Learning of deep convolutional neural network image classifiers in an image deformation model *

Michael Kohler [1] and Alisha Sänger [1,†]
[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: kohler@mathematik.tu-darmstadt.de, saenger@mathematik.tu-darmstadt.de*

July 2, 2026

Abstract
Image classification from independent and identically distributed random variables is considered. It is assumed that the images are generated by an image deformation model, where each image is a deformation of one of two given images. Overparametrized deep convolutional neural network image classifier are defined where the weights are learned by stochastic gradient descent. A theoretical result is presented which shows that the expected misclassification probability of the estimate converges to zero with rate $1/\sqrt{n}$ where n is the sample size.

AMS classification: Primary 62G05; secondary 62G20.

Key words and phrases: Convolutional neural networks, image classification, stochastic gradient descent, over-parametrization, rate of convergence.

1. Introduction

1.1. Image deformation

In this paper we use an image deformation model introduced in Chen, Langer and Schmidt-Hieber (2024) in order to investigate image classifiers. In this model it is assumed that the images are discrete observations of deformations of two continuous images described by mappings $f_0, f_1 : \mathbb{R}^2 \to [0, \infty)$. Here f_0 and f_1 characterize the grayscales of the images with smaller function values corresponding to darker pixels. The images are observed on a grid of size $d \times d$, and the observation at position $(i, j) \in \{1, \dots, d\} \times \{1, \dots, d\}$ is the average grayscale

$$\bar{f}_{i,j} = d^2 \cdot \int_{[(i-1)/d, i/d] \times [(j-1)/d, j/d]} f(u, v)\, du\, dv \tag{1}$$

of the image in the square $[(i-1)/d, i/d] \times [(j-1)/d, j/d]$ ($f \in \{f_0, f_1\}$). In our model we assume that we observe transformations of the images. More precisely, we let

*Running title: *Deep network classifiers*
†Corresponding author. Tel: +49-, Fax:+49

$A : \mathbb{R}^2 \to \mathbb{R}^2$ be an invertible mapping, and let $\eta \in (0, \infty)$ be a brightness factor. Then we assume that in our training data instead of (1) we observe

$$d^2 \cdot \eta \cdot \int_{[(i-1)/d, i/d] \times [(j-1)/d, j/d]} f(A(u, v))\, du\, dv \tag{2}$$

($(i, j) \in \{1, \dots, d\} \times \{1, \dots, d\}$) together with the information whether $f = f_0$ or $f = f_1$ holds for the point in our training data. And the aim is to predict for an additional observation (2) for $(i, j) \in \{1, \dots, d\} \times \{1, \dots, d\}$ whether $f = f_0$ or $f = f_1$ holds.

Possible examples for the above invertible mapping A include affine transformations

$$A(u, v) = \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix} \cdot \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} \tau \\ \tau' \end{pmatrix},$$

where $b_1, b_2, b_3, b_4, \tau, \tau' \in \mathbb{R}$ with $b_1 \cdot b_4 - b_2 \cdot b_3 \neq 0$. If we choose

$$\begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix} = \begin{pmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{pmatrix} \cdot \begin{pmatrix} \xi & 0 \\ 0 & \xi' \end{pmatrix} \tag{3}$$

we compose a scaling in x- and y-direction with a rotation and add a translation by τ, τ' along x- and y-direction respectively.

This leads to the following model for the generation of our data. Given are two functions $f_0, f_1 : \mathbb{R}^2 \to [0, \infty)$, some $\pi \in [0, 1]$, some class $\mathcal{A}$ of invertible transformations $A : \mathbb{R}^2 \to \mathbb{R}^2$, a probability distribution Q_1 on $\mathcal{A}$ and a probability distribution Q_2 on $(0, \infty]$. In a first step, independently a value $k \in \{0, 1\}$ is chosen from a Bernoulli distribution with probability of sucess π, an invertible transformation A is chosen from $\mathcal{A}$ using the probability distribution Q_1 and a brightness factor η is chosen from $(0, \infty)$ using the probability distribution Q_2. Then $X = (X_{i,j})_{i,j \in \{1,\dots,d\}} \in [0, 1]^{\{1,\dots,d\} \times \{1,\dots,d\}}$ is generated by

$$X_{i,j} = d^2 \cdot \eta \cdot \int_{[(i-1)/d, i/d] \times [(j-1)/d, j/d]} f_k(A(u, v))\, du\, dv, \tag{4}$$

and Y is set to the class of the image, i.e.,

$$Y = \begin{cases} 1, & \text{if } k = 1 \\ -1, & \text{if } k = 0. \end{cases}. \tag{5}$$

In this way independent and identically distributed data (X, Y), (X_1, Y_1), $\dots$, (X_n, Y_n) is generated. Given the training data

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}, \tag{6}$$

the task is to construct an estimate

$$f_n : [0, 1]^{\{1,\dots,d\} \times \{1,\dots,d\}} \to \{0, 1\}$$

such that the probability of misclassification

$$\mathbf{P}\{f_n(X) \neq Y | \mathcal{D}_n\} \tag{7}$$

is as small as possible.

1.2. Results in Chen, Langer and Schmidt-Hieber (2024)

The above image deformation model was introduced in Chen, Langer and Schmidt-Hieber (2024). There as theoretical benchmarks two different classifiers based on the one-nearest neighbor classifier are investigated. These classifiers are based on a transformation of the input space. It is shown that under a separation condition on the images perfect classification is achieved by these two approaches, i.e., the classifiers achieve zero probability of misclassification with high probability.

Furthermore, convolutional neural network (CNN) classifiers are analyzed in this model. Here the estimates are defined by empirical risk minimization based either on the 0-1-loss or the cross-entropy loss. A special topology of the CNN is proposed consisting of one convolutional layer followed by one max-pooling layer and a deep feedforward neural network, where the ReLU activation function is used as the activation function. Under a separation condition on the images it is shown that the misclassification probability of these estimates converges to zero with rate

$$O\left(\frac{\log n}{n}\right),$$

i.e., CNN classifiers achieve a parametric rate of convergence in this context.

In applications, it is not possible to minimize the empirical risk based either on the 0-1-loss or the cross-entropy loss for CNN classifiers, since the underlying minimization problem is nonlinear and non-convex and there are no feasible algorithms known for solving this problem. Instead usually one uses gradient descent (or one of its variants like stochastic gradient descent) to compute the corresponding CNN classifier. So the above results immediately lead to the question whether similar results can also be achieved for CNN classifiers computed by gradient descent, which is exactly the question which we will study in the sequel.

1.3. Main results

In the above introduced image deformation model we define estimates that consist of a linear combination of a large number of CNNs with ReLU activation function that are computed in parallel. Each of the CNNs in the linear combination consists of one convolutional layer followed by a max-pooling layer and a deep feedforward neural network. We propose a special way to initialize the weights of the network randomly, and train our estimates by minimizing the empirical logistic loss of a linear combination of networks via projected stochastic gradient descent with a special stepsize and a special number of gradient descent steps (see Section 2 and Theorem 1 for the details).

We analyze the rate of convergence of the estimates and provide under a separation condition the following upper bound on the misclassification probability in Theorem 1 below:

$$\mathbf{P}\left\{Y \neq \hat{C}_n(X)\right\} \leq c_1 \cdot \frac{(\log n)^{\frac{3}{2}} \cdot d^2 \cdot \log(d) \cdot |\mathcal{A}_d| \cdot (\log(|\mathcal{A}_d|))^2}{\sqrt{n}}. \tag{8}$$

Here $\mathcal{A}_d$ is a finite class of transformations used to approximate the transformations in our data generating model. So we are able to show a parametric rate of convergence also for estimates learned by gradient descent, but we get a weaker upper bound of $O((\log n)^{\frac{3}{2}}/\sqrt{n})$ than the upper bound $O((\log n)/n)$ proven by Chen, Langer and Schmidt-Hieber (2024) for the estimates defined by empirical risk minimization.

To establish our result, we present a general setting for estimates that consist of a convex combination of neural networks. In this setting we derive a general theorem that provides an upper bound on the expected logistic loss of neural networks trained via stochastic gradient descent to minimize the empirical logistic loss.

Furthermore, we implement the proposed network topology and study the finite sample size performance of the estimate on artificially generated samples from properly defined image deformation models and also on a real image classification data set. Here we see that indeed the probability of misclassification of the CNN classifier based on the topology and the initialization of the weights proposed in this article seems to converge to zero with increasing sample size, and it looks like the convergence takes place even faster than in our theoretical upper bound presented above.

1.4. Discussion of related results

Despite the astonishing accomplishments of deep learning and convolutional neural networks achieved by practitioners, there remain significant gaps in their theoretical foundation. In visual recognition applications, the data is notoriously high-dimensional, which makes the problems relatively hard to solve. In order to bridge the gap between empirical success and theoretical understanding, it is often assumed that the underlying functional relationship between inputs and outputs admits a low-dimensional representation (cf. e.g. Liu et al. (2021), Shen et al. (2021), Kohler, Krzyżak and Walter (2025), Kohler and Langer (2025)). Under this structural assumption, it is possible to derive convergence rates that depend only on the underlying low-dimensional representation, thereby potentially overcoming the limitations imposed by high-dimensional input spaces. In this work we study a binary classification problem, where each class is characterized by an underlying bivariate function and the observed images are discretizations thereof. In this setting, images of the same class are assumed to be deformed versions of a template image. This idea has been introduced by Grenander (1969), who characterizes the pattern recognition problem in terms of "pure images" and "deformation mechanisms". To solve the problem, one needs to find a recognition function that recovers the pattern class of the underlying pure image from a deformed image.

Convolutional neural networks (CNN) are commonly used to approximate this "recognition function". Practitioners have studied datasets subject to isolated deformations, e.g. Marcos, Volpi and Tuja (2016) attempt to encode rotation invariance by incorporating rotated versions of convolutional filters in their model and Jansson and Lindeberg (2022) deal with scale-invariance using a "scale-channel architecture". Cohen and Welling (2016) introduce "group convolutional" layers for discrete groups generated by translations, reflections and rotations. These layers are designed with a higher degree of weight sharing than standard convolutional layers, thereby enabling them to exploit larger symmetry

groups.

Kohler and Walter (2023) consider a binary pattern recognition problem, where the deformation mechanisms consist of rotations and establish theoretical convergence guarantees for the proposed empirical risk minimization-based network estimate, which are independent of the dimension of the image. They build upon the hierarchical max-pooling model for the a posteriori probability introduced in Kohler, Krzyżak and Walter (2022), extending it by rotating each subpart of an image by different angles.

However, in practice, it is not possible to compute the empirical risk minimizer. Therefore, gradient descent algorithms have been used to minimize the empirical risk. The convergence behavior of gradient descent has been studied intensively, e.g. in Arora et al. (2019), Du et al. (2019) and Bottou, Curtis and Nocedal (2018). A widely adopted approach to deal with large datasets is *stochastic* gradient descent (SGD), which is a drastic simplification in the sense that the exact computation of the gradient in each iteration is replaced by an estimate on the basis of a randomly chosen (sub-)sample. In order to limit the impact of the noisy gradient approximation on convergence speed Kohler, Krzyżak and Sänger (2025) impose further conditions on the iterates by using *projected* SGD. A linear combination of a large number of convolutional neural networks is considered. The projection step ensures that in this setting, the inner weights do not move too far away from their random initializations, while the outer weights are adjusted suitably by SGD.

Similar ideas have been applied to the analysis of the L_2-error of regression estimates with sigmoid activation function, which are learned by gradient descent in Braun et al. (2024) and Kohler and Krzyżak (2025). Here the fact that the internal weights do not change significantly during training is derived from properties of the sigmoid activation function together with the bounds imposed on the weights. Kohler, Krzyżak and Walter (2025) use the same approach to bound the optimization error of their CNN estimates for a binary image classification setting. However, this reasoning does not extend to estimates which employ the unbounded ReLU activation, which motivates the inclusion of a projection step in the gradient procedure, as described above, cf. also Kohler, Cho and Krzyżak (2025).

Gonon (2023) studies *random feature (neural) networks*, where the weights of the hidden layer are randomly initialized and subsequently held fixed, such that only the parameters of the outer layer are adjusted during the training of the network (cf. also Huang, Chen and Siew (2006)). This enables the authors to leverage convex optimization techniques. The so-called *lottery ticket hypothesis* offers a conceptually related viewpoint. Frankle and Carbin (2019) suggest that it is possible to identify sparse subnetworks which can be trained effectively when reset to their initial parameters through an iterative pruning process. These subnetworks are deemed "winning tickets" of the "initialization lottery". Da Cunha et al. (2022) derive a theoretical approximation result for this setting. Specifically the authors show that for a CNN with fixed weights, there exists a randomly initialized CNN, whose size scales logarithmically, which can be pruned to a network that approximates the original network well.

Besides the analysis of the learning algorithm, i.e. the optimization component, a theoretical analysis of deep learning should simultaneously incorporate approximation

and generalization aspects (cf. Kutyniok (2020)). Quite a lot of attention has been dedicated to approximation properties of deep neural networks in general (e.g. Yarotsky and Zhevnerchuk (2020), Kohler and Langer (2025), Lu et al. (2020), cf. Gühring, Raslan and Kutyniok (2022) for a comprehensive overview). Approximation guarantees for CNNs specifically have been established by Yarotsky (2022) and Zhou (2020). Yarotsky (2022) generalizes the universal approximation theorem from fully connected networks to CNN-like architectures and groups of translation.

Arora et al. (2018) obtain generalization bounds for fully connected networks via post-training compression schemes that reduce the effective number of parameters and show how this methodology extends to CNNs. Long and Sedghi (2020) consider a multiclass image classification problem and derive generalization bounds for CNNs that depend on the distance of the final network weights from the initial weights and the number of parameters, but not on the dimensions of the images. The central technique here are bounds on the covering number, borrowing from the analysis of fully-connected networks, cf. Bartlett, Foster and Telgarsky (2017).

1.5. Notation

The sets of natural numbers, real numbers and nonnegative real numbers are denoted by $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$, respectively. We define furthermore $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. For $z \in \mathbb{R}$, we denote the smallest integer greater than or equal to z by $\lceil z \rceil$, the largest integer less than or equal to z by $\lfloor z \rfloor$, and we set $z_+ = \max\{z, 0\}$. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$. For a closed and convex set $A \subseteq \mathbb{R}^d$ we denote by $\mathrm{Proj}_A\, x$ that element $\mathrm{Proj}_A\, x \in A$ with

$$\|x - \mathrm{Proj}_A\, x\| = \min_{z \in A} \|x - z\|.$$

For $f : \mathbb{R}^d \to \mathbb{R}$

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

is its supremum norm, and we set

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

for $A \subseteq \mathbb{R}^d$.

For $\mathbf{j} = (j^{(1)}, \dots, j^{(d)}) \in \mathbb{N}_0^d$ we write

$$\|\mathbf{j}\|_1 = j^{(1)} + \cdots + j^{(d)}$$

and for $f : \mathbb{R}^d \to \mathbb{R}$ we set

$$\partial^{\mathbf{j}} f = \frac{\partial^{\|\mathbf{j}\|_1} f}{(\partial x^{(1)})^{j^{(1)}} \dots (\partial x^{(d)})^{j^{(d)}}}.$$

Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$, let $x_1, \dots, x_n \in \mathbb{R}^d$, set $x_1^n = (x_1, \dots, x_n)$ and let $p \geq 1$. A finite collection $f_1, \dots, f_N : \mathbb{R}^d \to \mathbb{R}$ is called an L_p ε–packing in $\mathcal{F}$ on x_1^n if

$f_1, \dots, f_N \in \mathcal{F}$ and

$$\min_{1 \le i < j \le N} \left(\frac{1}{n} \sum_{k=1}^{n} |f_i(x_k) - f_j(x_k)|^p \right)^{1/p} \ge \varepsilon$$

hold. The L_p ε–packing number of $\mathcal{F}$ on x_1^n is the size N of the largest L_p ε–packing of $\mathcal{F}$ on x_1^n and is denoted by $\mathcal{M}_p(\varepsilon, \mathcal{F}, x_1^n)$.

For $z \in \mathbb{R}$ and $\beta > 0$ we define $T_\beta z = \max\{-\beta, \min\{\beta, z\}\}$. If $f : \mathbb{R}^d \to \mathbb{R}$ is a function then we set $(T_\beta f)(x) = T_\beta(f(x))$. And $\operatorname{sign}(z)$ is the sign of $z \in \bar{\mathbb{R}}$.

We define the discrete L^2-inner product for functions $g, h : [0,1]^2 \to \mathbb{R}$ by

$$\langle h, g \rangle_{2,d} := \frac{1}{d^2} \sum_{j,\ell=1}^{d} \overline{h}_{j,\ell} \overline{g}_{j,\ell}, \tag{9}$$

with $\overline{h}$ and $\overline{g}$ the pixel values defined in (1). The corresponding norm is then

$$\|g\|_{2,d} := \sqrt{\langle g, g \rangle_{2,d}} = \frac{\sqrt{\sum_{j,\ell=1}^{d} \overline{g}_{j,\ell}{}^2}}{d}. \tag{10}$$

1.6. Outline

The deep convolutional neural network classifiers considered in this work are introduced in Section 2. The main result, which establishes a bound on their misclassification probability, is stated in Section 3. In Section 4 we describe how the proposed estimator was implemented and conduct an empirical study to corroborate our theoretical findings. Section 5 presents the proof of the main result, starting with a general result in Section 5.1. The full proof of this general result is deferred to the appendix.

2. Definition of the estimate

Chen, Langer and Schmidt-Hieber (2024) used a single convolutional layer with many channels and a channel-wise defined max-pooling layer in order to decide for each of the two functions f_0 and f_1 and each invertible transformation how likely it is that the image is generated by applying this transformation to the function, and then applied a feedforward neural network to the results of these channels in order to decide whether the image was generated from f_0 or f_1. In Chen, Langer and Schmidt-Hieber (2024) empirical risk minimization was used to estimate the weights of this network from the observed data, which is not possible in applications. In the sequel we want to use stochastic gradient descent instead.

In a preprocessing step we replace each $X_s = ((X_s)_{i,j=1,\dots,d})$ by the normalized image

$$\bar{X}_s = ((\bar{X}_s)_{i,j=1,\dots,d}) \quad \text{where} \quad (\bar{X}_s)_{i,j} = \frac{(X_s)_{i,j}}{\sqrt{\sum_{k,l=1}^{d} |(X_s)_{k,l}|^2}}$$

such that all pixel values lie within the interval $[0,1]$. In the sequel we use a linear combination

$$f_n(x) = f_{(\mathbf{w},\boldsymbol{\vartheta})}(x) = \sum_{k=1}^{K_n} w_k \cdot \beta_n \cdot T_1 f_{\vartheta_k}(x) \tag{11}$$

of K_n networks f_{ϑ_k} $(k = 1, \ldots, K_n)$ which are computed in parallel and where

$$w_k \geq 0 \quad (k = 1, \ldots, K_n), \quad \sum_{k=1}^{K_n} |w_k| \leq 1 \quad \text{and} \quad \sum_{k=1}^{K_n} |w_k|^2 \leq \alpha_n. \tag{12}$$

In the following we chose the ReLU function $\sigma(x) = \max\{x, 0\}$ as activation function for the networks f_{ϑ_k}, which compute in the first layer a convolution with $2m$ channels

$$o^{(1)}_{(i,j),k,s} = \sigma\left(\sum_{\substack{r_1,r_2=1,\ldots,h:\\ i+r_1-1\leq d, j+r_2-1\leq d}} w_{k,s,r_1,r_2} \cdot \bar{x}_{i+r_1-1,j+r_2-1} \right) \quad (s = 1, \ldots, 2m), \tag{13}$$

and subsequently apply a channel-wise max-pooling to the results:

$$o^{(2)}_{k,s} = \max\left\{ o^{(1)}_{(i,j),k,s} \,:\, (i,j) \in \{1,\ldots,d\} \times \{1,\ldots,d\} \right\} \quad (s = 1, \ldots, 2m). \tag{14}$$

Then a fully connected feedforward neural network with $1 + 2\lceil \log_2 m \rceil$ many hidden layers and $4m$ neurons per hidden layer is applied to the $2m$ results of the max-pooling layer, thus we get the following for the first fully-connected layer $o^{(3)}_{k,r}$

$$o^{(3)}_{k,r} = \sigma\left(\sum_{t=1}^{2m} w^{(2)}_{k,r,t} \cdot o^{(2)}_{k,t} + w^{(2)}_{k,r,0} \right), \tag{15}$$

with $r = 1, \ldots, 4m$, which is followed by

$$o^{(2+s)}_{k,r} = \sigma\left(\sum_{t=1}^{4m} w^{(2+s-1)}_{k,r,t} \cdot o^{(2+s-1)}_{k,t} + w^{(2+s-1)}_{k,r,0} \right) \tag{16}$$

for $s = 2, \ldots, 1 + 2\lceil \log_2 m \rceil$ and $r = 1, \ldots, 4m$.
Finally we set

$$f_{\vartheta_k}(x) = \sum_{t=1}^{4m} w^{(2+1+2\lceil \log_2 m \rceil)}_{k,1,t} \cdot o^{(2+1+2\lceil \log_2 m \rceil)}_{k,t}. \tag{17}$$

Here ϑ_k is the vector consisting of all the

$$W = 2m \cdot h^2 + 4m \cdot (2m+1) + 2\lceil \log_2 m \rceil \cdot 4m \cdot (4m+1) + 4m$$

many weights of the k-th network and we choose ϑ_k from the set

$$\boldsymbol{\Theta} = [-1,1]^W.$$

The weights of the linear combination in (11) are initialized by setting

$$w_k^{(0)} = 0 \quad (k = 1, \dots, K_n).$$

All other weights, i.e. the weights of the convolution and fully connected layers

$$\vartheta_1^{(0)}, \dots, \vartheta_{K_n}^{(0)}$$

are chosen independently and uniformly from

$$\mathbf{\Theta}^0 = \{\vartheta \in \mathbf{\Theta} \quad : \quad \|\vartheta\|_\infty \leq B_n\},$$

such that they are also independent from (X, Y), (X_1, Y_1), ..., (X_n, Y_n), where B_n is a constant defined in Theorem 1.
After that we use stochastic gradient descent in order to minimize the empirical logistic risk

$$F_n((\mathbf{w}, \boldsymbol{\vartheta})) = \frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i \cdot f_{(\mathbf{w}, \boldsymbol{\vartheta})}(X_i)) \tag{18}$$

where

$$\varphi(x) = \log(1 + e^{-x})$$

is the logistic loss.

Set

$$\boldsymbol{\vartheta}^{(0)} = (\vartheta_1^{(0)}, \dots, \vartheta_{K_n}^{(0)}) \quad \text{and} \quad \mathbf{w}^{(0)} = (w_1^{(0)}, \dots, w_{K_n}^{(0)}),$$

assume that t_n/n is a natural number, and for $s \in \{1, \dots, t_n/n\}$ let

$$j_{(s-1)\cdot n}, \dots, j_{s \cdot n - 1}$$

be an arbitrary permutation of $1, \dots, n$. Choose a stepsize $\lambda_n > 0$ and set

$$\begin{aligned} \mathbf{w}^{(t+1)} &= \mathrm{Proj}_A \left(\mathbf{w}^{(t)} - \lambda_n \cdot \nabla_{\mathbf{w}} \varphi \left(Y_{j_t} \cdot f_{(\mathbf{w}^{(t)}, \boldsymbol{\vartheta}^{(t)})}(X_{j_t}) \right) \right), \\ \boldsymbol{\vartheta}^{(t+1)} &= \mathrm{Proj}_B \left(\boldsymbol{\vartheta}^{(t)} - \lambda_n \cdot \nabla_{\boldsymbol{\vartheta}} \varphi \left(Y_{j_t} \cdot f_{(\mathbf{w}^{(t)}, \boldsymbol{\vartheta}^{(t)})}(X_{j_t}) \right) \right) \end{aligned}$$

for $t = 0, \dots, t_n - 1$. Here A is the set of all $\mathbf{w}$ which satisfy (12), and

$$B = \left\{ \boldsymbol{\vartheta} \in \mathbf{\Theta}^{K_n} \, : \, \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^{(0)}\| \leq 1 \right\},$$

and Proj_A and Proj_B is the L_2 projection on A and B.

In order to compute the gradient with respect to the inner weights we use the following convention: We set

$$\sigma'(z) = \frac{\partial}{\partial z} \max\{z, 0\} = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{else} \end{cases}$$

$$\frac{\partial}{\partial z} T_1 z = \frac{\partial}{\partial z} \max\{-1, \min\{1, z\}\} = \begin{cases} 1, & \text{if } |z| \leq 1 \\ 0, & \text{else} \end{cases}$$

and

$$\frac{\partial}{\partial \vartheta_j} \max\left\{f_{\vartheta_1}(x), \ldots, f_{\vartheta_L}(x)\right\} = \frac{\partial}{\partial \vartheta_j} f_{\vartheta_l}(x)$$

where $l \in \{1, \ldots, L\}$ satisfies

$$\max\left\{f_{\vartheta_1}(x), \ldots, f_{\vartheta_L}(x)\right\} = f_{\vartheta_l}(x)$$

and

$$l = 1 \quad \text{or} \quad \max\left\{f_{\vartheta_1}(x), \ldots, f_{\vartheta_{l-1}}(x)\right\} < f_{\vartheta_l}(x).$$

Our estimate is then defined by

$$\hat{C}_n(x) = \operatorname{sign}(f_n(x)) = \operatorname{sign}(f_{(\hat{\mathbf{w}}, \boldsymbol{\vartheta}^{(t_n)})}(x)) \tag{19}$$

where

$$\hat{\mathbf{w}} = \frac{1}{t_n} \cdot \sum_{t=0}^{t_n - 1} \mathbf{w}^{(t)}. \tag{20}$$

3. Main result

The following assumptions are required for the statistical analysis of the model. First, we require that the underlying template functions fulfill a Lipschitz condition:

Assumption 1. *We have* $\operatorname{supp} f_k \subseteq [0,1]^2$ *for* $k \in \{0,1\}$. *Additionally,* f_0, f_1 *are Lipschitz continuous, in the sense that there exists a positive constant* C_L *such that for any real numbers* u, v, u', v',

$$\left|f_k(u,v) - f_k\left(u', v'\right)\right| \leq C_L \cdot \left(\left|u - u'\right| + \left|v - v'\right|\right), \quad k = 0, 1.$$

In order for our CNN classifier to detect an object within the image, the object has to remain visible, i.e. be fully contained in the image. Therefore we need to rule out deformations that move an object out of the bounds of the image. In other words, the support of the function representing the deformed image $f_k \circ A$ has to remain within $[0,1]^2$, which is ensured by the following assumption.

Assumption 2. *For any* $A = (a_1, a_2) \in \mathcal{A}$ *and for* $k \in \{0,1\}$, *we have* $\operatorname{supp} f_k \circ A \subseteq [0,1]^2$. *Further we assume that the identity is contained in the deformation class* $\mathcal{A}$, *and the partial derivatives of the functions* a_1, a_2 *on* $\mathbb{R}^2$ *exist and are bounded in supremum norm by a constant* $C_{\mathcal{A}}$.

Our third assumption helps us to control the complexity of our image classification problem.

Assumption 3. *The deformation class $\mathcal{A}$ contains a finite subset $\mathcal{A}_d$ such that for any $A \in \mathcal{A}$, there exists an $A' \in \mathcal{A}_d$ and indices $j, \ell \in \{1, \dots, d\}$ such that $A'(\cdot + j/d, \cdot + \ell/d)$ satisfies the conditions imposed in Assumption 2 and*

$$\left\| A'\left(\cdot + \frac{j}{d}, \cdot + \frac{\ell}{d}\right) - A \right\|_\infty \leq \frac{1}{d}.$$

$A_d \subseteq \mathcal{A}$ is obtained by discretizing the deformation class $\mathcal{A}$. If ν denotes the number of free parameters unrelated to image translation, $|\mathcal{A}_d|$ is of order d^ν (cf. Chen, Langer and Schmidt-Hieber (2024), Lemmas 3.3 -3.5). For example in (3), we have $\nu = 3$, since this model uses 3 parameters: the rotation angle γ and the two scaling factors ξ and ξ'.
For an appropriate binary classification task, the two classes we want to distinguish need to be sufficiently dissimilar. To be precise, we need to ensure that deformed functions of two different classes have a positive L^2-distance. The following quantity is introduced to formalize the required minimal separation between the template functions f_0, f_1 of the two classes:

$$\begin{aligned} &D := \max\{D(f_0, f_1), D(f_1, f_0)\} \\ &\text{with} \quad D(f, g) := \frac{\inf_{A, A' \in \mathcal{A},\, a, s, s' \in \mathbb{R}} \|af \circ A(\cdot + s, \cdot + s') - g \circ A'\|_{L^2(\mathbb{R}^2)}}{\|g\|_{L^2(\mathbb{R}^2)}} \end{aligned} \tag{21}$$

Our main result is the following bound on the probability of misclassification of our estimate.

Theorem 1. *Let $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$ be independent and identically distributed random variables with values in $[0,1]^{d \times d} \times \{-1, 1\}$, generated as described above in (4) to (6). Assume that Assumptions 1-3 hold, set $h = d$, $m = |\mathcal{A}_d|$ and*

$$B_n = 1$$

and assume $D^2 > c_2 \cdot d^{-1}$ for c_2 sufficiently large, assume that $K_n \in \mathbb{N}$ satisfies

$$\frac{K_n}{n \cdot \log(n) \cdot e^{c_3 \cdot d^2 \cdot m^2 \cdot (\log m)^3}} \to \infty \quad (n \to \infty)$$

and set

$$\alpha_n = \frac{1}{n \cdot d^4 \cdot e^{2 \cdot c_4 \cdot \log(m)^2}} \quad \text{and} \quad t_n = \left\lceil n^2 \cdot K_n \right\rceil.$$

Define the classifier $\hat{C}_n$ as in Section 2 with $\beta_n = c_5 \cdot \frac{d}{2} \cdot \log(n)$. Assume that the constants c_2, c_3, c_4 are sufficiently large.
There exists a constant $c_6 > 0$ such that we have for n sufficiently large

$$\mathbf{P}\left\{ Y \neq \hat{C}_n(X) \right\} \leq c_6 \cdot \frac{(\log n)^{\frac{3}{2}} \cdot d^2 \cdot \log(d) \cdot m \cdot (\log m)^2}{\sqrt{n}}. \tag{22}$$

Remark 1. *The cardinality of the deformation set $m = |\mathcal{A}_d|$ depends on d and thus the upper bound presented in the theorem above increase with the image resolution d. Despite the fact that higher resolutions produce clearer images and may therefore seem beneficial for classification and convergence, the separation quantity D defined in* (21) *is allowed to shrink as d increases. This suggests that higher resolutions can make deformed images harder to distinguish from one another.*

Remark 2. *In contrast to the present work, Chen, Langer and Schmidt-Hieber (2024) omit the training or optimization routine from their analysis, assuming ideal optimization and focusing on the empirical risk minimizer. Accounting for the training procedure comes at the cost of the rate stated in* (22) *being slower by a factor proportional to the reciprocal of the square root of the sample size, compared to Theorem 4.1 in Chen, Langer and Schmidt-Hieber (2024).*

4. Simulations

In this section we conduct an empirical investigation of the proposed network topology. The estimate $f_{(\hat{\mathbf{w}},\hat{\boldsymbol{\vartheta}}^{(t_n)})}$ described in Section 2 was implemented using Python 3.10 and PyTorch 2.1 (Ansel et al. (2024)). Specifically we used the Conv2d module for the convolutional layer and the Conv1d module with kernel-size parameter set to 1 and the groups-parameter set to K_n for the fully-connected layers. This enables us to compute the K_n parallel networks at the same time.

4.1. Deformation-based datasets

We analyze the empirical performance of the suggested network architecture on artificially generated deformation-based datasets. Here we choose template functions from the MNIST, FashionMNIST, American Sign Language (ASL)[‡] and CIFAR datsets. These datasets contain 28×28 or 32×32 grayscale images which depict handwritten digits, pieces of clothing, hand gestures corresponding to letters in the ASL alphabet or a selection of 100 animals/objects, respectively. Since we consider the task of binary classification, from each dataset we select two classes, resulting in the following four classification tasks:

a.) (MNIST): 0 vs. 4

b.) (FashionMNIST): shirt vs. dress

c.) (ASL): A vs. B

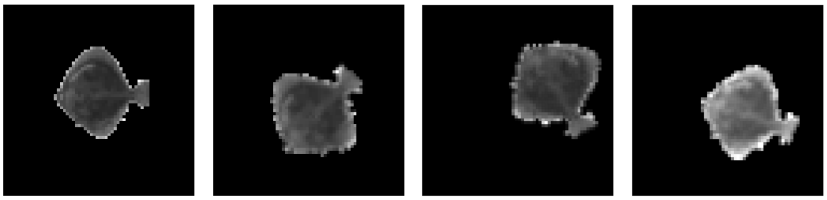
d.) (CIFAR-100): flatfish vs. fox

One image from each class is chosen as data generating template function (f_0 and f_1, respectively).

We generate training samples by applying random shifts (τ, τ'), scalings (ζ, ζ'), brightness changes (η) and rotations (γ) as in (3). Examples of such deformed images are depicted in Figure 1 and 5.

[‡] `https://www.kaggle.com/datasets/signnteam/asl-sign-language-pictures-minus-j-z`

(a) CIFAR class "flatfish"



(b) CIFAR class "fox"

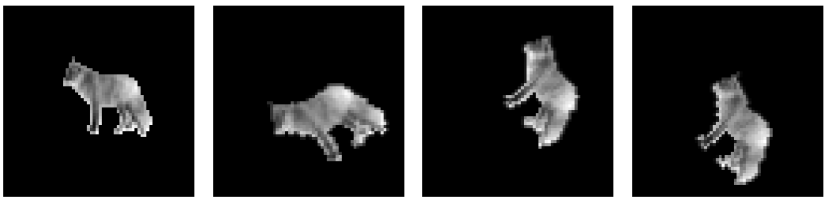


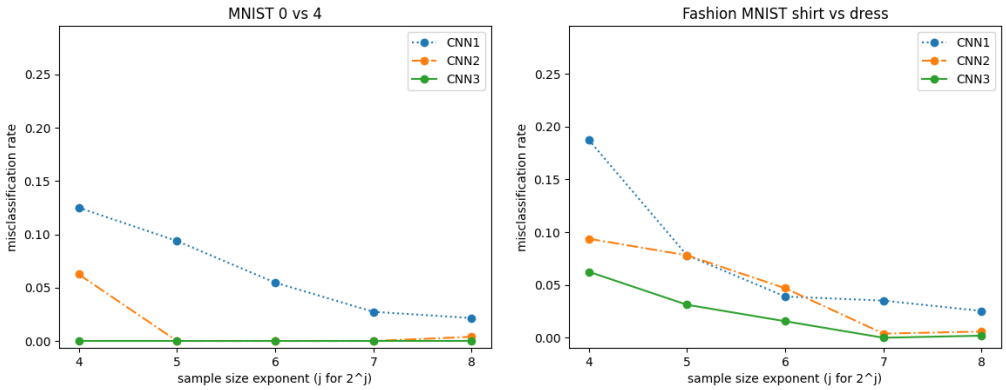
Figure 1: The last three images were generated by applying deformations to the first image of each row.

	K_n	m
CNN1	512	32
CNN2	1024	16
CNN3	2048	8

Table 1: Details of the CNN configurations.

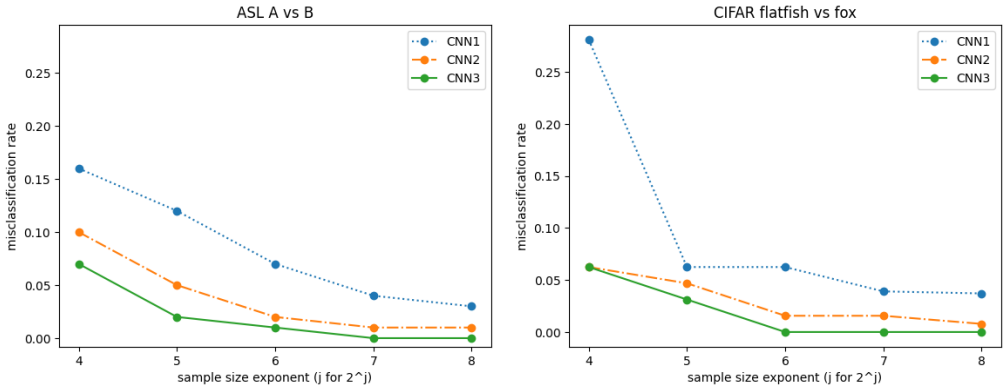
Since the parameter settings of Theorem 1 would result in a very computationally expensive estimate, we will analyze the estimate using smaller parameter values. We consider three different network configurations. For each of them we compute K_n networks in parallel, where each of the K_n sub-networks consists of one convolutional layer, followed by one max-pooling layer and $1 + 2\lceil \log_2(m) \rceil$ fully-connected layers with $4m$ neurons in each layer (cf. Section 2). The parameters K_n and m are chosen as detailed in Table 1. All three networks are trained using 50 epochs of the proposed projected stochastic gradient descent procedure. The training dataset consists of $n \in \{2^4, 2^5, 2^6, 2^7, 2^8\}$ samples, $n/2$ for each class. The validation dataset consists of 100 unseen (balanced) samples. Note that the deformation parameters must be chosen such that the support of the deformed image $f_k \circ A$ remains within the dimensions of the image. In particular, all deformation parameters were chosen uniformly within the following intervals: we have chosen $\tau, \tau' \in [-0.2 \cdot d, 0.2 \cdot d]$, scalings $\zeta, \zeta' \in [0.6, 1]$, brightness changes $\eta \in [0.5, 1.5]$ and rotations $\gamma \in [-60, 60]$.

Figure 2 shows a comparison of the misclassification rates for the three CNN configurations across the four datasets described above, evaluated at sample sizes $n \in \{2^4, 2^5, 2^6, 2^7, 2^8\}$. The graph shows the median misclassification rate over 20 repetitions of each setting. For each repetition at a fixed sample size the classifier was trained for 50 epochs using new randomly initialized parameters $(\mathbf{w}, \boldsymbol{\vartheta})$. The performance of



(a) (MNIST): 0 vs. 4

(b) (FashionMNIST): Shirt vs. Dress



(c) (ASL): A vs. B

(d) (CIFAR-100): Flatfish vs. Fox

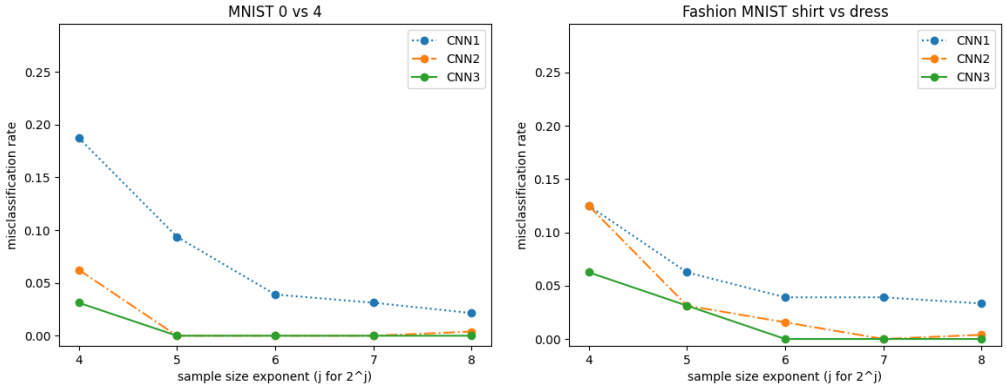
Figure 2: Comparison of the misclassification rate of three CNNs. Reported is the median of the error on the validation dataset with $N = 100$ over 20 repetitions.

the trained classifier was evaluated on the validation dataset containing $N = 100$ unseen samples.

In all four classification tasks, CNN2 and CNN3 significantly outperform CNN1. The difference between CNN2 and CNN3 is less pronounced. However, for sufficiently large sample sizes, CNN3 achieves zero error on the validation dataset in every task, while CNN2 retains an error in the range of $[0.01, 0.02]$.

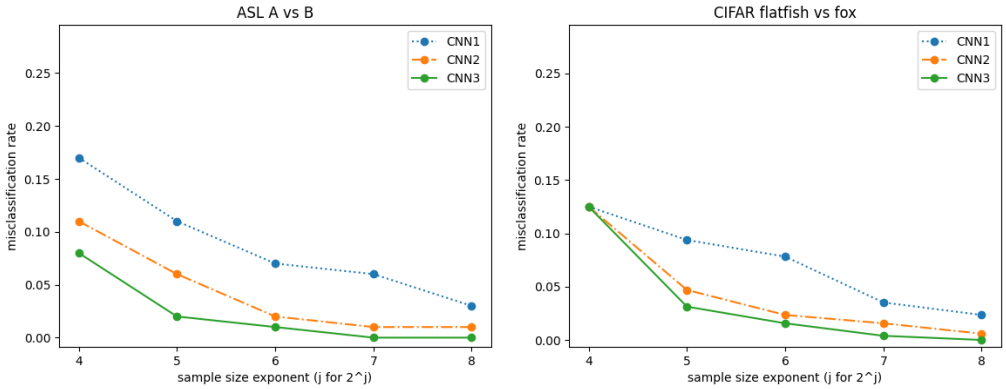
For the best performing CNN3 it looks like the misclassification probability indeed converges to zero for increasing sample size, and the rate of convergence seems to be even faster than in our theoretical result in Theorem 1.

The lower error on the validation dataset associated with larger K_n suggests that the estimator may be performing a form of representation "guessing" rather than actual learning. Through the outer weights $(w_k)_k$ our network can assign higher values to the networks f_{ϑ_k}, which best approximate the envisioned architecture (42).



(a) (MNIST) 0 vs 4

(b) (FashionMNIST) Shirt vs. Dress



(c) (ASL): A vs. B

(d) (CIFAR-100): flatfish vs. fox

Figure 3: Comparison of the misclassification rate of three CNNs, trained over 50 epochs of stochastic gradient descent. During training, the inner weights $(\vartheta_k)_k$ have been frozen. Reported is the median of the error on the validation dataset with $N = 100$ over 20 repetitions. In a) both CNN2 and CNN3 achieve zero misclassification, such that the lines overlap.

4.1.1. Influence of learning the inner weights

The decrease in error on the validation dataset observed for larger values of K_n raises the question of whether the learning of the inner weights $(\vartheta_k)_k$ is truly important. To investigate this, we repeated the experiment from the previous section while holding the inner weights $(\vartheta_k)_k$ fixed at their initial values $((\vartheta_0)_k)_k$. Figure 3 provides a comparison of the misclassification rates for the three CNN configurations across the four datasets in a setting analogous to that considered in Figure 2. The performance of the best performing CNN (CNN3) seems to be a little better for the estimate where the inner weights are learned, but in fact no significant differences are apparent with respect to the decay of the error on the validation dataset.

This finding is consistent with the fact that the projection step in the gradient descent

procedure restricts the extent to which the inner weights $(\vartheta_k^{(t)})_k$ can be modified. In our setting, learning the outer weights $(w_k)_k$ enables the selection of subnetworks with favorable initial parameters while still permitting small adjustments to the inner weights.

4.2. Performance on a dataset with unknown template functions

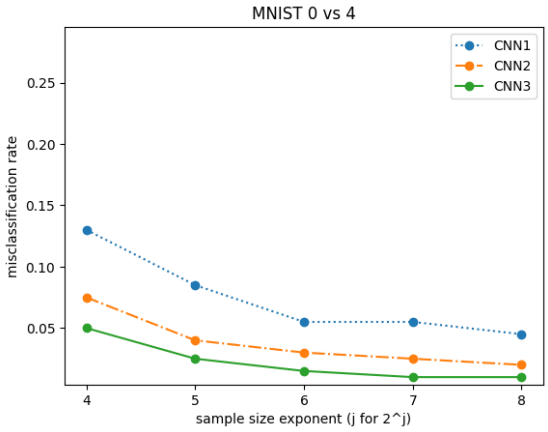


Figure 4: Comparison of the misclassification rate of three CNN architectures. Reported is the median of the error on the validation dataset over 25 repetitions.

In the previous section we considered an augmented dataset consisting of deformations of two template images (discrete observations of the template functions f_0 and f_1). However, in applications, it is not clear whether the dataset really consists of deformed template images. Thus we also want to analyze the proposed architecture on a non-augmented dataset, to see how well our method works in this setting. To do so we trained the CNNs proposed in the previous experiment on the MNIST dataset for 50 epochs and with balanced training sets of size $n \in \{2^4, \dots, 2^8\}$. The MNIST dataset contains images of handwritten digits. For the experiment we have chosen our samples from the classes 0 and 4. Analogously to the experiments in the previous section, CNN2 and CNN3 outperform CNN1 regarding the error on the validation dataset (cf. Figure 4). CNN2 and CNN3 achieve an error on the validation dataset of 0.02 and 0.01 respectively, compared to an error of 0.05 for CNN1. Note that CNN3 already performs well on the smallest dataset. It is plausible that the performance of these networks could be further improved by choosing higher values of K_n closer to the assumptions made in Theorem 1.

5. Proofs

5.1. A general result

In the following, we want to apply an adaptation of Theorem 1 from Kohler, Krzyżak and Sänger (2025). In order to state the theorem, we first introduce a more generalized

setting of estimates that consist of a convex combination of neural networks:
Consider a convolutional network from a given topology

$$f_\vartheta : [0,1]^{d\times d} \to \mathbb{R},$$

which is parametrized by the vector of weights $\vartheta \in \mathbf{\Theta}$, where the set of parameter values $\mathbf{\Theta}$ is closed and convex. In the sequel our objective will be to learn the parameter $\vartheta \in \mathbf{\Theta}$ from data $\mathcal{D}_n$ such that

$$\mathrm{sign}(f_\vartheta(x))$$

is a good classifier. In order to achieve this goal, we consider linear combinations of scaled and truncated convolutional neural networks $f_{\vartheta_k}(x)$ $(k = 1, \dots, K_n)$

$$f_{(\mathbf{w},\boldsymbol{\vartheta})}(x) = \sum_{k=1}^{K_n} w_k \cdot \beta_n \cdot T_1(f_{\vartheta_k}(x)), \tag{23}$$

where the vector of "outer" weights used for the linear combination $\mathbf{w} = (w_k)_{k=1,\dots,K_n}$ satisfies the following three conditions

$$w_k \geq 0 \quad (k = 1, \dots, K_n), \quad \sum_{k=1}^{K_n} w_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{K_n} w_k^2 \leq \alpha_n \tag{24}$$

for some $\alpha_n \in [0,1]$, where $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_{K_n}) \in \mathbf{\Theta}^{K_n}$ and where $\beta_n = c_3 \cdot \log n$. Observe that by choosing $\alpha_n = \frac{1}{N_n}$, $w_j = \frac{1}{N_n}$ $(j = 1, \dots, N_n)$, $\vartheta_j = \vartheta$ $(j = 1, \dots, N_n)$ and $w_k = 0$ for $k > N_n$ we get

$$f_{(\mathbf{w},\boldsymbol{\vartheta})}(x) = \beta_n \cdot T_1(f_\vartheta(x))$$

and in this way we can construct an estimate which satisfies

$$\mathrm{sign}(f_{(\mathbf{w},\boldsymbol{\vartheta})}(x)) = \mathrm{sign}(f_\vartheta(x))$$

for any $\vartheta \in \mathbf{\Theta}$.

Choosing large values for K_n, the number of parameters of the estimate described above will vastly exceed the sample size, i.e. we consider an over-parametrized estimate. The aim is to choose the tuple $(\mathbf{w}, \boldsymbol{\vartheta})$, such that it minimizes the logistic risk

$$F((\mathbf{w}, \boldsymbol{\vartheta})) = \mathbf{E}\left\{\varphi(Y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X))\right\},$$

where φ denotes the logistic loss

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

The logistic risk minimization is performed using the following stochastic gradient descent regime:
The initial values

$$\vartheta_1^{(0)}, \dots, \vartheta_{K_n}^{(0)} \tag{25}$$

are chosen uniformly from some closed and convex set $\mathbf{\Theta}^0 \subseteq \mathbf{\Theta}$ such that the random variables in (25) are independent and also independent from (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n). For the "outer" weights of this topology we use zero initialization, i.e.

$$w_k^{(0)} = 0 \quad (k = 1, \ldots, K_n).$$

Then we set

$$\boldsymbol{\vartheta}^{(0)} = (\vartheta_1^{(0)}, \ldots, \vartheta_{K_n}^{(0)}) \quad \text{and} \quad \mathbf{w}^{(0)} = (w_1^{(0)}, \ldots, w_{K_n}^{(0)})$$

and perform $t_n \in \mathbb{N}$ stochastic gradient descent steps. Here we assume that $t_n/n \in \mathbb{N}$ and for $s \in \{1, \ldots, t_n/n\}$ we let

$$j_{(s-1)\cdot n}, \ldots, j_{s\cdot n-1}$$

be an arbitrary permutation of $1, \ldots, n$. To ensure that the conditions we imposed in (24) remain satisfied, we will use projected stochastic gradient descent. Let $\lambda_n > 0$ be the stepsize and let A denote the set of all $\mathbf{w}$ which satisfy (24). We then set

$$B = \left\{ \boldsymbol{\vartheta} \in \mathbf{\Theta}^{K_n} \, : \, \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^{(0)}\| \leq 1 \right\},$$

to ensure the iterates $\boldsymbol{\vartheta}^{(t)}$ remain within a ball around $\boldsymbol{\vartheta}^{(0)}$ and

$$\begin{aligned} \mathbf{w}^{(t+1)} &= \mathrm{Proj}_A \left(\mathbf{w}^{(t)} - \lambda_n \cdot \nabla_{\mathbf{w}} \varphi \left(Y_{j_t} \cdot f_{(\mathbf{w}^{(t)}, \boldsymbol{\vartheta}^{(t)})}(X_{j_t}) \right) \right), \\ \boldsymbol{\vartheta}^{(t+1)} &= \mathrm{Proj}_B \left(\boldsymbol{\vartheta}^{(t)} - \lambda_n \cdot \nabla_{\boldsymbol{\vartheta}} \varphi \left(Y_{j_t} \cdot f_{(\mathbf{w}^{(t)}, \boldsymbol{\vartheta}^{(t)})}(X_{j_t}) \right) \right) \end{aligned}$$

for $t = 0, \ldots, t_n - 1$. Here Proj_A and Proj_B denote the L_2 projection on A and B. Our estimate is then defined by

$$f_n(x) = f_{(\hat{\mathbf{w}}, \boldsymbol{\vartheta}^{(t_n)})}(x) \tag{26}$$

where

$$\hat{\mathbf{w}} = \frac{1}{t_n} \cdot \sum_{t=0}^{t_n - 1} \mathbf{w}^{(t)}. \tag{27}$$

In this general setting, the next result introduces a bound on the logistic risk of the estimate introduced above.

Theorem 2. *Let (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n) be independent and identically distributed random variables with values in $[0, 1]^{d \times d} \times \{-1, 1\}$. Let $N_n, I_n, t_n \in \mathbb{N}$ and let $C_n, D_n \geq 0$. Set $\beta_n = c_7 \cdot \log n$,*

$$\alpha_n = \frac{1}{N_n}, \quad \lambda_n = \frac{1}{t_n}, \quad K_n \geq N_n \cdot I_n,$$

assume $\beta_n \geq 1$ and define the estimate f_n as above.

Assume that there exists $\tilde{\vartheta} \in \mathbf{\Theta}^0$, such that $f_{\tilde{\vartheta}}(X) = 0$, let $\mathbf{\Theta}^* \subset \mathbf{\Theta}^0$ and set

$$\bar{\mathbf{\Theta}} = \left\{ \vartheta \in \mathbf{\Theta} \, : \, \inf_{\tilde{\vartheta} \in \mathbf{\Theta}^0} \|\vartheta - \tilde{\vartheta}\| \leq 1 \right\}.$$

Assume

$$\|f_\vartheta - f_{\bar{\vartheta}}\|_{\infty,supp(X)} \leq C_n \cdot \|\vartheta - \bar{\vartheta}\|_\infty \quad \textit{for all } \vartheta, \bar{\vartheta} \in \bar{\mathbf{\Theta}}, \tag{28}$$

$$\kappa_n = \mathbf{P}\left\{\vartheta_1^{(0)} \in \mathbf{\Theta}^*\right\} > 0, \tag{29}$$

$$N_n \cdot (1-\kappa_n)^{I_n} \leq \frac{1}{n} \tag{30}$$

and

$$\|\nabla_{\mathbf{w}} \varphi(y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x))\| \leq D_n \tag{31}$$

for all $x \in [0,1]^{d\times d}$, $y \in \{-1,1\}$, $\mathbf{w} \in A$, $\boldsymbol{\vartheta} \in \bar{\mathbf{\Theta}}^{K_n}$.

Then we have

$$\begin{aligned}
&\mathbf{E}\left\{\varphi(Y \cdot f_n(X))\right\} - \min_{f:[0,1]^{d\times d}\to\bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\} \\
&\leq c_8 \cdot \left(\frac{\log n}{n} + \beta_n \cdot \sup_{x_1,\dots,x_n \in [0,1]^{d\times d}} \mathbf{E}\left\{ \left| \sup_{\vartheta\in\bar{\mathbf{\Theta}}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot \beta_n \cdot T_1 f_\vartheta(x_i) \right| \right\} + \beta_n \cdot \frac{C_n+1}{\sqrt{N_n}} + \frac{D_n^2}{t_n} \right. \\
&\quad + \frac{n \cdot \left(6 \cdot K_n \cdot \beta_n^2 + 8 \cdot (\beta_n+1) \cdot C_n \cdot \sup_{\substack{\mathbf{w}\in A, \boldsymbol{\vartheta}\in\bar{\Theta}^{K_n}, \\ y\in\{-1,1\}, x\in[0,1]^{d\times d}}} \|\nabla_{\boldsymbol{\vartheta}} \varphi(y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x))\|_\infty \right)}{t_n} \\
&\quad \left. + \sup_{\vartheta\in\mathbf{\Theta}^*} \mathbf{E}\left\{\varphi(Y \cdot \beta_n \cdot T_1 f_\vartheta(X))\right\} - \min_{f:[0,1]^{d\times d}\to\bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\} \right),
\end{aligned}$$

where $\epsilon_1, \dots, \epsilon_n$ *are independent and uniformly distributed on* $\{-1,1\}$ *(so-called Rademacher random variables).*

Proof. The proof is a modification of the proof of Theorem 1 presented in Kohler, Krzyżak and Sänger (2025). For the sake of completeness we present a complete proof of Theorem 2 in appendix B. □

5.1.1. Using the logistic risk for classification

Our next lemma will help us to derive from the bound on the logistic risk in Theorem 2 a bound on the probability of misclassification of the estimate.

Lemma 1. *Let φ be the logistic loss. Let $(X,Y),(X_1,Y_1),\dots,(X_n,Y_n)$, $\mathcal{D}_n, f_n$ and $\hat{C}_n$ be defined as in Section 2, and set*

$$f_\varphi^* = \arg\min_{f:[0,1]^{d\times d}\to\bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y\cdot f(X))\right\}, \qquad f^* = \arg\min_{f:[0,1]^{d\times d}\to\bar{\mathbb{R}}} \mathbf{P}\{f(X)\neq Y\}.$$

a) *Then*

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

holds.
b) *Assume that*

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

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

$$\mathbf{E}\left\{\varphi(Y\cdot f_\varphi^*(X))\right\}\leq c_9\cdot\tilde{F}_n\cdot e^{-\tilde{F}_n}$$

holds.

Proof.
a) This result follows from Lemma 1 b) Kohler and Langer (2025).
b) This result follows from Lemma 3 in Kim, Ohn and Kim (2021). □

5.2. Bound on the approximation error

In this subsection we study the approximation error of our estimate. The first result describes how the weights of the filter of our CNN classifier would ideally be chosen in the image deformation model.

Lemma 2. *Let f,g be two non-negative functions satisfying Assumption 1 with Lipschitz constant C_L and suppose that Assumptions 2, 3 hold for the deformation set $\mathcal{A}$.*
For any deformation $A\in\mathcal{A}$, let $\overline{\mathbf{X}}_{g\circ A} := \left(\bar{X}_{g\circ A,j,\ell}\right)_{j,\ell=1,\dots,d}$ with

$$\bar{X}_{g\circ A,j,\ell} := \frac{\overline{g\circ A}_{j,\ell}}{d\|g\circ A\|_{2,d}}$$

with $\overline{g\circ A}_{j,\ell}$ being the average intensity of $g\circ A$ on the square $[(j-1)/d, j/d]\times[(\ell-1)/d,\ell/d]$ as defined in (1), and set

$$\mathbf{w}_{(f\circ A)} := \left[\left(w_{(f\circ A)_{j,\ell}}\right)_{j,\ell=1,\dots,d}\right], \quad \text{where} \quad w_{(f\circ A)_{j,\ell}} := \frac{\overline{(f\circ A)}_{j,\ell}}{d\|(f\circ A)\|_{2,d}} \tag{32}$$

and $[W]$ *denotes the quadratic support of a matrix* $W \in \mathbb{R}^{d\times d}$*, i.e. the smallest square sub-matrix that contains all non-zero entries.*

Analogously to (13)*, set*

$$o^{(1)}_{\mathbf{w}_{(f\circ A)}(i,j)} = \sigma\left(\sum_{\substack{r_1,r_2=1,\dots,d:\\ i+r_1-1\le d, j+r_2-1\le d}} w_{(f\circ A)_{r1,r2}}\cdot \bar{x}_{i+r_1-1,j+r_2-1}\right) \quad (A\in\mathcal{A}_d), \tag{33}$$

and

$$o^{(2)}_{\mathbf{w}_{(f\circ A)}} = \max\left\{o^{(1)}_{\mathbf{w}_{(f\circ A)}(i,j)} \,:\, (i,j)\in\{1,\dots,d\}\times\{1,\dots,d\}\right\} \quad (A\in\mathcal{A}_d). \tag{34}$$

Then, there are constants $C_1(C_L, C_{\mathcal{A}}) \ge 0$ *and* $C_2(C_L, C_{\mathcal{A}}) \ge 0$ *such that for* $\mathcal{A}_d \subseteq \mathcal{A}$ *as defined by Assumption 3*

$$\max_{A'\in\mathcal{A}_d} o^{(2)}_{\mathbf{w}_{(g\circ A')}}(\overline{\mathbf{X}}_{g\circ A}) \ge 1 - \frac{C_1(C_L, C_{\mathcal{A}})}{d} \tag{35}$$

and with $D(f,g)$ *defined as in* (21)*, we have*

$$\max_{A'\in\mathcal{A}_d} o^{(2)}_{\mathbf{w}_{(f\circ A')}}(\overline{\mathbf{X}}_{g\circ A}) \le 1 - \frac{\max\{D^2(f,g), D^2(g,f)\}}{16C^2_{\mathcal{A}}C^2_L} + \frac{C_2(C_L, C_{\mathcal{A}})}{d}. \tag{36}$$

Proof. Refer to the proof of Proposition B.5 in Chen, Langer and Schmidt-Hieber (2024). □

Next we study how sensitive our deep convolutional neural network is to pertubations of its weights.

Lemma 3. *Let* $\mathcal{F}_n = \{f_\vartheta : \vartheta\in\mathbf{\Theta}\}$ *be the class of deep convolutional neural networks introduced in Section 2 (cf.*(13)*- (17)).*

Let $\vartheta, \bar{\vartheta}\in\mathbf{\Theta}$ *such that*

$$\left\|\vartheta - \bar{\vartheta}\right\|_\infty \le 1 \tag{37}$$

holds and all weights in f_ϑ *are bounded in absolute value by* $B_n \ge 1$*. Set* $L = 3 + 2\lceil\log_2 m\rceil$*. Then*

$$\|f_\vartheta - f_{\bar{\vartheta}}\|_{\infty,[0,1]^{d\times d}} \le (L-1)\cdot(4m+1)^{L-1}\cdot B_n^{L-1}\cdot h^2\cdot\|\vartheta-\bar{\vartheta}\|_\infty.$$

Proof. Let $o^{(l)}$ and $\bar{o}^{(l)}$ be defined as in Section 2 using the weights in ϑ and $\bar{\vartheta}$. Then $|\sigma(z)|\le|z|$ implies

$$|o^{(2)}_{k,s}(x)| \;\le\; h^2\cdot B_n\cdot 1.$$

Consequently for $r\in\{3,\dots,3+2\lceil\log_2 m\rceil\}$

$$|o^{(r)}_{k,s}(x)| \;\le\; (4m+1)\cdot B_n\cdot\max\{\|o^{(r-1)}(x)\|_\infty, 1\}$$

$$\leq \quad (4m+1)^{r-2} \cdot B_n^{r-2} \cdot h^2 \cdot B_n.$$

Using $|\sigma(z_1) - \sigma(z_2)| \leq |z_1 - z_2|$ we conclude

$$\begin{aligned}
&|o_{k,s}^{(1)}(x) - \bar{o}_{k,s}^{(1)}(x)| \\
&\leq \left| \sum_{\substack{r_1,r_2=1,\ldots,h: \\ i+r_1-1\leq d, j+r_2-1\leq d}} w_{k,s,r_1,r_2} \cdot x_{i+r_1-1,j+r_2-1} - \bar{w}_{k,s,r_1,r_2} \cdot x_{i+r_1-1,j+r_2-1} \right| \\
&\leq \sum_{\substack{r_1,r_2=1,\ldots,h: \\ i+r_1-1\leq d, j+r_2-1\leq d}} |w_{k,s,r_1,r_2} - \bar{w}_{k,s,r_1,r_2}| \cdot |x_{i+r_1-1,j+r_2-1}| \\
&\leq h^2 \cdot \|\vartheta - \bar{\vartheta}\|_\infty.
\end{aligned}$$

Since

$$|\max\{a_1, \ldots, a_n\} - \max\{b_1, \ldots, b_n\}| \leq \max\{|a_1 - b_1|, \ldots, |a_n - b_n|\},$$

it follows that

$$|o_{k,s}^{(2)}(x) - \bar{o}_{k,s}^{(2)}(x)| \leq h^2 \cdot \|\vartheta - \bar{\vartheta}\|_\infty.$$

Next we show

$$\left| o_{k,r}^{(2+s)} - \bar{o}_{k,r}^{(2+s)} \right| \leq s \cdot (4m+1)^s \cdot B_n^s \cdot \|\vartheta - \bar{\vartheta}\|_\infty \cdot h^2$$

for all $s \in \{0, 1, \ldots, 1 + 2\lceil \log_2 m \rceil\}$. Let $s \in \{1, \ldots, 1 + 2\lceil \log_2 m \rceil\}$ and assume that this assertion holds for $s-1$. Then

$$\begin{aligned}
&\left| o_{k,r}^{(2+s)} - \bar{o}_{k,r}^{(2+s)} \right| \\
&= \left| \sigma\left(\sum_{t=1}^{m_s} w_{k,r,t}^{(2+s-1)} \cdot o_{k,t}^{(2+s-1)} + w_{k,r,0}^{(2+s-1)} \right) - \sigma\left(\sum_{t=1}^{m_s} \bar{w}_{k,r,t}^{(2+s-1)} \cdot \bar{o}_{k,t}^{(2+s-1)} + \bar{w}_{k,r,0}^{(2+s-1)} \right) \right| \\
&\leq \left| \sum_{t=1}^{m_s} w_{k,r,t}^{(2+s-1)} \cdot o_{k,t}^{(2+s-1)} + w_{k,r,0}^{(2+s-1)} - \left(\sum_{t=1}^{m_s} \bar{w}_{k,r,t}^{(2+s-1)} \cdot \bar{o}_{k,t}^{(2+s-1)} + \bar{w}_{k,r,0}^{(2+s-1)} \right) \right| \\
&\leq \sum_{t=1}^{m_s} \left| w_{k,r,t}^{(2+s-1)} - \bar{w}_{k,r,t}^{(2+s-1)} \right| \cdot \left| o_{k,t}^{(2+s-1)} \right| + \left| w_{k,r,0}^{(2+s-1)} - \bar{w}_{k,r,0}^{(2+s-1)} \right| \\
&\quad + \sum_{t=1}^{m_s} \left| \bar{w}_{k,r,t}^{(2+s-1)} \right| \cdot \left| o_{k,t}^{(2+s-1)} - \bar{o}_{k,t}^{(2+s-1)} \right| \\
&\leq (4m+1) \cdot \|\vartheta - \bar{\vartheta}\|_\infty \cdot \max\{\|o^{(2+s-1)}(x)\|_\infty, 1\} + 4m \cdot B_n \cdot \left| o_{k,t}^{(2+s-1)} - \bar{o}_{k,t}^{(2+s-1)} \right| \\
&\leq (4m+1) \cdot \|\vartheta - \bar{\vartheta}\|_\infty \cdot (4m+1)^{s-1} \cdot B_n^{s-1} \cdot h^2 \cdot B_n \\
&\quad + 4m \cdot B_n \cdot (s-1) \cdot (4m+1)^{s-1} \cdot B_n^{s-1} \cdot \|\vartheta - \bar{\vartheta}\|_\infty \cdot h^2 \\
&\leq s \cdot (4m+1)^s \cdot B_n^s \cdot \|\vartheta - \bar{\vartheta}\|_\infty \cdot h^2.
\end{aligned}$$

For $f_\vartheta, f_{\bar\vartheta} \in \mathcal{F}$ this yields

$$
\begin{aligned}
&|f_\vartheta(x) - f_{\bar\vartheta}(x)| \\
&\leq \left| \sum_{t=1}^{4m} w_{1,t}^{(3+2\lceil \log_2 m \rceil)} \cdot o_t^{(3+2\lceil \log_2 m \rceil)} - \sum_{t=1}^{4m} \bar w_{1,t}^{(3+2\lceil \log_2 m \rceil)} \cdot \bar o_t^{(3+2\lceil \log_2 m \rceil)} \right| \\
&\leq \sum_{t=1}^{4m} \left| w_{1,t}^{(3+2\lceil \log_2 m \rceil)} - \bar w_{1,t}^{(3+2\lceil \log_2 m \rceil)} \right| \cdot \left| o_t^{(3+2\lceil \log_2 m \rceil)} \right| \\
&+ \sum_{t=1}^{4m} \left| \bar w_{1,t}^{(3+2\lceil \log_2 m \rceil)} \right| \cdot \left| o_t^{(3+2\lceil \log_2 m \rceil)} - \bar o_t^{(3+2\lceil \log_2 m \rceil)} \right| \\
&\leq 4m \cdot \Big(\|\vartheta - \bar\vartheta\|_\infty \cdot (4m+1)^{1+2\lceil \log_2 m \rceil} \cdot B_n^{2+2\lceil \log_2 m \rceil} \cdot h^2 \\
&\qquad\qquad + B_n \cdot \|o^{(3+2\lceil \log_2 m \rceil)} - \bar o^{(3+2\lceil \log_2 m \rceil)}\|_\infty \Big) \\
&\leq \|\vartheta - \bar\vartheta\|_\infty \cdot (4m+1)^{2+2\lceil \log_2 m \rceil} \cdot B_n^{2+2\lceil \log_2 m \rceil} \cdot h^2 \\
&\qquad + B_n \cdot (1 + 2\lceil \log_2 m \rceil) \cdot (4m+1)^{1+2\lceil \log_2 m \rceil} \cdot B_n^{1+2\lceil \log_2 m \rceil} \cdot \|\vartheta - \bar\vartheta\|_\infty \cdot h^2 \\
&\leq (2 + 2\lceil \log_2 m \rceil) \cdot (4m+1)^{2+2\lceil \log_2 m \rceil} \cdot B_n^{2+2\lceil \log_2 m \rceil} \cdot h^2 \cdot \|\vartheta - \bar\vartheta\|_\infty .
\end{aligned}
$$

□

Our next lemma will enable us to compute the maximum of the outputs of the max-pooling layer of our convolutional neural network via a feed-forward neural network.

Lemma 4. *There exists a network $f_{\max^r}$ with $1 + 2\lceil \log_2 r \rceil$ layers and $2r$ neurons, such that $f_{\max^r} \in [0,1]$,*

$$f_{\max^r}(\mathbf{x}) = \max\{x_1, \ldots, x_r\} \quad \text{for all} \quad \mathbf{x} = (x_1, \ldots, x_r) \in [0,1]^r$$

and all network parameters are bounded in absolute value by 1.

Proof. Refer to the proof of Lemma B.6 in Chen, Langer and Schmidt-Hieber (2024). □

5.3. Generalization error

In order to bound the generalization error of our estimate the following result will be useful.

Lemma 5. *Let $\{f_\vartheta : \vartheta \in \mathbf{\Theta}\}$ be defined as in Section 2 and let $\beta_n = c_1 \cdot \log n$. Then we have*

$$
\begin{aligned}
&\sup_{x_1,\ldots,x_n \in [0,1]^{d\times d}} \mathbf{E}\left\{ \left| \sup_{\vartheta \in \bar{\mathbf{\Theta}}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot \beta_n \cdot T_1 f_\vartheta(x_i) \right| \right\} \\
&\leq c_{10} \cdot \sqrt{\log(n) + \log(\beta_n)} \cdot m \cdot h \cdot L_{max} \cdot \frac{\sqrt{\log(m^2 \cdot h^2 \cdot L_{max})}}{\sqrt{n}}
\end{aligned}
$$

where $L_{max} = 1 + 2\lceil \log m \rceil$.

For the proof of Lemma 5 we need the following bound on the VC-dimension:

Lemma 6. *Let $\mathcal{F} := \{f_\vartheta : \vartheta \in \mathbf{\Theta}\}$ be the class of functions of Lemma 5 and set $L_{max} = 1 + 2\lceil \log m \rceil$. The VC-dimension of $\mathcal{F}^+$ is bounded from above by*

$$V_{\mathcal{F}^+} \leq c_{11} \cdot m^2 \cdot h^2 \cdot L_{max}^2 \cdot \log(m^2 \cdot d^2 \cdot L_{max}).$$

Proof. The result follows from the proof of Lemma 11 in Kohler, Krzyżak and Walter (2022) with $L_1 = 1$, $L_2 = 1 + 2\lceil \log_2(m) \rceil$, $k_{\max} = 4m$, $t = 1$, $d_1 = d_2 = d$ and $M_{\max} = h$. □

Proof of Lemma 5. The proof is analogous to the proof of Lemma 9 in Kohler, Krzyżak and Sänger (2025) and follows from Lemma 6 by applying standard techniques from VC theory. For the sake of completeness we present a detailed proof in appendix C. □

5.4. A bound on the gradient

The following bound on the gradient of our CNN classifier constitutes a preliminary step towards the proof of Theorem 1.

Lemma 7. *Let A, $\mathbf{\Theta}^0$ and $f_{(\mathbf{w},\boldsymbol{\vartheta})}$ be defined as in Section 2. Set*

$$\bar{\mathbf{\Theta}} = \left\{ \vartheta \in \mathbf{\Theta} \ : \ \inf_{\tilde{\vartheta} \in \mathbf{\Theta}^0} \|\vartheta - \tilde{\vartheta}\| \leq 1 \right\},$$

then

$$\sup_{\mathbf{w} \in A, \boldsymbol{\vartheta} \in \bar{\Theta}^{K_n}, y \in \{-1,1\}, x \in [0,1]^{d \times d}} \|\nabla_{\boldsymbol{\vartheta}} \varphi(y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x))\|_\infty \leq h^2 \cdot (B_n + 1)^{2L-2} \cdot (4m+1)^{2L-2},$$

where $L = 3 + 2\log\lceil m \rceil$.

Proof. The proof is a modification of Lemma 11 in Kohler, Krzyżak and Sänger (2025) with $M_{max} = M_1 = h$ and $L_n^{(1)} = 1$.

We have

$$f_{(\mathbf{w},\boldsymbol{\vartheta})}(x) = \sum_{k=1}^{K_n} w_k \cdot \beta_n \cdot T_1 f_{\vartheta_k}(x)$$

where

$$f_{\vartheta_k}(x) = \sum_{t=1}^{4m} w_{k,1,t}^{(2+1+2\lceil \log_2 m \rceil)} \cdot o_{k,t}^{(2+1+2\lceil \log_2 m \rceil)}$$

and

$$o_{k,r}^{(2+s)} = \sigma \left(\sum_{t=1}^{4m} w_{k,r,t}^{(2+s-1)} \cdot o_{k,t}^{(2+s-1)} + w_{k,r,0}^{(2+s-1)} \right)$$

for $s = 2, \ldots, 1 + 2\lceil \log_2 m \rceil, r = 1, \ldots, 4m$ and

$$o_{k,r}^{(3)} = \sigma\left(\sum_{t=1}^{2m} w_{k,r,t}^{(2)} \cdot o_{k,t}^{(2)} + w_{k,r,0}^{(2)}\right)$$

for $r = 1, \ldots, 4m$.

The input of the fully-connected layers, $o_{k,t}^{(2)}$, is obtained through the application of channel-wise max-pooling to the outputs of the convolutional layer:

$$o_{k,s}^{(2)} = \max\left\{o_{(i,j),s}^{(1)} \,:\, (i,j) \in \{1,\ldots,d\} \times \{1,\ldots,d\}\right\} \quad (s = 1, \ldots, 2m),$$

where

$$o_{(i,j),k,s}^{(1)} = \sigma\left(\sum_{\substack{r_1,r_2=1,\ldots,h:\\ i+r_1-1\leq d, j+r_2-1\leq d}} w_{k,s,r_1,r_2} \cdot \bar{x}_{i+r_1-1,j+r_2-1}\right) \quad (s = 1, \ldots, 2m).$$

Since $|\sigma(z)| \leq |z|$, we get

$$|o_{k,s}^{(2)}(x)| \leq h^2 \cdot (B_n + 1). \tag{38}$$

Set $L = 3 + 2\lceil \log_2 m \rceil$, let $\tilde{l} \in \{3, \ldots, 2 + 2\lceil \log_2 m \rceil\}$ and choose $\tilde{r}, \tilde{t} \in \{1, \ldots, 4m\}$. With the chain rule we get

$$\begin{aligned}
&\frac{\partial}{\partial w_{k,\tilde{r},\tilde{t}}^{(\tilde{l})}} f_{(\mathbf{w},\boldsymbol{\vartheta})}(x) \\
&= w_k \cdot \beta_n \cdot \frac{\partial}{\partial z} T_1 z\Big|_{z=f_{\vartheta_k}(x)} \cdot \sum_{t^{(L)}=1}^{4m} w_{k,1,t^{(L)}}^{(L)} \cdot \sigma'\left(\sum_{t=1}^{4m} w_{k,t^{(L)},t}^{(L-1)} \cdot o_{k,t}^{(L-1)} + w_{k,t^{(L)},0}^{(L-1)}\right) \\
&\cdot \sum_{t^{(L-1)}=1}^{4m} w_{k,t^{(L)},t^{(L-1)}}^{(L-1)} \cdot \sigma'\left(\sum_{t=1}^{4m} w_{k,t^{(L-1)},t}^{(L-2)} \cdot o_{k,t}^{(L-2)} + w_{k,t^{(L-1)},0}^{(L-2)}\right) \\
&\cdot \sum_{t^{(L-2)}=1}^{4m} w_{k,t^{(L-1)},t^{(L-2)}}^{(L-2)} \cdot \ldots \cdot \sum_{t^{(\tilde{l}+1)}=1}^{4m} w_{k,t^{(\tilde{l}+1)},\tilde{r}}^{(\tilde{l}+1)} \cdot \sigma'\left(\sum_{t=1}^{4m} w_{k,\tilde{r},t}^{(\tilde{l})} \cdot o_{k,t}^{(\tilde{l})} + w_{k,\tilde{r},0}^{(\tilde{l})}\right) \cdot o_{k,\tilde{t}}^{(\tilde{l})}.
\end{aligned}$$

Using (38) and that all weights are bounded in the absolute value by $B_n + 1$ we get

$$\left|o_{k,\tilde{t}}^{(\tilde{l})}\right| \leq (4m+1)^{\tilde{l}-2} \cdot (B_n+1)^{\tilde{l}-1} \cdot h^2$$

(cf., proof of Lemma 3) and using $|\sigma'(z)| \leq 1$, we conclude

$$\left|\frac{\partial}{\partial w_{k,r,t}^{(l)}} f_{(\mathbf{w},\boldsymbol{\vartheta})}(x)\right|$$

$$
\begin{aligned}
&\leq (B_n+1)\cdot\beta_n\cdot\underbrace{4m\cdot(B_n+1)\cdot 1\cdot\ldots\cdot 4m\cdot(B_n+1)\cdot 1}_{L-\tilde{l}\text{-times}}\cdot o_{k,\tilde{t}}^{(\tilde{l})}\\
&\leq (B_n+1)\cdot\beta_n\cdot(4m)^{L-\tilde{l}}\cdot(B_n+1)^{L-\tilde{l}}\cdot(4m+1)^{\tilde{l}-2}\cdot(B_n+1)^{\tilde{l}-1}\cdot h^2\\
&\leq \beta_n\cdot h^2\cdot(B_n+1)^{L}\cdot(4m+1)^{L-2}.
\end{aligned}
$$

For the weights of the convolutional layers we get for suitably chosen (random) $(i_0,j_0)\in\{1,\ldots,d\}\times\{1,\ldots,d\}$

$$
\begin{aligned}
&\frac{\partial}{\partial w_{k,\tilde{s},\tilde{r}_1,\tilde{r}_2}} f_{(\mathbf{w},\boldsymbol{\vartheta})}(x)\\
&= w_k\cdot\beta_n\cdot\frac{\partial}{\partial z}T_1 z\Big|_{z=f_{\vartheta_k}(x)}\cdot\sum_{t^{(L_n)}=1}^{4m} w_{k,1,t^{(L_n)}}^{(L_n)}\cdot\sigma'\left(\sum_{t=1}^{4m} w_{k,t^{(L_n)},t}^{(L_n-1)}\cdot o_{k,t}^{(L_n-1)}+w_{k,t^{(L_n)},0}^{(L_n-1)}\right)\\
&\cdot\sum_{t^{(L_n-1)}=1}^{4m} w_{k,t^{(L_n)},t^{(L_n-1)}}^{(L_n-1)}\cdot\sigma'\left(\sum_{t=1}^{4m} w_{k,t^{(L_n-1)},t}^{(L_n-2)}\cdot o_{k,t}^{(L_n-2)}+w_{k,t^{(L_n-1)},0}^{(L_n-2)}\right)\\
&\cdot\sum_{t^{(L_n-2)}=1}^{4m} w_{k,t^{(L_n-1)},t^{(L_n-2)}}^{(L_n-2)}\cdot\ldots\\
&\qquad\qquad\ldots\cdot\sigma'\left(\sum_{t=1}^{2m} w_{k,t^{(3)},t}^{(2)}\cdot o_{k,t}^{(2)}+w_{k,t^{(3)},0}^{(2)}\right)\\
&\cdot w_{k,t^{(3)},\tilde{s}}^{(2)}\cdot\sigma'\left(\sum_{\substack{r_1,r_2=1,\ldots,h:\\ i+r_1-1\leq d,j+r_2-1\leq d}} w_{k,\tilde{s},r_1,r_2}\cdot\bar{x}_{i+r_1-1,j+r_2-1}\right)\cdot\bar{x}_{i_0+\tilde{r}_1-1,j_0+\tilde{r}_2-1}
\end{aligned}
$$

Using $|\sigma'(z)|\leq 1$ and that all weights are bounded in the absolute value by B_n+1 we get

$$
\left|\frac{\partial}{\partial w_{k,s,r_1,r_2}} f_{(\mathbf{w},\boldsymbol{\vartheta})}(x)\right|\leq\beta_n\cdot(4m)^{L-2}\cdot(B_n+1)^{L-1}
$$

Analogously we can derive bounds on all the other partial derivatives occurring in the assertion (i.e. with respect to the weights w_k, the bias weights $w_{k,r,0}^{(l)}$ of the fully connected layers and the weights $w_{k,\tilde{r},\tilde{t}}^{(2)}$). □

5.4.1. Proof of Theorem 1

In the sequel we show

$$
\mathbf{E}\left\{\varphi(Y\cdot f_n(X))\right\}-\min_{f:[0,1]^{d\times d}\to\bar{\mathbb{R}}}\mathbf{E}\left\{\varphi(Y\cdot f(X))\right\}\leq c_{12}\cdot\frac{(\log n)^{\frac{3}{2}}\cdot d\cdot\log(d)\cdot m\cdot(\log m)^2}{\sqrt{n}}.. \tag{39}
$$

for n sufficiently large. This implies the assertion since by Lemma 1 a) and Lemma 1 b) we conclude from (39)

$$\begin{aligned}
&\mathbf{P}\left\{Y \neq \hat{C}_n(X)\right\} - \mathbf{P}\left\{Y \neq f^*(X)\right\} \\
&\leq 2 \cdot \left(\mathbf{E}\left\{\varphi(Y \cdot f_n(X))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_{\varphi^*}(X))\right\}\right) + 4 \cdot \frac{c_{13} \cdot \log n}{n^{1/2}} \\
&\leq c_{14} \cdot \frac{(\log n)^{\frac{3}{2}} \cdot d \cdot \log(d) \cdot m \cdot (\log m)^2}{\sqrt{n}}.
\end{aligned} \tag{40}$$

Here we have used the fact that

$$\max\left\{\frac{\mathbf{P}\{Y=1|X\}}{1-\mathbf{P}\{Y=1|X\}}, \frac{1-\mathbf{P}\{Y=1|X\}}{\mathbf{P}\{Y=1|X\}}\right\} > n^{1/2} \tag{41}$$

is equivalent to

$$|f_{\varphi^*}(X)| = \left|\log \frac{\mathbf{P}\{Y=1|X\}}{1-\mathbf{P}\{Y=1|X\}}\right| > \frac{1}{2} \cdot \log n$$

and that

$$\mathbf{P}\left\{\max\left\{\frac{\mathbf{P}\{Y=1|X\}}{1-\mathbf{P}\{Y=1|X\}}, \frac{1-\mathbf{P}\{Y=1|X\}}{\mathbf{P}\{Y=1|X\}}\right\} > n^{1/2}\right\} = 1$$

holds since

$$\begin{aligned}
\max\left\{\frac{\mathbf{P}\{Y=1|X\}}{1-\mathbf{P}\{Y=1|X\}}, \frac{1-\mathbf{P}\{Y=1|X\}}{\mathbf{P}\{Y=1|X\}}\right\} &= \max\left\{\frac{\mathbf{P}\{Y=1|X\}}{\mathbf{P}\{Y=0|X\}}, \frac{\mathbf{P}\{Y=0|X\}}{\mathbf{P}\{Y=1|X\}}\right\} \\
&= \max\left\{\frac{\mathbf{E}\left\{\mathbb{1}_{\{Y=1\}}|X\right\}}{\mathbf{E}\left\{\mathbb{1}_{\{Y=0\}}|X\right\}}, \frac{\mathbf{E}\left\{\mathbb{1}_{\{Y=0\}}|X\right\}}{\mathbf{E}\left\{\mathbb{1}_{\{Y=1\}}|X\right\}}\right\} \\
&= \max\left\{\frac{\mathbb{1}_{\{Y=1\}}}{\mathbb{1}_{\{Y=0\}}}, \frac{\mathbb{1}_{\{Y=0\}}}{\mathbb{1}_{\{Y=1\}}}\right\} = \infty,
\end{aligned}$$

where the second last equality holds since in our setting Y is a function of X and thus the random variable $\mathbb{1}_{\{Y=k\}}$ is measurable w.r.t. X.
Theorem 4.2 of Chen, Langer and Schmidt-Hieber (2024) states that under a subset of the assumptions of Theorem 1 a convolutional neural network classifier that achieves perfect classification, i.e. that recovers the correct label almost surely, can be constructed. Thus there exists $f^* : [0,1]^{d\times d} \to \{-1,1\}$ with $\mathbf{P}\{Y \neq f^*(X)\} = 0$, and (40) implies the assertion.

We use Lemma 2 to estimate this optimal network. Let $\hat{\mathbf{w}}_{(f\circ A)}$ denote the $d\times d$ matrix, which is obtained by extending $\mathbf{w}_{(f\circ A)}$ (cf. (32)) to a $d \times d$ matrix with zero-padding to the right and to the bottom. The idea is to aim for the following network architecture using two networks from Lemma 4 in parallel

$$f_{\vartheta^*}(X) = f_{\max^m}\left(\left(o^{(2)}_{\hat{\mathbf{w}}_{(f_1\circ A')}}(X)\right)_{A'\in\mathcal{A}_d}\right) - f_{\max^m}\left(\left(o^{(2)}_{\hat{\mathbf{w}}_{(f_0\circ A')}}(X)\right)_{A'\in\mathcal{A}_d}\right), \tag{42}$$

where $\left(o^{(2)}_{\hat{\mathbf{w}}_{(f_k\circ A')}}(X)\right)_{A'\in\mathcal{A}_d}$ for $k\in\{0,1\}$ is the $m=|\mathcal{A}_d|$-dimensional output of the max-pooling layer applied to the convolutional layer, as defined in (34).

Note that in (42) we can obtain $-f_{\max^m}(x_1,\dots,x_m)$ by multiplying the outer weights of the network introduced in Lemma 4 with -1.

Assume $D^2 > \frac{1}{d}16C_{\mathcal{A}}^2C_L^2(C_1(C_L,C_{\mathcal{A}})+C_2(C_L,C_{\mathcal{A}})+2)$ where $C_1(C_L,C_{\mathcal{A}})\geq 0$ and $C_2(C_L,C_{\mathcal{A}})\geq 0$ are the constants from Lemma 2. As a first case, assume that f_0 is the template function corresponding to the given image, i.e. consider

$$((\bar{X}_{f_0\circ A})_{j,\ell=1,\dots,d}) := \frac{\overline{f_0\circ A}_{j,\ell}}{d\|f_0\circ A\|_{2,d}},$$

with $A\in\mathcal{A}$ denoting the "true" deformation that generated the image. By assumption, the conditions of Lemma 2 are satisfied and we conclude that we have

$$\begin{aligned}
&f_{\max^m}\left(\left(o^{(2)}_{\hat{\mathbf{w}}_{(f_1\circ A')}}(\bar{X}_{f_0\circ A})\right)_{A'\in\mathcal{A}_d}\right) - f_{\max^m}\left(\left(o^{(2)}_{\hat{\mathbf{w}}_{(f_0\circ A')}}(\bar{X}_{f_0\circ A})\right)_{A'\in\mathcal{A}_d}\right)\\
&\leq 1-\frac{D^2}{16C_{\mathcal{A}}^2C_L^2}+\frac{C_2(C_L,C_{\mathcal{A}})}{d}-\left(1-\frac{C_1(C_L,C_{\mathcal{A}})}{d}\right)\\
&\leq -\frac{2}{d}. \qquad (43)
\end{aligned}$$

Analogously one can show that

$$f_{\max^m}\left(\left(o^{(2)}_{\hat{\mathbf{w}}_{(f_1\circ A')}}(\bar{X}_{f_1\circ A})\right)_{A'\in\mathcal{A}_d}\right) - f_{\max^m}\left(\left(o^{(2)}_{\hat{\mathbf{w}}_{(f_0\circ A')}}(\bar{X}_{f_1\circ A})\right)_{A'\in\mathcal{A}_d}\right) > \frac{2}{d}. \qquad (44)$$

Set

$$\delta_n = \frac{1}{d^3\cdot e^{c_{15}\cdot(\log m)^2}}$$

and choose

$$\Theta^* = \{\vartheta\in\Theta : \|\vartheta^*-\vartheta\|_\infty\leq\delta_n\}.$$

For rest of the proof we apply Theorem 2 with Θ^* as defined above and

$$\mathbf{\Theta}^0 = \{\vartheta\in\mathbf{\Theta} \quad : \quad \|\vartheta\|_\infty\leq B_n\}.$$

First we bound the last term on the right hand side of the bound in Theorem 2 and show

$$\sup_{\vartheta\in\mathbf{\Theta}^*}\mathbf{E}\{\varphi(Y\cdot\beta_n\cdot T_1 f_\vartheta(X))\} - \min_{f:[0,1]^{d\times d}\to\bar{\mathbb{R}}}\mathbf{E}\{\varphi(Y\cdot f(X))\}\leq\frac{1}{\sqrt{n}}.$$

For $\bar{\vartheta}\in\Theta^*$ we can use Lemma 3 to bound the maximum distance of our estimator $f_{\bar{\vartheta}}$ from f_{ϑ^*}:

$$\begin{aligned}
|f_{\bar{\vartheta}}(X)-f_{\vartheta^*}(X)| &\leq (L-1)\cdot(4m+1)^{L-1}\cdot B_n^{L-1}\cdot d^2\cdot\|\bar{\vartheta}-\vartheta^*\|_\infty\\
&\leq d^2\cdot e^{c_{16}\cdot(\log m)^2}\cdot\|\bar{\vartheta}-\vartheta^*\|_\infty
\end{aligned}$$

$$\leq \frac{1}{d}$$

provided $c_{15} \geq c_{16}$. Using this, (43) and (44) we get for any $A \in \mathcal{A}$

$$f_{\bar{\vartheta}}(\bar{X}_{f_1 \circ A}) > \frac{1}{d} \text{ and } f_{\bar{\vartheta}}(\bar{X}_{f_0 \circ A}) < -\frac{1}{d},$$

i.e. $\operatorname{sign}(f_\vartheta(X))$ recovers the correct label and we have $Y \cdot T_1 f_\vartheta(X) > \frac{1}{d}$. For $c_5 > 1$ this yields

$$\sup_{\vartheta \in \mathbf{\Theta}^*} \mathbf{E}\left\{\varphi(Y \cdot \beta_n \cdot T_1 f_\vartheta(X))\right\} \leq \log\left(1 + e^{-\frac{c_5 \cdot \log(n)}{d}}\right) \leq \frac{1}{\sqrt{n}}.$$

To bound κ_n, we need a bound on the total number of weights W of a network f_ϑ. Let W_r denote the number of weights up to layer r of our network. For the convolutional and the following max-pooling layer we have

$$W_2 = W_1 = h^2 \cdot 2m.$$

For the fully-connected sub-network we have

$$W_3 = W_2 + (2m+1) \cdot 4m,$$

and

$$W_t = W_{t-1} + (4m+1) \cdot 4m$$

for $3 < t \leq 3 + 2\lceil \log_2(m) \rceil$. Thus

$$W := W_{3+2\lceil \log_2(m) \rceil} + 4m = h^2 \cdot 2m + (2m+1) \cdot 4m + 4m \cdot (4m+1) \cdot 2\lceil \log_2(m) \rceil + 4m.$$

The definition of $\mathbf{\Theta}^*$ implies that

$$\kappa_n = \mathbf{P}\left\{\vartheta_1^{(0)} \in \mathbf{\Theta}^*\right\} \geq \left(\frac{1}{2 \cdot d^3 \cdot e^{c_{15} \cdot \log(m)^2}}\right)^W \geq e^{-c_{17} \cdot d^2 \cdot m^2 \cdot (\log m)^3} \geq 0$$

So if we choose

$$N_n = n \cdot d^4 \cdot e^{2 \cdot c_{16} \cdot \log(m)^2}, \quad I_n = (2 \cdot \log(n) + 2 \cdot c_{16} \cdot (\log m)^2 + 4\log(d)) \cdot e^{c_{17} \cdot d^2 \cdot m^2 \cdot (\log m)^3}$$

using $(1-z)^n \leq e^{-z}$ for $z \in \mathbb{R}$, we have

$$\begin{aligned}
&N_n \cdot (1-\kappa_n)^{I_n} \\
&\leq N_n \cdot \exp(-\kappa_n \cdot I_n) \\
&\leq n \cdot d^4 \cdot e^{2 \cdot c_{16} \cdot \log(m)^2} \cdot \exp(-(2 \cdot \log(n) + 2 \cdot c_{16} \cdot (\log m)^2 + 4\log(d))) \\
&\leq n \cdot d^4 \cdot e^{2 \cdot c_{16} \cdot \log(m)^2} \cdot \frac{1}{n^2} \cdot d^{-4} \cdot e^{-2 \cdot c_{16} \cdot \log(m)^2} \\
&\leq \frac{1}{n},
\end{aligned}$$

which is possible because

$$\begin{aligned}
&I_n \cdot N_n \\
&= n \cdot d^4 \cdot e^{2 \cdot c_{16} \cdot \log(m)^2} \cdot (2 \cdot \log(n) + 2 \cdot c_{16} \cdot (\log m)^2 + 4 \log(d)) \cdot e^{c_{17} \cdot d^2 \cdot m^2 \cdot (\log m)^3} \\
&\leq n \cdot \log(n) \cdot e^{c_{18} \cdot d^2 \cdot m^2 \cdot (\log m)^3} \leq K_n.
\end{aligned}$$

An application of Lemma 3 with $B_n = 1$ and $h = d$ yields that (28) is satisfied for

$$\begin{aligned}
C_n \quad &= (L-1) \cdot (4m+1)^{L-1} \cdot B_n^{L-1} \cdot h^2 \\
&\leq d^2 \cdot (2 + 2\lceil \log_2 m \rceil) \cdot (4m+1)^{2+2\lceil \log_2 m \rceil} \\
&\leq d^2 \cdot e^{c_{16} \cdot (\log m)^2},
\end{aligned}$$

and because of

$$\|\nabla_{\mathbf{w}} \varphi(y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x))\|^2 \leq \sum_{k=1}^{K_n} |1 \cdot \beta_n \cdot T_1 f_{\vartheta_k}(x)|^2 \leq K_n \cdot \beta_n^2$$

(31) is satisfied for

$$D_n = \sqrt{K}_n \cdot \beta_n.$$

By Lemma 7 we know

$$\begin{aligned}
\sup_{\mathbf{w} \in A, \boldsymbol{\vartheta} \in \bar{\boldsymbol{\Theta}}^{K_n}, y \in \{-1,1\}, x \in [0,1]^{d \times d}} \|\nabla_{\boldsymbol{\vartheta}} \varphi(y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x))\|_\infty \quad &\leq d^2 \cdot (B_n + 1)^{2L-2} \cdot (4m+1)^{2L-2}. \\
&\leq d^2 \cdot e^{c_{19} \cdot (\log m)^2}
\end{aligned}$$

and with $L = 3 + 2\lceil \log_2 m \rceil$ Lemma 5 yields the following upper bound for the Rademacher complexity

$$\begin{aligned}
&\sup_{x_1, \ldots, x_n \in [0,1]^{d \times d}} \mathbf{E} \left\{ \left| \sup_{\vartheta \in \bar{\boldsymbol{\Theta}}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot \beta_n \cdot T_1 f_\vartheta(x_i) \right| \right\} \\
&\leq c_{20} \cdot \sqrt{\log(n) + \log(\beta_n)} \cdot m \cdot d \cdot L \cdot \frac{\sqrt{\log(m^2 \cdot d^2 \cdot L)}}{\sqrt{n}} \\
&\leq c_{21} \cdot \frac{\sqrt{\log(n)}}{\sqrt{n}} \cdot m \cdot d \cdot \sqrt{\log(d)} \cdot (3 + 2\lceil \log_2 m \rceil) \cdot \sqrt{\log(m^2 \cdot d^2 \cdot (3 + 2\lceil \log_2 m \rceil))} \\
&\leq c_{22} \cdot \frac{\sqrt{\log(n)}}{\sqrt{n}} \cdot m \cdot (\log m)^2 \cdot d \cdot \log(d).
\end{aligned}$$

Application of Theorem 2 together with the above results yields

$$\mathbf{E}\left\{\varphi(Y \cdot f_n(X))\right\} - \min_{f:[0,1]^{d \times d} \to \bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\}$$

$$
\begin{aligned}
&\leq c_4 \cdot \left(\frac{\log n}{n} + \beta_n \cdot \sup_{x_1,\ldots,x_n \in [0,1]^{d\times d}} \mathbf{E}\left\{ \left| \sup_{\vartheta \in \tilde{\mathbf{\Theta}}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot \beta_n \cdot T_1 f_\vartheta(x_i) \right| \right\} + \beta_n \cdot \frac{C_n + 1}{\sqrt{N_n}} + \frac{D_n^2}{t_n} \right. \\
&\quad + \frac{n \cdot \left(6 \cdot K_n \cdot \beta_n^2 + 8 \cdot (\beta_n + 1) \cdot C_n \cdot \sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta}) \in \mathcal{W}, y \in \{-1,1\}, \\ x \in [0,1]^{d\times d}}} \|\nabla_{\boldsymbol{\vartheta}} \varphi(y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x))\|_\infty \right)}{t_n} \\
&\quad \left. + \sup_{\vartheta \in \mathbf{\Theta}^*} \mathbf{E}\left\{ \varphi(Y \cdot \beta_n \cdot T_1 f_\vartheta(X)) \right\} - \min_{f:[0,1]^{d\times d} \to \bar{\mathbb{R}}} \mathbf{E}\left\{ \varphi(Y \cdot f(X)) \right\} \right), \\
&\leq c_{23} \cdot \left(\frac{\log(n)}{n} + c_{24} \cdot \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}} \cdot m \cdot (\log m)^2 \cdot d^2 \cdot \log(d) + \frac{\beta_n}{\sqrt{n}} + \frac{K_n \cdot \beta_n^2}{K_n \cdot n^2} \right. \\
&\quad \left. + \frac{6 \cdot \beta_n^2}{n} + \frac{n \cdot \left(8 \cdot (\beta_n + 1) \cdot (d^2 \cdot e^{c_{16} \cdot (\log m)^2}) \cdot (d^2 \cdot e^{c_{25} \cdot (\log m)^2}) \right)}{n^3 \cdot \log(n) \cdot e^{c_{26} \cdot d^2 \cdot m^2 \cdot (\log m)^3}} + \frac{1}{\sqrt{n}} \right) \\
&\leq c_{27} \cdot \frac{(\log n)^{\frac{3}{2}} \cdot d^2 \cdot \log(d) \cdot m \cdot (\log m)^2}{\sqrt{n}}.
\end{aligned}
$$

□

References

Arora, S., Cohen, N., Golowich, N., Hu, W. (2019). A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks. *International Conference on Learning Representations.*

Arora, S., Ge, R., Neyshabur, B., Zhang, Y.. (2018). Stronger Generalization Bounds for Deep Nets via a Compression Approach. *Proceedings of the 35th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 80:254-263

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M. Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., & Chintala, S. (2024). PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24).* https://doi.org/10.1145/3620665.3640366

Bartlett, P. L., Foster, D. J., Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks. Proceedings of the 31st International *Conference on Neural Information Processing Systems*, 6241–6250.

Bottou, L., Curtis, F. E., Nocedal, J. (2018). Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2), 223–311

Braun, A., Kohler, M., Langer, S., and Walk, H. (2024). Convergence rates for shallow neural networks learned by gradient descent. *Bernoulli*, 30, pp. 475-502

Chen, J., Langer, S. and Schmidt-Hieber, J. (2024). A novel statistical approach to analyze image classification. Preprint, arXiv: 2206.02151v2. To appear in *Annals of Statistics*.

Cohen, T., Welling, M.. (2016). Group Equivariant Convolutional Networks. *Proceedings of The 33rd International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 48:2990-2999

da Cunha, A., Natale, E., and Viennot, L. (2022). Proving the Lottery Ticket Hypothesis for Convolutional Neural Networks. *International Conference on Learning Representations.*

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York, USA.

Du, S. S., Zhai, X., Poczos, B., Singh, A. (2019). Gradient Descent Provably Optimizes Over-parameterized Neural Networks. *International Conference on Learning Representations.*

Frankle, J., Carbin, M. (2019). The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. *International Conference on Learning Representations.*

Gonon, L. (2023). Random feature networks learn Black-Scholes type PDEs without curse of dimensionality. *Journal of Machine Learning Research*, 24, Article 189

Gühring, I., Raslan, M., Kutyniok, G. (2022). Expressivity of Deep Neural Networks. In P. Grohs & G. Kutyniok (Eds.), Mathematical Aspects of Deep Learning (pp. 149–199). chapter, Cambridge: Cambridge University Press.

Grenander, U. (1969). Foundations of Pattern Analysis. *Quarterly of Applied Mathematics*, 27(1), 1–55.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression.* Springer.

Huang, G. B., Chen, L., and Siew, C.-K. (2006). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks*, 17, pp. 879-892

Jansson, Y., Lindeberg, T. (2022). Scale-Invariant Scale-Channel Networks: Deep Networks That Generalise to Previously Unseen Scales. *Journal of Mathematical Imaging and Vision*, 64(5), 506–536.

Kim, Y., Ohn, I. and Kim, D. (2021). *Fast convergence rates of deep neural networks for classification.* Neural Netw., 138:179–197.

Kohler, M., and Krzyżak, A. (2025). Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent. *IEEE Transactions on Information Theory*, 71, pp. 6165-6182

Kohler, M., Cho, J., Krzyżak, A. (2025). On the rate of convergence of an over-parametrized deep neural network regression estimate with ReLU activation function learned by gradient descent. Electronic Journal of Statistics, 19(2).

Kohler, M., and Krzyżak, A. (2024). On the rate of convergence of an over-parametrized transformer classifier learned by gradient descent. Preprint, *arXiv: 2312.17007.*

Kohler, M., Krzyżak, A. and Sänger, A.. (2025). Learning of deep convolutional network image classifiers via stochastic gradient descent and over-parametrization. Preprint, arXiv: 2404.07128

Kohler, M., and Langer, S. (2025). Statistical theory for image classification using deep convolutional neural network with cross-entropy loss under the hierarchical max-pooling model. *Journal of Statistical Planning and Inference*, **234**. p. 106188

Kohler, M., Krzyżak, A., and Walter, B. (2022). On the rate of convergence of image classifiers based on convolutional neural networks. *Annals of the Institute of Statistical Mathematics*, **74**, pp. 1085-1108.

Kohler, M., Walter, B. (2023). Analysis of Convolutional Neural Network Image Classifiers in a Rotationally Symmetric Model. *IEEE Transactions on Information Theory*, 69(8).

Kohler, M., Krzyżak, A., Walter, B. (2025). Analysis of the rate of convergence of an over-parametrized convolutional neural network image classifier learned by gradient descent. *Journal of Statistical Planning and Inference*, 239(106291), 106291.

Kutyniok, G. (2020). Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function." *The Annals of Statistics*, 48(4).

Langer, S. (2020). Approximating smooth functions by deep neural networks with sigmoid activation function. *Journal of Multivariate Analysis* **182**, pp. 104696.

Liu, H., Chen, M., Zhao, T., Liao, W.. (2021). Besov Function Approximation and Binary Classification on Low-Dimensional Manifolds Using Convolutional Residual Networks. *Proceedings of the 38th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 139:6770-6780

Long, P. M., Sedghi, H. (2020). Generalization bounds for deep convolutional neural networks. *International Conference on Learning Representations.*

Lu, J., Shen, Z., Yang, H., Zhang, S. (2021). Deep Network Approximation for Smooth Functions. SIAM Journal on Mathematical Analysis, 53(5), 5465–5506.

Marcos, D., Volpi, M., Tuia, D. (2016). Learning rotation invariant convolutional filters for texture classification. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 2012–2017).

Shen, G., Jiao, Y., Lin, Y., Huang, J. (2021). Non-asymptotic Excess Risk Bounds for Classification with Deep Convolutional Neural Networks (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2105.00292

Yarotsky, D. Universal Approximations of Invariant Maps by Neural Networks. *Constr Approx* 55, 407–474 (2022).

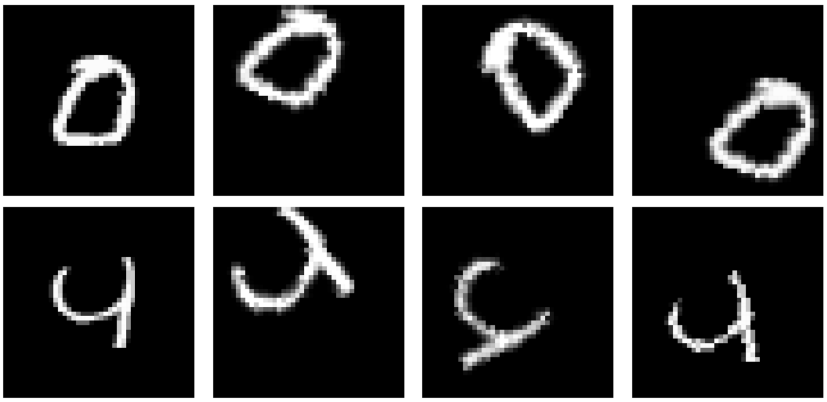
Yarotsky, D., Zhevnerchuk, A.: The phase diagram of approximation rates for deep neural networks. *Advances in Neural Information Processing Systems* 33, 13,005–13,015

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

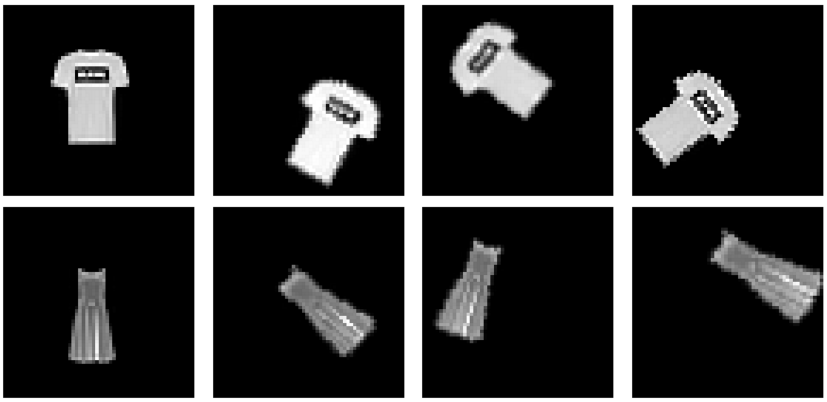
Zhou, D.-X. (2020). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2), 787–794.

A. Sample Images from datasets used for simulations

(a) MNIST class ”0“ vs ”4“



(b) Fashion MNIST class ”shirt“ vs ”dress“



(c) ASL class "a" vs "b"

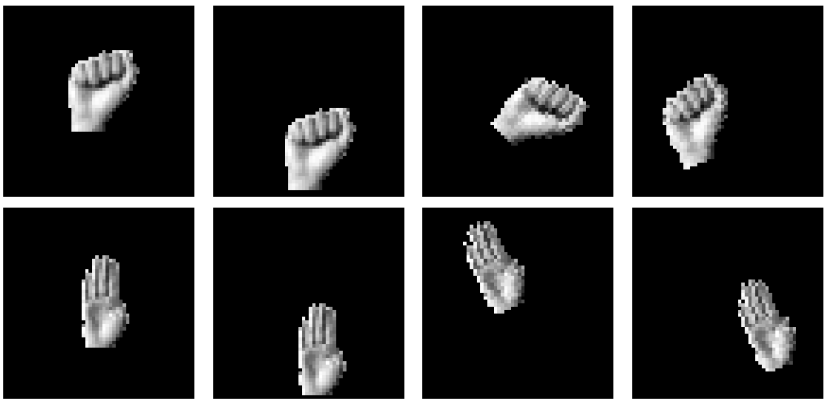


Figure 5: For each classification task, the last three images on the right were generated by applying deformations to the first image shown on the left.

B. Proof of Theorem 2

In the proof of Theorem 2 we will need the following auxiliary result.

Lemma 8. *Let $l_1, l_2, t_n \in \mathbb{N}$, let $D_n \geq 0$, let $A \subset \mathbb{R}^{l_1}$ be closed and convex, let $B \subseteq \mathbb{R}^{l_2}$ and let $F_t, F : \mathbb{R}^{l_1} \times \mathbb{R}^{l_2} \to \mathbb{R}_+$ $(t = 0, \ldots, t_n - 1)$ be functions such that for all $t \in \{0, \ldots, t_n - 1\}$*

$$u \mapsto F(u, v) \quad \textit{is differentiable and convex for all } v \in \mathbb{R}^{l_2},$$

$$u \mapsto F_t(u, v) \quad \textit{is differentiable for all } v \in \mathbb{R}^{l_2},$$

and

$$\|(\nabla_u F_t)(u, v)\| \leq D_n \tag{45}$$

for all $(u, v) \in A \times B$. Choose $(u_0, v_0) \in A \times B$, let $v_1, \ldots, v_{t_n} \in B$ and set

$$u_{t+1} = \mathrm{Proj}_A \left(u_t - \lambda \cdot (\nabla_u F_t)\left(u_t, v_t\right)\right) \quad (t = 0, \ldots, t_n - 1),$$

where

$$\lambda = \frac{1}{t_n}.$$

Let $u^* \in A$. Then it holds:

$$\begin{aligned}\frac{1}{t_n} \sum_{t=0}^{t_n-1} F(u_t, v_t) \quad &\leq \quad F(u^*, v_0) + \frac{1}{t_n} \sum_{t=1}^{t_n-1} |F(u^*, v_t) - F(u^*, v_0)| + \frac{\|u^* - u_0\|^2}{2} + \frac{D_n^2}{2 \cdot t_n} \\ &\quad + \frac{1}{t_n} \sum_{t=0}^{t_n-1} \left\langle (\nabla_u F)\left(u_t, v_t\right) - (\nabla_u F_t)\left(u_t, v_t\right), u_t - u^* \right\rangle.\end{aligned}$$

Proof. See Lemma 1 in Kohler, Krzyżak and Sänger (2025). □

Proof of Theorem 2. Let E_n be the event that there exist pairwise distinct $j_1, \ldots, j_{N_n} \in \{1, \ldots, K_n\}$ such that

$$\vartheta_{j_i}^{(0)} \in \mathbf{\Theta}^*$$

holds for all $i = 1, \ldots, N_n$. If E_n holds set

$$w_{j_i}^* = \frac{1}{N_n} \quad (i = 1, \ldots, N_n) \quad \text{and} \quad w_k^* = 0 \quad (k \in \{1, \ldots, K_n\} \setminus \{j_1, \ldots, j_{N_n}\})$$

and $\mathbf{w}^* = (w_k^*)_{k=1,\ldots,K_n}$, otherwise set $\mathbf{w}^* = 0$.

We will use the following error decomposition:

$$\begin{aligned}&\mathbf{E}\left\{\varphi(Y \cdot f_n(X))\right\} - \min_{f:[0,1]^{d\times d} \to \bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\} \\ &= \mathbf{E}\left\{\varphi(Y \cdot f_n(X)) \cdot 1_{E_n^c}\right\} \\ &\quad + \mathbf{E}\left\{\mathbf{E}\left\{\varphi(Y \cdot f_{(\hat{\mathbf{w}}, \boldsymbol{\vartheta}^{(t_n)})}(X)) \middle| \boldsymbol{\vartheta}^{(0)}, \mathcal{D}_n\right\} \cdot 1_{E_n}\right\}\end{aligned}$$

$$
\begin{aligned}
&\qquad -\mathbf{E}\left\{\frac{1}{t_n}\sum_{t=0}^{t_n-1}\mathbf{E}\left\{\varphi(Y\cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t_n)})}(X))\big|\boldsymbol{\vartheta}^{(0)},\mathcal{D}_n\right\}\cdot 1_{E_n}\right\}\\
&+\mathbf{E}\left\{\frac{1}{t_n}\sum_{t=0}^{t_n-1}\mathbf{E}\left\{\varphi(Y\cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t_n)})}(X))\big|\boldsymbol{\vartheta}^{(0)},\mathcal{D}_n\right\}\cdot 1_{E_n}\right\}\\
&\qquad -\mathbf{E}\left\{\frac{1}{t_n}\sum_{t=0}^{t_n-1}\mathbf{E}\left\{\varphi(Y\cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X))\big|\boldsymbol{\vartheta}^{(0)},\mathcal{D}_n\right\}\cdot 1_{E_n}\right\}\\
&+\mathbf{E}\left\{\frac{1}{t_n}\sum_{t=0}^{t_n-1}\mathbf{E}\left\{\varphi(Y\cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X))\big|\boldsymbol{\vartheta}^{(0)},\mathcal{D}_n\right\}\cdot 1_{E_n}\right\}\\
&\qquad -\min_{f:[0,1]^{d\times d}\to\bar{\mathbb{R}}}\mathbf{E}\left\{\varphi(Y\cdot f(X))\right\}\\
&=: T_{1,n}+T_{2,n}+T_{3,n}+T_{4,n}.
\end{aligned}
$$

In the *first step of the proof* we show

$$\mathbf{P}\{E_n^c\}\leq\frac{1}{n}. \tag{46}$$

To do this we consider a sequential choice of the initial weights $\vartheta_1^{(0)},\ \ldots,\ \vartheta_{K_n}^{(0)}$. By definition of κ_n we know that the probability that none of $\vartheta_1^{(0)},\ \ldots,\ \vartheta_{I_n}^{(0)}$ is contained in $\boldsymbol{\Theta}^*$ is given by

$$(1-\kappa_n)^{I_n}.$$

This implies that the probability that there exists $l\in\{1,\ldots,N_n\}$ such that none of $\vartheta_{(l-1)\cdot I_n+1}^{(0)},\ \ldots,\ \vartheta_{l\cdot I_n}^{(0)}$ is contained in $\boldsymbol{\Theta}^*$ is upper bounded by

$$N_n\cdot(1-\kappa_n)^{I_n}.$$

(30) implies

$$\mathbf{P}\{E_n^c\}\leq N_n\cdot(1-\kappa_n)^{I_n}\leq\frac{1}{n}.$$

In the *second step of the proof* we show

$$T_{1,n}\leq c_{10}\cdot\frac{(\log n)}{n}.$$

To do this, we observe that for $|z|\leq\beta_n$ and $n>1$ we have

$$\varphi(z)=\log(1+\exp(-z))\leq(\log 4)\cdot I_{\{z>-1\}}+\log(2\cdot\exp(-z))\cdot I_{\{z\leq-1\}}\leq 3+|z|\leq c_{11}\cdot\log n,$$

from which we can conclude by the first step of the proof

$$T_{1,n}\leq c_{11}\cdot(\log n)\cdot\mathbf{P}\{E_n^c\}\leq c_{11}\cdot\frac{\log n}{n}.$$

In the *third step of the proof* we show

$$T_{2,n} \quad \leq \quad 0.$$

This follows from the convexity of the logistic loss, which implies

$$\begin{aligned}
&\mathbf{E}\left\{\varphi(Y \cdot f_{(\hat{\mathbf{w}},\boldsymbol{\vartheta}^{(t_n)})}(X))\big|\boldsymbol{\vartheta}^{(0)},\mathcal{D}_n\right\} \\
&= \mathbf{E}\left\{\varphi(Y \cdot \frac{1}{t_n}\sum_{t=0}^{t_n-1} f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t_n)})}(X))\big|\boldsymbol{\vartheta}^{(0)},\mathcal{D}_n\right\} \\
&\leq \mathbf{E}\left\{\frac{1}{t_n}\sum_{t=0}^{t_n-1} \varphi(Y \cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t_n)})}(X))\big|\boldsymbol{\vartheta}^{(0)},\mathcal{D}_n\right\} \\
&= \frac{1}{t_n}\sum_{t=0}^{t_n-1} \mathbf{E}\left\{\varphi(Y \cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t_n)})}(X))\big|\boldsymbol{\vartheta}^{(0)},\mathcal{D}_n\right\}.
\end{aligned}$$

In the *fourth step of the proof* we show

$$T_{3,n} \leq 2 \cdot \beta_n \cdot \frac{C_n}{\sqrt{N_n}}.$$

Due to the fact that the logistic loss is Lipschitz continuous with Lipschitz constant 1 and by assumptions (24) and (28) we have

$$\begin{aligned}
&T_{3,n} \\
&\leq \frac{1}{t_n}\sum_{t=0}^{t_n-1} \mathbf{E}\left\{\left|\varphi(Y \cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t_n)})}(X)) - \varphi(Y \cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X))\right|\right\} \\
&\leq \frac{1}{t_n}\sum_{t=0}^{t_n-1} \mathbf{E}\left\{|f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t_n)})}(X) - f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X)|\right\} \\
&\leq \frac{1}{t_n}\sum_{t=0}^{t_n-1} \mathbf{E}\left\{\sum_{k=1}^{K_n} w_k^{(t)} \cdot \beta_n \cdot |T_1 f_{\vartheta_k^{(t_n)}}(X) - T_1 f_{\vartheta_k^{(t)}}(X)|\right\} \\
&\leq \frac{1}{t_n}\beta_n \cdot \sum_{t=0}^{t_n-1} \mathbf{E}\left\{\sum_{k=1}^{K_n} w_k^{(t)} \cdot |f_{\vartheta_k^{(t_n)}}(X) - f_{\vartheta_k^{(t)}}(X)|\right\} \\
&\leq \frac{1}{t_n} \cdot \beta_n \cdot \sum_{t=0}^{t_n-1} \mathbf{E}\left\{\sqrt{\sum_{k=1}^{K_n}(w_k^{(t)})^2} \cdot \sqrt{\sum_{k=1}^{K_n}|f_{\vartheta_k^{(t_n)}}(X) - f_{\vartheta_k^{(t)}}(X)|^2}\right\} \\
&\leq \frac{1}{t_n} \cdot \beta_n \cdot \sum_{t=0}^{t_n-1} \mathbf{E}\left\{\sqrt{\alpha_n} \cdot \sqrt{\sum_{k=1}^{K_n} C_n^2 \cdot \|\vartheta_k^{(t_n)} - \vartheta_k^{(t)}\|_\infty^2}\right\} \\
&\leq \frac{C_n}{\sqrt{N_n}} \cdot \beta_n \cdot \frac{1}{t_n}\sum_{t=0}^{t_n-1} \mathbf{E}\left\{\sqrt{\|\boldsymbol{\vartheta}^{(t_n)} - \boldsymbol{\vartheta}^{(t)}\|^2}\right\}
\end{aligned}$$

$$\leq 2 \cdot \beta_n \cdot \frac{C_n}{\sqrt{N_n}}.$$

In the *fifth step of the proof* we apply Lemma 8 to $T_{4,n}$. Set

$$F((\mathbf{w},\boldsymbol{\vartheta})) = \mathbf{E}\{\varphi(Y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X))\big|\boldsymbol{\vartheta}^{(0)}, \mathcal{D}_n\} \text{ and } F_t((\mathbf{w},\boldsymbol{\vartheta})) = \varphi(Y_{j_t} \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_{j_t})).$$

Then Lemma 8 implies

$$\begin{aligned}
T_{4,n} \;\leq\; & \mathbf{E}\left\{\mathbf{E}\left\{\varphi(Y \cdot f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(0)})}(X))\big|\boldsymbol{\vartheta}^{(0)}, \mathcal{D}_n\right\} \cdot 1_{E_n}\right\} - \min_{f:[0,1]^{d\times d}\to\bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\} \\
& + \frac{1}{t_n}\sum_{t=1}^{t_n-1} \mathbf{E}\Bigg\{\Bigg|\mathbf{E}\left\{\varphi(Y \cdot f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(t)})}(X))\big|\mathcal{D}_n, \boldsymbol{\vartheta}^{(0)}\right\} \\
& \qquad\qquad - \mathbf{E}\left\{\varphi(Y \cdot f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(0)})}(X))\big|\mathcal{D}_n, \boldsymbol{\vartheta}^{(0)}\right\}\Bigg| \cdot 1_{E_n}\Bigg\} + \frac{1}{2}\cdot\frac{1}{N_n} + \frac{D_n^2}{2\cdot t_n} \\
& + \frac{1}{t_n}\sum_{t=0}^{t_n-1} \mathbf{E}\left\{\left\langle (\nabla_{\mathbf{w}} F)\left(\mathbf{w}^{(t)}, \boldsymbol{\vartheta}^{(t)}\right) - \left(\nabla_{\mathbf{w}} F_t\right)\left(\mathbf{w}^{(t)}, \boldsymbol{\vartheta}^{(t)}\right), \mathbf{w}^{(t)} - \mathbf{w}^*\right\rangle \cdot 1_{E_n}\right\}.
\end{aligned}$$

In the *sixth step of the proof* we show

$$\begin{aligned}
& \mathbf{E}\left\{\mathbf{E}\left\{\varphi(Y \cdot f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(0)})}(X))\big|\boldsymbol{\vartheta}^{(0)}, \mathcal{D}_n\right\} \cdot 1_{E_n}\right\} - \min_{f:[0,1]^{d\times d}\to\bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\} \\
& \leq \sup_{\vartheta\in\boldsymbol{\Theta}^*} \mathbf{E}\left\{\varphi(Y \cdot \beta_n \cdot T_1 f_\vartheta(X))\right\} - \min_{f:[0,1]^{d\times d}\to\bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\}\Bigg).
\end{aligned}$$

This follows from the convexity of the logistic loss, which implies

$$\begin{aligned}
& \mathbf{E}\left\{\mathbf{E}\left\{\varphi(Y \cdot f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(0)})}(X))\big|\boldsymbol{\vartheta}^{(0)}, \mathcal{D}_n\right\} \cdot 1_{E_n}\right\} \\
& = \mathbf{E}\left\{\mathbf{E}\left\{\varphi\left(\frac{1}{N_n}\sum_{k=1}^{N_n} Y \cdot \beta_n \cdot T_1 f_{\vartheta_{j_k}^{(0)}}(X)\right)\big|\boldsymbol{\vartheta}^{(0)}, \mathcal{D}_n\right\} \cdot 1_{E_n}\right\} \\
& \leq \frac{1}{N_n}\sum_{k=1}^{N_n} \mathbf{E}\left\{\mathbf{E}\left\{\varphi\left(Y \cdot \beta_n \cdot T_1 f_{\vartheta_{j_k}^{(0)}}(X)\right)\big|\boldsymbol{\vartheta}^{(0)}, \mathcal{D}_n\right\} \cdot 1_{E_n}\right\} \\
& \leq \sup_{\vartheta\in\boldsymbol{\Theta}^*} \mathbf{E}\left\{\varphi(Y \cdot \beta_n \cdot T_1 f_\vartheta(X))\right\}.
\end{aligned}$$

In the *seventh step of the proof* we show

$$\begin{aligned}
& \frac{1}{t_n}\sum_{t=1}^{t_n-1} \mathbf{E}\left\{\left|\mathbf{E}\left\{\varphi\left(Y \cdot f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(t)})}(X)\right)\big|\mathcal{D}_n, \boldsymbol{\vartheta}^{(0)}\right\} - \mathbf{E}\left\{\varphi\left(Y \cdot f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(0)})}(X)\right)\big|\mathcal{D}_n, \boldsymbol{\vartheta}^{(0)}\right\}\right| \cdot 1_{E_n}\right\} \\
& \leq \beta_n \cdot \frac{C_n}{\sqrt{N_n}}.
\end{aligned}$$

The Lipschitz continuity of the logistic loss and assumption (28) imply

$$\begin{aligned}
&\frac{1}{t_n}\sum_{t=1}^{t_n-1}\mathbf{E}\left\{\left|\mathbf{E}\left\{\varphi\left(Y\cdot f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(t)})}(X)\right)\left|\mathcal{D}_n,\boldsymbol{\vartheta}^{(0)}\right.\right\}-\mathbf{E}\left\{\varphi\left(Y\cdot f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(0)})}(X)\right)\left|\mathcal{D}_n,\boldsymbol{\vartheta}^{(0)}\right.\right\}\right|\cdot 1_{E_n}\right\}\\
&\leq\frac{1}{t_n}\sum_{t=1}^{t_n-1}\mathbf{E}\left\{\left|\varphi\left(Y\cdot f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(t)})}(X)\right)-\varphi\left(Y\cdot f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(0)})}(X)\right)\right|\right\}\\
&\leq\frac{1}{t_n}\sum_{t=1}^{t_n-1}\mathbf{E}\left\{\left|f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(t)})}(X)-f_{(\mathbf{w}^*,\boldsymbol{\vartheta}^{(0)})}(X)\right|\right\}\\
&=\frac{1}{t_n}\sum_{t=1}^{t_n-1}\mathbf{E}\left\{\left|\sum_{k=1}^{K_n}w_k^*\cdot\beta_n\cdot\left(T_1f_{\vartheta_k^{(t)}}(X)-T_1f_{\vartheta_k^{(0)}}(X)\right)\right|\right\}\\
&\leq\frac{1}{t_n}\sum_{t=1}^{t_n-1}\mathbf{E}\left\{\beta_n\cdot\sqrt{\sum_{k=1}^{K_n}|w_k^*|^2}\cdot\sqrt{\sum_{k=1}^{K_n}\left(T_1f_{\vartheta_k^{(t)}}(X)-T_1f_{\vartheta_k^{(0)}}(X)\right)^2}\right\}\\
&\leq\frac{1}{t_n}\beta_n\cdot\sum_{t=1}^{t_n-1}\mathbf{E}\left\{\sqrt{\sum_{k=1}^{K_n}|w_k^*|^2}\cdot\sqrt{\sum_{k=1}^{K_n}\left(f_{\vartheta_k^{(t)}}(X)-f_{\vartheta_k^{(0)}}(X)\right)^2}\right\}\\
&\leq\frac{1}{t_n}\beta_n\cdot\sum_{t=1}^{t_n-1}\mathbf{E}\left\{\frac{1}{\sqrt{N_n}}\sqrt{\sum_{k=1}^{K_n}C_n^2\cdot\|\vartheta_k^{(t)}-\vartheta_k^{(0)}\|_\infty^2}\right\}\\
&\leq\beta_n\cdot\frac{1}{t_n}\sum_{t=1}^{t_n-1}\mathbf{E}\left\{\frac{C_n}{\sqrt{N_n}}\cdot\|\boldsymbol{\vartheta}^{(t)}-\boldsymbol{\vartheta}^{(0)}\|\right\}\leq\beta_n\cdot\frac{C_n}{\sqrt{N_n}}.
\end{aligned}$$

Let $\mathcal{W}$ be the set of all weight vectors $\mathbf{w}=((w_k)_{k=1,\dots,K_n},(\vartheta_k)_{k=1,\dots,K_n})$ which satisfy $\boldsymbol{\vartheta}=(\vartheta_k)_{k=1,\dots,K_n}\in\bar{\boldsymbol{\Theta}}^{K_n}$ and (12). Let (X_1',Y_1'), …, (X_n',Y_n'), ϵ_1, …, ϵ_n be random variables independent of $\boldsymbol{\vartheta}^{(0)}$ with $\mathbf{P}\{\epsilon_j=1\}=1/2=\mathbf{P}\{\epsilon_j=-1\}$ $(j=1,\dots,n)$ such that (X_1,Y_1), …, (X_n,Y_n), ϵ_1, …, ϵ_n, (X_1',Y_1'), …, (X_n',Y_n') are independent and (X_1,Y_1), …, (X_n,Y_n), (X_1',Y_1'), …, (X_n',Y_n') are identically distributed.

In the *eighth step of the proof* we show

$$\begin{aligned}
&\frac{1}{t_n}\sum_{t=0}^{t_n-1}\mathbf{E}\left\{\left\langle\left(\nabla_{\mathbf{w}}F\right)(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})-\left(\nabla_{\mathbf{w}}F_t\right)(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)}),\mathbf{w}^{(t)}-\mathbf{w}^*\right\rangle\cdot 1_{E_n}\right\}\\
&\leq 4\cdot\mathbf{E}\left\{\sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ k\in\{1,\dots,K_n\}}}\left(\frac{1}{n}\sum_{i=1}^n\epsilon_i\cdot\frac{1}{1+\exp(Y_i\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_i))}\cdot Y_i\cdot\beta_n\cdot T_1f_{\vartheta_k}(X_i)\right)\cdot 1_{E_n}\right\}\\
&\quad+\frac{n\cdot\left(6\cdot K_n\cdot\beta_n^2+4\cdot(\beta_n+1)\cdot C_n\cdot\sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ y\in\{-1,1\},x\in[0,1]^{d\times d}}}\|\nabla_{\boldsymbol{\vartheta}}\varphi(y\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x))\|_\infty\right)}{t_n}.
\end{aligned}$$

We have

$$\varphi'(z)=\frac{1}{1+\exp(-z)}\cdot(-\exp(-z))=\frac{-1}{1+\exp(z)},$$

which implies

$$\frac{\partial}{\partial w_j}\varphi\left(Y\cdot\sum_{k=1}^{K_n}w_k\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X)\right)=\frac{-Y\cdot\beta_n\cdot T_1 f_{\vartheta_j}(X)}{1+\exp\left(Y\cdot\sum_{k=1}^{K_n}w_k\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X)\right)}$$

and

$$\begin{aligned}
&\frac{\partial}{\partial w_j}\mathbf{E}\left\{\varphi\left(Y\cdot\sum_{k=1}^{K_n}w_k\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X)\right)\right\}\\
&=\lim_{h\to 0}\mathbf{E}\left\{\frac{\varphi\left(Y\cdot\sum_{k=1}^{K_n}(w_k+h\cdot I_{\{k=j\}})\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X)\right)-\varphi\left(Y\cdot\sum_{k=1}^{K_n}w_k\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X)\right)}{h}\right\}\\
&=\mathbf{E}\left\{\lim_{h\to 0}\frac{\varphi\left(Y\cdot\sum_{k=1}^{K_n}(w_k+h\cdot I_{\{k=j\}})\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X)\right)-\varphi\left(Y\cdot\sum_{k=1}^{K_n}w_k\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X)\right)}{h}\right\}\\
&=\mathbf{E}\left\{\frac{-Y\cdot\beta_n\cdot T_1 f_{\vartheta_j}(X)}{1+\exp\left(Y\cdot\sum_{k=1}^{K_n}w_k\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X)\right)}\right\},
\end{aligned}$$

where we have used

$$\begin{aligned}
&\left|\frac{\varphi\left(Y\cdot\sum_{k=1}^{K_n}(w_k+h\cdot I_{\{k=j\}})\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X)\right)-\varphi\left(Y\cdot\sum_{k=1}^{K_n}w_k\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X)\right)}{h}\right|\\
&=\beta_n\cdot|\varphi'(\xi)|\le\beta_n,
\end{aligned}$$

where ξ is determined according to the mean value theorem, and the dominated convergence theorem in order to interchange limits and expectations above.

Consequently,

$$\begin{aligned}
&\frac{1}{t_n}\sum_{t=0}^{t_n-1}\mathbf{E}\left\{\left\langle\left(\nabla_{\mathbf{w}}F\right)\left(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)}\right)-\left(\nabla_{\mathbf{w}}F_t\right)\left(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)}\right),\mathbf{w}^{(t)}-\mathbf{w}^*\right\rangle\cdot 1_{E_n}\right\}\\
&=\mathbf{E}\Bigg\{\frac{1}{t_n}\sum_{t=0}^{t_n-1}\sum_{k=1}^{K_n}\Bigg(\mathbf{E}\left\{\frac{-Y\cdot\beta_n\cdot T_1 f_{\vartheta_k^{(t)}}(X)}{1+\exp\left(Y\cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X)\right)}\middle|\mathcal{D}_n,\boldsymbol{\vartheta}^{(0)}\right\}\\
&\qquad\qquad-\frac{-Y_{j_t}\cdot\beta_n\cdot T_1 f_{\vartheta_k^{(t)}}(X_{j_t})}{1+\exp\left(Y_{j_t}\cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X_{j_t})\right)}\Bigg)\cdot(w_k^{(t)}-w_k^*)\cdot 1_{E_n}\Bigg\}\\
&=\frac{1}{t_n/n}\cdot\sum_{s=1}^{t_n/n}\mathbf{E}\Bigg\{\frac{1}{n}\sum_{t=(s-1)\cdot n}^{s\cdot n-1}\sum_{k=1}^{K_n}\Bigg(\mathbf{E}\left\{\frac{-Y\cdot\beta_n\cdot T_1 f_{\vartheta_k^{(t)}}(X)}{1+\exp\left(Y\cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X)\right)}\middle|\mathcal{D}_n,\boldsymbol{\vartheta}^{(0)}\right\}\\
&\qquad\qquad-\frac{-Y_{j_t}\cdot\beta_n\cdot T_1 f_{\vartheta_k^{(t)}}(X_{j_t})}{1+\exp\left(Y_{j_t}\cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X_{j_t})\right)}\Bigg)\cdot(w_k^{(t)}-w_k^*)\cdot 1_{E_n}\Bigg\}.
\end{aligned}$$

During n gradient descent steps the parameter $(\mathbf{w}, \boldsymbol{\vartheta})$ changes in supremum norm at most by

$$
\begin{aligned}
& n \cdot \lambda_n \cdot \max \left\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta}) \in \mathcal{W}, y \in \{-1,1\}, x \in [0,1]^{d \times d}} \| \nabla_{\mathbf{w}} \varphi(y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x)) \|_\infty, \right. \\
& \qquad \left. \sup_{(\mathbf{w},\boldsymbol{\vartheta}) \in \mathcal{W}, y \in \{-1,1\}, x \in [0,1]^{d \times d}} \| \nabla_{\boldsymbol{\vartheta}} \varphi(y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x)) \|_\infty \right\} \\
& \leq n \cdot \lambda_n \cdot \left(\beta_n + \sup_{(\mathbf{w},\boldsymbol{\vartheta}) \in \mathcal{W}, y \in \{-1,1\}, x \in [0,1]^{d \times d}} \| \nabla_{\boldsymbol{\vartheta}} \varphi(y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x)) \|_\infty \right) \\
& = \frac{n \cdot \left(\beta_n + \sup_{(\mathbf{w},\boldsymbol{\vartheta}) \in \mathcal{W}, y \in \{-1,1\}, x \in [0,1]^{d \times d}} \| \nabla_{\boldsymbol{\vartheta}} \varphi(y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x)) \|_\infty \right)}{t_n}.
\end{aligned}
$$

Using

$$
\begin{aligned}
& \sum_{k=1}^{K_n} \left(\mathbf{E} \left\{ \frac{-Y \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(t)}}(X)}{1 + \exp\left(Y \cdot f_{(\mathbf{w}^{(t)}, \boldsymbol{\vartheta}^{(t)})}(X) \right)} \middle| \mathcal{D}_n, \boldsymbol{\vartheta}^{(0)} \right\} \right. \\
& \qquad - \mathbf{E} \left\{ \frac{-Y \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(s \cdot n-1)}}(X)}{1 + \exp\left(Y \cdot f_{(\mathbf{w}^{(s \cdot n-1)}, \boldsymbol{\vartheta}^{(s \cdot n-1)})}(X) \right)} \middle| \mathcal{D}_n, \boldsymbol{\vartheta}^{(0)} \right\} \\
& \qquad - \frac{-Y_{j_t} \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(t)}}(X_{j_t})}{1 + \exp\left(Y_{j_t} \cdot f_{(\mathbf{w}^{(t)}, \boldsymbol{\vartheta}^{(t)})}(X_{j_t}) \right)} \\
& \qquad \left. + \frac{-Y_{j_t} \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(s \cdot n-1)}}(X_{j_t})}{1 + \exp\left(Y_{j_t} \cdot f_{(\mathbf{w}^{(s \cdot n-1)}, \boldsymbol{\vartheta}^{(s \cdot n-1)})}(X_{j_t}) \right)} \right) \cdot (w_k^{(t)} - w_k^*) \\
& \leq \sum_{k=1}^{K_n} \left(\mathbf{E} \left\{ \left| \frac{-Y \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(t)}}(X) + Y \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(s \cdot n-1)}}(X)}{1 + \exp\left(Y \cdot f_{(\mathbf{w}^{(t)}, \boldsymbol{\vartheta}^{(t)})}(X) \right)} \right| \right. \right. \\
& \qquad + \left| Y \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(s \cdot n-1)}}(X) \right| \cdot \left| \frac{1}{1 + \exp\left(Y \cdot f_{(\mathbf{w}^{(t)}, \boldsymbol{\vartheta}^{(t)})}(X) \right)} \right. \\
& \qquad \left. \left. - \frac{1}{1 + \exp\left(Y \cdot f_{(\mathbf{w}^{(s \cdot n-1)}, \boldsymbol{\vartheta}^{(s \cdot n-1)})}(X) \right)} \right| \middle| \mathcal{D}_n, \boldsymbol{\vartheta}^{(0)} \right\} \\
& \qquad + \left| \frac{-Y_{j_t} \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(s \cdot n-1)}}(X_{j_t}) + Y_{j_t} \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(t)}}(X_{j_t})}{1 + \exp\left(Y_{j_t} \cdot f_{(\mathbf{w}^{(s \cdot n-1)}, \boldsymbol{\vartheta}^{(s \cdot n-1)})}(X_{j_t}) \right)} \right| \\
& \qquad + \left| Y_{j_t} \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(t)}}(X_{j_t}) \right| \cdot \left| \frac{1}{1 + \exp\left(Y_{j_t} \cdot f_{(\mathbf{w}^{(t)}, \boldsymbol{\vartheta}^{(t)})}(X_{j_t}) \right)} \right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad - \frac{1}{1+\exp\left(Y_{j_t}\cdot f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X_{j_t})\right)}\Bigg|\Bigg)\cdot\left|w_k^{(t)}-w_k^*\right| \\
&\le \sum_{k=1}^{K_n}\Bigg(\mathbf{E}\left\{\left|T_1 f_{\vartheta_k^{(s\cdot n-1)}}(X)-T_1 f_{\vartheta_k^{(t)}}(X)\right|+\left|f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X)-f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X)\right|\Bigg|\mathcal{D}_n,\boldsymbol{\vartheta}^{(0)}\right\} \\
&\qquad +\left|T_1 f_{\vartheta_k^{(t)}}(X_{j_t})-T_1 f_{\vartheta_k^{(s\cdot n-1)}}(X_{j_t})\right| \\
&\qquad\quad +\left|f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X_{j_t})-f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X_{j_t})\right|\Bigg)\cdot\beta_n\cdot|w_k^{(t)}-w_k^*| \\
&\le \sum_{k=1}^{K_n}\Bigg(\mathbf{E}\left\{\left|f_{\vartheta_k^{(s\cdot n-1)}}(X)-f_{\vartheta_k^{(t)}}(X)\right|+\left|f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X)-f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X)\right|\Bigg|\mathcal{D}_n,\boldsymbol{\vartheta}^{(0)}\right\} \\
&\qquad +\left|f_{\vartheta_k^{(t)}}(X_{j_t})-f_{\vartheta_k^{(s\cdot n-1)}}(X_{j_t})\right| \\
&\qquad\quad +\left|f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X_{j_t})-f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X_{j_t})\right|\Bigg)\cdot\beta_n\cdot|w_k^{(t)}-w_k^*| \\
&\le 2\cdot\sum_{k=1}^{K_n}\left(C_n\cdot\|\vartheta_k^{(t)}-\vartheta_k^{(s\cdot n-1)}\|_\infty+C_n\cdot\max_j\|\vartheta_j^{(t)}-\vartheta_j^{(s\cdot n-1)}\|_\infty+K_n\cdot\frac{\beta_n\cdot n}{t_n}\right)\cdot\beta_n\cdot|w_k^{(t)}-w_k^*| \\
&\le 8\cdot\beta_n\cdot C_n\cdot\max_j\|\vartheta_j^{(t)}-\vartheta_j^{(s\cdot n-1)}\|_\infty+4\cdot K_n\cdot\frac{\beta_n^2\cdot n}{t_n} \\
&\le \frac{8\cdot\beta_n\cdot C_n\cdot n\cdot\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},y\in\{-1,1\},x\in[0,1]^{d\times d}}\|\nabla_{\boldsymbol{\vartheta}}\varphi(y\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x))\|_\infty+4\cdot K_n\cdot\beta_n^2\cdot n}{t_n},
\end{aligned}
$$

where the fourth inequality follows from

$$
\begin{aligned}
&\left|f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X)-f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X)\right| \\
&=\left|\sum_{k=1}^{K_n}w_k^{(t)}\cdot\beta_n\cdot T_1 f_{\vartheta_k^{(t)}}(X)-w_k^{(s\cdot n-1)}\cdot\beta_n\cdot T_1 f_{\vartheta_k^{(s\cdot n-1)}}(X)\right| \\
&\le\sum_{k=1}^{K_n}\beta_n\cdot\left|w_k^{(t)}\right|\cdot\left|T_1 f_{\vartheta_k^{(t)}}(X)-T_1 f_{\vartheta_k^{(s\cdot n-1)}}(X)\right|+\beta_n\cdot\left|w_k^{(t)}-w_k^{(s\cdot n-1)}\right|\cdot\left|T_1 f_{\vartheta_k^{(s\cdot n-1)}}(X)\right| \\
&\le C_n\cdot\beta_n\cdot\max_j\|\vartheta_j^{(t)}-\vartheta_j^{(s\cdot n-1)}\|_\infty+K_n\cdot\frac{n\cdot\beta_n}{t_n}\cdot\beta_n,
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{k=1}^{K_n}\Bigg(\mathbf{E}\left\{\frac{-Y\cdot\beta_n\cdot T_1 f_{\vartheta_k^{(s\cdot n-1)}}(X)}{1+\exp\left(Y\cdot f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X)\right)}\Bigg|\mathcal{D}_n,\boldsymbol{\vartheta}^{(0)}\right\} \\
&\qquad -\frac{-Y_{j_t}\cdot\beta_n\cdot T_1 f_{\vartheta_k^{(s\cdot n-1)}}(X_{j_t})}{1+\exp\left(Y_{j_t}\cdot f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X_{j_t})\right)}\Bigg)\cdot(w_k^{(t)}-w_k^{(s\cdot n-1)})
\end{aligned}
$$

$$\leq 2 \cdot \beta_n \cdot \sum_{k=1}^{K_n} |w_k^{(t)} - w_k^{(s\cdot n-1)}| \leq 2 \cdot K_n \cdot n \cdot \frac{\beta_n^2}{t_n}$$

we can conclude

$$\begin{aligned}
&\frac{1}{t_n/n} \cdot \sum_{s=1}^{t_n/n} \mathbf{E}\Bigg\{ \frac{1}{n} \sum_{t=(s-1)\cdot n}^{s\cdot n-1} \sum_{k=1}^{K_n} \Bigg(\mathbf{E}\left\{ \frac{-Y \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(t)}}(X)}{1+\exp\left(Y \cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X)\right)} \middle| \mathcal{D}_n, \boldsymbol{\vartheta}^{(0)} \right\} \\
&\qquad\qquad - \frac{-Y_{j_t} \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(t)}}(X_{j_t})}{1+\exp\left(Y_{j_t} \cdot f_{(\mathbf{w}^{(t)},\boldsymbol{\vartheta}^{(t)})}(X_{j_t})\right)} \Bigg) \cdot (w_k^{(t)} - w_k^*) \cdot 1_{E_n} \Bigg\} \\
&\leq \frac{1}{t_n/n} \cdot \sum_{s=1}^{t_n/n} \mathbf{E}\Bigg\{ \frac{1}{n} \sum_{t=(s-1)\cdot n}^{s\cdot n-1} \sum_{k=1}^{K_n} \Bigg(\mathbf{E}\left\{ \frac{-Y \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(s\cdot n-1)}}(X)}{1+\exp\left(Y \cdot f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X)\right)} \middle| \mathcal{D}_n, \boldsymbol{\vartheta}^{(0)} \right\} \\
&\qquad\qquad - \frac{-Y_{j_t} \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(s\cdot n-1)}}(X_{j_t})}{1+\exp\left(Y_{j_t} \cdot f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X_{j_t})\right)} \Bigg) \cdot (w_k^{(s\cdot n-1)} - w_k^*) \cdot 1_{E_n} \Bigg\} \\
&\quad + \frac{n \cdot \left(6 \cdot K_n \cdot \beta_n^2 + 8 \cdot \beta_n \cdot C_n \cdot \sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}, \\ y\in\{-1,1\}, x\in[0,1]^{d\times d}}} \|\nabla_{\boldsymbol{\vartheta}} \varphi(y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x))\|_\infty \right)}{t_n}.
\end{aligned}$$

We continue by deriving an upper bound on the first term of the sum of the right-hand side above. We have

$$\begin{aligned}
&\frac{1}{t_n/n} \cdot \sum_{s=1}^{t_n/n} \mathbf{E}\Bigg\{ \frac{1}{n} \sum_{t=(s-1)\cdot n}^{s\cdot n-1} \sum_{k=1}^{K_n} \Bigg(\mathbf{E}\left\{ \frac{-Y \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(s\cdot n-1)}}(X)}{1+\exp\left(Y \cdot f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X)\right)} \middle| \mathcal{D}_n, \boldsymbol{\vartheta}^{(0)} \right\} \\
&\qquad\qquad - \frac{-Y_{j_t} \cdot \beta_n \cdot T_1 f_{\vartheta_k^{(s\cdot n-1)}}(X_{j_t})}{1+\exp\left(Y_{j_t} \cdot f_{(\mathbf{w}^{(s\cdot n-1)},\boldsymbol{\vartheta}^{(s\cdot n-1)})}(X_{j_t})\right)} \Bigg) \cdot (w_k^{(s\cdot n-1)} - w_k^*) \cdot 1_{E_n} \Bigg\} \\
&\leq \frac{1}{t_n/n} \cdot \sum_{s=1}^{t_n/n} \mathbf{E}\Bigg\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}} \frac{1}{n} \sum_{t=(s-1)\cdot n}^{s\cdot n-1} \sum_{k=1}^{K_n} \Bigg(\mathbf{E}\left\{ \frac{-Y \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X)}{1+\exp\left(Y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X)\right)} \right\} \\
&\qquad\qquad - \frac{-Y_{j_t} \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X_{j_t})}{1+\exp\left(Y_{j_t} \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_{j_t})\right)} \Bigg) \cdot (w_k - w_k^*) \cdot 1_{E_n} \Bigg\} \\
&= \mathbf{E}\Bigg\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}} \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{K_n} \Bigg(\mathbf{E}\left\{ \frac{-Y \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X)}{1+\exp\left(Y \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X)\right)} \right\} \\
&\qquad\qquad - \frac{-Y_j \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X_j)}{1+\exp\left(Y_j \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j)\right)} \Bigg) \cdot (w_k - w_k^*) \cdot 1_{E_n} \Bigg\} \\
&= \mathbf{E}\Bigg\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}} \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{K_n} \Bigg(\mathbf{E}\left\{ \frac{-Y'_j \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X'_j)}{1+\exp\left(Y'_j \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X'_j)\right)} \middle| \mathcal{D}_n, \boldsymbol{\vartheta}^{(0)} \right\}
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad -\frac{-Y_j\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j)}{1+\exp\left(Y_j\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j)\right)}\Bigg)\cdot(w_k-w_k^*)\cdot 1_{E_n}\Bigg\}\\
&\le \mathbf{E}\Bigg\{\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{K_n}\Bigg(\frac{-Y_j'\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j')}{1+\exp\left(Y_j'\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j')\right)}\\
&\qquad\qquad -\frac{-Y_j\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j)}{1+\exp\left(Y_j\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j)\right)}\Bigg)\cdot(w_k-w_k^*)\cdot 1_{E_n}\Bigg\}\\
&= \mathbf{E}\Bigg\{\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{K_n}\epsilon_j\cdot\Bigg(\frac{-Y_j'\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j')}{1+\exp\left(Y_j'\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j')\right)}\\
&\qquad\qquad -\frac{-Y_j\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j)}{1+\exp\left(Y_j\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j)\right)}\Bigg)\cdot(w_k-w_k^*)\cdot 1_{E_n}\Bigg\}\\
&\le \mathbf{E}\Bigg\{\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{K_n}\epsilon_j\cdot\Bigg(\frac{-Y_j'\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j')}{1+\exp\left(Y_j'\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j')\right)}\Bigg)\cdot(w_k-w_k^*)\cdot 1_{E_n}\Bigg\}\\
&\quad +\mathbf{E}\Bigg\{\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{K_n}\epsilon_j\cdot\Bigg(\frac{Y_j\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j)}{1+\exp\left(Y_j\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j)\right)}\Bigg)\cdot(w_k-w_k^*)\cdot 1_{E_n}\Bigg\}\\
&= 2\cdot\mathbf{E}\Bigg\{\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{K_n}\epsilon_j\cdot\Bigg(\frac{Y_j\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j)}{1+\exp\left(Y_j\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j)\right)}\Bigg)\cdot(w_k-w_k^*)\cdot 1_{E_n}\Bigg\}\\
&\le 2\cdot\mathbf{E}\Bigg\{\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{K_n}\epsilon_j\cdot\frac{Y_j\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j)}{1+\exp\left(Y_j\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j)\right)}\cdot(w_k-w_k^*)_+\cdot 1_{E_n}\Bigg\}\\
&\quad +2\cdot\mathbf{E}\Bigg\{\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{K_n}(-\epsilon_j)\cdot\frac{Y_j\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j)}{1+\exp\left(Y_j\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j)\right)}\cdot(w_k^*-w_k)_+\cdot 1_{E_n}\Bigg\}\\
&\le 2\cdot\mathbf{E}\Bigg\{\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\sum_{k=1}^{K_n}\sup_{\substack{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})\in\mathcal{W},\\ \bar{k}\in\{1,\ldots,K_n\}}}\frac{1}{n}\sum_{j=1}^{n}\epsilon_j\cdot\frac{Y_j\cdot\beta_n\cdot T_1 f_{\bar{\vartheta}_{\bar{k}}}(X_j)}{1+\exp\left(Y_j\cdot f_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})}(X_j)\right)}\cdot(w_k-w_k^*)_+\cdot 1_{E_n}\Bigg\}\\
&\quad +2\cdot\mathbf{E}\Bigg\{\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\sum_{k=1}^{K_n}\sup_{\substack{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})\in\mathcal{W},\\ \bar{k}\in\{1,\ldots,K_n\}}}\frac{1}{n}\sum_{j=1}^{n}(-\epsilon_j)\cdot\frac{Y_j\cdot\beta_n\cdot T_1 f_{\bar{\vartheta}_{\bar{k}}}(X_j)}{1+\exp\left(Y_j\cdot f_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})}(X_j)\right)}\cdot(w_k^*-w_k)_+\cdot 1_{E_n}\Bigg\}\\
&\le 2\cdot\mathbf{E}\Bigg\{\sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ k\in\{1,\ldots,K_n\}}}\frac{1}{n}\sum_{j=1}^{n}\epsilon_j\cdot\frac{Y_j\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j)}{1+\exp\left(Y_j\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j)\right)}\cdot\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\sum_{k=1}^{K_n}(w_k-w_k^*)_+\cdot 1_{E_n}\Bigg\}\\
&\quad +2\cdot\mathbf{E}\Bigg\{\sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ k\in\{1,\ldots,K_n\}}}\frac{1}{n}\sum_{j=1}^{n}(-\epsilon_j)\cdot\frac{Y_j\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_j)}{1+\exp\left(Y_j\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j)\right)}\cdot\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\sum_{k=1}^{K_n}(w_k^*-w_k)_+\cdot 1_{E_n}\Bigg\}
\end{aligned}
$$

$$\leq 4 \cdot \mathbf{E}\Bigg\{ \sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ k\in\{1,\dots,K_n\}}} \frac{1}{n}\sum_{j=1}^{n} \epsilon_j \cdot \frac{Y_j \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X_j)}{1+\exp\left(Y_j \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_j)\right)} \Bigg\}.$$

Where in the second to last inequality we have used that

$$\sup_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})\in\mathcal{W},\bar{k}\in\{1,\dots,K_n\}} \frac{1}{n}\sum_{j=1}^{n}(\pm\epsilon_j) \cdot \frac{Y_j \cdot \beta_n \cdot T_1 f_{\bar{\vartheta}_{\bar{k}}}(X_j)}{1+\exp\left(Y_j \cdot f_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})}(X_j)\right)} \geq 0$$

since there exists $(\tilde{\mathbf{w}},\tilde{\boldsymbol{\vartheta}}) \in \mathcal{W}$ such that $f_{\tilde{\vartheta}_1}(X_j) = 0$ for $j \in \{1,\dots,n\}$.

In the *ninth step of the proof* we derive an upper bound on

$$\mathbf{E}\Bigg\{ \sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ k\in\{1,\dots,K_n\}}} \left(\frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot \frac{1}{1+\exp(Y_i \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_i))} \cdot Y_i \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X_i) \right) \Bigg\}.$$

To do this we use a contraction style argument. Because of the independence of the random variables we can compute the expectation by first computing the expectation with respect to ϵ_1 and then by computing the expectation with respect to all other random variables. This yields that the last term above is equal to

$$\begin{aligned}
&\frac{1}{2} \cdot \mathbf{E}\Bigg\{ \sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ k\in\{1,\dots,K_n\}}} \frac{1}{n}\Bigg(\sum_{i=2}^{n} \epsilon_i \cdot \frac{1}{1+\exp(Y_i \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_i))} \cdot Y_i \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X_i) \\
&\qquad\qquad + \frac{1}{1+\exp(Y_1 \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_1))} \cdot Y_1 \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X_1) \Bigg) \Bigg\} \\
&\quad + \frac{1}{2} \cdot \mathbf{E}\Bigg\{ \sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ k\in\{1,\dots,K_n\}}} \frac{1}{n}\Bigg(\sum_{i=2}^{n} \epsilon_i \cdot \frac{1}{1+\exp(Y_i \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_i))} \cdot Y_i \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X_i) \\
&\qquad\qquad - \frac{1}{1+\exp(Y_1 \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_1))} \cdot Y_1 \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X_1) \Bigg) \Bigg\} \\
&= \frac{1}{2} \cdot \mathbf{E}\Bigg\{ \sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta}),(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})\in\mathcal{W},\\ k,\bar{k}\in\{1,\dots,K_n\}}} \frac{1}{n}\Bigg(\sum_{i=2}^{n} \epsilon_i \cdot \frac{1}{1+\exp(Y_i \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_i))} \cdot Y_i \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X_i) \\
&\qquad\qquad + \frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot \frac{1}{1+\exp(Y_i \cdot f_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})}(X_i))} \cdot Y_i \cdot \beta_n \cdot T_1 f_{\bar{\vartheta}_{\bar{k}}}(X_i) \\
&\qquad\qquad + \frac{1}{1+\exp(Y_1 \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_1))} \cdot Y_1 \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X_1) \\
&\qquad\qquad - \frac{1}{1+\exp(Y_1 \cdot f_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})}(X_1))} \cdot Y_1 \cdot \beta_n \cdot T_1 f_{\bar{\vartheta}_{\bar{k}}}(X_1) \Bigg) \Bigg\} \\
&\leq \frac{1}{2} \cdot \mathbf{E}\Bigg\{ \sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta}),(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})\in\mathcal{W},\\ k,\bar{k}\in\{1,\dots,K_n\}}} \frac{1}{n}\Bigg(\sum_{i=2}^{n} \epsilon_i \cdot \frac{1}{1+\exp(Y_i \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_i))} \cdot Y_i \cdot \beta_n \cdot T_1 f_{\vartheta_k}(X_i)
\end{aligned}$$

$$+\sum_{i=2}^{n}\epsilon_i\cdot\frac{1}{1+\exp(Y_i\cdot f_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})}(X_i))}\cdot Y_i\cdot\beta_n\cdot T_1 f_{\bar{\vartheta}_{\bar{k}}}(X_i)$$

$$+\beta_n\cdot|f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_1)-f_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})}(X_1)|+\beta_n\cdot|T_1 f_{\vartheta_k}(X_1)-T_1 f_{\bar{\vartheta}_{\bar{k}}}(X_1)|\Bigg)\Bigg\}.$$

The sum inside the supremum above does not change its value if $((\mathbf{w},\boldsymbol{\vartheta}),k)$ is interchanged with $((\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}}),\bar{k})$. Consequently we can assume without loss of generality that $f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_1)-f_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})}(X_1)$ is positive or that it is negative. Set

$$\bar{\epsilon}_1=\bar{\epsilon}_1(\epsilon_1,X_1,Y_1,\ldots,X_n,Y_n)=\epsilon_1$$

if the functions $f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_1)-f_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})}(X_1)$ and $\beta_n\cdot T_1 f_{\vartheta_k}(X_1)-\beta_n\cdot T_1 f_{\bar{\vartheta}_{\bar{k}}}(X_1)$ which "attain" the above supremum have the same sign, and set it equal to $-\epsilon_1$ otherwise. Then the right-hand side above is equal to

$$\begin{aligned}
&\frac{1}{2}\cdot\mathbf{E}\Bigg\{\sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta}),(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})\in\mathcal{W},\\ k,\bar{k}\in\{1,\ldots,K_n\}}}\frac{1}{n}\Bigg(\sum_{i=2}^{n}\epsilon_i\cdot\frac{1}{1+\exp(Y_i\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_i))}\cdot Y_i\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_i)\\
&\qquad+\sum_{i=2}^{n}\epsilon_i\cdot\frac{1}{1+\exp(Y_i\cdot f_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})}(X_i))}\cdot Y_i\cdot\beta_n\cdot T_1 f_{\bar{\boldsymbol{\vartheta}}_{\bar{k}}}(X_i)\\
&\qquad+\beta_n\cdot\epsilon_1\cdot(f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_1)-f_{(\bar{\mathbf{w}},\bar{\boldsymbol{\vartheta}})}(X_1))+\bar{\epsilon}_1\cdot(\beta_n\cdot T_1 f_{\vartheta_k}(X_1)-\beta_n\cdot T_1 f_{\bar{\vartheta}_{\bar{k}}}(X_1))\Bigg)\Bigg\}\\
&\leq\mathbf{E}\Bigg\{\sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ k\in\{1,\ldots,K_n\}}}\frac{1}{n}\Bigg(\sum_{i=2}^{n}\epsilon_i\cdot\frac{1}{1+\exp(Y_i\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_i))}\cdot Y_i\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_i)\\
&\qquad+\beta_n\cdot\epsilon_1\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_1)+\bar{\epsilon}_1\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_1)\Bigg)\Bigg\},
\end{aligned}$$

where we have used that conditioned on $(X_1,Y_1),\ldots,(X_n,Y_n)$ the random vector $(\epsilon_1,\bar{\epsilon}_1)$ has the same distribution as the random vector $(-\epsilon_1,-\bar{\epsilon}_1)$.

Arguing in the same way for $k=2,\ldots,n$ we see that we can upper bound the term on the right-hand side above by

$$\begin{aligned}
&\mathbf{E}\Bigg\{\sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ k\in\{1,\ldots,K_n\}}}\frac{1}{n}\sum_{i=1}^{n}\Bigg(\beta_n\cdot\epsilon_i\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_i)+\bar{\epsilon}_i\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_i)\Bigg)\Bigg\}\\
&\leq\beta_n\cdot\mathbf{E}\Bigg\{\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(X_i)\Bigg\}\\
&\quad+\mathbf{E}\Bigg\{\sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ k\in\{1,\ldots,K_n\}}}\frac{1}{n}\sum_{i=1}^{n}\bar{\epsilon}_i\cdot\beta_n\cdot T_1 f_{\vartheta_k}(X_i)\Bigg\}\\
&\leq\beta_n\cdot\sup_{x_1,\ldots,x_n\in[0,1]^{d\times d}}\mathbf{E}\Bigg\{\sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x_i)\Bigg\}
\end{aligned}$$

$$
\begin{aligned}
&+ \sup_{x_1,\ldots,x_n\in[0,1]^{d\times d}} \mathbf{E}\left\{ \sup_{\substack{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},\\ k\in\{1,\ldots,K_n\}}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot \beta_n \cdot T_1 f_{\vartheta_k}(x_i) \right\} \\
&= \beta_n \cdot \sup_{x_1,\ldots,x_n\in[0,1]^{d\times d}} \mathbf{E}\left\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x_i) \right\} \\
&\quad + \sup_{x_1,\ldots,x_n\in[0,1]^{d\times d}} \mathbf{E}\left\{ \sup_{\vartheta\in\bar{\boldsymbol{\Theta}}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot \beta_n \cdot T_1 f_{\vartheta}(x_i) \right\},
\end{aligned}
$$

where the last equality follows from

$$
\{\beta_n \cdot T_1 f_{\vartheta_k} : (\mathbf{w},\boldsymbol{\vartheta}) \in \mathcal{W}, k \in \{1,\ldots,K_n\}\} = \{\beta_n \cdot T_1 f_{\vartheta_1} : (\mathbf{w},\boldsymbol{\vartheta}) \in \mathcal{W}\}. \tag{47}
$$

In the *tenth step of the proof* we show for $x_1, \ldots, x_n \in \mathbb{R}^{d\times d}$ arbitrary

$$
\mathbf{E}\left\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x_i) \right\} \leq \mathbf{E}\left\{ \sup_{\vartheta\in\bar{\boldsymbol{\Theta}}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot (\beta_n \cdot T_1 f_{\vartheta}(x_i)) \right| \right\}.
$$

We have

$$
\begin{aligned}
&\mathbf{E}\left\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot f_{(\mathbf{w},\boldsymbol{\vartheta})}(x_i) \right\} \\
&= \mathbf{E}\left\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot \sum_{j=1}^{K_n} w_j \cdot (\beta_n \cdot T_1 f_{\vartheta_j}(x_i)) \right\} \\
&= \mathbf{E}\left\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}} \sum_{j=1}^{K_n} w_j \cdot \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot (\beta_n \cdot T_1 f_{\vartheta_j}(x_i)) \right\} \\
&\leq \mathbf{E}\left\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}} \sum_{j=1}^{K_n} |w_j| \cdot \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot (\beta_n \cdot T_1 f_{\vartheta_j}(x_i)) \right| \right\} \\
&\leq \mathbf{E}\left\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W}} \sum_{j=1}^{K_n} |w_j| \cdot \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},k\in\{1,\ldots,K_n\}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot (\beta_n \cdot T_1 f_{\vartheta_k}(x_i)) \right| \right\} \\
&\leq 1 \cdot \mathbf{E}\left\{ \sup_{(\mathbf{w},\boldsymbol{\vartheta})\in\mathcal{W},k\in\{1,\ldots,K_n\}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot (\beta_n \cdot T_1 f_{\vartheta_k}(x_i)) \right| \right\} \\
&= \mathbf{E}\left\{ \sup_{\vartheta\in\bar{\boldsymbol{\Theta}}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot (\beta_n \cdot T_1 f_{\vartheta}(x_i)) \right| \right\},
\end{aligned}
$$

where the last equality followed from (47).

Summarizing the above results, the proof is complete. □

C. Supplement: Proof of Lemma 5

Set $\mathcal{F} = \{f_\vartheta : \vartheta \in \mathbf{\Theta}\}$. For $0 < \delta_n < 1$ and $x_1, \dots, x_n \in [0,1]^{d \times d}$ we have

$$\begin{aligned}
&\mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot \beta_n \cdot T_1(f(x_i))\right|\right\} \\
&= \beta_n \cdot \mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot T_1(f(x_i))\right|\right\} \\
&= \beta_n \cdot \int_0^1 \mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot T_1(f(x_i))\right| > t\right\} dt \\
&\leq \delta_n + \beta_n \cdot \int_{\delta_n}^1 \mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot T_1(f(x_i))\right| > t\right\} dt.
\end{aligned}$$

Applying a covering argument, which is frequently used in empirical process theory, cf. e.g. Kohler and Krzyżak (2024), we get that for any $t \geq \delta_n$ we have

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot T_1(f(x_i))\right| > t\right\} \\
&\leq \mathcal{M}_1\left(\frac{\delta_n}{2}, \{T_1 f : f \in \mathcal{F}\}, x_1^n\right) \cdot \sup_{f\in\mathcal{F}} \mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot T_1(f(x_i))\right| > \frac{t}{2}\right\}.
\end{aligned}$$

Application of Lemma 6 and Theorem 9.4 in Györfi et al. (2002) yields

$$\mathcal{M}_1\left(\frac{\delta_n}{2}, \{T_1 f : f \in \mathcal{F}\}, x_1^n\right) \leq c_{28} \cdot \left(\frac{c_{29}}{\delta_n}\right)^{c_{30}\cdot m^2\cdot h^2\cdot L_{max}^2 \cdot \log(m^2\cdot h^2\cdot L_{max})}.$$

By the inequality of Hoeffding and

$$|T_1(f(x))| \leq 1 \quad (x \in \mathbb{R}^d)$$

we have for any $f \in \mathcal{F}$

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot \beta_n \cdot T_1(f(x_i))\right| > t/2\right\} \leq 2 \cdot \exp\left(-\frac{2\cdot n\cdot (t/2)^2}{4}\right).$$

Hence we get

$$\begin{aligned}
&\sup_{x_1,\dots,x_n\in[0,1]^{d\times d}} \mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot \beta_n \cdot T_1(f(x_i))\right|\right\} \\
&\leq \delta_n + \beta_n \cdot \int_{\delta_n}^1 c_{28} \cdot \left(\frac{c_{29}}{\delta_n}\right)^{c_{30}\cdot m^2\cdot h^2\cdot L_{max}^2 \cdot \log(m^2\cdot h^2\cdot L_{max})} \cdot 2 \cdot \exp\left(-\frac{n\cdot\delta_n\cdot t}{8}\right) dt
\end{aligned}$$

$$\leq \delta_n + c_{28} \cdot \beta_n \cdot \left(\frac{c_{29}}{\delta_n}\right)^{c_{30} \cdot m^2 \cdot h^2 \cdot L_{max}^2 \cdot \log(m^2 \cdot h^2 \cdot L_{max})} \frac{16}{n \cdot \delta_n} \cdot \exp\left(-\frac{n \cdot \delta_n^2}{8}\right).$$

With

$$\delta_n = \sqrt{\frac{8}{n} \cdot \left(c_{30} \cdot m^2 \cdot h^2 \cdot L_{max}^2 \cdot \log(m^2 \cdot h^2 \cdot L_{max}) \cdot \log(c_{29} \cdot \sqrt{n}) + \log(\beta_n)\right)} \geq \frac{1}{\sqrt{n}}$$

(where the last inequality holds if we choose c_{29} and c_{30} large enough) we get the assertion. □