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 6, 2024

Problem 1

(a) One possible solution is

$$\begin{aligned}\Phi(x) &= \rho(x) - \rho(x-1)\\ &= \begin{pmatrix}1 & -1\end{pmatrix} \circ \rho \circ \left(\begin{pmatrix}1\\1\end{pmatrix} x + \begin{pmatrix}0\\-1\end{pmatrix} \right), \quad x \in \mathbb{R}.\end{aligned}$$

This network satisfies $\mathcal{L}(\Phi) = 2, \mathcal{W}(\Phi) = 2$, and $\mathcal{M}(\Phi) = 5$.

(b) We directly calculate the function f as

$$f(x) = \begin{cases} 0, & \text{for } x < 0 \\ 4x, & \text{for } 0 \leq x < \frac{1}{4} \\ -2x + \frac{3}{2}, & \text{for } \frac{1}{4} \leq x < \frac{3}{4} \\ 4x - 3, & \text{for } \frac{3}{4} \leq x < 1 \\ 1, & \text{for } x \geq 1. \end{cases}$$

The sketch of f is given below.

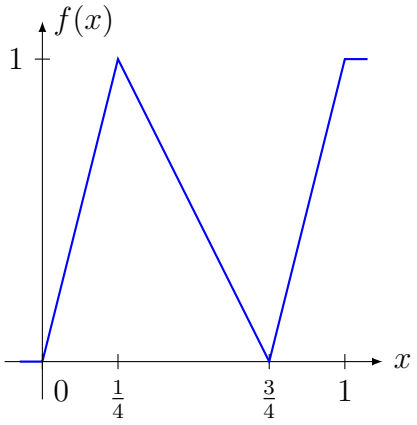

Figure 1: $f(x)$.

One possible solution for Φ^f is

$$\begin{aligned}\Phi^f(x) &= 4\,\rho(x) - 6\,\rho(x-1/4) + 6\,\rho(x-3/4) - 4\,\rho(x-1)\\ &= \begin{pmatrix}4 & -6 & 6 & -4\end{pmatrix} \circ \rho \circ \left(\begin{pmatrix}1\\1\\1\\1\end{pmatrix} x + \begin{pmatrix}0\\-1/4\\-3/4\\-1\end{pmatrix} \right), \quad x \in \mathbb{R},\end{aligned}$$

which satisfies $\mathcal{L}(\Phi^f) = 2$, $\mathcal{W}(\Phi^f) = 4$, $\mathcal{M}(\Phi^f) = 11$, and $\mathcal{B}(\Phi^f) = 6$.

(c) We first realize the function g as a σ-network according to

$$g(x, y) = (W_3 \circ \sigma \circ W_2 \circ \sigma \circ W_1)(x, y), \tag{1}$$

where

$$\begin{aligned} W_1(x, y) &= \begin{pmatrix} -2 & -1 \\ 0.5 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad x, y \in \mathbb{R}, \\ W_2(x) &= \begin{pmatrix} 1 & 1 \end{pmatrix} x, \quad x \in \mathbb{R}^2, \\ W_3(x) &= x, \quad x \in \mathbb{R}. \end{aligned}$$

Based on the result from subproblem (a), namely

$$\sigma(x) = \rho(x) - \rho(x - 1),$$

we can convert σ in (1) to obtain a ReLU network realizing $g(x, y)$ according to

$$\Phi^g(x, y) = (W_3' \circ \rho \circ W_2' \circ \rho \circ W_1')(x, y), \tag{2}$$

where

$$\begin{aligned} W_1'(x, y) &= \begin{pmatrix} -2 & -1 \\ -2 & -1 \\ 0.5 & -2 \\ 0.5 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \\ W_2'(x) &= \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \end{pmatrix} x + \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad x \in \mathbb{R}^4, \\ W_3'(x) &= \begin{pmatrix} 1 & -1 \end{pmatrix} x, \quad x \in \mathbb{R}^2. \end{aligned}$$

Inspection of (2) shows that $\mathcal{L}(\Phi^g) = 3$ as desired.

(d) We can express the operation $\vee$ by affine copies of ρ according to

$$x \vee y = \max\{x, y\} = x + \max\{0, y - x\} = \rho(x) - \rho(-x) + \rho(y - x), \quad \text{for } x, y \in \mathbb{R}.$$

One possible solution for $\Phi^\vee$ is hence

$$\Phi^\vee(x, y) = \begin{pmatrix} 1 & -1 & 1 \end{pmatrix} \circ \rho \circ \left(\begin{pmatrix} 1 & 0 \\ -1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \right), \quad x, y \in \mathbb{R}.$$

Likewise, for the operation $\wedge$, we have

$$x \wedge y = \min\{x, y\} = x - \max\{0, x - y\} = \rho(x) - \rho(-x) - \rho(x - y), \quad \text{for } x, y \in \mathbb{R}.$$

So we can choose $\Phi^\wedge$ according to

$$\Phi^\wedge(x, y) = \begin{pmatrix} 1 & -1 & -1 \end{pmatrix} \circ \rho \circ \left(\begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \right), \quad x, y \in \mathbb{R}.$$

(e) In (d), the "$\min$" operation can be realized according to

$$\min\{x, y\} = (W_2 \circ \rho \circ W_1)(x, y),$$

where

$$W_1(x, y) = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \quad x, y \in \mathbb{R},$$
$$W_2(x) = \begin{pmatrix} 1 & -1 & -1 \end{pmatrix} x, \quad x \in \mathbb{R}^3.$$

Using $z = \rho(z) - \rho(-z)$, for $z \in \mathbb{R}$, the ReLU network $W_4 \circ \rho \circ W_3$ with

$$W_3(x, y, z) = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad x, y, z \in \mathbb{R},$$
$$W_4(x) = \begin{pmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} x, \quad x \in \mathbb{R}^5.$$

satisfies

$$(W_4 \circ \rho \circ W_3)(x, y, z) = \begin{pmatrix} \min\{x, y\} \\ z \end{pmatrix}, \quad x, y, z \in \mathbb{R}.$$

Let $\Phi^h := W_2 \circ \rho \circ W_1 \circ W_4 \circ \rho \circ W_3$. Then $\Phi^h(x, y, z) = \min\{\min\{x, y\}, z\} = \min\{x, y, z\}$ as required.

Problem 2

(a) An ϵ-covering of $\mathcal{F}$ with respect to the metric ρ_∞ is a set $\{x_1, \ldots, x_N\} \subset \mathcal{F}$ such that for each $x \in \mathcal{F}$, there exists an $i \in \{1, \ldots, N\}$ so that $\rho_\infty(x, x_i) \leq \epsilon$. The ϵ-covering number $N(\epsilon; \mathcal{F}, \rho_\infty)$ is the cardinality of a smallest ϵ-covering of $\mathcal{F}$.

(b) For all $\theta, \theta' \in [0, 1]$, we have

$$\rho_\infty(f_{\theta,\theta'}, f_{0,0}) = \sup_{x\in[0,1]} |f_{\theta,\theta'}(x) - f_{0,0}(x)| = \max_{x\in[0,1]} |1 - e^{-\theta x} + \theta'| = \max_{x\in[0,1]} (1 - e^{-\theta x} + \theta') \leq 2.$$

Therefore, for $\epsilon \geq 2$, the ϵ-ball around $f_{0,0}$ contains all elements in $\mathcal{F}$. The singleton set $\{f_{0,0}\}$ constitutes an ϵ-covering of $\mathcal{F}$, which establishes

$$N(\epsilon; \mathcal{F}, \rho_\infty) = 1, \quad \text{for } \epsilon \geq 2.$$

(c) For given $\epsilon < 2$, for every $f_{\theta,\theta'} \in \mathcal{F}$, we can find (θ_i, θ'_j) in the set $\{(\theta_i, \theta'_j) : i, j = 0, 1, \ldots, T + 1\}$, such that $|\theta_i - \theta| \leq \epsilon/2$ and $|\theta'_j - \theta'| \leq \epsilon/2$. We then have

$$\begin{aligned}
\rho(f_{\theta,\theta'} - f_{\theta_i,\theta'_j}) &= \sup_{x\in[0,1]} |f_{\theta,\theta'}(x) - f_{\theta_i,\theta'_j}(x)| \\
&= \max_{x\in[0,1]} | - e^{-\theta x} + e^{-\theta_i x} + \theta' - \theta'_j| \\
&\leq |\theta' - \theta'_j| + \max_{x\in[0,1]} |e^{-\theta x} - e^{-\theta_i x}| \\
&\leq |\theta' - \theta'_j| + |\theta_i - \theta| \leq \epsilon,
\end{aligned}$$

where we used, for $\theta < \theta_i$,

$$\begin{aligned}
\max_{x\in[0,1]} |e^{-\theta x} - e^{-\theta_i x}| &= \max_{x\in[0,1]} e^{-\theta x} |1 - e^{-(\theta_i - \theta)x}| \\
&\leq \max_{x\in[0,1]} (1 - e^{-(\theta_i - \theta)x}) \\
&\leq \max_{x\in[0,1]} (\theta_i - \theta)x \\
&= |\theta_i - \theta|.
\end{aligned}$$

The case $\theta > \theta_i$ follows similarly. We conclude that the set $\{f_{\theta_i,\theta'_j} : i, j = 0, \ldots, T + 1\}$ constitutes an ϵ-covering of $\mathcal{F}$. An upper bound on the covering number is hence given by $N(\varepsilon; \mathcal{F}, \rho_\infty) \leq (T + 2)^2 \leq (\frac{1}{\epsilon} + 2)^2$.

(d) We construct an ϵ-packing as follows. Set $\theta_0 = 0$ and define $\theta_i = -\log(1 - \epsilon i)$ for all i such that $\theta_i \leq 1$. The largest index T so that this holds is given by $T = \left\lfloor \frac{1-e^{-1}}{\epsilon} \right\rfloor$. For $j = 0, 1, \ldots, \lfloor \frac{1}{\epsilon} \rfloor$, let $\theta'_j = j\epsilon$. Note that for any two distinct points (θ_i, θ'_j) and (θ_m, θ'_n), if $j \neq n$, we have

$$\rho(f_{\theta_i,\theta'_j}, f_{\theta_m,\theta'_n}) = \max_{x\in[0,1]} |f_{\theta_i,\theta'_j}(x) - f_{\theta_m,\theta'_n}(x)| \geq |f_{\theta_i,\theta'_j}(0) - f_{\theta_m,\theta'_n}(0)| = |\theta'_j - \theta'_n| \geq \epsilon,$$

and if $j = n$, then $i \neq m$, and we have

$$\rho(f_{\theta_i,\theta'_j}, f_{\theta_m,\theta'_n}) = \max_{x\in[0,1]} |f_{\theta_i,\theta'_j}(x) - f_{\theta_m,\theta'_n}(x)| \geq |f_{\theta_i,\theta'_j}(1) - f_{\theta_m,\theta'_n}(1)| = |\epsilon(i - m)| \geq \epsilon.$$

We can therefore conclude that the set $\{f_{\theta_i,\theta'_j} : i = 0, \ldots, T, j = 0, \ldots, \lfloor \frac{1}{\epsilon} \rfloor\}$ is an ϵ-packing and the packing number satisfies $M(\varepsilon; \mathcal{F}, \rho_\infty) \geq (T+1)(1 + \lfloor \frac{1}{\epsilon} \rfloor) \geq \frac{1-e^{-1}}{\epsilon^2}$.

(e) By subproblems (c) and (d) and Lemma 1 in the Handout, we obtain

$$\frac{1-e^{-1}}{(2\epsilon)^2} \leq M(2\varepsilon;\mathcal{F},\rho_\infty) \leq N(\varepsilon;\mathcal{F},\rho_\infty) \leq \left(\frac{1}{\epsilon}+2\right)^2,$$

which allows us to conclude that $\log N(\varepsilon;\mathcal{F},\rho_\infty) \asymp \log(1/\epsilon)$ as $\epsilon \to 0$.

Problem 3

(a)

$$\text{Softmax}^{(1)}(x) = \frac{\exp(x)}{\exp(x)} = 1, \tag{3}$$

for all $x \in \mathbb{R}$.

(b) Fix $x \in \mathbb{R}^n$ and $i \in \{1, \ldots, n\}$. For all $t \in \mathbb{R}$, $\exp(t) \geq 0$, so

$$\text{Softmax}_i^{(n)}(x) = \frac{\exp(x_i)}{\sum_{k=1}^n \exp(x_k)} \geq 0. \tag{4}$$

Moreover,

$$\text{Softmax}_i^{(n)}(x) = \frac{\exp(x_i)}{\sum_{k=1}^n \exp(x_k)} \leq \frac{\exp(x_i)}{\exp(x_i)} \leq 1. \tag{5}$$

(c) Let $n \in \mathbb{N}$, $i, j \in \{1, \ldots, n\}$, and $x \in \mathbb{R}^n$. Suppose that $i \neq j$.

$$\begin{aligned}
\partial_j \text{Softmax}_i^{(n)}(x) &= \partial_j \left(\frac{\exp(x_i)}{\sum_{k=1}^n \exp(x_k)} \right) && (6)\\
&= -\frac{\exp(x_i)\partial_j \left(\sum_{k=1}^n \exp(x_k)\right)}{\left(\sum_{k=1}^n \exp(x_k)\right)^2} && (7)\\
&= -\frac{\exp(x_i)\exp(x_j)}{\left(\sum_{k=1}^n \exp(x_k)\right)^2} && (8)\\
&= -\text{Softmax}_i^{(n)}(x)\,\text{Softmax}_j^{(n)}(x). && (9)
\end{aligned}$$

$$\begin{aligned}
\partial_i \text{Softmax}_i^{(n)}(x)_i &= \partial_i \left(\frac{\exp(x_i)}{\sum_{k=1}^n \exp(x_k)} \right) && (10)\\
&= \frac{\left(\sum_{k=1}^n \exp(x_k)\right) \partial_i(\exp(x_i)) - \exp(x_i)\partial_i \left(\sum_{k=1}^n \exp(x_k)\right)}{\left(\sum_{k=1}^n \exp(x_k)\right)^2} && (11)\\
&= \frac{\left(\sum_{k=1}^n \exp(x_k)\right) \exp(x_i) - \exp(x_i)^2}{\left(\sum_{k=1}^n \exp(x_k)\right)^2} && (12)\\
&= \frac{\exp(x_i)}{\sum_{k=1}^n \exp(x_k)} - \frac{\exp(x_i)^2}{\left(\sum_{k=1}^n \exp(x_k)\right)^2} && (13)\\
&= \text{Softmax}_i^{(n)}(x) - \text{Softmax}_i^{(n)}(x)\,\text{Softmax}_i^{(n)}(x). && (14)
\end{aligned}$$

Combining (14) for each $i \in \{1, \ldots, n\}$, we obtain the result in the hint, namely

$$\nabla \text{Softmax}^{(n)}(x) = \text{diag}\left(\text{Softmax}^{(n)}(x)\right) - \text{Softmax}^{(n)}(x)\,\text{Softmax}^{(n)}(x)^T. \tag{15}$$

By subproblem b), one has $0 \leq \text{Softmax}_i^{(n)}(x) \leq 1$, for all $x \in \mathbb{R}^n$, $i \in \{1, \ldots, n\}$. It follows that $-1 \leq \nabla \text{Softmax}_{i,j}^{(n)}(x) \leq 1$, for all $x \in \mathbb{R}^n$, $i, j \in \{1, \ldots, n\}$. Therefore, $\|\nabla \text{Softmax}^{(n)}(x)\|_\infty \leq 1$, for all $x \in \mathbb{R}^n$, as desired.

(d) Let $n \in \mathbb{N}$ and $x, y \in \mathbb{R}^n$, $x \neq y$. By Theorem 1 in the Handout, there exists $z \in \mathbb{R}^n$ such that

$$\|\text{Softmax}^{(n)}(y) - \text{Softmax}^{(n)}(x)\|_2 \leq \|\nabla \text{Softmax}^{(n)}(z)(y - x)\|_2. \tag{16}$$

If follows that

$$\begin{aligned}
\|\operatorname{Softmax}^{(n)}(y) - \operatorname{Softmax}^{(n)}(x)\|_\infty &\overset{(a)}{\leq} \|\operatorname{Softmax}^{(n)}(y) - \operatorname{Softmax}^{(n)}(x)\|_2 && (17)\\
&\overset{(16)}{\leq} \|\nabla \operatorname{Softmax}^{(n)}(z)(y-x)\|_2 && (18)\\
&\overset{(b)}{\leq} \sqrt{n}\|\nabla \operatorname{Softmax}^{(n)}(z)(y-x)\|_\infty && (19)\\
&\overset{(c)}{\leq} \sqrt{n}\, n\, \|\nabla \operatorname{Softmax}^{(n)}(z)\|_\infty \|y-x\|_\infty && (20)\\
&\overset{(d)}{\leq} n^{3/2}\|(y-x)\|_\infty, && (21)
\end{aligned}$$

where (a) and (b) follow from Lemma 2 in the Handout, (c) is by Lemma 3 in the Handout, and (d) is by subproblem (c). We conclude that

$$|\operatorname{Softmax}^{(n)}|_{Lip} := \sup_{\substack{x,y\in[-1,1]^d \\ x\neq y}} \frac{\|\operatorname{Softmax}^{(n)}(x) - \operatorname{Softmax}^{(n)}(y)\|_\infty}{\|x-y\|_\infty} \leq n^{3/2}. \tag{22}$$

(e) Let $x, y \in [-1,1]^d$.

$$\begin{aligned}
\|\phi(x) - \phi(y)\|_\infty &= \|\operatorname{Softmax}^{(n)}(A_2\rho(A_1x)) - \operatorname{Softmax}^{(n)}(A_2\rho(A_1y))\|_\infty && (23)\\
&\overset{(a)}{\leq} n^{3/2}\|A_2\rho(A_1x) - A_2\rho(A_1y)\|_\infty && (24)\\
&\overset{(b)}{\leq} n^{3/2}m\|A_2\|_\infty\|\rho(A_1x) - \rho(A_1y)\|_\infty && (25)\\
&\overset{(c)}{\leq} n^{3/2}m\|A_2\|_\infty\|A_1x - A_1y\|_\infty && (26)\\
&\overset{(d)}{\leq} n^{3/2}m\,d\|A_2\|_\infty\|A_1\|_\infty\|x-y\|_\infty, && (27)
\end{aligned}$$

where (a) is by subproblem (d), (b) follows from Lemma 3 in the Handout, (c) is a consequence of the ReLU function being 1-Lipschitz, and (d) follows from Lemma 3 in the Handout. In summary, we have established that

$$|\phi|_{Lip} \leq n^{3/2}m\,d\|A_2\|_\infty\|A_1\|_\infty. \tag{28}$$

Problem 4

(a) We use Definition 6 in the Handout to compute x_1, x_2, x_3, and x_4 as follows:

$$x_1 = (A_1 * K_1)_{3,2} = \sum_{\substack{p\in\{1,\dots,1\}\\ q\in\{1,\dots,3\}}} [A_1]_{3+p-1,2+q-1}[K_1]_{p,q} = \sum_{q\in\{1,\dots,3\}} [A_1]_{3,1+q}[K_1]_{1,q} \tag{29}$$

$$= [A_1]_{3,2}[K_1]_{1,1} + [A_1]_{3,3}[K_1]_{1,2} + [A_1]_{3,4}[K_1]_{1,3} = 1. \tag{30}$$

$$x_2 = (A_1 * K_1)_{4,2} = \sum_{\substack{p\in\{1,\dots,1\}\\ q\in\{1,\dots,3\}}} [A_1]_{4+p-1,2+q-1}[K_1]_{p,q} = \sum_{q\in\{1,\dots,3\}} [A_1]_{4,1+q}[K_1]_{1,q} \tag{31}$$

$$= [A_1]_{4,2}[K_1]_{1,1} + [A_1]_{4,3}[K_1]_{1,2} + [A_1]_{4,4}[K_1]_{1,3} = 1. \tag{32}$$

$$x_3 = (A_1 * K_2)_{2,4} = \sum_{\substack{p\in\{1,\dots,3\}\\ q\in\{1,\dots,1\}}} [A_1]_{2+p-1,4+q-1}[K_2]_{p,q} = \sum_{p\in\{1,\dots,3\}} [A_1]_{1+p,4}[K_2]_{p,1} \tag{33}$$

$$= [A_1]_{2,4}[K_2]_{1,1} + [A_1]_{3,4}[K_2]_{2,1} + [A_1]_{4,4}[K_2]_{3,1} = 2. \tag{34}$$

$$x_4 = (A_1 * K_2)_{3,4} = \sum_{\substack{p\in\{1,\dots,3\}\\ q\in\{1,\dots,1\}}} [A_1]_{3+p-1,4+q-1}[K_2]_{p,q} = \sum_{p\in\{1,\dots,3\}} [A_1]_{2+p,4}[K_2]_{p,1} \tag{35}$$

$$= [A_1]_{3,4}[K_2]_{1,1} + [A_1]_{4,4}[K_2]_{2,1} + [A_1]_{5,4}[K_2]_{3,1} = 3. \tag{36}$$

We deduce that

$$\|A_1 * K_1\|_\infty = 1,\ \|A_1 * K_2\|_\infty = 3. \tag{37}$$

It follows that

$$\phi(A_1) = (\|A_1 * K_1\|_\infty, \|A_1 * K_2\|_\infty) = (1, 3). \tag{38}$$

(b) As $\phi(A_1) = \phi(A_2)$, X is not in ϕ-general position. $\{X^+, X^-\}$ is ϕ-separable. Indeed, consider $w = (-1, 1)$. Then,

$$\langle \phi(A_1), w\rangle = \langle \phi(A_2), w\rangle = 2, \tag{39}$$

and

$$\langle \phi(A_3), w\rangle = \langle \phi(A_4), w\rangle = -2. \tag{40}$$

(c) $\|K_1\|_1 = \|K_2\|_1 = \frac{1}{11}\sum_{i=1}^{11} 1 = 1$. By Lemma 4 in the Handout, for all $A \in [0,1]^{28\times 28}$, $\|A * K_1\|_\infty \le \|A\|_\infty \|K_1\|_1 \le \|A\|_\infty$ and $\|A * K_2\|_\infty \le \|A\|_\infty \|K_2\|_1 \le \|A\|_\infty$. As $A \in [0,1]^{28\times 28}$, it follows that $\phi(A) \in [0,1]^2$, for all $A \in X$.

(d) In subproblem (a), we have seen that convolution with a matrix of ones lined up horizontally tends to produce maximal values if the image contains a horizontal line, and conversely, convolution with a matrix of ones lined up vertically tends to produce maximal values if the image contains a vertical line.

The red points represent the value of $\phi(A)$ for $A \in X^-$. X^- contains images of handwritten ones, which are approximately vertical lines. Hence, for $A \in X^-$,

$A * K_1$ yields small values (close to 0) while $A * K_2$ yields higher values (close to 1), and therefore $\|A * K_1\|_\infty \simeq 0$ while $\|A * K_2\|_\infty \simeq 1$. This explains why $\phi(A) \simeq (0, 1)$ for $A \in X^-$.

The blue points represent the value of $\phi(A)$ for $A \in X^+$. X^+ contains images of handwritten zeros, which are approximately a combination of horizontal and vertical lines. Hence, for $A \in X^+$, both $A * K_1$ and $A * K_2$ yield high values (close to 1), and therefore $\|A * K_1\|_\infty \simeq \|A * K_2\|_\infty \simeq 1$. This explains why $\phi(A) \simeq (1, 1)$ for $A \in X^+$.

(e) The distribution of blue and red points clearly overlap, so it is not possible to separate $\{\phi(X^+), \phi(X^-)\}$ with a line. Therefore, X is not ϕ-separable. We can design a map ϕ' which could lead to better separation properties as follows. ϕ is designed to "spot" the horizontal and vertical lines on images, through the convolution with K_1 and K_2. However, handwritten zeros and ones are not composed of horizontal and vertical lines: ones are not perfectly vertical, but rather deviate a little bit to the right, and zeros have many components, such as curved lines and edges. A strategy to get a better separation would be to add more coordinates to the output of ϕ to detect more diverse features. For example, one can choose $\phi' : \mathbb{R}^{28\times 28} \to \mathbb{R}^{12}$, according to

$$\phi'(A) = (A * K_i)_{i\in\{1,\dots,12\}}, \quad \forall A \in \mathbb{R}^{28\times 28}, \tag{41}$$

where the K_i are matrices. The first two coordinates would remain the same, but the next ones would use convolutions with other matrices, with the purpose of detecting inclined lines, curved lines and edges. Here is an example of the K_i, for $i = 3, \dots, 12$:

$$K_3 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad K_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \tag{42}$$

are designed to detect inclined lines,

$$K_5 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad K_6 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad K_7 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}, \quad K_8 = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

are designed to detect curved lines, and

$$K_9 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad K_{10} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad K_{11} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad K_{12} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \tag{43}$$

are designed to detect edges.