Supplementary material to "On the rate of convergence of image classifiers based on convolutional neural networks"

Michael Kohler, Adam Krzyżak and Benjamin Walter

This supplement contains an approximation result for convolutional neural networks in Section A, a bound on the covering number in Section B, and several auxiliary results from the literature in Section C. In Section D we provide a proof on the lower minimax rate described in Remark 6, and in Section E we explain how to design the network architecture proposed by the theory describe the choice of hyperparameters for the simulation studies.

A. An approximation result for convolutional neural networks

In this section we describe in Lemma 2 a connection between fully connected neural networks and convolutional neural networks, which will enbable us to derive in the proof of Theorem 1 an approximation result for the generalized hierarchical max-pooling model by the convolutional neural networks. Before we do this we present in Lemma 1 a bound on the error we make in case that we replace the functions g and $g_{k,s}$ in a hierarchical model by some approximations of them.

In the sequel $d_1, d_2 \in \mathbb{N}$ denote the image dimensions and furthermore let $l \in \mathbb{N}$ with $2^l \leq \min\{d_1, d_2\}$. We set $I = \{0, 1, \dots, 2^l - 1\} \times \{0, 1, \dots, 2^l - 1\}$ and assume

$$m : [0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \to [0,1]$$

satisfies a generalized hierarchical max-pooling model of level l and order $t \in \mathbb{N}$ with functions

$$g : \mathbb{R}^t \to [0,1] \quad \text{and} \quad g_{k,s}^{(a)} : \mathbb{R}^4 \to [0,1]$$

for $a \in \{1, \dots, t\}, k \in \{1, \dots, l\}$ and $s \in \{1, \dots, 4^{k-1}\}$. That is,

$$m(\mathbf{x}) = g(m_1(\mathbf{x}), \dots, m_t(\mathbf{x}))$$

and for all $a \in \{1, \dots, t\}$ it holds that

$$m_a(\mathbf{x}) = \max_{(i,j)\in\mathbb{Z}^2 \,:\, (i,j)+I\subseteq\{1,\dots,d_1\}\times\{1,\dots,d_2\}} f_a\left(\mathbf{x}_{(i,j)+I}\right)$$

where f_a satisfy

$$f_a = f_{l,1}^{(a)}$$

for some $f_{k,s}^{(a)} : \mathbb{R}^{\{1,\dots,2^k\}\times\{1,\dots,2^k\}} \to \mathbb{R}$ recursively defined by

$$f_{k,s}^{(a)}(\mathbf{x}) \quad = \quad g_{k,s}^{(a)}\big(f_{k-1,4\cdot(s-1)+1}^{(a)}(\mathbf{x}_{\{1,\dots,2^{k-1}\}\times\{1,\dots,2^{k-1}\}}),$$

$$f^{(a)}_{k-1,4\cdot(s-1)+2}(\mathbf{x}_{\{2^{k-1}+1,\dots,2^k\}\times\{1,\dots,2^{k-1}\}}),$$
$$f^{(a)}_{k-1,4\cdot(s-1)+3}(\mathbf{x}_{\{1,\dots,2^{k-1}\}\times\{2^{k-1}+1,\dots,2^k\}}),$$
$$f^{(a)}_{k-1,4\cdot s}(\mathbf{x}_{\{2^{k-1}+1,\dots,2^k\}\times\{2^{k-1}+1,\dots,2^k\}}))$$

for $k = 2, \dots, l, s = 1, \dots, 4^{l-k}$, and

$$f^{(a)}_{1,s}(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}) = g^{(a)}_{1,s}(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2})$$

for $s = 1, \dots, 4^{l-1}$. As already mentioned, we replace the functions g and $g_{k,s}$ by approximations of them. Therefore let

$$\bar{m} : [0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \to \mathbb{R} \tag{13}$$

be a function which satisfies a general hierarchical max-pooling model of level l and order t with functions

$$\bar{g} : \mathbb{R}^t \to \mathbb{R} \quad \text{and} \quad \bar{g}^{(a)}_{k,s} : \mathbb{R}^4 \to \mathbb{R}$$

for $a \in \{1, \dots, t\}, k \in \{1, \dots, l\}$ and $s \in \{1, \dots, 4^{k-1}\}$. Analogous to the above, we also define the functions $\bar{f}^{(a)}_{k,s} : \mathbb{R}^{\{1,\dots,2^k\}\times\{1,\dots,2^k\}} \to \mathbb{R}$ for all $a \in \{1, \dots, t\}, k \in \{1, \dots, l\}$ and $s \in \{1, \dots, 4^{k-1}\}$.

Lemma 1 *Assume that all restrictions $g^{(a)}_{k,s}|_{[-2,2]^4} : [-2,2]^4 \to [0,1]$ and $g|_{[-2,2]^t} : [-2,2]^t \to [0,1]$ are Lipschitz continuous (with respect to the Euclidean distance) with Lipschitz constant $C > 0$ for all $a \in \{1, \dots, t\}$, $k \in \{1, \dots, l\}$ and $s \in \{1, \dots, 4^{l-k}\}$. Furthermore, assume that for all $a \in \{1, \dots, t\}$, $k \in \{1, \dots, l\}$ and $s \in \{1, \dots, 4^{l-k}\}$*

$$\left\| \bar{g}^{(a)}_{k,s} \right\|_{[-2,2]^4,\infty} \leq 2. \tag{14}$$

Then for any $\mathbf{x} \in [0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}}$ it holds:

$$|m(\mathbf{x}) - \bar{m}(\mathbf{x})| \leq \sqrt{t} \cdot (2C+1)^l \cdot \max_{a\in\{1,\dots,t\}, j\in\{1,\dots,l\}, s\in\{1,\dots,4^{l-j}\}} \left\{ \|g^{(a)}_{j,s} - \bar{g}^{(a)}_{j,s}\|_{[-2,2]^4,\infty}, \|g - \bar{g}\|_{[-2,2]^t,\infty} \right\}.$$

Proof. Firstly, we show for any $a \in \{1, \dots, t\}$ that

$$|m_a(\mathbf{x}) - \bar{m}_a(\mathbf{x})| \leq (2C+1)^{l-1} \cdot \max_{j\in\{1,\dots,l\}, s\in\{1,\dots,4^{l-j}\}} \|g^{(a)}_{j,s} - \bar{g}^{(a)}_{j,s}\|_{[-2,2]^4,\infty}. \tag{15}$$

If $a_1, b_1, \dots, a_n, b_n \in \mathbb{R}$, then

$$|\max_{i=1,\dots,n} a_i - \max_{i=1,\dots,n} b_i| \leq \max_{i=1,\dots,n} |a_i - b_i|.$$

Indeed, in case $a_1 = \max_{i=1,\dots,n} a_i \geq \max_{i=1,\dots,n} b_i$ (which we can assume w.l.o.g.) we have

$$|\max_{i=1,\dots,n} a_i - \max_{i=1,\dots,n} b_i| = a_1 - \max_{i=1,\dots,n} b_i \leq a_1 - b_1 \leq \max_{i=1,\dots,n} |a_i - b_i|.$$

Consequently it suffices to show

$$\begin{aligned}&\max_{(i,j)\in\mathbb{Z}^2\,:\,(i,j)+I\subseteq\{1,\dots,d_1\}\times\{1,\dots,d_2\}}\left|f_a\left(\mathbf{x}_{(i,j)+I}\right)-\bar{f}_a\left(\mathbf{x}_{(i,j)+I}\right)\right|\\&\le(2C+1)^{l-1}\cdot\max_{j\in\{1,\dots,l\},s\in\{1,\dots,4^{l-j}\}}\|g_{j,s}^{(a)}-\bar{g}_{j,s}^{(a)}\|_{[-2,2]^4,\infty}.\end{aligned}$$

This in turn follows from

$$|f_{k,s}^{(a)}(\mathbf{x})-\bar{f}_{k,s}^{(a)}(\mathbf{x})|\le(2C+1)^{k-1}\cdot\max_{i\in\{1,\dots,k\},s\in\{1,\dots,4^{l-i}\}}\|g_{i,s}^{(a)}-\bar{g}_{i,s}^{(a)}\|_{[-2,2]^4,\infty}\tag{16}$$

for all $k\in\{1,\dots,l\}$, all $s\in\{1,\dots,4^{l-k}\}$ and all $\mathbf{x}\in[0,1]^{\{1,\dots,2^k\}\times\{1,\dots,2^k\}}$, which we show in the sequel by induction on k.

For $k=1$ and $s\in\{1,\dots,4^{l-1}\}$ we have

$$\begin{aligned}\left|f_{1,s}^{(a)}(\mathbf{x})-\bar{f}_{1,s}^{(a)}(\mathbf{x})\right|&=\left|g_{1,s}^{(a)}(x_{1,1},x_{1,2},x_{2,1},x_{2,2})-\bar{g}_{1,s}^{(a)}(x_{1,1},x_{1,2},x_{2,1},x_{2,2})\right|\\&\le\left\|g_{1,s}^{(a)}-\bar{g}_{1,s}^{(a)}\right\|_{[0,1]^4,\infty}.\end{aligned}$$

Assume now that (16) holds for some $k\in\{1,\dots,l-1\}$. The definition of $\bar{f}_{k,s}^{(a)}$ and inequality (14) imply that

$$\left|\bar{f}_{k,s}^{(a)}(\mathbf{x})\right|\le 2$$

for all $\mathbf{x}\in[0,1]^{\{1,\dots,2^k\}\times\{1,\dots,2^k\}}$ and $s\in\{1,\dots,4^{l-k}\}$. Then, the triangle inequality and the Lipschitz assumption on g imply

$$\begin{aligned}&|f_{k+1,s}^{(a)}(\mathbf{x})-\bar{f}_{k+1,s}^{(a)}(\mathbf{x})|\\&\le\Big|g_{k+1,s}^{(a)}\big(f_{k,4\cdot(s-1)+1}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{1,\dots,2^k\}}),f_{k,4\cdot(s-1)+2}^{(a)}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{1,\dots,2^k\}}),\\&\qquad f_{k,4\cdot(s-1)+3}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{2^k+1,\dots,2^{k+1}\}}),f_{k,4\cdot s}^{(a)}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{2^k+1,\dots,2^{k+1}\}})\big)\\&\quad-g_{k+1,s}^{(a)}\big(\bar{f}_{k,4\cdot(s-1)+1}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{1,\dots,2^k\}}),\bar{f}_{k,4\cdot(s-1)+2}^{(a)}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{1,\dots,2^k\}}),\\&\qquad\bar{f}_{k,4\cdot(s-1)+3}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{2^k+1,\dots,2^{k+1}\}}),\bar{f}_{k,4\cdot s}^{(a)}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{2^k+1,\dots,2^{k+1}\}})\big)\Big|\\&\quad+\Big|g_{k+1,s}^{(a)}\big(\bar{f}_{k,4\cdot(s-1)+1}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{1,\dots,2^k\}}),\bar{f}_{k,4\cdot(s-1)+2}^{(a)}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{1,\dots,2^k\}}),\\&\qquad\bar{f}_{k,4\cdot(s-1)+3}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{2^k+1,\dots,2^{k+1}\}}),\bar{f}_{k,4\cdot s}^{(a)}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{2^k+1,\dots,2^{k+1}\}})\big)\\&\quad-\bar{g}_{k+1,s}^{(a)}\big(\bar{f}_{k,4\cdot(s-1)+1}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{1,\dots,2^k\}}),\bar{f}_{k,4\cdot(s-1)+2}^{(a)}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{1,\dots,2^k\}}),\\&\qquad\bar{f}_{k,4\cdot(s-1)+3}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{2^k+1,\dots,2^{k+1}\}}),\bar{f}_{k,4\cdot s}^{(a)}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{2^k+1,\dots,2^{k+1}\}})\big)\Big|\\&\le C\cdot\Big(|f_{k,4\cdot(s-1)+1}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{1,\dots,2^k\}})-\bar{f}_{k,4\cdot(s-1)+1}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{1,\dots,2^k\}})|^2\\&\quad+|f_{k,4\cdot(s-1)+2}^{(a)}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{1,\dots,2^k\}})-\bar{f}_{k,4\cdot(s-1)+2}^{(a)}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{1,\dots,2^k\}})|^2\\&\quad+|f_{k,4\cdot(s-1)+3}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{2^k+1,\dots,2^{k+1}\}})-\bar{f}_{k,4\cdot(s-1)+3}^{(a)}(\mathbf{x}_{\{1,\dots,2^k\}\times\{2^k+1,\dots,2^{k+1}\}})|^2\end{aligned}$$

$$
\begin{aligned}
&+|f^{(a)}_{k,4\cdot s}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{2^k+1,\dots,2^{k+1}\}})-\bar{f}^{(a)}_{k,4\cdot s}(\mathbf{x}_{\{2^k+1,\dots,2^{k+1}\}\times\{2^k+1,\dots,2^{k+1}\}})|^2\Big)^{1/2}\\
&+\|g^{(a)}_{k+1,s}-\bar{g}^{(a)}_{k+1,s}\|_{[-2,2]^4,\infty}\\
\leq\ & (2\cdot C)\cdot(2C+1)^{k-1}\cdot \max_{i\in\{1,\dots,k\},s\in\{1,\dots,4^{l-i}\}}\|g^{(a)}_{i,s}-\bar{g}^{(a)}_{i,s}\|_{[-2,2]^4,\infty}\\
&+\|g^{(a)}_{k+1,s}-\bar{g}^{(a)}_{k+1,s}\|_{[-2,2]^4,\infty}\\
\leq\ & (2C+1)^k\cdot \max_{i\in\{1,\dots,k+1\},s\in\{1,\dots,4^{l-i}\}}\|g^{(a)}_{i,s}-\bar{g}^{(a)}_{i,s}\|_{[-2,2]^4,\infty}
\end{aligned}
$$

for all $\mathbf{x}\in[0,1]^{\{1,\dots,2^{k+1}\}\times\{1,\dots,2^{k+1}\}}$.

The definition of the functions $\bar{f}^{(a)}_{k,s}$ and inequality (14) imply that

$$
|\bar{m}_a(\mathbf{x})|\leq 2
$$

for all $\mathbf{x}\in[0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}}$ and $a\in\{1,\dots,t\}$. Then, the triangle inequality, the Lipschitz assumption on g and inequality (15) imply

$$
\begin{aligned}
&|g(m_1(\mathbf{x}),\dots,m_t(\mathbf{x}))-\bar{g}(\bar{m}_1(\mathbf{x}),\dots,\bar{m}_t(\mathbf{x}))|\\
&\leq |g(m_1(\mathbf{x}),\dots,m_t(\mathbf{x}))-g(\bar{m}_1(\mathbf{x}),\dots,\bar{m}_t(\mathbf{x}))|\\
&\quad+|g(\bar{m}_1(\mathbf{x}),\dots,\bar{m}_t(\mathbf{x}))-\bar{g}(\bar{m}_1(\mathbf{x}),\dots,\bar{m}_t(\mathbf{x}))|\\
&\leq C\cdot\left(|m_1(\mathbf{x})-\bar{m}_1(\mathbf{x})|^2+\dots+|m_t(\mathbf{x})-\bar{m}_t(\mathbf{x})|^2\right)^{1/2}+\|g-\bar{g}\|_{[-2,2]^t,\infty}\\
&\leq \sqrt{t}\cdot C\cdot(2C+1)^{l-1}\cdot\max_{a\in\{1,\dots,t\},j\in\{1,\dots,l\},s\in\{1,\dots,4^{l-j}\}}\|g^{(i)}_{j,s}-\bar{g}^{(i)}_{j,s}\|_{[-2,2]^4,\infty}\\
&\quad+\|g-\bar{g}\|_{[-2,2]^t,\infty}\\
&\leq\sqrt{t}\cdot(2C+1)^l\\
&\quad\cdot\max_{\substack{a\in\{1,\dots,t\},\\ j\in\{1,\dots,l\},s\in\{1,\dots,4^{l-j}\}}}\left\{\|g^{(a)}_{j,s}-\bar{g}^{(a)}_{j,s}\|_{[-2,2]^4,\infty},\|g-\bar{g}\|_{[-2,2]^t,\infty}\right\}
\end{aligned}
$$

for all $\mathbf{x}\in[0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}}$. □

In the next lemma, we show that a convolutional neural network $m_{net}\in\mathcal{F}(l_{net},\mathbf{k},\mathbf{M})$ can mimic a function $\bar{m}_a$ from the definition of $\bar{m}(\mathbf{x})=\bar{g}(\bar{m}_1(\mathbf{x}),\dots,\bar{m}_t(\mathbf{x}))$ (cf., equation (13)) if the functions $\bar{g}^{(a)}_{k,s}$ are standard feedforward neural networks.

Lemma 2 *Let $a\in\{1,\dots,t\}$ and assume that the functions*

$$
\bar{g}^{(a)}_{k,s}:\mathbb{R}^4\to\mathbb{R}
$$

in the definition the of $\bar{m}(\mathbf{x})=\bar{g}(\bar{m}_1(\mathbf{x}),\dots,\bar{m}_t(\mathbf{x}))$ (cf., equation (13)) are standard feedforward neural networks (defined as in equation (3)) with $L_{net}\in\mathbb{N}$ hidden layers and $r_{net}\in\mathbb{N}$ neurons per hidden layer and ReLU activation function for all $k\in\{1,\dots,l\}$ and $s\in\{1,\dots,4^{l-k}\}$. Set

$$
l_{net}=\frac{4^l-1}{3}\cdot L_{net}+l,
$$

$$k_s = \frac{2 \cdot 4^l + 4}{3} + r_{net} \quad (s = 1, \dots, l_{net}),$$

and set

$$M_s = 2^{\pi(s)} \quad \text{for } s \in \{1, \dots, l_{net}\},$$

where the function $\pi : \{1, \dots, l_{net}\} \to \{1, \dots, l\}$ *is defined by*

$$\pi(s) = \sum_{i=1}^{l} \mathbb{I}_{\left\{s \geq i + \sum_{r=l-i+1}^{l-1} 4^r \cdot L_{net}\right\}}.$$

Then there exists some $m_{net} \in \mathcal{F}(l_{net}, \mathbf{k}, \mathbf{M})$ *such that*

$$\bar{m}_a(\mathbf{x}) = m_{net}(\mathbf{x})$$

holds for all $\mathbf{x} \in [0,1]^{\{1,\dots,d_1\} \times \{1,\dots,d_2\}}$.

In order to prove Lemma 2 we will use the following auxiliary result.

Lemma 3 *Let* $\bar{g} : \mathbb{R}^4 \to \mathbb{R}$ *be a standard feedforward neural network (defined as in equation (3)) with* $L_{net} \in \mathbb{N}$ *hidden layers and* $r_{net} \in \mathbb{N}$ *neurons per hidden layer. Let* $d_1, d_2 \in \mathbb{N}$ *with* $d_1, d_2 > 1$ *and let* $\sigma(x) = \max\{x, 0\}$ *be the ReLU activation function. We assume that there is a convolutional neural network* $m_{net} \in \mathcal{F}(L, \mathbf{k}, \mathbf{M})$ *with* $L = r_0 + L_{net} + 1$ *convolutional layers and* $k_r = t + r_{net}$ *channels in the convolutional layer* r $(r = 1, \dots, r_0 + L_{net} + 1)$ *for* $r_0, t \in \mathbb{N}$, *and filter sizes* $M_1, \dots, M_{r_0 + L_{net} + 1} \in \mathbb{N}$ *with*

$$M_{r_0+1} = 2^k \text{ for some } k \in \mathbb{N} \text{ with } 2^k \leq \min\{d_1, d_2\}.$$

The convolutional neural network m_{net} *is given by its weight matrix*

$$\mathbf{w} = \left(w_{i,j,s_1,s_2}^{(r)}\right)_{1 \leq i,j \leq M_r, s_1 \in \{1,\dots,k_{r-1}\}, s_2 \in \{1,\dots,k_r\} r \in \{1,\dots,r_0+L_{net}+1\}}, \tag{17}$$

and its bias weights

$$\mathbf{w}_{bias} = \left(w_{s_2}^{(r)}\right)_{s_2 \in \{1,\dots,k_r\}, r \in \{1,\dots,r_0+L_{net}+1\}}. \tag{18}$$

Set $I^{(m)} = \{0, \dots, 2^m - 1\} \times \{0, \dots 2^m - 1\}$ *for* $m \in \mathbb{N}_0$. *Furthermore, let* $f_1, \dots, f_4 : [0,1]^{(1,1)+I^{(k-1)}} \to \mathbb{R}$ *be functions and let* $s_{2,1}, \dots, s_{2,10} \in \{1, \dots, t\}$. *Assume that the given convolutional neural network* m_{net} *satisfies the following four conditions for all* $(i_2, j_2) \in \{1, \dots, d_1 - 2^k + 1\} \times \{1, \dots, d_2 - 2^k + 1\}$:

$$o_{(i_2,j_2),s_{2,1}}^{(r_0)} - o_{(i_2,j_2),s_{2,2}}^{(r_0)} = f_1(\mathbf{x}_{(i_2,j_2)+I^{(k-1)}}), \tag{19}$$

$$o_{(i_2+2^{k-1},j_2),s_{2,3}}^{(r_0)} - o_{(i_2+2^{k-1},j_2),s_{2,4}}^{(r_0)} = f_2(\mathbf{x}_{(i_2+2^{k-1},j_2)+I^{(k-1)}}), \tag{20}$$

$$o_{(i_2,j_2+2^{k-1}),s_{2,5}}^{(r_0)} - o_{(i_2,j_2+2^{k-1}),s_{2,6}}^{(r_0)} = f_3(\mathbf{x}_{(i_2,j_2+2^{k-1})+I^{(k-1)}}) \tag{21}$$

and

$$o^{(r_0)}_{(i_2+2^{k-1},j_2+2^{k-1}),s_{2,7}} - o^{(r_0)}_{(i_2+2^{k-1},j_2+2^{k-1}),s_{2,8}} = f_4(\mathbf{x}_{(i_2+2^{k-1},j_2+2^{k-1})+I^{(k-1)}}). \tag{22}$$

Then we are able to modify the weights (17) *and* (18)

$$w^{(r)}_{t_1,t_2,s_1,s_2}, w^{(r)}_{s_2} \quad (s_1 \in \{1,\dots,t+r_{net}\}) \tag{23}$$

in layers $r \in \{r_0+1,\dots,r_0+L_{net}+1\}$ *and in channels* $s_2 \in \{s_{2,9}, s_{2,10}, t+1,\dots,t+r_{net}\}$ *such that*

$$\begin{aligned} & o^{(r_0+L_{net}+1)}_{(i_2,j_2),s_{2,9}} - o^{(r_0+L_{net}+1)}_{(i_2,j_2),s_{2,10}} \\ & = \bar{g}\Big(f_1(\mathbf{x}_{(i_2,j_2)+I^{(k-1)}}), f_2(\mathbf{x}_{(i_2+2^{k-1},j_2)+I^{(k-1)}}), \\ & \qquad f_3(\mathbf{x}_{(i_2,j_2+2^{k-1})+I^{(k-1)}}), f_4(\mathbf{x}_{(i_2+2^{k-1},j_2+2^{k-1})+I^{(k-1)}})\Big) \end{aligned}$$

holds for all $(i_2, j_2) \in \{1,\dots,d_1-2^k+1\} \times \{1,\dots,d_2-2^k+1\}$.

Remark 9. In the proof of Lemma 3 we only modify in layers $r_0+1,\dots,r_0+L_{net}+1$ the filters and bias weights (23) in channels

$$t+1,\dots,t+r_{net}$$

and in layer $r_0+L_{net}+1$ the filters and bias weights in channels

$$s_{2,9}, s_{2,10}.$$

This means that the calculation only takes place in these channels. The filter and bias weights in the remaining channels can therefore be arbitrary.
Proof. Let $(i_2, j_2) \in \{1,\dots,d_1-2^k+1\} \times \{1,\dots,d_2-2^k+1\}$ be arbitrary. We modify the weights (23) by using the weights of $\bar{g}$. Here we assume that $\bar{g}$ is given by

$$\bar{g}(\mathbf{x}) = \sum_{i=1}^{r_{net}} w^{(L_{net})}_{1,i} \cdot g^{(L_{net})}_i(\mathbf{x}) + w^{(L_{net})}_{1,0}$$

for $g^{(L_{net})}_i$'s recursively defined by

$$g^{(r)}_i(\mathbf{x}) = \sigma\left(\sum_{j=1}^{r_{net}} w^{(r-1)}_{i,j} \cdot g^{(r-1)}_j(\mathbf{x}) + w^{(r-1)}_{i,0}\right)$$

for $i \in \{1,\dots,r_{net}\}$, $r \in \{2,\dots,L_{net}\}$, and

$$g^{(1)}_i(\mathbf{x}) = \sigma\left(\sum_{j=1}^{4} w^{(0)}_{i,j} \cdot x^{(j)} + w^{(0)}_{i,0}\right) \quad (i \in \{1,\dots,r_{net}\}).$$

In layer r_0+1 we modify the weights (23) in channel $t+i$ by setting

$$w^{(r_0+1)}_{t_1,t_2,s,t+i}=0$$

for all $t_1,t_2 \notin \{1,1+2^{k-1}\}$ and all $s \notin \{s_{2,1},\dots,s_{2,8}\}$ and choose the only nonzero weights by

$$\begin{aligned}
&w^{(r_0+1)}_{1,1,s_{2,1},t+i}=w^{(0)}_{i,1}, && w^{(r_0+1)}_{1,1,s_{2,2},t+i}=-w^{(0)}_{i,1},\\
&w^{(r_0+1)}_{1+2^{k-1},1,s_{2,3},t+i}=w^{(0)}_{i,2}, && w^{(r_0+1)}_{1+2^{k-1},1,s_{2,4},t+i}=-w^{(0)}_{i,2},\\
&w^{(r_0+1)}_{1,1+2^{k-1},s_{2,5},t+i}=w^{(0)}_{i,3}, && w^{(r_0+1)}_{1,1+2^{k-1},s_{2,6},t+i}=-w^{(0)}_{i,3},\\
&w^{(r_0+1)}_{1+2^{k-1},1+2^{k-1},s_{2,7},t+i}=w^{(0)}_{i,4}, && w^{(r_0+1)}_{1+2^{k-1},1+2^{k-1},s_{2,8},t+i}=-w^{(0)}_{i,4}
\end{aligned}$$

and $w^{(r_0+1)}_{t+i}=w^{(0)}_{i,0}$ for $i\in\{1,\dots,r_{net}\}$. Then we calculate with the modified weights and the assumptions (19)–(22)

$$\begin{aligned}
o^{(r_0+1)}_{(i_2,j_2),t+i}=&\sigma\left(\sum_{s_1=1}^{t+r_{net}}\sum_{\substack{t_1,t_2\in\{1,\dots,M_{r_0+1}\}\\(i_2+t_1-1,j_2+t_2-1)\in D}} w^{(r_0+1)}_{t_1,t_2,s_1,t+i}\cdot o^{(r_0)}_{(i_2+t_1-1,j_2+t_2-1),s_1}+w^{(r_0+1)}_{t+i}\right)\\
=&\sigma\Bigg(w^{(0)}_{i,1}\cdot\big(o^{(r_0)}_{(i_2,j_2),s_{2,1}}-o^{(r_0)}_{(i_2,j_2),s_{2,2}}\big)\\
&\quad+w^{(0)}_{i,2}\cdot\big(o^{(r_0)}_{(i_2+2^{k-1},j_2),s_{2,3}}-o^{(r_0)}_{(i_2+2^{k-1},j_2),s_{2,4}}\big)\\
&\quad+w^{(0)}_{i,3}\cdot\big(o^{(r_0)}_{(i_2,j_2+2^{k-1}),s_{2,5}}-o^{(r_0)}_{(i_2,j_2+2^{k-1}),s_{2,6}}\big)\\
&\quad+w^{(0)}_{i,4}\cdot\big(o^{(r_0)}_{(i_2+2^{k-1},j_2+2^{k-1}),s_{2,7}}-o^{(r_0)}_{(i_2+2^{k-1},j_2+2^{k-1}),s_{2,8}}\big)+w^{(0)}_{i,0}\Bigg)\\
=&\sigma\Bigg(w^{(0)}_{i,1}f_1(\mathbf{x}_{(i_2,j_2)+I^{(k-1)}})+w^{(0)}_{i,2}f_2(\mathbf{x}_{(i_2+2^{k-1},j_2)+I^{(k-1)}})\\
&\quad+w^{(0)}_{i,3}f_3(\mathbf{x}_{(i_2,j_2+2^{k-1})+I^{(k-1)}})+w^{(0)}_{i,4}f_4(\mathbf{x}_{(i_2+2^{k-1},j_2+2^{k-1})+I^{(k-1)}})+w^{(0)}_{i,0}\Bigg)\\
=&g^{(1)}_i\Big(f_1(\mathbf{x}_{(i_2,j_2)+I^{(k-1)}}),f_2(\mathbf{x}_{(i_2+2^{k-1},j_2)+I^{(k-1)}}),\\
&\qquad f_3(\mathbf{x}_{(i_2,j_2+2^{k-1})+I^{(k-1)}}),f_4(\mathbf{x}_{(i_2+2^{k-1},j_2+2^{k-1})+I^{(k-1)}})\Big)
\end{aligned}\tag{24}$$

for $i\in\{1,\dots,r_{net}\}$. In layers $r\in\{r_0+2,\dots,r_0+L_{net}\}$ in channel $t+i$ we modify the weights (23) by setting

$$w^{(r)}_{t_1,t_2,s,t+i}=0$$

for all $(t_1,t_2)\neq(1,1)$ and all $s\in\{1,\dots,t\}$ and choose the only nonzero weights by

$$w_{1,1,t+j,t+i}^{(r)}=w_{i,j}^{(r-r_0-1)},\quad w_{t+i}^{(r)}=w_{i,0}^{(r-r_0-1)}\quad (j\in\{1,\dots,r_{net}\})$$

for $i\in\{1,\dots,r_{net}\}$. Thus we obtain

$$o_{(i_2,j_2),t+i}^{(r_0+r)}=\sigma\left(\sum_{j=1}^{r_{net}} w_{i,j}^{(r-1)}\cdot o_{(i_2,j_2),t+j}^{(r_0+r-1)}+w_{i,0}^{(r-1)}\right)$$

for $i\in\{1,\dots,r_{net}\}$ and $r\in\{2,\dots,L_{net}\}$. Then we get by equation (24) and the definition of $g_i^{(r)}$ that

$$\begin{aligned}o_{(i_2,j_2),t+i}^{(r_0+r)}=&g_i^{(r)}\Big(f_1(\mathbf{x}_{(i_2,j_2)+I^{(k-1)}}),f_2(\mathbf{x}_{(i_2+2^{k-1},j_2)+I^{(k-1)}}),\\&f_3(\mathbf{x}_{(i_2,j_2+2^{k-1})+I^{(k-1)}}),f_4(\mathbf{x}_{(i_2+2^{k-1},j_2+2^{k-1})+I^{(k-1)}})\Big)\end{aligned}$$

for $i\in\{1,\dots,r_{net}\}$ and $r\in\{2,\dots,L_{net}\}$. Now in layer $r_0+L_{net}+1$ in channels $s_{2,9},s_{2,10}\in\{1,\dots,t\}$ we modify the weights (23) by setting

$$w_{t_1,t_2,s,s_{2,9}}^{(r_0+L_{net}+1)}=w_{t_1,t_2,s,s_{2,10}}^{(r_0+L_{net}+1)}=0$$

for all $(t_1,t_2)\neq(1,1)$ and all $s\in\{1,\dots,t\}$ and choose the only nonzero weights by

$$\begin{aligned}w_{1,1,t+i,s_{2,9}}^{(r_0+L_{net}+1)}&=w_{1,i}^{(L_{net})},\quad w_{1,1,t+i,s_{2,10}}^{(r_0+L_{net}+1)}=-w_{1,i}^{(L_{net})},\\w_{s_{2,9}}^{(r_0+L_{net}+1)}&=w_{1,0}^{(L_{net})},\quad w_{s_{2,10}}^{(r_0+L_{net}+1)}=-w_{1,0}^{(L_{net})}\end{aligned}$$

for $i\in\{1,\dots,r_{net}\}$. Consequently, we get the following outputs:

$$\begin{aligned}o_{(i_2,j_2),s_{2,9}}^{(r_0+L_{net}+1)}&=\sigma\left(\sum_{i=1}^{r_{net}} w_{1,i}^{(L_{net})}\cdot o_{(i_2,j_2),t+i}^{(r_0+L_{net})}+w_{1,0}^{(L_{net})}\right)\\&=\sigma\Bigg(\bar{g}\Big(f_1(\mathbf{x}_{(i_2,j_2)+I^{(k-1)}}),f_2(\mathbf{x}_{(i_2+2^{k-1},j_2)+I^{(k-1)}}),\\&\qquad f_3(\mathbf{x}_{(i_2,j_2+2^{k-1})+I^{(k-1)}}),f_4(\mathbf{x}_{(i_2+2^{k-1},j_2+2^{k-1})+I^{(k-1)}})\Big)\Bigg)\end{aligned}$$

and

$$\begin{aligned}o_{(i_2,j_2),s_{2,10}}^{(r_0+L_{net}+1)}&=\sigma\left(\sum_{i=1}^{r_{net}} -w_{1,i}^{(L_{net})}\cdot o_{(i_2,j_2),t+i}^{(r_0+L_{net})}-w_{1,0}^{(L_{net})}\right)\\&=\sigma\Bigg(-\bar{g}\Big(f_1(\mathbf{x}_{(i_2,j_2)+I^{(k-1)}}),f_2(\mathbf{x}_{(i_2+2^{k-1},j_2)+I^{(k-1)}}),\\&\qquad f_3(\mathbf{x}_{(i_2,j_2+2^{k-1})+I^{(k-1)}}),f_4(\mathbf{x}_{(i_2+2^{k-1},j_2+2^{k-1})+I^{(k-1)}})\Big)\Bigg).\end{aligned}$$

Finally, we obtain

$$
\begin{aligned}
& o^{(r_0+L_{net}+1)}_{(i_2,j_2),s_{2,9}} - o^{(r_0+L_{net}+1)}_{(i_2,j_2),s_{2,10}} \\
& = \max\Bigg\{ \bar{g}\Big(f_1(\mathbf{x}_{(i_2,j_2)+I^{(k-1)}}), f_2(\mathbf{x}_{(i_2+2^{k-1},j_2)+I^{(k-1)}}), \\
& \qquad\qquad f_3(\mathbf{x}_{(i_2,j_2+2^{k-1})+I^{(k-1)}}), f_4(\mathbf{x}_{(i_2+2^{k-1},j_2+2^{k-1})+I^{(k-1)}})\Big), 0 \Bigg\} \\
& \quad - \max\Bigg\{ - g_{net}\Big(f_1(\mathbf{x}_{(i_2,j_2)+I^{(k-1)}}), f_2(\mathbf{x}_{(i_2+2^{k-1},j_2)+I^{(k-1)}}), \\
& \qquad\qquad f_3(\mathbf{x}_{(i_2,j_2+2^{k-1})+I^{(k-1)}}), f_4(\mathbf{x}_{(i_2+2^{k-1},j_2+2^{k-1})+I^{(k-1)}})\Big), 0 \Bigg\} \\
& = g_{net}\Big(f_1(\mathbf{x}_{(i_2,j_2)+I^{(k-1)}}), f_2(\mathbf{x}_{(i_2+2^{k-1},j_2)+I^{(k-1)}}), \\
& \qquad f_3(\mathbf{x}_{(i_2,j_2+2^{k-1})+I^{(k-1)}}), f_4(\mathbf{x}_{(i_2+2^{k-1},j_2+2^{k-1})+I^{(k-1)}})\Big).
\end{aligned}
$$

□

Proof of Lemma 2. In the proof we will use the network $f_{id} : \mathbb{R} \to \mathbb{R}$ defined by

$$f_{id}(x) = \sigma(x) - \sigma(-x) = \max\{x, 0\} - \max\{-x, 0\} = x,$$

which enables us to save a value computed in layer $r-1$ in channel s at position (i,j) by a difference of the outputs of two neurons in distinct channels s_1 and s_2 in layer r by

$$o^{(r)}_{(i,j),s_1} - o^{(r)}_{(i,j),s_2} = \sigma\Big(o^{(r-1)}_{(i,j),s}\Big) - \sigma\Big(-o^{(r-1)}_{(i,j),s}\Big) = o^{(r-1)}_{(i,j),s}. \tag{25}$$

Once a value has been saved in layer r by the difference of two neurons, it will be propagated analogously to the next layer $r+1$ by calculating

$$o^{(r+1)}_{(i,j),s_1} - o^{(r+1)}_{(i,j),s_2} = \sigma\Big(o^{(r)}_{(i,j),s_1} - o^{(r)}_{(i,j),s_2}\Big) - \sigma\Big(o^{(r)}_{(i,j),s_2} - o^{(r)}_{(i,j),s_1}\Big) = o^{(r)}_{(i,j),s_1} - o^{(r)}_{(i,j),s_2}. \tag{26}$$

In case we want to make use of equation (25) or equation (26), we have to choose the filters (and the bias weights) of the convolutional neural network in layer r in the channels s_1 and s_2 accordingly from the set $\{-1, 0, 1\}$.

With this approach of storing and propagating calculated values, the idea of our proof is to choose the filters (and the bias weights) such that our convolutional neural network saves in channels corresponding to position (i,j) the values of $x_{i,j}$, $\bar{f}_{1,s}(\mathbf{x}_{(i,j)+I^{(1)}})$ ($s = 1, \dots, 4^{l-1}$), $\bar{f}_{2,s}(\mathbf{x}_{(i,j)+I^{(2)}})$ ($s = 1, \dots, 4^{l-2}$), $\dots$, $\bar{f}_{l,s}(\mathbf{x}_{(i,j)+I^{(l)}})$ ($s = 1$), where we set

$$I^{(m)} = \{0, \dots, 2^m - 1\} \times \{0, \dots, 2^m - 1\}$$

for $m \in \mathbb{N}_0$. To do this we need two neurons for each of the above values, so altogether

$$2 \cdot (1 + 4^{l-1} + 4^{l-2} + \dots + 4^0) = 2 \cdot \left(1 + \frac{4^l - 1}{4 - 1}\right) = \frac{2 \cdot 4^l + 4}{3}$$

channels or neurons for each position (i,j). Furthermore, we will need r_{net} additional channels to compute the networks $\bar{g}_{k,s}$. So altogether we need

$$\frac{2 \cdot 4^l + 4}{3} + r_{net} = k_r$$

many channels in each convolutional layer r.

The convolutional neural network $m_{net} \in \mathcal{F}(l_{net}, \mathbf{k}, \mathbf{M})$, which we will construct to prove the assertion, has the parameters $l_{net}, \mathbf{k}$ and $\mathbf{M}$ of Lemma 2. We make use of the above idea by choosing the filters (and bias weights) of the convolutional neural m_{net} network so that it has the following property for any $k \in \{1, \dots, l\}$:

> For any $s \in \{1, \dots, 4^{l-k}\}$, $(i,j) \in \{1, \dots, d_1 - 2^k + 1\} \times \{1, \dots, d_2 - 2^k + 1\}$ and any $r \in \{4^{l-1} \cdot L_{net} + \cdots + 4^{l-k} \cdot L_{net} + k, \dots, l_{net}\}$ it holds that
>
> $$\begin{aligned} &o^{(r)}_{(i,j),2+2\cdot 4^{l-1}+\cdots+2\cdot 4^{l-k+1}+2\cdot s-1} - o^{(r)}_{(i,j),2+2\cdot 4^{l-1}+\cdots+2\cdot 4^{l-k+1}+2\cdot s} \\ &= \bar{f}_{k,s}(\mathbf{x}_{(i,j)+I^{(k)}}). \end{aligned} \tag{27}$$

Due to equation (26) is suffices to show equation (27) for $r = 4^{l-1} \cdot L_{net} + \cdots + 4^{l-k} \cdot L_{net} + k$. To construct our convolutional neural network m_{net} so that the above property (27) is fulfilled, we use an induction on k.

We start with $k = 1$. First we note that

$$\bar{f}_{1,s}(\mathbf{x}_{(i,j)+I^{(1)}}) = \bar{g}_{1,s}(x_{(i,j)}, x_{(i+1,j)}, x_{(i,j+1)}, x_{(i+1,j+1)})$$

for $s \in \{1, \dots, 4^{l-1}\}$ and $(i,j) \in \{1, \dots, d_1 - 1\} \times \{1, \dots, d_2 - 1\}$. So we have to compute the networks $\bar{g}_{1,1}, \dots, \bar{g}_{1,4^{l-1}}$ applied to the input of our convultional network. The idea is to use Lemma 3 for each network $\bar{g}_{1,s}$. Therefore, we first make sure that the assumptions (19)–(22) of Lemma 3 are fulfilled as we need them. In the first convolutional layer we copy $x_{i,j}$ in the first two channels using the weights as in equation (25), and we propagate these values in the successive layers using the weights as in equation (26). So after the first layer we have available the input in the first two channels in all convolutional layers, so that for all $r \in \{2, \dots, l_{net}\}$ and all $(i,j) \in \{1, \dots, d_1\} \times \{1, \dots, d_2\}$ it holds that

$$o^{(r)}_{(i,j),1} - o^{(r)}_{(i,j),2} = x_{(i,j)}.$$

For the filter size it holds that

$$M_r = 2 \quad (r \in \pi^{-1}(1) = \{1, 2, \dots, 4^{l-1} \cdot L_{net} + 1\}).$$

Starting already in parallel in the first layer, we compute successively the networks $\bar{g}_{1,1}, \dots, \bar{g}_{1,4^{l-1}}$ in layers

$$1, 2, \dots, 4^{l-1} \cdot L_{net} + 1$$

in the channels

$$\frac{2 \cdot 4^l + 4}{3} + 1, \frac{2 \cdot 4^l + 4}{3} + 2, \dots, \frac{2 \cdot 4^l + 4}{3} + r_{net}$$

for the computation of their hidden layers and the ouput layers in channels $2+1,\dots,2+2\cdot 4^{l-1}$ by applying Lemma 3 4^{l-1} times. We now describe how to use Lemma 3 to compute $\bar{g}_{1,s}$ ($s=1,\dots,4^{l-1}$). In particular, we specify how to choose the parameters $s_{2,1},\dots,s_{2,10}$ from Lemma 3. The computation of $\bar{g}_{1,s}$ takes place in layers

$$(s-1)\cdot L_{net}+1,\dots,s\cdot L_{net}$$

in channels

$$\frac{2\cdot 4^l+4}{3}+1,\frac{2\cdot 4^l+4}{3}+2,\dots,\frac{2\cdot 4^l+4}{3}+r_{net}$$

for the computation of its hidden layers and its output layer is computed in layer $s\cdot L_{net}+1$ in channels $s_{2,9}=2+2s-1$ and $s_{2,10}=2+2s$. As input the network $\bar{g}_{1,s}$ uses the first two channels for $s>1$ such that

$$s_{2,1}=s_{2,3}=s_{2,5}=s_{2,7}=1 \text{ and } s_{2,2}=s_{2,4}=s_{2,6}=s_{2,8}=2,$$

and in case $s=1$ it selects its input from the input of the convolutional network and then use a simple variation of Lemma 3 by adapting the assumptions (19)–(22). The computed function value of $\bar{g}_{1,s}$ is then saved in the two channels $s_{2,9}=2+2s-1$ and $s_{2,10}=2+2s$. Here we propagate again the value of these neurons successively to the next layer by using the weights as in equation (26). So after layer $4^{l-1}\cdot L_{net}+1$ we have available the values of all $\bar{f}_{1,s}$ in the channels $2+1,\ \dots,\ 2+2\cdot 4^{l-1}$, so that for any $s\in\{1,\dots,4^{l-1}\}$ and any $(i,j)\in\{1,\dots,d_1-1\}\times\{1,\dots,d_2-1\}$ it holds that

$$\begin{aligned} o_{(i,j),2+2\cdot s-1}^{(4^{l-1}\cdot L_{net}+1)}-o_{(i,j),2+2\cdot s}^{(4^{l-1}\cdot L_{net}+1)} &= \bar{g}_{1,s}(x_{(i,j)},x_{(i+1,j)},x_{(i,j+1)},x_{(i+1,j+1)}) \\ &= \bar{f}_{1,s}(\mathbf{x}_{(i,j)+I^{(1)}}). \end{aligned}$$

Thus property (27) holds for $k=1$.

Now we assume that equation (27) holds for $k\in\{1,\dots,l-1\}$. We use the values $\bar{f}_{k,s}(\mathbf{x}_{(i,j)+I^{(k)}})$, which are given by equation (27), to compute all values of

$$\begin{aligned} \bar{f}_{k+1,s}(\mathbf{x}_{(i,j)+I^{(k+1)}}) = \bar{g}_{k+1,s}\Big(&\bar{f}_{k,4\cdot(s-1)+1}(\mathbf{x}_{(i,j)+I^{(k)}}),\bar{f}_{k,4\cdot(s-1)+2}(\mathbf{x}_{(i+2^k,j)+I^{(k)}}), \\ &\bar{f}_{k,4\cdot(s-1)+3}(\mathbf{x}_{(i,j+2^k)+I^{(k)}}),\bar{f}_{k,4\cdot s}(\mathbf{x}_{(i+2^k,j+2^k)+I^{(k)}})\Big) \end{aligned}$$

for $s\in\{1,\dots 4^{l-(k+1)}\}$ using Lemma 3. We proceed similarly to the above case of $k=1$. For the filter size it holds that

$$M_r=2^{k+1}\quad (r\in\pi^{-1}(k+1)),$$

where $\pi^{-1}(k+1)$ is given by

$$\{4^{l-1}\cdot L_{net}+\dots+4^{l-k}\cdot L_{net}+(k+1),\dots,4^{l-1}\cdot L_{net}+\dots+4^{l-(k+1)}\cdot L_{net}+(k+1)\}.$$

By applying Lemma 3 $4^{l-(k+1)}$ times we compute successively the networks $\bar{g}_{k+1,1},\dots,\bar{g}_{k+1,4^{l-(k+1)}}$, in the corresponding layers

$$4^{l-1}\cdot L_{net}+\dots+4^{l-k}\cdot L_{net}+k+1,\dots,4^{l-1}\cdot L_{net}+\dots+4^{l-k}\cdot L_{net}+4^{l-(k+1)}\cdot L_{net}+k+1,$$

where the computation of their hidden layers takes place in channels

$$\frac{2 \cdot 4^l + 4}{3} + 1, \frac{2 \cdot 4^l + 4}{3} + 2, \dots, \frac{2 \cdot 4^l + 4}{3} + r_{net}$$

and the computation of their ouput layers takes place in channels

$$2 + 2 \cdot 4^{l-1} + \dots + 2 \cdot 4^{l-k} + 1, \dots, 2 + 2 \cdot 4^{l-1} + \dots + 2 \cdot 4^{l-(k+1)}.$$

As above we describe how to use Lemma 3 to compute $\bar{g}_{k+1,s}$ $(s = 1, \dots, 4^{l-(k+1)})$ and specify how to choose the parameters $s_{2,1}, \dots, s_{2,10}$ from Lemma 3. The computation of $\bar{g}_{k+1,s}$ $(s = 1, \dots, 4^{l-(k+1)})$ takes place in layers

$$4^{l-1} \cdot L_{net} + \dots + 4^{l-k} \cdot L_{net} + k + (s-1) \cdot L_{net} + 1, \dots, 4^{l-1} \cdot L_{net} + \dots + 4^{l-k} \cdot L_{net} + k + s \cdot L_{net}$$

in channels

$$\frac{2 \cdot 4^l + 4}{3} + 1, \frac{2 \cdot 4^l + 4}{3} + 2, \dots, \frac{2 \cdot 4^l + 4}{3} + r_{net}$$

for the computation of its hidden layers and its output layer is computed in layer

$$4^{l-1} \cdot L_{net} + \dots + 4^{l-k} \cdot L_{net} + k + s \cdot L_{net} + 1$$

in channels

$$s_{2,9} = 2 + 2 \cdot 4^{l-1} + \dots + 2 \cdot 4^{l-k} + 2s - 1 \tag{28}$$

and

$$s_{2,10} = 2 + 2 \cdot 4^{l-1} + \dots + 2 \cdot 4^{l-k} + 2s. \tag{29}$$

We choose

$$s_{2,m} = 2 + \left(\sum_{i=l-(k-1)}^{l-1} 2 \cdot 4^i \right) + 2 \cdot 4 \cdot (s-1) + m$$

for $m \in \{1, \dots, 8\}$, because then we have

$$o^{(r)}_{(i,j),s_{2,2\cdot m-1}} - o^{(r)}_{(i,j),s_{2,2\cdot m}} = \bar{f}_{k,4\cdot(s-1)+m}(\mathbf{x}_{(i,j)+I^{(k)}})$$

for $m \in \{1, \dots, 4\}$ and any $r \in \{4^{l-1} \cdot L_{net} + \dots + 4^{l-k} \cdot L_{net} + k, \dots, l_{net}\}$ and any $(i,j) \in \{1, \dots, d_1 - 2^k + 1\} \times \{1, \dots, d_2 - 2^k + 1\}$ due to the induction hypothesis. Then Lemma 3 let us choose the corresponding weights of the network m_{net} such that

$$\begin{aligned}
&o^{(4^{l-1} \cdot L_{net} + \dots + 4^{l-k} \cdot L_{net} + k + s \cdot L_{net} + 1)}_{(i,j),s_{2,9}} - o^{(4^{l-1} \cdot L_{net} + \dots + 4^{l-k} \cdot L_{net} + k + s \cdot L_{net} + 1)}_{(i,j),s_{2,10}} \\
&= \bar{g}_{k+1,s}\Big(\bar{f}_{k,4\cdot(s-1)+1}(\mathbf{x}_{(i,j)+I^{(k)}}), \bar{f}_{k,4\cdot(s-1)+2}(\mathbf{x}_{(i+2^k,j)+I^{(k)}}), \\
&\qquad \bar{f}_{k,4\cdot(s-1)+3}(\mathbf{x}_{(i,j+2^k)+I^{(k)}}), \bar{f}_{k,4\cdot s}(\mathbf{x}_{(i+2^k,j+2^k)+I^{(k)}})\Big) \\
&= \bar{f}_{k+1,s}(\mathbf{x}_{(i,j)+I^{(k+1)}}).
\end{aligned}$$

for any $(i,j) \in \{1,\dots,d_1-2^{k+1}+1\} \times \{1,\dots,d_2-2^{k+1}+1\}$. By propagating again the values of these neurons successively to the next layer we have available the values of all $\bar{f}_{k+1,s}$ after layer

$$4^{l-1} \cdot L_{net} + \cdots + 4^{l-k} \cdot L_{net} + 4^{l-(k+1)} \cdot L_{net} + k + 1$$

in the channels

$$2 + 2 \cdot 4^{l-1} + \cdots + 2 \cdot 4^{l-k} + 1, \dots, 2 + 2 \cdot 4^{l-1} + \cdots + 2 \cdot 4^{l-(k+1)}$$

so that for any $s \in \{1,\dots,4^{l-(k+1)}\}$ and any $(i,j) \in \{1,\dots,d_1-2^{k+1}-1\} \times \{1,\dots,d_2-2^{k+1}-1\}$ it holds that

$$o_{(i,j),2+2\cdot 4^{l-1}+\cdots+2\cdot 4^{l-k}+2s-1}^{(4^{l-1}\cdot L_{net}+\cdots+4^{l-k}\cdot L_{net}+4^{l-(k+1)}\cdot L_{net}+k+1)} - o_{(i,j),2+2\cdot 4^{l-1}+\cdots+2\cdot 4^{l-k}+2s}^{(4^{l-1}\cdot L_{net}+\cdots+4^{l-k}\cdot L_{net}+4^{l-(k+1)}\cdot L_{net}+k+1)}$$
$$= \bar{f}_{k+1,s}(\mathbf{x}_{(i,j)+I^{(k+1)}}).$$

So property (27) holds for all $k \in \{1,\dots,l\}$.

Hence in layer

$$l_{net} = 4^{l-1} \cdot L_{net} + 4^{l-2} \cdot L_{net} + \cdots + 4^0 \cdot L_{net} + l = \frac{4^l - 1}{3} \cdot L_{net} + l$$

we have by equation (27)

$$o_{(i,j),2+2\cdot 4^{l-1}+\cdots+2\cdot 4+1}^{(l_{net})} - o_{(i,j),2+2\cdot 4^{l-1}+\cdots+2\cdot 4+2}^{(l_{net})} = \bar{f}_{l,1}(\mathbf{x}_{(i,j)+I^{(l)}})$$

for all $(i,j) \in \{1,\dots,d_1-2^l+1\} \times \{1,\dots,d_2-2^l+1\}$. Now we choose the outer weights $\mathbf{w}_{out}$ of our convolutional neural network m_{net} such that

$$w_s = \begin{cases} 1, \text{ if } s = 2 + 2 \cdot 4^{l-1} + \cdots + 2 \cdot 4 + 1 \\ -1, \text{ if } s = 2 + 2 \cdot 4^{l-1} + \cdots + 2 \cdot 4 + 2 \\ 0, \text{ else.} \end{cases}$$

This implies that the output of our network is given by

$$m_{net}(\mathbf{x}) = \max \Bigg\{ o_{(i,j),2+2\cdot 4^{l-1}+\cdots+2\cdot 4^{l-l+1}+1}^{(l_{net})} - o_{(i,j),2+2\cdot 4^{l-1}+\cdots+2\cdot 4^{l-l+1}+2}^{(l_{net})} \quad :$$
$$(i,j) \in \{1,\dots,d_1-2^l+1\} \times \{1,\dots,d_2-2^l+1\} \Bigg\}$$
$$= \max \left\{ \bar{f}(\mathbf{x}_{(i,j)+I}) \quad : \quad (i,j) \in \mathbb{Z}_2, (i,j) + I \subseteq \{1,\dots,d_1\} \times \{1,\dots,d_2\} \right\}$$
$$= \bar{m}(\mathbf{x}).$$

□

B. A bound on the covering number

In this section, we present a result on the covering number of the class $\mathcal{F}_t\left(\mathbf{L}, \mathbf{k}^{(1)}, \mathbf{k}^{(2)}, \mathbf{M}\right)$ of convolutional neural networks.

Lemma 4 *Let $\sigma(x) = \max\{x, 0\}$ be the ReLU activation function, define*

$$\mathcal{F} := \mathcal{F}_t\left(\mathbf{L}, \mathbf{k}^{(1)}, \mathbf{k}^{(2)}, \mathbf{M}\right)$$

as in Section 2 and set

$$k_{max} = \max\left\{k_1^{(1)}, \dots, k_{L^{(1)}}^{(1)}, t, k_1^{(2)}, \dots, k_{L^{(2)}}^{(2)}\right\}, \quad M_{max} = \max\{M_1, \dots, M_{L^{(1)}}\}$$

and

$$L_{max} = \max\{L^{(1)}, L^{(2)}\}.$$

Assume $d_1 \cdot d_2 > 1$ and $c_4 \cdot \log n \geq 2$. Then we have for any $\epsilon \in (0,1)$:

$$\sup_{\mathbf{x}_1^n \in (\mathbb{R}^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}})^n} \log\left(\mathcal{N}_1\left(\epsilon, T_{c_4 \cdot \log n}\mathcal{F}, \mathbf{x}_1^n\right)\right)$$
$$\leq c_7 \cdot L_{max}^2 \cdot \log(L_{max} \cdot d_1 \cdot d_2) \cdot \log\left(\frac{c_4 \cdot \log n}{\epsilon}\right)$$

for some constant $c_7 > 0$ which depends only on k_{max} and M_{max}.

With the aim of proving Lemma 4, we first have to study the VC dimension of our function class $\mathcal{F}_t\left(\mathbf{L}, \mathbf{k}^{(1)}, \mathbf{k}^{(2)}, \mathbf{M}\right)$. For a class of subsets of $\mathbb{R}^d$, the VC dimension is defined as follows:

Definition 2 Let $\mathcal{A}$ be a class of subsets of $\mathbb{R}^d$ with $\mathcal{A} \neq \emptyset$ and $m \in \mathbb{N}$.

1. For $\mathbf{x}_1, ..., \mathbf{x}_m \in \mathbb{R}^d$ we define

$$s(\mathcal{A}, \{\mathbf{x}_1, ..., \mathbf{x}_m\}) := |\{A \cap \{\mathbf{x}_1, ..., \mathbf{x}_m\} \ : \ A \in \mathcal{A}\}|.$$

2. Then the mth **shatter coefficient** $S(\mathcal{A}, m)$ of $\mathcal{A}$ is defined by

$$S(\mathcal{A}, m) := \max_{\{\mathbf{x}_1, ..., \mathbf{x}_m\} \subset \mathbb{R}^d} s(\mathcal{A}, \{\mathbf{x}_1, ..., \mathbf{x}_m\}).$$

3. The **VC dimension** (Vapnik-Chervonenkis-Dimension) $V_{\mathcal{A}}$ of $\mathcal{A}$ is defined as

$$V_{\mathcal{A}} := \sup\{m \in \mathbb{N} \ : \ S(\mathcal{A}, m) = 2^m\}.$$

For a class of real-valued functions, we define the VC dimension as follows:

Definition 3 Let $\mathcal{H}$ denote a class of functions from $\mathbb{R}^d$ to $\{0, 1\}$ and let $\mathcal{F}$ be a class of real-valued functions.

1. For any non-negative integer m, we define the **growth function** of H as

$$\Pi_{\mathcal{H}}(m) \coloneqq \max_{\mathbf{x}_1,\dots,\mathbf{x}_m \in \mathbb{R}^d} |\{(h(\mathbf{x}_1),\dots,h(\mathbf{x}_m)) : h \in H\}|.$$

2. The **VC dimension** (Vapnik-Chervonenkis-Dimension) of $\mathcal{H}$ we define as

$$\mathrm{VCdim}(\mathcal{H}) \coloneqq \sup\{m \in \mathbb{N} : \Pi_{\mathcal{H}}(m) = 2^m\}.$$

3. For $f \in \mathcal{F}$ we denote $\mathrm{sgn}(f) \coloneqq \mathbb{I}_{\{f \geq 0\}}$ and $\mathrm{sgn}(\mathcal{F}) \coloneqq \{\mathrm{sgn}(f) : f \in \mathcal{F}\}$. Then the **VC dimension** of $\mathcal{F}$ is defined as

$$\mathrm{VCdim}(\mathcal{F}) \coloneqq \mathrm{VCdim}(\mathrm{sgn}(\mathcal{F})).$$

A connection between both definitions is given by the following lemma.

Lemma 5 *Suppose $\mathcal{F}$ is a class of real-valued functions on $\mathbb{R}^d$. Furthermore, we define*

$$\mathcal{F}^+ \coloneqq \{\{(\mathbf{x},y) \in \mathbb{R}^d \times \mathbb{R} : f(\mathbf{x}) \geq y\} : f \in \mathcal{F}\}$$

and define the class $\mathcal{H}$ of real-valued functions on $\mathbb{R}^d \times \mathbb{R}$ by

$$\mathcal{H} \coloneqq \{h((\mathbf{x},y)) = f(\mathbf{x}) - y : f \in \mathcal{F}\}.$$

Then, it holds that

$$V_{\mathcal{F}^+} = \mathrm{VCdim}(\mathcal{H}).$$

Proof. For all $(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_m,y_m) \in \mathbb{R}^d \times \mathbb{R}$ with $m \in \mathbb{N}$ it holds that

$$\begin{aligned}
&s(\mathcal{F}^+, \{(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_m,y_m)\}) \\
&= \left|\left\{A \cap \{(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_m,y_m)\} \ : \ A \in \mathcal{F}^+\right\}\right| \\
&= \left|\left\{\{(\mathbf{x},y) \in \mathbb{R}^d \times \mathbb{R} : f(\mathbf{x}) \geq y\} \cap \{(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_m,y_m)\} \ : \ f \in \mathcal{F}\right\}\right| \\
&= |\{\{(\mathbf{x},y) \in \{(x_1,y_1),\dots,(x_m,y_m)\} : f(\mathbf{x}) \geq y\} \ : \ f \in \mathcal{F}\}| \\
&= |\{\{i \in \{1,\dots,m\} : f(\mathbf{x}_i) \geq y_i\} \ : \ f \in \mathcal{F}\}| \\
&= |\{(\mathrm{sgn}(f(\mathbf{x}_1) - y_1),\dots,\mathrm{sgn}(f(\mathbf{x}_m) - y_m)) : f \in \mathcal{F}\}| \\
&= |\{(\mathrm{sgn}(h(\mathbf{x}_1,y_1)),\dots,\mathrm{sgn}(h(\mathbf{x}_m,y_m))) : h \in \mathcal{H}\}|.
\end{aligned}$$

It follows that

$$S(\mathcal{F}^+, m) = \Pi_{\mathcal{H}}(m)$$

holds for all $m \in \mathbb{N}$, which implies

$$V_{\mathcal{F}^+} = \mathrm{VCdim}(\mathcal{H}).$$

□

In order to bound the VC dimension of our function class, we need the following auxiliary result about the number of possible sign vectors attained by polynomials of bounded degree.

Lemma 6 *Suppose $W \leq m$ and let $f_1, ..., f_m$ be polynomials of degree at most D in W variables. Define*

$$K \coloneqq |\{(\mathrm{sgn}(f_1(\mathbf{a})), \dots, \mathrm{sgn}(f_m(\mathbf{a}))) : \mathbf{a} \in \mathbb{R}^W\}|.$$

Then we have

$$K \leq 2 \cdot \left(\frac{2 \cdot e \cdot m \cdot D}{W} \right)^W .$$

Proof. See Theorem 8.3 in Anthony and Bartlett (1999). □

To get an upper bound for the VC dimension of our function class $\mathcal{F}_t\left(\mathbf{L}, \mathbf{k}^{(1)}, \mathbf{k}^{(2)}, \mathbf{M}\right)$ defined as in Section 2 we will use a modification of Theorem 6 in Bartlett et al. (2019).

Lemma 7 *Let $\sigma(x) = \max\{x, 0\}$ be the ReLU activation function, define*

$$\mathcal{F} \coloneqq \mathcal{F}_t \left(\mathbf{L}, \mathbf{k}^{(1)}, \mathbf{k}^{(2)}, \mathbf{M} \right)$$

as in Section 2, set

$$k_{max} = \max\left\{k_1^{(1)}, \dots, k_{L^{(1)}}^{(1)}, t, k_1^{(2)}, \dots, k_{L^{(2)}}^{(2)}\right\}, \quad M_{max} = \max\{M_1, \dots, M_{L^{(1)}}\}$$

and

$$L_{max} = \max\{L^{(1)}, L^{(2)}\}.$$

Assume $d_1 \cdot d_2 > 1$. Then, we have

$$V_{\mathcal{F}^+} \leq c_{10} \cdot L_{max}^2 \cdot \log_2(L_{max} \cdot d_1 \cdot d_2)$$

for some constant $c_{10} > 0$ which depends only on k_{max} and M_{max}.

Proof. We want to use Lemma 5 to bound $\mathcal{V}_{\mathcal{F}^+}$ by $\mathrm{VCdim}(\mathcal{H})$, where $\mathcal{H}$ is the class of real-valued functions on $[0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \times \mathbb{R}$ defined by

$$\mathcal{H} \coloneqq \{h((\mathbf{x}, y)) = f(\mathbf{x}) - y : f \in \mathcal{F}\}.$$

Let $h \in \mathcal{H}$. Then h depends on t convolutional neural networks

$$f_1, \dots, f_t \in \mathcal{F}(L^{(1)}, \mathbf{k}^{(1)}, \mathbf{M})$$

and one standard feedforward neural network $g \in \mathcal{G}_t(L^{(2)}, \mathbf{k}^{(2)})$ such that

$$h((\mathbf{x}, y)) = g \circ (f_1, \dots, f_t)(\mathbf{x}) - y$$

Each one of the convolutional neural networks $f_1, \dots, f_t$ depends on a weight matrix

$$\mathbf{w}^{(b)} = \left(w_{i,j,s_1,s_2}^{(b,r)} \right)_{1 \leq i,j \leq M_r, s_1 \in \{1,\dots,k_{r-1}^{(1)}\}, s_2 \in \{1,\dots,k_r^{(1)}\}, r \in \{1,\dots,L^{(1)}\}},$$

the weights

$$\mathbf{w}_{bias}^{(b)} = \left(w_{s_2}^{(b,r)}\right)_{s_2 \in \{1,\dots,k_r^{(1)}\}, r \in \{1,\dots,L^{(1)}\}}$$

for the bias in each channel and each convolutional layer, the output weights

$$\mathbf{w}_{out}^{(b)} = (w_s^{(b)})_{s \in \{1,\dots,k_{L^{(1)}}^{(1)}\}}$$

for $b \in \{1, \dots, t\}$. The standard feedforward neural network $g \in \mathcal{G}_t(L^{(2)}, \mathbf{k}^{(2)})$ depends on the inner weigths

$$w_{i,j}^{(r-1)}$$

for $j \in \{0, \dots, k_{r-1}^{(2)}\}$, $i \in \{1, \dots, k_r^{(2)}\}$ and $r \in \{1, \dots, L^{(2)}\}$ and the outer weights

$$w_i^{(L^{(2)})}$$

for $i \in \{0, \dots, k_{L^{(2)}}^{(2)}\}$ (where $k_0^{(2)} = t$).

We set

$$\mathbf{k} = (k_0, \dots, k_{L^{(1)}+L^{(2)}+1}) = (1, k_1^{(1)}, \dots, k_{L^{(1)}}^{(1)}, t, k_1^{(2)}, \dots, k_{L^{(2)}}^{(2)})$$

and count the number of weights used up to layer $r \in \{1, \dots, L^{(1)}\}$ in the convolutional part by

$$W_r \coloneqq t \cdot \left(\sum_{s=1}^{r} M_s^2 \cdot k_s \cdot k_{s-1} + \sum_{s=1}^{r} k_s \right),$$

for $r \in \{1, \dots, L^{(1)}\}$ (where we set $W_0 \coloneqq 0$) and

$$W_{L^{(1)}+1} \coloneqq W_{L^{(1)}} + t \cdot k_{L^{(1)}}.$$

We continue in the part of the standard feedforward neural network by counting the weights used up to layer $r \in \{1, \dots, L^{(2)}\}$ by

$$W_{L^{(1)}+1+r} = W_{L^{(1)}+r} + \left(k_{L^{(1)}+r} + 1\right) \cdot k_{L^{(1)}+r+1}$$

and denote the total number of weights by

$$\begin{aligned} W &= W_{L^{(1)}+L^{(2)}+2} \\ &= W_{L^{(1)}+L^{(2)}+1} + k_{L^{(1)}+L^{(2)}+1} + 1 \\ &\leq L^{(1)} \cdot t \cdot \left(M_{max}^2 \cdot k_{max}^2 + k_{max} \right) + t \cdot k_{max} \\ &\quad + L^{(2)} \cdot ((k_{max} + 1) \cdot k_{max}) + k_{max} + 1 \\ &\leq L^{(1)} \cdot t \cdot \left(M_{max}^2 \cdot (k_{max} + 1) \cdot k_{max} \right) \\ &\quad + L^{(2)} \cdot ((k_{max} + 1) \cdot k_{max}) \\ &\quad + 2 \cdot t \cdot (k_{max} + 1) \\ &\leq (L^{(1)} + L^{(2)} + 2) \cdot t \cdot M_{max}^2 \cdot (k_{max} + 1) \cdot k_{max} \\ &\leq 2 \cdot (L^{(1)} + L^{(2)} + 2) \cdot t \cdot M_{max}^2 \cdot k_{max}^2. \end{aligned} \tag{30}$$

We define $I^{(0)} = \emptyset$ and for $r \in \{1, \ldots, L^{(1)} + L^{(2)} + 2\}$ we define the index sets

$$I^{(r)} = \{1, \ldots, W_r\}.$$

Furthermore, we define a sequence of vectors containing the weights used up to layer $r \in \{1, \ldots, L^{(1)}\}$ in the convolutional part by

$$\mathbf{a}_{I^{(r)}} := \Big(\mathbf{a}_{I^{(r-1)}}, w_{1,1,1,1}^{(1,r)}, \ldots, w_{M_r, M_r, k_{r-1}, k_r}^{(1,r)}, w_1^{(1,r)}, \ldots, w_{k_r}^{(1,r)}, \\ \ldots, w_{1,1,1,1}^{(t,r)}, \ldots, w_{M_r, M_r, k_{r-1}, k_r}^{(t,r)}, w_1^{(t,r)}, \ldots, w_{k_r}^{(t,r)}\Big) \in \mathbb{R}^{W_r}$$

(where $\mathbf{a}_\emptyset$ denotes the empty vector),

$$\mathbf{a}_{I^{(L^{(1)}+1)}} := (\mathbf{a}_{I^{(L^{(1)})}}, w_1^{(1)}, \ldots, w_{k_{L^{(1)}}}^{(1)}, \ldots, w_1^{(t)}, \ldots, w_{k_{L^{(1)}}}^{(t)}) \in \mathbb{R}^{W_{L^{(1)}+1}},$$

and by continuing with the part of the standard feedforward neural network we get for $r \in \{1, \ldots, L^{(2)}\}$

$$\mathbf{a}_{I^{(r+L^{(1)}+1)}} := \left(\mathbf{a}_{I^{(r+L^{(1)})}}, w_{1,0}^{(r-1)}, \ldots, w_{k_{r+L^{(1)}+1}, k_{r+L^{(1)}}}^{(r-1)}\right) \in \mathbb{R}^{W_{r+L^{(1)}+1}}$$

and

$$\mathbf{a} := \left(\mathbf{a}_{I^{(L^{(1)}+L^{(2)}+1)}}, w_0^{(L^{(2)})}, \ldots, w_{L^{(2)}}^{(L^{(2)})}\right) \in \mathbb{R}^W.$$

With this notation we can write

$$\mathcal{H} = \{(\mathbf{x}, y) \mapsto h((\mathbf{x}, y), \mathbf{a}) : \mathbf{a} \in \mathbb{R}^W\}$$

and for $b \in \{1, \ldots, t\}$

$$\mathcal{F}(L^{(1)}, \mathbf{k}^{(1)}, \mathbf{M}) = \{\mathbf{x} \mapsto f_b(\mathbf{x}, \mathbf{a}) : \mathbf{a} \in \mathbb{R}^W\},$$

where the convolutional networks $f_1, \ldots, f_t \in \mathcal{F}(L^{(1)}, \mathbf{k}^{(1)}, \mathbf{M})$, as described above, each depends only on $W_{L^{(1)}+1}/t$ variables of $\mathbf{a}$. To get an upper bound for the VC-dimension of $\mathcal{H}$, we will bound the growth function $\Pi_{\operatorname{sgn}(\mathcal{H})}(m)$. In the following we assume that m is a positive integer with

$$m \geq W \tag{31}$$

since this will allow us several uses of Lemma 6. To bound the growth function $\Pi_{\operatorname{sgn}(\mathcal{H})}(m)$, we fix the input values

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in [0,1]^{\{1,\ldots,d_1\} \times \{1,\ldots,d_2\}} \times \mathbb{R}$$

and consider $h \in \mathcal{H}$ as a function of the weight vector $\mathbf{a} \in \mathbb{R}^W$ of h

$$\mathbf{a} \mapsto h((\mathbf{x}_k, y_k), \mathbf{a}) = g \circ (f_1, \ldots, f_t)(\mathbf{x}_k, \mathbf{a}) - y_k = h_k(\mathbf{a})$$

for any $k \in \{1, \ldots, m\}$. Then, an upper bound for

$$K := |\{(\operatorname{sgn}(h_1(\mathbf{a})), \ldots, \operatorname{sgn}(h_m(\mathbf{a}))) : \mathbf{a} \in \mathbb{R}^W\}|$$

implies an upper bound for the growth function $\Pi_{\mathrm{sgn}(\mathcal{H})}(m)$. For any partition

$$\mathcal{S} = \{S_1, \ldots, S_M\}$$

of $\mathbb{R}^W$ it holds that

$$K \leq \sum_{i=1}^{M} |\{(\mathrm{sgn}(h_1(\mathbf{a})), \ldots, \mathrm{sgn}(h_m(\mathbf{a})) : \mathbf{a} \in S_i\}|. \tag{32}$$

We will construct a partition $\mathcal{S}$ of $\mathbb{R}^W$ such that within each region $S \in \mathcal{S}$, the functions $h_k(\cdot)$ are all fixed polynomials of bounded degree for $k \in \{1, \ldots, m\}$, so that each summand of equation (32) can be bounded via Lemma 6. We do this in two steps.

In the first step we construct a partition $\mathcal{S}^{(1)}$ of $\mathbb{R}^W$ such that within each $S \in \mathcal{S}^{(1)}$ the t convolutional neural networks $f_{1,k}(\mathbf{a}), \ldots, f_{t,k}(\mathbf{a})$ are all fixed polynomials with dergee of at most $L^{(1)} + 1$ for all $k \in \{1, \ldots, m\}$, where we denote

$$f_{b,k}(\mathbf{a}) = f_b(\mathbf{x}_k, \mathbf{a})$$

for $b \in \{1, \ldots, t\}$. We define

$$D = \{1, \ldots, d_1\} \times \{1, \ldots, d_2\}.$$

For $b \in \{1, \ldots, t\}$ we have

$$f_{b,k}(\mathbf{a}) = \max\left\{ \sum_{s=1}^{k_{L^{(1)}}} w_s^{(b)} \cdot o_{(i,j),b,s,\mathbf{x}_k}^{(L^{(1)})}(\mathbf{a}_{I^{(L^{(1)})}}) : (i,j) \in D \right\},$$

where $o_{(i,j),b,s_2,\mathbf{x}}^{(L^{(1)})}(\mathbf{a}_{I^{(L^{(1)})}})$ is recursively defined by

$$\begin{aligned} & o_{(i,j),b,s_2,\mathbf{x}}^{(r)}(\mathbf{a}_{I^{(r)}}) \\ & = \sigma\left(\sum_{s_1=1}^{k_{r-1}} \sum_{\substack{t_1,t_2 \in \{1,\ldots,M_r\} \\ (i+t_1-1,j+t_2-1) \in D}} w_{t_1,t_2,s_1,s_2}^{(b,r)} \cdot o_{(i+t_1-1,j+t_2-1),b,s_1,\mathbf{x}}^{(r-1)}(\mathbf{a}_{I^{(r-1)}}) + w_{s_2}^{(b,r)} \right) \end{aligned}$$

for $(i,j) \in D$ and $r \in \{1, \ldots, L^{(1)}\}$, and by

$$o_{(i,j),b,1,\mathbf{x}}^{(0)}(\mathbf{a}_{I^{(0)}}) = x_{i,j} \quad \text{for } (i,j) \in D.$$

Firstly, we construct a partition $\mathcal{S}_{L^{(1)}} = \{S_1, \ldots, S_M\}$ of $\mathbb{R}^W$ such that within each $S \in \mathcal{S}_{L^{(1)}}$

$$o_{(i,j),b,s,\mathbf{x}_k}^{(L^{(1)})}(\mathbf{a}_{I^{(L^{(1)})}})$$

is a fixed polynomial for all $k \in \{1, \ldots, m\}$, $s \in \{1, \ldots, k_L\}$, $b \in \{1, \ldots, t\}$ and $(i,j) \in D$ with degree of at most $L^{(1)}$ in the $W_{L^{(1)}}$ variables $\mathbf{a}_{I^{(L^{(1)})}}$ of $\mathbf{a} \in S$. We construct the partition $\mathcal{S}_{L^{(1)}}$ iteratively layer by layer, by creating a sequence $\mathcal{S}_0, \ldots, \mathcal{S}_{L^{(1)}}$, where each $\mathcal{S}_r$ is a partition of $\mathbb{R}^W$ with the following properties:

1. We have $|\mathcal{S}_0| = 1$ and, for each $r \in \{1, \dots, L^{(1)}\}$,

$$\frac{|\mathcal{S}_r|}{|\mathcal{S}_{r-1}|} \leq 2 \left(\frac{2 \cdot e \cdot t \cdot k_r \cdot d_1 \cdot d_2 \cdot m \cdot r}{W_r} \right)^{W_r}, \tag{33}$$

2. For each $r \in \{0, \dots, L^{(1)}\}$, and each element $S \in \mathcal{S}_r$, each $(i,j) \in D$, each $s \in \{1, \dots, k_r\}$, each $k \in \{1, \dots, m\}$, and each $b \in \{1, \dots, t\}$ when $\mathbf{a}$ varies in S,

$$o^{(r)}_{(i,j),b,s,\mathbf{x}_k}(\mathbf{a}_{I^{(r)}})$$

is a fixed polynomial function in the W_r variables $\mathbf{a}_{I^{(r)}}$ of $\mathbf{a}$, of total degree no more than r.

We define $\mathcal{S}_0 \coloneqq \{\mathbb{R}^W\}$. Since

$$o^{(0)}_{(i,j),b,s,\mathbf{x}_k}(\mathbf{a}_{I^{(0)}}) = (x_k)_{i,j}$$

is a constant polynomial, property 2 above is satisfied for $r = 0$. Now suppose that $\mathcal{S}_0, \dots, \mathcal{S}_{r-1}$ have been defined, and we want to define $\mathcal{S}_r$. For $S \in \mathcal{S}_{r-1}$ let

$$p_{(i,j),b,s_1,\mathbf{x}_k,S}(\mathbf{a}_{I^{(r-1)}})$$

denote the function $o^{(r-1)}_{(i,j),b,s_1,\mathbf{x}_k}(\mathbf{a}_{I^{(r-1)}})$, when $\mathbf{a} \in S$. By induction hypothesis

$$p_{(i,j),b,s_1,\mathbf{x}_k,S}(\mathbf{a}_{I^{(r-1)}})$$

is a polynomial with total degree no more than $r-1$, and depends on the W_{r-1} variables $\mathbf{a}_{I^{(r-1)}}$ of $\mathbf{a}$ for any $b \in \{1, \dots, t\}$, $k \in \{1, \dots, m\}$, $(i,j) \in D$ and $s_1 \in \{1, \dots, k_{r-1}\}$. Hence for any $b \in \{1, \dots, t\}$ $k \in \{1, \dots, m\}$, $(i,j) \in D$ and $s_2 \in \{1, \dots, k_r\}$

$$\sum_{s_1=1}^{k_{r-1}} \sum_{\substack{t_1,t_2 \in \{1,\dots,M_r\} \\ (i+t_1-1,j+t_2-1) \in D}} w^{(r)}_{t_1,t_2,s_1,s_2} \cdot p_{(i+t_1-1,j+t_2-1),b,s_1,\mathbf{x}_k,S}(\mathbf{a}_{I^{(r-1)}}) + w^{(b,r)}_{s_2}$$

is a polynomial in the W_r variables $\mathbf{a}_{I^{(r)}}$ of $\mathbf{a}$ with total degree no more than r. Because of condition (31) we have $t \cdot k_r \cdot m \cdot d_1 \cdot d_2 \geq W_r$. Hence, by Lemma 6, the collection of polynomials

$$\left\{ \sum_{s_1=1}^{k_{r-1}} \sum_{\substack{t_1,t_2 \in \{1,\dots,M_r\} \\ (i+t_1-1,j+t_2-1) \in D}} w^{(b,r)}_{t_1,t_2,s_1,s_2} \cdot p_{(i+t_1-1,j+t_2-1),b,s_1,\mathbf{x}_k,S}(\mathbf{a}_{I^{(r-1)}}) + w^{(b,r)}_{s_2} : \right.$$
$$\left. b \in \{1, \dots, t\}, k \in \{1, \dots, m\}, (i,j) \in D, s_2 \in \{1, \dots, k_r\} \right\}$$

attains at most

$$\Pi \coloneqq 2\left(\frac{2\cdot e\cdot t\cdot k_r\cdot m\cdot d_1\cdot d_2\cdot r}{W_r}\right)^{W_r}$$

distinct sign patterns when $\mathbf{a}_{I(r)} \in \mathbb{R}^{W_r}$ and therefore the above collection of polynomials also attains at most Π distinct sign patterns when $\mathbf{a}$ varies in $\mathbb{R}^W$ since the above polynomials depend only on the W_r variables $\mathbf{a}_{I(r)}$ of $\mathbf{a}$. Therefore, we can partition $S \subset \mathbb{R}^W$ into Π subregions, such that all the polynomials don't change their signs within each subregion. Doing this for all regions $S \in \mathcal{S}_{r-1}$ we get our required partition $\mathcal{S}_r$ by assembling all of these subregions. In particular, property 1 (inequality (33)) is then satisfied.

Fix some $S' \in \mathcal{S}_r$. Notice that, when $\mathbf{a}$ varies in S', all the polynomials

$$\left\{\sum_{s_1=1}^{k_{r-1}} \sum_{\substack{t_1,t_2\in\{1,\dots,M_r\}\\(i+t_1-1,j+t_2-1)\in D}} w^{(b,r)}_{t_1,t_2,s_1,s_2}\cdot p_{(i+t_1-1,j+t_2-1),b,s_1,\mathbf{x}_k,S}(\mathbf{a}_{I(r-1)}) + w^{(b,r)}_{s_2} : \right.$$
$$\left. b\in\{1,\dots,t\}, k\in\{1,\dots,m\}, (i,j)\in D, s_2\in\{1,\dots,k_r\}\right\}$$

don't change their signs, hence

$$\begin{aligned}
&o^{(r)}_{(i,j),b,s_2,\mathbf{x}_k}(\mathbf{a}_{I(r)})\\
&=\sigma\left(\sum_{s_1=1}^{k_{r-1}} \sum_{\substack{t_1,t_2\in\{1,\dots,M_r\}\\(i+t_1-1,j+t_2-1)\in D}} w^{(b,r)}_{t_1,t_2,s_1,s_2}\cdot o^{(r-1)}_{(i+t_1-1,j+t_2-1),b,s_1,\mathbf{x}_k}(\mathbf{a}_{I(r-1)}) + w^{(b,r)}_{s_2}\right)\\
&=\max\left\{\sum_{s_1=1}^{k_{r-1}} \sum_{\substack{t_1,t_2\in\{1,\dots,M_r\}\\(i+t_1-1,j+t_2-1)\in D}} w^{(b,r)}_{t_1,t_2,s_1,s_2}\cdot o^{(r-1)}_{(i+t_1-1,j+t_2-1),b,s_1,\mathbf{x}_k}(\mathbf{a}_{I(r-1)}) + w^{(b,r)}_{s_2}, 0\right\}
\end{aligned}$$

is either a polynomial of degree no more than r in the W_r variables $\mathbf{a}_{I(r)}$ of $\mathbf{a}$ or a constant polynomial with value 0 for all $(i,j)\in D$, $b\in\{1,\dots,t\}$, $s_2\in\{1,\dots,k_r\}$ and $k\in\{1,\dots,m\}$. Hence, property 2 is also satisfied and we are able to construct our desired partition $\mathcal{S}_{L^{(1)}}$. Because of inequality (33) of property 1 it holds that

$$|\mathcal{S}_{L^{(1)}}| \le \prod_{r=1}^{L^{(1)}} 2\left(\frac{2\cdot e\cdot t\cdot k_r\cdot d_1\cdot d_2\cdot m\cdot r}{W_r}\right)^{W_r}.$$

For any $(i,j) \in D$, $b \in \{1,\dots,t\}$ and $k \in \{1,\dots,m\}$, we define

$$f_{(i,j),b,\mathbf{x}_k}(\mathbf{a}_{I(L^{(1)}+1)}) \coloneqq \sum_{s_2=1}^{k_{L^{(1)}}} w_{s_2}^{(b)} \cdot o_{(i,j),b,s_2,\mathbf{x}_k}^{(L^{(1)})}(\mathbf{a}_{I(L^{(1)})}).$$

For any fixed $S \in \mathcal{S}_{L^{(1)}}$, let $p_{(i,j),b,S,\mathbf{x}_k}(\mathbf{a}_{I(L^{(1)}+1)})$ denote the function $f_{(i,j),b,\mathbf{x}_k}(\mathbf{a}_{I(L^{(1)}+1)})$, when $\mathbf{a} \in S$. By construction of $\mathcal{S}_{L^{(1)}}$ this is a polynomial of degree no more than $L^{(1)}+1$ in the $W_{L^{(1)}+1}$ variables $\mathbf{a}_{I(L^{(1)}+1)}$ of $\mathbf{a}$. Because of condition (31) we have $t \cdot d_1^2 \cdot d_2^2 \cdot m \geq W_{L^{(1)}+1}$. Hence, by Lemma 6, the collection of polynomials

$$\Big\{ p_{(i_1,j_1),b,S,\mathbf{x}_k}(\mathbf{a}_{I(L^{(1)}+1)}) - p_{(i_2,j_2),b,S,\mathbf{x}_k}(\mathbf{a}_{I(L^{(1)}+1)}) : \\ (i_1,j_1),(i_2,j_2) \in D, (i_1,j_1) \neq (i_2,j_2), b \in \{1,\dots,t\}, k \in \{1,\dots,m\} \Big\}$$

attains at most

$$\Delta \coloneqq 2 \left(\frac{2 \cdot e \cdot t \cdot d_1^2 \cdot d_2^2 \cdot m \cdot (L^{(1)}+1)}{W_{L^{(1)}+1}} \right)^{W_{L^{(1)}+1}}$$

distinct sign patterns when $\mathbf{a}_{I(L^{(1)}+1)} \in \mathbb{R}^{W_{L^{(1)}+1}}$ and therefore the above collection of polynomials also attains at most Δ distinct sign patterns when $\mathbf{a}$ varies in $\mathbb{R}^W$ since the above polynomials depend only on the $W_{L^{(1)}+1}$ variables $\mathbf{a}_{I(L^{(1)}+1)}$ of $\mathbf{a}$. Therefore, we can partition $S \subset \mathbb{R}^W$ into Δ subregions, such that all the polynomials don't change their signs within each subregion. Doing this for all regions $S \in \mathcal{S}_{L^{(1)}}$ we get our required partition $\mathcal{S}^{(1)}$ by assembling all of these subregions. For the size of our partition $\mathcal{S}^{(1)}$ we get

$$|\mathcal{S}^{(1)}| \leq \prod_{r=1}^{L^{(1)}} 2 \cdot \left(\frac{2 \cdot t \cdot e \cdot k_r \cdot d_1 \cdot d_2 \cdot m \cdot r}{W_r} \right)^{W_r} \cdot 2 \cdot \left(\frac{2 \cdot e \cdot t \cdot d_1^2 \cdot d_2^2 \cdot m \cdot (L^{(1)}+1)}{W_{L^{(1)}+1}} \right)^{W_{L^{(1)}+1}}.$$

Fix some $S' \in \mathcal{S}^{(1)}$. Notice that, when $\mathbf{a}$ varies in S', all the polynomials

$$\Big\{ p_{(i_1,j_1),b,S,\mathbf{x}_k}(\mathbf{a}_{I(L^{(1)}+1)}) - p_{(i_2,j_2),b,S,\mathbf{x}_k}(\mathbf{a}_{I(L^{(1)}+1)}) : \\ (i_1,j_1),(i_2,j_2) \in D, (i_1,j_1) \neq (i_2,j_2), b \in \{1,\dots,t\}, k \in \{1,\dots,m\} \Big\}$$

don't change their signs. Hence, there is a permutation $\pi^{(b,k)}$ of the set

$$\{1,\dots,d_1 - M_{L^{(1)}} + 1\} \times \{1,\dots,d_2 - M_{L^{(1)}} + 1\}$$

for any $b \in \{1,\dots,t\}$ and $k \in \{1,\dots,m\}$ such that

$$f_{\pi^{(b,k)}((1,1)),b,\mathbf{x}_k}(\mathbf{a}_{I(L^{(1)}+1)}) \geq \dots \geq f_{\pi^{(b,k)}((d_1 - M_{L^{(1)}}+1, d_2 - M_{L^{(1)}}+1)),b,\mathbf{x}_k}(\mathbf{a}_{I(L^{(1)}+1)})$$

for $\mathbf{a} \in S'$ and any $k \in \{1,\dots,m\}$ and $b \in \{1,\dots,t\}$. Therefore, it holds that

$$f_{b,k}(\mathbf{a}) = \max \left\{ f_{(1,1),b,\mathbf{x}_k}\left(\mathbf{a}_{I(L^{(1)}+1)}\right), \dots, f_{\left(d_1 - M_{L^{(1)}}+1, d_2 - M_{L^{(1)}}+1\right),b,\mathbf{x}_k}\left(\mathbf{a}_{I(L^{(1)}+1)}\right) \right\}$$

$$= f_{\pi^{(b,k)}((1,1)),b,\mathbf{x}_k}(\mathbf{a}_{I(L^{(1)}+1)}),$$

for $\mathbf{a} \in S'$. Since $f_{\pi^{(b,k)}((1,1)),b,\mathbf{x}_k}(\mathbf{a}_{I(L^{(1)}+1)})$ is a polynomial within S', also $f_{b,k}(\mathbf{a})$ is a polynomial within S' with degree no more than $L^{(1)}+1$ and in the $W_{L^{(1)}+1}$ variables $\mathbf{a}_{I(L^{(1)}+1)}$ of $\mathbf{a} \in \mathbb{R}^W$.

In the second step we construct the partition $\mathcal{S}$ starting from partition $\mathcal{S}^{(1)}$ such that within each region $S \in \mathcal{S}$ the functions $h_k(\cdot)$ are all fixed polynomials of degree of at most $L^{(1)}+L^{(2)}+2$ for $k \in \{1,\dots,m\}$. We have

$$h_k(\mathbf{a}) = \sum_{i=1}^{k_{L^{(1)}+L^{(2)}+1}} w_i^{(L^{(2)})} \cdot g_{i,k}^{(L^{(2)})}\left(\mathbf{a}_{I(L^{(1)}+L^{(2)}+1)}\right) + w_0^{(L^{(2)})} - y_k$$

where the $g_{i,k}^{(L^{(2)})}$ are recursively defined by

$$g_{i,k}^{(r)}\left(\mathbf{a}_{I(L^{(1)}+r+1)}\right) = \sigma\left(\sum_{j=1}^{k_{L^{(1)}+r}} w_{i,j}^{(r-1)} g_{j,k}^{(r-1)}(\mathbf{a}_{I(L^{(1)}+r)})\right)$$

for $r \in \{1,\dots,L^{(2)}\}$ and

$$g_{i,k}^{(0)}(\mathbf{a}_{I(L^{(1)}+1)}) = f_{i,k}(\mathbf{a})$$

for $i \in \{1,\dots,k_{L^{(1)}+1}\}$ ($k_{L^{(1)}+1} = t$). As above we construct the partition $\mathcal{S}$ iteratively layer by layer, by creating a sequence $\mathcal{S}_0,\dots,\mathcal{S}_{L^{(2)}}$, where each $\mathcal{S}_r$ is a partition of $\mathbb{R}^W$ with the following porperties:

1. We set $\mathcal{S}_0 = \mathcal{S}^{(1)}$ and, for each $r \in \{1,\dots,L^{(2)}\}$,

$$\frac{|\mathcal{S}_r|}{|\mathcal{S}_{r-1}|} \le 2\left(\frac{2 \cdot e \cdot k_{L^{(1)}+r+1} \cdot m \cdot (L^{(1)}+r+1)}{W_{L^{(1)}+r+1}}\right)^{W_{L^{(1)}+r+1}}, \tag{34}$$

2. For each $r \in \{0,\dots,L^{(2)}\}$, and each element $S \in \mathcal{S}_r$, each $i \in \{1,\dots,k_{L^{(1)}+r+1}\}$, and each $k \in \{1,\dots,m\}$ when $\mathbf{a}$ varies in S,

$$g_{i,k}^{(r)}(\mathbf{a}_{I(L^{(1)}+r+1)})$$

   is a fixed polynomial function in the $W_{L^{(1)}+r+1}$ variables $\mathbf{a}_{I(L^{(1)}+r+1)}$ of $\mathbf{a}$, of total degree no more than $L^{(1)}+r+1$.

As we have already shown in step 1, property 2 above is satisfied for $r=0$. Now suppose that $\mathcal{S}_0,\dots,\mathcal{S}_{r-1}$ have been defined, and we want to define $\mathcal{S}_r$. For $S \in \mathcal{S}_{r-1}$ and $j \in \{1,\dots,k_{L^{(1)}+r}\}$ let $p_{j,k,S}(\mathbf{a}_{I(L^{(1)}+r)})$ denote the function $g_{j,k}^{(r-1)}(\mathbf{a}_{I(L^{(1)}+r)})$, when $\mathbf{a} \in S$. By induction hypothesis $p_{j,k,S}(\mathbf{a}_{I(L^{(1)}+r)})$ is a polynomial with total degree no more than $L^{(1)} + r$, and depends on the $W_{L^{(1)}+r}$ variables $\mathbf{a}_{I(L^{(1)}+r)}$ of $\mathbf{a}$. Hence for any $k \in \{1, \ldots, m\}$ and $i \in \{1, \ldots, k_{L^{(1)}+r+1}\}$

$$\sum_{j=1}^{k_{L^{(1)}+r}} w_{(i,j)}^{(r-1)} \cdot p_{j,k,S}(\mathbf{a}_{I(L^{(1)}+r)}) + w_{i,0}^{(r-1)}$$

is a polynomial in the $W_{L^{(1)}+r+1}$ variables $\mathbf{a}_{I(L^{(1)}+r+1)}$ variables of $\mathbf{a}$ with total degree no more than $L^{(1)} + r + 1$. Because of condition (31) we have $k_{L^{(1)}+r+1} \cdot m \geq W_{L^{(1)}+r+1}$. Hence, by Lemma 6, the collection of polynomials

$$\left\{ \sum_{j=1}^{k_{L^{(1)}+r}} w_{(i,j)}^{(r-1)} \cdot p_{j,k,S}(\mathbf{a}_{I(L^{(1)}+r)}) + w_{i,0}^{(r-1)} : k \in \{1, \ldots, m\}, i \in \{1, \ldots, k_{L^{(1)}+r+1}\} \right\}$$

attains at most

$$\Pi := 2 \left(\frac{2 \cdot e \cdot k_{L^{(1)}+r+1} \cdot m \cdot (L^{(1)} + r + 1)}{W_{L^{(1)}+r+1}} \right)^{W_{L^{(1)}+r+1}}$$

distinct sign patterns when $\mathbf{a}_{I(L^{(1)}+r+1)} \in \mathbb{R}^{W_{L^{(1)}+r+1}}$ and therefore the above collection of polynomials also attains at most Π distinct sign patterns when $\mathbf{a}$ varies in $\mathbb{R}^W$ since the above polynomials depend only on the $W_{L^{(1)}+r+1}$ variables $\mathbf{a}_{I(L^{(1)}+r+1)}$ of $\mathbf{a}$. Therefore, we can partition $S \subset \mathbb{R}^W$ into Π subregions, such that all the polynomials don't change their signs within each subregion. Doing this for all regions $S \in \mathcal{S}_{r-1}$ we get our required partition $\mathcal{S}_r$ by assembling all of these subregions. In particular property 1 is then satisfied. In order to see that condition 2 is also satisfied, we can proceed analogously to step 1. Hence, when $\mathbf{a}$ varies in $S \in \mathcal{S}$ the function

$$h_k(\mathbf{a}) = \sum_{i=1}^{k_{L^{(1)}+L^{(2)}+1}} w_i^{(L)} \cdot g_{i,k}^{(L^{(2)})} \left(\mathbf{a}_{I(L^{(1)}+L^{(2)}+1)} \right) + w_0^{(L)} - y_k$$

is a polynomial of degree no more than $L^{(1)} + L^{(2)} + 2$ in the W variables of $\mathbf{a} \in \mathbb{R}^W$ for any $k \in \{1, \ldots, m\}$. For the size of our partition $\mathcal{S}$ we get

$$\begin{aligned}
|\mathcal{S}| \leq & \prod_{r=1}^{L^{(1)}} 2 \cdot \left(\frac{2 \cdot e \cdot t \cdot k_r \cdot d_1 \cdot d_2 \cdot m \cdot r}{W_r} \right)^{W_r} \cdot 2 \cdot \left(\frac{2 \cdot e \cdot d_1^2 \cdot d_2^2 \cdot m \cdot (L^{(1)} + 1)}{W_{L^{(1)}+1}} \right)^{W_{L^{(1)}+1}} \\
& \cdot \prod_{r=1}^{L^{(2)}} 2 \cdot \left(\frac{2 \cdot e \cdot k_{L^{(1)}+r+1} \cdot m \cdot (L^{(1)} + r + 1)}{W_{L^{(1)}+r+1}} \right)^{W_{L^{(1)}+r+1}} \\
\leq & \prod_{r=1}^{L^{(1)}+L^{(2)}+1} 2 \cdot \left(\frac{2 \cdot e \cdot t \cdot k_r \cdot d_1^2 \cdot d_2^2 \cdot m \cdot r}{W_r} \right)^{W_r}
\end{aligned}$$

By condition (31) and another application of Lemma 6 it holds for any $S' \in \mathcal{S}$ that

$$\begin{aligned}&|\{(\operatorname{sgn}(h_1(\mathbf{a})), \dots, \operatorname{sgn}(h_m(\mathbf{a}))) : \mathbf{a} \in S'\}| \\ &\leq 2 \cdot \left(\frac{2 \cdot e \cdot m \cdot (L^{(1)} + L^{(2)} + 2)}{W} \right)^W .\end{aligned}$$

Now we are able to bound K via equation (32) and because K is an upper bound for the growth function we set $k_{L^{(1)}+L^{(2)}+2} = 1$ and get

$$\begin{aligned}\Pi_{\operatorname{sgn}(\mathcal{H})}(m) \leq & \prod_{r=1}^{L^{(1)}+L^{(2)}+2} 2 \cdot \left(\frac{2 \cdot e \cdot t \cdot k_r \cdot d_1^2 \cdot d_2^2 \cdot r \cdot m}{W_r} \right)^{W_r} \\ \leq & 2^{L^{(1)}+L^{(2)}+2} \cdot \left(\frac{\sum_{r=1}^{L^{(1)}+L^{(2)}+2} 2 \cdot e \cdot t \cdot k_r \cdot d_1^2 \cdot d_2^2 \cdot r \cdot m}{\sum_{r=1}^{L^{(1)}+L^{(2)}+2} W_r} \right)^{\sum_{r=1}^{L^{(1)}+L^{(2)}+2} W_r} \\ = & 2^{L^{(1)}+L^{(2)}+2} \cdot \left(\frac{R \cdot m}{\sum_{r=1}^{L^{(1)}+L^{(2)}+2} W_r} \right)^{\sum_{r=1}^{L^{(1)}+L^{(2)}+2} W_r} , \qquad (35)\end{aligned}$$

with $R \coloneqq 2 \cdot e \cdot t \cdot d_1^2 \cdot d_2^2 \cdot \sum_{r=1}^{L^{(1)}+L^{(2)}+2} k_r \cdot r$. In the third row we used the weighted AM-GM inequality. Without loss of generality, we can assume that $\operatorname{VCdim}(\mathcal{H}) \geq \sum_{r=1}^{L^{(1)}+L^{(2)}+2} W_r$ because in the case $\operatorname{VCdim}(\mathcal{H}) < \sum_{r=1}^{L^{(1)}+L^{(2)}+2} W_r$ we have

$$\begin{aligned}\operatorname{VCdim}(\mathcal{H}) &< (L^{(1)} + L^{(2)} + 2) \cdot W \\ &\overset{(30)}{\leq} 2 \cdot (L^{(1)} + L^{(2)} + 2)^2 \cdot t \cdot M_{max}^2 \cdot k_{max}^2 \\ &\leq c_{10} \cdot L_{max}^2\end{aligned}$$

for some constant $c_{10} > 0$ which only depends on M_{max} and k_{max} and get the assertion by Lemma 5. Hence we get by the definition of the VC–dimension and inequality (35) (which only holds for $m \geq W$)

$$2^{\operatorname{VCdim}(\mathcal{H})} = \Pi_{\operatorname{sgn}(\mathcal{H})}(\operatorname{VCdim}(\mathcal{H})) \leq 2^{L^{(1)}+L^{(2)}+2} \cdot \left(\frac{R \cdot \operatorname{VCdim}(\mathcal{H})}{\sum_{r=1}^{L^{(1)}+L^{(2)}+2} W_r} \right)^{\sum_{r=1}^{L^{(1)}+L^{(2)}+2} W_r} .$$

Since

$$R \geq 2 \cdot e \cdot t \cdot d_1^2 \cdot d_2^2 \cdot \sum_{r=1}^{1+1+2} r \geq 2 \cdot e \cdot t \cdot d_1^2 \cdot d_2^2 \cdot 10 \geq 16$$

Lemma 8 below (with parameters R, $m = \operatorname{VCdim}(\mathcal{H})$, $w = \sum_{r=1}^{L^{(1)}+L^{(2)}+2} W_r$ and $L = L^{(1)} + L^{(2)} + 2$) implies that

$$\operatorname{VCdim}(\mathcal{H}) \leq (L^{(1)} + L^{(2)} + 2) + \left(\sum_{r=1}^{L^{(1)}+L^{(2)}+2} W_r \right) \cdot \log_2(2 \cdot R \cdot \log_2(R))$$

$$
\begin{aligned}
&\leq (L^{(1)} + L^{(2)} + 2) + (L^{(1)} + L^{(2)} + 2) \cdot W \\
&\qquad \cdot \log_2(2 \cdot (2 \cdot e \cdot t \cdot d_1^2 \cdot d_2^2 \cdot (L^{(1)} + L^{(2)} + 2) \cdot k_{max})^2) \\
&\leq 2 \cdot (L^{(1)} + L^{(2)} + 2) \cdot W \cdot \log_2 \left(\left(2 \cdot e \cdot t \cdot (L^{(1)} + L^{(2)} + 2) \cdot k_{max} \cdot d_1 \cdot d_2 \right)^4 \right) \\
&\overset{(30)}{\leq} 16 \cdot t \cdot (L^{(1)} + L^{(2)} + 2)^2 \cdot k_{\max}^2 \cdot M_{\max}^2 \\
&\qquad \cdot \log_2 \left(2 \cdot e \cdot t \cdot (L^{(1)} + L^{(2)} + 2) \cdot k_{max} \cdot d_1 \cdot d_2 \right) \\
&\leq c_{10} \cdot L_{max}^2 \cdot \log_2(L_{max} \cdot d_1 \cdot d_2),
\end{aligned}
$$

for some constant $c_{10} > 0$ which only depends on $k_{\max}$ and $M_{\max}$. In the third row we used equation (30) for the total number of weights W. Now we make use of Lemma 5 and finally get

$$
V_{\mathcal{F}^+} \leq c_{10} \cdot L_{max}^2 \cdot \log_2(L_{max} \cdot d_1 \cdot d_2).
$$

□

Lemma 8 *Suppose that $2^m \leq 2^L \cdot (m \cdot R/w)^w$ for some $R \geq 16$ and $m \geq w \geq L \geq 0$. Then,*

$$
m \leq L + w \cdot \log_2(2 \cdot R \cdot \log_2(R)).
$$

Proof. See Lemma 16 in Bartlett et al. (2019). □

Proof of Lemma 4. Using Lemma 7 and

$$
V_{T_{c_4 \cdot \log n}\mathcal{F}^+} \leq V_{\mathcal{F}^+},
$$

we can conclude from this together with Lemma 9.2 and Theorem 9.4 in Györfi et al. (2002)

$$
\begin{aligned}
&\mathcal{N}_1 \left(\epsilon, T_{c_4 \cdot \log n}\mathcal{F}, \mathbf{x}_1^n \right) \\
&\leq 3 \cdot \left(\frac{4e \cdot c_4 \cdot \log n}{\epsilon} \cdot \log \frac{6e \cdot c_4 \cdot \log n}{\epsilon} \right)^{V_{T_{c_4 \cdot \log n}\mathcal{F}^+}} \\
&\leq 3 \cdot \left(\frac{6e \cdot c_4 \cdot \log n}{\epsilon} \right)^{2 \cdot c_{10} \cdot L_{max}^2 \cdot \log(L_{max} \cdot d_1 \cdot d_2)}.
\end{aligned}
$$

This completes the proof of Lemma 4. □

C. Auxiliary results

In this section we present several auxiliary results from the literature which we have used in the proof of Theorem 1. Our first result is a bound on the expected L_2 error of the (truncated) least squares regression estimate.

Lemma 9 *Let* $(\mathbf{X}, Y)$, $(\mathbf{X}_1, Y_1)$, ..., $(\mathbf{X}_n, Y_n)$ *be independent and identically distributed* $\mathbb{R}^d \times \mathbb{R}$*-valued random variables. Assume that the distribution of* $(\mathbf{X}, Y)$ *satisfies*

$$\mathbf{E}\{\exp(c_3 \cdot Y^2)\} < \infty$$

for some constant $c_3 > 0$ *and that the regression function* $m(\cdot) = \mathbf{E}\{Y|\mathbf{X} = \cdot\}$ *is bounded in absolute value. Let* $\tilde{m}_n$ *be the least squares estimate*

$$\tilde{m}_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(\mathbf{X}_i)|^2$$

based on some function space $\mathcal{F}_n$ *consisting of functions* $f : \mathbb{R}^d \to \mathbb{R}$ *and set* $m_n = T_{c_4 \cdot \log(n)} \tilde{m}_n$ *for some constant* $c_4 > 0$*. Then* m_n *satisfies*

$$\begin{aligned}
&\mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\
&\leq \frac{c_5 \cdot (\log(n))^2 \cdot \sup_{\mathbf{x}_1^n \in (\mathbb{R}^d)^n} \left(\log \left(\mathcal{N}_1 \left(\frac{1}{n \cdot c_4 \log(n)}, T_{c_4 \log(n)} \mathcal{F}_n, \mathbf{x}_1^n \right) \right) + 1 \right)}{n} \\
&\quad + 2 \cdot \inf_{f \in \mathcal{F}_n} \int |f(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x})
\end{aligned}$$

for $n > 1$ *and some constant* $c_5 > 0$*, which does not depend on* n *or the parameters of the estimate.*

Proof. This result follows in a straightforward way from the proof of Theorem 1 in Bagirov et al. (2009). A complete proof can be found in the supplement of Bauer and Kohler (2019). □

Our second auxiliary result is an approximation result for (p, C)–smooth functions by very deep feedforward neural networks.

Lemma 10 *Let* $d \in \mathbb{N}$*, let* $f : \mathbb{R}^d \to \mathbb{R}$ *be* (p, C)*–smooth for some* $p = q + s$*,* $q \in \mathbb{N}_0$ *and* $s \in (0, 1]$*, and* $C > 0$*. Let* $M \in \mathbb{N}$ *with* $M > 1$ *sufficiently large, where*

$$M^{2p} \geq c_5 \cdot \left(\max \left\{ 2, \sup_{\substack{\mathbf{x} \in [-2,2]^d \\ (l_1, \ldots, l_d) \in \mathbb{N}^d \\ l_1 + \cdots + l_d \leq q}} \left| \frac{\partial^{l_1 + \cdots + l_d} f}{\partial^{l_1} x^{(1)} \ldots \partial^{l_d} x^{(d)}} (\mathbf{x}) \right| \right\} \right)^{4(q+1)}$$

must hold for some sufficiently large constant $c_5 \geq 1$*. Let* $\sigma : \mathbb{R} \to \mathbb{R}$ *be the ReLU activation function*

$$\sigma(x) = \max\{x, 0\}$$

and let $L, r \in \mathbb{N}$ *such that*

(i)

$$
\begin{aligned}
L \geq & 5M^d + \left\lceil \log_4 \left(M^{2p+4\cdot d\cdot(q+1)} \cdot e^{4(q+1)\cdot(M^d-1)} \right) \right\rceil \\
& \cdot \lceil \log_2(\max\{d,q\}+2) \rceil + \lceil \log_4(M^{2p}) \rceil
\end{aligned}
$$

(ii)

$$
r \geq 132 \cdot 2^d \cdot \lceil e^d \rceil \cdot \binom{d+q}{d} \cdot \max\{q+1, d^2\}
$$

hold. Then there exists a feedforward neural network

$$
f_{net} \in \mathcal{G}_d(L, \mathbf{k})
$$

with $\mathbf{k} = (k_1, \dots, k_L)$ *and* $k_1 = \dots = k_L = r$ *such that*

$$
\begin{aligned}
& \sup_{\mathbf{x}\in[-2,2]^d} |f(\mathbf{x}) - f_{net}(\mathbf{x})| \\
& \leq c_6 \cdot \left(\max \left\{ 2, \sup_{\substack{\mathbf{x}\in[-2,2]^d \\ (l_1,\dots,l_d)\in\mathbb{N}^d \\ l_1+\dots+l_d \leq q}} \left| \frac{\partial^{l_1+\dots+l_d} f}{\partial^{l_1} x^{(1)} \dots \partial^{l_d} x^{(d)}}(\mathbf{x}) \right| \right\} \right)^{4(q+1)} \cdot M^{-2p}.
\end{aligned}
$$

Proof. See Theorem 2 in Kohler and Langer (2021). An alternative proof of a closely related result can be found in Yarotsky and Zhevnerchuk (2019), see Theorem 4.1 therein. □

D. A minimax lower bound

In this section, we show that the rate of convergence of our truncated least squares estimate introduced in the proof of Theorem 1, up to a logarithmic factor, is in some sense an optimal minimax rate of convergence. To show the upper bound (6) in Theorem 1, we used equation (2) and then derived an upper bound on

$$
\mathbf{E} \left\{ \int |\eta_n(\mathbf{x}) - \eta(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\}, \tag{36}
$$

where $\eta_n(\cdot, \mathcal{D}_n)$ is a (truncated) estimate of the a posteriori probability η. Thus, we solve the classification problem via regression estimation. In our theorem we have made the assumption that $(\mathbf{X}, Y)$ is a $[0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \times \{0,1\}$-valued random variable and the a posteriri probability satifies a generalized hierarchical max-pooling model given by the functions

$$
\left\{ g_{k,s}^{(a)} : \mathbb{R}^4 \to [0,1] \right\}_{k=1,\dots,l,s=1,\dots,4^{l-k},a=1,\dots,d^*} \quad \text{and} \quad g : \mathbb{R}^{d^*} \to [0,1]
$$

However, we have shown our upper bound for (36) for a more general case in which $(\mathbf{X}, Y)$ is a $[0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \times \mathbb{R}$-valued random variable satisfying

$$\mathbf{E}\left\{\exp\left(c_3 \cdot Y^2\right)\right\} < \infty$$

for a constant $c_3 > 0$, where the regression function $\eta(\mathbf{x}) = \mathbf{E}\{Y|\mathbf{X} = \mathbf{x}\}$ satisfies a generalized hierarchical max-pooling model with functions

$$\left\{g_{k,s}^{(a)} : \mathbb{R}^4 \to [-1/2, 3/2]\right\}_{k=1,\dots,l, s=1,\dots,4^{l-k}, a=1,\dots,d^*} \quad \text{and} \quad g : \mathbb{R}^{d^*} \to [-1/2, 3/2] \quad (37)$$

such that

$$g([0,1]^{d^*}) \subseteq [0,1] \quad \text{and} \quad g_{k,s}^{(a)}([0,1]^4) \subseteq [0,1].$$

The aim of this section is to show that for this more general class of distributions

$$\max\left\{n^{-\frac{2\cdot p_1}{2\cdot p_1+4}}, n^{-\frac{2\cdot p_2}{2\cdot p_2+d^*}}\right\}$$

is also a lower minimax rate of convergence, which means that our estimate for this class has an optimal rate of convergence up to the logarithmic factor.

In the following, for $d \in \mathbb{N}$ and $A \subseteq \mathbb{R}^d$, $\mathcal{H}_A^{(p,C)}$ denote the class of all (p, C)-smooth functions $h : \mathbb{R}^d \to [0,1]$ with *supp*$(h) \subseteq A$ and $\mathcal{G}_d^{(p,C)}$ denote the class of all (p, C)-smooth functions $g : \mathbb{R}^d \to [-1/2, 3/2]$ with $g([0,1]^d) \subseteq [0,1]$. Furthermore, $\mathcal{H}_{l,d^*}^{(p_1,p_2)}$ denote the class of all real-valued functions on $[0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}}$ that satisfy a generalized hierarchical max-pooling model of order d^* and level l such that the functions (37) are choosen from $\mathcal{G}_4^{(p_1,C_1)}$ and $\mathcal{G}_{d^*}^{(p_2,C_2)}$ for some $C_1, C_2 > 0$, respectively.

Lemma 11 *Let $p_1, p_2 \in [1, \infty)$ and $d_1, d_2, d^*, l \in \mathbb{N}$ with $d_1, d_2 > 1$, and $\sqrt{d^*} \leq 2^l \leq \min\{d_1, d_2\}$.*
a) *We set $I_1 = \{1,2\} \times \{1,2\}$ and define the subset $\mathcal{A}_1 \subset [0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}}$ by*

$$\mathcal{A}_1 = \left\{(a_{\mathbf{i}})_{\mathbf{i}\in\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \ : \ a_{\mathbf{i}} \in [0,1] \ (\mathbf{i} \in I_1), \ a_{\mathbf{i}} = 0 \ (\mathbf{i} \notin I_1)\right\}.$$

For each $h \in \mathcal{H}_{[0,1]^4}^{(p_1,C_1)}$ there exist $\eta_h \in \mathcal{H}_{l,d^*}^{(p_1,p_2)}$ such that

$$\eta_h(\mathbf{x}) = h(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}) \quad \left(\mathbf{x} \in \mathcal{A}_1\right).$$

b) *Let $I_2 = \{\mathbf{i}_1, \dots, \mathbf{i}_{d^*}\} \subseteq \{1, \dots, 2^l\} \times \{1, \dots, 2^l\}$ such that $\mathbf{i}_1, \dots, \mathbf{i}_{d^*}$ are pairwise distinct. For $j \in \{1, \dots, d^*\}$ we define the interval*

$$A_j = \left[\frac{3\cdot j - 2}{3\cdot d^*}, \frac{3\cdot j - 1}{3\cdot d^*}\right], \tag{38}$$

and define the subset $\mathcal{A}_2 \subseteq [0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}}$ by

$$\mathcal{A}_2 = \left\{(a_{\mathbf{i}})_{\mathbf{i}\in\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \ : \ a_{\mathbf{i}_j} \in A_j \ (j \in \{1, \dots, d^*\}), a_{\mathbf{i}} = 0 \ (\mathbf{i} \notin I_2)\right\}.$$

For each $h \in \mathcal{H}_{A_1\times\cdots\times A_{d^*}}^{(p_2,C_2)}$ there exist $\eta_h \in \mathcal{H}_{l,d^*}^{(p_1,p_2)}$ such that

$$\eta_h(\mathbf{x}) = h(x_{\mathbf{i}_1}, \dots, x_{\mathbf{i}_{d^*}}) \quad \left(\mathbf{x} \in \mathcal{A}_2\right).$$

Proof. a) Let $h \in \mathcal{H}^{(p_1,C_1)}_{[0,1]^4}$. We define $\eta_h \in \mathcal{H}^{(p_1,p_2)}_{l,d^*}$ by choosing the functions

$$g \in \mathcal{G}^{(p_2,C_2)}_{d^*} \quad \text{and} \quad \left\{ g^{(a)}_{k,s} \in \mathcal{G}^{(p_1,C_1)}_{4} \right\}_{k=1,\dots,l,s=1,\dots,4^{l-k},a=1,\dots,d^*}$$

of the corresponding generalized hierarchical max-pooling model as follows:

1. Set $g^{(a)}_{1,s} = h$ for all $s = 1, \dots, 4^{l-1}$ and $a = 1, \dots, d^*$.
2. Choose $g^{(a)}_{k,s} \in \mathcal{G}^{(p_1,C_1)}_{4}$ such that
$$g^{(a)}_{k,s}(\mathbf{x}) = x_1 \tag{39}$$
for $\mathbf{x} \in [0,1]^4$ and all $k = 2, \dots, l$, $s = 1, \dots, 4^{l-k}$ and $a = 1, \dots, d^*$.
3. Choose $g \in \mathcal{G}^{(p_2,C_2)}_{d^*}$ such that $g(\mathbf{x}) = x_1$ for $\mathbf{x} \in [0,1]^{d^*}$.

By using equation (39) and since $supp(h) \subseteq (0,1)^4$ we get

$$\begin{aligned} m_a(\mathbf{x}) &= \max_{(i,j)\in\{1,\dots,d_1-2^l+1\}\times\{1,\dots,d_2-2^l+1\}} f^{(a)}_{l,1}(\mathbf{x}_{\{i,\dots,i+2^l-1\}\times\{j,\dots,j+2^l-1\}}) \\ &= \max_{(i,j)\in\{1,\dots,d_1-2^l+1\}\times\{1,\dots,d_2-2^l+1\}} h(x_{i,j}, x_{i,j+1}, x_{i+1,j}, x_{i+1,j+1}) \\ &= \max\{h(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}), 0\} \\ &= h(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}) \end{aligned}$$

for $\mathbf{x} \in \mathcal{A}_1$ and therefore we have

$$\begin{aligned} \eta_h(\mathbf{x}) &= g(m_1(\mathbf{x}), \dots, m_{d^*}(\mathbf{x})) \\ &= m_1(\mathbf{x}) \\ &= \max_{(i,j)\in\{1,\dots,d_1-2^l+1\}\times\{1,\dots,d_2-2^l+1\}} f^{(1)}_{l,1}(\mathbf{x}_{\{i,\dots,i+2^l-1\}\times\{j,\dots,j+2^l-1\}}) \\ &= h(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}) \end{aligned}$$

for all $\mathbf{x} \in \mathcal{A}_1$.
b) Let $h \in \mathcal{H}^{(p_2,C_2)}_{A_1\times\dots\times A_{d^*}}$ and define $\eta_h \in \mathcal{H}^{(p_1,p_2)}_{l,d^*}$ as follows:

1. We choose $j(k,s,a) \in \{1,\dots,4\}$ and $g^{(a)}_{k,s} \in \mathcal{G}^{(p_1,C_1)}_{4}$ for some $C_1 > 0$ sufficiently large with
$$g^{(a)}_{k,s}(\mathbf{x}) = \begin{cases} x_{j(k,s,a)} & , \text{ if } x_{j(k,s,a)} \in A_a = \left[\frac{3\cdot a-2}{3\cdot d^*}, \frac{3\cdot a-1}{3\cdot d^*}\right] \\ 0 & , \text{ if } x_{j(k,s,a)} \in [0,1] \setminus \left(\frac{3\cdot a-3}{3\cdot d^*}, \frac{3\cdot a}{3\cdot d^*}\right) \end{cases}$$
for $\mathbf{x} \in [0,1]^4$ and all $k = 1, \dots, l$ and $s = 1, \dots, 4^{l-k}$ such that
$$f^{(a)}_{l,1}(\mathbf{x}) = \begin{cases} x_{\mathbf{i}_a} & , \text{ if } x_{\mathbf{i}_a} \in A_a \\ 0 & , \text{ if } x_{\mathbf{i}_a} \in [0,1] \setminus \left(\frac{3\cdot a-3}{3\cdot d^*}, \frac{3\cdot a}{3\cdot d^*}\right) \end{cases} \tag{40}$$
for $\mathbf{x} \in [0,1]^{\{1,\dots,2^l\}\times\{1,\dots,2^l\}}$ and all $a \in \{1,\dots,d^*\}$.

2. We set $g = h$.

For $\mathbf{x} \in \mathcal{A}_2$ equation (40) yields

$$\begin{aligned}
m_a(\mathbf{x}) &= \max_{(i_2,j_2)\in\{1,\dots,d_1-2^l+1\}\times\{1,\dots,d_2-2^l+1\}} f_{l,1}^{(a)}(\mathbf{x}_{\{i_2,\dots,i_2+2^l-1\}\times\{j_2,\dots,j_2+2^l-1\}}) \\
&= \max\{x_{\mathbf{i}_a}, 0\} \\
&= x_{\mathbf{i}_a}
\end{aligned}$$

for $a \in \{1, \dots, d^*\}$, which imply

$$\eta_h(\mathbf{x}) = h(x_{\mathbf{i}_1}, \dots, x_{\mathbf{i}_{d^*}}).$$

□

To show the lower minimax rate of convergence, we use the following lemma, which is a modification of Theorem 3.2 in Györfi et al. (2002).

Lemma 12 *Let $d \in \mathbb{N}$, let $\epsilon > 0$, and let $a_1, \dots, a_d \in \mathbb{R}$. Define the cube*

$$A = [a_1, a_1 + \epsilon] \times \cdots \times [a_d, a_d + \epsilon] \subset \mathbb{R}^d$$

and let $\mathcal{D}_A^{(p,C)}$ be the class of distributions of $(\mathbf{X}, Y)$ such that:

1. *$\mathbf{X}$ is uniformly distributed on A,*
2. *$Y = \eta(\mathbf{X}) + N$, where $\mathbf{X}$ and N are independent and N is standard normal, and $\eta \in \mathcal{H}_A^{(p,C)}$.*

Then

$$\liminf_{n\to\infty} \inf_{\eta_n} \sup_{(X,Y)\in\mathcal{D}_A^{(p,C)}} \frac{\mathbf{E}\left\{\int |\eta_n(\mathbf{x}) - g(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x})\right\}}{n^{-\frac{2\cdot p}{2\cdot p+d}}} \geq C_1 > 0.$$

In the proof of Lemma 12 we follow the proof of Theorem 3.2 in Györfi et al. (2002).

Proof. Firstly, we define a subclass of $\mathcal{D}_A^{(p,C)}$ for sufficiently large n. Therefore set

$$M_n = \left\lceil (C^2 \cdot n)^{\frac{1}{2p+d}} \right\rceil$$

and partition the cube A into M_n^d equal sized cubes $\{A_{n,j}\}_{j\in\{1,\dots,M_n^d\}}$ with side length ϵ/M_n and centers $\{\mathbf{a}_{n,j}\}_{j\in\{1,\dots,M_n^d\}}$. Set $k = \lfloor p \rfloor$, $\beta = p - k$, and let $g : \mathbb{R}^d \to \mathbb{R}_+$ be a bounded and $(p, (\epsilon^p/2) \cdot 2^{\beta-1} \cdot C)$-smooth function with *supp*$(g) \subset (-1/2, 1/2)$ and $\int g^2(\mathbf{x})d\mathbf{x} > 0$. For $j \in \{1, \dots, M_n^d\}$ we define the functions $g_{n,j} : \mathbb{R}^d \to \mathbb{R}_+$ by

$$g_{n,j}(\mathbf{x}) = M_n^{-p} \cdot g\left(\frac{M_n}{\epsilon} \cdot (\mathbf{x} - \mathbf{a}_{n,j})\right) \quad (\mathbf{x} \in \mathbb{R}^d)$$

and for $\mathbf{c}_n = (c_{n,1}, \dots, c_{n,M_n^d}) \in \{-1, 1\}^{M_n^d} =: \mathbb{C}_n$ we define the function

$$m^{(\mathbf{c}_n)}(\mathbf{x}) = \sum_{j=1}^{M_n^d} (1 + c_{n,j}) \cdot g_{n,j}(\mathbf{x}) \quad (\mathbf{x} \in \mathbb{R}^d).$$

Next, we show that $m^{(\mathbf{c}_n)}$ is a (p,C)-smooth function. Let $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{i=1}^d \alpha_i = k$ and set $D^{\boldsymbol{\alpha}} = \frac{\partial^k}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$. For $j \in \{1, \dots, M_n^d\}$ and $\mathbf{x}, \mathbf{z} \in A_{n,j}$ we have

$$
\begin{aligned}
&|D^{\boldsymbol{\alpha}} m^{(\mathbf{c}_n)}(\mathbf{x}) - D^{\boldsymbol{\alpha}} m^{(\mathbf{c}_n)}(\mathbf{z})| \\
&= |1 + c_{n,j}| \cdot |D^{\boldsymbol{\alpha}} g_{n,j}(\mathbf{x}) - D^{\boldsymbol{\alpha}} g_{n,j}(\mathbf{z})| \\
&\leq 2 \cdot M_n^{-p} \cdot \frac{M_n^k}{\epsilon^k} \cdot \frac{\epsilon^p}{2} \cdot 2^{\beta-1} \cdot C \cdot \left\| \frac{M_n}{\epsilon} \cdot (\mathbf{x} - \mathbf{a}_{n,j}) - \frac{M_n}{\epsilon} \cdot (\mathbf{z} - \mathbf{a}_{n,j}) \right\|^\beta \\
&= 2^{\beta-1} \cdot C \cdot \|\mathbf{x} - \mathbf{z}\|^\beta \\
&\leq C \cdot \|\mathbf{x} - \mathbf{z}\|^\beta.
\end{aligned}
$$

Now let $i, j \in \{1, \dots, M_n^d\}$ with $i \neq j$, and $\mathbf{x} \in A_{n,i}$ and $\mathbf{z} \in A_{n,j}$. We choose $\bar{\mathbf{x}}$ on the boundary of $A_{n,i}$ and $\bar{\mathbf{z}}$ on the boundary of $A_{n,j}$ such that $\|\mathbf{x} - \bar{\mathbf{x}}\| + \|\mathbf{z} - \bar{\mathbf{z}}\| \leq \|\mathbf{x} - \mathbf{z}\|$ and get

$$
\begin{aligned}
&|D^{\boldsymbol{\alpha}} m^{(\mathbf{c}_n)}(\mathbf{x}) - D^{\boldsymbol{\alpha}} m^{(\mathbf{c}_n)}(\mathbf{z})| \\
&= |(1 + c_{n,i}) \cdot D^{\boldsymbol{\alpha}} g_{n,i}(\mathbf{x}) - (1 + c_{n,j}) \cdot D^{\boldsymbol{\alpha}} g_{n,j}(\mathbf{z})| \\
&\leq |(1 + c_{n,i}) \cdot D^{\boldsymbol{\alpha}} g_{n,i}(\mathbf{x})| + |(1 + c_{n,j}) \cdot D^{\boldsymbol{\alpha}} g_{n,j}(\mathbf{z})| \\
&= |1 + c_{n,i}| \cdot |D^{\boldsymbol{\alpha}} g_{n,i}(\mathbf{x}) - \underbrace{D^{\boldsymbol{\alpha}} g_{n,i}(\bar{\mathbf{x}})}_{=0}| + |1 + c_{n,j}| \cdot |D^{\boldsymbol{\alpha}} g_{n,j}(\mathbf{z}) - \underbrace{D^{\boldsymbol{\alpha}} g_{n,j}(\bar{\mathbf{z}})}_{=0}| \\
&\leq 2^{\beta-1} \cdot C \cdot (\|\mathbf{x} - \bar{\mathbf{x}}\|^\beta + \|\mathbf{z} - \bar{\mathbf{z}}\|^\beta) \\
&= 2^\beta \cdot C \cdot \left(\frac{1}{2} \cdot \|\mathbf{x} - \bar{\mathbf{x}}\|^\beta + \frac{1}{2} \cdot \|\mathbf{z} - \bar{\mathbf{z}}\|^\beta \right) \\
&\leq 2^\beta \cdot C \cdot \left(\frac{1}{2} \cdot \|\mathbf{x} - \bar{\mathbf{x}}\| + \frac{1}{2} \cdot \|\mathbf{z} - \bar{\mathbf{z}}\| \right)^\beta \\
&\leq C \cdot \|\mathbf{x} - \mathbf{z}\|^\beta,
\end{aligned}
$$

where we used that $u \mapsto u^\beta$ is a concave function on $\mathbb{R}_+$. Therefore the class $\tilde{\mathcal{D}}_n^{(p,C)}$ of distributions of $(\mathbf{X}, Y)$ with

1. $\mathbf{X}$ is uniformly distributed on A,
2. $Y = m^{(\mathbf{c}_n)}(\mathbf{X}) + N$ for $\mathbf{c}_n \in \mathbb{C}_n$, where $\mathbf{X}$ and N are independent and N is standard normal.

is a subclass of $\mathcal{D}_A^{(p,C)}$ for sufficiently large n. Therefore it is sufficient to show that

$$
\liminf_{n \to \infty} \inf_{\eta_n} \sup_{(X,Y) \in \tilde{\mathcal{D}}_n^{(p,C)}} \frac{M_n^{2p}}{C^2} \cdot \mathbf{E} \left\{ \int |\eta_n(\mathbf{x}) - m^{(\mathbf{c}_n)}(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} > 0.
$$

Now, let η_n be an arbitrary estimate. Since the supports of the functions $g_{n,j}$ are disjoint by construction, $\{g_{n,j}\}_{j=1,\dots,M_n^d}$ is an orthogonal system with respect to the L_2

inner product. Hence the orthogonal projection $\hat{m}_n$ of m_n to $\{m^{(\mathbf{c}_n)}\}_{\mathbf{c}_n \in \mathbb{C}_n}$ is given by

$$\hat{m}_n(\mathbf{x}) = \sum_{j=1}^{M_n^d} \hat{c}_{n,j} \cdot g_{n,j}(\mathbf{x}),$$

where

$$\hat{c}_{n,j} = \frac{\int_{A_{n,j}} m_n(\mathbf{x}) \cdot g_{n,j}(\mathbf{x}) d\mathbf{x}}{\int_{A_{n,j}} g_{n,j}^2(\mathbf{x}) d\mathbf{x}}.$$

For an arbitrary $\mathbf{c}_n \in \mathbb{C}_n$ we have

$$\begin{aligned}
&\int \left| m_n(\mathbf{x}) - m^{(\mathbf{c}_n)}(\mathbf{x}) \right|^2 d\mathbf{x} \\
&\geq \int \left| \hat{m}_n(\mathbf{x}) - m^{(\mathbf{c}_n)}(\mathbf{x}) \right|^2 d\mathbf{x} \\
&= \int \Big| \sum_{j=1}^{M_n^d} \left(\hat{c}_{n,j} \cdot g_{n,j}(\mathbf{x}) - (1 + c_{n,j}) \cdot g_{n,j}(\mathbf{x}) \right) \Big|^2 d\mathbf{x} \\
&= \sum_{j=1}^{M_n^d} \int_{A_{n,j}} \left| \hat{c}_{n,j} \cdot g_{n,j}(\mathbf{x}) - (1 + c_{n,j}) \cdot g_{n,j}(\mathbf{x}) \right|^2 d\mathbf{x} \\
&= \sum_{j=1}^{M_n^d} \left| \hat{c}_{n,j} - (1 + c_{n,j}) \right|^2 \int_{A_{n,j}} g_{n,j}^2(\mathbf{x}) d\mathbf{x} \\
&= \int g^2(\mathbf{x}) d\mathbf{x} \cdot \frac{\epsilon^d}{M_n^{2p+d}} \cdot \sum_{j=1}^{M_n^d} \left| \hat{c}_{n,j} - (1 + c_{n,j}) \right|^2.
\end{aligned}$$

We set

$$\tilde{c}_{n,j} = \begin{cases} 1, & \text{if } \hat{c}_{n,j} \geq 1 \\ -1, & \text{if } \hat{c}_{n,j} < 1. \end{cases}$$

Because of

$$\left| \hat{c}_{n,j} - (1 + c_{n,j}) \right| \geq \mathbb{I}_{\{\tilde{c}_{n,j} \neq c_{n,j}\}}$$

we get

$$\begin{aligned}
&\int \left| m_n(\mathbf{x}) - m^{(c_n)}(\mathbf{x}) \right|^2 d\mathbf{x} \\
&\geq \int g^2(\mathbf{x}) d\mathbf{x} \cdot \frac{\epsilon^d}{M_n^{2p+d}} \cdot \sum_{j=1}^{M_n^d} \mathbb{I}_{\{\tilde{c}_{n,j} \neq c_{n,j}\}} \\
&= \frac{C^2}{M_n^{2p}} \cdot \underbrace{\int g^2(\mathbf{x}) d\mathbf{x} \cdot \frac{\epsilon^d}{C^2}}_{>0} \cdot \frac{1}{M_n^d} \cdot \sum_{j=1}^{M_n^d} \mathbb{I}_{\{\tilde{c}_{n,j} \neq c_{n,j}\}}
\end{aligned}$$

and therefore it suffices to show that

$$\liminf_{n\to\infty} \inf_{\tilde{\mathbf{c}}_n} \sup_{\mathbf{c}_n\in\mathbb{C}_n} \frac{1}{M_n^d} \sum_{j=1}^{M_n^d} \mathbf{P}\{\tilde{c}_{n,j} \neq c_{n,j}\} > 0. \tag{41}$$

The proof of inequality (41) can be found in the proof of Theorem 3.2 in Györfi et al. (2002). $\square$

Lemma 13 *Let $p_1, p_2 \in [1,\infty)$ and $d_1, d_2, d^*, l \in \mathbb{N}$ with $d_1, d_2 > 1$, and $\sqrt{d^*} \leq 2^l \leq \min\{d_1, d_2\}$. Let $\mathcal{D}$ be the class of distributions of a $[0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \times \mathbb{R}$-valued random variable $(\mathbf{X}, Y)$ such that:*

1. *$\mathbf{E}\left\{\exp\left(c_3 \cdot Y^2\right)\right\} < \infty$,*
2. *$\eta(\cdot) = \mathbf{E}\{Y|\mathbf{X} = \cdot\} \in \mathcal{H}_{l,d^*}^{(p_1,p_2)}$,*

where $\mathcal{H}_{l,d^*}^{(p_1,p_2)}$ is defined as above. Then we have

$$\liminf_{n\to\infty} \inf_{\eta_n} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbf{E}\left\{\int |\eta_n(\mathbf{x}) - \eta(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x})\right\}}{\max\left\{n^{-\frac{2\cdot p_1}{2\cdot p_1+4}}, n^{-\frac{2\cdot p_2}{2\cdot p_2+d^*}}\right\}} \geq C > 0.$$

Proof. The idea of the proof is to use Lemma 1 to find corresponding subclasses of $\mathcal{D}$ which allow us to reduce the assertion to the case of Lemma 12. Set $I_1 = \{1,2\}\times\{1,2\}$, let $I_2 = \{\mathbf{i}_1, \dots, \mathbf{i}_{d^*}\} \subseteq \{1,\dots,2^l\}\times\{1,\dots,2^l\}$ such that $\mathbf{i}_1,\dots,\mathbf{i}_{d^*}$ are pairwise distinct and set $D = \{1,\dots,d_1\}\times\{1,\dots,d_2\}$. We define

$$A^{I_2} = \{(a_{\mathbf{i}})_{\mathbf{i}\in I_2} \ : \ a_{\mathbf{i}_j} \in A_j \ (j \in \{1,\dots,d^*\})\} \subset [0,1]^{I_2}$$

for $A_j \subset [0,1]$ $(j \in \{1,\dots,d^*\})$ defined as in equation (38) and set

$$\mathcal{H}_1 = \left\{\eta_h \in \mathcal{H}_{l,d^*}^{(p_1,p_2)} \ : \ h \in \mathcal{H}_{[0,1]^4}^{(p_1,C_1)}\right\} \text{ and } \mathcal{H}_2 = \left\{\eta_h \in \mathcal{H}_{l,d^*}^{(p_1,p_2)} \ : \ h \in \mathcal{H}_{A_1\times\dots\times A_{d^*}}^{(p_2,C_2)}\right\},$$

where the functions $\eta_h \in \mathcal{H}_{l,d^*}^{(p_1,p_2)}$ are defined as in Lemma 11 a) and b), respectively. We then define two subclasses $\mathcal{D}_1$ and $\mathcal{D}_2$ of $\mathcal{D}$ as follows. Let $\mathcal{D}_1$ the class of distributions of $(\mathbf{X}, Y)$ such that:

1. $\mathbf{X}_{I_1}$ is uniformly distributed on $[0,1]^{I_1}$ and $\mathbf{X}_{D\setminus I_1}$ is concentrated on $\{0\}^{D\setminus I_1}$,
2. $Y = \eta(\mathbf{X}) + N$, where $\mathbf{X}$ and N are independent and N is standard normal, and $\eta \in \mathcal{H}_1$,

and let $\mathcal{D}_2$ the class of distributions of $(\mathbf{X}, Y)$ such that:

1. $\mathbf{X}_{I_2}$ is uniformly distributed on A^{I_2} and $\mathbf{X}_{D\setminus I_2}$ is concentrated on $\{0\}^{D\setminus I_2}$,

2. $Y = \eta(\mathbf{X}) + N$, where $\mathbf{X}$ and N are independent and N is standard normal, and $\eta \in \mathcal{H}_2$.

Furthermore, let $\mathcal{D}^{(I_1)}$ be the class of distributions of $(\mathbf{X}_{I_1}, Y)$ such that

1. $\mathbf{X}_{I_1}$ is uniformly distributed on $[0,1]^{I_1}$
2. $Y = \eta(X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2}) + N$, where $\mathbf{X}_{I_1}$ and N are independent and N is standard normal, and $\eta \in \mathcal{H}^{(p_1, C_1)}_{[0,1]^4}$,

and let $\mathcal{D}^{(I_2)}$ be the class of distributions of $(\mathbf{X}_{I_2}, Y)$ such that

1. $\mathbf{X}_{I_2}$ is uniformly distributed on A^{I_2}
2. $Y = \eta(X_{\mathbf{i}_1}, \dots, X_{\mathbf{i}_{d^*}}) + N$, where $\mathbf{X}_{I_2}$ and N are independent and N is standard normal, and $\eta \in \mathcal{H}^{(p_2, C_2)}_{A_1 \times \dots \times A_{d^*}}$,

By Lemma 11 and Fubini's Theorem we get

$$\liminf_{n\to\infty} \inf_{\eta_n} \sup_{(\mathbf{X},Y)\in\mathcal{D}} \frac{\mathbf{E}\left\{\int |\eta_n(\mathbf{x}) - \eta(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x})\right\}}{\max\left\{n^{-\frac{2\cdot p_1}{2\cdot p_1+4}}, n^{-\frac{2\cdot p_2}{2\cdot p_2+d^*}}\right\}}$$

$$\geq \min\left\{\liminf_{n\to\infty} \inf_{\eta_n} \sup_{(\mathbf{X},Y)\in\mathcal{D}} \frac{\mathbf{E}\left\{\int |\eta_n(\mathbf{x}) - \eta(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x})\right\}}{n^{-\frac{2\cdot p_1}{2\cdot p_1+4}}},\right.$$
$$\left.\liminf_{n\to\infty} \inf_{\eta_n} \sup_{(\mathbf{X},Y)\in\mathcal{D}} \frac{\mathbf{E}\left\{\int |\eta_n(\mathbf{x}) - \eta(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x})\right\}}{n^{-\frac{2\cdot p_2}{2\cdot p_2+d^*}}}\right\}$$

$$\geq \min\left\{\liminf_{n\to\infty} \inf_{\eta_n} \sup_{(\mathbf{X},Y)\in\mathcal{D}_1} \frac{\mathbf{E}\left\{\int |\eta_n(\mathbf{x}) - \eta(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x})\right\}}{n^{-\frac{2\cdot p_1}{2\cdot p_1+4}}},\right.$$
$$\left.\liminf_{n\to\infty} \inf_{\eta_n} \sup_{(\mathbf{X},Y)\in\mathcal{D}_2} \frac{\mathbf{E}\left\{\int |\eta_n(\mathbf{x}) - \eta(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x})\right\}}{n^{-\frac{2\cdot p_2}{2\cdot p_2+d^*}}}\right\}$$

$$= \min\left\{\liminf_{n\to\infty} \inf_{\eta_n} \sup_{(\mathbf{X},Y)\in\mathcal{D}_1}\right.$$
$$\frac{\mathbf{E}_{\mathcal{D}_n^{(I_1)}}\left\{\int |\eta_n((\mathbf{x}_{I_1}, 0, \dots, 0)) - g^{(1)}_{1,1}(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}))|^2 \mathbf{P}_{\mathbf{X}_{I_1}}(d\mathbf{x}_{I_1})\right\}}{n^{-\frac{2\cdot p_1}{2\cdot p_1+4}}},$$
$$\liminf_{n\to\infty} \inf_{\eta_n} \sup_{(\mathbf{X},Y)\in\mathcal{D}_2}$$
$$\left.\frac{\mathbf{E}_{\mathcal{D}_n^{(I_2)}}\left\{\int |\eta_n((\mathbf{x}_{I_2}, 0, \dots, 0)) - g(x_{\mathbf{i}_1}, \dots, x_{\mathbf{i}_{d^*}})|^2 \mathbf{P}_{\mathbf{X}_{I_2}}(d\mathbf{x}_{I_2})\right\}}{n^{-\frac{2\cdot p_2}{2\cdot p_2+d^*}}}\right\}$$

$$\geq \min\left\{\liminf_{n\to\infty} \inf_{\eta_n} \sup_{(\mathbf{X}_{I_1},Y)\in\mathcal{D}^{(I_1)}}\right.$$

$$\frac{\mathbf{E}_{\mathcal{D}_n^{(I_1)}} \left\{ \int |\eta_n(\mathbf{x}_{I_1}) - \eta(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2})|^2 \mathbf{P}_{\mathbf{X}_{I_1}}(d\mathbf{x}_{I_1}) \right\}}{n^{-\frac{2 \cdot p_1}{2 \cdot p_1 + 4}}},$$

$$\left. \liminf_{n \to \infty} \inf_{\eta_n} \sup_{(\mathbf{X}_{I_2}, Y) \in \mathcal{D}_2^{(I_2)}} \frac{\mathbf{E}_{\mathcal{D}_n^{(I_2)}} \left\{ \int |\eta_n(\mathbf{x}_{I_2}) - \eta(x_{\mathbf{i}_1}, \ldots, x_{\mathbf{i}_{d^*}})|^2 \mathbf{P}_{\mathbf{X}_{I_2}}(d\mathbf{x}_{I_2}) \right\}}{n^{-\frac{2 \cdot p_2}{2 \cdot p_2 + d^*}}} \right\},$$

where in the last two lines η_n is an estimator depending on a n-sized sample of the random vector $(\mathbf{X}_{I_1}, Y)$ and $(\mathbf{X}_{I_2}, Y)$, respectively. For $i \in \{1,2\}$ the subscript $\mathcal{D}_n^{(I_i)}$ in $\mathbf{E}_{\mathcal{D}_n^{(I_i)}}$ indicates that the expectation is taken with respect to a n-sized sample of the random vector $(\mathbf{X}_{I_i}, Y)$ instead of the n-sized sample $\mathcal{D}_n$ of $(\mathbf{X}, Y)$. The assertion follows by Lemma 12. □

E. Design of the network architecture and choice of hyperparameters

In this section, we first describe how, derived from our theory, we used the class of convolutional neural networks introduced in Section 2 in the simulation study in Section 4 and Section 5. We then explain which hyperparameters the other methods use and list the parameter sets for the adaptive choices of hyperparameters in Table 3.

The class of convolutional neural networks introduced in Section 2 depends on the parameters t, $\mathbf{L} = (L^{(1)}, L^{(2)})$, $\mathbf{k}^{(1)}$, $\mathbf{k}^{(2)}$ and $\mathbf{M}$. In Theorem 1, some of these parameters depend on the level l and the order d^* of the generalized hierarchical max-pooling model. Therefore, we adaptively choose (using the splitting of the sample technique as descripted in Section 4) these two parameters from the parameter sets shown in Table 3. As in our theoretical result the filter sizes M_r have the values $2^1, 2^2, \ldots, 2^l$ for $r \in \{1, \ldots, L^{(1)}\}$, where the filter sizes grow with increasing r. To simplify the architecture of our classifier, each value of the filter sizes is repeated L_n times. The number of layers in the convolutional part is then given by $L^{(1)} = L_n \cdot l$. Furthermore, as in our theoretical result, we choose $k^{(1)}$ channels in each layer in the convolutional part and $k^{(2)}$ neurons in each layer of the fully connected neural network part, i e., we have

$$\mathbf{k}^{(1)} = (k_1^{(1)}, \ldots, k_1^{(1)}) \quad \text{and} \quad \mathbf{k}^{(2)} = (k_1^{(2)}, \ldots, k_1^{(2)}).$$

The parameter sets from which we adaptively choose the parameters and the resulting network parameters (derived from our theoretical result) are shown in Table 3. Next, we describe the hyperparameters of the other methods, whose parameters are adaptively choosen from the parameter sets from Table 3. The connected standard feedforward neural network (abbr. *neural-s*) has L hidden layers and k neurons per layer. Our k_n-nearest neighbor classification estimate (abbreviated *neighbor*) has only the parameter k_n. For our random forest classifier (abbr. *rand-f*), we choose N_{leaves} as the maximum number of leaf nodes and N_{trees} as the number of trees in the forest. Both support vector machine approaches, *svm-p* and *svm-rbf*, have a parameter C that controls the importance of the regularization term and a parameter γ that represents the kernel

	choice of hyperparameters	
approach	*adaptively choosen parameters*	*resulting parameters*
neural-c	$l \in \{2, 3, 4\}$, $d^* \in \{1, 2\}$, $L_n \in \{1, 2, 3\}$ $k^{(1)} \in \{2, 4, 8\}$, $k^{(2)} \in \{5, 10\}$	$L^{(1)} = l \cdot L_n$, $L^{(2)} = L_n$ $M_{(r-1)\cdot L_n+1}, \dots, M_{r\cdot L_n} = 2^r$ for $r = 1, \dots, l$, $t = d^*$
neural-s	$L \in \{1, 2, \dots, 8\}$, $k \in \{10, 20, 50, 100, 200\}$	
neighbor	$k_n \in \{1, 2, 3\} \cup \{2, 4, 8, 12, 16, \dots, 4 \cdot \lfloor \frac{n_l}{4} \rfloor\}$	
rand-f	$N_{leaves} \in \{8, 16, 32\}$, $N_{trees} \in \{50, 100, 200\}$	
svm-p	$d \in \{1, 2, 3, 4\}$, $C \in \{10^{-2}, 10^{-1}, 1, 10\}$ $\gamma \in \{10^{-2}, 10^{-1}, 1, 10\}$	
svm-rbf	$C \in \{10^{-2}, 10^{-1}, 1, 10\}$, $\gamma \in \{10^{-2}, 10^{-1}, 1, 10\}$	

Table 3: Parameter sets for the choice of the hyperparameters.

coefficient. The polynomial kernel of the support vector machine (abbr. *svm-p*) has a degree of d.

References

Anthony, M., Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, Cambridge.

Bagirov, A. M., Clausen, C., Kohler, M. (2009). Estimation of a regression function by maxima of minima of linear functions. *IEEE Transactions on Information Theory*, 55, 833–845.

Bartlett, P. L., Harvey, N., Liaw, C., Mehrabian, A. (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20, 1–17.

Bauer, B., Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics*, 47, 2261–2285.

Györfi, L., Kohler, M., Krzyzak, A., Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York.

Kohler, M., Langer, S. (2021). On the rate of convergence of fully connected very deep neural network regression estimates. *Annals of Statistics*, 49, 2231–2249.

Yarotsky, D., Zhevnerchuk, A. (2019). The phase diagram of approximation rates for deep neural networks. arXiv: 1906.09477.