Foundations of Deep Learning **2022–2023 Spring Semester**

Assignment 2: Expressiveness

Due: May 21, 2023

Important Guidelines

1. Work can be done in groups of up to three students..
2. Solutions must be submitted in digital format (PDF file). It is highly recommended to typeset the solutions in LaTeX or Word. If you choose to submit scanned handwritten solutions, please make sure they are clearly written and the scan is of high quality.
3. Please include your ID numbers in the PDF report.

Part 1 - Boolean AND-OR Networks

1. **(12 points)** In class we established expressive efficiency by showing that the XOR_d function is realizable by a deep network of width $\bar{B} = d$, but requires width $B \geq 2^{d-1}$ in order to be realized by the shallow network. Is it possible to define a function that provides a stronger result, in the sense that the lower bound on B is larger than 2^{d-1}? Prove your answer.
2. **(13 points)** Prove that with polynomial width $\bar{B}$, the deep network cannot realize all possible functions. In particular derive an exponential (in d) lower bound on $\bar{B}$ required in order for its hypotheses space to be equal to $\mathcal{Y}^{\mathcal{X}}$.

Part 2 - Fully Connected ReLU Networks with 1D Input

1. **(13 points)** Prove the following proposition from class, and explain why the condition $B \geq 2$ there is needed.
   Proposition: A shallow network of width $B \geq 2$ can realize any piecewise linear mapping with $\leq B$ pieces. Conversely, any mapping realizable by such network is piecewise linear with $\leq B + 1$ pieces.
2. **(12 points)** In the context of the expressive efficiency with inapproximability analysis delivered in class, derive a lower bound on the number of intervals in $S_>$ and $S_<$ that a piecewise linear mapping with $\leq B + 1$ pieces must miss.
3. **(13 points)** Modify the universality and expressive efficiency analyses given in class so that they apply to leaky ReLU activation — $\sigma(z) = \max\{az, z\}$ for some $a \in (0, 1)$ — instead of ReLU.
4. **(Bonus 10 points)** Prove that the shallow and deep networks are $\mathcal{F}$ -universal in the sense of $d(\cdot, \cdot)$, where $\mathcal{F} \subseteq \mathbb{R}^{\mathbb{R}}$ is the set of Riemann integrable functions and $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\geq 0}$ is defined by:
   $$d(f_1, f_2) := \int_0^1 |f_1(x) - f_2(x)| \, dx$$
   (in class we saw a proof for the more restricted case where $\mathcal{F}$ is the set of continuous functions).

Part 3 - Convolutional Arithmetic Circuits

1. **(5 points)** Prove the following proposition from class.
Proposition: Consider the setting and notation of the "Expressiveness 2" lecture notes. The tensor (function) generated by the deep convolutional arithmetic circuit described in class is given by:

$$\begin{aligned}
\phi^{1,j,\gamma} &= \sum\nolimits_{\alpha=1}^{r_0} a_\alpha^{1,j,\gamma} \cdot \underline{a}^{0,2j-1,\alpha} \otimes \underline{a}^{0,2j,\alpha} \quad , \; j = 1, 2, \ldots, \frac{N}{2} \; , \; \gamma = 1, 2, \ldots, r_1 \\
&\vdots \\
\phi^{l,j,\gamma} &= \sum\nolimits_{\alpha=1}^{r_{l-1}} a_\alpha^{l,j,\gamma} \cdot \underbrace{\phi^{l-1,2j-1,\alpha}}_{\text{order } 2^{l-1}} \otimes \underbrace{\phi^{l-1,2j,\alpha}}_{\text{order } 2^{l-1}} \quad , \; j = 1, 2, \ldots, \frac{N}{2^l} \; , \; \gamma = 1, 2, \ldots, r_l \\
&\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad , \\
\phi^{L-1,j,\gamma} &= \sum\nolimits_{\alpha=1}^{r_{L-2}} a_\alpha^{L-1,j,\gamma} \cdot \underbrace{\phi^{L-2,2j-1,\alpha}}_{\text{order } \frac{N}{4}} \otimes \underbrace{\phi^{L-2,2j,\alpha}}_{\text{order } \frac{N}{4}} \quad , \; j = 1, 2 \; , \; \gamma = 1, 2, \ldots, r_{L-2} \\
\mathcal{A} &= \sum\nolimits_{\alpha=1}^{r_{L-1}} a_\alpha^{L} \cdot \underbrace{\phi^{L-1,1,\alpha}}_{\text{order } \frac{N}{2}} \otimes \underbrace{\phi^{L-1,2,\alpha}}_{\text{order } \frac{N}{2}}
\end{aligned}$$

where $\otimes$ stands for outer product.

Important: A proof of this proposition appears in the class notes. Feel free to consult it. The goal is for you to rewrite it in your own words so you better understand the equivalence between Hierarchical Tucker decomposition and the deep convolutional arithmetic circuit described in class.

2. **(10 points)** Prove the following lemma from class.
Lemma (*"matricization of outer product = Kronecker product of matricizations"*): Let T and $\bar{T}$ be tensors of orders n and $\bar{n}$ respectively. Let $I \subset [n + \bar{n}]$ and denote by $I - n$ the set obtained by subtracting n from each element of I. Then:

$$[\![T \otimes \bar{T}]\!]_I = [\![T]\!]_{I \cap [n]} \odot [\![\bar{T}]\!]_{I - n \cap [\bar{n}]}$$

3. **(10 points)** Modify the expressive efficiency analyses given in class to the case in which the deep network has $\log_4 N$ hidden layers (assume N is a power of 4), with pooling windows of size 4 (in class we treated $\log_2 N$ hidden layers with pooling windows of size 2).

4. **(12 points)** Consider the following "quadrant" partition of the N input elements:

$$I_{quad} = \left\{1, 2, \ldots, \frac{1}{4}N, \frac{1}{2}N + 1, \frac{1}{2}N + 2, \ldots, \frac{3}{4}N\right\}$$

$$I_{quad}^C = \left\{\frac{1}{4}N + 1, \frac{1}{4}N + 2, \ldots, \frac{1}{2}N, \frac{3}{4}N + 1, \frac{3}{4}N + 2, \ldots, N\right\}$$

Prove that under this partition, the separation rank of a function realized by the deep network (with $L = \log_2 N$ hidden layers) is no greater than $r_{L-1} \cdot r_{L-2}^2$, where r_l stands for the width of hidden layer l.