Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Prof. Dr. Helmut Bölcskei
Chair for Mathematical Information Science

Exam on Neural Network Theory
February 11, 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 6 pages including this page. Please verify that you have received all 6 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 a solution to a subproblem, you may, whenever applicable, nonetheless assume its conclusion in the ensuing subproblems.

After the exam:

6. 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.
7. 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.
8. 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)

For $a, b \in \mathbb{R}$ with $a < b$, let $\mathbb{I}_{[a,b)} : \mathbb{R} \to \{0, 1\}$ denote the indicator function of the interval $[a, b)$, defined as

$$\mathbb{I}_{[a,b)}(x) := \begin{cases} 1, & x \in [a, b) \\ 0, & x \notin [a, b) \end{cases}.$$

The goal of this problem is to approximate indicator functions by ReLU networks.

(a) (2 points) For $t \in \mathbb{R}$, let $H_t \colon \mathbb{R} \to \{0, 1\}$ denote the Heaviside function with jump at t, given by

$$H_t(x) := \begin{cases} 0, & x < t \\ 1, & x \geq t \end{cases}, \qquad x \in \mathbb{R}.$$

Let $a, b \in \mathbb{R}$ with $a < b$. Write $\mathbb{I}_{[a,b)}$ as a linear combination of Heaviside functions.

(b) (5 points) For $t \in \mathbb{R}$, $\ell \in \mathbb{N}$, let $G_{t,\ell} \colon \mathbb{R} \to [0, 1]$ be given by

$$G_{t,\ell}(x) := \begin{cases} 0, & x \leq t - 2^{-\ell} \\ 2^{\ell}(x - (t - 2^{-\ell})), & t - 2^{-\ell} < x \leq t \\ 1, & x > t \end{cases}.$$

Realize $G_{t,\ell}$ as a ReLU neural network $\Phi_{t,\ell}$ with $\mathcal{L}(\Phi_{t,\ell}) = 2$. Specify $\Phi_{t,\ell}$, $\mathcal{W}(\Phi_{t,\ell})$, $\mathcal{M}(\Phi_{t,\ell})$, and $\mathcal{B}(\Phi_{t,\ell})$.

(c) (8 points) Let $t \in \mathbb{R}$ and let $\ell \in \mathbb{N}$. Show that

$$\|G_{t,\ell} - H_t\|_{L^2(\mathbb{R})} \leq \tfrac{1}{\sqrt{3}} 2^{-\frac{\ell}{2}}.$$

(d) (4 points) Let $a, b \in \mathbb{R}$ with $a < b$ and $\ell \in \mathbb{N}$. Use Lemma 1 in the Handout to establish the existence of a ReLU network $\Phi_{a,b,\ell} \in \mathcal{N}_{1,1}$ satisfying

$$\|\Phi_{a,b,\ell} - \mathbb{I}_{[a,b)}\|_{L^2(\mathbb{R})} \leq \tfrac{2}{\sqrt{3}} 2^{-\frac{\ell}{2}}.$$

Hint: *Specify $\Phi_{a,b,\ell}$ in terms of networks $\Phi_{t,\ell}$ as derived in subproblem (b).*

(e) (6 points) Let $a, b \in \mathbb{R}$ with $a < b$ and $\varepsilon \in (0, \frac{1}{2})$. Find a ReLU network $\Psi_{a,b,\varepsilon}$ satisfying

$$\|\Psi_{a,b,\varepsilon} - \mathbb{I}_{[a,b)}\|_{L^2(\mathbb{R})} \leq \varepsilon.$$

Specify $\Psi_{a,b,\varepsilon}$, $\mathcal{L}(\Psi_{a,b,\varepsilon})$, $\mathcal{B}(\Psi_{a,b,\varepsilon})$, and $\mathcal{W}(\Psi_{a,b,\varepsilon})$ as well as an upper bound on $\mathcal{M}(\Psi_{a,b,\varepsilon})$.

Hint: *Make use of the result in subproblem (c) and Lemma 1 in the Handout and take ℓ to depend on ε.*

Problem 2 (25 points)

For $a, b \in [0, 1)$ with $a < b$, let $\mathbb{I}_{[a,b)} \colon [0, 1) \to \{0, 1\}$ denote the indicator function of the interval $[a, b)$, defined as

$$\mathbb{I}_{[a,b)}(x) := \begin{cases} 1, & x \in [a, b) \\ 0, & x \notin [a, b) \end{cases}.$$

Note that here the indicator function is defined on the domain $[0, 1)$.
For $k \in \mathbb{N}$, let

$$S_k := \left\{ h_c = \sum_{j=1}^{k} c_j \mathbb{I}_{\left[\frac{j-1}{k}, \frac{j}{k}\right)} \colon c = (c_1, \dots, c_k) \in [0, 1]^k \right\}$$

denote the set of step functions on $[0, 1)$ with k steps of length $\frac{1}{k}$ and height in $[0, 1]$.

(a) (4 points) Let $k \in \mathbb{N}$ and $c^1, c^2 \in [0, 1]^k$. Show that $\|h_{c^1} - h_{c^2}\|_{L^\infty([0,1))} = \|c^1 - c^2\|_\infty$.

(b) (8 points) Let $k = 2$, $\varepsilon = \frac{1}{3}$, and $X = \{h_{c^1}, h_{c^2}, h_{c^3}, h_{c^4}\}$ with

$$c^1 = (\tfrac{1}{3}, \tfrac{1}{3}), \quad c^2 = (\tfrac{1}{3}, \tfrac{2}{3}), \quad c^3 = (\tfrac{2}{3}, \tfrac{1}{3}), \quad c^4 = (\tfrac{2}{3}, \tfrac{2}{3}).$$

Show that X is a $\frac{1}{3}$-covering of S_2 with respect to the metric $\rho_\infty(f, g) := \|f - g\|_{L^\infty([0,1))}$.

(c) (8 points) Show that $N(\frac{1}{3}; S_2, \rho_\infty) = 4$.

Hint: *You may use, without proof, that* $M(\frac{2}{3}; S_2, \rho_\infty) \leq N(\frac{1}{3}; S_2, \rho_\infty)$.

(d) (5 points) Let $k \in \mathbb{N}$. Show that $N(\frac{1}{2}; S_k, \rho_\infty) = 1$.

Problem 3 (25 points)

(a) (4 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) (3 points) Consider $X_2 = \{-3\pi/2, -\pi/2, \pi/2, 3\pi/2\}$. Is the dichotomy

$$\{X_2^+ = \{-3\pi/2, -\pi/2\}, X_2^- = \{\pi/2, 3\pi/2\}\},$$

homogeneously linearly separable? Justify your answer.

(c) (8 points) Let $\phi_1(x) = (\cos(x), \sin(x))$. Show that the dichotomy $\{X_2^+, X_2^-\}$ from subproblem (b) is not ϕ_1-separable and find a function $f : \mathbb{R} \mapsto \mathbb{R}$ such that $\{X_2^+, X_2^-\}$ is ϕ_2-separable with $\phi_2(x) = (\cos(x), \sin(x), f(x))$.

(d) (5 points) Consider the class of functions

$$\mathcal{F} := \left\{ f : \mathbb{R}^3 \mapsto \{0,1\} : f(x_1, x_2, x_3) = \text{sgn}\Big(\sum_{i=1}^{3} a_i x_i\Big), (a_1, a_2, a_3) \in \mathbb{R}^3 \right\},$$

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

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

Find a subset of $\mathbb{R}^3$ with three elements that can be shattered by $\mathcal{F}$, and justify your answer.

Hint: *Please see Definition 6 in the Handout for the definition of shattering. You can use the equivalent definition of shattering effected by Lemma 2 in the Handout.*

(e) (5 points) Let $d \in \mathbb{N}$ and $\mathcal{G}$ be a class of $\{0, 1\}$-valued functions on $\mathbb{R}^d$. Suppose that the growth function of $\mathcal{G}$ satisfies $\Pi_{\mathcal{G}}(m) \leq 4m^2$, for all $m \in \mathbb{N}$. Show that VC($\mathcal{G}$) ≤ 8.

Hint: *Please see Definitions 5 and 6 in the Handout for the definition of the growth function and of VC dimension, respectively.*

Problem 4 (25 points)

Fix $W \in \mathbb{N}$ with $W \geq 3$. Let

$$\mathcal{F}(W) = \{\Phi : \mathbb{R} \mapsto \mathbb{R} : \Phi \text{ is a ReLU network with } \mathcal{L}(\Phi) = 2, \mathcal{W}(\Phi) \leq W\}$$

be the class of single-hidden-layer ReLU networks with width at most W, and let

$$\begin{aligned}\text{sgn}(\mathcal{F}(W)) = \{g : \mathbb{R} \mapsto \{0,1\} : \ &\text{there exists } \Phi \in \mathcal{F}(W) \\ &\text{such that } g(x) = \text{sgn}(\Phi(x)), x \in \mathbb{R}\},\end{aligned}$$

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

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

In this problem, we study the VC dimension of the class sgn$(\mathcal{F}(W))$.

(a) (4 points) Show that every network Φ in $\mathcal{F}(W)$ can be written as

$$\sum_{i=1}^{w} a_i \rho(s_i(x - b_i)) + c, \ x \in \mathbb{R}, \tag{1}$$

for some $w \in \mathbb{N}$, $a_1, \ldots, a_w, b_1, \ldots, b_w, c \in \mathbb{R}$, $s_1, \ldots, s_w \in \{-1, 1\}$ such that $w \leq W$ and $b_1 \leq b_2 \leq b_3 \leq \cdots \leq b_w$. Here, ρ is the ReLU activation function $\rho \colon \mathbb{R} \mapsto \mathbb{R}$ given by $\rho(x) := \max(x, 0), x \in \mathbb{R}$.

(b) (4 points) A function $f : \mathbb{R} \mapsto \mathbb{R}$ is said to be affine on a set $X \subset \mathbb{R}$ if there exist $u, v \in \mathbb{R}$ such that $f(x) = ux + v$, for all $x \in X$. Suppose that a network $\Phi \in \mathcal{F}(W)$ is represented in the form (1) according to subproblem (a). Show that Φ is affine on each of the $(w+1)$ intervals $(-\infty, b_1]$, $[b_1, b_2]$, $[b_2, b_3]$, …, $[b_{w-1}, b_w]$, and $[b_w, \infty)$.

(c) (4 points) Suppose that x_1, x_2, x_3 are real numbers with $x_1 \leq x_2 \leq x_3$ and $f : \mathbb{R} \mapsto \mathbb{R}$ is affine on $[x_1, x_3]$. Show that if sgn$(f(x_1))$ = sgn$(f(x_3))$, then necessarily sgn$(f(x_1))$ = sgn$(f(x_2))$ = sgn$(f(x_3))$.

(d) (4 points) Show that for all $n \in \mathbb{N}$ and $(x_i)_{i=1}^n \in \mathbb{R}^n$ with $n \geq 2W + 3$ and $x_1 < x_2 < \cdots < x_n$, there does not exist a ReLU network $\Phi \in \mathcal{F}(W)$ such that

$$\text{sgn}(\Phi(x_i)) = \begin{cases} 0, & \text{if } i \text{ is odd}, \\ 1, & \text{if } i \text{ is even}, \end{cases}$$

for $i = 1, \ldots, n$.

Hint: *Use the results from subproblems (a), (b), (c) and the pigeonhole principle in Lemma 3 in the Handout.*

(e) (2 points) Use the result from subproblem (d) to show that

$$\text{VC}(\text{sgn}(\mathcal{F}(W))) \leq 2W + 2.$$

(f) (7 points) Show that for every $(z_i)_{i=1}^{W} \in \mathbb{R}^W$, there exists a ReLU network $\Phi \in \mathcal{F}(W)$ such that $\Phi(i) = z_i$, for $i = 1, \ldots, W$. Then, use this result to establish that $\mathrm{VC}(\mathrm{sgn}(\mathcal{F}(W))) \geq W$.

Hint: *Work with expression* (1) *with $w = W$, $s_i = 1$, and $b_i = i$, for $i = 1, \ldots, W$.*

You can get partial credit by solving this subproblem for the special case $W = 3$.

ETH
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
February 11, 2022

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

$$\|x\|_\infty := \max_{j \in \{1,\ldots,n\}} |x_j|.$$

For $X, Y \subseteq \mathbb{R}$, $f\colon X \to Y$, 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 (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 $\{x_1, \ldots, x_N\} \subseteq \mathcal{C}$ such that for each $x \in \mathcal{C}$, there exists an $i \in \{1, \ldots, N\}$ 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 3 (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 $\{x_1, \ldots, x_N\} \subseteq \mathcal{C}$ such that $\rho(x_i, x_j) > \varepsilon$, for all distinct i, j. The ε-packing number $M(\varepsilon; \mathcal{X}, \rho)$ is the cardinality of the largest ε-packing.

Definition 4 (ReLU network)**.** Let $L \in \mathbb{N}$ and $N_0, N_1, \ldots, N_L \in \mathbb{N}$. 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 $\ell \in \{1, 2, \ldots, L\}$, $W_\ell\colon \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, W_\ell(x) := A_\ell x + b_\ell$ 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., $\rho(x_1, \ldots, x_N) := (\rho(x_1), \ldots, \rho(x_N))$. We denote by $\mathcal{N}_{d,d'}$ the set of all ReLU 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 network Φ:

- the *connectivity* $\mathcal{M}(\Phi)$ is the total number of non-zero entries in the matrices A_ℓ, $\ell \in \{1, 2, \ldots, L\}$, and the vectors b_ℓ, $\ell \in \{1, 2, \ldots, L\}$,
- *depth* $\mathcal{L}(\Phi) := L$,
- *width* $\mathcal{W}(\Phi) := \max_{\ell=0,\ldots,L} N_\ell$,
- *weight magnitude* $\mathcal{B}(\Phi) := \max_{\ell=1,\ldots,L} \max\{\|A_\ell\|_\infty, \|b_\ell\|_\infty\}$.

Lemma 1. Let $c_1, c_2 \in \mathbb{R}$, and $\Phi_1, \Phi_2 \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi_1) = \mathcal{L}(\Phi_2)$. There exists a network $\Phi \in \mathcal{N}_{1,1}$ satisfying

$$\Phi(x) = c_1\Phi_1(x) + c_2\Phi_2(x), \quad \text{for all } x \in \mathbb{R},$$

$\mathcal{L}(\Phi) = \mathcal{L}(\Phi_1)$, $\mathcal{B}(\Phi) = \max\{|c_1|\mathcal{B}(\Phi_1), |c_2|\mathcal{B}(\Phi_2)\}$, $\mathcal{W}(\Phi) \leq \mathcal{W}(\Phi_1)+\mathcal{W}(\Phi_2)$, and $\mathcal{M}(\Phi) \leq \mathcal{M}(\Phi_1) + \mathcal{M}(\Phi_2)$.

Definition 5 (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 6 (Shattering and VC dimension)**.** Let $\mathcal{F}$ be a class of $\{0,1\}$-valued functions on a domain $\mathcal{X}$. Suppose that $X = \{x_1, x_2, \ldots, x_N\}$ 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 2 (Equivalent definition of shattering)**.** Let $\mathcal{F}$ be a class of $\{0,1\}$-valued functions on a domain $\mathcal{X}$. Suppose that $X = \{x_1, x_2, \ldots, x_N\}$ 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 $f \in \mathcal{F}$ such that $f(x_i) = y_i$, $i = 1, \ldots, N$.

Lemma 3 (The pigeonhole principle)**.** Suppose $X, S_1, \ldots, S_n$ are sets such that $X \subset \cup_{i=1}^n S_i$. Then, there exists an $i \in \{1\ldots, n\}$ so that $|X \cap S_i| \geq \frac{|X|}{n}$.