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 6, 2024

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 7 pages including this page. Please verify that you have received all 7 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 the solution to a subproblem, you may, whenever applicable, nonetheless assume its conclusion in the ensuing subproblems.
6. All results in the Handout can be used without proof.

After the exam:

7. 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.
8. 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.
9. 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)

The clipped ReLU function $\sigma : \mathbb{R} \to \mathbb{R}$ is defined according to

$$\sigma(x) = \begin{cases} 0, & \text{for } x < 0, \\ x, & \text{for } 0 \leq x \leq 1, \\ 1, & \text{for } x > 1. \end{cases}$$

(a) (4 points) Realize the clipped ReLU function through a ReLU network Φ (see the Handout for the definition of a ReLU network). Specify $\mathcal{L}(\Phi)$, $\mathcal{W}(\Phi)$, and $\mathcal{M}(\Phi)$.

(b) (6 points) Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \sigma(4x) - \sigma(2x - 1/2) + \sigma(4x - 3).$$

Sketch the function f. Find a ReLU network Φ^f satisfying $\Phi^f(x) = f(x)$, for all $x \in \mathbb{R}$, with $\mathcal{L}(\Phi^f) = 2$, and specify $\mathcal{W}(\Phi^f)$, $\mathcal{M}(\Phi^f)$, and $\mathcal{B}(\Phi^f)$.

(c) (4 points) The two-dimensional function $g : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$g(x, y) = \sigma(\sigma(-2x - y + 1) + \sigma(0.5x - 2y)).$$

Find a ReLU network Φ^g satisfying $\Phi^g(x, y) = g(x, y)$, for all $x, y \in \mathbb{R}$, with $\mathcal{L}(\Phi^g) = 3$.
Hint: *Use the result from subproblem (a).*

(d) (6 points) Define the operations $\wedge$ and $\vee$ on $\mathbb{R}$ according to

$$x \vee y := \max\{x, y\}$$
$$x \wedge y := \min\{x, y\}.$$

Find ReLU networks $\Phi^\vee$ and $\Phi^\wedge$ satisfying $\Phi^\vee(x, y) = x \vee y$ and $\Phi^\wedge(x, y) = x \wedge y$, for all $x, y \in \mathbb{R}$.

(e) (5 points) The three-dimensional function $h : \mathbb{R}^3 \to \mathbb{R}$ is given by

$$h(x, y, z) = \min\{x, y, z\}.$$

Find a ReLU network Φ^h satisfying $\Phi^h(x, y, z) = h(x, y, z)$, for all $x, y, z \in \mathbb{R}$.

Hint: *Write h in the form of nested minima, i.e.,* $\min\{x, y, z\} = \min\{\min\{x, y\}, z\}$.

Problem 2 (25 points)

Consider the following parametric class of functions

$$\mathcal{F} = \{f_{\theta,\theta'} : [0,1] \to \mathbb{R} | \theta, \theta' \in [0,1]\},$$

where for $\theta, \theta' \in [0,1]$, we set $f_{\theta,\theta'}(x) := 1 - e^{-\theta x} + \theta', x \in [0,1]$. We consider covering numbers and packing numbers with respect to the metric

$$\rho_\infty(f, g) := \sup_{x \in [0,1]} |f(x) - g(x)|.$$

(a) (4 points) State the definition of an ϵ-covering of $\mathcal{F}$ with respect to the metric ρ_∞ and of the corresponding ϵ-covering number $N(\varepsilon; \mathcal{F}, \rho_\infty)$.

(b) (5 points) Show that, for all $\epsilon \geq 2$, it holds that

$$N(\varepsilon; \mathcal{F}, \rho_\infty) = 1.$$

(c) (6 points) For $\epsilon < 2$, construct an ϵ-covering for the class $\mathcal{F}$ as follows. Set $T = \lfloor \frac{1}{\epsilon} \rfloor$, and for $i, j = 0, 1, \ldots, T$, define $\theta_i = \epsilon i$ and $\theta'_j = \epsilon j$. By also adding the points $\theta_{T+1} = 1$ and $\theta'_{T+1} = 1$, we obtain a collection $\{(\theta_i, \theta'_j) : i, j = 0, 1, \ldots, T+1\}$ contained within $[0,1]^2$ of cardinality $(T+2)^2$. Show that the associated functions $\{f_{\theta_i, \theta'_j} : i, j = 0, 1, \ldots, T+1\}$ constitute an ϵ-covering of $\mathcal{F}$. Determine an upper bound on the ϵ-covering number $N(\varepsilon; \mathcal{F}, \rho_\infty)$ as a function of ϵ.

Hint: *You can use without proof that* $1 - e^{-x} \leq x$, *for* $x \in [0,1]$.

(d) (6 points) For $\epsilon < 2$, construct an ϵ-packing for the class $\mathcal{F}$ with respect to the metric ρ_∞. Find a lower bound on the ϵ-packing number $M(\varepsilon; \mathcal{F}, \rho_\infty)$ in terms of ϵ.

(e) (4 points) Show that the metric entropy of the class $\mathcal{F}$ with respect to the metric ρ_∞ satisfies

$$\log N(\varepsilon; \mathcal{F}, \rho_\infty) \asymp \log(1/\epsilon), \quad \text{as } \epsilon \to 0^1.$$

[1] One writes $f \asymp g$, if $f = O(g)$ and $g = O(f)$. One writes $f = O(g)$, if $\limsup_{\epsilon \to 0} \left| \frac{f(\epsilon)}{g(\epsilon)} \right| < \infty$.

Problem 3 (20 points)

In this problem, you will investigate how ReLU networks are used in classification tasks, when there are more than two classes. The usual approach is to compose the ReLU network with a Softmax function, defined as follows.

Definition 1. Let $n \in \mathbb{N}$. The Softmax function is defined as

$$\begin{aligned} \mathrm{Softmax}^{(n)} : \quad \mathbb{R}^n &\to \mathbb{R}^n \\ x &\mapsto \frac{\exp(x)}{\sum_{j=1}^n \exp(x_j)}, \end{aligned} \tag{1}$$

where $\exp(x) := (\exp(x_1), \dots, \exp(x_n))$.

You will specifically study the Lipschitz constant of ReLU networks composed with the Softmax function. We next define the Lipschitz constant.

Definition 2. Let $f : \mathbb{R}^d \mapsto \mathbb{R}^k$. The Lipschitz constant of f is defined as

$$|f|_{Lip} := \sup_{\substack{x,y \in [-1,1]^d \\ x \neq y}} \frac{\|f(x) - f(y)\|_\infty}{\|x - y\|_\infty}. \tag{2}$$

(a) (2 points) Compute $\mathrm{Softmax}^{(1)}$.

(b) (2 points) Let $n \in \mathbb{N}$. Show that for all $x \in \mathbb{R}^n$, and all $i \in \{1, \dots, n\}$, the individual components $\mathrm{Softmax}_i^{(n)}(x)$ of the Softmax function satisfy

$$0 \leq \mathrm{Softmax}_i^{(n)}(x) \leq 1. \tag{3}$$

(c) (8 points) Let $n \in \mathbb{N}$ and $x \in \mathbb{R}^n$. Show that $\left\| \nabla \mathrm{Softmax}^{(n)} \right\|_\infty \leq 1$, where ∇ is per Definition 5 in the Handout.

Hint: First prove that

$$\nabla \mathrm{Softmax}^{(n)}(x) = \mathrm{diag}\left(\mathrm{Softmax}^{(n)}(x)\right) - \mathrm{Softmax}^{(n)}(x)\, \mathrm{Softmax}^{(n)}(x)^T, \tag{4}$$

for all $x \in \mathbb{R}^n$, where diag is defined in Definition 4 in the Handout.

(d) (4 points) Let $n \in \mathbb{N}$, $i \in \{1, \dots, n\}$, and $x \in \mathbb{R}^n$. Show that

$$|\mathrm{Softmax}^{(n)}|_{Lip} \leq n^{3/2}. \tag{5}$$

Hint: Use Theorem 1, and Lemmata 2 and 3 in the Handout.

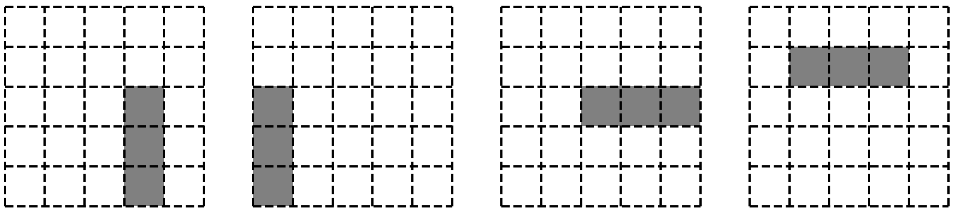
(e) (4 points) Let $n, m, d \in \mathbb{N}$, and let $\phi = W_2 \circ \rho \circ W_1$ be a RELU network with $W_1 : \mathbb{R}^d \to \mathbb{R}^m$, $W_1(x) := A_1 x$ and $W_2 : \mathbb{R}^m \to \mathbb{R}^n$, $W_2(x) := A_2 x$, where $A_1 \in \mathbb{R}^{m \times d}$ and $A_2 \in \mathbb{R}^{n \times m}$. Show that

$$\left| \mathrm{Softmax}^{(n)} \circ \phi \right|_{Lip} \leq n^{3/2} m d \|A_1\|_\infty \|A_2\|_\infty, \tag{6}$$

where $f \circ g$ stands for the concatenation of the functions f and g. *Hint: Use Lemmata 2 and 3 in the Handout, along with (5).*

Problem 4 (30 points)

In this problem, you will investigate how 2-D convolution can be used to classify images. Let us consider grayscale images, of size 5×5 pixels, containing either a vertical line or a horizontal line of 3 pixels, randomly positioned in the image:



These images are represented by 5×5 matrices, with entries equal to 0 corresponding to white pixels and entries equal to 1 corresponding to gray pixels. The next 4 matrices respectively represent the 4 images above.

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{7}$$

We consider the set $X := \{A_1, A_2, A_3, A_4\}$, along with the dichotomy $X^+ := \{A_1, A_2\}$, $X^- := \{A_3, A_4\}$, which separates horizontal from vertical lines. We define the map $\phi : \mathbb{R}^{5\times5} \to \mathbb{R}^2$ according to

$$\phi(A) = (\|A * K_1\|_\infty, \|A * K_2\|_\infty), \quad \text{for all } A \in \mathbb{R}^{5\times5}, \tag{8}$$

where $*$ is the convolution product as per Definition 6 in the Handout, and

$$K_1 := \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}, \quad K_2 := \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \tag{9}$$

It can be shown that

$$A_1 * K_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & x_1 & 1 \\ 0 & x_2 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad \text{and } A_1 * K_2 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & x_3 & 0 \\ 0 & 0 & 0 & x_4 & 0 \end{pmatrix}, \tag{10}$$

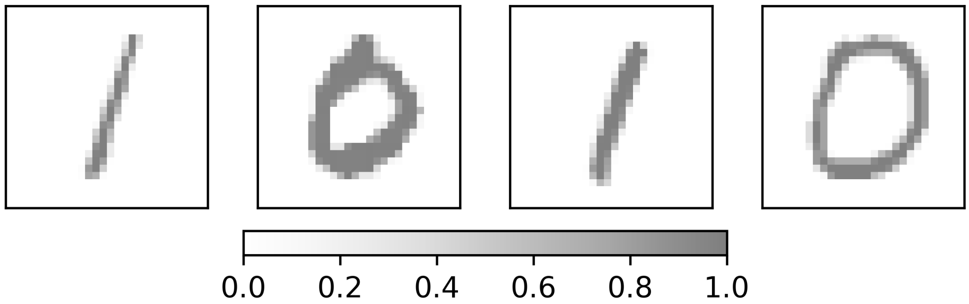
where $x_1, x_2, x_3, x_4 \in \mathbb{R}$.

(a) (6 points) Show that $x_1 = x_2 = 1$, $x_3 = 2$, and $x_4 = 3$. Deduce that $\phi(A_1) = (1, 3)$.

It can be shown that $\phi(A) = (1, 3)$ for all $A \in X^+$, and $\phi(A) = (3, 1)$ for all $A \in X^-$.

(b) (6 points) Is X in ϕ-general position? Is $\{X^+, X^-\}$ ϕ-separable ? If yes, characterize a corresponding separating surface.

The 0-1 MNIST dataset is a set of 1000 28×28 images of handwritten zeros and 1000 28×28 images of handwritten ones, as depicted in the examples below.



These images are represented by 28×28 matrices with values in $[0, 1]$ (the darkest gray pixels correspond to a value of 1, and the white pixels correspond to a value of 0). The set of all images is denoted by X, the set of images of handwritten zeros is denoted by X^+, and the set of images of handwritten ones is denoted by X^-. We define the map $\phi : \mathbb{R}^{28\times 28} \to \mathbb{R}^2$, aiming to separate $\{X^+, X^-\}$, according to

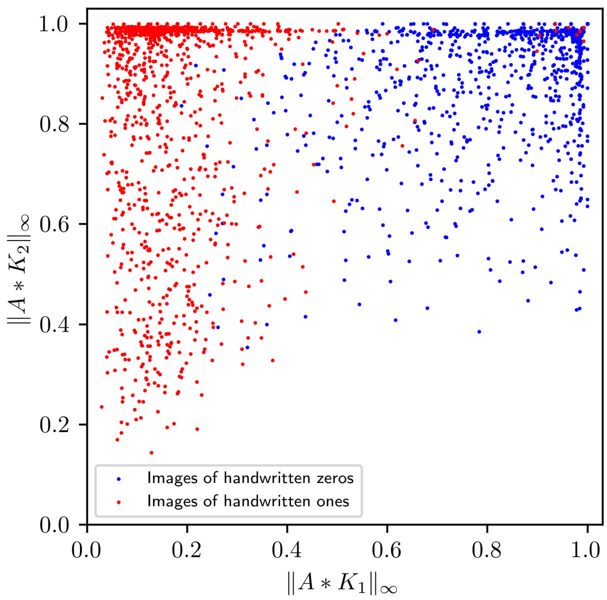
$$\phi(A) = (\|A * K_1\|_\infty,\ \|A * K_2\|_\infty)\,, \quad \text{for all } A \in \mathbb{R}^{28\times 28}, \tag{11}$$

where

$$K_1 = \frac{1}{11}\begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{1\times 11}, \quad K_2 = \frac{1}{11}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{11\times 1}. \tag{12}$$

(c) (4 points) Show that $\phi(A) \in [0, 1]^2$, for all $A \in \mathbb{R}^{28\times 28}$.

In the next picture we display $\phi(A)$ for all $A \in X$, where the blue points are for $A \in X^+$, and the red points are for $A \in X^-$.



The next two questions leave a lot of room for creativity. Any initiative will be rewarded with points, and the full grade can be obtained without answering the questions completely.

(d) (7 points) Explain why the red points tend to accumulate around $(0, 1)$, and the blue points tend to accumulate around $(1, 1)$.

(e) (7 points) Explain why $\{X^+, X^-\}$ is not ϕ-separable. Propose a strategy to define another mapping $\phi : \mathbb{R}^{28\times 28} \to \mathbb{R}^k$, where $k \in \mathbb{N}$, which, potentially, would lead to $\{X^+, X^-\}$ being ϕ-separable. You can freely choose the value of k. No proof is expected, but rather creative and well-justified propositions. Drawings and schemes are encouraged.

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 6, 2024

Definition 1. Let $n, m \in \mathbb{N}$, $x \in \mathbb{R}^n$, and $A \in \mathbb{R}^{m\times n}$. We define

$$\|x\|_\infty := \max_{i=1,\dots,n} |x_i|, \tag{1}$$

$$\|x\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2}, \tag{2}$$

$$\|A\|_\infty := \max_{\substack{i=1,\dots,m\\ j=1,\dots,n}} |A_{i,j}| \tag{3}$$

$$\|A\|_1 := \sum_{\substack{i=1,\dots,m\\ j=1,\dots,n}} |A_{i,j}|. \tag{4}$$

Definition 2 (ReLU network)**.** Let $L \in \mathbb{N}$ and $N_0, N_1, \dots, N_L \in \mathbb{N}$. 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 $\ell \in \{1, 2, \dots, L\}$, $W_\ell \colon \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, W_\ell(x) := A_\ell x + b_\ell$ 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., $\rho(x_1, \dots, x_N) := (\rho(x_1), \dots, \rho(x_N))$. 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, $\ell \in \{1, 2, \dots, L\}$, and the vectors b_ℓ, $\ell \in \{1, 2, \dots, L\}$,
- *depth* $\mathcal{L}(\Phi) := L$,
- *width* $\mathcal{W}(\Phi) := \max_{\ell=0,\dots,L} N_\ell$,
- *weight magnitude* $\mathcal{B}(\Phi) := \max_{\ell=1,\dots,L} \max\{\|A_\ell\|_\infty, \|b_\ell\|_\infty\}$.

Lemma 1. Let $(\mathcal{X}, \rho)$ be a metric space and $\mathcal{C}$ a compact set in $\mathcal{X}$. For all $\epsilon > 0$, the packing and covering number are related according to

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

Definition 3. Let $n \in \mathbb{N}$ and $x \in \mathbb{R}^n$. $A = \operatorname{diag}(x)$ denotes the $\mathbb{R}^{n\times n}$ matrix whose entries are all equal to 0, except on the main diagonal, which is given by $A_{i,i} = x_i$, for all $i \in \{1, \dots, n\}$.

Lemma 2. Let $n \in \mathbb{N}$ and $x \in \mathbb{R}^n$. Then,

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty. \tag{5}$$

Lemma 3. Let $n, m \in \mathbb{N}$, $x \in \mathbb{R}^n$, and $A \in \mathbb{R}^{m \times n}$. Then,

$$\|Ax\|_\infty \leq n\|A\|_\infty \|x\|_\infty. \tag{6}$$

Definition 4. Let $n, m \in \mathbb{N}$, and let $f : \mathbb{R}^n \to \mathbb{R}^m$. For $j \in \{1, \ldots, m\}$, we define $f_j : \mathbb{R}^n \to \mathbb{R}$ to be the function corresponding to the j-th coordinate of f, i.e.,

$$f(x) =: (f_1(x), \ldots, f_m(x)), \ \forall x \in \mathbb{R}^n. \tag{7}$$

Definition 5. Let $n, m \in \mathbb{N}$, and let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable function. For $j \in \{1, \ldots, n\}$, $i \in \{1, \ldots, m\}$, we define $\partial_j f_i$ to be the j-th partial derivative of f_i. Further, for $j \in \{1, \ldots, n\}$, we define $\partial_j f : \mathbb{R}^n \to \mathbb{R}^m$ as

$$\partial_j f := (\partial_j f_1, \ldots, \partial_j f_m). \tag{8}$$

Moreover, we define $\nabla f : \mathbb{R}^n \to \mathbb{R}^{m \times n}$ as

$$\nabla f := \begin{pmatrix} \partial_1 f_1 & \cdots & \partial_n f_1 \\ \partial_1 f_2 & \cdots & \partial_n f_2 \\ \vdots & \ddots & \vdots \\ \partial_1 f_m & \cdots & \partial_n f_m \end{pmatrix}. \tag{9}$$

Theorem 1. (Generalized Multivariate Mean Value Theorem) Let $n, m \in \mathbb{N}$, and let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable function. Then, for all $x, y \in \mathbb{R}^n$, with $x_i < y_i$, for all $i \in \{1, \ldots, n\}$, there exists $z \in (x_1, y_1) \times (x_2, y_2) \times \cdots \times (x_n, y_n)$ such that

$$\|f(y) - f(x)\|_2 \leq \|\nabla f(z)(y - x)\|_2. \tag{10}$$

Definition 6. Let $k, \ell, n, m \in \mathbb{N}$ be such that $k < n$ and $\ell < m$. Let $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{k \times \ell}$. The convolution product $A * B \in \mathbb{R}^{(n-k+1) \times (m-\ell+1)}$ is defined as

$$(A * B)_{i,j} = \sum_{\substack{p \in \{1,\ldots,k\} \\ q \in \{1,\ldots,\ell\}}} A_{i+p-1,j+q-1} B_{p,q}, \tag{11}$$

for all $i \in \{1, \ldots, n-k+1\}$, $j \in \{1, \ldots, m-\ell+1\}$.

We here show, by way of an example, how to compute a convolution product as defined in (11). Consider the matrices

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{4 \times 4}, \text{ and } B = \begin{pmatrix} 1 & 1 \end{pmatrix} \in \mathbb{R}^{1 \times 2}. \tag{12}$$

We want to compute $A * B$. First, note that $A * B \in \mathbb{R}^{(4-1+1) \times (4-2+1)} = \mathbb{R}^{4 \times 3}$. Now, to compute $(A * B)_{1,1}$, we apply (11):

$$(A * B)_{1,1} = \sum_{\substack{p \in \{1,\ldots,1\} \\ q \in \{1,\ldots,2\}}} A_{1+p-1,1+q-1} B_{p,q} = \sum_{q \in \{1,\ldots,2\}} A_{1,q} B_{1,q} = A_{1,1} B_{1,1} + A_{1,2} B_{1,2} = 0. \tag{13}$$

We continue with $(A * B)_{1,2}$:

$$(A * B)_{2,1} = \sum_{\substack{p \in \{1,\ldots,1\} \\ q \in \{1,\ldots,2\}}} A_{2+p-1,1+q-1} B_{p,q} = \sum_{q \in \{1,\ldots,2\}} A_{2,q} B_{1,q} = A_{2,1} B_{1,1} + A_{2,2} B_{1,2} = 2. \tag{14}$$

Continuing this procedure, we find that

$$A * B = \begin{pmatrix} 0 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{15}$$

Lemma 4. Let k, ℓ, n, and $m \in \mathbb{N}$ be such that $k < n$ and $\ell < m$. Let $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{k \times \ell}$. Then,

$$\|A * B\|_\infty \leq \|A\|_\infty \|B\|_1. \tag{16}$$