Estimation of hierarchical composition models by deep neural network regression estimates learned by gradient descent *

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

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

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

February 26, 2025

Abstract

Nonparametric regression with random design is considered. The L_2 error with integration with respect to the design measure is used as the error criterion. Over-parametrized deep neural network regression estimates with logistic activation function are defined, where all weights are learned by gradient descent. It is shown that the estimates are able to adapt to hierarchical composition models, i.e., in case that the regression function satisfies such a model the estimates achieve a rate of convergence which is nearly optimal for this model and hence are able to circumvent the curse of dimensionality.

AMS classification: Primary 62G08; secondary 62G20.

Key words and phrases: Deep neural networks, dimension reduction, gradient descent, over-parametrization, rate of convergence, regression estimation.

1 Introduction

Motivated by the great success of deep learning in applications, there is an increasing interest in understanding theoretically why these estimate are so successful in practice. In this context these estimates are often studied in the field of nonparametric regression.

In nonparametric regression, an $\mathbb{R}^d \times \mathbb{R}$–valued random vector with $\mathbf{E}Y^2 < \infty$ is given and the problem of predicing the value of Y given the observed value of X is considered. If the main goal of the analysis is the minimization of the expected squared error of prediction, then the task is to find a function $f^* : \mathbb{R}^d \to \mathbb{R}$ such that its so-called L_2 risk

$$\mathbf{E}\{|Y - f^*(X)|^2\} \tag{1}$$

*Running title: *Estimation of hierarchical composition models by deep neural networks*

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

is as small as possible.

Let $m : \mathbb{R}^d \to \mathbb{R}$ defined by $m(x) = \mathbf{E}\{Y|X = x\}$ be the so–called regression function corresponding to (X, Y). Then for any measurable function $f : \mathbb{R}^d \to \mathbb{R}$

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

holds (cf., e.g., Section 1.1 in Györfi et al. (2002)), which implies that the regression function is the optimal predictor and that the so-called L_2 error

$$\int |f(x) - m(x)^2 \mathbf{P}_X(dx) \tag{3}$$

describes how far the L_2 risk of a function f is away from its optimal value $\mathbf{E}\{|Y - m(X)|^2\}$.

In applications usually the distribution of (X, Y) and hence also the regression function is unknown. But often it is possible to observe a sample of this distribution, and the task is then to estimate the corresponding regression function. Formally, this can be described as follows: Given a data set

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \tag{4}$$

where (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n) are independent and identically distributed, construct an estimate

$$m_n(\cdot) = m_n(\cdot, \mathcal{D}_n) : \mathbb{R}^d \to \mathbb{R} \tag{5}$$

such that its L_2 error

$$\int |m_n(x) - m(x)^2 \mathbf{P}_X(dx) \tag{6}$$

is as small as possible.

It is well-known that without regularity assumptions on the underlying distribution, in particular on the smoothness of the regression function, nontrivial results about the rate of convergence of (6) towards zero cannot be derived (cf., e.g., Chapter 3 in Györfi et al. (2002) and Devroye (1982)). Stone (1982) considered regression functions which are (p, C)–smooth according to the following definition.

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

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

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

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

$$n^{-\frac{2p}{2p+d}}. \tag{7}$$

If p is small compared to d then this rate of convergence converges to zero rather slowly (so-called curse of dimensionality). Since this rate of convergence is optimal, it can not be improved in general. The only way to circumvent this curse of dimensionality is to use additional assumptions on the structure of the regression function in order to be able to derive better rates.

Stone (1985) showed that in case of additive models, where the regression function is assumed to be a sum of d univariate functions applied to the d components of x, suitably defined estimates achieve the rate of convergence

$$n^{-\frac{2p}{2p+1}}.$$

More generally, it was shown in Stone (1994) that in case that the regression function satisfies an interaction model, where the regression function is given by a sum of functions applied to subsets consisting of d^* of the d components of x, suitably defined estimates achieve a rate of convergence of order

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

Other classical assumptions which lead to a dimension reduction include single index models (cf., Härdle, Hall and Ichimura (1993), Härdle and Stoker (1989), Yu and Ruppert (2002) and Kong and Xia (2007)) or projection pursuit (cf, Friedman and Stuetzle (1981)).

For least squares neural network regression estimates it has been shown that these estimates can achieve a dimension reduction under rather general assumptions, which is one possibility to explain theoretically the success of deep learning in practice. In its most general form, which includes additive models, interaction models, single index models and projection pursuit models, the regression function is assumed to be a composition of functions either depending only on a few components or being rather smooth. This assumption can be formalized as follows:

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

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

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

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

It was shown in Schmidt-Hieber (2020) that suitably defined least squares neural network regression estimates achieve up to a logarithmic factor a rate of convergence of

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

in case that the regression function satisfies a hierarchical composition model of some finite level with order and smoothness constraint $\mathcal{P}$ (for related results see Kohler and Krzyżak (2017), Bauer and Kohler (2019) and Kohler and Langer (2021)).

These results are valid for least squares neural network estimates, which are not computable in practice. In practice neural network estimates are learned by gradient descent, so if one wants to explain the success of deep neural networks in practice theoretically by showing a dimension reduction in case that the regression function satisfies a hierarchical composition model, it is necessary to show that such a dimension reduction also takes place in case that the estimate is learned via gradient descent.

1.1 Main result in this article

In this article we show that the expected L_2 error of suitably defined neural network regression estimates m_n with logistic activation function, where all parameters are learned by gradient descent, satisfies for any $\epsilon \in (0,1)$

$$\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_1 \cdot \max_{(p,K)\in\mathcal{P}} n^{-\frac{2p}{2p+K}+\epsilon}$$

in case that the regression function satisfies a hierarchical composition model of some finite level with order and smoothness constraint $\mathcal{P}$. Here the neural network depends on (in the sample size) polynomially many weights, however the number of gradient descent steps is chosen to be exponentially large, and during gradient descent exponentially many pruning steps are needed (such that only varying subsets of the network are considered during gradient descent). The pruning steps are here used to simplify the optimization of the neural networks.

1.2 Discussion of related results

Inspired by immense success of deep learning in applications a lot of effort has been recently devoted to analyze deep learning theoretically. The researchers focused on approximation and estimation capabilities of deep network estimates as well as their efficient implementations, i.e., on optimization. For some recent approximation results we refer the reader to Yarotsky (2018), Yarotsky and Zhevnerchute (2019), Lu et al. (2020), Langer (2021) and the literature cited therein. These papers demonstrate that smooth functions can be approximated well by deep neural networks having appropriate topology and they specify the numbers of nonzero weights necessary to approximate smooth function up to some given error.

In practice, functions which one wants to approximate have to be estimated from the observed data, which are usually contaminated by random errors. It has been studied in

the literature how well deep networks learned from such noisy data generalize on a new independent test data. Such results have been achieved by means of the classical VC theory, e.g. by bounding the VC dimension of classes of neural networks, see Bartlett et al. (2019), or in case of over-parametrized deep neural networks, in which the number of free parameters learned from the observed data significantly exceeds the sample size, by bounding the Rademacher complexity, see, e.g., Liang, Rakhlin and Sridharan (2015), Golowich, Rakhlin and Shamir (2019), Lin and Zhang (2019), Wang and Ma (2022) and the literature cited therein. By putting together these results one could handle the error of the least squares regression estimates. As it was demonstrated in the papers by Kohler and Krzyżak (2017), Bauer and Kohler (2019), Schmidt-Hieber (2020) and Kohler and Langer (2021) the least squares regression estimates using deep networks are able to achieve a dimension reduction for estimated functions satisfying a hierarchical composition model, i.e., whenever estimated functions are compositions of smooth functions either depending only on a few components or being rather smooth. This property follows from the network structure of deep networks implying that composition of networks is itself a deep network. Thus, any approximation result obtained for some functions by using deep networks can be extended to an approximation result for composition of such functions by a deep network representing a composition of the approximating networks. Here the number of weights and the depth of the network determining the VC dimension and consequently the complexity of the network provided that it is not over-parametrized (cf., Bartlett et al. (2019)), do not change a lot. Consequently such networks have approximation properties and complexity of a network for low dimensional predictors and can thus achieve a dimension reduction.

Quite a large number of interesting results on optimization of deep neural networks have recently appeared in the literature obtained, see, e.g., Zou et al. (2018), Du et al. (2019), Allen-Zhu, Li and Song (2019) and Kawaguchi and Huang (2019), where authors applied gradient descent to over-parameterized deep neural networks and analyzed the results. These papers demonstrated that this leads to neural networks which (globally) minimize the empirical risk, but unfortunately, the corresponding estimates do not perform well on a new independent data, see Kohler and Krzyżak (2021).

The aforementioned theoretical results do not provide clear guidance to practitioners applying deep neural networks, where it is essential to control simultaneously all three errors, i.e., the approximation, generalization and optimization errors (cf., Kutyniok (2020)). None of the works mentioned thus far deal with all these three errors together.

There are situations where approximation, estimation and optimization errors are investigated simultaneously in some equivalent models of deep learning. The best known approach in this domain is the neural tangent kernel approach proposed by Jacot, Gabriel and Hongler (2020). In this approach a kernel estimate is studied in lieu of neural network estimate and the error of the kernel estimate is used to bound the error of the neural network estimate, see Hanin and Nica (2019) and the literature cited therein for related work. Nitanda and Suzuki (2021) observed that in most studies on the neural tangent kernel the equivalence to deep neural networks holds only pointwise rather than for the global L_2 error, hence we cannot draw conclusions about the behavior of the L_2 error of the deep neural network from these results. Nitanda and Suzuki (2021) were able to

analyze the global error of over-parametrized shallow neural networks learned by gradient descent. However, the use of the neural tangent kernel implies that the smoothness condition imposed on the function to be estimated needs to be defined in terms of a norm involving the kernel, which does not lead to the standard classical smoothness conditions, making it difficult to interpret the obtained results. Furthermore, their result did not specify the number of neurons that shallow neural network must posses, it only implied that the number of neurons must be sufficiently large. Thus it is not clear whether the number of neurons should grow, e.g., exponentially in the sample size or not. Another estimation approach studied in some asymptotically equivalent model is the mean field approach, see Mei, Montanari, and Nguyen (2018), Chizat and Bach (2018) or Nguyen and Pham (2020). The problem with this approach is that it is unclear how close the behaviour of the deep networks in the equivalent model mimics their behaviour in the applications, because they are based on some approximation of the application using e.g. some asymptotic expansions.

The results of this paper follow the statistical theory for deep neural networks developed by Braun et al. (2023), Drews and Kohler (2022, 2023), Kohler and Krzyżak (2022, 2023) and Kohler (2024).

1.3 Notation

The sets of natural numbers, real numbers and nonnegative real numbers are denoted by $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$, respectively. For $z \in \mathbb{R}$, we denote the smallest integer greater than or equal to z by $\lceil z \rceil$. And the largest integer less than or equal to z is denoted by $\lfloor z \rfloor$. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$. For a closed and convex set $A \subseteq \mathbb{R}^d$ we denote by $Proj_A x$ that element $Proj_A x \in A$ with

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

The diameter of a set $A \subseteq \mathbb{R}^d$ (w.r.t. the Euclidean norm) is denoted by $diam(A)$. For $f : \mathbb{R}^d \to \mathbb{R}$

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

is its supremum norm, and we set

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

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

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

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

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

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

1.4 Outline

Section 2 contains the definition of the estimate. The main result is presented in Section 3 and proven in Section 4.

2 Definition of the estimate

Throughout the paper we let $\sigma(x) = 1/(1+e^{-x})$ be the logistic squasher, i.e., we use the so–called logistic activation function.

In the sequel we will define hierarchically composed neural networks, which use as building blocks neural networks with the following special topology: Let $K, L, r \in \mathbb{N}$ be parameters of our class of neural networks. We consider neural networks which consist of K fully connected neural networks of depth L and width r computed in parallel and compute a linear combination of the outputs of these K neural networks. The weights in the k-th such network are denoted by $(w_{k,i,j}^{(l)})_{i,j,l}$, where $w_{k,i,j}^{(l)}$ is the weight between neuron j in layer l and neuron i in layer $l+1$. Formally we set for $x \in \mathbb{R}^d$

$$f_{\mathbf{w}}(x) = \sum_{k=1}^{K} w_{k,1,1}^{(L)} \cdot f_{k,1}^{(L)}(x) \tag{8}$$

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

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

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

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

for some $w_{k,i,0}^{(0)}, \ldots, w_{k,i,d}^{(0)} \in \mathbb{R}$. Let $\mathcal{F}_{d,K,L,r}$ be the set of all neural networks (8) of the above form.

In the sequel we will estimate a regression function which satisfies a hierarchical composition model by a recursively defined function $h : \mathbb{R}^d \to \mathbb{R}$. Here we will compose functions of $\mathcal{F}_{d,K,L,r}$ with different values for K, L, r and d.

Let $l \in \mathbb{N}$ be the depth of the hierarchical composition of neural networks. We define $h : \mathbb{R}^d \to \mathbb{R}$ recursively by

$$h(x) = h_1^{(l)}(x) \quad (x \in \mathbb{R}^d) \tag{11}$$

where

$$h_i^{(s)}(x) = g_{NN,i,s}(h^{(s-1)}_{\sum_{r=1}^{i-1} K_{r,s}+1}(x), \ldots, h^{(s-1)}_{\sum_{r=1}^{i-1} K_{r,s}+K_{i,s}}(x)) \tag{12}$$

for some

$$g_{NN,i,s} \in \mathcal{F}_{K_{i,s},K_n^{(s)},L_s,r_{i,s}}$$

in case $s \in \{1,\dots,l\}$ and $i \in \{1,\dots,N_s\}$, and

$$h_i^{(0)}(x) = g_{NN,i,0}(x^{(1)},\dots,x^{(d)}) \tag{13}$$

for some

$$g_{NN,i,0} \in \mathcal{F}_{d,K_n^{(0)},L_0,r_{i,0}}$$

in case $i \in \{1,\dots,N_0\}$. Here $N_s \in \mathbb{N}$ is the number of functions $g_{NN,i,s}$ at level s which is given by

$$N_l = 1 \quad \text{and} \quad N_s = \sum_{r=1}^{N_{s+1}} K_{r,s} \quad \text{for } s \in \{0,\dots,l-1\},$$

and $K_n^{(s)}, L_s, r_{i,s}, K_{i,s} \in \mathbb{N}$ are parameters of the estimate.

In order to learn an estimate of the above type from the data using gradient descent we proceed as follows:

We start with an initialization of the weights where all the weights of the hierarchically composed networks $g_{NN,i,s}$ above are initialized independently from each other as follows: We choose $w_{k,1,1}^{(L_s)}$ uniformly distributed on $[-c_{3,n}, c_{3,n}]$ in case $s < l$, and in case $s = l$ we set $w_{k,1,1}^{(L_l)} = 0$ for $k \in \{1,\dots,K_n^{(l)}\}$. We choose $w_{k,i,j}^{(t)}$ uniformly distributed on $[-c_{2,n}, c_{2,n}]$ if $t \in \{1,\dots,L_s - 1\}$, and we choose $w_{k,i,j}^{(0)}$ uniformly distributed on

$$[-c_{1,i,s,n}, c_{1,i,s,n}].$$

Here $c_{1,i,s,n}, c_{2,n}, c_{3,n} > 0$ are parameters of the estimate, and the random values are defined such that all components of $\mathbf{w}$ are independent, where $\mathbf{w}$ is the weight vector containing all weights of the network as its components.

If we introduce after each of the above networks except the network on the highest level an additional layer with the identity function as the activation function, we can describe this deep network by a network of depth

$$L = \sum_{s=0}^{l-1}(L_s + 1) + L_l = \sum_{s=0}^{l} L_s + l$$

as follows:

$$f_{\mathbf{w}}(x) = h_1^{(l)}(x) = \sum_{j \in \{1,\dots,k_L\} : (L,1,j) \in I} w_{1,j}^{(L)} \cdot f_j^{(L)}(x) = \sum_{k=1}^{K_n^{(l)}} w_{1,k}^{(L)} \cdot f_k^{(L)}(x), \tag{14}$$

where

$$f_i^{(s)}(x) = \sigma_s \left(\sum_{j \in \{0,\dots,k_{s-1}\} : (s-1,i,j) \in I} w_{i,j}^{(s-1)} \cdot f_j^{(s-1)}(x) \right) \quad \text{for } s \in \{1,\dots,L\} \text{ and } i > 0 \tag{15}$$

and

$$f_i^{(1)}(x) = \sigma_1\left(\sum_{j\in\{0,\dots,k_0\}:\,(0,i,j)\in I} w_{i,j}^{(0)}\cdot f_j^{(0)}(x)\right) = \sigma\left(\sum_{j=1}^{d} w_{i,j}^{(0)}\cdot x^{(j)} + w_{i,0}^{(0)}\right) \quad \text{for } i>0. \tag{16}$$

The activation functions depend here on the layer and are given by

$$\sigma_s(x) = \begin{cases} x & \text{if } s\in\{L_0+1, L_0+L_1+2.\dots, L_0+\dots+L_{l-1}+l\},\\ \frac{1}{1+e^{-x}} & \text{elsewhere.}\end{cases}$$

And we have used the abbreviations

$$f_0^{(s-1)}(x)=1 \quad \text{for } s\in\{1,\dots,L\} \quad \text{and} \quad f_j^{(0)}(x)=x^{(j)} \quad \text{for } j\in\{1,\dots,d\},$$

and $k_s\in\mathbb{N}$ and $I\subseteq\{0,\dots,L\}\times\mathbb{N}\times\mathbb{N}$ are implicitly defined by (11)–(16).

This gives us our initial weight vector $\mathbf{w}^{(0)}$ and our initial estimate $f_{\mathbf{w}^{(0)}}$.

After that we perform $t_n\in\mathbb{N}$ gradient descent steps each with a step size $\lambda_n>0$ and an additional projection step. More precisely, we let $\mathbf{W}$ be the set of all weight vectors $\mathbf{w}=(w_{i,j}^{(s)})_{s,i,j:0\le s\le L}$ which satisfy

$$\sum_{k=1}^{K_n^{(l)}} |w_{1,k}^{(L)}|^2 \le \alpha_n \quad \text{and} \quad \|(w_{i,j}^{(s)})_{i,j,s:(s,i,j)\in I,s<L} - ((\mathbf{w}^{(0)})_{i,j}^{(s)})_{i,j,s:(s,i,j)\in I,s<L}\| \le \delta_n, \tag{17}$$

where $\alpha_n,\delta_n\ge 0$ are parameters of the estimate. We choose a stepsize $\lambda_n\ge 0$ and a number $t_n\in\mathbb{N}$ of gradient descent steps and we set

$$\mathbf{w}^{(t)} = Proj_{\mathbf{W}}\left(\mathbf{w}^{(t-1)} - \lambda_n\cdot\nabla_{\mathbf{w}}F_n(\mathbf{w}^{(t-1)})\right) \quad (t=1,\dots,t_n), \tag{18}$$

where

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

is the empirical L_2 risk of $f_{\mathbf{w}}$.

During gradient descent (and also directly after the initialization) we apply the following pruning step (which simplifies the optimization during the computation of the estimate): For

$$t\in\{0, s_n, 2\cdot s_n,\dots\}$$

we select in the output level of all $g_{NN,i,s}$ with $s<l$ randomly $\bar{K}_n^{(s,i)}$ of the $K_n^{(s)}$ weights using the uniform distribution. We ignore until the next pruning step all weight vectors not chosen together with all the weights of the in parallel computed completely connected small networks for which they are the top weights in all computations (also in the projection step, so we compute the norm in the projection step using only a subset of the weights). And directly after the selection of the weights we project the weight vector of the chosen subnetwork towards the corresponding subvector of weights from $\mathbf{w}^{(0)}$ in

our standard way. This random selection of the weights is done independently for all $g_{NN,i,s}$. Furthermore, we keep the values of all weights not chosen during one pruning step constant until the next pruning step.

Finally we define our estimate as a truncated version of the neural network with that weight vector $\mathbf{w}^{(\hat{t})}$ for which the empirical L_2 risk was minimal during the training, i.e., we set

$$m_n(x) = T_{\beta_n}(f_{\mathbf{w}^{(\hat{t})}}(x)) \quad \text{where} \quad \hat{t} = \arg \min_{t \in \{0,1,\ldots,t_n-1\}} F_n(\mathbf{w}^{(t)}) \tag{19}$$

and $\beta_n = c_2 \cdot \log n$.

3 Main result

Our main result is the following bound on the expected L_2 error of this estimate.

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

$$\mathbf{E}\{\exp(c_3 \cdot Y^2)\} < \infty \tag{20}$$

holds for some $c_3 > 0$. Let $K_{s,r} \in \mathbb{N}$ with $K_{0,r} = d$ and set $N_l = 1$ and $N_s = \sum_{r=1}^{N_{s+1}} K_{r,s}$ for $s = 0, \ldots, l-1$. Assume that the regression function $m : \mathbb{R}^d \to \mathbb{R}$ is given by

$$m(x) = h_1^{(l)}(x),$$

where

$$h_i^{(s)}(x) = g_{i,s}\left(h^{(s-1)}_{\sum_{r=1}^{i-1} K_{s-1,r}+1}(x), \ldots, h^{(s-1)}_{\sum_{r=1}^{i-1} K_{s-1,r}+K_{s-1,i}}(x)\right)$$

for $s \in \{1, \ldots, L\}$, $i \in \{1, \ldots, N_s\}$,

$$h_i^{(0)}(x) = g_{i,0}(x)$$

for $i \in \{1, \ldots, N_0\}$. Assume that

$$g_{i,s} : \mathbb{R}^{K_{i,s}} \to \mathbb{R}$$

are $(p_{i,s}, C_{i,s})$–smooth for some $p_{i,s} \geq 1$, $C_{i,s} > 0$ for all i, s. Let $r_{max} = \max_{i,s} r_{i,s}$ and $K_{max} = \max_{i,s} K_{i,s}$. Set $\beta_n = c_2 \cdot \log n$ for some $c_2 > 0$ satisfying $c_2 \cdot c_3 \geq 3$. Set

$$K_n^{(s)} = K_n \quad (s = 0, \ldots, l), \quad L_s = \bar{L} \quad (s = 0, \ldots, l) \quad \text{and} \quad r_{i,s} = 2 \cdot (\lceil 2p_{i,s} + K_{i,s} \rceil)^2,$$

where $K_n \in \mathbb{N}$ satisfies

$$\frac{K_n}{n^{(3 \cdot l + 10) \cdot (r_{max}+1)^2 \cdot (\bar{L}+K_{max})+7}} \to \infty \quad (n \to \infty)$$

and

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

for some $\kappa > 0$, and where

$$\bar{L} = \max_{r,s} \lceil \log_2(p_{r,s} + K_{r,s}) \rceil.$$

Set

$$\bar{K}_n^{(s,i)} = \left\lceil c_4 \cdot n^{\frac{K_{i,s}}{2p_{i,s}+K_{i,s}}} \right\rceil$$

for $s = 1, \ldots, l-1$. Set

$$c_{1,i,s,n} = c_5 \cdot n^{\frac{1}{2p_{i,s}+K_{i,s}}} \cdot \log n, \quad c_{2,n} = c_6, \quad c_{3,n} = n, \quad \delta_n = \frac{1}{n^{3l+6}}, \quad \alpha_n = \frac{1}{n^2},$$

$$\lambda_n = \frac{1}{n^4 \cdot K_n^3}, \quad s_n = n^4 \cdot K_n^3 \quad \textit{and} \quad t_n = \lceil e^{(\log n)^2 \cdot n} \rceil.$$

Let $\sigma(x) = 1/(1+e^{-x})$ be the logistic squasher, let $c_6 > 0$ be sufficiently large, and define the estimate m_n as above.

Then we have for any $\epsilon > 0$:

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_7 \cdot \max_{i,s} n^{-\frac{2p_{i,s}}{2p_{i,s}+K_{i,s}}+\epsilon}.$$

Remark 1. In the assumption on the regression function in Theorem 1 we can use projections for the functions $h_i^{(0)}$, and since these projections are (p, C)–smooth for any $p > 0$ they will have no influence on the upper bound on the rate of convergence in Theorem 1.

Remark 2. The rate of convergence in Theorem 1 is optimal up to the ϵ factor in the exponent. This factor ϵ appears due to our use of metric entropy bounds for bounding the complexity of the over-parametrized deep neural networks. It is an open problem whether one can show the same result without the ϵ factor in the exponent.

Remark 3. The result in Theorem 1 is valid for the logistic squasher. In our proof the smoothness of the activation function is crucial in order to be able to apply the metric entropy bounds mentioned in Remark 2, hence the result does not hold for the ReLU activation function and it is an open problem whether one can show a similar result for the ReLU activation function.

The estimate above depends on the structure of the hierarchical composition model. Since in practice this structure will usually be unknown, any estimate using this structure in its definition cannot be applied directly. Of course, one can consider the whole structure of the network as a parameter of the network and use a standard technique (like splitting of the sample or cross validation (cf., e.g., Chapters 7 and 8 in Györfi et al. (2002)) to choose this parameter in a data dependent way. However, in the case of a hierarchical composition model there are too many values for the parameter to be considered to apply this in practice.

In the sequel we describe a more specific assumption from Kohler and Krzyżak (2017) for which least squares neural network regression estimates can achieve a dimension reduction (cf., Kohler and Krzyżak (2017) and Bauer and Kohler (2019)) and show that Theorem 1 can also be applied to this special situation. This assumption will depend on much less parameters so that they can in principle be chosen by splitting of the sample or cross validation. Our next definition describes this more specific assumption.

Definition 3 *Let $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ and $m : \mathbb{R}^d \to \mathbb{R}$.*
a) *We say that m satisfies a* **generalized hierarchical interaction model of order d^* and level 0**, *if there exist $a_1, \dots, a_{d^*} \in \mathbb{R}^d$ and $f : \mathbb{R}^{d^*} \to \mathbb{R}$ such that*

$$m(x) = f(a_1^T x, \dots, a_{d^*}^T x) \quad \text{for all } x \in \mathbb{R}^d.$$

b) *We say that m satisfies a* **generalized hierarchical interaction model of order d^* and level $l+1$**, *if there exist $K \in \mathbb{N}$, $g_k : \mathbb{R}^{d^*} \to \mathbb{R}$ $(k = 1, \dots, K)$ and $f_{1,k}, \dots, f_{d^*,k} : \mathbb{R}^d \to \mathbb{R}$ $(k = 1, \dots, K)$ such that $f_{1,k}, \dots, f_{d^*,k}$ $(k = 1, \dots, K)$ satisfy a generalized hierarchical interaction model of order d^* and level l and*

$$m(x) = \sum_{k=1}^{K} g_k \left(f_{1,k}(x), \dots, f_{d^*,k}(x) \right) \quad \text{for all } x \in \mathbb{R}^d.$$

c) *We say that the* **generalized hierarchical interaction model** *defined above* **is (p, C)-smooth**, *if all functions occurring in its definition are (p, C)–smooth according to Definition 1.*

So let us assume from now on that the regression function $m : \mathbb{R}^d \to \mathbb{R}$ satisfies a (p, C)–smooth generalized hierarchical interaction model of order d^* and finite level $\bar{l}$. Let $I \in \mathbb{N}$ be the maximal number K which occurs in part b) of the definition of this generalized hierarchical interaction model. We choose the topology of our neural network estimate by (11)-(13), where we use the following values for the parameters: We set $l = 2 \cdot \bar{l} + 1$,

$$K_{i,s} = \begin{cases} I & \text{if } s \in \{3, 5, \dots, 2 \cdot \bar{l} + 1\}, \\ d^* & \text{if } s \in \{1, 2, 4, 6, \dots, 2 \cdot \bar{l}\}, \\ d & \text{if } s = 0, \end{cases}$$

$K_n^{(s)} = K_n$ (where the value of K_n will be chosen in Corollary 1 below),

$$L_s = L = \max \left\{ \left\lceil \log_2 \left(\frac{p \cdot d}{d^*} + d \right) \right\rceil, \left\lceil \log_2 \left(\frac{p \cdot I}{d^*} + I \right) \right\rceil \right\}$$

and

$$r_{i,s} = \begin{cases} 2 \cdot (\lceil 2 \cdot \frac{p \cdot I}{d^*} + I \rceil)^2 & \text{if } s \in \{3, 5, \dots, 2 \cdot \bar{l} + 1\}, \\ 2 \cdot (\lceil 2 \cdot p + d^* \rceil)^2 & \text{if } s \in \{1, 2, 4, 6, \dots, 2 \cdot \bar{l}\}, \\ 2 \cdot (\lceil 2 \cdot \frac{p \cdot d}{d^*} + d \rceil)^2 & \text{if } s = 0. \end{cases}$$

Furthermore we set

$$\bar{K}_n^{(i,s)} = \left\lceil c_9 \cdot n^{\frac{d^*}{2p+d^*}} \right\rceil,$$

$$c_{1,i,s,n} = \begin{cases} c_{10} \cdot n^{\frac{d^*}{2I \cdot p + d^* \cdot I}} \cdot \log n & \text{if } s \in \{3, 5, \ldots, 2 \cdot \bar{l} + 1\}, \\ c_{11} \cdot n^{\frac{1}{2p+d^*}} \cdot \log n & \text{if } s \in \{1, 2, 4, 6, \ldots, 2 \cdot \bar{l}\}, \\ c_{12} \cdot n^{\frac{d^*}{2 \cdot d \cdot p + d^* \cdot d}} & \text{if } s = 0, \end{cases}$$

and choose $c_{2,n}$, $c_{3,n}$, δ_n, α_n, λ_n, s_n and t_n as in Theorem 1. Let σ be the logistic squasher and define the estimate m_n as in Section 2.

Corollary 1 *Let $n \in \mathbb{N}$, let (X, Y), (X_1, Y_n), ..., (X_n, Y_n) be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$–valued random variables such that supp(X) is bounded and that (20) holds for some $c_3 > 0$. Assume that the regression functions satisfy a (p, C)–smooth generalized hierarchical interaction model of order $d^* \in \{1, \ldots, d\}$ and finite level $\bar{l}$ and define the estimate m_n as above, where $K_n \in \mathbb{N}$ is chosen such that*

$$\frac{K_n}{n^{(6 \cdot \bar{l} + 13) \cdot (2 \cdot (2 \cdot p \cdot I + 2 \cdot p \cdot d + I + d)^2 + 1) \cdot (\bar{L} + d + I) + 7}} \to \infty \quad (n \to \infty)$$

and

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

for some $\kappa > 0$ hold. Assume $c_2 \cdot c_3 \geq 3$.

Then we have for any $\epsilon > 0$:

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_{13} \cdot n^{-\frac{2p}{2p+d^*} + \epsilon}.$$

Proof. The generalized hierarchical interaction model in Corollary 1 can be represented as the model for m in Theorem 1 if we choose $l = 2 \cdot \bar{l} + 1$ and

$$h_i^{(0)}(x) = a_i^T x,$$

$$h_i^{(1)}(x) = f_i(h_{j_{1,i,1}}^{(0)}(x), \ldots, h_{j_{d^*,i,1}}^{(0)}(x)),$$

$$h_i^{(2 \cdot s)}(x) = g_i(h_{j_{1,i,2s}}^{(2 \cdot s - 1)}(x), \ldots, h_{j_{d^*,i,2s}}^{(2 \cdot s - 1)}(x)) \quad (s = 1, \ldots, \bar{l})$$

and

$$h_i^{(2 \cdot s + 1)}(x) = \sum_{k=1}^{I} h_{j_{k,i,2s+1}}^{(2 \cdot s)}(x) \quad (s = 1, \ldots, \bar{l}),$$

where we have used that we can extend the last sum to I terms by just adding zeros.

The projections at level 0 and the sums at levels 3, 5, ... are arbitrary smooth, so in particular we can assume that they are $(\bar{p}_1, C)$–smooth and $(\bar{p}_1, C)$–smooth, respectively, with $\bar{p}_1 = p \cdot d / d^*$ and $\bar{p}_2 = p \cdot I / d^*$.

The parameters of our estimate are chosen such that the assumptions of Theorem 1 are satisfied. Hence we can conclude from Theorem 1

$$\begin{aligned}\mathbf{E}\int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx) &\leq c_{14}\cdot \max\left\{n^{-\frac{2\bar{p}_1}{2\bar{p}_1+d}+\epsilon}, n^{-\frac{2p}{2p+d^*}+\epsilon}, n^{-\frac{2\bar{p}_2}{2\bar{p}_2+I}+\epsilon}\right\} \\ &= c_{15}\cdot n^{-\frac{2p}{2p+d^*}+\epsilon}.\end{aligned}$$

□

Remark 4. The results above require an exponential large number of gradient descent steps (in the sample size). It is an open problem whether one can show a similar result for the number of gradient descent steps growing only polynomially.

Remark 5. It follows from the proof of Theorem 1 that it also holds if only the outer weights of the deep neural network in our estimate are learned by gradient descent (and the weights on all levels $s < L$ use during gradient descent always their initial value).

4 Proof of Theorem 1

4.1 Neural network optimization

Our first lemma is our main tool to analyze the gradient descent.

Lemma 1 *Let $d_1, d_2 \in \mathbb{N}$, let $C_n, D_n \geq 0$, let $A \subset \mathbb{R}^{d_1}$ and $B \subseteq \mathbb{R}^{d_2}$ be closed and convex, and let $F : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}_+$ be a function such that*

$$u \mapsto F(u,v) \quad \textit{is differentiable and convex for all } v \in \mathbb{R}^{d_1}$$

and

$$\|(\nabla_u F)(u,v)\| \leq C_n \tag{21}$$

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

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

where

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

Let $u^* \in A$, $v^* \in B$, and assume

$$|F(u^*, v_t) - F(u^*, v^*)| \leq D_n \cdot \|u^*\| \cdot \|v_t - v^*\| \tag{22}$$

for all $t = 1, \ldots, t_n$. Then it holds:

$$\min_{t=0,\ldots,t_n-1} F(u_t, v_t) \leq F(u^*, v^*) + D_n \cdot \|u^*\| \cdot diam(B) + \frac{\|u^* - u_0\|^2}{2} + \frac{C_n^2}{2 \cdot t_n}.$$

Proof. The result follows in a straightforward way from the proof of Lemma 1 in Kohler and Krzyżak (2023). For the sake of completeness we repeat the proof below.

In the *first step of the proof* we show

$$\frac{1}{t_n}\sum_{t=0}^{t_n-1} F(u_t,v_t) \le \frac{1}{t_n}\sum_{t=0}^{t_n-1} F(u^*,v_t) + \frac{\|u^*-u_0\|^2}{2} + \frac{1}{2\cdot t_n^2}\sum_{t=0}^{t_n-1}\|(\nabla_u F)(u_t,v_t)\|^2. \quad (23)$$

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

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

This implies

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

In the *second step of the proof* we show the assertion.

Using the result of step 1 we get

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

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

By (22) we get

$$\begin{aligned} \frac{1}{t_n} \sum_{t=0}^{t_n-1} |F(u^*, v_t) - F(u^*, v^*)| &\leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} D_n \cdot \|u^*\| \cdot \|v_t - v^*\| \\ &\leq D_n \cdot \|u^*\| \cdot diam(B). \end{aligned}$$

And by (21) we get

$$\frac{1}{2 \cdot t_n^2} \sum_{t=0}^{t_n-1} \|(\nabla_u F)(u_t, v_t)\|^2 \leq \frac{1}{2 \cdot t_n^2} \sum_{t=0}^{t_n-1} C_n^2 = \frac{C_n^2}{2 \cdot t_n}.$$

Summarizing the above results, the proof is complete. □

Next we prove two results which will help us to verify the assumptions of Lemma 1. First we consider (22).

Lemma 2 *Let $d, J_n, K_n \in \mathbb{N}$, and for $\mathbf{w} = ((w_k)_{k=1,\ldots,K_n}, \mathbf{v})$ with $w_k \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^{J_n}$ let $f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ be a (deep) neural network with weight vector $\mathbf{w}$ given by*

$$f_{\mathbf{w}}(x) = \sum_{k=1}^{K_n} w_k \cdot f_{\mathbf{v},k}^{(L)}(x) \quad (x \in \mathbb{R}^n),$$

where $|f_{\mathbf{v},k}^{(L)}(x)| \leq 1$ for all $x \in \mathbb{R}^d$ $(k = 1, \ldots, K_n)$. Set $u = (w_k)_{k=1,\ldots,K_n}$, $f_{u,v}(x) = f_{\mathbf{w}}(x)$ and

$$F(u, v) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{u,v}(X_i)|^2.$$

Let $\beta_n \geq 0$, $C_n, E_n, \tilde{K}_n \in \mathbb{N}$ and assume

$$|Y_i| \leq \beta_n \quad (i = 1, \ldots, n), \tag{24}$$

$$\sum_{k=1}^{K_n} |w_k| \leq E_n, \tag{25}$$

$$|\{k \in \{1, \ldots, K_n\} \quad : \quad w_k \neq 0\}| \leq \tilde{K}_n \tag{26}$$

and

$$|f_{\mathbf{v},k}^{(L)}(x) - f_{\mathbf{v}^{(0)},k}^{(L)}(x)| \leq C_n \cdot \|\mathbf{v} - \mathbf{v}^{(0)}\| \tag{27}$$

for all $x \in \{X_1, \ldots, X_n\}$. Then

$$|F(u, v) - F(u, v^{(0)})| \leq 2 \cdot (\beta_n + E_n) \cdot C_n \cdot \sqrt{\tilde{K}_n} \cdot \|u\| \cdot \|\mathbf{v} - \mathbf{v}^{(0)}\|.$$

Proof. Because of (25) we have

$$|f_{u,v}(X_i)| \leq E_n \quad \text{and} \quad |f_{u,v^{(0)}}(X_i)| \leq E_n.$$

This implies

$$\begin{aligned}
&|F(u,v) - F(u,v^{(0)})| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_{u,v}(X_i) + Y_i - f_{u,v^{(0)}}(X_i)) \cdot (f_{u,v}(X_i) - f_{u,v^{(0)}}(X_i)) \right| \\
&\leq (2 \cdot \beta_n + 2 \cdot E_n) \cdot \frac{1}{n} \sum_{i=1}^{n} |f_{u,v}(X_i) - f_{u,v^{(0)}}(X_i)| \\
&\leq (2 \cdot \beta_n + 2 \cdot E_n) \cdot \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K_n} |w_k| \cdot |f^{(L)}_{\mathbf{v},k}(X_i) - f^{(L)}_{(\mathbf{v}^{(0)}),k}(X_i)| \\
&\leq (2 \cdot \beta_n + 2 \cdot E_n) \cdot \frac{1}{n} \sum_{i=1}^{n} \sqrt{\sum_{k=1}^{K_n} |w_k|^2} \cdot \sqrt{\sum_{k=1,\ldots,K_n : w_k \neq 0} |f^{(L)}_{\mathbf{v},k}(X_i) - f^{(L)}_{\mathbf{v}^{(0)},k}(X_i)|^2} \\
&\leq (2 \cdot \beta_n + 2 \cdot E_n) \cdot \frac{1}{n} \sum_{i=1}^{n} \sqrt{\sum_{k=1}^{K_n} |w_k|^2} \cdot \sqrt{\sum_{k=1,\ldots,K_n : w_k \neq 0} C_n^2 \cdot \|\mathbf{v} - \mathbf{v}^{(0)}\|^2} \\
&= 2 \cdot (\beta_n + E_n) \cdot C_n \cdot \|u\| \cdot \sqrt{\tilde{K}_n} \cdot \|\mathbf{v} - \mathbf{v}^{(0)}\|.
\end{aligned}$$

□

Lemma 2 requires (27), for which we will use our next lemma.

Lemma 3 *Define*

$$f_{\mathbf{w}}(x) = h_1^{(l)}(x) = \sum_{k=1}^{K_n^{(l)}} w_{1,k}^{(L)} \cdot f_k^{(L)}(x)$$

by (11)–(16) and set $f^{(L)}_{\mathbf{v},k}(x) = f_k^{(L)}(x)$. *Let* $\alpha, A_n, B_n, E_n \geq 1$. *Assume that the weight vectors* $\mathbf{w}$ *of* $g_{NN,i,s}$ *defined in (12) and (13) satisfy*

$$|w_{k,r,j}^{(0)}| \leq A_n \quad \textit{for } j > 0, \tag{28}$$

$$|w_{k,r,j}^{(t)}| \leq B_n \quad \textit{for } j > 0 \quad \textit{and} \quad t = 1, \ldots, L_s - 1, \tag{29}$$

and

$$\sum_{k=1}^{K_n^{(s)}} |w_{k,1,1}^{(L_s)}| \leq E_n \quad \textit{in case } s < l. \tag{30}$$

Then we have for any $x \in [-\alpha, \alpha]^d$

$$|f^{(L)}_{\mathbf{v}_1,k}(x) - f^{(L)}_{\mathbf{v}_2,k}(x)| \leq c_{16} \cdot \sqrt{\max_{s=0,\ldots,l-1} K_n^{(s)}} \cdot A_n^l \cdot B_n^{L-2l-1} \cdot E_n^l \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|.$$

Proof. For $s \notin \{L_0+1, L_0+L_1+2, \ldots, L_0+\cdots+L_{l-1}+l\}$ we have

$$|f_i^{(s)}(x)| = \left|\sigma\left(\sum_{j\in\{0,\ldots,k_{s-1}\}\,:\,(s-1,i,j)\in I} w_{i,j}^{(s-1)} \cdot f_j^{(s-1)}(x)\right)\right| \leq 1.$$

For $s \in \{L_0+1, L_0+L_1+2, \ldots, L_0+\cdots+L_{l-1}+l\}$ we can conclude from (30)

$$\begin{aligned} |f_i^{(s)}(x)| &= \left|\sum_{j\in\{0,\ldots,k_{s-1}\}\,:\,(s-1,i,j)\in I} w_{i,j}^{(s-1)} \cdot f_j^{(s-1)}(x)\right| \\ &\leq \sum_{j\in\{0,\ldots,k_{s-1}\}\,:\,(s-1,i,j)\in I} |w_{i,j}^{(s-1)}| \leq E_n. \end{aligned}$$

Using that σ_s is Lipschitz continuous with Lipschitz constant one we get

$$\begin{aligned} |f_{\mathbf{v}_1,i}^{(s)}(x) - f_{\mathbf{v}_2,i}^{(s)}(x)| &\leq \Bigg|\sum_{j\in\{0,\ldots,k_{s-1}\}\,:\,(s-1,i,j)\in I} (\mathbf{v}_1)_{i,j}^{(s-1)} \cdot f_{\mathbf{v}_1,j}^{(s-1)}(x) \\ &\qquad - \sum_{j\in\{0,\ldots,k_{s-1}\}\,:\,(s-1,i,j)\in I} (\mathbf{v}_2)_{i,j}^{(s-1)} \cdot f_{\mathbf{v}_2,j}^{(s-1)}(x)\Bigg| \\ &\leq \left|\sum_{j\in\{0,\ldots,k_{s-1}\}\,:\,(s-1,i,j)\in I} ((\mathbf{v}_1)_{i,j}^{(s-1)} - (\mathbf{v}_2)_{i,j}^{(s-1)}) \cdot f_{\mathbf{v}_1,j}^{(s-1)}(x)\right| \\ &\quad + \left|\sum_{j\in\{0,\ldots,k_{s-1}\}\,:\,(s-1,i,j)\in I} (\mathbf{v}_2)_{i,j}^{(s-1)} \cdot (f_{\mathbf{v}_1,j}^{(s-1)}(x) - f_{\mathbf{v}_2,j}^{(s-1)}(x))\right| \\ &\leq \sum_{j\in\{0,\ldots,k_{s-1}\}\,:\,(s-1,i,j)\in I} \left|(\mathbf{v}_1)_{i,j}^{(s-1)} - (\mathbf{v}_2)_{i,j}^{(s-1)}\right| \cdot \max_k |f_{\mathbf{v}_1,k}^{(s-1)}(x)| \\ &\quad + \sum_{j\in\{0,\ldots,k_{s-1}\}\,:\,(s-1,i,j)\in I} \left|(\mathbf{v}_2)_{i,j}^{(s-1)}\right| \cdot \max_k |f_{\mathbf{v}_1,k}^{(s-1)}(x) - f_{\mathbf{v}_2,k}^{(s-1)}(x)| \\ &\leq \sqrt{|\{j \in \{0,\ldots,k_{s-1}\} \,:\, (s-1,i,j)\in I\}|} \\ &\quad \cdot \sqrt{\sum_{j\in\{0,\ldots,k_{s-1}\}\,:\,(s-1,i,j)\in I} \left|(\mathbf{v}_1)_{i,j}^{(s-1)} - (\mathbf{v}_2)_{i,j}^{(s-1)}\right|^2} \cdot \max_k |f_{\mathbf{v}_1,k}^{(s-1)}(x)| \\ &\quad + \sum_{j\in\{0,\ldots,k_{s-1}\}\,:\,(s-1,i,j)\in I} \left|(\mathbf{v}_2)_{i,j}^{(s-1)}\right| \cdot \max_k |f_{\mathbf{v}_1,k}^{(s-1)}(x) - f_{\mathbf{v}_2,k}^{(s-1)}(x)|. \end{aligned}$$

In case $s = 1$ we have

$$\max_k |f_{\mathbf{v}_1,k}^{(s-1)}(x)| \leq \alpha$$

and

$$\max_k |f_{\mathbf{v}_1,k}^{(s-1)}(x) - f_{\mathbf{v}_2,k}^{(s-1)}(x)| = 0,$$

from which we can conclude

$$|f^{(1)}_{\mathbf{v}_1,i}(x) - f^{(1)}_{\mathbf{v}_2,i}(x)| \leq \sqrt{d+1} \cdot \alpha \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|.$$

In case $s \in \{2, 3, \ldots, L_0, L_0+3, L_0+4, \ldots, L_0+L_1+1, \ldots, L_0+\cdots+L_{l-1}+l+2, \ldots, L_0+\cdots+L_l+l\}$ we have

$$|\{j \in \{0, \ldots, k_{s-1}\} : (s-1, i, j) \in I\}| \leq r+1, \tag{31}$$

$$\max_k |f^{(s-1)}_{\mathbf{v}_1,k}(x)| \leq 1 \tag{32}$$

and

$$\sum_{j \in \{0,\ldots,k_{s-1}\} : (s-1,i,j) \in I} \left|(\mathbf{v}_2)^{(s-1)}_{i,j}\right| \leq (r+1) \cdot B_n, \tag{33}$$

from which we can conclude

$$|f^{(s)}_{\mathbf{v}_1,i}(x) - f^{(s)}_{\mathbf{v}_2,i}(x)| \leq \sqrt{r+1} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\| + (r+1) \cdot B_n \cdot \max_k |f^{(s-1)}_{\mathbf{v}_1,k}(x) - f^{(s-1)}_{\mathbf{v}_2,k}(x)|.$$

In case $s \in \{L_0+1, L_0+L_1+2, \ldots, L_0+\cdots+L_{l-1}+l\}$ (32) holds, and in addition we have

$$|\{j \in \{0, \ldots, k_{s-1}\} : (s-1, i, j) \in I\}| \leq \max_{s=0,\ldots,l-1} K^{(s)}_n$$

and

$$\sum_{j \in \{0,\ldots,k_{s-1}\} : (s-1,i,j) \in I} \left|(\mathbf{v}_2)^{(s-1)}_{i,j}\right| \leq E_n,$$

from which we can conclude

$$|f^{(s)}_{\mathbf{v}_1,i}(x) - f^{(s)}_{\mathbf{v}_2,i}(x)| \leq \sqrt{\max_{s=0,\ldots,l-1} K^{(s)}_n} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\| + E_n \cdot \max_k |f^{(s-1)}_{\mathbf{v}_1,k}(x) - f^{(s-1)}_{\mathbf{v}_2,k}(x)|.$$

In case $s \in \{L_0+2, L_0+L_1+3, \ldots, L_0+\ldots, L_{l-1}+l+1\}$ we have

$$\{j \in \{0, \ldots, k_{s-1}\} : (s-1, i, j) \in I\}| \leq K_{i,s}+1,$$

$$\max_k |f^{(s-1)}_{\mathbf{v}_1,k}(x)| \leq E_n$$

and

$$\sum_{j \in \{0,\ldots,k_{s-1}\} : (s-1,i,j) \in I} \left|(\mathbf{v}_2)^{(s-1)}_{i,j}\right| \leq (K_{i,s}+1) \cdot A_n,$$

from which we can conclude

$$\begin{aligned}
&|f^{(s)}_{\mathbf{v}_1,i}(x) - f^{(s)}_{\mathbf{v}_2,i}(x)| \\
&\leq \sqrt{K_{i,s}+1} \cdot E_n \cdot \|\mathbf{v}_1 - \mathbf{v}_2\| + (K_{i,s}+1) \cdot A_n \cdot \max_k |f^{(s-1)}_{\mathbf{v}_1,k}(x) - f^{(s-1)}_{\mathbf{v}_2,k}(x)|.
\end{aligned}$$

Applying these inequalities recursively we get

$$|f^{(1)}_{\mathbf{v}_1,i}(x) - f^{(1)}_{\mathbf{v}_2,i}(x)| \le c_{17} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|,$$

$$|f^{(2)}_{\mathbf{v}_1,i}(x) - f^{(2)}_{\mathbf{v}_2,i}(x)| \le c_{18} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\| + c_{19} \cdot B_n \cdot c_{17} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\| \le c_{20} \cdot B_n \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|,$$

$$|f^{(L_0)}_{\mathbf{v}_1,i}(x) - f^{(L_0)}_{\mathbf{v}_2,i}(x)| \le c_{21} \cdot B_n^{L_0-1} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|,$$

$$\begin{aligned}
&|f^{(L_0+1)}_{\mathbf{v}_1,i}(x) - f^{(L_0+1)}_{\mathbf{v}_2,i}(x)| \\
&\le \sqrt{\max_{s=0,\dots,l-1} K_n^{(s)}} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\| + E_n \cdot c_{21} \cdot B_n^{L_0-1} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\| \\
&\le c_{22} \cdot \sqrt{\max_{s=0,\dots,l-1} K_n^{(s)}} \cdot E_n \cdot B_n^{L_0-1} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|,
\end{aligned}$$

$$\begin{aligned}
&|f^{(L_0+2)}_{\mathbf{v}_1,i}(x) - f^{(L_0+2)}_{\mathbf{v}_2,i}(x)| \\
&\le c_{23} \cdot E_n \cdot \|\mathbf{v}_1 - \mathbf{v}_2\| + c_{24} \cdot A_n \cdot c_{22} \cdot \sqrt{\max_{s=0,\dots,l-1} K_n^{(s)}} \cdot E_n \cdot B_n^{L_0-1} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\| \\
&\le c_{25} \cdot \sqrt{\max_{s=0,\dots,l-1} K_n^{(s)}} \cdot A_n \cdot E_n \cdot B_n^{L_0-1} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|,
\end{aligned}$$

$$\begin{aligned}
&|f^{(L_0+L_1+1)}_{\mathbf{v}_1,i}(x) - f^{(L_0+L_1+1)}_{\mathbf{v}_2,i}(x)| \\
&\le c_{26} \cdot \sqrt{\max_{s=0,\dots,l-1} K_n^{(s)}} \cdot A_n \cdot E_n \cdot B_n^{L_0+L_1-2} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|,
\end{aligned}$$

$$\begin{aligned}
&|f^{(L_0+L_1+2)}_{\mathbf{v}_1,i}(x) - f^{(L_0+L_1+2)}_{\mathbf{v}_2,i}(x)| \\
&\le c_{27} \cdot \sqrt{\max_{s=0,\dots,l-1} K_n^{(s)}} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\| \\
&\quad + E_n \cdot c_{26} \cdot \sqrt{\max_{s=0,\dots,l-1} K_n^{(s)}} \cdot A_n \cdot E_n \cdot B_n^{L_0+L_1-2} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\| \\
&\le c_{28} \cdot \sqrt{\max_{s=0,\dots,l-1} K_n^{(s)}} \cdot A_n \cdot E_n^2 \cdot B_n^{L_0+L_1-2} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|
\end{aligned}$$

and finally

$$\begin{aligned}
|f^{(L)}_{\mathbf{v}_1,i}(x) - f^{(L)}_{\mathbf{v}_2,i}(x)| &= |f^{(L_0+\dots+L_l+l)}_{\mathbf{v}_1,i}(x) - f^{(L_0+\dots+L_l+l)}_{\mathbf{v}_2,i}(x)| \\
&\le c_{29} \cdot \sqrt{\max_{s=0,\dots,l-1} K_n^{(s)}} \cdot A_n^l \cdot E_n^l \cdot B_n^{L_0+\dots+L_l-l-1} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|.
\end{aligned}$$

□

Finally we present a result which will help us to verify (21).

Lemma 4 *Let $F(u,v) = F(\mathbf{w})$ be defined as in Lemma 2 and assume that (24) and (25) hold. Then*

$$\left\| \nabla_{(w_k)_{k=1,\dots,K_n}} F(\mathbf{w}) \right\| \leq 2 \cdot (\beta_n + E_n) \cdot \sqrt{K_n}.$$

Proof. We have

$$\begin{aligned}
&\left\| \nabla_{(w_k)_{k=1,\dots,K_n}} F(\mathbf{w}) \right\|^2 \\
&= \sum_{k=1}^{K_n} \left| \frac{1}{n} \sum_{i=1}^{n} 2 \cdot (f_{\mathbf{w}}(X_i) - Y_i) \cdot f_{\mathbf{v},k}^{(L)}(X_i) \right|^2 \\
&\leq \sum_{k=1}^{K_n} \left| \frac{1}{n} \sum_{i=1}^{n} 2 \cdot (E_n + \beta_n) \cdot 1 \right|^2 \\
&= 4 \cdot K_n \cdot (E_n + \beta_n)^2.
\end{aligned}$$

□

4.2 Neural network approximation

In this subsection we study the approximation properties of our hierarchical space of deep neural networks. Our starting point is the following result from Kohler (2024).

Lemma 5 *Let $d \in \mathbb{N}$, $p = q + \beta$ where $\beta \in (0,1]$ and $q \in \mathbb{N}_0$, $C > 0$, $A \geq 1$ and $A_n, B_n, \gamma_n^* \geq 1$. For $L, r \in \mathbb{N}$ let $\mathcal{F}$ be the set of all networks $f_{\mathbf{w}}$ defined by (8)–(10) with logistic squasher and K replaced by r, where the weight vector satisfies*

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

for all $l \in \{1, \dots, L-1\}$, all i, j and all $k = 1, \dots, r$, and for $L, r, K \in \mathbb{N}$ set

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

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

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

and set

$$A_n = A \cdot K \cdot \log K, \quad B_n = c_{30} \quad \text{and} \quad \gamma_n^* = c_{31} \cdot K^{q+d}.$$

Assume $K \geq c_{32}$ for c_{32} sufficiently large. Then there exists for any (p,C)–smooth $f : \mathbb{R}^d \to \mathbb{R}$ a neural network $h \in \mathcal{H}$ such that

$$\sup_{x \in [-A,A)^d} |f(x) - h(x)| \leq \frac{c_{33}}{K^p}.$$

Proof. See Theorem 3 in Kohler (2024). □

Our next result extends Lemma 5 to the hierarchical spaces of neural networks introduced in Section 2.

Lemma 6 *Let $K_{s,r} \in \mathbb{N}$ with $K_{0,r} = d$. Set $N_l = 1$ and $N_s = \sum_{r=1}^{N_{s+1}} K_{r,s}$ for $s \in \{0, \dots, l-1\}$. Define $m : \mathbb{R}^d \to \mathbb{R}$ by*

$$m(x) = h_1^{(l)}(x),$$

where

$$h_i^{(s)}(x) = g_{i,s}(h^{(s-1)}_{\sum_{r=1}^{i-1} K_{s-1,r}+1}(x), \dots, h^{(s-1)}_{\sum_{r=1}^{i-1} K_{s-1,r}+K_{s-1,i}}(x))$$

for $s \in \{1, \dots, L\}$, $i \in \{1, \dots, N_s\}$,

$$h_i^{(0)}(x) = g_{i,0}(x)$$

for $i \in \{1, \dots, N_0\}$. Assume that

$$g_{i,s} : \mathbb{R}^{K_{i,s}} \to \mathbb{R}$$

are $(p_{i,s}, C_{i,s})$-smooth for some $p_{i,s} \geq 1$, $C_{i,s} > 0$ for all i, s.

Let $\mathcal{H}$ be the set of all neural networks defined by (11)-(13) with $K_n^{(s)}$ replaced by $K_n^{(s,i)} \in \mathbb{N}$, $L_s = \bar{L} = \max_{r,t} \lceil \log_2(p_{r,t} + K_{r,t}) \rceil$, $r_{i,s} = 2 \cdot (\lceil (2p_{i,s} + K_{i,s}) \rceil)^2$ and where for all i, s and each function from $g_{NN,i,s} \in \mathcal{F}_{K_{i,s}, K_n^{(s,i)}, L_s, r_{i,s}}$ the weight constraints

$$|w_{k,i,j}^{(0)}| \leq c_{34} \cdot (K_n^{(s,i)})^{1/K_{i,s}} \cdot \log K_n^{(s,i)}, \quad |w_{k,i,j}^{(l)}| \leq c_{35} \quad \text{and} \quad |w_{k,1,1}^{(L_s)}| \leq c_{36} \cdot (K_n^{(s,i)})^{\frac{p_{i,s}+K_{i,s}}{K_{i,s}}}$$

are satisfied for $l = 1, \dots, L_s - 1$.

Then there exists $h \in \mathcal{H}$ such that for sufficiently large $K_n^{(s,i)}$ it holds

$$\sup_{x \in [-A,A]^d} |h(x) - m(x)| \leq c_{37} \cdot \max_{i,s} \frac{1}{(K_n^{(s,i)})^{p_{i,s}/K_{i,s}}}.$$

Proof. Choose $\bar{A} \geq A$ such that

$$h_{i,s}(x) \in [-\bar{A}, \bar{A}]$$

holds for all $x \in [-A, A]^d$ and all i, s, which is possible because of the continuity of the $g_{i,s}$. For i, s let $\hat{g}_{i,s} \in \mathcal{F}_{K_{i,s}, K_n^{(s,i)}, L_s, r_{i,s}}$ be the neural network approximation of $g_{i,s}$ defined in Lemma 5 which satisfies

$$\sup_{x \in [-\bar{A}-1, \bar{A}+1)^{K_{i,s}}} |\hat{g}_{i,s}(x) - g_{i,s}(x)| \leq \frac{c_{38}}{(K_n^{(s,i)})^{p_{i,s}/K_{i,s}}} \tag{34}$$

for some $c_{38} \geq 1$. Let $C_{Lip} \geq 1$ be an upper bound on the Lipschitz constant of $g_{i,s}$ on $[-\bar{A}, \bar{A}]^{K_{i,s}}$ for all i, s (which exists because of $p_{i,s} \geq 1$). W.l.o.g. we assume in the sequel that $K_n^{(s,i)}$ is so large that

$$\frac{2^{l-1} \cdot C_{Lip}^{l-1} \cdot c_{38}}{(K_n^{(s,i)})^{p_{i,s}/K_{i,s}}} < 1$$

holds for all i, s.

Define

$$\hat{h}_1^{(l)}$$

recursively by

$$\hat{h}_i^{(s)}(x) = \hat{g}_{i,s}(\hat{h}^{(s-1)}_{\sum_{r=1}^{i-1} K_{s-1,r}+1}(x), \ldots, \hat{h}^{(s-1)}_{\sum_{r=1}^{i-1} K_{s-1,r}+K_{s-1,i}}(x))$$

for $s \in \{1, \ldots, L\}$, $i \in \{1, \ldots, N_s\}$ and

$$\hat{h}_i^{(0)}(x) = \hat{g}_{i,0}(x)$$

for $i \in \{1, \ldots, N_0\}$.

Then $\hat{h}_1^{(l)}$ is a function from $\mathcal{H}$ which satisfies the conditions of Lemma 6, hence it suffices to show

$$\sup_{x \in [-A,A]^d} |\hat{h}_1^{(l)}(x) - m(x)| \leq c_{39} \cdot \max_{i,s} \frac{1}{(K_n^{(s,i)})^{p_{i,s}/K_{i,s}}}.$$

To do this, we show recursively

$$\sup_{x \in [-A,A]^d} |\hat{h}_j^{(s)}(x) - h_j^{(s)}(x)| \leq 2^s \cdot C_{Lip}^s \cdot c_{38} \cdot \max_{i,t:t \leq s} \frac{1}{(K_n^{(t,i)})^{p_{i,t}/K_{i,t}}} \tag{35}$$

for all $j \in \{1, \ldots, N_s\}$ and all $s \in \{0, 1, \ldots, l\}$.

By construction (35) holds for $s = 0$ (cf., (34)). So assume now that (35) holds for some $s \in \{0, \ldots, l-1\}$. Then $\hat{h}_i^{(s)}(x) \in [-\bar{A}-1, \bar{A}+1)$, the choice of $\hat{g}_{i,s+1}$, the Lipschitz smoothness of $g_{i,s+1}$ (which holds because of $p_{i,s+1} \geq 1$) and our induction assumption imply for any $x \in [-A, A]^d$ and any i

$$\begin{aligned}
&|\hat{h}_i^{(s+1)}(x) - h_i^{(s+1)}(x)| \\
&= \Big| \hat{g}_{i,s+1}(\hat{h}^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+1}(x), \ldots, \hat{h}^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+K_{s,i}}(x)) \\
&\qquad\qquad - g_{i,s+1}(h^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+1}(x), \ldots, h^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+K_{s,i}}(x)) \Big| \\
&\leq \Big| \hat{g}_{i,s+1}(\hat{h}^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+1}(x), \ldots, \hat{h}^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+K_{s,i}}(x))
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad -g_{i,s+1}(\hat{h}^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+1}(x),\dots,\hat{h}^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+K_{s,i}}(x))\Big| \\
&+\Big|g_{i,s+1}(\hat{h}^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+1}(x),\dots,\hat{h}^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+K_{s,i}}(x)) \\
&\qquad\qquad -g_{i,s+1}(h^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+1}(x),\dots,h^{(s)}_{\sum_{r=1}^{i-1} K_{s,r}+K_{s,i}}(x))\Big| \\
&\leq c_{38}\cdot\frac{1}{(K_n^{(s+1,i)})^{p_{i,s+1}/K_{i,s+1}}}+C_{Lip}\cdot\max_{k=\sum_{r=1}^{i-1}K_{s,r}+1,\dots,\sum_{r=1}^{i-1}K_{s,r}+K_{s,i}}|\hat{h}_k^{(s)}(x)-h_k^{(s)}(x)| \\
&\leq c_{38}\cdot\frac{1}{(K_n^{(s+1,i)})^{p_{i,s+1}/K_{i,s+1}}}+C_{Lip}\cdot 2^s\cdot C_{Lip}^s\cdot c_{38}\cdot\max_{i,t:t\leq s}\frac{1}{(K_n^{(t,i)})^{p_{i,t}/K_{i,t}}} \\
&\leq 2^{s+1}\cdot C_{Lip}^{s+1}\cdot c_{38}\cdot\max_{i,t:t\leq s+1}\frac{1}{(K_n^{(t,i)})^{p_{i,t}/K_{i,t}}}.
\end{aligned}
$$

□

4.3 Neural network generalization

Next we derive a bound on the covering number of hierarchically defined spaces of neural networks. We do this by composing coverings, and in order to be able to prove that the composed covering is a covering of the hierarchically defined function space, we will use the following generalization of a supremum norm cover.

Definition 4 *Let $\epsilon > 0$ and $\delta \geq 0$, let $A \in \mathbb{R}_+$, let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$, and let $f_1, \dots, f_n : \mathbb{R}^d \to \mathbb{R}$. $\{f_1, \dots, f_n\}$ is called an ϵ-$\|\cdot\|_{\infty,[-A,A]^d}$-**cover of $\mathcal{F}$ for δ perturbated data**, if for any $f \in \mathcal{F}$ there exists $i \in \{1, \dots, n\}$ such that*

$$\sup_{x\in[-A,A]^d,\tilde{x}\in\mathbb{R}^d:\|x-\tilde{x}\|_\infty\leq\delta}|f(x)-f_i(\tilde{x})|<\epsilon.$$

*The ϵ-$\|\cdot\|_{\infty,[-A,A]^d}$-**covering number of $\mathcal{F}$ for δ perturbated data***

$$\mathcal{N}_{\|\cdot\|_{\infty,[-A,A]^d},\delta}(\epsilon,\mathcal{F})$$

is the minimal $n \in \mathbb{N}$ such that a ϵ-$\|\cdot\|_{\infty,[-A,A]^d}$-cover of $\mathcal{F}$ for δ perturbated data of size n exists.

As our next result shows, the above introduced covering number is especially suited for hierarchical compositions of function spaces.

Lemma 7 *Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^K \to \mathbb{R}$, let $\alpha \geq 1$, let $\mathcal{G}_1, \dots, \mathcal{G}_K$ be sets of functions $g : \mathbb{R}^d \to [-\alpha, \alpha]$, and let $\mathcal{H}$ be the set of all functions*

$$h(x) = f(g_1(x), \dots, g_K(x)) \quad (x \in \mathbb{R}^d)$$

for some $f \in \mathcal{F}$, $g_1 \in \mathcal{G}_1$, $\ldots$, $g_K \in \mathcal{G}_K$. *Then we have for any* $\epsilon, \eta > 0$ *and any* $\delta > 0$

$$\mathcal{N}_{\|\cdot\|_{\infty,[-A,A]^d},\delta}(\epsilon, \mathcal{H}) \leq \mathcal{N}_{\|\cdot\|_{\infty,[-\alpha,\alpha]^K},\eta}(\epsilon, \mathcal{F}) \cdot \prod_{k=1}^{K} \mathcal{N}_{\|\cdot\|_{\infty,[-A,A]^d},\delta}(\eta, \mathcal{G}_k).$$

Proof. Let $f_1, \ldots, f_n : \mathbb{R}^K \to \mathbb{R}$ be an ϵ-$\|\cdot\|_{\infty,[-\alpha,\alpha]^K}$-cover of minimal size of $\mathcal{F}$ for η perturbated data, and for $k \in \{1, \ldots, K\}$ let $g_{k,1}, \ldots, g_{k,n_k} : \mathbb{R}^d \to \mathbb{R}$ be an η-$\|\cdot\|_{\infty,[-A,A]^d}$-cover of minimal size of $\mathcal{G}_k$ for δ perturbated data. In the sequel we show that the set of all functions

$$h(x) = f_i(g_{1,j_1}(x), \ldots, g_{K,j_K}(x)) \quad (x \in \mathbb{R}^d)$$

with $i \in \{1, \ldots, n\}$, $j_1 \in \{1, \ldots, n_1\}$, $\ldots$, $j_K \in \{1, \ldots, n_K\}$ is an ϵ-$\|\cdot\|_{\infty,[-A,A]^d}$-cover of $\mathcal{H}$ for δ perturbated data. From this we get the assertion, because from this we can conclude

$$\mathcal{N}_{\|\cdot\|_{\infty,[-A,A]^d},\delta}(\epsilon, \mathcal{H}) \leq n \cdot \prod_{k=1}^{K} n_k = \mathcal{N}_{\|\cdot\|_{\infty,[-\alpha,\alpha]^K},\eta}(\epsilon, \mathcal{F}) \cdot \prod_{k=1}^{K} \mathcal{N}_{\|\cdot\|_{\infty,[-A,A]^d},\delta}(\eta, \mathcal{G}_k).$$

In order to show that the functions h defined above are an ϵ-$\|\cdot\|_{\infty,[-A,A]^d}$-cover of $\mathcal{H}$ for δ perturbated data, let $h \in \mathcal{H}$ be arbitrary. Then h is given by

$$h(x) = f(g_1(x), \ldots, g_K(x)) \quad (x \in \mathbb{R}^d)$$

for some $f \in \mathcal{F}$, $g_1 \in \mathcal{G}_1$, $\ldots$, $g_K \in \mathcal{G}_K$. Choose $i \in \{1, \ldots, n\}$, $j_1 \in \{1, \ldots, n_1\}$, $\ldots$, $j_K \in \{1, \ldots, n_K\}$ such that

$$\sup_{x \in [-\alpha,\alpha]^K, \tilde{x} \in \mathbb{R}^K : \|x - \tilde{x}\|_\infty \leq \eta} |f(x) - f_i(\tilde{x})| < \epsilon$$

and

$$\sup_{x \in [-A,A]^d, \tilde{x} \in \mathbb{R}^d : \|x - \tilde{x}\|_\infty \leq \delta} |g_k(x) - g_{k,j_k}(\tilde{x})| < \eta$$

hold for all $k \in \{1, \ldots, K\}$.

Then we get for any $x \in [-A, A]^d$ and $\tilde{x} \in \mathbb{R}^d$ with $\|x - \tilde{x}\|_\infty \leq \delta$

$$g_k(x) \in [-\alpha, \alpha] \quad \text{and} \quad |g_k(x) - g_{k,j_k}(\tilde{x})| < \eta \quad \text{for all } k = 1, \ldots, K$$

(where the first relation holds by the definition of $\mathcal{G}_k$), from which we conclude

$$|f(g_1(x), \ldots, g_K(x)) - f_i(g_{1,j_1}(\tilde{x}), \ldots, g_{K,j_K}(\tilde{x}))| < \epsilon.$$

□

Our next lemma generalizes the result from Lemma 12 in Kohler (2024) to the more complex notion of a covering introduced above.

Lemma 8 *Let $\alpha \geq 1$ and let $A, B, C \geq 1$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be k-times differentiable such that all derivatives up to order k are bounded on $\mathbb{R}$. Let $\mathcal{F}$ be the set of all functions $f_{\mathbf{w}}$ defined by (8)-(10) where the weight vector $\mathbf{w}$ satisfies*

$$\sum_{j=1}^{K} |w_{j,1,1}^{(L)}| \leq C, \tag{36}$$

$$|w_{k,i,j}^{(l)}| \leq B \quad (k \in \{1, \ldots, K\}, i, j \in \{1, \ldots, r\}, l \in \{1, \ldots, L-1\}) \tag{37}$$

and

$$|w_{k,i,j}^{(0)}| \leq A \quad (k \in \{1, \ldots, K\}, i \in \{1, \ldots, r\}, j \in \{1, \ldots, d\}). \tag{38}$$

Let $\epsilon, \delta \in (0, 1]$ and assume

$$\delta \leq \frac{c_{43} \cdot \epsilon}{d \cdot A \cdot B^{L-1} \cdot C} \tag{39}$$

for some suitably small constant $c_{43} > 0$. Then we have

$$\begin{aligned} &\mathcal{N}_{\|\cdot\|_{\infty,[-\alpha,\alpha]^d},\delta}(\epsilon, \mathcal{F}) \\ &\leq \left(c_{44} \cdot \frac{A^{k-1} \cdot B^{(L-1)\cdot(k-1)} \cdot C}{\epsilon} \right)^{c_{45} \cdot \alpha^d \cdot A^d \cdot B^{(L-1)\cdot d} \cdot \left(\frac{C}{\epsilon}\right)^{d/k}}. \end{aligned}$$

Proof. It is shown in the proof of Lemma 12 in Kohler (2024) that for any $f_{\mathbf{w}} \in \mathcal{F}$, any $x \in [-2\alpha, 2\alpha]^d$ and any $s_1, \ldots, s_k \in \{1, \ldots, d\}$

$$\left| \frac{\partial^k f_{\mathbf{w}}}{\partial x^{(s_1)} \ldots \partial x^{(s_k)}}(x) \right| \leq c_{46} \cdot C \cdot B^{(L-1)\cdot k} \cdot A^k \tag{40}$$

holds.

Partition $[-2 \cdot \alpha, 2 \cdot \alpha]^d$ into at most

$$K = \left\lceil \left(c_{47} \cdot \frac{2 \cdot (4\alpha)^k \cdot A^k \cdot B^{(L-1)\cdot k} \cdot C}{\epsilon} \right)^{d/k} \right\rceil$$

many cubes of side length

$$\begin{aligned} \eta &= \frac{4 \cdot \alpha}{\lfloor K^{1/d} \rfloor} \leq \frac{4\alpha}{K^{1/d} - 1} \\ &\leq \frac{4\alpha}{\left(c_{47} \cdot \frac{2 \cdot (4\alpha)^k \cdot A^k \cdot B^{(L-1)\cdot k} \cdot C}{\epsilon} \right)^{1/k} - 1} \\ &\leq \frac{4\alpha}{\frac{1}{2} \cdot \left(c_{47} \cdot \frac{2 \cdot (4\alpha)^k \cdot A^k \cdot B^{(L-1)\cdot k} \cdot C}{\epsilon} \right)^{1/k}} \\ &\leq \left(\frac{1}{c_{47}/2^{k+1}} \cdot \frac{\epsilon/4}{A^k \cdot B^{(L-1)\cdot k} \cdot C} \right)^{1/k}. \end{aligned} \tag{41}$$

Here the second inequality holds if $c_{47} \geq 1$, since

$$\frac{2 \cdot (4\alpha)^k \cdot A^k \cdot B^{(L-1)\cdot k} \cdot C}{\epsilon} \geq 2 \cdot 4^k \geq 2^k.$$

Let $f \in \mathcal{F}$, let C be any cube of the above partition, and let p be the Taylor polynomial of total degree $k-1$ of f around the center x_C of C. By using a standard bound on the remainder of the multivariate Taylor polynomial it is possible to show that for any $\tilde{x}$, which is in supremum norm not further away of the center of C than $\eta/2$, we have for c_{47} sufficiently large

$$\begin{aligned}
&|f(\tilde{x}) - p(\tilde{x})| \\
&\leq \Bigg| \sum_{\substack{j_1,\ldots,j_d \in \mathbb{N}_0: \\ j_1+\cdots+j_d=k}} \frac{k}{j_1! \cdot \cdots \cdot j_d!} \int_0^1 (1-t)^{k-1} \cdot \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \ldots \partial^{j_d} x^{(d)}}(x_c - t \cdot (\tilde{x} - x_c))\, dt \\
&\qquad\qquad \cdot (\tilde{x}^{(1)} - x_C^{(1)})^{j_1} \cdot \cdots \cdot (\tilde{x}^{(d)} - x_C^{(d)})^{j_d} \Bigg| \\
&\leq c_{48} \cdot c_{46} \cdot C \cdot B^{(L-1)\cdot k} \cdot A^k \cdot \left(\frac{\eta}{2}\right)^k \leq \frac{\epsilon}{4}.
\end{aligned}$$

Furthermore we have for any $x \in [-\alpha, \alpha]^d$ and any $\tilde{x} \in \mathbb{R}^d$ with $\|x - \tilde{x}\|_\infty \leq \delta$

$$\tilde{x} \in [-2\alpha, 2\alpha]^d,$$

hence $\tilde{x}$ is contained in one set of the above partition, and it holds

$$\begin{aligned}
|f(x) - f(\tilde{x})| &\leq \sum_{i=1}^{d} |f(\tilde{x}^{(1)}, \ldots, \tilde{x}^{(i-1)}, x^{(i)}, \ldots, x^{(d)}) - f(\tilde{x}^{(1)}, \ldots, \tilde{x}^{(i)}, x^{(i+1)}, \ldots, x^{(d)})| \\
&\leq \sum_{i=1}^{d} \left| \frac{\partial f}{\partial x^{(i)}} \right|_{\infty, [-2\alpha, 2\alpha]^d} \cdot |x^{(i)} - \tilde{x}^{(i)}| \leq d \cdot c_{46} \cdot C \cdot B^{(L-1)} \cdot A \cdot \delta \leq \frac{\epsilon}{4}
\end{aligned}$$

provided $c_{43} < 1/(4 \cdot c_{46})$ (cf. (39)).

Let $\mathcal{H}$ be the set of all piecewise polynomials (with respect to the above partition of $[-2 \cdot \alpha, 2 \cdot \alpha]^d$) of total degree $k-1$ with the coefficients bounded in absolute value by $c_{46} \cdot C \cdot B^{(L-1)\cdot k-1} \cdot A^{k-1}$. This set contains all piecewise Taylor polynomials of the above form. And for any $f \in \mathcal{F}$ we can find $h \in \mathcal{H}$ such that for any $x \in [-\alpha, \alpha]^d$ and any $\tilde{x} \in \mathbb{R}^d$ with $\|x - \tilde{x}\|_\infty \leq \delta$ it holds

$$\begin{aligned}
|f(x) - h(\tilde{x})| &= |f(x) - p(\tilde{x})| \\
&\leq |f(x) - f(\tilde{x})| + |f(\tilde{x}) - p(\tilde{x})| \\
&\leq \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2}.
\end{aligned}$$

If we discretize in $\mathcal{H}$ all the

$$c_{49} \cdot K$$

many coefficients, which all take on values in an interval of length

$$2 \cdot c_{50} \cdot C \cdot B^{(L-1)\cdot(k-1)} \cdot A^{k-1},$$

by a grid of size $\epsilon/(2 \cdot c_{51})$, then

$$\left| \sum_{\substack{j_1,\dots,j_d \in \mathbb{N}_0: \\ j_1+\cdots+j_d \le k-1}} a_{j_1,\dots,j_d} \cdot (x^{(1)})^{j_1} \cdot \dots \cdot (x^{(d)})^{j_d} - \sum_{\substack{j_1,\dots,j_d \in \mathbb{N}_0: \\ j_1+\cdots+j_d \le k-1}} b_{j_1,\dots,j_d} \cdot (x^{(1)})^{j_1} \cdot \dots \cdot (x^{(d)})^{j_d} \right|$$
$$\le d^k \cdot (\max\{\|x\|_\infty, 1\})^{k-1} \cdot \max_{\substack{j_1,\dots,j_d \in \mathbb{N}_0: \\ j_1+\cdots+j_d \le k-1}} |a_{j_1,\dots,j_d} - b_{j_1,\dots,j_d}|$$

implies that the resulting set is (for $c_{51} > d^k \cdot (2 \cdot \alpha)^{k-1}$) an $\epsilon/2$-$\| \cdot \|_{\infty,[-2\cdot\alpha,2\cdot\alpha]^d}$-covering of $\mathcal{H}$. Since the number of functions in this covering does not exceed

$$\left(\frac{2 \cdot c_{50} \cdot C \cdot B^{(L-1)\cdot(k-1)} \cdot A^{k-1}}{\epsilon/(2 \cdot c_{51})} \right)^{c_{49} \cdot K},$$

we get the desired cover of $\mathcal{F}$. □

4.4 Proof of Theorem 1

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

$$K_n^{(s,i)} = \left\lceil c_{52} \cdot n^{\frac{K_{i,s}}{2p_{i,s}+K_{i,s}}} \right\rceil \quad (s = 0, \dots, l),$$

hence

$$\bar{K}_n^{(s,i)} = K_n^{(s,i)} \quad \text{for } s = 0, \dots, l-1.$$

Set

$$\bar{K}_n^{(l,1)} = n^5 \cdot K_n^{(l,1)}.$$

Let $\mathbf{w}^*$ be a weight vector of a neural network in Lemma 6 which approximates m, where for the network approximating $g_{1,l}$ all the in parallel computed neural networks are repeated n^5 times with values $w_{j,1,1}^{(L_l)} = w_{1,j}^{(L)}$ replaced by $w_{j,1,1}^{(L_l)}/n^5$. Then the corresponding network

$$f_{\mathbf{w}^*}(x)$$

satisfies

$$\sup_{x \in [-A,A]^d} |f_{\mathbf{w}^*}(x) - m(x)| \le c_{37} \cdot \max_{i,s} \frac{1}{(K_n^{(s,i)})^{p_{i,s}/K_{i,s}}} \le c_{53} \cdot \max_{i,s} \frac{1}{n^{\frac{p_{i,s}}{2p_{i,s}+K_{i,s}}}}.$$

The weight vectors corresponding to $g_{NN,i,s}$ of this hierarchically composed neural network satisfy

$$(\mathbf{w}^*)_{k,1,1}^{(L_s)} \in [-c_{3,n}, c_{3,n}], \quad (\mathbf{w}^*)_{k,r,j}^{(l)} \in [-c_{2,n}, c_{2,n}] \quad (l = 1, \dots, L_s - 1)$$

and

$$(\mathbf{w}^*)_{k,r,j}^{(0)} \in [-c_{1,i,s,n}, c_{1,i,s,n}].$$

Furthermore, the weights in the last layer of the hierarchical composed networks satisfy

$$\sum_{k=1}^{n^5 \cdot K_n^{(l,1)}} |(\mathbf{w}^*)_{1,k}^{(L)}|^2 \leq n^5 \cdot K_n^{(l,1)} \cdot \left(\frac{n}{n^5}\right)^2 \leq \frac{1}{n^2} = \alpha_n.$$

Set

$$\epsilon_n = \frac{1}{n^{3 \cdot l + 9}}.$$

Let $A_{n,1}$ be the event that the weight vector $\mathbf{w}^{(0)}$ satisfies firstly in each of the hierarchically combined networks $g_{NN,i,s}$ with $s \in \{0, \ldots, l-1\}$

$$|(\mathbf{w}^{(0)})_{j_{t,i,s},k,j}^{(r)} - (\mathbf{w}^*)_{t,k,j}^{(r)}| \leq \epsilon_n \quad \text{for all } r \in \{0, \ldots, L_s\}, t \in \{1, \ldots, \bar{K}_n^{(s,i)}\}$$

for some pairwise distinct $j_{1,i,s}, \ldots, j_{\bar{K}_n^{(s,i)},i,s} \in \{1, \ldots, K_n\}$ and that it also satisfies in $g_{1,l}$

$$|(\mathbf{w}^{(0)})_{j_{t,1,l},k,j}^{(r)} - (\mathbf{w}^*)_{t,k,j}^{(r)}| \leq \epsilon_n \quad \text{for all } r \in \{0, \ldots, L_l - 1\}, t \in \{1, \ldots, \bar{K}_n^{(l,1)}\}$$

for some pairwise distinct $j_{1,1,l}, \ldots, j_{\bar{K}_n^{(l,1)},1,l} \in \{1, \ldots, K_n\}$. Let $A_{n,2}$ be the event that in case that $A_{n,1}$ holds at some pruning step simultaneously all the weights

$$w_{j_{1,i,s},1,1}^{(L_s)}, \ldots, w_{j_{\bar{K}_n^{(s,i)},i,s},1,1}^{(L_s)}$$

from $g_{i,s}$ are chosen for all i and all $s = 0, \ldots, l-1$. And let $A_{n,3}$ be the event that

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

holds. Let A_n be the event that $A_{n,1}$, $A_{n,2}$ and $A_{n,3}$ hold simultaneously.

In the sequel we decompose the L_2 error of m_n in a sum of several terms. Set

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

We have

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&= \left(\mathbf{E}\left\{|m_n(X) - Y|^2 | \mathcal{D}_n\right\} - \mathbf{E}\{|m(X) - Y|^2\}\right) \cdot 1_{A_n} + \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{A_n^c} \\
&= \Big[\mathbf{E}\left\{|m_n(X) - Y|^2 | \mathcal{D}_n\right\} - \mathbf{E}\{|m(X) - Y|^2\} \\
&\qquad - \left(\mathbf{E}\left\{|m_n(X) - T_{\beta_n} Y|^2 | \mathcal{D}_n\right\} - \mathbf{E}\{|m_{\beta_n}(X) - T_{\beta_n} Y|^2\}\right)\Big] \cdot 1_{A_n}
\end{aligned}$$

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

In the remainder of the proof we bound

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

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

In the *first step of the proof* we show

$$\mathbf{E}T_{j,n}\leq c_{54}\cdot\frac{\log n}{n}\quad\text{for } j\in\{1,3\}.$$

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

In the *second step of the proof* we show

$$\mathbf{E}T_{5,n}\leq c_{55}\cdot\frac{(\log n)^2}{n}.$$

The definition of m_n implies $\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\leq 4\cdot c_2^2\cdot(\log n)^2$, hence it suffices to show

$$\mathbf{P}(A_n^c)\leq\frac{c_{56}}{n^2}. \tag{42}$$

To do this, we consider separately in each of the hierarchically composed subnetworks a sequential choice of the weights of the K_n fully connected neural networks. The probability that the weights in the first of the K_n in parallel computed neural networks in $g_{NN,i,s}$ differ in all components at most by ϵ_n from the weights in the first of the $\bar{K}_n^{(s,i)}$ in parallel computed neural networks in the corresponding network in our hierarchically composed network with good approximation properties constructed above is for large n bounded from below by

$$\left(\frac{\epsilon_n}{2\cdot c_{3,n}}\right)\cdot\left(\frac{\epsilon_n}{2\cdot c_{2,n}}\right)^{r_{max}\cdot(r_{max}+1)\cdot(L_{max}-1)}\cdot\left(\frac{\epsilon_n}{2\cdot c_{1,i,s,n}}\right)^{r_{max}\cdot(K_{max}+1)}$$

$$\geq n^{-(3\cdot l+10)\cdot(r_{max}+1)^2\cdot(L_{max}+K_{max})}$$

where

$$K_{max} = \max_{i,s} K_{i,s}, \quad r_{max} = \max_{i,s} r_{i,s} \quad \text{and} \quad L_{max} = \max_{s} L_s = \bar{L}.$$

Hence the probability that none of the first $n^{(3\cdot l+10)\cdot(r_{max}+1)^2\cdot(L_{max}+K_{max})+1}$ neural networks satisfies this condition is bounded from above by

$$\begin{aligned}
&(1 - n^{-(3\cdot l+10)\cdot(r_{max}+1)^2\cdot(L_{max}+K_{max})})^{n^{(3\cdot l+10)\cdot(r_{max}+1)^2\cdot(L_{max}+K_{max})+1}} \\
&\leq \left(\exp\left(-n^{-(3\cdot l+10)\cdot(r_{max}+1)^2\cdot(L_{max}+K_{max})}\right)\right)^{n^{(3\cdot l+10)\cdot(r_{max}+1)^2\cdot(L_{max}+K_{max})+1}} \\
&= \exp(-n).
\end{aligned}$$

And since there are only finitely many of these subnetworks the probability that in any of these subnetworks none of the first $n^{(3\cdot l+10)\cdot(r_{max}+1)^2\cdot(L_{max}+K_{max})+1}$ neural networks satisfies this condition is for large n bounded from above by

$$c_{57} \cdot \exp(-n).$$

Since we have $K_n \geq n^{(3\cdot l+10)\cdot(r_{max}+1)^2\cdot(L_{max}+K_{max})+1} \cdot \max_{i,s} \bar{K}_n^{(s,i)}$ for n large we can successively use the same construction for all of the weights in any of the subnetworks and we can conclude: The probability that there exist (s,i) and $k \in \{1,\ldots,\bar{K}_n^{(s,i)}\}$ such that none of the K_n weight vectors of the network corresponding to $g_{NN,i,s}$ differs in all components by at most ϵ_n from $(w_{k,i,j}^{(r)})_{i,j,r:r\leq L_s}$ (or $(w_{k,i,j}^{(r)})_{i,j,r:r<L_s}$ in case $s = l$) is for large n bounded from above by

$$c_{58} \cdot \max_{i,s} \bar{K}_n^{(s,i)} \cdot \exp(-n) \leq n^6 \cdot c_{59} \cdot \exp(-n) \leq \frac{1}{n^2}.$$

This implies for large n

$$\begin{aligned}
\mathbf{P}(A_{n,1}^c) + \mathbf{P}(A_{n,3}^c) &\leq \frac{1}{n^2} + \mathbf{P}\{\max_{i=1,\ldots,n} |Y_i| > \sqrt{\beta_n}\} \leq \frac{1}{n^2} + n \cdot \mathbf{P}\{|Y| > \sqrt{\beta_n}\} \\
&\leq \frac{1}{n^2} + n \cdot \frac{\mathbf{E}\{\exp(c_3 \cdot Y^2)}{\exp(c_3 \cdot \beta_n)} \leq \frac{c_{33}}{n^2},
\end{aligned}$$

where the last inequality follows from the assumption $c_2 \cdot c_3 \geq 3$. Furthermore, the probability that during one pruning step $A_{n,1}$ holds and that in all hierarchical composed networks just the subsets are chosen where according to $A_{n,1}$ the best approximating weights are approximated with an error at most ϵ_n is for large n bounded from below by

$$n^6 \cdot c_{59} \cdot \exp(-n) \cdot \left(\frac{1}{\binom{n^\kappa}{c_{61}\cdot n}}\right)^{c_{62}} \geq n^6 \cdot c_{59} \cdot \exp(-n) \cdot \frac{1}{(n^\kappa)^{c_{63}\cdot n}} \geq e^{-c_{64}\cdot n\cdot\log n}.$$

Since we perform at least $\lfloor t_n/s_n \rfloor$ of these pruning steps independently, we can conclude

$$\mathbf{P}\{A_{n,2}^c\} \leq \left(1 - e^{-c_{64}\cdot n\cdot\log n}\right)^{t_n/s_n - 1} \leq \frac{c_{65}}{n^2}.$$

Hence we have shown

$$\mathbf{P}\{A_n^c\} \leq \mathbf{P}(A_{n,1}^c) + \mathbf{P}(A_{n,2}^c) + \mathbf{P}(A_{n,3}^c) \leq \frac{c_{66}}{n^2}.$$

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

$$\mathbf{E}T_{2,n} \leq c_{67} \cdot \max_{i,s} n^{-\frac{2p_{i,s}}{2p_{i,s}+K_{i,s}}+\epsilon}.$$

Let $\mathcal{W}_n$ be the set of all weight vectors of the hierarchically composed neural network defined by (11)–(13) with $K_{i,s}$, $K_n^{(s)}$, L_s and $r_{i,s}$ chosen as in Theorem 1, where for each of the subnetwork $g_{NN,i,s}$ the weights $(w_{i,j,k}^{(l)})_{i,j,k,l}$ satisfy

$$|w_{1,1,k}^{(L_s)}| \leq c_{68} \cdot n \quad (k = 1, \ldots, K_n),$$

$$|w_{i,j,k}^{(l)}| \leq c_{69} \quad (l = 1, \ldots, L_s - 1)$$

and

$$|w_{i,j,k}^{(0)}| \leq c_{70} \cdot n^{\frac{1}{2p_{i,s}+K_{i,s}}} \cdot \log n.$$

The initialization of $\mathbf{w}^{(0)}$ together with (17) implies

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

Hence, for any $u > 0$ we get

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

where

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

By Lemma 7 and Lemma 8 we get for any $x_1, \ldots, x_n \in supp(X)$

$$\begin{aligned}
&\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n} \cdot f : f \in \mathcal{F}_n\right\}, x_1^n\right) \leq \mathcal{N}_1\left(\delta \cdot \beta_n, \mathcal{F}_n, x_1^n\right) \\
&\leq \mathcal{N}_{\|\cdot\|_{\infty,[-A,A]^d},0}\left(\delta \cdot \beta_n, \mathcal{F}_n\right) \\
&\leq \prod_{i,s}\left(\frac{c_{71} \cdot n^{c_{72}}}{\delta \cdot \beta_n / n^{c_{73} \cdot l}}\right)^{c_{74} \cdot (n^{1/(2p_{i,s}+K_{i,s})} \cdot \log n)^{K_{i,s}} \cdot (c_{69})^{(L_s-1)\cdot K_{i,s}} \cdot \left(\frac{K_n \cdot c_{68} \cdot n}{\beta_n \cdot c_{76} \cdot \delta / n^{c_{77} \cdot l}}\right)^{K_{i,s}/k}}.
\end{aligned}$$

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

$$\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n} \cdot f : f \in \mathcal{F}_n\right\}, x_1^n\right) \le c_{79} \cdot \prod_{i,s} n^{c_{80} \cdot n^{\frac{K_{i,s}}{2p_{i,s}+K_{i,s}}+\epsilon/2}}.$$

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

$$\mathbf{P}\{T_{2,n} > u\} \le 14 \cdot c_{79} \cdot \prod_{i,s} n^{c_{80} \cdot n^{\frac{K_{i,s}}{2p_{i,s}+K_{i,s}}+\epsilon/2}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot u\right).$$

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

$$\begin{aligned}
\mathbf{E}\{T_{2,n}\} &\le \epsilon_n + \int_{\epsilon_n}^{\infty} \mathbf{P}\{T_{2,n} > u\}\, du \\
&\le \epsilon_n + 14 \cdot c_{79} \cdot \prod_{i,s} n^{c_{80} \cdot n^{\frac{K_{i,s}}{2p_{i,s}+K_{i,s}}+\epsilon/2}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot \epsilon_n\right) \cdot \frac{5136 \cdot \beta_n^2}{n}.
\end{aligned}$$

Setting

$$\epsilon_n = \frac{5136 \cdot \beta_n^2}{n} \cdot \log \prod_{i,s} n^{c_{80} \cdot n^{\frac{K_{i,s}}{2p_{i,s}+K_{i,s}}+\epsilon/2}} = c_{81} \cdot \sum_{i,s} \frac{n^{\frac{K_{i,s}}{2p_{i,s}+K_{i,s}}+\epsilon/2} \cdot (\log n)^3}{n}$$

yields the assertion of the third step of the proof.

In the *fourth step of the proof* we show

$$\mathbf{E}\{T_{4,n}\} \le c_{82} \cdot \max_{i,s} n^{-\frac{2p_{i,s}}{2p_{i,s}+K_{i,s}}}.$$

Using

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

we get

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

Let $T \in \mathbb{N}$ be a random number such that on A_n the pruning step number T leads to the choice of the right subsets in $A_{n,3}$. By definition of $\hat{t}$ we have

$$F_n(\mathbf{w}^{(\hat{t})}) = \min_{t=0,\ldots,t_n-1} F_n(\mathbf{w}^{(t)}) \leq \min_{t=T\cdot s_n,\ldots,T\cdot s_n+s_n-1} F_n(\mathbf{w}^{(t)}),$$

hence

$$T_{4,n}/2 \leq \Big[\min_{t=T\cdot s_n,\ldots,T\cdot s_n+s_n-1} F_n(\mathbf{w}^{(t)}) - \frac{1}{n}\sum_{i=1}^{n} |m(X_i)-Y_i|^2 \Big] \cdot 1_{A_n}.$$

Next we apply Lemma 1 with

$$F(u,v) = F_n((w_{1,k}^{(L)})_k, ((w_{i,j}^{(s)})_{i,j,s:s<L})),$$

$$u^* = (((\mathbf{w}^*)_{k,1}^{(L)})_k \quad \text{and} \quad v^* = (((\mathbf{w}^*)_{i,j}^{(s)})_{i,j,s:s<L}).$$

By Lemma 4 we know

$$\|\nabla_{(w_k)_{k=1,\ldots,K_n}} F_n(\mathbf{w})\| \leq 2\cdot(\beta_n + \sqrt{K_n}\cdot\sqrt{\alpha_n})\cdot\sqrt{K_n} \leq K_n,$$

hence assumption (21) of Lemma 1 is satisfied for

$$C_n = K_n.$$

In order to determine the value of D_n in assumption (22) of Lemma 1 we apply Lemma 2 and Lemma 3. Because of the pruning steps during the computation of our estimates the occuring weights satisfies

$$\sum_{k=1}^{K_n^{(s)}} |w_{k,1,1}^{(L_s)}| \leq \max_i \bar{K}_n^{(i,s)} \cdot \max_{k=1,\ldots,K_n^{(s)}} |w_{k,1,1}^{(L_s)}| \leq n^2,$$

and by Lemma 3 we can conclude

$$|f_{\mathbf{v}_1,k}^{(L)}(x) - f_{\mathbf{v}_2,k}^{(L)}(x)| \leq c_{16}\cdot\sqrt{\max_{i,s:s=0,\ldots,l-1} \bar{K}_n^{(i,s)}}\cdot n^l \cdot c_{84}^{L-2l-1}\cdot n^{2l}\cdot\|\mathbf{v}_1-\mathbf{v}_2\|.$$

Application of Lemma 2 yields that assumption (22) of Lemma 1 is satisfied for

$$D_n = 2\cdot(\beta_n + K_n^{(l,1)}\cdot n)\cdot c_{16}\cdot c_{84}^{L-2l-1}\cdot n^{3l+1/2}\cdot\sqrt{n^5\cdot K_n^{(l,1)}} \leq c_{85}\cdot n^{3l+6},$$

where we have used

$$\sum_{k=1}^{K_n} |(\mathbf{w}^*)_{1,k}^{(L)}| \leq K_n^{(l,1)}\cdot n.$$

Furthermore, if we set in $((\mathbf{w}^*)_{k,i,j}^{(l)})_{k,i,j,l:l<L}$ all those components equal to the values in $((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l<L}$, where $(\mathbf{w}^*)_{k,1,1}^{(L_s)} = 0$, which does not change the value of $F_n(((\mathbf{w}^*)_{k,1}^{(L)})_k, ((\mathbf{w}^*)_{i,j}^{(s)})_{i,j,s:s<L}))$, we have

$$\|((\mathbf{w}^*)_{i,j}^{(s)})_{i,j,s:s<L} - ((\mathbf{w}^{(0)})_{i,j}^{(s)})_{i,j,l:s<L}\| \leq c_{86}\cdot n\cdot\epsilon_n \leq \delta_n.$$

Application of Lemma 1 yields

$$\begin{aligned} T_{4,n}/2 &\leq \Big[F_n(((\mathbf{w}^*)_{k,1}^{(L)})_k, ((\mathbf{w}^*)_{i,j}^{(s)})_{i,j,s:s<L})) + c_{87} \cdot n^{3 \cdot l+6} \cdot \sqrt{\alpha_n} \cdot \delta_n \\ &\quad + \frac{\alpha_n}{2} + \frac{C_n^2}{2 \cdot s_n} - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \Big] \cdot 1_{A_n} \\ &\leq \Big[F_n(((\mathbf{w}^*)_{k,1}^{(L)})_k, (((\mathbf{w}^*)_{k,i,j}^{(l)})_{k,i,j,l:l<L})) - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \Big] \cdot 1_{A_n} + \frac{c_{88}}{n}. \end{aligned}$$

This implies

$$\begin{aligned} &\mathbf{E}\{T_{4,n}/2\} \\ &\leq \mathbf{E}\left\{ \Big[\frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{\mathbf{w}^*}(X_i)|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \Big] \cdot 1_{A_n} \right\} + \frac{c_{88}}{n} \\ &\leq \mathbf{E}\left\{ \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{\mathbf{w}^*}(X_i)|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right\} \\ &\quad + \sqrt{\mathbf{E}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right|^2 \right\}} \cdot \sqrt{\mathbf{P}(A_n^c)} + \frac{c_{88}}{n} \\ &\leq \int |f_{\mathbf{w}^*}(x) - m(x)|^2 \mathbf{P}_X(dx) + \frac{c_{89}}{n} \\ &\leq c_{90} \cdot \max_{i,s} n^{-\frac{2p_{i,s}}{2p_{i,s}+K_{i,s}}}. \end{aligned}$$

□

References

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

[2] Bartlett, P., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight VC-dimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research* **20**, pp. 1-17.

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

[4] Braun, A., Kohler, M., Langer, S., and Walk, H. (2023). Convergence rates for shallow neural networks learned by gradient descent. Accepted for publication in *Bernoulli*. Preprint, *arXiv: 2107.09550.*

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

[6] Devroye, L. (1982). Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **61**, pp. 467-481.

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

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

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

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

[11] Golowich, N., Rakhlin, A., and Shamir, O. (2019). Size-Independent sample complexity of neural networks. Preprint, *arXiv: 1712.06541.*

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

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

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

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

[16] Jacot, A., Gabriel, F., and Hongler, C. (2020). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *arXiv: 1806.07572v4.*

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

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

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

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

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

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

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

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

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

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

[27] Liang, T., Rakhlin, A., and Sridharan, K. (2015). Learning with square loss: localization through offset Rademacher complexity. Preprint, *arXiv: 1502.06134.*

[28] Lin, S., and Zhang, J. (2019). Generalization bounds for convolutional neural networks. Preprint, *arXiv: 1910.01487.*

[29] Lu, J., Shen, Z., Yang, H., and Zhang, S. (2020). Deep network approximation for smooth functions. *arxiv: 2001.03040.*

[30] Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. In *Proceedings of the National Academy of Sciences*, **115**, pp. E7665-E7671.

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

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

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

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

[37] Wang, M., and Ma, C. (2022). Generalization error bounds for deep neural network trained by SGD. Preprint, *arXiv: 2206.03299v1.*

[38] Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. Preprint, *arXiv: 1802.03620.*

[39] Yarotsky, D., and Zhevnerchuk, A. (2019). The phase diagram of approximation rates for deep neural networks. Preprint, *arXiv: 1906.09477.*

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

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