On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent *

Selina Drews† and Michael Kohler
Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: drews@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de

August 22, 2022

Abstract
Estimation of a multivariate regression function from independent and identically distributed data is considered. An estimate is defined which fits a deep neural network consisting of a large number of fully connected neural networks, which are computed in parallel, via gradient descent to the data. The estimate is over-parametrized in the sense that the number of its parameters is much larger than the sample size. It is shown that in case of a suitable random initialization of the network, a suitable small stepsize of the gradient descent, and a number of gradient descent steps which is slightly larger than the reciprocal of the stepsize of the gradient descent, the estimate is universally consistent in the sense that its expected L_2 error converges to zero for all distributions of the data where the response variable is square integrable.

AMS classification: Primary 62G08; secondary 62G20.

Key words and phrases: neural networks, nonparametric regression, over-parametrization, universal consistency.

1 Introduction

Deep neural networks belong nowadays to the most promising approaches in many different applications. They have been successfully applied, e.g., in image classification (cf., e.g., Krizhevsky, Sutskever and Hinton (2012)), text classification (cf., e.g., Kim (2014)), machine translation (cf., e.g., Wu et al. (2016)) or mastering of games (cf., e.g., Silver et al. (2017)).

In the last few years various results concerning the approximation power of deep neural networks (cf., e.g., Yarotsky (2017), Yarotsky and Zhevnerchuck (2020), Lu et al. (2020), Langer (2021b) and the literature cited therein) or concerning the statistical risk of corresponding least squares estimates (cf., e.g., Bauer and Kohler (2019), Kohler

*Running title: *Consistent neural networks*
†Corresponding author. Tel: +49-6151-16-23372, Fax:+49-6151-16-23381

and Krzyżak (2017), Schmidt-Hieber (2020), Kohler and Langer (2021), Langer (2021a), Imaizumi and Fukumizu (2019), Suzuki (2018), Suzuki and Nitanda (2019), and the literature cited therein) have been derived.

The above results ignore two important features of the typical application of deep neural networks: Firstly, in practice the estimates are computed using gradient descent and not (as in the theoretical results above) the principle of least squares. And secondly, often the applied neural networks are over-parametrized in the sense that the number of parameters is much larger than their sample size.

These two principles contradict classical theory. Nevertheless, to some surprise, they work very well in practice. In Bartlett, Montanari and Rakhlin (2021), the theory of this observation is explored in more detail. To do this, they put forward two hypotheses. The first hypothesis concerns the *tractability via over-parametrization.* It is conjectured that even if the objective function is non-convex, the hardness of the optimization problem depends on the relationship between the dimension of the parameter space (the number of optimization variables) and the sample size. Thus the tractability is given if and only if a model is chosen that is over-parametrized. This is in contrast to the classical assumption that statistical learning is achieved by restricting to linearly parameterized classes of functions and convex objectives. The second hypothesis concerns *generalization via implicit regularization.* In classical theory, one wanted to avoid over-parametrized neural networks and therefore restricted them to an under-parametrized regime or a suitable regularizing regime. It was assumed that a method that has too many degrees of freedom by perfectly interpolating noisy data cannot have a good generalization. However, it was observed in practice that over-parametrized models generalize well. This is very interesting since empirical evidence shows that an optimization task is simplified if the model is sufficiently over-parametrized.

Bartlett, Montanari and Rakhlin (2021) suggest that deep learning models can be divided into a simple component and a spiky component. The simple component is useful for prediction, and the spiky component is useful for overfitting. If the model is suitably over-parametrized, this interpolation does not affect the prediction accuracy.

Nonconvex empirical risk minimization problems in a linear regime are solved efficiently by gradient methods. In a linear regime, a parameterized function can be approximated exactly by its linearization over an initial parameter vector. Bartlett, Montanari and Rakhlin (2021) were able to show that for a suitable parameterization and initialization, a gradient method remains in the linear regime. Further, it leads to linear convergence of the empirical risk and to a solution whose prediction is well approximated by the linearization of the initialization. Especially, they showed that for two-layer networks in the linear regime, a suitably large over-parametrization together with a suitable initialization is sufficient. This theory is not able to capture training schemes in which the weights genuinely change.

One approach beyond the linear regime considered in Bartlett, Montanari and Rakhlin (2021) is the mean field limit. Here, the weights move in a nontrivial way during training, even though the network is infinitely wide. Using mean field theory, global convergence results can be proved for two-layer (Mei, Montanari, and Nguyen (2018), Chizat and Bach (2018)) and multi-layer neural networks (Nguyen and Pham (2020)).

Another approach is to consider the linearized evolution as a Taylor expansion of the first order, and then construct higher-order approximations. Other approaches are considered in Dyer and Gur-Ar (2019) and Hanin and Nica (2019).

It is well-known that gradient descent leads to a small empirical L_2 risk in over-parametrized neural networks, see, e.g., Allen-Zhu, Li and Song (2019), Kawaguchi and Huang (2019) and the literature cited therein. However, such results are in general not useful for the proof of the consistency of corresponding estimates, because it was shown in Kohler and Krzyżak (2021) that any estimate which interpolates the training data does not generalize well in a sense that its error does not converge to zero for sample size tending to infinity in case of a general design measure.

In the current paper, we analyze deep neural networks in the context of nonparametric regression. Here we consider an $\mathbb{R}^d \times \mathbb{R}$–valued random vector (X, Y) with $\mathbf{E}Y^2 < \infty$, where X is the so–called observation vector and Y is the so-called response variable. We assume that a sample of (X, Y), i.e., a data set

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}, \tag{1}$$

where (X, Y), (X_1, Y_1), ..., (X_n, Y_n) are i.i.d., is available. We are searching for an estimator

$$m_n(\cdot) = m_n(\cdot, \mathcal{D}_n) : \mathbb{R}^d \to \mathbb{R}$$

of the so–called regression function $m : \mathbb{R}^d \to \mathbb{R}$, $m(x) = \mathbf{E}\{Y|X = x\}$ such that the so–called L_2 error

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)$$

is "small" (cf., e.g., Györfi et al. (2002) for a systematic introduction to nonparametric regression and a motivation for the L_2 error).

In Section 2 we introduce an over-parametrized deep neural network estimate, where the weights are learned by gradient descent. Our main result is that in case we initialize our starting weights randomly in a proper way, and proceed with a suitable number of gradient descent steps with a sufficiently small constant stepsize, this estimate m_n is universally consistent in the sense that

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \to 0 \quad (n \to \infty)$$

holds for every distribution of (X, Y) with $\mathbf{E}Y^2 < \infty$.

For many years it is well-known that universally consistent regression estimates exist, see Stone (1977) for the first result in this respect and Györfi et al. (2002) for an extensive overview of such results. So it is not surprising that deep neural networks have this property, too. However, our main result presents interesting aspects of the application of gradient descent to deep neural networks which are useful for proving such universal consistency: Firstly, the over-parametrization is useful in our result since it ensures that a finite subset of the initially chosen inner weights have nice properties. Secondly, due to the fact that we use a small stepsize, gradient descent applied to a properly regularized empirical L_2 risk is able to adjust the outer weights in an optimal way. And thirdly, due to the fact that the number of gradient descent steps is only slightly larger than the reciprocal of the stepsize, the inner weights do not change drastically during our learning. Altogether this enables our estimate to perform a kind of representation guessing instead of representation learning.

In our proofs we use techniques that have been introduced in Braun et al. (2021) in the context of the analysis of gradient descent of neural networks with one hidden layer. These techniques have been also applied in Kohler and Krzyżak (2022) to analyze the performance of over-parametrized neural networks with one hidden layer. But in contrast to Kohler and Krzyżak (2022) we do not control the complexity of our estimate by using a strong regularization term. Instead we combine the techniques introduced in Braun et al. (2021) with the approach of Li, Gu and Ding (2021), which suggested to analyze the complexity of over-parametrized neural networks with metric entropy bounds.

Throughout this paper we will use the following notation: The sets of natural numbers, real numbers and nonnegative real numbers are denoted by $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$, respectively. For $z \in \mathbb{R}$, we denote the smallest integer greater than or equal to z by $\lceil z \rceil$. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$. For $f : \mathbb{R}^d \to \mathbb{R}$

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

is its supremum norm. 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–cover of $\mathcal{F}$ on x_1^n if for any $f \in \mathcal{F}$ there exists $i \in \{1, \dots, N\}$ such that

$$\left(\frac{1}{n} \sum_{k=1}^{n} |f(x_k) - f_i(x_k)|^p \right)^{1/p} < \varepsilon.$$

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

If A is a subset of $\mathbb{R}^d$ and $x \in \mathbb{R}^d$, then we set $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ otherwise. 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 and $\mathcal{F}$ is a set of such functions, then we set $(T_\beta f)(x) = T_\beta(f(x))$ and

$$T_\beta \mathcal{F} = \{T_\beta f \ : \ f \in \mathcal{F}\}.$$

In Section 2 we define our estimate. In Section 3 we present our main result concerning the universal consistency of our deep neural network estimate learned by gradient descent. The proof of the main result is given in Section 4.

2 Definiton of the estimate

Let $\sigma(x) = 1/(1 + e^{-x})$ be the logistic squasher. In the sequel we use a network topology where we compute a linear combination of K_n fully connected neural networks with L

layers and r neurons per layer, i.e., we define our neural network as follows: We set

$$f_{\mathbf{w}}(x) = \sum_{j=1}^{K_n} w_{1,1,j}^{(L)} \cdot f_{j,1}^{(L)}(x) \tag{2}$$

for some $w_{1,1,1}^{(L)}, \dots, w_{1,1,K_n}^{(L)} \in \mathbb{R}$, where $f_{j,1}^{(L)}$ are recursively defined by

$$f_{k,i}^{(l)}(x) = \sigma\left(\sum_{j=1}^{r} w_{k,i,j}^{(l-1)} \cdot f_{k,j}^{(l-1)}(x) + w_{k,i,0}^{(l-1)}\right) \tag{3}$$

for some $w_{k,i,0}^{(l-1)}, \dots, w_{k,i,r}^{(l-1)} \in \mathbb{R}$ $(l = 2, \dots, L)$ and

$$f_{k,i}^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)}\right) \tag{4}$$

for some $w_{k,i,0}^{(0)}, \dots, w_{k,i,d}^{(0)} \in \mathbb{R}$.

The above neural network consists of K_n fully connected neural networks with depth L, which are computed in parallel. These networks have r neurons in all layers except for the last layer, where they only have one neuron. In the k-th such network we denote the output of neuron i in the l-th layer by $f_{k,i}^{(l)}$, and the weight between neuron j in the $(l-1)$-th layer and neuron i in the l-th layer is denoted by $w_{k,i,j}^{(l-1)}$. The number of weights of the above neural network is given by

$$K_n \cdot (1 + (L-1) \cdot r \cdot (r+1) + r \cdot (d+1)).$$

In order to learn the weight vector $\mathbf{w} = (w_{k,i,j}^{(l)})_{k,i,j,l}$ of our neural network we apply gradient descent to a properly regularized empirical L_2 risk of our estimate. We initialize $\mathbf{w}^{(0)}$ by setting

$$w_{1,1,j}^{(L)} = 0 \quad \text{for } j = 1, \dots, K_n, \tag{5}$$

and by choosing all others weights randomly such that all weights $w_{k,i,j}^{(l)}$ are independently randomly chosen, and such that all weights $w_{k,i,j}^{(l)}$ with $1 \leq l < L$ are uniformly distributed on $[-20d \cdot (\log n)^2, 20d \cdot (\log n)^2]$, and all weights $w_{k,i,j}^{(0)}$ are uniformly distributed on $[-n^{\tau}, n^{\tau}]$ for some fixed $0 < \tau < 1/(d+1)$. Then we set $\alpha_n = c_1 \cdot \log n$, and compute

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)}) \quad (t = 0, \dots, t_n - 1)$$

where

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}}(X_i) - Y_i|^2 \cdot 1_{[-\alpha_n, \alpha_n]^d}(X_i) + c_2 \cdot \sum_{j=1}^{K_n} |w_{1,1,j}^{(L)}|^2$$

is the regularized empirical L_2 risk of the network $f_{\mathbf{w}}$ on the training data. The step size $\lambda_n > 0$ and the number t_n of gradient descent steps will be chosen in Theorem 1 below.

The estimate is defined by

$$m_n(x) = (T_{\beta_n} f_{\mathbf{w}^{(t_n)}}(x)) \cdot 1_{[-\alpha_n,\alpha_n]^d}(x),$$

i.e., as an estimate we use the neural network with the weight vector which we get after t_n gradient descent steps, and truncate this function on height $-\beta_n$ and β_n, and set it equal to zero outside of a cube. Here we set $\beta_n = c_3 \cdot \log n$.

Because of (5) we have

$$F_n(\mathbf{w}^{(0)}) = \frac{1}{n} \sum_{i=1}^{n} |Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i).$$

3 Main result

Our main result is the following theorem.

Theorem 1 *Let σ be the logistic squasher, and let $K_n, L, r \in \mathbb{N}$ and $\tau \in \mathbb{R}_+$. Assume $L \geq 2$, $r \geq 2d$, $0 < \tau < 1/(d+1)$*

$$\frac{K_n}{n^\kappa} \to 0 \quad (n \to \infty) \tag{6}$$

for some $\kappa > 0$,

$$\frac{K_n}{n^r \cdot \log n} \to \infty \quad (n \to \infty), \tag{7}$$

and set $\alpha_n = c_1 \cdot \log n$, $\beta_n = c_3 \cdot \log n$,

$$\lambda_n = \frac{1}{L_n} \quad \text{and} \quad t_n = \lceil c_4 \cdot L_n \cdot \log n \rceil \tag{8}$$

for some $L_n > 0$ which satisfies

$$L_n \geq (\log n)^{10 \cdot L + 10} \cdot K_n^{3/2}. \tag{9}$$

Let the estimate m_n be defined as in Section 2. Then we have

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \to 0 \quad (n \to \infty)$$

for every distribution of (X, Y) with $\mathbf{E}Y^2 < \infty$.

Remark 1. Condition (7) implies that K_n is asymptotically larger than n^r, consequently the number of parameters of our estimate is much larger than the sample size and our estimate is over-parametrized. Nevertheless it generalizes well on new independent data since its expected L_2 error converges to zero for sample size tending to infinity. In its definition we add a regularization term to the empirical L_2 risk, however, this regularization term is not really used to control the complexity of our estimate, it is only used to help us in analyzing the gradient descent. We control the complexity of our estimate by imposing bounds on the absolute values of the initial weights and by requiring that the number of gradient descent steps is not much larger than the reciprocal of the stepsize. In particular, the condition $0 < \tau < 1/(d+1)$ controls the range $[-n^\tau, n^\tau]$ of the weights $w_{k,i,j}^{(0)}$, and all initial weights of level $1, \dots, L-1$ are bounded in absolute value by some logarithmic term.

Remark 2. We need only a single initialization of our random starting weights in Theorem 1. This is due to the over-parametrization, which enables us to show that even with one single initialization there exists with probability close to one a finite subset of our K_n fully connected neural networks where the initial inner weights have some nice property.

Remark 3. In the proof of Theorem 1 we use that the inner weights do not change much during gradient descent and that gradient descent is able to find proper values for the outer weights of our network. In this sense our network is not based on representation learning, instead it is using a representation guessing.

4 Proofs

4.1 Auxiliary results

In this subsection we present various auxiliary results which we will need in the proof of Theorem 1.

Lemma 1 *Let $F : \mathbb{R}^K \to \mathbb{R}_+$ be a nonnegative differentiable function. Let $t \in \mathbb{N}$, $L > 0$, $\mathbf{a}_0 \in \mathbb{R}^K$ and set*

$$\lambda = \frac{1}{L}$$

and

$$\mathbf{a}_{k+1} = \mathbf{a}_k - \lambda \cdot (\nabla_{\mathbf{a}} F)(\mathbf{a}_k) \quad (k \in \{0, 1, \dots, t-1\}).$$

Assume

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\| \leq \sqrt{2 \cdot t \cdot L \cdot \max\{F(\mathbf{a}_0), 1\}} \tag{10}$$

for all $\mathbf{a} \in \mathbb{R}^K$ with $\|\mathbf{a} - \mathbf{a}_0\| \leq \sqrt{2 \cdot t \cdot \max\{F(\mathbf{a}_0), 1\}/L}$, and

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}) - (\nabla_{\mathbf{a}} F)(\mathbf{b})\| \leq L \cdot \|\mathbf{a} - \mathbf{b}\| \tag{11}$$

for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$ satisfying

$$\|\mathbf{a} - \mathbf{a}_0\| \leq \sqrt{8 \cdot \frac{t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}} \quad \textit{and} \quad \|\mathbf{b} - \mathbf{a}_0\| \leq \sqrt{8 \cdot \frac{t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}}. \tag{12}$$

Then we have

$$\|\mathbf{a}_k - \mathbf{a}_0\| \leq \sqrt{2 \cdot \frac{k}{L} \cdot (F(\mathbf{a}_0) - F(\mathbf{a}_k))} \quad \textit{for all } k \in \{1, \dots, t\},$$

$$\sum_{k=0}^{s-1} \|\mathbf{a}_{k+1} - \mathbf{a}_k\|^2 \leq \frac{2}{L} \cdot (F(\mathbf{a}_0) - F(\mathbf{a}_s)) \quad \textit{for all } s \in \{1, \dots, t\}$$

and

$$F(\mathbf{a}_k) \leq F(\mathbf{a}_{k-1}) - \frac{1}{2L} \cdot \|\nabla_{\mathbf{a}} F(\mathbf{a}_{k-1})\|^2 \quad \textit{for all } k \in \{1, \dots, t\}.$$

Proof. The result follows from Lemma 2 in Braun et al. (2021) and its proof. □

Lemma 2 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be bounded and differentiable, and assume that its derivative is bounded. Let $\alpha_n \geq 1$, $t_n \geq L_n$, $\gamma_n^* \geq 1$, $B_n \geq 1$, $r \geq 2d$,*

$$|w_{1,1,k}^{(L)}| \leq \gamma_n^* \quad (k = 1, \dots, K_n), \tag{13}$$

$$|w_{k,i,j}^{(l)}| \leq B_n \quad \textit{for } l = 1, \dots, L-1 \tag{14}$$

and

$$\|\mathbf{w} - \mathbf{v}\|_\infty^2 \leq \frac{2t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}. \tag{15}$$

Then we have

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w})\| \leq c_5 \cdot K_n^{3/2} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2 \cdot \sqrt{\frac{t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}}.$$

Proof. We have

$$\begin{aligned}
&\|\nabla_{\mathbf{w}} F_n(\mathbf{w})\|^2 \\
&= \sum_{k,i,j,l} \left(\frac{2}{n} \sum_{s=1}^n (Y_s - f_{\mathbf{w}}(X_s)) \cdot 1_{[-\alpha_n, \alpha_n]^d}(X_s) \cdot \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(X_s) \right. \\
&\qquad \left. + \frac{\partial}{\partial w_{k,i,j}^{(l)}} \left(c_2 \cdot \sum_{r=1}^{K_n} |w_{1,1,r}^{(L)}|^2 \right) \right)^2 \\
&\leq 8 \cdot \sum_{k,i,j,l} \frac{1}{n} \sum_{s=1}^n (Y_s - f_{\mathbf{w}}(X_s))^2 \cdot 1_{[-\alpha_n, \alpha_n]^d}(X_s) \cdot \left(\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \\
&\qquad + 8 \cdot c_2^2 \cdot K_n \cdot (\gamma_n^*)^2 \\
&\leq c_6 \cdot K_n \cdot L \cdot r^2 \cdot d \cdot \max_{k,i,j,l,s} \left(\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \cdot 1_{[-\alpha_n, \alpha_n]^d}(X_s) \\
&\qquad \cdot \frac{1}{n} \sum_{s=1}^n (Y_s - f_{\mathbf{w}}(X_s))^2 \cdot 1_{[-\alpha_n, \alpha_n]^d}(X_s) + 8 \cdot c_2^2 \cdot K_n \cdot (\gamma_n^*)^2.
\end{aligned}$$

The chain rule implies

$$\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(x) = \sum_{s_{l+2}=1}^{r} \cdots \sum_{s_{L-1}=1}^{r} f_{k,j}^{(l)}(x) \cdot \sigma' \left(\sum_{t=1}^r w_{k,i,t}^{(l)} \cdot f_{k,t}^{(l)}(x) + w_{k,i,0}^{(l)} \right)$$

$$
\begin{aligned}
&\cdot w^{(l+1)}_{k,s_{l+2},i} \cdot \sigma' \left(\sum_{t=1}^{r} w^{(l+1)}_{k,s_{l+2},t} \cdot f^{(l+1)}_{k,t}(x) + w^{(l+1)}_{k,s_{l+2},0} \right) \cdot w^{(l+2)}_{k,s_{l+3},s_{l+2}} \\
&\cdot \sigma' \left(\sum_{t=1}^{r} w^{(l+2)}_{k,s_{l+3},t} \cdot f^{(l+2)}_{k,t}(x) + w^{(l+2)}_{k,s_{l+3},0} \right) \cdots w^{(L-2)}_{k,s_{L-1},s_{L-2}} \\
&\cdot \sigma' \left(\sum_{t=1}^{r} w^{(L-2)}_{k,s_{L-1},t} \cdot f^{(L-2)}_{k,t}(x) + w^{(L-2)}_{k,s_{L-1},0} \right) \cdot w^{(L-1)}_{k,1,s_{L-1}} \\
&\cdot \sigma' \left(\sum_{t=1}^{r} w^{(L-1)}_{k,1,t} \cdot f^{(L-1)}_{k,t}(x) + w^{(L-1)}_{k,1,0} \right) \cdot w^{(L)}_{1,1,k}, \qquad (16)
\end{aligned}
$$

where we have used the abbreviations

$$
f^{(0)}_{k,j}(x) = \begin{cases} x^{(j)} & \text{if } j \in \{1, \dots, d\} \\ 1 & \text{if } j = 0 \end{cases}
$$

and

$$
f^{(l)}_{k,0}(x) = 1 \quad (l = 1, \dots, L-1).
$$

Using the assumptions of Lemma 2 we can conclude

$$
\max_{k,i,j,l,s} \left(\frac{\partial f_{\mathbf{w}}}{\partial w^{(l)}_{k,i,j}}(X_s) \right)^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \le c_6 \cdot r^{2L} \cdot \max\{\|\sigma'\|^{2L}_\infty, 1\} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2.
$$

By the Lipschitz continuity of σ together with the assumptions of Lemma 2 we get for any $x \in [-\alpha_n, \alpha_n]^d$

$$
|f_{\mathbf{w}}(x) - f_{\mathbf{v}}(x)| \le 2 \cdot K_n \cdot \max\{\|\sigma'\|^L_\infty, 1\} \cdot \gamma_n^* \cdot (2r+1)^L \cdot B_n^L \cdot \alpha_n \cdot \max\{\|\sigma\|_\infty, 1\} \cdot \|\mathbf{w} - \mathbf{v}\|_\infty.
$$

(cf., e.g., Lemma 5 in Kohler and Krzyżak (2021) for a related proof). This implies

$$
\begin{aligned}
&\frac{1}{n} \sum_{s=1}^{n} (Y_s - f_{\mathbf{w}}(X_s))^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \\
&\le 2 \cdot F_n(\mathbf{v}) + \frac{2}{n} \sum_{s=1}^{n} (f_{\mathbf{v}}(X_s) - f_{\mathbf{w}}(X_s))^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \\
&\le 2 \cdot F_n(\mathbf{v}) + 8 \cdot \max\{\|\sigma'\|^{2L}_\infty, 1\} \cdot K_n^2 \cdot \gamma_n^{*2} \cdot (2r+1)^{2L} \cdot B_n^{2L} \cdot \alpha_n^2 \cdot \max\{\|\sigma\|_\infty, 1\}^2 \\
&\quad \cdot \frac{2t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}.
\end{aligned}
$$

Summarizing the above results, the proof is complete. □

Lemma 3 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be bounded and differentiable, and assume that its derivative is Lipschitz continuous and bounded. Let $\alpha_n \ge 1$, $t_n \ge L_n$, $\gamma_n^* \ge 1$, $B_n \ge 1$, $r \ge 2d$ and assume*

$$
|\max\{(\mathbf{w}_1)^{(L)}_{1,1,k}, (\mathbf{w}_2)^{(L)}_{1,1,k}\}| \le \gamma_n^* \quad (k = 1, \dots, K_n), \qquad (17)
$$

$$|\max\{(\mathbf{w}_1)^{(l)}_{k,i,j},(\mathbf{w}_2)^{(l)}_{k,i,j}\}|\leq B_n \quad \textit{for } l=1,\dots,L-1 \tag{18}$$

and

$$\|\mathbf{w}_2-\mathbf{v}\|^2\leq 8\cdot\frac{t_n}{L_n}\cdot\max\{F_n(\mathbf{v}),1\}. \tag{19}$$

Then we have

$$\begin{aligned}&\|(\nabla_{\mathbf{w}}F_n)(\mathbf{w}_1)-(\nabla_{\mathbf{w}}F_n)(\mathbf{w}_2)\|\\ &\leq c_7\cdot\max\{\sqrt{F_n(\mathbf{v})},1\}\cdot(\gamma_n^*)^2\cdot B_n^{3L}\cdot\alpha_n^3\cdot K_n^{3/2}\cdot\sqrt{\frac{t_n}{L_n}}\cdot\|\mathbf{w}_1-\mathbf{w}_2\|.\end{aligned}$$

Proof. We have

$$\begin{aligned}
&\|\nabla_{\mathbf{w}}F_n(\mathbf{w}_1)-\nabla_{\mathbf{w}}F_n(\mathbf{w}_2)\|^2\\
&=\sum_{k,i,j,l}\Bigg(\frac{2}{n}\sum_{s=1}^n(Y_s-f_{\mathbf{w}_1}(X_s))\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)\cdot\frac{\partial f_{\mathbf{w}_1}}{\partial w^{(l)}_{k,i,j}}(X_s)\\
&\qquad\qquad+\frac{\partial}{\partial w^{(l)}_{k,i,j}}\left(c_2\cdot\sum_{r=1}^{K_n}|(\mathbf{w}_1)^{(L)}_{1,1,r}|^2\right)\\
&\quad-\sum_{k,i,j,l}\Bigg(\frac{2}{n}\sum_{s=1}^n(Y_s-f_{\mathbf{w}_2}(X_s))\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)\cdot\frac{\partial f_{\mathbf{w}_2}}{\partial w^{(l)}_{k,i,j}}(X_s)\\
&\qquad\qquad+\frac{\partial}{\partial w^{(l)}_{k,i,j}}\left(c_2\cdot\sum_{r=1}^{K_n}|(\mathbf{w}_2)^{(L)}_{1,1,r}|^2\right)\Bigg)^2\\
&\leq 16\cdot\sum_{k,i,j,l}\left(\frac{1}{n}\sum_{s=1}^n(f_{\mathbf{w}_2}(X_s)-f_{\mathbf{w}_1}(X_s))\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)\cdot\frac{\partial f_{\mathbf{w}_1}}{\partial w^{(l)}_{k,i,j}}(X_s)\right)^2\\
&\quad+16\cdot\sum_{k,i,j,l}\left(\frac{1}{n}\sum_{s=1}^n(Y_s-f_{\mathbf{w}_2}(X_s))\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)\cdot\left(\frac{\partial f_{\mathbf{w}_1}}{\partial w^{(l)}_{k,i,j}}(X_s)-\frac{\partial f_{\mathbf{w}_2}}{\partial w^{(l)}_{k,i,j}}(X_s)\right)\right)^2\\
&\quad+8\cdot c_2^2\cdot\|\mathbf{w}_1-\mathbf{w}_2\|^2\\
&\leq 16\cdot\sum_{k,i,j,l}\max_{s=1,\dots,n}\left(\frac{\partial f_{\mathbf{w}_1}}{\partial w^{(l)}_{k,i,j}}(X_s)\right)^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)\\
&\qquad\qquad\cdot\frac{1}{n}\sum_{s=1}^n(f_{\mathbf{w}_2}(X_s)-f_{\mathbf{w}_1}(X_s))^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)\\
&\quad+16\cdot\frac{1}{n}\sum_{s=1}^n(Y_s-f_{\mathbf{w}_2}(X_s))^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)\\
&\qquad\qquad\cdot\sum_{k,i,j,l}\max_{s=1,\dots,n}\left(\frac{\partial f_{\mathbf{w}_1}}{\partial w^{(l)}_{k,i,j}}(X_s)-\frac{\partial f_{\mathbf{w}_2}}{\partial w^{(l)}_{k,i,j}}(X_s)\right)^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)
\end{aligned}$$

$$+8 \cdot c_2^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|_\infty^2.$$

From the proof of Lemma 2 we can conclude

$$\begin{aligned}
&\sum_{k,i,j,l} \max_{s=1,\dots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \\
&\leq c_8 \cdot K_n \cdot L \cdot r^2 \cdot d \cdot r^{2L} \cdot \max\{\|\sigma'\|_\infty^{2L}, 1\} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2,
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(X_s) - f_{\mathbf{w}_1}(X_s))^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \\
&\leq 4 \cdot \max\{\|\sigma'\|_\infty^{2L}, 1\} \cdot K_n^2 \cdot (2r+1)^{2L} \cdot (\gamma_n^*)^2 \cdot B_n^{2L} \cdot \alpha_n^2 \cdot \max\{\|\sigma\|_\infty, 1\}^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n} \sum_{s=1}^n (Y_s - f_{\mathbf{w}_2}(X_s))^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \\
&\leq 2 \cdot F_n(\mathbf{v}) + 8 \cdot \max\{\|\sigma'\|_\infty^{2L}, 1\} \cdot K_n^2 \cdot (2r+1)^{2L} \cdot (\gamma_n^*)^2 \cdot B_n^{2L} \cdot \alpha_n^2 \cdot \max\{\|\sigma\|_\infty, 1\}^2 \\
&\quad \cdot \frac{8 t_n}{L_n} \cdot \max\{F_n(v), 1\}.
\end{aligned}$$

So it remains to bound

$$\sum_{k,i,j,l} \max_{s=1,\dots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s).$$

By (16) we know that

$$\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(x)$$

is for fixed $x \in [-\alpha_n, \alpha_n]^d$ a sum of products of Lipschitz continuous functions (considered as functions of $\mathbf{w}$). Arguing as in the proof of Lemma 6 in Kohler and Krzyżak (2021) we can show that we have for any $x \in [-\alpha_n, \alpha_n]^d$

$$\left| \frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(x) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(x) \right| \leq c_9 \cdot B_n^{2L} \cdot \gamma_n^* \cdot \alpha_n \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|,$$

which implies

$$\begin{aligned}
&\sum_{k,i,j,l} \max_{s=1,\dots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \\
&\leq c_{10} \cdot K_n \cdot L \cdot r^2 \cdot d \cdot B_n^{4L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^4 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2.
\end{aligned}$$

Summarizing the above results we get the assertion. □

Lemma 4 *Let $\alpha \geq 1$, $\beta > 0$ and let $A, B, C \geq 1$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be k-times differentiable such that all derivatives up to order k are bounded on $\mathbb{R}$. Let $\mathcal{F}$ be the set of all functions $f_{\mathbf{w}}$ defined by (2)–(4) where the weight vector $\mathbf{w}$ satsifies*

$$\sum_{j=1}^{K_n} |w_{1,1,j}^{(L)}| \leq C, \tag{20}$$

$$|w_{k,i,j}^{(l)}| \leq B \quad (k \in \{1, \dots, K_n\}, i, j \in \{1, \dots, r\}, l \in \{1, \dots, L-1\}) \tag{21}$$

and

$$|w_{k,i,j}^{(0)}| \leq A \quad (k \in \{1, \dots, K_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d\}). \tag{22}$$

Then we have for any $1 \leq p < \infty$, $0 < \epsilon < \beta$ and $x_1^n \in \mathbb{R}^d$

$$\mathcal{N}_p\left(\epsilon, \{T_\beta f \cdot 1_{[-\alpha,\alpha]^d} \,:\, f \in \mathcal{F}\}, x_1^n\right) \leq \left(c_{11} \cdot \frac{\beta^p}{\epsilon^p}\right)^{c_{12} \cdot \alpha^d \cdot B^{(L-1)\cdot d} \cdot A^d \cdot \left(\frac{C}{\epsilon}\right)^{d/k} + c_{13}}.$$

Proof. In the *first step* of the proof we show for any $f_{\mathbf{w}} \in \mathcal{F}$, any $x \in \mathbb{R}^d$ and any $s_1, \dots, s_k \in \{1, \dots, d\}$

$$\left| \frac{\partial^k f_{\mathbf{w}}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \right| \leq c_{14} \cdot C \cdot B^{(L-1)\cdot k} \cdot A^k =: c. \tag{23}$$

The definition of $f_{\mathbf{w}}$ implies

$$\frac{\partial^k f_{\mathbf{w}}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) = \sum_{j=1}^{K_n} w_{1,1,j}^{(L)} \cdot \frac{\partial^k f_{j,1}^{(L)}(x)}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x),$$

hence (23) is implied by

$$\left| \frac{\partial^k f_{j,1}^{(L)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \right| \leq c_{15} \cdot B^{(L-1)\cdot k} \cdot A^k. \tag{24}$$

We have

$$\begin{aligned} \frac{\partial f_{k,i}^{(l)}}{\partial x^{(s)}}(x) &= \sigma'\left(\sum_{t=1}^r w_{k,i,t}^{(l-1)} \cdot f_{k,t}^{(l-1)}(x) + w_{k,i,0}^{(l-1)}\right) \cdot \sum_{j=1}^r w_{k,i,j}^{(l-1)} \cdot \frac{\partial f_{k,j}^{(l-1)}}{\partial x^{(s)}}(x) \\ &= \sum_{j=1}^r w_{k,i,j}^{(l-1)} \cdot \sigma'\left(\sum_{t=1}^r w_{k,i,t}^{(l-1)} \cdot f_{k,t}^{(l-1)}(x) + w_{k,i,0}^{(l-1)}\right) \cdot \frac{\partial f_{k,j}^{(l-1)}}{\partial x^{(s)}}(x) \end{aligned}$$

and

$$\frac{\partial f_{k,i}^{(1)}}{\partial x^{(s)}}(x) = \sigma'\left(\sum_{j=1}^{d} w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)}\right) \cdot w_{k,i,s}^{(0)}.$$

By the product rule of derivation we can conclude for $l > 1$ that

$$\frac{\partial^k f_{k,i}^{(l)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \tag{25}$$

is a sum of at most $r \cdot (r+k)^{k-1}$ terms of the form

$$w \cdot \sigma^{(s)}\left(\sum_{j=1}^{r} w_{k,i,j}^{(l-1)} \cdot f_{k,j}^{(l-1)}(x) + w_{k,i,0}^{(l-1)}\right) \cdot \frac{\partial^{t_1} f_{k,j_1}^{(l-1)}}{\partial x^{(r_{1,1})} \dots \partial x^{(r_{1,t_1})}}(x) \cdot \dots \cdot \frac{\partial^{t_s} f_{k,j_s}^{(l-1)}}{\partial x^{(r_{s,1})} \dots \partial x^{(r_{s,t_s})}}(x)$$

where we have $s \in \{1, \dots, k\}$, $|w| \leq B^s$ and $t_1 + \dots + t_s = k$. Furthermore

$$\frac{\partial^k f_{k,i}^{(1)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x)$$

is a given by

$$\prod_{j=1}^{k} w_{k,i,s_j}^{(0)} \cdot \sigma^{(k)}\left(\sum_{t=1}^{d} w_{k,i,t}^{(0)} \cdot x^{(t)} + w_{k,i,0}^{(0)}\right).$$

Because of the boundedness of the derivatives of σ we can conclude from (22)

$$\left|\frac{\partial^k f_{k,i}^{(1)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x)\right| \leq c_{16} \cdot A^k$$

for all $k \in \mathbb{N}$ and $s_1, \dots, s_k \in \{1, \dots, d\}$.

Recursively we can conclude from the above representation of (25) that we have

$$\left|\frac{\partial^k f_{k,i}^{(l)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x)\right| \leq c_{17} \cdot B^{(l-1)\cdot k} \cdot A^k.$$

Setting $l = L$ we get (24).

In the *second step* of the proof we show

$$\mathcal{N}_p\left(\epsilon, \{T_\beta f \cdot 1_{[-\alpha,\alpha]^d} : f \in \mathcal{F}\}, x_1^n\right) \leq \mathcal{N}_p\left(\frac{\epsilon}{2}, T_\beta \mathcal{G} \circ \Pi, x_1^n\right), \tag{26}$$

where $\mathcal{G}$ is the set of all polynomials of degree less than or equal to $k-1$ which vanish outside of $[-\alpha, \alpha]^d$ and Π is the family of all partitions of $\mathbb{R}^d$ which consist of a partition of $[-\alpha, \alpha]^d$ into K many cubes of sidelenght

$$\left(c_{19} \cdot \frac{\epsilon}{c}\right)^{1/k}$$

where $c_{19} = c_{19}(d, k)$ is a suitable small constant greater than zero and the additional set $\mathbb{R}^d \setminus [-\alpha, \alpha]^d$.

A standard bound on the remainder of a multivariate Taylor polynomial together with (23) shows that for each $f_{\mathbf{w}}$ we can find $g \in \mathcal{G} \circ \Pi$ such that

$$|f_{\mathbf{w}}(x) - g(x)| \leq \frac{\epsilon}{2}$$

holds for all $x \in [-\alpha, \alpha]^d$, which implies (26).

In the *third step* of the proof we show the assertion of Lemma 4. Since $\mathcal{G} \circ \Pi$ is a linear vector space of dimension less than or equal to

$$c_{20} \cdot \alpha^d \cdot \left(\frac{c}{\epsilon}\right)^{d/k}$$

we conclude from Theorem 9.4 and Theorem 9.5 in Györfi et al. (2002),

$$\mathcal{N}_p(\frac{\epsilon}{2}, T_\beta \mathcal{G} \circ \Pi, x_1^n) \leq 3\left(\frac{2e(2\beta)^p}{(\epsilon/2)^p} \log\left(\frac{3e(2\beta)^p}{(\epsilon/2)^p}\right)\right)^{c_{20} \cdot \alpha^d \cdot \left(\frac{c}{\epsilon}\right)^{d/k} + 1}.$$

Together with (26) this implies the assertion. □

Lemma 5 *Let σ be the logistic squasher and let $0 < \delta \leq 1$, $1 \leq \alpha_n \leq \log n$, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ with*

$$v^{(l)} - u^{(l)} \geq 2\delta \quad \textit{for } l \in \{1, \dots, d\}$$

and $x \in [-\alpha_n, \alpha_n]^d$. Let $L, r, n \in \mathbb{N}$ with $L \geq 2$, $r \geq 2 \cdot d$, $n \geq 8d$ and $n \geq \exp(r+1)$. Let

$$f_{\mathbf{w}}(x) = f_{1,1}^{(L)}(x)$$

where $f_{k,i}^{(l)}(x)$ are recursively defined by (3) and (4).

Assume

$$w_{1,j,j}^{(0)} = \frac{4d \cdot (\log n)^2}{\delta} \quad \textit{and} \quad w_{1,j,0}^{(0)} = -\frac{4d \cdot (\log n)^2}{\delta} \cdot u^{(j)} \quad \textit{for } j \in \{1, \dots, d\}, \tag{27}$$

$$w_{1,j+d,j}^{(0)} = -\frac{4d \cdot (\log n)^2}{\delta} \quad \textit{and} \quad w_{1,j+d,0}^{(0)} = \frac{4d \cdot (\log n)^2}{\delta} \cdot v^{(j)} \quad \textit{for } j \in \{1, \dots, d\}, \tag{28}$$

$$w_{1,s,t}^{(0)} = 0 \quad \textit{if } s \leq 2d, s \neq t, s \neq t+d \textit{ and } t > 0, \tag{29}$$

$$w_{1,1,t}^{(1)} = 8 \cdot (\log n)^2 \quad \textit{for } t \in \{1, \dots, 2d\}, \tag{30}$$

$$w_{1,1,0}^{(1)} = -8 \cdot (\log n)^2 \left(2d - \frac{1}{2}\right), \tag{31}$$

$$w_{1,1,t}^{(1)} = 0 \quad \textit{for } t > 2d, \tag{32}$$

$$w_{1,1,1}^{(l)} = 6 \cdot (\log n)^2 \quad \textit{for } l \in \{2, \ldots, L\}, \tag{33}$$

$$w_{1,1,0}^{(l)} = -3 \cdot (\log n)^2 \quad \textit{for } l \in \{2, \ldots, L\} \tag{34}$$

and

$$w_{1,1,t}^{(l)} = 0 \quad \textit{for } t > 1 \textit{ and } l \in \{2, \ldots, L\}. \tag{35}$$

Let $\bar{\mathbf{w}}$ *be such that*

$$|\bar{w}_{1,i,j}^{(l)} - w_{1,i,j}^{(l)}| \leq \log n \quad \textit{for all } l = 0, \ldots, L-1. \tag{36}$$

Then, we have

$$f_{\bar{\mathbf{w}}}(x) \geq 1 - \frac{1}{n} \textit{ if } x \in [u^{(1)} + \delta, v^{(1)} - \delta] \times \cdots \times [u^{(d)} + \delta, v^{(d)} - \delta]$$

and

$$f_{\bar{\mathbf{w}}}(x) \leq \frac{1}{n} \textit{ if } x^{(i)} \notin [u^{(i)} - \delta, v^{(i)} + \delta] \textit{ for some } i \in \{1, \ldots, d\}.$$

Proof. At the beginning we define

$$\bar{f}_{k,i}^{(l)}(x) = \sigma\left(\sum_{j=1}^{r} \bar{w}_{k,i,j}^{(l-1)} \cdot \bar{f}_{k,j}^{(l-1)}(x) + \bar{w}_{k,i,0}^{(l-1)}\right)$$

for $l = 2, \ldots, L$ and

$$\bar{f}_{k,i}^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} \bar{w}_{k,i,j}^{(0)} \cdot x^{(j)} + \bar{w}_{k,i,0}^{(0)}\right).$$

In the *first step* of the proof we show

$$f_{\bar{\mathbf{w}}}(x) \geq 1 - \frac{1}{n} \text{ for all } x \in [u^{(1)} + \delta, v^{(1)} - \delta] \times \cdots \times [u^{(d)} + \delta, v^{(d)} - \delta].$$

Let $x \in [u^{(1)} + \delta, v^{(1)} - \delta] \times \cdots \times [u^{(d)} + \delta, v^{(d)} - \delta]$. Then we get for any $i \in \{1, \ldots, d\}$ by (27), (29) and (36)

$$\begin{aligned}
&\sum_{j=1}^{d} \bar{w}_{1,i,j}^{(0)} \cdot x^{(j)} + \bar{w}_{1,i,0}^{(0)} \\
&= \sum_{j=1}^{d} (\bar{w}_{1,i,j}^{(0)} - w_{1,i,j}^{(0)}) \cdot x^{(j)} + (\bar{w}_{1,i,0}^{(0)} - w_{1,i,0}^{(0)}) + \sum_{j=1}^{d} w_{1,i,j}^{(0)} \cdot x^{(j)} + w_{1,i,0}^{(0)} \\
&\geq -d(\log n) \cdot \alpha_n - \log n + \frac{4d(\log n)^2}{\delta}(u^{(i)} + \delta) - \frac{4d(\log n)^2}{\delta} u^{(i)}
\end{aligned}$$

$$
\begin{aligned}
&= 3d(\log n)^2 - \log n \\
&\geq \log n
\end{aligned}
$$

and for any $i \in \{d+1, \dots, 2d\}$ by (28), (29) and (36)

$$
\begin{aligned}
&\sum_{j=1}^{d} \bar{w}_{1,i,j}^{(0)} \cdot x^{(j)} + \bar{w}_{1,i,0}^{(0)} \\
&= \sum_{j=1}^{d} (\bar{w}_{1,i,j}^{(0)} - w_{1,i,j}^{(0)}) \cdot x^{(j)} + (\bar{w}_{1,i,0}^{(0)} - w_{1,i,0}^{(0)}) + \sum_{j=1}^{d} w_{1,i,j}^{(0)} \cdot x^{(j)} + w_{1,i,0}^{(0)} \\
&\geq -d(\log n) \cdot \alpha_n - \log n - \frac{4d(\log n)^2}{\delta}(v^{(i-d)} - \delta) + \frac{4d(\log n)^2}{\delta} v^{(i-d)} \\
&= 3d(\log n)^2 - \log n \\
&\geq \log n.
\end{aligned}
$$

This implies

$$
\bar{f}_{1,i}^{(1)}(x) \geq \sigma(\log n) = 1 - \frac{1}{n+1} \geq 1 - \frac{1}{n}
$$

for any $i \in \{1, \dots, 2d\}$.

Using (30)-(32) and $|\sigma(u)| \leq 1$ for any $u \in \mathbb{R}$ we get similarly as above

$$
\begin{aligned}
&\sum_{j=1}^{r} \bar{w}_{1,1,j}^{(1)} \cdot \bar{f}_{1,j}^{(1)}(x) + \bar{w}_{1,1,0}^{(1)} \\
&\geq -(r+1)\log n + \sum_{j=1}^{r} w_{1,1,j}^{(1)} \cdot \bar{f}_{1,j}^{(1)}(x) + w_{1,1,0}^{(1)} \\
&= -(r+1)\log n + \sum_{j=1}^{2d} w_{1,1,j}^{(1)} \cdot \bar{f}_{1,j}^{(1)}(x) + \sum_{j=2d+1}^{r} w_{1,1,j}^{(1)} \cdot \bar{f}_{1,j}^{(1)}(x) + w_{1,1,0}^{(1)} \\
&\geq -(r+1)\log n + 2d \cdot 8(\log n)^2 \left(1 - \frac{1}{n}\right) - 8(\log n)^2 \left(2d - \frac{1}{2}\right) \\
&= -(r+1)\log n + 8(\log n)^2 \left(\frac{1}{2} - \frac{2d}{n}\right) \\
&\geq \log n.
\end{aligned}
$$

Therefore, we obtain

$$
\bar{f}_{1,1}^{(2)}(x) \geq 1 - \frac{1}{n}.
$$

With the same argument as above and with (33)-(35), we can recursively conclude for $l = 3, \dots, L$ that we have

$$
\begin{aligned}
&\sum_{j=1}^{r} \bar{w}_{1,1,j}^{(l-1)} \cdot \bar{f}_{1,j}^{(l-1)}(x) + \bar{w}_{1,1,0}^{(l-1)} \\
&\geq -(r+1)\log n + \sum_{j=1}^{r} w_{1,1,j}^{(l-1)} \cdot \bar{f}_{1,j}^{(l-1)}(x) + w_{1,1,0}^{(l-1)} \\
&\geq -(r+1)\log n + 6(\log n)^2 \left(1 - \frac{1}{n}\right) - 3(\log n)^2 \\
&= -(r+1)\log n + 3(\log n)^2 - \frac{6}{n}(\log n)^2 \\
&\geq 2 \cdot (\log n)^2 - \frac{6}{n}(\log n)^2 \\
&\geq \log n.
\end{aligned}
$$

Therefore, we obtain

$$
\bar{f}_{1,1}^{(l)}(x) \geq 1 - \frac{1}{n}
$$

for $l = 3, \dots, L$.

This implies

$$
f_{\bar{\mathbf{w}}}(x) \geq 1 - \frac{1}{n} \text{ if } x \in [u^{(1)} + \delta, v^{(1)} - \delta] \times \cdots \times [u^{(d)} + \delta, v^{(d)} - \delta].
$$

In the *second step* of the proof we show

$$
f_{\bar{\mathbf{w}}}(x) \leq \frac{1}{n} \text{ if } x^{(i)} \notin [u^{(i)} - \delta, v^{(i)} + \delta] \text{ for some } i \in \{1, \dots, d\}.
$$

Let $i \in \{1, \dots, d\}$ and assume $x^{(i)} \notin [u^{(i)} - \delta, v^{(i)} + \delta]$. Then, we can argue similarly as above. In case $x^{(i)} < u^{(i)} - \delta$ for some $i \in \{1, \dots, d\}$ we obtain by (27), (29) and (36)

$$
\begin{aligned}
&\sum_{j=1}^{d} \bar{w}_{1,i,j}^{(0)} \cdot x^{(j)} + \bar{w}_{1,i,0}^{(0)} \\
&\leq d(\log n) \cdot \alpha_n + \log n + \frac{4d(\log n)^2}{\delta}(u^{(i)} - \delta) - \frac{4d(\log n)^2}{\delta} u^{(i)} \\
&\leq -3d(\log n)^2 + \log n \\
&\leq -\log n
\end{aligned}
$$

and in case $x^{(i)} > v^{(i)} + \delta$ for some $i \in \{1, \dots, d\}$ we obtain by (28), (29) and (36)

$$
\begin{aligned}
&\sum_{j=1}^{d} \bar{w}_{1,i+d,j}^{(0)} \cdot x^{(j)} + \bar{w}_{1,i+d,0}^{(0)} \\
&\leq d(\log n) \cdot \alpha_n + \log n - \frac{4d(\log n)^2}{\delta}(v^{(i)} + \delta) + \frac{4d(\log n)^2}{\delta} v^{(i)}
\end{aligned}
$$

$$\leq -3d(\log n)^2 + \log n$$
$$\leq -\log n.$$

Together with the logistic squasher it holds

$$\bar{f}_{1,i}^{(1)}(x) \leq \sigma(-\log n) = \frac{1}{n+1} \leq \frac{1}{n}$$

for some $i \in \{1, ..., 2d\}$.

Similar to above we get with (30)-(32) and (36)

$$\begin{aligned}
&\sum_{j=1}^{r} \bar{w}_{1,1,j}^{(1)} \cdot \bar{f}_{1,j}^{(1)}(x) + \bar{w}_{1,1,0}^{(1)} \\
&\leq (r+1)\log n + \sum_{j=1}^{r} w_{1,1,j}^{(1)} \cdot \bar{f}_{1,j}^{(1)}(x) + w_{1,1,0}^{(1)} \\
&\leq (r+1)\log n + (2d-1) \cdot 8(\log n)^2 + 8(\log n)^2 \cdot \frac{1}{n} - 8(\log n)^2 \left(2d - \frac{1}{2}\right) \\
&= (r+1)\log n + 8(\log n)^2 \left(\frac{1}{n} - \frac{1}{2}\right) \\
&\leq -\log n.
\end{aligned}$$

From this we obtain

$$\bar{f}_{1,1}^{(2)}(x) \leq \frac{1}{n}.$$

Using (33)-(36) we can argue similarly as above and conclude recursively for $l = 3, \ldots, L$

$$\begin{aligned}
&\sum_{j=1}^{r} \bar{w}_{1,1,j}^{(l-1)} \cdot \bar{f}_{1,j}^{(l-1)}(x) + \bar{w}_{1,1,0}^{(l-1)} \\
&\leq (r+1)\log n + \sum_{j=1}^{r} w_{1,1,j}^{(l-1)} \cdot \bar{f}_{1,j}^{(l-1)}(x) + w_{1,1,0}^{(l-1)} \\
&\leq (r+1)\log n + 6 \cdot (\log n)^2 \cdot \frac{1}{n} - 3(\log n)^2 \\
&\leq (\log n)^2 + 6 \cdot (\log n)^2 \cdot \frac{1}{n} - 3(\log n)^2 \\
&= -2 \cdot (\log n)^2 + 6 \cdot (\log n)^2 \cdot \frac{1}{n} \\
&\leq -\log n.
\end{aligned}$$

So we get $\bar{f}_{1,1}^{(l)}(x) \leq \frac{1}{n}$ for $l = 3, \ldots, L$. Therefore,

$$f_{\bar{\mathbf{w}}}(x) \leq \frac{1}{n} \text{ if } x^{(i)} \notin [u^{(i)} - \delta, v^{(i)} + \delta] \text{ for some } i \in \{1, \ldots, d\}$$

holds.

This yields the assertion. □

Lemma 6 *Let* $0 < \delta \leq 1$, $1 \leq \alpha_n \leq \log n$ *and let* σ *be the logistic squasher, let* $m : \mathbb{R}^d \to \mathbb{R}$ *be Lipschitz continuous with Lipschitz constant* C_{Lip}*, let* $L, r, n \in \mathbb{N}$ *with* $L \geq 2$, $r \geq 2d$, $n \geq 8d$, $n \geq \exp(r+1)$ *and let* $K \in \mathbb{N}$ *with* $K^d \leq K_n$*. Furthermore define* $f_{\bar{\mathbf{w}}}$ *by*

$$f_{\bar{\mathbf{w}}}(x) = \sum_{j=1}^{K_n} \bar{w}_{1,1,j}^{(L)} \cdot \bar{f}_{j,1}^{(L)}(x)$$

for some $\bar{w}_{1,1,1}^{(L)}, \ldots, \bar{w}_{1,1,K_n}^{(L)} \in \mathbb{R}$*, where* $\bar{f}_{j,1}^{(L)}$ *are recursively defined by*

$$\bar{f}_{k,i}^{(l)}(x) = \sigma\left(\sum_{j=1}^{r} \bar{w}_{k,i,j}^{(l-1)} \cdot \bar{f}_{k,j}^{(l-1)}(x) + \bar{w}_{k,i,0}^{(l-1)}\right)$$

for some $\bar{w}_{k,i,0}^{(l-1)}, \ldots, \bar{w}_{k,i,r}^{(l-1)} \in \mathbb{R}$ $(l = 2, \ldots, L)$ *and*

$$\bar{f}_{k,i}^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} \bar{w}_{k,i,j}^{(0)} \cdot x^{(j)} + \bar{w}_{k,i,0}^{(0)}\right).$$

Choose $\mathbf{w}$ *such that*

$$w_{j_k,j,j}^{(0)} = \frac{4d \cdot (\log n)^2}{\delta} \quad \text{and} \quad w_{j_k,j,0}^{(0)} = \frac{-4d \cdot (\log n)^2 \cdot u_k^{(j)}}{\delta} \quad \text{for } j \in \{1, \ldots, d\}, \tag{37}$$

$$w_{j_k,j+d,j}^{(0)} = \frac{-4d \cdot (\log n)^2}{\delta} \quad \text{and} \quad w_{j_k,j+d,0}^{(0)} = \frac{4d \cdot (\log n)^2 \cdot v_k^{(j)}}{\delta} \quad \text{for } j \in \{1, \ldots, d\}, \tag{38}$$

$$w_{j_k,s,t}^{(0)} = 0 \quad \text{if } s \leq 2d, s \neq t, s \neq t+d \text{ and } t > 0, \tag{39}$$

$$w_{j_k,1,t}^{(1)} = 8 \cdot (\log n)^2 \quad \text{for } t \in \{1, \ldots, 2d\}, \tag{40}$$

$$w_{j_k,1,0}^{(1)} = -8 \cdot (\log n)^2 \left(2d - \frac{1}{2}\right) \tag{41}$$

$$w_{j_k,1,t}^{(1)} = 0 \quad \text{for } t > 2d, \tag{42}$$

$$w_{j_k,1,1}^{(l)} = 6 \cdot (\log n)^2 \quad \text{for } l \in \{2, \ldots, L\}, \tag{43}$$

$$w_{j_k,1,0}^{(1)} = -3(\log n)^2 \quad \text{for } l \in \{2, \ldots, L\} \tag{44}$$

and

$$w_{j_k,1,t}^{(l)} = 0 \quad \text{for } t > 1 \text{ and } l \in \{2, \ldots, L\} \tag{45}$$

for all $k \in \{1, \ldots, K^d\}$.

Let $a_1, \ldots, a_d, b_1, \ldots, b_d \in [-\alpha_n, \alpha_n]^d$ *with* $b_i - a_i = \Delta$ *for all* $i \in \{1, \ldots, d\}$ *and* $\Delta \in \mathbb{R}_+$*. Then there exist*

$$\alpha_1, \ldots, \alpha_{K^d} \in [-\|m\|_\infty, \|m\|_\infty] \tag{46}$$

and $u_1, v_1, \ldots, u_{K^d}, v_{K^d} \in [a_1, b_1) \times \cdots \times [a_d, b_d)$ *such that for all pairwise distinct* $j_1, \ldots, j_{K^d} \in \{1, \ldots, K_n\}$ *the inequality*

$$|f_{\bar{\mathbf{w}}}(x) - m(x)| \leq c_{22} \cdot \left(C_{Lip} \cdot \frac{\Delta}{K} + K^d \cdot \frac{1}{n} \right) \tag{47}$$

holds for all $x \in [a_1, b_1] \times \cdots \times [a_d, b_d]$ *which are not contained in*

$$\bigcup_{j \in \{0,1,\ldots,K\}} \bigcup_{i \in \{1,\ldots,d\}} \left\{ x \in \mathbb{R}^d : \left| x^{(i)} - \left(a_i + j \cdot \frac{b_i - a_i}{K} \right) \right| < \delta \right\} \tag{48}$$

and for all weight vectors $\bar{\mathbf{w}}$ *which satisfy*

$$\bar{w}_{1,1,j_k}^{(L)} = \alpha_k \quad (k \in \{1, \ldots, K^d\}), \quad \bar{w}_{1,1,k}^{(L)} = 0 \quad (k \notin \{j_1, \ldots, j_{K^d}\}) \tag{49}$$

and

$$|w_{j_s,k,i}^{(l)} - \bar{w}_{j_s,k,i}^{(l)}| \leq \log n \quad \textit{for all } l \in \{0, \ldots, L-1\}, s \in \{1, \ldots, K^d\}. \tag{50}$$

For $\delta \leq \frac{\Delta}{K}$ *and* $x \in \mathbb{R}^d$ *we get additionally*

$$|f_{\bar{\mathbf{w}}}(x)| \leq \|m\|_\infty \cdot \left(3^d + \frac{K^d}{n} \right). \tag{51}$$

Proof. We partition $[a_1, b_1) \times \cdots \times [a_d, b_d)$ into K^d equivolume cubes of side length $\frac{\Delta}{K}$. For comprehensibility, we number these cubes C_i by $i \in \{1, \ldots, K^d\}$, such that C_i corresponds to the cube

$$[u_i^{(1)}, v_i^{(1)}) \times \cdots \times [u_i^{(d)}, v_i^{(d)}).$$

Since m is Lipschitz continuous with Lipschitz constant C_{Lip}, m can be approximated by an approximand S that is piecewise constant on each cube. S can be expressed in the form

$$S(x) = \sum_{i \in \{1,\ldots,K^d\}} \alpha_i \cdot \mathbb{1}_{C_i}(x)$$

where $\alpha_i = m(z_i)$ for $i \in \{1, \ldots, K^d\}$ and z_i is the center of the cube C_i.

$S(x)$ has value $m(z_i)$ for $x \in C_i$. Therefore it follows from the Lipschitz continuity of m

$$|S(x) - m(x)| \leq C_{Lip} \cdot \|z_i - x\|_2$$

if $x \in C_i$ for some $i \in \{1, \ldots, K^d\}$. Since every cube has a side length of $\frac{\Delta}{K}$ we obtain

$$|S(x) - m(x)| \leq C_{Lip} \cdot \sqrt{d} \cdot \frac{\Delta}{K} \text{ for } x \in [a_1, b_1) \times \cdots \times [a_d, b_d).$$

Now, because of (49) and the definitions of S and $f_{\bar{\mathbf{w}}}(x)$, we have

$$S(x) - f_{\bar{\mathbf{w}}}(x) = \sum_{k=1}^{K^d} \alpha_k \left(\mathbb{1}_{C_k}(x) - \bar{f}_{j_k,1}^{(L)}(x) \right).$$

Application of Lemma 5 yields

$$|S(x) - f_{\bar{\mathbf{w}}}(x)| \le c_{23} \cdot K^d \cdot \frac{1}{n}$$

for all $x \in [a_1, b_1) \times \cdots \times [a_d, b_d)$ which are not contained in (48).

From this we obtain

$$\left|f_{\bar{\mathbf{w}}}(x) - m(x)\right| = \left|f_{\bar{\mathbf{w}}}(x) - S(x)\right| + \left|S(x) - m(x)\right| \le c_{24} \left(K^d \cdot \frac{1}{n} + C_{Lip} \cdot \frac{\Delta}{K}\right)$$

for all $x \in [a_1, b_1) \times \cdots \times [a_d, b_d)$ which are not contained in (48). This implies (47).

To show inequality (51), we assume that $x \in \mathbb{R}^d$. Then for $\delta \le \frac{\Delta}{K}$ and every fixed x we get

$$\begin{aligned} |f_{\bar{\mathbf{w}}}(x)| &= \sum_{k=1}^{K^d} \bar{w}_{1,1,j_k}^{(L)} \cdot \bar{f}_{j_k,1}^{(L)}(x) \\ &= \sum_{k \in \{1,\dots,K^d\} \,:\, x \in [u_k^{(1)} - \delta, v_k^{(1)} + \delta] \times \cdots \times [u_k^{(d)} - \delta, v_k^{(d)} + \delta]} \bar{w}_{1,1,j_k}^{(L)} \cdot \bar{f}_{j_k,1}^{(L)}(x) \\ &\quad + \sum_{k \in \{1,\dots,K^d\} \,:\, x \notin [u_k^{(1)} - \delta, v_k^{(1)} + \delta] \times \cdots \times [u_k^{(d)} - \delta, v_k^{(d)} + \delta]} \bar{w}_{1,1,j_k}^{(L)} \cdot \bar{f}_{j_k,1}^{(L)}(x) \end{aligned}$$

There are at most 3^d many cubes $[u_i^{(1)} - \delta, v_i^{(1)} + \delta] \times \cdots \times [u_i^{(d)} - \delta, v_i^{(d)} + \delta]$ which contain x. Together with the definition of $\bar{w}_{1,1,j_k}$ for $k \in \{1, \dots, K^d\}$ we get

$$\sum_{k \in \{1,\dots,K^d\} \,:\, x \in [u_k^{(1)} - \delta, v_k^{(1)} + \delta] \times \cdots \times [u_k^{(d)} - \delta, v_k^{(d)} + \delta]} \bar{w}_{1,1,j_k}^{(L)} \cdot \bar{f}_{j_k,1}^{(L)}(x) \le 3^d \cdot \|m\|_\infty.$$

By Lemma 5 we know that $\bar{f}_{j_k,1}^{(L)}(x) \le \frac{1}{n}$ for $k \in \{1, \dots, K^d\}$ if $x \notin [u_k^{(1)} - \delta, v_k^{(1)} + \delta] \times \cdots \times [u_k^{(d)} - \delta, v_k^{(d)} + \delta]$. Thus we get together with the definition of $w_{1,1,j_k}$ for $k \in \{1, \dots, K^d\}$

$$\sum_{k \in \{1,\dots,K^d\} \,:\, x \notin [u_k^{(1)} - \delta, v_k^{(1)} + \delta] \times \cdots \times [u_k^{(d)} - \delta, v_k^{(d)} + \delta]} w_{1,1,j_k}^{(L)} \cdot f_{j_k,1}^{(L)}(x) \le K^d \cdot \frac{1}{n} \cdot \|m\|_\infty.$$

This results in

$$|f_{\bar{\mathbf{w}}}(x)| \le 3^d \cdot \|m\|_\infty + K^d \cdot \frac{1}{n} \cdot \|m\|_\infty$$

for $x \in \mathbb{R}^d$.

□

Lemma 7 *Let σ be the logistic squasher, let $1 \le \alpha_n \le \log n$, let $m : \mathbb{R}^d \to \mathbb{R}$ be Lipschitz continuous as well as bounded, let $L, r, n \in \mathbb{N}$ with $L \ge 2$, $r \ge 2d$, $n \ge 8d$ and*

$n \geq \exp(r+1)$ and let $K \in \mathbb{N}$ with $2 \leq K \leq \alpha_n - 1$ and $(K^2+1)^{3d} \leq K_n$. Choose $\mathbf{w}$ *such that*

$$w^{(0)}_{j_k,j,j} = 4d \cdot K^2 \cdot (\log n)^2 \quad \textit{and} \quad w^{(0)}_{j_k,j,0} = -4d \cdot K^2 \cdot (\log n)^2 \cdot u^{(j)}_k \quad \textit{for } j \in \{1,\dots,d\}, \tag{52}$$

$$w^{(0)}_{j_k,j+d,j} = -4d \cdot K^2 \cdot (\log n)^2 \quad \textit{and} \quad w^{(0)}_{j_k,j+d,0} = 4d \cdot K^2 \cdot (\log n)^2 \cdot v^{(j)}_k \quad \textit{for } j \in \{1,\dots,d\}, \tag{53}$$

$$w^{(0)}_{j_k,s,t} = 0 \quad \textit{if } s \leq 2d, s \neq t, s \neq t+d \textit{ and } t > 0, \tag{54}$$

$$w^{(1)}_{j_k,1,t} = 8 \cdot (\log n)^2 \quad \textit{for } t \in \{1,\dots,2d\}), \tag{55}$$

$$w^{(1)}_{j_k,1,0} = -8 \cdot (\log n)^2 \left(2d - \frac{1}{n}\right) \tag{56}$$

$$w^{(1)}_{j_k,1,t} = 0 \quad \textit{for } t > 2d, \tag{57}$$

$$w^{(l)}_{j_k,1,1} = 6 \cdot (\log n)^2 \quad \textit{for } l \in \{2,\dots,L\}, \tag{58}$$

$$w^{(l)}_{j_k,1,0} = -3 \cdot (\log n)^2 \quad \textit{for } l \in \{2,\dots,L\} \tag{59}$$

and

$$w^{(l)}_{j_k,1,t} = 0 \quad \textit{for } t > 1 \textit{ and } l \in \{2,\dots,L\} \tag{60}$$

for all $k \in \{1,\dots,(K^2+1)^{3d}\}$.

Then there exists

$$\alpha_1,\dots,\alpha_{(K^2+1)^{3d}} \in \left[-\frac{\|m\|_\infty}{(K^2+1)^{2d}}, \frac{\|m\|_\infty}{(K^2+1)^{2d}}\right] \tag{61}$$

and $u_1, v_1, \dots, u_{(K^2+1)^{3d}}, v_{(K^2+1)^{3d}} \in [-K - \frac{2}{K}, K]^d$ such that for all pairwise distinct $j_1,\dots,j_{(K^2+1)^{3d}} \in \{1,\dots,K_n\}$

$$\begin{aligned} &\int |f_{\bar{\mathbf{w}}}(x) - m(x)|^2 \mathbf{P}_X(dx) \\ &\leq c_{25} \cdot \left(\frac{1}{K} + \frac{K^{12d}}{n^2} + \left(\frac{K^{6d}}{n} + 1\right)^2 \cdot \mathbf{P}_X(\mathbb{R}^d \setminus [-K,K]^d) \right) \end{aligned} \tag{62}$$

holds for all weight vectors $\bar{\mathbf{w}}$ which satisfy

$$\bar{w}^{(L)}_{1,1,j_k} = \alpha_k \quad (k \in \{1,\dots,(K^2+1)^{3d}\}), \quad \bar{w}^{(L)}_{1,1,k} = 0 \quad (k \notin \{j_1,\dots,j_{(K^2+1)^{3d}}\}) \tag{63}$$

and

$$|w^{(l)}_{j_s,k,i} - \bar{w}^{(l)}_{j_s,k,i}| \leq \log n \quad \textit{for all } l \in \{0,\dots,L-1\}, s \in \{1,\dots,(K^2+1)^{3d}\}. \tag{64}$$

Proof. We subdivide $[-K - \frac{2}{K}, K]^d$ in $(K^2+1)^d$ cubes of side length $\frac{2}{K}$. We number these cubes C_i by $i \in \{1, \dots, (K^2+1)^d\}$, such that C_i corresponds to the cube

$$[u_i^{(1)}, v_i^{(1)}) \times \cdots \times [u_i^{(d)}, v_i^{(d)}).$$

Let C_{Lip} be the Lipschitz constant of m. Then by Lemma 6 applied to $m/(K^2+1)^{2d}$ and $\delta = \frac{1}{K^2}$ we know that

$$\left| f_{\bar{\mathbf{w}}}(x) - \frac{1}{(K^2+1)^{2d}} \cdot m(x) \right| \leq c_{22} \cdot \left(\frac{C_{Lip}}{(K^2+1)^{2d}} \cdot \frac{2}{K} + (K^2+1)^d \cdot \frac{1}{n} \right)$$

holds for all $x \in [-K - \frac{2}{K}, K]^d$ which are not contained in

$$A := \bigcup_{i \in \{0,1,\dots,K+1\}} \bigcup_{j \in \{1,\dots,d\}} \left\{ x \in \mathbb{R}^d : \left| x^{(j)} - \left(-K - \frac{2}{K} + i \cdot \frac{2}{K} \right) \right| < \delta \right\}. \tag{65}$$

Next we repeat the whole construction $(K^2+1)^{2d}$ many times, which results in an approximation $f_{\bar{\mathbf{w}}}$ of

$$(K^2+1)^{2d} \cdot \frac{1}{(K^2+1)^{2d}} \cdot m(x)$$

which satisfies

$$|f_{\bar{\mathbf{w}}}(x) - m(x)| \leq c_{26} \cdot \left(\frac{1}{K} + (K^2+1)^{3d} \cdot \frac{1}{n} \right) \tag{66}$$

outside of A.

Now we want to move the grid so that $[-K, K]^d$ is always covered. We slightly shift the whole grid of cubes along the j-th component by modifying all $u_i^{(j)}, v_i^{(j)}$ by the same additional summand. This summand is chosen from the set

$$\left\{ k \cdot \frac{2}{K^2} \quad : \quad k = 0, 1, \dots, K-1 \right\}$$

for fixed $j \in \{1, \dots, d\}$. In this way we can construct K different versions of $f_{\bar{\mathbf{w}}}$ that still satisfy (66) for all $x \in [-K, K]^d$ up to corresponding versions of A.

Since we shift the grid of cubes we obtain for fixed $j \in \{1, \dots, d\}$ K disjoint versions of $\bigcup_{i \in \{0,1,\dots,K+1\}} \left\{ x \in \mathbb{R}^d : \left| x^{(j)} - \left(-K - \frac{2}{K} + i \cdot \frac{2}{K} \right) \right| < \delta \right\}$. Because the sum of $\mathbf{P}_X$-measures of these K disjoint sets is less than or equal to one, at least one of them must have measure less than or equal to $\frac{1}{K}$. Consequently we can shift the u_i and v_i such that

$$\mathbf{P}_X(A) = \sum_{j \in \{1,\dots,d\}} \frac{1}{K} = \frac{d}{K}$$

holds.

Now we have found a shifted version of the grid such that the set (65) has a measure less than or equal to $\frac{d}{K}$. By (66) we know that $|f_{\bar{\mathbf{w}}}(x) - m(x)| \leq c_{27} \cdot \left(\frac{1}{K} + \frac{K^{6d}}{n} \right)$ holds for $x \in [-K,K]^d \setminus A$. From this together with the second assertion from Lemma 6 we obtain

$$
\begin{aligned}
&\int |f_{\bar{w}}(x) - m(x)|^2 \, \mathbf{P}_X(dx) \\
&= \int_{[-K,K]^d \setminus A} |f_{\bar{w}}(x) - m(x)|^2 \, \mathbf{P}_X(dx) + \int_A |f_{\bar{w}}(x) - m(x)|^2 \, \mathbf{P}_X(dx) \\
&\quad + \int_{\mathbb{R}^d \setminus [-K,K]^d} |f_{\bar{w}}(x) - m(x)|^2 \, \mathbf{P}_X(dx) \\
&\leq c_{27}^2 \left(\frac{1}{K} + \frac{K^{6d}}{n} \right)^2 + c_{28} \left(3^d + \frac{K^{6d}}{n} \right)^2 \cdot \frac{d}{K} \\
&\quad + c_{29} \left(3^d + \frac{K^{6d}}{n} \right)^2 \cdot \mathbf{P}_X(\mathbb{R}^d \setminus [-K,K]^d),
\end{aligned}
$$

which implies the assertion. □

In order to be able to formulate our next auxiliary result we need the following notation: Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, let $K \in \mathbb{N}$, let $B_1, \ldots, B_K : \mathbb{R}^d \to \mathbb{R}$ and let $c_2 > 0$. In the next lemma we consider the problem to minimize

$$
F(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^n |\sum_{k=1}^K a_k \cdot B_k(x_i) - y_i|^2 + c_2 \cdot \sum_{k=1}^{K_n} a_k^2, \tag{67}
$$

where $\mathbf{a} = (a_1, \ldots, a_K)^T$, by gradient descent. To do this, we choose $\mathbf{a}^{(0)} \in \mathbb{R}^K$ and set

$$
\mathbf{a}^{(t+1)} = \mathbf{a}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)}) \tag{68}
$$

for some properly chosen $\lambda_n > 0$.

Lemma 8 *Let F be defined by (67) and choose $\mathbf{a}_{opt}$ such that*

$$
F(\mathbf{a}_{opt}) = \min_{\mathbf{a} \in \mathbb{R}^K} F(\mathbf{a}).
$$

Then for any $\mathbf{a} \in \mathbb{R}^K$ we have

$$
\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2 \geq 4 \cdot c_2 \cdot (F(\mathbf{a}) - F(\mathbf{a}_{opt})).
$$

Proof. The proof is a modification of the proof of Lemma 3 in Braun, Kohler and Walk (2019).

Set

$$
\mathbf{E} = c_2 \cdot \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix},
$$

$$\mathbf{B} = (B_j(x_i))_{1 \le i \le n, 1 \le j \le K} \quad \text{and} \quad \mathbf{A} = \frac{1}{n} \cdot \mathbf{B}^T \cdot \mathbf{B} + c_2 \cdot \mathbf{E}.$$

Then $\mathbf{A}$ is positive definite and hence regular, from which we can conclude

$$\begin{aligned} F(\mathbf{a}) &= \frac{1}{n} \cdot (\mathbf{B} \cdot \mathbf{a} - \mathbf{y})^T \cdot (\mathbf{B} \cdot \mathbf{a} - \mathbf{y}) + c_2 \cdot \mathbf{a}^T \cdot \mathbf{E} \cdot \mathbf{a} \\ &= \mathbf{a}^T \mathbf{A} \mathbf{a} - 2\mathbf{y}^T \frac{1}{n} \mathbf{B} \mathbf{a} + \frac{1}{n} \mathbf{y}^T \mathbf{y} \\ &= (\mathbf{a} - \mathbf{A}^{-1} \frac{1}{n} \mathbf{B}^T \mathbf{y})^T \mathbf{A} (\mathbf{a} - \mathbf{A}^{-1} \frac{1}{n} \mathbf{B}^T \mathbf{y}) + F(\mathbf{a}_{opt}), \end{aligned}$$

where

$$F(\mathbf{a}_{opt}) = \frac{1}{n} \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \cdot \frac{1}{n} \cdot \mathbf{B} \cdot \mathbf{A}^{-1} \cdot \frac{1}{n} \cdot \mathbf{B}^T \mathbf{y}.$$

Using

$$\mathbf{b}^T \mathbf{A} \mathbf{b} \ge c_2 \cdot \mathbf{b}^T \mathbf{E} \mathbf{b} = c_2 \cdot \mathbf{b}^T \mathbf{b}$$

and $\mathbf{A}^T = \mathbf{A}$ we conclude

$$\begin{aligned} & F(\mathbf{a}) - F(\mathbf{a}_{opt}) \\ &= ((\mathbf{A}^{1/2})^T (\mathbf{a} - \mathbf{A}^{-1} \frac{1}{n} \mathbf{B}^T \mathbf{y}))^T \mathbf{A}^{1/2} (\mathbf{a} - \mathbf{A}^{-1} \frac{1}{n} \mathbf{B}^T \mathbf{y}) \\ &\le \frac{1}{c_2} \cdot ((\mathbf{A}^{1/2})^T (\mathbf{a} - \mathbf{A}^{-1} \frac{1}{n} \mathbf{B}^T \mathbf{y}))^T \mathbf{A} \mathbf{A}^{1/2} (\mathbf{a} - \mathbf{A}^{-1} \frac{1}{n} \mathbf{B}^T \mathbf{y}) \\ &= \frac{1}{c_2} \cdot ((\mathbf{A})^T (\mathbf{a} - \mathbf{A}^{-1} \frac{1}{n} \mathbf{B}^T \mathbf{y}))^T \mathbf{A} (\mathbf{a} - \mathbf{A}^{-1} \frac{1}{n} \mathbf{B}^T \mathbf{y}) \\ &= \frac{1}{c_2} \cdot (\mathbf{A}\mathbf{a} - \frac{1}{n} \mathbf{B}^T \mathbf{y})^T (\mathbf{A}\mathbf{a} - \frac{1}{n} \mathbf{B}^T \mathbf{y}) \\ &= \frac{1}{4 \cdot c_2} \cdot (2\mathbf{A}\mathbf{a} - \frac{2}{n} \mathbf{B}^T \mathbf{y})^T (2\mathbf{A}\mathbf{a} - \frac{2}{n} \mathbf{B}^T \mathbf{y}) \\ &= \frac{1}{4 \cdot c_2} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2, \end{aligned}$$

where the last equality follows from

$$(\nabla_{\mathbf{a}} F)(\mathbf{a}) = \nabla_{\mathbf{a}} \left(\mathbf{a}^T \mathbf{A} \mathbf{a} - 2\mathbf{y}^T \frac{1}{n} \mathbf{B} \mathbf{a} + \frac{1}{n} \mathbf{y}^T \mathbf{y} \right) = 2\mathbf{A}\mathbf{a} - \frac{2}{n} \mathbf{B}^T \mathbf{y}.$$

□

4.2 Proof of Theorem 1

Let $\epsilon > 0$ and $K \in \mathbb{N}$ be arbitrary. W.l.o.g. we assume $K_n \ge (K^2 + 1)^{3d}$. Choose a Lipschitz continuous and bounded function $\bar{m} : \mathbb{R}^d \to \mathbb{R}$ such that

$$\int |\bar{m}(x) - m(x)|^2 \mathbf{P}_X(dx) \le \epsilon. \tag{69}$$

Let A_n be the event that firstly the weight vector $\mathbf{w}^{(0)}$ satisfies

$$|(\mathbf{w}^{(0)})^{(l)}_{j_s,k,i} - \mathbf{w}^{(l)}_{j_s,k,i}| \leq \log n \quad \text{for all } l \in \{0, \dots, L-1\}, s \in \{1, \dots, (K^2+1)^{3d}\}$$

for some weight vector $\mathbf{w}$ which satisfies the conditions (52)–(60) of Lemma 7 for $\bar{m}$ and some $j_1, \dots, j_{(K^2+1)^{3d}} \in \{1, \dots, K_n\}$, and that secondly

$$\frac{1}{n} \sum_{i=1}^{n} Y_i^2 \leq \beta_n^3$$

holds. Define the weight vectors $(\mathbf{w}^*)^{(t)}$ by

$$((\mathbf{w}^*)^{(t)})^{(l)}_{k,i,j} = (\mathbf{w}^{(t)})^{(l)}_{k,i,j} \quad \text{for all } l = 0, \dots, L-1$$

and

$$((\mathbf{w}^*)^{(t)})^{(L)}_{1,1,j_k} = \alpha_k \quad \text{for all } k = 1, \dots, (K^2+1)^{3d}$$

and

$$((\mathbf{w}^*)^{(t)})^{(L)}_{1,1,k} = 0 \quad \text{for all } k \notin \{j_1, \dots, j_{(K^2+1)^{3d}}\}$$

where α_k is chosen as in Lemma 7 in case that A_n holds and where $\alpha_1 = \dots = \alpha_{(K^2+1)^{3d}} = 0$ in case that A_n does not hold.

In the *first step of the proof* we decompose the L_2 error of m_n in a sum of several terms. We have

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&= (\mathbf{E}\{|m_n(X) - Y|^2|\mathcal{D}_n\} - \mathbf{E}\{|m(X) - Y|^2\}) \cdot 1_{A_n} + \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{A_n^c} \\
&= (\mathbf{E}\{|m_n(X) - Y|^2|\mathcal{D}_n\} \\
&\qquad\qquad -\mathbf{E}\{|m_n(X) - Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X)|\mathcal{D}_n\}) \cdot 1_{A_n} \\
&\quad +(\mathbf{E}\{|m_n(X) - Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X)|\mathcal{D}_n\} \\
&\qquad\qquad -(1+\epsilon) \cdot \mathbf{E}\{|m_n(X) - T_{\beta_n}Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X)|\mathcal{D}_n\}) \cdot 1_{A_n} \\
&\quad +((1+\epsilon) \cdot \mathbf{E}\{|m_n(X) - T_{\beta_n}Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X)|\mathcal{D}_n\} \\
&\qquad\qquad -(1+\epsilon) \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - T_{\beta_n}Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i)) \cdot 1_{A_n} \\
&\quad +((1+\epsilon) \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - T_{\beta_n}Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i) \\
&\qquad\qquad -(1+\epsilon) \cdot \frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}^{(t_n)}}(X_i) - T_{\beta_n}Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i)) \cdot 1_{A_n} \\
&\quad +(1+\epsilon) \cdot (\frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}^{(t_n)}}(X_i) - T_{\beta_n}Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i)
\end{aligned}$$

$$
\begin{aligned}
&-(1+\epsilon)\cdot\frac{1}{n}\sum_{i=1}^{n}|f_{\mathbf{w}^{(t_n)}}(X_i)-Y_i|^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i))\cdot 1_{A_n}\\
&+((1+\epsilon)^2\cdot\frac{1}{n}\sum_{i=1}^{n}|f_{\mathbf{w}^{(t_n)}}(X_i)-Y_i|^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i)-\mathbf{E}\{|m(X)-Y|^2\})\cdot 1_{A_n}\\
&+\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\cdot 1_{A_n^c}\\
=&\sum_{j=1}^{7}T_{j,n}.
\end{aligned}
$$

In the *second step of the proof* we show

$$\limsup_{n\to\infty}\mathbf{E}T_{j,n}\le 0\quad\text{for } j\in\{1,2,5\}.$$

Because of $\alpha_n\to\infty$ $(n\to\infty)$ and $\mathbf{E}Y^2<\infty$ it holds

$$\mathbf{E}T_{1,n}=\mathbf{E}\{|Y|^2\cdot 1_{\mathbb{R}^d\setminus[-\alpha_n,\alpha_n]^d}(X)\}\to 0\quad(n\to\infty).$$

Using $(a+b)^2\le(1+\epsilon)\cdot a^2+(1+\frac{1}{\epsilon})\cdot b^2$ $(a,b\in\mathbb{R})$ we get

$$\mathbf{E}T_{2,n}\le\left(1+\frac{1}{\epsilon}\right)\cdot\mathbf{E}\{|T_{\beta_n}Y-Y|^2\}$$

and

$$\mathbf{E}T_{5,n}\le(1+\epsilon)\cdot\left(1+\frac{1}{\epsilon}\right)\cdot\mathbf{E}\{|T_{\beta_n}Y-Y|^2\}.$$

Because of $\beta_n\to\infty$ $(n\to\infty)$ and $\mathbf{E}Y^2<\infty$ this implies the assertion of the second step.

In the *third step of the proof* we show

$$\limsup_{n\to\infty}\mathbf{E}T_{4,n}\le 0.$$

If $|y|\le\beta_n$ then it holds for any $z\in\mathbb{R}$

$$|T_{\beta_n}z-y|\le|z-y|.$$

This implies

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-T_{\beta_n}Y_i|^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i)\\
&=\frac{1}{n}\sum_{i=1}^{n}|T_{\beta_n}f_{\mathbf{w}^{(t_n)}}(X_i)-T_{\beta_n}Y_i|^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i)\\
&\le\frac{1}{n}\sum_{i=1}^{n}|f_{\mathbf{w}^{(t_n)}}(X_i)-T_{\beta_n}Y_i|^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i),
\end{aligned}
$$

hence $T_{4,n} \leq 0$ holds, which implies the assertion of the third step.

In the *fourth step of the proof* we show that the assumptions of Lemma 1 are satisfied if A_n holds. If A_n holds, then we have

$$F_n(\mathbf{w}^{(0)}) = \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \leq \beta_n^3.$$

Hence if the assumptions of the two conditions, which we have to show in Lemma 1, hold for $\mathbf{a}_0 = \mathbf{w}^{(0)}$, then we can conclude from the random initialization of $\mathbf{w}^{(0)}$ that (13), (14), (17) and (18) hold with $\gamma_n^* = c_{23} \cdot (\log n)^2$ and $B_n = c_{24} \cdot (\log n)^2$. From this and Lemma 2 and Lemma 3 and the assumptions on L_n and t_n in Theorem 1 we get that (10) and (11) hold.

In the *fifth step of the proof* we show

$$\mathbf{P}(A_n^c) \leq \frac{c_{30}}{\beta_n^3}.$$

To do this, we bound the probability that the weight vector $\mathbf{w}^{(0)}$ does not satisfy the first condition in the definition of the event A_n by considering a sequential choice of the weights in the K_n fully connected neural networks which we compute in parallel. In a single fully connected neural network the probability that all $(L-1) \cdot r \cdot (r+1) + r \cdot (d+1)$ weights satisfy the conditions of Lemma 7 for j_1 is bounded from below by

$$\left(\frac{2 \cdot \log n}{40d \cdot (\log n)^2} \right)^{(L-1) \cdot r \cdot (r+1)} \cdot \left(\frac{2 \cdot \log n}{2n^\tau} \right)^{r \cdot (d+1)}.$$

Hence in the first $K_n/(K^2+1)^{3d}$ many fully connected networks the probability that the condition for j_1 is never satisfied is bounded from above by

$$\left(1 - \left(\frac{1}{20d \cdot \log n} \right)^{(L-1) \cdot r \cdot (r+1)} \cdot \left(\frac{\log n}{n^\tau} \right)^{r \cdot (d+1)} \right)^{K_n/(K^2+1)^{3d}}.$$

This implies that all conditions of Lemma 7 are satisfied outside of an event of probability

$$(K^2+1)^{3d} \cdot \left(1 - \left(\frac{1}{20d \cdot \log n} \right)^{(L-1) \cdot r \cdot (r+1)} \cdot \left(\frac{\log n}{n^\tau} \right)^{r \cdot (d+1)} \right)^{K_n/(K^2+1)^{3d}},$$

which is for large n less than

$$(K^2+1)^{3d} \cdot \left(1 - \frac{1}{n^r} \right)^{n^r \cdot \log n} \leq (K^2+1)^{3d} \cdot e^{-\log n} = \frac{(K^2+1)^{3d}}{n} \leq \frac{1}{2} \cdot \frac{c_{30}}{\beta_n^3}$$

because of (7) and $0 < \tau < 1/(d+1)$. Furthermore, we can conclude from Markov's inequality

$$\mathbf{P}\left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i^2 > \beta_n^3 \right\} \leq \frac{\mathbf{E}\left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \right\}}{\beta_n^3} = \frac{\mathbf{E}\{Y^2\}}{\beta_n^3}.$$

In the *sixth step of the proof* we show

$$\limsup_{n\to\infty} \mathbf{E} T_{3,n} \leq 0.$$

We have

$$\begin{aligned}
&\frac{1}{1+\epsilon} \cdot \mathbf{E}\left\{T_{3,n}\right\} \\
&\leq \int_0^{4\cdot\beta_n^2} \mathbf{P}\Bigg\{ (\mathbf{E}\{|m_n(X) - T_{\beta_n}Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X)|\mathcal{D}_n\} \\
&\qquad -\frac{1}{n}\sum_{i=1}^n |m_n(X_i) - T_{\beta_n}Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i)) \cdot 1_{A_n} > t \Bigg\}\, dt \\
&\leq \frac{1}{n^{1/4}} + \int_{1/n^{1/4}}^{4\cdot\beta_n^2} \mathbf{P}\Bigg\{ (\mathbf{E}\{|m_n(X) - T_{\beta_n}Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X)|\mathcal{D}_n\} \\
&\qquad\qquad -\frac{1}{n}\sum_{i=1}^n |m_n(X_i) - T_{\beta_n}Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i)) \cdot 1_{A_n} > t \Bigg\}\, dt.
\end{aligned}$$

We want to derive a bound for the above probability. For this we can assume without loss of generality that A_n holds. Due to the fourth step of the proof it is then possible to apply Lemma 1 and to conclude

$$\|\mathbf{w}^{(t_n)} - \mathbf{w}^{(0)}\|_\infty \leq \|\mathbf{w}^{(t_n)} - \mathbf{w}^{(0)}\| \leq c_{31} \cdot (\log n)^2.$$

From this and the choice of $\mathbf{w}^{(0)}$ we can conclude that m_n is contained in the function space

$$\{T_{\beta_n} f \cdot 1_{[-\alpha_n,\alpha_n]^d} \ : \ f \in \mathcal{F}\}$$

where $\mathcal{F}$ is defined as in Lemma 4 with $C = c_{32} \cdot \sqrt{K_n} \cdot (\log n)^2$, $B = c_{33} \cdot (\log n)^2$ and $A = c_{34} \cdot n^\tau$. Application of Lemma 4 together with standard bounds of empirical process theory (cf., Theorem 9.1 in Györfi et al. (2002)) yields

$$\begin{aligned}
&\mathbf{P}\Bigg\{ (\mathbf{E}\{|m_n(X) - T_{\beta_n}Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X)|\mathcal{D}_n\} \\
&\qquad\qquad -\frac{1}{n}\sum_{i=1}^n |m_n(X_i) - T_{\beta_n}Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i)) \cdot 1_{A_n} > t \Bigg\} \\
&\leq 8 \cdot \left(c_{35} \cdot \frac{\beta_n}{t/8} \right)^{c_{36}\cdot(\log n)^{c_{37}} \cdot n^{\tau\cdot d} \cdot \left(\frac{c_{38}\cdot\sqrt{K_n}\cdot(\log n)^2}{t/8} \right)^{d/k} + c_{39}} \cdot \exp\left(-\frac{n \cdot t^2}{128 \beta_n^4} \right).
\end{aligned}$$

By choosing k large enough the right-hand side above is for $t > 1/n^{1/4}$ bounded from above by

$$c_{40} \cdot \exp\left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \leq c_{41} \cdot \exp\left(-\frac{\sqrt{n}}{256 \cdot \beta_n^4} \right).$$

Consequently, we get

$$\mathbf{E}\left\{T_{3,n}\right\} \leq (1+\epsilon) \cdot \left(\frac{1}{n^{1/4}} + 4\beta_n^2 \cdot c_{42} \cdot \exp\left(-\frac{\sqrt{n}}{256 \beta_n^4} \right) \right) \to 0 \quad (n \to \infty).$$

In the *seventh step of the proof* we show

$$\limsup_{n \to \infty} \mathbf{E}\{T_{7,n}\} \leq 2\epsilon.$$

W.l.o.g. we assume $\|\bar{m}\|_\infty \leq \beta_n$. The boundedness of m_n by β_n together with the fifth step imply

$$\begin{aligned}
\mathbf{E}\{T_{7,n}\} &\leq 2 \cdot 4\beta_n^2 \cdot \mathbf{P}(A_n^c) + 2 \int |\bar{m}(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq 2 \cdot 4\beta_n^2 \cdot \frac{c_{30}}{\beta_n^3} + 2\epsilon.
\end{aligned}$$

Because of $\beta_n \to \infty$ $(n \to \infty)$ this implies the assertion of the seventh step.

In the *eighth step of the proof* we bound

$$\mathbf{E} T_{6,n}.$$

If A_n holds, then we can apply Lemma 1, which together with Lemma 8 (which we can apply if we use that the norm of the gradient of our nonlinear function is larger than the sum of the squares of the partial derivatives with respect to the weights corresponding only to the last layer L) yields

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}^{(t_n)}}(X_i) - Y_i|^2 \cdot 1_{[-\alpha_n, \alpha_n]^d}(X_i) \\
&\leq \frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}^{(t_n)}}(X_i) - Y_i|^2 \cdot 1_{[-\alpha_n, \alpha_n]^d}(X_i) + c_2 \cdot \sum_{k=1}^{K_n} ((\mathbf{w}^{(t_n)})_{1,1,k}^{(L)})^2 \\
&= F_n(\mathbf{w}^{(t_n)}) \\
&\leq F_n(\mathbf{w}^{(t_n - 1)}) - \frac{1}{2L_n} \cdot \|\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t_n-1)})\|^2 \\
&\leq F_n(\mathbf{w}^{(t_n - 1)}) - \frac{1}{2L_n} \cdot 4 \cdot c_2 \cdot (F_n(\mathbf{w}^{(t_n-1)}) - F_n((\mathbf{w}^*)^{(t_n-1)})) \\
&= \left(1 - \frac{2 \cdot c_2}{L_n}\right) \cdot F_n(\mathbf{w}^{(t_n-1)}) + \frac{2 \cdot c_2}{L_n} \cdot F_n((\mathbf{w}^*)^{(t_n-1)}) \\
&\leq \left(1 - \frac{2 \cdot c_2}{L_n}\right)^2 \cdot F_n(\mathbf{w}^{(t_n-2)}) + \frac{2 \cdot c_2}{L_n} \cdot F_n((\mathbf{w}^*)^{(t_n-1)}) \\
&\qquad + \frac{2 \cdot c_2}{L_n} \cdot \left(1 - \frac{2 \cdot c_2}{L_n}\right) F_n((\mathbf{w}^*)^{(t_n-2)}) \\
&\leq \dots
\end{aligned}$$

$$\leq \left(1-\frac{2\cdot c_2}{L_n}\right)^{t_n}\cdot F_n(\mathbf{w}^{(0)})+\sum_{k=1}^{t_n}\frac{2\cdot c_2}{L_n}\cdot\left(1-\frac{2\cdot c_2}{L_n}\right)^{k-1}F_n((\mathbf{w}^*)^{(t_n-k)}).$$

This implies

$$\begin{aligned}
&\mathbf{E}\left\{T_{6,n}\right\}\\
&\leq (1+\epsilon)^2\cdot\left(1-\frac{2\cdot c_2}{L_n}\right)^{t_n}\cdot\mathbf{E}\{Y^2\}\\
&\quad+(1+\epsilon)^2\cdot\sum_{k=1}^{t_n}\frac{2\cdot c_2}{L_n}\cdot\left(1-\frac{2\cdot c_2}{L_n}\right)^{k-1}\Big(\mathbf{E}\{F_n((\mathbf{w}^*)^{(t_n-k)})\cdot 1_{A_n}\}\\
&\qquad\qquad-\mathbf{E}\{|m(X)-Y|^2\}\cdot\mathbf{P}(A_n)\Big)\\
&\quad+((1+\epsilon)^2-1)\cdot\mathbf{E}\{|m(X)-Y|^2\}.
\end{aligned}$$

We have

$$\left(1-\frac{2\cdot c_2}{L_n}\right)^{t_n}\leq\exp\left(\frac{-2\cdot c_2\cdot t_n}{L_n}\right)\leq\exp\left(-c_{43}\cdot\log n\right)\to 0\quad(n\to\infty)$$

and

$$\begin{aligned}
&\mathbf{E}\{F_n((\mathbf{w}^*)^{(t_n-k)})\cdot 1_{A_n}\}-\mathbf{E}\{|m(X)-Y|^2\}\cdot\mathbf{P}(A_n)\\
&=\mathbf{E}\Big\{\frac{1}{n}\sum_{i=1}^n|f_{((\mathbf{w}^*)^{(t_n-k)})}(X_i)-Y_i|^2\cdot\mathbb{1}_{[-\alpha_n,\alpha_n]^d}(X_i)\cdot 1_{A_n}+c_2\cdot\sum_{j=1}^{K_n}|(\mathbf{w}^*)^{(L)}_{1,1,j}|^2\Big\}\\
&\quad-\mathbf{E}\{|m(X)-Y|^2\}\cdot\mathbf{P}(A_n)\\
&=\Bigg(\mathbf{E}\Big\{\frac{1}{n}\sum_{i=1}^n|f_{((\mathbf{w}^*)^{(t_n-k)})}(X_i)-Y_i|^2\cdot\mathbb{1}_{[-\alpha_n,\alpha_n]^d}(X_i)\cdot 1_{A_n}\Big\}\\
&\qquad-(1+\epsilon)\cdot\mathbf{E}\Big\{\frac{1}{n}\sum_{i=1}^n|f_{((\mathbf{w}^*)^{(t_n-k)})}(X_i)-T_{\beta_n}Y_i|^2\cdot\mathbb{1}_{[-\alpha_n,\alpha_n]^d}(X_i)\cdot 1_{A_n}\Big\}\Bigg)\\
&\quad+(1+\epsilon)\cdot\Bigg(\mathbf{E}\Big\{\frac{1}{n}\sum_{i=1}^n|f_{((\mathbf{w}^*)^{(t_n-k)})}(X_i)-T_{\beta_n}Y_i|^2\cdot\mathbb{1}_{[-\alpha_n,\alpha_n]^d}(X_i)\cdot 1_{A_n}\Big\}\\
&\qquad-\mathbf{E}\Big\{\mathbf{E}\Big\{|f_{((\mathbf{w}^*)^{(t_n-k)})}(X)-T_{\beta_n}Y|^2\cdot\mathbb{1}_{[-\alpha_n,\alpha_n]^d}(X)\cdot 1_{A_n}\Big|\mathcal{D}_n\Big\}\Big\}\Bigg)\\
&\quad+(1+\epsilon)\cdot\Bigg(\mathbf{E}\Big\{|f_{((\mathbf{w}^*)^{(t_n-k)})}(X)-T_{\beta_n}Y|^2\cdot\mathbb{1}_{[-\alpha_n,\alpha_n]^d}(X)\cdot 1_{A_n}\Big\}\\
&\qquad-(1+\epsilon)^2\cdot\mathbf{E}\Big\{|f_{((\mathbf{w}^*)^{(t_n-k)})}(X)-Y|^2\cdot\mathbb{1}_{[-\alpha_n,\alpha_n]^d}(X)\cdot 1_{A_n}\Big\}\Bigg)\\
&\quad+(1+\epsilon)^2\cdot\Bigg(\mathbf{E}\Big\{|f_{((\mathbf{w}^*)^{(t_n-k)})}(X)-Y|^2\cdot\mathbb{1}_{[-\alpha_n,\alpha_n]^d}(X)\cdot 1_{A_n}\Big\}
\end{aligned}$$

$$-\mathbf{E}\{|m(X)-Y|^2\}\cdot\mathbf{P}(A_n)\Big)$$

$$+((1+\epsilon)^2-1)\cdot\mathbf{E}\{|m(X)-Y|^2\}+c_2\cdot\sum_{j=1}^{K_n}|(\mathbf{w}^*)^{(L)}_{1,1,j}|^2$$

$$=T_{8,n}+T_{9,n}+T_{10,n}+T_{11,n}+T_{12,n}+T_{13,n}.$$

Similar to the second step, we obtain

$$\limsup_{n\to\infty} T_{8,n}\le 0 \quad \text{and} \quad \limsup_{n\to\infty} T_{10,n}\le 0.$$

According to the assertion of Lemma 6 we know that $f_{\mathbf{w}^*}$ is bounded. Thus, we get as in the sixth step

$$\limsup_{n\to\infty} T_{9,n}\le 0.$$

The choice of $\bar{m}$ and Lemma 7 imply

$$\begin{aligned}
&T_{11,n}/(1+\epsilon)^2\\
&\le \mathbf{E}\Big\{\int |f_{((\mathbf{w}^*)^{(t_n-k)})}(x)-m(x)|^2\mathbf{P}_X(dx)\cdot 1_{A_n}\Big\}\\
&\le 2\mathbf{E}\Big\{\int |f_{((\mathbf{w}^*)^{(t_n-k)})}(x)-\bar{m}(x)|^2\mathbf{P}_X(dx)\cdot 1_{A_n}\Big\}+2\int|\bar{m}(x)-m(x)|^2\mathbf{P}_X(dx)\\
&\le c_{44}\cdot\left(\frac{1}{K}+\frac{K^{12d}}{n^2}+\left(\frac{K^{6d}}{n}+1\right)^2\mathbf{P}_X(\mathbb{R}^d\setminus[-K,K]^d)\right)+2\epsilon,
\end{aligned}$$

from which we can conclude

$$\limsup_{n\to\infty} T_{11,n}\le c_{44}\cdot\left(\frac{1}{K}+\mathbf{P}_X(\mathbb{R}^d\setminus[-K,K]^d)\right)+2\epsilon\cdot(1+\epsilon)^2.$$

Furthermore we obtain by Lemma 7

$$T_{13,n}\le c_{45}\cdot(K^2+1)^{3d}\cdot\left(\frac{1}{(K^2+1)^{2d}}\right)^2.$$

Summarizing the above results yields

$$\begin{aligned}
&\limsup_{n\to\infty}\sum_{k=1}^{t_n}\frac{2\cdot c_2}{L_n}\cdot\left(1-\frac{2\cdot c_2}{L_n}\right)^{(k-1)}\Big(\mathbf{E}\{F_n((\mathbf{w}^*)^{(t_n-k)})\cdot 1_{A_n}\}\\
&\qquad -\mathbf{E}\{|m(X)-Y|^2\}\cdot\mathbf{P}(A_n)\Big)\\
&\le c_{46}\cdot\left(\frac{1}{K}+\mathbf{P}_X(\mathbb{R}^d\setminus[-K,K]^d)+\epsilon\right)+c_{45}\cdot(K^2+1)^{3d}\cdot\left(\frac{1}{(K^2+1)^{2d}}\right)^2
\end{aligned}$$

and

$$\limsup_{n\to\infty} \mathbf{E}\left\{T_{6,n}\right\} \leq (1+\epsilon)^2 \left(c_{46} \cdot \left(\frac{1}{K} + \mathbf{P}_X\left(\mathbb{R}^d \setminus [-K,K]^d\right) + \epsilon \right) + c_{47} \cdot \frac{1}{(K^2+1)^d} \right) + ((1+\epsilon)^2 - 1) \cdot \mathbf{E}\{|m(X)-Y|^2\}.$$

In the *ninth step of the proof* we finish the proof of Theorem 1. The results of steps 1,2,3,6,7 and 8 imply for $K \to \infty$

$$\limsup_{n\to\infty} \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_{48} \cdot \epsilon.$$

With $\epsilon \to 0$ we get the assertion. □

References

[1] Allen-Zhu, Z., Li, Y. and Song, Z. (2019). A Convergence Theory for Deep Learning via Over-Parameterization *Proceedings of the 36th International Conference on Machine Learning,* **97**, pp. 242-252

[2] Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *arXiv: 2103.09177v1.*

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

[4] Braun, A., Kohler, M., and Walk, H. (2019). On the rate of convergence of a neural network regression estimate learned by gradient descent. Preprint, arXiv: 1912.03921.

[5] Braun, A., Kohler, M., Langer, S., and Walk, H. (2021). The smoking gun: statistical theory improves neural network estimates. Preprint, arXiv: 2107.09550.

[6] Chizat, L. and Bach, F. (2018). On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. arXiv:1805.09545.

[7] Dyer, E. and Gur-Ari, G. (2019). Asymptotics of Wide Networks from Feynman Diagrams. arXiv: 1909.11304.

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

[9] Hanin, B. and Nica, M. (2019). Finite Depth and Width Corrections to the Neural Tangent Kernel. arXiv: 1909.05989.

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

[11] Kawaguchi, K. and Huang, J. (2019) Gradient Descent Finds Global Minima for Generalizable Deep Neural Networks of Practical Sizes *arxiv: 1908.02419*

[12] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. arXiv: 1408.5882.

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

[14] Kohler, M., and Krzyżak, A. (2021). Over-parametrized deep neural networks minimizing the empirical risk do not generalize well. *Bernoulli*, **27**, pp. 2564-2597.

[15] Kohler, M., and Krzyżak, A. (2022). Over-parametrized neural networks learned by gradient descent can generalize especially well. Submitted for publication.

[16] Kohler, M., and Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates using ReLU activation functions. *Annals of Statistics*, **49**, pp. 2231-2249.

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

[18] Langer, S. (2021a). Analysis of the rate of convergence of fully connected deep neural network regression estimates with smooth activation function, *Journal of Multivariate Analysis*, **182**, pp. 104695

[19] Langer, S. (2021b). Approximating smooth functions by deep neural networks with sigmoid activation function, *Journal of Multivariate Analysis*, **182**, pp. 104696

[20] Li, G., Gu, Y. and Ding, J. (2021). The Rate of Convergence of Variation-Constrained Deep Neural Networks. arXiv: 2106.12068

[21] Lu, J., Shen, Z., Yang, H. and Zhang, S. (2020) Deep Network Approximation for Smooth Functions. arxiv: 2001.03040

[22] Mei, S., Montanari, A. and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. In *Proceedings of the National Academy of Sciences*, **115**, pp. E7665-E7671.

[23] Nguyen, P.-M. and Pham, H. T. (2020). A Rigorous Framework for the Mean Field Limit of Multilayer Neural Networks arXiv:2001.1144.

[24] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function (with discussion). *Annals of Statistics* **48**, pp. 1875–1897.

[25] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Huber, T., et al. (2017). Mastering the game of go without human knowledge. *Nature* **550**, pp. 354-359.

[26] Suzuki, T. (2018). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. arXiv: 1810.08033.

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

[28] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**, pp. 595-645.

[29] Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M., Macherey, W., Krikum, M., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv: 1609.08144*.

[30] Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks, *Neural Networks*, **94**, pp. 103–114.

[31] Yarotsky, D. and Zhevnerchuk, A. (2020). The phase diagram of approximation rates for deep neural networks. In *Advances in Neural Information Processing Systems*, **33**, pp. 13005–13015.