Foundations of Deep Learning **Fall 2024**

Exercise 3: Complexity Theory

Lecturer: Aurelien Lucchi

Problem 1 (Barron norm of Gaussian measure):

Consider the density function of a Gaussian random variable, $f(x) = (2\pi\sigma^2)^{d/2}\exp(-\frac{\|x\|^2}{2\sigma^2})$. We will calculate the complexity measure $\int \left\|\widehat{\nabla f}(\boldsymbol{\omega})\right\| d\boldsymbol{\omega}$ that appears in Barron's theorem discussed in class.

a) Use the following facts: (Why are they true?)

- For f defined as above, we have $|\hat{f}(\boldsymbol{\omega})| = \exp(-2\pi^2\sigma^2\|\boldsymbol{\omega}\|^2)$;
- (Cauchy-Schwartz inequality in expected value) $\mathbb{E}[\|X\|] \leq \left(\mathbb{E}\left[\|X\|^2\right]\right)^{\frac{1}{2}}$ where X is a random vector.

Show that $\int \|\boldsymbol{\omega}\||\hat{f}(\boldsymbol{\omega})|d\boldsymbol{\omega} \leq Z\left(\int \|\boldsymbol{\omega}\|^2 Z^{-1}\hat{f}(\boldsymbol{\omega})d\boldsymbol{\omega}\right)^{1/2}$ where $Z = (2\pi\sigma^2)^{-d/2}$.

b) Using the fact that $\int g(\boldsymbol{\omega})Z^{-1}\hat{f}(\boldsymbol{\omega})d\boldsymbol{\omega}$ is the expectation of the function $g(\boldsymbol{\omega})$ with respect to the density $\mathcal{N}\left(0, \frac{1}{4\pi^2\sigma^2}\right)$, show that $\int \|\boldsymbol{\omega}\||\hat{f}(\boldsymbol{\omega})|d\boldsymbol{\omega} \leq \sqrt{\frac{d}{4\pi^2\sigma^2}} \cdot (2\pi\sigma^2)^{-d/2}$.

c) Since $\int \left\|\widehat{\nabla f}(\boldsymbol{\omega})\right\| d\boldsymbol{\omega} = 2\pi \int \|\boldsymbol{\omega}\| \cdot |\hat{f}(\boldsymbol{\omega})|d\boldsymbol{\omega}$, what do you conclude about the complexity measure $\int \left\|\widehat{\nabla f}(\boldsymbol{\omega})\right\| d\boldsymbol{\omega}$ when d is very large?

Problem 2 (Maurey's lemma):

Let $X = \mathbb{E}V$ be given, with V a random vector supported on a subset S of the event space, and let $V_1, \dots V_k$ be i.i.d. realization of V.

a) Show that

$$\mathbb{E}_{V_i}\left\|X - \frac{1}{k}\sum_{i=1}^{k} V_i\right\|^2 = \frac{1}{k}\mathbb{E}_V\|V - X\|^2.$$

b) Show that

$$\frac{1}{k}\mathbb{E}_V\|V - X\|^2 \leq \frac{\sup_{U\in S}\|U\|^2}{k}.$$

Problem 3 (Number of affine pieces in a ReLU network):

We will prove the lemma that bounds the number of affine pieces in a ReLU network stated in the lecture notes. Denote by $\delta_A(f)$ be the (minimum) number of affine pieces of a piece-wisely linear continuous function $f : \mathbb{R} \to \mathbb{R}$. Show that

a) $\delta_A(\sum_k a_k f_k + b_k) \leq \sum_k \delta_A(f_k)$, for any finite sequence of piece-wisely linear continuous functions f_k, and for any real sequences a_k and b_k.

b) Let $g_i : \mathbb{R} \to \mathbb{R}$ denote the output of some node in layer i in a ReLU network with L layers of widths $(m_1, \dots, m_L)$ as a function of the input. Using induction on i, show that the number of affine pieces $\delta_A(g_i)$ satisfies

$$\delta_A(g_i) \leq 2^i \prod_{j<i} m_j.$$

c) Recall that $g_{L+1} : \mathbb{R} \to \mathbb{R}$ is a ReLU network with L layers of widths $(m_1, \dots, m_L)$ such that $m = \sum_{i=1}^{L} m_i$. The number of affine pieces in g_{L+1} satisfies

$$\delta_A(g_{L+1}) \leq \left(\frac{2m}{L}\right)^L.$$

d) What value of L maximizes the upper bound $\left(\frac{2m}{L}\right)^L$? Although taking this value of L gives a large value of $\delta_A(g_{L+1})$, give a reason why in practice we usually pick a value much smaller than this.