Statistical theory for image classification using deep convolutional neural networks with cross-entropy loss [1]

Michael Kohler and Sophie Langer[2]
Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: kohler@mathematik.tu-darmstadt.de, langer@mathematik.tu-darmstadt.de

November 26, 2020

Abstract
Convolutional neural networks learned by minimizing the cross-entropy loss are nowadays the standard for image classification. Till now, the statistical theory behind those networks is lacking. We analyze the rate of convergence of the misclassification risk of the estimates towards the optimal misclassification risk. Under suitable assumptions on the smoothness and structure of the aposteriori probability it is shown that these estimates achieve a rate of convergence which is independent of the dimension of the image. The study shed light on the good performance of CNNs learned by cross-entropy loss and partly explains their success in practical applications.

AMS classification: Primary 62G05; secondary 62G20.

Key words and phrases: Cross-entropy loss, curse of dimensionality, convolutional neural networks, image classification, rate of convergence.

1 Introduction

Deep convolutional neural networks (CNNs) have achieved remarkable success in various applications, especially in visual recognition tasks, see, e.g., LeCun, Bengio and Hinton (2015), Krizhevsky, Sutskever and Hinton (2012), Schmidhuber (2015), Rawat and Wang (2017). Recently it was shown in Kohler, Krzyżak and Walter (2020) that such networks applied to image classification learned by minimizing the squared loss achieve a dimension reduction provided suitable assumptions on the smoothness and structure of the aposteriori probability holds. In practice CNNs are often learned by minimizing the cross-entropy loss. The aim of this article is to show that these networks also achieve a dimension reduction in image classification.

[1]Running title: *Statistical theory for image classification*
[2]Corresponding author. Tel: +49-6151-16-23371

1.1 Image classification

To achieve this goal we study image classification which we formalize as follows: Let $d_1, d_2 \in \mathbb{N}$ and let $(\mathbf{X}, Y)$, $(\mathbf{X}_1, Y_1)$, $\dots$, $(\mathbf{X}_n, Y_n)$ be independent and identically distributed random variables with values in

$$[0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \times \{-1,1\}.$$

Here we use the notation

$$[0,1]^J = \{(a_j)_{j\in J} \,:\, a_j \in [0,1] \quad (j \in J)\}$$

for a nonempty and finite index set J, and we describe a (random) image from (random) class $Y \in \{-1,1\}$ by a (random) matrix $\mathbf{X}$ with d_1 columns and d_2 rows, which contains at position (i,j) the grey scale value of the pixel of the image at the corresponding position. Our aim is to predict Y given $\mathbf{X}$. Therefore we define a classifier as a function $f : [0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \to \mathbb{R}$ and predict the value $+1$ when $f(\mathbf{X}) \geq 0$ and -1 when $f(\mathbf{X}) < 0$. Let

$$\eta(\mathbf{x}) = \mathbf{P}\{Y = 1|\mathbf{X} = \mathbf{x}\} \quad (\mathbf{x} \in [0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}}) \tag{1}$$

be the so–called aposteriori probability. Our aim is to find a classifier which predicts the right class with high probability. The so-called *prediction error* of our classifier is measured by

$$\mathbf{P}(Yf(\mathbf{X}) \leq 0).$$

It is well-known, that *Bayes'* rule

$$f^*(\mathbf{x}) = \begin{cases} 1, & \text{if } \eta(\mathbf{x}) \geq \frac{1}{2} \\ -1, & \text{elsewhere} \end{cases}$$

minimizes the prediction error

$$\mathbf{P}(Yf(\mathbf{X}) \leq 0),$$

i.e.

$$\min_{f:[0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \to \{-1,1\}} \mathbf{P}\{f(\mathbf{X}) \neq Y\} = \mathbf{P}\{f^*(\mathbf{X}) \neq Y\}$$

holds (cf., e.g., Theorem 2.1 in Devroye, Györfi and Lugosi (1996)). Because we do not know the distribution of $(\mathbf{X}, Y)$, we cannot find f^*. Instead we estimate f^* by using the training data

$$\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}.$$

A popular approach is estimating f^* by the empirical risk minimization, i.e.,

$$f_n = \arg\min_{f\in\mathcal{C}_n} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f(\mathbf{X}_i) \neq Y_i\},$$

where $\mathcal{C}_n$ is a given class of classifiers. In practice f_n is not computational feasible, since minimizing the empirical risk with the 0/1 loss over $\mathcal{C}_n$ is NP hard (Bartlett, Jordan and McAuliffe (2006)). By replacing the number of misclassifications by a surrogate loss φ, one can overcome computational problems. Instead of a class of classifiers $\mathcal{C}_n$, we consider a class of real-valued functions $\mathcal{F}_n$. For a given loss φ we are searching for an estimate $\hat{f}_n \in \mathcal{F}_n$ such that the surrogate empirical risk

$$\frac{1}{n}\sum_{i=1}^{n}\varphi\left(Y_i\hat{f}_n(\mathbf{X}_i)\right)$$

is small. There are different loss functions to choose (see Friedman, Hastie and Tibishirani (2009) for an overview). A wide variety of classification methods are based on the idea to replace the 0/1 loss by some kind of convex surrogate loss. In particular, AdaBoost (Friedmann, Hastie and Tibshirani (2000)) employs the *exponential* loss $\exp(-z)$, while support vector machines often use the so–called *hinge* loss of the form $\max(1-z,0)$ (Vapnik (1998)) and logistic regression applies the *log* loss $\varphi(x) = \log(1+\exp(-x))$ (Friedman, Hastie and Tibishirani (2009)). In the context of CNNs and image classification it is a standard to use *cross-entropy* loss or *log* loss. That is why we use this loss function in the following. Cross-entropy loss is Fisher consistent, i.e.

$$f^* = \operatorname{sgn}\left(\arg\min_{\text{for all } f}\mathbf{E}(\varphi(Yf(\mathbf{X})))\right),$$

where

$$\operatorname{sgn}(z) = \begin{cases} 1 & \text{for } z \geq 0 \\ -1 & \text{for } z < 0, \end{cases}$$

which follows since

$$f_\varphi^* := \arg\min_{\text{for all } f}\mathbf{E}(\varphi(Yf(\mathbf{X}))) = \log\frac{\eta(\mathbf{x})}{1-\eta(\mathbf{x})}$$

(see Friedman, Hastie and Tibishirani (2009)). According to this, we define $\hat{C}_n(\mathbf{x}) = \operatorname{sgn}\hat{f}_n(\mathbf{x})$ as a classifier, where $\hat{f}_n$ minimizes the cross-entropy loss over the function space $\mathcal{F}_n$. As function space $\mathcal{F}_n$ we choose a class of CNNs, which is defined in Section 2. Our aim is to construct our classifier $\hat{C}_n$ such that its misclassification risk

$$\mathbf{P}\{\hat{C}_n(\mathbf{X}) \neq Y|\mathcal{D}_n\}$$

is as small as possible. To analyze the performance of the classifier we derive a bound on the expected difference of the misclassification risk of $\hat{C}_n$ and the optimal misclassification risk, i.e., we derive an upper bound on

$$\mathbf{E}\left\{\mathbf{P}\{\hat{C}_n(\mathbf{X}) \neq Y|\mathcal{D}_n\} - \min_{f:[0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}}\to\{-1,1\}}\mathbf{P}\{f(\mathbf{X}) \neq Y\}\right\}$$
$$= \mathbf{P}\{\hat{C}_n(\mathbf{X}) \neq Y\} - \mathbf{P}\{f^*(\mathbf{X}) \neq Y\}.$$

1.2 Rate of convergence

In order to derive nontrivial rate of convergence results on the difference between the misclassification risk of any estimate and the minimal possible value it is necessary to restrict the class of distributions (cf., Cover (1968) and Devroye (1982)). As in Kohler, Krzyżak and Walter (2020) we use for this the following assumptions on the structure and the smoothness of the aposteriori probability. For the formulation of the definition we need the definition of (p, C)-smoothness, which is the following:

Definition 1 *Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $0 < s \leq 1$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-smooth, if for every $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = q$ the partial derivative $\frac{\partial^q f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ exists and satisfies*

$$\left| \frac{\partial^q f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(\mathbf{x}) - \frac{\partial^q f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(\mathbf{z}) \right| \leq C \cdot \|\mathbf{x} - \mathbf{z}\|^s$$

for all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$.

For our next definition we also need the following notation: For $M \subseteq \mathbb{R}^d$ and $\mathbf{x} \in \mathbb{R}^d$ we define

$$\mathbf{x} + M = \{\mathbf{x} + \mathbf{z} \,:\, \mathbf{z} \in M\}.$$

For $I \subseteq \{1, \ldots, d_1\} \times \{1, \ldots, d_2\}$ and $\mathbf{x} = (x_i)_{i \in \{1, \ldots, d_1\} \times \{1, \ldots, d_2\}} \in [0, 1]^{\{1, \ldots, d_1\} \times \{1, \ldots, d_2\}}$ we set

$$\mathbf{x}_I = (x_i)_{i \in I}.$$

Definition 2 *Let $d_1, d_2 \in \mathbb{N}$ with $d_1, d_2 > 1$ and $m : [0, 1]^{\{1, \ldots, d_1\} \times \{1, \ldots, d_2\}} \to \mathbb{R}$.*
a) *We say that m satisfies a* **max-pooling model with index set**

$$I \subseteq \{0, \ldots, d_1 - 1\} \times \{0, \ldots, d_2 - 1\},$$

if there exist a function $f : [0, 1]^{(1,1)+I} \to \mathbb{R}$ such that

$$m(\mathbf{x}) = \max_{(i,j) \in \mathbb{Z}^2 \,:\, (i,j)+I \subseteq \{1, \ldots, d_1\} \times \{1, \ldots, d_2\}} f\left(\mathbf{x}_{(i,j)+I}\right) \quad (\mathbf{x} \in [0, 1]^{\{1, \ldots, d_1\} \times \{1, \ldots, d_2\}}).$$

b) *Let $I = \{0, \ldots, 2^l - 1\} \times \{0, \ldots, 2^l - 1\}$ for some $l \in \mathbb{N}$. We say that*

$$f : [0, 1]^{\{1, \ldots, 2^l\} \times \{1, \ldots, 2^l\}} \to \mathbb{R}$$

satisfies a **hierarchical model of level** *l, if there exist functions*

$$g_{k,s} : \mathbb{R}^4 \to [0, 1] \quad (k = 1, \ldots, l, s = 1, \ldots, 4^{l-k})$$

such that we have

$$f = f_{l,1}$$

for some $f_{k,s} : [0,1]^{\{1,\dots,2^k\}\times\{1,\dots,2^k\}} \to \mathbb{R}$ *recursively defined by*

$$\begin{aligned} f_{k,s}(\mathbf{x}) \quad = \quad & g_{k,s}\big(f_{k-1,4\cdot(s-1)+1}(\mathbf{x}_{\{1,\dots,2^{k-1}\}\times\{1,\dots,2^{k-1}\}}), \\ & f_{k-1,4\cdot(s-1)+2}(\mathbf{x}_{\{2^{k-1}+1,\dots,2^k\}\times\{1,\dots,2^{k-1}\}}), \\ & f_{k-1,4\cdot(s-1)+3}(\mathbf{x}_{\{1,\dots,2^{k-1}\}\times\{2^{k-1}+1,\dots,2^k\}}), \\ & f_{k-1,4\cdot s}(\mathbf{x}_{\{2^{k-1}+1,\dots,2^k\}\times\{2^{k-1}+1,\dots,2^k\}})\big) \\ & \qquad \left(\mathbf{x} \in [0,1]^{\{1,\dots,2^k\}\times\{1,\dots,2^k\}}\right) \end{aligned}$$

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

$$f_{1,s}(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}) = g_{1,s}(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}) \quad (x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2} \in [0,1])$$

for $s = 1, \dots, 4^{l-1}$.
c) *We say that* $m : [0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \to \mathbb{R}$ *satisfies a* **hierarchical max-pooling model of level** l *(where* $2^l \leq \min\{d_1, d_2\}$*), if* m *satisfies a max-pooling model with index set*

$$I = \left\{0, \dots, 2^l - 1\right\} \times \left\{0, \dots, 2^l - 1\right\}$$

and the function $f : [0,1]^{(1,1)+I} \to \mathbb{R}$ *in the definition of this max-pooling model satisfies a hierarchical model with level* l.
d) *We say that the hierarchical max-pooling model* $m : [0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \to \mathbb{R}$ *of level* l *is* (p, C)*–smooth if all functions* $g_{k,s}$ *in the definition of the functions* m *are* (p, C)*–smooth for some* $C > 0$.

This definition is motivated by the following observation: Human beings often decide, whether a given image contains some object, i.e. a car, or not by scanning subparts of the image and checking, whether the searched object is on this subpart. For each subpart the human estimates a probability that the searched object is on it. The probability that the whole image contains the object is then simply the maximum of the probabilities for each subpart of the image. This leads to the definition of a max-pooling model for the aposteriori probability.
Additionally, the probability that a subpart contains the searched object is composed by several decisions, if parts of the searched objects are identifiable. This motivates the hierarchical structure of our model.

1.3 Main result

The goal of this paper is to study CNNs minimized by cross-entropy loss from a statistical point of view. We show that (under suitable assumptions on the aposteriori probability) we can derive convergence results for CNN classifiers (with ReLU activation function) learned by cross-entropy loss, which are independent of the input dimension of the image. In particular, we show in Theorem 1 a) below that in case that the aposteriori probability satisfies a (p, C)–smooth hierarchical max-pooling model, the expected misclassification risk of the estimate converges toward the minimal possible value with rate

$$n^{-\min\left\{\frac{p}{4p+8}, \frac{1}{8}\right\}}$$

(up to some logarithmic factor). If in addition the conditional class probabilities of most data are sufficiently close to either 1 or zero, the rate can be further improved to

$$n^{-\min\left\{\frac{p}{2p+4}, \frac{1}{4}\right\}}$$

(see Theorem 1 b)). Since human beings are quite good in recognizing images, this assumption is not unusual in the context of image classification. In both cases, our classifier circumvents the curse of dimensionality in image classification. These results shed light on the good performance of CNN classifiers in visual recognition tasks and partly explain their success from a theoretical point of view.

1.4 Discussion of related results

CNNs trained with logistic loss have achieved remarkable success in various visual recognition tasks, cf., e.g., LeCun et al. (1998), LeCun, Bengio and Hinton (2015), Krizhevsky, Sutskever and Hinton (2012), He et al. (2016) and the literature cited therein.
But, as already mentioned in Rawat and Wang (2017) and Kohler, Krzyżak and Walter (2020) there is a lack of mathematical understanding. There are only a very few papers analyzing the performance of CNNs from a theoretical point of view. Oono and Suzuki (2019) (and the literature cited therein) showed that properly defined CNNs are able to mimic feedforward deep neural networks (DNNs) and therefore derive similar rate of convergence results. Unfortunately, those results do not demonstrate situations, where CNNs outperform simple feedfoward DNNs, which is the case in many practical applications, especially in image classification. Lin and Zhang (2019) derived generalization bounds for CNNs. In case of overparametrized CNNs, e.g. Du et al. (2019) could show that the gradient descent is able to find the global minimum of the empirical loss function. But, as shown in Kohler and Krzyżak (2019), overparametrized DNNs minimizing the empirical loss do not, in general, generalize well. Yarotsky (2018) obtained interesting approximation properties of deep CNNs, but, unfortunately, only in an abstract setting, where it is unclear how to apply those results. Kohler, Krzyżak and Walter (2020) analyzed CNNs in the context of image classification and showed that in case that the aposteriori probability satisfies a generalized hierarchical max-pooling model with smoothness constraint p_1 and p_2 (see Definition 1 in Kohler, Krzyżak and Walter (2020)), suitable defined CNNs achieve a rate of convergence which does not depend on the input dimension of the image. In this result, the CNNs are learned by the squared loss. As, e.g. experimental results in Golik, Doetsch and Ney (2013) show, DNNs learned by cross entropy loss allow to find a better local optimum than the squared loss criterion. Thus the CNNs learned by cross entropy loss are of higher practical relevance.
Cross entropy loss or, more general, convex surrogate loss functions have been studied in

Bartlett, Jordan and McAuliffe (2006) and Zhang (2004). Bartlett, Jordan and McAuliffe (2006) showed that for convex loss functions satisfying a certain uniform strict convexity condition the rate of convergence can be strictly faster than the classical $n^{-1/2}$, depending on the strictness of convexity of ϕ and the complexity of class of classifiers. Zhang (2004) analyzed how close the optimal Bayes error rate can be approximately reached using a classification algorithm that computes a classifier by minimizing a convex upper bound of the classification error function. Some results of this article (see Lemma 1) are also used in our analysis. In Lemma 1 b) we derive a modification of Zhang's bound which enables us to derive better rate of convergence under proper assumptions on the aposteriori probability.
Much more theoretical results are known for simple feedforward DNNs. Under suitable compository assumptions on the structure of the regression function, those networks are able to circumvent the curse of dimensionality (cf., Kohler and Krzyżak (2017), Bauer and Kohler (2019), Schmidt-Hieber (2020), Kohler and Langer (2020), Suzuki and Nitanda (2019) and Langer (2020)). Imaizumi und Fukumizu (2019) derived results concerning estimation by neural networks of piecewise polynomial regression functions with partitions having rather general smooth boundaries. Eckle and Schmidt-Hieber (2019) and Kohler, Krzyżak and Langer (2019) analyzed regression functions which have the form of a common statistical model, i.e., which have the form of multivariate adaptive regression splines (MARS), and showed that in this case the convergence rate by DNNs can also be improved. Kim, Ohn and Kim (2019) analyzed classification problems with standard feedforward DNNs and derived fast rate of convergence for DNNs learned by cross-entropy under the condition that the conditional class probabilities of most data are sufficiently close to either 1 or zero. The condition formulated in this article is also used in the analysis of our CNNs.

1.5 Notation

Throughout the paper, the following notation is used: The sets of natural numbers, natural numbers including 0, integers and real numbers are denoted by $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{Z}$ and $\mathbb{R}$, respectively. For $z \in \mathbb{R}$, we denote the smallest integer greater than or equal to z by $\lceil z \rceil$. Let $D \subseteq \mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be a real-valued function defined on $\mathbb{R}^d$. We write $\mathbf{x} = \arg\min_{\mathbf{z} \in D} f(\mathbf{z})$ if $\min_{\mathbf{z} \in \mathcal{D}} f(\mathbf{z})$ exists and if $\mathbf{x}$ satisfies $\mathbf{x} \in D$ and $f(\mathbf{x}) = \min_{\mathbf{z} \in \mathcal{D}} f(\mathbf{z})$. For $f : \mathbb{R}^d \to \mathbb{R}$

$$\|f\|_\infty = \sup_{\mathbf{x} \in \mathbb{R}^d} |f(\mathbf{x})|$$

is its supremum norm, and the supremum norm of f on a set $A \subseteq \mathbb{R}^d$ is denoted by

$$\|f\|_{A,\infty} = \sup_{\mathbf{x} \in A} |f(\mathbf{x})|.$$

Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$, let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ and set $\mathbf{x}_1^n = (\mathbf{x}_1, \dots, \mathbf{x}_n)$. A finite collection $f_1, \dots, f_N : \mathbb{R}^d \to \mathbb{R}$ is called an $\varepsilon-$ cover of $\mathcal{F}$ on $\mathbf{x}_1^n$ if for any $f \in \mathcal{F}$

there exists $i \in \{1, \dots, N\}$ such that

$$\frac{1}{n} \sum_{k=1}^{n} |f(\mathbf{x}_k) - f_i(\mathbf{x}_k)| < \varepsilon.$$

The ε–covering number of $\mathcal{F}$ on $\mathbf{x}_1^n$ is the size N of the smallest ε–cover of $\mathcal{F}$ on $\mathbf{x}_1^n$ and is denoted by $\mathcal{N}_1(\varepsilon, \mathcal{F}, \mathbf{x}_1^n)$.

For $z \in \mathbb{R}$ and $\beta > 0$ we define $T_\beta z = \max\{-\beta, \min\{\beta, z\}\}$. Throughout the remainder of this paper

$$\varphi(z) = \log(1 + \exp(-z))$$

denotes the cross entropy or logistic loss.

1.6 Outline of the paper

In Section 2 the CNN image classifiers used in this paper are defined. The main result is presented in Section 3 and proven in Section 4.

2 Definition of the estimates

The architecture of CNN is inspired by the natural visual perception mechansim of the humans. The first modern framework was published by LeCun et al. (1989), called LeNet-5, which could classify handwritten digits. Even though today exist numerous variants of CNN, the basic components are still the same, namely convolutional, pooling and fully-connected layers. The convolutional layers aims to learn feature representations of the inputs.

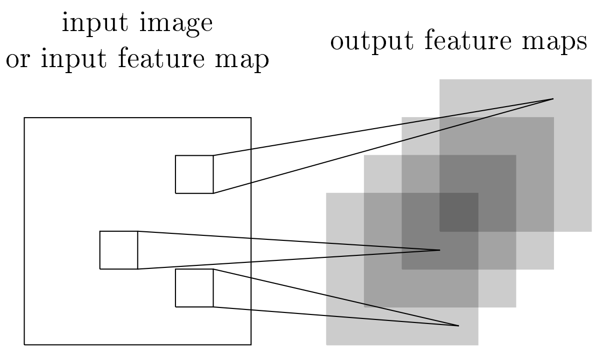

Figure 1: Illustration of a convolutional layer

As shown in Figure 1 each convolutional layer l ($l \in \{1, \dots, L\}$) consists of $k_l \in \mathbb{N}$ channels (also called feature maps) and the convolution in layer l is performed by using a window of values of layer $l-1$ of size $M_l \in \{1, \dots, \min\{d_1, d_2\}\}$. Specifically, each neuron of a channel is connected to a region of neighboring neurons in the previous layer. A new channel can be obtained by first convolving the input with a weight matrix (so–called filter) and then applying an element-wise nonlinear activation function $\sigma : \mathbb{R} \to \mathbb{R}$ on the convolved result. As activation function we choose the ReLU function $\sigma(x) = \max\{x, 0\}$ in the following. The weight matrix is defined by

$$\mathbf{w} = \left(w^{(l)}_{i,j,s_1,s_2} \right)_{1 \le i,j \le M_l, s_1 \in \{1,\dots,k_{l-1}\}, s_2 \in \{1,\dots,k_l\}, l \in \{1,\dots,L\}}.$$

Furthermore we need some weights

$$\mathbf{w}_{bias} = (w^{(l)}_{s_2})_{s_2 \in \{1,\dots,k_r\}, l \in \{1,\dots,L\}}$$

for the bias of the channels and some output weights

$$\mathbf{w}_{out} = (w_s)_{s \in \{1,\dots,k_L\}}.$$

Mathematically, the channel value at location (i,j) in the k_l-th channel of layer l is calculated by:

$$o^{(l)}_{(i,j),s_2} = \sigma \left(\sum_{s_1=1}^{k_l-1} \sum_{\substack{t_1,t_2 \in \{1,\dots,M_l\} \\ (i+t_1-1,j+t_2-1) \in D}} w^{(l)}_{t_1,t_2,s_1,s_2} o^{(l-1)}_{(i+t_1-1,j+t_2-1),s_1} + w^{(l)}_{s_2} \right), \tag{2}$$

where $D = \{1, \dots, d_1\} \times \{1, \dots, d_2\}$ and

$$o^{(0)}_{(i,j),1} = x_{i,j} \quad \text{for } i \in \{1,\dots,d_1\} \text{ and } j \in \{1,\dots,d_2\}.$$

Here one may see that weights generating the feature map $o^{(l)}_{(:,:),s_2}$ are shared, which has the advantage that it can reduce the model complexity and the duration of the networks' training. In our network, only in the last step a max-pooling layer is applied to the values of the last convolutional layer L. Thus, the output of the network is given by a real-valued function on $[0,1]^{\{1,\dots,d_1\} \times \{1,\dots,d_2\}}$ of the form

$$f_{\mathbf{w},\mathbf{w}_{bias},\mathbf{w}_{out}}(\mathbf{x}) = \max \left\{ \sum_{s_2=1}^{k_L} w_{s_2} \cdot o^{(L)}_{(i,j),s_2} : i \in \{1,\dots,d_1 - M_L + 1\} \right.$$
$$\left. , j \in \{1,\dots,d_2 - M_L + 1\} \right\}.$$

Our class of convolutional neural networks with parameters L, $\mathbf{k} = (k_1,\dots,k_L)$ and $\mathbf{M} = (M_1,\dots,M_L)$ is defined by $\mathcal{F}^{CNN}_{L,\mathbf{k},\mathbf{M}}$. As in Kohler, Krzyżak and Walter (2020) we use a so-called zero padding in the definition of the index set D in (2). Thus, the size of a channel is the same as in the previous layer (see Kohler, Krzyżak and Walter (2020) for a further illustration). Our final estimate is a composition of a convolutional neural network out of the class $\mathcal{F}^{CNN}_{L,\mathbf{k},\mathbf{M}}$ and a fully-connected neural network, which is defined as follows: The output of this network is produced by a function $g : \mathbb{R} \to \mathbb{R}$ of the form

$$g(x) = \sum_{i=1}^{k_L} w^{(L)}_i g^{(L)}_i(x) + w^{(L)}_0, \tag{3}$$

where $w_0^{(L)}, \ldots, w_{k_L}^{(L)} \in \mathbb{R}$ denote the output weights and for $i \in \{1, \ldots, k_L\}$ the $g_i^{(L)}$ are recursively defined by

$$g_i^{(r)}(x) = \sigma\left(\sum_{j=1}^{k_{r-1}} w_{i,j}^{(r-1)} g_j^{(r-1)}(x) + w_{i,0}^{(r-1)}\right)$$

for $w_{i,0}^{(r-1)}, \ldots, w_{i,k_{r-1}}^{(r-1)} \in \mathbb{R}$, $i \in \{1, \ldots, k_r\}$, $r \in \{1, \ldots, L\}$, $k_0 = 1$ and

$$g_1^{(0)}(x) = x.$$

Here the function $\sigma : \mathbb{R} \to \mathbb{R}$ denotes again the ReLU activation function

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

We define the function class of all real-valued functions on $\mathbb{R}$ of the form (3) with parameters L and $\mathbf{k} = (k_1, \ldots, k_L)$ by $\mathcal{F}_{L,\mathbf{k}}^{FNN}$.

Our final function class $\mathcal{F}_n$ is then of the form

$$\mathcal{F}_n = \left\{ g \circ f : f \in \mathcal{F}_{L_n^{(1)}, \mathbf{k}^{(1)}, \mathbf{M}}^{CNN}, g \in \mathcal{F}_{L_n^{(2)}, \mathbf{k}^{(2)}}^{FNN}, \|g \circ f\|_\infty \le \beta_n \right\},$$

which depends on the parameters

$$\mathbf{L} = (L_n^{(1)}, L_n^{(2)}),\ \mathbf{k}^{(1)} = \left(k_1^{(1)}, \ldots, k_{L_1}^{(1)}\right),\ \mathbf{k}^{(2)} = \left(k_1^{(2)}, \ldots, k_{L_2}^{(2)}\right),\ \mathbf{M} = (M_1, \ldots, M_{L_n^{(2)}})$$

and $\beta_n = c_1 \cdot \log n$. Let

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-Y_i \cdot f(\mathbf{X}_i))$$

be the CNN, which minimizes cross entropy loss on the trainings data $\mathcal{D}_n$. We define our CNN image classifier by

$$\hat{C}_n(\mathbf{x}) = \operatorname{sgn}(\hat{f}_n(\mathbf{x})).$$

3 Main results

Our main result is the following theorem, which presents two upper bounds on the distance between the expected misclassification risk of our CNN classifier and the optimal misclassification risk.

Theorem 1 *Let $p \ge 1$ and $C > 0$ be arbitrary. Assume that the aposteriori probability $\eta(\mathbf{x}) = \mathbf{P}\{Y = 1|\mathbf{X} = \mathbf{x}\}$ satisfies a (p, C)–smooth hierarchical max-pooling model of finite level l and supp$(\mathbf{P_X}) \subseteq [0,1]^{d_1 \times d_2}$. Set*

$$L_n^{(1)} = \frac{4^l - 1}{3} \cdot \lceil c_3 \cdot n^{2/(2p+4)} \rceil + l \quad \textit{and} \quad L_n^{(2)} = \lceil c_2 \cdot n^{1/4} \rceil,$$

$$M_s = 2^{\pi(s)} \quad (s = 1, \dots, L_n^{(1)}),$$

where the function $\pi : \{1, \dots, L_n^{(1)}\} \to \{1, \dots, l\}$ *is defined by*

$$\pi(s) = \sum_{i=1}^{l} \mathbb{1}_{\{s \geq i + \sum_{r=l-i+1}^{l-1} 4^r \cdot \lceil c_3 \cdot n^{2/(2p+4)} \rceil\}},$$

choose $\mathbf{k}^{(1)} = (c_4, \dots, c_4) \in \mathbb{N}^{L_n^{(1)}}$ *and* $\mathbf{k}^{(2)} = (c_5, \dots, c_5) \in \mathbb{N}^{L_n^{(2)}}$*, and define the estimate* $\hat{C}_n$ *as in Section 2. Assume that the constants* $c_2, \dots, c_5$ *are sufficiently large.*
a) *There exists a constant* $c_6 > 0$ *such that we have for any* $n > 1$

$$\mathbf{P}\left\{Y \neq \hat{C}_n(\mathbf{X})\right\} - \mathbf{P}\{Y \neq f^*(\mathbf{X})\} \leq c_6 \cdot (\log n) \cdot n^{-\min\{\frac{p}{4p+8}, \frac{1}{8}\}}.$$

b) *If, in addition,*

$$\mathbf{P}\left\{X : |f_\varphi^*(\mathbf{X})| > \frac{1}{2} \cdot \log n\right\} \geq 1 - \frac{1}{\sqrt{n}} \quad (n \in \mathbb{N}) \tag{4}$$

holds, then there exists a constant $c_7 > 0$ *such that we have for any* $n > 1$

$$\mathbf{P}\left\{Y \neq \hat{C}_n(\mathbf{X})\right\} - \mathbf{P}\{Y \neq f^*(\mathbf{X})\} \leq c_7 \cdot (\log n)^2 \cdot n^{-\min\{\frac{p}{2p+4}, \frac{1}{4}\}}.$$

Remark 1. The rate of convergence in Theorem 1 does not depend on the dimension $d_1 \cdot d_2$ of the predictor variable, hence under the assumptions on the structure of the aposteriori probabilities in Theorem 1 our convolutional neural network classifier is able to circumvent the curse of dimensionality.
Remark 2. Assumption (4) requires that with high probability the aposteriori probability is very close to zero or very close to one, and hence the optimal classification rule makes only a very small error. This is in particular realistic for many applications in image classification, where often there is not much doubt about the class of objects (cf., Kim, Ohn and Kim (2019)).
Remark 3. The definition of the parameter $L_n^{(2)}$ of the estimate in Theorem 1 depends on the smoothness and the level of the hierarchical max-pooling model for the aposteriori probability, which are usually unknown in applications. In this case it is possible to define this parameter in a data-dependent way, e.g., by using a splitting of the sample approach (cf., e.g., Chapter 7 in Györif et al. (2002)).

4 Proofs

Lemma 1 *Let* φ *be the logistic loss. Let* $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ *and* $\mathcal{D}_n, \hat{f}_n$*,* $\hat{C}_n$ *and* f_φ^* *as in Section 1.*
a) *Then*

$$\mathbf{P}\left\{Y \neq \hat{C}_n(\mathbf{X}) | \mathcal{D}_n\right\} - \mathbf{P}\{Y \neq f^*(\mathbf{X})\}$$

$$\leq \frac{1}{\sqrt{2}} \cdot \left(\mathbf{E}\left\{ \varphi(Y \cdot \hat{C}_n(\mathbf{X})) | \mathcal{D}_n \right\} - \mathbf{E}\left\{ \varphi(Y \cdot f_\varphi^*(\mathbf{X})) \right\} \right)^{1/2}$$

holds.
b) *Then*

$$\mathbf{P}\left\{ Y \neq \hat{C}_n(\mathbf{X}) | \mathcal{D}_n \right\} - \mathbf{P}\left\{ Y \neq f^*(\mathbf{X}) \right\}$$
$$\leq 2 \cdot \left(\mathbf{E}\left\{ \varphi(Y \cdot \hat{C}_n(\mathbf{X})) | \mathcal{D}_n \right\} - \mathbf{E}\left\{ \varphi(Y \cdot f_\varphi^*(\mathbf{X})) \right\} \right) + 4 \cdot \mathbf{E}\left\{ \varphi(Y \cdot f_\varphi^*(\mathbf{X})) \right\}.$$

holds.
c) *Assume that*

$$\mathbf{P}\left\{ \mathbf{X} : |f_\varphi^*(\mathbf{X})| > \tilde{F}_n \right\} \geq 1 - e^{-\tilde{F}_n}$$

for a given sequence $\{\tilde{F}_n\}_{n \in \mathbb{N}}$ *with* $F_n \to \infty$. *Then*

$$\mathbf{E}\left\{ \varphi(Y \cdot f_\varphi^*(\mathbf{X})) \right\} \leq c_8 \cdot \tilde{F}_n \cdot e^{-\tilde{F}_n}$$

holds.

Proof. a) This result follows from Theorem 2.1 in Zhang (2004), where we choose $s = 2$ and $c = 2^{-1/2}$.
b) Set $\bar{f}_n(\mathbf{x}) = 1/(1 + \exp(-\hat{f}_n(\mathbf{x}))$. Then we have

$$\mathbf{P}\left\{ Y \neq \mathrm{sgn}(\hat{f}_n(\mathbf{X})) | \mathcal{D}_n \right\} - \mathbf{P}\left\{ Y \neq f^*(\mathbf{X}) \right\}$$
$$\leq \int \Big((1 - \eta(\mathbf{x})) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) \geq 0\}}(\mathbf{x}) + \eta(\mathbf{x}) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) < 0\}}(\mathbf{x})$$
$$- (1 - \eta(\mathbf{x})) \cdot \mathbb{1}_{\{\eta(\mathbf{x}) \geq \frac{1}{2}\}}(\mathbf{x}) + \eta(\mathbf{x}) \cdot \mathbb{1}_{\{\eta(\mathbf{x}) < \frac{1}{2}\}}(\mathbf{x}) \Big) \mathbf{P}_{\mathbf{X}}(d\mathbf{x})$$
$$= \int \Big((1 - \eta(\mathbf{x})) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) \geq 0\}}(\mathbf{x}) + \eta(\mathbf{x}) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) < 0\}}(\mathbf{x})$$
$$- (1 - \bar{f}_n(\mathbf{x})) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) \geq 0\}}(\mathbf{x}) - \bar{f}_n(\mathbf{x}) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) < 0\}}(\mathbf{x})$$
$$+ (1 - \bar{f}_n(\mathbf{x})) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) \geq 0\}}(\mathbf{x}) + \bar{f}_n(\mathbf{x}) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) < 0\}}(\mathbf{x})$$
$$- (1 - \bar{f}_n(\mathbf{x})) \cdot \mathbb{1}_{\{\eta(\mathbf{x}) \geq \frac{1}{2}\}}(\mathbf{x}) - \bar{f}_n(\mathbf{x}) \cdot \mathbb{1}_{\{\eta(x) < \frac{1}{2}\}}(\mathbf{x})$$
$$+ (1 - \bar{f}_n(\mathbf{x})) \cdot \mathbb{1}_{\{\eta(\mathbf{x}) \geq \frac{1}{2}\}}(\mathbf{x}) + \bar{f}_n(\mathbf{x}) \cdot \mathbb{1}_{\{\eta(\mathbf{x}) < \frac{1}{2}\}}(\mathbf{x})$$
$$- (1 - \eta(\mathbf{x})) \cdot \mathbb{1}_{\{\eta(\mathbf{x}) \geq \frac{1}{2}\}}(\mathbf{x}) - \eta(x) \cdot \mathbb{1}_{\{\eta(\mathbf{x}) < \frac{1}{2}\}}(x) \Big) \mathbf{P}_{\mathbf{X}}(d\mathbf{x})$$
$$= \int \big((1 - \eta(\mathbf{x}) - 1 + \bar{f}_n(\mathbf{x})) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) \geq 0\}}(\mathbf{x}) + (\eta(\mathbf{x}) - \bar{f}_n(\mathbf{x})) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) < 0\}}(\mathbf{x})$$
$$+ (1 - \bar{f}_n(\mathbf{x}) - 1 + \eta(\mathbf{x})) \cdot \mathbb{1}_{\{\eta(\mathbf{x}) \geq \frac{1}{2}\}}(\mathbf{x}) + (\bar{f}_n(\mathbf{x}) - \eta(\mathbf{x})) \cdot \mathbb{1}_{\{\eta(\mathbf{x}) < \frac{1}{2}\}}(\mathbf{x})$$
$$+ (1 - \bar{f}_n(\mathbf{x})) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) \geq 0\}}(\mathbf{x}) + \bar{f}_n(\mathbf{x}) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x}) < 0\}}(\mathbf{x})$$
$$- (1 - \bar{f}_n(\mathbf{x})) \cdot \mathbb{1}_{\{\eta(\mathbf{x}) \geq \frac{1}{2}\}}(\mathbf{x}) - \bar{f}_n(\mathbf{x}) \cdot \mathbb{1}_{\{\eta(\mathbf{x}) < \frac{1}{2}\}}(\mathbf{x}) \big) \mathbf{P}_{\mathbf{X}}(d\mathbf{x})$$

$$\leq 2 \cdot \mathbf{E}\{|\bar{f}_n(\mathbf{X}) - \eta(\mathbf{X})||\mathcal{D}_n\}. \tag{5}$$

Here the last inequality follows since

$$\hat{f}_n(\mathbf{x}) \geq 0 \text{ implies } \bar{f}_n(\mathbf{x}) \geq \frac{1}{2} \text{ and}$$
$$\hat{f}_n(\mathbf{x}) < 0 \text{ implies } \bar{f}_n(\mathbf{x}) < \frac{1}{2}$$

and consequently we have

$$\begin{aligned} &(1-\bar{f}_n(\mathbf{x})) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x})\geq 0\}}(\mathbf{x}) + \bar{f}_n(\mathbf{x}) \cdot \mathbb{1}_{\{\hat{f}_n(\mathbf{x})<0\}}(\mathbf{x}) \\ =& \min\{1-\bar{f}_n(\mathbf{x}), \bar{f}_n(\mathbf{x})\} \\ \leq& (1-\bar{f}_n(\mathbf{x})) \cdot \mathbb{1}_{\{\eta(\mathbf{x})\geq \frac{1}{2}\}}(\mathbf{x}) + \bar{f}_n(\mathbf{x}) \cdot \mathbb{1}_{\{\eta(\mathbf{x})<\frac{1}{2}\}}(\mathbf{x}). \end{aligned}$$

Furthermore we can bound (5) by

$$2 \cdot \mathbf{E}\left\{\left|\varphi\left(Y \cdot \log \frac{\bar{f}_n(\mathbf{X})}{1-\bar{f}_n(\mathbf{X})}\right) - \varphi\left(Y \cdot \log \frac{\eta(\mathbf{X})}{1-\eta(\mathbf{X})}\right)\right| \middle| \mathcal{D}_n\right\}. \tag{6}$$

Here we used that for $g(z) = \log \frac{z}{1-z}$ with $z \in (0,1)$,

$$\begin{aligned} h_1(z) = \varphi(1 \cdot g(z)) &= \log\left(1 + \exp\left(-\log \frac{z}{1-z}\right)\right) \\ &= \log\left(1 + \frac{1-z}{z}\right) \\ &= \log\left(\frac{1}{z}\right) = -\log(z) \end{aligned}$$

and

$$\begin{aligned} h_2(z) = \varphi(-1 \cdot g(z)) &= \log\left(1 + \frac{z}{1-z}\right) \\ &= \log\left(\frac{1}{1-z}\right) \\ &= -\log(1-z), \end{aligned}$$

we have

$$h_1'(z) = -\frac{1}{z} \text{ and } h_2'(z) = \frac{1}{1-z}$$

and consequently

$$|h_1'(z)| = \frac{1}{|z|} \geq 1 \text{ and } |h_2'(z)| = \frac{1}{|1-z|} \geq 1 \text{ for } z \in (0,1).$$

Using mean value theorem it follows for $j \in \{1, 2\}$ and $z_1, z_2 \in [0, 1]$

$$|h_j(z_1) - h_j(z_2)| \geq 1 \cdot |z_1 - z_2|.$$

Since

$$|a - b| \leq |a| + |b| = a + b \quad \text{for } a, b \geq 0$$

we can finally bound (6) by

$$\begin{aligned}
&2 \cdot \mathbf{E}\left\{ \left| \varphi\left(Y \cdot \log \frac{\bar{f}_n(\mathbf{X})}{1 - \bar{f}_n(\mathbf{X})}\right) - \varphi\left(Y \cdot \log \frac{\eta(\mathbf{X})}{1 - \eta(\mathbf{X})}\right) \right| \middle| \mathcal{D}_n \right\} \\
&\leq 2 \cdot \mathbf{E}\left\{ \varphi\left(Y \cdot \hat{f}_n(\mathbf{X})\right) + \varphi\left(Y \cdot f_\varphi^*(\mathbf{X})\right) \middle| \mathcal{D}_n \right\} \\
&= 2 \cdot \mathbf{E}\left\{ \varphi\left(Y \cdot \hat{f}_n(\mathbf{X})\right) - \varphi\left(Y \cdot f_\varphi^*(\mathbf{X})\right) \middle| \mathcal{D}_n \right\} + 4 \cdot \mathbf{E}\left\{ \varphi\left(Y \cdot f_\varphi^*(\mathbf{X})\right) | \mathcal{D}_n \right\}.
\end{aligned}$$

c) This result follows from Lemma 3 in Kim, Ohn and Kim (2019). □

Lemma 2 *Let φ be the logistic loss. Define $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ and $\mathcal{D}_n, \hat{f}_n$ and f_φ^* as in Section 1. Let $\mathcal{F}_n$ be a function space consisting of functions $f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$. Then*

$$\begin{aligned}
&\mathbf{E}\left\{ \varphi(Y \cdot \hat{f}_n(\mathbf{X})) | \mathcal{D}_n \right\} - \mathbf{E}\left\{ \varphi(Y \cdot f_\varphi^*(\mathbf{X})) \right\} \\
&\leq 2 \cdot \sup_{f \in \mathcal{F}_n} \left| \mathbf{E}\left\{ \varphi(Y \cdot f(\mathbf{X})) \right\} - \frac{1}{n} \sum_{i=1}^n \varphi(Y_i \cdot f(\mathbf{X}_i)) \right| \\
&\quad + \inf_{f \in \mathcal{F}_n} \mathbf{E}\left\{ \varphi(Y \cdot f(\mathbf{X})) \right\} - \mathbf{E}\left\{ \varphi(Y \cdot f_\varphi^*(\mathbf{X})) \right\}.
\end{aligned}$$

Proof. This result is the standard error bound for empirical risk minimization. For the sake of completeness we present nevertheless a complete proof.

Let $f \in \mathcal{F}_n$ be arbitrary. Then the definition of $\hat{f}_n$ implies

$$\begin{aligned}
&\mathbf{E}\left\{ \varphi(Y \cdot \hat{f}_n(\mathbf{X})) | \mathcal{D}_n \right\} - \mathbf{E}\left\{ \varphi(Y \cdot f_\varphi^*(\mathbf{X})) \right\} \\
&\leq \mathbf{E}\left\{ \varphi(Y \cdot \hat{f}_n(\mathbf{X})) | \mathcal{D}_n \right\} - \frac{1}{n} \sum_{i=1}^n \varphi(Y_i \cdot \hat{f}_n(\mathbf{X}_i)) \\
&\quad + \frac{1}{n} \sum_{i=1}^n \varphi(Y_i \cdot \hat{f}_n(\mathbf{X}_i)) - \frac{1}{n} \sum_{i=1}^n \varphi(Y_i \cdot f(\mathbf{X}_i)) \\
&\quad + \frac{1}{n} \sum_{i=1}^n \varphi(Y_i \cdot f(\mathbf{X}_i)) - \mathbf{E}\left\{ \varphi(Y \cdot f(\mathbf{X})) \right\} \\
&\quad + \mathbf{E}\left\{ \varphi(Y \cdot f(\mathbf{X})) \right\} - \mathbf{E}\left\{ \varphi(Y \cdot f_\varphi^*(\mathbf{X})) \right\} \\
&\leq \mathbf{E}\left\{ \varphi(Y \cdot \hat{f}_n(\mathbf{X})) | \mathcal{D}_n \right\} - \frac{1}{n} \sum_{i=1}^n \varphi(Y_i \cdot \hat{f}_n(\mathbf{X}_i))
\end{aligned}$$

$$
\begin{aligned}
&+0 + \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i \cdot f(\mathbf{X}_i)) - \mathbf{E}\left\{\varphi(Y \cdot f(\mathbf{X}))\right\} \\
&+\mathbf{E}\left\{\varphi(Y \cdot f(\mathbf{X}))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(\mathbf{X}))\right\} \\
\leq\ & 2 \cdot \sup_{g \in \mathcal{F}_n} \left| \mathbf{E}\left\{\varphi(Y \cdot g(\mathbf{X}))\right\} - \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i \cdot g(\mathbf{X}_i)) \right| \\
&+\mathbf{E}\left\{\varphi(Y \cdot f(\mathbf{X}))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(\mathbf{X}))\right\}.
\end{aligned}
$$

□

Lemma 3 *Let φ be the logistic loss and $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be independent and identically distributed $\mathbb{R}^{d_1 \times d_2} \times \mathbb{R}$-valued random variables. Let $\mathcal{F}_n$ be defined as in Section 2. Then*

$$
\mathbf{E}\left\{ \sup_{f \in \mathcal{F}_n} \left| \mathbf{E}\left\{\varphi(Y \cdot f(\mathbf{X}))\right\} - \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i \cdot f(\mathbf{X}_i)) \right| \right\} \leq c_9 \cdot (\log n)^2 \cdot \frac{\max\{L_n^{(1)}, L_n^{(2)}\}}{\sqrt{n}}.
$$

In order to prove Lemma 3 we need the following bound on the covering number of $\mathcal{F}_n$.

Lemma 4 *Let $\sigma(x) = \max\{x, 0\}$ be the ReLU activation function, define $\mathcal{F}_n$ as in Section 2 and set*

$$
k_{max} = \max\left\{k_1^{(1)}, \ldots, k_{L_n^{(1)}}^{(1)}, k_1^{(2)}, \ldots, k_{L_n^{(2)}}^{(2)}\right\}, \quad M_{max} = \max\{M_1, \ldots, M_{L_n^{(2)}}\}
$$

and

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

Assume $d_1 \cdot d_2 > 1$ and $\beta_n = c_1 \cdot \log n \geq 2$. Then we have for any $\epsilon \in (0, 1)$:

$$
\begin{aligned}
&\sup_{\mathbf{x}_1^n \in (\mathbb{R}^{\{1,\ldots,d_1\} \times \{1,\ldots,d_2\}})^n} \log\left(\mathcal{N}_1\left(\epsilon, \mathcal{F}_n, \mathbf{x}_1^n\right)\right) \\
&\leq c_{10} \cdot L_{max}^2 \cdot \log(L_{max} \cdot d_1 \cdot d_2) \cdot \log\left(\frac{c_1 \cdot \log n}{\epsilon}\right)
\end{aligned}
$$

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

Proof. See Lemma 7 in Kohler, Krzyżak and Walter (2020). □

Proof of Lemma 3. Since $f \in \mathcal{F}_n$ satisfies $\|f\|_\infty \leq \beta_n = c_1 \cdot \log n$, we have for any $\mathbf{x} \in \mathbb{R}^{d_1 \times d_2}$ and $y \in \{-1, 1\}$

$$
\varphi(y \cdot f(\mathbf{x})) = \log(1 + e^{-y \cdot f(\mathbf{x})}) \leq \log(1 + e^{|f(\mathbf{x})|}) \leq \log(1 + e^{\beta_n}) = c_{11} \cdot \log n.
$$

Set

$$
\mathbf{Z} = (\mathbf{X}, Y), \mathbf{Z}_1 = (\mathbf{X}_1, Y_1), \ldots, \mathbf{Z}_n = (\mathbf{X}_n, Y_n),
$$

and

$$\mathcal{H}_n = \{h : \mathbb{R}^{d_1 \times d_2} \times \mathbb{R} \to \mathbb{R} : \exists f \in \mathcal{F}_n \text{ such that } h(\mathbf{x}, y) = \varphi(y \cdot f(\mathbf{x}))\}.$$

By Theorem 9.1 in Györfi et al. (2002) one has, for arbitrary $\epsilon > 0$

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{f \in \mathcal{F}_n} \left| \mathbf{E}\{\varphi(Y \cdot f(\mathbf{X}))\} - \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i \cdot f(\mathbf{X}_i)) \right| > \epsilon \right\} \\
&= \mathbf{P}\left\{\sup_{f \in \mathcal{H}_n} \left| \mathbf{E}h(\mathbf{Z}) - \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{Z}_i) \right| > \epsilon \right\} \\
&\leq 8\mathbf{E}\left\{\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{H}_n, \mathbf{Z}_1^n\right)\right\} e^{-\frac{n\epsilon^2}{128 \cdot c_{11}^2 \cdot (\log n)^2}}.
\end{aligned}$$

Let $h_i(\mathbf{x}, y) = \varphi(y \cdot f_i(\mathbf{x}))$ $((\mathbf{x}, y) \in \mathbb{R}^{d_1 \times d_2} \times \{-1, 1\})$ for some $f_i : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$. Then it follows with the Lipschitz continuity of φ, that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} |h_1(\mathbf{Z}_i) - h_2(\mathbf{Z}_i)| \\
&= \frac{1}{n}\sum_{i=1}^{n} |\varphi(Y_i \cdot f_1(\mathbf{X}_i)) - \varphi(Y_i \cdot f_2(\mathbf{X}_i))| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} |Y_i f_1(\mathbf{X}_i) - Y_i f_2(\mathbf{X}_i)| \\
&= \frac{1}{n}\sum_{i=1}^{n} |f_1(\mathbf{X}_i) - f_2(\mathbf{X}_i)|.
\end{aligned}$$

Thus, if $\{f_1, \ldots, f_\ell\}$ is a ϵ-cover of $\mathcal{F}_n$, then $\{h_1, \ldots, h_\ell\}$ is a ϵ-cover of $\mathcal{H}_n$. Then

$$\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{H}_n, \mathbf{Z}_1^n\right) \leq \mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{F}_n, \mathbf{X}_1^n\right).$$

Set $L_{max} = \max\{L_n^{(1)}, L_n^{(2)}\}$. Lemma 4 implies

$$\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{F}_n, X_1^n\right) \leq \left(\frac{c_{12} \cdot \log n}{\epsilon}\right)^{c_{10} \cdot L_{\max}^2 \cdot \log(L_{\max} \cdot d_1 \cdot d_2)}$$

and

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{f \in \mathcal{F}_n} \left| \mathbf{E}\{\varphi(Y \cdot f(\mathbf{X}))\} - \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i \cdot f(\mathbf{X}_i)) \right| > \epsilon \right\} \\
&\leq 8 \cdot \left\{\left(\frac{c_{12} \cdot \log n}{\epsilon}\right)^{c_{10} \cdot L_{\max}^2 \cdot \log(L_{\max} \cdot d_1 \cdot d_2)}\right\} e^{-\frac{n\epsilon^2}{c_{13}(\log n)^2}}.
\end{aligned}$$

Set

$$\begin{aligned}\epsilon_n &= \sqrt{\frac{1}{2} \cdot \frac{c_{13} \cdot (\log n)^2}{n} \cdot \log \left(8 \left\{ \left(\frac{c_{12} \cdot \log n}{1/n} \right)^{c_{10} \cdot L_{\max}^2 \cdot \log(L_{\max} \cdot d_1 \cdot d_2)} \right\} \right)} \\ &= c_{14} \cdot (\log n)^2 \cdot \frac{\max\{L_n^{(1)}, L_n^{(2)}\}}{\sqrt{n}} \geq \frac{1}{n}.\end{aligned}$$

Here we have used for the last inequality that w.l.o.g. we can assume $\min\{c_{10}, c_{12}, c_{13}\} \geq 1$. Then

$$\begin{aligned}&\mathbf{E}\left\{ \sup_{f \in \mathcal{F}_n} \left| \mathbf{E}\left\{\varphi(Y \cdot f(\mathbf{X}))\right\} - \frac{1}{n} \sum_{i=1}^n \varphi(Y_i \cdot f(\mathbf{X}_i)) \right| \right\} \\ &= \epsilon_n + \int_{\epsilon_n}^\infty 8 \left\{ \left(\frac{c_{12} \cdot \log n}{t} \right)^{c_{10} \cdot L_{\max}^2 \cdot \log(L_{\max} \cdot d_1 \cdot d_2)} \right\} e^{-\frac{nt^2}{c_{13} \cdot (\log n)^2}} \, dt \\ &= \epsilon_n + \int_{\epsilon_n}^\infty 8 \left\{ \left(\frac{c_{12} \cdot \log n}{t} \right)^{c_{10} \cdot L_{\max}^2 \cdot \log(L_{\max} \cdot d_1 \cdot d_2)} \right\} e^{-\frac{nt^2}{2c_{13} \cdot (\log n)^2}} e^{-\frac{nt^2}{2c_{13} \cdot (\log n)^2}} \, dt \\ &\leq \epsilon_n + \int_{\epsilon_n}^\infty e^{-\frac{nt^2}{2 \cdot c_{13} \cdot (\log n)^2}} \, dt \\ &\leq \epsilon_n + \int_{\epsilon_n}^\infty e^{-\frac{n \cdot \epsilon_n \cdot t}{2 \cdot c_{13} \cdot (\log n)^2}} \, dt \\ &\leq c_{15} \cdot (\log n)^2 \cdot \frac{\max\{L_n^{(1)}, L_n^{(2)}\}}{\sqrt{n}}.\end{aligned}$$

□

Lemma 5 *Let $p \geq 1$ and $C > 0$ be arbitrary. Assume that $\eta : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ satisfies a (p, C)–smooth hierarchical max-pooling model. Let $\mathcal{F}_n$ be the set of all CNNs with ReLU activation function, which have $L_n^{(1)}$ convolutional layers with c_4 neurons in each layer, where c_4 is sufficiently large, one max pooling layer and $L_n^{(2)}$ additional layers with 7 neurons per layer. Furthermore assume $(L_n^{(1)})^{2p/d} \geq c_{16} \cdot L_n^{(2)}$. Then*

$$\inf_{f \in \mathcal{F}_n} \mathbf{E}\left\{\varphi(Y \cdot f(\mathbf{X}))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(\mathbf{X}))\right\} \leq c_{17} \cdot \left(\frac{\log L_n^{(2)}}{L_n^{(2)}} + \frac{1}{(L_n^{(1)})^{2p/4}} \right).$$

In order to prove Lemma 5 we need the following four auxiliary results.

Lemma 6 *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 $A \geq 1$ and $M \in \mathbb{N}$ sufficiently large (independent of the size of A, but*

$$M \geq 2 \text{ and } M^{2p} \geq c_{18} \cdot \left(\max\left\{ A, \|f\|_{C^q([-A,A]^d)} \right\} \right)^{4(q+1)},$$

where

$$\|f\|_{C^q([-A,A]^d)} = \max_{\substack{\alpha_1,\dots,\alpha_d\in\mathbb{N}_0,\\ \alpha_1+\dots+\alpha_d\le q}} \left\| \frac{\partial^q f}{\partial x_1^{\alpha_1}\dots\partial x_d^{\alpha_d}} \right\|_{\infty,[-A,A]^d},$$

must hold for some sufficiently large constant $c_{18} \ge 1$*).*
a) Let $L, r \in \mathbb{N}$ *be such that*

1. $L \ge 5 + \lceil \log_4(M^{2p}) \rceil \cdot (\lceil \log_2(\max\{q,d\}+1\}) \rceil + 1)$
2. $r \ge 2^d \cdot 64 \cdot \binom{d+q}{d} \cdot d^2 \cdot (q+1) \cdot M^d$

hold. There exists a feedforward neural network $f_{net,wide}$ *with ReLU activation function, L hidden layers and r neurons per hidden layer such that*

$$\|f - f_{net,wide}\|_{\infty,[-A,A]^d} \le c_{19} \cdot \left(\max\left\{A, \|f\|_{C^q([-A,A]^d)}\right\}\right)^{4(q+1)} \cdot M^{-2p}. \tag{7}$$

b) Let $L, r \in \mathbb{N}$ *be such that*

1. $L \ge 5M^d + \left\lceil \log_4 \left(M^{2p+4\cdot d\cdot(q+1)} \cdot e^{4\cdot(q+1)\cdot(M^d-1)} \right) \right\rceil \cdot \lceil \log_2(\max\{q,d\}+1) \rceil + \lceil \log_4(M^{2p}) \rceil$
2. $r \ge 132 \cdot 2^d \cdot \lceil e^d \rceil \cdot \binom{d+q}{d} \cdot \max\{q+1, d^2\}$

hold. There exists a feedforward neural network $f_{net,deep}$ *with ReLU activation function, L hidden layers and r neurons per hidden layer such that (7) holds with* $f_{net,wide}$ *replaced by* $f_{net,deep}$*.*

Proof. See Theorem 2 in Kohler and Langer (2020). □

Lemma 7 *Set*

$$f(z) = \begin{cases} \infty & , z = 1 \\ \log \frac{z}{1-z} & , 0 < z < 1 \\ -\infty & , z = 0 \end{cases}$$

and let $K \in \mathbb{N}$ *with* $K \ge 6$*. Let* $\eta : \mathbb{R}^d \to [0,1]$ *and let* $\bar{g} : \mathbb{R}^d \to \mathbb{R}$ *such that* $\|\bar{g} - \eta\|_\infty \le \epsilon$ *for some* $0 \le \epsilon \le 1/K$*. Then there exists a neural network* $\bar{f} : \mathbb{R} \to \mathbb{R}$ *with ReLU activation function,* $K+3$ *hidden layers with* 7 *neurons per layer, which is bounded in absolute value by* $\log(K+1)$ *and which satisfies*

$$\sup_{\mathbf{x}\in\mathbb{R}^{d_1\times d_2}} \left(\left|\eta(\mathbf{x}) \cdot \left(\varphi(\bar{f}(\bar{g}(\mathbf{x})) - \varphi(f(\eta(\mathbf{x}))\right)\right| + \left|(1-\eta(\mathbf{x})) \cdot \left(\varphi(-\bar{f}(\bar{g}(\mathbf{x})) - \varphi(-f(\eta(\mathbf{x}))\right)\right| \right)$$
$$\le c_{20} \cdot \left(\frac{\log K}{K} + \epsilon \right).$$

Proof. For $k \in \{-1, 0, \dots, K+1\}$ define

$$B_k(z) = \begin{cases} 0 & , z < \frac{k-1}{K} \\ K \cdot (z - \frac{k-1}{K}) & , \frac{k-1}{K} \leq z < \frac{k}{K} \\ K \cdot (\frac{k+1}{K} - z) & , \frac{k}{K} \leq z < \frac{k+1}{K} \\ 0 & , z \geq \frac{k+1}{K}, \end{cases}$$

(which implies $B_k(k/K) = 1$ and $B_k(j/K) = 0$ for $j \in \mathbb{Z} \setminus \{k\}$) and set

$$\begin{aligned} \bar{f}(z) &= f(1/K) \cdot (B_{-1}(z) + B_0(z)) + \sum_{k=1}^{K-1} f(k/K) \cdot B_k(z) + f(1 - 1/K) \cdot (B_K(z) + B_{K+1}(z)) \\ &=: \sum_{k=-1}^{K+1} a_k \cdot B_k(z). \end{aligned}$$

Then $\bar{f}$ interpolates the points $(-1/K, f(1/K))$, $(0, f(1/K))$, $(1/K, f(1/K))$, $(2/K, f(2/K))$, $\dots$, $((K-1)/K, f((K-1)/K))$, $(1, f((K-1)/K))$ and $(1 + 1/K, f((K-1)/K))$, is zero outside of $(-2/K, 1 + 2/K)$ and is linear on each interval $[k/K, (k+1)/K]$ $(k \in \{-2, \dots, K+1\})$. Because of

$$B_k(z) = \sigma\left(K \cdot \left(z - \frac{k-1}{K}\right)\right) - 2 \cdot \sigma\left(K \cdot \left(z - \frac{k}{K}\right)\right) + \sigma\left(K \cdot \left(z - \frac{k+1}{K}\right)\right)$$

$\bar{f}$ can be computed by a neural network with ReLU activation function and $K+3$ hidden layers with 7 neurons per layer. To see this we use that we have

$$\sigma(x) - \sigma(-x) = \max\{x, 0\} - \max\{-x, 0\} = x \quad \text{for } x \in \mathbb{R},$$

which enables us to compute $\bar{f}$ recursively as follows:

$$\bar{f}(x) = \bar{f}_1^{K+3} - \bar{f}_2^{K+3}$$

where

$$\begin{aligned} \bar{f}_1^l &= \sigma(\bar{f}_1^{l-1} - \bar{f}_2^{l-1}) + a_{l-2} \cdot B_{l-2}(\bar{f}_3^{l-1} - \bar{f}_4^{l-1}) \quad \text{for } l \in \{2, \dots, K+3\}, \\ \bar{f}_2^l &= \sigma(-\bar{f}_1^{l-1} + \bar{f}_2^{l-1}) \quad \text{for } l \in \{2, \dots, K+3\}, \\ \bar{f}_3^l &= \sigma(\bar{f}_3^{l-1} - \bar{f}_4^{l-1}) \quad \text{for } l \in \{2, \dots, K+3\}, \\ \bar{f}_4^l &= \sigma(-\bar{f}_3^{l-1} + \bar{f}_4^{l-1}) \quad \text{for } l \in \{2, \dots, K+3\} \end{aligned}$$

and

$$\bar{f}_1^1 = a_{-1} \cdot B_{-1}(x), \bar{f}_2^1 = 0, \bar{f}_3^1 = \sigma(x) \quad \text{and} \quad \bar{f}_4^1 = \sigma(-x).$$

By induction it is easy to see that the above recursion implies

$$\bar{f}_3^l - \bar{f}_4^l = x \quad \text{and} \quad \bar{f}_1^l - \bar{f}_2^l = \sum_{k=-1}^{l-2} a_k \cdot B_k(x)$$

for $l \in \{1, \ldots, K+3\}$.

Set

$$h_1(z) = \varphi(f(z)) = \log\left(1 + \exp\left(-\log\frac{z}{1-z}\right)\right) = \log\left(1 + \frac{1-z}{z}\right) = -\log z$$

and

$$h_2(z) = \varphi(-f(z)) = \log\left(1 + \exp\left(\log\frac{z}{1-z}\right)\right) = -\log(1-z).$$

First we consider the case $\eta(\mathbf{x}) \in [0, 2/K]$, which implies $f(\eta(\mathbf{x})) \leq f(2/K) = -\log(K/2-1) < 0$. In this case we have $-1/K \leq \bar{g}(\mathbf{x}) \leq 3/K$ and $-\log(K-1) = f(1/K) \leq \bar{f}(\bar{g}(\mathbf{x})) \leq f(3/K) = -\log(K/3-1)$. Consequently we get

$$\begin{aligned} \left|\eta(\mathbf{x}) \cdot \left(\varphi(\bar{f}(\bar{g}(\mathbf{x})) - \varphi(f(\eta(\mathbf{x}))\right)\right| &\leq \eta(\mathbf{x}) \cdot \varphi(\bar{f}(\bar{g}(\mathbf{x})) + \eta(\mathbf{x}) \cdot h_1(\eta(\mathbf{x})) \\ &\leq \frac{2}{K} \cdot \log(1 + \exp(\log(K-1))) + \eta(\mathbf{x}) \cdot \log(\frac{1}{\eta(\mathbf{x})}) \\ &\leq 4 \cdot \frac{\log K}{K} \end{aligned}$$

(where we have used $z \cdot \log(1/z) \leq (2/K) \cdot \log(K/2)$ for $0 < z < 2/K$) and

$$\begin{aligned} &\left|(1-\eta(\mathbf{x})) \cdot \left(\varphi(-\bar{f}(\bar{g}(\mathbf{x})) - \varphi(-f(\eta(\mathbf{x}))\right)\right| \\ &\leq \varphi(-\bar{f}(\bar{g}(\mathbf{x})) + \varphi(-f(\eta(\mathbf{x})) \\ &= \log(1 + \exp(\bar{f}(\bar{g}(\mathbf{x})))) + \log(1 + \exp(f(\eta(\mathbf{x})))) \\ &\leq \log(1 + \exp(-\log(K/3-1))) + \log(1 + \exp(-\log(K/2-1))) \\ &\leq 2 \cdot \exp(-\log(K/3-1)) = \frac{6}{K-3}. \end{aligned}$$

Similarly we get in case $\eta(\mathbf{x}) \geq 1 - 2/K$

$$\begin{aligned} &\left|\eta(\mathbf{x}) \cdot \left(\varphi(\bar{f}(\bar{g}(\mathbf{x})) - \varphi(f(\eta(\mathbf{x}))\right)\right| + \left|(1-\eta(\mathbf{x})) \cdot \left(\varphi(-\bar{f}(\bar{g}(\mathbf{x})) - \varphi(-f(\eta(\mathbf{x}))\right)\right| \\ &\leq 12 \cdot \frac{\log K}{K-3}. \end{aligned}$$

Hence it suffices to show

$$\begin{aligned} &\sup_{\substack{\mathbf{x} \in \mathbb{R}^d, \\ \eta(\mathbf{x}) \in [2/K, 1-2/K]}} \Bigg(\left|\eta(\mathbf{x}) \cdot \left(\varphi(\bar{f}(\bar{g}(\mathbf{x})) - \varphi(f(\eta(\mathbf{x}))\right)\right| \\ &\qquad\qquad + \left|(1-\eta(\mathbf{x})) \cdot \left(\varphi(-\bar{f}(\bar{g}(\mathbf{x})) - \varphi(-f(\eta(\mathbf{x}))\right)\right| \Bigg) \\ &\leq c_{21} \cdot (\frac{\log K}{K} + \epsilon). \end{aligned}$$

By the monotony of f, $|f'(z)| = \frac{1}{z \cdot (1-z)} \geq 1$ for $z \in (0,1)$, the mean value theorem and the definition of $\bar{f}$ we can conclude that for any $\mathbf{x} \in \mathbb{R}^d$ with $\eta(\mathbf{x}) \in [2/K, 1-2/K]$ we find $\xi_x, \delta_{\mathbf{x}} \in \mathbb{R}$ with $|\xi_x| \leq \frac{1}{K}$, $|\delta_{\mathbf{x}}| \leq \frac{1}{K} + \epsilon$ and $\eta(\mathbf{x}) + \delta_{\mathbf{x}} \in [1/K, 1-1/K]$ such that

$$\bar{f}(\bar{g}(\mathbf{x})) = f(\bar{g}(\mathbf{x}) + \xi_x) = f(\eta(\mathbf{x}) + \delta_{\mathbf{x}}). \tag{8}$$

This implies

$$
\begin{aligned}
&\sup_{\substack{\mathbf{x}\in\mathbb{R}^d,\\ \eta(\mathbf{x})\in[2/K,1-2/K]}} \Bigg(\left|\eta(\mathbf{x})\cdot\left(\varphi(\bar{f}(\bar{g}(\mathbf{x}))-\varphi(f(\eta(\mathbf{x}))\right)\right| \\
&\qquad + \left|(1-\eta(\mathbf{x}))\cdot\left(\varphi(-\bar{f}(\bar{g}(\mathbf{x}))-\varphi(-f(\eta(\mathbf{x}))\right)\right| \Bigg) \\
&= \sup_{\substack{\mathbf{x}\in\mathbb{R}^d,\\ \eta(\mathbf{x})\in[2/K,1-2/K]}} \Bigg(|\eta(\mathbf{x})|\cdot|h_1(\eta(\mathbf{x})+\delta_{\mathbf{x}})-h_1(\eta(\mathbf{x}))| \\
&\qquad + |1-\eta(\mathbf{x})|\cdot|h_2(\eta(\mathbf{x})+\delta_{\mathbf{x}})-h_2(\eta(\mathbf{x}))| \Bigg).
\end{aligned}
$$

Consequently it suffices to show that there exists a constant c_{22} such that we have for any $z\in[2/K,1-2/K]$ and any $\delta\in\mathbb{R}$ with $|\delta|\leq\frac{1}{K}+\epsilon$ and $z+\delta\in[1/K,1-1/K]$

$$
|z|\cdot|h_1(z+\delta)-h_1(z)|\leq c_{22}\cdot\left(\frac{1}{K}+\epsilon\right) \tag{9}
$$

and

$$
|1-z|\cdot|h_2(z+\delta)-h_2(z)|\leq c_{22}\cdot\left(\frac{1}{K}+\epsilon\right). \tag{10}
$$

We have

$$
h_1'(z)=-\frac{1}{z}.
$$

By the mean value theorem we get for some $\xi\in[\min\{z+\delta,z\},\max\{z+\delta,z\}]$

$$
|z|\cdot|h_1(z+\delta)-h_1(z)|=|z|\cdot\frac{1}{|\xi|}\cdot|\delta|\leq 4\cdot|\delta|\leq 4\cdot\left(\frac{1}{K}+\epsilon\right),
$$

where we have used that $z,z+\delta\in[1/K,1-1/K]$ and $\delta\leq 2/K$ imply $4|\xi|\geq|z|$.

In the same way we get

$$
h_2'(z)=\frac{1}{1-z}
$$

and

$$
|1-z|\cdot|h_2(z+\delta)-h_2(z)|=|1-z|\cdot\frac{1}{|1-\xi|}\cdot|\delta|\leq 4\cdot|\delta|\leq 4\cdot\left(\frac{1}{K}+\epsilon\right).
$$

□

Lemma 8 *Let $d_1,d_2,l\in\mathbb{N}$ with $2^l\leq\min\{d_1,d_2\}$ and set $I=\{0,1,\dots,2^l-1\}\times\{0,1,\dots,2^l-1\}$. Define m and $\bar{m}$ by*

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

and

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

where f and $\bar{f}$ satisfy

$$f = f_{l,1} \quad \text{and} \quad \bar{f} = \bar{f}_{l,1}$$

for some $f_{k,s}, \bar{f}_{k,s} : \mathbb{R}^{\{1,\dots,2^k\}\times\{1,\dots,2^k\}} \to \mathbb{R}$ recursively defined by

$$\begin{aligned} f_{k,s}(\mathbf{x}) \;=\; g_{k,s}\big(& f_{k-1,4\cdot(s-1)+1}(\mathbf{x}_{\{1,\dots,2^{k-1}\}\times\{1,\dots,2^{k-1}\}}),\\ & f_{k-1,4\cdot(s-1)+2}(\mathbf{x}_{\{2^{k-1}+1,\dots,2^k\}\times\{1,\dots,2^{k-1}\}}),\\ & f_{k-1,4\cdot(s-1)+3}(\mathbf{x}_{\{1,\dots,2^{k-1}\}\times\{2^{k-1}+1,\dots,2^k\}}),\\ & f_{k-1,4\cdot s}(\mathbf{x}_{\{2^{k-1}+1,\dots,2^k\}\times\{2^{k-1}+1,\dots,2^k\}})\big) \end{aligned}$$

and

$$\begin{aligned} \bar{f}_{k,s}(x) \;=\; \bar{g}_{k,s}\big(& \bar{f}_{k-1,4\cdot(s-1)+1}(\mathbf{x}_{\{1,\dots,2^{k-1}\}\times\{1,\dots,2^{k-1}\}}),\\ & \bar{f}_{k-1,4\cdot(s-1)+2}(\mathbf{x}_{\{2^{k-1}+1,\dots,2^k\}\times\{1,\dots,2^{k-1}\}}),\\ & \bar{f}_{k-1,4\cdot(s-1)+3}(\mathbf{x}_{\{1,\dots,2^{k-1}\}\times\{2^{k-1}+1,\dots,2^k\}}),\\ & \bar{f}_{k-1,4\cdot s}(\mathbf{x}_{\{2^{k-1}+1,\dots,2^k\}\times\{2^{k-1}+1,\dots,2^k\}})\big) \end{aligned}$$

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

$$f_{1,s}(x_{1,1},x_{1,2},x_{2,1},x_{2,2}) = g_{1,s}(x_{1,1},x_{1,2},x_{2,1},x_{2,2})$$

and

$$\bar{f}_{1,s}(x_{1,1},x_{1,2},x_{2,1},x_{2,2}) = \bar{g}_{1,s}(x_{1,1},x_{1,2},x_{2,1},x_{2,2})$$

for $s = 1,\dots,4^{l-1}$. Assume that all functions $g_{k,s} : \mathbb{R}^4 \to [0,1]$ are Lipschitz continuous regarding the Euclidean distance with Lipschitz constant $C > 0$ and for all $k \in \{1,\dots,l\}$ and $s \in \{1,\dots,4^{l-k}\}$ we assume that

$$\|\bar{g}_{k,s}\|_{[-2,2]^4,\infty} \leq 2. \tag{11}$$

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

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

Proof. The assertion follows from Lemma 4 in Kohler, Krzyżak and Walter (2020). □

Lemma 9 *Let $d_1, d_2, l \in \mathbb{N}$ with $2^l \leq \min\{d_1, d_2\}$. For $k \in \{1,\dots,l\}$ and $s \in \{1,\dots,4^{l-k}\}$ let*

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

be defined by a feedforward neural network with $L_{net} \in \mathbb{N}$ hidden layers and $r_{net} \in \mathbb{N}$ neurons per hidden layer and ReLU activation function. Set

$$I = \left\{0,\dots,2^l-1\right\} \times \left\{0,\dots,2^l-1\right\}$$

and define $\bar{m} : [0,1]^{\{1,\dots,d_1\}\times\{1,\dots,d_2\}} \to \mathbb{R}$ *by*

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

where $\bar{f}$ *satisfies*

$$\bar{f} = \bar{f}_{l,1}$$

for some $\bar{f}_{k,s} : [0,1]^{\{1,\dots,2^k\}\times\{1,\dots,2^k\}} \to \mathbb{R}$ *recursively defined by*

$$\begin{aligned} \bar{f}_{k,s}(\mathbf{x}) \;\; = \;\; & \bar{g}_{net,k,s}\big(\bar{f}_{k-1,4\cdot(s-1)+1}(\mathbf{x}_{\{1,\dots,2^{k-1}\}\times\{1,\dots,2^{k-1}\}}), \\ & \bar{f}_{k-1,4\cdot(s-1)+2}(\mathbf{x}_{\{2^{k-1}+1,\dots,2^k\}\times\{1,\dots,2^{k-1}\}}), \\ & \bar{f}_{k-1,4\cdot(s-1)+3}(\mathbf{x}_{\{1,\dots,2^{k-1}\}\times\{2^{k-1}+1,\dots,2^k\}}), \\ & \bar{f}_{k-1,4\cdot s}(\mathbf{x}_{\{2^{k-1}+1,\dots,2^k\}\times\{2^{k-1}+1,\dots,2^k\}})\big) \end{aligned}$$

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

$$\bar{f}_{1,s}(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}) = \bar{g}_{net,1,s}(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2})$$

for $s = 1,\dots,4^{l-1}$. *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

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

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*

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

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

Proof. See Lemma 5 in Kohler, Krzyżak and Walter (2020). □

Proof of Lemma 5. For each $g_{k,s}$ in the hierarchical max–pooling model for η we select an approximating neural network from Lemma 6 a) which approximates $g_{k,s} : \mathbb{R}^4 \to \mathbb{R}$ up to an error of order $(L_n^{(1)})^{-2p/4}$. Then we use Lemma 9 to generate with these networks a convolutional neural network, and combine this network with the feedforward neural network with $L_n^{(2)}$ layers and 7 neurons per layer of Lemma 7. For the corresponding network $h \in \mathcal{F}_n$ we get by Lemma 7 and Lemma 8

$$\sup_{x\in\mathbb{R}^{d_1\times d_2}} \left(\left|\eta(\mathbf{x}) \cdot \left(\varphi(h(\mathbf{x})) - \varphi(f_\varphi^*(\eta(\mathbf{x}))\right)\right| + \left|(1-\eta(\mathbf{x})) \cdot \left(\varphi(-h(\mathbf{x})) - \varphi(-f_\varphi^*(\eta(\mathbf{x}))\right)\right|\right)$$

$$\leq c_{23} \cdot \left(\frac{\log L_n^{(2)}}{L_n^{(2)}} + \frac{1}{(L_n^{(1)})^{2p/4}} \right).$$

Because of

$$\begin{aligned}
&\inf_{f \in \mathcal{F}_n} \mathbf{E}\left\{\varphi(Y \cdot f(\mathbf{X}))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(\mathbf{X}))\right\} \\
&\leq \mathbf{E}\left\{\varphi(Y \cdot h(\mathbf{X}))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(\mathbf{X}))\right\} \\
&= \int (\eta(\mathbf{x}) \cdot \varphi(h(\mathbf{x})) + (1-\eta(\mathbf{x})) \cdot \varphi(-h(\mathbf{x})))\mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\
&\quad - \int (\eta(\mathbf{x}) \cdot \varphi(f_\varphi^*(\mathbf{x})) + (1-\eta(\mathbf{x})) \cdot \varphi(-f_\varphi^*(\mathbf{x})))\mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\
&\leq \sup_{\mathbf{x} \in [0,1]^{d_1 \times d_2}} \Big(\left|\eta(\mathbf{x}) \cdot \left(\varphi(h(\mathbf{x})) - \varphi(f_\varphi^*(\eta(\mathbf{x}))\right)\right| \\
&\quad + \left|(1-\eta(\mathbf{x})) \cdot \left(\varphi(-h(\mathbf{x})) - \varphi(-f_\varphi^*(\eta(\mathbf{x}))\right)\right| \Big)
\end{aligned}$$

this implies the assertion.

□

Proof of Theorem 1. In the sequel we show

$$\mathbf{E}\left\{\varphi(Y \cdot \operatorname{sgn}(\hat{f}_n(\mathbf{X})))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(\mathbf{X}))\right\} \leq c_{24} \cdot (\log n)^2 \cdot n^{-\min\{\frac{p}{2p+4}, \frac{1}{4}\}}. \tag{12}$$

This implies the assertion, because by Lemma 1 a) we can conclude from (12)

$$\begin{aligned}
&\mathbf{P}\left\{Y \neq \operatorname{sgn}(\hat{f}_n(\mathbf{X}))|\right\} - \mathbf{P}\left\{Y \neq f^*(\mathbf{X})\right\} \\
&\leq \mathbf{E}\left\{ \frac{1}{\sqrt{2}} \cdot \left(\mathbf{E}\left\{\varphi(Y \cdot \operatorname{sgn}(\hat{f}_n(\mathbf{X})))|\mathcal{D}_n\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(\mathbf{X}))\right\} \right)^{1/2} \right\} \\
&\leq \frac{1}{\sqrt{2}} \cdot \sqrt{\mathbf{E}\left\{\varphi(Y \cdot \operatorname{sgn}(\hat{f}_n(\mathbf{X})))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(\mathbf{X}))\right\}} \\
&\leq c_{25} \cdot (\log n) \cdot n^{-\min\{\frac{p}{4p+8}, \frac{1}{8}\}}.
\end{aligned}$$

And from Lemma 1 b), (4) and Lemma 1 c) we can conclude from (12)

$$\begin{aligned}
&\mathbf{P}\left\{Y \neq \operatorname{sgn}(\hat{f}_n(\mathbf{X}))|\right\} - \mathbf{P}\left\{Y \neq f^*(\mathbf{X})\right\} \\
&\leq 2 \cdot \left(\mathbf{E}\left\{\varphi(Y \cdot \operatorname{sgn}(\hat{f}_n(\mathbf{X})))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(\mathbf{X}))\right\} \right) + 4 \cdot \frac{c_{26} \cdot \log n}{\sqrt{n}} \\
&\leq c_{27} \cdot (\log n)^2 \cdot n^{-\min\{\frac{p}{2p+4}, \frac{1}{4}\}}.
\end{aligned}$$

So it suffices to prove (12). Application of Lemma 2, Lemma 3 and Lemma 5 yields

$$\mathbf{E}\left\{\varphi(Y \cdot \hat{f}_n(\mathbf{X}))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(\mathbf{X}))\right\}$$

$$
\begin{aligned}
&\le 2 \cdot \mathbf{E}\left\{ \sup_{f \in \mathcal{F}_n} \left| \mathbf{E}\left\{\varphi(Y \cdot f(\mathbf{X}))\right\} - \frac{1}{n}\sum_{i=1}^n \varphi(Y_i \cdot f(\mathbf{X}_i)) \right| \right\} \\
&\quad + \inf_{f \in \mathcal{F}_n} \mathbf{E}\left\{\varphi(Y \cdot f(\mathbf{X}))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(\mathbf{X}))\right\} \\
&\le c_{28} \cdot (\log n)^2 \cdot \frac{\max\{L_n^{(1)}, L_n^{(2)}\}}{\sqrt{n}} + c_{29} \cdot \left(\frac{\log L_n^{(2)}}{L_n^{(2)}} + \frac{1}{(L_n^{(1)})^{2p/4}} \right) \\
&\le c_{28} \cdot (\log n)^2 \cdot \frac{L_n^{(2)}}{\sqrt{n}} + c_{29} \cdot \frac{\log L_n^{(2)}}{L_n^{(2)}} + c_{28} \cdot (\log n)^2 \cdot \frac{L_n^{(1)}}{\sqrt{n}} + c_{29} \cdot \frac{1}{(L_n^{(1)})^{2p/4}} \\
&\le c_{30} \cdot (\log n)^2 \cdot n^{-\min\{\frac{p}{2p+4}, \frac{1}{4}\}}.
\end{aligned}
$$

□

References

[1] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification and risk bounds. *Journal of the American Statistical Association*, **101**, pp. 138–156.

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

[3] Cover, T. M. (1968). Rates of convergence of nearest neighbor procedures. In *Proceedings of the Hawaii International Conference on Systems Sciences*, pp. 413-415, Honolulu, HI.

[4] Devroye, L. (1982). Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **61**, pp. 467-481.

[5] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory Of Pattern Recognition.* Springer.

[6] Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. arXiv: 1811.03804.

[7] Eckle, K., and Schmidt-Hieber, J. (2019). A comparison of deep networks with ReLU activation function and linear spline-type methods. *Neural Networks*, **110**, pp. 232-242.

[8] Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, **28**, pp. 337–407.

[9] Friedman, J., Hastie, T. and Tibshirani, R. (2009). *The Elements of Statistical Learning.* Springer.

[10] Golik, P., Doetsch, P. and Ney, H. (2013). Cross-entropy vs. squared error training: a theoretical and experimental comparison. *Interspeech*, pp. 1756 - 1760.

[11] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution–Free Theory of Nonparametric Regression*. Springer.

[12] He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recoginition. *Proceedings of the IEEE conference on computer vision and pattern recoginition*, pp. 770-778.

[13] Imaizumi, M., and Fukamizu, K. (2019). Deep neural networks learn non-smooth functions effectively. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019).* Naha, Okinawa, Japan.

[14] Kim, Y., Ohn, I, and Kim, D. (2019). Fast convergence rates of deep neural networks for classification. arXiv: 1812.03599.

[15] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira et al. (Eds.), *Advances In Neural Information Processing Systems* **25**, pp. 1097-1105. Red Hook, NY: Curran.

[16] Kohler, M., and Krzyżak, A. (2017). Nonparametric regression based on hierarchical interaction models. *IEEE Transaction on Information Theory*, **63**, pp. 1620-1630.

[17] Kohler, M., and Krzyżak, A. (2019). Over-parametrized deep neural networks minimizing the empirical risk do not generalize well. arXiv:1912.03925.

[18] Kohler, M., Krzyżak, A., and Langer, S. (2019). Estimation of a function of low local dimensionality by deep neural networks. arXiv: 1908.11140.

[19] Kohler, M., and Langer, S. (2020). On the rate of convergence of fully connected deep neural network regression estimates. To appear in *Ann. Stat.* arXiv: 1908.11133

[20] Kohler, M., Krzyżak, A., and Walter, B. (2020). On the rate of convergence of image classifiers based on convolutional neural networks. arXiv: 2003.01526

[21] Langer, S. (2020). Analysis of the rate of convergence of fully connected deep neural network regression estimates with smooth activation function. arXiv: 2010.06168

[22] LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, **521**, pp.436-444.

[23] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, **1**, pp. 541-551.

[24] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**, pp.2278-2324.

[25] Lin, S., and Zhang, J. (2019). Generalization bounds for convolutional neural networks. arXiv: 1910.01487.

[26] Oono, K., and Suzuki, T. (2019) Approximation and Non-parametric Estimation of ResNet-type Convolutional Neural Networks. arXiv: 1903.10047

[27] Rawat, W., and Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, **29**, pp. 2352-2449.

[28] Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks*, **61**, pp. 85-117.

[29] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, **4**, pp. 1875-1897.

[30] Suzuki, T., and Nitanda, A. (2019). Deep learning is adaptive to intrinsic diemsnionality of model smoothness in anisotropic Besov space. arXiv: 1910.12799.

[31] Vapnik, V. N. (1998). Statistical Learning Theory. *Wiley*.

[32] Yarotsky, D. (2018). Universal approximations of invariant maps by neural networks. arXiv: 1804.10306.

[33] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, **32**, pp. 56 - 134.