Over-parametrized neural networks learned by gradient descent can generalize especially well *

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

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

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

February 20, 2022

Abstract

Estimation of univariate regression function by a neural network with one hidden layer is considered, where the weight vector is determined by applying gradient descent to a regularized empirical L_2 risk. Here the number of hidden neurons is chosen much larger than the sample size. It is shown that the estimate nevertheless generalizes well in case that the Fourier transform of the regression function decays suitably fast, and that in this case over-parametrization leads to a particular good rate of convergence.

AMS classification: Primary 62G08; secondary 62G20.

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

1. Introduction

1.1. Scope of this article

In the last decade deep learning was successfully applied in many areas. Deep convolutional networks have been applied in image classification by Krizhevsky, Sutskever and Hinton (2012), in language processing by Ni et al. (2021), in machine translation by Wu et al. (2016), in medical diagnosis by Mondal et al. (2021), and in many other areas. Despite impressive successes in applications there are very few theoretical studies explaining reasons for strong performance of deep networks in practice. Recently, several theoretical studies investigating deep neural networks appeared, see, e.g., Kohler and Krzyzak (2017), Bauer and Kohler (2019), Schmidt-Hieber (2020), Kohler and Langer (2021), Suzuki and Nitanda (2019) and Kohler and Krzyżak (2021).

Backpropagation is the most common method for training neural networks in practice. Braun et al. (2021) analyze the L_2 error of neural network regression estimates with

*Running title: *Over-parametrized neural networks*

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

one hidden layer. They showed that in the class of regression functions having Fourier transform decreasing suitably fast, a neural network estimate whose weights are initialized randomly according to a proper uniform distributions and then are learned by the gradient descent, achieves a rate of convergence of $1/\sqrt{n}$ (up to a logarithmic factor). Kohler and Krzyżak (2021) demonstrated that over-parametrized deep neural networks with the sigmoidal squasher interpolating the data do not generalize well on a new data, i. e., the networks which minimize the empirical risk do not achieve the optimal minimax rate of convergence for estimation of smooth regression functions and for design points having discrete distribution. In the present paper we show that over-parametrization of one hidden layer neural network with properly regularized L_2 risk trained by the gradient descent achieves the rate of convergence $n^{-2/3}$ in one dimensional case in the class of regression functions having Fourier transform decreasing suitably fast. So in this case over-parametrization leads to a better rate of convergence than in Braun et al. (2021).

1.2. Regression estimation

In this paper we study neural network regression estimates in connection with nonparametric regression. To do this we consider an $\mathbb{R}^d \times \mathbb{R}$–valued random vector (X, Y), where X is the so–called observation vector and Y is the so-called response. Assume the condition $\mathbf{E}\{Y^2\} < \infty$. We are interested in the functional correlation between the response Y and the observation vector X. In applications the distribution of (X, Y) is unknown, therefore we want to recover the functional correlation between X and Y using only a sample of (X, Y), i.e., a data set

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

where (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n) are i.i.d. We are searching for an estimate

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

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

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

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

1.3. Neural networks

Neural networks try to mimic the human brain in order to define classes of functions. The starting point is a very simple model of a nerve cell, in which some kind of thresholding is applied to a linear combination of the outputs of other nerve cells. This leads to functions of the form

$$f(x) = \sigma\left(\sum_{j=1}^{d} w_j \cdot x^{(j)} + w_0\right) \quad (x = (x^{(1)}, \ldots, x^{(d)})^T \in \mathbb{R}^d),$$

where we call $w_0, \ldots, w_d \in \mathbb{R}$ weights of the neuron and where we call $\sigma : \mathbb{R} \to \mathbb{R}$ an activation function. Traditionally, so–called squashing functions are chosen as activation functions, which are nondecreasing and satisfy $\lim_{x \to -\infty} \sigma(x) = 0$ and $\lim_{x \to \infty} \sigma(x) = 1$. An example of a squashing function is the sigmoidal or logistic squasher

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \quad (x \in \mathbb{R}). \tag{2}$$

Recently, also unbounded activation functions have been used, e.g., the ReLU activation function

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

The simplest form of neural networks are shallow networks, i.e., neural networks with one hidden layer, in which a simple linear combination of the above neurons is used to define a function $f : \mathbb{R}^d \to \mathbb{R}$ by

$$f(x) = \sum_{k=1}^{K} \alpha_k \cdot \sigma \left(\sum_{j=1}^{d} \beta_{k,j} \cdot x^{(j)} + \beta_{k,0} \right) + \alpha_0. \tag{3}$$

Here $K \in \mathbb{N}$ is the number of hidden neurons, and the weights $\alpha_k \in \mathbb{R}$ $(k = 0, \ldots, K)$, $\beta_{k,j} \in \mathbb{R}$ $(k = 1, \ldots, K, j = 0, \ldots, d)$ can be adapted to the data (1) in order to define an estimate of the regression function. This can be achieved by, for example, applying the principle of least squares, i.e., by defining the regression estimate m_n by

$$m_n(\cdot) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|^2, \tag{4}$$

where $\mathcal{F}$ is the set of all functions of the form (3) with a fixed number of neurons K and fixed activation function σ.

The rate of convergence of the shallow neural networks regression estimates has been analyzed in Barron (1994) and McCaffrey and Gallant (1994). Barron (1994) proved a dimensionless rate of $n^{-1/2}$ (up to some logarithmic factor), provided the Fourier transform of the regression function has a finite first moment, which basically requires that the function becomes smoother with increasing dimension d of X. McCaffrey and Gallant (1994) showed a rate of $n^{-\frac{2p}{2p+d+5}+\varepsilon}$ in case of a (p, C)-smooth regression function, but their study was restricted to the use of a certain cosine squasher as activation function.

In deep learning neural networks with several hidden layers are used to define classes of functions. Here, the neurons are arranged in $L \in \mathbb{N}$ layers, where the $k_s \in \mathbb{N}$ neurons in layer $s \in \{2, \ldots, L\}$ get the output of the k_{s-1} neurons in layer $s-1$ as input, and where the neurons in the first layer are applied to the d components of the input. We denote the weight between neuron j in layer $s-1$ and neuron i in layer s by $w_{i,j}^{(s)}$. This leads to the following recursive definition of a neural network with L layers and k_s neurons in layer $s \in \{1, \ldots, L\}$:

$$f(x) = \sum_{i=1}^{k_L} w_{1,i}^{(L)} f_i^{(L)}(x) + w_{1,0}^{(L)} \tag{5}$$

for some $w_{1,0}^{(L)}, \ldots, w_{1,k_L}^{(L)} \in \mathbb{R}$ and for $f_i^{(L)}$'s recursively defined by

$$f_i^{(s)}(x) = \sigma\left(\sum_{j=1}^{k_{s-1}} w_{i,j}^{(s-1)} f_j^{(s-1)}(x) + w_{i,0}^{(s-1)}\right) \tag{6}$$

for some $w_{i,0}^{(s-1)}, \ldots, w_{i,k_{s-1}}^{(s-1)} \in \mathbb{R}$, $s \in \{2, \ldots, L\}$, and

$$f_i^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} w_{i,j}^{(0)} x^{(j)} + w_{i,0}^{(0)}\right) \tag{7}$$

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

The rate of convergence of least squares estimates based on multilayer neural networks has been analyzed in Kohler and Krzyżak (2017), Imaizumi and Fukamizu (2018), Bauer and Kohler (2019), Kohler, Krzyżak and Langer (2019), Suzuki and Nitanda (2019), Schmidt-Hieber (2020) and Kohler and Langer (2021). One of the main results achieved in this context shows that neural networks can achieve some kind of dimension reduction, provided the regression function is a composition of (sums of) functions, where each of the function is a function of at most $d^* < d$ variables (see Kohler and Langer (2020) for a motivation of such a function class). In Kohler and Krzyżak (2017) it was shown that in this case suitably defined least squares estimates based on multilayer neural networks achieve the rate of convergence $n^{-2p/(2p+d^*)}$ (up to some logarithmic factor) for $p \leq 1$. This result also holds for $p > 1$ provided the squashing function is suitably smooth as was shown in Bauer and Kohler (2019). Schmidt-Hieber (2020) showed the surprising result that this is also true for neural networks which use the non-smooth ReLU activation function. In Kohler and Langer (2021) it was shown that such results also hold for very simply constructed fully connected feedforward neural networks. Kohler, Krzyżak and Langer (2019) considered regression functions with a low local dimensionality and demonstrated that neural networks are also able to circumvent the curse of dimensionality in this context. Results regarding the estimation of regression functions which are piecewise polynomials having partitions with rather general smooth boundaries by neural networks have been derived in Imaizumi and Fukamizu (2018). That neural networks can also achieve a dimension reduction in Besov spaces was shown in Suzuki and Nitanda (2019).

1.4. Gradient descent

In Subsection 1.3 the neural network regression estimates are defined as nonlinear least squares estimates, i.e., as functions which minimize the empirical L_2 risk over nonlinear classes of neural networks. In practice, it is usually not possible to find the global minimum of the empirical L_2 risk over a nonlinear class of neural networks and we try to find a local minimum using, for instance, the gradient descent algorithm (so-called backpropagation).

Denote by $f_{net,\mathbf{w}}$ the neural network defined by (5)–(7) with weight vector

$$\mathbf{w} = (w_{j,k}^{(s)})_{s=0,\dots,L,j=1,\dots,k_{s+1},k=0,\dots,k_s}$$

(where we have set $k_0 = d$ and $k_{L+1} = 1$), and set

$$F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{net,\mathbf{w}}(X_i)|^2. \tag{8}$$

In backpropagation gradient descent is used to minimize (8) with respect to $\mathbf{w}$. Here, set

$$\mathbf{w}(0) = \mathbf{v} \tag{9}$$

for some (usually randomly chosen) initial weight vector $\mathbf{v}$ and define

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \lambda_n \cdot \nabla_{\mathbf{w}} F(\mathbf{w}(t)) \tag{10}$$

for $t = 0, 1, \dots, t_n - 1$, where $\lambda_n > 0$ is the stepsize and $t_n \in \mathbb{N}$ is the number of performed gradient descent steps. The estimate is then defined by

$$m_n(\cdot) = f_{net,\mathbf{w}(t_n)}(\cdot). \tag{11}$$

1.5. Main results

We study the rate of convergence of the univariate neural network regression estimate with one hidden layer where the weights of the network are determined by the gradient descent of a regularized empirical L_2 risk. The number of neurons is chosen much larger than the sample size. We assume that the Fourier transform

$$\hat{F} : \mathbb{R} \to \mathbb{C}, \quad \hat{F}(\omega) = \frac{1}{(2\pi)^{1/2}} \cdot \int_{\mathbb{R}} e^{-i \cdot \omega \cdot x} \cdot m(x)\, dx$$

of the regression function satisfies

$$|\hat{F}(\omega)| \leq \frac{c_1}{|\omega|^2 \cdot (\log |\omega|)^2} \quad \text{for all } \omega \in \mathbb{R} \text{ with } |\omega| \geq 2$$

for some $c_1 > 0$. We show that the neural network estimate with the logistic squasher activation function generalizes well regardless of the number of hidden neurons (as long as this number is bounded by some polynomial in the sample size), if the initial weights of the neural network are chosen from some uniform distribution, if a suitable penalty is added to the empirical L_2 risk during backpropagation and if a suitable number of gradient descent steps is performed. In particular, we show that any value of the number of hidden neurons larger than $n^{2/3}$ leads under the above assumptions to the rate of convergence $n^{-2/3}$ (up to some logarithmic factor), which improves the rate of convergence achieved in Braun et al. (2021) in case $K_n \approx \sqrt{n}$.

1.6. Discussion of related results

Study of deep learning has been very active field of research in recent years, see Berner et al. (2021) for a recent survey of progress in mathematics of deep learning. A large number of results were recently obtained for neural network regression estimates learned by the gradient descent. Braun et al. (2021) showed rate of convergence $1/\sqrt{n}$ (up to a logarithmic factor) for regression functions that have Fourier transforms with polynomially decreasing tails (an assumption slightly stronger than the finite first moment of the Fourier transform assumption of Barron (1993)).

Many recent papers tried to demonstrate theoretically that backpropagation learning works for deep neural networks. The most popular approach which emerged in this context is so–called landscape approach. Choromanska et al. (2015) used random matrix theory to derive a heuristic argument showing that the risk of most of the local minima of the empirical L_2 risk $F_n(\mathbf{w})$ is not much larger than the risk of the global minimum. This claim was validated for neural networks with special activation function by, e.g., Arora et al. (2018), Kawaguchi (2016), and Du and Lee (2018), which have analyzed gradient descent for neural networks with a linear or quadratic activation function. No good approximation results exist for such neural networks, and consequently one cannot deduce from these results good rates of convergence for neural network regression estimates. Du et al. (2018) analyzed gradient descent learning for neural networks with one hidden layer and Gaussian inputs. As they used the expected gradient instead of the gradient in their gradient descent routine, one cannot apply their results to derive the rate of convergence for neural network regression estimates learned by the gradient descent. Liang et al. (2018) applied gradient descent to a modified loss function in classification, where it is assumed that the data can be interpolated by a neural network. Neural tangent kernel networks (NTK) were introduced by Jacot, Gabriel and Honger (2020). They showed that in the infinite-width limit case NTK converges to a deterministic limit kernel which stays constant during Gaussian descent training of the random weights initialized with the Gaussian distributions. These results were extended by Huang, Du and Xu (2020) to orthogonal initialization which was shown to speed up training of fully connected deep networks. Nitanda and Suzuki (2017) obtained global convergence rate for the averaged stochastic gradient descent for over-parametrized shallow neural networks.

Recently it was shown in several papers, see, e.g., Allen-Zhu, Li and Song (2019), Kawaguchi and Huang (2019) and the literature cited therein, that the gradient descent leads to a small empirical L_2 risk in over-parametrized neural networks. Here the results in Allen-Zhu, Li and Song (2019) are proven for the ReLU activation function and neural networks with a polynomial size in the sample size. The neural networks in Kawaguchi and Huang (2019) use squashing activation functions and are much smaller (in fact, they require only a linear size in the sample size). In contrast to Allen-Zhu, Li and Song (2019) there the learning rate is set to zero for all neurons except for neurons in the output layer and consequently in different layers of the network different learning rates are used. Actually, they compute a linear least squares estimate with the gradient descent, which is not used in practice. It was shown in Kohler and Krzyżak (2021) that any estimate which interpolates the training data does not generalize well in a sense

that it can, in general, not achieve the optimal minimax rate of convergence in case of a general design measure.

In recent survey paper Bartlett et al. (2021) conjectured that over-parametrization allows gradient descent to find interpolating solutions which implicitly impose regularization, and that over-parametrization leads to benign overfiting. For related results involving the truncated Hilbert kernel regression estimate refer to Belkin et al. (2019) and to Wyner et al. (2017) for the results involving AdaBoost and random forests. Linear regression in overfitting regime has been also considered in Bartlett et al. (2020). Benign over-parametrization in shallow ReLU networks has been analyzed by Wang and Lin (2021). They showed L_2 error rate of $\sqrt{\log n/n}$ for over-parametrized neural network when the number of hidden neurons exceeds the number of samples. In the present paper we show that over-parametrization in learning the regularized L_2 risk by the gradient descent leads to excellent generalization.

1.7. Notation

Throughout the paper, the following notation is used: The sets of natural numbers, natural numbers including 0, real numbers, nonegative real numbers and complex numbers are denoted by $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{C}$, respectively. For $z \in \mathbb{R}$, we denote the smallest integer greater than or equal to z by $\lceil z \rceil$ and the largest integer smaller or equal to z by $\lfloor z \rfloor$. Let $D \subseteq \mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be a real-valued function defined on $\mathbb{R}^d$. We write $x = \arg\min_{z \in D} f(z)$ if $\min_{z \in \mathcal{D}} f(z)$ exists and if x satisfies $x \in D$ and $f(x) = \min_{z \in \mathcal{D}} f(z)$. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$. For $f : \mathbb{R}^d \to \mathbb{R}$

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

is its supremum norm. Furthermore we set

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

for $A \subseteq \mathbb{R}^d$. S_r denotes the ball with radius r in $\mathbb{R}^d$ and center 0 (with respect to the Euclidean norm). Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$, let $x_1, \dots, x_n \in \mathbb{R}^d$, set $x_1^n = (x_1, \dots, x_n)$ and let $p \geq 1$. A finite collection $f_1, \dots, f_N : \mathbb{R}^d \to \mathbb{R}$ is called an L_p ε–cover of $\mathcal{F}$ on x_1^n if for any $f \in \mathcal{F}$ there exists $i \in \{1, \dots, N\}$ such that

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

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

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

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

1.8. Outline

In Section 2 we define our estimate. In Section 3 we present our main result concerning the rate of convergence of a neural network estimate with one hidden layer learned by gradient descent. The proof of the main result is given in Section 4. In the appendix we present the proof of an auxiliary result from empirical process theory applied in the proof of our main result.

2. An over-parametrized neural network regression estimator

We consider neural networks with one hidden layer defined by

$$f_{net,\mathbf{w}}(x) = w_{1,0}^{(1)} + \sum_{j=1}^{K_n} w_{1,j}^{(1)} \cdot \sigma(\sum_{k=1}^{d} w_{j,k}^{(0)} \cdot x^{(k)} + w_{j,0}^{(0)}) = \alpha_0 + \sum_{j=1}^{K_n} \alpha_j \cdot \sigma(\beta_j^T \cdot x + \gamma_j) \quad (12)$$

where $K_n \in \mathbb{N}$ is the number of hidden neurons,

$$\alpha_i \in \mathbb{R} \quad (i = 0, \dots, K_n), \quad \beta_i = (\beta_{i,1}, \dots, \beta_{i,d})^T \in \mathbb{R}^d \quad (i = 1, \dots, K_n),$$

$$\gamma_i \in \mathbb{R} \quad (i = 1, \dots, K_n)$$

and

$$\mathbf{w} = (w_{j,k}^{(l)})_{j,k,l} = (\alpha, \beta, \gamma) = (\alpha_0, \alpha_1, \dots, \alpha_{K_n}, \beta_1, \dots, \beta_{K_n}, \gamma_1, \dots, \gamma_{K_n})$$

is the vector of the $1 + K_n \cdot (d + 2)$ many weights of the neural network $f_{net,\mathbf{w}}$. In the sequel we use the logistic squasher

$$\sigma(x) = 1/(1 + e^{-x}) \quad (13)$$

as the activation function.

We will learn the weight vector w by applying gradient descent to the regularized empirical L_2 risk

$$F_n(w) = \frac{1}{n} \sum_{i=1}^{n} |f_{net,w}(X_i) - Y_i|^2 + c_2 \cdot \left(\frac{\alpha_0^2}{n^{2/3}} + \frac{K_n}{n^{2/3}} \cdot \sum_{k=1}^{K_n} \alpha_k^2 \right). \quad (14)$$

To do this, we initialize $w(0)$ randomly (independent from $\mathcal{D}_n$) as follows: We set

$$B_n = 12 \cdot n \cdot K_n$$

and choose β_k uniformly distributed on $\{x \in \mathbb{R}^d : \|x\| = B_n\}$ and γ_k uniformly distributed on $[-B_n, B_n]$ such that $\beta_1, \dots, \beta_{K_n}, \gamma_1, \dots, \gamma_{K_n}$ are independent, we choose $\alpha_k = 0$ $(k = 0, \dots, K_n)$ and then we set

$$\mathbf{w}(0) = (\alpha_0, \dots, \alpha_{K_n}, \beta_1, \dots, \beta_{K_n}, \gamma_0, \dots, \gamma_{K_n}).$$

Next we compute

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \lambda_n \cdot \nabla_w F_n(\mathbf{w}(t)) \quad (15)$$

for $t = 0, 1, \ldots, t_n - 1$. Here $t_n \in \mathbb{N}$ is the number of gradient descent steps which we perform.

Our estimate is then defined by

$$\tilde{m}_n(x) = f_{net,\mathbf{w}(t_n)}(x) \tag{16}$$

and

$$m_n(x) = T_{\beta_n} \tilde{m}_n(x), \tag{17}$$

where $T_\beta z = \max\{\min\{z, \beta\}, -\beta\}$ is a truncation operator and $\beta_n = c_3 \cdot \log n$.

3. Main results

Theorem 1 *Assume $d = 1$. Let (X, Y) be an $[0, 1] \times \mathbb{R}$–valued random vector such that*

$$\mathbf{E}\left\{e^{c_4 \cdot Y^2}\right\} < \infty \tag{18}$$

holds for some constant $c_4 > 0$ and assume that the corresponding regression function $m(x) = \mathbf{E}\{Y|X = x\}$ is bounded, satisfies

$$\int_{\mathbb{R}} |m(x)| \, dx < \infty,$$

and that its Fourier transform $\hat{F}$ satisfies

$$|\hat{F}(\omega)| \leq \frac{c_1}{|\omega|^2 \cdot (\log |\omega|)^2} \quad \text{for all } \omega \in \mathbb{R} \text{ with } |\omega| \geq 2 \tag{19}$$

for some $c_1 > 0$. Choose $c_5 > 0$, let $c_6 > 0$ be sufficiently large, and set

$$K_n = n^{c_5}, \quad L_n = c_6 \cdot (\log n)^5 \cdot n^{2/3} \cdot K_n, \quad \lambda_n = \frac{1}{L_n}, \quad B_n = 12 \cdot n \cdot K_n$$

and

$$t_n = \lceil (\log n)^2 \cdot n^{2/3} \cdot L_n \rceil,$$

let σ be the logistic squasher and define the estimate as in Section 2. Then one has for n sufficiently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_7 \cdot (\log n)^8 \cdot n^{-\min\left\{c_5, \frac{2}{3}\right\}}.$$

Remark 1. In Theorem 1 the number of parameters of the network can be arbitrarily large, as long as it is bounded by a polynomial in the sample size n. In particular, it may be much larger than the sample size n, and nevertheless the estimate generalizes well.

Remark 2. In Braun et al. (2021) neural network estimates with one hidden layer learned by gradient descent have been analyzed and the generalization error was bounded

using the classical Vapnik-Chervonenkis theory. There the optimal rate of convergence of (up to some logarithmic factor) $n^{-1/2}$ was shown and this rate has occurred for K_n of order $\sqrt{n}$. In contrast, Theorem 1 above shows that in case of a proper regularization of the empirical L_2 risk any value of $K_n = n^{c_5}$ with $c_5 > 1/2$ leads (up to some logarithmic factor) to the better rate of convergence of $n^{-\min\{2/3, c_5\}}$. In particular, as soon as we choose K_n larger than $n^{2/3}$ we get (up to some logarithmic factor) the rate of convergence $n^{-2/3}$. So in our theoretical setting the over-parametrization improves up to some point the rate of convergence, and from this point on any further over-parametrization achieves this better rate (as soon as it is not too large, i.e., as long as the number K_n of hidden neurons is a polynomial in the sample size n).

4. Proofs

4.1. Auxiliary results concerning the estimation error

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

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

and

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

Assume

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\| \leq \sqrt{2 \cdot t \cdot L \cdot \max\{F(\mathbf{a}_0), 1\}} \tag{20}$$

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

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}) - (\nabla_{\mathbf{a}} F)(\mathbf{b})\| \leq L \cdot \|\mathbf{a} - \mathbf{b}\| \tag{21}$$

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

$$\|\mathbf{a} - \mathbf{a}_0\| \leq \sqrt{8 \cdot \frac{t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}} \quad \text{and} \quad \|\mathbf{b} - \mathbf{a}_0\| \leq \sqrt{8 \cdot \frac{t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}}. \tag{22}$$

Then we have

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

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

and

$$F(\mathbf{a}_k) \leq F(\mathbf{a}_{k-1}) \quad \text{for all } k \in \{1, \ldots, t\}.$$

Proof. See Lemma 2 in Braun et al. (2021). □

Lemma 2 *Assume* $supp(X) \subseteq [0,1]^d$, $\gamma_n^* \geq 1$, $2 \cdot t_n \geq L_n$, $c_2^2 \leq n^{4/3}/16$ *and*

$$|w_{1,k}^{(1)}| \leq \gamma_n^* \quad (k = 1, \ldots, K_n) \quad \textit{and} \quad \|\mathbf{w} - \mathbf{v}\|^2 \leq \frac{2t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}. \tag{23}$$

Then we have with probability one

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w})\| \leq 26 \cdot d \cdot (\gamma_n^*)^2 \cdot K_n^{3/2} \cdot \sqrt{\frac{t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}}.$$

Proof. Using $(a+b)^2 \leq 2 \cdot a^2 + 2 \cdot b^2$ $(a, b \in \mathbb{R})$ we get

$$\begin{aligned}
&\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w})\|^2 \\
&= \sum_{j,k,l} \left(\frac{2}{n} \sum_{i=1}^n (Y_i - f_{net,\mathbf{w}}(X_i)) \cdot \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}}(X_i) \right. \\
&\quad \left. + \frac{\partial}{\partial w_{j,k}^{(l)}} \left(\frac{c_2}{n^{2/3}} \cdot (w_{1,0}^{(1)})^2 + \frac{c_2 \cdot K_n}{n^{2/3}} \cdot \sum_{l=1}^{K_n} (w_{1,l}^{(1)})^2 \right) \right)^2 \\
&\leq \sum_{j,k,l} \frac{8}{n} \sum_{i=1}^n (Y_i - f_{net,\mathbf{w}}(X_i))^2 \cdot \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}}(X_i) \right)^2 + \frac{8 \cdot c_2^2 \cdot K_n^2}{n^{4/3}} \cdot \sum_{k=0}^{K_n} (w_{1,k}^{(1)})^2.
\end{aligned}$$

From this we get the assertion as in the proof of Lemma 5 in Braun et al. (2021). □

Lemma 3 *Assume* $supp(X) \subseteq [0,1]^d$, $\gamma_n^* \geq 1$, $t_n \geq L_n$ *and*

$$\max\left\{|(\mathbf{w}_2)_{1,k}^{(1)}|, |v_{1,k}^{(1)}|\right\} \leq \gamma_n^* \quad (k = 1, \ldots, K_n) \quad \textit{and} \quad \|\mathbf{w}_2 - \mathbf{v}\|^2 \leq 8 \cdot \frac{t_n}{L_n} \cdot \max\{F(\mathbf{v}), 1\}. \tag{24}$$

Then we have with probability one

$$\begin{aligned}
&\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}_2)\| \\
&\leq 131 \cdot d^{3/2} \cdot \max\{\sqrt{F_n(\mathbf{v})}, 1\} \cdot \max\{1, c_2\} \cdot (\gamma_n^*)^2 \cdot K_n \cdot \sqrt{\frac{t_n}{L_n}} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|.
\end{aligned}$$

Proof. We have

$$\begin{aligned}
&\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F_n(\mathbf{w}_2)\|^2 \\
&= \sum_{j,k,l} \left(\frac{2}{n} \sum_{i=1}^n (Y_i - f_{net,\mathbf{w}_1}(X_i)) \cdot \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_1}(X_i) \right. \\
&\qquad \left. + \frac{\partial}{\partial w_{j,k}^{(l)}} \left(\frac{c_2}{n^{2/3}} \cdot ((\mathbf{w}_1)_{1,0}^{(1)})^2 + \frac{c_2 \cdot K_n}{n^{2/3}} \cdot \sum_{l=1}^{K_n} ((\mathbf{w}_1)_{1,l}^{(1)})^2 \right) \right. \\
&\quad \left. - \frac{2}{n} \sum_{i=1}^n (Y_i - f_{net,\mathbf{w}_2}(X_i)) \cdot \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_2}(X_i) \right.
\end{aligned}$$

$$
\begin{aligned}
&\qquad -\frac{\partial}{\partial w_{j,k}^{(l)}} \left(\frac{c_2}{n^{2/3}} \cdot ((\mathbf{w}_2)_{1,0}^{(1)})^2 + \frac{c_2 \cdot K_n}{n^{2/3}} \cdot \sum_{l=1}^{K_n} ((\mathbf{w}_2)_{1,l}^{(1)})^2 \right) \Bigg)^2 \\
&\leq 4 \cdot \sum_{j,k,l} \Bigg(\frac{2}{n} \sum_{i=1}^{n} (Y_i - f_{net,\mathbf{w}_1}(X_i)) \cdot \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_1}(X_i) \\
&\qquad -\frac{2}{n} \sum_{i=1}^{n} (Y_i - f_{net,\mathbf{w}_2}(X_i)) \cdot \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_1}(X_i) \Bigg)^2 \\
&\quad +4 \cdot \sum_{j,k,l} \Bigg(\frac{2}{n} \sum_{i=1}^{n} (Y_i - f_{net,\mathbf{w}_2}(X_i)) \cdot \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_1}(X_i) \\
&\qquad -\frac{2}{n} \sum_{i=1}^{n} (Y_i - f_{net,\mathbf{w}_2}(X_i)) \cdot \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_2}(X_i) \Bigg)^2 \\
&\qquad +8 \cdot \frac{c_2^2 \cdot K_n^2}{n^{4/3}} \cdot \sum_{k=0}^{K_n} |(\mathbf{w}_1)_{1,k}^{(1)} - (\mathbf{w}_2)_{1,k}^{(1)}|^2.
\end{aligned}
$$

From this we get the assertion as in the proof of Lemma 6 in Braun et al. (2021). □

Lemma 4 *Let $\beta_n = c_3 \cdot \log(n)$ for some suitably large constant $c_3 > 0$. Assume that the distribution of (X, Y) satisfies (18) for some constant $c_4 > 0$ and that the regression function m is bounded in absolute value. Let A_n be an arbitrary event. Let $\mathcal{F}_n$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$ and assume that on the event A_n the estimate m_n satisfies*

$$m_n = T_{\beta_n} \tilde{m}_n$$

for some $\tilde{m}_n$ which satisfies

$$\tilde{m}_n(\cdot) = \tilde{m}_n(\cdot, (X_1, Y_1), \ldots, (X_n, Y_n)) \in \mathcal{F}_n.$$

Let $\delta_n > c_8 \cdot (\log n)^2/n$ be such that we have for any $\delta > \delta_n/(8 \cdot \beta_n^2)$

$$c_9 \cdot \frac{\sqrt{n} \cdot \delta}{\beta_n} \geq \int_{c_{10} \cdot \delta}^{\sqrt{\delta}} (\log \mathcal{N}_2(u \cdot \beta_n, T_{\beta_n} \mathcal{F}_n, x_1^n))^{1/2} \, du.$$

Then m_n satisfies

$$
\begin{aligned}
&\mathbf{E}\Bigg(\Bigg(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\quad -2 \cdot \left(\frac{1}{n} \sum_{i=1}^{n} |\tilde{m}_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| \leq \beta_n \, (j \in \{1, \ldots, n\})\}} - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \Bigg) \cdot 1_{A_n} \Bigg) \\
&\leq c_{11} \cdot \left(\delta_n + \frac{(\log n)^2}{n} \right)
\end{aligned}
$$

for $n > 1$ and some constant $c_{46} > 0$, which does not depend on n or β_n.

Proof. This lemma follows in a straightforward way from the proof of Theorem 1 in Bagirov, Clausen and Kohler (2009). A complete proof is given in the appendix. □

Lemma 5 *Let $p \geq 1$, $L, V > 0$, $K \in \mathbb{N}$, let $\sigma : \mathbb{R} \to [0,1]$ be a continuous squashing function, and let $\mathcal{F}$ be the set of all functions*

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = c_0 + \sum_{k=1}^{K} c_k \cdot \sigma(a_k \cdot x + b_k)$$

where

$$a_1, \ldots, a_K, b_1, \ldots b_K, c_0, c_1, \ldots, c_K \in \mathbb{R} \quad \textit{with} \quad \sum_{k=1}^{K} |c_k| \leq V$$

are arbitrary. Then we have for any $0 < \delta < 2 \cdot L$ and $x_1^n \in [0,1]^n$

$$\mathcal{N}_p\left(3 \cdot \delta, T_L \mathcal{F}, x_1^n\right) \leq \left(\left(n + \frac{V}{\delta} + 1\right) \cdot \frac{2 \cdot L}{\delta}\right)^{\frac{V}{\delta}+2}.$$

Proof. Let

$$f(x) = c_0 + \sum_{k=1}^{K} c_k \cdot \sigma(a_k \cdot x + b_k).$$

Then we can rewrite f as

$$f(x) = c_0 + \sum_{k=1,\ldots,K: a_k = 0} c_k \cdot \sigma(b_k) + \sum_{k=1,\ldots,K: a_k \neq 0} c_k \cdot \sigma(a_k \cdot x + b_k),$$

hence we can assume without loss of generality that in the definition of $\mathcal{F}$ all a_k are nonzero.

Choose $u_j \in \mathbb{R}$ $(j = 1, \ldots, J = \lceil V/\delta \rceil - 1)$ such that

$$\sigma(u_j) = j \cdot \frac{\delta}{V} \quad (j = 1, \ldots, J).$$

Then $u_1 < u_2 < \cdots < u_J$ holds and for $k \in \{1, \ldots, J-1\}$ and $x \in (u_k, u_{k+1}]$ we have

$$k \cdot \frac{\delta}{V} \leq \sigma(x) \leq (k+1) \cdot \frac{\delta}{V}, \quad k \cdot \frac{\delta}{V} \leq \sum_{j=1}^{J} \frac{\delta}{V} \cdot 1_{[u_j, \infty)}(x) \leq (k+1) \cdot \frac{\delta}{V}$$

and

$$k \cdot \frac{\delta}{V} \leq \sum_{j=1}^{J} \frac{\delta}{V} \cdot 1_{(u_j, \infty)}(x) \leq (k+1) \cdot \frac{\delta}{V}.$$

From this we can conclude

$$\sup_{x \in \mathbb{R}} \left| \sigma(x) - \sum_{j=1}^{J} \frac{\delta}{V} \cdot 1_{[u_j, \infty)}(x) \right| \leq \frac{\delta}{V}$$

and

$$\sup_{x\in\mathbb{R}}\left|\sigma(x)-\sum_{j=1}^{J}\frac{\delta}{V}\cdot 1_{(u_j,\infty)}(x)\right|\leq\frac{\delta}{V}.$$

This implies

$$\begin{aligned}
&\left|c_0+\sum_{k=1}^{K}c_k\cdot\sigma(a_k\cdot x+b_k)-c_0-\sum_{k=1}^{K}c_k\cdot\left(1_{\{a_k>0\}}\cdot\sum_{j=1}^{J}\frac{\delta}{V}\cdot 1_{[u_j,\infty)}(a_k\cdot x+b_k)\right.\right.\\
&\qquad\left.\left.+1_{\{a_k<0\}}\cdot\sum_{j=1}^{J}\frac{\delta}{V}\cdot 1_{(u_j,\infty)}(a_k\cdot x+b_k)\right)\right|\\
&\leq\sum_{k=1}^{K}|c_k|\cdot 1_{\{a_k>0\}}\cdot\left|\sigma(a_k\cdot x+b_k)-\sum_{j=1}^{J}\frac{\delta}{V}\cdot 1_{[u_j,\infty)}(a_k\cdot x+b_k)\right|\\
&\quad+\sum_{k=1}^{K}|c_k|\cdot 1_{\{a_k<0\}}\cdot\left|\sigma(a_k\cdot x+b_k)-\sum_{j=1}^{J}\frac{\delta}{V}\cdot 1_{(u_j,\infty)}(a_k\cdot x+b_k)\right|\\
&\leq\sum_{k=1}^{K}|c_k|\cdot\frac{\delta}{V}\leq\delta.
\end{aligned}$$

In case $a_k>0$ we have

$$1_{[u_j,\infty)}(a_k\cdot x+b_k)=1_{[u_j-b_k,\infty)}(a_k\cdot x)=1_{[(u_j-b_k)/a_k,\infty)}(x),$$

and in case $a_k<0$ we have

$$1_{(u_j,\infty)}(a_k\cdot x+b_k)=1_{(u_j-b_k,\infty)}(a_k\cdot x)=1_{(-\infty,(u_j-b_k)/a_k)}(x)=1-1_{[(u_j-b_k)/a_k,\infty)}(x).$$

Using this together with

$$\sum_{j=1}^{J}\left|c_k\cdot\frac{\delta}{V}\right|=|c_k|\cdot J\cdot\frac{\delta}{V}\leq|c_k|$$

we can rewrite

$$c_0+\sum_{k=1}^{K}c_k\cdot\left(1_{\{a_k>0\}}\cdot\sum_{j=1}^{J}\frac{\delta}{V}\cdot 1_{[u_j,\infty)}(a_k\cdot x+b_k)+1_{\{a_k<0\}}\cdot\sum_{j=1}^{J}\frac{\delta}{V}\cdot 1_{(u_j,\infty)}(a_k\cdot x+b_k)\right)$$

on $[0,1]$ as

$$g(x)=\bar{c}_0+\sum_{k=1}^{\bar{K}}\bar{c}_k\cdot 1_{[\bar{a}_k,\infty)}(x)$$

where $\bar{K}=K\cdot J$, $0\leq\bar{a}_1<\bar{a}_2<\cdots<\bar{a}_{\bar{K}}\leq 1$ and where $\bar{c}_0,\ldots,\bar{c}_{\bar{K}}\in\mathbb{R}$ satisfy

$$\sum_{k=1}^{\bar{K}}|\bar{c}_k|\leq V.$$

Let $\mathcal{G}$ be the set of all functions g of the above form. Then we have shown that for any $f \in \mathcal{F}$ there exists $g \in \mathcal{G}$ such that

$$\sup_{x\in[0,1]} |f(x) - g(x)| \leq \delta.$$

Hence it suffices to show that we have for any $0 < \delta < 2 \cdot L$ and $x_1^n \in [0,1]^n$

$$\mathcal{N}_p\left(2 \cdot \delta, T_L\mathcal{G}, x_1^n\right) \leq \left(\left(n + \frac{V}{\delta} + 1\right) \cdot \frac{2 \cdot L}{\delta}\right)^{\frac{V}{\delta}+2},$$

which we will show next.

Let $g \in \mathcal{F}$ be arbitrary, i.e., assume that $g : [0,1] \to \mathbb{R}$ is given by

$$g(x) = \bar{c}_0 + \sum_{k=1}^{\bar{K}} \bar{c}_k \cdot 1_{[\bar{a}_k,\infty)}(x)$$

where

$$0 \leq \bar{a}_1 < \bar{a}_2 < \cdots < \bar{a}_{\bar{K}} \leq 1$$

and

$$\bar{c}_0, \ldots, \bar{c}_{\bar{K}} \in \mathbb{R} \quad \text{with} \quad \sum_{k=1}^{\bar{K}} |\bar{c}_k| \leq V.$$

Choose $N \in \mathbb{N}$ and $t_1, \ldots, t_{N+1} \in \mathbb{R}$ with

$$t_1 = 0 < t_1 < \cdots < t_N \leq 1 \leq t_{N+1}$$

such that

$$|g(t_{i+1}-) - g(t_i)| < \delta \quad (i = 1, \ldots, N) \tag{25}$$

and

$$|g(t_{i+1}) - g(t_i)| \geq \delta \quad (i = 1, \ldots, N-1)$$

hold. Then we have

$$\sum_{i=1}^{N-1} |g(t_{i+1}) - g(t_i)| \quad \leq \quad \sum_{k=1}^{\bar{K}} |\bar{c}_k| \leq V,$$

which implies

$$(N-1) \cdot \delta \leq \sum_{i=1}^{N-1} |g(t_{i+1}) - g(t_i)| \leq V$$

and

$$N \leq \frac{V}{\delta} + 1.$$

Let $\mathcal{G}_N$ be the set of all piecewise constant functions $g : [0,1] \to \mathbb{R}$ which are piecewise constant with respect to a partition of $[0,1]$ into $N+1$ intervals. By (25) we know

$$\inf_{\bar{g} \in \mathcal{G}_N} \|g - \bar{g}\|_{\infty,[0,1]} < \delta.$$

Together with Lemma 13.1 and Example 13.1 in Györfi et al. (2002) this implies

$$\begin{aligned}
\mathcal{N}_p(2 \cdot \delta, T_L\mathcal{G}, x_1^n) &\leq \mathcal{N}_p(\delta, T_L\mathcal{G}_N, x_1^n) \\
&\leq \binom{n+N+1-1}{n} \cdot \left(\frac{2 \cdot L}{\delta}\right)^{N+1} \\
&\leq (n+N)^N \cdot \left(\frac{2 \cdot L}{\delta}\right)^{N+1} \\
&= \left(\left(n + \frac{V}{\delta} + 1\right) \cdot \frac{2 \cdot L}{\delta}\right)^{\frac{V}{\delta}+2}.
\end{aligned}$$

□

4.2. Auxiliary results concerning the approximation error

Lemma 6 *Let $r \geq 1$ and $\tilde{K}_n \in \mathbb{N}$ with*

$$\tilde{K}_n \cdot \exp\left(-\frac{(\log n)^2}{c_{12} \cdot r}\right) \leq \frac{1}{4},$$

and set $K_n = 8 \cdot (\lceil \log n \rceil)^3 \cdot \tilde{K}_n$. Let $m : \mathbb{R}^d \to \mathbb{R}$ be a function with

$$\int_{\mathbb{R}^d} |m(x)|\, dx < \infty, \tag{26}$$

and assume that the Fourier transform $\hat{F}$ of m satisfies

$$\int_{\mathbb{R}^d} \|\omega\| \cdot \sup_{\tilde{\omega} \in \mathbb{R}^d : \|\tilde{\omega}\| = \|\omega\|} |\hat{F}(\tilde{\omega})|\, d\omega < \infty. \tag{27}$$

Let (X,Y), (X_1,Y_1), ..., (X_n,Y_n) be $\mathbb{R}^d \times \mathbb{R}$–valued random variables with $supp(\mathbf{P}_X) \subseteq S_r$, and let W_1, ..., W_{K_n}, T_1, ..., T_{K_n} be independent random variables, independent from (X,Y), (X_1,Y_1), ..., (X_n,Y_n), such that W_1, ..., W_{K_n} are uniformly distributed on $\{x \in \mathbb{R}^d : \|x\| = 1\}$ and T_1, ..., T_{K_n} are uniformly distributed on $[-r,r]$. Then for n sufficiently large there exist (random)

$$\alpha_0 \in [-c_{13}, c_{13}] \quad \text{and} \quad \alpha_1, \ldots, \alpha_{K_n} \in \left[-\frac{c_{13}}{\tilde{K}_n}, \frac{c_{13}}{\tilde{K}_n}\right],$$

which are independent of (X,Y), (X_1,Y_1), ..., (X_n,Y_n), such that outside of an event with probability less than or equal to

$$\frac{1}{n}$$

one has

$$\int_{S_r} \left| m(x) - \alpha_0 - \sum_{k=1}^{K_n} \alpha_k \cdot 1_{[0,\infty)} \left(W_k^T \cdot x + T_k \right) \right|^2 \mathbf{P}_X(dx) \leq \frac{c_{14}}{\tilde{K}_n} \tag{28}$$

and

$$\min_{\substack{i=1,\ldots,n, k=1,\ldots,K_n: \\ \alpha_k \neq 0}} |W_k^T X_i + T_k| \geq \delta_n, \tag{29}$$

where

$$\delta_n = \frac{r}{16 \cdot n \cdot \tilde{K}_n \cdot (\lceil \log n \rceil)^2}.$$

Proof. The proof is an modification of the proof of Lemma 1 in Braun et al. (2021). The assertion depends only on the joint distribution of

$$(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n), (T_1, W_1), \ldots, (T_{K_n}, W_{K_n}).$$

In the sequel we construct $(T_1, W_1), \ldots, (T_{K_n}, W_{K_n})$ in a special way such that this joint distribution remains fixed and that the assertion holds.

To do this, define $g : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ by

$$g(t, \omega) = \frac{1}{2 \cdot r} \cdot 1_{[-r,r]}(t) \cdot c_{15} \cdot \|\omega\| \cdot \sup_{\tilde{\omega} \in \mathbb{R}^d : \|\tilde{\omega}\| = \|\omega\|} |\hat{F}(\tilde{\omega})|$$

where $c_{15} > 0$ is chosen such that g is a density, i.e., we have

$$c_{15} = \frac{1}{\int_{\mathbb{R}^d} \|\omega\| \cdot \sup_{\tilde{\omega} \in \mathbb{R}^d : \|\tilde{\omega}\| = \|\omega\|} |\hat{F}(\tilde{\omega})| \, d\omega}.$$

Set $\bar{K}_n = 8 \cdot \lceil \log n \rceil \cdot \tilde{K}_n$. Let $A_{1,1} = (T_{1,1}, W_{1,1})$, $A_{1,2} = (T_{1,2}, W_{1,2})$, $\ldots$, $A_{2,1} = (T_{2,1}, W_{2,1})$, $A_{2,2} = (T_{2,2}, W_{2,2})$, $\ldots$, $A_{\bar{K}_n,1} = (T_{\bar{K}_n,1}, W_{\bar{K}_n,1})$, $A_{\bar{K}_n,2} = (T_{\bar{K}_n,2}, W_{\bar{K}_n,2})$, $\ldots$, be independent and identically distributed random variables with density g, which are independent from (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n), and let $U_{1,1}$, $U_{2,1}$, $\ldots$, $U_{\bar{K}_n,1}$, $U_{1,2}$, $U_{2,2}$, $\ldots$, $U_{\bar{K}_n,2}$, $\ldots$ be independent uniformly on $[0, 1]$ distributed random variables, which are independent of all other random variables.

Since g is a product of the density of an uniform distribution on $[-r, r]$ and a radially symmetric density on $\mathbb{R}^d$, we can assume without loss of generality that (T_1, W_1), $\ldots$, (T_{K_n}, W_{K_n}) are in fact given by

$$\begin{aligned}
&((-1) \cdot T_{1,1}, \frac{W_{1,1}}{\|W_{1,1}\|}), \ldots, ((-1) \cdot T_{1,(\lceil \log n \rceil)^2}, \frac{W_{1,(\lceil \log n \rceil)^2}}{\|W_{1,(\lceil \log n \rceil)^2}\|}), \ldots, \\
&((-1) \cdot T_{\bar{K}_n,1}, \frac{W_{\bar{K}_n,1}}{\|W_{\bar{K}_n,1}\|}), \ldots, ((-1) \cdot T_{\bar{K}_n,(\lceil \log n \rceil)^2}, \frac{W_{\bar{K}_n,(\lceil \log n \rceil)^2}}{\|W_{\bar{K}_n,C_n}\|}),
\end{aligned}$$

hence it suffices to show that there exist (random)

$$\alpha_0 \in [-c_{13}, c_{13}] \quad \text{and} \quad \alpha_{1,1}, \ldots, \alpha_{1,(\lceil \log n \rceil)^2}, \ldots, \alpha_{\bar{K}_n,1}, \ldots \alpha_{\bar{K}_n,(\lceil \log n \rceil)^2} \in \left[-\frac{c_{13}}{\tilde{K}_n}, \frac{c_{13}}{\tilde{K}_n} \right],$$

which are independent of (X,Y), (X_1,Y_1), $\ldots$, (X_n,Y_n), such that outside of an event with probability less than or equal to $1/n$ we have

$$\int_{S_r} \left| m(x) - \alpha_0 - \sum_{k=1}^{\bar{K}_n} \sum_{j=1}^{(\lceil \log n \rceil)^2} \alpha_{k,j} \cdot 1_{[0,\infty)} \left(\frac{W_{k,j}^T}{\|W_{k,j}\|} \cdot x - T_{k,j} \right) \right|^2 \mathbf{P}_X(dx) \leq \frac{c_{14}}{\tilde{K}_n} \tag{30}$$

and

$$\min_{\substack{i=1,\ldots,n, k=1,\ldots,\bar{K}_n, j=1,\ldots,(\lceil \log n \rceil)^2: \\ \alpha_{k,j} \neq 0}} |\frac{W_{k,j}^T}{\|W_{k,j}\|} \cdot X_i - T_{k,j}| \geq \delta_n. \tag{31}$$

Define

$$h(t,\omega) = \sin(\|\omega\| \cdot t + \theta(\omega)) \cdot \|\omega\| \cdot |\hat{F}(\omega)|,$$

where $\theta(\omega) \in [0, 2\pi)$,

$$\hat{F}(\omega) = e^{i \cdot \theta(\omega)} \cdot |\hat{F}(\omega)|,$$

and

$$f(t,\omega) = \begin{cases} c_{16} \cdot |h(t,\omega)| & \text{if } ((t,\omega) \in [-r,r] \times \mathbb{R}^d, \\ 0 & \text{if } ((t,\omega) \in (\mathbb{R} \setminus [-r,r]) \times \mathbb{R}^d, \end{cases}$$

where

$$c_{16} = \frac{1}{\int_{[-r,r] \times \mathbb{R}^d} |h(t,\omega)| d(t,\omega)}.$$

Then f is a density on $\mathbb{R} \times \mathbb{R}^d$ and

$$\begin{aligned} 0 &< \frac{1}{2 \cdot r \cdot \int_{\mathbb{R}^d} \|\omega\| \cdot |\hat{F}(\omega)| \, d\omega} \leq c_{16} \\ &= \frac{1}{\int_{[-r,r] \times \mathbb{R}^d} |\sin(\|\omega\| \cdot t + \theta(\omega))| \cdot \|\omega\| \cdot |\hat{F}(\omega)| d(t,\omega)} < \infty. \end{aligned}$$

By the definitions of f and g we know

$$f(t,\omega) \leq 2 \cdot r \cdot \frac{1}{2 \cdot r} \cdot 1_{[-r,r]}(t) \cdot c_{16} \cdot \|\omega\| \cdot \sup_{\tilde{\omega} \in \mathbb{R}^d : \|\tilde{\omega}\| = \|\omega\|} |\hat{F}(\tilde{\omega})| = c_{17} \cdot g(t,\omega) \quad ((t,\omega) \in \mathbb{R} \times \mathbb{R}^d), \tag{32}$$

where

$$c_{17} = 2 \cdot r \cdot \frac{c_{16}}{c_{15}}.$$

From the properties of the Fourier transform and the fact that m is real–valued, it follows that

$$\begin{aligned} & m(x) - m(0) \\ &= Re \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left(e^{i\omega^T x} - e^{i\omega^T 0} \right) \cdot \hat{F}(\omega) \, d\omega \\ &= Re \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left(e^{i\omega^T x} - 1 \right) \cdot e^{i \cdot \theta(\omega)} \cdot |\hat{F}(\omega)| \, d\omega \end{aligned}$$

$$
\begin{aligned}
&= Re\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\left(e^{i(\omega^T x+\theta(\omega))}-e^{i\cdot\theta(\omega)}\right)\cdot|\hat{F}(\omega)|\,d\omega\\
&= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\left(\cos(\omega^T x+\theta(\omega))-\cos(\theta(\omega))\right)\cdot|\hat{F}(\omega)|\,d\omega\\
&= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\frac{\left(\cos\left(\|\omega\|\cdot\frac{\omega^T x}{\|\omega\|}+\theta(\omega)\right)-\cos(\theta(\omega))\right)}{\|\omega\|}\cdot\|\omega\|\cdot|\hat{F}(\omega)|\,d\omega\\
&= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\int_{\frac{\omega^T x}{\|\omega\|}}^{0}\sin\left(\|\omega\|\cdot t+\theta(\omega)\right)\,dt\cdot\|\omega\|\cdot|\hat{F}(\omega)|\,d\omega.
\end{aligned}
$$

Assume $\|x\|\le r$, which implies

$$
|\frac{\omega^T x}{\|\omega\|}|\le\frac{\|\omega\|\cdot\|x\|}{\|\omega\|}\le r.
$$

By considering the cases $\frac{\omega^T x}{\|\omega\|}\ge 0$ and $\frac{\omega^T x}{\|\omega\|}<0$ separately, we get

$$
\begin{aligned}
&\int_{\frac{\omega^T x}{\|\omega\|}}^{0}\sin\left(\|\omega\|\cdot t+\theta(\omega)\right)\,dt\\
&=-\int_0^r 1_{[0,\infty)}\left(\frac{\omega^T x}{\|\omega\|}-t\right)\cdot\sin\left(\|\omega\|\cdot t+\theta(\omega)\right)\,dt\\
&\quad+\int_{-r}^0 1_{(0,\infty)}\left(t-\frac{\omega^T x}{\|\omega\|}\right)\cdot\sin\left(\|\omega\|\cdot t+\theta(\omega)\right)\,dt\\
&=-\int_0^r 1_{[0,\infty)}\left(\frac{\omega^T x}{\|\omega\|}-t\right)\cdot\sin\left(\|\omega\|\cdot t+\theta(\omega)\right)\,dt\\
&\quad+\int_{-r}^0\left(1-1_{[0,\infty)}\left(\frac{\omega^T x}{\|\omega\|}-t\right)\right)\cdot\sin\left(\|\omega\|\cdot t+\theta(\omega)\right)\,dt\\
&=\int_{-r}^0\sin\left(\|\omega\|\cdot t+\theta(\omega)\right)\,dt\\
&\quad-\int_{-r}^r 1_{[0,\infty)}\left(\frac{\omega^T x}{\|\omega\|}-t\right)\cdot\sin\left(\|\omega\|\cdot t+\theta(\omega)\right)\,dt.
\end{aligned}
$$

Consequently we get for any $x\in\mathbb{R}^d$, $\|x\|\le r$

$$
\begin{aligned}
m(x) &= m(0)+\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\int_{\frac{\omega^T x}{\|\omega\|}}^{0}\sin\left(\|\omega\|\cdot t+\theta(\omega)\right)\,dt\cdot\|\omega\|\cdot|\hat{F}(\omega)|\,d\omega\\
&= m(0)+\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\int_{-r}^{0}\sin\left(\|\omega\|\cdot t+\theta(\omega)\right)\,dt\cdot\|\omega\|\cdot|\hat{F}(\omega)|\,d\omega\\
&\quad-\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\int_{-r}^{r}1_{[0,\infty)}\left(\frac{\omega^T x}{\|\omega\|}-t\right)\cdot\sin\left(\|\omega\|\cdot t+\theta(\omega)\right)\,dt\cdot\|\omega\|\cdot|\hat{F}(\omega)|\,d\omega\\
&= c_{18}-\frac{1}{(2\pi)^{d/2}}\cdot\int_{[-r,r]\times\mathbb{R}^d}1_{[0,\infty)}\left(\frac{\omega^T x}{\|\omega\|}-t\right)\cdot h(t,\omega)\,d(t,\omega)
\end{aligned}
$$

$$= c_{18} - \frac{1}{(2\pi)^{d/2}} \cdot \frac{1}{c_{16}} \cdot \int_{[-r,r]\times\mathbb{R}^d} 1_{[0,\infty)}\left(\frac{\omega^T x}{\|\omega\|} - t\right) \cdot sgn(h(t,\omega)) \cdot f(t,\omega)\, d(t,\omega).$$

Here c_{18} is a constant which is bounded independent of r since

$$\begin{aligned}
&|\int_{\mathbb{R}^d}\int_{-r}^{0} \sin\left(\|\omega\| \cdot t + \theta(\omega)\right)\, dt \cdot \|\omega\| \cdot |\hat{F}(\omega)|\, d\omega| \\
&= |\int_{\mathbb{R}^d} ((-1)\cdot\cos\left(\|\omega\|\cdot 0 + \theta(\omega)\right) + \cos\left(\|\omega\|\cdot r + \theta(\omega)\right)) \cdot |\hat{F}(\omega)|\, d\omega| \\
&\leq 2 \cdot \int_{\mathbb{R}^d} |\hat{F}(\omega)|\, d\omega < \infty
\end{aligned}$$

(where the last inequality followed from (26)).

For $k \in \{1,\ldots,\bar{K}_n\}$ let $j_k \in \mathbb{N}$ be the minimal $j \in \mathbb{N}$ which satisfies

$$U_{k,j} \leq \frac{f(A_{k,j})}{c_{17}\cdot g(A_{k,j})}.$$

For any $i \in \mathbb{N}$ we have

$$\begin{aligned}
&\mathbf{P}\left\{U_{k,i} \leq \frac{f(A_{k,i})}{c_{17}\cdot g(A_{k,i})}\right\} = \mathbf{E}\left\{\mathbf{P}\left\{U_{k,i} \leq \frac{f(A_{k,i})}{c_{17}\cdot g(A_{k,i})}\big|A_{k,i}\right\}\right\} = \mathbf{E}\left\{\frac{f(A_{k,i})}{c_{17}\cdot g(A_{k,i})}\right\} \\
&= \int_{\mathbb{R}\times\mathbb{R}^d} \frac{f(z)}{c_{17}\cdot g(z)} \cdot g(z)\, dz = \int_{\mathbb{R}\times\mathbb{R}^d} f(z)\, dz \cdot \frac{1}{c_{17}} = \frac{1}{c_{17}},
\end{aligned}$$

hence

$$\mathbf{P}\left\{\exists i \in \mathbb{N} : U_{k,i} \leq \frac{f(A_{k,i})}{c_{17}\cdot g(A_{k,i})}\right\} = \sum_{i=1}^{\infty}\left(1-\frac{1}{c_{17}}\right)^{i-1}\cdot\frac{1}{c_{17}} = 1.$$

Consequently $j_k \in \mathbb{N}$ exists with probability one.

Furthermore we have for any measurable $B \subseteq \mathbb{R}\times\mathbb{R}^d$

$$\begin{aligned}
\mathbf{P}\{A_{k,j_k} \in B\} &= \sum_{i=1}^{\infty} \mathbf{P}\{j_k = i, A_{k,i} \in B\} \\
&= \sum_{i=1}^{\infty}\left(1-\frac{1}{c_{17}}\right)^{i-1}\cdot\mathbf{P}\left\{U_{k,i} \leq \frac{f(A_{k,i})}{c_{17}\cdot g(A_{k,i})}, A_{k,i} \in B\right\} \\
&= \sum_{i=1}^{\infty}\left(1-\frac{1}{c_{17}}\right)^{i-1}\cdot\mathbf{E}\left\{\frac{f(A_{k,i})}{c_{17}\cdot g(A_{k,i})}\cdot 1_B(A_{k,i})\right\} \\
&= \sum_{i=1}^{\infty}\left(1-\frac{1}{c_{17}}\right)^{i-1}\cdot\int_{\mathbb{R}\times\mathbb{R}^d}\frac{f(z)}{c_{17}\cdot g(z)}\cdot I_B(z)\cdot g(z)\, dz \\
&= \int_B f(z)\, dz.
\end{aligned}$$

Hence for any $l \in \{1, \dots, 8 \cdot \lceil \log n \rceil\}$ the random elements $(A_{k,j_k})_{k=(l-1)\cdot \tilde{K}_n+1,\dots,(l-1)\cdot \tilde{K}_n+\tilde{K}_n}$ are independent with density f and

$$\mathbf{P}\{j_k = i\} = \left(1 - \frac{1}{c_{17}}\right)^{i-1} \cdot \frac{1}{c_{17}}$$

holds for any $i \in \mathbb{N}$, $k \in \{1, \dots, \bar{K}_n\}$.

Set

$$f_{l,\tilde{K}_n}(x) = c_{18} + \frac{1}{\tilde{K}_n} \cdot \sum_{k=(l-1)\cdot \tilde{K}_n+1}^{(l-1)\cdot \tilde{K}_n+\tilde{K}_n} \frac{(-1)}{(2\pi)^{d/2} \cdot c_{16}} \cdot sign(h(T_{k,j_k}, W_{k,j_k})) \cdot 1_{[0,\infty)} \left(\frac{W_{k,j_k}^T \cdot x}{\|W_{k,j_k}\|} - T_{k,j_k} \right).$$

For any $x \in S_r$ we have that the random variables

$$Z_k = \frac{(-1)}{(2\pi)^{d/2} \cdot c_{16}} \cdot sign(h(T_{k,j_k}, W_{k,j_k})) \cdot 1_{[0,\infty)} \left(\frac{W_{k,j_k}^T \cdot x}{\|W_{k,j_k}\|} - T_{k,j_k} \right)$$

$(k = (l-1) \cdot \tilde{K}_n + 1, \dots, (l-1) \cdot \tilde{K}_n + \tilde{K}_n)$ are independent and identically distributed with expectation equal to $m(x) - c_{18}$. Consequently we have

$$\begin{aligned}
\mathbf{E} \int_{S_r} |m(x) - f_{l,\tilde{K}_n}(x)|^2 \mathbf{P}_X(dx) &= \int_{S_r} \mathbf{E} |m(x) - f_{l,\tilde{K}_n}(x)|^2 \mathbf{P}_X(dx) \\
&= \int_{S_r} \mathbf{Var} \left(\frac{1}{\tilde{K}_n} \cdot \sum_{k=(l-1)\cdot \tilde{K}_n+1}^{(l-1)\cdot \tilde{K}_n+\tilde{K}_n} Z_k \right) \mathbf{P}_X(dx) \\
&= \frac{1}{\tilde{K}_n} \cdot \int_{S_r} \mathbf{Var}\left(Z_1\right) \mathbf{P}_X(dx) \\
&\leq \frac{1}{\tilde{K}_n} \cdot \int_{S_r} \mathbf{E}\left(Z_1^2\right) \mathbf{P}_X(dx) \\
&\leq \frac{1}{\tilde{K}_n} \cdot \frac{1}{(2\pi)^d \cdot c_{16}^2} \cdot \mathbf{E} \left\{ 1_{[0,\infty)} \left(\frac{W_{1,j_1}^T \cdot x}{\|W_{1,j_1}\|} - T_{1,j_1} \right) \right\} \leq \frac{c_{19}}{\tilde{K}_n}.
\end{aligned}$$

By Markov inequality this implies that we have for any $l \in \{1, \dots, (\lceil \log n \rceil)^2\}$ at least with probability $1/2$

$$\int_{S_r} |m(x) - f_{l,\tilde{K}_n}(x)|^2 \mathbf{P}_X(dx) \leq \frac{2 \cdot c_{19}}{\tilde{K}_n}.$$

For any $l \in \{1, \dots, 8 \cdot \lceil \log n \rceil\}$ the probability that $j_k > (\lceil \log n \rceil)^2$ holds for some $k \in \{(l-1) \cdot \tilde{K}_n + 1, \dots, l \cdot \tilde{K}_n\}$ is bounded from above by

$$\tilde{K}_n \cdot \left(1 - \frac{1}{c_{17}}\right)^{(\lceil \log n \rceil)^2} \leq \tilde{K}_n \cdot \exp\left(- \frac{(\lceil \log n \rceil)^2}{c_{17}} \right) \leq \frac{1}{4}.$$

Furthermore we have

$$\begin{aligned}
&\mathbf{P}\Bigg\{\exists i \in \{1,\dots,n\}\, \exists k \in \{(l-1)\cdot \tilde{K}_n + 1,\dots, l\cdot \tilde{K}_n : \\
&\qquad\qquad j_k \le (\lceil \log n \rceil)^2 \quad \text{and} \quad \left| \frac{W_{k,j_k}^T \cdot X_i}{\|W_{k,j_k}\|} - T_{k,j_k} \right| < \delta_n \Bigg\} \\
&\le \mathbf{P}\Bigg\{\exists i \in \{1,\dots,n\}\, \exists k \in \{(l-1)\cdot \tilde{K}_n + 1,\dots, l\cdot \tilde{K}_n\, \exists j \in \{1,\dots,(\lceil \log n \rceil)^2\} : \\
&\qquad\qquad \left| \frac{W_{k,j}^T \cdot X_i}{\|W_{k,j}\|} - T_{k,j} \right| < \delta_n \Bigg\} \\
&\le n \cdot \tilde{K}_n \cdot (\lceil \log n \rceil)^2 \cdot \frac{2\cdot \delta_n}{r} \le \frac{1}{8}.
\end{aligned}$$

Hence for any $l \in \{1,\dots,8\cdot\lceil \log n\rceil\}$ at least with probability $1/8$ we know that

$$\int_{S_r} |m(x) - f_{l,\tilde{K}_n}(x)|^2 \mathbf{P}_X(dx) \le \frac{2\cdot c_{19}}{\tilde{K}_n},$$

$$j_k \le (\lceil \log n \rceil)^2 \text{ for all } k \in \{(l-1)\cdot \tilde{K}_n + 1,\dots, l\cdot \tilde{K}_n\}$$

and

$$\min_{i\in\{1,\dots,n\}, k\in\{(l-1)\cdot \tilde{K}_n+1,\dots,l\cdot\tilde{K}_n\}} \left| \frac{W_{k,j_k}^T \cdot X_i}{\|W_{k,j_k}\|} - T_{k,j_k} \right| \ge \delta_n$$

hold. Consequently it holds at least with probability

$$1 - \left(1 - \frac{1}{8}\right)^{8\cdot(\lceil \log n\rceil)} \ge 1 - \exp\left(-\frac{1}{8}\cdot 8 \cdot (\lceil \log n\rceil)\right) \ge 1 - \frac{1}{n}$$

that there exists $l \in \{1,\dots,8\cdot(\lceil \log n\rceil)\}$ such that we have

$$\int_{S_r} |m(x) - f_{l,\tilde{K}_n}(x)|^2 \mathbf{P}_X(dx) \le \frac{2\cdot c_{19}}{\tilde{K}_n},$$

$$j_k \le (\lceil \log n \rceil)^2 \text{ for all } k \in \{(l-1)\cdot \tilde{K}_n + 1,\dots, l\cdot \tilde{K}_n\}$$

and

$$\min_{i\in\{1,\dots,n\}, k\in\{(l-1)\cdot \tilde{K}_n+1,\dots,l\cdot\tilde{K}_n, j\in\{1,\dots,(\lceil \log n\rceil)^2\}} \left| \frac{W_{k,j_k}^T \cdot x}{\|W_{k,j_k}\|} - T_{k,j_k} \right| \ge \delta_n.$$

Let $l \in \{1,\dots,(\lceil \log n\rceil)^2\}$ be minimal such that the above three properties hold and set

$$\alpha_{k,j_k} = -\frac{1}{(2\pi)^{d/2}\cdot c_{16}} \cdot \frac{1}{\tilde{K}_n} \cdot sign(h(T_{k,j_k}, W_{k,j_k})) \in \left[-c_{13}\cdot\frac{1}{\tilde{K}_n}, c_{13}\cdot\frac{1}{\tilde{K}_n}\right]$$

for all $k \in \{(l-1)\cdot \tilde{K}_n + 1,\dots, l\cdot \tilde{K}_n\}$ and set all other $\alpha_{k,j} = 0$. (In case that the above two properties do not hold for all $l \in \{1,\dots,8\cdot(\lceil \log n\rceil)\}$ set $\alpha_{k,j} = 0$ for all k and j.)

Then (30) and (31) hold with probability at least $1 - n$. $\square$

4.3. Auxiliary results concerning the optimization error

Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, let $K \in \mathbb{N}$, let $B_1, \ldots, B_K : \mathbb{R}^d \to \mathbb{R}$ and let $c_2 > 0$. In this subsection we consider the problem to minimize

$$F(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} |a_0 + \sum_{k=1}^{K} a_k \cdot B_k(x_i) - y_i|^2 + c_2 \cdot \left(\frac{a_0^2}{n^{2/3}} + \frac{K_n}{n^{2/3}} \cdot \sum_{k=1}^{K_n} a_k^2 \right), \tag{33}$$

where $\mathbf{a} = (a_0, \ldots, a_K)^T$, by gradient descent. To do this, we choose $\mathbf{a}^{(0)} \in \mathbb{R}^K$ and set

$$\mathbf{a}^{(t+1)} = \mathbf{a}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)}) \tag{34}$$

for some properly chosen $\lambda_n > 0$.

Lemma 7 *Let F be defined by (33) and choose $\mathbf{a}_{opt}$ such that*

$$F(\mathbf{a}_{opt}) = \min_{\mathbf{a} \in \mathbb{R}^{K+1}} F(\mathbf{a}).$$

Then for any $\mathbf{a} \in \mathbb{R}^{K+1}$ we have

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2 \geq \frac{4 \cdot c_2}{n^{2/3}} \cdot (F(\mathbf{a}) - F(\mathbf{a}_{opt})).$$

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

Set $B_0(x) = 1$,

$$\mathbf{E} = \frac{c_2}{n^{2/3}} \cdot \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & K_n & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & K_n \end{pmatrix},$$

$$\mathbf{B} = (B_j(x_i))_{1 \leq i \leq n, 0 \leq j \leq K} \quad \text{and} \quad \mathbf{A} = \frac{1}{n} \cdot \mathbf{B}^T \cdot \mathbf{B} + \frac{c_2}{n^{2/3}} \cdot \mathbf{E}.$$

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

$$\begin{aligned} F(\mathbf{a}) &= \frac{1}{n} \cdot (\mathbf{B} \cdot \mathbf{a} - \mathbf{y})^T \cdot (\mathbf{B} \cdot \mathbf{a} - \mathbf{y}) + \frac{c_2}{n^{2/3}} \cdot \mathbf{a}^T \cdot \mathbf{E} \cdot \mathbf{a} \\ &= \mathbf{a}^T \mathbf{A} \mathbf{a} - 2 \mathbf{y}^T \frac{1}{n} \mathbf{B} \mathbf{a} + \frac{1}{n} \mathbf{y}^T \mathbf{y} \end{aligned}$$

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

where

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

Using

$$\mathbf{b}^T\mathbf{A}\mathbf{b} \geq \frac{c_2}{n^{2/3}} \cdot \mathbf{b}^T\mathbf{E}\mathbf{b} \geq \frac{c_2}{n^{2/3}} \cdot \mathbf{b}^T\mathbf{b}$$

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

$$\begin{aligned}
&F(\mathbf{a}) - F(\mathbf{a}_{opt}) \\
&= ((\mathbf{A}^{1/2})^T(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}))^T\mathbf{A}^{1/2}(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}) \\
&\leq \frac{n^{2/3}}{c_2} \cdot ((\mathbf{A}^{1/2})^T(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}))^T\mathbf{A}\mathbf{A}^{1/2}(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}) \\
&= \frac{n^{2/3}}{c_2} \cdot ((\mathbf{A})^T(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}))^T\mathbf{A}(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}) \\
&= \frac{n^{2/3}}{c_2} \cdot (\mathbf{A}\mathbf{a} - \frac{1}{n}\mathbf{B}^T\mathbf{y})^T(\mathbf{A}\mathbf{a} - \frac{1}{n}\mathbf{B}^T\mathbf{y}) \\
&= \frac{n^{2/3}}{4 \cdot c_2} \cdot (2\mathbf{A}\mathbf{a} - \frac{2}{n}\mathbf{B}^T\mathbf{y})^T(2\mathbf{A}\mathbf{a} - \frac{2}{n}\mathbf{B}^T\mathbf{y}) \\
&= \frac{n^{2/3}}{4 \cdot c_2} \cdot \|(\nabla_{\mathbf{a}}F)(\mathbf{a})\|^2,
\end{aligned}$$

where the last equality follows from

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

□

4.4. Proof of Theorem 1

Our proof is an extension of the proof of Theorem 1 in Braun et al. (2021).

W.l.o.g. we assume $\|m\|_\infty \leq \beta_n$. Set $\tilde{K}_n = \lfloor K_n/(8 \cdot (\lceil \log n \rceil)^3) \rfloor$ and let A_n be the event that $|Y_i| \leq \beta_n$ holds for all $i = 1, \ldots, n$ and that there exist (random)

$$\alpha_0 \in [-c_{13}, c_{13}] \quad \text{and} \quad \alpha_1, \ldots, \alpha_{K_n} \in \left[-\frac{c_{13}}{\tilde{K}_n}, \frac{c_{13}}{\tilde{K}_n}\right],$$

which are independent of (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n), such that

$$\int_{[0,1]^d} \left| m(x) - \alpha_0 - \sum_{k=1}^{K_n} \alpha_k \cdot 1_{[0,\infty)}\left(W_k^T \cdot x + T_k\right) \right|^2 \mathbf{P}_X(dx) \leq \frac{c_{14}}{\tilde{K}_n} \tag{35}$$

and

$$\min_{\substack{i=1,\dots,n,k=1,\dots,K_n: \\ \alpha_k \neq 0}} \left| (\beta_k^{(0)})^T X_i + \gamma_k^{(0)} \right| \geq \delta_n \tag{36}$$

hold for $\delta_n = \frac{B_n}{16 \cdot n \cdot \tilde{K}_n \cdot (\lceil \log n \rceil)^2} \geq 6 \cdot \log n$.

We have

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&= \mathbf{E} \left(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{A_n} \right) + \mathbf{E} \left(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{A_n^c} \right) \\
&\leq \mathbf{E} \left(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{A_n} \right) + 4\beta_n^2 \cdot \mathbf{P}(A_n^c) \\
&= \mathbf{E} \Bigg(\Bigg(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\quad -2 \cdot \left(\frac{1}{n} \sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| \leq \beta_n \, (j \in \{1,\dots,n\})\}} - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right) \Bigg) \cdot 1_{A_n} \Bigg) \\
&\quad +2 \cdot \mathbf{E} \left(\left(\frac{1}{n} \sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| \leq \beta_n \, (j \in \{1,\dots,n\})\}} - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right) \cdot 1_{A_n} \right) \\
&\quad +4\beta_n^2 \cdot \mathbf{P}(A_n^c) \\
&=: T_{1,n} + T_{2,n} + T_{3,n}.
\end{aligned}$$

In the remainder of the proof we derive bounds on $T_{i,n}$ for $i \in \{1,2,3\}$.

In the *first step of the proof* we show that we have on A_n

$$F_n(\mathbf{w}(0)) \leq c_{20} \cdot (\log n)^2.$$

On A_n it holds

$$F_n(\mathbf{w}(0)) \quad = \quad \frac{1}{n} \sum_{i=1}^n |0 - Y_i|^2 + c_2 \cdot \left(\frac{0^2}{n^{2/3}} + \frac{K_n}{n^{2/3}} \cdot \sum_{k=1}^{K_n} 0^2 \right) \leq c_3^2 \cdot (\log n)^2.$$

In the *second step of the proof* we show that we have on A_n

$$\sum_{k=1}^{K_n} |\alpha_k^{(t_n)}| \leq c_{21} \cdot (\log n) \cdot n^{1/3}.$$

To do this, we show first that on A_n the assumptions (20) and (21) of Lemma 1 hold, i.e., we show that on A_n we have

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w})\| \leq \sqrt{2 \cdot t_n \cdot L_n \cdot \max\{F_n(\mathbf{w}(0)), 1\}} \tag{37}$$

for all $\mathbf{w}$ with

$$\|\mathbf{w} - \mathbf{w}(0)\| \leq \sqrt{2 \cdot t_n \cdot \max\{F_n(\mathbf{w}(0)), 1\}/L_n}$$

and

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}_2)\| \leq L_n \cdot \|\mathbf{w}_1 - \mathbf{w}_2\| \tag{38}$$

for all $\mathbf{w}_1, \mathbf{w}_2$ satisfying

$$\|\mathbf{w}_1 - \mathbf{w}(0)\| \leq \sqrt{8 \cdot \frac{t_n}{L_n} \cdot \max\{F_n(\mathbf{w}(0)), 1\}}$$

and

$$\|\mathbf{w}_2 - \mathbf{w}(0)\| \leq \sqrt{8 \cdot \frac{t_n}{L_n} \cdot \max\{F_n(\mathbf{w}(0)), 1\}}.$$

By the result of the first step we know that we have $F_n(\mathbf{w}(0)) \leq c_{20} \cdot (\log n)^2$. Together with

$$|w_{1,k}^{(1)}| \leq \|\mathbf{w} - \mathbf{w}(0)\| + 0 \quad (k = 1, \ldots, K_n)$$

(where we have used that the initial weights satisfy $\alpha_k = 0$ $(k = 0, \ldots, K_n)$) and

$$\frac{t_n}{L_n} \leq 2 \cdot n^{2/3} \cdot (\log n)^2$$

this implies that in order to prove (37) we can assume that the assumptions of Lemma 2 hold with $\gamma_n^* = c_{21} \cdot n^{1/3} \cdot (\log n)^2$. From this and

$$c_{22} \cdot (\log n)^4 \cdot n^{2/3} \cdot K_n^{3/2} \leq L_n$$

we can derive (37) by an application of Lemma 2. In the same way we can prove (38) by applying Lemma 3 with $\gamma_n^* = c_{21} \cdot n^{1/3} \cdot (\log n)^2$.

From this we can conclude by Lemma 1 and the first step of the proof that we have on A_n

$$F(\mathbf{w}(t_n)) \leq F(\mathbf{w}(0)) \leq c_3^2 \cdot (\log n)^2, \tag{39}$$

which implies

$$\begin{aligned}
\left(\sum_{k=1}^{K_n} |\alpha_k^{(t_n)}|\right)^2 &\leq K_n \cdot \sum_{k=1}^{K_n} |\alpha_k^{(t_n)}|^2 \\
&\leq \frac{n^{2/3}}{c_2} \cdot c_2 \cdot \left(\frac{(\alpha_0^{(t_n)})^2}{n^{2/3}} + \frac{K_n}{n^{2/3}} \cdot \sum_{k=1}^{K_n} (\alpha_k^{(t_n)})^2\right) \\
&\leq \frac{n^{2/3}}{c_2} \cdot F(\mathbf{w}(t_n)) \leq \frac{n^{2/3}}{c_2} \cdot F(\mathbf{w}(0)) \\
&\leq \frac{c_3^2}{c_2} \cdot n^{2/3} \cdot (\log n)^2.
\end{aligned}$$

In the *third step of the proof* we show

$$T_{1,n} \leq c_{23} \cdot (\log n)^4 \cdot n^{-2/3}. \tag{40}$$

By the results of the second step we know that on A_n it holds

$$\tilde{m}_n(x) = \alpha_0^{(t_n)} + \sum_{k=1}^{K_n} \alpha_k^{(t_n)} \cdot \sigma(\beta_k^{(t_n)} \cdot x + \gamma_k^{(t_n)}) \in \mathcal{F},$$

where $\mathcal{F}$ is defined as in Lemma 5 with $V = c_{21} \cdot (\log n) \cdot n^{1/3}$. By Lemma 5 we know

$$\begin{aligned}
&\log \mathcal{N}_2\left(\delta, T_{\beta_n}\mathcal{F}, x_1^n\right) \\
&\leq \left(\frac{c_{21} \cdot (\log n) \cdot n^{1/3}}{\delta} + 2\right) \cdot \log\left(\left(n + \frac{c_{21} \cdot (\log n) \cdot n^{1/3}}{\delta/3} + 1\right) \cdot \frac{2 \cdot \beta_n}{\delta/3}\right).
\end{aligned}$$

Consequently,

$$c_9 \cdot \frac{\sqrt{n} \cdot \delta}{\beta_n} \geq \int_{c_{10} \cdot \delta}^{\sqrt{\delta}} (\log \mathcal{N}_2(u, T_{\beta_n}\mathcal{F}, x_1^n))^{1/2}\, du$$

is for $\delta \geq c_{23} \cdot n^{-2/3}/(8 \cdot \beta_n^2)$ implied by

$$\frac{\sqrt{n} \cdot \delta}{\beta_n} \geq c_{24} \cdot (\log n) \cdot n^{1/6} \cdot \sqrt{\delta},$$

which in turn is implied by

$$\delta \geq c_{25} \cdot (\log n)^4 \cdot n^{-2/3}.$$

Application of Lemma 4 yields the assertion.

In the *fourth step of the proof* we show

$$T_{3,n} \leq c_{26} \cdot \frac{(\log n)^2}{n}. \tag{41}$$

By Lemma 6 (applied with $r = 1$) we get

$$\begin{aligned}
T_{3,n} &\leq 4 \cdot \beta_n^2 \cdot \left(\frac{1}{n} + \mathbf{P}\left\{|Y_i| > \beta_n \text{ for some } i \in \{1, \ldots, n\}\right\}\right) \\
&= c_{27} \cdot \frac{(\log n)^2}{n} + 4 \cdot \beta_n^2 \cdot \mathbf{P}\left\{|Y_i| > \beta_n \text{ for some } i \in \{1, \ldots, n\}\right\}.
\end{aligned}$$

Using (18), which implies

$$\begin{aligned}
\mathbf{P}\left\{|Y_i| > \beta_n \text{ for some } i \in \{1, \ldots, n\}\right\} &\leq n \cdot \mathbf{P}\left\{\exp(c_4 \cdot Y^2) > \exp(c_4 \cdot \beta_n^2)\right\} \\
&\leq n \cdot \frac{\mathbf{E}\left\{\exp(c_4 \cdot Y^2)\right\}}{\exp\left(c_4 \cdot c_3^2 \cdot (\log n)^2\right)} \leq \frac{c_{28}}{n^3}, \qquad (42)
\end{aligned}$$

we get (41).

In the *fifth step of the proof* we derive an upper bound on $T_{2,n}$ by a sum of several terms. For that, let $\alpha_0 \ldots \alpha_{K_n}$ be defined as in the definition of the event A_n and define on $[0,1]$ a piecewise constant approximation of m by

$$f(x) = \alpha_0 + \sum_{k=1}^{K_n} \alpha_k \cdot 1_{[0,\infty)}\left(\sum_{j=1}^{d} w_{k,j}^{(0)} \cdot x^{(j)} + w_{k,0}^{(0)}\right).$$

(In case that A_n does not hold set $\alpha_0 = \alpha_1 = \cdots = \alpha_{K_n} = 0$.) Set

$$f^*(x) = \alpha_0 + \sum_{k=1}^{K_n} \alpha_k \cdot \sigma \left(\sum_{j=1}^{d} w_{k,j}^{(0)} \cdot x^{(j)} + w_{k,1}^{(0)} \right)$$

For $g(x) = \alpha_0 + \sum_{k=1}^{K_n} \alpha_k \cdot \sigma(\beta_k^T \cdot x + \gamma_k)$ we define

$$\mathrm{pen}(g) = c_2 \cdot \left(\frac{\alpha_0^2}{n^{2/3}} + \frac{K_n}{n^{2/3}} \cdot \sum_{k=1}^{K_n} \alpha_k^2 \right).$$

We have

$$\begin{aligned}
&\frac{1}{2} \cdot T_{2,n} \\
&= \mathbf{E} \left(\left(\frac{1}{n} \sum_{i=1}^{n} |\tilde{m}_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| \leq \beta_n \, (j \in \{1,\ldots,n\})\}} - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \cdot 1_{A_n} \right) \\
&\leq \mathbf{E} \Bigg(\Bigg(\frac{1}{n} \sum_{i=1}^{n} |\tilde{m}_n(X_i) - Y_i|^2 + c_2 \cdot \left(\frac{(\alpha_0^{(t_n)})^2}{n^{2/3}} + \frac{K_n}{n^{2/3}} \cdot \sum_{k=1}^{K_n} (\alpha_k^{(t_n)})^2 \right) \\
&\quad - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \Bigg) \cdot 1_{A_n} \Bigg) \\
&= \mathbf{E} \left(\left(F(\alpha^{(t_n)}, \beta^{(t_n)}, \gamma^{(t_n)}) - \frac{1}{n} \sum_{i=1}^{n} |Y_i - f^*(X_i)|^2 - pen(f^*) \right) \cdot 1_{A_n} \right) \\
&\quad + \mathbf{E} \left(\left(\frac{1}{n} \sum_{i=1}^{n} (|Y_i - f^*(X_i)|^2 - \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|^2 \right) \cdot 1_{A_n} \right) \\
&\quad + \mathbf{E} \left(\left(\frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \cdot 1_{A_n} \right) \\
&\quad + \mathbf{E} \left(pen(f^*) \cdot 1_{A_n} \right) \\
&=: T_{5,n} + T_{6,n} + T_{7,n} + T_{8,n}.
\end{aligned}$$

In the *sixth step of the proof* we show

$$T_{8,n} \leq c_{29} \cdot (\log n)^6 \cdot n^{-2/3}. \tag{43}$$

On A_n we have

$$\begin{aligned}
pen(f^*) \cdot 1_{A_n} &= c_2 \cdot \left(\frac{\alpha_0^2}{n^{2/3}} + \frac{K_n}{n^{2/3}} \cdot \sum_{k=1}^{K_n} \alpha_k^2 \right) \leq \frac{c_2 \cdot c_{13}^2}{n^{2/3}} + \frac{c_2 \cdot K_n}{n^{2/3}} \cdot \sum_{k=1}^{K_n} \frac{c_{13}^2}{\tilde{K}_n^2} \\
&\leq c_{29} \cdot \frac{(\log n)^6}{n^{2/3}},
\end{aligned}$$

which implies (43).

In the *seventh step of the proof* we show

$$T_{7,n} \le c_{30} \cdot \frac{(\log n)^3}{K_n} + c_{31} \cdot \frac{(\log n)^8}{n^{3/2}}. \tag{44}$$

Let $\tilde{A}_n$ be the event that there exists (random)

$$\alpha_0 \in [-c_{13}, c_{13}] \quad \text{and} \quad \alpha_1, \dots, \alpha_{K_n} \in \left[-\frac{c_{13}}{\tilde{K}_n}, \frac{c_{13}}{\tilde{K}_n} \right],$$

which are independent of (X, Y), (X_1, Y_1), $\dots$, (X_n, Y_n), such that (35) and (36) hold for $\delta_n = \frac{B_n}{16 \cdot n \cdot \tilde{K}_n \cdot (\lceil \log n \rceil)^2} \ge 6 \cdot \log n$. By the Cauchy-Schwarz inequality, by (42), by (18) (which implies $\mathbf{E}Y^4 < \infty$), and by conditioning inside the expectation on the random variables $w_{k,0}^{(0)}, \dots, w_{k,d}^{(0)}$ $(k = 1, \dots, K_n)$ and $\alpha_0, \dots, \alpha_{K_n}$, which are independent of $\mathcal{D}_n$, we get

$$\begin{aligned}
T_{7,n} &\le \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right) \cdot 1_{\tilde{A}_n} \Bigg) \\
&\quad + \sqrt{\mathbf{E}\left\{ \left(\frac{1}{n} \sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right)^2 \right\}} \cdot \sqrt{\mathbf{E}\{(1_{A_n} - 1_{\tilde{A}_n})^2\}} \\
&\le \mathbf{E}\left\{ \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{\tilde{A}_n} \right\} \\
&\quad + c_{32} \cdot (\log n)^6 \cdot \sqrt{\mathbf{P}\{|Y_i| > \beta_n \text{ for some } i \in \{1, \dots, n\}\}} \\
&\le \mathbf{E}\left\{ \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{\tilde{A}_n} \right\} + c_{33} \cdot \frac{(\log n)^6}{n^{3/2}}.
\end{aligned}$$

On $\tilde{A}_n$ we have by (35)

$$\int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \le c_{34} \cdot \frac{1}{\tilde{K}_n} \le c_{35} \cdot \frac{(\log n)^3}{K_n},$$

which implies (44).

In the *eighth step of the proof* we show

$$T_{6,n} \le c_{36} \cdot \frac{1}{n}. \tag{45}$$

We have

$$\begin{aligned}
T_{6,n} &= \mathbf{E}\left(\left(\frac{1}{n} \sum_{i=1}^n (|Y_i - f^*(X_i)|^2 - \frac{1}{n} \sum_{i=1}^n |Y_i - f(X_i)|^2 \right) \cdot 1_{A_n} \right) \\
&\le \frac{1}{n} \sum_{i=1}^n \mathbf{E}\left(|f(X_i) - f^*(X_i)| \cdot |2Y_i - f(X_i) - f^*(X_i)| \cdot 1_{A_n} \right)
\end{aligned}$$

$$\leq \quad \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\Big(|f(X_i)-f^*(X_i)|\cdot(2\beta_n+c_{37}\cdot(\log n)^3)\cdot 1_{A_n}\Big).$$

On A_n we have by $|\sigma(x)-1_{[0,\infty)}(x)|\leq e^{-|x|}$ (cf., Lemma 9 a) in Braun et al. (2021)), (36) and $\delta_n\geq 6\cdot\log n$

$$\begin{aligned}
|f^*(X_i)-f(X_i)| &= |\sum_{k=0}^{K_n}\alpha_k\cdot\sigma((\beta_k^{(0)})^T\cdot X_i+\gamma_k^{(0)})-\sum_{k=0}^{K_n}\alpha_k\cdot\mathbf{1}_{[0,\infty)}((\beta_k^{(0)})^T\cdot X_i+\gamma_k^{(0)})| \\
&\leq \sum_{k=0}^{K_n}|\alpha_k|\cdot\exp\Big(-|(\beta_k^{(0)})^T\cdot X_i+\gamma_k^{(0)})|\Big) \\
&\leq \sum_{k=0}^{K_n}|\alpha_k|\cdot\exp\left(-\delta_n\right) \\
&\leq 2\cdot(\log n)^4\cdot\exp\left(-6\cdot\log n\right) \\
&= \frac{2\cdot(\log n)^4}{n^6}.
\end{aligned}$$

Hence, we have

$$T_{6,n}\leq\frac{2\cdot(\log n)^4}{n^6}\cdot(2\beta_n+c_{78}\cdot(\log n)^4)\leq c_{36}\cdot\frac{1}{n},$$

which implies (45).

In the *ninth step of the proof* we show that we have on A_n for any $t\in\{1,\ldots,t_n-1\}$

$$\begin{aligned}
&F(\alpha^{(t+1)},\beta^{(t+1)},\gamma^{(t+1)})-\frac{1}{n}\sum_{i=1}^{n}|Y_i-f^*(X_i)|^2-\mathrm{pen}(f^*) \\
&\leq\left(1-\frac{2\cdot c_2}{L_n\cdot n^{2/3}}\right)\cdot\left(F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})-\frac{1}{n}\sum_{i=1}^{n}|Y_i-f^*(X_i)|^2-\mathrm{pen}(f^*)\right) \\
&\quad+\frac{2\cdot c_2}{L_n\cdot n^{2/3}}\cdot\Big(F(\alpha^*,\beta^{(0)},\gamma^{(0)})-F(\alpha^*,\beta^{(t)},\gamma^{(t)})\Big), \qquad (46)
\end{aligned}$$

where $\alpha^*=(\alpha_k)_{k=0,\ldots,K_n}$ with α_k defined as in the fifth step of the proof. On A_n we know from the second step of the proof that Lemma 1 holds, and from the proof of Lemma 1 (cf., proof of Lemma 2 in Braun et al. (2021)) we know

$$F(\alpha^{(t+1)},\beta^{(t+1)},\gamma^{(t+1)})\leq F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})-\frac{1}{2L_n}\cdot\|\nabla_{(\alpha,\beta,\gamma)}F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})\|^2.$$

By Lemma 7 we know

$$\begin{aligned}
&\|\nabla_{(\alpha,\beta,\gamma)}F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})\|^2 \\
&=\sum_{k=0}^{K_n}\left|\frac{\partial}{\partial\alpha_k}F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})\right|^2+\sum_{k=1}^{K_n}\sum_{j=1}^{d}\left|\frac{\partial}{\partial\beta_{k,j}}F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})\right|^2
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{k=1}^{K_n}\left|\frac{\partial}{\partial\gamma_k}F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})\right|^2\\
&\geq\sum_{k=0}^{K_n}\left|\frac{\partial}{\partial\alpha_k}F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})\right|^2\\
&=\|\nabla_\alpha F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})\|^2\\
&\geq\frac{4\cdot c_2}{n^{2/3}}\cdot\left(F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})-\min_{\mathbf{a}\in\mathbb{R}^{K_n+1}}F(\mathbf{a},\beta^{(t)},\gamma^{(t)})\right)\\
&\geq\frac{4\cdot c_2}{n^{2/3}}\cdot\left(F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})-F(\alpha^*,\beta^{(t)},\gamma^{(t)})\right).
\end{aligned}
$$

Consequently we get

$$
\begin{aligned}
&F(\alpha^{(t+1)},\beta^{(t+1)},\gamma^{(t+1)})-\frac{1}{n}\sum_{i=1}^n|Y_i-f^*(X_i)|^2-\mathrm{pen}(f^*)\\
&=F(\alpha^{(t+1)},\beta^{(t+1)},\gamma^{(t+1)})-F(\alpha^*,\beta^{(0)},\gamma^{(0)})\\
&\leq F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})-\frac{1}{2L_n}\cdot\|\nabla_{(\alpha,\beta,\gamma)}F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})\|^2-F(\alpha^*,\beta^{(0)},\gamma^{(0)})\\
&\leq F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})-\frac{1}{2L_n}\cdot\frac{4\cdot c_2}{n^{2/3}}\cdot\left(F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})-F(\alpha^*,\beta^{(t)},\gamma^{(t)})\right)\\
&\quad-F(\alpha^*,\beta^{(0)},\gamma^{(0)})\\
&=\left(1-\frac{2\cdot c_2}{L_n\cdot n^{2/3}}\right)\cdot\left(F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})-F(\alpha^*,\beta^{(0)},\gamma^{(0)})\right)\\
&\quad+\frac{2\cdot c_2}{L_n\cdot n^{2/3}}\cdot\left(F(\alpha^*,\beta^{(0)},\gamma^{(0)})-F(\alpha^*,\beta^{(t)},\gamma^{(t)})\right)\\
&=\left(1-\frac{2\cdot c_2}{L_n\cdot n^{2/3}}\right)\cdot\left(F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})-\frac{1}{n}\sum_{i=1}^n|Y_i-f^*(X_i)|^2-\mathrm{pen}(f^*)\right)\\
&\quad+\frac{2\cdot c_2}{L_n\cdot n^{2/3}}\cdot\left(F(\alpha^*,\beta^{(0)},\gamma^{(0)})-F(\alpha^*,\beta^{(t)},\gamma^{(t)})\right),
\end{aligned}
$$

which implies (46).

In the *tenth step of the proof* we show that we have on A_n

$$
\begin{aligned}
&F(\alpha^*,\beta^{(0)},\gamma^{(0)})-F(\alpha^*,\beta^{(t)},\gamma^{(t)})\\
&\leq c_{38}\cdot(\beta_n+(\log n)^3)\cdot\sum_{\substack{k=1,\ldots,K_n:\\ \alpha_k\neq 0}}(\sum_{j=1}^d|\beta_{k,j}^{(t)}-\beta_{k,j}^{(0)}|+|\gamma_k^{(t)}-\gamma_k^{(0)}|).
\end{aligned}\tag{47}
$$

On A_n we have $|Y_i|\leq\beta_n$ for $i=1,\ldots,n$. From this, $X_i\in[0,1]$ *a.s.* and Lipschitz continuity of the logistic squasher we get

$$
F(\alpha^*,\beta^{(0)},\gamma^{(0)})-F(\alpha^*,\beta^{(t)},\gamma^{(t)})
$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}|Y_i - f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(X_i)|^2 - \frac{1}{n}\sum_{i=1}^{n}|Y_i - f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(X_i)|^2$$

$$\leq c_{38}\cdot(\beta_n + (\log n)^3)\cdot \sum_{\substack{k=1,\ldots,K_n:\\ \alpha_k\neq 0}} (\sum_{j=1}^{d}|\beta_{k,j}^{(t)} - \beta_{k,j}^{(0)}| + |\gamma_k^{(t)} - \gamma_k^{(0)}|),$$

which implies (47).

In the *eleventh step of the proof* we show that on A_n we have for any $t \in \{1,\ldots,t_n-1\}$

$$\sum_{\substack{k=1,\ldots,K_n:\\ \alpha_k\neq 0}} (\sum_{j=1}^{d}|\beta_{k,j}^{(t)} - \beta_{k,j}^{(0)}| + |\gamma_k^{(t)} - \gamma_k^{(0)}|) \leq c_{39}\cdot\frac{1}{n^2}. \tag{48}$$

For this we show

$$\min_{\substack{i=1,\ldots,n,k=1,\ldots,K_n:\\ \alpha_k\neq 0}} \left|(\beta_k^{(s)})^T\cdot X_i + \gamma_k^{(s)}\right| \geq \frac{\delta_n}{2} \tag{49}$$

and

$$\sum_{\substack{k=1,\ldots,K_n:\\ \alpha_k\neq 0}} (\sum_{j=1}^{d}|\beta_{k,j}^{(s)} - \beta_{k,j}^{(0)}| + |\gamma_k^{(s)} - \gamma_k^{(0)}|) \leq s\cdot c_3\cdot c_{21}\cdot(\log n)^2\cdot\frac{1}{L_n}\cdot\exp(-\delta_n/2) \tag{50}$$

for all $s \in \{0,\ldots,t_n\}$ by induction on s. Observe that (50) directly implies (48).

Trivially, (49) and (50) hold on A_n for $s=0$. Next we show that if (49) and (50) hold for some $s \in \{1,\ldots,t_n-1\}$, then (50) also holds for $s+1$. To do this, set

$$(\bar{\beta},\bar{\gamma}) = (\beta,\gamma) - \lambda_n\cdot\nabla_{(\beta,\gamma)}F((\alpha,\beta,\gamma)).$$

Using

$$|\sigma'(x)| = |\sigma(x)\cdot(1-\sigma(x))| \leq \min\{|\sigma(x)|,|1-\sigma(x)|\} \leq |\sigma(x) - 1_{[0,\infty)}(x)|$$

(where the first inequality holds due to $\sigma(x) \in [0,1]$) we can conclude from $|\sigma(x) - 1_{[0,\infty)}(x)| \leq e^{-|x|}$ (cf., Lemma 9 a) in Braun et al. (2021)) that

$$\begin{aligned}\max_{i=1,\ldots,n}\left|\sigma'\left(\beta_k^T\cdot X_i + \gamma_k\right)\right| &\leq \max_{i=1,\ldots,n}\exp\left(-\left|\beta_k^T\cdot X_i + \gamma_k\right|\right)\\ &= \exp\left(-\min_{i=1,\ldots,n}\left\{\left|\beta_k^T\cdot X_i + \gamma_k\right|\right\}\right).\end{aligned}$$

As a consequence, we get for $k \in \{1,\ldots,K_n\}$ and $j \in \{1,\ldots,d\}$ by the Cauchy-Schwarz inequality

$$\left|\frac{\partial F}{\partial \beta_{k,j}}(\alpha,\beta,\gamma)\right|$$

$$
\begin{aligned}
&= \left| \frac{2}{n} \sum_{i=1}^{n} (f_{net,(\alpha,\beta,\gamma)}(X_i) - Y_i) \cdot \alpha_k \cdot \sigma'\left(\beta_k^T \cdot X_i + \gamma_k\right) \cdot X_i^{(j)} \right| \\
&\leq 2 \cdot |\alpha_k| \cdot \frac{1}{n} \sum_{i=1}^{n} |f_{net,(\alpha,\beta,\gamma)}(X_i) - Y_i| \cdot |X_i^{(j)}| \cdot \left|\sigma'\left(\beta_k^T \cdot X_i + \gamma_k\right)\right| \\
&\leq 2 \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} |f_{net,(\alpha,\beta,\gamma)}(X_i) - Y_i|^2} \cdot |\alpha_k| \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} |\sigma'(\beta_k^T \cdot X_i + \gamma_k)|^2} \\
&\leq 2 \cdot \sqrt{F(\alpha,\beta,\gamma)} \cdot |\alpha_k| \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} |\sigma'(\beta_k^T \cdot X_i + \gamma_k)|^2} \\
&\leq 2 \cdot \sqrt{F(\alpha,\beta,\gamma)} \cdot |\alpha_k| \cdot \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \beta_k^T \cdot X_i + \gamma_k \right| \right\} \right).
\end{aligned}
$$

Hence, we have shown

$$
\begin{aligned}
&|\bar{\beta}_{k,j} - \beta_{k,j}| \\
&= \lambda_n \cdot \left| \frac{\partial F}{\partial \beta_{k,j}}((\alpha,\beta,\gamma)) \right| \\
&\leq \lambda_n \cdot 2 \cdot \sqrt{F((\alpha,\beta,\gamma))} \cdot |\alpha_k| \cdot \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \beta_k^T \cdot X_i + \gamma_k \right| \right\} \right)
\end{aligned}
$$

for any $k \in \{1, \ldots, K_n\}$.

For γ_j we get in a similar fashion

$$
\begin{aligned}
|\bar{\gamma_k} - \gamma_k| &= \lambda_n \cdot \left| \frac{\partial F}{\partial \gamma_k}(\alpha,\beta,\gamma) \right| \\
&\leq \lambda_n \cdot 2 \cdot \sqrt{F(\alpha,\beta,\gamma)} \cdot 1 \cdot |\alpha_k| \cdot \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \beta_k^T \cdot X_i + \gamma_k \right| \right\} \right)
\end{aligned}
$$

for any $k \in \{1, \ldots, K_n\}$.

Using

$$
F(\alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}) \leq F(\mathbf{w}(0)) \leq c_3^2 \cdot (\log n)^2
$$

(cf., (39)), (49) and (50) we can conclude

$$
\begin{aligned}
&\sum_{\substack{k=1,\ldots,K_n: \\ \alpha_k \neq 0}} (\sum_{j=1}^{d} |\beta_{k,j}^{(s+1)} - \beta_{k,j}^{(0)}| + |\gamma_k^{(s+1)} - \gamma_k^{(0)}|) \\
&\leq s \cdot c_3 \cdot c_{21} \cdot (\log n)^2 \cdot \frac{1}{L_n} \cdot \exp(-\delta_n/2) + \sum_{\substack{k=1,\ldots,K_n: \\ \alpha_k \neq 0}} (\sum_{j=1}^{d} |\beta_{k,j}^{(s+1)} - \beta_{k,j}^{(s)}| + |\gamma_k^{(s+1)} - \gamma_k^{(s)}|) \\
&\leq s \cdot c_3 \cdot c_{21} \cdot (\log n)^2 \cdot \frac{1}{L_n} \cdot \exp(-\delta_n/2) + c_3 \cdot c_{21} \cdot (\log n)^2 \cdot \frac{1}{L_n} \cdot \exp(-\delta_n/2)
\end{aligned}
$$

$$\leq (s+1) \cdot c_3 \cdot c_{21} \cdot (\log n)^2 \cdot \frac{1}{L_n} \cdot \exp(-\delta_n/2).$$

Furthermore, if (49) holds for some $s \in \{1, \ldots, t_n - 1\}$ and (50) holds for $s+1$, then (49) also holds for $s+1$, since

$$\begin{aligned}
&\min_{\substack{i=1,\ldots,n,k=1,\ldots,K_n: \\ \alpha_k \neq 0}} \left| \sum_{j=1}^{d} \beta_{k,j}^{(s+1)} \cdot X_i^{(j)} + \gamma_k^{(s+1)} \right| \\
&\geq \min_{\substack{i=1,\ldots,n,k=1,\ldots,K_n: \\ \alpha_k \neq 0}} \left| \sum_{j=1}^{d} \beta_{k,j}^{(0)} \cdot X_i^{(j)} + \gamma_k^{(0)} \right| \\
&\quad - \max_{\substack{i=1,\ldots,n,k=1,\ldots,K_n: \\ \alpha_k \neq 0}} \left(\sum_{j=1}^{d} |\beta_{k,j}^{(s+1)} - \beta_{k,j}^{(0)}||X_i^{(j)}| + |\gamma_k^{(s+1)} - \gamma_k^{(0)}| \right) \\
&\geq \delta_n - \max_{\substack{i=1,\ldots,n,k=1,\ldots,K_n: \\ \alpha_k \neq 0}} \left(\sum_{j=1}^{d} |\beta_{k,j}^{(s+1)} - \beta_{k,j}^{(0)}| + |\gamma_k^{(s+1)} - \gamma_k^{(0)}| \right) \\
&\geq \delta_n - c_3 \cdot c_{21} \cdot (\log n)^2 \cdot \frac{t_n}{L_n} \cdot \exp\left(-\delta_n/2\right) \\
&\geq \delta_n - c_3 \cdot c_{21} \cdot (\log n)^2 \cdot \frac{t_n}{L_n} \cdot \exp\left(-6 \cdot \log n\right) \\
&\geq \frac{\delta_n}{2}.
\end{aligned}$$

In the *twelfth step of the proof* we finish the proof by showing

$$T_{5,n} \leq c_{40} \cdot \frac{(\log n)^4}{n}. \tag{51}$$

Applying the result of the ninth step recursively together with the results of the steps ten and eleven we get that we have on A_n

$$\begin{aligned}
&T_{5,n} = F(\alpha^{(t_n)}, \beta^{(t_n)}, \gamma^{(t_n)}) - \frac{1}{n} \sum_{i=1}^{n} |Y_i - f^*(X_i)|^2 - \mathrm{pen}(f^*) \\
&\leq \left(1 - \frac{2 \cdot c_2}{L_n \cdot n^{2/3}}\right)^{t_n} \cdot \left(F(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}) - \frac{1}{n} \sum_{i=1}^{n} |Y_i - f^*(X_i)|^2 - \mathrm{pen}(f^*) \right) \\
&\quad + t_n \cdot \frac{2 \cdot c_2}{L_n \cdot n^{2/3}} \cdot c_{41} \cdot (\beta_n + (\log n)^3) \cdot c_{42} \cdot \frac{1}{n^2} \\
&\leq \exp\left(-\frac{2 \cdot c_2}{L_n \cdot n^{2/3}} \cdot t_n\right) \cdot F(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}) + c_{43} \cdot (\log n)^2 \cdot \frac{\beta_n + (\log n)^3}{n} \\
&\leq \exp(-c_{44} \cdot (\log n)^2) \cdot c_{45} \cdot (\log n)^2 + c_{46} \cdot (\log n)^2 \cdot \frac{\beta_n + (\log n)^3}{n^2} \leq c_{40} \cdot \frac{(\log n)^4}{n}.
\end{aligned}$$

Summarizing the above results we get the assertion. □

References

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

[2] Arora, S., Cohen, N., Golowich, N., and Hu, W. (2018). A convergence analysis of gradient descent for deep linear neural networks. *International Conference on Learning Representations (ICLR 2019).* New Orleans, Louisiana.

[3] Bagirov, A. M., Clausen, C., and Kohler, M. (2009). Estimation of a regression function by maxima of minima of linear functions. *IEEE Transactions on Information Theory* **55**, pp. 833-845.

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

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

[6] Bartlett, P. L., Long, P. M., and Lugosi, G. (2020). Beningn overfitting in linear regression. *Proceedings of the National Academy of Sciences*, **117**, pp. 30063-30070.

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

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

[9] Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2019). Does data interpolation contradict statistical optimality? *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1611-1619.

[10] Berner, J., Grohs, P., Kutyniok, G., and Petersen, P. (2021). The modern mathematics of deep learning. *arXiv: 2105.04026v1.*

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

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

[13] Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2015). The loss surface of multilayer networks. International Conference on Articial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. Proceeding of Machine Learning Research, volume 38, pp. 192-204.

[14] Du, S., and Lee, J. (2018). On the power of over-parametrization in neural networks with quadratic activation. *Proceedings of the 35th International Conference on Machine Learning (PMLR 2018)*, **80**, pp. 1329-1338. Stockholm, Sweden.

[15] Du, S., Lee, J., Tian, Y., Poczos, B., and Singh, A. (2018). Gradient descent learns one-hidden-layer CNN: don't be afraid of spurious local minima. *Proceedings of the 35th International Conference on Machine Learning (PMLR 2018)*, **80**, pp. 1339-1348. Stockholm, Sweden.

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

[17] Huang, W., Du, W., and Xu, Y.D. (2021). On the neural tangent kernel of deep networks with orthogonal initialization. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*. pp. 2577-2583. Montreal, Canada.

[18] Imaizumi, M., and Fukamizu, K. (2019). Deep neural networks learn non-smooth functions effectively. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*. pp. 869-878. Naha, Okinawa, Japan.

[19] Jacot, A., Gabriel, F., und Hongler, C. (2020). Neural tangent kernel: convergence and generalization in neural networks. *arXiv: 1806.07572v4.*

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

[21] Kawaguchi, K, and Huang, J. (2019). Gradient descent finds global minima for generalizable deep neural networks of practical sizes. *arXiv: 1908.02419v1.*

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

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

[24] Kohler, M., Krzyżak, A., and Langer, S. (2019). Estimation of a function of low local dimensionality by deep neural networks, *IEEE Transaction on Information Theory* (to appear).

[25] Kohler, M., and Langer, S. (2020). Discussion of "Nonparametric regression using deep neural networks with ReLU activation function". *Annals of Statistics* **48**, pp. 1906-1910.

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

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

[28] Liang, S., Sun, R., Lee, J., and Srikant, R. (2018). Adding one neuron can eliminate all bad local minima. *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018)*, pp. 4355 - 4365. Montreal, Canada.

[29] McCaffrey, D. F., and Gallant, A. R. (1994). Convergence rates for single hidden layer feedforward networks. *Neural Networks* **7**, pp. 147-158.

[30] Mondal, M. R. H., Bharati, S., and Podder, P. (2021) Diagnosis of covid-19 using machine learning and deep learning: a review. *Current Medical Imaging*, **17**, pp. 1403-1418.

[31] Ni, J., Young, T., Pandelea, V., Xue, F., Adiga, V., and Cambria, E. (2021). Recent advances in deep learning based dialogue systems: a systematic survey. *ArXiv: 2105.04387v4*.

[32] Nitanda, A., and Suzuki, T. (2017). Stochastic particle gradient descent for infinite ensembles. *arXiv:1712.05438*.

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

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

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

[36] Wang, H. and Lin, W. (2021). Harmless overparametrization in two-layer neural networks. *arXiv: 2106.04795v1*.

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

[38] Wyner, J. A., Olson, M., Bleich, J., and Mease, D. (2017) Explaining the success of AdaBost and random forest as interpolating classifiers. *The Journal of Machine Learning Research*, **18**, pp. 1558-1590.

A. Proof of Lemma 4

In the proof we use the following error decomposition:

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

$$
\begin{aligned}
&\quad -2 \cdot \left(\frac{1}{n} \sum_{i=1}^{n} |\tilde{m}_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| \le \beta_n \, (j \in \{1,\dots,n\})\}} - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \Bigg) \cdot 1_{A_n} \Bigg) \\
&= \Big[\mathbf{E}\Big\{ |m_n(X) - Y|^2 | \mathcal{D}_n \Big\} - \mathbf{E}\Big\{ |m(X) - Y|^2 \Big\} \\
&\quad - \Big(\mathbf{E}\Big\{ |m_n(X) - T_{\beta_n} Y|^2 | \mathcal{D}_n \Big\} - \mathbf{E}\Big\{ |m_{\beta_n}(X) - T_{\beta_n} Y|^2 \Big\} \Big) \Big] \cdot 1_{A_n} \\
&+ \Bigg[\mathbf{E}\Big\{ |m_n(X) - T_{\beta_n} Y|^2 | \mathcal{D}_n \Big\} - \mathbf{E}\Big\{ |m_{\beta_n}(X) - T_{\beta_n} Y|^2 \Big\} \\
&\quad -2 \cdot \frac{1}{n} \sum_{i=1}^{n} \Big(|m_n(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \Big) \Bigg] \cdot 1_{A_n} \\
&+ \Bigg[2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - T_{\beta_n} Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \\
&\quad - \left(2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \Bigg] \cdot 1_{A_n} \\
&+ \Bigg[2 \left(\frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \\
&\quad -2 \cdot \left(\frac{1}{n} \sum_{i=1}^{n} |\tilde{m}_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| \le \beta_n \, (j \in \{1,\dots,n\})\}} - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \Bigg] \cdot 1_{A_n} \\
&= \sum_{i=1}^{4} T_{i,n} \cdot 1_{A_n},
\end{aligned}
$$

where $T_{\beta_n} Y$ is the truncated version of Y and m_{β_n} is the regression function of $T_{\beta_n} Y$, i.e.,

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

We start with bounding $T_{1,n} \cdot 1_{A_n}$. By using $a^2 - b^2 = (a-b)(a+b)$ we get

$$
\begin{aligned}
T_{1,n} &= \mathbf{E}\Big\{ |m_n(X) - Y|^2 - |m_n(X) - T_{\beta_n} Y|^2 \Big| \mathcal{D}_n \Big\} \\
&\quad - \mathbf{E}\Big\{ |m(X) - Y|^2 - |m_{\beta_n}(X) - T_{\beta_n} Y|^2 \Big\} \\
&= \mathbf{E}\Big\{ (T_{\beta_n} Y - Y)(2 m_n(X) - Y - T_{\beta_n} Y) \Big| \mathcal{D}_n \Big\} \\
&\quad - \mathbf{E}\Big\{ \Big((m(X) - m_{\beta_n}(X)) + (T_{\beta_n} Y - Y) \Big) \Big(m(X) + m_{\beta_n}(X) - Y - T_{\beta_n} Y \Big) \Big\} \\
&= T_{5,n} + T_{6,n}.
\end{aligned}
$$

With the Cauchy-Schwarz inequality and

$$
I_{\{|Y| > \beta_n\}} \le \frac{\exp(c_4/2 \cdot |Y|^2)}{\exp(c_4/2 \cdot \beta_n^2)} \tag{52}
$$

we conclude

$$\begin{aligned}
|T_{5,n}| &\leq \sqrt{\mathbf{E}\big\{|T_{\beta_n}Y - Y|^2\big\}} \cdot \sqrt{\mathbf{E}\big\{|2m_n(X) - Y - T_{\beta_n}Y|^2 \big| \mathcal{D}_n\big\}} \\
&\leq \sqrt{\mathbf{E}\big\{|Y|^2 \cdot I_{\{|Y|>\beta_n\}}\big\}} \cdot \sqrt{\mathbf{E}\big\{2 \cdot |2m_n(X) - T_{\beta_n}Y|^2 + 2 \cdot |Y|^2 \big| \mathcal{D}_n\big\}} \\
&\leq \sqrt{\mathbf{E}\left\{|Y|^2 \cdot \frac{\exp(c_4/2 \cdot |Y|^2)}{\exp(c_4/2 \cdot \beta_n^2)}\right\}} \\
&\qquad \cdot \sqrt{\mathbf{E}\big\{2 \cdot |2m_n(X) - T_{\beta_n}Y|^2 \big| \mathcal{D}_n\big\} + 2\mathbf{E}\big\{|Y|^2\big\}} \\
&\leq \sqrt{\mathbf{E}\Big\{|Y|^2 \cdot \exp(c_4/2 \cdot |Y|^2)\Big\}} \cdot \exp\left(-\frac{c_4 \cdot \beta_n^2}{4}\right) \cdot \sqrt{2(3\beta_n)^2 + 2\mathbf{E}\big\{|Y|^2\big\}}.
\end{aligned}$$

With $x \leq \exp(x)$ for $x \in \mathbb{R}$ we get

$$|Y|^2 \leq \frac{2}{c_4} \cdot \exp\left(\frac{c_4}{2} \cdot |Y|^2\right) \tag{53}$$

and hence $\sqrt{\mathbf{E}\Big\{|Y|^2 \cdot \exp(c_4/2 \cdot |Y|^2)\Big\}}$ is bounded by

$$\mathbf{E}\left(\frac{2}{c_4} \cdot \exp\left(c_4/2 \cdot |Y|^2\right) \cdot \exp(c_4/2 \cdot |Y|^2)\right) \leq \mathbf{E}\left(\frac{2}{c_4} \cdot \exp\left(c_4 \cdot |Y|^2\right)\right) \leq c_{47}$$

which is less than infinity by the assumptions of the lemma. Furthermore the third term is bounded by $\sqrt{18\beta_n^2 + c_{48}}$ because

$$\mathbf{E}(|Y|^2) \leq \mathbf{E}(1/c_4 \cdot \exp(c_4 \cdot |Y|^2) \leq c_{49} < \infty, \tag{54}$$

which follows again as above. With the setting $\beta_n = c_3 \cdot \log(n)$ it follows for some constants $c_{50}, c_{51} > 0$ that

$$|T_{5,n}| \leq \sqrt{c_{47}} \cdot \exp\left(-c_{50} \cdot (\log n)^2\right) \cdot \sqrt{(18 \cdot c_3^2 \cdot (\log n)^2 + c_{49})} \leq c_{51} \cdot \frac{\log(n)}{n}.$$

From the Cauchy-Schwarz inequality we get

$$\begin{aligned}
|T_{6,n}| \leq & \sqrt{2 \cdot \mathbf{E}\Big\{|(m(X) - m_{\beta_n}(X))|^2\Big\} + 2 \cdot \mathbf{E}\Big\{|(T_{\beta_n}Y - Y)|^2\Big\}} \\
& \cdot \sqrt{\mathbf{E}\left\{\Big|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n}Y\Big|^2\right\}},
\end{aligned}$$

where we can bound the second factor on the right-hand side in the above inequality in the same way we have bounded the second factor from $T_{5,n}$, because by assumption

$||m||_\infty$ is finite and furthermore m_{β_n} is bounded by β_n. Thus we get for some constant $c_{52} > 0$

$$\sqrt{\mathbf{E}\left\{\left|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n}Y\right|^2\right\}} \leq c_{52} \cdot \log(n).$$

Next we consider the first term. With Jensen's inequality it follows that

$$\mathbf{E}\left\{|m(X) - m_{\beta_n}(X)|^2\right\} \quad \leq \quad \mathbf{E}\left\{\mathbf{E}\left(|Y - T_{\beta_n}Y|^2 \middle| X\right)\right\} = \mathbf{E}\left\{|Y - T_{\beta_n}Y|^2\right\}.$$

Hence we get

$$|T_{6,n}| \quad \leq \quad \sqrt{4 \cdot \mathbf{E}\left\{|Y - T_{\beta_n}Y|^2\right\}} \cdot c_{52} \cdot \log(n)$$

and therefore with the calculations from $T_{5,n}$ it follows that $T_{6,n} \leq c_{53} \cdot \log(n)/n$ for some constant $c_{53} > 0$. Altogether we get

$$T_{1,n} \cdot 1_{A_n} \leq |T_{5,n}| + |T_{6,n}| \leq c_{54} \cdot \frac{\log(n)}{n}$$

for some constant $c_{54} > 0$.
Next we consider $T_{2,n} \cdot 1_{A_n}$ and conclude for $t > 0$

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

Application of Theorem 19.3 in Györfi et al. (2002) (with $K_1 = 4$ and $K_2 = 16$) and the relation $\mathcal{N}_2\left(\delta, \left\{\frac{1}{\beta_n}g : g \in \mathcal{G}\right\}, x_1^n\right) \leq \mathcal{N}_2\left(\delta \cdot \beta_n, \mathcal{G}, x_1^n\right)$ for an arbitrary function space $\mathcal{G}$ and $\delta > 0$ imply that we have for any $t > \delta_n$

$$\mathbf{P}\{T_{2,n} \cdot 1_{A_n} > t\} \leq 60 \cdot \exp\left(-\frac{n}{10^8 \cdot \beta_n^2} \cdot t\right).$$

This implies

$$\begin{aligned}&\mathbf{E}(T_{2,n} \cdot 1_{A_n}) \\ &\leq \delta_n + \int_{\delta_n}^{\infty} \mathbf{P}\{T_{2,n} \cdot 1_{A_n} > t\} dt \\ &\leq \delta_n + 60 \cdot \exp\left(-\frac{n}{10^8 \cdot \beta_n^2} \cdot \delta_n\right) \cdot \frac{10^8 \cdot \beta_n^2}{n} \leq c_{55} \cdot \delta_n.\end{aligned}$$

By bounding $T_{3,n} \cdot 1_{A_n}$ similarly to $T_{1,n} \cdot 1_{A_n}$ we get

$$\mathbf{E}(T_{3,n} \cdot 1_{A_n}) \quad \leq \quad c_{56} \cdot \frac{\log(n)}{n}$$

for some large enough constant $c_{56} > 0$ and hence we get in total

$$\mathbf{E}\left(\sum_{i=1}^{3} T_{i,n} \cdot 1_{A_n}\right) \quad \leq \quad c_{57} \cdot \left(\delta_n + \frac{(\log n)^2}{n}\right)$$

for some sufficient large constant $c_{57} > 0$.

We finish the proof by bounding $T_{4,n} \cdot 1_{A_n}$. We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| \leq \beta_n \, (j \in \{1,\ldots,n\})\}} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| > \beta_n \text{ for some } j \in \{1,\ldots,n\}\}} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} |\tilde{m}_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| \leq \beta_n \, (j \in \{1,\ldots,n\})\}} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| > \beta_n \text{ for some } j \in \{1,\ldots,n\}\}}
\end{aligned}$$

where we have used that $|T_\beta z - y| \leq |z - y|$ holds for $|y| \leq \beta$. Consequently we have

$$\begin{aligned}
&\mathbf{E}\{T_{4,n} \cdot 1_{A_n}\} \\
&\leq \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| > \beta_n \text{ for some } j \in \{1,\ldots,n\}\}}\right\} \\
&= \mathbf{E}\left\{|m_n(X_1) - Y_1|^2 \cdot 1_{\{|Y_j| > \beta_n \text{ for some } j \in \{1,\ldots,n\}\}}\right\} \\
&\leq \sqrt{\mathbf{E}\left\{|m_n(X_1) - Y_1|^4\right\}} \cdot \sqrt{\mathbf{P}\{|Y_j| > \beta_n \text{ for some } j \in \{1,\ldots,n\}\}} \\
&\leq c_{58} \cdot \beta_n^2 \cdot \frac{1}{n},
\end{aligned}$$

where the last inequality followed from the proof of Theorem 1 (cf., (42)) and (18) and (53), which imply

$$\begin{aligned}
\mathbf{E}\left\{|m_n(X_1) - Y_1|^4\right\} \quad &\leq \quad 16 \cdot (\beta_n^4 + \mathbf{E}\{Y^4\})) \leq 16 \cdot \left(\beta_n^4 + \frac{4}{c_4^2} \cdot \mathbf{E}\left\{\exp\left(c_4 \cdot Y^2\right)\right\}\right) \\
&\leq \quad c_{59} \cdot \beta_n^4.
\end{aligned}$$

In combination with the other considerations in the proof this implies the assertion of Lemma 4. □