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 17, 2020

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 ETHZ "Disziplinarordnung RSETHZ 361.1" 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 and your Legi-number below.
3. Please place an identification document on 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. All sheets, including those containing problem statements, must be handed in. Please sign this page.

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

Legi-No.: ....

Signature: ....

Problem 1

Let $\boldsymbol{b}_1, \dots, \boldsymbol{b}_q \in \mathbb{R}^n$ with $q > n$. For $k \in \{n, \dots, q\}$, set $\boldsymbol{B}_k = (\boldsymbol{b}_1 \dots \boldsymbol{b}_k)$ and suppose that $\boldsymbol{B}_n$ is invertible. Define the function $g_{\boldsymbol{B}_n} : \mathbb{R}^n \to \mathbb{R}$ according to

$$g_{\boldsymbol{B}_n}(\boldsymbol{x}) = \begin{cases} \frac{1}{|\det(\boldsymbol{B}_n)|}, & \text{if } \boldsymbol{x} \in \Big\{ \boldsymbol{B}_n \boldsymbol{y} : \boldsymbol{y} \in \big[-\frac{1}{2}, \frac{1}{2}\big]^n \Big\} \\ 0, & \text{else.} \end{cases}$$

The box spline $g_{\boldsymbol{B}_q}(\boldsymbol{x})$ is defined recursively by setting

$$g_{\boldsymbol{B}_{n+j}}(\boldsymbol{x}) = \int_{-\frac{1}{2}}^{\frac{1}{2}} g_{\boldsymbol{B}_{n+j-1}}(\boldsymbol{x} - t\boldsymbol{b}_{n+j}) \,\mathrm{d}t, \quad \text{for } j = 1, \dots, q-n.$$

Box splines are generally used for multivariate approximation/interpolation. The purpose of this problem is to construct a positive semidefinite kernel function from $g_{\boldsymbol{B}_q}(\boldsymbol{x})$.

You will need the following mathematical preliminaries. The n-dimensional Fourier transform $\hat{h}$ of a function $h \colon \mathbb{R}^n \to \mathbb{R}$ is defined according to

$$\hat{h}(\xi) = \int_{\mathbb{R}^n} h(\boldsymbol{x}) e^{-i\xi^{\mathrm{T}}\boldsymbol{x}} \,\mathrm{d}\boldsymbol{x}$$

with its inverse transform

$$h(\boldsymbol{x}) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \hat{h}(\xi) e^{i\xi^{\mathrm{T}}\boldsymbol{x}} \,\mathrm{d}\xi.$$

Here, i is the imaginary unit and ξ^{T} denotes the transpose of the vector ξ. The convolution of two functions $f, g \colon \mathbb{R}^n \to \mathbb{R}$ is defined as

$$(f \star g)(\boldsymbol{x}) = \int_{\mathbb{R}^n} f(\boldsymbol{y}) g(\boldsymbol{x} - \boldsymbol{y}) \,\mathrm{d}\boldsymbol{y}.$$

You can assume that all functions in this problem satisfy the necessary integrability conditions for the n-dimensonal (inverse) Fourier transform and the convolution to be well defined. You are also allowed, whenever required, to exchange the order of integration without justification.

(a) Suppose that $q = n + 1$. Provide an explicit expression for the Fourier transform $\hat{g}_{\boldsymbol{B}_{n+1}}(\xi)$ of the box spline $g_{\boldsymbol{B}_{n+1}}(\boldsymbol{x})$.

(b) Show that for general $q > n$ the Fourier transform $\hat{g}_{\boldsymbol{B}_q}(\xi)$ of the box spline $g_{\boldsymbol{B}_q}(\boldsymbol{x})$ is given by

$$\hat{g}_{\boldsymbol{B}_q}(\xi) = \prod_{i=1}^{q} \left(\frac{2}{\xi^{\mathrm{T}}\boldsymbol{b}_i} \sin\left(\frac{\xi^{\mathrm{T}}\boldsymbol{b}_i}{2} \right) \right).$$

Hint. *Use the result from subproblem (a) for $q = n + 1$ and proceed by induction over q.*

(c) Compute the Fourier transform $\hat{k}(\xi)$ of the function $k(\boldsymbol{x}) = (g_{\boldsymbol{B}_q} \star g_{\boldsymbol{B}_q})(\boldsymbol{x})$.

(d) Let $k(\boldsymbol{x})$ be the function defined in subproblem (c) and set $K(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{x} - \boldsymbol{y})$, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Prove that for every $k \in \mathbb{N}$ and all $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k \in \mathbb{R}^n$, the $k \times k$ Gramian matrix

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

is positive semidefinite, i.e., $K(\boldsymbol{x}, \boldsymbol{y})$ is a positive semidefinite kernel.

Hint. *Express $k(\boldsymbol{x})$ in terms of its Fourier transform $\hat{k}(\xi)$.*

Problem 2

In this problem, you will explicitly solve the support vector machine algorithm to construct the best possible straight line separating two sets of vectors in $\mathbb{R}^2$. Specifically, consider the vectors

$$\boldsymbol{x}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{x}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{x}_3 = \begin{pmatrix} 1 \\ \lambda \end{pmatrix},$$

where $\lambda > 0$ is a positive parameter, and let $y_1 = 1$, $y_2 = 1$, and $y_3 = -1$.

(a) Consider the optimization problem

$$\min_{\boldsymbol{w} \in \mathbb{R}^2, b \in \mathbb{R}} \tfrac{1}{2}\|\boldsymbol{w}\|_2^2$$
$$\text{such that } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle_2 - b) \geq 1, \quad \text{for } i = 1, 2, 3.$$

Write down the Lagrange function $L(\boldsymbol{w}, b, \boldsymbol{c})$ corresponding to this optimization problem, where $\boldsymbol{c} = (c_1\ c_2\ c_3)^\top$ is the vector containing the Lagrange multipliers $c_1, c_2, c_3 \in \mathbb{R}$.

(b) Consider the Lagrange function $L(\boldsymbol{w}, b, \boldsymbol{c})$ from subproblem (a). The corresponding Lagrange dual function has the form

$$g(\boldsymbol{c}) = \boldsymbol{a}^\top \boldsymbol{c} - \frac{1}{2}\boldsymbol{c}^\top \boldsymbol{A}\, \boldsymbol{c},$$

where $\boldsymbol{a} \in \mathbb{R}^3$ and $\boldsymbol{A} \in \mathbb{R}^{3\times 3}$. Compute $\boldsymbol{a}$ and $\boldsymbol{A}$ explicitly.

(c) Consider the Lagrange dual function $g(\boldsymbol{c})$ with explicit $\boldsymbol{a}$ and $\boldsymbol{A}$ from subproblem (b). For general $\lambda > 0$, solve the corresponding Lagrange dual problem

$$\max_{\boldsymbol{c} \in \mathbb{R}^3} g(\boldsymbol{c})$$
$$\text{such that } c_i \geq 0, \quad \text{for } i = 1, 2, 3 \ \text{ and } \ \sum_{i=1}^{3} c_i y_i = 0.$$

Hint. *The function $-g(\boldsymbol{c})$ is convex. To solve the Lagrange dual problem, you can therefore consider the Lagrange function*

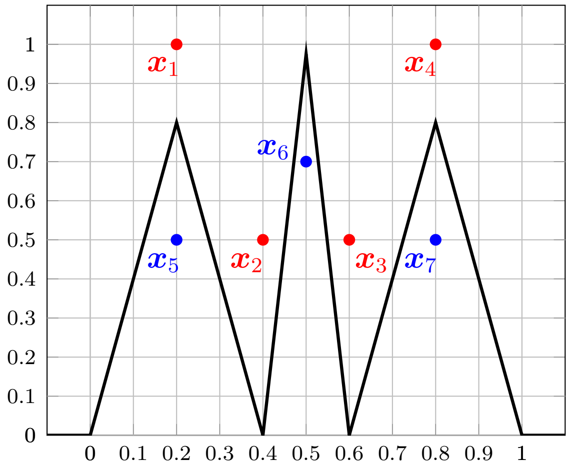
$$\tilde{L}(\boldsymbol{c}, \boldsymbol{\mu}, \gamma) = -g(\boldsymbol{c}) - \boldsymbol{\mu}^\top \boldsymbol{c} + \gamma\left(\sum_{i=1}^{3} c_i y_i\right),$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^\top \in \mathbb{R}^3$ and $\gamma \in \mathbb{R}$ are Lagrange multipliers, and solve the KKT conditions. Consider the two cases $\lambda \in (0, 1)$ and $\lambda \geq 1$ separately and use the ansatz: $c_i > 0$ if and only if $\boldsymbol{x}_i$ is a support vector. The support vectors can best be identified from a picture containing $\boldsymbol{x}_1, \boldsymbol{x}_2$, and $\boldsymbol{x}_3$ and the separating straight line between $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ and $\{\boldsymbol{x}_3\}$ of largest possible margins for the specific choices of $\lambda \in (0, 1)$ and $\lambda \geq 1$.

(d) For general $\lambda \in (0, 1)$, compute a solution $(\boldsymbol{w}^*, b^*)$ of the optimization problem in subproblem (a) and write down the expression for the corresponding hard margin binary classifier.

Problem 3

Consider the following figure depicting the vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4\}$ in red and the vectors $\{\boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7\}$ in blue together with a corresponding class-separating piecewise linear function $f(x)$ (thick black line):



(a) Realize $f(x)$ through a depth-2 ReLU network (see the Handout for the definition of a depth-L ReLU network). Specify the width and the connectivity of the resulting network.

(b) Recall the sawtooth function $g(x) = \rho(2x) - \rho(4x - 2) + \rho(2x - 2)$ studied in class. Plot the function $f(g(x))$. Determine the value of x the function $f(g(x))$ is symmetric around.

(c) In subproblem (a) the class-separating function was realized by a ReLU network of depth 2. Using the sawtooth function from subproblem (b), find a deeper ReLU network of width 3 that realizes $f(x)$. Specify the depth and the connectivity of this network.
Hint. *Exploit the mirror symmetry of $f(x)$ around the point $x = 0.5$ and follow the idea underlying subproblem (b).*

Problem 4

(a) Let $(\mathcal{X}, \rho)$ be a metric space and $\mathcal{C}$ a compact set in $\mathcal{X}$.

(i) State the definitions of an ϵ-covering of $\mathcal{C}$ and of the ϵ-covering number $N(\epsilon; \mathcal{C}, \rho)$.

(ii) State the definitions of an ϵ-packing of $\mathcal{C}$ and of the ϵ-packing number $M(\epsilon; \mathcal{C}, \rho)$.

(iii) Order the following four quantities

$$N(\epsilon; \mathcal{C}, \rho), N(2\epsilon; \mathcal{C}, \rho), M(\epsilon; \mathcal{C}, \rho), M(2\epsilon; \mathcal{C}, \rho).$$

(b) Fix $n \in \mathbb{N}$, $\epsilon < 2^{-n}$, and consider the interval $I_n := [-2^{-n}, 2^{-n}] \in \mathbb{R}$ equipped with the metric $\rho_1(x, x') = |x - x'|$. Let $K := \log_2(1/\epsilon)$ and $L := \lceil 2^{K-n} \rceil$.

(i) Construct an ϵ-covering $A_n(\epsilon)$ of the interval I_n as follows. Divide I_n into L sub-intervals of equal length and show that the corresponding interval centers constitute an ϵ-covering of I_n.

(ii) Construct a 2ϵ-packing $P_n(\epsilon)$ of the interval I_n such that $|A_n| = |P_n|$.
Hint. *Divide I_n into $L - 1$ sub-intervals and keep in mind that $|A_n| = |P_n|$.*

(iii) Compute $N(\epsilon; I_n, \rho_1)$.

(c) Let $\mathcal{C} = \{f : \mathbb{N} \to \mathbb{R};\ f(n) \in [-2^{-n}, 2^{-n}],\ \forall n \in \mathbb{N}\}$ be a set of sequences in the space of bounded sequences equipped with the metric $\rho_2(f, g) = \sup_{n \in \mathbb{N}} |f(n) - g(n)|$.

(i) For $\epsilon \leq 1/2$, construct an ϵ-covering of $\mathcal{C}$.

(ii) Show that for every $\epsilon \leq 1/2$,

$$N(\epsilon; \mathcal{C}, \rho_2) \leq \left(\frac{1}{\epsilon}\right)^{\frac{1}{2}\log_2(1/\epsilon) + C},$$

for some $C > 0$, which does not depend on ϵ.

Hints.

(i) Use the result from subproblem (b.i) for $\epsilon < 2^{-n}$. For $\epsilon \geq 2^{-n}$, you can use, without proof, that $A_n(\epsilon) = \{0\}$ constitutes an ϵ-covering with $N(\epsilon; I_n, \rho_1) = 1$.

(ii) You may use, without proof, that $\lceil 2^{K-n} \rceil \leq 2^{\lceil K \rceil - n}$, for $n \leq \lceil K \rceil - 1$.