Annals of the Institute of Statistical Mathematics (2024) 76:361–391
https://doi.org/10.1007/s10463-024-00898-6

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

Selina Drews[1] · **Michael Kohler**[1]

Received: 13 October 2022 / Revised: 11 September 2023 / Accepted: 28 December 2023 /
Published online: 8 April 2024

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

Keywords Neural networks · Nonparametric regression · Over-parametrization · Universal consistency

1 Introduction

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

✉ Selina Drews
drews@mathematik.tu-darmstadt.de

Michael Kohler
kohler@mathematik.tu-darmstadt.de

[1] Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany

In the last few years, various results concerning the approximation power of deep neural networks [cf., e.g., (Yarotsky 2017; Yarotsky and Zhevnerchuk 2020; Lu et al. 2020; Langer 2021b) and the literature cited therein] or concerning the statistical risk of corresponding least squares estimates [cf., e.g., (Bauer and Kohler 2019; Kohler and Krzyżak 2017; Schmidt-Hieber 2020; Kohler and Langer 2021; Langer 2021a; Imaizumi and Fukamizu 2019; Suzuki 2018; Suzuki and Nitanda 2019) and the literature cited therein] have been derived.

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

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

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

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

One approach beyond the linear regime considered in Bartlett et al. (2021) is the mean field limit. Here, the weights move in a nontrivial way during training, even

though the network is infinitely wide. Using mean field theory, global convergence results can be proved for two-layer (Mei et al. 2018; Chizat and Bach 2018) and multi-layer neural networks (Nguyen and Pham 2020).

Another approach is to consider the linearized evolution as a Taylor expansion of the first order and then construct higher-order approximations. Other approaches are considered in Dyer and Gur-Ari (2019), Hanin and Nica (2019) and Yang and Hu (2020).

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

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

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

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

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

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

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

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

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

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

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

For many years, it is well-known that universally consistent regression estimates exist, see Stone (1977) for the first result in this respect and Györfi et al. (2002) for an extensive overview of such results. So, it is not surprising that deep neural networks have this property, too. However, our main result presents

interesting aspects of the application of gradient descent to deep neural networks which are useful for proving such universal consistency: Firstly, the over-parametrization is useful in our result since it ensures that a finite subset of the initially chosen inner weights has nice properties. Secondly, due to the fact that we use a small stepsize, gradient descent applied to a properly regularized empirical L_2 risk is able to adjust the outer weights in an optimal way. And thirdly, due to the fact that the number of gradient descent steps is only slightly larger than the reciprocal of the stepsize, the inner weights do not change drastically during our learning. Altogether this enables our estimate to perform a kind of representation guessing instead of representation learning.

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

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

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

is its supremum norm.

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

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

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

If A is a subset of $\mathbb{R}^d$ and $x \in \mathbb{R}^d$, then we set $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ otherwise. For $z \in \mathbb{R}$ and $\beta > 0$ we define $T_\beta z = \max\{-\beta, \min\{\beta, z\}\}$. If $f : \mathbb{R}^d \to \mathbb{R}$ is a function and $\mathcal{F}$ is a set of such functions, then we set $(T_\beta f)(x) = T_\beta(f(x))$ and

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

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

2 Definition of the estimate

Let $\sigma(x) = 1/(1 + e^{-x})$ be the logistic squasher. In the sequel, we use a network topology where we compute a linear combination of K_n fully connected neural networks with L layers and r neurons per layer, i.e., we define our neural network as follows: We set

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

The estimate is defined by

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

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

Because of (5), we have

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

3 Main result

Our main result is the following theorem.

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

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

for some $\kappa > 0$,

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

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

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

for some $L_n > 0$ which satisfies

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

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

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

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

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

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

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

4 Proofs

4.1 Auxiliary results

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

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

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

and

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

Assume

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

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

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

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

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

Then, we have

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

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

and

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

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

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

$$\sum_{j=1}^{K_n} |w_{1,1,j}^{(L)}| \leq C, \tag{12}$$

$$|w_{k,i,j}^{(l)}| \leq B \quad (k \in \{1, \dots, K_n\}, i, j \in \{1, \dots, r\}, l \in \{1, \dots, L-1\}) \tag{13}$$

and

$$|w_{k,i,j}^{(0)}| \leq A \quad (k \in \{1, \dots, K_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d\}). \tag{14}$$

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

$$\mathcal{N}_p\left(\epsilon, \{T_\beta f \cdot 1_{[-\alpha,\alpha]^d} \ : \ f \in \mathcal{F}\}, x_1^n\right) \leq \left(c_5 \cdot \frac{\beta^p}{\epsilon^p}\right)^{c_6 \cdot \alpha^d \cdot B^{(L-1)\cdot d} \cdot A^d \cdot \left(\frac{C}{\epsilon}\right)^{d/k} + c_7}.$$

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

$$\left| \frac{\partial^k f_{\mathbf{w}}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \right| \leq c_8 \cdot C \cdot B^{(L-1)\cdot k} \cdot A^k =: c. \tag{15}$$

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

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

hence (15) is implied by

$$\left| \frac{\partial^k f_{j,1}^{(L)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \right| \leq c_9 \cdot B^{(L-1)\cdot k} \cdot A^k. \tag{16}$$

We have

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

and

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

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

$$\frac{\partial^k f_{k,i}^{(l)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \tag{17}$$

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

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

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

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

is a given by

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

Because of the boundedness of the derivatives of σ we can conclude from (14)

$$\left|\frac{\partial^k f_{k,i}^{(1)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x)\right| \leq c_{10} \cdot A^k$$

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

Recursively we can conclude from the above representation of (17) that we have

$$\left|\frac{\partial^k f_{k,i}^{(l)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x)\right| \leq c_{11} \cdot B^{(l-1)\cdot k} \cdot A^k.$$

Setting $l = L$ we get (16).

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

$$\mathcal{N}_p\left(\epsilon, \{T_\beta f \cdot 1_{[-\alpha,\alpha]^d} \ : \ f \in \mathcal{F}\}, x_1^n\right) \leq \mathcal{N}_p\left(\frac{\epsilon}{2}, T_\beta \mathcal{G} \circ \Pi, x_1^n\right), \tag{18}$$

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

$$\left(c_{12} \cdot \frac{\epsilon}{c}\right)^{1/k}$$

where $c_{12} = c_{12}(d, k)$ is a suitable small constant greater than zero and the additional set $\mathbb{R}^d \backslash [-\alpha, \alpha]^d$.

A standard bound on the remainder of a multivariate Taylor polynomial together with (15) shows that for each $f_{\mathbf{w}}$ we can find $g \in \mathcal{G} \circ \Pi$ such that

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

holds for all $x \in [-\alpha, \alpha]^d$, which implies (18).

In the *third step* of the proof we show the assertion of Lemma 2. Since $\mathcal{G} \circ \Pi$ is a linear vector space of dimension less than or equal to

$$c_{13} \cdot \alpha^d \cdot \left(\frac{c}{\epsilon}\right)^{d/k}$$

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

$$\mathcal{N}_p\left(\frac{\epsilon}{2}, T_\beta \mathcal{G} \circ \Pi, x_1^n\right) \leq 3\left(\frac{2e(2\beta)^p}{(\epsilon/2)^p} \log\left(\frac{3e(2\beta)^p}{(\epsilon/2)^p}\right)\right)^{c_{13}\cdot\alpha^d\cdot\left(\frac{c}{\epsilon}\right)^{d/k}+1}.$$

Together with (18), this implies the assertion. □

Lemma 3 *Let σ be the logistic squasher and let $0 < \delta \leq 1, 1 \leq \alpha_n \leq \log n$, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ with*

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

and $x \in [-\alpha_n, \alpha_n]^d$. Let $L, r, n \in \mathbb{N}$ with $L \geq 2$, $r \geq 2d$, $n \geq 8d$ and $n \geq \exp(r+1)$. Consider $f_{1,1}^{(L)}(x)$ where $f_{k,j}^{(l)}(x)$ are recursively defined by (3) and (4).

Assume

$$w_{1,j,j}^{(0)} = \frac{4d(\log n)^2}{\delta} \quad \text{for } j \in \{1, \ldots, d\}, \tag{19}$$

$$w_{1,j,0}^{(0)} = -\frac{4d(\log n)^2}{\delta} \cdot u^{(j)} \quad \text{for } j \in \{1, \ldots, d\}, \tag{20}$$

$$w_{1,j+d,j}^{(0)} = -\frac{4d(\log n)^2}{\delta} \quad \text{for } j \in \{1, \ldots, d\}, \tag{21}$$

$$w_{1,j+d,0}^{(0)} = \frac{4d(\log n)^2}{\delta} \cdot v^{(j)} \quad \text{for } j \in \{1, \ldots, d\}, \tag{22}$$

$$w_{1,s,t}^{(0)} = 0 \quad \text{if } s \leq 2d, s \neq t, s \neq t+d \quad \text{and} \quad t > 0, \tag{23}$$

$$w_{1,1,t}^{(1)} = 8 \cdot (\log n)^2 \quad \text{for } t \in \{1, \ldots, 2d\}, \tag{24}$$

$$w_{1,1,0}^{(1)} = -8 \cdot (\log n)^2 \Big(2d - \frac{1}{2}\Big), \tag{25}$$

$$w_{1,1,t}^{(1)} = 0 \quad \text{for } t > 2d, \tag{26}$$

$$w_{1,1,1}^{(l)} = 6 \cdot (\log n)^2 \quad \text{for } l \in \{2, \ldots, L-1\}, \tag{27}$$

$$w_{1,1,0}^{(l)} = -3 \cdot (\log n)^2 \quad \text{for } l \in \{2, \ldots, L-1\} \tag{28}$$

and

$$w_{1,1,t}^{(l)} = 0 \quad \text{for } t > 1 \quad \text{and} \quad l \in \{2, \ldots, L-1\}. \tag{29}$$

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

$$|\bar{w}_{1,i,j}^{(l)} - w_{1,i,j}^{(l)}| \leq \log n \quad \text{for all } l = 0, \dots, L-1. \tag{30}$$

We denote by $\bar{f}_{1,1}^{(L)}(x)$ *the neural network corresponding to* $\bar{\mathbf{w}}$ *where*

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

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

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

Then, we have

$$\bar{f}_{1,1}^{(L)}(x) \geq 1 - \frac{1}{n} \quad \text{if } x \in [u^{(1)} + \delta, v^{(1)} - \delta] \times \cdots \times [u^{(d)} + \delta, v^{(d)} - \delta]$$

and

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

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

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

Let $x \in [u^{(1)} + \delta, v^{(1)} - \delta] \times \cdots \times [u^{(d)} + \delta, v^{(d)} - \delta]$. Then, we get for any $i \in \{1, \dots, d\}$ by (19), (20), (23) and (30)

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

and for any $i \in \{d+1, \dots, 2d\}$ by (21)–(23) and (30)

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

This implies

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

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

Using (24)–(26), (30) and $|\sigma(u)| \leq 1$ for any $u \in \mathbb{R}$, we get similarly as above

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

Therefore, we obtain

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

With the same argument as above and with (27)-(30), we can recursively conclude for $l = 3, \dots, L$ that we have

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

Therefore, we obtain

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

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

This implies

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

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

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

Let $i \in \{1, \dots, d\}$ and assume $x^{(i)} \notin [u^{(i)} - \delta, v^{(i)} + \delta]$. Then, we can argue similarly as above. In case $x^{(i)} < u^{(i)} - \delta$ for some $i \in \{1, \dots, d\}$ we obtain by (19), (20), (23) and (30)

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

and in case $x^{(i)} > v^{(i)} + \delta$ for some $i \in \{1, \dots, d\}$ we obtain by (21)–(23) and (30)

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

Together with the logistic squasher it holds

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

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

Similar to above we get with (24)–(26) and (30)

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

From this, we obtain

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

Using (27)–(30), we can argue similarly as above and conclude recursively for $l = 3, \ldots, L$

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

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

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

holds. This yields the assertion. □

Lemma 4 *Let $0 < \delta \leq 1$, $1 \leq \alpha_n \leq \log n$ and let σ be the logistic squasher, let $m : \mathbb{R}^d \to \mathbb{R}$ be Lipschitz continuous with Lipschitz constant C_{Lip}, let $L, r, n \in \mathbb{N}$ with $L \geq 2$, $r \geq 2d$, $n \geq 8d$, $n \geq \exp(r+1)$ and let $K \in \mathbb{N}$ with $K^d \leq K_n$. Furthermore, define $f_{\bar{\mathbf{w}}}$ by*

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

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

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

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

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

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

$$w_{k,j,j}^{(0)} = \frac{4d(\log n)^2}{\delta} \quad \text{for } j \in \{1, \dots, d\}, \tag{31}$$

$$w_{k,j,0}^{(0)} = \frac{-4d(\log n)^2}{\delta} \cdot u_k^{(j)} \quad \text{for } j \in \{1, \dots, d\}, \tag{32}$$

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

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

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

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

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

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

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

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

and

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

for all $k \in \{1, \dots, K^d\}$. *Furthermore, set*

$$w_{k,s,t}^{(l)} = 0 \quad \text{for } k \notin \{1, \dots, K^d\}.$$

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

$$\alpha_1, \dots, \alpha_{K^d} \in [-\|m\|_\infty, \|m\|_\infty] \tag{42}$$

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

$$|f_{\bar{\mathbf{w}}}(x) - m(x)| \leq c_{14} \cdot \left(C_{Lip} \cdot \frac{\Delta}{K} + K^d \cdot \frac{1}{n}\right) \tag{43}$$

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

$$\bigcup_{j \in \{0,1,\dots,K\}} \bigcup_{i \in \{1,\dots,d\}} \left\{ x \in \mathbb{R}^d : \left| x^{(i)} - \left(a_i + j \cdot \frac{b_i - a_i}{K}\right) \right| < \delta \right\} \tag{44}$$

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

$$\bar{w}_{1,1,j_k}^{(L)} = \alpha_k \quad (k \in \{1, \dots, K^d\}), \quad \bar{w}_{1,1,k}^{(L)} = 0 \quad (k \notin \{j_1, \dots, j_{K^d}\}) \tag{45}$$

and

$$|w_{s,k,i}^{(l)} - \bar{w}_{j_s,k,i}^{(l)}| \leq \log n \quad \text{for all } l \in \{0, \dots, L-1\}, \quad s \in \{1, \dots, K^d\}. \tag{46}$$

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

$$|f_{\bar{\mathbf{w}}}(x)| \leq \|m\|_\infty \cdot \left(3^d + \frac{K^d}{n}\right). \tag{47}$$

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

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

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

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

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

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

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

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

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

Now, because of (45) and the definitions of S and $f_{\bar{\mathbf{w}}}(x)$, we have

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

Application of Lemma 3 yields

$$|S(x) - f_{\bar{\mathbf{w}}}(x)| \leq c_{15} \cdot K^d \cdot \frac{1}{n}$$

for all $x \in [a_1, b_1) \times \cdots \times [a_d, b_d)$ which are not contained in (44).

From this, we obtain

$$\big|f_{\bar{\mathbf{w}}}(x) - m(x)\big| = \big|f_{\bar{\mathbf{w}}}(x) - S(x)\big| + \big|S(x) - m(x)\big| \leq c_{16} \Big(K^d \cdot \frac{1}{n} + C_{Lip} \cdot \frac{\Delta}{K} \Big)$$

for all $x \in [a_1, b_1) \times \cdots \times [a_d, b_d)$ which are not contained in (44). This implies (43).

To show inequality (47), we assume that $x \in \mathbb{R}^d$. Then for $\delta \leq \frac{\Delta}{K}$ and every fixed x, we get

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

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

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

By Lemma 3, we know that $\bar{f}^{(L)}_{j_k,1}(x) \le \frac{1}{n}$ for $k \in \{1, \dots, K^d\}$ if $x \notin [u_k^{(1)} - \delta, v_k^{(1)} + \delta] \times \cdots \times [u_k^{(d)} - \delta, v_k^{(d)} + \delta]$. Thus, we get together with the definition of $w_{1,1,k}$ for $k \in \{1, \dots, K^d\}$

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

This results in

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

for $x \in \mathbb{R}^d$. □

Lemma 5 *Let σ be the logistic squasher, let $1 \le \alpha_n \le \log n$, let $m : \mathbb{R}^d \to \mathbb{R}$ be Lipschitz continuous as well as bounded, let $L, r, n \in \mathbb{N}$ with $L \ge 2$, $r \ge 2d$, $n \ge 8d$ and $n \ge \exp(r + 1)$ and let $K \in \mathbb{N}$ with $2 \le K \le \alpha_n - 1$ and $(K^2 + 1)^{3d} \le K_n$. Given $u_1, v_1, \dots, u_{(K^2+1)^{3d}}, v_{(K^2+1)^{3d}} \in [-K - \frac{2}{K}, K]^d$, choose $\mathbf{w}$ such that*

$$
w^{(0)}_{k,j,j} = 4d \cdot K^2 \cdot (\log n)^2 \quad \text{for } j \in \{1, \dots, d\}, \tag{48}
$$

$$
w^{(0)}_{k,j,0} = -4d \cdot K^2 \cdot (\log n)^2 \cdot u_k^{(j)} \quad \text{for } j \in \{1, \dots, d\}, \tag{49}
$$

$$
w^{(0)}_{k,j+d,j} = -4d \cdot K^2 \cdot (\log n)^2 \quad \text{for } j \in \{1, \dots, d\}, \tag{50}
$$

$$
w^{(0)}_{k,j+d,0} = 4d \cdot K^2 \cdot (\log n)^2 \cdot v_k^{(j)} \quad \text{for } j \in \{1, \dots, d\}, \tag{51}
$$

$$
w^{(0)}_{k,s,t} = 0 \quad \text{if } s \le 2d, s \ne t, s \ne t + d \quad \text{and} \quad t > 0, \tag{52}
$$

$$w_{k,1,t}^{(1)} = 8 \cdot (\log n)^2 \quad \text{for } t \in \{1, \dots, 2d\}), \tag{53}$$

$$w_{k,1,0}^{(1)} = -8 \cdot (\log n)^2 \left(2d - \frac{1}{n}\right) \tag{54}$$

$$w_{k,1,t}^{(1)} = 0 \quad \text{for } t > 2d, \tag{55}$$

$$w_{k,1,1}^{(l)} = 6 \cdot (\log n)^2 \quad \text{for } l \in \{2, \dots, L-1\}, \tag{56}$$

$$w_{k,1,0}^{(l)} = -3 \cdot (\log n)^2 \quad \text{for } l \in \{2, \dots, L-1\} \tag{57}$$

and

$$w_{k,1,t}^{(l)} = 0 \quad \text{for } t > 1 \quad \text{and} \quad l \in \{2, \dots, L-1\} \tag{58}$$

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

$$w_{k,s,t}^{(l)} = 0 \quad \text{for } k \notin \{1, \dots, (K^2+1)^{3d}\}.$$

Then there exists

$$\alpha_1, \dots, \alpha_{(K^2+1)^{3d}} \in \left[-\frac{\|m\|_\infty}{(K^2+1)^{2d}}, \frac{\|m\|_\infty}{(K^2+1)^{2d}} \right] \tag{59}$$

such that for all pairwise distinct $j_1, \dots, j_{(K^2+1)^{3d}} \in \{1, \dots, K_n\}$

$$\begin{aligned} &\int \left| f_{\bar{\mathbf{w}}}(x) - m(x) \right|^2 \mathbf{P}_X(dx) \\ &\leq c_{17} \cdot \left(\frac{1}{K} + \frac{K^{12d}}{n^2} + \left(\frac{K^{6d}}{n} + 1 \right)^2 \cdot \mathbf{P}_X(\mathbb{R}^d \setminus [-K, K]^d) \right) \end{aligned} \tag{60}$$

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

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

and

$$|w_{s,k,i}^{(l)} - \bar{w}_{j_s,k,i}^{(l)}| \leq \log n \quad \text{for all } l \in \{0, \dots, L-1\}, \quad s \in \{1, \dots, (K^2+1)^{3d}\}. \tag{61}$$

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

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

Let C_{Lip} be the Lipschitz constant of m. Then by Lemma 4 applied to $m/(K^2+1)^{2d}$ and $\delta = \frac{1}{K^2}$, we know that

$$\left| f_{\bar{\mathbf{w}}}(x) - \frac{1}{(K^2+1)^{2d}} \cdot m(x) \right| \leq c_{14} \cdot \left(\frac{C_{Lip}}{(K^2+1)^{2d}} \cdot \frac{2}{K} + (K^2+1)^d \cdot \frac{1}{n} \right)$$

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

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

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

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

which satisfies

$$\left| f_{\bar{\mathbf{w}}}(x) - m(x) \right| \leq c_{18} \cdot \left(\frac{1}{K} + (K^2+1)^{3d} \cdot \frac{1}{n} \right) \tag{62}$$

outside of A.

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

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

for fixed $j \in \{1, \ldots, d\}$. In this way, we can construct K different versions of $f_{\bar{\mathbf{w}}}$ that still satisfy (62) for all $x \in [-K, K]^d$ up to corresponding versions of A.

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

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

holds.

Now, we have found a shifted version of the grid such that the set A has a measure less than or equal to $\frac{d}{K}$. By (62) we know that $|f_{\bar{\mathbf{w}}}(x) - m(x)| \leq c_{19} \cdot \left(\frac{1}{K} + \frac{K^{6d}}{n} \right)$ holds for $x \in [-K, K]^d \setminus A$. From this together with the second assertion from Lemma 4, we obtain

$$
\begin{aligned}
&\int |f_{\bar{w}}(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&= \int_{[-K,K]^d \setminus A} |f_{\bar{w}}(x) - m(x)|^2 \mathbf{P}_X(dx) + \int_A |f_{\bar{w}}(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\quad + \int_{\mathbb{R}^d \setminus [-K,K]^d} |f_{\bar{w}}(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq c_{19}^2 \left(\frac{1}{K} + \frac{K^{6d}}{n} \right)^2 + c_{20} \left(3^d + \frac{K^{6d}}{n} \right)^2 \cdot \frac{d}{K} \\
&\quad + c_{21} \left(3^d + \frac{K^{6d}}{n} \right)^2 \cdot \mathbf{P}_X(\mathbb{R}^d \setminus [-K,K]^d),
\end{aligned}
$$

which implies the assertion. □

4.2 Proof of Theorem 1

Let $\epsilon > 0$ and $K \in \mathbb{N}$ be arbitrary. W.l.o.g., we assume $K_n \geq (K^2+1)^{3d}$. By Jensen's inequality for conditional expectation, we obtain

$$
\mathbf{E}\{|m(X)|^2\} = \mathbf{E}\{|\mathbf{E}\{Y|X\}|^2\} \leq \mathbf{E}\{\mathbf{E}\{|Y|^2|X\}\} = \mathbf{E}\{Y^2\} < \infty.
$$

Choose a Lipschitz continuous and bounded function $\bar{m} : \mathbb{R}^d \to \mathbb{R}$ such that

$$
\int |\bar{m}(x) - m(x)|^2 \mathbf{P}_X(dx) \leq \epsilon. \tag{63}
$$

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

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

for some weight vector $\mathbf{w}$ which satisfies the conditions (48)–(58) of Lemma 5 for $\bar{m}$ and some $j_1,\dots,j_{(K^2+1)^{3d}} \in \{1,\dots,K_n\}$, and that secondly

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

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

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

and

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

and

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

where α_k is chosen as in Lemma 5 in case that A_n holds and where $\alpha_1 = ... = \alpha_{(K^2+1)^{3d}} = 0$ in case that A_n does not hold.

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

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

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

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

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

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

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

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

and

$$\begin{aligned}\mathbf{E}T_{5,n} &\leq (1+\epsilon) \cdot \mathbf{E}\left\{\left(1 + \frac{1}{\epsilon}\right) \cdot \frac{1}{n} \sum_{i=1}^{n} |Y_i - T_{\beta_n} Y_i|^2 \cdot 1_{[-\alpha_n, \alpha_n]^d}(X_i) \cdot 1_{A_n}\right\} \\ &\leq (1+\epsilon) \cdot \left(1 + \frac{1}{\epsilon}\right) \cdot \mathbf{E}\{|T_{\beta_n} Y - Y|^2\}.\end{aligned}$$

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

In the *third step of the proof,* we show

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

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

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

This implies

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

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

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

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

Hence, if the conditions required by Lemma 1, are satisfied for $\mathbf{a}_0 = \mathbf{w}^{(0)}$, we can conclude from the random initialization of $\mathbf{w}^{(0)}$ that (64) and (65) from Lemma 6 in the supplement, as well as (68) and (69) from Lemma 7 in the supplement hold with $\gamma_n^* = c_{22} \cdot (\log n)^2$ and $B_n = c_{23} \cdot (\log n)^2$. From this, Lemma 6 and Lemma 7 in the supplement and the assumptions on L_n and t_n in Theorem 1 we get that (10) and (11) hold.

In the *fifth step of the proof,* we show

$$\mathbf{P}(A_n^c) \leq \frac{c_{24}}{\beta_n^3}.$$

To do this, we bound the probability that the weight vector $\mathbf{w}^{(0)}$ does not satisfy the first condition in the definition of the event A_n by considering a sequential choice of the weights in the K_n fully connected neural networks which we compute in parallel. In a single fully connected neural network, the probability that all $(r+1)+(L-2)\cdot r\cdot(r+1)+r\cdot(d+1)$ weights satisfy the conditions of Lemma 5 for j_1 is bounded from below by

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

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

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

This implies that all conditions of Lemma 5 are satisfied outside of an event of probability

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

which is for large n less than

$$(K^2+1)^{3d}\cdot\left(1-\frac{1}{n^r}\right)^{n^r\cdot\log n}\leq(K^2+1)^{3d}\cdot e^{-\log n}=\frac{(K^2+1)^{3d}}{n}\leq\frac{1}{2}\cdot\frac{c_{25}}{\beta_n^3}$$

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

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

In the *sixth step of the proof*, we show

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

We have

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

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

$$\|\mathbf{w}^{(t_n)} - \mathbf{w}^{(0)}\|_\infty \leq \|\mathbf{w}^{(t_n)} - \mathbf{w}^{(0)}\| \leq c_{26} \cdot (\log n)^2.$$

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

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

where $\mathcal{F}$ is defined as in Lemma 2 with $C = c_{27} \cdot \sqrt{K_n} \cdot (\log n)^2$, $B = c_{28} \cdot (\log n)^2$ and $A = c_{29} \cdot n^\tau$. Application of Lemma 2 together with standard bounds of empirical process theory [cf., Theorem 9.1 in Györfi et al. (2002)] yields

$$\begin{aligned}
&\mathbf{P}\Bigg\{ (\mathbf{E}\{|m_n(X) - T_{\beta_n} Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X) | \mathcal{D}_n\} \\
&\qquad - \frac{1}{n} \sum_{i=1}^n |m_n(X_i) - T_{\beta_n} Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i)) \cdot 1_{A_n} > t \Bigg\} \\
&\quad \leq 8 \cdot \left(c_{30} \cdot \frac{\beta_n}{t/8} \right)^{c_{31} \cdot (\log n)^{c_{32}} \cdot n^{\tau \cdot d} \cdot \left(\frac{c_{33} \cdot \sqrt{K_n} \cdot (\log n)^2}{t/8} \right)^{d/k} + c_{34}} \cdot \exp\left(-\frac{n \cdot t^2}{128 \beta_n^4} \right).
\end{aligned}$$

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

$$c_{35} \cdot \exp\left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \leq c_{36} \cdot \exp\left(-\frac{\sqrt{n}}{256 \cdot \beta_n^4} \right).$$

Consequently, we get

$$\mathbf{E}\{T_{3,n}\} \leq (1+\epsilon) \cdot \left(\frac{1}{n^{1/4}} + 4\beta_n^2 \cdot c_{37} \cdot \exp\left(-\frac{\sqrt{n}}{256\beta_n^4} \right) \right) \to 0 \quad (n \to \infty).$$

In the *seventh step of the proof* we show

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

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

$$\mathbf{E}\{T_{7,n}\} \leq 2 \cdot 4\beta_n^2 \cdot \mathbf{P}(A_n^c) + 2 \int |\bar{m}(x) - m(x)|^2 \mathbf{P}_X(dx) \leq 2 \cdot 4\beta_n^2 \cdot \frac{c_{24}}{\beta_n^3} + 2\epsilon.$$

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

In the *eighth step of the proof*, we bound

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

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

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} |f_{\mathbf{w}^{(t_n)}}(X_i) - Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i) \\
&\quad \leq \frac{1}{n}\sum_{i=1}^{n} |f_{\mathbf{w}^{(t_n)}}(X_i) - Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i) + c_2 \cdot \sum_{k=1}^{K_n} ((\mathbf{w}^{(t_n)})_{1,1,k}^{(L)})^2 = F_n(\mathbf{w}^{(t_n)}) \\
&\quad \leq F_n(\mathbf{w}^{(t_n-1)}) - \frac{1}{2L_n} \cdot \|\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t_n-1)})\|^2 \\
&\quad \leq F_n(\mathbf{w}^{(t_n-1)}) - \frac{1}{2L_n} \cdot 4 \cdot c_2 \cdot (F_n(\mathbf{w}^{(t_n-1)}) - F_n((\mathbf{w}^*)^{(t_n-1)})) \\
&\quad = \left(1 - \frac{2 \cdot c_2}{L_n}\right) \cdot F_n(\mathbf{w}^{(t_n-1)}) + \frac{2 \cdot c_2}{L_n} \cdot F_n((\mathbf{w}^*)^{(t_n-1)}) \\
&\quad \leq \left(1 - \frac{2 \cdot c_2}{L_n}\right)^2 \cdot F_n(\mathbf{w}^{(t_n-2)}) + \frac{2 \cdot c_2}{L_n} \cdot F_n((\mathbf{w}^*)^{(t_n-1)}) \\
&\qquad + \frac{2 \cdot c_2}{L_n} \cdot \left(1 - \frac{2 \cdot c_2}{L_n}\right) F_n((\mathbf{w}^*)^{(t_n-2)}) \\
&\quad \leq \dots \\
&\quad \leq \left(1 - \frac{2 \cdot c_2}{L_n}\right)^{t_n} \cdot F_n(\mathbf{w}^{(0)}) + \sum_{k=1}^{t_n} \frac{2 \cdot c_2}{L_n} \cdot \left(1 - \frac{2 \cdot c_2}{L_n}\right)^{k-1} F_n((\mathbf{w}^*)^{(t_n-k)}).
\end{aligned}$$

This implies

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

We have

$$\left(1 - \frac{2 \cdot c_2}{L_n}\right)^{t_n} \le \exp\left(\frac{-2 \cdot c_2 \cdot t_n}{L_n}\right) \le \exp\left(-c_{38} \cdot \log n\right) \to 0 \quad (n \to \infty)$$

and

$$\begin{aligned}
&\mathbf{E}\{F_n((\mathbf{w}^*)^{(t_n-k)}) \cdot 1_{A_n}\} - \mathbf{E}\{|m(X) - Y|^2\} \cdot \mathbf{P}(A_n) \\
&= \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} |f_{((\mathbf{w}^*)^{(t_n-k)})}(X_i) - Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i) \cdot 1_{A_n} \right. \\
&\quad \left. + c_2 \cdot \sum_{j=1}^{K_n} |(\mathbf{w}^*)_{1,1,j}^{(L)}|^2 \right\} - \mathbf{E}\{|m(X) - Y|^2\} \cdot \mathbf{P}(A_n) \\
&= \left(\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} |f_{((\mathbf{w}^*)^{(t_n-k)})}(X_i) - Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i) \cdot 1_{A_n}\right\}\right. \\
&\quad \left. - (1+\epsilon) \cdot \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} |f_{((\mathbf{w}^*)^{(t_n-k)})}(X_i) - T_{\beta_n} Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i) \cdot 1_{A_n}\right\}\right) \\
&\quad + (1+\epsilon) \cdot \left(\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} |f_{((\mathbf{w}^*)^{(t_n-k)})}(X_i) - T_{\beta_n} Y_i|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_i) \cdot 1_{A_n}\right\}\right. \\
&\quad \left. - \mathbf{E}\left\{\mathbf{E}\left\{|f_{((\mathbf{w}^*)^{(t_n-k)})}(X) - T_{\beta_n} Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X) \cdot 1_{A_n} \middle| \mathcal{D}_n\right\}\right\}\right) \\
&\quad + (1+\epsilon) \cdot \left(\mathbf{E}\left\{|f_{((\mathbf{w}^*)^{(t_n-k)})}(X) - T_{\beta_n} Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X) \cdot 1_{A_n}\right\}\right. \\
&\quad \left. - (1+\epsilon)^2 \cdot \mathbf{E}\left\{|f_{((\mathbf{w}^*)^{(t_n-k)})}(X) - Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X) \cdot 1_{A_n}\right\}\right) \\
&\quad + (1+\epsilon)^2 \cdot \left(\mathbf{E}\left\{|f_{((\mathbf{w}^*)^{(t_n-k)})}(X) - Y|^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X) \cdot 1_{A_n}\right\}\right. \\
&\quad \left. - \mathbf{E}\{|m(X) - Y|^2\} \cdot \mathbf{P}(A_n)\right) \\
&\quad + ((1+\epsilon)^2 - 1) \cdot \mathbf{E}\{|m(X) - Y|^2\} + c_2 \cdot \sum_{j=1}^{K_n} |(\mathbf{w}^*)_{1,1,j}^{(L)}|^2 \\
&= T_{8,n} + T_{9,n} + T_{10,n} + T_{11,n} + T_{12,n} + T_{13,n}.
\end{aligned}$$

Similar to the second step, we obtain

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

According to the assertion of Lemma 4, we know that $f_{\mathbf{w}^*}$ is bounded. Thus, we get as in the sixth step

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

The choice of $\bar{m}$ and Lemma 5 imply

$$\begin{aligned}
&T_{11,n}/(1+\epsilon)^2 \\
&\quad \le \mathbf{E}\Big\{ \int |f_{((\mathbf{w}^*)^{(t_n-k)})}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{A_n} \Big\} \\
&\quad \le 2 \cdot \mathbf{E}\Big\{ \int |f_{((\mathbf{w}^*)^{(t_n-k)})}(x) - \bar{m}(x)|^2 \mathbf{P}_X(dx) \cdot 1_{A_n} \Big\} \\
&\qquad + 2 \int |\bar{m}(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\quad \le c_{39} \cdot \left(\frac{1}{K} + \frac{K^{12d}}{n^2} + \left(\frac{K^{6d}}{n} + 1 \right)^2 \mathbf{P}_X(\mathbb{R}^d \setminus [-K,K]^d) \right) + 2\epsilon,
\end{aligned}$$

from which we can conclude

$$\limsup_{n\to\infty} T_{11,n} \le c_{39} \cdot (1+\epsilon)^2 \Big(\frac{1}{K} + \mathbf{P}_X(\mathbb{R}^d \setminus [-K,K]^d) \Big) + 2\epsilon \cdot (1+\epsilon)^2.$$

Furthermore, we obtain by Lemma 5

$$T_{13,n} \le c_{40} \cdot (K^2+1)^{3d} \cdot \left(\frac{1}{(K^2+1)^{2d}} \right)^2.$$

Summarizing the above results yields

$$\begin{aligned}
&\limsup_{n\to\infty} \sum_{k=1}^{t_n} \frac{2 \cdot c_2}{L_n} \left(1 - \frac{2 \cdot c_2}{L_n} \right)^{(k-1)} \\
&\quad \cdot \Big(\mathbf{E}\{ F_n((\mathbf{w}^*)^{(t_n-k)}) \cdot 1_{A_n} \} - \mathbf{E}\{ |m(X) - Y|^2 \} \cdot \mathbf{P}(A_n) \Big) \\
&\quad \le c_{41} \cdot \Big(\frac{1}{K} + \mathbf{P}_X(\mathbb{R}^d \setminus [-K,K]^d) + \epsilon \Big) + c_{40} \cdot (K^2+1)^{3d} \cdot \left(\frac{1}{(K^2+1)^{2d}} \right)^2
\end{aligned}$$

and

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbf{E}\{ T_{6,n} \} \\
&\quad \le (1+\epsilon)^2 \left(c_{41} \cdot \Big(\frac{1}{K} + \mathbf{P}_X\big(\mathbb{R}^d \setminus [-K,K]^d\big) + \epsilon \Big) + c_{42} \cdot \frac{1}{(K^2+1)^d} \right) \\
&\qquad + ((1+\epsilon)^2 - 1) \cdot \mathbf{E}\{ |m(X) - Y|^2 \}.
\end{aligned}$$

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

$$\limsup_{n\to\infty} \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_{43} \cdot \epsilon.$$

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

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-024-00898-6.

Acknowledgements The authors would like to thank an anonymous referee for many invaluable comments which helped to improve an early version of this manuscript.

References

Allen-Zhu, Z., Li, Y., Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th international conference on machine learning, 97*, 242–252.

Bartlett, P. L., Montanari, A., Rakhlin, A. (2021). Deep learning: A statistical viewpoint. *Acta Numerica, 30*, 87–201.

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

Braun, A., Kohler, M., Langer, S., Walk, H. (2021). Convergence rates for shallow neural networks learned by gradient descent. *Bernoulli, 30*, 475–502.

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

Dyer, E., Gur-Ari, G.(2019). Asymptotics of wide networks from Feynman diagrams. arXiv:1909.11304.

Györfi, L., Kohler, M., Krzyżak, A., Walk, H. (2002). Why is nonparametric regression important? In *Springer series in statistics, A distribution-free theory of nonparametric regression* (pp. 1–16). New York: Springer.

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

Imaizumi, M., Fukamizu, K. (2019). Deep neural networks learn non-smooth functions effectively. In *The 22nd international conference on artificial intelligence and statistics* (pp. 869–878).

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

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv:1408.5882.

Kohler, M., Krzyżak, A. (2017). Nonparametric regression based on hierarchical interaction models. *IEEE Transaction on Information Theory, 63*, 1620–1630.

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

Kohler, M., Krzyżak, A. (2022). Over-parametrized neural networks learned by gradient descent can generalize especially well. Submitted for publication. https://www2.mathematik.tu-darmstadt.de/~kohler/preprint22_01.pdf.

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

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Association for Computing Machinery, 60*, 84–90.

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

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

Li, G., Gu, Y., Ding, J. (2021). The rate of convergence of variation-constrained deep neural networks. arXiv:2106.12068.
Lu, J., Shen, Z., Yang, H., Zhang, S.(2020). Deep network approximation for smooth functions. arXiv: 2001.03040
Mei, S., Montanari, A., Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences, 115*, E7665–E7671.
Nguyen, P.-M., Pham, H. T. (2020). A rigorous framework for the mean field limit of multilayer neural networks. arXiv:2001.1144.
Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function (with discussion). *Annals of Statistics, 48*, 1875–1897.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T. P., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature, 550*, 354–359.
Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics, 5*, 595–645.
Suzuki, T. (2018). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. arXiv:1810.08033.
Suzuki, T., Nitanda, A. (2019). Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. arXiv:1910.12799.
Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144.
Yang, G., Hu, J. E. (2020). Feature learning in infinite-width neural networks. arXiv:2011.14522.
Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks, 94*, 103–114.
Yarotsky, D., Zhevnerchuk, A. (2020). The phase diagram of approximation rates for deep neural networks. *Advances in Neural Information Processing Systems, 33*, 13005–13015.