Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent *

Michael Kohler[1] and Adam Krzyżak[2,†]

[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: kohler@mathematik.tu-darmstadt.de*

[2] *Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8, email: krzyzak@cs.concordia.ca*

March 21, 2023

Abstract

Estimation of a regression function from independent and identically distributed random variables is considered. The L_2 error with integration with respect to the design measure is used as an error criterion. Over-parametrized deep neural network estimates are defined where all the weights are learned by the gradient descent. It is shown that the expected L_2 error of these estimates converges to zero with the rate close to $n^{-1/(1+d)}$ in case that the regression function is Hölder smooth with Hölder exponent $p \in [1/2, 1]$. In case of an interaction model where the regression function is assumed to be a sum of Hölder smooth functions where each of the functions depends only on d^* many of d components of the design variable, it is shown that these estimates achieve the corresponding d^*-dimensional rate of convergence.

AMS classification: Primary 62G08; secondary 62G20.

Key words and phrases: neural networks, nonparametric regression, over-parametrization, rate of convergence.

1 Introduction

1.1 Deep learning

Deep learning, i.e., the fitting of deep neural networks to data, has achieved tremendous success in various applications in the past ten years. Deep neural networks are nowadays the most successful methods 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)).

*Running title: *Over-parametrized deep neural networks*

†Corresponding author. Tel: +1-514-848-2424 ext. 3007, Fax:+1-514-848-2830

Motivated by this huge success in applications there is also an increasing interest in theoretical properties of the estimates based on the deep neural networks. Here in the past years various impressive results concerning the rate of convergence of the least squares regression estimates based on the deep neural networks have been derived (cf. , e.g., Bauer and Kohler (2019), Schmidt-Hieber (2020), Kohler and Langer (2021), and the literature cited therein). But these results ignore two important properties of the deep neural network estimates applied in practice: Firstly, the most successful estimates are usually over-parametrized in the sense that the number of parameters of these estimates is much larger than the sample size. And secondly, these estimates are learned by the gradient descent applied to randomly initialized weights of the neural network. This motivates the question: If we define an over-parametrized deep neural network estimate by randomly initializing its weights and by performing a suitable number of gradient descent steps, does the resulting estimate then have nice theoretical properties? The purpose of this article is to give results that partially answer this question.

1.2 Nonparametric regression

We study deep neural networks in the context of nonparametric regression. Here, (X, Y) is an $\mathbb{R}^d \times \mathbb{R}$–valued random vector (X, Y) with $\mathbf{E}Y^2 < \infty$, and $m(x) = \mathbf{E}\{Y|X = x\}$ is the corresponding regression function $m : \mathbb{R}^d \to \mathbb{R}$. Given a sample of (X, Y), i.e., a data set

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

where (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n) are i.i.d., the goal is to construct an estimator

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

of the corresponding regression function $m(x) = \mathbf{E}\{Y|X = x\}$ 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).

We are interested to investigate for given estimates m_n how quickly the expected L_2 error

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

converges to zero. It is well-known, that without regularity assumptions on the smoothness of m it is not possible to derive nontrivial asymptotic bounds on (2) (cf., Theorem 7.2 and Problem 7.2 in Devroye, Györfi and Lugosi (1996) and Section 3 in Devroye and Wagner (1980)). In order to formulate such regularity assumptions we will use in this paper the notion of (p, C)–smoothness, which we introduce next.

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

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

for all $x, z \in \mathbb{R}^d$, where $\| \cdot \|$ denotes the Euclidean norm.

For $p \leq 1$ (p, C)–smoothness means that the function is Hölder-smooth with exponent p and Hölder-constant C.

1.3 Main results

Based on the previous work of the authors and their co-authors we introduce in this article a technique to analyze the rate of convergence of over-parametrized deep neural network estimates learned by gradient descent. The key ingredients in this theory are as follows: We control *generalization* ability of the estimate by using a metric entropy bound of a class of neural networks with bounded weights (cf., Lemma 4 below), by imposing proper bounds on the weights during initialization, and by choosing the number of gradient steps and the step size properly. We analyze *optimization* of the empirical L_2 risk during gradient descent by introducing a proper regularization, and by optimization of the outer weights during the gradient descent. Here we show at the same time that due to our restrictions on the number of gradient descent steps and the step size the inner weights do not change much (cf., Lemma 1 below). And we control *approximation* by using the over-parametrization and our special topology of the networks to show that with high probability a subset of the initial inner weights has nice properties and that this in turn leads to good approximation properties of the corresponding networks as soon as the outer weights are suitably chosen.

This new approach of analyzing over-parametrized deep neural networks is illustrated by analyzing the rate of convergence of an estimate introduced in Section 2 as follows: We choose topology of the network where the output of the network is defined as a linear combination of a huge number of fully connected deep neural networks of constant widths and depth. We introduce special initialization of the weights, where the output weights are zero and all inner weights are generated with uniform distribution. Then we perform a suitable large number of gradient descent steps with a suitably small step size. We show that the expected L_2 error of the truncated version of the resulting estimate converges to zero with the rate of convergence close to

$$n^{-\frac{1}{1+d}}$$

in case that the regression function is (p, C)-smooth for some $1/2 \leq p \leq 1$. Furthermore, we show that in case that the regression function is the sum of Hölder smooth functions where each of the functions depends only of d^* of d components of X, our estimate achieves the rate of convergence close to

$$n^{-\frac{1}{1+d^*}}.$$

1.4 Discussion of related results

In the last six years various results concerning the rate of convergence of the least squares regression estimates based on deep neural networks have been shown. One of the main achievement in this area is derivation of good rates of convergence for such estimates in case that the regression function is a composition of functions where each of these functions depends only on a few of its components. This was first shown in Kohler and Krzyżak (2017) for (p, C)–smooth regression functions with $p \leq 1$. Bauer and Kohler (2019) showed that such results also hold in case $p > 1$ provided the activation function of the network is sufficiently smooth. The surprising fact that such results also hold for the non-smooth ReLU activation function was shown in Schmidt-Hieber (2020). Kohler and Langer (2021) proved that such result can be also derived without imposing a sparsity constraint on the underlying networks. Suzuki (2018) and Suzuki and Nitanda (2019) proved corresponding results under weaker smoothness assumptions then in the above papers. That the least squares estimates based on the deep neural networks are able to adapt to some kind of local dimension of the regression function was shown in Kohler, Krzyżak and Langer (2022) and Eckle and Schmidt-Hieber (2019). For further results on least squares estimates based on deep neural networks we refer to Imaizumi and Fukamizu (2019) and Langer (2021a) and the literature cited therein. Various approximation results for deep neural network can be found in Lu et al. (2020), Yarotsky (2017),Yarotsky and Zhevnerchuk (2020) and Langer (2021b).

For neural networks with one hidden layer Braun et al. (2021) analyzed the gradient descent. There it was shown that in case of a proper initialization estimates learned by the gradient descent can achieve a rate of convergence of order $1/\sqrt{n}$ (up to a logarithmic factor) in case that the Fourier transform of the regression functions decays suitably fast. This decay of the Fourier transform is related to the classical results of Barron (1993, 1994) for the least squares estimate, where also a similar dimension-free rate of convergence was proven in case that the first moment of the Fourier transform of the regression function is finite. In case $d = 1$ the above estimate of Braun et al. (2021) was analyzed in Kohler and Krzyżak (2022) in an over-parametrized setting. By controlling the complexity of the estimate via strong regularization it was possible in this article to show that over-parametrization leads to an improved rate of convergence in case $d = 1$.

A review of various results on over-parametrized deep neural network estimates learned by gradient descent can be found in Bartlett, Montanari and Rakhlin (2021). These results usually analyze the estimates in some asymptotically equivalent models (like the mean field approach in Mei, Montanari, and Nguyen (2018), Chizat and Bach (2018) or Nguyen and Pham (2020) or the neural tangent approach in Hanin and Nica (2019)). In contrast, we analyze directly the expected L_2 error of the estimate in a standard regression model.

It is well-known that gradient descent applied to deep neural network can lead to estimates which minimizes the empirical L_2-risk (cf., e.g., Allen-Zhu, Li and Song (2019), Kawaguchi and Huang (2019) and the literature cited therein). However, as was shown in Kohler and Krzyżak (2022) such estimates, in general, do not perform well on data, which is independent of the training data. In this article we avoid this problem by restricting

the number of gradient descent steps and the step sizes and by imposing bounds on the absolute values of the initial weights.

Our approach is related to Drews and Kohler (2022), where the universal consistence of over-parametrized deep neural network estimate learned by gradient descent was shown. In fact, we generalize the results there such that we are able to analyze the rate of convergence of the estimates.Iin particular, this requires more precise analysis of the approximation error, which we do by using a multiscale approximation in Lemma 7 below.

Our bound on the covering number of over-parametrized deep neural networks is based on the corresponding result in Li, Gu and Ding (2021).

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. 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 $\mathcal{N}_p(\varepsilon, \mathcal{F}, x_1^n)$.

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 $(T_\beta f)(x) = T_\beta (f(x))$.

1.6 Outline

The over-parametrized deep neural network estimates considered in this paper are introduced in Section 2. The main results are presented in Section 3. Section 4 contains the proofs.

2 Definition of the estimate

Throughout the paper we let $\sigma(x) = 1/(1 + e^{-x})$ be the logistic squasher and we define the topology of our neural networks as follows: We let $K_n, L, r \in \mathbb{N}$ be parameters of our estimate and using these parameters we set

$$f_{\mathbf{w}}(x) = \sum_{j=1}^{K_n} w_{1,1,j}^{(L)} \cdot f_{j,1}^{(L)}(x) \tag{3}$$

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

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

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

$$f_{k,i}^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)}\right) \tag{5}$$

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

This means that we consider neural networks which consist of K_n fully connected neural networks of depth L and width r computed in parallel and compute a linear combination of the outputs of these K_n neural networks. The weights in the k-th such network are denoted by $(w_{k,i,j}^{(l)})_{i,j,l}$, where $w_{k,i,j}^{(l)}$ is the weight between neuron j in layer l and neuron i in layer $l+1$.

We initialize the weights $\mathbf{w}^{(0)} = ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l}$ as follows: We set

$$(\mathbf{w}^{(0)})_{1,1,k}^{(L)} = 0 \quad (k = 1, \ldots, K_n), \tag{6}$$

we choose $(\mathbf{w}^{(0)})_{k,i,j}^{(l)}$ uniformly distributed on $[-c_1 \cdot (\log n)^2, c_1 \cdot (\log n)^2]$ if $l \in \{1, \ldots, L-1\}$, and we choose $(\mathbf{w}^{(0)})_{k,i,j}^{(0)}$ uniformly distributed on $[-c_2 \cdot (\log n)^2 \cdot n^\tau, c_2 \cdot (\log n)^2 \cdot n^\tau]$, where $\tau > 0$ is a parameter of the estimate. Here the random values are defined such that all components of $\mathbf{w}^{(0)}$ are independent.

After initialization of the weights we perform $t_n \in \mathbb{N}$ gradient descent steps each with a step size $\lambda_n > 0$. Here we try to minimize the regularized empirical L_2 risk

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{\mathbf{w}}(X_i)|^2 + c_3 \cdot \sum_{k=1}^{K_n} |w_{1,1,k}^{(L)}|^2. \tag{7}$$

To do this we set

$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \lambda_n \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t-1)}) \quad (t = 1, \ldots, t_n). \tag{8}$$

Finally we define our estimate as a truncated version of the neural network with weight vector $\mathbf{w}^{(t_n)}$, i.e., we set

$$m_n(x) = T_{\beta_n}(f_{\mathbf{w}^{(t_n)}}(x)) \tag{9}$$

where $\beta_n = c_4 \cdot \log n$ and $T_\beta z = \max\{\min\{z, \beta\}, -\beta_n\}$ for $z \in \mathbb{R}$ and $\beta > 0$.

3 Main results

3.1 A general theorem

Our first result is a general theorem which we will apply in the next two subsections in order to analyze the rate of convergence of our over-parametrized deep neural network estimate.

Theorem 1 *Let $n \in \mathbb{N}$, let (X, Y), (X_1, Y_n), ..., (X_n, Y_n) be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$–valued random variables such that supp(X) is bounded and*

$$\mathbf{E}\left\{e^{c_5 \cdot Y^2}\right\} < \infty \tag{10}$$

holds and that the corresponding regression function $m(x) = \mathbf{E}\{Y|X = x\}$ is bounded.

Let $\sigma(x) = 1/(1 + e^{-x})$ be the logistic squasher, let $K_n, L, r, t_n \in \mathbb{N}$, $\lambda_n, \tau > 0$ and define the estimate m_n as in Section 2. Let $\tilde{K}_n \in \{1, \ldots, K_n\}$,

$$w_{k,i,j}^{(l)} \in [-c_1 \cdot (\log n)^2, c_1 \cdot (\log n)^2] \quad (l = 1, \ldots, L, k = 1, \ldots \tilde{K}_n)$$

and

$$w_{k,i,j}^{(0)} \in [-c_2 \cdot (\log n)^2 \cdot n^\tau, c_2 \cdot (\log n)^2 \cdot n^\tau] \quad (k = 1, \ldots, \tilde{K}_n).$$

Assume

$$\left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,j}^{(L)}(x) \right| \leq \beta_n \quad (x \in supp(X)) \tag{11}$$

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

Assume furthermore

$$\frac{K_n}{n^\kappa} \to 0 \quad (n \to \infty) \tag{12}$$

for some $\kappa > 0$ and

$$\frac{K_n}{\tilde{K}_n \cdot n^{r \cdot (d+1) \cdot \tau + 1}} \to \infty \quad (n \to \infty), \tag{13}$$

and that t_n, λ_n are given by

$$t_n = \lceil c_6 \cdot L_n \cdot \log n \rceil \quad \textit{and} \quad \lambda_n = \frac{1}{L_n} \tag{14}$$

for some $L_n > 0$ which satisfies

$$L_n \geq (\log n)^{10 \cdot L + 10} \cdot K_n^{3/2}. \tag{15}$$

Assume

$$2 \cdot c_3 \cdot c_6 \geq 1, \quad c_4 \cdot c_5 \geq 1 \quad \textit{and} \quad 4 \cdot c_4 \cdot c_6 \leq 1. \tag{16}$$

Then we have for any $\epsilon > 0$

$$\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_7 \cdot \left(\frac{n^{\tau \cdot d + \epsilon}}{n} + \sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 \right.$$
$$\left. + \sup_{\substack{(\bar{w}_{i,j,k}^{(l)})_{i,j,k,l}: \\ |\bar{w}_{i,j,k}^{(l)} - w_{i,j,k}^{(l)}| \leq \log n \ (l=0,\ldots,L-1)}} \int |\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x) - m(x)|^2 \mathbf{P}_X(dx) \right).$$

Remark 1. The upper bound on the expected L_2 error of our neural network estimates corresponds to the usual bounds for least squares estimates (cf., e.g., Theorem 11.5 in Györfi et al. (2002)). The term

$$\frac{n^{\tau \cdot d + \epsilon}}{n}$$

is used to bound the estimation error, which comes from the fact that we minimize the empirical L_2 risk and not the L_2 risk during gradient descent. And

$$\sup_{\substack{(\bar{w}_{i,j,k}^{(l)})_{i,j,k,l}: \\ |\bar{w}_{i,j,k}^{(l)} - w_{i,j,k}^{(l)}| \leq \log n \ (l=0,\ldots,L-1)}} \int |\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x) - m(x)|^2 \mathbf{P}_X(dx)$$

is used to bound the approximation error, which occurs since we restrict our estimate to our class of neural networks. Observe that here we can choose $\mathbf{w}$ optimally in view of the upper bound in Theorem 1, so in fact our approximation error also includes a minimum over all possible weights $\mathbf{w}$. The additional term

$$\sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2$$

is needed to bound an additional error due to gradient descent.

Remark 2. Condition (16) is e.g. satisfied, if we set

$$c_4 = \frac{1}{c_5}, \quad c_6 = \frac{c_5}{4} \quad \text{and} \quad c_3 = \frac{1}{8 \cdot c_5}.$$

3.2 Rate of convergence for (p, C)–smooth regression functions

The main challenge in deriving a rate of convergence from Theorem 1 above is to derive an approximation result for a smooth regression function and our neural networks such that the bounds on the weights of the networks are small. The first result in this respect is shown in our next theorem.

Theorem 2 *Let* $n \in \mathbb{N}$, *let* (X, Y), (X_1, Y_n), $\ldots$, (X_n, Y_n) *be independent and identically distributed* $\mathbb{R}^d \times \mathbb{R}$ *valued random variables which satisfy supp*$(X) \subseteq [0,1]^d$ *and (10).*

Assume that the corresponding regression function $m(x) = \mathbf{E}\{Y|X = x\}$ *is* (p, C)*-smooth for some* $1/2 \leq p \leq 1$ *and some* $C > 0$.

Let $\sigma(x) = 1/(1 + e^{-x})$ *be the logistic squasher, let* $L, r \in \mathbb{N}$ *with* $L \geq 2$ *and* $r \geq 2d$, *set*

$$K_n = n^{6d+r+2},$$

$$\tau = \frac{1}{1+d}$$

and

$$t_n = \lceil c_6 \cdot L_n \cdot \log n \rceil \quad \textit{and} \quad \lambda_n = \frac{1}{L_n}$$

for some $L_n > 0$ *which satisfies (15). Define the estimate as in Section 2 and assume that (16) holds.*

Then we have for any $\epsilon > 0$

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_8 \cdot n^{-\frac{1}{1+d}+\epsilon}.$$

Remark 3. According to Stone (1982), the optimal minimax L_2 rate of convergence in case of (p, C)–smooth regression function is

$$n^{-\frac{2p}{2p+d}}.$$

For $p = 1/2$ our estimate achieves a rate of convergence, which is arbitrary close to this rate of convergence. For $p > 1/2$ our derived rate of convergence is not optimal. We conjecture this is a consequence of our proof and not of property of the estimate, but it is an open problem to prove this.

3.3 Rate of convergence in an interaction model

In this subsection we assume that the regression function satisfies

$$m(x) = \sum_{I \subseteq \{1,\ldots,d\} : |I| = d^*} m_I(x_I),$$

where $1 \leq d^* < d$, $m_I : \mathbb{R}^{d^*} \to \mathbb{R}$ $(I \subseteq \{1, \ldots, d\}, |I| = d^*)$ are (p, C)-smooth functions and we use the notation

$$x_I = (x^{(j_1)}, \ldots, x^{(j_{d^*})})$$

for $I = \{j_1, \ldots, j_{d^*}\}$. Our aim is the modify the estimate of Subsection 3.2 such that it achieves in this case the d^*-dimensional rate of convergence.

To achieve this, we define

$$f_{\mathbf{w}}(x) = \sum_{I \subseteq \{1,\ldots,d\} : |I| = d^*} f_{\mathbf{w}_I}(x_I)$$

where $f_{\mathbf{w}_I}$ is defined by (3)–(5) with d replaced by d^* and weight vector $\mathbf{w}_I$, and

$$\mathbf{w} = (\mathbf{w}_I)_{I \subseteq \{1,\dots,d\}, |I|=d^*} \cdot$$

We initialize the weights $\mathbf{w}^{(0)} = (((\mathbf{w}_I^{(0)})_{k,i,j}^{(l)})_{k,i,j,l})_{I \subseteq \{1,\dots,d\}, |I|=d^*}$ as follows: We set

$$(\mathbf{w}_I^{(0)})_{1,1,k}^{(L)} = 0 \quad (k = 1, \dots, K_n, I \subseteq \{1, \dots, d\}, |I| = d^*),$$

we choose $(\mathbf{w}_I^{(0)})_{k,i,j}^{(l)}$ uniformly distributed on $[-c_1 \cdot (\log n)^2, c_1 \cdot (\log n)^2]$ if $l \in \{1, \dots, L-1\}$, and we choose $(\mathbf{w}_I^{(0)})_{k,i,j}^{(0)}$ uniformly distributed on $[-c_2 \cdot (\log n)^2 \cdot n^\tau, c_2 \cdot (\log n)^2 \cdot n^\tau]$, where $\tau > 0$ is a parameter of the estimate defined in Theorem 3 below ($I \subseteq \{1, \dots, d\}, |I| = d^*$). Here the random values are defined such that all components of $\mathbf{w}^{(0)}$ are independent.

After initialization of the weights we perform $t_n \in \mathbb{N}$ gradient descent steps each with a step size $\lambda_n > 0$. Here we try to minimize the regularized empirical L_2 risk

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{\mathbf{w}}(X_i)|^2 + c_3 \cdot \sum_{I \subseteq \{1,\dots,d\} : |I|=d^*} \sum_{k=1}^{K_n} |(\mathbf{w}_I)_{1,1,k}^{(L)}|^2.$$

To do this we set

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

Finally we define our estimate as a truncated version of the neural network with weight vector $\mathbf{w}^{(t_n)}$, i.e., we set

$$m_n(x) = T_{\beta_n}(f_{\mathbf{w}^{(t_n)}}(x))$$

where $\beta_n = c_4 \cdot \log n$

Theorem 3 *Let $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$, $1/2 \leq p \leq 1$, $C > 0$, let $n \in \mathbb{N}$, let (X, Y), (X_1, Y_n), …, (X_n, Y_n) be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ valued random variables such that supp$(X) \subseteq [0,1]^d$ and (10) holds. Assume that the corresponding regression function $m(x) = \mathbf{E}\{Y|X = x\}$ satisfies*

$$m(x) = \sum_{I \subseteq \{1,\dots,d\} : |I|=d^*} m_I(x_I) \quad (x \in [0,1]^d)$$

for some (p, C)–smooth functions $m_I : \mathbb{R}^{d^*} \to \mathbb{R}$ ($I \subseteq \{1, \dots, d\}, |I| = d^*$).

Let $\sigma(x) = 1/(1 + e^{-x})$ be the logistic squasher, let $L, r \in \mathbb{N}$ with $L \geq 2$ and $r \geq 2d^*$, set

$$K_n = n^{6d^*+r+2},$$

$$\tau = \frac{1}{1 + d^*}$$

and

$$t_n = \lceil c_6 \cdot L_n \cdot \log n \rceil \quad \text{and} \quad \lambda_n = \frac{1}{L_n}$$

for some $L_n > 0$ which satisfies (15). Define the estimate as above.

Then we have for any $\epsilon > 0$

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_9 \cdot n^{-\frac{1}{1+d^*}+\epsilon}.$$

Remark 4. The rate of convergence derived in Theorem 3 does not depend on d, hence under the above assumption on the regression function our estimate is able to circumvent the curse of dimensionality. That this is possible is well-known (cf., Stone (1994) and the literature cited therein), however our result is the first result which shows that this is also possible for (over-parametrized) neural network estimates learned by the gradient descent.

4 Proofs

4.1 Auxiliary results for the proof of Theorem 1

In this section we present five auxiliary results which we will use in the proof of Theorem 1. Our first auxiliary result will play key role in the analysis of gradient descent.

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

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

Assume

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\| \leq \sqrt{2 \cdot t \cdot L \cdot \max\{F(\mathbf{a}_0), 1\}} \tag{17}$$

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

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}) - (\nabla_{\mathbf{a}} F)(\mathbf{b})\| \leq L \cdot \|\mathbf{a} - \mathbf{b}\| \tag{18}$$

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

$$\|\mathbf{a} - \mathbf{a}_0\| \leq \sqrt{8 \cdot \frac{t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}} \quad \text{and} \quad \|\mathbf{b} - \mathbf{a}_0\| \leq \sqrt{8 \cdot \frac{t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}}. \tag{19}$$

Then we have

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

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

and

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

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

Our next auxiliary result will help us to show that assumption (17) is satisfied in the proof of Theorem 1.

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

$$|w_{1,1,k}^{(L)}| \leq \gamma_n^* \quad (k = 1, \ldots, K_n), \tag{20}$$

$$|w_{k,i,j}^{(l)}| \leq B_n \quad \textit{for } l = 1, \ldots, L-1 \tag{21}$$

and

$$\|\mathbf{w} - \mathbf{v}\|_\infty^2 \leq \frac{2t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}. \tag{22}$$

Then we have

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w})\| \leq c_{10} \cdot K_n^{3/2} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2 \cdot \sqrt{\frac{t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}}.$$

Proof. See Lemma 2 in Drews and Kohler (2022). □

Our third auxiliary result will help us to show that assumption (18) is satisfied in the proof of Theorem 1.

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

$$|\max\{(\mathbf{w}_1)_{1,1,k}^{(L)}, (\mathbf{w}_2)_{1,1,k}^{(L)}\}| \leq \gamma_n^* \quad (k = 1, \ldots, K_n), \tag{23}$$

$$|\max\{(\mathbf{w}_1)_{k,i,j}^{(l)}, (\mathbf{w}_2)_{k,i,j}^{(l)}\}| \leq B_n \quad \textit{for } l = 1, \ldots, L-1 \tag{24}$$

and

$$\|\mathbf{w}_2 - \mathbf{v}\|^2 \leq 8 \cdot \frac{t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}. \tag{25}$$

Then we have

$$\begin{aligned}
&\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}_2)\| \\
&\leq c_{11} \cdot \max\{\sqrt{F_n(\mathbf{v})}, 1\} \cdot (\gamma_n^*)^2 \cdot B_n^{3L-1} \cdot \alpha_n^3 \cdot K_n^{3/2} \cdot \sqrt{\frac{t_n}{L_n}} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|.
\end{aligned}$$

Proof. See Lemma 3 in Drews and Kohler (2022). □

Our fourth auxiliary result uses a metric entropy bound in order to control the complexity of a set of over-parametrized deep neural networks.

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 (3)–(5) where the weight vector $\mathbf{w}$ satsifies*

$$\sum_{j=1}^{K_n} |w_{1,1,j}^{(L)}| \leq C, \tag{26}$$

$$|w_{k,i,j}^{(l)}| \leq B \quad (k \in \{1, \ldots, K_n\}, i, j \in \{1, \ldots, r\}, l \in \{1, \ldots, L-1\}) \tag{27}$$

and

$$|w_{k,i,j}^{(0)}| \leq A \quad (k \in \{1, \ldots, K_n\}, i \in \{1, \ldots, r\}, j \in \{1, \ldots, d\}). \tag{28}$$

Then we have for any $1 \leq p < \infty$, $0 < \epsilon < \beta$ and $x_1^n \in [-\alpha, \alpha]^d$

$$\begin{aligned} &\mathcal{N}_p\left(\epsilon, \{T_\beta f \,:\, f \in \mathcal{F}\}, x_1^n\right) \\ &\leq \left(c_{12} \cdot \frac{\beta^p}{\epsilon^p}\right)^{c_{13} \cdot \alpha^d \cdot A^d \cdot B^{(L-1)\cdot d} \left(\frac{C}{\epsilon}\right)^{d/k} + c_{14}}. \end{aligned}$$

Proof. Standard application of the chain rule for derivation together with the above bounds on the weight vector $\mathbf{w}$ shows that we have for any $x \in \mathbb{R}^d$ and any $s_1, \ldots, s_k \in \{1, \ldots, d\}$

$$\left| \frac{\partial^k f_{\mathbf{w}}}{\partial x^{(s_1)} \ldots \partial x^{(s_k)}}(x) \right| \leq c_{15} \cdot C \cdot B^{(L-1)\cdot k} \cdot A^k =: c$$

(cf., e.g., proof of Lemma 4 in Drews and Kohler (2022)).

Let $\mathcal{G} \circ \Pi$ be the set of all piecewise polynomials of total degree less than k with respect to a partition Π of $[-\alpha, \alpha]^d$ into cubes of sidelength

$$\left(c_{16} \cdot \frac{\epsilon}{c}\right)^{1/k},$$

where $c_{16} = c_{16}(d, k)$ is a suitable small constant greater zero. Then a standard bound on the remainder of a multivariate Taylor polynomial shows that for each $f_{\mathbf{w}}$ we can find $g \in \mathcal{G} \circ \Pi$ such that

$$|f_{\mathbf{w}}(x) - g(x)| \leq \frac{1}{2}$$

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

$$\mathcal{N}_p\left(\epsilon, \{T_\beta f \,:\, f \in \mathcal{F}\}, x_1^n\right) \leq \mathcal{N}_p\left(\frac{\epsilon}{2}, \{T_\beta g \,:\, g \in \mathcal{G}\}, x_1^n\right).$$

$\mathcal{G}$ is a linear vector space of dimension less than or equal to

$$c_{17} \cdot \alpha^d \cdot \left(\frac{c}{\epsilon}\right)^{d/k},$$

from which we get the assertion by an application of Theorems 9.4 and 9.5 in Györfi et al. (2002). □

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 $c_3 > 0$. In the next lemma we consider the problem to minimize

$$F(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} |\sum_{k=1}^{K} a_k \cdot B_k(x_i) - y_i|^2 + c_3 \cdot \sum_{k=1}^{K_n} a_k^2, \tag{29}$$

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

$$\mathbf{a}^{(t+1)} = \mathbf{a}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)}) \tag{30}$$

for some properly chosen $\lambda_n > 0$.

Lemma 5 *Let F be defined by (29) 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

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2 \geq 4 \cdot c_3 \cdot (F(\mathbf{a}) - F(\mathbf{a}_{opt})).$$

Proof. See Lemma 8 in Drews and Kohler (2022). □

4.2 Proof of Theorem 1

In the proof we combine ideas from the proof of Theorem 1 in Drews and Kohler (2022) with ideas from the proof of Lemma 1 in Bauer and Kohler (2019).

W.l.o.g. we assume throughout the proof that n is sufficiently large and that $\|m\|_\infty \leq \beta_n$ holds. Let A_n be the event that firstly the weight vector $\mathbf{w}^{(0)}$ satisfies

$$|(\mathbf{w}^{(0)})_{j_s,k,i}^{(l)} - \mathbf{w}_{j_s,k,i}^{(l)}| \leq \log n \quad \text{for all } l \in \{0, \ldots, L-1\}, s \in \{1, \ldots, \tilde{K}_n\}$$

for some pairwise distinct $j_1, \ldots, j_{\tilde{K}_n} \in \{1, \ldots, K_n\}$ and such that secondly

$$\max_{i=1,\ldots,n} |Y_i| \leq \sqrt{\beta_n}$$

holds.

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

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

and

$$((\mathbf{w}^*)^{(t)})_{1,1,j_k}^{(L)} = w_{1,1,k}^{(L)} \quad \text{for all } k = 1, \ldots, \tilde{K}_n$$

and

$$((\mathbf{w}^*)^{(t)})_{1,1,k}^{(L)} = 0 \quad \text{for all } k \notin \{j_1, \ldots, j_{\tilde{K}_n}\}.$$

We decompose the L_2 error of m_n in a sum of several terms. Set

$$m_{\beta_n}(x) = \mathbf{E}\{T_{\beta_n} Y | X = x\}.$$

We have

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&= \left(\mathbf{E}\left\{|m_n(X) - Y|^2 | \mathcal{D}_n\right\} - \mathbf{E}\{|m(X) - Y|^2\}\right) \cdot 1_{A_n} + \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{A_n^c} \\
&= \Big[\mathbf{E}\left\{|m_n(X) - Y|^2 | \mathcal{D}_n\right\} - \mathbf{E}\{|m(X) - Y|^2\} \\
&\qquad - \left(\mathbf{E}\left\{|m_n(X) - T_{\beta_n} Y|^2 | \mathcal{D}_n\right\} - \mathbf{E}\{|m_{\beta_n}(X) - T_{\beta_n} Y|^2\}\right)\Big] \cdot 1_{A_n} \\
&\quad + \Big[\mathbf{E}\left\{|m_n(X) - T_{\beta_n} Y|^2 | \mathcal{D}_n\right\} - \mathbf{E}\{|m_{\beta_n}(X) - T_{\beta_n} Y|^2\} \\
&\qquad -2 \cdot \frac{1}{n} \sum_{i=1}^n \left(|m_n(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2\right)\Big] \cdot 1_{A_n} \\
&\quad + \Big[2 \cdot \frac{1}{n} \sum_{i=1}^n |m_n(X_i) - T_{\beta_n} Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \\
&\qquad - \left(2 \cdot \frac{1}{n} \sum_{i=1}^n |m_n(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2\right)\Big] \cdot 1_{A_n} \\
&\quad + \Big[2 \cdot \frac{1}{n} \sum_{i=1}^n |m_n(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2\Big] \cdot 1_{A_n} \\
&\quad + \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{A_n^c} \\
&=: \sum_{j=1}^5 T_{j,n}.
\end{aligned}$$

In the reminder of the proof we bound

$$\mathbf{E} T_{j,n}$$

for $j \in \{1, \ldots, 5\}$.

In the *first step of the proof* we show

$$\mathbf{E} T_{1,n} \leq c_{18} \cdot \frac{\log n}{n}.$$

This follows as in the proof of Lemma 1 in Bauer and Kohler (2019).

In the *second step of the proof* we show

$$\mathbf{E} T_{3,n} \leq c_{19} \cdot \frac{\log n}{n}.$$

Again this follows from the proof of Lemma 1 in Bauer and Kohler (2019).

In the *third step of the proof* we show

$$\mathbf{E}T_{5,n} \leq c_{20} \cdot \frac{(\log n)^2}{n}.$$

The definition of m_n implies $\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq 4 \cdot c_4^2 \cdot (\log n)^2$, hence it suffices to show

$$\mathbf{P}(A_n^c) \leq \frac{c_{21}}{n}. \tag{31}$$

To do this, we consider sequential choice of the weights of the K_n fully connected neural networks. Probability that the weights in the first of these networks differ in all components at most by $\log n$ from $w_{1,i,j}^{(l)}$ $(l = 0, \ldots, L-1)$ is for large n bounded below by

$$\left(\frac{\log n}{2 \cdot c_1 \cdot (\log n)^2}\right)^{r \cdot (r+1) \cdot (L-1)} \cdot \left(\frac{\log n}{2 \cdot c_2 \cdot n^\tau}\right)^{r \cdot (d+1)} \quad \geq \quad n^{-r \cdot (d+1) \cdot \tau - 0.5}.$$

Hence probability that none of the first $n^{r \cdot (d+1) \cdot \tau + 1}$ neural networks satisfies this condition is for large n bounded above by

$$\begin{aligned} (1 - n^{-r \cdot (d+1) \cdot \tau - 0.5})^{n^{r \cdot (d+1) \cdot \tau + 1}} &\leq \left(\exp\left(-n^{-r \cdot (d+1) \cdot \tau - 0.5}\right)\right)^{n^{r \cdot (d+1) \cdot \tau + 1}} \\ &= \exp(-n^{0.5}). \end{aligned}$$

Since we have $K_n \geq n^{r \cdot (d+1) \cdot \tau + 1} \cdot \tilde{K}_n$ for n large we can successively use the same construction for all of $\tilde{K}_n$ weights and we can conclude: Probability that there exists $k \in \{1, \ldots, \tilde{K}_n\}$ such that none of the K_n weight vectors of the fully connected neural network differs by at most $\log n$ from $(w_{i,j,k}^{(l)})_{i,j,l}$ is for large n bounded from above by

$$\tilde{K}_n \cdot \exp(-n^{0.5}) \leq n^\kappa \cdot \exp(- \cdot n^{0.5}) \leq \frac{c_{22}}{n}.$$

This implies for large n

$$\begin{aligned} \mathbf{P}(A_n^c) &\leq \frac{c_{22}}{n} + \mathbf{P}\{\max_{i=1,\ldots,n} |Y_i| > \sqrt{\beta_n}\} \leq \frac{c_{22}}{n} + n \cdot \mathbf{P}\{|Y| > \sqrt{\beta_n}\} \\ &\leq \frac{c_{22}}{n} + n \cdot \frac{\mathbf{E}\{\exp(c_5 \cdot Y^2)}{\exp(c_5 \cdot \beta_n)} \leq \frac{c_{19}}{n}, \end{aligned}$$

where the last inequlity holds because of (10) and $c_4 \cdot c_5 \geq 1$.

Let $\epsilon > 0$ be arbitrary. In the *fourth step of the proof* we show

$$\mathbf{E}T_{2,n} \leq c_{23} \cdot \frac{n^{\tau \cdot d + \epsilon}}{n}.$$

Let $\mathcal{W}_n$ be the set of all weight vectors $(w_{i,j,k}^{(l)})_{i,j,k,l}$ which satisfy

$$|w_{1,1,k}^{(L)}| \leq (c_1 + 1) \cdot (\log n)^2 \quad (k = 1, \ldots, K_n),$$

$$|w_{i,j,k}^{(l)}| \leq (c_1+1)\cdot(\log n)^2 \quad (l=1,\ldots,L-1)$$

and

$$|w_{i,j,k}^{(0)}| \leq (c_2+1)\cdot(\log n)^2\cdot n^{\tau}.$$

By Lemma 1, Lemma 2 and Lemma 3 we can conclude that on A_n we have

$$\|\mathbf{w}^{(t)} - \mathbf{w}^{(0)}\| \leq \log n \quad (t=1,\ldots,t_n). \tag{32}$$

This follows from the fact that on A_n we have

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

and that

$$\frac{2\cdot t_n}{L_n}\cdot\beta_n \leq 4\cdot c_4\cdot c_6\cdot(\log n)^2 \leq (\log n)^2.$$

Together with the initial choice of $\mathbf{w}^{(0)}$ this implies that on A_n we have

$$\mathbf{w}^{(t)} \in \mathcal{W}_n \quad (t=0,\ldots,t_n).$$

Hence, for any $u>0$ we get

$$\begin{aligned}
&\mathbf{P}\{T_{2,n} > u\} \\
&\leq \mathbf{P}\Bigg\{\exists f\in\mathcal{F}_n : \mathbf{E}\left(\left|\frac{f(X)}{\beta_n} - \frac{T_{\beta_n}Y}{\beta_n}\right|^2\right) - \mathbf{E}\left(\left|\frac{m_{\beta_n}(X)}{\beta_n} - \frac{T_{\beta_n}Y}{\beta_n}\right|^2\right) \\
&\qquad\qquad -\frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{f(X_i)}{\beta_n} - \frac{T_{\beta_n}Y_i}{\beta_n}\right|^2 - \left|\frac{m_{\beta_n}(X_i)}{\beta_n} - \frac{T_{\beta_n}Y_i}{\beta_n}\right|^2\right)\Bigg\} \\
&\qquad > \frac{1}{2}\cdot\left(\frac{u}{\beta_n^2} + \mathbf{E}\left(\left|\frac{f(X)}{\beta_n} - \frac{T_{\beta_n}Y}{\beta_n}\right|^2\right) - \mathbf{E}\left(\left|\frac{m_{\beta_n}(X)}{\beta_n} - \frac{T_{\beta_n}Y}{\beta_n}\right|^2\right)\right),
\end{aligned}$$

where

$$\mathcal{F}_n = \left\{T_{\beta_n} f_{\mathbf{w}} \quad : \quad \mathbf{w}\in\mathcal{W}_n\right\}.$$

By Lemma 4 we get

$$\begin{aligned}
&\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n}\cdot f : f\in\mathcal{F}_n\right\}, x_1^n\right) \leq \mathcal{N}_1\left(\delta\cdot\beta_n, \mathcal{F}_n, x_1^n\right) \\
&\leq \left(\frac{c_{24}}{\delta}\right)^{c_{25}\cdot(\log n)^{2d} n^{\tau\cdot d}\cdot(\log n)^{2\cdot(L-1)\cdot d}\cdot\left(\frac{K_n\cdot(\log n)^2}{\beta_n\cdot\delta}\right)^{d/k} + c_{26}}.
\end{aligned}$$

By choosing k large enough we get for $\delta > 1/n^2$

$$\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n}\cdot f : f\in\mathcal{F}_n\right\}, x_1^n\right) \leq c_{27}\cdot n^{c_{28}\cdot n^{\tau\cdot d+\epsilon/2}}.$$

This together with Theorem 11.4 in Györfi et al. (2002) leads for $u \geq 1/n$ to

$$\mathbf{P}\{T_{2,n} > u\} \leq 14 \cdot c_{27} \cdot n^{c_{28} \cdot n^{\tau \cdot d + \epsilon/2}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot u\right).$$

For $\epsilon_n \geq 1/n$ we can conclude

$$\begin{aligned}
\mathbf{E}\{T_{2,n}\} &\leq \epsilon_n + \int_{\epsilon_n}^{\infty} \mathbf{P}\{T_{2,n} > u\}\, du \\
&\leq \epsilon_n + 14 \cdot c_{27} \cdot n^{c_{28} \cdot n^{\tau \cdot d + \epsilon/2}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot \epsilon_n\right) \cdot \frac{5136 \cdot \beta_n^2}{n}.
\end{aligned}$$

Setting

$$\epsilon_n = \frac{5136 \cdot \beta_n^2}{n} \cdot c_{28} \cdot n^{\tau \cdot d + \epsilon/2} \cdot \log n$$

yields the assertion of the fourth step of the proof.

In the *fifth step of the proof* we show

$$\begin{aligned}
&\mathbf{E}\{T_{4,n}\} \\
&\leq c_{29} \cdot \Bigg(\sup_{\substack{(\bar{w}_{i,j,k}^{(l)})_{i,j,k,l}: \\ |\bar{w}_{i,j,k}^{(l)} - w_{i,j,k}^{(l)}| \leq \log n \ (l=0,\ldots,L-1)}} \int |\sum_{k=1}^{\tilde{K}_n} w_{1,1,k} \cdot f_{\bar{\mathbf{w}},k,j}^{(L)}(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\qquad + \sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 + \frac{(\log n)^2}{n} + \frac{n^{\tau \cdot d + \epsilon}}{n} \Bigg).
\end{aligned}$$

Using

$$|T_{\beta_n} z - y| \leq |z - y| \quad \text{for } |y| \leq \beta_n$$

we get

$$\begin{aligned}
&T_{4,n}/2 \\
&= \Big[\frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2\Big] \cdot 1_{A_n} \\
&\leq \Big[\frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}^{(t_n)}}(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2\Big] \cdot 1_{A_n} \\
&\leq \Big[F_n(\mathbf{w}^{(t_n)}) - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2\Big] \cdot 1_{A_n}.
\end{aligned}$$

Application of Lemma 1 (which is applicable on A_n because of Lemma 2 and Lemma 3) implies that this in turn is less than

$$\Big[F_n(\mathbf{w}^{(t_n - 1)}) - \frac{1}{2L_n} \cdot \|\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t_n - 1)})\|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2\Big] \cdot 1_{A_n}.$$

Since the sum of squares of all partial derivatives is at least as large as the sum of squares of the partial derivatives with respect to the outer weights $w_{1,1,k}^{(L)}$ $(k = 1, \dots, K_n)$, we can upper bound this in turn via Lemma 5 by

$$\begin{aligned}
&\big[F_n(\mathbf{w}^{(t_n-1)}) - \frac{1}{2L_n} \cdot 4 \cdot c_3 \cdot (F_n(\mathbf{w}^{(t_n-1)}) - F_n((\mathbf{w}^*)^{(t_n-1)}) \\
&\qquad\qquad -\frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\Big] \cdot 1_{A_n} \\
&= \Big[\Big(1 - \frac{2 \cdot c_3}{L_n}\Big) \cdot F_n(\mathbf{w}^{(t_n-1)}) + \frac{2 \cdot c_3}{L_n} \cdot F_n((\mathbf{w}^*)^{(t_n-1)}) - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\Big] \cdot 1_{A_n}.
\end{aligned}$$

Applying this argument repeatedly shows that

$$\begin{aligned}
&T_{4,n}/2 \\
&\leq \Big[\Big(1 - \frac{2 \cdot c_3}{L_n}\Big)^{t_n} \cdot F_n(\mathbf{w}^{(0)}) + \sum_{k=1}^{t_n} \frac{2 \cdot c_3}{L_n} \cdot \Big(1 - \frac{2 \cdot c_3}{L_n}\Big)^{k-1} F_n((\mathbf{w}^*)^{(t_n-k)}) \\
&\qquad\qquad -\frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\Big] \cdot 1_{A_n}.
\end{aligned}$$

This implies

$$\begin{aligned}
&\mathbf{E}\{T_{4,n}/2\} \\
&\leq \Big(1 - \frac{2 \cdot c_3}{L_n}\Big)^{t_n} \cdot \mathbf{E}\{Y^2\} + \sum_{k=1}^{t_n} \frac{2 \cdot c_3}{L_n} \cdot \Big(1 - \frac{2 \cdot c_3}{L_n}\Big)^{k-1} \cdot \\
&\quad \mathbf{E}\Bigg(\Bigg(\frac{1}{n}\sum_{i=1}^{n} |\sum_{k=1}^{K_n} w_{1,1,k}^{(L)} \cdot f^{(L)}_{(\mathbf{w}^*)^{(t_n-k)},j,1}(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\Bigg) \cdot 1_{A_n}\Bigg) \\
&\quad + c_3 \cdot \sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 \\
&\leq \Big(1 - \frac{2 \cdot c_3}{L_n}\Big)^{t_n} \cdot \mathbf{E}\{Y^2\} + c_3 \cdot \sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 \\
&\quad + \sum_{k=1}^{t_n} \frac{2 \cdot c_3}{L_n} \cdot \Big(1 - \frac{2 \cdot c_3}{L_n}\Big)^{k-1} \cdot 2 \cdot \\
&\qquad \Bigg(\sup_{\substack{(\bar{w}_{i,j,k}^{(l)})_{i,j,k,l}: \\ |\bar{w}_{i,j,k}^{(l)} - w_{i,j,k}^{(l)}| \leq \log n \, (l=0,\dots,L-1)}} \int |\sum_{k=1}^{\tilde{K}_n} w_{1,1,k} \cdot f^{(L)}_{\bar{\mathbf{w}},k,1}(x) - m(x)|^2 \mathbf{P}_X(dx)\Bigg) \\
&\quad + \sum_{k=1}^{t_n} \frac{2 \cdot c_3}{L_n} \cdot \Big(1 - \frac{2 \cdot c_3}{L_n}\Big)^{k-1} \cdot
\end{aligned}$$

$$\mathbf{E}\Bigg(\Bigg(\frac{1}{n}\sum_{i=1}^{n}|\sum_{k=1}^{K_n} w_{1,1,k}\cdot f^{(L)}_{(\mathbf{w}^*)^{(t_n-k)},j,1}(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2$$
$$-2\cdot\left(\mathbf{E}\{|\sum_{k=1}^{K_n} w_{1,1,k}\cdot f^{(L)}_{(\mathbf{w}^*)^{(t_n-k)},j,1}(X)-Y|^2|\mathcal{D}_n\}-\mathbf{E}\{|m(X)-Y|^2\}\right)\Bigg)\cdot 1_{A_n}\Bigg),$$

where the last inequality followed from (32). Arguing as in the beginning of the proof (and using in particularly the arguments from Steps 1, 2 and 4) we get

$$\mathbf{E}\Bigg(\Bigg(\frac{1}{n}\sum_{i=1}^{n}|\sum_{k=1}^{K_n} w_{1,1,k}\cdot f^{(L)}_{(\mathbf{w}^*)^{(t_n-k)},j,1}(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2$$
$$-2\cdot\left(\mathbf{E}\{|\sum_{k=1}^{K_n} w_{1,1,k}\cdot f^{(L)}_{(\mathbf{w}^*)^{(t_n-k)},j,1}(X)-Y|^2|\mathcal{D}_n\}-\mathbf{E}\{|m(X)-Y|^2\}\right)\Bigg)\cdot 1_{A_n}\Bigg)$$
$$\leq c_{30}\cdot\frac{(\log n)^2}{n}+c_{31}\cdot\frac{n^{\tau\cdot d+\epsilon}}{n}.$$

From this we conclude

$$\mathbf{E}\{T_{4,n}/2\}$$
$$\leq\left(1-\frac{2\cdot c_3}{L_n}\right)^{t_n}\cdot\mathbf{E}\{Y^2\}$$
$$+4\cdot\left(\sup_{\substack{(\bar{w}^{(l)}_{i,j,k})_{i,j,k,l}:\\ |\bar{w}^{(l)}_{i,j,k}-w^{(l)}_{i,j,k}|\leq\log n\ (l=0,\ldots,L-1)}}\int|\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}\cdot f^{(L)}_{\tilde{\mathbf{w}},k,j}(x)-m(x)|^2\mathbf{P}_X(dx)\right)$$
$$+c_3\cdot\sum_{k=1}^{\tilde{K}_n}|w^{(L)}_{1,1,k}|^2+c_{30}\cdot\frac{(\log n)^2}{n}+c_{31}\cdot\frac{n^{\tau\cdot d+\epsilon}}{n}.$$

The definition of t_n together with (16) implies

$$\left(1-\frac{2\cdot c_3}{L_n}\right)^{t_n}\cdot\mathbf{E}\{Y^2\}\leq\exp\left(-\frac{2\cdot c_3}{L_n}\cdot t_n\right)\cdot\mathbf{E}\{Y^2\}\leq\frac{c_{32}}{n}.$$

Summarizing the above results we get the assertion. □

4.3 Auxiliary results for the proof of Theorem 2

Lemma 6 *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*

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

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

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

where $f_{k,i}^{(l)}(x)$ *are recursively defined by (4) and (5).*

Assume

$$w_{1,j,j}^{(0)} = \frac{4d \cdot (\log n)^2}{\delta} \quad \textit{and} \quad w_{1,j,0}^{(0)} = -\frac{4d \cdot (\log n)^2}{\delta} \cdot u^{(j)} \quad \textit{for } j \in \{1, \dots, d\}, \tag{33}$$

$$w_{1,j+d,j}^{(0)} = -\frac{4d \cdot (\log n)^2}{\delta} \quad \textit{and} \quad w_{1,j+d,0}^{(0)} = \frac{4d \cdot (\log n)^2}{\delta} \cdot v^{(j)} \quad \textit{for } j \in \{1, \dots, d\}, \tag{34}$$

$$w_{1,s,t}^{(0)} = 0 \quad \textit{if } s \leq 2d, s \neq t, s \neq t + d \textit{ and } t > 0, \tag{35}$$

$$w_{1,1,t}^{(1)} = 8 \cdot (\log n)^2 \quad \textit{for } t \in \{1, \dots, 2d\}, \tag{36}$$

$$w_{1,1,0}^{(1)} = -8(\log n)^2 \left(2d - \frac{1}{2}\right), \tag{37}$$

$$w_{1,1,t}^{(1)} = 0 \quad \textit{for } t > 2d, \tag{38}$$

$$w_{1,1,1}^{(l)} = 6 \cdot (\log n)^2 \quad \textit{for } l \in \{2, \dots, L\}, \tag{39}$$

$$w_{1,1,0}^{(l)} = -3 \cdot (\log n)^2 \quad \textit{for } l \in \{2, \dots, L\} \tag{40}$$

and

$$w_{1,1,t}^{(l)} = 0 \quad \textit{for } t > 1 \textit{ and } l \in \{2, \dots, L\}. \tag{41}$$

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

$$|\bar{w}_{1,i,j}^{(l)} - w_{1,i,j}^{(l)}| \leq \log n \quad \textit{for all } l = 0, \dots, L-1. \tag{42}$$

Then, we have

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

and

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

Proof. The result follows from the proof of Lemma 5 in Drews and Kohler (2022). In fact, in this proof it is shown that

$$\sum_{j=1}^{r} \bar{w}_{1,1,j}^{(L-1)} \cdot \bar{f}_{1,j}^{(L-1)}(x) + \bar{w}_{1,1,0}^{(l-1)} \geq 2(\log n)^2 - \frac{6}{n}(\log n)^2$$

holds if $x \in [u^{(1)} + \delta, v^{(1)} - \delta] \times \cdots \times [u^{(d)} + \delta, v^{(d)} - \delta]$, and that

$$\sum_{j=1}^{r} \bar{w}_{1,1,j}^{(L-1)} \cdot \bar{f}_{1,j}^{(L-1)}(x) + \bar{w}_{1,1,0}^{(l-1)} \leq -2(\log n)^2 + 6 \log n \cdot \frac{1}{n}$$

holds if $x^{(i)} \notin [u^{(i)} - \delta, v^{(i)} + \delta]$ for some $i \in \{1, \dots, d\}$. Because of $n \geq 6$ and $n \geq e^s$ we have

$$2(\log n)^2 - \frac{6}{n}(\log n)^2 \geq s \cdot \log n$$

and

$$-2(\log n)^2 + 6 \log n \cdot \frac{1}{n} \leq -s \cdot \log n,$$

from which we get the assertion as in the proof of Lemma 5 in Drews and Kohler (2022). □

In our next lemma we use a multiscale approximation in order to approximate a Lipschitz continuous function by a deep neural network.

Lemma 7 *Let $1/2 \leq p \leq 1$, $C > 0$, let $f : \mathbb{R}^d \to \mathbb{R}$ be a (p, C)–smooth function and let X be an $\mathbb{R}^d$-valued random variable with $supp(X) \subseteq [0,1]^d$. Let $l \in \mathbb{N}$, $0 < \delta < 1/2$ with*

$$c_{33} \cdot \delta \leq \frac{1}{2^l} \leq c_{34} \cdot \delta \tag{43}$$

and let $L, r, s \in \mathbb{N}$ with

$$L \geq 2 \quad \textit{and} \quad r \geq 2d$$

and let

$$\tilde{K}_n \geq \left(l \cdot (2^l + 1)^{2d} + 1 \right)^3$$

Then there exist

$$w_{k,i,j}^{(l)} \in [-c_1 \cdot (\log n)^2, c_1 \cdot (\log n)^2] \quad (l = 1, \dots, L, k = 1, \dots \tilde{K}_n)$$

and

$$w_{k,i,j}^{(0)} \in \left[-\frac{8 \cdot d \cdot (\log n)^2}{\delta}, \frac{8 \cdot d \cdot (\log n)^2}{\delta} \right] \quad (k = 1, \dots, \tilde{K}_n).$$

such that for all $\bar{\mathbf{w}}$ satisfying $|\bar{w}_{i,j,k}^{(l)} - w_{i,j,k}^{(l)}| \leq \log n$ $(l = 0, \dots, L-1)$ we have for n sufficiently large

$$\begin{aligned} &\int |\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x) - f(x)|^2 \mathbf{P}_X(dx) \\ &\qquad \leq c_{35} \cdot \left(l^2 \cdot \delta + \delta^{2p} + \frac{l \cdot (2^l + 1)^{2d}}{n^s} \right), \end{aligned} \tag{44}$$

$$|\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x)| \leq c_{36} \cdot \left(1 + \frac{(2^l + 1)^{2d}}{n^s} \right) \quad (x \in [0,1]^d) \tag{45}$$

and

$$\sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 \leq \frac{c_{37}}{2^{2 \cdot d \cdot l}}. \tag{46}$$

Proof. In the proof we use a multiscale approximation of f by piecewise constant functions and use Lemma 6 in order to approximate the piecewise constant functions by a linear combination of neural networks.

In order to construct the multiscale approximation of f by piecewise constant functions we use a sequence of coverings $\mathcal{P}^{(0)} = \{[0,1]^d\}$, $\mathcal{P}^{(1)}$, $\ldots$, $\mathcal{P}^{(l)}$ of $[0,1]^d$ with the following properties:

1. $\mathcal{P}^{(k)}$ consists of $(2^k+1)^d$ many pairwise disjoint cubes of side length $1/2^k$ $(k = 1,\ldots,l)$.
2. $[0,1]^d \subseteq \cup_{A\in\mathcal{P}^{(k)}} A$
3. $$\mathbf{P}_X\left(\cup_{A\in\mathcal{P}^{(k)}} A_{border,\delta}\right) \leq 4d\cdot 2^k\cdot\delta, \tag{47}$$
   where
   $$\begin{aligned} A_{border,\delta} &= [u^{(1)}-\delta, v^{(1)}+\delta]\times\cdots\times[u^{(d)}-\delta, v^{(d)}+\delta] \\ &\quad \backslash [u^{(1)}+\delta, v^{(1)}-\delta]\times\cdots\times[u^{(d)}+\delta, v^{(d)}-\delta] \end{aligned}$$
   for
   $$A = [u^{(1)}, v^{(1)}]\times\cdots\times[u^{(d)}, v^{(d)}].$$

We can ensure (47) by shifting a partition of

$$\left[-\frac{1}{2^k}, 1\right]^d$$

consisting of $(2^k+1)^d$ many cubes of side length $1/2^k$ separately in each component by multiples of $2\cdot\delta$ less than or equal to $1/2^k$, which gives us for each component

$$\left\lfloor \frac{1}{2\cdot\delta}\cdot\frac{1}{2^k}\right\rfloor$$

disjoint sets of which at least one must have $\mathbf{P}_X$-measure less than or equal to

$$\frac{1}{\lfloor\frac{1}{2\cdot\delta}\cdot\frac{1}{2^k}\rfloor} \leq \frac{1}{\frac{1}{2\cdot\delta}\cdot\frac{1}{2^k}-1} \leq \frac{2\cdot\delta\cdot 2^k}{1-2\cdot\delta\cdot 2^k} \leq 4\cdot\delta\cdot 2^k$$

in case $2\cdot\delta\cdot 2^k \leq 1/2$, which we can assume w.l.o.g.

For $x\in[0,1]^d$ denote by $z_{\mathcal{P}^{(k)}}(x)$ the center z_A of the set $A\in\mathcal{P}^{(k)}$ which contains x.

We approximate

$$f(x)$$

by

$$f(z_{\mathcal{P}^{(l)}}(x))$$

$$= f(z_{\mathcal{P}^{(0)}}(x)) + \sum_{k=1}^{l} (f(z_{\mathcal{P}^{(k)}}(x)) - f(z_{\mathcal{P}^{(k-1)}}(x)))$$

$$= f(z_{\mathcal{P}^{(0)}}(x)) + \sum_{k=1}^{l} \sum_{A_1 \in \mathcal{P}^{(k)}, A_2 \in \mathcal{P}^{(k-1)}: A_1 \cap A_2 \neq \emptyset} (f(z_{A_1}) - f(z_{A_2})) \cdot 1_{A_1 \cap A_2}(x). \quad (48)$$

For a d-dimensional rectangle R let

$$f_{net,R,\delta}$$

be the neural network from Lemma 6 which approximates

$$1_R.$$

(In case that 2δ is less than the minimal side length of R, Lemma 6 does not imply that $f_{net,R,\delta}$ is in the inner part of R close to one).

We approximate $f(z_{\mathcal{P}^{(l)}}(x))$ by

$$f(z_{\mathcal{P}^{(0)}}(x)) \cdot f_{net,[-1,2]^d,\delta}(x)$$
$$+ \sum_{k=1}^{l} \sum_{A_1 \in \mathcal{P}^{(k)}, A_2 \in \mathcal{P}^{(k-1)}: A_1 \cap A_2 \neq \emptyset} (f(z_{A_1}) - f(z_{A_2})) \cdot f_{net,A_1 \cap A_2,\delta}(x)$$

and let $\mathbf{w}$ be the weights of the above neural network. Observe that this network consists of

$$1 + \sum_{k=1}^{l} |\{A_1 \in \mathcal{P}^{(k)}, A_2 \in \mathcal{P}^{(k-1)} \, : \, A_1 \cap A_2 \neq \emptyset\}| \leq 1 + l \cdot (2^l + 1)^{2d}$$

many fully connected neural networks which are computed in parallel. Denote the neural network where the weights of $f_{net,R,\delta}$ are replaced by the corresponding weights of $\bar{\mathbf{w}}$ by $f_{net,\bar{\mathbf{w}},R,\delta}$. We then set

$$\begin{aligned} f_{net}(x) &= \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x) \\ &= f(z_{\mathcal{P}^{(0)}}(x)) \cdot f_{net,\bar{\mathbf{w}},[-1,2]^d,\delta}(x) \\ &\quad + \sum_{k=1}^{l} \sum_{A_1 \in \mathcal{P}^{(k)}, A_2 \in \mathcal{P}^{(k-1)}: A_1 \cap A_2 \neq \emptyset} (f(z_{A_1}) - f(z_{A_2})) \cdot f_{net,\bar{\mathbf{w}},A_1 \cap A_2,\delta}(x). \end{aligned}$$

Since we have $\delta < 1/2$ we know by Lemma 6 that $f_{net,\bar{\mathbf{w}},[-1,2]^d,\delta}(x) \geq 1 - 1/n^s$ holds for all $x \in [0,1]^d$. We have

$$\int |f_{net}(x) - f(x)|^2 \mathbf{P}_X(dx)$$
$$\leq 2 \cdot \int |f(z_{\mathcal{P}^{(l)}}(x)) - f(x)|^2 \mathbf{P}_X(dx) + 2 \cdot \int |f_{net}(x) - f(z_{\mathcal{P}^{(l)}}(x))|^2 \mathbf{P}_X(dx).$$

By Lemma 6, (48) and the (p, C)–smoothness of f, which implies

$$|f(z_{\mathcal{P}^{(l)}}(x)) - f(x)| \leq c_{38} \cdot \left(\frac{1}{2^l}\right)^p \quad \text{for all } x \in [0,1]^d$$

and

$$|f(z_{A_1}) - f(z_{A_2})| \leq c_{39} \cdot \left(\frac{1}{2^k}\right)^p \quad \text{for all } A_1 \in \mathcal{P}^{(k)}, A_2 \in \mathcal{P}^{(k-1)} \text{ with } A_1 \cap A_2 \neq \emptyset, \quad (49)$$

we can bound the last sum above by

$$\begin{aligned}
&c_{40} \cdot \left(\frac{1}{2^l}\right)^{2p} + c_{41} \cdot \left(\frac{1}{n^s}\right)^2 \\
&\quad +4 \cdot l \cdot \sum_{k=1}^{l} \int \Big| \sum_{A_1 \in \mathcal{P}^{(k)}, A_2 \in \mathcal{P}^{(k-1)} : A_1 \cap A_2 \neq \emptyset} (f(z_{A_1}) - f(z_{A_2})) \\
&\qquad\qquad \cdot (1_{A_1 \cap A_2}(x) - f_{net, A_1 \cap A_2, \delta}(x)) \Big|^2 \mathbf{P}_X(dx) \\
&\leq c_{40} \cdot \left(\frac{1}{2^l}\right)^{2p} + c_{41} \cdot \left(\frac{1}{n^s}\right)^2 + 4 \cdot l \cdot \sum_{k=1}^{l} \left(c_{42} \cdot \left(\frac{1}{2^k}\right)^{2p} \cdot \mathbf{P}_X(\cup_{A \in \mathcal{P}^{(k)} \cup \mathcal{P}^{(k-1)}} A_{border,\delta}) \right. \\
&\quad \left. +c_{43} \cdot \left(\left(\frac{1}{2^k}\right)^p \cdot \frac{1}{n^s} \right)^2 \cdot (2^k + 1)^{2d} \right) \\
&\leq c_{44} \cdot \left(\left(\frac{1}{2^l}\right)^{2p} + l^2 \cdot \delta + l \cdot (2^l + 1)^{2d} \cdot \frac{1}{n^s} \right).
\end{aligned}$$

By (43) this implies (44).

Next we prove (45). By construction we know that $f_{net,R,\delta}$ is bounded in absolute value by one, which implies

$$\begin{aligned}
&\left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x) \right| \\
&\leq |f(z_{\mathcal{P}^{(0)}}(x))| + \sum_{k=1}^{l} \sum_{A_1 \in \mathcal{P}^{(k)}, A_2 \in \mathcal{P}^{(k-1)} : A_1 \cap A_2 \neq \emptyset} |f(z_{A_1}) - f(z_{A_2})| \cdot f_{net,\bar{\mathbf{w}}, A_1 \cap A_2, \delta}(x).
\end{aligned}$$

Using (49), Lemma 5 and that for each $k \in \{1, \ldots, l\}$ there are at most c_{45} many $A_1 \in \mathcal{P}^{(k)}$, $A_2 \in \mathcal{P}^{(k-1)}$ such that

$$A_1 \cap A_2 \neq \emptyset \quad \text{and} \quad x \in (A_1 \cap A_2) \cup (A_1 \cap A_2)_{border,\delta},$$

we can bound the term on the right-hand side above by

$$\|f\|_\infty + \sum_{k=1}^{l} \left(c_{45} \cdot c_{39} \cdot \left(\frac{1}{2^k}\right)^p + c_{39} \cdot \left(\frac{1}{2^k}\right)^p \cdot (2^k + 1)^{2d} \cdot \frac{1}{n^s} \right)$$

$$\leq c_{46} \cdot \left(1 + \frac{(2^l+1)^{2d}}{n^s}\right).$$

There at most $l \cdot (2^l+1)^{2d}+1$ many output weights of the above neural network are all bounded in absolute value by a constant, which implies that if we use $\tilde{K}_n = l \cdot (2^l+1)^{2d}+1$ we will get

$$\sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 \leq c_{47} \cdot \tilde{K}_n.$$

In order to get a smaller upper bound, we repeat the whole construction $(l \cdot (2^l+1)^{2d}+1)^2$ many times, each time with output weights divided by $(l \cdot (2^l+1)^{2d}+1)^2$. The above proof implies that the linear combination of these $(l \cdot (2^l+1)^{2d}+1)^3$ many neural networks still satisfies (44) and (45) (here we use that in each of the networks we can use the same exception sets where $f_{net,R,\delta}$ is not accurate). This results in

$$\sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 \leq \sum_{k=1}^{(l\cdot(2^l+1)^{2d}+1)^3} \left(\frac{c_{47}}{(l \cdot (2^l+1)^{2d}+1)^2}\right)^2 \leq \frac{c_{48}}{2^{2d \cdot l}}.$$

□

4.4 Proof of Theorem 2

Set

$$l = \lfloor \frac{1}{1+d} \cdot \log n \rfloor$$

(which implies that that (43) holds for $\delta = n^{-1/(1+d)}$) and $\tilde{K}_n = (l \cdot (2^l+1)^{2d}+1)^3$. By applying Lemma 7 with a sufficiently large s together with Theorem 1 we get for n sufficiently large

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_7 \cdot \left(\frac{n^{\frac{1}{1+d} \cdot d + \epsilon}}{n} + \sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 \right. \\
&\qquad + \sup_{\substack{(\bar{w}_{i,j,k}^{(l)})_{i,j,k,l}: \\ |\bar{w}_{i,j,k}^{(l)} - w_{i,j,k}^{(l)}| \leq \log n \, (l=0,\ldots,L-1)}} \left. \int |\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \\
&\leq c_{49} \cdot \left(\frac{n^{\frac{1}{1+d} \cdot d + \epsilon}}{n} + n^{-\frac{2d}{d+1}} + (\log n)^2 \cdot n^{-1/(1+d)} + n^{-\frac{2p}{d+1}} \right) \\
&\leq c_{50} \cdot n^{-\frac{1}{1+d}+\epsilon}.
\end{aligned}$$

□

4.5 Auxiliary results for the proof of Theorem 3

Lemma 8 *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*

$$f_{\mathbf{w}}(x) = \sum_{I \subseteq \{1,\dots,d\} : |I| = d^*} f_{\mathbf{w}_I}(x_I)$$

where $f_{\mathbf{w}_I}$ as defined by (3)–(5) with d replaced by d^* and weight vector $\mathbf{w}_I$,

$$\mathbf{w} = (\mathbf{w}_I)_{I \subseteq \{1,\dots,d\} : |I| = d^*},$$

and where for any $I \subseteq \{1, \dots, d\}$ with $|I| = d^*$ the weight vector $\mathbf{w}_I$ satisfies

$$\sum_{j=1}^{K_n} |(\mathbf{w}_I)_{1,1,j}^{(L)}| \leq C, \tag{50}$$

$$|(\mathbf{w}_I)_{k,i,j}^{(l)}| \leq B \quad (k \in \{1, \dots, K_n\}, i, j \in \{1, \dots, r\}, l \in \{1, \dots, L-1\}) \tag{51}$$

and

$$|(\mathbf{w}_I)_{k,i,j}^{(0)}| \leq A \quad (k \in \{1, \dots, K_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d\}). \tag{52}$$

Then we have for any $1 \leq p < \infty$, $0 < \epsilon < \beta$ and $x_1^n \in [-\alpha, \alpha]^d$

$$\begin{aligned} &\mathcal{N}_p\left(\epsilon, \{T_\beta f \,:\, f \in \mathcal{F}\}, x_1^n\right) \\ &\leq \left(c_{51} \cdot \frac{\beta^p}{\epsilon^p}\right)^{c_{52} \cdot \alpha^{d^*} \cdot A^{d^*} \cdot B^{(L-1)\cdot d^*} \left(\frac{C}{\epsilon}\right)^{d^*/k} + c_{53}}. \end{aligned}$$

Proof. By the proof of Lemma 4 we have for any $I \subseteq \{1, \dots, d\}$ with $|I| = d^*$, $x \in \mathbb{R}^{d^*}$ and any $s_1, \dots, s_k \in \{1, \dots, d\}$

$$\left| \frac{\partial^k f_{\mathbf{w}_I}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \right| \leq c_{54} \cdot C \cdot B^{(L-1)\cdot k} \cdot A^k =: c.$$

For $I \subseteq \{1, \dots, d\}$ with $|I| = d^*$ let $\mathcal{G} \circ \Pi_I$ be the set of all piecewise polynomials of total degree less than k with respect to a partition Π_I of $[-\alpha, \alpha]^{d^*}$ into cubes of sidelength

$$\left(c_{55} \cdot \frac{\epsilon}{c}\right)^{1/k},$$

where $c_{55} = c_{55}(d, k)$ is a suitable constant greater zero. Then a standard bound on the remainder of a multivariate Taylor polynomial shows that for each $f_{\mathbf{w}_I}$ we can find $g_{\mathbf{w}_I} \in \mathcal{G} \circ \Pi_I$ such that

$$|f_{\mathbf{w}_I}(x) - g_{\mathbf{w}_I}(x)| \leq \frac{\epsilon}{2 \cdot \binom{d}{d^*}}$$

holds for all $x \in [-\alpha,\alpha]^{d^*}$. Let $\mathcal{H}$ be the set of all functions of the form

$$h(x) = \sum_{I \subseteq \{1,\dots,d\}\,:\,|I|=d^*} g_{\mathbf{w}_I}(x_I) \quad (x \in [-\alpha,\alpha]^d)$$

($g_I \in \mathcal{G} \circ \Pi_I, I \subseteq \{1,\dots,d\}, |I| = d^*$). Then we have

$$\mathcal{N}_p\left(\epsilon, \{T_\beta f \,:\, f \in \mathcal{F}\}, x_1^n\right) \le \mathcal{N}_p\left(\frac{\epsilon}{2}, \{T_\beta h \,:\, h \in \mathcal{H}\}, x_1^n\right).$$

$\mathcal{H}$ is a linear vector space of dimension less than or equal to

$$c_{56} \cdot \binom{d}{d^*} \cdot \alpha^{d^*} \cdot \left(\frac{c}{\epsilon}\right)^{d^*/k},$$

from which we get the assertion by an application of Theorems 9.4 and 9.5 in Györfi et al. (2002). □

4.6 Proof of Theorem 3

Set

$$l = \lfloor \frac{1}{1+d^*} \log n \rfloor$$

(which implies that that (44) holds for $\delta = n^{-1/(1+d^*)}$ and $\tilde{K}_n = (l \cdot (2^l+1)^{2d^*}+1)^3$.

By applying Lemma 7 with a sufficiently large s together with Theorem 1 and Lemma 8 we get

$$\begin{aligned}
&\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \le c_7 \cdot \left(\frac{n^{\frac{1}{1+d^*}\cdot d^* + \epsilon}}{n} + \sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 \right. \\
&\qquad \left. + \sup_{\substack{(\bar{w}_{i,j,k}^{(l)})_{i,j,k,l}: \\ |\bar{w}_{i,j,k}^{(l)} - w_{i,j,k}^{(l)}| \le \log n \; (l=0,\dots,L-1)}} \int |\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,j}^{(L)}(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \\
&\le c_{57} \cdot \left(\frac{n^{\frac{1}{1+d^*}\cdot d^* + \epsilon}}{n} + n^{-\frac{2p}{d^*+1}} + (\log n)^2 \cdot n^{-1/(1+d^*)} \right) \\
&\le c_{58} \cdot n^{-\frac{1}{1+d^*}+\epsilon}.
\end{aligned}$$

□

References

[1] Allen-Zhu, Z., Li, Y., and Liang, Y. (2019). Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pages 6155-6166.

[2] Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, **39**, pp. 930-944.

[3] Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*, **14**, pp. 115-133.

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

[5] 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.

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

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

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

[9] Devroye, L., and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, **8**, pp. 231-239.

[10] Drews, S. and Kohler, M. (2022). On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent, arXiv:2208.14283.

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

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

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

[14] 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.

[15] Kawaguchi, K. (2016). Deep learning without poor local minima. *30th Conference on Neural Information Processing Systems (NIPS 2016)*. Barcelona, Spain.

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

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

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

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

[20] Kohler, M., Krzyżak, A., and Langer, S. (2022). Estimation of a function of low local dimensionality by deep neural networks, *IEEE Transaction on Information Theory*, **68**, pp. 4032-4041.

[21] 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.

[22] 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.

[23] 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

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

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

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

[27] 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.

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

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

[30] 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.

[31] 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.

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

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

[34] Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics*, **25**, pp. 118-184.

[35] 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*.

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

[37] 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.