Analysis of the expected L_2 error of an over-parametrized deep neural network estimate learned by gradient descent without regularization

*

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

November 27, 2023

Abstract
Recent results show that estimates defined by over-parametrized deep neural networks learned by applying gradient descent to a regularized empirical L_2 risk are universally consistent and achieve good rates of convergence. In this paper, we show that the regularization term is not necessary to obtain similar results. In the case of a suitably chosen initialization of the network, a suitable number of gradient descent steps, and a suitable step size we show that an estimate without a regularization term is universally consistent for bounded predictor variables. Additionally, we show that if the regression function is Hölder smooth with Hölder exponent $1/2 \leq p \leq 1$, the L_2 error converges to zero with a convergence rate of approximately $n^{-1/(1+d)}$. Furthermore, in case of an interaction model, where the regression function consists of a sum of Hölder smooth functions with d^* components, a rate of convergence is derived which does not depend on the input dimension d.

AMS classification: Primary 62G08; secondary 62G20.

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

1. Introduction

1.1. Deep learning

In the last decade, deep learning, i.e. the application of deep neural networks to data, has gained a lot of attraction in practical applications as well as in theoretical considerations and achieves impressive results. For instance, in natural science AlphaFold can predict protein structures (c.f. Billings et al. (2019)) or AlphaZero is able to outperform humans in three popular board games (c.f. McGrath et al. (2022)). Furthermore, the chatbot

*Running title: *Over-parametrized deep neural network estimates*
†Corresponding author. Tel: +49-6151-16-23372, Fax:+49-6151-16-23381

ChatGPT is one of the most capable language models nowadays, consisting of 175 billion parameters (c.f. Zong and Krishnamachari (2022)). Another example is the text-to-image model DALL-E-2, which consists of 3.5 billion parameters (c.f. Hunt (2023)). These observations suggest that a large number of parameters is very useful in practical applications of deep learning.
From a theoretical point of view this great success cannot yet be fully explained. However, there exist already some results concerning deep neural networks. For example, results on the rate of convergence of least squares estimates based on deep neural networks have been derived (cf., e.g., Bauer and Kohler (2019), Schmidt-Hieber (2020), Kohler and Langer (2021)).

Nevertheless, the above results neglect two features which have turned out to be very important for practical applications. First, in contrast to the results above, in practice estimates are computed using gradient descent instead of least squares. The second property they ignore is that practical applications often use over-parametrized neural networks. A network is called over-parametrized if the number of its parameters is much larger than the sample size.

In this paper we consider a suitable over-parametrized deep neural network estimate computed by gradient descent and analyze its statistical performance in a nonparametric regression setting.

1.2. Nonparametric regression

We analyze deep neural network estimates in the context of nonparametric regression. To do this we consider a random $\mathbb{R}^d \times \mathbb{R}$-valued vector (X, Y) with $\mathbf{E}Y^2 < \infty$. Here we are interested in the dependence of the so-called *response variable* Y on the value of the *observation vector* X. We assume that it is possible to observe data of (X, Y). This dataset is given by

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

where $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed (i.i.d.). Our aim is to construct an estimate $m_n(x) := m_n(x, \mathcal{D}_n)$ of the corresponding regression function $m : \mathbb{R}^d \to \mathbb{R}$ with $m(x) = \mathbf{E}\{Y|X = x\}$ such that the L_2 error

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

is small (cf. e.g. Györfi et al. (2002) for a detailed introduction to nonparametric regression).

There exist different modes of convergence. An important property an estimate should satisfy is to be universally consistent. Universal consistency means that the L_2 error converges to zero for all distributions of (X, Y). But it does not provide any information about a rate of convergence of the L_2 error. As shown in Theorem 7.2 and Problem 7.2 in Devroye, Györfi and Lugosi (1996) a rate of convergence cannot be derived in general since there always exists a distribution such that the L_2 error converges to zero arbitrarily slowly. Thus, in order to derive non-trivial results about the rate of convergence, we need

to restrict the class of regression functions. To do this we use the following definition of (p, C) 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 $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = q$ the partial derivative $\frac{\partial^q m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ exists and satisfies*

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

for all $x, z \in \mathbb{R}^d$, where $\| \cdot \|$ denotes the Euclidean norm. For $p \leq 1$, the function m is called Hölder-smoooth with exponent p and Hölder-constant C.

Stone (1982) showed that the optimal minimax rate of convergence in nonparametric regression for (p, C)-smooth functions is given by $n^{-2p/(2p+d)}$.

1.3. Main results

The goal of this paper is to study over-parametrized deep neural networks trained by gradient descent from a statistical point of view.

In this context we define an estimate which fits an over-parametrized deep neural network via gradient descent to the data. The output of the network is defined as a linear combination of a large number of fully connected deep neural networks. The main feature of this work is that we do not need a regularization term in the empirical L_2 risk. To derive the results for the empirical L_2 risk without a regularization term, we use a new approach to analyze the optimization error. Previous methods used that the gradient of the empirical L_2 risk is Lipschitz continuous (Braun et al. (2024)) or they used the convexity of the empirical L_2 risk (Yehudai and Shamir (2022)) to analyze the optimization error. In our work, we combine these two techniques (cf. Lemma 1 below).

Firstly, we prove that for a bounded support of the input data the estimate is universally consistent. Secondly, we analyze the rate of convergence of the over-parametrized deep neural network estimate. In Theorem 2 below, we show that if the regression function is (p, C)-smooth with $p \in [1/2, 1]$, then the expected L_2 error of the truncated estimate tends to zero, with a rate of convergence close to

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

Furthermore, in case of an interaction model, i.e. if the regression function is a sum of (p, C)-smooth functions where each function depends on at most d^* of the d components of X, the estimate achieves a rate of convergence which is close to

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

and does not depend on the input dimension d.

1.4. Discussion of related results

From a theoretical point of view, there have been many interesting results in the past years. For instance, in the case of a suitably defined least squares estimate based on a multilayer neural network, Kohler and Krzyżak (2017) were able to derive a rate of convergence of $n^{-\frac{2p}{2p+d^*}}$ for (p, C)-smooth regression functions, which satisfy a model where the functions consist of a composition of functions applied to at most d^* components of its inputs, with $p \leq 1$. This rate does not depend on the input dimension d. Bauer and Kohler (2019) were able to show the same assertion for $p > 1$, provided that the activation function is sufficiently smooth. Under the additional condition of the so-called *sparsity* of the networks, Schmidt-Hieber (2020) showed the same rate of convergence for neural network estimates with ReLU activation function in case that the regression function satisfies a kind of hierarchical composition model. The property of sparsity is not necessary to obtain this rate of convergence for a regression function satisfying a hierarchical composition model (Kohler and Langer (2021)). Suzuki (2018) and Suzuki and Nitanda (2021) were able to show that this dimensional reduction can be achieved even under weaker assumptions on the smoothness of the regression function. For this purpose they considered deep learning with ReLU activation function for functions in Besov spaces.

In classical machine learning theory, the goal was to avoid over-parametrized neural networks. It was thought that these networks lead to an overfitting of the weights to the data and therefore generalize poorly to new data. Therefore, it was desired to obtain a bias-variance trade-off. The model should be complex enough to represent the structure of the underlying data, but simple enough to avoid overfitting. However, in practical applications, over-parametrized neural networks are often used very successfully and achieve high accuracy on new test data. This phenomenon has been studied recently by many different researchers (c.f., e.g. Belkin et al. (2019), Frei, Chatterji and Bartlett (2022) and the literature cited therein).

Belkin et al. (2019) were able to reconcile classical theory with new findings in practical applications and identified a pattern of performance dependence on unseen model capacity data and the underlying mechanism of emergence. This dependence is represented by the so-called *double descent curve*. This curve shows that if the model capacity is greater than the interpolation threshold, the test risk will decrease again.

In recent years, much work has focused on the capacity of neural networks. They tried to understand the ability of neural networks to adapt to the training data either on a finite data set (Bubeck et al. (2020)) or with respect to neural networks trained by gradient methods (Daniely (2019), Daniely (2020)) or in neural tangent training (Montanari and Zhong (2020)). In addition Bartlett et al. (2020) and Belkin, Rakhlin and Tsybakov (2019) have already shown that over-parametrized neural networks can achieve good rates of convergence.

In practical applications the weights of neural network estimates are computed by gradient descent. So from a theoretical point of view it is very interesting to derive theoretical results for such estimates. Nitanda and Suzuki (2021) showed that the averaged stochastic gradient descent for over-parametrized neural networks with one hidden layer

can achieve the optimal minimax rate of convergence in a neural tangent kernel setting where the smoothness of the regression function is measured by the neural tangent kernel. Furthermore, it was shown that the (stochastic) gradient descent can find a global minimum of the empirical L_2 risk for suitable over-parametrized deep neural networks (Allen-Zhu, Li and Song (2019), Kawaguchi and Huang (2020), Arora et al. (2019), Du et al. (2018)).

However, Kohler and Krzyżak (2021) were able to exhibit that over-parametrized deep neural networks minimizing the empirical L_2 risk do not generalize well on new data, in the sense that the networks that minimize the empirical risk do not achieve the optimal minimax rate of convergence.

Braun et al. (2024) showed that under the assumption that the Fourier transform of the regression function decays fast enough, a suitably initialized estimate learned by gradient descent achieves (up to a logarithmic factor) a rate of convergence of $n^{-\frac{1}{2}}$ which does not depend on the input dimension d. This result is related to a classical result of Barron (1994). He was able to derive a similar rate of convergence for a least squares neural network estimate in case that the Fourier transform has a finite first moment, which requires that the function becomes smoother with increasing dimension. Kohler and Krzyżak (2022b) analyzed the same estimates in an over-parametrized setting and were able to derive an improved rate of convergence of $n^{-\frac{2}{3}}$ for $d = 1$ using a suitably regularized L_2 risk.

The property of over-parametrization allows the gadient descent to find interpolating solutions that implicitly impose a regularization. This over-parametrization leads to benign overfitting (Bartlett, Montanari and Rakhlin (2021)).

Drews and Kohler (2022) showed the property of universal consistency for over-parametrized deep neural network estimates computed by minimizing the L_2 risk with gradient descent. For these estimates, that are computed minimizing a regularized L_2 risk via gradient descent, Kohler and Krzyżak (2022a) were able to derive a rate of convergence close to $n^{-\frac{1}{1+d}}$ for suitably smooth regression functions. For this, they considered three key ingredients. Firstly, they used that with high probability a subset of the inner initial weights has good properties. For a suitable choice of the outer weights this causes a good approximation property. Secondly, for a suitably chosen number of gradient steps, a suitably chosen step size and suitably chosen bounds for the initial weights, they were able to use a metric entropy bound to control the generalizability of the estimate. Thirdly, they analyzed the optimization of the empirical L_2 risk by optimizing the outer weights during the gradient descent and by using a regularization term.

Kohler and Krzyżak (2022a) were also able to show a dimension-independent rate of convergence for interaction models close to $n^{-\frac{1}{1+d^*}}$.

Our results are an extension of the results in Drews and Kohler (2022) and Kohler and Krzyżak (2022a). We show that the regularization term is not necessary. The estimate is universally consistent even by minimizing the empirical L_2 risk without a regularization term and achieves similar rates of convergence.

1.5. Notation

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

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

as the supremum norm. Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$, let $x_1, \dots, x_n \in \mathbb{R}^d$, set $x_1^n = (x_1, \dots, x_n)$ and let $p \geq 1$. A finite collection $f_1, \dots, f_N : \mathbb{R}^d \to \mathbb{R}$ is called an L_p ε–cover of $\mathcal{F}$ on x_1^n if for any $f \in \mathcal{F}$ there exists $i \in \{1, \dots, N\}$ such that

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

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

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

For $z \in \mathbb{R}$ and $\beta > 0$ we define $T_\beta z = \max\{-\beta, \min\{\beta, z\}\}$. If $f : \mathbb{R}^d \to \mathbb{R}$ is a function and $\mathcal{F}$ is a set of such functions, then we set $(T_\beta f)(x) = T_\beta\left(f(x)\right)$ and

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

1.6. Outline

In Section 2 we define the estimate. The main results concerning the universal consistency and the rate of convergence of the deep neural network estimate learned by gradient descent are presented in Section 3. The proofs of the main results are given in Section 4.

2. Definiton of the estimate

In order to define the estimate, let $\sigma(x) = 1/(1 + e^{-x})$ be the logistic squasher. For the neural network topology, we choose a linear combination of K_n fully connected neural networks with L layers and r neurons per layer. The output of the network is defined by

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

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

$$f_{k,i}^{(l)}(x) = \sigma\left(\sum_{j=1}^{r} w_{k,i,j}^{(l-1)} \cdot f_{k,j}^{(l-1)}(x) + w_{k,i,0}^{(l-1)} \right) \tag{2}$$

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

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

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

We denote by $f_{k,i}^{(l)}$ the output of neuron i in the l-th layer of the k-th fully connected network. By $w_{k,i,j}^{(l-1)}$ we denote the weight between neuron j in the $(l-1)$-th layer and neuron i in the l-th layer of the k-th fully connected network.

The number of weights of the neural network is given by

$$W_n = K_n \cdot (1 + (r+1) + (L-2) \cdot r \cdot (r+1) + r \cdot (d+1)).$$

To obtain the estimate, we want to learn the weights using gradient descent. For this we initialize the weights $\mathbf{w}^{(0)} = ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l}$ by setting

$$(\mathbf{w}^{(0)})_{1,1,k}^{(L)} = 0 \quad \text{for } k = 1, \dots, K_n, \tag{4}$$

and by choosing all components of $\mathbf{w}^{(0)}$ such that they are independent. We choose the weights $(\mathbf{w}^{(0)})_{k,i,j}^{(l)}$ with $l \in \{1, \dots, L-1\}$ uniformly distributed on $[-20d \cdot (\log n)^2, 20d \cdot (\log n)^2]$ and $(\mathbf{w}^{(0)})_{k,i,j}^{(0)}$ uniformly distributed on $[-8d \cdot (\log n)^2 \cdot n^\tau, 8d \cdot (\log n)^2 \cdot n^\tau]$ for some fixed $\tau > 0$.

Set

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

and compute

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)})$$

for $t = 0, \dots, t_n - 1$ where

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}}(X_i) - Y_i|^2$$

is the empirical L_2 risk of the network $f_{\mathbf{w}}$ on the training data. The number of gradient descent steps t_n will be chosen in Section 3 below.

The estimate is then defined by

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

where $\beta_n = c_1 \cdot \log n$.

Because of (4) we have

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

3. Main results

3.1. Universal consistency

Our first result is the following theorem which presents the universal consistency for bounded X of an estimate learned by gradient descent using the empirical L_2 risk without a regularization term.

Theorem 1. *Let $\sigma(x) = 1/(1+e^{-x})$ be the logistic squasher, and let $K_n, L, r \in \mathbb{N}$ and $\tau \in \mathbb{R}_+$. Furthermore, assume that $L \geq 2$, $r \geq 2d$, $\tau = 1/(d+1)$,*

$$\frac{K_n}{n^\kappa} \to 0 \quad (n \to \infty) \tag{5}$$

for some $\kappa > 0$,

$$\frac{K_n}{n^{r+2}} \to \infty \quad (n \to \infty), \tag{6}$$

and $\beta_n = c_1 \cdot \log n$ for some $c_1 > 0$. Let the estimate m_n be defined as in Section 2 with

$$t_n = \lceil c_2 \cdot L_n \rceil \tag{7}$$

for some $c_2 \geq 1$ and $L_n > 0$ which satisfies

$$L_n \geq K_n^{3/2} \cdot (\log n)^{6L+5}.$$

Then we have

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

for every distribution of (X, Y) where $supp(X)$ is bounded and $\mathbf{E}Y^2 < \infty$.

Remark 1. *The estimate is over-parametrized in the sense that the number of its parameters is much larger than the sample size. This is due to condition (6), which requires that K_n is asymptotically larger than $n^{r+2} \geq n^{2d+2}$. Theorem 1 shows that using the empirical L_2 risk without a regularization term, a suitable large number of steps, and a step size which is equal to the reciprocal of the number of steps provides a good generalization of new independent data.*

Remark 2. *The proof uses that the inner weights are chosen with high probability such that for properly chosen outer weights of the neural networks, the corresponding neural network has a small empirical L_2 risk. Hence, this neural network is based on representation guessing instead of representation learning.*

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

Besides the result about the universal consistency of the estimate, it is interesting how fast the expected L_2 error converges to 0. Therefore, in the next theorem a rate of convergence for (p, C)-smooth functions is derived.

Theorem 2. *Let $n \in \mathbb{N}$ and let $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ valued random variables which satisfy $supp(X) \subseteq [0,1]^d$ and*

$$\mathbf{E}\left\{e^{c_3 \cdot Y^2}\right\} < \infty \tag{8}$$

for some $c_3 > 0$. Assume that the corresponding regression function $m(x) = \mathbf{E}\{Y|X = x\}$ is (p, C)-smooth for some $1/2 \leq p \leq 1$ and some $C > 0$.

Let $\sigma(x) = 1/(1 + e^{-x})$ be the logistic squasher, let $L, r \in \mathbb{N}$ with $L \geq 2$ and $r \geq 2d$. Set $\beta_n = c_1 \cdot \log n$ for some $c_1 > 0$,

$$K_n = n^{6d+r+2},$$
$$\tau = \frac{1}{1+d}.$$

Define the estimate m_n as in Section 2 with

$$t_n = \lceil c_4 \cdot L_n \rceil \tag{9}$$

for some $c_4 \geq 1$ and $L_n > 0$ which satisfies

$$L_n \geq K_n^{3/2} \cdot (\log n)^{6L+2}$$

and assume that

$$c_1 \cdot c_3 \geq 2. \tag{10}$$

Then we have for any $\epsilon > 0$

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_5 \cdot n^{-\frac{1}{1+d}+\epsilon}.$$

Remark 3. *According to Stone (1982) the optimal minimax rate of convergence for (p, C)-smooth functions is $n^{-\frac{2p}{2p+d}}$. Thus, the rate of convergence derived from Theorem 2 is almost optimal for $p = \frac{1}{2}$. Unfortunately, if $p > \frac{1}{2}$, the rate of convergence derived above is not almost optimal. We assume that this is not a property of the estimate, but rather a consequence of our proof.*

3.3. Rate of convergence in an interaction model

The aim of this subsection is to modify the estimate defined above in order to obtain a rate of convergence that does not depend on d. For this we assume that the regression function satisfies

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

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

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

for $I = \{j_1, \dots, j_{d^*}\}$.

The neural network is then given by

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

where $f_{\mathbf{w}_I}$ is defined by (1)–(3) with d replaced by d^* and weight vector $\mathbf{w}_I$, where

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

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

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

and by choosing all weights $(\mathbf{w}_I^{(0)})_{k,i,j}^{(l)}$ such that they are independent. We choose the weights $(\mathbf{w}_I^{(0)})_{k,i,j}^{(l)}$ with $l \in \{1, \dots, L-1\}$ uniformly distributed on the interval $[-20d^* \cdot (\log n)^2, 20d^* \cdot (\log n)^2]$ and we choose $(\mathbf{w}_I^{(0)})_{k,i,j}^{(0)}$ uniformly distributed on the interval $[-8d \cdot (\log n)^2 \cdot n^\tau, 8d \cdot (\log n)^2 \cdot n^\tau]$ where $\tau = \frac{1}{1+d^*}$ ($I \subseteq \{1, \dots, d\}$ with $|I| = d^*$).

Similar as above we define the estimate as follows: Set

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

and compute

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)})$$

for $t = 0, \dots, t_n - 1$ where

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}}(X_i) - Y_i|^2$$

is the empirical L_2 risk of the network $f_{\mathbf{w}}$ on the training data. The number of gradient descent steps t_n will be chosen in Theorem 3 below.

The estimate is then defined by

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

where $\beta_n = c_1 \cdot \log n$.

Theorem 3. *Let $d \in \mathbb{N}, d^* \in \{1, \dots, d\}$, $1/2 \leq p \leq 1$. Furthermore, let $C > 0$, $n \in \mathbb{N}$ and let $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$ be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ valued random variables such that supp$(X) \subseteq [0,1]^d$ and*

$$\mathbf{E}\{e^{c_3 \cdot Y^2}\} < \infty$$

holds for some $c_3 > 0$. *Assume that the corresponding regression function* $m(x) = \mathbf{E}\{Y|X = x\}$ *satisfies*

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

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

Let $\sigma(x) = 1/(1 + e^{-x})$ *be the logistic squasher. Let* $L, r \in \mathbb{N}$ *with* $L \geq 2$ *and* $r \geq 2d^*$. *Set* $\beta_n = c_1 \cdot \log n$,

$$K_n = n^{6d^*+r+2} \qquad \text{and} \qquad \tau = \frac{1}{1+d^*}.$$

Define the estimate m_n *as in Section 3.3 with*

$$t_n = \lceil c_6 \cdot L_n \rceil$$

for some $c_6 \geq 1$ *and* $L_n > 0$ *which satisfies*

$$L_n \geq K_n^{3/2} \cdot (\log n)^{6L+2}$$

and assume that

$$c_1 \cdot c_3 \geq 2. \tag{11}$$

Then we have for any $\epsilon > 0$

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_7 \cdot n^{-\frac{1}{1+d^*}+\epsilon}.$$

Remark 4. *The optimal minimax rate of convergence* $n^{-\frac{2p}{2p+d}}$ *derived by Stone (1982) suffers from the curse of dimensionality. This means that if* d *is very large compared to* p, *the rate of convergence becomes extremely slow. However, Stone (1994) has already established that under appropriate assumptions it is possible to circumvent the curse of dimensionality. In Theorem 3, we were able to show that under the above assumptions on the regression function, the over-parametrized neural network estimate can also circumvent the curse of dimensionality.*

4. Proofs

4.1. Auxiliary results for the proof of Theorem 1

In this section we will present auxiliary results that are necessary for the proof of Theorem 1. The first auxiliary result enables us to analyze the gradient descent.

Lemma 1. *Let $d_1, d_2 \in \mathbb{N}$, let $(u_0, v_0) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, let $F : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}_+$ be a continuously differentiable function and set*

$$A = \left\{(u, v) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \,:\, \|(u, v) - (u_0, v_0)\| \leq 2 \cdot \sqrt{F(u_0, v_0)} + 1\right\}.$$

Let $u^* \in \mathbb{R}^{d_1}$ and assume that for some $D_n, L_n > 0$ it holds

$$u \mapsto F(u, v) \quad \text{is convex for all } v \in \mathbb{R}^{d_2},$$

$$\|(\nabla_{(u,v)} F)(u, v)\| \leq L_n \tag{12}$$

for all $(u, v) \in A$,

$$\|(\nabla_{(u,v)} F)(u_1, v_1) - (\nabla_{(u,v)} F)(u_2, v_2)\| \leq L_n \cdot \|(u_1, v_1) - (u_2, v_2)\| \tag{13}$$

for all $(u_1, v_1), (u_2, v_2) \in A$ and

$$|F(u^*, v) - F(u^*, v_0)| \leq D_n \cdot \|u^*\| \cdot \|v - v_0\| \tag{14}$$

for all $v \in \{\tilde{v} \,:\, \|\tilde{v} - v_0\| \leq \sqrt{2 \cdot F(u_0, v_0)}\}$.

Set

$$u_{t+1} = u_t - \lambda \cdot (\nabla_u F)(u_t, v_t),$$
$$v_{t+1} = v_t - \lambda \cdot (\nabla_v F)(u_t, v_t)$$

for $t = 0, 1, \ldots, t_n - 1$, where

$$t_n \geq L_n \qquad \text{and} \qquad \lambda = \frac{1}{t_n}.$$

Then we have

$$F(u_{t_n}, v_{t_n}) \leq F(u^*, v_0) + D_n \cdot \|u^*\| \cdot \sqrt{2 \cdot F(u_0, v_0)} + \frac{\|u^* - u_0\|^2}{2} + \frac{F(u_0, v_0)}{t_n}.$$

Furthermore, even if (14) does not hold, we have

$$\|u_t - u_0\| \leq \sqrt{2 \cdot F(u_0, v_0)} \quad \text{and} \quad \|v_t - v_0\| \leq \sqrt{2 \cdot F(u_0, v_0)}$$

for all $t = 0, 1, \ldots, t_n$.

Proof. In the *first step of the proof* we show

$$\frac{1}{t_n} \sum_{t=0}^{t_n - 1} F(u_t, v_t) \leq \frac{1}{t_n} \sum_{t=0}^{t_n - 1} F(u^*, v_t) + \frac{\|u^* - u_0\|^2}{2} + \frac{1}{2 \cdot t_n} \sum_{t=0}^{t_n - 1} \lambda \cdot \|(\nabla_u F)(u_t, v_t)\|^2. \tag{15}$$

By convexity of $u \mapsto F(u, v_t)$ we have

$$\begin{aligned} &F(u_t, v_t) - F(u^*, v_t) \\ &\leq \langle (\nabla_u F)(u_t, v_t), u_t - u^* \rangle \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2\cdot\lambda}\cdot 2\cdot\langle\lambda\cdot(\nabla_u F)(u_t,v_t), u_t-u^*\rangle\\
&\le \frac{1}{2\cdot\lambda}\cdot\left(-\|u_t-u^*-\lambda\cdot(\nabla_u F)(u_t,v_t)\|^2+\|u_t-u^*\|^2+\|\lambda\cdot(\nabla_u F)(u_t,v_t)\|^2\right)\\
&= \frac{1}{2\cdot\lambda}\cdot\left(\|u_t-u^*\|^2-\|u_{t+1}-u^*\|^2+\lambda^2\cdot\|(\nabla_u F)(u_t,v_t)\|^2\right).
\end{aligned}
$$

This implies

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

In the *second step of the proof* we show that

$$
\begin{aligned}
&\|(\nabla_{(u,v)}F)((u_t,v_t)+\tau\cdot((u_{t+1},v_{t+1})-(u_t,v_t)))-(\nabla_{(u,v)}F)(u_t,v_t)\|\\
&\le L_n\cdot\tau\cdot\|(u_{t+1},v_{t+1})-(u_t,v_t)\|
\end{aligned}
\tag{16}
$$

for all $\tau\in[0,1]$ implies

$$
F(u_{t+1},v_{t+1})-F(u_t,v_t)\le-\frac{1}{2}\cdot\lambda\cdot\|(\nabla_u F)(u_t,v_t)\|^2-\frac{1}{2}\cdot\lambda\cdot\|(\nabla_v F)(u_t,v_t)\|^2. \tag{17}
$$

The function

$$
H:[0,1]\to\mathbb{R},\quad H(\tau)=F((u_t,v_t)+\tau\cdot((u_{t+1},v_{t+1})-(u_t,v_t)))
$$

is continuously differentiable. By the fundamental theorem of calculus, assumption (16) and $\lambda\le 1/L_n$ we get

$$
\begin{aligned}
&F(u_{t+1},v_{t+1})-F(u_t,v_t)=H(1)-H(0)=\int_0^1 H'(\tau)d\tau\\
&=\int_0^1(\nabla_{(u,v)}F)((u_t,v_t)+\tau\cdot((u_{t+1},v_{t+1})-(u_t,v_t)))\cdot((u_{t+1},v_{t+1})-(u_t,v_t))d\tau\\
&=\int_0^1\left((\nabla_{(u,v)}F)((u_t,v_t)+\tau\cdot((u_{t+1},v_{t+1})-(u_t,v_t)))-(\nabla_{(u,v)}F)(u_t,v_t)\right)
\end{aligned}
$$

$$
\begin{aligned}
&\cdot((u_{t+1},v_{t+1})-(u_t,v_t))d\tau\\
&+\int_0^1(\nabla_{(u,v)}F)(u_t,v_t)\cdot((u_{t+1},v_{t+1})-(u_t,v_t))d\tau\\
\le&\int_0^1\big\|(\nabla_{(u,v)}F)((u_t,v_t)+\tau\cdot((u_{t+1},v_{t+1})-(u_t,v_t)))-(\nabla_{(u,v)}F)(u_t,v_t)\big\|\\
&\cdot\|(u_{t+1},v_{t+1})-(u_t,v_t)\|d\tau\\
&+(\nabla_{(u,v)}F)(u_t,v_t)\cdot((u_{t+1},v_{t+1})-(u_t,v_t))\\
\le&\int_0^1 L_n\cdot\tau\cdot\|(u_{t+1},v_{t+1})-(u_t,v_t)\|^2d\tau\\
&+(\nabla_{(u,v)}F)(u_t,v_t)\cdot((u_{t+1},v_{t+1})-(u_t,v_t))\\
=&\frac{1}{2}\cdot L_n\cdot\|(u_{t+1},v_{t+1})-(u_t,v_t)\|^2+(\nabla_{(u,v)}F)(u_t,v_t)\cdot((u_{t+1},v_{t+1})-(u_t,v_t))\\
=&\frac{1}{2}\cdot L_n\cdot\big(\lambda^2\cdot\|(\nabla_uF)(u_t,v_t)\|^2+\lambda^2\cdot\|(\nabla_vF)(u_t,v_t)\|^2\big)-\lambda\cdot\|(\nabla_uF)(u_t,v_t)\|^2\\
&-\lambda\cdot\|(\nabla_vF)(u_t,v_t)\|^2\\
=&\lambda\cdot\left(\frac{1}{2}\cdot L_n\cdot\lambda-1\right)\cdot\|(\nabla_uF)(u_t,v_t)\|^2+\lambda\cdot\left(\frac{1}{2}\cdot L_n\cdot\lambda-1\right)\cdot\|(\nabla_vF)(u_t,v_t)\|^2\\
\le&-\frac{1}{2}\cdot\lambda\cdot\|(\nabla_uF)(u_t,v_t)\|^2-\frac{1}{2}\cdot\lambda\cdot\|(\nabla_vF)(u_t,v_t)\|^2.
\end{aligned}
$$

In the *third step of the proof* we show

$$F(u_1,v_1)-F(u_0,v_0)\le-\frac{1}{2}\cdot\lambda\cdot\|(\nabla_uF)(u_0,v_0)\|^2-\frac{1}{2}\cdot\lambda\cdot\|(\nabla_vF)(u_0,v_0)\|^2.$$

By (12) we know

$$\|(\nabla_{(u,v)}F)(u_0,v_0)\|\le L_n,$$

which implies for any $\tau\in[0,1]$

$$\|(u_0,v_0)+\tau\cdot((u_1,v_1)-(u_0,v_0))-(u_0,v_0)\|\le\lambda\cdot L_n.$$

Consequently we can conclude from (13) that (16) holds for $t=0$, from which we get the assertion of step 3 by applying the result from step 2.

In the *fourth step of the proof* we show that by induction on t, that

$$F(u_{t+1},v_{t+1})-F(u_t,v_t)\le-\frac{1}{2}\cdot\lambda\cdot\|(\nabla_uF)(u_t,v_t)\|^2-\frac{1}{2}\cdot\lambda\cdot\|(\nabla_vF)(u_t,v_t)\|^2$$

holds for all $t\in\{0,1,\dots,t_n-1\}$, and that

$$\|u_t-u_0\|\le\sqrt{2\cdot F(u_0,v_0)}\quad\text{and}\quad\|v_t-v_0\|\le\sqrt{2\cdot F(u_0,v_0)}$$

hold for all $t\in\{0,1,\dots,t_n\}$.

For $t=0$ the assertion follows from step 3. So assume now that the assertion holds for some $t\in\{0,1,\dots,t_n-1\}$. Then (12) implies

$$\|(\nabla_{(u,v)}F)(u_t,v_t)\|\le L_n,$$

which implies for any $\tau \in [0,1]$

$$\|(u_t, v_t) + \tau \cdot ((u_{t+1}, v_{t+1}) - (u_t, v_t)) - (u_0, v_0)\| \leq \|(u_t, v_t) - (u_0, v_0)\| + \lambda \cdot L_n$$
$$\leq \sqrt{\|u_t - u_0\|^2 + \|v_t - v_0\|^2} + \lambda \cdot L_n \leq 2 \cdot \sqrt{F(u_0, v_0)} + \lambda \cdot L_n.$$

Consequently we can conclude from (13) that (16) holds, from which we get by step 2

$$F(u_{t+1}, v_{t+1}) - F(u_t, v_t) \leq -\frac{1}{2} \cdot \lambda \cdot \|(\nabla_u F)(u_t, v_t)\|^2 - \frac{1}{2} \cdot \lambda \cdot \|(\nabla_v F)(u_t, v_t)\|^2.$$

Furthermore, we have

$$\begin{aligned}
&\|u_{t+1} - u_0\| \\
&\leq \sum_{s=0}^{t} \|u_{s+1} - u_s\| \\
&\leq \sqrt{(t+1) \cdot \sum_{s=0}^{t} \|u_{s+1} - u_s\|^2} \\
&= \sqrt{(t+1) \cdot \sum_{s=0}^{t} \lambda^2 \cdot \|(\nabla_u F)(u_s, v_s)\|^2} \\
&\leq \sqrt{2 \cdot (t+1) \cdot \lambda \cdot \sum_{s=0}^{t} (F(u_s, v_s) - F(u_{s+1}, v_{s+1}))} \\
&\leq \sqrt{2 \cdot (t+1) \cdot \frac{1}{t_n} \cdot F(u_0, v_0)} \\
&\leq \sqrt{2 \cdot F(u_0, v_0)}
\end{aligned}$$

and

$$\begin{aligned}
&\|v_{t+1} - v_0\| \\
&\leq \sum_{s=0}^{t} \|v_{s+1} - v_s\| \\
&\leq \sqrt{(t+1) \cdot \sum_{s=0}^{t} \|v_{s+1} - v_s\|^2} \\
&= \sqrt{(t+1) \cdot \sum_{s=0}^{t} \lambda^2 \cdot \|(\nabla_v F)(u_s, v_s)\|^2} \\
&\leq \sqrt{2 \cdot (t+1) \cdot \lambda \cdot \sum_{s=0}^{t} (F(u_s, v_s) - F(u_{s+1}, v_{s+1}))}
\end{aligned}$$

$$\leq \sqrt{2 \cdot (t+1) \cdot \frac{1}{t_n} \cdot F(u_0, v_0)}$$
$$\leq \sqrt{2 \cdot F(u_0, v_0)}.$$

In the *fifth step of the proof* we show the assertion. By the fourth step we know that F is monotonically decreasing hence it holds

$$F(u_{t_n}, v_{t_n}) \leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} F(u_t, v_t).$$

Then we can conclude from the first step of the proof

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

By (14) and the fourth step of the proof we get

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

And as in the fourth step of the proof we get

$$\sum_{t=0}^{t_n-1} \lambda \cdot \|(\nabla_u F)(u_t, v_t)\|^2 \leq 2 \cdot \sum_{t=0}^{t_n-1} (F(u_t, v_t) - F(u_{t+1}, v_{t+1})) \leq 2 \cdot F(u_0, v_0).$$

Summarizing the above results, the proof is complete. □

With the following two results we can show that the assumptions (12) and (13) in the proof of Theorem 1 are satisfied.

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

$$|w_{1,1,k}^{(L)}| \leq \gamma_n^* \quad \textit{for } k = 1, \ldots, K_n, \tag{18}$$

$$|w_{k,i,j}^{(l)}| \leq B_n \quad \textit{for } l = 1, \ldots, L-1 \tag{19}$$

and

$$\|\mathbf{w} - \mathbf{v}\|_\infty^2 \leq \frac{8 t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}. \tag{20}$$

Then we have for $X_1, \dots, X_n \in [-\alpha_n, \alpha_n]^d$

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w})\| \leq c_8 \cdot K_n^{3/2} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2 \cdot \sqrt{\frac{t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}}.$$

Proof. The proof follows from Lemma 2 in Drews and Kohler (2022). For sake of completeness the proof is given in the appendix. □

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

$$|\max\{(\mathbf{w}_1)_{1,1,k}^{(L)}, (\mathbf{w}_2)_{1,1,k}^{(L)}\}| \leq \gamma_n^* \quad \text{for } k = 1, \dots, K_n, \tag{21}$$

$$|\max\{(\mathbf{w}_1)_{k,i,j}^{(l)}, (\mathbf{w}_2)_{k,i,j}^{(l)}\}| \leq B_n \quad \text{for } l = 1, \dots, L-1 \tag{22}$$

and

$$\|\mathbf{w}_2 - \mathbf{v}\|^2 \leq 8 \cdot \frac{t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}. \tag{23}$$

Then we have for $X_1, \dots, X_n \in [-\alpha_n, \alpha_n]^d$

$$\begin{aligned}&\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}_2)\| \\ &\leq c_9 \cdot \max\{\sqrt{F_n(\mathbf{v})}, 1\} \cdot (\gamma_n^*)^2 \cdot B_n^{3L} \cdot \alpha_n^3 \cdot K_n^{3/2} \cdot \sqrt{\frac{t_n}{L_n}} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|.\end{aligned}$$

Proof. The proof follows from Lemma 3 in Drews and Kohler (2022). For the sake of completeness the complete proof is given in the appendix. □

The next auxiliary result uses a metric entropy bound to control the complexity of a set of over-parametrized deep neural networks.

Lemma 4. *Let* $\alpha \geq 1$, $\beta > 0$ *and let* $A, B, C \geq 1$. *Let* $\sigma : \mathbb{R} \to \mathbb{R}$ *be* k*-times differentiable such that all derivatives up to order* k *are bounded on* $\mathbb{R}$. *Let* $\mathcal{F}$ *be the set of all functions* $f_{\mathbf{w}}$ *defined by (1)–(3) where the weight vector* $\mathbf{w}$ *satisfies*

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

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

and

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

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

$$\begin{aligned}&\mathcal{N}_p\left(\epsilon, \{T_\beta f \cdot 1_{[-\alpha,\alpha]^d} : f \in \mathcal{F}\}, x_1^n\right) \\ &\leq \left(c_{10} \cdot \frac{\beta^p}{\epsilon^p}\right)^{c_{11} \cdot \alpha^d \cdot B^{(L-1)\cdot d} \cdot A^d \cdot \left(\frac{C}{\epsilon}\right)^{d/k} + c_{12}}.\end{aligned}$$

Proof. See Lemma 4 in Drews and Kohler (2022). □

The next lemma gives a bound on the error of the approximation of a Lipschitz continuous and bounded function by an over-parametrized deep neural network.

Lemma 5. *Let σ be the logistic squasher, let $1 \leq \alpha_n \leq \log n$, let $m : \mathbb{R}^d \to \mathbb{R}$ be Lipschitz continuous as well as bounded, let $L, r, n \in \mathbb{N}$ with $L \geq 2$, $r \geq 2d$, $n \geq 8d$ and $n \geq \exp(r+1)$ and let $K, N_n \in \mathbb{N}$ with $2 \leq K \leq \alpha_n - 1$ and $N_n \cdot (K^2+1)^d \leq K_n$. Given $u_1, v_1, \ldots, u_{N_n(K^2+1)^d}, v_{N_n(K^2+1)^d} \in [-K - \frac{2}{K}, K]^d$, choose $\mathbf{w}$ such that*

$$w^{(0)}_{k,j,j} = 4d \cdot K^2 \cdot (\log n)^2 \quad \text{and} \quad w^{(0)}_{k,j,0} = -4d \cdot K^2 \cdot (\log n)^2 \cdot u^{(j)}_k \quad \text{for } j \in \{1, \ldots, d\}, \tag{27}$$

$$w^{(0)}_{k,j+d,j} = -4d \cdot K^2 \cdot (\log n)^2 \quad \text{and} \quad w^{(0)}_{k,j+d,0} = 4d \cdot K^2 \cdot (\log n)^2 \cdot v^{(j)}_k \quad \text{for } j \in \{1, \ldots, d\}, \tag{28}$$

$$w^{(0)}_{k,s,t} = 0 \quad \text{if } s \leq 2d, s \neq t, s \neq t+d \text{ and } t > 0, \tag{29}$$

$$w^{(1)}_{k,1,t} = 8 \cdot (\log n)^2 \quad \text{for } t \in \{1, \ldots, 2d\}, \tag{30}$$

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

$$w^{(1)}_{k,1,t} = 0 \quad \text{for } t > 2d, \tag{32}$$

$$w^{(l)}_{k,1,1} = 6 \cdot (\log n)^2 \quad \text{for } l \in \{2, \ldots, L\}, \tag{33}$$

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

and

$$w^{(l)}_{k,1,t} = 0 \quad \text{for } t > 1 \text{ and } l \in \{2, \ldots, L\} \tag{35}$$

for all $k \in \{1, \ldots, N_n \cdot (K^2+1)^d\}$.

Then there exists $u_1, v_1, \ldots, u_{N_n(K^2+1)^d}, v_{N_n(K^2+1)^d} \in [-K - \frac{2}{K}, K]^d$ and

$$\alpha_1, \ldots, \alpha_{N_n(K^2+1)^d} \in \left[-\frac{\|m\|_\infty}{N_n}, \frac{\|m\|_\infty}{N_n}\right] \tag{36}$$

such that for all pairwise distinct $j_1, \ldots, j_{N_n(K^2+1)^d} \in \{1, \ldots, K_n\}$

$$\begin{aligned} &\int |f_{\bar{\mathbf{w}}}(x) - m(x)|^2 \, \mathbf{P}_X(dx) \\ &\leq c_{13} \cdot \left(\frac{1}{K} + \frac{N_n^2 \cdot K^{4d}}{n^2} + \left(\frac{K^{2d}}{n} + 1\right)^2 \cdot \mathbf{P}_X(\mathbb{R}^d \setminus [-K, K]^d) \right) \end{aligned} \tag{37}$$

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

$$\bar{w}^{(L)}_{1,1,j_k} = \alpha_k \ (k \in \{1, \ldots, N_n \cdot (K^2+1)^d\}), \quad \bar{w}^{(L)}_{1,1,k} = 0 \ (k \notin \{j_1, \ldots, j_{N_n(K^2+1)^d}\}) \tag{38}$$

and

$$|w_{s,k,i}^{(l)} - \bar{w}_{j_s,k,i}^{(l)}| \leq \log n \quad \textit{for all } l \in \{0,\dots,L-1\}, s \in \{1,\dots,N_n \cdot (K^2+1)^d\}. \tag{39}$$

Additionally we get

$$\|f_{\bar{\mathbf{w}}}\|_\infty \leq c_{14} \cdot \left(3^d + \frac{(K^2+1)^d}{n}\right) \tag{40}$$

where c_{14} depends on $\|m\|_\infty$.

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

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

Let C_{Lip} be the Lipschitz constant of m.

We apply Lemma 6 from Drews and Kohler (2022) with $a_i = -K - \frac{2}{K}$, $b_i = K$ and K^2+1 instead of K to m/N_n and $\delta = \frac{1}{K^2}$. This results in

$$\left|f_{\bar{\mathbf{w}}}(x) - \frac{1}{N_n} \cdot m(x)\right| \leq c_{15} \cdot \left(\frac{C_{Lip}}{N_n} \cdot \frac{2}{K} + (K^2+1)^d \cdot \frac{1}{n}\right)$$

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

$$A := \bigcup_{i \in \{0,1,\dots,K^2+1\}} \bigcup_{j \in \{1,\dots,d\}} \left\{x \in \mathbb{R}^d : \left|x^{(j)} - \left(-K - \frac{2}{K} + i \cdot \frac{2}{K}\right)\right| < \delta\right\}. \tag{41}$$

We repeat the whole construction N_n many times. Thus we obtain an approximation $f_{\bar{\mathbf{w}}}$ of

$$N_n \cdot \frac{1}{N_n} \cdot m(x)$$

which satisfies

$$|f_{\bar{\mathbf{w}}}(x) - m(x)| \leq c_{16} \cdot \left(\frac{1}{K} + N_n \cdot (K^2+1)^d \cdot \frac{1}{n}\right) \tag{42}$$

for $x \notin A$.

Next we shift the grid along the j-th component so that $[-K,K]^d$ is always covered. This means we modify all $u_i^{(j)}, v_i^{(j)}$ by the same additional summand which is chosen from the set

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

for fixed $j \in \{1,\dots,d\}$. Thus we obtain K different versions of $f_{\bar{\mathbf{w}}}$ that still satisfy (42) for all $x \in [-K,K]^d$ up to corresponding versions of A.

Since we shift the grid of cubes we obtain for fixed $j \in \{1,\dots,d\}$ K disjoint versions of $\bigcup_{i \in \{0,1,\dots,K^2+1\}} \left\{x \in \mathbb{R}^d : \left|x^{(j)} - \left(-K - \frac{2}{K} + i \cdot \frac{2}{K}\right)\right| < \delta\right\}$. The sum of $\mathbf{P}_X$-measures

of these K disjoint sets is less than or equal to one. Therefore at least one of them must have measure less than or equal to $\frac{1}{K}$. Consequently we can shift u_i and v_i so that

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

holds.

Now we have shown that there exists a shifted version of the grid such that the set A has a measure less than or equal to $\frac{d}{K}$. By inequality (42) we get that $|f_{\bar{\mathbf{w}}}(x) - m(x)| \leq c_{16} \cdot \left(\frac{1}{K} + \frac{N_n \cdot K^{2d}}{n} \right)$ holds for $x \in [-K, K]^d \setminus A$.

From the second assertion of Lemma 6 in Drews and Kohler (2022) we obtain

$$|f_{\bar{\mathbf{w}}}(x)| \leq \|m\|_\infty \cdot \left(3^d + (K^2+1)^d \cdot \frac{1}{n} \right)$$

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

Summarizing the above results we get

$$\begin{aligned}
&\int |f_{\bar{\mathbf{w}}}(x) - m(x)|^2 \, \mathbf{P}_X(dx) \\
&= \int_{[-K,K]^d \setminus A} |f_{\bar{\mathbf{w}}}(x) - m(x)|^2 \, \mathbf{P}_X(dx) + \int_A |f_{\bar{\mathbf{w}}}(x) - m(x)|^2 \, \mathbf{P}_X(dx) \\
&\quad + \int_{\mathbb{R}^d \setminus [-K,K]^d} |f_{\bar{\mathbf{w}}}(x) - m(x)|^2 \, \mathbf{P}_X(dx) \\
&\leq c_{16}^2 \left(\frac{1}{K} + \frac{N_n \cdot K^{2d}}{n} \right)^2 + c_{17} \left(3^d + \frac{K^{2d}}{n} \right)^2 \cdot \frac{d}{K} \\
&\quad + c_{18} \left(3^d + \frac{K^{2d}}{n} \right)^2 \cdot \mathbf{P}_X(\mathbb{R}^d \setminus [-K,K]^d),
\end{aligned}$$

which implies the assertion.

□

4.2. Proof of Theorem 1

Proof. Let $\epsilon > 0$, $K \in \mathbb{N}$ arbitrary and $N_n = \lceil c_{19} \cdot (\log n)^{18L} \rceil$. Furthermore, let $\bar{m} : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz continuous and bounded function such that

$$\int |\bar{m}(x) - m(x)|^2 \mathbf{P}_X(dx) \leq \epsilon. \tag{43}$$

We denote by A_n the event that firstly there exists pairwise disjoint $j_1, \dots, j_{N_n(K^2+1)^d}$ such that the weight vector $\mathbf{w}^{(0)}$ satisfies

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

for some weight vector $\mathbf{w}$ which satisfies the conditions (27)–(35) of Lemma 5 for $\bar{m}$ and that secondly the inequality

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

holds.

In case that A_n holds α_k is chosen as in Lemma 5 for $\bar{m}$, otherwise we set $\alpha_1 = ... = \alpha_{N_n(K^2+1)^d} = 0$. Then we define the weight vectors $\mathbf{w}^*$ for given $\mathbf{w}$ by

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

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

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

and $(\mathbf{w}^*)^{(0)}$ by

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

$$((\mathbf{w}^*)^{(0)})_{1,1,j_k}^{(L)} = \alpha_k \quad \text{for all } k = 1, \dots, N_n \cdot (K^2+1)^d,$$

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

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

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

$$= \sum_{j=1}^{6} T_{j,n}.$$

In the *second step of the proof* we show

$$\limsup_{n\to\infty} \mathbf{E} T_{j,n} \leq 0 \quad \text{for } j \in \{1,4\}.$$

By using $(a+b)^2 \leq (1+\epsilon)\cdot a^2 + (1+\frac{1}{\epsilon})\cdot b^2$ for $a,b \in \mathbb{R}$ we obtain

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

and

$$\mathbf{E} T_{4,n} \leq (1+\epsilon) \cdot \left(1+\frac{1}{\epsilon}\right) \cdot \mathbf{E}\{|T_{\beta_n} Y - Y|^2\}.$$

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

In the *third step of the proof* we show

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

If $|y| \leq \beta_n$ then it holds

$$|T_{\beta_n} z - y| \leq |z-y|$$

for any $z \in \mathbb{R}$. This implies

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - T_{\beta_n} Y_i|^2 &= \frac{1}{n}\sum_{i=1}^{n} |T_{\beta_n} f_{\mathbf{w}^{(t_n)}}(X_i) - T_{\beta_n} Y_i|^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n} |f_{\mathbf{w}^{(t_n)}}(X_i) - T_{\beta_n} Y_i|^2.\end{aligned}$$

Thus the assertion of the third step holds.

In the *fourth step of the proof* we show that the assumptions (12) - (14) of Lemma 1 are satisfied which means that

$$\|(\nabla_{\mathbf{w}} F)(\mathbf{w})\| \leq L_n$$

for all $\mathbf{w} \in S := \left\{\mathbf{v} \,:\, \|\mathbf{v} - \mathbf{w}^{(0)}\| \leq 2 \cdot \sqrt{F(\mathbf{w}^{(0)})} + 1\right\}$,

$$\|(\nabla_{\mathbf{w}} F)(\mathbf{w}) - (\nabla_{\mathbf{w}} F)(\tilde{\mathbf{w}})\| \leq L_n \cdot \|\mathbf{w} - \tilde{\mathbf{w}}\|$$

for all $\mathbf{w}, \tilde{\mathbf{w}} \in S$ and

$$\begin{aligned}&|F(\mathbf{w}^*) - F((\mathbf{w}^*)^{(0)})| \\ &\leq D_n \cdot \|((\mathbf{w}^*)^{(L)}_{1,1,k})_{k=1,\ldots,K_n}\| \cdot \|(\mathbf{w}^{(l)}_{i,j,k})_{i,j,k,l:l<L} - ((\mathbf{w}^{(0)})^{(l)}_{i,j,k})_{i,j,k,l:l<L}\|\end{aligned}$$

for all

$$(\mathbf{w}_{i,j,k}^{(l)})_{i,j,k,l:l<L} \in \tilde{S}$$
$$:= \left\{ (\tilde{\mathbf{w}}_{i,j,k}^{(l)})_{i,j,k,l:l<L} : \|(\tilde{\mathbf{w}}_{i,j,k}^{(l)})_{i,j,k,l:l<L} - ((\mathbf{w}^{(0)})_{i,j,k}^{(l)})_{i,j,k,l:l<L}\| \leq \sqrt{2 \cdot F(\mathbf{w}^{(0)})} \right\}$$

hold, if A_n holds.

If A_n holds, then we have

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

Let $\mathbf{w} \in S$. Then we can conclude that

$$\begin{aligned} \|(\mathbf{w}_{i,j,k}^{(l)})_{i,j,k,l:1\leq l<L}\|_\infty &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\| + \|((\mathbf{w}^{(0)})_{i,j,k}^{(l)})_{i,j,k,l:1\leq l<L}\|_\infty \\ &\leq 2 \cdot \sqrt{F_n(\mathbf{w}^{(0)})} + 1 + c_{20} \cdot (\log n)^2 \\ &\leq c_{21} \cdot (\log n)^2 \end{aligned}$$

and

$$\begin{aligned} \|(\mathbf{w}_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\|_\infty &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\| + \|((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\|_\infty \\ &\leq 2 \cdot \sqrt{F_n(\mathbf{w}^{(0)})} + 1 \\ &\leq c_{22} \cdot (\log n)^{3/2}. \end{aligned}$$

Hence (18)-(23) are satisfied for $B_n = c_{21} \cdot (\log n)^2$ and $\gamma_n^* = c_{22} \cdot (\log n)^{3/2}$. By Lemma 2 and Lemma 3 we get for $\alpha_n = c_{23}$ that (12) and (13) are satisfied provided that $L_n \geq K_n^{3/2} \cdot (\log n)^{6L+5}$.

It remains to show that (14) holds. Let $\mathbf{w}$ such that $(\mathbf{w}_{i,j,k}^{(l)})_{i,j,k,l:l<L} \in \tilde{S}$. Then we have

$$\begin{aligned} &|F_n(\mathbf{w}^*) - F_n((\mathbf{w}^*)^{(0)})| \\ &= \left| \frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}^*}(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^{n} |f_{(\mathbf{w}^*)^{(0)}}(X_i) - Y_i|^2 \right| \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(f_{\mathbf{w}^*}(X_i) - Y_i + f_{(\mathbf{w}^*)^{(0)}}(X_i) - Y_i \right) \left(f_{\mathbf{w}^*}(X_i) - f_{(\mathbf{w}^*)^{(0)}}(X_i) \right) \\ &\leq \left(\frac{1}{n} \sum_{i=1}^{n} \left(f_{\mathbf{w}^*}(X_i) - Y_i + f_{(\mathbf{w}^*)^{(0)}}(X_i) - Y_i \right)^2 \right)^{1/2} \\ &\qquad \left(\frac{1}{n} \sum_{i=1}^{n} \left(f_{\mathbf{w}^*}(X_i) - f_{(\mathbf{w}^*)^{(0)}}(X_i) \right)^2 \right)^{1/2} \end{aligned}$$

$$\leq \left(\frac{2}{n} \sum_{i=1}^{n} \left(f_{\mathbf{w}^*}(X_i) + f_{(\mathbf{w}^*)^{(0)}}(X_i) \right)^2 + \frac{8}{n} \sum_{i=1}^{n} Y_i^2 \right)^{1/2}$$
$$\left(\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \sum_{k=1}^{K_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(X_i) - f_{(\mathbf{w}^*)^{(0)},k,1}^{(L)}(X_i) \right|^2 \right) \right)^{1/2}.$$

For the first term we get

$$\left(\frac{2}{n} \sum_{i=1}^{n} \left(f_{\mathbf{w}^*}(X_i) + f_{(\mathbf{w}^*)^{(0)}}(X_i) \right)^2 + \frac{8}{n} \sum_{i=1}^{n} Y_i^2 \right)^{1/2}$$
$$= \left(\frac{2}{n} \sum_{i=1}^{n} \left(f_{(\mathbf{w}^*)}(X_i) - f_{(\mathbf{w}^*)^{(0)}}(X_i) + 2 \cdot f_{(\mathbf{w}^*)^{(0)}}(X_i) \right)^2 + \frac{8}{n} \sum_{i=1}^{n} Y_i^2 \right)^{1/2}$$
$$\leq \left(\frac{2}{n} \sum_{i=1}^{n} 2 \cdot \left(f_{\mathbf{w}^*}(X_i) - f_{(\mathbf{w}^*)^{(0)}}(X_i) \right)^2 + \frac{2}{n} \sum_{i=1}^{n} 8 \cdot f_{(\mathbf{w}^*)^{(0)}}(X_i)^2 + \frac{8}{n} \sum_{i=1}^{n} Y_i^2 \right)^{1/2}$$
$$\leq \left(4 \cdot \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \sum_{k=1}^{K_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(X_i) - f_{(\mathbf{w}^*)^{(0)},k,1}^{(L)}(X_i) \right|^2 \right) \right.$$
$$\left. + 16 \cdot \frac{1}{n} \sum_{i=1}^{n} f_{(\mathbf{w}^*)^{(0)}}(X_i)^2 + \frac{8}{n} \sum_{i=1}^{n} Y_i^2 \right)^{1/2}$$
$$\leq \left(4 \cdot \sum_{k=1}^{K_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \max_{i=1,\ldots,n} \sum_{k=1}^{K_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(X_i) - f_{(\mathbf{w}^*)^{(0)},k,1}^{(L)}(X_i) \right|^2 \right.$$
$$\left. + 16 \cdot \frac{1}{n} \sum_{i=1}^{n} f_{(\mathbf{w}^*)^{(0)}}(X_i)^2 + \frac{8}{n} \sum_{i=1}^{n} Y_i^2 \right)^{1/2}.$$

By Lemma 5 for $\bar{m}$ we get $|\alpha_k| \leq \frac{c_{24}}{N_n}$ and

$$f_{(\mathbf{w}^*)^{(0)}}(X_i) \leq c_{25} \cdot \left(3^d + \frac{(K^2+1)^d}{n} \right).$$

From the proof of Lemma 2 we know that

$$|f_{\mathbf{w}^*,k,1}^{(L)}(x) - f_{(\mathbf{w}^*)^{(0)},k,1}^{(L)}(x)|$$
$$\leq c_{26} \cdot (\log n)^{2L} \cdot \max_{i,j,s:s<L} |(\mathbf{w}^*)_{k,i,j}^{(s)} - ((\mathbf{w}^*)^{(0)})_{k,i,j}^{(s)}| \tag{44}$$

holds. Therefore we obtain for $(\mathbf{w}_{i,j,k}^{(l)})_{i,j,k,l:l<L} \in \tilde{S}$

$$\max_{i=1,\ldots,n} \sum_{k=1}^{K_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(X_i) - f_{(\mathbf{w}^*)^{(0)},k,1}^{(L)}(X_i) \right|^2$$

$$
\begin{aligned}
&\leq c_{27} \cdot (\log n)^{4L} \cdot \sum_{k=1}^{K_n} \max_{i,j,s:s<L} |(\mathbf{w}^*)_{k,i,j}^{(s)} - ((\mathbf{w}^*)^{(0)})_{k,i,j}^{(s)}|^2 \\
&\leq c_{27} \cdot (\log n)^{4L} \cdot \|(\mathbf{w}_{i,j,k}^{(l)})_{i,j,k,l;l<L} - ((\mathbf{w}^{(0)})_{i,j,k}^{(l)})_{i,j,k,l;l<L}\|^2 \\
&\leq c_{27} \cdot (\log n)^{4L+3}.
\end{aligned}
$$

From this together with the definition of $\mathbf{w}^*$ we can conclude

$$
\begin{aligned}
&\left(\frac{2}{n} \sum_{i=1}^{n} \left(f_{\mathbf{w}^*}(X_i) + f_{(\mathbf{w}^*)^{(0)}}(X_i) \right)^2 + \frac{8}{n} \sum_{i=1}^{n} Y_i^2 \right)^{1/2} \\
&\leq \left(4 \cdot \sum_{k=1}^{K_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot c_{27} \cdot (\log n)^{4L+3} + 16 \cdot \frac{1}{n} \sum_{i=1}^{n} f_{(\mathbf{w}^*)^{(0)}}(X_i)^2 + \frac{8}{n} \sum_{i=1}^{n} Y_i^2 \right)^{1/2} \\
&\leq \left(c_{28} \cdot \frac{N_n \cdot (K^2+1)^d}{N_n^2} \cdot (\log n)^{4L+3} + \left(c_{29} \cdot \left(3^d + \frac{(K^2+1)^d}{n} \right) \right)^2 + 8 \cdot c_1^3 \cdot (\log n)^3 \right)^{1/2} \\
&\leq c_{30} \left(\frac{(K^2+1)^d}{N_n} \cdot (\log n)^{4L+3} + \left(\frac{(K^2+1)^d}{n} \right)^2 + (\log n)^3 \right)^{1/2}.
\end{aligned}
$$

Furthermore, due to (44) we have

$$
\begin{aligned}
&\left(\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \sum_{k=1}^{K_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(X_i) - f_{(\mathbf{w}^*)^{(0)},k,1}^{(L)}(X_i) \right|^2 \right) \right)^{1/2} \\
&\leq \left(\sum_{k=1}^{K_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \sum_{k=1}^{K_n} \max_{i=1,\ldots,n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(X_i) - f_{(\mathbf{w}^*)^{(0)},k,1}^{(L)}(X_i) \right|^2 \right)^{1/2} \\
&\leq \|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\| \cdot c_{31} \cdot (\log n)^{2L} \cdot \|(\mathbf{w}_{i,j,k}^{(l)})_{i,j,k,l:l<L} - ((\mathbf{w}^{(0)})_{i,j,k}^{(l)})_{i,j,k,l:l<L}\|.
\end{aligned}
$$

This yields

$$
\begin{aligned}
&|F_n((\mathbf{w}^*)^{(t)}) - F_n((\mathbf{w}^*)^{(0)})| \\
&\leq c_{32} \left(\frac{(K^2+1)^{d/2}}{N_n^{1/2}} \cdot (\log n)^{4L+3/2} + \frac{(K^2+1)^d \cdot (\log n)^{2L}}{n} + (\log n)^{2L+3/2} \right) \\
&\qquad \cdot \|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\| \cdot \|(\mathbf{w}_{i,j,k}^{(l)})_{i,j,k,l:l<L} - ((\mathbf{w}^{(0)})_{i,j,k}^{(l)})_{i,j,k,l:l<L}\|.
\end{aligned}
$$

Thus (14) is satisfied with

$$
D_n = c_{32} \cdot (\log n)^{2L} \cdot \left(\frac{(K^2+1)^{d/2}}{N_n^{1/2}} \cdot (\log n)^{2L+3/2} + \frac{(K^2+1)^d}{n} + (\log n)^{3/2} \right).
$$

For n large we get

$$
D_n \leq c_{33} \cdot (\log n)^{4L+2}.
$$

In the *fifth step of the proof* we show

$$\mathbf{P}(A_n^c) \leq \frac{c_{34}}{(\log n)^3}.$$

To show this, we first bound the probability that the weight vector $\mathbf{w}^{(0)}$ does not satisfy the first condition of the event A_n. For this, we consider a sequential choice of weights in the K_n fully connected neural networks. Each of these K_n fully connected neural networks contains $(r+1)+(L-2)\cdot r\cdot(r+1)+r\cdot(d+1)$ weights. Therefore, the probability that all these weights never satisfy condition (39) for $s=1$ is bounded from below by

$$\left(\frac{\log n}{40d\cdot(\log n)^2}\right)^{(r+1)+(L-2)\cdot r\cdot(r+1)}\cdot\left(\frac{\log n}{16d\cdot(\log n)^2\cdot n^\tau}\right)^{r\cdot(d+1)}.$$

Consequently, the probability that this condition is never satisfied in the first $n^{r(d+1)\tau+1}$ many fully connected neural networks for j_1 is for large n bounded from above by

$$\left(1-\left(\frac{1}{40d\cdot\log n}\right)^{(r+1)+(L-2)\cdot r\cdot(r+1)}\cdot\left(\frac{1}{16d\cdot\log n\cdot n^\tau}\right)^{r\cdot(d+1)}\right)^{n^{r(d+1)\tau+1}}$$
$$\leq\left(1-n^{-r(d+1)\tau-0{,}5}\right)^{n^{r(d+1)\tau+1}}.$$

Because of condition (6) we have $K_n \geq N_n\cdot(K^2+1)^d\cdot n^{r(d+1)\tau+1}$ for large n. This implies that for large n condition (39) is satisfied outside of an event of probability

$$\begin{aligned}
&N_n\cdot(K^2+1)^d\cdot\left(1-n^{-r(d+1)\tau-0.5}\right)^{n^{r(d+1)\tau+1}}\\
&\leq N_n\cdot(K^2+1)^d\cdot\left(\exp\left(-n^{-r(d+1)\tau-0.5}\right)\right)^{n^{r(d+1)\tau+1}}\\
&\leq N_n\cdot(K^2+1)^d\cdot\exp\left(-n^{0.5}\right)\\
&\leq n^\kappa\cdot\exp\left(-n^{0.5}\right)\\
&\leq\frac{c_{35}}{n}.
\end{aligned}$$

Then we obtain for large n by Markov's inequality

$$\begin{aligned}
\mathbf{P}(A^c) &\leq \frac{c_{35}}{n}+\mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^n Y_i^2>\beta_n^3\right\}\\
&\leq\frac{c_{35}}{n}+\frac{\mathbf{E}\{\frac{1}{n}\sum_{i=1}^n Y_i^2\}}{\beta_n^3}\\
&\leq\frac{c_{35}}{n}+\frac{\mathbf{E}\{Y^2\}}{\beta_n^3}
\end{aligned}$$

$$\leq \frac{c_{36}}{(\log n)^3}$$

where the last inequality holds since $\mathbf{E}Y^2 < \infty$.

In the *sixth step of the proof* we show

$$\limsup_{n\to\infty} \mathbf{E}T_{2,n} \leq 0.$$

We have

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

W.l.o.g. we can assume that A_n holds. Hence, by Lemma 1, it follows that

$$\|((\mathbf{w}^{(t_n)})_{i,j,k}^{(l)})_{i,j,k,l:l<L} - ((\mathbf{w}^{(0)})_{i,j,k}^{(l)})_{i,j,k,l:l<L}\| \leq c_{37} \cdot (\log n)^{3/2}$$

and

$$\|((\mathbf{w}^{(t_n)})_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\| \leq c_{38} \cdot (\log n)^{3/2}.$$

hold. Consequently, we obtain

$$\begin{aligned}
&\|((\mathbf{w}^{(t_n)})_{i,j,k}^{(l)})_{i,j,k,l:1\leq l<L}\|_\infty \\
&\leq \|((\mathbf{w}^{(t_n)})_{i,j,k}^{(l)})_{i,j,k,l:1\leq l<L} - ((\mathbf{w}^{(0)})_{i,j,k}^{(l)})_{i,j,k,l:1\leq l<L}\|_\infty + \|((\mathbf{w}^{(0)})_{i,j,k}^{(l)})_{i,j,k,l:1\leq l<L}\|_\infty \\
&\leq c_{37} \cdot (\log n)^{3/2} + c_{39} \cdot (\log n)^2
\end{aligned}$$

and

$$\begin{aligned}
&\|((\mathbf{w}^{(t_n)})_{i,j,k}^{(0)})_{i,j,k}\|_\infty \\
&\leq \|((\mathbf{w}^{(t_n)})_{i,j,k}^{(0)})_{i,j,k} - ((\mathbf{w}^{(0)})_{i,j,k}^{(0)})_{i,j,k}\|_\infty + \|((\mathbf{w}^{(0)})_{i,j,k}^{(0)})_{i,j,k}\|_\infty \\
&\leq c_{38} \cdot (\log n)^{3/2} + c_{40} \cdot (\log n)^2 \cdot n^\tau.
\end{aligned}$$

This implies that m_n is contained in the function space

$$\{T_{\beta_n} f \,:\, f \in \mathcal{F}\}$$

where $\mathcal{F}$ is defined as in Lemma 4 with $C = c_{41} \cdot K_n \cdot (\log n)^{3/2}$, $B = c_{42} \cdot (\log n)^2$ and $A = c_{43} \cdot (\log n)^2 \cdot n^{\tau}$. Thus, with Lemma 4 and standard bounds of empirical process theory (cf., Theorem 9.1 in Györfi et al. (2002)), it follows

$$\begin{aligned}
&\mathbf{P}\left\{ (\mathbf{E}\{|m_n(X) - T_{\beta_n} Y|^2(X)|\mathcal{D}_n\} - \frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - T_{\beta_n} Y_i|^2) \cdot 1_{A_n} > t \right\} \\
&\leq 8 \cdot \left(c_{44} \cdot \frac{\beta_n}{t/8} \right)^{c_{45} \cdot (\log n)^{c_{46}} \cdot n^{\tau \cdot d} \cdot \left(\frac{c_{47} \cdot K_n \cdot (\log n)^{3/2}}{t/8} \right)^{d/k} + c_{48}} \cdot \exp\left(-\frac{n \cdot t^2}{128 \cdot \beta_n^4} \right).
\end{aligned}$$

Using (5) and $\tau = \frac{1}{d+1}$ we get for $k > (\kappa+1) \cdot d \cdot (d+1) \cdot (d+2)$ and $t > n^{\frac{-1}{4(d+2)}}$ that the left hand side above is for n large enough less than or equal to

$$\begin{aligned}
&8 \cdot \left(c_{44} \cdot \frac{\beta_n}{t/8} \right)^{c_{45} \cdot (\log n)^{c_{46}} \cdot n^{\tau \cdot d} \cdot \left(\frac{c_{47} \cdot K_n \cdot (\log n)^{3/2}}{t/8} \right)^{d/k} + c_{48}} \\
&\qquad \cdot \exp\left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \cdot \exp\left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \\
&\leq \exp\left(c_{49} \cdot (\log n)^{c_{46}} \cdot n^{\tau \cdot d + (\kappa+1) \cdot \frac{d}{k}} \cdot \log\left(c_{44} \cdot \frac{\beta_n}{t/8} \right) \right) \\
&\qquad \cdot \exp\left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \cdot \exp\left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \\
&\leq \exp\left(c_{50} \cdot \left((\log n)^{c_{51}} \cdot n^{\frac{d}{d+1} + (\kappa+1)\frac{d}{k}} - \frac{n^{\frac{2d+3}{2(d+2)}}}{(\log n)^4} \right) \right) \cdot \exp\left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \\
&\leq \exp\left(c_{50} \cdot \left((\log n)^{c_{51}} \cdot n^{\frac{d+1}{d+2}} - \frac{n^{\frac{2d+3}{2(d+2)}}}{(\log n)^4} \right) \right) \cdot \exp\left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \\
&= \exp\left(c_{50} \cdot n^{\frac{d+1}{d+2}} \cdot \left((\log n)^{c_{50}} - \frac{n^{\frac{1}{2(d+2)}}}{(\log n)^4} \right) \right) \cdot \exp\left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \\
&\leq c_{52} \cdot \exp\left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \\
&\leq c_{53} \cdot \exp\left(-\frac{n^{\frac{2d+3}{2(d+2)}}}{256 \cdot \beta_n^4} \right)
\end{aligned}$$

holds. Therefore, we obtain

$$\mathbf{E}\left\{T_{2,n}\right\} \leq (1+\epsilon) \cdot \left(n^{\frac{-1}{4(d+2)}} + 4\beta_n^2 \cdot c_{54} \cdot \exp\left(-\frac{n^{\frac{2d+3}{2(d+2)}}}{256 \cdot \beta_n^4} \right) \right) \to 0 \quad (n \to \infty).$$

In the *seventh step of the proof* we show

$$\limsup_{n \to \infty} \mathbf{E}\{T_{6,n}\} \leq 0.$$

Due to the assertion of the fifth step together with the integrability of m we get

$$\begin{aligned}\mathbf{E}\{T_{6,n}\} &\leq \left(2 \cdot \int |m_n(x)|^2 \, \mathbf{P}_X(dx) + 2 \cdot \int |m(x)|^2 \, \mathbf{P}_X(dx)\right) \cdot \mathbf{P}(A_n^c) \\ &\leq 2 \cdot (\beta_n^2 + c_{55}) \cdot \frac{c_{36}}{(\log n)^3}.\end{aligned}$$

This implies the assertion of the seventh step.

In the *eighth step of the proof* we bound

$$\mathbf{E}T_{5,n}.$$

If A_n holds, then as shown in step four, we can apply Lemma 1 with

$$D_n \leq c_{35} \cdot (\log n)^{4L+2}.$$

This together with the definition of $\mathbf{w}^*$ and $(\mathbf{w}^*)^{(0)}$ yields

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} |f_{\mathbf{w}^{(t_n)}}(X_i) - Y_i|^2 = F_n\left(\mathbf{w}^{(t_n)}\right) \\ &\leq F_n\left((\mathbf{w}^*)^{(0)}\right) + D_n \cdot \|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\| \cdot \sqrt{2 \cdot F_n\left(\mathbf{w}^{(0)}\right)} \\ &\qquad + \frac{\|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\ldots,K_n} - ((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\|^2}{2} + \frac{F_n\left(\mathbf{w}^{(0)}\right)}{t_n} \\ &\leq F_n\left((\mathbf{w}^*)^{(0)}\right) + c_{56} \cdot (\log n)^{4L+2} \cdot \frac{N_n^{1/2} \cdot (K^2+1)^{d/2}}{N_n} \cdot (\log n)^{3/2} \\ &\qquad + \frac{N_n \cdot (K^2+1)^d}{2 \cdot N_n^2} + \frac{c_1^3 \cdot (\log n)^3}{t_n} \\ &\leq F_n\left((\mathbf{w}^*)^{(0)}\right) + c_{57} \cdot \frac{(K^2+1)^d}{{N_n}^{1/2}} \cdot (\log n)^{4L+4}.\end{aligned}$$

Thus we obtain

$$\begin{aligned}&\mathbf{E}\left\{T_{5,n}\right\} \\ &= (1+\epsilon)^2 \cdot \mathbf{E}\left\{\left(F_n\left(\mathbf{w}^{(t_n)}\right) - \mathbf{E}\{|m(X)-Y|^2\}\right) \cdot 1_{A_n}\right\} \\ &\qquad + ((1+\epsilon)^2 - 1) \cdot \mathbf{E}\left\{\mathbf{E}\{|m(X)-Y|^2\} \cdot 1_{A_n}\right\} \\ &\leq (1+\epsilon)^2 \cdot \left(\mathbf{E}\left\{F_n\left((\mathbf{w}^*)^{(0)}\right) \cdot 1_{A_n} + c_{57}\left(\frac{(K^2+1)^d}{{N_n}^{1/2}} \cdot (\log n)^{4L+4}\right) \cdot 1_{A_n}\right\}\right. \\ &\qquad \left. - \mathbf{E}\{|m(X)-Y|^2\} \cdot \mathbf{P}(A_n)\right) + ((1+\epsilon)^2 - 1) \cdot \mathbf{E}\{|m(X)-Y|^2\}.\end{aligned}$$

Let $\tilde{A}_n$ be the event where the weight vector $\mathbf{w}^{(0)}$ satisfies

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

for some weight vector $\mathbf{w}$ which satisfies conditions (27)–(35) of Lemma 5 for $\bar{m}$. Then we get from the fifth step of the proof

$$\mathbf{P}(\tilde{A}_n) - \mathbf{P}(A_n) \le \mathbf{P}\left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i^2 > \beta_n^3 \right\} \le \frac{c_{58}}{(\log n)^3}.$$

This together with the fact that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent of $\tilde{A}_n$ yields

$$\begin{aligned}
&\mathbf{E}\{F_n((\mathbf{w}^*)^{(0)}) \cdot 1_{A_n}\} - \mathbf{E}\{|m(X) - Y|^2\} \cdot \mathbf{P}(A_n) \\
&\le \mathbf{E}\Big\{ \frac{1}{n} \sum_{i=1}^{n} |f_{((\mathbf{w}^*)^{(0)})}(X_i) - Y_i|^2 \cdot 1_{\tilde{A}_n} \Big\} - \mathbf{E}\{|m(X) - Y|^2\} \cdot \mathbf{P}(\tilde{A}_n) \\
&\qquad + \mathbf{E}\{|m(X) - Y|^2\} \cdot (\mathbf{P}(\tilde{A}_n) - \mathbf{P}(A_n)) \\
&\le \mathbf{E}\left\{ \mathbf{E}\Big\{ \frac{1}{n} \sum_{i=1}^{n} |f_{((\mathbf{w}^*)^{(0)})}(X_i) - Y_i|^2 \cdot 1_{\tilde{A}_n} \Big| (\mathbf{w}^*)^{(0)} \Big\} - \mathbf{E}\{|m(X) - Y|^2\} \cdot 1_{\tilde{A}_n} \right\} \\
&\qquad + \frac{c_{59}}{(\log n)^3} \\
&= \mathbf{E}\left\{ \left(\mathbf{E}\Big\{ \frac{1}{n} \sum_{i=1}^{n} |f_{((\mathbf{w}^*)^{(0)})}(X_i) - Y_i|^2 \Big| (\mathbf{w}^*)^{(0)} \Big\} - \mathbf{E}\{|m(X) - Y|^2\} \right) \cdot 1_{\tilde{A}_n} \right\} \\
&\qquad + \frac{c_{59}}{(\log n)^3} \\
&\le \mathbf{E}\left\{ \int |f_{((\mathbf{w}^*)^{(0)})}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{\tilde{A}_n} \right\} + \frac{c_{59}}{(\log n)^3}.
\end{aligned}$$

Because of the choice of $\bar{m}$ and Lemma 5 we get for K such large that $supp(X) \subseteq [-K, K]^d$

$$\begin{aligned}
&\mathbf{E}\Big\{ \int |f_{((\mathbf{w}^*)^{(0)})}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{\tilde{A}_n} \Big\} \\
&\le 2 \cdot \mathbf{E}\Big\{ \int |f_{((\mathbf{w}^*)^{(0)})}(x) - \bar{m}(x)|^2 \mathbf{P}_X(dx) \cdot 1_{\tilde{A}_n} \Big\} + 2 \int |\bar{m}(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\le c_{60} \cdot \left(\frac{1}{K} + \frac{N_n^2 \cdot K^{4d}}{n^2} + \left(\frac{K^{2d}}{n} + 1 \right)^2 \cdot \mathbf{P}_X(\mathbb{R}^d \setminus [-K, K]^d) \right) + 2\epsilon \\
&\le c_{61} \cdot \left(\frac{1}{K} + \frac{N_n^2 \cdot K^{4d}}{n^2} \right) + 2\epsilon.
\end{aligned}$$

Due to the definition of N_n we obtain

$$\frac{(K^2+1)^d}{{N_n}^{1/2}} \cdot (\log n)^{4L+4} \to 0 \qquad (n \to \infty).$$

Summarizing the above results yields

$$\limsup_{n \to \infty} \mathbf{E}\{T_{5,n}\} \le c_{62} \cdot (1+\epsilon)^2 \cdot \left(\frac{1}{K} + 2\epsilon \right) + ((1+\epsilon)^2 - 1) \cdot \mathbf{E}\{|m(X) - Y|^2\}.$$

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

$$\begin{aligned}&\limsup_{n\to\infty} \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)\\ &\leq c_{63} \cdot \left((1+\epsilon)^2 \cdot 2\epsilon + ((1+\epsilon)^2 - 1) \cdot \mathbf{E}\{|m(X) - Y|^2\}\right).\end{aligned}$$

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

4.3. Auxiliary results for the proof of Theorem 2

The following theorem is crucial for proving Theorem 2 and Theorem 3. It is needed to analyze the rate of convergence of the over-parametrized deep neural network estimate.

Theorem 4. *Let $n \in \mathbb{N}$ with $n \geq 2$, let $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$-valued random variables such that supp(X) is bounded, that*

$$\mathbf{E}\left\{e^{c_3 \cdot Y^2}\right\} < \infty \tag{45}$$

holds and that the corresponding regression function $m(x) = \mathbf{E}\{Y|X = x\}$ is bounded. Let $\sigma(x) = 1/(1+e^{-x})$ be the logistic squasher, let $K_n, L, r, t_n \in \mathbb{N}$, $M_n \geq 1$ and $\lambda_n, \tau > 0$. Let $\tilde{K}_n \in \{1, \ldots, K_n\}$,

$$w_{k,i,j}^{(l)} \in [-20d \cdot (\log n)^2, 20d \cdot (\log n)^2] \quad \text{for } l = 1, \ldots, L,\ k = 1, \ldots \tilde{K}_n$$

and

$$w_{k,i,j}^{(0)} \in [-8d \cdot (\log n)^2 \cdot n^\tau, 8d \cdot (\log n)^2 \cdot n^\tau] \quad \text{for } k = 1, \ldots, \tilde{K}_n.$$

Assume that

$$\sqrt{\sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2} \leq M_n, \tag{46}$$

and that

$$\left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x) \right| \leq \beta_n \tag{47}$$

holds for $x \in$ supp(X) and for all $\bar{\mathbf{w}}$ which satisfy

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

Assume furthermore

$$\frac{K_n}{n^\kappa} \to 0 \qquad (n \to \infty) \tag{48}$$

for some $\kappa > 0$,

$$\frac{K_n}{\tilde{K}_n \cdot n^{r\cdot(d+1)\cdot\tau+1}} \to \infty \qquad (n \to \infty) \tag{49}$$

Define the estimate m_n *as in Section 2 with*

$$\lambda_n = \frac{1}{t_n} \quad \textit{and} \quad t_n = \lceil c_{64} \cdot L_n \rceil \tag{50}$$

for some $c_{64} \geq 1$ *and for some* $L_n > 0$ *which satisfies*

$$L_n \geq K_n^{3/2} \cdot (\log n)^{6L+2},$$

and assume

$$c_1 \cdot c_3 \geq 2. \tag{51}$$

Then we have for any $\epsilon > 0$

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_{65} \cdot \Bigg(\frac{n^{\tau\cdot d+\epsilon}}{n} + M_n^2 \cdot (\log n)^{4L+3/2}$$

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

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

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

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

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

holds. Then we define the weight vectors $\mathbf{w}^*$ for given $\tilde{\mathbf{w}}$ by

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

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

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

and $(\mathbf{w}^*)^{(0)}$ by

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

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

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

In the following we set

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

Furthermore, we assume w.l.o.g. that n is sufficiently large and that $\|m\|_\infty \leq \beta_n$ holds.

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

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

In the *second step of the proof* we show

$$\mathbf{E} T_{1,n} \leq c_{66} \cdot \frac{\log n}{n} \quad \text{and} \quad \mathbf{E} T_{3,n} \leq c_{67} \cdot \frac{\log n}{n}.$$

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

In the *third step of the proof* we show

$$\mathbf{E} T_{5,n} \leq c_{68} \cdot \frac{(\log n)^2}{n}.$$

Due to the definition of m_n and the assumption that $\|m\|_\infty \leq \beta_n$, it holds

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq 4 \cdot {c_1}^2 \cdot (\log n)^2.$$

Thus it suffices to show

$$\mathbf{P}(A_n^c) \leq \frac{c_{69}}{n}. \tag{52}$$

To do this, we first bound the probability that the weights in the first of the K_n fully connected neural networks differ by at most $\log n$ from $(\mathbf{w}_{1,i,j}^{(l)})_{i,j,l:l<L}$ in all components. For large n, this probability is bounded from below by

$$\left(\frac{\log n}{40d \cdot (\log n)^2}\right)^{r\cdot(r+1)\cdot(L-2)} \cdot \left(\frac{\log n}{16d \cdot (\log n)^2 \cdot n^\tau}\right)^{r\cdot(d+1)} \quad \geq \quad n^{-r\cdot(d+1)\cdot\tau-0.5}.$$

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

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

Assumption (49) implies $K_n \geq n^{r\cdot(d+1)\cdot\tau+1} \cdot \tilde{K}_n$ for large n. Thus we can apply this construction successively for all $\tilde{K}_n$ weights $((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{i,j,l:l<L}$. The probability that there exists $k \in \{1, \ldots, \tilde{K}_n\}$ such that none of the K_n weight vectors of the fully connected neural network differs from $(\mathbf{w}_{k,i,j}^{(l)})_{i,j,l:l<L}$ by at most $\log n$ is then for large n bounded from above by

$$\tilde{K}_n \cdot \exp(-n^{0.5}) \leq n^\kappa \cdot \exp(-n^{0.5}) \leq \frac{c_{70}}{n}.$$

Hence, for large n it is

$$\begin{aligned}\mathbf{P}(A_n^c) &\leq \frac{c_{70}}{n} + \mathbf{P}\{\max_{i=1,\ldots,n} |Y_i| > \sqrt{\beta_n}\} \leq \frac{c_{70}}{n} + n \cdot \mathbf{P}\{|Y| > \sqrt{\beta_n}\} \\ &\leq \frac{c_{70}}{n} + n \cdot \frac{\mathbf{E}\{\exp(c_3 \cdot Y^2)\}}{\exp(c_3 \cdot \beta_n)} \leq \frac{c_{71}}{n},\end{aligned}$$

where the last inequality holds because of (45) and (51).

In the *fourth step of the proof* we show that the assumptions (12) - (14) of Lemma 1 are satisfied which means that

$$\|(\nabla_{\mathbf{w}} F)(\mathbf{w})\| \leq L_n$$

for all $\mathbf{w} \in S := \left\{\mathbf{v} : \|\mathbf{v} - \mathbf{w}^{(0)}\| \leq 2 \cdot \sqrt{F(\mathbf{w}^{(0)})} + 1\right\}$,

$$\|(\nabla_{\mathbf{w}} F)(\mathbf{w}) - (\nabla_{\mathbf{w}} F)(\bar{\mathbf{w}})\| \leq L_n \cdot \|\mathbf{w} - \bar{\mathbf{w}}\|$$

for all $\mathbf{w}, \bar{\mathbf{w}} \in S$ and

$$\begin{aligned}
&|F(\mathbf{w}^*) - F((\mathbf{w}^*)^{(0)})| \\
&\leq D_n \cdot \|((\mathbf{w}^*)^{(L)}_{1,1,k})_{k=1,\dots,K_n}\| \cdot \|(\mathbf{w}^{(l)}_{i,j,k})_{i,j,k,l:l<L} - ((\mathbf{w}^{(0)})^{(l)}_{i,j,k})_{i,j,k,l:l<L}\|
\end{aligned}$$

for all

$$\begin{aligned}
&(\mathbf{w}^{(l)}_{i,j,k})_{i,j,k,l:l<L} \in \tilde{S} \\
&\quad := \left\{ (\bar{\mathbf{w}}^{(l)}_{i,j,k})_{i,j,k,l:l<L} \,:\, \|(\bar{\mathbf{w}}^{(l)}_{i,j,k})_{i,j,k,l:l<L} - ((\mathbf{w}^{(0)})^{(l)}_{i,j,k})_{i,j,k,l:l<L}\| \leq \sqrt{2 \cdot F(\mathbf{w}^{(0)})} \right\}
\end{aligned}$$

hold, if A_n holds.

If A_n holds, then we have

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

Let $\mathbf{w} \in S$. Then we have

$$\begin{aligned}
\|(\mathbf{w}^{(l)}_{i,j,k})_{i,j,k,l:1\leq l<L}\|_\infty &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\| + \|((\mathbf{w}^{(0)})^{(l)}_{i,j,k})_{i,j,k,l:1\leq l<L}\|_\infty \\
&\leq 2 \cdot \sqrt{F_n(\mathbf{w}^{(0)})} + 1 + c_{72} \cdot (\log n)^2 \\
&\leq c_{73} \cdot (\log n)^2
\end{aligned}$$

and

$$\begin{aligned}
\|(\mathbf{w}^{(L)}_{1,1,k})_{k=1,\dots,K_n}\|_\infty &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\| + \|((\mathbf{w}^{(0)})^{(L)}_{1,1,k})_{k=1,\dots,K_n}\|_\infty \\
&\leq 2 \cdot \sqrt{F_n(\mathbf{w}^{(0)})} + 1 \\
&\leq c_{74} \cdot (\log n)^{1/2}.
\end{aligned}$$

Hence (18)-(23) are satisfied for $B_n = c_{73} \cdot (\log n)^2$ and $\gamma_n^* = c_{74} \cdot (\log n)^{1/2}$. By Lemma 2 and Lemma 3 we get that (12) and (13) are satisfied provided that $L_n \geq K_n^{3/2} \cdot (\log n)^{6L+2}$.

Furthermore, let $\tilde{\mathbf{w}}$ such that $(\tilde{\mathbf{w}}^{(l)}_{i,j,k})_{i,j,k,l:l<L} \in \tilde{S}$. Then we obtain as in the proof of Theorem 1

$$\begin{aligned}
&|F_n(\mathbf{w}^*) - F_n((\mathbf{w}^*)^{(0)})| \\
&\leq \left(\frac{2}{n} \sum_{i=1}^{n} \left(f_{\mathbf{w}^*}(X_i) + f_{(\mathbf{w}^*)^{(0)}}(X_i) \right)^2 + \frac{8}{n} \sum_{i=1}^{n} Y_i^2 \right)^{1/2} \\
&\qquad \left(\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K_n} |(\mathbf{w}^*)_{1,1,k}|^2 \cdot \sum_{k=1}^{K_n} \left| f^{(L)}_{\mathbf{w}^*,k,1}(X_i) - f^{(L)}_{(\mathbf{w}^*)^{(0)},k,1}(X_i) \right|^2 \right) \right)^{1/2}.
\end{aligned}$$

With

$$\left(\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K_n}|(\mathbf{w}^*)_{1,1,k}|^2\cdot\sum_{k=1}^{K_n}\left|f^{(L)}_{\mathbf{w}^*,k,1}(X_i)-f^{(L)}_{(\mathbf{w}^*)^{(0)},k,1}(X_i)\right|^2\right)\right)^{1/2}$$
$$\leq \|((\mathbf{w}^*)^{(L)}_{1,1,k})_{k=1,\dots,K_n}\|\cdot c_{75}\cdot(\log n)^{2L}\cdot\|(\tilde{\mathbf{w}}^{(l)}_{i,j,k})_{i,j,k,l:l<L}-((\tilde{\mathbf{w}}^{(0)})^{(l)}_{i,j,k})_{i,j,k,l:l<L}\|$$

and since $(\tilde{\mathbf{w}}^{(l)}_{i,j,k})_{i,j,k,l:l<L}\in\tilde{S}$ we get

$$\left(\frac{2}{n}\sum_{i=1}^{n}\left(f_{\mathbf{w}^*}(X_i)+f_{(\mathbf{w}^*)^{(0)}}(X_i)\right)^2+\frac{8}{n}\sum_{i=1}^{n}Y_i^2\right)^{1/2}$$
$$\leq\left(4\cdot\sum_{k=1}^{K_n}\left|(\mathbf{w}^*)^{(L)}_{1,1,k}\right|^2\cdot\max_{i=1,\dots,n}\sum_{k=1}^{K_n}\left|f^{(L)}_{\mathbf{w}^*,k,1}(X_i)-f^{(L)}_{(\mathbf{w}^*)^{(0)},k,1}(X_i)\right|^2\right.$$
$$\left.+16\cdot\frac{1}{n}\sum_{i=1}^{n}f_{(\mathbf{w}^*)^{(0)}}(X_i)^2+\frac{8}{n}\sum_{i=1}^{n}Y_i^2\right)^{1/2}$$
$$\leq\left(4\cdot\sum_{k=1}^{K_n}\left|(\mathbf{w}^*)^{(L)}_{1,1,k}\right|^2\cdot c_{76}\cdot(\log n)^{4L}\cdot\|((\mathbf{w}^*)^{(l)}_{i,j,k})_{i,j,k,l:l<L}-(((\mathbf{w}^*)^{(0)})^{(l)}_{i,j,k})_{i,j,k,l:l<L}\|^2\right.$$
$$\left.+16\cdot\sum_{i=1}^{n}f_{(\mathbf{w}^*)^{(0)}}(X_i)^2+\frac{8}{n}\sum_{i=1}^{n}Y_i^2\right)^{1/2}$$
$$\leq\left(c_{77}\cdot M_n^2\cdot(\log n)^{4L+1}+16\cdot\beta_n^2+8\cdot\beta_n\right)^{1/2}$$
$$\leq c_{78}\cdot M_n\cdot(\log n)^{2L+1}.$$

Summarizing these steps yields

$$|F_n(\mathbf{w}^*)-F_n((\mathbf{w}^*)^{(0)})|$$
$$\leq c_{79}\cdot(\log n)^{4L+1}\cdot M_n\cdot\|((\mathbf{w}^*)^{(L)}_{1,1,k})_{k=1,\dots,K_n}\|\cdot\|(\tilde{\mathbf{w}}^{(l)}_{i,j,k})_{i,j,k,l:l<L}-((\tilde{\mathbf{w}}^{(0)})^{(l)}_{i,j,k})_{i,j,k,l:l<L}\|$$

Thus (14) is satisfied with

$$D_n=c_{79}\cdot M_n\cdot(\log n)^{4L+1}.$$

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

$$\mathbf{E}T_{2,n}\leq c_{80}\cdot\frac{n^{\tau\cdot d+\epsilon}}{n}.$$

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

$$|w^{(L)}_{1,1,k}|\leq c_{81}\cdot(\log n)^2\quad(k=1,\dots,K_n),$$

$$|w_{i,j,k}^{(l)}| \leq (20d+1)\cdot(\log n)^2 \quad (l=1,\ldots,L-1)$$

and

$$|w_{i,j,k}^{(0)}| \leq (8d+1)\cdot(\log n)^2 \cdot n^{\tau}.$$

From Lemma 1 we know for n large that

$$\|((\mathbf{w}^{(t_n)})_{1,1,k}^{(L)})_{k=1,\ldots,K_n} - ((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\| \leq \sqrt{2F_n(\mathbf{w}^{(0)})} \leq (\log n)^2$$

and

$$\|((\mathbf{w}^{(t_n)})_{i,j,k}^{(l)})_{i,j,k,l:l<L} - ((\mathbf{w}^{(0)})_{i,j,k}^{(l)})_{i,j,k,l:l<L}\| \leq \sqrt{2F_n(\mathbf{w}^{(0)})} \leq (\log n)^2.$$

This and the initial choice of $\mathbf{w}^{(0)}$ imply that we have on A_n for n large

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

Thus, for any $u>0$ we get

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

where

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

Application of Lemma 4 yields for $x_1^n \in supp(X)$ and $\alpha = c_{82}$

$$\begin{aligned}
&\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n}\cdot f : f \in \mathcal{F}_n\right\}, x_1^n\right) \leq \mathcal{N}_1\left(\delta\cdot\beta_n, \mathcal{F}_n, x_1^n\right) \\
&\leq \left(\frac{c_{83}\cdot\beta_n}{\delta\cdot\beta_n}\right)^{c_{84}\cdot(\log n)^{2d}\cdot n^{\tau\cdot d}\cdot(\log n)^{2\cdot(L-1)\cdot d}\cdot\left(\frac{K_n\cdot(\log n)^2}{\delta\cdot\beta_n}\right)^{d/k}+c_{85}}.
\end{aligned}$$

For $\delta > 1/n$ and k large enough, we obtain

$$\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n}\cdot f : f \in \mathcal{F}_n\right\}, x_1^n\right) \leq c_{86}\cdot n^{c_{87}\cdot n^{\tau\cdot d+\epsilon/2}}.$$

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

$$\mathbf{P}\{T_{2,n} > u\} \leq 14\cdot c_{86}\cdot n^{c_{87}\cdot n^{\tau\cdot d+\epsilon/2}}\cdot\exp\left(-\frac{n}{5136\cdot\beta_n^2}\cdot u\right).$$

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

$$\begin{aligned}\mathbf{E}\{T_{2,n}\} &\leq \epsilon_n + \int_{\epsilon_n}^{\infty} \mathbf{P}\{T_{2,n} > u\}\, du \\ &\leq \epsilon_n + 14 \cdot c_{86} \cdot n^{c_{87} \cdot n^{\tau \cdot d + \epsilon/2}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot \epsilon_n\right) \cdot \frac{5136 \cdot \beta_n^2}{n}.\end{aligned}$$

Setting

$$\epsilon_n = \frac{5136 \cdot \beta_n^2}{n} \cdot c_{87} \cdot n^{\tau \cdot d + \epsilon/2} \cdot \log n$$

yields the assertion of the fifth step of the proof.

In the *sixth step of the proof* we show

$$\begin{aligned}&\mathbf{E}\{T_{4,n}\} \\ &\leq c_{88} \cdot \Bigg(\sup_{\substack{(\bar{w}_{i,j,k}^{(l)})_{i,j,k,l}: \\ |\bar{w}_{i,j,k}^{(l)} - w_{i,j,k}^{(l)}| \leq \log n \ (l=0,\ldots,L-1)}} \int |\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x) - m(x)|^2 \mathbf{P}_X(dx) \\ &\qquad + c_{89} \cdot \frac{\log n}{n} + c_{90} \cdot \frac{n^{\tau \cdot d + \epsilon}}{n} + c_{91} \cdot M_n^2 \cdot (\log n)^{4L+3/2}.\end{aligned}$$

Since

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

we obtain

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

The application of Lemma 1 implies

$$\begin{aligned}&\mathbf{E}\{T_{4,n}/2\} \\ &\leq \mathbf{E}\Bigg\{\Bigg[F_n\left((\mathbf{w}^*)^{(0)}\right) + D_n \cdot \|(((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\| \cdot \sqrt{2 \cdot F_n\left(\mathbf{w}^{(0)}\right)} \\ &\quad + \frac{\|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\ldots,K_n} - (((\mathbf{w}^*)^{(0)})_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\|^2}{2} + \frac{F_n\left(\mathbf{w}^{(0)}\right)}{t_n} \\ &\quad - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\Bigg] \cdot 1_{A_n}\Bigg\}\end{aligned}$$

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

$$+\mathbf{E}\Bigg\{ \Bigg(F_n\left((\mathbf{w}^*)^{(0)}\right) + D_n \cdot \|(((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\| \cdot \sqrt{2 \cdot F_n\left(\mathbf{w}^{(0)}\right)}$$

$$+\frac{\|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\ldots,K_n} - (((\mathbf{w}^*)^{(0)})_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\|^2}{2} + \frac{F_n\left(\mathbf{w}^{(0)}\right)}{t_n}$$

$$-\frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2$$

$$-2 \cdot \left(\mathbf{E}\{|\sum_{k=1}^{K_n} (\mathbf{w}^*)_{1,1,k}^{(L)} \cdot f_{(\mathbf{w}^*),j,1}^{(L)}(X) - Y|^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\} - \mathbf{E}\{|m(X) - Y|^2\} \right) \Bigg) 1_{A_n} \Bigg\}.$$

Due to step 4 we have

$$D_n = c_{79} \cdot M_n \cdot (\log n)^{4L+1}.$$

Using the same arguments as in step 2 and 5 of the proof we obtain

$$\mathbf{E}\Bigg\{ \Bigg(F_n((\mathbf{w}^*)^{(0)}) - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2$$

$$+c_{92} \cdot M_n \cdot (\log n)^{4L+1} \cdot \|(((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\| \cdot \sqrt{2 \cdot F_n\left(\mathbf{w}^{(0)}\right)}$$

$$+\frac{\|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\ldots,K_n} - (((\mathbf{w}^*)^{(0)})_{1,1,k}^{(L)})_{k=1,\ldots,K_n}\|^2}{2} + \frac{F_n\left(\mathbf{w}^{(0)}\right)}{t_n}$$

$$-2 \cdot \left(\mathbf{E}\{|\sum_{k=1}^{K_n} (\mathbf{w}^*)_{1,1,k}^{(L)} \cdot f_{(\mathbf{w}^*),j,1}^{(L)}(X) - Y|^2 | \mathcal{D}_n, \mathbf{w}^{(0)}\} - \mathbf{E}\{|m(X) - Y|^2\} \right) \Bigg) 1_{A_n} \Bigg\}$$

$$\leq c_{93} \cdot \frac{\log n}{n} + c_{93} \cdot M_n^2 \cdot (\log n)^{4L+3/2} + \frac{M_n^2}{2} + \frac{c_1 \cdot \log n}{t_n} + c_{94} \cdot \frac{n^{\tau \cdot d + \epsilon}}{n}.$$

This implies

$$\mathbf{E}\{T_{4,n}/2\}$$

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

$$+c_{91} \cdot \frac{\log n}{n} + c_{92} \cdot \frac{n^{\tau \cdot d + \epsilon}}{n} + c_{95} \cdot M_n^2 \cdot (\log n)^{4L+3/2}.$$

Summarizing the above results we get the assertion.

□

To prove Theorem 2, we use the following lemma, which provides another bound on the approximation error. Furthermore, it ensures that the outer weights are sufficiently small.

Lemma 6. *Let $1/2 \leq p \leq 1$, $C > 0$, let $f : \mathbb{R}^d \to \mathbb{R}$ be a (p, C)–smooth function and let X be a $\mathbb{R}^d$-valued random variable with $supp(X) \subseteq [0,1]^d$. Let $l \in \mathbb{N}$, $0 < \delta < 1/2$ with*

$$c_{96} \cdot \delta \leq \frac{1}{2^l} \leq c_{97} \cdot \delta \tag{53}$$

and let $L, r, s \in \mathbb{N}$ with $L \geq 2$ and $r \geq 2d$. Furthermore, let

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

Then there exist

$$w_{k,i,j}^{(l)} \in [-20d \cdot (\log n)^2, 20d \cdot (\log n)^2] \quad \text{for } l = 1, \ldots, L, k = 1, \ldots \tilde{K}_n$$

and

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

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

$$\begin{aligned} &\int |\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x) - f(x)|^2 \mathbf{P}_X(dx) \\ &\qquad \leq c_{98} \cdot \left(l^2 \cdot \delta + \delta^{2p} + \frac{l \cdot (2^l + 1)^{2d}}{n^s} \right), \end{aligned} \tag{54}$$

$$|\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(x)| \leq c_{99} \cdot \left(1 + \frac{(2^l + 1)^{2d}}{n^s} \right) \qquad (x \in [0,1]^d) \tag{55}$$

and

$$\sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 \leq \frac{c_{100}}{2^{2 \cdot d \cdot l}}. \tag{56}$$

Proof. The proof follows from the proof of Lemma 7 in Kohler and Krzyżak (2022a). □

4.4. Proof of Theorem 2

Let $l = \lfloor \frac{1}{1+d} \log_2 n \rfloor$. Then condition (53) holds for $\delta = n^{-1/(1+d)}$. Set

$$\tilde{K}_n = (l \cdot (2^l + 1)^{2d} + 1)^3 \approx c_{101} \cdot (\log n)^3 \cdot n^{\frac{6d}{1+d}}, \qquad N_n = n^{c_{102}}$$

and define the weight vector $\mathbf{w}$ as in Lemma 6. Then we obtain by Lemma 6 that assumption (46) is satisfied for $M_n = \frac{c_{103}}{n^{d/(d+1)}}$. Assumption (47) follows directly from (55) of Lemma 6 for s sufficiently large.

By Theorem 4, $\tau = \frac{1}{1+d}$ and Lemma 6 where s is sufficiently large we get for large n

$$\begin{aligned}
&\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \le c_{104} \cdot \left(\frac{n^{\frac{1}{1+d}\cdot d+\epsilon}}{n} + c_{105} \cdot \frac{(\log n)^{4L+3/2}}{n^{\frac{2d}{d+1}}} \right. \\
&\qquad \left. + \sup_{\substack{(\bar{w}_{i,j,k}^{(l)})_{i,j,k,l}: \\ |\bar{w}_{i,j,k}^{(l)} - w_{i,j,k}^{(l)}| \le \log n \ (l=0,\dots,L-1)}} \int |\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\tilde{\mathbf{w}},k,1}^{(L)}(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \\
&\le c_{106} \cdot \left(\frac{n^{\frac{1}{1+d}\cdot d+\epsilon}}{n} + \frac{(\log n)^{4L+3/2}}{n^{\frac{2d}{1+d}}} + \frac{(\log n)^2}{n^{\frac{1}{1+d}}} + \frac{1}{n^{\frac{2p}{1+d}}} + \frac{\log n \cdot n^{\frac{2d}{1+d}}}{n^s} \right) \\
&\le c_{107} \cdot n^{-\frac{1}{1+d}+\epsilon}
\end{aligned}$$

for s sufficiently large. □

4.5. Auxiliary results for the proof of Theorem 3

In order to prove Theorem 3, we use the following lemma, which controls the complexity of a set of over-parametrized deep neural networks for interaction models.

Lemma 7. *Let $\alpha \ge 1$, $\beta > 0$ and let $A, B, C \ge 1$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be k-times differentiable such that all derivatives up to order k are bounded on $\mathbb{R}$. Let $\mathcal{F}$ be the set of all functions*

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

where $f_{\mathbf{w}_I}$ is defined by (1)–(3) with d replaced by d^* and weight vector $\mathbf{w}_I$,

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

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

$$\sum_{j=1}^{K_n} |(\mathbf{w}_I)_{1,1,j}^{(L)}| \le C, \tag{57}$$

$$|(\mathbf{w}_I)_{k,i,j}^{(l)}| \le B \quad (k \in \{1, \dots, K_n\}, i, j \in \{1, \dots, r\}, l \in \{1, \dots, L-1\}) \tag{58}$$

and

$$|(\mathbf{w}_I)_{k,i,j}^{(0)}| \le A \quad (k \in \{1, \dots, K_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d\}). \tag{59}$$

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

$$\mathcal{N}_p\left(\epsilon, \{T_\beta f \,:\, f \in \mathcal{F}\}, x_1^n\right)$$

$$\leq \left(c_{108} \cdot \frac{\beta^p}{\epsilon^p} \right)^{c_{109} \cdot \alpha^{d^*} \cdot A^{d^*} \cdot B^{(L-1) \cdot d^*} \left(\frac{C}{\epsilon} \right)^{d^*/k} + c_{110}}.$$

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

4.6. Proof of Theorem 3

Let $l = \lfloor \frac{1}{1+d^*} \log_2 n \rfloor$ and $N_n = n^{c_{111}}$. Furthermore, let

$$\tilde{K}_n = \binom{d}{d^*} (l \cdot (2^l + 1)^{2d^*} + 1)^3 \approx c_{112} \cdot (\log n)^3 \cdot n^{\frac{6d^*}{1+d^*}}.$$

Then (53) holds for $\delta = n^{-1/(1+d^*)}$. Define the weight vector $\mathbf{w}$ such that the components are chosen according to Lemma 6 so that they approximate m_I.

Assumption (46) and (47) of Theorem 4 are satisfied for $\mathbf{w}$ since

$$\sum_{I \subseteq \{1,\dots,d\} \, : \, |I| = d^*} \sum_{k=1}^{\tilde{K}_n} |(\mathbf{w}_I)_{1,1,k}^{(L)}|^2 \leq c_{113} \cdot n^{-\frac{2d^*}{d^*+1}}$$

and

$$\sum_{I \subseteq \{1,\dots,d\} \, : \, |I| = d^*} \left| \sum_{k=1}^{\tilde{K}_n} (\mathbf{w}_I)_{1,1,k}^{(L)} \cdot f_{(\tilde{\mathbf{w}}_I)_{k,1}}^{(L)}(x) \right| \leq c_{114} \cdot \left(1 + \frac{n^{2d^*/(1+d^*)}}{n^s} \right)$$

hold.

Then we obtain by application of Lemma 7 the assertion of Theorem 4 for interaction models. Therefore the proof of Theorem 3 follows similarly to the proof of Theorem 2. By applying the assertions of Theorem 4 and Lemma 6 we get for $\tau = \frac{1}{1+d^*}$ and s sufficiently large

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_{115} \cdot \left(\frac{n^{\frac{1}{1+d^*} \cdot d^* + \epsilon}}{n} + c_{113} \cdot n^{-\frac{2d^*}{1+d^*}} \cdot (\log n)^{4L+3/2} \right. \\
&\qquad \left. + \sup_{\substack{(\bar{w}_{i,j,k}^{(l)})_{i,j,k,l} : \\ |\bar{w}_{i,j,k}^{(l)} - w_{i,j,k}^{(l)}| \leq \log n \, (l=0,\dots,L-1)}} \int |\sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\tilde{\mathbf{w}},k,1}^{(L)}(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \\
&\leq c_{116} \cdot \left(\frac{n^{\frac{1}{1+d^*} \cdot d^* + \epsilon}}{n} + \frac{(\log n)^{4L+3/2}}{n^{\frac{2d^*}{1+d^*}}} + \frac{(\log n)^2}{n^{\frac{1}{1+d^*}}} + \frac{1}{n^{\frac{2p}{1+d^*}}} + \frac{\log n \cdot n^{\frac{2d^*}{1+d^*}}}{n^s} \right) \\
&\leq c_{117} \cdot n^{-\frac{1}{1+d^*} + \epsilon}.
\end{aligned}$$

□

References

[1] Allen-Zhu, Li and Song. "A Convergence Theory for Deep Learning via Over-Parameterization". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 242–252. URL: https://proceedings.mlr.press/v97/allen-zhu19a.html.

[2] Arora, Cohen, Golowich, and Hu. *A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks*. Preprint. 2019. DOI: 10.48550/ARXIV.1810.02281. arXiv: 1810.02281. URL: https://arxiv.org/abs/1810.02281.

[3] Barron. "Approximation and Estimation Bounds for Artificial Neural Networks". In: *Machine Learning* 14 (Jan. 1994), pp. 115–133. DOI: 10.1007/BF00993164.

[4] Bartlett, Long, Lugosi, and Tsigler. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070. DOI: 10.1073/pnas.1907378117.

[5] Bartlett, Montanari and Rakhlin. "Deep learning: a statistical viewpoint". In: *Acta numerica* 30 (2021), pp. 87–201. DOI: 10.1017/S0962492921000027.

[6] Bauer and Kohler. "On as a remedy for the curse of dimensionality in nonparametric regression". In: *The Annals of Statistics* 47.4 (2019), pp. 2261 –2285. DOI: 10.1214/18-AOS1747.

[7] Belkin, Hsu, Ma, and Mandal. "Reconciling modern machine learning practice and the classical bias variance trade off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854. DOI: 10.1073/pnas.1903070116.

[8] Belkin, Rakhlin and Tsybakov. "Does data interpolation contradict statistical optimality?" In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1611–1619. URL: https://proceedings.mlr.press/v89/belkin19a.html.

[9] Billings, Hedelius, Millecam, Wingate, and Corte. "ProSPr: Democratized Implementation of Alphafold Protein Distance Prediction Network". In: (2019). DOI: 10.1101/830273.

[10] Braun, Kohler, Langer, and Walk. "Convergence rates for shallow neural networks learned by gradient descent". In: *Bernoulli* 30.1 (2024), pp. 475 –502. DOI: 10.3150/23-BEJ1605.

[11] Bubeck, Eldan, Lee, and Mikulincer. *Network size and weights size for memorization with two-layers neural networks*. Preprint. 2020. arXiv: 2006.02855.

[12] Daniely. *Neural Networks Learning and Memorization with (almost) no Over-Parameterization*. Preprint. 2019. arXiv: 1911.09873.

[13] Daniely. *Memorizing Gaussians with no over-parameterizaion via gradient decent on neural networks*. Preprint. 2020. arXiv: 2003.12895.

[14] Devroye, Györfi and Lugosi. *A probabilistic theory of pattern recognition*. Vol. 31. Applications of mathematics. New York, 1996. URL: http://scans.hebis.de/HEBCGI/show.pl?04874027_toc.pdf.

[15] Drews and Kohler. *On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent*. Preprint. 2022. arXiv: 2208.14283.

[16] Du, Lee, Tian, Singh, and Poczos. "Gradient Descent Learns One-hidden-layer CNN: Don't be Afraid of Spurious Local Minima". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1339–1348. URL: https://proceedings.mlr.press/v80/du18b.html.

[17] Frei, Chatterji and Bartlett. "Benign Overfitting without Linearity: Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 2668–2703. URL: https://proceedings.mlr.press/v178/frei22a.html.

[18] Györfi, Kohler, Krzyżak, and Walk. "A distribution-free theory of nonparametric regression". In: Springer series in statistics (2002). URL: http://scans.hebis.de/HEBCGI/show.pl?30930960_toc.pdf.

[19] Hunt. "Could artificial intelligence win the next Weather Photographer of the Year competition?" In: *Weather* 78.4 (2023), pp. 108–112. DOI: https://doi.org/10.1002/wea.4348.

[20] Kawaguchi and Huang. *Gradient Descent Finds Global Minima for Generalizable Deep Neural Networks of Practical Sizes*. Preprint. 2020. arXiv: 1908.02419.

[21] Kohler and Krzyżak. "Nonparametric Regression Based on Hierarchical Interaction Models". In: *IEEE Transactions on Information Theory* 63.3 (2017), pp. 1620–1630. DOI: 10.1109/TIT.2016.2634401.

[22] Kohler and Krzyżak. "Over-parametrized deep neural networks minimizing the empirical risk do not generalize well". In: *Bernoulli* 27.4 (2021), pp. 2564 –2597. DOI: 10.3150/21-BEJ1323.

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

[24] Kohler and Krzyżak. *Over-parametrized neural networks learned by gradient descent can generalize especially well*. Preprint. 2022. arXiv: 2003.12895.

[25] Kohler and Langer. "On the rate of convergence of fully connected very deep neural network regression estimates". In: *The Annals of Statistics* 49.4 (2021), pp. 2231–2249. DOI: 10.1214/20-AOS2034.

[26] McGrath, Kapishnikov, Tomašev, Pearce, Wattenberg, Hassabis, Kim, Paquet, and Kramnik. "Acquisition of chess knowledge in AlphaZero". In: *Proceedings of the National Academy of Sciences* 119.47 (2022), e2206625119. DOI: 10.1073/pnas.2206625119.

[27] Montanari and Zhong. *The Interpolation Phase Transition in Neural Networks: Memorization and Generalization under Lazy Training*. Preprint. 2020. arXiv: 2007.12826.

[28] Nitanda and Suzuki. *Optimal Rates for Averaged Stochastic Gradient Descent under Neural Tangent Kernel Regime*. Preprint. 2021. arXiv: 2006.12297.

[29] Schmidt-Hieber. "Nonparametric regression using deep neural networks with ReLU activation function". In: *The Annals of Statistics* 48.4 (2020). DOI: 10.1214/19-aos1875.

[30] Stone. "Optimal Global Rates of Convergence for Nonparametric Regression". In: *The Annals of Statistics* 10.4 (1982), pp. 1040 –1053. DOI: 10.1214/aos/1176345969.

[31] Stone. "The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation". In: *The Annals of Statistics* 22.1 (1994), pp. 118 –171. DOI: 10.1214/aos/1176325361.

[32] Suzuki. *Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality*. Preprint. 2018. arXiv: 1810.08033.

[33] Suzuki and Nitanda. *Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space*. Preprint. 2021. arXiv: 1910.12799.

[34] Yehudai and Shamir. *On the Power and Limitations of Random Features for Understanding Neural Networks*. Preprint. 2022. arXiv: 1904.00687.

[35] Zong and Krishnamachari. *A survey on GPT-3*. Preprint. 2022. arXiv: 2212.00857.

A. Proof of Lemma 2

Proof. We have

$$\begin{aligned}
\|\nabla_{\mathbf{w}} F_n(\mathbf{w})\|^2 &= \sum_{k,i,j,l} \left(\frac{2}{n} \sum_{s=1}^{n} (Y_s - f_{\mathbf{w}}(X_s)) \cdot \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \\
&\leq 4 \cdot \sum_{k,i,j,l} \frac{1}{n} \sum_{s=1}^{n} (Y_s - f_{\mathbf{w}}(X_s))^2 \cdot \left(\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \\
&\leq c_{118} \cdot K_n \cdot L \cdot r^2 \cdot d \cdot \max_{k,i,j,l,s} \left(\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \cdot \frac{1}{n} \sum_{s=1}^{n} (Y_s - f_{\mathbf{w}}(X_s))^2.
\end{aligned}$$

The partial derivative of f is given by

$$\begin{aligned}\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(x) = \sum_{s_{l+2}=1}^{r} \cdots \sum_{s_{L-1}=1}^{r} & f_{k,j}^{(l)}(x) \cdot \sigma'\left(\sum_{t=1}^{r} w_{k,i,t}^{(l)} \cdot f_{k,t}^{(l)}(x) + w_{k,i,0}^{(l)}\right) \\ & \cdot w_{k,s_{l+2},i}^{(l+1)} \cdot \sigma'\left(\sum_{t=1}^{r} w_{k,s_{l+2},t}^{(l+1)} \cdot f_{k,t}^{(l+1)}(x) + w_{k,s_{l+2},0}^{(l+1)}\right) \cdot w_{k,s_{l+3},s_{l+2}}^{(l+2)} \\ & \cdot \sigma'\left(\sum_{t=1}^{r} w_{k,s_{l+3},t}^{(l+2)} \cdot f_{k,t}^{(l+2)}(x) + w_{k,s_{l+3},0}^{(l+2)}\right) \cdots w_{k,s_{L-1},s_{L-2}}^{(L-2)} \\ & \cdot \sigma'\left(\sum_{t=1}^{r} w_{k,s_{L-1},t}^{(L-2)} \cdot f_{k,t}^{(L-2)}(x) + w_{k,s_{L-1},0}^{(L-2)}\right) \cdot w_{k,1,s_{L-1}}^{(L-1)} \\ & \cdot \sigma'\left(\sum_{t=1}^{r} w_{k,1,t}^{(L-1)} \cdot f_{k,t}^{(L-1)}(x) + w_{k,1,0}^{(L-1)}\right) \cdot w_{1,1,k}^{(L)}, \end{aligned} \tag{60}$$

where we have used the abbreviations

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

and

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

Together with (18) and (19) we obtain

$$\max_{k,i,j,l,s} \left(\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(X_s)\right)^2 \leq c_{119} \cdot r^{2L} \cdot \max\{\|\sigma'\|_\infty^{2L}, 1\} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2.$$

In the next step of the proof we want to show, that

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

Let $\bar{f}_{k,i}^{(l)}$ be defined by

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

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

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

First, we show by induction that

$$|f_{i,j}^{(l)}(x) - \bar{f}_{i,j}^{(l)}(x)| \leq \max\{\|\sigma'\|_\infty^l, 1\} \cdot (2r+1)^l \cdot B_n^l \cdot \alpha_n$$

$$\cdot \max_{i,j,s:s<L} |w_{k,i,j}^{(s)} - v_{k,i,j}^{(s)}| \cdot \max\{\|\sigma\|_\infty, 1\} \tag{61}$$

holds for $l = 1, \ldots, L$ and $x \in [-\alpha_n, \alpha_n]^d$.

The function σ is differentiable and its derivative is bounded, hence σ is Lipschitz continuous with Lipschitz constant $\|\sigma'\|_\infty$. This implies

$$\begin{aligned}
|f_{i,j}^{(1)}(x) - \bar{f}_{i,j}^{(1)}(x)| &\le \|\sigma'\|_\infty \cdot \left(\sum_{j=1}^{d} |w_{k,i,j}^{(0)} - v_{k,i,j}^{(0)}| \cdot |x^{(j)}| + |w_{k,i,0}^{(0)} - v_{k,i,0}^{(0)}| \right) \\
&\le \|\sigma'\|_\infty \cdot (2r+1) \cdot \alpha_n \cdot \max_{i,j,s:s<L} |w_{k,i,j}^{(s)} - v_{k,i,j}^{(s)}|.
\end{aligned}$$

Assume (61) holds for some $l-1$ with $l = 2, \ldots, L-1$. Then we have

$$\begin{aligned}
&\left| f_{i,j}^{(l)}(x) - \bar{f}_{i,j}^{(l)}(x) \right| \\
&\le \|\sigma'\|_\infty \cdot \left(\sum_{j=1}^{r} \left| w_{k,i,j}^{(l-1)} \right| \cdot \left| f_{k,j}^{(l-1)}(x) - \bar{f}_{k,j}^{(l-1)}(x) \right| \right. \\
&\qquad \left. + \sum_{j=1}^{r} \left| w_{k,i,j}^{(l-1)} - v_{k,i,j}^{(l-1)} \right| \cdot \left| \bar{f}_{k,j}^{(l-1)}(x) \right| + \left| w_{k,i,0}^{(l)} - v_{k,i,0}^{(l-1)} \right| \right) \\
&\le \|\sigma'\|_\infty \cdot \left(r \cdot B_n \cdot \max_{j=1,\ldots,r} \left| f_{k,j}^{(l-1)}(x) - \bar{f}_{k,j}^{(l-1)}(x) \right| \right. \\
&\qquad \left. + (r+1) \cdot \max_{i,j,s:s<L} \left| w_{k,i,j}^{(s)} - v_{k,i,j}^{(s)} \right| \cdot \max\{\|\sigma\|_\infty, 1\} \right) \\
&\le \max\{\|\sigma'\|_\infty^l, 1\} \cdot (2r+1)^l \cdot B_n^l \cdot \alpha_n \cdot \max_{i,j,s:s<L} \left| w_{k,i,j}^{(s)} - v_{k,i,j}^{(s)} \right| \cdot \max\{\|\sigma\|_\infty, 1\}.
\end{aligned}$$

This implies

$$\begin{aligned}
&|f_{\mathbf{w}}(x) - f_{\mathbf{v}}(x)| \\
&= \left| \sum_{j=1}^{K_n} w_{1,1,j}^{(L)} \cdot f_{j,1}^{(L)}(x) - \sum_{j=1}^{K_n} v_{1,1,j}^{(L)} \cdot \bar{f}_{j,1}^{(L)}(x) \right| \\
&\le \left| \sum_{j=1}^{K_n} w_{1,1,j}^{(L)} \left(f_{j,1}^{(L)}(x) - \bar{f}_{j,1}^{(L)}(x) \right) \right| + \left| \sum_{j=1}^{K_n} \left(w_{1,1,j}^{(L)} - v_{1,1,j}^{(L)} \right) \cdot \bar{f}_{j,1}^{(L)}(x) \right| \\
&\le K_n \cdot \max_j \left| w_{1,1,j}^{(L)} \right| \cdot \max_j \left| f_{j,1}^{(L)}(x) - \bar{f}_{j,1}^{(L)}(x) \right| \\
&\qquad + K_n \cdot \max_j \left| w_{1,1,j}^{(L)} - v_{1,1,j}^{(L)} \right| \cdot \max\{\|\sigma\|_\infty, 1\} \\
&\le 2 \cdot K_n \cdot \gamma_n^* \cdot \max\{\|\sigma'\|_\infty^L, 1\} \cdot (2r+1)^L \cdot B_n^L \cdot \alpha_n \cdot \|\mathbf{w} - \mathbf{v}\|_\infty \cdot \max\{\|\sigma\|_\infty, 1\}.
\end{aligned}$$

Together with assumption (20) we can conclude

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

By summarizing the above results we obtain the assertion. □

B. Proof of Lemma 3

Proof. We have

$$\begin{aligned}
&\|\nabla_{\mathbf{w}}F_n(\mathbf{w}_1) - \nabla_{\mathbf{w}}F_n(\mathbf{w}_2)\|^2 \\
&= \sum_{k,i,j,l}\left(\frac{2}{n}\sum_{s=1}^{n}(Y_s - f_{\mathbf{w}_1}(X_s))\cdot\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s)\right. \\
&\quad \left.-\left(\frac{2}{n}\sum_{s=1}^{n}(Y_s - f_{\mathbf{w}_2}(X_s))\cdot\frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(X_s)\right)\right)^2 \\
&\leq 8\cdot\sum_{k,i,j,l}\left(\frac{1}{n}\sum_{s=1}^{n}(f_{\mathbf{w}_2}(X_s) - f_{\mathbf{w}_1}(X_s))\cdot\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s)\right)^2 \\
&\quad +8\cdot\sum_{k,i,j,l}\left(\frac{1}{n}\sum_{s=1}^{n}(Y_s - f_{\mathbf{w}_2}(X_s))\cdot\left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(X_s)\right)\right)^2 \\
&\leq 8\cdot\sum_{k,i,j,l}\max_{s=1,\ldots,n}\left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s)\right)^2\cdot\frac{1}{n}\sum_{s=1}^{n}(f_{\mathbf{w}_2}(X_s) - f_{\mathbf{w}_1}(X_s))^2 \\
&\quad +8\cdot\frac{1}{n}\sum_{s=1}^{n}(Y_s - f_{\mathbf{w}_2}(X_s))^2\cdot\sum_{k,i,j,l}\max_{s=1,\ldots,n}\left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(X_s)\right)^2.
\end{aligned}$$

From the proof of Lemma 2 we can conclude

$$\begin{aligned}
&\sum_{k,i,j,l}\max_{s=1,\ldots,n}\left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s)\right)^2 \\
&\leq c_{120}\cdot K_n\cdot L\cdot r^2\cdot d\cdot r^{2L}\cdot\max\{\|\sigma'\|_\infty^{2L},1\}\cdot B_n^{2L}\cdot(\gamma_n^*)^2\cdot\alpha_n^2,
\end{aligned}$$

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

and

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

So it remains to bound

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

By (60) we know that

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

for fixed $x \in [-\alpha_n, \alpha_n]^d$ is a sum of at most r^{L-2} products where each product contains at most $2L+1$ factors. Each of these products contains at most L factors, that are bounded in absolute value by B_n except the last one, which is bounded in absolute value by γ_n^*. Considered as a function in $\mathbf{w}$, these products are Lipschitz continuous with a Lipschitz constant bounded by 1.

According to the proof of Lemma 2 we know that $f_{k,j}^{(l)}(x)$, which is either bounded by $\|\sigma\|_\infty$ or α_n, is Lipschitz continuous with a Lipschitz constant bounded by $\max\{\|\sigma'\|_\infty^l, 1\} \cdot \max\{\|\sigma\|_\infty, 1\} \cdot (2r+1)^l \cdot B_n^l \cdot \alpha_n$. The remaining at most L factors are bounded by $\max\{\|\sigma'\|_\infty, 1\}$ with a Lipschitz constant bounded by $c_{121} \cdot (2r+1)^L \cdot B_n^L \cdot \alpha_n \cdot \max\{\|\sigma\|_\infty, 1\}$.

The assertion follows from the following result: If $g_1, \ldots, g_s : \mathbb{R} \to \mathbb{R}$ are Lipschitz continuous functions with Lipschitz constants $C_{Lip,g_1}, \ldots, C_{Lip,g_s}$, then

$$\prod_{l=1}^{s} g_l \text{ and } \sum_{l=1}^{s} g_l$$

are Lipschitz continuous functions with Lipschitz constant bounded by

$$\sum_{l=1}^{s} C_{Lip,g_l} \cdot \prod_{k \in \{1,\ldots,s\}\setminus\{l\}} \|g_k\|_\infty \leq s \cdot \max_l C_{Lip,g_l} \cdot \prod_{k \in \{1,\ldots,s\}\setminus\{l\}} \|g_k\|_\infty$$

and by

$$\sum_{l=1}^{s} C_{lip,g_l} \leq s \cdot \max_L C_{Lip,g_l}$$

respectively.

This together with the fact that σ and σ' are bounded yields

$$\sum_{k,i,j,l} \max_{s=1,\ldots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \leq c_{122} \cdot K_n \cdot B_n^{4L} \cdot \alpha_n^4 \cdot (\gamma_n^*)^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2.$$

Summarizing the above results we get the assertion of Lemma 3.

□