Rate of convergence of over-parametrized deep neural network regression estimates learned by stochastic 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*

December 18, 2024

Abstract
Nonparametric regression with random design is considered. The L_2 error with integration with respect to the design measure is used as error criterion. Over-parametrized deep neural network estimates are defined with logistic activation function where all parameters are learned by stochastic gradient descent. It is shown that the estimates achieve a nearly optimal rate of convergence in case that the regression function is (p, C)–smooth. In case that the regression function satisfies a projection pursuit model or more generally a hierarchical composition model the estimate achieves a rate of convergence which does not depend on the input dimension.

AMS classification: Primary 62G08; secondary 62G20.

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

1 Introduction

As demonstrated by the recent nobel prize in chemistry one half of which was awarded to Demis Hassabis and John Jumper, who have developed with AlphaFold a deep learning model able to predict protein structures (cf., e.g., Billings et al. (2019)), deep learning had a major impact on modern science. This is thanks to its tremendous success in applications, which include, besides the above mentioned application in chemistry, also applications in image classification (cf., e.g., Krizhevsky, Sutskever and Hinton (2012)), language recognition (cf., e.g., Kim (2014)), machine translation (cf., e.g., Wu et al. (2016)), mastering of games (cf., e.g., Silver et al. (2017)) or simulation of human conversation (cf., e.g., Zong and Krishnamachari (2022)). Given this large success in

*Running title: *Over-parametrized deep neural networks learned by SGD*
†Corresponding author. Tel: +1-514-848-2424 ext. 3007, Fax:+1-514-848-2830

applications, there is also an increasing interest in theoretical understanding of deep learning. And this is where our article makes its contributions.

We study deep learning in the context of nonparametric regression. Here we are given an $\mathbb{R}^d \times \mathbb{R}$–valued random vector (X, Y) with $\mathbf{E}Y^2 < \infty$, and our goal is to predict the value of Y given the value of X. Let $m(x) = \mathbf{E}\{Y|X = x\}$ be the regression function. Then any measurable $f : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$\mathbf{E}\{|f(X) - Y|^2\} = \mathbf{E}\{|m(X) - Y|^2\} + \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \tag{1}$$

(cf., e.g., Section 1.1 in Györfi et al. (2002)), hence in view of minimizing the L_2 risk (1) of f the regression function m is the optimal predictor, and the L_2 error

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

describes how far the L_2 risk of a function f is away from its optimal value.

In applications the distribution of (X, Y) and hence also the corresponding regression function m is typically unknown. But often it is possible to observe data from the underlying distribution, and the task is to use this data to estimate the unknown regression function. In view of minimization of the L_2 risk of the estimate, here it is natural to use the L_2 error as an error criterion.

In order to introduce this problem formally, let (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n) be independent and identically distributed. In nonparametric regression the data set

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \tag{3}$$

is given, and the task is to construct an estimate

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

such that its L_2 error

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

is small. A systematic introduction to nonparametric regression, its estimates and known results can be found, e.g., in Györfi et al. (2002).

In deep learning the regression function is estimated by fitting a deep neural network to the data. Such a deep neural network depends on an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, e.g., the logistic activation function

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{4}$$

or the ReLU activation function

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

In its simplest form (multilayer feedforward neural network) it has a number $L \in \mathbb{N}$ of hidden layers (depth of the network) and a number $r \in \mathbb{N}$ of hidden neurons per layer (width of the network) and it is recursively defined by

$$f_{\mathbf{w}}(x) = \sum_{k=1}^{r} w_{1,k}^{(L)} \cdot f_{\mathbf{w},k,1}^{(L)}(x) + w_{1,0}^{(0)},$$

where

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

for $l \in \{2, \ldots, L\}$, and

$$f_{\mathbf{w},i}^{(1)}(x) = \sigma\left(\sum_{k=1}^{d} w_{i,k}^{(0)} \cdot x^{(i)} + w_{i,0}^{(0)}\right).$$

Here

$$\mathbf{w} = \left(w_{k,i}^{(l)}\right)_{k,i,l} \in \mathbb{R}^{r\cdot(d+1)+(L-2)\cdot r\cdot(r+1)+2\cdot(r+1)}$$

is the vector of the weights of the network, and one constructs neural network regression estimates by fitting these weights to the data, i.e., by using the data to select the weights such that the resulting neural network is a good approximation of the regression function.

The simplest approach is to use the principle of the least squares and to define the regression estimate by

$$m_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2,$$

where $\mathcal{F}_n$ is the set of all neural networks with depth L_n, width r_n and some given activation function. Here the so–called empirical L_2 risk is minimized over the set of these networks.

The rate of convergence of the least squares estimates based on multilayer neural networks has been analyzed in Kohler and Krzyżak (2017), Imaizumi and Fukamizu (2018), Bauer and Kohler (2019), Suzuki and Nitanda (2019), Schmidt-Hieber (2020) and Kohler and Langer (2021). One of the main results achieved in this context shows that neural networks can achieve some kind of dimension reduction under rather general assumptions. The most general form goes back to Schmidt-Hieber (2020). In order to formulate it we need the following notion of smoothness.

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 $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^{d} \alpha_j = q$ the partial derivative $\partial^q m / (\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d})$ exists and satisfies*

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

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

Stone (1982) showed that in case of a (p, C)–smooth regression function the optimal Minimax rate of convergence for the expected L_2 error is

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

This rate suffers from the so–called curse of dimensionality: If the dimension d is large compared to the smoothness p of the regression function, the exponent will be close to zero and the rate of convergence will be rather slow. Since this rate is optimal, the only way to circumvent this is to impose additional assumptions on the structure of the regression function. Such constraints resulted in, e.g., additive models (cf., e.g., Stone (1985)), interaction models (cf., e.g., Stone (1994)), single index models (cf., e.g., Härdle, Hall and Ichimura (1993), Härdle and Stoker (1989), Yu and Ruppert (2002) and Kong and Xia (2007)) or projection pursuit (cf, e.g., Friedman and Stuetzle (1981)), where corresponding low dimensional rates of convergence can be achieved (cf., e.g., Stone (1985, 1994) and Chapter 22 in Györfi et al. (2002)).

Schmidt-Hieber (2020) used an assumption of the following form to achieve a dimension reduction for the least squares neural networks.

Definition 2 *Let $d \in \mathbb{N}$ and $m : \mathbb{R}^d \to \mathbb{R}$ and let $\mathcal{P}$ be a subset of $(0, \infty) \times \mathbb{N}$.*
a) *We say that m satisfies a hierarchical composition model of level 0 with order and smoothness constraint $\mathcal{P}$, if there exists a $K \in \{1, \dots, d\}$ such that*

$$m(\mathbf{x}) = x^{(K)} \quad \text{for all } \mathbf{x} = (x^{(1)}, \dots, x^{(d)})^\top \in \mathbb{R}^d.$$

b) *We say that m satisfies a hierarchical composition model of level $l+1$ with order and smoothness constraint $\mathcal{P}$, if there exist $(p, K) \in \mathcal{P}$, $C > 0$, $g : \mathbb{R}^K \to \mathbb{R}$ and $f_1, \dots, f_K : \mathbb{R}^d \to \mathbb{R}$, such that g is (p, C)–smooth, $f_1, \dots, f_K$ satisfy a hierarchical composition model of level l with order and smoothness constraint $\mathcal{P}$ and*

$$m(\mathbf{x}) = g(f_1(\mathbf{x}), \dots, f_K(\mathbf{x})) \quad \text{for all } \mathbf{x} \in \mathbb{R}^d.$$

Note that this assumption is more general then the assumption used in additive models, interaction models, single index models or projection pursuit models.

Schmid-Hieber (2020) showed that suitable the least squares neural network regression estimates achieve (up to some logarithmic factor) a rate of convergence of order

$$\max_{(p,K)\in\mathcal{P}} n^{-\frac{2p}{2p+K}}$$

in case that the regression function satisfies a hierarchical composition model of some finite level with order and smoothness constraint $\mathcal{P}$. Since this rate of convergence does not depend on the dimension d of X, this results shows that the least squares neural network regression estimates are able to circumvent the curse of dimensionality in case that the regression function satisfies a hierarchical composition model.

The least squares neural network estimates described above cannot be used in practice, since the minimization of the empirical L_2 risk with respect to the weights of the neural

network is a nonlinear minimization problem, and for solving this minimization problem no feasible algorithm is known. In practice usually gradient descent (and its variants) are applied to solve this problem approximately. To do this, one chooses (usually randomly) a starting vector $\mathbf{w}^{(0)}$ for the weights and then makes $t_n \in \mathbb{N}$ gradient descent steps

$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \lambda_n \cdot \nabla_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}^{(t-1)}}(X_i) - Y_i|^2 \quad (t = 1, \ldots, t_n)$$

with some stepsize $\lambda > 0$. Then the estimate is defined by

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

A modification of the above gradient descent is stochastic gradient descent, where one selects (e.g., randomly) for each gradient descent step one data point (X_{i_t}, Y_{i_t}) and updates the weight vector by

$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \lambda_n \cdot \nabla_{\mathbf{w}} |f_{\mathbf{w}^{(t-1)}}(X_{i_t}) - Y_{i_t}|^2 \quad (t = 1, \ldots, t_n).$$

In this case one does not need to store the whole sample at once in the memory, which is an advantage for large data sets.

As was shown in Zou et al. (2018), Du et al. (2019), Allen-Zhu, Li and Song (2019) and Kawaguchi and Huang (2019), in case of over-parameterized deep neural networks (which have much more weights than there are data points) the application of gradient descent to over-parameterized deep neural networks leads to neural networks which minimize the empirical L_2 risk. Unfortunately, as was shown in Kohler and Krzyżak (2021), in general the corresponding estimates do not behave well on the new independent data.

In order to get a good behaviour of the estimate on the new independent data, it is necessary to study simultaneously the approximation error, the generalization error and the optimization error (cf., e.g., Kutyniok (2020)). There exist various approaches where these three components are studied simultaneously in some equivalent models of deep learning. The most prominent approach here is the neural tangent kernel setting proposed by Jacot, Gabriel and Hongler (2020). Here instead of a neural network estimate a kernel estimate is studied and its error is used to bound the error of the neural network estimate. For further results in this context see Hanin and Nica (2019) and the literature cited therein. As was pointed out in Nitanda and Suzuki (2021) in most studies in the neural tangent kernel setting the equivalence to deep neural networks holds only pointwise and not for the global L_2 error, hence from these result it is not clear how the L_2 error of the deep neural network estimate behaves. Nitanda and Suzuki (2021) were able to analyze the global error of an over-parametrized shallow neural network learned by gradient descent based on this approach. However, due to the use of the neural tangent kernel, also the smoothness assumption of the function to be estimated has to be defined with the aid of a norm involving the kernel, which does not lead to classical smoothness conditions, which makes it hard to understand the meaning of the results. Furthermore, their result did not specify how many neurons the shallow neural network must have, it was only shown that the results hold if the number of neurons is sufficiently large, and it

is not clear whether it must grow, e.g., exponentially in the sample size or not. Another approach where the estimate is studied in some asymptotically equivalent model is the mean field approach, cf., Mei, Montanari, and Nguyen (2018), Chizat and Bach (2018) or Nguyen and Pham (2020).

In a standard statistical setting all three of the above aspects have been studied simultaneously in Drews and Kohler (2023, 2024), Kohler and Krzyżak (2022, 2023) and Kohler (2024) in case of over-parametrized deep neural network regression estimates learned by gradient descent. Here the rate of convergence of the estimate in case of (p, C)–smooth regression function was analyzed, and in case of interaction models it was shown that these estimates achieve a dimension reduction. The basic idea in the proofs of these results is that for smooth activation functions the inner weights do not change much during learning if the stepsizes are sufficiently small and it was shown that at the same time the outer weights will be chosen suitably by gradient descent. It can be shown that the rates of convergence in these articles can also be achieved if only the weights of the output layer are changed during gradient descent and all other weights retain their initial values. This approach is related to the so–called random feature networks, where the inner weights are not learned at all and gradient descent is applied only to the weights in the output level, cf., e.g., Huang, Chen and Siew (2006) and Rahimi and Recht (2008a, 2008b, 2009).

In this article we use a similar approach and apply it to stochastic gradient descent. We define in Section 2 a special topology, where we compute a linear combination of many fully connected neural networks with logistic activation function in parallel and apply stochastic gradient descent together with suitable projection operators applied to the weights (cf., Section 2 for the details) in order to learn the weights. We show for suitably chosen parameters of the estimate three different results for the rate of convergence of the estimate: We show

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_1 \cdot n^{-\frac{2p}{2p+d}+\delta} \tag{5}$$

in case of a (p, C)–smooth regression function, and then

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_2 \cdot (\log n)^{3/2} \cdot n^{-\frac{p}{2p+1}+\delta} \tag{6}$$

in case of a regression function which satisfies a (p, C)–smooth projection pursuit model, and finally we show

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_3 \cdot (\log n)^{3/2} \cdot \max_{(p,K)\in\mathcal{P}} n^{-\frac{p}{2p+K}+\delta} \tag{7}$$

in case of a regression function which satisfies a hierarchical composition model. Here $\delta \in (0, 1)$ is arbitrary and the constants c_1, c_2 and c_3 depend on δ. In the first two results the number of weights of the neural networks and the number of gradient descent steps are bounded by a polynomial in the sample size, but in the third result both are required to grow exponential in the sample size.

1.1 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 a closed and convex set $A \subseteq \mathbb{R}^d$ we denote by $Proj_A x$ that element $Proj_A x \in A$ with

$$\|x - Proj_A x\| = \min_{z \in A} \|x - z\|.$$

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

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

is its supremum norm, and we set

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

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

$$\|f\|_{C^q(A)} := \max \left\{ \left\| \frac{\partial^{j_1 + \cdots + j_d}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}} \right\|_{\infty, A} : j_1 + \cdots + j_d \leq q,\, j_1, \dots, j_d \in \mathbb{N}_0 \right\}$$

for $A \subseteq \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$.

A finite collection $f_1, \dots, f_N : \mathbb{R}^d \to \mathbb{R}$ is called an L_p ε–covering of $\mathcal{F}$ on x_1^n if for all $f \in \mathcal{F}$

$$\min_{1 \leq j \leq N} \left(\frac{1}{n} \sum_{k=1}^{n} |f(x_k) - f_j(x_k)|^p \right)^{1/p} \leq \varepsilon$$

hold. The L_p ε–covering number of $\mathcal{F}$ on x_1^n is the size N of the smallest L_p ε–covering 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 then we set $(T_\beta f)(x) = T_\beta (f(x))$.

1.2 Outline

The estimates are defined in Section 2. Section 3 contains the main results. The proofs are given in Section 4.

2 Definition of the estimate

Throughout the paper we let $\sigma(x) = 1/(1 + e^{-x})$ be the logistic activation function. We define the topology of our neural networks as follows: We let $K_n, L_n, r_n \in \mathbb{N}$ be parameters of our estimate. We consider neural networks which consist of K_n fully connected neural networks of depth L_n and width r_n computed in parallel. The output of our network is then a linear combination of the outputs of these K_n neural networks.

In these networks we will denote the weights in the k-th network by $(w^{(l)}_{k,i,j})_{i,j,l}$. More precisely, $w^{(l)}_{k,i,j}$ will be the weight between neuron j in layer l and neuron i in layer $l+1$.

Formally we define this network by setting

$$f_{\mathbf{w}}(x) = \sum_{k=1}^{K_n} w^{(L_n)}_{k,1,1} \cdot f^{(L_n)}_{k,1}(x) \tag{8}$$

for some $w^{(L_n)}_{1,1,1}, \dots, w^{(L_n)}_{K_n,1,1} \in \mathbb{R}$, where $f^{(L_n)}_{k,1} = f^{(L_n)}_{\mathbf{w},k,1}$ are recursively defined by

$$f^{(l)}_{k,i}(x) = f^{(l)}_{\mathbf{w},k,i}(x) = \sigma\left(\sum_{j=1}^{r_n} w^{(l-1)}_{k,i,j} \cdot f^{(l-1)}_{k,j}(x) + w^{(l-1)}_{k,i,0}\right) \tag{9}$$

for some $w^{(l-1)}_{k,i,0}, \dots, w^{(l-1)}_{k,i,r_n} \in \mathbb{R}$ $(l = 2, \dots, L_n)$ and

$$f^{(1)}_{k,i}(x) = f^{(1)}_{\mathbf{w},k,i}(x) = \sigma\left(\sum_{j=1}^{d} w^{(0)}_{k,i,j} \cdot x^{(j)} + w^{(0)}_{k,i,0}\right) \tag{10}$$

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

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

$$(\mathbf{w}^{(0)})^{(L_n)}_{k,1,1} = 0 \quad (k = 1, \dots, K_n), \tag{11}$$

we choose $(\mathbf{w}^{(0)})^{(l)}_{k,i,j}$ uniformly distributed on $[-c_{1,n}, c_{1,n}]$ if $l \in \{1, \dots, L_n - 1\}$, and we choose $(\mathbf{w}^{(0)})^{(0)}_{k,i,j}$ uniformly distributed on $[-c_{2,n}, c_{2,n}]$, where $c_{1,n}, c_{2,n} > 0$ are parameters of the estimate. Here the random values are defined such that all components of $\mathbf{w}^{(0)}$ are independent.

Then we perform $t_n \in \mathbb{N}$ stochastic gradient descent steps starting with

$$\mathbf{w}^{(0)}.$$

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

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

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

$$\begin{aligned}\left(\mathbf{w}^{(t+1)})^{(L_n)}_{k,1,1}\right)_k \;=\; & Proj_A\Bigg(\left((\mathbf{w}^{(t)})^{(L_n)}_{k,1,1}\right)_k \\ & -\lambda_n \cdot \nabla_{(\mathbf{w}^{(L_n)}_{k,1,1})_k}\left(Y_{j_t} - f_{\mathbf{w}^{(t)}}(X_{j_t})\right)^2\Bigg),\end{aligned}$$

$$\Big((\mathbf{w}^{(t+1)})_{k,i,j}^{(l)}\Big)_{k,i,j,l:l<L_n} = Proj_B\Bigg(\Big(\mathbf{w}^{(t)}\Big)_{k,i,j}^{(l)})_{k,i,j,l:l<L_n} - \lambda_n \cdot \nabla_{((\mathbf{w}^{(t)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n}} \Big(Y_{j_t} - f_{\mathbf{w}^{(t)}}(X_{j_t})\Big)^2\Bigg)$$

for $t = 0, \ldots, t_n - 1$. Here A is the set of all weight vectors $(\mathbf{w}_{k,1,1}^{(L_n)})_k$ which satisfy

$$\sum_{k=1}^{K_n} |\mathbf{w}_{k,1,1}^{(L_n)}| \leq \gamma_n \quad \text{and} \quad \sum_{k=1}^{K_n} |\mathbf{w}_{k,1,1}^{(L_n)}|^2 \leq \alpha_n$$

and B is the set of all weight vectors $(\mathbf{w}_{k,i,j}^{(l)})_{k,i,j,l:l<L_n}$ which satisfy

$$\Big\|(\mathbf{w}_{k,i,j}^{(l)})_{k,i,j,l:l<L_n} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n}\Big\| \leq c_{3,n}.$$

Again γ_n, α_n and $c_{3,n}$ are parameters of the estimate.

Our estimate is then defined by

$$m_n(x) = \frac{1}{t_n} \sum_{t=0}^{t_n-1} T_{\beta_n} f_{\mathbf{w}^{(t)}}(x) \quad (x \in \mathbb{R}^d), \tag{12}$$

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

3 Main results

In our first result we analyze the rate of convergence of our estimate in case of a (p, C)–smooth regression function.

Theorem 1 *Let $p, C > 0$. Choose $L, r \in \mathbb{N}$ with*

$$L \geq \log_2(p + d) \quad \textit{and} \quad r \geq 2 \cdot (2p + d) \cdot (p + d), \tag{13}$$

let $K_n \in \mathbb{N}$ be such that

$$\frac{K_n}{n^{10 \cdot r^2 \cdot L}} \to \infty \quad (n \to \infty), \tag{14}$$

and set

$$L_n = L, \quad r_n = r, \quad \gamma_n = c_5 \cdot n^2, \quad \alpha_n = \frac{c_6}{n^6}, \quad t_n = \lceil c_7 \cdot n^5 \cdot K_n \rceil, \quad \lambda_n = \frac{1}{t_n},$$

$$c_{1,n} = n^{1/(2p+d)} \cdot (\log n)^2, \quad c_{2,n} = c_8, \quad c_{3,n} = c_9 \cdot \log n$$

and define the estimate as in Section 2.

Assume that the distribution of (X, Y) satisfies supp$(\mathbf{P}_X)$ bounded,

$$\mathbf{E}\Big\{e^{c_{10} \cdot Y^2}\Big\} < \infty \tag{15}$$

for some $c_{10} > 0$ and $m(\cdot) = \mathbf{E}(Y|X = \cdot)$ (p, C)–smooth, and assume that c_8 is sufficiently large. Then we have for any $\delta > 0$ that for n sufficiently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_1 \cdot n^{-\frac{2p}{2p+d}+\delta}$$

holds for some constant $c_1 > 0$ which depends on δ.

In our second contribution we consider regression functions which satisfy a projection pursuit model.

Theorem 2 *Let $\delta \in (0, 1/4)$. Let $K_n \in \mathbb{N}$ be such that*

$$\frac{K_n}{n^{(22p+50)\cdot((p+1)^2+d+3)^2+2}} \to \infty \quad (n \to \infty) \tag{16}$$

and set

$$L_n = L = 3, \quad r_n = r = \max\{(\lceil p \rceil + 1)^2, 4\}, \gamma_n = c_{11} \cdot n^{\frac{1}{2\cdot(2p+1)}+\delta}, \alpha_n = c_{12} \cdot \frac{1}{n^{4p+30}},$$

$$t_n = \lceil n^5 \cdot K_n \rceil, \lambda_n = \frac{1}{t_n}, c_{1,n} = c_{2,n} = c_{13} \cdot n^{p+5}, c_{3,n} = \log n$$

and define the estimate as in Section 2.

Assume that the distribution of (X, Y) satisfies supp$(\mathbf{P}_X)$ bounded, assumption (15) and

$$m(x) = \sum_{k=1}^{K} m_k(b_k^t x) \quad (x \in \mathbb{R}^d)$$

for some $K \in \mathbb{N}$, $b_k \in \mathbb{R}^d$ and some (p, C)–smooth functions $m_k : \mathbb{R} \to \mathbb{R}$ $(k = 1, \ldots, K)$, which satisfy

$$\max_{s \in \mathbb{N}_0 : s \leq p} \|m_k^{(s)}\|_\infty \leq c_{14} \quad (k = 1, \ldots, K).$$

Then we have that for n sufficiently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_2 \cdot (\log n)^{3/2} \cdot n^{-\frac{p}{2p+1}+\delta}$$

holds for some constant $c_2 > 0$ which depends on δ.

Next we formulate a result concerning the estimation of a regression function which satisfies a hierarchical composition model with smoothness and order constraint $\mathcal{P} \subseteq [1, \infty) \times \mathbb{N}$. Let $\mathcal{H}(\ell, \mathcal{P})$ be the set of all functions which satisfy a hierarchical composition model of level l with order and smoothness constraint $\mathcal{P}$. In order to compute a function $h_1^{(\ell)} \in \mathcal{H}(\ell, \mathcal{P})$, one has to compute different hierarchical composition models of some level i $(i \in \{1, \ldots, \ell - 1\})$. Let $\tilde{N}_i$ denote the number of hierarchical composition models of level i, needed to compute $h_1^{(\ell)}$. We denote in the following by

$$h_j^{(i)} : \mathbb{R}^d \to \mathbb{R} \tag{17}$$

the j–th hierarchical composition model of some level i ($j \in \{1,\dots,\tilde{N}_i\}, i \in \{1,\dots,\ell\}$), that applies a $(p_j^{(i)}, C)$–smooth function $g_j^{(i)} : \mathbb{R}^{K_j^{(i)}} \to \mathbb{R}$ with $p_j^{(i)} = q_j^{(i)} + s_j^{(i)}$, $q_j^{(i)} \in \mathbb{N}_0$ and $s_j^{(i)} \in (0,1]$, where $(p_j^{(i)}, K_j^{(i)}) \in \mathcal{P}$. The computation of $h_1^{(\ell)}(x)$ can then be recursively described as follows:

$$h_j^{(i)}(x) = g_j^{(i)}\left(h^{(i-1)}_{\sum_{t=1}^{j-1} K_t^{(i)}+1}(x), \dots, h^{(i-1)}_{\sum_{t=1}^{j} K_t^{(i)}}(x)\right) \tag{18}$$

for $j \in \{1,\dots,\tilde{N}_i\}$ and $i \in \{2,\dots,\ell\}$ and

$$h_j^{(1)}(x) = g_j^{(1)}\left(x^{\left(\pi(\sum_{t=1}^{j-1} K_t^{(1)}+1)\right)}, \dots, x^{\left(\pi(\sum_{t=1}^{j} K_t^{(1)})\right)}\right) \tag{19}$$

for some function $\pi : \{1,\dots,\tilde{N}_1\} \to \{1,\dots,d\}$.

Theorem 3 *Assume that the distribution of (X,Y) satisfies supp$(\mathbf{P}_X)$ bounded, assumption (15) and that the regression function $m(\cdot) = \mathbf{E}(Y|X=\cdot)$ satisfies some hierarchical composition model with order and smoothness constraint $\mathcal{P}$ described as above, where $|\mathcal{P}| < \infty$. Assume that the functions $g_j^{(i)}$ are Lipschitz continuous with Lipschitz constant $C_{Lip} \geq 1$ (i.e., $p_j^{(i)} \geq 1$ holds for all i, j) and satisfy*

$$\|g^{(I)}\|_{C^{q_j^{(i)}}(\mathbb{R}^d)} \leq c_{15}$$

for some $c_{15} > 0$. Denote by $K_{max} = \max_{i,j} K_j^{(i)}$ the maximal input dimension and by $p_{max} = \max_{i,j} p_j^{(i)}$ the maximal smoothness of the functions $g_j^{(i)}$.

Let $K_n \in \mathbb{N}$ be such that

$$\frac{K_n}{e^{(\log n)^2 \cdot n}} \to \infty \quad (n \to \infty). \tag{20}$$

Let $\delta \in (0,1)$ be arbitrary and set

$$L_n = L = l \cdot (8 + \lceil \log_2(\max K_{max}, p_{max}+1) \rceil) + 1, \quad r_n = \left\lceil c_{16} \cdot \max_{i,j} n^{\frac{K_j^{(i)}}{2\cdot(2p_j^{(i)}+K_j^{(i)})}} \right\rceil,$$

$$\gamma_n = c_{17} \cdot n^{\delta}, \quad \alpha_n = \frac{c_{18}}{n^{2L\cdot(5p_{max}+K_{max}+6)}}, \quad t_n = \lceil c_{19} \cdot n^3 \cdot K_n \rceil, \quad \lambda_n = \frac{1}{t_n},$$

$$c_{1,n} = c_{2,n} = c_{20} \cdot n^{5p_{max}+K_{max}+5}, \quad c_{3,n} = c_{21} \cdot \log n,$$

and define the estimate as in Section 2.

Then we have that for n sufficiently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_3 \cdot (\log n)^{3/2} \cdot \max_{(p,K)\in\mathcal{P}} n^{-\frac{p}{2p+K}+\delta}$$

holds for some constant $c_3 > 0$ which depends on δ.

Remark 1. The network in Theorem 1 has only polynomially many weights (in the sample size), it is trained by stochastic gradient descent, and it achieves a nearly optimal rate of convergence in case that the regression function is (p, C)–smooth. For gradient descent a similar result was shown in Kohler (2024).

Remark 2. The rate of convergence in Theorem 2 is not optimal, the optimal rate of convergence for a (p, C)–smooth projection pursuit regression model should be close to $n^{-2p/(2p+1)}$ instead of $n^{-p/(2p+1)}$ as in Theorem 2. However, Theorem 2 is the first result which shows that with a polynomial size network (stochastic) gradient descent can achieve a dimension reduction for projection pursuit.

Remark 3. As in Theorem 2 the rate of convergence in Theorem 3 is not optimal, however, Theorem 3 is the first result showing that (stochastic) gradient descent can achieve a dimension reduction in case of a hierarchical composition model. In contrast to Theorem 2 it is required in more general setting of Theorem 3 that the network be of exponential size (in the sample size).

4 Proofs

4.1 A general result

In the proofs of our main results we will apply the following general result. This theorem is an adaption to regression of Theorem 1 in Kohler, Krzyżak and Sänger (2024), which deals with pattern recognition.

Theorem 4 *Assume that (X, Y) satisfies $\mathbf{E}\{Y|X = x\}$ is bounded and $\mathbf{E}\{e^{c_{10} \cdot Y^2}\} < \infty$. Let $\mathbf{\Theta}$ be the set of all weight vectors $\mathbf{w} = (w_{i,j,k})^{(l)})_{i,j,k,l}$ which satisfy*

$$|w_{i,j,k}^{(L_n)}| \leq \gamma_n, \quad |w_{i,j,k}^{(l)}| \leq c_{2,n} + c_{3,n}, \quad |w_{i,j,k}^{(0)}| \leq c_{1,n} + c_{3,n}$$

$(l \in \{1, \ldots, L_n - 1\})$ for some γ_n, $c_{1,n}$, $c_{2,n}$, $c_{3,n} \geq 0$, Let $C_n, D_n \geq 0$. Assume

$$\sum_{k=1}^{K_n} |f_{\mathbf{w},k,1}^{(L_n)}(x) - f_{\bar{\mathbf{w}},k,1}^{(L_n)}(x)|^2 \leq C_n^2 \cdot \left\| (\mathbf{w}_{k,i,j}^{(l)})_{k,i,j,l:l<L_n} - (\bar{\mathbf{w}}_{k,i,j}^{(l)})_{k,i,j,l:l<L_n} \right\|^2 \tag{21}$$

for all $\mathbf{w}, \bar{\mathbf{w}} \in \mathbf{\Theta}$ and all $x \in supp(X)$. Define the estimate as in Section 2 with

$$\lambda_n = \frac{1}{t_n}$$

and assume that on the event $\{\max_{i=1,\ldots,n} |Y_i| \leq \beta_n\}$

$$\left\| \nabla_{(\mathbf{w}_{k,1,1}^{(L_n)})_k} (Y_{j_t} - f_{\mathbf{w}^{(t)}}(X_{j_t}))^2 \right\|_\infty \leq D_n \tag{22}$$

holds a.s. for all $t = 0, \ldots, t_n - 1$. Let E_n be an event which depends only on $\mathbf{w}^{(0)}$, and let $(\mathbf{w}^*)_{k,1,1}^{(L_n)} \in \mathbb{R}$ $(k = 1, \ldots, K_n)$ be such that

$$\sum_{k=1}^{K_n} |(\mathbf{w}^*)_{k,1,1}^{(L_n)}| \leq \gamma_n \quad \textit{and} \quad \sum_{k=1}^{K_n} |(\mathbf{w}^*)_{k,1,1}^{(L_n)}|^2 \leq \alpha_n.$$

Then

$$\begin{aligned}
&\mathbf{E}\int |m_n(x)-m(x)|^2\mathbf{P}_X(dx)\\
&\leq c_{22}\cdot\left(\frac{\beta_n^2}{n^5}+\beta_n^2\cdot\sqrt{\mathbf{P}(E_n^c)}+\mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\mathbf{E}\{|T_{\beta_n}f_{\mathbf{w}}(X)-Y|^2-|m(X)-Y|^2\}\right.\\
&\qquad\qquad\left.-\frac{2}{n}\sum_{i=1}^{n}\left(|T_{\beta_n}f_{\mathbf{w}}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\Big)\Big\}\right.\\
&+\mathbf{E}\left\{\int |f_{(((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k,((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x)-m(x)|^2\mathbf{P}_X(dx)\cdot 1_{E_n}\right\}\\
&\left.+\beta_n\cdot n\cdot\lambda_n\cdot K_n\cdot D_n+(\beta_n+\gamma_n)\cdot\sqrt{\alpha_n}\cdot C_n\cdot c_{3,n}+\sum_{k=1}^{K_n}|(\mathbf{w}^*)_{k,1,1}^{(L_n)}|^2+\frac{K_n\cdot D_n^2}{t_n}\right).
\end{aligned}$$

In the proof of Theorem 4 we will apply the following lemma, which will help us to bound the optimization error for stochastic gradient descent.

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

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

and

$$\|(\nabla_u F_t)(u,v)\|\leq D_n \tag{23}$$

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

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

where

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

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

$$\frac{1}{t_n}\sum_{t=0}^{t_n-1}F_t(u_t,v_t)\quad\leq\quad\frac{1}{t_n}\sum_{t=1}^{t_n-1}F_t(u^*,v_0)+\frac{1}{t_n}\sum_{t=1}^{t_n-1}|F_t(u^*,v_t)-F_t(u^*,v_0)|+\frac{\|u^*-u_0\|^2}{2}+\frac{D_n^2}{2\cdot t_n}.$$

Proof. By convexity of $u\mapsto F_t(u,v_t)$ and because of $u^*\in A$ we have

$$\begin{aligned}
&F_t(u_t,v_t)-F_t(u^*,v_t)\\
&\leq<(\nabla_u F_t)(u_t,v_t),u_t-u^*>\\
&=\frac{1}{2\cdot\lambda}\cdot 2\cdot<\lambda\cdot(\nabla_u F_t)(u_t,v_t),u_t-u^*>\\
&=\frac{1}{2\cdot\lambda}\cdot\left(-\|u_t-u^*-\lambda\cdot(\nabla_u F_t)(u_t,v_t)\|^2+\|u_t-u^*\|^2+\|\lambda\cdot(\nabla_u F_t)(u_t,v_t)\|^2\right)
\end{aligned}$$

$$\leq \frac{1}{2\cdot\lambda}\cdot\left(-\|Proj_A(u_t-\lambda\cdot(\nabla_u F_t)(u_t,v_t))-u^*\|^2+\|u_t-u^*\|^2+\lambda^2\cdot\|(\nabla_u F_t)(u_t,v_t)\|^2\right)$$

$$= \frac{1}{2\cdot\lambda}\cdot\left(\|u_t-u^*\|^2-\|u_{t+1}-u^*\|^2+\lambda^2\cdot\|(\nabla_u F_t)(u_t,v_t)\|^2\right).$$

This implies

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

Using the above result and (23) we get

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

□

Proof of Theorem 4. In the *first step of the proof* we upper bound the expected L_2 error of the estimate by a sum of several terms.

Let $\bar{E}_n$ be the event that E_n and $\{\max_{i=1,\ldots,n}|Y_i|\leq\beta_n\}$ hold. W.l.o.g. we assume $\|m\|_\infty\leq\beta_n$. We have

$$\begin{aligned}
&\mathbf{E}\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\\
&\leq\mathbf{E}\left\{\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\cdot 1_{\bar{E}_n}\right\}+4\cdot\beta_n^2\cdot\mathbf{P}\{\bar{E}_n^c\}\\
&\leq\mathbf{E}\left\{\left(\mathbf{E}\{|m_n(X)-Y|^2|\mathcal{D}_n,\mathbf{w}^{(0)}\}-\mathbf{E}\{|m(X)-Y|^2\}\right)\cdot 1_{\bar{E}_n}\right\}+4\cdot\beta_n^2\cdot\mathbf{P}\{\bar{E}_n^c\}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}\left\{\left(\mathbf{E}\{|\frac{1}{t_n}\sum_{t=0}^{t_n-1} T_{\beta_n} f_{\mathbf{w}^{(t)}}(X) - Y|^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\} - \mathbf{E}\{|m(X)-Y|^2\}\right)\cdot 1_{\bar{E}_n}\right\} \\
&\quad +4\cdot\beta_n^2\cdot\mathbf{P}\{\bar{E}_n^c\} \\
&\leq \mathbf{E}\left\{\frac{1}{t_n}\sum_{t=0}^{t_n-1}\left(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}^{(t)}}(X) - Y|^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\} - \mathbf{E}\{|m(X)-Y|^2\}\right)\cdot 1_{\bar{E}_n}\right\} \\
&\quad +4\cdot\beta_n^2\cdot\mathbf{P}\{\bar{E}_n^c\} \\
&= \mathbf{E}\Bigg\{\frac{1}{t_n}\sum_{t=0}^{t_n-1}\Bigg(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}^{(t)}}(X) - Y|^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\} - \mathbf{E}\{|m(X)-Y|^2\} \\
&\qquad\qquad -2\cdot\left(|T_{\beta_n} f_{\mathbf{w}^{(t)}}(X_{j_t}) - Y_{j_t}|^2 - |m(X_{j_t})-Y_{j_t}|^2\right)\Bigg)\cdot 1_{\bar{E}_n}\Bigg\} \\
&\quad +2\cdot\mathbf{E}\Bigg\{\frac{1}{t_n}\sum_{t=0}^{t_n-1}\left(|T_{\beta_n} f_{\mathbf{w}^{(t)}}(X_{j_t}) - Y_{j_t}|^2 - |m(X_{j_t})-Y_{j_t}|^2\right)\cdot 1_{\bar{E}_n}\Bigg\} \\
&\quad +4\cdot\beta_n^2\cdot\mathbf{P}\{\bar{E}_n^c\} \\
&= \frac{1}{t_n/n}\sum_{s=1}^{t_n/n}\mathbf{E}\Bigg\{\frac{1}{n}\sum_{t=(s-1)\cdot n}^{s\cdot n-1}\Bigg(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}^{(t)}}(X) - Y|^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\} - \mathbf{E}\{|m(X)-Y|^2\} \\
&\qquad\qquad -2\cdot\left(|T_{\beta_n} f_{\mathbf{w}^{(t)}}(X_{j_t}) - Y_{j_t}|^2 - |m(X_{j_t})-Y_{j_t}|^2\right)\Bigg)\cdot 1_{\bar{E}_n}\Bigg\} \\
&\quad +2\cdot\mathbf{E}\Bigg\{\frac{1}{t_n}\sum_{t=0}^{t_n-1}\left(|T_{\beta_n} f_{\mathbf{w}^{(t)}}(X_{j_t}) - Y_{j_t}|^2 - |m(X_{j_t})-Y_{j_t}|^2\right)\cdot 1_{\bar{E}_n}\Bigg\} \\
&\quad +4\cdot\beta_n^2\cdot\mathbf{P}\{\bar{E}_n^c\} \\
&= \frac{1}{t_n/n}\sum_{s=1}^{t_n/n}\mathbf{E}\Bigg\{\frac{1}{n}\sum_{t=(s-1)\cdot n}^{s\cdot n-1}\Bigg(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}^{(s\cdot n)}}(X) - Y|^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\} - \mathbf{E}\{|m(X)-Y|^2\} \\
&\qquad\qquad -2\cdot\left(|T_{\beta_n} f_{\mathbf{w}^{(s\cdot n)}}(X_{j_t}) - Y_{j_t}|^2 - |m(X_{j_t})-Y_{j_t}|^2\right)\Bigg)\cdot 1_{\bar{E}_n}\Bigg\} \\
&\quad +\frac{1}{t_n/n}\sum_{s=1}^{t_n/n}\mathbf{E}\Bigg\{\frac{1}{n}\sum_{t=(s-1)\cdot n}^{s\cdot n-1}\Bigg(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}^{(t)}}(X) - Y|^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\} - \\
&\qquad\qquad \mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}^{(s\cdot n)}}(X) - Y|^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\}\Bigg)\cdot 1_{\bar{E}_n}\Bigg\} \\
&\quad +2\cdot\frac{1}{t_n/n}\sum_{s=1}^{t_n/n}\mathbf{E}\Bigg\{\frac{1}{n}\sum_{t=(s-1)\cdot n}^{s\cdot n-1}\Bigg(|T_{\beta_n} f_{\mathbf{w}^{(s\cdot n)}}(X_{j_t}) - Y_{j_t}|^2
\end{aligned}
$$

$$-|T_{\beta_n} f_{\mathbf{w}^{(t)}}(X_{j_t}) - Y_{j_t}|^2 \Big) \cdot 1_{\bar{E}_n} \Big\}$$

$$+2 \cdot \mathbf{E}\Big\{ \frac{1}{t_n} \sum_{t=0}^{t_n-1} \big(|T_{\beta_n} f_{\mathbf{w}^{(t)}}(X_{j_t}) - Y_{j_t}|^2 - |m(X_{j_t}) - Y_{j_t}|^2 \big) \cdot 1_{\bar{E}_n} \Big\}$$

$$+4 \cdot \beta_n^2 \cdot \mathbf{P}\{\bar{E}_n^c\}$$

$$=: T_{1,n} + T_{2,n} + T_{3,n} + T_{4,n} + T_{5,n}.$$

In the remainder of the proof we bound $T_{j,n}$ for $j \in \{1, \ldots, 5\}$.

In the *second step of the proof* we show

$$T_{5,n} \le c_{23} \cdot \beta_n^2 \cdot \Big(\mathbf{P}(E_n^c) + \frac{1}{n^{10}} \Big).$$

This follows from

$$\mathbf{P}\{\bar{E}_n^c\} \le \mathbf{P}\{E_n^c\} + \mathbf{P}\{\max_{i=1,\ldots,n} |Y_i| > \beta_n\}$$

and

$$\mathbf{P}\{\max_{i=1,\ldots,n} |Y_i| > \beta_n\} \le n \cdot \mathbf{P}\{\exp(c_{10} \cdot |Y|^2) > \exp(c_{10} \cdot \beta_n^2)\} \le n \cdot \frac{\mathbf{E}\{\exp(c_{10} \cdot |Y|^2)\}}{\exp(c_{10} \cdot \beta_n^2)} \le \frac{c_{24}}{n^{10}}.$$

In the *third step of the proof* we show

$$T_{1,n} \le c_{25} \cdot \Bigg(\mathbf{E}\Big\{ \sup_{\mathbf{w} \in \mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X) - Y|^2 - |m(X) - Y|^2\} - \frac{2}{n} \sum_{i=1}^{n} \big(|T_{\beta_n} f_{\mathbf{w}}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \big) \Big) \Big\} + \beta_n^2 \cdot \sqrt{\mathbf{P}\{E_n^c\}} + \frac{1}{n^5} \Bigg).$$

By the definition of the estimate we know

$$\mathbf{w}^{(s \cdot n)} \in \mathbf{\Theta}$$

and

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

for all $s \in \{1, \ldots, t_n/n\}$. Hence

$$T_{1,n}$$

$$= \frac{1}{t_n/n} \sum_{s=1}^{t_n/n} \mathbf{E}\Bigg\{ \frac{1}{n} \sum_{t=(s-1) \cdot n}^{s \cdot n - 1} \Bigg(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}^{(s \cdot n)}}(X) - Y|^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\} - \mathbf{E}\{|m(X) - Y|^2\}$$

$$-2 \cdot \big(|T_{\beta_n} f_{\mathbf{w}^{(s \cdot n)}}(X_{j_t}) - Y_{j_t}|^2 - |m(X_{j_t}) - Y_{j_t}|^2 \big) \Bigg) \Bigg\}$$

$$
\begin{aligned}
&-\frac{1}{t_n/n}\sum_{s=1}^{t_n/n}\mathbf{E}\Bigg\{\frac{1}{n}\sum_{t=(s-1)\cdot n}^{s\cdot n-1}\Bigg(\mathbf{E}\{|T_{\beta_n}f_{\mathbf{w}^{(s\cdot n)}}(X)-Y|^2|\mathcal{D}_n,\mathbf{w}^{(0)}\}-\mathbf{E}\{|m(X)-Y|^2\}\\
&\qquad\qquad -2\cdot\left(|T_{\beta_n}f_{\mathbf{w}^{(s\cdot n)}}(X_{j_t})-Y_{j_t}|^2-|m(X_{j_t})-Y_{j_t}|^2\right)\Bigg)\cdot 1_{\bar{E}_n}\Bigg\}\\
&\le \mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\mathbf{E}\{|T_{\beta_n}f_{\mathbf{w}}(X)-Y|^2-|m(X)-Y|^2\}\\
&\qquad -\frac{2}{n}\sum_{i=1}^{n}\left(|T_{\beta_n}f_{\mathbf{w}}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\Big)\Big\}\\
&\quad +\frac{1}{t_n/n}\sum_{s=1}^{t_n/n}\mathbf{E}\Bigg\{\frac{1}{n}\sum_{t=(s-1)\cdot n}^{s\cdot n-1}\left(\mathbf{E}\{|m(X)-Y|^2\}+2\cdot|T_{\beta_n}f_{\mathbf{w}^{(s\cdot n)}}(X_{j_t})-Y_{j_t}|^2\right)\cdot 1_{\bar{E}_n^c}\Bigg\}\\
&\le \mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\mathbf{E}\{|T_{\beta_n}f_{\mathbf{w}}(X)-Y|^2-|m(X)-Y|^2\}\\
&\qquad -\frac{2}{n}\sum_{i=1}^{n}\left(|T_{\beta_n}f_{\mathbf{w}}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\Big)\Big\}\\
&\quad +\mathbf{E}\{|m(X)-Y|^2\}\cdot\mathbf{P}\{\bar{E}_n^c\}+2\cdot\sqrt{\max_{\substack{i=1,\dots,n,\\ s=1,\dots,t_n/n}}\mathbf{E}\{|T_{\beta_n}f_{\mathbf{w}^{(s\cdot n)}}(X_i)-Y_i|^4\}}\cdot\sqrt{\mathbf{P}\{\bar{E}_n^c\}}\\
&\le \mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\mathbf{E}\{|T_{\beta_n}f_{\mathbf{w}}(X)-Y|^2-|m(X)-Y|^2\}\\
&\qquad -\frac{2}{n}\sum_{i=1}^{n}\left(|T_{\beta_n}f_{\mathbf{w}}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\Big)\Big\}\\
&\quad +c_{26}\cdot\mathbf{P}\{\bar{E}_n^c\}+c_{27}\cdot\beta_n^2\cdot\sqrt{\mathbf{P}\{\bar{E}_n^c\}}\\
&\le c_{25}\cdot\Bigg(\mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\mathbf{E}\{|T_{\beta_n}f_{\mathbf{w}}(X)-Y|^2-|m(X)-Y|^2\}\\
&\qquad -\frac{2}{n}\sum_{i=1}^{n}\left(|T_{\beta_n}f_{\mathbf{w}}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\Big)\Big\}+\beta_n^2\cdot\sqrt{\mathbf{P}\{E_n^c\}}+\frac{1}{n^5}\Bigg),
\end{aligned}
$$

where the last inequality follows from the proof of the assertion of the second step.

In the *fourth step of the proof* we show

$$
T_{2,n}\le c_{28}\cdot\beta_n\cdot\left(n\cdot\lambda_n\cdot K_n\cdot D_n+\sqrt{\alpha_n}\cdot C_n\cdot c_{3,n}\right).
$$

We have

$$
\begin{aligned}
&T_{2,n}\\
&=\frac{1}{t_n/n}\sum_{s=1}^{t_n/n}\mathbf{E}\Bigg\{\frac{1}{n}\sum_{t=(s-1)\cdot n}^{s\cdot n-1}\mathbf{E}\Big\{(T_{\beta_n}f_{\mathbf{w}^{(t)}}(X)+T_{\beta_n}f_{\mathbf{w}^{(s\cdot n)}}(X)-2Y)
\end{aligned}
$$

$$\cdot (T_{\beta_n} f_{\mathbf{w}^{(t)}}(X) - T_{\beta_n} f_{\mathbf{w}^{(s\cdot n)}}(X)) \Big| \mathcal{D}_n, \mathbf{w}^{(0)} \Big\} \Bigg) \cdot 1_{\bar{E}_n} \Bigg\}$$

$$\leq \frac{1}{t_n/n} \sum_{s=1}^{t_n/n} \mathbf{E} \Bigg\{ \frac{1}{n} \sum_{t=(s-1)\cdot n}^{s\cdot n-1} \sqrt{\mathbf{E}\{(T_{\beta_n} f_{\mathbf{w}^{(t)}}(X) + T_{\beta_n} f_{\mathbf{w}^{(s\cdot n)}}(X) - 2Y)^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\}}$$

$$\cdot \sqrt{\mathbf{E}\{(T_{\beta_n} f_{\mathbf{w}^{(t)}}(X) - T_{\beta_n} f_{\mathbf{w}^{(s\cdot n)}}(X))^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\} \cdot 1_{\bar{E}_n}} \Bigg\}$$

$$\leq c_{29} \cdot \beta_n \cdot \frac{1}{t_n/n} \sum_{s=1}^{t_n/n} \frac{1}{n} \sum_{t=(s-1)\cdot n}^{s\cdot n-1} \sqrt{\mathbf{E}\{(T_{\beta_n} f_{\mathbf{w}^{(t)}}(X) - T_{\beta_n} f_{\mathbf{w}^{(s\cdot n)}}(X))^2\} \cdot 1_{\bar{E}_n}}.$$

Using that on $\bar{E}_n$ we have

$$\begin{aligned}
&(T_{\beta_n} f_{\mathbf{w}^{(t)}}(X) - T_{\beta_n} f_{\mathbf{w}^{(s\cdot n)}}(X))^2 \\
&\leq (f_{\mathbf{w}^{(t)}}(X) - f_{\mathbf{w}^{(s\cdot n)}}(X))^2 \\
&= \left(\sum_{k=1}^{K_n} (\mathbf{w}^{(t)})_{k,1,1}^{(L_n)} \cdot f_{\mathbf{w}^{(t)},k,1}^{(L_n)}(X) - \sum_{k=1}^{K_n} (\mathbf{w}^{(s\cdot n)})_{k,1,1}^{(L_n)} \cdot f_{\mathbf{w}^{(s\cdot n)},k,1}^{(L_n)}(X) \right)^2 \\
&\leq 2 \cdot \left(\sum_{k=1}^{K_n} (\mathbf{w}^{(t)})_{k,1,1}^{(L_n)} \cdot f_{\mathbf{w}^{(t)},k,1}^{(L_n)}(X) - \sum_{k=1}^{K_n} (\mathbf{w}^{(s\cdot n)})_{k,1,1}^{(L_n)} \cdot f_{\mathbf{w}^{(t)},k,1}^{(L_n)}(X) \right)^2 \\
&\quad + 2 \cdot \left(\sum_{k=1}^{K_n} (\mathbf{w}^{(s\cdot n)})_{k,1,1}^{(L_n)} \cdot f_{\mathbf{w}^{(t)},k,1}^{(L_n)}(X) - \sum_{k=1}^{K_n} (\mathbf{w}^{(s\cdot n)})_{k,1,1}^{(L_n)} \cdot f_{\mathbf{w}^{(s\cdot n)},k,1}^{(L_n)}(X) \right)^2 \\
&\leq 2 \cdot \left(\sum_{k=1}^{K_n} |(\mathbf{w}^{(t)})_{k,1,1}^{(L_n)} - (\mathbf{w}^{(s\cdot n)})_{k,1,1}^{(L_n)}| \right)^2 \\
&\quad + 2 \cdot \sum_{k=1}^{K_n} |(\mathbf{w}^{(s\cdot n)})_{k,1,1}^{(L_n)}|^2 \cdot \sum_{k=1}^{K_n} \left| f_{\mathbf{w}^{(t)},k,1}^{(L_n)}(X) - f_{\mathbf{w}^{(s\cdot n)},k,1}^{(L_n)}(X) \right|^2 \\
&\leq 2 \cdot (|t - s\cdot n| \cdot \lambda_n \cdot K_n \cdot D_n)^2 \\
&\quad + 2 \cdot \alpha_n \cdot C_n^2 \cdot \|((\mathbf{w}^{(t)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n} - ((\mathbf{w}^{(s\cdot n)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n}\|^2 \\
&\leq 2 \cdot (|t - s\cdot n| \cdot \lambda_n \cdot K_n \cdot D_n)^2 + 2 \cdot \alpha_n \cdot C_n^2 \cdot 4 \cdot c_{3,n}^2
\end{aligned}$$

we get the assertion.

In the *fifth step of the proof* we show

$$T_{3,n} \leq c_{30} \cdot \beta_n \cdot (n \cdot \lambda_n \cdot K_n \cdot D_n + \sqrt{\alpha_n} \cdot C_n \cdot c_{3,n}).$$

Arguing as in the fourth step of the proof we get

$T_{3,n}$

$$\leq 2 \cdot \frac{1}{t_n/n} \sum_{s=1}^{t_n/n} \mathbf{E}\left\{ \frac{1}{n} \sum_{t=(s-1)\cdot n+1}^{s\cdot n} c_{31} \cdot \beta_n \cdot |T_{\beta_n} f_{\mathbf{w}^{(s\cdot n)}}(X_{j_t}) - T_{\beta_n} f_{\mathbf{w}^{(t)}}(X_{j_t})| \cdot 1_{\bar{E}_n} \right\}$$
$$\leq c_{30} \cdot \beta_n \cdot (n \cdot \lambda_n \cdot K_n \cdot D_n + \sqrt{\alpha_n} \cdot C_n \cdot c_{3,n}) .$$

In the *sixth step of the proof* we show

$$T_{4,n}$$
$$\leq c_{31} \cdot \left(\mathbf{E}\left\{ \int |f_{(((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k, ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{E_n} \right\} \right.$$
$$\left. + (\beta_n + \gamma_n) \cdot \sqrt{\alpha_n} \cdot C_n \cdot c_{3,n} + \sum_{k=1}^{K_n} |(\mathbf{w}^*)_{k,1,1}^{(L_n)}|^2 + \frac{K_n \cdot D_n^2}{t_n} \right).$$

To do this we apply Lemma 1. Because of $|T_{\beta_n} z - y| \leq |z - y|$ for $|y| \leq \beta_n$ the definition of $\bar{E}_n$ implies

$$T_{4,n}$$
$$\leq 2 \cdot \mathbf{E}\left\{ \frac{1}{t_n} \sum_{t=0}^{t_n-1} \left(|f_{\mathbf{w}^{(t)}}(X_{j_t}) - Y_{j_t}|^2 - |m(X_{j_t}) - Y_{j_t}|^2 \right) \cdot 1_{\bar{E}_n} \right\}$$
$$= 2 \cdot \mathbf{E}\left\{ \frac{1}{t_n} \sum_{t=0}^{t_n-1} F_t\left(((\mathbf{w}^{(t)})_{k,1,1}^{(L_n)})_k, ((\mathbf{w}^{(t)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n} \right) \cdot 1_{\bar{E}_n} \right\}$$

where

$$F_t\left((\mathbf{w}_{k,1,1}^{(L_n)})_k, (\mathbf{w}_{k,i,j}^{(l)})_{k,i,j,l:l<L_n} \right) = |f_{((\mathbf{w}_{k,1,1}^{(L_n)})_k, (\mathbf{w}_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(X_{j_t}) - Y_{j_t}|^2 - |m(X_{j_t}) - Y_{j_t}|^2$$

is a convex and differentiable function of its first argument with 2-norm of the gradient bounded by $\sqrt{K_n} \cdot D_n$. Application of Lemma 1 with

$$u_t = ((\mathbf{w}^{(t)})_{k,1,1}^{(L_n)})_k, \quad v_t = ((\mathbf{w}^{(t)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n} \quad \text{and} \quad u^* = ((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k$$

yields

$$T_{4,n}$$
$$\leq 2 \cdot \mathbf{E}\left\{ \frac{1}{t_n} \sum_{t=0}^{t_n-1} F_t\left(u^*, v_0\right) \cdot 1_{\bar{E}_n} \right\} + 2 \cdot \mathbf{E}\left\{ \frac{1}{t_n} \sum_{t=0}^{t_n-1} |F_t(u^*, v_t) - F_t(u^*, v_0)| \cdot 1_{\bar{E}_n} \right\} + \|u^*\|^2$$
$$+ \frac{K_n \cdot D_n^2}{2 \cdot t_n}.$$

Arguing as in the fifth step of the proof we get

$$\mathbf{E}\left\{ \frac{1}{t_n} \sum_{t=0}^{t_n-1} |F_t(u^*, v_t) - F_t(u^*, v_0)| \cdot 1_{\bar{E}_n} \right\} \leq c_{32} \cdot (\beta_n + \gamma_n) \cdot \sqrt{\alpha_n} \cdot C_n \cdot c_{3,n}.$$

So it remains to bound

$$\mathbf{E}\left\{\frac{1}{t_n}\sum_{t=0}^{t_n-1} F_t\left(u^*, v_0\right)\cdot 1_{\bar{E}_n}\right\}.$$

Since $\bar{E}_n \subseteq E_n$ we get

$$\begin{aligned}
&\mathbf{E}\left\{\frac{1}{t_n}\sum_{t=0}^{t_n-1} F_t\left(u^*, v_0\right)\cdot 1_{\bar{E}_n}\right\}\\
&\leq \mathbf{E}\left\{\frac{1}{t_n}\sum_{t=0}^{t_n-1} \mathbf{E}\left\{F_t\left(u^*, v_0\right)\cdot 1_{\bar{E}_n}\middle| \mathbf{w}^{(0)}, \mathcal{D}_n\right\}\right\}\\
&\leq \mathbf{E}\left\{\int |f_{((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k,((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x) - m(x)|^2 \mathbf{P}_X(dx)\cdot 1_{\bar{E}_n}\right\}\\
&\leq \mathbf{E}\left\{\int |f_{((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k,((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x) - m(x)|^2 \mathbf{P}_X(dx)\cdot 1_{E_n}\right\}.
\end{aligned}$$

Summarizing the above results the proof is complete. □

4.2 Proof of Theorem 1

4.2.1 Auxiliary results

In order to apply Theorem 4 in the proof of Theorem 1 we will need the following auxiliary results.

Lemma 2 *Let σ be the logistic activation function. Let $a, B_n \geq 1$, $L_n, r_n \in \mathbb{N}$ and define the deep neural network $f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ with weight vector $\mathbf{w}$ by (8)-(10). Assume that the weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ satisfy*

$$|w_{k,i,j}^{(l)}| \leq B_n$$

for all $l \in \{1, \ldots, L_n - 1\}$. Then we have for any $x \in [-a, a]^d$ and any $k \in \{1, \ldots, K_n\}$

$$\begin{aligned}
&\left|f_{\mathbf{w}_1,k,1,1}^{(L_n)}(x) - f_{\mathbf{w}_2,k,1,1}^{(L_n)}(x)\right|\\
&\leq a\cdot(\max\{2r_n, d\}+1)^{L_n}\cdot B_n^{L_n-1}\cdot \|((\mathbf{w}_1)_{k,i,j}^{(l)})_{i,j,l:l<L_n} - ((\mathbf{w}_2)_{k,i,j}^{(l)})_{i,j,l:l<L_n}\|_\infty.
\end{aligned}$$

Proof. We show

$$\begin{aligned}
&|f_{\mathbf{w}_1,k,j}^{(l)}(x) - f_{\mathbf{w}_2,k,j}^{(l)}(x)|\\
&\leq a\cdot(\max\{2r_n, d\}+1)^{l}\cdot B_n^{l-1}\cdot \|((\mathbf{w}_1)_{k,\bar{i},\bar{j}}^{(\bar{l})})_{\bar{i},\bar{j},\bar{l}} - ((\mathbf{w}_2)_{k,\bar{i},\bar{j}}^{(\bar{l})})_{\bar{i},\bar{j},\bar{l}}\|_\infty
\end{aligned}$$

for $l \in \{1, \ldots, L_n\}$ by induction on l.

For $l = 1$ we use that σ is Lipschitz continuous with Lipschitz constant 1 and get

$$|f_{\mathbf{w}_1,k,i}^{(1)}(x) - f_{\mathbf{w}_2,k,i}^{(1)}(x)|$$

$$
\begin{aligned}
&\leq \left| \sigma\left(\sum_{j=1}^{d} (\mathbf{w}_1)_{k,i,j}^{(0)} \cdot x^{(j)} + (\mathbf{w}_1)_{k,i,0}^{(0)} \right) - \sigma\left(\sum_{j=1}^{d} (\mathbf{w}_2)_{k,i,j}^{(0)} \cdot x^{(j)} + (\mathbf{w}_2)_{k,i,0}^{(0)} \right) \right| \\
&\leq \left| \sum_{j=1}^{d} (\mathbf{w}_1)_{k,i,j}^{(0)} \cdot x^{(j)} + (\mathbf{w}_1)_{k,i,0}^{(0)} - \sum_{j=1}^{d} (\mathbf{w}_2)_{k,i,j}^{(0)} \cdot x^{(j)} - (\mathbf{w}_2)_{k,i,0}^{(0)} \right| \\
&\leq \sum_{j=1}^{d} |(\mathbf{w}_1)_{k,i,j}^{(0)} - (\mathbf{w}_2)_{k,i,j}^{(0)}| \cdot a + |(\mathbf{w}_1)_{k,i,0}^{(0)} - (\mathbf{w}_2)_{k,i,0}^{(0)}| \\
&\leq a \cdot (\max\{2r_n, d\} + 1)^1 \cdot B_n^{1-1} \cdot \|((\mathbf{w}_1)_{k,\bar{i},\bar{j}}^{(\bar{l})})_{\bar{i},\bar{j},\bar{l}} - ((\mathbf{w}_2)_{k,\bar{i},\bar{j}}^{(\bar{l})})_{\bar{i},\bar{j},\bar{l}}\|_\infty.
\end{aligned}
$$

Assume next that the assertion holds for some $l \in \{1, \ldots, L_n - 1\}$. Then

$$
\begin{aligned}
&|f_{\mathbf{w}_1,k,i}^{(l+1)}(x) - f_{\mathbf{w}_2,k,i}^{(l+1)}(x)| \\
&\leq \left| \sigma\left(\sum_{j=1}^{r_n} (\mathbf{w}_1)_{k,i,j}^{(l)} \cdot f_{\mathbf{w}_1,k,j}^{(l)}(x) + (\mathbf{w}_1)_{k,i,0}^{(l)} \right) - \sigma\left(\sum_{j=1}^{r_n} (\mathbf{w}_2)_{k,i,j}^{(l)} \cdot f_{\mathbf{w}_2,k,j}^{(l)}(x) + (\mathbf{w}_2)_{k,i,0}^{(l)} \right) \right| \\
&\leq \left| \sum_{j=1}^{r_n} (\mathbf{w}_1)_{k,i,j}^{(l)} \cdot f_{\mathbf{w}_1,k,j}^{(l)}(x) + (\mathbf{w}_1)_{k,i,0}^{(l)} - \sum_{j=1}^{r_n} (\mathbf{w}_2)_{k,i,j}^{(l)} \cdot f_{\mathbf{w}_2,k,j}^{(l)}(x) - (\mathbf{w}_2)_{k,i,0}^{(l)} \right| \\
&\leq \sum_{j=1}^{r_n} |(\mathbf{w}_1)_{k,i,j}^{(l)} \cdot f_{\mathbf{w}_1,k,j}^{(l)}(x) - (\mathbf{w}_2)_{k,i,j}^{(l)} \cdot f_{\mathbf{w}_2,k,j}^{(l)}(x)| + |(\mathbf{w}_1)_{k,i,0}^{(l)} - (\mathbf{w}_2)_{k,i,0}^{(l-1)}| \\
&\leq \sum_{j=1}^{r_n} |(\mathbf{w}_1)_{k,i,j}^{(l)} - (\mathbf{w}_2)_{k,i,j}^{(l)}| + \sum_{j=1}^{r_n} |(\mathbf{w}_2)_{k,i,j}^{(l)}| \cdot |f_{\mathbf{w}_1,k,j}^{(l)}(x) - f_{\mathbf{w}_2,k,j}^{(l)}(x)| \\
&\quad + |(\mathbf{w}_1)_{k,i,0}^{(l)} - (\mathbf{w}_2)_{k,i,0}^{(l-1)}| \\
&\leq (r_n + 1) \cdot \|((\mathbf{w}_1)_{k,\bar{i},\bar{j}}^{(\bar{l})})_{\bar{i},\bar{j},\bar{l}} - ((\mathbf{w}_2)_{k,\bar{i},\bar{j}}^{(\bar{l})})_{\bar{i},\bar{j},\bar{l}}\|_\infty \\
&\quad + r_n \cdot B_n \cdot a \cdot (\max\{2r_n, d\} + 1)^l \cdot B_n^{l-1} \cdot \|((\mathbf{w}_1)_{k,\bar{i},\bar{j}}^{(\bar{l})})_{\bar{i},\bar{j},\bar{l}} - ((\mathbf{w}_2)_{k,\bar{i},\bar{j}}^{(\bar{l})})_{\bar{i},\bar{j},\bar{l}}\|_\infty \\
&\leq a \cdot (\max\{2r_n, d\} + 1)^{l+1} \cdot B_n^l \cdot \|((\mathbf{w}_1)_{k,\bar{i},\bar{j}}^{(\bar{l})})_{\bar{i},\bar{j},\bar{l}} - ((\mathbf{w}_2)_{k,\bar{i},\bar{j}}^{(\bar{l})})_{\bar{i},\bar{j},\bar{l}}\|_\infty.
\end{aligned}
$$

□

Lemma 3 *Let σ be the logistic activation function. Let $a, B_n \geq 1$, $L_n, r_n \in \mathbb{N}$ and define the deep neural network $f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ with weight vector $\mathbf{w}$ by (8)-(10). Assume that the weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ satisfy*

$$|w_{k,i,j}^{(l)}| \leq B_n$$

for all $l \in \{1, \ldots, L_n - 1\}$. Then we have for any $x \in [-a, a]^d$

$$
\begin{aligned}
&\sum_{k=1}^{K_n} \left| f_{\mathbf{w}_1,k,1,1}^{(L_n)}(x) - f_{\mathbf{w}_2,k,1,1}^{(L_n)}(x) \right|^2 \\
&\leq a^2 \cdot (\max\{2r_n, d\} + 1)^{2L_n} \cdot B_n^{2L_n - 2} \cdot \|((\mathbf{w}_1)_{k,i,j}^{(l)})_{k,i,j,l:l<L_n} - ((\mathbf{w}_2)_{k,i,j}^{(l)})_{k,i,j,l:l<L_n}\|^2.
\end{aligned}
$$

Proof. By Lemma 2 we get

$$\begin{aligned}
&\sum_{k=1}^{K_n} \left| f_{\mathbf{w}_1,k,1,1}^{(L_n)}(x) - f_{\mathbf{w}_2,k,1,1}^{(L_n)}(x) \right|^2 \\
&\leq \sum_{k=1}^{K_n} a^2 \cdot (\max\{2r_n, d\} + 1)^{2L_n} \cdot B_n^{2L_n - 2} \cdot \|((\mathbf{w}_1)_{k,i,j}^{(l)})_{i,j,l:l<L_n} - ((\mathbf{w}_2)_{k,i,j}^{(l)})_{i,j,l:l<L_n}\|_\infty^2 \\
&\leq \sum_{k=1}^{K_n} a^2 \cdot (\max\{2r_n, d\} + 1)^{2L_n} \cdot B_n^{2L_n - 2} \cdot \|((\mathbf{w}_1)_{k,i,j}^{(l)})_{i,j,l:l<L_n} - ((\mathbf{w}_2)_{k,i,j}^{(l)})_{i,j,l:l<L_n}\|^2 \\
&= a^2 \cdot (\max\{2r_n, d\} + 1)^{2L_n} \cdot B_n^{2L_n - 2} \cdot \sum_{k=1}^{K_n} \|((\mathbf{w}_1)_{k,i,j}^{(l)})_{i,j,l:l<L_n} - ((\mathbf{w}_2)_{k,i,j}^{(l)})_{i,j,l:l<L_n}\|^2 \\
&= a^2 \cdot (\max\{2r_n, d\} + 1)^{2L_n} \cdot B_n^{2L_n - 2} \cdot \|((\mathbf{w}_1)_{k,i,j}^{(l)})_{k,i,j,l:l<L_n} - ((\mathbf{w}_2)_{k,i,j}^{(l)})_{k,i,j,l:l<L_n}\|^2.
\end{aligned}$$

□

In order to bound the approximation error in the proof of Theorem 1 we will apply the following result.

Lemma 4 *Let $d \in \mathbb{N}$, $p = q + \beta$ where $\beta \in (0, 1]$ and $q \in \mathbb{N}_0$, $C > 0$, $A \geq 1$ and $A_n, B_n, \gamma_n^* \geq 1$. For $L, r, K \in \mathbb{N}$ let $\mathcal{F}$ be the set of all networks $f_{\mathbf{w}}$ defined by (8)–(10) with K_n replaced by 1, L_n replaced by L and r_n replaced by r, where the weight vector satisfies*

$$|w_{1,i,j}^{(0)}| \leq A_n, \quad |w_{1,i,j}^{(l)}| \leq B_n \quad \text{and} \quad |w_{1,i,j}^{(L)}| \leq \gamma_n^*$$

for all $l \in \{1, \ldots, L-1\}$ and all i, j, and set

$$\mathcal{H} = \left\{ \sum_{k=1}^{K^d} f_k \quad : \quad f_k \in \mathcal{F} \quad (k = 1, \ldots, K) \right\}.$$

Let $L, r \in \mathbb{N}$ with

$$L \geq \lceil \log_2(q + d) \rceil \quad \text{and} \quad r \geq 2 \cdot (2p + d) \cdot (q + d),$$

and set

$$A_n = A \cdot K \cdot \log K, \quad B_n = c_{34} \quad \text{and} \quad \gamma_n^* = c_{35} \cdot K^{q+d}$$

with $c_{34}, c_{35} \geq 0$ sufficiently large. Assume $K \geq c_{36}$ for $c_{36} > 0$ sufficiently large. Then there exists for any (p, C)–smooth $f : \mathbb{R}^d \to \mathbb{R}$ a neural network $h \in \mathcal{H}$ such that

$$\sup_{x \in [-A,A)^d} |f(x) - h(x)| \leq \frac{c_{37}}{K^p}.$$

Proof. See Lemma 2 in Kohler (2024). □

The generalization error in the proof of Theorem 1 will be bounded by using the following metric entropy bound for deep neural networks with smooth activation function.

Lemma 5 *Let $\alpha, \beta \geq 1$ 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 (8)–(10) where the weight vector $\mathbf{w}$ satisfies*

$$\sum_{j=1}^{K_n} |w_{1,1,j}^{(L)}| \leq C, \tag{24}$$

$$|w_{k,i,j}^{(l)}| \leq B \quad (k \in \{1, \ldots, K_n\}, i, j \in \{1, \ldots, r\}, l \in \{1, \ldots, L-1\}) \tag{25}$$

and

$$|w_{k,i,j}^{(0)}| \leq A \quad (k \in \{1, \ldots, K_n\}, i \in \{1, \ldots, r\}, j \in \{1, \ldots, d\}). \tag{26}$$

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

$$\mathcal{N}_p\left(\epsilon, \{T_\beta f \cdot 1_{[-\alpha,\alpha]^d} : f \in \mathcal{F}\}, x_1^n\right) \leq \left(c_{38} \cdot \frac{\beta^p}{\epsilon^p}\right)^{c_{39} \cdot \alpha^d \cdot B^{(L-1)\cdot d} \cdot A^d \cdot \left(\frac{C}{\epsilon}\right)^{d/k} + c_{40}}.$$

Proof. See Lemma 4 in Drews and Kohler (2024). □

4.2.2 Proof of Theorem 1

The assertion follows more or less directly from Theorem 4 by using arguments as in the proof of Theorem 1 in Kohler (2024). For the sake of completeness we nevertheless present the complete proof.

W.l.o.g. we assume throughout the proof that n is sufficiently large and that $\|m\|_\infty \leq \beta_n$ holds. Let $A > 0$ with $supp(X) \subseteq [-A, A]^d$. Set

$$\tilde{K}_n = \left\lceil c_{41} \cdot n^{\frac{d}{2p+d}} \right\rceil$$

and

$$N_n = \left\lceil c_{42} \cdot n^9 \right\rceil$$

and let $\bar{\mathbf{w}}$ be a weight vector of a neural networks where the results of $N_n \cdot \tilde{K}_n$ in parallel computed neural networks with L hidden layers and r neurons per layer are computed such that the corresponding network

$$f_{\bar{\mathbf{w}}}(x) = \sum_{k=1}^{N_n \cdot \tilde{K}_n} (\bar{\mathbf{w}})_{1,1,k} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x)$$

satisfies

$$\sup_{x \in [-A,A]^d} |f_{\bar{\mathbf{w}}}(x) - m(x)| \leq \frac{c_{43}}{\tilde{K}_n^{p/d}} \tag{27}$$

and

$$|(\bar{\mathbf{w}})_{k,1,1}^{(L)}| \leq \frac{c_{44} \cdot \tilde{K}_n^{(q+d)/d}}{N_n} \quad (k = 1, \ldots, N_n \cdot \tilde{K}_n)$$

and

$$|(\bar{\mathbf{w}})_{k,i,j}^{(l)}| \leq c_{45}$$

for $l \in \{1, \ldots, L-1\}$ and

$$|(\bar{\mathbf{w}})_{k,i,j}^{(0)}| \leq c_{46} \cdot \tilde{K}_n^{1/d} \cdot \log(\tilde{K}_n).$$

Note that such a network exists according Lemma 4 if we repeat in the outer sum of the function space $\mathcal{H}$ each of the f_k's in Lemma 4 N_n–times with outer weights divided by N_n. By construction, the outer weights of this network satisfy

$$\sum_{k=1}^{N_n \cdot \tilde{K}_n} |(\bar{\mathbf{w}})_{k,1,1}^{(L)}| \leq \tilde{K}_n \cdot c_{44} \cdot \tilde{K}_n^{\frac{q+d}{d}} \leq \gamma_n$$

and

$$\sum_{k=1}^{N_n \cdot \tilde{K}_n} |(\bar{\mathbf{w}})_{k,1,1}^{(L)}|^2 \leq \frac{\tilde{K}_n^{(2q+3d)/d}}{N_n} \leq \alpha_n.$$

Set

$$\epsilon_n = \frac{c_{47}}{n^3}.$$

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

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

for some pairwise distinct $j_1, \ldots, j_{N_n \cdot \tilde{K}_n} \in \{1, \ldots, K_n\}$. Define

$$(\mathbf{w}^*)_{j_k,1,1}^{(L)} = \bar{\mathbf{w}}_{k,1,1}^{(L)} \quad (k = 1, \ldots, N_n \cdot \tilde{K}_n)$$

and

$$(\mathbf{w}^*)_{k,1,1}^{(L)} = 0 \quad (k \in \{1, \ldots, K_n\} \setminus \{j_1, \ldots, j_{N_n \cdot \tilde{K}_n}\}.$$

Next we check the assumptions of Theorem 4. By construction of our estimate its weights satisfy the constraints

$$|w_{i,j,k}^{(L_n)}| \leq \gamma_n = c_5 \cdot n^2, \quad |w_{i,j,k}^{(l)}| \leq c_{2,n} + c_{3,n} \leq c_{48} \cdot \log n$$

and

$$|w_{i,j,k}^{(0)}| \leq c_{1,n} + c_{3,n} \leq c_{49} \cdot (\log n)^2 \cdot n^{\frac{1}{2p+d}}.$$

By Lemma 3 we know that (21) holds with

$$C_n^2 = c_{50} \cdot B_n^{2L} = c_{50} \cdot (c_{2,n} + c_{3,n})^{2L} \leq c_{51} \cdot (\log n)^{2L}.$$

And on the event $\{\max_{i=1,\ldots,n} |Y_i| \leq \beta_n\}$ we have

$$\left\| \nabla_{(\mathbf{w}_{k,1,1}^{(L_n)})_k} (Y_{j_t} - f_{\mathbf{w}^{(t)}}(X_{j_t}))^2 \right\|_\infty \quad \leq \quad \max_{k=1,\ldots,K_n} 2 \cdot (\beta_n + |f_{\mathbf{w}^{(t)}}(X_{j_t})|) \cdot |f_{\mathbf{w}^{(t)},1,k}(X_{j_t})|$$

$$\leq \quad 2 \cdot (\beta_n + \gamma_n) \cdot 1,$$

so (22) holds with

$$D_n = 2 \cdot (\beta_n + \gamma_n) \leq c_{52} \cdot n^2.$$

Application of Theorem 4 yields

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq c_{53} \cdot \left(\frac{\beta_n^2}{n^5} + \beta_n^2 \cdot \sqrt{\mathbf{P}(E_n^c)} + \mathbf{E} \Big\{ \sup_{\mathbf{w} \in \mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X) - Y|^2 - |m(X) - Y|^2\} \right. \\
&\qquad \left. - \frac{2}{n} \sum_{i=1}^{n} \left(|T_{\beta_n} f_{\mathbf{w}}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \Big) \Big\} \right. \\
&\quad + \mathbf{E} \left\{ \int |f_{((\mathbf{w}^*)^{(L)}, ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{E_n} \right\} \\
&\quad \left. + \beta_n \cdot n \cdot \lambda_n \cdot K_n \cdot D_n + (\beta_n + \gamma_n) \cdot \sqrt{\alpha_n} \cdot C_n \cdot c_{3,n} + \|(\mathbf{w}^*)^{(L)}\|^2 + \frac{K_n \cdot D_n^2}{t_n} \right).
\end{aligned}$$

Next we bound $\mathbf{P}(E_n^c)$. To do this, we consider a sequential choice of the weights of the K_n fully connected neural networks. The probability that the weights in the first of these networks differ in all components at most by ϵ_n from $(\bar{\mathbf{w}})_{1,i,j}^{(l)}$ $(l = 0, \ldots, L-1)$ is for large n bounded from below by

$$\begin{aligned}
&\left(\frac{c_{93}}{2 \cdot c_2 \cdot n^3} \right)^{r \cdot (r+1) \cdot (L-2) + r + 1} \cdot \left(\frac{c_{93}}{2 \cdot (\log n)^2 \cdot n^{1/(2p+d)} \cdot n^3} \right)^{r \cdot (d+1)} \\
&\geq n^{-r \cdot (r+1) \cdot (L-2) \cdot 3 - 3 \cdot (r+1) - 4 \cdot r \cdot (d+1) - 0.5}.
\end{aligned}$$

Hence probability that none of the first $n^{r \cdot (r+1) \cdot (L-2) \cdot 3 + 3 \cdot (r+1) + 4 \cdot r \cdot (d+1) + 1}$ neural networks satisfies this condition is for large n bounded above by

$$\begin{aligned}
&(1 - n^{-r \cdot (r+1) \cdot (L-2) \cdot 3 - 3 \cdot (r+1) - 4 \cdot r \cdot (d+1) - 0.5})^{n^{r \cdot (r+1) \cdot (L-2) \cdot 3 + 3 \cdot (r+1) + 4 \cdot r \cdot (d+1) + 1}} \\
&\leq \left(\exp\left(-n^{-r \cdot (r+1) \cdot (L-2) \cdot 3 - 3 \cdot (r+1) - 4 \cdot r \cdot (d+1) - 0.5} \right) \right)^{n^{r \cdot (r+1) \cdot (L-2) \cdot 3 + 3 \cdot (r+1) + 4 \cdot r \cdot (d+1) + 1}} \\
&= \exp(-n^{0.5}).
\end{aligned}$$

Since we have $K_n \geq n^{r \cdot (r+1) \cdot (L-2) \cdot 3 + 3 \cdot (r+1) + 4 \cdot r \cdot (d+1) + 1} \cdot N_n \cdot \tilde{K}_n$ for n large we can successively use the same construction for all of $N_n \cdot \tilde{K}_n$ weights and we can conclude: The probability that there exists $k \in \{1, \ldots, N_n \cdot \tilde{K}_n\}$ such that none of the K_n weight vectors of the fully connected neural network differs by at most ϵ_n from $((\bar{\mathbf{w}})_{i,j,k}^{(l)})_{i,j,l}$ is for large n bounded from above by

$$N_n \cdot \tilde{K}_n \cdot \exp(-n^{0.5}) \leq c_{54} \cdot n^{10} \cdot \exp(-n^{0.5}) \leq \frac{c_{55}}{n^2}.$$

This proves

$$\mathbf{P}\{E_n^c\} \leq \frac{c_{55}}{n^2}.$$

Next we bound

$$\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X)-Y|^2-|m(X)-Y|^2\}-\frac{2}{n}\sum_{i=1}^n \big(|T_{\beta_n} f_{\mathbf{w}}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\big)\Big)\Big\}.$$

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

$$\begin{aligned}
&\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X)-Y|^2-|m(X)-Y|^2\}-\frac{2}{n}\sum_{i=1}^n \big(|T_{\beta_n} f_{\mathbf{w}}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\big)\Big)\Big\}\\
&\leq \mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X)-T_{\beta_n}Y|^2-|m_{\beta_n}(X)-T_{\beta_n}Y|^2\}\\
&\qquad\qquad -\frac{2}{n}\sum_{i=1}^n \big(|T_{\beta_n} f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2-|m_{\beta_n}(X_i)-T_{\beta_n}Y_i|^2\big)\Big)\Big\}\\
&\quad +\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X)-Y|^2-|m(X)-Y|^2\}\\
&\qquad\qquad -\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X)-T_{\beta_n}Y|^2-|m_{\beta_n}(X)-T_{\beta_n}Y|^2\}\Big)\Big\}\\
&\quad +\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big(\frac{2}{n}\sum_{i=1}^n \big(|T_{\beta_n} f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2-|m_{\beta_n}(X_i)-T_{\beta_n}Y_i|^2\big)\\
&\qquad\qquad -\frac{2}{n}\sum_{i=1}^n \big(|T_{\beta_n} f_{\mathbf{w}}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\big)\Big)\Big\}\\
&=: T_{1,n}+T_{2,n}+T_{3,n}.
\end{aligned}$$

As in the proof of Lemma 1 in Bauer and Kohler (2019) we get

$$T_{2,n}+T_{3,n} \leq c_{56}\cdot\frac{(\log n)^2}{n}.$$

Next we show

$$\mathbf{E}T_{1,n} \leq c_{57}\cdot\frac{n^{d/(2p+d)+\delta}}{n}.$$

Let $\delta_n \geq 1/n$. Then

$$\begin{aligned}
&\mathbf{E}\{T_{1,n}\}\\
&\leq \int_0^\infty \mathbf{P}\{T_{1,n} > t\}\,dt\\
&\leq \delta_n + \int_{\delta_n}^\infty \mathbf{P}\Big\{\exists \mathbf{w}\in\mathbf{\Theta}: \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X)-T_{\beta_n}Y|^2-|m_{\beta_n}(X)-T_{\beta_n}Y|^2\}\\
&\qquad\qquad -\frac{2}{n}\sum_{i=1}^n \big(|T_{\beta_n} f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2-|m_{\beta_n}(X_i)-T_{\beta_n}Y_i|^2\big)\Big) > t\Big\}\,dt.
\end{aligned}$$

By Theorem 11.4 in Györfi et al. (2002) we get for $t > 1/n$

$$\begin{aligned}
&\mathbf{P}\Big\{\exists \mathbf{w} \in \mathbf{\Theta} : \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X) - T_{\beta_n} Y|^2 - |m_{\beta_n}(X) - T_{\beta_n} Y|^2\} \\
&\qquad\qquad -\frac{2}{n}\sum_{i=1}^{n}\left(|T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2\right)\Big) > t\Big\} \\
&= \mathbf{P}\Big\{\exists \mathbf{w} \in \mathbf{\Theta} : \Big(\mathbf{E}\{|\frac{T_{\beta_n} f_{\mathbf{w}}(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}|^2 - |\frac{m_{\beta_n}(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}|^2\} \\
&\qquad\qquad -\frac{2}{n}\sum_{i=1}^{n}\left(|\frac{T_{\beta_n} f_{\mathbf{w}}(X_i)}{\beta_n} - \frac{T_{\beta_n} Y_i}{\beta_n}|^2 - |\frac{m_{\beta_n}(X_i)}{\beta_n} - \frac{T_{\beta_n} Y_i}{\beta_n}|^2\right)\Big) > \frac{t}{\beta_n^2}\Big\} \\
&\leq 14 \cdot \sup_{x_1^n \in supp(X)} \mathcal{N}_2\left(\frac{1}{80 \cdot \beta_n^2 \cdot n}, \left\{\frac{1}{\beta_n} \cdot f_{\mathbf{w}} \,:\, \mathbf{w} \in \mathbf{\Theta}\right\}, x_1^n\right) \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot t\right).
\end{aligned}$$

Application of Lemma 5 with $A_n = c_{58} \cdot (\log n)^2 \cdot n^{\frac{1}{2p+d}}$, $B_n = c_{59} \cdot \log n$ and $C_n = \gamma_n = c_{60} \cdot n^2$ yields for k large enough

$$\begin{aligned}
&\sup_{x_1^n \in supp(X)} \mathcal{N}_2\left(\frac{1}{80 \cdot \beta_n^2 \cdot n}, \left\{\frac{1}{\beta_n} \cdot T_{\beta_n} f_{\mathbf{w}} \,:\, \mathbf{w} \in \mathbf{\Theta}\right\}, x_1^n\right) \\
&\leq \sup_{x_1^n \in supp(X)} \mathcal{N}_2\left(\frac{1}{80 \cdot \beta_n \cdot n}, \{T_{\beta_n} f_{\mathbf{w}} \,:\, \mathbf{w} \in \mathbf{\Theta}\}, x_1^n\right) \\
&\leq c_{61} \cdot n^{c_{62} \cdot n^{\frac{d}{2p+d}+\delta/2}}.
\end{aligned}$$

Hence

$$\mathbf{E}\{T_{21n}\} \;\leq\; \delta_n + 14 \cdot c_{61} \cdot n^{c_{62} \cdot n^{\frac{d}{2p+d}+\delta/2}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot \delta_n\right) \cdot \frac{5136 \cdot \beta_n^2}{n}.$$

Setting

$$\delta_n = \frac{5136 \cdot \beta_n^2}{n} \cdot c_{62} \cdot n^{\frac{d}{2p+d}+\delta/2} \cdot \log n$$

we get

$$\mathbf{E}T_{1,n} \leq c_{63} \cdot \frac{n^{d/(2p+d)+\delta}}{n} = c_{63} \cdot n^{-\frac{2p}{2p+d}+\delta}.$$

This proves

$$\begin{aligned}
&\mathbf{E}\Big\{\sup_{\mathbf{w} \in \mathbf{\Theta}}\Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X) - Y|^2 - |m(X) - Y|^2\} \\
&\qquad\qquad -\frac{2}{n}\sum_{i=1}^{n}\left(|T_{\beta_n} f_{\mathbf{w}}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right)\Big)\Big\} \\
&\leq c_{64} \cdot n^{-2p/(2p+d)+\delta}.
\end{aligned}$$

So it remains to bound

$$\mathbf{E}\left\{\int |f_{(((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k, ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{E_n}\right\}.$$

We have

$$\begin{aligned}
&\mathbf{E}\left\{\int |f_{(((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k,((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x)-m(x)|^2\mathbf{P}_X(dx)\cdot 1_{E_n}\right\}\\
&\le 2\cdot\int |f_{\bar{\mathbf{w}}}(x)-m(x)|^2\mathbf{P}_X(dx)\\
&\quad +2\cdot\mathbf{E}\left\{\int |f_{\bar{\mathbf{w}}}(x)-f_{(((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k,((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x)|^2\mathbf{P}_X(dx)\cdot 1_{E_n}\right\}.
\end{aligned}$$

By (27) we know

$$\int |f_{\bar{\mathbf{w}}}(x)-m(x)|^2\mathbf{P}_X(dx)\le\frac{c_{43}^2}{\tilde{K}_n^{2p/d}}\le c_{65}\cdot n^{-\frac{2p}{2p+d}}.$$

And using that on E_n we have

$$\begin{aligned}
&|f_{\bar{\mathbf{w}}}(x)-f_{(((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k,((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x)|\\
&=\left|\sum_{k=1}^{N_n\cdot\tilde{K}_n}(\bar{\mathbf{w}})_{1,1,k}\cdot(f_{\bar{\mathbf{w}},k,1}^{(L)}(x)-f_{\mathbf{w}^{(0)},j_k,1}^{(L)}(x))\right|\\
&\le\sum_{k=1}^{N_n\cdot\tilde{K}_n}|(\bar{\mathbf{w}})_{1,1,k}^{(L)}|\cdot|f_{\bar{\mathbf{w}},k,1}^{(L)}(x)-f_{\mathbf{w}^{(0)},j_k,1}^{(L)}(x)|\\
&\le\gamma_n\cdot\max_{k=1,\dots,N_n\cdot\tilde{K}_n}|f_{\bar{\mathbf{w}},k,1}^{(L)}(x)-f_{\mathbf{w}^{(0)},j_k,1}^{(L)}(x)|\\
&\le\gamma_n\cdot c_{66}\cdot\epsilon_n\le\frac{c_{67}}{n}
\end{aligned}$$

(where the third inequality followed from Lemma 2) we get

$$\mathbf{E}\left\{\int |f_{\mathbf{w}^*}(x)-f_{(u^*,((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x)|^2\mathbf{P}_X(dx)\cdot 1_{E_n}\right\}\le\frac{c_{68}}{n}.$$

Summarizing the above results we get

$$\begin{aligned}
&\mathbf{E}\int |m_n(x)-m(x)|^2\mathbf{P}_X(dx)\\
&\le c_{69}\cdot\left(\frac{\beta_n^2}{n^5}+\frac{\beta_n^2}{n}+n^{-2p/(2p+d)+\delta}+n^{-\frac{2p}{2p+d}}+\frac{1}{n}\right.\\
&\qquad\left.+\beta_n\cdot\frac{1}{n^2}+(\beta_n+\gamma_n)\cdot(\log n)^{L+1}\cdot\frac{1}{n^3}+\frac{1}{n^6}+\frac{1}{n}\right)\\
&\le c_{70}\cdot n^{-\frac{2p}{2p+d}+\delta}.
\end{aligned}$$

□

4.3 Proof of Theorem 2

4.3.1 Auxiliary results

The next lemma will be used in order to bound the approximation error in Theorem 2.

Lemma 6 *Let $a \geq 1$ and $p = q + s$ for some $q \in \mathbb{N}_0$ and $s \in (0, 1]$, and let $C > 0$. Let $m : \mathbb{R}^d \to \mathbb{R}$ be a (p, C)-smooth function, which satisfies*

$$\max_{\substack{j_1,\ldots,j_d \in \{0,1,\ldots,q\},\\ j_1+\cdots+j_d \leq q}} \left\| \frac{\partial^{j_1+\cdots+j_d} m}{\partial^{j_1} x^{(1)} \cdots \partial^{j_d} x^{(d)}} \right\|_{\infty,[-2a,2a]^d} \leq c_{71}. \tag{28}$$

Let ν be an arbitrary probability measure on $\mathbb{R}^d$. Let $N \in \mathbb{N}_0$ be chosen such that $N \geq q$ and let $\sigma : \mathbb{R} \to [0, 1]$ be the logistic activation function. Then, for any $\eta \in (0, 1)$ and $M \in \mathbb{N}$ sufficiently large (independent of the size of a and η, but $a \leq M$ must hold), a neural network of the type

$$t(x) = \sum_{i=1}^{\binom{d+N}{d} \cdot (N+1) \cdot (M+1)^d} \mu_i \cdot \sigma\left(\sum_{l=1}^{4d} \lambda_{i,l} \cdot \sigma\left(\sum_{v=1}^{d} \theta_{i,l,v} \cdot x^{(v)} + \theta_{i,l,0} \right) + \lambda_{i,0} \right) \tag{29}$$

exists such that

$$|t(x) - m(x)| \leq c_{72} \cdot a^{N+q+3} \cdot M^{-p}$$

holds for all $x \in [-a, a]^d$ up to a set of ν-measure less than or equal to η. The weights of $t(x)$ can be bounded by

$$\begin{aligned} |\mu_i| &\leq c_{73} \cdot a^q \cdot M^{N \cdot p} \\ |\lambda_{i,l}| &\leq M^{d+p \cdot (N+2)} \\ |\theta_{i,l,v}| &\leq 6 \cdot d \cdot \frac{1}{\eta} \cdot M^{d+p \cdot (2N+3)+1} \end{aligned}$$

for all $i \in \left\{1, \ldots, \binom{d+N}{N} \cdot (N+1) \cdot (M+1)^d\right\}$, $l \in \{0, \ldots, 4d\}$, and $v \in \{0, \ldots, d\}$.

Proof. See Theorem 2 in Bauer and Kohler (2019). □

The neural network in the lemma above has large outer weights. In order to construct a neural network with smaller outer weights we will compose it with the network in the next lemma.

Lemma 7 *Let σ be the logistic activation function, let $t_\sigma \in \mathbb{R}$ be such that $\sigma'(t_\sigma) \neq 0$. Then for any $N \in \mathbb{N}$ with $N > 1$ there exist*

$$\alpha_j, \beta_j \in \mathbb{R} \quad (j = 0, \ldots, N-1)$$

such that for any $R > 0$

$$f_{id}(x) = \frac{R}{\sigma'(t_\sigma)} \cdot \sum_{j=0}^{N-1} \alpha_j \cdot \sigma\left(\frac{\beta_j \cdot x}{R} + t_\sigma \right)$$

satisfies for all $A > 0$ *and all* $x \in [-A, A]$*:*

$$|f_{id}(x) - x| \leq c_{74} \cdot \frac{A^N}{R^{N-1}}$$

for some $c_{74} = c_{54}(N, \sigma^{(\prime)}(t_\sigma), \|\sigma^{(N)}\|_\infty, \alpha_0, \ldots, \alpha_{N-1}, \beta_0, \ldots, \beta_{N-1}) \geq 0$*.*

Proof. The proof is based on a modification of the proof of Theorem 2 in Scarselli and Tsoi (1998) presented in the proof of Lemma 9 in Kohler (2024).

Let $\beta_j \in \mathbb{R}$ $(j = 0, \ldots, N-1)$ be pairwise distinct. Then the vectors

$$\mathbf{v}_l = (\beta_0^l, \ldots, \beta_{N-1}^l)^T \quad (l = 0, \ldots, N-1)$$

are linearly independent since

$$\sum_{l=0}^{N-1} \alpha_l \cdot \mathbf{v}_l = 0$$

implies that the polynomial

$$p(x) = \sum_{l=0}^{N-1} \alpha_l \cdot x^l$$

of degree $N-1$ has the N roots $\beta_0, \ldots, \beta_{N-1}$, which is possible only in case $\alpha_0 = \cdots = \alpha_{N-1} = 0$. Hence we can choose $\alpha_0, \ldots, \alpha_{N-1} \in \mathbb{R}$ such that

$$\alpha_0 \cdot \mathbf{v}_0 + \cdots + \alpha_{N-1} \cdot \mathbf{v}_{N-1}$$

is equal to the second unit vector in $\mathbb{R}^N$, which implies

$$\sum_{j=0}^{N-1} \alpha_j \cdot \beta_j^l = \begin{cases} 1, & \text{if } l = 1 \\ 0, & \text{if } l \in \{0, \ldots, N-1\} \setminus \{1\}. \end{cases} \tag{30}$$

Using these values for the α_j and β_j, a Taylor expansion of

$$u \mapsto \sigma(u + t_\sigma)$$

around t_σ of order $N-1$ implies

$$\begin{aligned} f_{id}(x) &= \frac{R}{\sigma'(t_\sigma)} \cdot \sum_{j=0}^{N-1} \alpha_j \cdot \left(\sum_{l=0}^{N-1} \frac{\sigma^{(l)}(t_\sigma)}{l!} \cdot \left(\frac{\beta_j \cdot x}{R} \right)^l + \frac{\sigma^{(N)}(\xi_j)}{N!} \cdot \left(\frac{\beta_j \cdot x}{R} \right)^N \right) \\ &= \frac{R}{\sigma'(t_\sigma)} \cdot \sum_{l=0}^{N-1} \frac{\sigma^{(l)}(t_\sigma)}{l!} \cdot \left(\frac{x}{R} \right)^l \cdot \left(\sum_{j=0}^{N-1} \alpha_j \cdot \beta_j^l \right) \\ &\quad + \frac{R}{\sigma'(t_\sigma)} \cdot \sum_{j=0}^{N-1} \alpha_j \cdot \frac{\sigma^{(N)}(\xi_j)}{N!} \cdot \left(\frac{\beta_j \cdot x}{R} \right)^N \end{aligned}$$

$$= \quad x + \frac{1}{\sigma'(t_\sigma)} \cdot \sum_{j=0}^{N-1} \alpha_j \cdot \frac{\sigma^{(N)}(\xi_j)}{N!} \cdot \frac{\beta_j^N \cdot x^N}{R^{N-1}},$$

where the last equality follows from (30). Hence

$$|f_{id}(x) - x| \leq \left| \frac{1}{\sigma'(t_\sigma)} \cdot \sum_{j=0}^{N-1} \alpha_j \cdot \frac{\sigma^{(N)}(\xi_j)}{N!} \cdot \beta_j^N \right| \cdot \frac{|x|^N}{R^{N-1}} \leq c_{74} \cdot \frac{|x|^N}{R^{N-1}} \leq c_{74} \cdot \frac{A^N}{R^{N-1}}.$$

□

In the next lemma we bound the generalization error of the estimate by a Rademacher complexity.

Lemma 8 *Assume that (X, Y) satisfies m bounded and assumption (15). Let $\mathbf{\Theta}$ be a set of weight vectors $\mathbf{w} = (w_{i,j,k}^{(l)})_{i,j,k,l}$ of the neural networks $f_{\mathbf{w}}$ defined by (8)–(10), where all weight vectors satisfy*

$$\sum_{k=1}^{K_n} |w_{k,1,1}^{(L_n)}| \leq \gamma_n \tag{31}$$

for some $\gamma_n \geq 0$. Set $\beta_n = const \cdot \log n$. Then

$$\begin{aligned} &\mathbf{E}\Big\{ \sup_{\mathbf{w} \in \mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X) - Y|^2 - |m(X) - Y|^2\} \\ &\qquad\qquad - \frac{1}{n} \sum_{i=1}^n \left(|T_{\beta_n} f_{\mathbf{w}}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \Big) \Big\} \\ &\leq \frac{c_{75}}{\sqrt{n}} + 8 \cdot \beta_n \cdot \gamma_n \cdot \mathbf{E}\Big\{ \sup_{\mathbf{w} \in \mathbf{\Theta}, k \in \{1,\ldots,K_n\}} \Big| \frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot f_{\mathbf{w},k,1}^{(L_n)}(X_i) \Big| \Big\}, \end{aligned}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher random variables which are independent from $X_1, \ldots, X_n$.

Proof. Choose random variables $(X_1', Y_1'), \ldots, (X_n', Y_n')$, $\epsilon_1, \ldots, \epsilon_n$ such that

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_1', Y_1'), \ldots, (X_n', Y_n\prime), \epsilon_1, \ldots, \epsilon_n$$

are independent,

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_1', Y_1'), \ldots, (X_n', Y_n'),$$

are identically distributed and

$$\mathbf{P}\{\epsilon_i = 1\} = \mathbf{P}\{\epsilon_i = -1\} = \frac{1}{2} \quad (i = 1, \ldots, n).$$

We use the error decomposition

$$\mathbf{E}\Big\{ \sup_{\mathbf{w} \in \mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X) - Y|^2 - |m(X) - Y|^2\}$$

$$
\begin{aligned}
&\qquad\qquad -\frac{1}{n}\sum_{i=1}^{n}\left(|T_{\beta_n} f_{\mathbf{w}}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\Big)\Big\}\\
&= \mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X)-T_{\beta_n}Y|^2\}-\frac{1}{n}\sum_{i=1}^{n}|T_{\beta_n} f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2\Big)\Big\}\\
&\quad +\mathbf{E}\Big\{-\mathbf{E}\{|m(X)-Y|^2\}+\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\Big)\Big\}\\
&\quad +\mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X)-Y|^2-|T_{\beta_n} f_{\mathbf{w}}(X)-T_{\beta_n}Y|^2\}\Big)\Big\}\\
&\quad +\mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\frac{1}{n}\sum_{i=1}^{n}\left(-|T_{\beta_n} f_{\mathbf{w}}(X_i)-Y_i|^2+|T_{\beta_n} f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2\right)\Big)\Big\}\\
&=: T_{1,n}+T_{2,n}+T_{3,n}+T_{4,n}.
\end{aligned}
$$

We have

$$
\begin{aligned}
T_{2,n} &\leq \sqrt{\mathbf{E}\Big\{\Big|-\mathbf{E}\{|m(X)-Y|^2\}+\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\Big)\Big|^2\Big\}}\\
&= \sqrt{\mathbf{Var}\Big\{\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\Big\}}\\
&\leq \frac{c_{76}}{\sqrt{n}},
\end{aligned}
$$

and as in the proof of Lemma 1 in Bauer and Kohler (2019) we get

$$
T_{3,n}+T_{4,n}\leq c_{77}\cdot\frac{(\log n)^2}{n}.
$$

Hence it suffices to show

$$
T_{1,n}\leq 8\cdot\beta_n\cdot\gamma_n\cdot\mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta},k\in\{1,\ldots,K_n\}}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot f_{\mathbf{w},k,1}^{(L_n)}(X_i)\Big|\Big\}. \tag{32}
$$

We have

$$
\begin{aligned}
&\mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X)-T_{\beta_n}Y|^2\}-\frac{1}{n}\sum_{i=1}^{n}|T_{\beta_n} f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2\Big)\Big\}\\
&= \mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\mathbf{E}\{\frac{1}{n}\sum_{i=1}^{n}|T_{\beta_n} f_{\mathbf{w}}(X_i')-T_{\beta_n}Y_i'|^2\big|(X_1,Y_1),\ldots(X_n,Y_n)\}\\
&\qquad\qquad -\frac{1}{n}\sum_{i=1}^{n}|T_{\beta_n} f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2\Big)\Big\}\\
&= \mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\mathbf{E}\{\Big(\frac{1}{n}\sum_{i=1}^{n}|T_{\beta_n} f_{\mathbf{w}}(X_i')-T_{\beta_n}Y_i'|^2
\end{aligned}
$$

$$
\begin{aligned}
&\qquad\qquad -\frac{1}{n}\sum_{i=1}^{n}|T_{\beta_n}f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2\Big)\big|(X_1,Y_1),\dots(X_n,Y_n)\big\}\Big\}\\
&\le \mathbf{E}\Big\{\mathbf{E}\big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\frac{1}{n}\sum_{i=1}^{n}|T_{\beta_n}f_{\mathbf{w}}(X_i')-T_{\beta_n}Y_i'|^2\\
&\qquad\qquad -\frac{1}{n}\sum_{i=1}^{n}|T_{\beta_n}f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2\Big)\big|(X_1,Y_1),\dots(X_n,Y_n)\big\}\Big\}\\
&= \mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\frac{1}{n}\sum_{i=1}^{n}|T_{\beta_n}f_{\mathbf{w}}(X_i')-T_{\beta_n}Y_i'|^2-\frac{1}{n}\sum_{i=1}^{n}|T_{\beta_n}f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2\Big)\Big\}.
\end{aligned}
$$

The joint distribution of $(X_1,Y_1),\dots,(X_n,Y_n),(X_1',Y_1'),\dots,(X_n',Y_n')$ does not change if one (randomly) interchanges components of $(X_1,Y_1),\dots,(X_n,Y_n)$ and $(X_1',Y_1'),\dots,(X_n',Y_n')$. Consequently the right hand-side above is equal to

$$
\begin{aligned}
&\mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot\big(|T_{\beta_n}f_{\mathbf{w}}(X_i')-T_{\beta_n}Y_i'|^2-|T_{\beta_n}f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2\big)\Big)\Big\}\\
&\le \mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot|T_{\beta_n}f_{\mathbf{w}}(X_i')-T_{\beta_n}Y_i'|^2\Big)\Big\}\\
&\quad+\mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\frac{1}{n}\sum_{i=1}^{n}(-\epsilon_i)\cdot|T_{\beta_n}f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2\Big)\Big\}\\
&= 2\cdot\mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot|T_{\beta_n}f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2\Big)\Big\}.
\end{aligned}
$$

Next we use a contraction style argument. Due to the independence of the data we can compute the expectation above in such a way that we first compute the expectation with respect to ϵ_1 and then with respect to all other random variables. This implies that the right-hand side above is equal to

$$
\begin{aligned}
&2\cdot\mathbf{E}\Big\{\frac{1}{2}\cdot\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot|T_{\beta_n}f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2+\frac{1}{n}\cdot1\cdot|T_{\beta_n}f_{\mathbf{w}}(X_1)-T_{\beta_n}Y_1|^2\Big)\\
&\qquad+\frac{1}{2}\cdot\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot|T_{\beta_n}f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2+\frac{1}{n}\cdot(-1)\cdot|T_{\beta_n}f_{\mathbf{w}}(X_1)-T_{\beta_n}Y_1|^2\Big)\Big\}\\
&= \mathbf{E}\Big\{\sup_{\mathbf{w},\bar{\mathbf{w}}\in\mathbf{\Theta}}\Big(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot|T_{\beta_n}f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2+\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot|T_{\beta_n}f_{\bar{\mathbf{w}}}(X_i)-T_{\beta_n}Y_i|^2\\
&\qquad\qquad+\frac{1}{n}\cdot|T_{\beta_n}f_{\mathbf{w}}(X_1)-T_{\beta_n}Y_1|^2-\frac{1}{n}\cdot|T_{\beta_n}f_{\bar{\mathbf{w}}}(X_1)-T_{\beta_n}Y_1|^2\Big)\Big\}\\
&\le \mathbf{E}\Big\{\sup_{\mathbf{w},\bar{\mathbf{w}}\in\mathbf{\Theta}}\Big(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot|T_{\beta_n}f_{\mathbf{w}}(X_i)-T_{\beta_n}Y_i|^2+\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot|T_{\beta_n}f_{\bar{\mathbf{w}}}(X_i)-T_{\beta_n}Y_i|^2\\
&\qquad\qquad+\frac{1}{n}\cdot4\cdot\beta_n\cdot|T_{\beta_n}f_{\mathbf{w}}(X_1)-T_{\beta_n}f_{\bar{\mathbf{w}}}(X_1)|\Big)\Big\}
\end{aligned}
$$

$$\leq \mathbf{E}\Big\{ \sup_{\mathbf{w},\bar{\mathbf{w}}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\bar{\mathbf{w}}}(X_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \cdot 4 \cdot \beta_n \cdot |f_{\mathbf{w}}(X_1) - f_{\bar{\mathbf{w}}}(X_1)|\Big)\Big\}.$$

The term in the above supremum is for fixed X_i, Y_i, ϵ_i symmetric in $\mathbf{w}$ and $\bar{\mathbf{w}}$, hence

$$\begin{aligned}
&\sup_{\mathbf{w},\bar{\mathbf{w}}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\bar{\mathbf{w}}}(X_i) - T_{\beta_n} Y_i|^2 \\
&\qquad + \frac{1}{n} \cdot 4 \cdot \beta_n \cdot |f_{\mathbf{w}}(X_1) - f_{\bar{\mathbf{w}}}(X_1)|\Big) \\
&= \sup_{\mathbf{w},\bar{\mathbf{w}}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\bar{\mathbf{w}}}(X_i) - T_{\beta_n} Y_i|^2 \\
&\qquad + \frac{1}{n} \cdot 4 \cdot \beta_n \cdot 1 \cdot (f_{\mathbf{w}}(X_1) - f_{\bar{\mathbf{w}}}(X_1))\Big) \\
&= \sup_{\mathbf{w},\bar{\mathbf{w}}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\bar{\mathbf{w}}}(X_i) - T_{\beta_n} Y_i|^2 \\
&\qquad + \frac{1}{n} \cdot 4 \cdot \beta_n \cdot (-1) \cdot (f_{\mathbf{w}}(X_1) - f_{\bar{\mathbf{w}}}(X_1))\Big)
\end{aligned}$$

This yields

$$\begin{aligned}
&\mathbf{E}\Big\{ \sup_{\mathbf{w},\bar{\mathbf{w}}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\bar{\mathbf{w}}}(X_i) - T_{\beta_n} Y_i|^2 \\
&\qquad + \frac{1}{n} \cdot 4 \cdot \beta_n \cdot |f_{\mathbf{w}}(X_1) - f_{\bar{\mathbf{w}}}(X_1)|\Big)\Big\} \\
&= \mathbf{E}\Big\{ \frac{1}{2} \cdot \sup_{\mathbf{w},\bar{\mathbf{w}}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\bar{\mathbf{w}}}(X_i) - T_{\beta_n} Y_i|^2 \\
&\qquad + 4 \cdot \beta_n \cdot \frac{1}{n} \cdot 1 \cdot (f_{\mathbf{w}}(X_1) - f_{\bar{\mathbf{w}}}(X_1))\Big) \\
&\qquad \frac{1}{2} \cdot \sup_{\mathbf{w},\bar{\mathbf{w}}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\bar{\mathbf{w}}}(X_i) - T_{\beta_n} Y_i|^2 \\
&\qquad + 4 \cdot \beta_n \cdot \frac{1}{n} \cdot (-1) \cdot (f_{\mathbf{w}}(X_1) - f_{\bar{\mathbf{w}}}(X_1))\Big)\Big\} \\
&= \mathbf{E}\Big\{ \sup_{\mathbf{w},\bar{\mathbf{w}}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\bar{\mathbf{w}}}(X_i) - T_{\beta_n} Y_i|^2 \\
&\qquad + 4 \cdot \beta_n \cdot \frac{1}{n} \cdot \epsilon_1 \cdot (f_{\mathbf{w}}(X_1) - f_{\bar{\mathbf{w}}}(X_1))\Big)\Big\} \\
&\leq \mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 + 4 \cdot \beta_n \cdot \frac{1}{n} \cdot \epsilon_1 \cdot (f_{\mathbf{w}}(X_1))\Big)\Big\}
\end{aligned}$$

$$
\begin{aligned}
&+\mathbf{E}\Big\{ \sup_{\bar{\mathbf{w}}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\bar{\mathbf{w}}}(X_i) - T_{\beta_n} Y_i|^2 + 4\cdot\beta_n \cdot \frac{1}{n}\cdot(-\epsilon_1)\cdot f_{\bar{\mathbf{w}}}(X_1)\Big)\Big\} \\
&= 2\cdot\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=2}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 + 4\cdot\beta_n \cdot \frac{1}{n}\cdot\epsilon_1\cdot f_{\mathbf{w}}(X_1)\Big)\Big\} \\
&= \dots \\
&\le 2\cdot\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big(\frac{1}{n}\sum_{i=3}^{n} \epsilon_i \cdot |T_{\beta_n} f_{\mathbf{w}}(X_i) - T_{\beta_n} Y_i|^2 + 4\cdot\beta_n \cdot \frac{1}{n}\cdot\epsilon_2\cdot f_{\mathbf{w}}(X_2) \\
&\qquad\qquad +4\cdot\beta_n\cdot\frac{1}{n}\cdot\epsilon_1\cdot f_{\mathbf{w}}(X_1)\Big)\Big\} \\
&\le \dots \\
&\le 2\cdot\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big(\frac{4\cdot\beta_n}{n}\sum_{i=1}^{n} \epsilon_i \cdot f_{\mathbf{w}}(X_i)\Big)\Big\}.
\end{aligned}
$$

By the definition of $f_{\mathbf{w}}$ and (31) the right-hand side above is bounded by

$$
\begin{aligned}
&2\cdot\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big|\frac{4\cdot\beta_n}{n}\sum_{i=1}^{n} \epsilon_i \cdot \sum_{k=1}^{K_n} w_{k,1,1}^{(L_n)} \cdot f_{\mathbf{w},k,1}^{(L_n)}(X_i)\Big|\Big\} \\
&= 2\cdot\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big|\sum_{k=1}^{K_n} w_{k,1,1}^{(L_n)} \cdot \frac{4\cdot\beta_n}{n}\sum_{i=1}^{n} \epsilon_i \cdot f_{\mathbf{w},k,1}^{(L_n)}(X_i)\Big|\Big\} \\
&\le 2\cdot\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \sum_{k=1}^{K_n} |w_{k,1,1}^{(L_n)}| \cdot \Big|\frac{4\cdot\beta_n}{n}\sum_{i=1}^{n} \epsilon_i \cdot f_{\mathbf{w},k,1}^{(L_n)}(X_i)\Big|\Big\} \\
&\le 2\cdot\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \gamma_n \cdot \Big|\frac{4\cdot\beta_n}{n}\sum_{i=1}^{n} \epsilon_i \cdot f_{\mathbf{w},k,1}^{(L_n)}(X_i)\Big|\Big\} \\
&= 8\cdot\beta_n\cdot\gamma_n\cdot\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot f_{\mathbf{w},k,1}^{(L_n)}(X_i)\Big|\Big\}.
\end{aligned}
$$

□

4.3.2 Proof of Theorem 2

The regression function is given by

$$
m(x) = \sum_{k=1}^{K} m_k(b_k^t x) \quad (x\in\mathbb{R}^d)
$$

where $K\in\mathbb{N}$, $b_k\in\mathbb{R}^d$ and $m_k : \mathbb{R}\to\mathbb{R}$ $(k=1,\dots,K)$ are (p,C)–smooth functions. Let $p = q+s$ for some $q\in\mathbb{N}$ and $s\in(0,1]$.

We start the proof by constructing neural networks which approximate the functions m_k $(k=1,\dots,K)$. To do this we use Lemma 6. Let

$$
A = \sup\{b_k^t x \,:\, x\in \mathit{supp}(X),\, k\in\{1,\dots,K\}\}.
$$

Set

$$M_n = \lceil c_{78} \cdot n^{\frac{1}{2 \cdot (2p+1)}} \rceil,$$

$$\eta_n = \frac{1}{n^2}$$

and let ν_k be the measure defined by

$$\nu_k(B) = \mathbf{P}\{b_k^t X \in B\}$$

for a Borel set $B \subseteq \mathbb{R}$.

By Lemma 6 (applied with $d = 1$, $N = q$ and $M = M_n$) there exists a neural network

$$g_{k,\mathbf{w}_k}(x) = \sum_{i=1}^{(q+1)^2 \cdot (M_n+1)} \mu_i \cdot \sigma\left(\sum_{l=1}^{4} \lambda_{i,l} \cdot \sigma\left(\theta_{i,l,1} \cdot x + \theta_{i,l,0} \right) + \lambda_{i,0} \right)$$

which approximates m_k on $[-A, A]$ outside of a set, which has ν_k-measure at most η_n, with an error bounded in absolute value from above by

$$c_{79} \cdot M_n^{-p} = c_{80} \cdot n^{-\frac{p}{2 \cdot (2p+1)}},$$

where all its weights are bounded in absolute value by

$$c_{81} \cdot \frac{1}{\eta_n} \cdot M_n^{p \cdot (2p+3)+2} \leq c_{82} \cdot n^{p+4}.$$

It follows from the proof of Theorem 2 in Bauer and Kohler (2019) that we can rewrite $g_{k,\mathbf{w}_k}$ in the form

$$g_{k,\mathbf{w}_k}(x) = \sum_{s=1}^{M_n+1} \sum_{i=1}^{(q+1)^2} \mu_{s,i} \cdot \sigma\left(\sum_{l=1}^{4} \lambda_{s,i,l} \cdot \sigma\left(\theta_{s,i,l,1} \cdot x + \theta_{s,i,l,0} \right) + \lambda_{s,i,0} \right)$$

such that the above approximation result still holds, such that the weights are bounded as before and such that in addition it holds

$$\left| \sum_{i=1}^{(q+1)^2} \mu_{s,i} \cdot \sigma\left(\sum_{l=1}^{4} \lambda_{s,i,l} \cdot \sigma\left(\theta_{s,i,l,1} \cdot x + \theta_{s,i,l,0} \right) + \lambda_{s,i,0} \right) \right| \leq c_{83} \cdot (\|m_k\|_\infty + 1) \leq c_{84}$$

for all $s = 1, \ldots, M_n + 1$ on $[-A, A]$ outside of the above set which has ν_k-measure at most η_n.

Set

$$R_n = c_{85} \cdot n^{\delta},$$

and let $N \in \mathbb{N}$ be such that

$$(N-1) \cdot \delta \geq 2.$$

Let f_{id} be the network from Lemma 7 which satisfies

$$|f_{id}(x) - x| \leq c_{86} \cdot \frac{c_{87}^N}{R_n^{N-1}} \leq c_{88} \cdot \frac{1}{n^2}$$

for $|x| \leq c_{84}$. We will approximate $m_k(b_k^t x)$ by

$$\sum_{s=1}^{M_n+1} f_{id}\left(\sum_{i=1}^{(q+1)^2} \mu_{s,i} \cdot \sigma\left(\sum_{l=1}^{4} \lambda_{s,i,l} \cdot \sigma\left(\theta_{s,i,l,1} \cdot b_k^t x + \theta_{s,i,l,0}\right) + \lambda_{s,i,0}\right)\right).$$

By construction we have

$$\begin{aligned}
&\left|\sum_{s=1}^{M_n+1} f_{id}\left(\sum_{i=1}^{(q+1)^2} \mu_{s,i} \cdot \sigma\left(\sum_{l=1}^{4} \lambda_{s,i,l} \cdot \sigma\left(\theta_{s,i,l,1} \cdot b_k^t x + \theta_{s,i,l,0}\right) + \lambda_{s,i,0}\right)\right) - m_k(b_k^t x)\right| \\
&\leq \frac{c_{89}}{n^2} \cdot (M_n + 1) \\
&\quad + \left|\sum_{s=1}^{M_n+1} \sum_{i=1}^{(q+1)^2} \mu_{s,i} \cdot \sigma\left(\sum_{l=1}^{4} \lambda_{s,i,l} \cdot \sigma\left(\theta_{s,i,l,1} \cdot b_k^t x + \theta_{s,i,l,0}\right) + \lambda_{s,i,0}\right) - m_k(b_k^t x)\right| \\
&\leq c_{90} \cdot n^{-\frac{p}{2\cdot(2p+1)}}
\end{aligned}$$

on $supp(X)$ outside of a set of $\mathbf{P}_X$ measure η_n and all weights of this network are bounded in absolute value by $c_{82} \cdot n^{p+4}$ and its outer weights are bounded in absolute value by R_n. So if we sum these networks up for $k = 1, \ldots, K$ we approximate m on $supp(X)$ outside of a set of $\mathbf{P}_X$ measure

$$K \cdot \eta_n \leq \frac{c_{91}}{n^2}$$

with a pointwise error of at most $c_{92} \cdot n^{-\frac{p}{2\cdot(2p+1)}}$, and the weights of the corresponding networks are bounded as above.

Set

$$N_n = n^{12p+33}$$

In order to get smaller outer weights we repeat this whole network N_n times with outer weights divided by N_n and sum the resulting N_n networks up which yields the same approximation result as above. In this way we construct a weight vector $\bar{\mathbf{w}}$ of a network

$$f_{\bar{\mathbf{w}}}(x) = \sum_{k=1}^{N_n \cdot K \cdot (M_n+1) \cdot N} \bar{\mathbf{w}}_{k,1,1}^{(3)} \cdot f_{\bar{\mathbf{w}},k,1}^{(3)}(x)$$

where each of the $f_{\bar{\mathbf{w}},k,1}^{(3)}$ is a network with $L = 3$ layers and at most

$$\max\{(q+1)^2, 4\}$$

neurons per hidden layer, and all weights bounded in absolute value by $c_{82} \cdot n^{p+4}$ and with outer weights bounded in absolute value by

$$c_{93} \cdot \frac{R_n}{N_n}.$$

Furthermore it satisfies

$$|f_{\bar{\mathbf{w}}}(x) - m(x)| \leq c_{94} \cdot n^{-\frac{p}{2\cdot(2p+1)}}$$

for $x \in supp(X)$ outside of a set of $\mathbf{P}_X$ measure c_{95}/n^2.

Set

$$\epsilon_n = \frac{1}{n^{6p+11}}$$

and let E_n be the event that there exist pairwise distinct $j_1, \ldots, j_{N_n \cdot K \cdot (M_n+1) \cdot N} \in \{1, \ldots, K_n\}$ such that

$$\left\| ((\mathbf{w}^{(0)})^{(l)}_{j_k,i,j})_{i,j,l:l<L} - (\bar{\mathbf{w}}^{(l)}_{k,i,j})_{i,j,l:l<L} \right\|_\infty \leq \epsilon_n$$

holds for all $k \in \{1, \ldots, N_n \cdot K \cdot (M_n + 1) \cdot N\}$.

If E_n holds, then set

$$(\mathbf{w}^*)^{(L)}_{j_k,1,1} = (\bar{\mathbf{w}})^{(L)}_{k,1,1} \quad \text{for } k = 1, \ldots, N_n \cdot K \cdot (M_n + 1) \cdot N,$$

and set

$$(\mathbf{w}^*)^{(L)}_{j_k,1,1} = 0 \quad \text{for } k \in \{1, \ldots, K_n\} \setminus \{j_1, \ldots, j_{N_n \cdot K \cdot (M_n+1) \cdot N}\}.$$

If E_n does not hold, then set $\mathbf{w}^* = 0$.

Then

$$\sum_{k=1}^{K_n} |(\mathbf{w}^*)^{(L)}_{k,1,1}| \leq c_{96} \cdot n^{\frac{1}{2\cdot(2p+1)}+\delta} \leq \gamma_n$$

and

$$\sum_{k=1}^{K_n} |(\mathbf{w}^*)^{(L)}_{k,1,1}|^2 \leq N_n \cdot K \cdot (M_n + 1) \cdot N \cdot \left(\frac{c_{97} \cdot n^\delta}{N_n} \right)^2 \leq \alpha_n$$

hold.

By Lemma 3 we know that assumption (21) of Theorem 4 is satisfied for

$$C_n = c_{98} \cdot \left(n^{p+4}\right)^{L-1} \leq c_{99} \cdot n^{2p+8}.$$

Furthermore, on the event $\{\max_{i=1,\ldots,n} |Y_i| \leq \beta_n\}$

$$\left\| \nabla_{(\mathbf{w}^{(L_n)}_{k,1,1})_k} (Y_{j_t} - f_{\mathbf{w}^{(t)}}(X_{j_t}))^2 \right\|_\infty \leq 2 \cdot (\beta_n + \gamma_n) \leq c_{100} \cdot n^{\frac{1}{2p+1}+\delta}$$

holds, hence assumption (22) of Theorem 4 is satisfied for

$$D_n = c_{101} \cdot n^{\frac{1}{2p+1}+\delta}.$$

Application of Theorem 4 yields

$$\begin{aligned}
&\mathbf{E}\int |m_n(x)-m(x)|^2\mathbf{P}_X(dx)\\
&\le c_{102}\cdot\left(\frac{\beta_n^2}{n^5}+\beta_n^2\cdot\sqrt{\mathbf{P}(E_n^c)}+\mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\mathbf{E}\{|T_{\beta_n}f_{\mathbf{w}}(X)-Y|^2-|m(X)-Y|^2\}\right.\\
&\qquad\qquad\left.-\frac{2}{n}\sum_{i=1}^n\big(|T_{\beta_n}f_{\mathbf{w}}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\big)\Big)\Big\}\right.\\
&\left.+\mathbf{E}\left\{\int |f_{(((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k,((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x)-m(x)|^2\mathbf{P}_X(dx)\cdot 1_{E_n}\right\}+\frac{1}{n}\right).
\end{aligned}$$

Hence it suffices to show

$$\mathbf{P}(E_n^c)\le\frac{c_{103}}{n}, \tag{33}$$

$$\begin{aligned}
&\mathbf{E}\Big\{\sup_{\mathbf{w}\in\mathbf{\Theta}}\Big(\mathbf{E}\{|T_{\beta_n}f_{\mathbf{w}}(X)-Y|^2-|m(X)-Y|^2\}\\
&\qquad\qquad-\frac{2}{n}\sum_{i=1}^n\big(|T_{\beta_n}f_{\mathbf{w}}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\big)\Big)\Big\}\\
&\le c_{104}\cdot(\log n)^{3/2}\cdot n^{-\frac{p}{2p+1}+\delta}
\end{aligned} \tag{34}$$

and

$$\mathbf{E}\left\{\int |f_{(((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k,((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x)-m(x)|^2\mathbf{P}_X(dx)\cdot 1_{E_n}\right\}\le c_{105}\cdot n^{-\frac{p}{2p+1}}. \tag{35}$$

Proof of (33): In order to bound $\mathbf{P}(E_n^c)$ we consider a sequential choice of the weights of the K_n networks

$$f_{\mathbf{w}^{(0)},k,1}^{(L)}\quad(k=1,\dots,K_n).$$

The probability that none of the

$$r\cdot(d+1)+(L-2)\cdot r\cdot(r+1)+(r+1)\le 3((p+1)^2+d+3)^2$$

weights of the first of these networks differs from the corresponding weight in

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

by more than ϵ_n is for large n bounded from above by

$$\left(\frac{\epsilon_n}{2\cdot c_{106}\cdot n^{p+5}}\right)^{3((p+1)^2+d+3)^2}\le n^{-(21p+48)\cdot((p+1)^2+d+3)^2-0.5}.$$

Hence we get that the probability that none of the

$$\lceil n^{(21p+48)\cdot((p+1)^2+d+3)^2+1}\rceil$$

many networks

$$f^{(L)}_{\mathbf{w}^{(0)},k,1} \quad (k = 1, \ldots, \lceil n^{(21p+48)\cdot((p+1)^2+d+3)^2+1} \rceil)$$

differ in all weights from the corresponding weight in

$$f^{(L)}_{\tilde{\mathbf{w}},1,1}$$

by at most ϵ_n is bounded from above by

$$\left(1 - n^{-(21p+48)\cdot((p+1)^2+d+3)^2-0.5}\right)^{n^{(21p+48)\cdot((p+1)^2+d+3)^2+1}} \leq e^{-0.5\cdot n}.$$

Since we have for n large

$$N_n \cdot K \cdot (M_n + 1) \cdot N \cdot \lceil n^{(21p+48)\cdot((p+1)^2+d+3)^2+1} \rceil \leq K_n,$$

we can conclude: The probability that for any $k \in \{1, \ldots, N_n \cdot K \cdot (M_n + 1) \cdot N\}$ in all of the networks

$$f^{(L)}_{\mathbf{w}^{(0)},j,1} \quad (j = (k-1) \cdot \lceil n^{(21p+48)\cdot((p+1)^2+d+3)^2+1} \rceil + 1, \ldots, k \cdot \lceil n^{(21p+48)\cdot((p+1)^2+d+3)^2+1} \rceil$$

at least one of the weights differs from the corresponding weight in

$$f^{(L)}_{\tilde{\mathbf{w}},k,1}$$

by more than ϵ_n is bounded from above by

$$N_n \cdot K \cdot (M_n + 1) \cdot N \cdot e^{-0.5\cdot n} \leq \frac{c_{107}}{n}.$$

Since $\mathbf{P}\{E_n^c\}$ is upper bounded by this probability, this implies (33).
Proof of (34): By Lemma 8 we have

$$\begin{aligned}
&\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X) - Y|^2 - |m(X) - Y|^2\} \\
&\qquad\qquad -\frac{1}{n}\sum_{i=1}^{n} \left(|T_{\beta_n} f_{\mathbf{w}}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \Big) \Big\} \\
&\leq \frac{c_{108}}{\sqrt{n}} + c_{109} \cdot \beta_n \cdot c_{110} \cdot n^{\frac{1}{2\cdot(2p+1)}+\delta} \cdot \mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta}, k\in\{1,\ldots,K_n\}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot f^{(L)}_{\mathbf{w},k,1}(X_i) \right| \Big\}.
\end{aligned}$$

By discretizing the

$$r \cdot (d+1) + (L-2) \cdot r \cdot (r+1) + (r+1) \leq 3((p+1)^2 + d + 3)^2$$

many coefficients in each of the classes

$$\{f^{(L)}_{\mathbf{w},k,1} \quad : \quad \mathbf{w} \in \mathbf{\Theta}\} \quad (k = 1, \ldots, K_n)$$

on a grid of length

$$\Delta_n = 2 \cdot c_{111} \cdot n^{p+5}$$

and grid size less than or equal to

$$\frac{1/n}{c_{112} \cdot n^{2p+10}},$$

which leads for all k to the same set of functions, we get (by Lemma 2) a $1/n$-supremum norm cover of

$$\{f^{(L)}_{\mathbf{w},k,1} \quad : \quad \mathbf{w} \in \mathbf{\Theta},\, k \in \{1, \ldots, K_n\}\}$$

of size

$$\left(\frac{\Delta_n}{\frac{1/n}{c_{113} \cdot n^{2p+10}}} \right)^{3((p+1)^2+d+3)^2} \leq c_{114} \cdot n^{(9p+48)\cdot((p+1)^2+d+3)^2}.$$

Together with the union bound and the inequality of Hoeffding (cf., e.g., Lemma A.3 in Györfi et al. (2002)) this implies for $\delta_n \geq 2/n$

$$\begin{aligned}
&\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta},k\in\{1,\ldots,K_n\}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot f_{\mathbf{w},k,1}(X_i) \right| \Big\} \\
&\leq \delta_n + \int_{\delta_n}^{\infty} \mathbf{P}\left\{ \sup_{\mathbf{w}\in\mathbf{\Theta},k\in\{1,\ldots,K_n\}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot f_{\mathbf{w},k,1}(X_i) \right| > t \right\} dt \\
&\leq \delta_n + c_{114} \cdot n^{(9p+48)\cdot((p+1)^2+d+3)^2} \cdot \int_{\delta_n}^{\infty} 2 \cdot e^{-\frac{n\cdot(t/2)^2}{4}} dt \\
&\leq \delta_n + c_{114} \cdot n^{(9p+48)\cdot((p+1)^2+d+3)^2} \cdot \int_{\delta_n}^{\infty} 2 \cdot e^{-\frac{n\cdot t\cdot\delta_n}{8}} dt \\
&\leq \delta_n + c_{114} \cdot n^{(9p+48)\cdot((p+1)^2+d+3)^2} \cdot \frac{16}{n\cdot\delta_n} \cdot e^{-\frac{n\cdot\delta_n^2}{8}}.
\end{aligned}$$

With

$$\delta_n = \sqrt{\frac{16}{n} \cdot (9p+48) \cdot ((p+1)^2 + d + 3)^2 \cdot \log n}$$

we get

$$\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta},k\in\{1,\ldots,K_n\}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot f_{\mathbf{w},k,1}(X_i) \Big\} \leq c_{115} \cdot \sqrt{\log n} \cdot \frac{1}{\sqrt{n}},$$

which implies (34).

Proof of (35): The definition of $((\mathbf{w}^*)^{(L)}_{k,1,1})_k$ implies

$$\begin{aligned}
&|f_{(((\mathbf{w}^*)^{(Ln)}_{k,1,1})_k,((\mathbf{w}^{(0)})^{(l)}_{k,i,j})_{k,i,j,l:l<Ln})}(x) - m(x)|^2 \\
&= \Big| \sum_{k=1}^{N_n\cdot K\cdot(M_n+1)\cdot N} \bar{\mathbf{w}}^{(L)}_{k,1,1} \cdot f^{(L)}_{\mathbf{w}^{(0)},j_k,1}(x) - m(x) \Big|^2
\end{aligned}$$

$$\leq 2 \cdot | \sum_{k=1}^{N_n \cdot K \cdot (M_n+1) \cdot N} \bar{\mathbf{w}}_{k,1,1}^{(L)} \cdot f_{\mathbf{w}^{(0)},j_k,1}^{(L)}(x) - f_{\bar{\mathbf{w}}}(x)|^2 + 2 \cdot |f_{\bar{\mathbf{w}}}(x) - m(x)|^2$$
$$\leq 2 \cdot | \sum_{k=1}^{N_n \cdot K \cdot (M_n+1) \cdot N} \bar{\mathbf{w}}_{k,1,1}^{(L)} \cdot f_{\mathbf{w}^{(0)},j_k,1}^{(L)}(x) - f_{\bar{\mathbf{w}}}(x)|^2 + 2 \cdot \left(c_{116} \cdot \max_{i,j} n^{\frac{-p}{2 \cdot (2p+1)}} \right)^2,$$

where the last inequality holds outside of a set of $\mathbf{P}_X$ measure $1/n^2$. On E_n we get by Lemma 2 that this in turn is bounded from above by

$$\begin{aligned}
&2 \cdot \sum_{k=1}^{N_n \cdot K \cdot (M_n+1) \cdot N} |\bar{\mathbf{w}}_{k,1,1}^{(L)}|^2 \cdot \sum_{k=1}^{N_n \cdot K \cdot (M_n+1) \cdot N} |f_{\mathbf{w}^{(0)},j_k,1}^{(L)}(x) - f_{\bar{\mathbf{w}},k,1}^{(L)}(x)|^2 \\
&\quad + 2 \cdot \left(c_{116} \cdot n^{-\frac{p}{2 \cdot (2p+1)}} \right)^2 \\
&\leq 2 \cdot \alpha_n \cdot N_n \cdot K \cdot (M_n + 1) \cdot N \cdot c_{117} \cdot n^{4p+16} \cdot \epsilon_n^2 + 2 \cdot \left(c_{116} \cdot n^{-\frac{p}{2 \cdot (2p+1)}} \right)^2 \\
&\leq 2 c_{117} \cdot K \cdot N \cdot n^{12p+20} \cdot \epsilon_n^2 + 2 \cdot \left(c_{116} \cdot n^{-\frac{p}{2 \cdot (2p+1)}} \right)^2 \\
&\leq c_{118} \cdot n^{\frac{-p}{2p+1}}.
\end{aligned}$$

Since

$$|f_{(((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k,((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x)| \leq \gamma_n = c_{11} \cdot n^{\frac{1}{2 \cdot (2p+1)}+\delta}$$

we can conclude

$$\begin{aligned}
&\mathbf{E} \left\{ \int |f_{(((\mathbf{w}^*)_{k,1,1}^{(L_n)})_k,((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L_n})}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{E_n} \right\} \\
&\leq c_{118} \cdot n^{-\frac{p}{2p+1}} + c_{11} \cdot n^{\frac{1}{2 \cdot (2p+1)}+\delta} \cdot \frac{1}{n^2} \\
&\leq c_{105} \cdot n^{-\frac{p}{2p+1}}.
\end{aligned}$$

□

4.4 Proof of Theorem 3

4.4.1 Auxiliary results

In the proof of Theorem 3 we will apply the following result in order to bound the approximation error of the estimate.

Lemma 9 *Let $m : \mathbb{R}^d \to \mathbb{R}$ be contained in the class $\mathcal{H}(l, \mathcal{P})$ for some $l \in \mathbb{N}$ and $\mathcal{P} \subseteq [1, \infty) \times \mathbb{N}$. Describe m as in Theorem 3, and assume that the functions $g_j^{(i)}$ are Lipschitz continuous with Lipschitz constant $C_{Lip} \geq 1$ and satisfy*

$$\|g^{(I)}\|_{C^{q_j^{(i)}}([-a,a]^d)} \leq c_{119}$$

for some $c_{119} > 0$. *Denote by* $K_{max} = \max_{i,j} K_j^{(i)}$ *the maximal input dimension and by* $p_{max} = \max_{i,j} p_j^{(i)}$ *the maximal smoothness of the functions* $g_j^{(i)}$. *Then, for any* $a \geq 1$ *and* $M_{j,i} \in \mathbb{N}$ *sufficiently large a neural network* $f_{\mathbf{w}}$ *with logistic activation function and*

$$L = l \cdot (8 + \lceil \log_2(\max\{K_{max}, p_{max} + 1\}) \rceil)$$

layers and

$$r = \max_{i \in \{1,\dots,l\}} \sum_{j=1}^{\tilde{N}_i} 29 \binom{K_j^{(i)} + q_j^{(i)}}{q_j^{(i)}} \cdot (K_j^{(i)})^2 \cdot (q_j^{(i)} + 1) \cdot M_{j,i}^{K_j^{(i)}}$$

neurons per layers, where the weight vector $\mathbf{w}$ *satisfies*

$$\|\mathbf{w}\|_\infty \leq c_{120} \cdot a^{24} \cdot e^{12 \cdot 2^{2(K_{max}+1)+1} \cdot a \cdot K_{max}} \cdot max_{j,i} M_{j,i}^{20 p_{max} + 4K_{max} + 20},$$

exists such that

$$\|f_{\mathbf{w}} - m\|_{\infty,[-a,a]^d} \leq c_{121} \cdot a^{5 \cdot p_{max} + 3} \cdot \max_{i,j} M_{j,i}^{-2 \cdot p_j^{(i)}}.$$

Proof. See Theorem 2 in Langer (2021). □

4.4.2 Proof of Theorem 3

Set

$$M_{i,j} = \left\lceil c_{122} \cdot n^{\frac{1}{2 \cdot (2p_j^{(i)} + K_j^{(i)})}} \right\rceil,$$

$$R_n = c_{123} \cdot n^{\delta}$$

and choose $N \in \mathbb{N}$ so large that

$$(N-1) \cdot \delta \geq \frac{1}{2}.$$

Let $g_{\mathbf{w}}$ be the neural network from Lemma 9 with

$$\bar{L} = l \cdot (8 + \lceil \log_2(\max\{K_{max}, p_{max} + 1\}) \rceil)$$

layers and

$$r = \max_{i \in \{1,\dots,l\}} \sum_{j=1}^{\tilde{N}_i} 29 \binom{K_j^{(i)} + q_j^{(i)}}{q_j^{(i)}} \cdot (K_j^{(i)})^2 \cdot (q_j^{(i)} + 1) \cdot M_{j,i}^{K_j^{(i)}} \leq \left\lceil c_{16} \cdot \max_{i,j} n^{\frac{K_j^{(i)}}{2 \cdot (2p_j^{(i)} + K_j^{(i)})}} \right\rceil = r_n$$

neurons per layers, where all the weights are bounded in absolute value by

$$c_{125} \cdot \max_{j,i} M_{j,i}^{20 p_{max} + 4K_{max} + 20} \leq c_{126} \cdot n^{5 p_{max} + K_{max} + 5}, \tag{36}$$

which satisfies

$$\|g_{\mathbf{w}} - m\|_{\infty,supp(X)} \leq c_{127} \cdot \max_{i,j} n^{\frac{-p_j^{(i)}}{2p_j^{(i)}+K_j^{(i)}}}.$$

Let f_{id} be the network from Lemma 7 where the N outer weights are bounded in absolute value by

$$c_{128} \cdot R_n = c_{129} \cdot n^{\delta}, \tag{37}$$

which satisfies

$$|f_{id}(x) - x| \leq c_{130} \cdot \frac{1}{R_n^{N-1}} = c_{131} \cdot n^{-\delta\cdot(N-1)} \leq c_{131} \cdot n^{-1/2}$$

for all $x \in [-\|m\|_\infty - 1, \|m\|_\infty + 1]$.

Then the neural network

$$f_{id} \circ g_{\mathbf{w}}$$

has $L = \bar{L} + 1$ layers with (for n large enough) at most r_n neurons, all its weights are bounded in absolute value by (36) and its outer weights are bounded in absolute value by (37). Furthermore, for n large enough we have

$$\|g_{\mathbf{w}} - m\|_{\infty,supp(X)} \leq 1,$$

which implies

$$\begin{aligned}
\|f_{id} \circ g_{\mathbf{w}} - m\|_{\infty,supp(X)} &\leq \|f_{id} \circ g_{\mathbf{w}} - g_{\mathbf{w}}\|_{\infty,supp(X)} + \|g_{\mathbf{w}} - m\|_{\infty,supp(X)} \\
&\leq c_{131} \cdot n^{-1/2} + c_{132} \cdot \max_{i,j} n^{\frac{-p_j^{(i)}}{2p_j^{(i)}+K_j^{(i)}}} \\
&\leq c_{133} \cdot \max_{i,j} n^{\frac{-p_j^{(i)}}{2p_j^{(i)}+K_j^{(i)}}}.
\end{aligned}$$

Set

$$N_n = n^{2L\cdot(5p_{max}+K_{max}+8)}$$

and let

$$f_{\bar{\mathbf{w}}}(x) = \sum_{k=1}^{N\cdot N_n} \bar{\mathbf{w}}_{k,1,1}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x)$$

be a linear combination of N_n of the neural networks $f_{id} \circ g_{\mathbf{w}}$ computed in parallel (with different weights), where the weights in the linear combination of these networks are all equal to $1/N_n$. Consequently,

$$|\bar{\mathbf{w}}_{k,1,1}^{(L)}| \leq \frac{c_{129} \cdot n^{\delta}}{N_n} \quad (k = 1, \ldots, N \cdot N_n),$$

$$\sum_{k=1}^{N\cdot N_n} |\bar{\mathbf{w}}_{k,1,1}^{(L)}| \leq c_{129} \cdot n^{\delta}$$

for some $c_{129} = c_{129}(\delta)$ and

$$\|f_{\bar{\mathbf{w}}} - m\|_{\infty, supp(X)} = \|f_{id} \circ g_{\mathbf{w}} - m\|_{\infty, supp(X)} \leq c_{133} \cdot \max_{i,j} n^{\frac{-p_j^{(i)}}{2p_j^{(i)} + K_j^{(i)}}}$$

hold.

Set

$$\epsilon_n = \frac{1}{n^{L/2 + (10p_{max} + 2K_{max} + 10) \cdot L + 3}}$$

and let E_n be the event that there exist pairwise distinct $j_1, \ldots, j_{N \cdot N_n} \in \{1, \ldots, K_n\}$ such that

$$\left\| ((\mathbf{w}^{(0)})_{j_k,i,j}^{(l)})_{i,j,l:l<L} - (\bar{\mathbf{w}}_{k,i,j}^{(l)})_{k,i,j,l:l<L} \right\|_\infty \leq \epsilon_n$$

holds for all $k \in \{1, \ldots, N \cdot N_n\}$.

If E_n holds, then set

$$(\mathbf{w}^*)_{j_k,1,1}^{(L)} = (\bar{\mathbf{w}})_{k,1,1}^{(L)} \quad \text{for } k = 1, \ldots, N \cdot N_n,$$

and set

$$(\mathbf{w}^*)_{j_k,1,1}^{(L)} = 0 \quad \text{for } k \in \{1, \ldots, K_n\} \setminus \{j_1, \ldots, j_{N \cdot N_n}\}.$$

If E_n does not hold, then set $\mathbf{w}^* = 0$.

Then

$$\sum_{k=1}^{K_n} |(\mathbf{w}^*)_{k,1,1}^{(L)}| \leq c_{129} \cdot n^\delta \leq \gamma_n$$

and

$$\sum_{k=1}^{K_n} |(\mathbf{w}^*)_{k,1,1}^{(L)}|^2 \leq N \cdot N_n \cdot \left(\frac{c_{129} \cdot n^\delta}{N_n} \right)^2 = N \cdot c_{129}^2 \cdot \frac{n^{2\delta}}{N_n} \leq \alpha_n$$

hold.

By Lemma 3 we know that assumption (21) of Theorem 4 is satisfied for

$$C_n = c_{133} \cdot r_n^L \cdot \left(n^{5p_{max} + K_{max} + 5} \right)^{L-1} \leq c_{134} \cdot n^{L \cdot (5p_{max} + K_{max} + 5)}.$$

Furthermore, on the event $\{\max_{i=1,\ldots,n} |Y_i| \leq \beta_n\}$

$$\left\| \nabla_{(\mathbf{w}_{k,1,1}^{(L_n)})_k} (Y_{j_t} - f_{\mathbf{w}^{(t)}}(X_{j_t}))^2 \right\|_\infty \leq 2 \cdot (\beta_n + c_{17} \cdot n^\delta) \leq c_{135} \cdot n^\delta$$

holds, hence assumption (22) of Theorem 4 is satisfied for

$$D_n = c_{135} \cdot n^\delta.$$

Application of Theorem 4 yields

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

$$\leq c_{136} \cdot \left(\frac{\beta_n^2}{n^5} + \beta_n^2 \cdot \sqrt{\mathbf{P}(E_n^c)} + \mathbf{E}\Big\{ \sup_{\mathbf{w} \in \mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X) - Y|^2 - |m(X) - Y|^2\} \right.$$

$$\left. - \frac{2}{n} \sum_{i=1}^{n} \left(|T_{\beta_n} f_{\mathbf{w}}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \Big) \Big\} \right.$$

$$+ \mathbf{E} \left\{ \int |f_{(((\mathbf{w}^*)_{k,1,1}^{(L)})_k, ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L})}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{E_n} \right\}$$

$$\left. + \frac{1}{n} + (\beta_n + n^{\delta}) \cdot \frac{1}{n^{L \cdot (5p_{max} + K_{max} + 6)}} \cdot n^{L \cdot (5p_{max} + K_{max} + 5)} \cdot \log n + \frac{n^{2\delta}}{N_n} + \frac{n^{2\delta}}{n^3} \right).$$

Hence it suffices to show

$$\mathbf{P}(E_n^c) \leq \frac{c_{137}}{n}, \tag{38}$$

$$\mathbf{E}\Big\{ \sup_{\mathbf{w} \in \mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X) - Y|^2 - |m(X) - Y|^2\}$$

$$- \frac{2}{n} \sum_{i=1}^{n} \left(|T_{\beta_n} f_{\mathbf{w}}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \Big) \Big\}$$

$$\leq c_{138} \cdot (\log n)^{3/2} \cdot \max_{i,j} n^{\frac{-p_j^{(i)}}{2p_j^{(i)} + K_j^{(i)}} + \delta} \tag{39}$$

and

$$\mathbf{E} \left\{ \int |f_{(((\mathbf{w}^*)_{k,1,1}^{(L)})_k, ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L})}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{E_n} \right\} \leq c_{139} \cdot \max_{i,j} n^{\frac{-p_j^{(i)}}{2p_j^{(i)} + K_j^{(i)}}}. \tag{40}$$

Proof of (38): In order to bound $\mathbf{P}(E_n^c)$ we consider a sequential choice of the weights of the K_n networks

$$f_{\mathbf{w}^{(0)},k,1}^{(L)} \quad (k = 1, \ldots, K_n).$$

The probability that none of the

$$r_n \cdot (d+1) + (L-2) \cdot r_n \cdot (r_n + 1) + (r_n + 1) \leq c_{140} \cdot n$$

weights of the first of these networks differs from the corresponding weight in

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

by more than ϵ_n is bounded from above by

$$\left(\frac{\epsilon_n}{2 \cdot c_{20} \cdot n^{5p_{max} + K_{max} + 5}} \right)^{c_{140} \cdot n} \leq e^{-c_{141} \cdot (\log n) \cdot n}.$$

Hence we get that the probability that none of the

$$\lceil n \cdot e^{c_{141} \cdot (\log n) \cdot n} \rceil$$

many networks

$$f^{(L)}_{\mathbf{w}^{(0)},k,1} \quad (k = 1, \dots, \lceil n \cdot e^{c_{141} \cdot (\log n) \cdot n} \rceil)$$

differ in all weights from the corresponding weight in

$$f^{(L)}_{\bar{\mathbf{w}},1,1}$$

by at most ϵ_n is bounded from above by

$$\left(1 - e^{-c_{141} \cdot (\log n) \cdot n}\right)^{n \cdot e^{c_{141} \cdot (\log n) \cdot n}} \leq e^{-n}.$$

Since we have for n large

$$N \cdot N_n \cdot \lceil n \cdot e^{c_{141} \cdot (\log n) \cdot n} \rceil \leq K_n,$$

we can conclude: The probability that for any $k \in \{1, \dots, N \cdot N_n\}$ in all of the networks

$$f^{(L)}_{\mathbf{w}^{(0)},j,1} \quad (j = (k-1) \cdot \lceil n \cdot e^{c_{141} \cdot (\log n) \cdot n} \rceil + 1, \dots, k \cdot \lceil n \cdot e^{c_{141} \cdot (\log n) \cdot n} \rceil)$$

at least one of the weights differs from the corresponding weight in

$$f^{(L)}_{\bar{\mathbf{w}},k,1}$$

by more than ϵ_n is bounded from above by

$$N \cdot N_n \cdot e^{-n} \leq \frac{c_{142}}{n}.$$

Since $\mathbf{P}\{E_n^c\}$ is upper bounded by this probability, this implies (38).
Proof of (39): By Lemma 8 we have

$$\mathbf{E}\Big\{ \sup_{\mathbf{w} \in \mathbf{\Theta}} \Big(\mathbf{E}\{|T_{\beta_n} f_{\mathbf{w}}(X) - Y|^2 - |m(X) - Y|^2\} - \frac{1}{n} \sum_{i=1}^{n} \left(|T_{\beta_n} f_{\mathbf{w}}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \Big) \Big\}$$
$$\leq \frac{c_{143}}{\sqrt{n}} + c_{144} \cdot \beta_n \cdot c_{145} \cdot n^{\delta} \cdot \mathbf{E}\Big\{ \sup_{\mathbf{w} \in \mathbf{\Theta}, k \in \{1, \dots, K_n\}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot f_{\mathbf{w},k,1}(X_i) \right| \Big\}.$$

By discretizing the

$$r_n \cdot (d+1) + (L-2) \cdot r_n \cdot (r_n + 1) + (r_n + 1) \leq c_{146} \cdot \max_{i,j} n^{\frac{K_j^{(i)}}{2p_j^{(i)} + K_j^{(i)}}}$$

many weights in each of the classes

$$\{f^{(L)}_{\mathbf{w},k,1} \quad : \quad \mathbf{w} \in \mathbf{\Theta}\} \quad (k = 1, \dots, K_n)$$

by creating a grid in the intervals of length

$$\Delta_n = 2 \cdot c_{147} \cdot n^{5p_{max} + K_{max} + 5}$$

with grid size

$$\frac{1/n}{c_{148} \cdot n^{L/2+L\cdot(5p_{max}+K_{max}+5)}},$$

which leads for all k to the same set of functions, we get (cf., Lemma 2) a $1/n$-supremum norm cover of

$$\{f^{(L)}_{\mathbf{w},k,1} \quad : \quad \mathbf{w} \in \mathbf{\Theta},\, k \in \{1,\ldots,K_n\}\}$$

of size at most

$$e^{c_{149}\cdot(\log n)\cdot\max_{i,j} n^{\frac{K_j^{(i)}}{2p_j^{(i)}+K_j^{(i)}}}}.$$

Together with the union bound and the inequality of Hoeffding (cf., e.g., Lemma A.3 in Györfi et al. (2002)) this implies for $\delta_n \geq 2/n$

$$\begin{aligned}
&\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta},k\in\{1,\ldots,K_n\}} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot f^{(L)}_{\mathbf{w},k,1}(X_i) \right| \Big\} \\
&\leq \delta_n + \int_{\delta_n}^\infty \mathbf{P}\left\{ \sup_{\mathbf{w}\in\mathbf{\Theta},k\in\{1,\ldots,K_n\}} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot f^{(L)}_{\mathbf{w},k,1}(X_i) \right| > t \right\} dt \\
&\leq \delta_n + 2\cdot e^{c_{149}\cdot(\log n)\cdot\max_{i,j} n^{\frac{K_j^{(i)}}{2p_j^{(i)}+K_j^{(i)}}}} \int_{\delta_n}^\infty 2\cdot e^{-\frac{2\cdot n\cdot(t/2)^2}{4}} dt \\
&\leq \delta_n + e^{c_{149}\cdot(\log n)\cdot\max_{i,j} n^{\frac{K_j^{(i)}}{2p_j^{(i)}+K_j^{(i)}}}} \int_{\delta_n}^\infty 2\cdot e^{-\frac{n\cdot t\cdot\delta_n}{8}} dt \\
&\leq \delta_n + e^{c_{149}\cdot(\log n)\cdot\max_{i,j} n^{\frac{K_j^{(i)}}{2p_j^{(i)}+K_j^{(i)}}}} \frac{16}{n\cdot\delta_n} e^{-\frac{n\cdot\delta_n^2}{8}}.
\end{aligned}$$

With

$$\delta_n = \sqrt{\frac{8}{n}\cdot c_{149}\cdot(\log n)\cdot\max_{i,j} n^{\frac{K_j^{(i)}}{2p_j^{(i)}+K_j^{(i)}}} + 8\cdot\frac{\log n}{n}}$$

we get

$$\mathbf{E}\Big\{ \sup_{\mathbf{w}\in\mathbf{\Theta},k\in\{1,\ldots,K_n\}} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot f^{(L)}_{\mathbf{w},k,1}(X_i) \right| \Big\} \leq c_{150}\cdot\sqrt{\log n}\cdot\max_{i,j} n^{\frac{-p_j^{(i)}}{2p_j^{(i)}+K_j^{(i)}}}$$

which implies (39).

Proof of (40): The definition of $((\mathbf{w}^*)^{(L)}_{k,1,1})_k$ implies that on the event E_n we have

$$|f_{(((\mathbf{w}^*)^{(L)}_{k,1,1})_k,((\mathbf{w}^{(0)})^{(l)}_{k,i,j})_{k,i,j,l:l<L})}(x) - m(x)|^2$$

$$
\begin{aligned}
&= |\sum_{k=1}^{N\cdot N_n} \bar{\mathbf{w}}_{k,1,1}^{(L)} \cdot f_{\mathbf{w}^{(0)},j_k,1}^{(L)}(x) - m(x)|^2 \\
&\leq 2 \cdot |\sum_{k=1}^{N\cdot N_n} \bar{\mathbf{w}}_{k,1,1}^{(L)} \cdot f_{\mathbf{w}^{(0)},j_k,1}^{(L)}(x) - f_{\bar{\mathbf{w}}}(x)|^2 + 2 \cdot |f_{\bar{\mathbf{w}}}(x) - m(x)|^2 \\
&\leq 2 \cdot |\sum_{k=1}^{N\cdot N_n} \bar{\mathbf{w}}_{k,1,1}^{(L)} \cdot f_{\mathbf{w}^{(0)},j_k,1}^{(L)}(x) - f_{\bar{\mathbf{w}}}(x)|^2 + 2 \cdot \left(c_{151} \cdot \max_{i,j} n^{\frac{-p_j^{(i)}}{2\cdot(2p_j^{(i)}+K_j^{(i)})}} \right)^2 .
\end{aligned}
$$

On E_n we get by Lemma 2 that this in turn is bounded from above by

$$
\begin{aligned}
&2 \cdot \alpha_n \cdot c_{152} \cdot \left(n^{L/2} \cdot n^{L\cdot(5p_{max}+K_{max}+5)} \right)^2 \cdot \epsilon_n^2 + 2 \cdot \left(c_{151} \cdot \max_{i,j} n^{\frac{-p_j^{(i)}}{2\cdot(2p_j^{(i)}+K_j^{(i)})}} \right)^2 \\
&\leq c_{153} \cdot \max_{i,j} n^{\frac{-p_j^{(i)}}{2p_j^{(i)}+K_j^{(i)}}},
\end{aligned}
$$

which implies (40).

□

References

[1] Allen-Zhu, Z., Li, Y., und Song, Z. (2019). A convergence theory for deep learning via over-parameterization. *Proceedings of the 36th International Conference on Machine Learning (PMLR 2019)*, Long Beach, California, **97**, pp. 242-252.

[2] Bauer, B., and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics* **4**, pp. 2261–2285.

[3] Billings, W., Hedelius, B., Millecam, T., Wingate, D., and Corte, D. (2019). ProSPr: Democratized Implementation of Alphafold Protein Distance Prediction Network. BioRxiv, doi: 10.1101/830273.

[4] Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. Preprint, *arXiv: 1805.09545.*

[5] Drews, S., and Kohler, M. (2023). Analysis of the expected L_2 error of an over-parametrized deep neural network estimate learned by gradient descent without regularization. Preprint.

[6] Drews, S., and Kohler, M. (2024). On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent. *Annals of the Institue of Statistical Mathematics* **76**, pp. 361-391.

[7] Du, S., Lee, J., Li, H., Wang, L., und Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. International Conference on Machine Learning, Preprint, *arXiv: 1811.03804*.

[8] Friedman, J. H., and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, **76**, pp. 817-823.

[9] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics, Springer-Verlag, New York.

[10] Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, **21**, pp. 157-178.

[11] Härdle, W., and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, **84**, pp 986-995.

[12] Hanin, B., and Nica, M. (2019). Finite depth and width corrections to the neural tangent kernel. *arXiv: 1909.05989.*

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

[14] Imaizumi, M., and Fukamizu, K. (2018). 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] Jacot, A., Gabriel, F., and Hongler, C. (2020). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *arXiv: 1806.07572v4.*

[16] Kim, Y. (2014). Convolutional neural networks for sentence classification. Preprint, *arXiv: 1408.5882.*

[17] Kawaguchi, K, and Huang, J. (2019). Gradient descent finds global minima for generalizable deep neural networks of practical sizes. *57th IEEE Annual Allerton Conference on Communication, Control, and Computing*, Allerton, IL, pp. 92-99.

[18] Kohler, M. (2024). On the rate of convergence of deep neural network regression estimates learned by gradient descent. Preprint.

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

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

[21] Kohler, M., and Krzyżak, A. (2022). Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent. Preprint, *arXiv: 2210.01443.*

[22] Kohler, M., and Krzyżak, A. (2023). On the rate of convergence of an over-parametrized deep neural network regression estimate with ReLU activation function learned by gradient descent. Preprint.

[23] Kohler, M., Krzyżak, A., and Sänger (2024). Learning of deep convolutional network image classifiers via stochastic gradient descent and over-parametrization. Preprint.

[24] 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.

[25] Kong, E., and Xia, Y. (2007). Variable selection for the single-index model. *Biometrika*, **94**, pp. 217-229.

[26] 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* Red Hook, NY: Curran. **25**, pp. 1097-1105.

[27] Kutyniok, G. (2020). Discussion of "Nonparametric regression using deep neural networks with ReLU activation function". *Annals of Statistics* **48**, pp. 1902–1905.

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

[29] 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.

[30] Nguyen, P.-M., and Pham, H. T. (2020). A rigorous framework for the mean field limit of multilayer neural networks. Preprint, *arXiv: 2001.1144.*

[31] Nitanda, A., and Suzuki, T. (2021). Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *arXiv: 2006.12297.*

[32] Rahimi, A., and Recht, B. (2008a). Random features for large-scale kernel machines. In *Advances in Neural Information Procesing Systems*, pp. 1177-1184.

[33] Rahimi, A., and Recht, B. (2008b). Uniform approximation of function with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 555-561, IEEE.

[34] Rahimi, A., and Recht, B. (2009). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurman, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, Curran Associates, Inc. **21**, pp. 1313-1320.

[35] Scarselli, F., and Tsoi, A. C. (1998). Universal Approximation Using Feedforward Neural Networks: A Survey of Some Existing Methods, and Some New Results. *Neural Networks*, **11**, pp. 15-37.

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

[37] 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.

[38] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, pp. 1040-1053.

[39] Stone, C. J. (1985). Additive regression and other nonparametric models, *Annals of Statistics*, **13**, pp. 689-705.

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

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

[42] 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. Preprint, *arXiv: 1609.08144*.

[43] Yu, Y., and Ruppert, D. (2002). Penalized Spline Estimation for Partially Linear Single-Index Models. *Journal of the American Statistical Association*, **97**, pp. 1042-1054.

[44] Zong, M., and Krishnamachari, B.(2022). A survey on GPT-3. *arXiv: 2212.00857*

[45] Zou, D., Cao, Y., Zhou, D., und Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep ReLU networks. Preprint, *arXiv: 1811.08888*.