ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Prof. Dr. Helmut Bölcskei
Chair for Mathematical Information Science

Exam on Neural Network Theory
August 29, 2022

Please note:

- Exam duration: 180 minutes
- Maximum number of points: 100
- You are not allowed to use any printed or handwritten material (i.e., books, lecture and discussion session notes, summaries), computers, tablets, smart phones or other electronic devices.
- Your solutions should be explained in detail and your handwriting needs to be clean and legible.
- Please do not use red or green pens. You may use pencils.
- Please note that the "ETH Zurich Ordinance on Disciplinary Measures" applies.

Before you start:

1. The problem statements consist of 5 pages including this page. Please verify that you have received all 5 pages.
2. Please fill in your name, student ID card number and sign below.
3. Please place your student ID card at the front of your desk so we can verify your identity.

During the exam:

4. For your solutions, please use only the empty sheets provided by us. Should you need additional sheets, please let us know.
5. Each problem consists of several subproblems. If you do not provide the solution to a subproblem, you may, whenever applicable, nonetheless assume its conclusion in the ensuing subproblems.
6. All results in the Handout can be used without proof.

After the exam:

7. Please write your name on every solution sheet and prepare all sheets in a pile. All sheets, including those containing problem statements, must be handed in.
8. Please clean up your desk and remain seated and silent until you are allowed to leave the room in a staggered manner row by row.
9. Please avoid crowding and leave the building by the most direct route.

Family name: .... First name: ....

Student ID card No.: ....

Signature: ....

Problem 1 (25 points)

Let $f_1\colon [0,1] \to [0,1]$ be given by

$$f_1(x) := \begin{cases} 0, & x \in [0, \frac{1}{4}] \\ 2(x - \frac{1}{4}), & x \in (\frac{1}{4}, \frac{3}{4}) \\ 1, & x \in [\frac{3}{4}, 1] \end{cases},$$

and, for $n \geq 2$, $n \in \mathbb{N}$, let $f_n := f_1 \circ f_{n-1}$.

(a) (3 points) Find a ReLU neural network Φ_1 satisfying $\Phi_1(x) = f_1(x)$, for all $x \in [0,1]$, and specify $\mathcal{L}(\Phi_1)$, $\mathcal{W}(\Phi_1)$, $\mathcal{M}(\Phi_1)$, and $\mathcal{B}(\Phi_1)$.

(b) (5 points) Find a ReLU neural network Φ_3 satisfying $\Phi_3(x) = f_3(x)$, for all $x \in [0,1]$, with $\mathcal{B}(\Phi_3) \leq 2$ and specify $\mathcal{L}(\Phi_3)$, $\mathcal{W}(\Phi_3)$, and $\mathcal{M}(\Phi_3)$.

(c) (6 points) Show that, for $n \in \mathbb{N}$ and $x \in [0,1]$,

$$f_n(x) = \begin{cases} 0, & x \in [0, \frac{1}{2} - 2^{-(n+1)}] \\ 2^n(x - (\frac{1}{2} - 2^{-(n+1)})), & x \in [\frac{1}{2} - 2^{-(n+1)}, \frac{1}{2} + 2^{-(n+1)}] \\ 1, & x \in [\frac{1}{2} + 2^{-(n+1)}, 1] \end{cases}.$$

(d) (5 points) Let $H\colon [0,1] \to [0,1]$ be given by

$$H(x) := \begin{cases} 0, & x \in [0, \frac{1}{2}] \\ 1, & x \in (\frac{1}{2}, 1] \end{cases}.$$

Show that, for every $\varepsilon \in (0, \frac{1}{2})$, there exists a ReLU neural network Ψ_ε satisfying

$$\|H - \Psi_\varepsilon\|_{L^2([0,1])} \leq \varepsilon.$$

(e) (2 points) Show that $\rho(x) + \rho(-x) = |x|$, for $x \in \mathbb{R}$, where $\rho(x) := \max\{0, x\}$ is the ReLU activation function.

(f) (4 points) Let $d \in \mathbb{N}$. Realize $g\colon \mathbb{R}^d \to \mathbb{R}, x \mapsto \|x\|_1$ as a ReLU neural network Γ and specify $\mathcal{L}(\Gamma)$, $\mathcal{W}(\Gamma)$, $\mathcal{M}(\Gamma)$, and $\mathcal{B}(\Gamma)$.

Problem 2 (25 points)

For $n, d \in \mathbb{N}$, let $\mathcal{C}_{n,d} \subset L^\infty(\mathbb{R}^d)$ be given by

$$\mathcal{C}_{n,d} := \{\mathbb{I}_k \colon k \in \{0, \dots, n-1\}^d\},$$

where, for $k \in \{0, \dots, n-1\}^d$, we denote the indicator function of the d-dimensional cube $\bigtimes_{j=1}^d [k_j, k_j + 1) \subseteq \mathbb{R}^d$ by

$$\mathbb{I}_k(x) := \begin{cases} 1, & x \in \bigtimes_{j=1}^d [k_j, k_j + 1) \\ 0, & \text{else} \end{cases}.$$

We consider covering numbers and packing numbers with respect to the metric

$$\rho_\infty(f, g) := \|f - g\|_{L^\infty(\mathbb{R}^d)}.$$

(a) (2 points) Show that, for $n, d \in \mathbb{N}$ and $k, k' \in \{0, \dots, n-1\}^d$, the metric ρ_∞ satisfies

$$\rho_\infty(\mathbb{I}_k, \mathbb{I}_{k'}) = \begin{cases} 0, & k = k' \\ 1, & k \neq k' \end{cases}.$$

(b) (5 points) Show that, for $n, d \in \mathbb{N}$ and $\varepsilon \in (0, \infty)$, the ε-covering numbers of the set $\mathcal{C}_{n,d}$ with respect to the metric ρ_∞ satisfy

$$N(\varepsilon; \mathcal{C}_{n,d}, \rho_\infty) = \begin{cases} 1, & \varepsilon \geq 1 \\ n^d, & \varepsilon < 1 \end{cases}.$$

For $n, d \in \mathbb{N}$, let $\mathcal{C}^*_{n,d} \subset L^\infty(\mathbb{R}^d)$ be given by

$$\mathcal{C}^*_{n,d} := \{\alpha \mathbb{I}_k \colon k \in \{0, \dots, n-1\}^d, \alpha \in [0, 1]\}.$$

(c) (9 points) Show that there exists a constant $b \in \mathbb{R}_+$ such that, for all $n, d \in \mathbb{N}$ and $\varepsilon \in (0, \frac{1}{2})$,

$$N(\varepsilon; \mathcal{C}^*_{n,d}, \rho_\infty) \leq b\, n^d\, \varepsilon^{-1}.$$

(d) (9 points) Show that there exists a constant $a \in \mathbb{R}_+$ such that, for all $n, d \in \mathbb{N}$ and $\varepsilon \in (0, \frac{1}{2})$,

$$M(\varepsilon; \mathcal{C}^*_{n,d}, \rho_\infty) \geq a\, n^d\, \varepsilon^{-1}.$$

Problem 3 (30 points)

(a) (5 points) Let X_1 be a finite subset of $\mathbb{R}^d$, $d \in \mathbb{N}$, let $\{X_1^+, X_1^-\}$ be a dichotomy of X_1, and consider the mapping $\phi : \mathbb{R}^d \mapsto \mathbb{R}^m$, $m \in \mathbb{N}$. State the definition for the dichotomy $\{X_1^+, X_1^-\}$ to be homogeneously linearly separable and the definition for it to be ϕ-separable.

(b) (6 points) Consider the set $X_2 = \{(-1,0),(1,0),(0,1),(0,-1)\}$. Show that the dichotomy

$$\{X_2^+ = \{(-1,0),(1,0)\}, X_2^- = \{(0,1),(0,-1)\}\},$$

is not homogeneously linearly separable and find a function $\phi : \mathbb{R}^2 \mapsto \mathbb{R}$ such that $\{X_2^+, X_2^-\}$ is ϕ-separable.

(c) (6 points) Consider the class of functions

$$\mathcal{F} := \left\{ f : \mathbb{R} \mapsto \{0,1\} : f(x) = \text{sgn}(\sin(kx+b)), (k,b) \in \mathbb{R}^2 \right\},$$

where sgn : $\mathbb{R} \mapsto \{0,1\}$ is given by

$$\text{sgn}(x) := \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases}$$

Show that $\mathcal{F}$ shatters the set $\{0,1\}$.

(d) (5 points) Let $\mathcal{G}$ be a class of $\{0,1\}$-valued functions on $\mathbb{R}$. Suppose that the growth function of $\mathcal{G}$ satisfies $\Pi_{\mathcal{G}}(1) = 1$. Show that $\Pi_{\mathcal{G}}(N) = 1$, for all $N \in \mathbb{N}$.

(e) (8 points) Show that the VC dimension of the class of functions $\mathcal{F}$ from subproblem (c) satisfies

$$\text{VC}(\mathcal{F}) = \infty.$$

Problem 4 (20 points)

Consider the family of 1-Lipschitz continuous functions on $[0,1]$ given by

$$H^1([0,1]) := \{f : \mathbb{R} \mapsto \mathbb{R} : f \text{ is continuous}, |f(x) - f(y)| \leq |x - y|, \forall x, y \in [0,1]\}.$$

In this problem, we study the fundamental limit of ReLU neural network approximation of functions in $H^1([0,1])$, using a VC dimension upper bound for ReLU neural networks.

(a) (5 points) Let $p : [0,1] \mapsto \mathbb{R}$ be defined according to

$$p(x) := \begin{cases} \frac{1}{4} - x, & \text{for } x \in \left[0, \frac{1}{2}\right], \\ x - \frac{3}{4}, & \text{for } x \in \left[\frac{1}{2}, 1\right]. \end{cases}$$

Plot p and show that $p \in H^1([0,1])$.

(b) (4 points) For $n \in \mathbb{N}$ and $y = (y_0, \dots, y_n) \in \{0,1\}^{n+1}$, show the existence of a function $h_y \in H^1([0,1])$ such that

$$h_y\Big(\frac{i}{n}\Big) = \frac{2y_i - 1}{2n}, \text{ for } i = 0, \dots, n, \tag{1}$$

and

$$\text{sgn}\Big(h_y\Big(\frac{i}{n}\Big)\Big) = y_i, \text{ for } i = 0, \dots, n, \tag{2}$$

where sgn : $\mathbb{R} \mapsto \{0,1\}$ is defined as

$$\text{sgn}(x) := \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases}$$

(c) (4 points) For $n \in \mathbb{N}$ and $y = (y_0, \dots, y_n) \in \{0,1\}^{n+1}$, let $h_y \in H^1([0,1])$ be a function satisfying (1) and (2) from subproblem (b). Show that for every function $g : [0,1] \mapsto \mathbb{R}$ satisfying $\sup_{x \in [0,1]} |h_y(x) - g(x)| \leq \frac{1}{4n}$, it holds that

$$\text{sgn}\Big(g\Big(\frac{i}{n}\Big)\Big) = y_i, \text{ for } i = 0, \dots, n. \tag{3}$$

(d) (7 points) Fix $W, L \in \mathbb{N}$ with $L \geq 2$. Consider the set of ReLU neural networks

$$\mathcal{N}(W, L) = \{\Phi \in \mathcal{N}_{1,1} : \mathcal{L}(\Phi) \leq L \text{ and } \mathcal{W}(\Phi) \leq W\},$$

and define

$$\text{sgn}(\mathcal{N}(W, L)) = \{\text{sgn} \circ \Phi : \Phi \in \mathcal{N}(W, L)\}.$$

It is known from the literature that the VC dimension of the class sgn$(\mathcal{N}(W, L))$ satisfies

$$\text{VC}(\text{sgn}(\mathcal{N}(W, L))) \leq CW^2L^2(\log(W) + \log(L)), \tag{4}$$

for some constant C not depending on W, L. Show that there exists a function $h \in H^1([0,1])$ such that for all ReLU neural networks $\Phi \in \mathcal{N}(W, L)$,

$$\sup_{x \in [0,1]} |h(x) - \Phi(x)| > \frac{1}{4CW^2L^2(\log(W) + \log(L))}.$$

Hint: *Use the results from subproblems (b) and (c).*

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Prof. Dr. Helmut Bölcskei
Chair for Mathematical Information Science

Handout for Exam on Neural Network Theory
August 29, 2022

Definition 1 (Norms; Cartesian product)**.** For $n \in \mathbb{N}$, $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, we define

$$\|x\|_1 := \sum_{j=1}^{n} |x_j|.$$

For $X, Y \subseteq \mathbb{R}^d$, $f\colon X \to Y$, we define

$$\|f\|_{L^2(X)} := \left(\int_X |f(x)|^2 \mathrm{d}x \right)^{\frac{1}{2}}$$

and

$$\|f\|_{L^\infty(X)} := \sup_{x \in X} |f(x)|.$$

Definition 2 (Cartesian product)**.** For $d \in \mathbb{N}$ and sets $S_j \subseteq \mathbb{R}$, $j \in \{1, \dots, d\}$, we define their Cartesian product as

$$\bigtimes_{j=1}^{d} S_j := \{x = (x_1, \dots, x_d) \in \mathbb{R}^d \colon x_j \in S_j, \text{ for } j \in \{1, \dots, d\}\}.$$

In the case of $S_j = S$, for all $j \in \{1, \dots, d\}$, we write $S^d := \bigtimes_{j=1}^{d} S$.

Definition 3 (Covering and covering number)**.** Let $(\mathcal{X}, \rho)$ be a metric space. An ε-covering of a compact set $\mathcal{C} \subseteq \mathcal{X}$ with respect to the metric ρ is a set $\{x_1, \dots, x_N\} \subseteq \mathcal{C}$ such that for each $x \in \mathcal{C}$, there exists an $i \in \{1, \dots, N\}$ so that $\rho(x, x_i) \leq \varepsilon$. The ε-covering number $N(\varepsilon; \mathcal{C}, \rho)$ is the cardinality of the smallest ε-covering.

Definition 4 (Packing and packing number)**.** Let $(\mathcal{X}, \rho)$ be a metric space. An ε-packing of a compact set $\mathcal{C} \subseteq \mathcal{X}$ with respect to the metric ρ is a set $\{x_1, \dots, x_N\} \subseteq \mathcal{C}$ such that $\rho(x_i, x_j) > \varepsilon$, for all $i, j \in \{1, \dots, N\}$ with $i \neq j$. The ε-packing number $M(\varepsilon; \mathcal{C}, \rho)$ is the cardinality of the largest ε-packing.

Definition 5 (ReLU neural network)**.** Let $L \in \mathbb{N}$ and $N_0, N_1, \dots, N_L \in \mathbb{N}$. A ReLU neural network Φ is a map $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ given by

$$\Phi = \begin{cases} W_1, & L = 1, \\ W_2 \circ \rho \circ W_1, & L = 2, \\ W_L \circ \rho \circ W_{L-1} \circ \rho \circ \cdots \circ \rho \circ W_1, & L \geq 3, \end{cases}$$

where, for $\ell \in \{1, \dots, L\}$, $W_\ell \colon \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, W_\ell(x) := A_\ell x + b_\ell$ are the associated affine transformations with matrices $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and (bias) vectors $b_\ell \in \mathbb{R}^{N_\ell}$, and the ReLU activation function $\rho\colon \mathbb{R} \to \mathbb{R}$, $\rho(x) := \max\{x, 0\}$ acts component-wise, i.e., $\rho(x_1, \dots, x_N) := (\rho(x_1), \dots, \rho(x_N))$. We denote by $\mathcal{N}_{d,d'}$ the set of all ReLU neural networks with input dimension $N_0 = d$ and output dimension $N_L = d'$. Moreover, we define the following quantities related to the notion of size of the ReLU neural network Φ:

- the *connectivity* $\mathcal{M}(\Phi)$ is the total number of non-zero entries in the matrices A_ℓ, $\ell \in \{1, \dots, L\}$, and the vectors b_ℓ, $\ell \in \{1, \dots, L\}$,
- *depth* $\mathcal{L}(\Phi) := L$,
- *width* $\mathcal{W}(\Phi) := \max_{\ell=0,\dots,L} N_\ell$,
- *weight magnitude* $\mathcal{B}(\Phi) := \max_{\ell=1,\dots,L} \max\{\|A_\ell\|_\infty, \|b_\ell\|_\infty\}$.

Definition 6 (Growth function)**.** Let $\mathcal{F}$ be a class of $\{0,1\}$-valued functions on a domain $\mathcal{X}$. We define the growth function of $\mathcal{F}$, $\Pi_\mathcal{F} : \mathbb{N} \mapsto \mathbb{N}$, as

$$\Pi_\mathcal{F}(N) = \max\{|\mathcal{F}_{|X}| : X \subseteq \mathcal{X}, |X| = N\},$$

where $\mathcal{F}_{|X} = \{f|_X : f \in \mathcal{F}\}$, for $X \subset \mathcal{X}$, and $f|_X : X \mapsto \{0,1\}$ is the restriction of f to X, given by $f|_X(x) = f(x)$, for all $x \in X$.

Definition 7 (Shattering and VC dimension)**.** Let $\mathcal{F}$ be a class of $\{0,1\}$-valued functions on a domain $\mathcal{X}$. Suppose that $X = \{x_1, \dots, x_N\}$ is a subset of $\mathcal{X}$. We say that $\mathcal{F}$ shatters X if $|\mathcal{F}|_X| = 2^N$. The VC dimension of $\mathcal{F}$ is the size of the largest subset of $\mathcal{X}$ shattered by $\mathcal{F}$, or, equivalently, the largest value of N for which the growth function $\Pi_\mathcal{F}(N)$ equals 2^N. Formally,

$$\begin{aligned} \text{VC}(\mathcal{F}) &= \max\{|X| : X \subset \mathcal{X}, \mathcal{F} \text{ shatters } X\} \\ &= \max\left\{N \in \mathbb{N} : \Pi_\mathcal{F}(N) = 2^N\right\}. \end{aligned}$$

Lemma 1 (Equivalence to shattering)**.** Let $\mathcal{F}$ be a class of $\{0,1\}$-valued functions on a domain $\mathcal{X}$. Suppose that $X = \{x_1, \dots, x_N\}$ is a subset of $\mathcal{X}$. The set X is shattered by $\mathcal{F}$ if and only if for every $(y_i)_{i=1}^N \in \{0,1\}^N$, there exists a function $f \in \mathcal{F}$ so that $f(x_i) = y_i$, $i = 1, \dots, N$.