Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Prof. Dr. Helmut Bölcskei
Chair for Mathematical Information Science

Solutions to the **Exam on Neural Network Theory** **February 11, 2022**

Problem 1

(a) $I_{a,b} = H_a - H_b$

(b) $\Phi_{t,\ell}(x) = 2^\ell \rho(x - (t - 2^{-\ell})) - 2^\ell \rho(x - t)$,
$\mathcal{W}(\Phi_{t,\ell}) = 2$, $\mathcal{B}(\Phi_{t,\ell}) = \max\{2^\ell, |t|, |t - 2^{-\ell}|\}$, and

$$\mathcal{M}(\Phi_{t,\ell}) = \begin{cases} 5, & t \in \{0, 2^{-\ell}\} \\ 6, & \text{else} \end{cases}.$$

(c) Observe that

$$\begin{aligned} \|G_{t,\ell} - H_t\|_{L^2(\mathbb{R})}^2 &= \int_{t-2^{-\ell}}^{t} |G_{t,\ell}(x) - H_t(x)|^2 \mathrm{d}x \\ &= \int_{t-2^{-\ell}}^{t} |2^\ell(x - (t - 2^{-\ell}))|^2 \mathrm{d}x \\ &= \int_0^{2^{-\ell}} |2^\ell x|^2 \mathrm{d}x \\ &= \int_0^{2^{-\ell}} 2^{2\ell} x^2 \mathrm{d}x \\ &= 2^{2\ell} \left(\tfrac{1}{3} x^3 \Big|_0^{2^{-\ell}} \right) \\ &= \tfrac{1}{3} 2^{-\ell}. \end{aligned}$$

Taking square roots now yields the desired result.

(d) Let $\Phi_{a,\ell}$ and $\Phi_{b,\ell}$ be the networks obtained by setting $t = a$ and $t = b$, respectively, in the network $\Phi_{t,\ell}$ from subproblem (b). Take, in accordance with Lemma 1 in the Handout, $\Phi_{a,b,\ell}$ to be the network satisfying $\Phi_{a,b,\ell}(x) := \Phi_{a,\ell}(x) - \Phi_{b,\ell}(x)$, for all $x \in \mathbb{R}$. Using the results from subproblems (a) and (c), it follows that

$$\begin{aligned} \|\Phi_{a,b,\ell} - I_{a,b}\|_{L^2(\mathbb{R})} &= \|(\Phi_{a,\ell} - \Phi_{b,\ell}) - (H_a - H_b)\|_{L^2(\mathbb{R})} \\ &= \|(G_{a,\ell} - G_{b,\ell}) - (H_a - H_b)\|_{L^2(\mathbb{R})} \\ &\leq \|G_{a,\ell} - H_a\|_{L^2(\mathbb{R})} + \|G_{b,\ell} - H_b\|_{L^2(\mathbb{R})} \\ &\leq \tfrac{2}{\sqrt{3}} 2^{-\frac{\ell}{2}}. \end{aligned}$$

(e) Take $\Psi_{a,b,\varepsilon} = \Phi_{a,b,\ell_\varepsilon}$ for $\ell_\varepsilon = \lceil 2\log_2(\frac{2}{\sqrt{3}}\varepsilon^{-1})\rceil$ with $\Phi_{a,b,\ell_\varepsilon}$ as defined in subproblem (d). Consequently, we have

$$\|\Psi_{a,b,\varepsilon} - I_{a,b}\|_{L^2(\mathbb{R})} = \|\Phi_{a,b,\ell_\varepsilon} - I_{a,b}\|_{L^2(\mathbb{R})} \leq \tfrac{2}{\sqrt{3}} 2^{-\frac{\ell_\varepsilon}{2}} \leq \varepsilon.$$

Application of Lemma 1 in the Handout yields

$$\begin{aligned}
\mathcal{L}(\Psi_{a,b,\varepsilon}) &= 2 \\
\mathcal{B}(\Psi_{a,b,\varepsilon}) &= \max\{2^{\ell_\varepsilon}, |a|, |b|, |a - 2^{-\ell_\varepsilon}|, |b - 2^{-\ell_\varepsilon}|\} \\
\mathcal{M}(\Psi_{a,b,\varepsilon}) &\leq 12.
\end{aligned}$$

Problem 2

(a) It holds that

$$\|h_{c^1} - h_{c^2}\|_{L^\infty([0,1))} = \left\| \sum_{j=1}^{k} (c_j^1 - c_j^2) \mathbb{I}_{\left[\frac{j-1}{k}, \frac{j}{k}\right)} \right\|_{L^\infty([0,1))} = \max_{j=1,\dots,k} |c_j^1 - c_j^2| = \|c^1 - c^2\|_\infty.$$

(b) First, note that, for every $x \in [0,1]$, we have $|x - \frac{1}{3}| \leq \frac{1}{3}$ or $|x - \frac{2}{3}| \leq \frac{1}{3}$ (or both). Now, let $h \in S_2$. There exists $c \in [0,1]^2$ such that $h = h_c$. Consequently, there exists $i \in \{1,2,3,4\}$ so that

$$\|h - h_{c^i}\|_{L^\infty([0,1))} = \|h_c - h_{c^i}\|_{L^\infty([0,1))} = \|c - c^i\|_\infty \leq \tfrac{1}{3}.$$

(c) Let $Y = \{h_{e^1}, h_{e^2}, h_{e^3}, h_{e^4}\}$ with $e^1 = (\frac{1}{8}, \frac{1}{8})$, $e^2 = (\frac{1}{8}, \frac{7}{8})$, $e^3 = (\frac{7}{8}, \frac{1}{8})$, and $e^4 = (\frac{7}{8}, \frac{7}{8})$. For $i, j \in \{1,2,3,4\}$ with $i \neq j$, we have

$$\|h_{e^i} - h_{e^j}\|_{L^\infty([0,1))} = \|e^i - e^j\|_\infty = |\tfrac{7}{8} - \tfrac{1}{8}| = \frac{3}{4} > \tfrac{2}{3}.$$

This implies that Y is a $\frac{2}{3}$-packing of S_2. Combining the hint with the result from subproblem (b) completes the proof.

(d) For $c := (\frac{1}{2}, \dots, \frac{1}{2}) \in [0,1]^k$, we get $h_c(x) = \frac{1}{2}$, for all $x \in [0,1)$. As $\max_{x \in [0,1]} |x - \frac{1}{2}| \leq \frac{1}{2}$, the singleton $\{h_c\}$ is a $\frac{1}{2}$-covering of S_k with respect to the metric ρ_∞. The claim now follows by noting that an empty set cannot be a covering for a non-empty set.

Problem 3

(a) The dichotomy $\{X_1^+, X_1^-\}$ is said to be homogeneously linearly separable if there exists a nonzero vector $w_1 \in \mathbb{R}^d$ such that

$$\begin{aligned}\langle x, w_1\rangle &> 0, \text{ for all } x \in X_1^+,\\ \langle x, w_1\rangle &< 0, \text{ for all } x \in X_1^-,\end{aligned}$$

and it is said to be ϕ-separable if there exists a nonzero vector $w_2 \in \mathbb{R}^m$ such that

$$\begin{aligned}\langle \phi(x), w_2\rangle &> 0, \text{ for all } x \in X_1^+,\\ \langle \phi(x), w_2\rangle &< 0, \text{ for all } x \in X_1^-.\end{aligned}$$

(b) The dichotomy is homogeneously linearly separable. Let $w = -1$. Then $\langle x, w\rangle > 0$ for all $x \in X_2^+ = \{-3\pi/2, -\pi/2\}$, and $\langle x, w\rangle < 0$ for all $x \in X_2^- = \{\pi/2, 3\pi/2\}$.

(c) Suppose, for the sake of contradiction, that $\{X_2^+, X_2^-\}$ is ϕ_1-separable. Then, there exists a nonzero vector $w = (u, v)$ such that

$$\begin{aligned}\langle \phi_1(x), (u, v)\rangle &> 0, \text{ for all } x \in X_2^+,\\ \langle \phi_1(x), (u, v)\rangle &< 0, \text{ for all } x \in X_2^-,\end{aligned}$$

which amounts to

$$\begin{aligned}\langle \phi_1(-3\pi/2), (u, v)\rangle &= \quad v > 0, &(1)\\ \langle \phi_1(-\pi/2), (u, v)\rangle &= -v > 0, &(2)\\ \langle \phi_1(\pi/2), (u, v)\rangle &= \quad v < 0, &(3)\\ \langle \phi_1(3\pi/2), (u, v)\rangle &= -v < 0. &(4)\end{aligned}$$

Relations (1) and (3) can not hold simultaneously, which establishes the desired contradiction.

Let $f(x) = x$, $x \in \mathbb{R}$, and hence $\phi_2 = (\cos(x), \sin(x), x)$, $x \in \mathbb{R}$, and let $w = (0, 0, -1)$. Then, we have

$$\begin{aligned}\langle \phi_2(-3\pi/2),\, w\rangle &= \quad 3\pi/2 > 0, &(5)\\ \langle \phi_2(-\pi/2),\, w\rangle &= \quad \pi/2 > 0, &(6)\\ \langle \phi_2(\pi/2),\, w\rangle &= -\pi/2 < 0, &(7)\\ \langle \phi_2(3\pi/2),\, w\rangle &= -3\pi/2 < 0, &(8)\end{aligned}$$

and therefore the dichotomy $\{X_2^+ = \{-3\pi/2, -\pi/2\}, X_2^- = \{\pi/2, 3\pi/2\}\}$ is ϕ_2-separable.

(d) Let $x_1 = (1, 0, 0)$, $x_2 = (0, 1, 0)$, $x_3 = (0, 0, 1)$, and set $X = \{x_1, x_2, x_3\}$. Then, for every $(y_1, y_2, y_3) \in \{0, 1\}^3$, there exists an $f \in \mathcal{F}$ such that

$$f(x_i) = y_i, \quad i = 1, 2, 3,$$

namely $f(z_1, z_2, z_3) = \text{sgn}(\sum_{i=1}^3 (2y_i - 1)z_i)$, $(z_1, z_2, z_3) \in \mathbb{R}^3$. It therefore follows from Lemma 2 in the Handout that $\mathcal{F}$ shatters X.

(e) Suppose $N \in \mathbb{N}$ such that $\Pi_{\mathcal{G}}(N) = 2^N$. We have $2^N = \Pi_{\mathcal{G}}(N) \leq 4N^2$, where the inequality is by the assumption $\Pi_{\mathcal{G}}(m) \leq 4m^2$, for all $m \in \mathbb{N}$. It follows that $N - 2 - 2\log_2(N) \leq 0$. Let $g(x) = x - 2 - 2\log_2(x)$. We have $g(8) = 8 - 2 - 2\log_2(8) = 0$, and

$$g'(x) = 1 - \frac{2}{x\ln(2)} > 0,$$

for all $x \geq 8$. It hence follows that g is strictly increasing on $[8, \infty)$, and $g(x) > 0$, for all $x \in (8, \infty)$. Since $g(N) \leq 0$, we must have $N \notin (8, \infty)$, i.e., $N \leq 8$. Then, by the definition of VC dimension,

$$\text{VC}(\mathcal{G}) = \max\{N \in \mathbb{N} : \Pi_{\mathcal{G}}(N) = 2^N\} \leq 8.$$

Problem 4

(a) Suppose that $\Phi \in \mathcal{F}(W)$. By definition, $\Phi = W_2 \circ \rho \circ W_1$ for some $W_1(x) = (d_1, \ldots, d_n)^T x + (e_1, \ldots, e_n)^T$, $x \in \mathbb{R}$, and $W_2(x) = (f_1, \ldots, f_n)x + g$, $x \in \mathbb{R}^n$, with $n \leq W$. We have

$$\Phi(x) = \sum_{i=1}^{n} f_i \rho(d_i x + e_i) + g \tag{9}$$

$$= \sum_{\substack{i=1,\ldots,n \\ d_i \neq 0}} f_i |d_i| \rho\left(\frac{d_i}{|d_i|}\left(x - \left(-\frac{e_i}{d_i}\right)\right)\right) + \left(\sum_{\substack{i=1,\ldots,n \\ d_i = 0}} f_i \rho(e_i) + g\right), \tag{10}$$

where in (10) we used $\rho(uv) = u\rho(v)$, for all $u \geq 0$ and $v \in \mathbb{R}$. Let $\mathcal{I} = \{i \in \{1, \ldots, n\} : d_i \neq 0\}$, $w = |\mathcal{I}|$, and $k : \{1, \ldots, w\} \mapsto \mathcal{I}$ be an ordering of $\mathcal{I}$ such that $\left(-\frac{e_{k(i)}}{d_{k(i)}}\right)_{i=1}^{w}$ is non-decreasing. Set $a_i = f_{k(i)}|d_{k(i)}|$, $b_i = \left(-\frac{e_{k(i)}}{d_{k(i)}}\right)$, $s_i = \frac{d_{k(i)}}{|d_{k(i)}|} \in \{0, 1\}$, for $i = 1, \ldots, w$, and $c = \sum_{\substack{i=1,\ldots,n \\ d_i = 0}} f_i \rho(e_i) + g$. Then, $\Phi(x)$ can be written as

$$\sum_{i=1}^{w} a_i \rho(s_i(x - b_i)) + c, \; x \in \mathbb{R}, \tag{11}$$

such that $b_1 \leq b_2 \leq b_3 \leq \cdots \leq b_w$, and $w \leq n \leq W$.

(b) For all $i \in \{1, \ldots, w\}$, as the function $x \mapsto a_i \rho(s_i(x - b_i))$ is affine on $(-\infty, b_i]$ and $[b_i, \infty)$, and each of the intervals $(-\infty, b_1]$, $[b_1, b_2]$, $[b_2, b_3]$, $\ldots$, $[b_{w-1}, b_w]$, and $[b_w, \infty)$ is contained in either $(-\infty, b_i]$ or $[b_i, \infty)$, we have that the function $x \mapsto a_i \rho(s_i(x - b_i))$ is affine on each of the intervals $(-\infty, b_1]$, $[b_1, b_2]$, $[b_2, b_3]$, $\ldots$, $[b_{w-1}, b_w]$, and $[b_w, \infty)$. Since affinity is preserved under addition, the sum $x \mapsto \sum_{i=1}^{w} a_i \rho(s_i(x - b_i)) + c$ is also affine on each of the intervals $(-\infty, b_1]$, $[b_1, b_2]$, $[b_2, b_3]$, $\ldots$, $[b_{w-1}, b_w]$, and $[b_w, \infty)$.

(c) Suppose first that $\text{sgn}(f(x_1)) = \text{sgn}(f(x_3)) = 1$, which implies $f(x_1) \geq 0$ and $f(x_3) \geq 0$. Since f is affine on $[x_1, x_3]$, it attains its minimum and maximum on $[x_1, x_3]$ at its boundary points, i.e., at either x_1 or x_3. This then implies $f(x_2) \geq \min\{f(x_1), f(x_3)\} \geq 0$ and hence $\text{sgn}(f(x_2)) = 1 = \text{sgn}(f(x_1)) = \text{sgn}(f(x_3))$. The statement for the case $\text{sgn}(f(x_1)) = \text{sgn}(f(x_3)) = 0$ can be established similarly.

(d) Suppose, for the sake of contradiction, that there exists a set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}$ of n elements with $n \geq 2W + 3$ and $x_1 < x_2 < x_3 < \cdots < x_n$ and 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 $i = 1, \ldots, n$. According to the result from subproblem (a), Φ can be written as

$$\sum_{i=1}^{w} a_i \rho(s_i(x - b_i)) + c, \; x \in \mathbb{R}, \tag{12}$$

where $w \leq W$, $a_1, \ldots, a_w, b_1, \ldots, b_w, c \in \mathbb{R}$, $s_1, \ldots, s_w \in \{-1, 1\}$ and $b_1 \leq b_2 \leq \cdots \leq b_w$. Let $S_1 = (-\infty, b_1]$, $S_i = [b_{i-1}, b_i]$, for $i = 2, \ldots, w$, and $S_{w+1} = [b_w, \infty)$.

Then $X \subset \cup_{i=1}^{w+1} S_i$. By the pigeonhole principle, there exists an interval S_i, $i \in \{1, \ldots, w+1\}$, such that $|X \cap S_i| \geq \frac{|X|}{w+1} \geq \frac{2W+3}{W+1} > 2$, which implies $|X \cap S_i| \geq 3$. We can hence assume, without loss of generality, that $\{x_j, x_{j+1}, x_{j+2}\} \subset S_i$ for some $j \in \{1, \ldots, n-2\}$. We have

$$\text{sgn}(\Phi(x_j)) = \text{sgn}(\Phi(x_{j+2})) \neq \text{sgn}(\Phi(x_{j+1})) \tag{13}$$

by assumption. According to subproblem (b), Φ is affine on S_i and therefore affine on its subinterval $[x_j, x_{j+2}]$. Then, according to the result from subproblem (c) with $\text{sgn}(\Phi(x_j)) = \text{sgn}(\Phi(x_{j+2}))$, we must have $\text{sgn}(\Phi(x_j)) = \text{sgn}(\Phi(x_{j+2})) = \text{sgn}(\Phi(x_{j+1}))$, which contradicts (13).

(e) Suppose, for the sake of contradiction, that $\text{VC}(\text{sgn}(\mathcal{F}(W))) \geq 2W + 3$. Then, by definition, there exists a subset $X = \{x_1, \ldots, x_n\}$, with $n \geq 2W + 3$ and $x_1 < x_2 < \cdots < x_n$, such that $\text{sgn}(\mathcal{F}(W))$ shatters X. In particular, this together with Lemma 2 in the Handout implies the existence of a network $\Phi \in \mathcal{F}(W)$ such that, for $i = 1, \ldots, n$,

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

The existence of this Φ, however, contradicts the result from subproblem (d).

(f) For every $\mathbf{z} = (z_i)_{i=1}^W \in \mathbb{R}^W$, we define $\Phi_{\mathbf{z}} : \mathbb{R} \mapsto \mathbb{R}$ as

$$\begin{aligned} \Phi_{\mathbf{z}}(x) =& z_1(1 - \rho(x-1) + \rho(x-2)) \\ &+ \sum_{i=2}^{W-1} z_i(\rho(x-(i-1)) - 2\rho(x-i) + \rho(x-(i+1))) \\ &+ z_W(\rho(x-(W-1))) \qquad (14) \\ =& z_1 + (z_2 - z_1)\rho(x-1) + \sum_{i=2}^{W-1}(z_{i-1} - 2z_i + z_{i+1})\rho(x-i) \\ &+ z_{W-1}\rho(x-W), \qquad (15) \end{aligned}$$

such that $\Phi_{\mathbf{z}}(i) = z_i$, for all $i \in \{1, \ldots, W\}$, according to expression (14), and $\Phi_{\mathbf{z}} \in \mathcal{F}(W)$, according to expression (15).

For every $\mathbf{y} = (y_i)_{i=1}^W \in \{0,1\}^W$, let $\mathbf{z} = (z_i)_{i=1}^W = (2y_i - 1)_{i=1}^W$. We have that $x \mapsto \text{sgn}(\Phi_{\mathbf{z}}(x)) \in \text{sgn}(\mathcal{F}(W))$, and for $i = 1, \ldots, W$,

$$\begin{aligned} \text{sgn}(\Phi_{\mathbf{z}}(i)) =& \text{sgn}(2y_i - 1) \qquad (16) \\ =& y_i, \qquad (17) \end{aligned}$$

where (16) follows from $\Phi_{\mathbf{z}}(i) = z_i$, $i = 1, \ldots, W$, and in (17) we used $\text{sgn}(2y-1) = y$ for $y \in \{0, 1\}$. It follows from Lemma 2 in the Handout that $\text{sgn}(\mathcal{F}(W))$ shatters $\{1, \ldots, W\}$, and therefore, by the definition of VC dimension, we get

$$\text{VC}(\text{sgn}(\mathcal{F}(W))) \geq |\{1, \ldots, W\}| = W.$$