entropy loss is a non-convex problem.

Suppose that the activation is not assumed to be ReLU but is some unspecified function ρ differentiable almost everywhere. How will the above formula change?

Problem 3. *Wasserstein distance $W_2(P,Q)$ between two distributions P and Q on $\mathbb{R}^d$ is defined as $W_2(\mu,\nu) = \inf \mathrm{E}\left[\|X-Y\|^2\right]^{1/2}$, where the infimum is over all random $(X,Y) \in \mathbb{R}^d \times \mathbb{R}^d$ with $X \sim P$ and $Y \sim Q$. Consider the case when the distributions μ and ν are two standard normal distributions on $\mathbb{R}$, that is, $P = \mathcal{N}\left(\mu_1, \sigma_1^2\right)$ and $Q = \mathcal{N}\left(\mu_2, \sigma_2^2\right)$. Prove that $W_2(P,Q)^2 = (\mu_1-\mu_2)^2 + (\sigma_1-\sigma_2)^2$.*

Problem 4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function which is l-strongly convex. Prove that*

$$f(Y) \geq f(X) + (Y-X)^T \nabla f(X) + \frac{l}{2}\|X-Y\|^2 \qquad \forall X, Y \in \mathbb{R}^n$$

Problem 5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function. In class we defined f to be L-smooth if $\|\nabla f(X) - \nabla f(Y)\| \leq L\|X-Y\|$, $\forall X, Y \in \mathbb{R}^n$. Prove that for L-smooth functions*

$$-LI \preceq \nabla^2 f(X) \preceq LI \qquad \forall X \in \mathbb{R}^n.$$