ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Prof. Dr. Helmut Bölcskei
Chair for Mathematical Information Sciences

Solutions to the Exam on Neural Network Theory February 6, 2020

Problem 1

(a) We have to establish that $K(\boldsymbol{x},\boldsymbol{y}) = K(\boldsymbol{y},\boldsymbol{x})$, which follows from

$$\begin{aligned} K(\boldsymbol{x},\boldsymbol{y}) &= \alpha x_1^2 y_1^2 + \beta x_2^2 y_2^2 \\ &= \alpha y_1^2 x_1^2 + \beta y_2^2 x_2^2 \\ &= K(\boldsymbol{y},\boldsymbol{x}), \quad \text{for all } \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^2. \end{aligned}$$

(b) Let $\varepsilon > 0$ and $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^2$. We have to show that there exists a $\delta = \delta(\varepsilon,\boldsymbol{x},\boldsymbol{y}) > 0$ such that if $\boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^2$ satisfy

$$\left\| \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} - \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} \right\|_2 < \delta, \tag{1}$$

then $|K(\boldsymbol{x},\boldsymbol{y}) - K(\boldsymbol{u},\boldsymbol{v})| < \varepsilon$. Since

$$\begin{aligned} |K(\boldsymbol{x},\boldsymbol{y}) - K(\boldsymbol{u},\boldsymbol{v})| &= \big|\alpha x_1^2 y_1^2 + \beta x_2^2 y_2^2 - \alpha u_1^2 v_1^2 - \beta u_2^2 v_2^2\big| \\ &\leq \alpha\big|x_1^2 y_1^2 - u_1^2 v_1^2\big| + \beta\big|x_2^2 y_2^2 - u_2^2 v_2^2\big|, \end{aligned}$$

it is sufficient to choose $\delta > 0$ such that

$$\big|x_1^2 y_1^2 - u_1^2 v_1^2\big| < \frac{\varepsilon}{2\alpha}$$

and

$$\big|x_2^2 y_2^2 - u_2^2 v_2^2\big| < \frac{\varepsilon}{2\beta}.$$

Suppose that (1) holds. This implies that

$$|x_1 - u_1| < \delta, \quad |y_1 - v_1| < \delta, \quad |x_2 - u_2| < \delta, \quad \text{and} \quad |y_2 - v_2| < \delta.$$

It now follows that

$$\begin{aligned} \big|x_1^2 y_1^2 - u_1^2 v_1^2\big| &= \big|x_1^2 y_1^2 - x_1^2 v_1^2 + x_1^2 v_1^2 - u_1^2 v_1^2\big| \\ &\leq x_1^2\big|y_1^2 - v_1^2\big| + v_1^2\big|x_1^2 - u_1^2\big| \\ &= x_1^2|(y_1 + v_1)(y_1 - v_1)| + v_1^2|(x_1 + u_1)(x_1 - u_1)| \\ &\leq x_1^2(|y_1| + |v_1|)\delta + v_1^2(|x_1| + |u_1|)\delta \\ &\leq \big(x_1^2(2|y_1| + \delta) + (|y_1| + \delta)^2(2|x_1| + \delta)\big)\delta \\ &< \frac{\varepsilon}{2\alpha} \end{aligned}$$

provided that $\delta < C_1$ with

$$C_1 = \min\left\{1, \frac{\varepsilon}{2\alpha(x_1^2(2|y_1| + 1) + (|y_1| + 1)^2(2|x_1| + 1))}\right\}.$$

Swapping the roles of x_1 and x_2, y_1 and y_2, u_1 and u_2, v_1 and v_2, and α and β and using the same line of arguments yields

$$|x_2^2 y_2^2 - u_2^2 v_2^2| < \frac{\varepsilon}{2\beta}$$

provided that $\delta < C_2$ with

$$C_2 = \min\left\{1, \frac{\varepsilon}{2\beta(x_2^2(2|y_2| + 1) + (|y_2| + 1)^2(2|x_2| + 1))}\right\}.$$

We conclude that if $\delta < \min\{C_1, C_2\}$, then

$$\left\| \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} - \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} \right\|_2 < \delta$$

implies $|K(\boldsymbol{x}, \boldsymbol{y}) - K(\boldsymbol{u}, \boldsymbol{v})| < \varepsilon$, which establishes continuity of $K(\boldsymbol{x}, \boldsymbol{y})$.

(c) We have to show that for every $k \in \mathbb{N}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^2$, the $k \times k$ Gramian matrix

$$\boldsymbol{K}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) = \begin{pmatrix} K(\boldsymbol{x}_1, \boldsymbol{x}_1) & K(\boldsymbol{x}_1, \boldsymbol{x}_2) & \ldots & K(\boldsymbol{x}_1, \boldsymbol{x}_k) \\ K(\boldsymbol{x}_2, \boldsymbol{x}_1) & K(\boldsymbol{x}_2, \boldsymbol{x}_2) & \ldots & K(\boldsymbol{x}_2, \boldsymbol{x}_k) \\ \vdots & \vdots & \ldots & \vdots \\ K(\boldsymbol{x}_k, \boldsymbol{x}_1) & K(\boldsymbol{x}_k, \boldsymbol{x}_2) & \ldots & K(\boldsymbol{x}_k, \boldsymbol{x}_k) \end{pmatrix}$$

is positive semidefinite. To this end, consider the mapping $\Phi\colon \mathbb{R}^2 \to \mathbb{R}^2$ defined as

$$\Phi(\boldsymbol{x}) = \begin{pmatrix} \sqrt{\alpha} x_1^2 \\ \sqrt{\beta} x_2^2 \end{pmatrix}$$

and note that $K(\boldsymbol{x}, \boldsymbol{y}) = \Phi^\mathsf{T}(\boldsymbol{x})\Phi(\boldsymbol{y})$. It follows that, for every $k \in \mathbb{N}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^2$, we have

$$\begin{aligned} \boldsymbol{c}^\mathsf{T} \boldsymbol{K}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)\, \boldsymbol{c} &= \sum_{i,j=1}^{k} c_i c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= \sum_{i,j=1}^{k} c_i c_j \Phi^\mathsf{T}(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j) \\ &= \left(\sum_{i=1}^{k} c_i \Phi(\boldsymbol{x}_i)\right)^\mathsf{T} \left(\sum_{j=1}^{k} c_j \Phi(\boldsymbol{x}_j)\right) \\ &\geq 0, \quad \text{for all } \boldsymbol{c} = (c_1 \ldots c_k)^\mathsf{T} \in \mathbb{R}^k, \end{aligned}$$

which proves that $K(\boldsymbol{x}, \boldsymbol{y})$ is positive semidefinite.

(d) For every $\boldsymbol{y} \in \mathbb{R}^2$, we define the function $K_{\boldsymbol{y}} \colon \mathbb{R}^2 \to \mathbb{R}$ according to $K_{\boldsymbol{y}}(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{y})$ and set

$$\mathcal{H}_0 = \operatorname{span}\{K_{\boldsymbol{y}} : \boldsymbol{y} \in \mathbb{R}^2\}.$$

Let $\boldsymbol{e}_1 = (1 \ 0)^\mathsf{T}$ and $\boldsymbol{e}_2 = (0 \ 1)^\mathsf{T}$ with corresponding

$$K_{\boldsymbol{e}_1}(\boldsymbol{x}) = K(\boldsymbol{e}_1, \boldsymbol{x}) = \alpha x_1^2$$

and

$$K_{\boldsymbol{e}_2}(\boldsymbol{x}) = K(\boldsymbol{e}_2, \boldsymbol{x}) = \beta x_2^2,$$

respectively. Since

$$\begin{aligned} K_{\boldsymbol{a}}(\boldsymbol{x}) &= \alpha x_1^2 a_1^2 + \beta x_2^2 a_2^2 \\ &= a_1^2 K_{\boldsymbol{e}_1}(\boldsymbol{x}) + a_2^2 K_{\boldsymbol{e}_2}(\boldsymbol{x}), \quad \text{for all } \boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \in \mathbb{R}^2, \end{aligned}$$

it follows that $\mathcal{H}_0 = \operatorname{span}\{K_{\boldsymbol{e}_1}, K_{\boldsymbol{e}_2}\}$. We conclude that the reproducing kernel Hilbert space $\mathcal{H}_K$ corresponding to the kernel $K(\boldsymbol{x}, \boldsymbol{y})$ is given by

$$\mathcal{H}_K = \overline{\mathcal{H}_0} = \overline{\operatorname{span}\{K_{\boldsymbol{e}_1}, K_{\boldsymbol{e}_2}\}} = \operatorname{span}\{K_{\boldsymbol{e}_1}, K_{\boldsymbol{e}_2}\}.$$

(e) Consider $K_{\boldsymbol{e}_1}$ and $K_{\boldsymbol{e}_2}$ from subproblem (d). Using the reproducing property of $\mathcal{H}_K$, we have

$$\begin{aligned} \langle K_{\boldsymbol{e}_1}, K_{\boldsymbol{e}_1} \rangle_{\mathcal{H}_K} &= K(\boldsymbol{e}_1, \boldsymbol{e}_1) = \alpha, \\ \langle K_{\boldsymbol{e}_2}, K_{\boldsymbol{e}_2} \rangle_{\mathcal{H}_K} &= K(\boldsymbol{e}_2, \boldsymbol{e}_2) = \beta, \end{aligned}$$

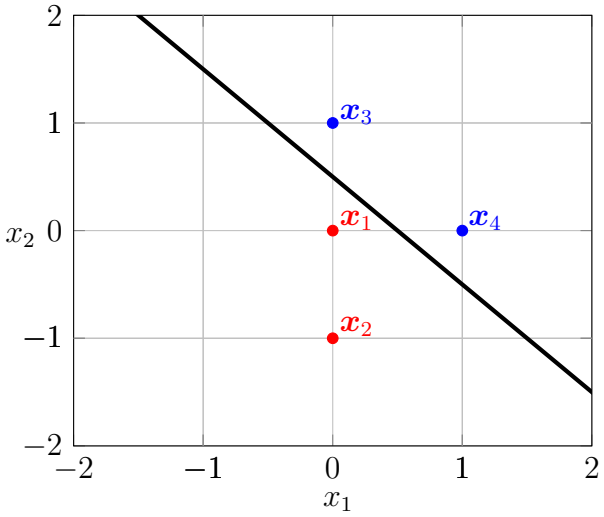
and

$$\langle K_{\boldsymbol{e}_1}, K_{\boldsymbol{e}_2} \rangle_{\mathcal{H}_K} = K(\boldsymbol{e}_1, \boldsymbol{e}_2) = 0.$$

This implies that $\{K_{\boldsymbol{e}_1}/\sqrt{\alpha}, K_{\boldsymbol{e}_2}/\sqrt{\beta}\}$ is an orthonormal basis for the reproducing kernel Hilbert space $\mathcal{H}_K = \operatorname{span}\{K_{\boldsymbol{e}_1}, K_{\boldsymbol{e}_2}\}$ corresponding to the kernel $K(\boldsymbol{x}, \boldsymbol{y})$.

Problem 2

(a) The following figure depicts $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ (red points) and $\{\boldsymbol{x}_3, \boldsymbol{x}_4\}$ (blue points) together with a separating straight line (thick black line):



(b) The Lagrange function $L(\boldsymbol{w}, b, \boldsymbol{c})$ id given by

$$L(\boldsymbol{w}, b, \boldsymbol{c}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{4} c_i(1 - y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle_2 - b)). \tag{2}$$

(c) Setting $\nabla_{\boldsymbol{w}} L(\boldsymbol{w}, b, \boldsymbol{c}) = 0$ and $\nabla_b L(\boldsymbol{w}, b, \boldsymbol{c}) = 0$ yields

$$\boldsymbol{w} = \sum_{i=1}^{4} c_i y_i \boldsymbol{x}_i \tag{3}$$

and

$$\sum_{i=1}^{4} c_i y_i = 0, \tag{4}$$

respectively. Using (3) and (4) in (2) results in

$$\begin{aligned} g(\boldsymbol{c}) &= \min_{\boldsymbol{w} \in \mathbb{R}^2, b \in \mathbb{R}} L(\boldsymbol{w}, b, \boldsymbol{c}) \\ &= \sum_{i=1}^{4} c_i - \frac{1}{2} \sum_{i,j=1}^{4} c_i y_i c_j y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_2 \\ &= \boldsymbol{a}^\mathsf{T} \boldsymbol{c} - \frac{1}{2} \boldsymbol{c}^\mathsf{T} \boldsymbol{A}\, \boldsymbol{c} \end{aligned}$$

with $\boldsymbol{a} = (1 \;\; 1 \;\; 1 \;\; 1)^\mathsf{T}$ and

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

(d) It follows from subproblem (c) that the Lagrange dual function $g(\boldsymbol{c})$ can be written as

$$g(\boldsymbol{c}) = c_1 + c_2 + c_3 + c_4 - \frac{1}{2}(c_2^2 + c_3^2 + 2c_2c_3 + c_4^2).$$

Using the constraint

$$\sum_{i=1}^{4} c_i y_i = c_1 + c_2 - c_3 - c_4 = 0, \tag{5}$$

we can write

$$g(\boldsymbol{c}) = 2(c_3 + c_4) - \frac{1}{2}(c_3^2 + c_4^2) - \frac{1}{2}(c_2^2 + 2c_2c_3).$$

It follows that

$$\begin{aligned} g(\boldsymbol{c}) &\leq 2(c_3 + c_4) - \frac{1}{2}(c_3^2 + c_4^2) \\ &\leq 4 \end{aligned}$$

with equalities in both steps for $\tilde{c}_2 = 0$ and $\tilde{c}_3 = \tilde{c}_4 = 2$. Using (5) again, we get $\tilde{c}_1 = \tilde{c}_3 + \tilde{c}_4 - \tilde{c}_2 = 4$. The solution of the Lagrange dual problem is therefore given by $\tilde{\boldsymbol{c}} = (4 \quad 0 \quad 2 \quad 2)^\mathsf{T}$. A vector $\boldsymbol{x}_i$ is a support vector if and only if the corresponding $\tilde{c}_i$ is strictly positive. The support vectors are therefore given by $\{\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_4\}$.

(e) Using (3) and the solution $\tilde{\boldsymbol{c}} = (4 \;\; 0 \;\; 2 \;\; 2)^\mathsf{T}$ from subproblem (d), we obtain the following solution $\tilde{\boldsymbol{w}}$ of the optimization problem in subproblem (b):

$$\tilde{\boldsymbol{w}} = 4\boldsymbol{x}_1 - 2\boldsymbol{x}_3 - 2\boldsymbol{x}_4 = -2 \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The solution $\tilde{b}$ can be obtained from

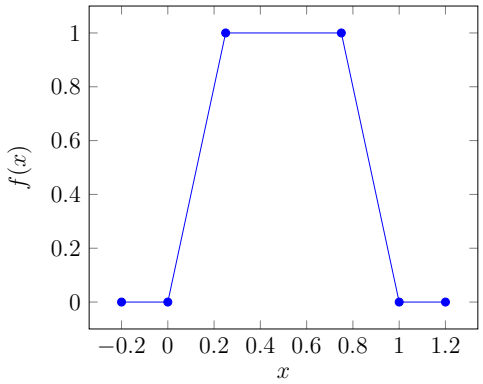
$$\begin{aligned} \tilde{b} &= \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_1 \rangle_2 - y_1 \\ &= -1, \end{aligned}$$

which follows from the fact that for the support vector $\boldsymbol{x}_1$ we have $\tilde{c}_1 > 0$ and, therefore, the corresponding inequality constraint $y_1(\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_1 \rangle_2 - \tilde{b}) \geq 1$ must be satisfied with equality. The hard margin binary classifier $g_{\text{hm}}(\boldsymbol{x})$ therefore has the following form:

$$\begin{aligned} g_{\text{hm}}(\boldsymbol{x}) &= (\langle \tilde{\boldsymbol{w}}, \boldsymbol{x} \rangle_2 - \tilde{b}) \\ &= -2 \begin{pmatrix} 1 & 1 \end{pmatrix} \boldsymbol{x} + 1. \end{aligned}$$

Problem 3

(a) The following figure depicts $f(x)$.



(b) Denote the ReLU function as $\rho(x) = \max\{0, x\}$. The function $f(x)$ can be realized as a linear combination of shifted ReLU functions according to

$$f(x) = \rho(4x) - \rho(4x-1) - \rho(4x-3) + \rho(4x-4).$$

This function can be realized through a depth-2 ReLU network according to

$$\Phi(x) = W_2(\rho(W_1(x)))$$

with

$$W_1(x) = \begin{pmatrix} 4 \\ 4 \\ 4 \\ 4 \end{pmatrix} x + \begin{pmatrix} 0 \\ -1 \\ -3 \\ -4 \end{pmatrix}, \quad W_2(x) = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}.$$

The network Φ has depth 2, width 4, and connectivity 11.

(c) One way to realize $h(x)$ through a ReLU-network is according to

$$\begin{aligned} h(x) = f(4x) + f(4x-3) = \rho(16x) - \rho(16x-1) - \rho(16x-3) + \rho(16x-4) \\ + \rho(16x-12) - \rho(16x-13) - \rho(16x-15) + \rho(16x-16). \end{aligned}$$

This corresponds to the depth-2 ReLU-network

$$\Phi_1(x) = W_2(\rho(W_1(x)))$$

with

$$W_1(x) = \begin{pmatrix} 16 \\ 16 \\ 16 \\ 16 \\ 16 \\ 16 \\ 16 \\ 16 \end{pmatrix} x + \begin{pmatrix} 0 \\ -1 \\ -3 \\ -4 \\ -12 \\ -13 \\ -15 \\ -16 \end{pmatrix}, \quad W_2(x) = \begin{pmatrix} 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{pmatrix}.$$

The network Φ_1 has depth 2, width 8, and connectivity 23.

(d) An alternative way to realize $h(x)$ is to note that $h(x) = f(f(x))$. The function $f(x)$ is realized through the network $\Phi(x) = W_2(\rho(W_1(x)))$ in subproblem (b), therefore $h(x) = f(f(x))$ can be realized through the network

$$\Phi_2(x) = \Phi(\Phi(x)) = W_2(\rho(W_1(W_2(\rho(W_1(x)))))) = W_3'(\rho(W_2'(\rho(W_1'(x)))))$$

with

$$W_1'(x) = W_1(x) = \begin{pmatrix} 4 \\ 4 \\ 4 \\ 4 \end{pmatrix} x + \begin{pmatrix} 0 \\ -1 \\ -3 \\ -4 \end{pmatrix}, \quad W_3'(x) = W_2(x) = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix},$$

$$W_2'(x) = W_1(W_2(x)) = \begin{pmatrix} 4 \\ 4 \\ 4 \\ 4 \end{pmatrix} \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} 0 \\ -1 \\ -3 \\ -4 \end{pmatrix}$$

$$= \begin{pmatrix} 4 & -4 & -4 & 4 \\ 4 & -4 & -4 & 4 \\ 4 & -4 & -4 & 4 \\ 4 & -4 & -4 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} 0 \\ -1 \\ -3 \\ -4 \end{pmatrix}.$$

The network Φ_2 has depth 3, width 4, and connectivity 30.

Problem 4

(a) An ϵ-covering of the compact set $\mathcal{C} \subseteq \mathcal{X}$ with respect to the metric ρ is a set $\{x_1, \ldots, x_N\} \subset \mathcal{C}$ such that for each $x \in \mathcal{C}$, there exists an $i \in [1, N]$ such that $\rho(x, x_i) \leq \epsilon$. The ϵ-covering number $N(\epsilon; \mathcal{C}, \rho)$ is the cardinality of the smallest ϵ-covering.

An ϵ-packing of a compact set $\mathcal{C} \subseteq \mathcal{X}$ with respect to the metric ρ is a set $\{x_1, \ldots, x_N\} \subset \mathcal{C}$ such that $\rho(x_i, x_j) > \epsilon$, for all distinct i, j. The ϵ-packing number $M(\epsilon; \mathcal{C}, \rho)$ is the cardinality of the largest ϵ-packing.

(b) Let $\{x_1, \ldots, x_N\}$ be an ϵ-packing of $\mathcal{C}$. The ℓ_2-balls centered at the x_i and of radius $\epsilon/2$ are referred to as packing balls. As the distance between each pair in the packing is at least ϵ, the packing balls are mutually disjoint. Therefore, the maximum volume covered by the packing balls is $M(\epsilon; \mathcal{C}, \|\cdot\|_2)\ c(\epsilon/2)^d$, where c is a constant. Since each point in the packing must be in $\mathcal{C}$, all the packing balls are, indeed, contained in an ℓ_2-ball of radius $1 + \epsilon/2$. Hence, we get, for all $\epsilon \leq 2$,

$$M(\epsilon; \mathcal{C}, \|\cdot\|_2)\ c\ (\epsilon/2)^d \leq c\ (1 + \epsilon/2)^d \implies M(\epsilon; \mathcal{C}, \|\cdot\|_2) \leq \left(\frac{2}{\epsilon} + 1\right)^d \leq \left(\frac{4}{\epsilon}\right)^d.$$