

Learning of deep neural network regression estimates using gradient descent with pruning *

Michael Kohler¹, Adam Krzyżak^{2,†} and Vincent Molinero Römer¹

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

August 26, 2026

Abstract

Estimation of a regression function from independent and identically distributed data is considered. The L_2 error with integration with respect to the design variable is used as the error criterion. An initially randomly pruned fully connected deep neural network with logistic squasher as activation function is fitted to the data via gradient descent, using a data-dependent choice of non-constant stepsizes during gradient descent. It is shown that this network achieves (up to a logarithmic factor) the optimal minimax rate of convergence in case that the regression function is (p, C) -smooth. Here the estimate is able to circumvent the curse of dimensionality provided the predictors are concentrated in the neighborhood of a low dimensional manifold. The finite sample size performance of the estimate is illustrated by applying it to the simulated data.

AMS classification: Primary 62G08; secondary 62G20.

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

1 Introduction

1.1 Scope of this paper

Deep learning is currently changing the world. It has recently achieved tremendous success in various applications like image classification (cf., e.g., Krizhevsky, Sutskever and Hinton (2012)), in language recognition (cf., e.g., Kim (2014)), in machine translation (cf., e.g., Wu et al. (2016)), in game playing (cf., e.g., Silver et al. (2017)) or in simulation of human conversation (cf., e.g., Minaee et al. (2025)). Nowadays it is also applied in mathematics in various fields outside of statistics, e.g., Kohler et al. (2026) applies deep

*Running title: *Learning of deep neural networks*

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

learning in efficiently solving partial differential equations occurring in modelling of the diffusion MRIs.

Motivated by this huge success of deep learning in applications there has been also an increasing interest in theoretical foundations of deep learning. Often deep learning is studied theoretically in some equivalent model like the neural tangent kernel (cf., e.g., Jacot, Gabriel and Hongler (2018)) or the mean-field theory (cf., e.g. Mei, Montanari and Nguyen (2018)). In this article we use a different approach: we study deep learning in the context of nonparametric regression directly without using any asymptotic description. Here we analyze simultaneously the approximation error, the generalization error and the optimization error of the estimate and use our results to propose a special topology for the deep neural network, a special initialization scheme for its initial weights, a special choice for the non-constant step sizes during gradient descent, and a special choice for the number of gradient descent steps, and derive results concerning the rate of convergence of this estimate.

1.2 Neural networks

In order to introduce our theoretically motivated definition of deep neural network regression estimates let us first describe fully connected deep neural networks. These networks depend on an activation function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, e.g., the so-called logistic squasher

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

or the famous ReLU activation function

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

A fully connected deep neural network with L layers and r_n neurons per layer is defined as a function

$$f_{\mathbf{w}}(x) = \sum_{k=1}^{r_n} w_{1,k}^{(L)} \cdot f_{\mathbf{w},k}^{(L)}(x), \quad (2)$$

where

$$f_{\mathbf{w},i}^{(l)}(x) = \sigma \left(\sum_{k=1}^{r_n} w_{i,k}^{(l-1)} \cdot f_{\mathbf{w},k}^{(l-1)}(x) + w_{i,0}^{(l-1)} \right) \quad (3)$$

for $i \in \{1, \dots, r_n\}$ and $l \in \{2, \dots, L\}$, and

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

for $i \in \{1, \dots, r_n\}$. Here $w_{i,k}^{(l)}$ is the weight between neuron k in layer l and neuron i in layer $l+1$ ($l \in \{1, \dots, L-1\}$), and

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

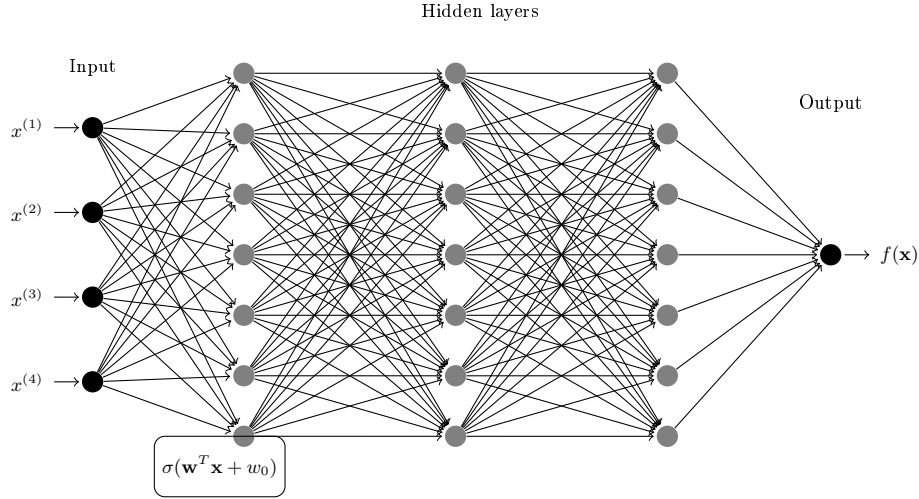


Figure 1: A fully connected network of the class $\mathcal{F}(3, 7)$

is the vector of the weights of the network. We denote the class all such networks by $\mathcal{F}(L, r_n)$, where L is the depth of the network and r_n is the width of the network. A schematic representation of such a network in case $L = 3$ and $r_n = 7$ can be found in Figure 1.

In the sequel we want to use such networks to define nonparametric regression estimates.

1.3 Nonparametric regression

In nonparametric regression (X, Y) , $(X_1, Y_1), \dots$ are independent and identically distributed (i.i.d.) $\mathbb{R}^d \times \mathbb{R}$ valued random variables with $\mathbf{E}Y^2 < \infty$, and given the i.i.d. sample

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

of (X, Y) the task is to construct an estimate

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

of the regression function

$$m : \mathbb{R}^d \rightarrow \mathbb{R}, \quad m(x) = \mathbf{E}\{Y|X = x\}$$

such that the so-called L_2 error

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

is "small". For an introduction to nonparametric regression and for a motivation of the L_2 error criterion we refer to Györfi et al. (2002).

1.4 Neural network regression estimates

To construct a neural network regression estimate, one chooses a class of neural networks $f_{\mathbf{w}}$ (e.g., the class $\mathcal{F}(L, r_n)$ of fully connected neural networks with depth L and width r_n), and minimizes the empirical L_2 risk

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}}(X_i) - Y_i|^2 \quad (5)$$

with respect to the weight vector $\mathbf{w}$ of the neural network. It is well-known that the least squares neural network regression estimates, where one minimizes (5) over suitably defined spaces of deep neural network, achieve nice rate of convergence results (cf., e.g., Bauer and Kohler (2019), Schmidt-Hieber (2020), Kohler and Langer (2021) or Jiao et al. (2023), and the literature cited therein). Unfortunately, in applications these least squares estimates are not applicable: Since (5) is a nonlinear and nonconvex function of $\mathbf{w}$, it is not possible to minimize (5) exactly. In practice, gradient descent (or one of its variants like stochastic gradient descent) is used to minimize it approximately. Here one starts with a random initialization $\mathbf{w}^{(0)}$ of the weight vector, and then computes

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_t \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}) \quad \text{for } t = 0, \dots, t_n - 1, \quad (6)$$

i.e., one performs t_n gradient descent steps with step sizes $\lambda_0, \dots, \lambda_{t_n-1} \geq 0$.

In order to define this estimate, one has to decide which activation function one wants to use, what kind of topology of the network one wants to use, how the initialization of the weights is done, and how many gradient descent steps with what kind of step sizes one wants to perform. All of these decisions influence the performance of the estimate in applications, and it is not clear what the right choices are.

In this article we use statistical theory to provide answers to these questions. Next we quickly describe the results we get.

1.5 Main results

We will use the logistic squasher (1) as the activation function. As far as the topology of the network is concerned we start with a fully connected network which we initially randomly prune by randomly choosing for each hidden neurons of the layers $2, \dots, L$ r neurons from the previous layer, use only the connections to the randomly selected neurons and ignore all the rest. See Figure 2 for an illustration of a pruning of the network from Figure 1.

We will use uniform distributions in order to initialize our weights. Our choice is motivated by results concerning the approximation properties and the generalization error of neural networks, and we choose the weights between the input and the first hidden layer uniformly distributed on same large interval $[-A, A]$, and the weights between all the hidden layers uniformly distributed on some moderate interval $[-B, B]$. All output weights will be initially set to zero.

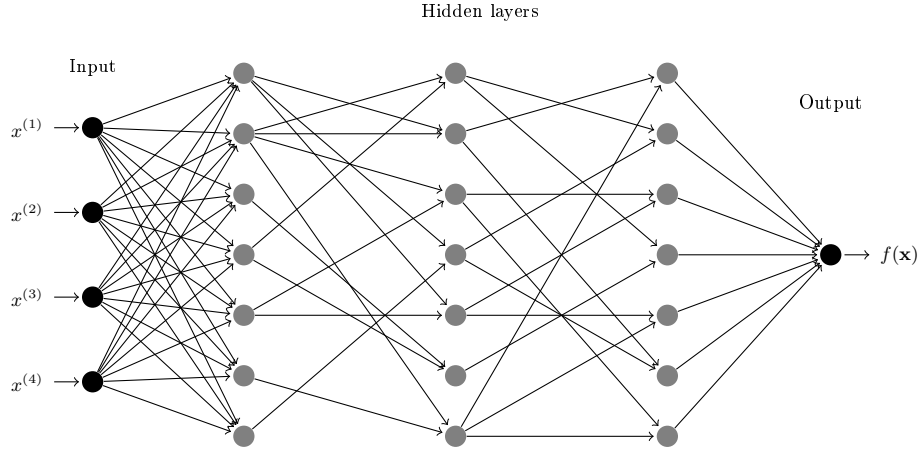


Figure 2: The above network after random pruning with $r = 2$

Our basic idea for choosing the step sizes is to choose the maximal λ_t such that

$$F_n(\mathbf{w}^{(t+1)}) \leq F_n(\mathbf{w}^{(t)}) - \frac{\lambda_t}{2} \cdot \left\| \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}) \right\|^2 \quad (t = 0, 1, \dots, t_n - 1) \quad (7)$$

holds, which leads to a version of the well-known Armijo line search procedure from optimization for the choice of the step sizes during gradient descent (cf., Armijo (1966)). It turns out that in order to be able to show a nice error bound for the estimate it suffices to impose the following additional conditions on the non-constant step sizes $\lambda_0, \dots, \lambda_{t_n-1} \geq 0$ and by assuming that the number of gradient descent steps t_n satisfy:

$$1 \leq \sum_{s=0}^{t_n-1} \lambda_s \leq 2 \quad \text{and} \quad \max_{s=0, \dots, t_n-1} \lambda_s \leq \frac{1}{n}.$$

An algorithm which computes data-dependent non-constant step sizes satisfying the above conditions is presented below as Algorithm 1. Here we stop checking condition (7) as soon as the step size is below some threshold, since in this case the step size will fulfill condition (7) automatically with high probability as we will see in the proof of our main results.

Our estimate leads to simple rules for otherwise difficult decisions in the applications of deep neural network regression estimates concerning the topology of the network, the initialization and the choices of the parameters of the gradient descent. The topology of our network is a simple fully connected neural network which is initially randomly pruned. After pruning it is still drastically overparameterized in the sense that it has many more parameters than data points, but due to our special structure of the pruning, the special initialization of the weights, and the special choice of the step sizes and the number of gradient descent steps our estimate does not overfit the data.

We are able to derive bounds on the expected L_2 error of this estimate which imply that it achieves asymptotically up to some logarithmic factor the optimal minimax rate of

```

Data:  $(x_1, y_1), \dots, (x_n, y_n)$ 
 $L, r_n, r, A, B$ 
 $\lambda_{min} = \frac{1}{(\log n)^{2L+3} \cdot r_n^{3/2}}, c_1 = 1$ 
begin
   $t = 0$ 
   $\mathbf{w}^{(0)} = \text{InitializeWeights}(L, r_n, r, A, B)$ 
  repeat
     $\lambda_t = \frac{2}{n}$ 
    repeat
       $\lambda_t = \lambda_t / 2$ 
       $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_t \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})$ 
    until  $F_n(\mathbf{w}^{(t+1)}) \leq F_n(\mathbf{w}^{(t)}) - \frac{\lambda_t}{2} \cdot \|\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})\|^2$  or  $\lambda_t \leq \lambda_{min}$ ;
     $t = t + 1$ 
  until  $\sum_{s=0}^{t-1} \lambda_s \geq c_1$ ;
   $t_n = t$ 
  Result:  $f_{\mathbf{w}^{(t_n)}}$ 
end

```

Algorithm 1: Pseudo code for the data-dependent choice of non-constant step sizes for gradient descent.

convergence for estimation of smooth functions. In addition, we show that the estimate is able to circumvent the so-called curse of dimensionality in case the predictor is concentrated in the neighborhood of a low dimensional manifold. Furthermore, we illustrate the finite sample size behaviour of our estimate (and in particular our data-dependent choice of the non-constant step sizes) by applying it to simulated data from an univariate regression problem.

The results in this paper are based on earlier theoretical results developed by the authors and co-authors (cf., Subsection 1.6). The main new contributions in this article are:

1. We extend the previous theory such that it covers more general network topologies (initially randomly pruned fully connected deep neural networks instead of linear combinations of small parallel deep neural networks) and non-constant step sizes during gradient descent.
2. We prove rates of convergence which are optimal up to logarithmic factor instead of optimal up to small constants in the exponent of the rates (like, e.g., in Kohler (2026)).
3. We extend previous known results from manifold data to the case that the predictor is contained in a neighborhood of a manifold which is important since it occurs often in applications that the data is close to manifolds, but not necessarily contained in a manifold (cf., e.g., Carlsson (2009)).

1.6 Discussion of related results

In this paper we use a version of the Armijo line-search during gradient descent. Concerning results of this method in connection with optimization of general functions we refer to Vaswani and Babanezhod (2026) and the literature cited therein. Applications of it towards deep learning are described in Kenneweg, Kenneweg and Hammer (2024).

We also use pruning, which is a standard method in learning of deep neural networks. A review of various pruning methods can be found in Cheng, Zhang and Shi (2024). In our paper we apply pruning before the training of the networks, so we basically use it to define a (random) class of sparse deep neural networks. This helps us to simultaneously control the approximation error, the generalization error and the optimization error of our estimate.

Deep learning has been studied from several complementary mathematical perspectives. Approximation-theoretic results quantify the expressive power of the neural networks. Examples include the classical universal approximation theorem for sigmoidal networks (cf. Cybenko (1989) and Hornik, Stinchcombe and White (1989)) and approximation results for deep neural networks like Yarotsky (2017, 2018), Yarotsky and Zhevnerchuck (2019), Langer (2020) or Lu et al. (2020).

Often least squares neural network estimates have been analyzed. Early works for shallow (one hidden layer) neural networks include Barron (1994). The rate of convergence results for deep neural networks which include dimension reduction in case that the regression function is a composition of low-dimensional smooth functions have been shown by Kohler and Krzyżak (2017), Bauer and Kohler (2019), Schmidt-Hieber (2020) and Kohler and Langer (2021). Kohler, Langer and Reif (2023) demonstrated that least squares neural network estimates can achieve a dimension reduction in case that the predictor is contained in a low dimensional manifold. In practice, data is often just close, but not exactly contained in a low dimensional manifold. In this case Jiao et al. (2023) showed that the deep neural network least squares estimates can also achieve a low dimensional rate of convergence and can hence circumvent the curse of dimensionality. Our second main result can be considered as a generalization of this result to the case that the estimate is learned by gradient descent (as it is always done in applications since there least squares estimates are computationally infeasible).

Another line of work studies the optimization of deep neural networks. E.g., Du et al. (2019) and Allen-Zhu, Li and Song (2019) consider gradient-based algorithms for the minimization of the highly non-convex empirical loss of an overparameterized network. Statistical analyses, in turn, investigate the population prediction error and its dependence on the sample size and the regularity of the regression function. These questions are related and need to be studied simultaneously, since a small approximation error does not imply that an optimization algorithm finds a good network, and a small empirical training error does not by itself imply good generalization. In fact, Kohler and Krzyżak (2021) show that overparameterized networks can attain an almost minimal empirical risk without achieving an optimal nonparametric convergence rate.

One popular approach to study optimization and generalization simultaneously is the neural tangent kernel (NTK) introduced by Jacot, Gabriel and Hongler (2018) for shallow

neural networks. They showed that, in the infinite-width limit, the training dynamics of the network output are governed by a deterministic kernel that remains constant during gradient flow. Building on this result, Lee et al. (2019) showed that the parameters and outputs of a sufficiently wide network trained by gradient descent remain close to those of the first-order Taylor expansion around the initialization. Their result therefore provides an explicit linear-model interpretation of the NTK regime and extends the analysis to discrete gradient descent.

The NTK approach has been used to obtain the optimization and generalization results for overparameterized networks. Arora et al. (2019), for instance, study both optimization and generalization through a kernel associated with the network. In the setting of nonparametric regression, Nitanda and Suzuki (2021) derive minimax optimal convergence rates for averaged stochastic gradient descent in the NTK regime. Nguyen and Mücke (2024) obtain refined bounds for early-stopped gradient descent and quantify the network width required to approximate the limiting kernel estimator.

Another asymptotic description is given by mean-field theory. For a wide shallow network, the empirical distribution of the neuron parameters is represented by a probability measure. Under an appropriate mean-field scaling, its evolution is described by a nonlinear partial differential equation or a Wasserstein gradient flow, see Mei, Montanari and Nguyen (2018), Chizat and Bach (2018), and Sirignano and Spiliopoulos (2020). In contrast to the classical NTK regime, the parameter distribution can evolve nonlinearly, so the mean-field model can describe changes of the features during training. Mei, Misiakiewicz and Montanari (2019) additionally derive finite-particle approximation bounds and show how a kernel regime can be recovered as a special limit of the mean-field description. Mean-field frameworks have also been extended to multilayer networks by Nguyen and Pham (2023).

Both of those approaches provide powerful descriptions of limiting training dynamics often restricted to shallow neural networks. Our approach focuses on deep neural networks with a specific network architecture and logistic activation function. Here we study our network estimator directly without asymptotic descriptions, which allows for a statistically guided construction of concrete regression estimates.

The works most closely related to the present paper analyze neural network regression estimators trained by gradient descent directly in a statistical regression setting. Braun et al. (2024) derive a convergence rate of order $n^{-1/2}$, up to a logarithmic factor, for shallow neural network estimators with logistic activation and a suitable random initialization, under an appropriate decay condition on the Fourier transform of the regression function. Drews and Kohler (2024) establish universal consistency for an overparameterized deep neural network estimator. Quantitative bounds on the expected L_2 error and corresponding convergence rates are obtained by Kohler and Krzyżak (2025a, 2025b), Drews and Kohler (2026), and Kohler (2026). The analyses in these works jointly control approximation, optimization, and generalization for concrete finite-width neural network estimators trained by gradient descent. Our article relies in particular on new techniques which have been introduced in Kohler (2026): The approximation of the gradient descent by the gradient descent applied to linear Taylor polynomials of the deep neural network and approximation theorems for deep neural network with bounded weights, which en-

able the use of metric entropy bounds to control the generalization error of the estimates. Kohler, Krzyżak and Molinero Römer (2026) derive the rate of convergence results for deep neural network regression estimates learned by gradient descent in case of dependent data. There it is also shown that these estimates achieve a dimension reduction in case that the predictor is exactly contained on a low dimensional manifold. All rates of convergence results shown in the above papers are only optimal up to some arbitrarily small additive term in the exponent of the rate of convergence, which is in contrast to the results for the least squares neural network estimates where it is usually shown that the rates of convergence are optimal up to some logarithmic factor.

The present paper continues this line of research. Its main distinguishing features are a network topology obtained by random pruning of an initially fully connected architecture, the data-dependent selection of potentially non constant step sizes and the dimension reduction in case of predictors contained in a neighborhood of a manifold. Moreover, for (p, C) -smooth regression functions (where, roughly speaking, the regression function is p -times continuously differentiable, cf., Definition 1 below), the resulting estimator achieves the minimax rate $n^{-2p/(2p+d)}$ up to a logarithmic factor.

1.7 Notation

The sets of natural numbers, real numbers and nonnegative real numbers are denoted by $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$, respectively. For $z \in \mathbb{R}$, we denote the smallest integer greater than or equal to z by $\lceil z \rceil$. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$, and the scalar product of $x, y \in \mathbb{R}^d$ is denoted by $\langle x, y \rangle$. And $\|x\|_\infty$ denotes the supremum norm of $x \in \mathbb{R}^d$. For $f : \mathbb{R}^d \rightarrow \mathbb{R}$

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

is its supremum norm.

If X is an $\mathbb{R}^d$ valued random variable, we denote its support by

$$\text{supp}(\mathbf{P}_X) = \left\{ x \in \mathbb{R}^d : \mathbf{P}\{\|X - x\| > \epsilon\} > 0 \text{ for all } \epsilon > 0 \right\}.$$

Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$. A finite collection $f_1, \dots, f_N : \mathbb{R}^d \rightarrow \mathbb{R}$ is called an L_p ϵ -covering of $\mathcal{F}$ on x_1^n if for all $f \in \mathcal{F}$

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

holds. 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 $\mathcal{N}_p(\epsilon, \mathcal{F}, x_1^n)$.

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

1.8 Outline

Section 2 contains the definition of the estimate. The main results are presented in Section 3. In Section 4 we illustrate the finite sample size performance of the estimate by applying it to simulated data. The proofs are given in Section 5.

2 Definition of the estimate

Throughout this paper we use the so-called logistic squasher (1) as our activation function. For $L, r_n \in \mathbb{N}$ let $\mathcal{F}(L, r_n)$ be the class of all multilayer deep neural networks with L layers and r_n neurons per layer and logistic squasher as activation function defined by (2)–(4). In the sequel we describe how our estimate fits such a network to the data.

We start by initializing the weight vector $\mathbf{w}$ of this network. To do this, we set all output weights to zero, i.e., we set

$$(\mathbf{w}^{(0)})_{1,i}^{(L)} = 0 \quad \text{for } i = 1, \dots, r_n. \quad (8)$$

Next we initialize all inner weights of the network by choosing their values uniformly distributed from some interval $[-B, B]$ (where $B > 0$ is a parameter of the estimate), i.e., we choose

$$(\mathbf{w}^{(0)})_{i,j}^{(l)} \sim \text{Unif}([-B, B]) \quad \text{for } l = 1, \dots, L-1, \quad (9)$$

and we initialize all input weights of the network by choosing their values uniformly distributed from some interval $[-A_n, A_n]$ (where $A_n > 0$ is another parameter of the estimate), i.e., we choose

$$(\mathbf{w}^{(0)})_{i,j}^{(0)} \sim \text{Unif}([-A_n, A_n]). \quad (10)$$

Here all random values are chosen from the independent distributions.

After that we randomly prune the network depending on some $r \in \mathbb{N}$, where we assume that $r_n \geq r$. To do this, we choose for each neuron of the network (except for the input neurons) r neurons from the previous level, and set all weights connecting neurons from the previous level and the considered neuron to zero except the r neurons selected. Here the r neurons are randomly selected from the uniform distribution on the r_n neurons of the previous level, and all considered distributions are independent and also independent from the distribution used to define the initial values of the weights. And for the rest of the computation of the estimate we keep all those weights, which we initially have set to zero, at the value zero and do not change them.

Next we train the resulting network using gradient descent. Here we perform $t_n \in \mathbb{N}$ gradient descent steps with step sizes $\lambda_0, \dots, \lambda_{t_n-1} > 0$. To define this formally, let $\mathbf{w}$ and $\mathbf{w}^{(0)}$ be vectors of weights which contain only those weights which have not been set to zero during the pruning step. Let $f_{\mathbf{w}}$ be the network corresponding to the weight vector $\mathbf{w}$, and denote its empirical L_2 risk by

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}}(X_i) - Y_i|^2. \quad (11)$$

Then we compute

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_t \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}) \quad \text{for } t = 0, \dots, t_n - 1, \quad (12)$$

where the step sizes and the number of gradient descent steps are chosen as in Algorithm 1. We define our estimate as truncated network with weight vector $\mathbf{w}^{(t_n)}$. More precisely, we set $\beta_n = c_2 \cdot \log n$ and

$$m_n(x) = T_{\beta_n} f_{\mathbf{w}^{(t_n)}}(x) \quad (x \in \mathbb{R}^d), \quad (13)$$

where $T_{\beta} z = \max\{\min\{z, \beta\}, -\beta\}$ for $z \in \mathbb{R}$ and $\beta \geq 0$.

3 Main results

In our main results we will impose the following smoothness assumption on the regression function.

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

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

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

Let $p, C, c_3, \alpha > 0$. We will impose the following assumptions on the data to which we apply our estimate.

(A1) $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$ are independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ -valued random vectors.

(A2) $\mathbf{P}\{X \in [-\alpha, \alpha]^d\} = 1$.

(A3) $\mathbf{E}\{e^{c_3 \cdot Y^2}\} < \infty$.

(A4) The regression function $m(x) = \mathbf{E}\{Y|X = x\}$ is (p, C) -smooth.

Under the above assumptions we can prove the following bound on the expected L_2 error of our estimate.

Theorem 1 Let $p, C, c_3, \alpha > 0$ be arbitrary and assume that (A1)–(A4) hold. Choose $c_2 > 0$ such that $c_2 \cdot c_3 \geq 2$, let c_4, c_5 and c_6 be sufficiently large and choose $r \in \mathbb{N}$ such that

$$r \geq 4 \cdot (p + d)^2,$$

and let L, r_n, A_n, B be such that

$$L \geq \lceil \log_2(p+d) \rceil, B = c_4, A_n = c_5 \cdot (\log n) \cdot n^{\frac{1}{2p+d}}, \quad (14)$$

$$\frac{r_n}{n^{c_7}} \rightarrow 0 \quad (n \rightarrow \infty) \quad (15)$$

for some $c_7 > 0$ and

$$\frac{r_n}{n^{8 \cdot L \cdot r^{L-1} \cdot (d+1) + 10}} \rightarrow \infty \quad (n \rightarrow \infty) \quad (16)$$

hold. Set $\beta_n = c_2 \cdot \log n$ and define the estimate as in Section 2 with step sizes and number of gradient descent steps chosen data-dependent as in Algorithm 1. Then for any $n \in \mathbb{N}$ it holds

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_6 \cdot (\log n)^{4 \cdot d \cdot L + 3} \cdot n^{-\frac{2p}{2p+d}}.$$

It follows from Stone (1982) that the above rate of convergence is optimal up to some logarithmic factor. For d large this error bound will converge rather slowly to zero. Since the above rate of convergence is optimal up to a logarithmic factor, the only way to circumvent the so-called curse of dimensionality is to impose additional assumptions on the distribution of (X, Y) .

Next we show that our estimate achieves a better rate of convergence in the case of manifold data, where we assume that the predictor X takes on values on some manifold in $\mathbb{R}^d$ of dimension $d^* < d$, which we define as follows:

Definition 2 Let $\mathcal{M} \subseteq \mathbb{R}^d$ be compact and let $d^* \in \{1, \dots, d\}$.

a) We say that $U_1, \dots, U_s$ is an open covering of $\mathcal{M}$, if $U_1, \dots, U_s \subset \mathbb{R}^d$ are open (with respect to the Euclidean topology on $\mathbb{R}^d$) and satisfy

$$\mathcal{M} \subseteq \bigcup_{l=1}^s U_l.$$

b) We say that

$$\psi_1, \dots, \psi_s : [0, 1]^{d^*} \rightarrow \mathbb{R}^d$$

are bi-Lipschitz functions, if there exists $0 < C_{\psi,1} \leq C_{\psi,2} < \infty$ such that

$$C_{\psi,1} \cdot \|x_1 - x_2\| \leq \|\psi_l(x_1) - \psi_l(x_2)\| \leq C_{\psi,2} \cdot \|x_1 - x_2\| \quad (17)$$

holds for any $x_1, x_2 \in [0, 1]^{d^*}$ and any $l \in \{1, \dots, s\}$.

c) We say that $\mathcal{M}$ is a d^* -dimensional Lipschitz-manifold if there exist bi-Lipschitz functions $\psi_i : [0, 1]^{d^*} \rightarrow \mathbb{R}^d$ ($i \in \{1, \dots, s\}$), and an open covering $U_1, \dots, U_s$ of $\mathcal{M}$ such that

$$\psi_l \left((0, 1)^{d^*} \right) = \mathcal{M} \cap U_l$$

holds for all $l \in \{1, \dots, s\}$. Here we call $\psi_1, \dots, \psi_s$ the parametrizations of the manifold.

In applications, often the data is just close to a manifold, but not contained in it (cf., e.g., Carlsson (2009)), so it is more reasonable to assume that the data is contained in a neighborhood of a manifold. To describe this, we use the next definition.

Definition 3 *Let $\delta \geq 0$. We say that the support of X is contained in an approximate d^* -dimensional manifold of order δ , if there exists a d^* -dimensional Lipschitz manifold $\mathcal{M}$ such that the support of X is contained in the δ -neighborhood of $\mathcal{M}$, i.e., such that*

$$\text{supp}(\mathbf{P}_X) \subseteq \mathcal{M}_\delta := \left\{ x \in \mathbb{R}^d : \inf\{\|x - y\| : y \in \mathcal{M}\} \leq \delta \right\}.$$

Our main result in this setting is the following theorem.

Theorem 2 *Let $p, C, c_3, \alpha > 0$ be arbitrary. Choose $c_2 > 0$ such that $c_2 \cdot c_3 \geq 2$, let c_4 and c_5 be sufficiently large and choose $L, r \in \mathbb{N}, B \in \mathbb{R}$ such that*

$$r \geq 4 \cdot (p + d)^2, L \geq \lceil \log_2(p + d) \rceil, B = c_4, A_n = c_5 \cdot (\log n) \cdot n^{\frac{1}{2p+d^*}},$$

and let r_n be such that (15) and (16) hold. Set $\beta_n = c_2 \cdot \log n$ and define the estimate as in Section 2 with step sizes and number of gradient descent steps chosen data-dependent as in Algorithm 1. Then for any $n \in \mathbb{N}$ and any distribution of (X, Y) where (A1), (A3) and (A4) hold and where the support of X is contained in an approximate d^ -dimensional manifold of order*

$$\delta_n = n^{-\frac{1}{2p+d^*}} \cdot (\log n)^{-4L-1}$$

we have

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_6 \cdot (\log n)^{4 \cdot d^* \cdot L + d + 3} \cdot n^{-\frac{2p}{2p+d^*}}.$$

Remark 1. Our estimate uses a polynomial number of neurons (cf., (15)) and a data-dependent number of gradient descent steps determined by Algorithm 1. Since the step size is in each step bounded from below by $\lambda_{\min}/2$, the definition of $\lambda_{\min}$ and (15) imply that also the number of gradient descent steps is bounded from above by some polynomial in the sample size.

4 Application to simulated data

In order to illustrate the finite sample size performance of our newly proposed estimate, we apply it to the following univariate regression problem introduced in Chapter 1 in Györfi et al. (2002).

We define the distribution of an $\mathbb{R} \times \mathbb{R}$ valued random vector (X, Y) by

$$Y = m(X) + \sigma \cdot s(X) \cdot N,$$

Value of c_1	median L_2 error (IQR)
0.1	0.0095 (0.0030)
0.2	0.0068 (0.0029)
0.3	0.0069 (0.0037)
0.4	0.0074 (0.0038)
0.5	0.0065 (0.0024)
0.6	0.0077 (0.0037)
0.7	0.0080 (0.0029)
0.8	0.0080 (0.0040)
0.9	0.0089 (0.0041)
1.0	0.0086 (0.0029)

Table 1: Median L_2 errors (and IQRs) in 50 simulations with $\sigma = 1$, $n = 100$, $r_n = 2000$, $L = 4$, $r = 8$, $A = 1000$ and $B = 20$.

where X is standard normal restricted to $[-1, 1]$,

$$m(x) = \begin{cases} (x+2)^2/2 & \text{if } -1 \leq x < -0.5, \\ x/2 + 0.875 & \text{if } -0.5 \leq x < 0, \\ -5 \cdot (x-0.2)^2 + 1.075 & \text{if } 0 \leq x < 0.5, \\ x + 0.125 & \text{if } 0.5 \leq x \leq 1, \end{cases}$$

$$s(x) = 0.2 - 0.1 \cdot \cos(2 \cdot \pi \cdot x),$$

$\sigma \in \{0.5, 1\}$ and N is a standard normally distributed random variable independent of X .

We construct an independent sample $\mathcal{D}_n$ of sample size $n \in \{100, 200, 400\}$ from this distribution, and apply our estimate defined as in Algorithm 1 to estimate the regression function m from this sample.

In our first simulation we analyze the stopping criteria in the loop in Algorithm 1 where we stop whenever

$$\sum_{s=0}^{t-1} \lambda_s \geq c_1,$$

and we consider $c_1 > 0$ as a parameter of the estimate. We apply our estimate, which we have implemented in R, for various values of this parameter to 50 independent data sets with $\sigma = 1$ and sample size $n = 100$, and compute numerically the L_2 errors of the estimates. The medians and the IQRs of the L_2 errors are reported in Table 1.

We see that c_1 acts like a smoothing parameter for the estimate: if it is too small or too large the error gets larger than if it maintains a moderate value. Since we do not know what the optimal value for this constant is in an application, we consider it together with the value of A as smoothing parameters (for B we will use a moderately large fixed value in the sequel), and will choose both values by splitting of the sample.

We apply our estimate (NNprun) defined in Algorithm 1, which we have implemented in R, with parameters $r_n = 2000$, $L = 4$, $r = 8$ and $B = 20$. We split the data into a learning sample of size $n_l = 0.8 * n$ and a testing sample of size $n_t = 0.2 * n$, compute the estimate for $A \in \{50, 100, 200, 500\}$ and all numbers t of gradient descent steps with $\sum_{s=0}^{t-1} \lambda_s \leq 1$, and choose the value of A and the number of gradient descent steps by minimizing the empirical L_2 risk on the testing data.

We compare our estimate with feedforward neural network estimates NN1, NN3 and NN5 with 1, 3 and 5 hidden layers, respectively. Each layer has a common width r_n and we use the logistic squasher as activation function as well as a linear output layer with a bias term. We again split the data in a learning sample of size $n_l = 0.8 * n$ and a testing sample of size $n_t = 0.2 * n$. Here the number $r_n \in \{10, 25, 50, 100, 200\}$ of hidden neurons and the number $t_n \in \{500, 1000, 1500, 2000, 5000\}$ of gradient descent steps are jointly selected to minimize the empirical L_2 risk on the testing data. All network weights are initialized using the Glorot uniform initialization, whereas all bias terms are initialized at zero. The estimates are implemented in Python using the package *tensorflow* and trained by gradient descent using full-batch Adam with a base learning rate of 0.001.

Furthermore, we compare the estimate with a smoothing spline estimate *smooth-spline* as implemented in R by the procedure *Tps()* from the library *fields*. Here the smoothing parameter of this estimate is chosen data dependent by generalized cross validation as implemented in *Tps()*.

As before we apply the estimates to 50 independent samples with $n \in \{100, 200, 400\}$ and $\sigma \in \{0.5, 1\}$, compute for each estimate numerically its L_2 errors on these 50 samples and report the medians and the IQRs in Table 2.

σ	0.5			1		
<i>sample size</i>	$n = 100$	$n = 200$	$n = 400$	$n = 100$	$n = 200$	$n = 400$
<i>approach</i>	median (IQR)	median (IQR)	median (IQR)	median (IQR)	median (IQR)	median (IQR)
NNprun	0.0040 (0.0022)	0.0025 (0.0006)	0.0018 (0.0004)	0.0086 (0.0046)	0.0056 (0.0022)	0.0035 (0.0012)
NN1	0.0194 (0.0019)	0.0181 (0.0022)	0.0179 (0.0015)	0.0199 (0.0023)	0.0193 (0.0014)	0.0178 (0.0032)
NN3	0.0045 (0.0042)	0.0024 (0.0014)	0.0019 (0.0008)	0.0136 (0.0114)	0.0067 (0.0061)	0.0037 (0.0030)
NN5	0.0044 (0.0027)	0.0025 (0.0017)	0.0014 (0.0005)	0.0146 (0.0093)	0.0075 (0.0079)	0.0039 (0.0018)
smooth- spline	0.0035 (0.0016)	0.0023 (0.0079)	0.0016 (0.0003)	0.0076 (0.0047)	0.0053 (0.0027)	0.0033 (0.0012)

Table 2: Median and IQR of the L_2 errors of the estimates in the univariate regression problem

As we can see from Table 2, our estimate leads in most cases to results which are either better than or rather close to the standard fully connected neural networks NN1, NN3

and NN5, and for large values of n it achieves error bounds similar to the smoothing spline estimate, which can be considered as gold standard for univariate regression estimation problems. This shows that (at least in this example) our completely theoretically motivated deep neural network estimate with data-dependent non-constant step sizes during gradient achieves a reasonable performance.

5 Proofs

5.1 Neural network approximation

We will use the following approximation result from Kohler (2026) to show that our randomly initialized weights lead with high probability to networks with good approximation properties if we properly modify the weights in the output layer.

Lemma 1 *Let $d \in \mathbb{N}$, $p = q + s$ where $s \in (0, 1]$ and $q \in \mathbb{N}_0$, $C > 0$, $\bar{\alpha} \geq 1$ and $A_n, B_n, \gamma_n^* \geq 1$. Let σ be the logistic squasher. For $L, r, K \in \mathbb{N}$ let $\mathcal{F}$ be the set of all networks $f_{\mathbf{w}}$ defined by*

$$f_{\mathbf{w}}(x) = \sum_{j=1}^r w_{1,1,j}^{(L)} \cdot f_{j,1}^{(L)}(x) \quad (18)$$

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

$$f_{k,i}^{(l)}(x) = f_{\mathbf{w},k,i}^{(l)}(x) = \sigma \left(\sum_{j=1}^r w_{k,i,j}^{(l-1)} \cdot f_{k,j}^{(l-1)}(x) + w_{k,i,0}^{(l-1)} \right) \quad (19)$$

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

$$f_{k,i}^{(1)}(x) = f_{\mathbf{w},k,i}^{(1)}(x) = \sigma \left(\sum_{j=1}^d w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)} \right) \quad (20)$$

for some $w_{k,i,0}^{(0)}, \dots, w_{k,i,d}^{(0)} \in \mathbb{R}$, where the weight vector satisfies

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

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

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

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

$$L \geq \lceil \log_2(q + d) \rceil \quad \text{and} \quad r \geq 4 \cdot (p + d) \cdot (q + d),$$

and set

$$A_n = \bar{\alpha} \cdot K \cdot \log K, \quad B_n = c_8 \quad \text{and} \quad \gamma_n^* = c_9 \cdot K^{q+d}.$$

Assume $K \geq c_{10}$ for $c_{10} > 0$ sufficiently large. Then there exists for any (p, C) -smooth $f : \mathbb{R}^d \rightarrow \mathbb{R}$ a neural network $h \in \mathcal{H}$ such that

$$\sup_{x \in [-\bar{\alpha}, \bar{\alpha}]^d} |f(x) - h(x)| \leq \frac{c_{11}}{K^p}.$$

Proof. See Theorem 3 in Kohler (2026). $\square$

In the sequel we show that our initial network contains with high probability subnetworks close to the subnetworks occurring in the network in Lemma 1. Our main result with this respect is the following result.

Lemma 2 *Let $p, C, \alpha > 0$ and assume that (A4) holds. Choose the parameters of the estimate as in Theorem 1. Let $\mathbf{w}^{(0)}$ be the randomly initialized and pruned weight vector of the network. Then outside of an event whose probability is bounded from above by*

$$n^{c_{12}} \cdot \exp(-(\log n)^2)$$

there exists (random) $w_1^, \dots, w_{r_n}^* \in \mathbb{R}$ such that*

$$\sup_{x \in [-\alpha, \alpha]^d} \left| \sum_{k=1}^{r_n} w_k^* \cdot f_{\mathbf{w}^{(0)}, k}^{(L)}(x) - m(x) \right| \leq c_{13} \cdot n^{-\frac{p}{2p+d}}$$

and

$$\sum_{k=1}^{r_n} |w_k^*|^2 \leq \frac{c_{14} \cdot n^{4L \cdot r^{L-1} \cdot (d+1) + 4}}{r_n}.$$

In the proof we will need the following auxiliary result.

Lemma 3 *Let $\alpha, B \geq 1$, and $L, r \in \mathbb{N}$, let σ be the logistic squasher and define $f_{\mathbf{w}, 1}^{(L)}$ by (3) and (4). Assume that $\mathbf{w}$ and $\tilde{\mathbf{w}}$ satisfy*

$$|\mathbf{w}_{i,j}^{(l)}| \leq B \quad (l = 1, \dots, L-1) \tag{21}$$

and

$$\max_{i=1, \dots, r_n} |\{1 \leq j \leq r_n : w_{i,j}^{(l)} \neq 0 \text{ or } \tilde{w}_{i,j}^{(l)} \neq 0\}| \leq r \quad (l = 1, \dots, L-1). \tag{22}$$

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

$$|f_{\mathbf{w}, 1}^{(L)}(x) - f_{\tilde{\mathbf{w}}, 1}^{(L)}(x)| \leq (d+1) \cdot \alpha \cdot (2r+1)^{L-1} \cdot B^{L-1} \cdot \|\mathbf{w} - \tilde{\mathbf{w}}\|_{\infty}.$$

Proof. The logistic squasher is Lipschitz continuous with Lipschitz constant 1, which implies

$$\begin{aligned} |f_{\mathbf{w},i}^{(1)}(x) - f_{\tilde{\mathbf{w}},i}^{(1)}(x)| &\leq \left| \sum_{j=1}^d (w_{i,j}^{(0)} - \tilde{w}_{i,j}^{(0)}) \cdot x^{(j)} + w_{i,0}^{(0)} - \tilde{w}_{i,0}^{(0)} \right| \\ &\leq (d+1) \cdot \alpha \cdot \|\mathbf{w} - \tilde{\mathbf{w}}\|_\infty \end{aligned}$$

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

$$\begin{aligned} |f_{\mathbf{w},i}^{(l)}(x) - f_{\tilde{\mathbf{w}},i}^{(l)}(x)| &\leq \left| \sum_{j=1}^{r_n} w_{i,j}^{(l-1)} \cdot f_{\mathbf{w},j}^{(l-1)}(x) + w_{i,0}^{(l-1)} - \sum_{j=1}^{r_n} \tilde{w}_{i,j}^{(l-1)} \cdot f_{\tilde{\mathbf{w}},j}^{(l-1)}(x) - \tilde{w}_{i,0}^{(l-1)} \right| \\ &\leq \left| \sum_{j=1}^{r_n} w_{i,j}^{(l-1)} \cdot (f_{\mathbf{w},j}^{(l-1)}(x) - f_{\tilde{\mathbf{w}},j}^{(l-1)}(x)) \right| + \left| \sum_{j=1}^{r_n} (w_{i,j}^{(l-1)} - \tilde{w}_{i,j}^{(l-1)}) \cdot f_{\tilde{\mathbf{w}},j}^{(l-1)}(x) \right| \\ &\quad + |w_{i,0}^{(l-1)} - \tilde{w}_{i,0}^{(l-1)}| \\ &= \left| \sum_{j=1, \dots, r_n: w_{i,j}^{(l-1)} \neq 0} w_{i,j}^{(l-1)} \cdot (f_{\mathbf{w},j}^{(l-1)}(x) - f_{\tilde{\mathbf{w}},j}^{(l-1)}(x)) \right| \\ &\quad + \left| \sum_{j=1, \dots, r_n: w_{i,j}^{(l-1)} - \tilde{w}_{i,j}^{(l-1)} \neq 0} (w_{i,j}^{(l-1)} - \tilde{w}_{i,j}^{(l-1)}) \cdot f_{\tilde{\mathbf{w}},j}^{(l-1)}(x) \right| \\ &\quad + |w_{i,0}^{(l-1)} - \tilde{w}_{i,0}^{(l-1)}| \\ &\leq r \cdot B \cdot \max_{j=1, \dots, r_n} |f_{\mathbf{w},j}^{(l-1)}(x) - f_{\tilde{\mathbf{w}},j}^{(l-1)}(x)| + (r+1) \cdot \|\mathbf{w} - \tilde{\mathbf{w}}\|_\infty \end{aligned}$$

where the last inequality follows from (21) and (22). Using induction we conclude for any $l \in \{1, \dots, L\}$

$$|f_{\mathbf{w},i}^{(l)}(x) - f_{\tilde{\mathbf{w}},i}^{(l)}(x)| \leq (d+1) \cdot \alpha \cdot (2r+1)^{l-1} \cdot B^{l-1} \cdot \|\mathbf{w} - \tilde{\mathbf{w}}\|_\infty,$$

which implies the assertion. $\square$

Proof of Lemma 2. W.l.o.g. we assume that n is so large that $A_n \geq B$ holds. Set $\tilde{K}_n = \lceil n^{\frac{1}{2p+d}} \rceil$ and

$$I_n = \lceil \frac{r_n}{n^{4L \cdot r^{L-1} \cdot (d+1) + 1}} \rceil.$$

In the *first step of the proof* we show that there exists a neural network

$$f_{\mathbf{w}}(x) = \sum_{j=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_j \cdot f_{\mathbf{w}_{j,1}}^{(L)}(x) \quad (23)$$

where $f_{\mathbf{w}_{j,1}}^{(L)}$ are recursively defined by (19) and (20) such that

$$\sup_{x \in [-\alpha, \alpha]^d} \left| \sum_{j=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_j \cdot f_{\mathbf{w}_{j,1}}^{(L)}(x) - m(x) \right| \leq c_{15} \cdot n^{-\frac{p}{2p+d}} \quad \text{and} \quad \sum_{j=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_j^2 \leq c_{16} \cdot \frac{n^3}{I_n}.$$

To show this we use Lemma 1 with $\bar{\alpha} = \alpha + 1$. This implies that if we use a linear combination of $r \cdot \tilde{K}_n^d$ fully connected networks of depth L and width r with a single neuron as output, then the corresponding network will approximate m in supremum norm on the cube $[-\alpha, \alpha]^d$ up to an error of order

$$\frac{1}{\tilde{K}_n^p} \leq n^{-\frac{p}{2p+d}}. \quad (24)$$

Here the outer weights are bounded in absolute value by

$$c_{17} \cdot \tilde{K}_n^{q+d} \leq c_{18} \cdot n.$$

Hence if we use a sum of I_n of these networks where we divide all outer weights by I_n , then the corresponding network (23) still approximates m with a supremum norm error of order (24), and in addition it holds

$$\sum_{j=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_j^2 \leq I_n \cdot r \cdot \tilde{K}_n^d \cdot \left(\frac{c_{19} \cdot n}{I_n} \right)^2 = c_{20} \cdot r \cdot \tilde{K}_n^d \cdot n^2 \cdot \frac{1}{I_n} \leq c_{21} \cdot \frac{n^3}{I_n}.$$

Set

$$\epsilon_n = \frac{1}{n^3} \quad \text{and} \quad N_n = 4^{L \cdot r^{L-1} \cdot (d+1)} \cdot \left[n^{4 \cdot L \cdot r^{L-1} \cdot (d+1)} \cdot (\log n)^2 \right].$$

In the *second step of the proof* we show that outside of an event whose probability is bounded from above by

$$I_n \cdot r \cdot \tilde{K}_n^d \cdot \exp \left(-N_n \cdot \left(\frac{1}{4} \cdot \frac{\epsilon_n}{A_n} \right)^{L \cdot r^{L-1} \cdot (d+1)} \right)$$

there exist (random) indices $j_1, \dots, j_{I_n \cdot r \cdot \tilde{K}_n^d} \in \{1, \dots, r_n\}$ such that

$$\sup_{x \in [-\alpha, \alpha]^d} |f_{\mathbf{w}_k, k, 1}^{(L)}(x) - f_{\mathbf{w}^{(0)}, j_k}^{(L)}(x)| \leq (d+1) \cdot \alpha \cdot (2r+1)^{L-1} \cdot B^{L-1} \cdot \epsilon_n$$

$$(k = 1, \dots, I_n \cdot r \cdot \tilde{K}_n^d).$$

(Observe that here $\mathbf{w}^{(0)}$ is the initial weight of the pruned network.) Note that we can rewrite all $f_{\mathbf{w}_k, k, 1}^{(L)}$ such that in each layer $l \in \{1, \dots, L-1\}$ each neuron is connected to exactly one neuron of layer $l+1$ (which is possible by increasing the width of the network) and represent it by a network $\bar{f}_{\bar{\mathbf{w}}_k, k}^{(L)}$ defined by (3) and (4). This implies in particular that each neuron in layer $l \in \{2, \dots, L\}$ is connected to r different neurons in layer $l-1$, which all have no connections to other neurons in layer l than to this neuron. Because $f_{\mathbf{w}_k, k, 1}^{(L)}$ has width r , the newly constructed network $\bar{f}_{\bar{\mathbf{w}}_k, k}^{(L)}$ has the property that its weight vector $\bar{\mathbf{w}}_k$ satisfies

$$\max_{i=1, \dots, r_n} |\{1 \leq j \leq r_n : (\bar{\mathbf{w}}_k)_{i,j}^{(l)} \neq 0\}| \leq r \quad (l = 1, \dots, L-1).$$

Let $\bar{f}_{\bar{\mathbf{w}}_k,k}^{(L)}$ be networks of that type which satisfy $\bar{f}_{\bar{\mathbf{w}}_k,k}^{(L)}(x) = f_{\mathbf{w}_k,k,1}^{(L)}(x)$ ($x \in \mathbb{R}^d$) for all $k = 1, \dots, I_n \cdot r \cdot \tilde{K}_n^d$.

Because of Lemma 3 it suffices to show that there exist random indices $j_1, \dots, j_{I_n \cdot r \cdot \tilde{K}_n^d} \in \{1, \dots, r_n\}$ such that all the weights in $\bar{f}_{\bar{\mathbf{w}}_k,k}^{(L)}$ have distance at most ϵ_n to the corresponding weights in $f_{\mathbf{w}^{(0)},j_k}^{(L)}$. (Here we consider in $f_{\mathbf{w}^{(0)},j_k}^{(L)}$ only those weights which have not been set to zero during the initial pruning.)

To show this, we consider a special way to choose the initial network. We consider the initialization of $f_{\mathbf{w},j}^{(L)}$ successively for $j = 1, \dots, r_n$. When we initialize $f_{\mathbf{w},j}^{(L)}$ for some fixed j , we do this in a top down manner starting at level L : We select randomly k neurons of the previous level which we connect with the current neuron, select randomly their weights from the corresponding uniform distributions and then consider all neurons of the previous level which have yet not been considered during determination of the weights of $f_{\mathbf{w},1}^{(L)}, \dots, f_{\mathbf{w},j-1}^{(L)}$. Since in our initialization in Section 2 the pruning is done independently from the selection of the values of the weights and also the initialization and pruning are done independently for each weight and each neuron, this leads to the same distribution of the random initial network.

If we assume that during consideration of neuron j in each layer at least half of the neurons have not been considered before, then the probability that $f_{\mathbf{w}^{(0)},j}^{(L)}$ has the same topology as $\bar{f}_{\bar{\mathbf{w}}_k,k}^{(L)}$ and simultaneously that all corresponding weights are at a distance at most ϵ_n from one another is lower bounded by

$$\left(\prod_{l=1}^{L-1} \left(\frac{1}{2} \cdot \frac{\epsilon_n}{2B} \right)^{r^l} \right) \cdot \left(\frac{1}{2} \cdot \frac{\epsilon_n}{2A_n} \right)^{r^{L-1} \cdot (d+1)} \geq \left(\frac{1}{2} \cdot \frac{\epsilon_n}{2A_n} \right)^{L \cdot r^{L-1} \cdot (d+1)}.$$

Here the factor $1/2$ takes into account that we have to choose neurons which have not been chosen in the previous steps, i.e., we have to choose neurons from a set of neurons which contains at least half of all possible r_n neurons, which occurs with probability at least $1/2$ since we choose the neurons from an uniform distribution on the set of all neurons. Consequently the probability that none of the first N_n networks $f_{\mathbf{w}^{(0)},j,1}^{(L)}$ ($j = 1, \dots, N_n$) has the same topology as $\bar{f}_{\bar{\mathbf{w}}_k,k}^{(L)}$ and simultaneously that not all corresponding weights are at a distance at most ϵ_n from one another is bounded from above by

$$\left(1 - \left(\frac{1}{4} \cdot \frac{\epsilon_n}{A_n} \right)^{L \cdot r^{L-1} \cdot (d+1)} \right)^{N_n} \leq \exp \left(-N_n \cdot \left(\frac{1}{4} \cdot \frac{\epsilon_n}{A_n} \right)^{L \cdot r^{L-1} \cdot (d+1)} \right)$$

provided during selection of each of the N_n networks there are in all layers at least half of the neurons not considered.

Since w.l.o.g.

$$r_n \geq 2 \cdot N_n \cdot I_n \cdot r \cdot \tilde{K}_n^d \cdot r^L,$$

the last condition is satisfied if we consider successively all networks $\bar{f}_{\bar{\mathbf{w}}_k,k}^{(L)}$ ($k = 1, \dots, I_n \cdot r \cdot \tilde{K}_n^d$) and we can conclude: The probability that there exists $k \in \{1, \dots, I_n \cdot r \cdot \tilde{K}_n^d\}$ such

that none of the networks $f_{\mathbf{w}^{(0)},j,1}^{(L)}$ has the same topology as $\bar{f}_{\bar{\mathbf{w}}_k,k}^{(L)}$ and simultaneously that not all corresponding weights are at a distance at most ϵ_n from one another is upper bounded by

$$I_n \cdot r \cdot \tilde{K}_n^d \cdot \exp \left(-N_n \cdot \left(\frac{1}{4} \cdot \frac{\epsilon_n}{A_n} \right)^{L \cdot r^{L-1} \cdot (d+1)} \right).$$

In the *third step of the proof* we show the assertion. Let

$$f_{\mathbf{w}}(x) = \sum_{j=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_j \cdot f_{\mathbf{w},j,1}^{(L)}(x)$$

be the neural network of step 1, and let $j_1, \dots, j_{I_n \cdot r \cdot \tilde{K}_n^d} \in \{1, \dots, K_n\}$ be the random indices from step 2, and set

$$\mathbf{w}_j^* = \begin{cases} w_k, & \text{if } j = j_k \text{ for some } k \in \{1, \dots, I_n \cdot r \cdot \tilde{K}_n^d\}, \\ 0, & \text{else.} \end{cases}$$

Then

$$\sum_{k=1}^{r_n} |w_k^*|^2 = \sum_{j=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_j^2 \leq c_{22} \cdot \frac{n^3}{I_n} \leq \frac{c_{23} \cdot n^{4L \cdot r^{L-1} \cdot (d+1) + 4}}{r_n},$$

and outside of an event whose probability is bounded from above by

$$I_n \cdot r \cdot \tilde{K}_n^d \cdot \exp \left(-N_n \cdot \left(\frac{1}{4} \cdot \frac{\epsilon_n}{A_n} \right)^{L \cdot r^{L-1} \cdot (d+1)} \right) \leq n^{c_{24}} \cdot \exp(-(\log n)^2)$$

we have

$$\begin{aligned} & \sup_{x \in [-\alpha, \alpha]^d} \left| \sum_{k=1}^{r_n} w_k^* \cdot f_{\mathbf{w}^{(0)},k}^{(L)}(x) - m(x) \right| \\ & \leq \sup_{x \in [-\alpha, \alpha]^d} \left| \sum_{k=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_k \cdot f_{\mathbf{w}^{(0)},j_k}^{(L)}(x) - \sum_{k=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_k \cdot f_{\mathbf{w}_k,k,1}^{(L)}(x) \right| \\ & \quad + \sup_{x \in [-\alpha, \alpha]^d} \left| \sum_{k=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_k \cdot f_{\mathbf{w}_k,k,1}^{(L)}(x) - m(x) \right| \\ & \leq I_n \cdot r \cdot \tilde{K}_n^d \cdot \frac{c_{25} \cdot n}{I_n} \cdot (d+1) \cdot \alpha \cdot (2r+1)^{l-1} \cdot B^{l-1} \cdot \epsilon_n + c_{26} \cdot n^{-\frac{p}{2p+d}} \\ & \leq c_{27} \cdot n^{-\frac{p}{2p+d}}, \end{aligned}$$

which implies the assertion. □

5.2 Neural network generalization

Our next lemma is our main tool to analyze the generalization error of our estimate. It extends the previous result of Kohler (2026) to the more general topology considered in our paper, and sharpens the bound there by computing explicitly the constant in the exponent of the upper bound on the covering number, which will allow us in the proof of the main result to choose some parameter k as a function of the sample size. The main idea of the proof is to approximate over-parametrized deep neural networks with bounded weights by the piecewise Taylor polynomials and use the fact that the covering number of these piecewise Taylor polynomials can be easily bounded.

Lemma 4 *Let $L, r_n, r \in \mathbb{N}$, let $\alpha, \beta, A, B, C \geq 1$. Let σ be the logistic squasher and let $\mathcal{F}$ be the set of all neural networks $f_{\mathbf{w}}$ defined by (2)–(4), where the weight vector $\mathbf{w}$ satisfies*

$$\sum_{i=1}^{r_n} |w_{1,i}^{(L)}| \leq C, \quad (25)$$

$$|w_{i,j}^{(l)}| \leq B \quad \text{for all } i, j = 1, \dots, r_n, l = 1, \dots, L-1, \quad (26)$$

$$|w_{i,j}^{(0)}| \leq A \quad \text{for all } i = 1, \dots, r_n, j = 1, \dots, d \quad (27)$$

and

$$|\{j \in \{1, \dots, r_n\} : w_{i,j}^{(l)} \neq 0\}| \leq r \quad \text{for all } i = 1, \dots, r_n, l = 1, \dots, L-1. \quad (28)$$

Then we have for any $0 < \epsilon \leq 1$, any $x_1, \dots, x_n \in [-\alpha, \alpha]^d$ and any $k \in \mathbb{N}$

$$\mathcal{N}(\epsilon, T_\beta \mathcal{F}, x_1^n) \leq 3 \left(\frac{12e \cdot \beta}{\epsilon} \right)^{2 \cdot (4ed)^d \cdot 2^{L \cdot d} \cdot (r+k)^{3 \cdot d \cdot L} \cdot \alpha^d \cdot A^d \cdot B^{(L-1) \cdot d} \cdot \left(\frac{C}{\epsilon}\right)^{\frac{d}{k} + 2}}.$$

Proof. In the *first step of the proof* we show for any $s \in \mathbb{N}$

$$\max_{x \in \mathbb{R}} |\sigma^{(s)}(x)| \leq 2^{s-1} \cdot (s+1)!$$

To do this we show by induction

$$\sigma^{(s)}(x) = \sum_{l=1}^{s+1} a_{s,l} \cdot (\sigma(x))^l$$

for some $a_{s,l} \in \mathbb{R}$ which satisfy

$$\max_{l=1, \dots, s+1} |a_{s,l}| \leq 2^{s-1} \cdot s!$$

For $s = 1$ we have

$$\sigma'(x) = (-1) \cdot (1+e^{-x})^{-2} \cdot e^{-x} \cdot (-1) = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \sigma(x) \cdot (1-\sigma(x)) = \sigma(x) - (\sigma(x))^2,$$

so the assertion trivially holds.

Assume now that the assertion holds for some $s \in \mathbb{N}$. Then

$$\begin{aligned}
& \sigma^{(s+1)}(x) \\
&= \sum_{l=1}^{s+1} a_{s,l} \cdot l \cdot (\sigma(x))^{l-1} \cdot \sigma'(x) \\
&= \sum_{l=1}^{s+1} a_{s,l} \cdot l \cdot (\sigma(x))^{l-1} \cdot (\sigma(x) - (\sigma(x))^2) \\
&= \sum_{l=1}^{s+1} a_{s,l} \cdot l \cdot (\sigma(x))^l - \sum_{l=1}^{s+1} a_{s,l} \cdot l \cdot (\sigma(x))^{l+1} \\
&= a_{s,1} \cdot \sigma(x) + \sum_{l=2}^{s+1} (a_{s,l} \cdot l - a_{s,l-1} \cdot (l-1)) \cdot (\sigma(x))^l - a_{s,s+1} \cdot (s+1) \cdot (\sigma(x))^{s+2},
\end{aligned}$$

which implies

$$\sigma^{(s+1)}(x) = \sum_{l=1}^{s+2} a_{s+1,l} \cdot (\sigma(x))^s$$

where $a_{s+1,l}$ satisfy

$$\max_{l=1,\dots,s+2} |a_{s+1,l}| \leq 2 \cdot (s+1) \cdot \max_{l=1,\dots,s+1} |a_{s,l}| \leq 2 \cdot (s+1) \cdot 2^{s-1} \cdot s! = 2^s \cdot (s+1)!$$

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

$$\left| \frac{\partial^k f_{\mathbf{w}}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \right| \leq (r+k)^{3 \cdot k \cdot L} \cdot \left(2^{k-1}\right)^L \cdot C \cdot B^{(L-1) \cdot k} \cdot A^k =: c. \quad (29)$$

Because of (25) and

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

it suffices to show for all $i \in \{1, \dots, r_n\}$

$$\left| \frac{\partial^k f_{\mathbf{w},i}^{(L)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \right| \leq (r+k)^{3 \cdot k \cdot L} \cdot \left(2^{k-1}\right)^L \cdot B^{(L-1) \cdot k} \cdot A^k. \quad (30)$$

To show (30) we will demonstrate for any $i \in \{1, \dots, r_n\}$, any $s \in \{1, \dots, L\}$, any $k \in \mathbb{N}$ and any $s_1, \dots, s_k \in \{1, \dots, d\}$

$$\left| \frac{\partial^k f_{\mathbf{w},i}^{(s)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \right| \leq (r+k)^{3 \cdot k \cdot s} \cdot \left(2^{k-1}\right)^s \cdot B^{(s-1) \cdot k} \cdot A^k. \quad (31)$$

For $s = 1$ we have

$$\frac{\partial^k f_{\mathbf{w},i}^{(1)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) = \prod_{j=1}^k w_{i,s_j}^{(0)} \cdot \sigma^{(k)} \left(\sum_{t=1}^d w_{i,t}^{(0)} \cdot x^{(t)} + w_{i,0}^{(0)} \right)$$

from which we get (31) for all $k \in \mathbb{N}$ by (27) and the assertion of the first step of the proof.

Let $l \in \{2, \dots, L\}$ and assume that (31) holds for $s = 1, \dots, l-1$, all $i \in \{1, \dots, r_n\}$, all $k \in \mathbb{N}$ and all $s_1, \dots, s_k \in \{1, \dots, d\}$. We have

$$\begin{aligned} & \frac{\partial f_{\mathbf{w},i}^{(l)}}{\partial x^{(s)}}(x) \\ &= \sigma' \left(\sum_{t=1}^{r_n} w_{i,t}^{(l-1)} \cdot f_{\mathbf{w},t}^{(l-1)}(x) + w_{i,0}^{(l-1)} \right) \cdot \sum_{j=1}^{r_n} w_{i,j}^{(l-1)} \cdot \frac{\partial f_{\mathbf{w},j}^{(l-1)}}{\partial x^{(s)}}(x) \\ &= \sum_{j=1, \dots, r_n: w_{i,j}^{(l-1)} \neq 0} w_{i,j}^{(l-1)} \cdot \sigma' \left(\sum_{t=1}^{r_n} w_{i,t}^{(l-1)} \cdot f_{\mathbf{w},t}^{(l-1)}(x) + w_{i,0}^{(l-1)} \right) \cdot \frac{\partial f_{\mathbf{w},j}^{(l-1)}}{\partial x^{(s)}}(x). \end{aligned}$$

Because of (28) the sum on the right hand side above has at most r terms. Using the product rule of derivation we can conclude for any $k \in \mathbb{N}$ that

$$\frac{\partial^k f_{\mathbf{w},i}^{(l)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \quad (32)$$

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

$$\begin{aligned} & w \cdot \sigma^{(s)} \left(\sum_{j=1}^{r_n} w_{i,j}^{(l-1)} \cdot f_{\mathbf{w},j}^{(l-1)}(x) + w_{i,0}^{(l-1)} \right) \cdot \frac{\partial^{t_1} f_{\mathbf{w},j_1}^{(l-1)}}{\partial x^{(r_{1,1})} \dots \partial x^{(r_{1,t_1})}}(x) \cdot \\ & \dots \cdot \frac{\partial^{t_s} f_{\mathbf{w},j_s}^{(l-1)}}{\partial x^{(r_{s,1})} \dots \partial x^{(r_{s,t_s})}}(x) \end{aligned} \quad (33)$$

where we have $s \in \{1, \dots, k\}$, $|w| \leq B^s$ and $t_1 + \dots + t_s = k$. The induction hypothesis and the first step of the proof imply that the absolute value of (33) is upper bounded by

$$\begin{aligned} & B^s \cdot 2^{s-1} \cdot (s+1)! \cdot \prod_{j=1}^s (r+t_j)^{3 \cdot t_j \cdot (l-1)} \cdot (2^{t_j-1})^{l-1} \cdot B^{(l-2) \cdot t_j} \cdot A^{t_j} \\ & \leq B^k \cdot 2^{k-1} \cdot (k+1)! \cdot \prod_{j=1}^s (r+k)^{3 \cdot t_j \cdot (l-1)} \cdot (2^{t_j-1})^{l-1} \cdot B^{(l-2) \cdot t_j} \cdot A^{t_j} \\ & \leq (r+k)^{3 \cdot k \cdot (l-1)} \cdot (k+1)! \cdot (2^{k-1})^l \cdot B^{k \cdot (l-1)} \cdot A^k, \end{aligned}$$

hence

$$\left| \frac{\partial^k f_{\mathbf{w},i}^{(l)}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x) \right|$$

$$\begin{aligned}
&\leq (r+k)^k \cdot (r+k)^{3 \cdot k \cdot (l-1)} \cdot (k+1)! \cdot (2^{k-1})^l \cdot B^{k \cdot (l-1)} \cdot A^k \\
&\leq (r+k)^{3 \cdot k \cdot l} \cdot (2^{k-1})^l \cdot B^{k \cdot (l-1)} \cdot A^k,
\end{aligned}$$

where we have used

$$(k+1)! \leq (k+1)^{k+1} \leq (r+k)^{2k}.$$

Let $\mathcal{G}$ be the set of all polynomials of degree less than or equal to $k-1$, let Π be a partition of $[-\alpha, \alpha]^d$ into cubes of sidelength

$$\left(\frac{k!}{2d^k} \cdot \frac{\epsilon}{c} \right)^{1/k} \leq 1,$$

and let $\mathcal{G} \circ \Pi$ be the class of all functions whose restriction on each cube in Π lies in $\mathcal{G}$ and which are zero outside of $[-\alpha, \alpha]^d$.

In the *fourth step of the proof* we show

$$\mathcal{N}_p(\epsilon, \{T_\beta f : f \in \mathcal{F}\}, x_1^n) \leq \mathcal{N}_p\left(\frac{\epsilon}{2}, T_\beta \mathcal{G} \circ \Pi, x_1^n\right). \quad (34)$$

Let $(Tf_{\mathbf{w}})_{k-1,u}$ be the multivariate Taylor polynomial of $f_{\mathbf{w}}$ of degree $k-1$ around $u \in \mathbb{R}^d$, i.e.

$$\begin{aligned}
&(Tf_{\mathbf{w}})_{k-1,u}(x) \\
&= \sum_{\substack{j_1, \dots, j_d \in \mathbb{N}_0, \\ j_1 + \dots + j_d \leq k-1}} \frac{1}{j_1! \dots j_d!} \cdot \frac{\partial^{j_1 + \dots + j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \cdot (x^{(1)} - u^{(1)})^{j_1} \dots (x^{(d)} - u^{(d)})^{j_d}.
\end{aligned}$$

For each $I \in \Pi$ fix some $u \in I$. By the multivariate Taylor theorem we get as in the proof of Lemma 1 in Kohler (2014) for each $x \in I$

$$\begin{aligned}
&|f_{\mathbf{w}}(x) - (Tf_{\mathbf{w}})_{k-1,u}(x)| \\
&= \left| f_{\mathbf{w}}(x) - (Tf_{\mathbf{w}})_{k-2,u}(x) \right. \\
&\quad \left. - \sum_{\substack{j_1, \dots, j_d \in \mathbb{N}_0, \\ j_1 + \dots + j_d = k-1}} \frac{1}{j_1! \dots j_d!} \cdot \frac{\partial^{j_1 + \dots + j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \cdot (x^{(1)} - u^{(1)})^{j_1} \dots (x^{(d)} - u^{(d)})^{j_d} \right| \\
&= \left| \sum_{\substack{j_1, \dots, j_d \in \mathbb{N}_0, \\ j_1 + \dots + j_d = k-1}} \frac{k-1}{j_1! \dots j_d!} \cdot \int_0^1 (1-t)^{k-2} \cdot \frac{\partial^{j_1 + \dots + j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u + t \cdot (x - u)) dt \right. \\
&\quad \left. \cdot (x^{(1)} - u^{(1)})^{j_1} \dots (x^{(d)} - u^{(d)})^{j_d} \right. \\
&\quad \left. - \sum_{\substack{j_1, \dots, j_d \in \mathbb{N}_0, \\ j_1 + \dots + j_d = k-1}} \frac{1}{j_1! \dots j_d!} \cdot \frac{\partial^{j_1 + \dots + j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \cdot (x^{(1)} - u^{(1)})^{j_1} \dots (x^{(d)} - u^{(d)})^{j_d} \right| \\
&= \left| \sum_{\substack{j_1, \dots, j_d \in \mathbb{N}_0, \\ j_1 + \dots + j_d = k-1}} \frac{k-1}{j_1! \dots j_d!} \cdot (x^{(1)} - u^{(1)})^{j_1} \dots (x^{(d)} - u^{(d)})^{j_d} \right.
\end{aligned}$$

$$\begin{aligned}
& \cdot \int_0^1 (1-t)^{k-2} \cdot \left(\frac{\partial^{j_1+\dots+j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u+t \cdot (x-u)) - \frac{\partial^{j_1+\dots+j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \right) dt \Big| \\
& \leq \sum_{\substack{j_1, \dots, j_d \in \mathbb{N}_0, \\ j_1 + \dots + j_d = k-1}} \frac{k-1}{j_1! \dots j_d!} \cdot |x^{(1)} - u^{(1)}|^{j_1} \dots |x^{(d)} - u^{(d)}|^{j_d} \\
& \quad \cdot \int_0^1 (1-t)^{k-2} \cdot \left| \frac{\partial^{j_1+\dots+j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u+t \cdot (x-u)) - \frac{\partial^{j_1+\dots+j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \right| dt.
\end{aligned}$$

By the mean value theorem and (29) we get for any $j_1, \dots, j_d \in \mathbb{N}_0$ with $j_1 + \dots + j_d = k-1$

$$\begin{aligned}
& \left| \frac{\partial^{j_1+\dots+j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u+t \cdot (x-u)) - \frac{\partial^{j_1+\dots+j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \right| \\
& \leq \sum_{i=1}^d \left| \frac{\partial^{j_1+\dots+j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u^{(1)}, \dots, u^{(i-1)}, u^{(i)} + t \cdot (x^{(i)} - u^{(i)}), \dots, u^{(d)} + t \cdot (x^{(d)} - u^{(d)})) \right. \\
& \quad \left. - \frac{\partial^{j_1+\dots+j_d} f_{\mathbf{w}}}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u^{(1)}, \dots, u^{(i)}, u^{(i+1)} + t \cdot (x^{(i+1)} - u^{(i+1)}), \dots, u^{(d)} + t \cdot (x^{(d)} - u^{(d)})) \right| \\
& \leq \sum_{i=1}^d c \cdot |u^{(i)} + t \cdot (x^{(i)} - u^{(i)}) - u^{(i)}| \\
& = c \cdot t \cdot \sum_{i=1}^d |x^{(i)} - u^{(i)}|.
\end{aligned}$$

Using

$$\int_0^1 (k-1) \cdot t \cdot (1-t)^{k-2} dt = \frac{1}{k}$$

we get

$$\begin{aligned}
& |f_{\mathbf{w}}(x) - (Tf_{\mathbf{w}})_{k-1,u}(x)| \\
& \leq \sum_{\substack{j_1, \dots, j_d \in \mathbb{N}_0, \\ j_1 + \dots + j_d = k-1}} \frac{k-1}{j_1! \dots j_d!} \cdot c \cdot \sum_{i=1}^d |x^{(i)} - u^{(i)}| \cdot |x^{(1)} - u^{(1)}|^{j_1} \dots |x^{(d)} - u^{(d)}|^{j_d} \\
& \quad \cdot \int_0^1 t \cdot (1-t)^{k-2} dt \\
& = \frac{c}{k} \cdot \sum_{i=1}^d |x^{(i)} - u^{(i)}| \cdot \frac{1}{(k-1)!} \cdot \sum_{\substack{j_1, \dots, j_d \in \mathbb{N}_0, \\ j_1 + \dots + j_d = k-1}} \frac{(k-1)!}{j_1! \dots j_d!} \cdot |x^{(1)} - u^{(1)}|^{j_1} \dots |x^{(d)} - u^{(d)}|^{j_d} \\
& = \frac{c}{k!} \cdot \sum_{i=1}^d |x^{(i)} - u^{(i)}| \cdot \left(\sum_{j=1}^d |x^{(j)} - u^{(j)}| \right)^{k-1} \\
& = \frac{c}{k!} \cdot \left(\sum_{i=1}^d |x^{(i)} - u^{(i)}| \right)^k
\end{aligned}$$

$$\leq c \cdot \frac{d^k}{k!} \cdot \|x - u\|_\infty^k \leq c \cdot \frac{d^k}{k!} \cdot \frac{k!}{2 \cdot d^k} \cdot \frac{\epsilon}{c} = \frac{\epsilon}{2},$$

where for the proof of the second equality above we have used the multinomial theorem. By repeating this argument for every cube $I \in \Pi$ we see 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 $x \in \mathbb{R}^d$, which implies (34).

In the *third step* of the proof we show the assertion of Lemma 4. $\mathcal{G} \circ \Pi$ is a linear vector space of dimension less than or equal to

$$k^d \cdot \left(\frac{4\alpha}{\left(\frac{k!}{2d^k} \cdot \frac{\epsilon}{c}\right)^{1/k}} \right)^d = 4^d \cdot 2^{d/k} \cdot d^d \cdot \frac{k^d}{(k!)^{d/k}} \cdot \alpha^d \cdot \left(\frac{c}{\epsilon}\right)^{\frac{d}{k}}.$$

Since

$$e^k \geq \frac{k^k}{k!}, \quad i.e., \quad k! \geq \frac{k^k}{e^k}$$

this in turn is bounded from above by

$$(4ed)^d \cdot 2^{d/k} \cdot \alpha^d \cdot \left(\frac{c}{\epsilon}\right)^{\frac{d}{k}}.$$

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

$$\mathcal{N}\left(\frac{\epsilon}{2}, T_\beta \mathcal{G} \circ \Pi, x_1^n\right) \leq 3 \left(\frac{2e(2\beta)}{(\epsilon/2)} \log \left(\frac{3e(2\beta)}{(\epsilon/2)} \right) \right)^{(4ed)^d \cdot 2^{d/k} \cdot \alpha^d \cdot \left(\frac{c}{\epsilon}\right)^{\frac{d}{k}+1}}.$$

We have

$$\begin{aligned} c^{\frac{d}{k}} &= \left((r+k)^{3 \cdot k \cdot L} \cdot \left(2^{k-1}\right)^L \right)^{\frac{d}{k}} \cdot C^{\frac{d}{k}} \cdot B^{(L-1) \cdot d} \cdot A^d \\ &= (r+k)^{3 \cdot d \cdot L} \cdot 2^{L \cdot d} \cdot 2^{-L \cdot \frac{d}{k}} \cdot A^d \cdot B^{(L-1) \cdot d} \cdot C^{\frac{d}{k}}, \end{aligned}$$

hence

$$\mathcal{N}\left(\frac{\epsilon}{2}, T_\beta \mathcal{G} \circ \Pi, x_1^n\right) \leq 3 \left(\frac{12e \cdot \beta}{\epsilon} \right)^{2 \cdot (4ed)^d \cdot 2^{L \cdot d} \cdot (r+k)^{3 \cdot d \cdot L} \cdot \alpha^d \cdot A^d \cdot B^{(L-1) \cdot d} \cdot \left(\frac{C}{\epsilon}\right)^{\frac{d}{k}+2}}.$$

Together with (34) this implies the assertion. □

5.3 Neural network optimization

Our next lemma is our main tool for analyzing the optimization error of the estimate. It extends results from Kohler (2026) and Kohler, Krzyżak and Molinero Römer (2026) to the case of non-constant step sizes.

Lemma 5 *Let $d, J_n \in \mathbb{N}$, and for*

$$\mathbf{w} = (w_1, \dots, w_{J_n}) \in \mathbb{R}^{J_n}$$

let $f_{\mathbf{w}} : \mathbb{R}^d \rightarrow \mathbb{R}$ be a (deep) neural network with weight vector $\mathbf{w}$. Let F_n be defined by (11), let $\mathbf{w}^{(0)} \in \mathbb{R}^{J_n}$, $t_n \in \mathbb{N}$ and $\lambda_0, \dots, \lambda_{t_n-1} > 0$, and set

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_t \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}) \quad \text{for } t = 0, \dots, t_n - 1. \quad (35)$$

Let $\mathbf{w}^ \in \mathbb{R}^{J_n}$. Assume*

$$F_n(\mathbf{w}^{(s+1)}) \leq F_n(\mathbf{w}^{(s)}) - \frac{\lambda_s}{2} \cdot \left\| \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)}) \right\|^2 \quad (36)$$

for $s = 0, 1, \dots, t_n - 1$,

$$\sum_{j=1}^{J_n} \left| \frac{\partial}{\partial w^{(j)}} f_{\mathbf{w}_1}(x) - \frac{\partial}{\partial w^{(j)}} f_{\mathbf{w}_2}(x) \right|^2 \leq C_n^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2 \quad (37)$$

for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^{J_n}$ with $\max\{\|\mathbf{w}^ - \mathbf{w}_1\|, \|\mathbf{w}^* - \mathbf{w}_2\|\} \leq \|\mathbf{w}^* - \mathbf{w}^{(0)}\|$ and all $x \in \{X_1, \dots, X_n\}$,*

$$|Y_i| \leq \beta_n \quad (i = 1, \dots, n), \quad (38)$$

$$|f_{\mathbf{w}^*}(X_i)| \leq \beta_n \quad (i = 1, \dots, n) \quad (39)$$

and

$$C_n \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 \leq \beta_n. \quad (40)$$

Then

$$\begin{aligned} F_n(\mathbf{w}^{(t_n)}) &\leq F_n(\mathbf{w}^*) + \left(5 \cdot \beta_n \cdot C_n + \frac{1}{2 \cdot \sum_{s=0}^{t_n-1} \lambda_s} \right) \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 \\ &\quad + \frac{\max_{s=0, \dots, t_n-1} \lambda_s}{\sum_{s=0}^{t_n-1} \lambda_s} \cdot F_n(\mathbf{w}^{(0)}). \end{aligned}$$

In the proof of Lemma 5 we will need the following auxiliary result.

Lemma 6 *Let $\mathbf{w}^{(0)}, \mathbf{w}^* \in \mathbb{R}^{J_n}$, define $\mathbf{w}^{(t+1)}$ by (35) for $t = 0, \dots, t_n - 1$ and assume that (36) holds for $s = 0, \dots, t - 1$. Let*

$$f_{lin, \mathbf{w}, \tilde{\mathbf{w}}}(x) = f_{\mathbf{w}}(x) + \sum_{j=1}^{J_n} \frac{\partial f_{\mathbf{w}}(x)}{\partial w^{(j)}} \cdot (\tilde{\mathbf{w}}^{(j)} - \mathbf{w}^{(j)}) \quad (41)$$

be the linear Taylor polynomial of $f_{\mathbf{w}}(x)$ around $\mathbf{w}$, set

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

and assume

$$|f_{lin,\mathbf{w},\mathbf{w}^*}(x) - f_{\mathbf{w}^*}(x)| \leq C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2 \quad (42)$$

for all $x \in \{X_1, \dots, X_n\}$ and all $\mathbf{w} \in \mathbb{R}^{J_n}$ with $\|\mathbf{w}^* - \mathbf{w}\| \leq \|\mathbf{w}^* - \mathbf{w}^{(0)}\|$, and

$$F_n(\mathbf{w}^{(s+1)}) \geq F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) \quad (43)$$

for $s = 0, 1, \dots, t-1$. Then

$$\|\mathbf{w}^{(s)} - \mathbf{w}^*\| \leq \|\mathbf{w}^{(0)} - \mathbf{w}^*\| \quad \text{for } s = 1, \dots, t.$$

Proof. The proof is a straightforward modification of the proof of Lemma 2 in Kohler, Krzyżak and Molinero Römer (2026). For the sake of completeness we nevertheless present a complete proof.

Let $s \in \{0, 1, \dots, t-1\}$ be arbitrary. We have

$$F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}) = F_n(\mathbf{w}^{(s)}) \quad \text{and} \quad \nabla_{\mathbf{w}} F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}) = \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)})$$

and (since $\mathbf{w} \mapsto F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w})$ is convex and differentiable)

$$F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}) - F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) \leq \langle \nabla_{\mathbf{w}} F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}), \mathbf{w}^{(s)} - \mathbf{w}^* \rangle,$$

which implies

$$\begin{aligned} \langle \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)}), \mathbf{w}^{(s)} - \mathbf{w}^* \rangle &= \langle \nabla_{\mathbf{w}} F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}), \mathbf{w}^{(s)} - \mathbf{w}^* \rangle \\ &\geq F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}) - F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) \\ &= F_n(\mathbf{w}^{(s)}) - F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*). \end{aligned}$$

Hence

$$\begin{aligned} &\|\mathbf{w}^{(s+1)} - \mathbf{w}^*\|^2 \\ &= \left\| \mathbf{w}^{(s)} - \lambda_s \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)}) - \mathbf{w}^* \right\|^2 \\ &= \left\| \mathbf{w}^{(s)} - \mathbf{w}^* \right\|^2 - 2 \cdot \lambda_s \cdot \langle \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)}), \mathbf{w}^{(s)} - \mathbf{w}^* \rangle + \lambda_s^2 \cdot \left\| \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)}) \right\|^2 \\ &\leq \left\| \mathbf{w}^{(s)} - \mathbf{w}^* \right\|^2 - 2 \cdot \lambda_s \left(F_n(\mathbf{w}^{(s)}) - F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) \right) \\ &\quad + 2 \cdot \lambda_s \cdot (F_n(\mathbf{w}^{(s)}) - F_n(\mathbf{w}^{(s+1)})) \\ &= \left\| \mathbf{w}^{(s)} - \mathbf{w}^* \right\|^2 - 2 \cdot \lambda_s \left(F_n(\mathbf{w}^{(s+1)}) - F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) \right) \\ &\leq \left\| \mathbf{w}^{(s)} - \mathbf{w}^* \right\|^2. \end{aligned}$$

Here we have used (36) in the first inequality and (43) in the second inequality. $\square$

Proof of Lemma 5. The proof is an extension of the proof of Lemma 1 in Kohler, Krzyżak and Molinero Römer (2026).

Set

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

where $f_{lin,\mathbf{w},\mathbf{w}^*}$ is defined by (41).

In the *first step of the proof* we show that (37) implies (42), and consequently the assumptions of Lemma 6 are satisfied under the assumptions of Lemma 5 provided (43) holds.

To do this, we proceed as in the proof of Lemma 1 in Kohler (2026). Let $\mathbf{w} \in \mathbb{R}^{J_n}$ with $\|\mathbf{w}^* - \mathbf{w}\| \leq \|\mathbf{w}^* - \mathbf{w}^{(0)}\|$. Set

$$H(s) = f_{\mathbf{w}+s(\mathbf{w}^*-\mathbf{w})}(x) \quad \text{for } s \in [0, 1].$$

Then

$$\max\{\|\mathbf{w}^* - \mathbf{w}\|, \|\mathbf{w}^* - (\mathbf{w} + s \cdot (\mathbf{w}^* - \mathbf{w}))\|\} = \max\{\|\mathbf{w}^* - \mathbf{w}\|, (1-s) \cdot \|\mathbf{w}^* - \mathbf{w}\|\} \\ \leq \|\mathbf{w}^* - \mathbf{w}^{(0)}\|$$

for all $s \in [0, 1]$, hence we can conclude from (37)

$$\begin{aligned} & |f_{\mathbf{w}^*}(x) - f_{lin,\mathbf{w},\mathbf{w}^*}(x)| \\ &= |f_{\mathbf{w}^*}(x) - f_{\mathbf{w}}(x) - \sum_{j=1}^{J_n} \frac{\partial f_{\mathbf{w}}(x)}{\partial \mathbf{w}^{(j)}} \cdot ((\mathbf{w}^*)^{(j)} - \mathbf{w}^{(j)})| \\ &= |H(1) - H(0) - \sum_{j=1}^{J_n} \frac{\partial f_{\mathbf{w}}(x)}{\partial \mathbf{w}^{(j)}} \cdot ((\mathbf{w}^*)^{(j)} - \mathbf{w}^{(j)})| \\ &= \left| \int_0^1 H'(s) ds - \sum_{j=1}^{J_n} \frac{\partial f_{\mathbf{w}}(x)}{\partial \mathbf{w}^{(j)}} \cdot ((\mathbf{w}^*)^{(j)} - \mathbf{w}^{(j)}) \right| \\ &= \left| \int_0^1 \sum_{j=1}^{J_n} \frac{\partial f_{\mathbf{w}+s(\mathbf{w}^*-\mathbf{w})}(x)}{\partial \mathbf{w}^{(j)}} \cdot ((\mathbf{w}^*)^{(j)} - \mathbf{w}^{(j)}) ds - \sum_{j=1}^{J_n} \frac{\partial f_{\mathbf{w}}(x)}{\partial \mathbf{w}^{(j)}} \cdot ((\mathbf{w}^*)^{(j)} - \mathbf{w}^{(j)}) \right| \\ &= \left| \int_0^1 \sum_{j=1}^{J_n} \left(\frac{\partial f_{\mathbf{w}+s(\mathbf{w}^*-\mathbf{w})}(x)}{\partial \mathbf{w}^{(j)}} - \frac{\partial f_{\mathbf{w}}(x)}{\partial \mathbf{w}^{(j)}} \right) \cdot ((\mathbf{w}^*)^{(j)} - \mathbf{w}^{(j)}) ds \right| \\ &\leq \int_0^1 \sum_{j=1}^{J_n} \left| \frac{\partial f_{\mathbf{w}+s(\mathbf{w}^*-\mathbf{w})}(x)}{\partial \mathbf{w}^{(j)}} - \frac{\partial f_{\mathbf{w}}(x)}{\partial \mathbf{w}^{(j)}} \right| \cdot |(\mathbf{w}^*)^{(j)} - \mathbf{w}^{(j)}| ds \\ &\leq \int_0^1 \sqrt{\sum_{j=1}^{J_n} \left| \frac{\partial f_{\mathbf{w}+s(\mathbf{w}^*-\mathbf{w})}(x)}{\partial \mathbf{w}^{(j)}} - \frac{\partial f_{\mathbf{w}}(x)}{\partial \mathbf{w}^{(j)}} \right|^2} \cdot \|\mathbf{w}^* - \mathbf{w}\| ds \end{aligned}$$

$$\begin{aligned}
&\leq \int_0^1 \sqrt{C_n^2 \cdot \|\mathbf{w} + s \cdot (\mathbf{w}^* - \mathbf{w}) - \mathbf{w}\|^2} \cdot \|\mathbf{w}^* - \mathbf{w}\| ds \\
&\leq C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2 \cdot \int_0^1 s ds = \frac{1}{2} \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2.
\end{aligned}$$

In the *second step of the proof* we show that for any $\mathbf{w} \in \mathbb{R}^{J_n}$ with $\|\mathbf{w}^* - \mathbf{w}\| \leq \|\mathbf{w}^* - \mathbf{w}^{(0)}\|$, we have

$$|F_{n,lin,\mathbf{w}}(\mathbf{w}^*) - F_n(\mathbf{w}^*)| \leq 5 \cdot \beta_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2. \quad (44)$$

Using (37)–(40) and the first step of the proof we get

$$\begin{aligned}
&|F_{n,lin,\mathbf{w}}(\mathbf{w}^*) - F_n(\mathbf{w}^*)| \\
&= \left| \frac{1}{n} \sum_{i=1}^n |Y_i - f_{lin,\mathbf{w},\mathbf{w}^*}(X_i)|^2 - \frac{1}{n} \sum_{i=1}^n |Y_i - f_{\mathbf{w}^*}(X_i)|^2 \right| \\
&\leq \frac{1}{n} \sum_{i=1}^n |2 \cdot Y_i - 2 \cdot f_{\mathbf{w}^*}(X_i) + f_{\mathbf{w}^*}(X_i) - f_{lin,\mathbf{w},\mathbf{w}^*}(X_i)| \cdot |f_{lin,\mathbf{w},\mathbf{w}^*}(X_i) - f_{\mathbf{w}^*}(X_i)| \\
&\leq \frac{1}{n} \sum_{i=1}^n (2 \cdot \beta_n + 2 \cdot \beta_n + C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2) \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2 \\
&\leq 5 \cdot \beta_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2.
\end{aligned}$$

In the *third step of the proof* we show that the assertion holds in case that we have for some $s \in \{0, 1, \dots, t_n - 1\}$

$$F_n(\mathbf{w}^{(s+1)}) < F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*). \quad (45)$$

So assume that (45) holds for some $s \in \{0, 1, \dots, t_n - 1\}$. By choosing s minimal with this property we can assume

$$F_n(\mathbf{w}^{(t+1)}) \geq F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*)$$

for all $t \in \{0, 1, \dots, s - 1\}$. By Lemma 6 we can conclude

$$\|\mathbf{w}^{(s)} - \mathbf{w}^*\| \leq \|\mathbf{w}^{(0)} - \mathbf{w}^*\|.$$

By using (36) (which implies $F_n(\mathbf{w}^{(t_n)}) = \min_{t=0,\dots,t_n} F_n(\mathbf{w}^{(t)})$) and the result of the second step of the proof we get

$$\begin{aligned}
F_n(\mathbf{w}^{(t_n)}) &\leq F_n(\mathbf{w}^{(s+1)}) \\
&< F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) \\
&= F_n(\mathbf{w}^*) + F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) - F_n(\mathbf{w}^*) \\
&\leq F_n(\mathbf{w}^*) + 5 \cdot \beta_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}^{(s)}\|^2 \\
&\leq F_n(\mathbf{w}^*) + 5 \cdot \beta_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2.
\end{aligned}$$

In the *fourth step of the proof* we show the assertion in case that (45) does not hold for all $s \in \{0, 1, \dots, t_n - 1\}$.

In this case we have

$$F_n(\mathbf{w}^{(s+1)}) \geq F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*)$$

for all $s \in \{0, 1, \dots, t_n - 1\}$, so we can conclude from Lemma 6

$$\|\mathbf{w}^{(t)} - \mathbf{w}^*\| \leq \|\mathbf{w}^{(0)} - \mathbf{w}^*\|$$

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

Using (36) (which implies $F_n(\mathbf{w}^{(t_n)}) = \min_{t=0,\dots,t_n} F_n(\mathbf{w}^{(t)})$) and the result of the second step of the proof we conclude

$$\begin{aligned} & F_n(\mathbf{w}^{(t_n)}) - F_n(\mathbf{w}^*) \\ & \leq \frac{1}{\sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \lambda_t \cdot F_n(\mathbf{w}^{(t)}) - F_n(\mathbf{w}^*) \\ & = \frac{1}{\sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \lambda_t \cdot \left(F_n(\mathbf{w}^{(t)}) - F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*) \right) \\ & \quad + \frac{1}{\sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \lambda_t \cdot \left(F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*) - F_n(\mathbf{w}^*) \right) \\ & \leq \frac{1}{\sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \lambda_t \cdot \left(F_n(\mathbf{w}^{(t)}) - F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*) \right) \\ & \quad + \frac{1}{\sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \lambda_t \cdot 5 \cdot \beta_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}^{(t)}\|^2 \\ & \leq \frac{1}{\sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \lambda_t \cdot \left(F_n(\mathbf{w}^{(t)}) - F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*) \right) + 5 \cdot \beta_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2. \end{aligned}$$

Since

$$F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^{(t)}) = F_n(\mathbf{w}^{(t)}) \quad \text{and} \quad \nabla_{\mathbf{w}} F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^{(t)}) = \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})$$

and $\mathbf{w} \mapsto F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w})$ is convex and differentiable we get furthermore

$$\begin{aligned} & \frac{1}{\sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \lambda_t \cdot \left(F_n(\mathbf{w}^{(t)}) - F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*) \right) \\ & = \frac{1}{\sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \lambda_t \cdot \left(F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^{(t)}) - F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*) \right) \\ & \leq \frac{1}{\sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \lambda_t \cdot \langle \nabla_{\mathbf{w}} F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* \rangle \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \lambda_t \cdot \langle \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* \rangle \\
&= \frac{1}{2 \cdot \sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} 2 \cdot \langle \lambda_t \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* \rangle \\
&= \frac{1}{2 \cdot \sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \left(\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t)} - \mathbf{w}^* - \lambda_t \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})\|^2 \right. \\
&\quad \left. + \lambda_t^2 \cdot \|\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})\|^2 \right) \\
&\leq \frac{1}{2 \cdot \sum_{s=0}^{t_n-1} \lambda_s} \sum_{t=0}^{t_n-1} \left(\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 \right) \\
&\quad + \frac{\max_{s=0, \dots, t_n-1} \lambda_s}{\sum_{s=0}^{t_n-1} \lambda_s} \cdot \sum_{t=0}^{t_n-1} \frac{\lambda_t}{2} \cdot \|\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})\|^2 \\
&\leq \frac{1}{2 \cdot \sum_{s=0}^{t_n-1} \lambda_s} \cdot \left(\|\mathbf{w}^{(0)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t_n)} - \mathbf{w}^*\|^2 \right) \\
&\quad + \frac{\max_{s=0, \dots, t_n-1} \lambda_s}{\sum_{s=0}^{t_n-1} \lambda_s} \cdot \sum_{t=0}^{t_n-1} (F_n(\mathbf{w}^{(t)}) - F_n(\mathbf{w}^{(t+1)})) \\
&\leq \frac{1}{2 \cdot \sum_{s=0}^{t_n-1} \lambda_s} \cdot \|\mathbf{w}^{(0)} - \mathbf{w}^*\|^2 + \frac{\max_{s=0, \dots, t_n-1} \lambda_s}{\sum_{s=0}^{t_n-1} \lambda_s} \cdot F_n(\mathbf{w}^{(0)}),
\end{aligned}$$

where the second to last inequality follows from assumption (36). $\square$

Our next lemma will enable us to show in the proofs of our main results that the weights during gradient descent do not move too far away from the initial weights. Together with weight bounds from the initialization this will enable us to compute upper bounds on the weights needed for the application of Lemma 4.

Lemma 7 *Let F_n be defined by (11), set*

$$\mathbf{w}^{(s+1)} = \mathbf{w}^{(s)} - \lambda_s \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)}) \quad \text{for } s = 0, \dots, t-1$$

and assume

$$F_n(\mathbf{w}^{(s+1)}) \leq F_n(\mathbf{w}^{(s)}) - \frac{\lambda_s}{2} \cdot \left\| \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)}) \right\|^2 \quad (46)$$

for $s = 0, 1, \dots, t-1$. Then

$$\|\mathbf{w}^{(t)} - \mathbf{w}^{(0)}\| \leq \sqrt{\sum_{s=1}^t \lambda_{s-1}} \cdot \sqrt{2 \cdot F_n(\mathbf{w}^{(0)})}.$$

Proof. Using the triangle inequality, the Cauchy-Schwartz inequality and (46) we get

$$\|\mathbf{w}^{(t)} - \mathbf{w}^{(0)}\|$$

$$\begin{aligned}
&\leq \sum_{s=1}^t \|\mathbf{w}^{(s)} - \mathbf{w}^{(s-1)}\| \\
&= \sum_{s=1}^t \sqrt{\lambda_{s-1}} \cdot \sqrt{\lambda_{s-1}} \cdot \|\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s-1)})\| \\
&\leq \sqrt{\sum_{s=1}^t \lambda_{s-1}} \cdot \sqrt{\sum_{s=1}^t \lambda_{s-1} \cdot \|\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s-1)})\|^2} \\
&\leq \sqrt{\sum_{s=1}^t \lambda_{s-1}} \cdot \sqrt{\sum_{s=1}^t 2 \cdot (F_n(\mathbf{w}^{(s-1)}) - F_n(\mathbf{w}^{(s)}))} \\
&\leq \sqrt{\sum_{s=1}^t \lambda_{s-1}} \cdot \sqrt{2 \cdot F_n(\mathbf{w}^{(0)})}.
\end{aligned}$$

□

In the sequel we derive auxiliary results which we will use to verify the assumptions of Lemma 5 in the proof of Theorem 1. Therefore we introduce some additional notation to describe our initially pruned deep neural network. We describe the connections in our pruned neural network by the set

$$\begin{aligned}
\mathcal{I} \subseteq & \cup_{l=1}^{L-1} \{(l, i, k) : i \in \{1, \dots, r_n\}, k \in \{0, \dots, r_n\}\} \\
& \cup \{(0, i, k) : i \in \{1, \dots, r_n\}, k \in \{0, \dots, d\}\} \cup \{(L, 1, k) : k \in \{0, \dots, r_n\}\}.
\end{aligned}$$

Here $(l, i, k) \in \mathcal{I}$ means that neuron k of level l is connected to neuron i of level $l+1$. Using this set the pruned neural network is described by

$$f_{\mathbf{w}}(x) = \sum_{k=1}^{r_n} w_{1,k}^{(L)} \cdot f_{\mathbf{w},k}^{(L)}(x), \quad (47)$$

$$f_{\mathbf{w},i}^{(l)}(x) = \sigma \left(\sum_{k=1, \dots, r_n: (l-1, i, k) \in \mathcal{I}} w_{i,k}^{(l-1)} \cdot f_{\mathbf{w},k}^{(l-1)}(x) + w_{i,0}^{(l-1)} \cdot 1_{\{(l-1, i, 0) \in \mathcal{I}\}} \right) \quad (48)$$

for $i \in \{1, \dots, r_n\}$ and $l \in \{2, \dots, L\}$, and

$$f_{\mathbf{w},i}^{(1)}(x) = \sigma \left(\sum_{k=1}^d w_{i,k}^{(0)} \cdot x^{(i)} + w_{i,0}^{(0)} \right) \quad (49)$$

for $i \in \{1, \dots, r_n\}$.

Our next lemma concerns inequality (37).

Lemma 8 *Let σ be the logistic squasher. Let $a, B_n, \gamma_n^* \geq 1$, $L, r_n, r \in \mathbb{N}$ and define the deep neural network $f_{\mathbf{w}} : \mathbb{R}^d \rightarrow \mathbb{R}$ with weight vector $\mathbf{w}$ by (47)–(49) where $\mathcal{I}$ satisfies*

$$|\{1 \leq k \leq r_n : (l, i, k) \in \mathcal{I}\}| \leq r \quad \text{for all } l \in \{1, \dots, L-1\}, i \in \{1, \dots, r_n\}. \quad (50)$$

Assume that the weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ satisfy

$$\sum_{k=1}^{r_n} |w_{1,k}^{(L)}|^2 \leq (\gamma_n^*)^2 \quad \text{and} \quad |w_{i,j}^{(l)}| \leq B_n$$

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

$$\sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \left| \frac{\partial}{\partial w_{i,j}^{(l)}} f_{\mathbf{w}_1}(x) - \frac{\partial}{\partial w_{i,j}^{(l)}} f_{\mathbf{w}_2}(x) \right|^2 \leq c_{28} \cdot r_n \cdot B_n^{4L} \cdot (\gamma_n^*)^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2$$

for some $c_{28} = c_{28}(d, L, r, a) > 0$.

Proof. The proof is a modification of the proof of Lemma 2 in Kohler (2026).

We have

$$\begin{aligned} & \sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \left| \frac{\partial}{\partial w_{i,j}^{(l)}} f_{\mathbf{w}_1}(x) - \frac{\partial}{\partial w_{i,j}^{(l)}} f_{\mathbf{w}_2}(x) \right|^2 \\ & \leq \sum_{k=1}^{r_n} |f_{\mathbf{w}_1,k}^{(L)}(x) - f_{\mathbf{w}_2,k}^{(L)}(x)|^2 \\ & \quad + \sum_{i,j,l:l < L, (l,i,j) \in \mathcal{I} \text{ or } l=0} \left| \sum_{k=1}^{r_n} \left((\mathbf{w}_1)_{1,k}^{(L)} \cdot \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_1,k}^{(L)}(x) - (\mathbf{w}_2)_{1,k}^{(L)} \cdot \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_2,k}^{(L)}(x) \right) \right|^2 \\ & \leq \sum_{k=1}^{r_n} |f_{\mathbf{w}_1,k}^{(L)}(x) - f_{\mathbf{w}_2,k}^{(L)}(x)|^2 \\ & \quad + r_n \cdot \sum_{k=1}^{r_n} \sum_{i,j,l:l < L, (l,i,j) \in \mathcal{I} \text{ or } l=0} \left| (\mathbf{w}_1)_{1,k}^{(L)} \cdot \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_1,k}^{(L)}(x) - (\mathbf{w}_2)_{1,k}^{(L)} \cdot \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_2,k}^{(L)}(x) \right|^2 \\ & \leq \sum_{k=1}^{r_n} |f_{\mathbf{w}_1,k}^{(L)}(x) - f_{\mathbf{w}_2,k}^{(L)}(x)|^2 \\ & \quad + 2 \cdot r_n \cdot \sum_{k=1}^{r_n} \sum_{i,j,l:l < L, (l,i,j) \in \mathcal{I} \text{ or } l=0} \left| (\mathbf{w}_1)_{1,k}^{(L)} - (\mathbf{w}_2)_{1,k}^{(L)} \right|^2 \cdot \left| \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_1,k}^{(L)}(x) \right|^2 \\ & \quad + 2 \cdot r_n \cdot \sum_{k=1}^{r_n} \sum_{i,j,l:l < L, (l,i,j) \in \mathcal{I} \text{ or } l=0} \left| (\mathbf{w}_2)_{1,k}^{(L)} \right|^2 \cdot \left| \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_1,k}^{(L)}(x) - \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_2,k}^{(L)}(x) \right|^2 \\ & = \sum_{k=1}^{r_n} |f_{\mathbf{w}_1,k}^{(L)}(x) - f_{\mathbf{w}_2,k}^{(L)}(x)|^2 \\ & \quad + 2 \cdot r_n \cdot \sum_{k=1}^{r_n} \left| (\mathbf{w}_1)_{1,k}^{(L)} - (\mathbf{w}_2)_{1,k}^{(L)} \right|^2 \cdot \sum_{i,j,l:l < L, (l,i,j) \in \mathcal{I} \text{ or } l=0} \left| \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_1,k}^{(L)}(x) \right|^2 \end{aligned}$$

$$+2 \cdot r_n \cdot \sum_{k=1}^{r_n} \left| (\mathbf{w}_2)_{1,k}^{(L)} \right|^2 \cdot \sum_{i,j,l:l < L, (l,i,j) \in \mathcal{I} \text{ or } l=0} \left| \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_1,k}^{(L)}(x) - \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_2,k}^{(L)}(x) \right|^2.$$

By construction of our randomly pruned weight vector we know that $f_{\mathbf{w},k}^{(L)}$ depends on at most

$$\sum_{l=1}^{L-1} r^{L-l-1} \cdot (r+1) + r^{L-1} \cdot (d+1) \leq L \cdot (r+d)^L$$

many weights $w_{i,j}^{(l)}$, hence for any fixed $k \in \{1, \dots, r_n\}$ there are at most $L \cdot (r+d)^L$ many terms in the sums

$$\sum_{i,j,l:l < L, (l,i,j) \in \mathcal{I} \text{ or } l=0} \left| \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_1,k}^{(L)}(x) \right|^2$$

and

$$\sum_{i,j,l:l < L, (l,i,j) \in \mathcal{I} \text{ or } l=0} \left| \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_1,k}^{(L)}(x) - \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w}_2,k}^{(L)}(x) \right|^2$$

which are not zero.

The chain rule implies

$$\begin{aligned} \frac{\partial f_{\mathbf{w},k}^{(L)}}{\partial w_{i,j}^{(l)}}(x) &= \sum_{s_{l+2}=1, \dots, r_n: (l+1, s_{l+2}, i) \in \mathcal{I}} \sum_{s_{l+3}=1, \dots, r_n: (l+2, s_{l+3}, s_{l+2}) \in \mathcal{I}} \dots \\ &\quad \sum_{s_{L-1}=1, \dots, r_n: (L-2, s_{L-1}, s_{L-2}) \in \mathcal{I}} f_{\mathbf{w},j}^{(l)}(x) \cdot \sigma' \left(\sum_{t=0, \dots, r_n: (l, i, t) \in \mathcal{I}} w_{i,t}^{(l)} \cdot f_{\mathbf{w},t}^{(l)}(x) \right) \\ &\quad \cdot w_{s_{l+2},i}^{(l+1)} \cdot \sigma' \left(\sum_{t=0, \dots, r_n: (l+1, s_{l+2}, t) \in \mathcal{I}} w_{s_{l+2},t}^{(l+1)} \cdot f_{\mathbf{w},t}^{(l+1)}(x) \right) \cdot w_{s_{l+3},s_{l+2}}^{(l+2)} \\ &\quad \cdot \sigma' \left(\sum_{t=0, \dots, r_n: (l+2, s_{l+3}, t) \in \mathcal{I}} w_{s_{l+3},t}^{(l+2)} \cdot f_{\mathbf{w},t}^{(l+2)}(x) \right) \cdot \dots \cdot w_{s_{L-1},s_{L-2}}^{(L-2)} \\ &\quad \cdot \sigma' \left(\sum_{t=0, \dots, r_n: (L-2, s_{L-1}, t) \in \mathcal{I}} w_{s_{L-1},t}^{(L-2)} \cdot f_{\mathbf{w},t}^{(L-2)}(x) \right) \cdot w_{k,s_{L-1}}^{(L-1)} \\ &\quad \cdot \sigma' \left(\sum_{t=0, \dots, r_n: (L-1, 1, t) \in \mathcal{I}} w_{k,t}^{(L-1)} \cdot f_{\mathbf{w},t}^{(L-1)}(x) \right), \end{aligned} \tag{51}$$

where we have used the abbreviations

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

and

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

If $f_{i,1}, \dots, f_{i,L}$ are real-valued functions defined on $\mathbb{R}^{J_n}$ where $f_{i,l}$ is bounded in absolute value by $B_{i,l} \geq 1$ and Lipschitz continuous (w.r.t. $\|\cdot\|_\infty$) with Lipschitz constant $C_{i,l} \geq 1$ ($i = 1, \dots, r$) then

$$\begin{aligned} & \left| \sum_{i=1}^r \prod_{l=1}^L f_{i,l}(\mathbf{w}_1) - \sum_{i=1}^r \prod_{l=1}^L f_{i,l}(\mathbf{w}_2) \right| \\ & \leq \sum_{i=1}^r \sum_{j=1}^L \prod_{l=1}^{j-1} |f_{i,l}(\mathbf{w}_1)| \cdot |f_{j,l}(\mathbf{w}_1) - f_{j,l}(\mathbf{w}_2)| \prod_{l=j+1}^L |f_{i,l}(\mathbf{w}_2)| \\ & \leq r \cdot L \cdot \max_{i=1,\dots,r} \prod_{l=1}^L B_{i,l} \cdot \max_{i=1,\dots,r} \max_{l=1,\dots,L} C_{i,l} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|_\infty. \end{aligned}$$

Using this, (50),

$$0 \leq \sigma(x) \leq 1 \quad \text{and} \quad |\sigma'(x)| = |\sigma(x) \cdot (1 - \sigma(x))| \leq 1$$

and

$$\begin{aligned} & |f_{\mathbf{w}_1,j}^{(l)}(x) - f_{\mathbf{w}_2,j}^{(l)}(x)| \\ & \leq c_{29} \cdot a \cdot (\max\{2r, d\} + 1)^l \cdot B_n^{l-1} \cdot \|((\mathbf{w}_1)_{i,j}^{(\bar{l})})_{i,j,\bar{l}} - ((\mathbf{w}_2)_{i,j}^{(\bar{l})})_{i,j,\bar{l}}\|_\infty \\ & \leq c_{29} \cdot a \cdot (\max\{2r, d\} + 1)^l \cdot B_n^{l-1} \cdot \|((\mathbf{w}_1)_{i,j}^{(\bar{l})})_{i,j,\bar{l}} - ((\mathbf{w}_2)_{i,j}^{(\bar{l})})_{i,j,\bar{l}}\| \end{aligned}$$

(cf., proof of Lemma 3) we get

$$\begin{aligned} & \sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \left| \frac{\partial}{\partial w_{k,i,j}^{(l)}} f_{\mathbf{w}_1}(x) - \frac{\partial}{\partial w_{k,i,j}^{(l)}} f_{\mathbf{w}_2}(x) \right|^2 \\ & \leq c_{30} \cdot a^2 \cdot (2r + d)^{2L} \cdot B_n^{2L} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2 \\ & \quad + c_{31} \cdot r_n \cdot L \cdot (r + d)^L \cdot r^{2L} \cdot a^2 \cdot B_n^{2L} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2 \\ & \quad + c_{32} \cdot r_n \cdot (\gamma_n^*)^2 \cdot L \cdot (r + d)^L \cdot r^{2L} \cdot (3L)^2 \cdot a^4 \cdot B_n^{4L} \cdot (2r + d)^{2L} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2 \\ & \leq c_{33} \cdot r_n \cdot (\gamma_n^*)^2 \cdot L^3 \cdot (2r + d)^{5L} \cdot B_n^{4L} \cdot a^4 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2. \end{aligned}$$

□

Our next two lemmata will be needed to verify assumption (36) of Lemma 5 in the proof of our main results.

Lemma 9 *Let σ be the logistic squasher. Let $a, \beta_n, B_n, \gamma_n^* \geq 1$, $L, r_n, r \in \mathbb{N}$ and define the deep neural network $f_{\mathbf{w}} : \mathbb{R}^d \rightarrow \mathbb{R}$ with weight vector $\mathbf{w}$ by (47)–(49) where $\mathcal{I}$ satisfies (50), and assume*

$$X_i \in [-a, a]^d \quad \text{and} \quad |Y_i| \leq \beta_n \quad (i = 1, \dots, n)$$

and

$$\sum_{k=1}^{r_n} |w_{1,k}^{(L)}|^2 \leq (\gamma_n^*)^2 \quad \text{and} \quad |w_{i,j}^{(l)}| \leq B_n$$

for all $l \in \{1, \dots, L-1\}$. Assume

$$r_n \cdot \gamma_n^* \geq \beta_n.$$

Then

$$\|\nabla_{\mathbf{w}} F_n(\mathbf{w})\| \leq c_{34} \cdot r_n^{3/2} \cdot (\gamma_n^*)^2 \cdot B_n^L$$

for some $c_{34} = c_{34}(d, L, r, a) > 0$.

Proof. The proof is a modification of Lemma 3 in Kohler (2026).

We have

$$|f_{\mathbf{w}}(x)| \leq r_n \cdot \gamma_n^*,$$

which implies

$$\begin{aligned} & \|\nabla_{\mathbf{w}} F_n(\mathbf{w})\|^2 \\ &= \sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \left| \frac{1}{n} \sum_{s=1}^n 2 \cdot (Y_s - f_{\mathbf{w}}(X_s)) \cdot \frac{\partial}{\partial w_{i,j}^{(l)}} f_{\mathbf{w}}(X_s) \cdot (-1) \right|^2 \\ &\leq 16 \cdot (r_n \cdot \gamma_n^*)^2 \cdot \sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \max_{s=1,\dots,n} \left| \frac{\partial}{\partial w_{i,j}^{(l)}} f_{\mathbf{w}}(X_s) \right|^2 \\ &\leq 16 \cdot (r_n \cdot \gamma_n^*)^2 \cdot \left(\sum_{k=1}^{r_n} \max_{s=1,\dots,n} |f_{\mathbf{w},k}^{(L)}(X_s)|^2 \right. \\ &\quad \left. + \sum_{i,j,l:l < L, (l,i,j) \in \mathcal{I} \text{ or } l=0} \max_{s=1,\dots,n} \left| \sum_{k=1}^{r_n} \mathbf{w}_{1,k}^{(L)} \cdot \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w},k}^{(L)}(X_s) \right|^2 \right) \\ &\leq 16 \cdot (r_n \cdot \gamma_n^*)^2 \cdot \left(\sum_{k=1}^{r_n} \max_{s=1,\dots,n} |f_{\mathbf{w},k}^{(L)}(X_s)|^2 \right. \\ &\quad \left. + r_n \cdot \sum_{k=1}^{r_n} \sum_{i,j,l:l < L, (l,i,j) \in \mathcal{I} \text{ or } l=0} \max_{s=1,\dots,n} \left| \mathbf{w}_{1,k}^{(L)} \cdot \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w},k}^{(L)}(X_s) \right|^2 \right) \\ &= 16 \cdot (r_n \cdot \gamma_n^*)^2 \cdot \left(\sum_{k=1}^{r_n} \max_{s=1,\dots,n} |f_{\mathbf{w},k}^{(L)}(X_s)|^2 \right. \\ &\quad \left. + r_n \cdot \sum_{k=1}^{r_n} |\mathbf{w}_{1,k}^{(L)}|^2 \cdot \sum_{i,j,l:l < L, (l,i,j) \in \mathcal{I} \text{ or } l=0} \max_{s=1,\dots,n} \left| \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w},k}^{(L)}(X_s) \right|^2 \right) \\ &\leq 16 \cdot (r_n \cdot \gamma_n^*)^2 \cdot \left(r_n \cdot 1 \right) \end{aligned}$$

$$\begin{aligned}
& +r_n \cdot (\gamma_n^*)^2 \cdot L \cdot (r+d)^L \cdot \max_{s=1,\dots,n} \max_{k=1,\dots,r_n} \max_{\substack{i,j,l: \\ l < L, (l,i,j) \in \mathcal{I}}} \left| \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w},k}^{(L)}(X_s) \right|^2 \Bigg) \\
& \leq 16 \cdot (r_n \cdot \gamma_n^*)^2 \cdot (r_n \cdot 1 + r_n \cdot (\gamma_n^*)^2 \cdot L \cdot (r+d)^L \cdot r^{2L} \cdot a^2 \cdot B_n^{2L}),
\end{aligned}$$

where the last inequality follows from (51), the assumptions on the weights and the bounds on the logistic squasher mentioned in the proof of Lemma 8. $\square$

Lemma 10 *Let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ be the logistic squasher, let $L, r_n, r \in \mathbb{N}$, let $f_{\mathbf{w}}$ be defined by (47)–(49) where $\mathcal{I}$ satisfies (50), and let F_n be defined by (11). Let $a \geq 1$, $\gamma_n^* \geq 1$, $B_n \geq 1$, and assume $X_i \in [-a, a]^d$ and $|Y_i| \leq \beta_n$ ($i = 1, \dots, n$),*

$$\max\left\{\sum_{k=1}^{r_n} |(\mathbf{w}_1)_{1,k}^{(L)}|^2, \sum_{k=1}^{r_n} |(\mathbf{w}_2)_{1,k}^{(L)}|^2\right\} \leq (\gamma_n^*)^2, \quad (52)$$

$$\max\{|(\mathbf{w}_1)_{i,j}^{(l)}|, |(\mathbf{w}_2)_{i,j}^{(l)}|\} \leq B_n \quad \text{for } l = 1, \dots, L-1 \quad (53)$$

and

$$r_n \cdot \gamma_n^* \geq \beta_n. \quad (54)$$

Then we have

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}_2)\| \leq c_{35} \cdot r_n^{3/2} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|$$

for some $c_{35} = c_{35}(d, L, r, a) > 0$.

Proof. The proof is a modification of the proof of Lemma 5 in Kohler (2026).

We have

$$\begin{aligned}
& \|\nabla_{\mathbf{w}} F_n(\mathbf{w}_1) - \nabla_{\mathbf{w}} F_n(\mathbf{w}_2)\|^2 \\
& = \sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \left(\frac{2}{n} \sum_{s=1}^n (f_{\mathbf{w}_1}(X_s) - Y_s) \cdot \frac{\partial f_{\mathbf{w}_1}}{\partial w_{i,j}^{(l)}}(X_s) \right. \\
& \quad \left. - \frac{2}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(X_s) - Y_s) \cdot \frac{\partial f_{\mathbf{w}_2}}{\partial w_{i,j}^{(l)}}(X_s) \right)^2 \\
& \leq 16 \cdot \sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \max_{s=1,\dots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{i,j}^{(l)}}(X_s) \right)^2 \cdot \frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(X_s) - f_{\mathbf{w}_1}(X_s))^2 \\
& \quad + 16 \cdot \frac{1}{n} \sum_{s=1}^n (Y_s - f_{\mathbf{w}_2}(X_s))^2 \cdot \sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{i,j}^{(l)}}(X_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{i,j}^{(l)}}(X_s) \right)^2.
\end{aligned}$$

From the proof of Lemma 9 we can conclude

$$\sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \max_{s=1,\dots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{i,j}^{(l)}}(X_s) \right)^2$$

$$\begin{aligned}
&\leq \sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \max_{s=1,\dots,n} \left(\sum_{k=1}^{r_n} \mathbf{w}_{1,k}^{(L)} \cdot \frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w},k}^{(L)}(X_s) \right)^2 \\
&\leq \sum_{k=1}^{r_n} |\mathbf{w}_{1,k}^{(L)}|^2 \cdot \sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \sum_{k=1}^{r_n} \max_{s=1,\dots,n} \left(\frac{\partial}{\partial \mathbf{w}_{i,j}^{(l)}} f_{\mathbf{w},k}^{(L)}(X_s) \right)^2 \\
&\leq c_{36} \cdot (\gamma_n^*)^2 \cdot L \cdot (r+d)^L \cdot r_n \cdot r^{2L} \cdot B_n^{2L} \cdot a^2,
\end{aligned}$$

from the proof of Lemma 8 we know

$$\begin{aligned}
&\frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(X_s) - f_{\mathbf{w}_1}(X_s))^2 \\
&\leq \frac{2}{n} \sum_{s=1}^n \sum_{k=1}^{r_n} r_n \cdot |(\mathbf{w}_2)_{1,k}^{(L)}|^2 \cdot (f_{\mathbf{w}_2,k}^{(L)}(X_s) - f_{\mathbf{w}_1,k}^{(L)}(X_s))^2 \\
&\quad + \frac{2}{n} \sum_{s=1}^n \sum_{k=1}^{r_n} r_n \cdot |(\mathbf{w}_2)_{1,k}^{(L)} - (\mathbf{w}_1)_{1,k}^{(L)}|^2 \cdot |f_{\mathbf{w}_1,k}^{(L)}(X_s)|^2 \\
&\leq c_{37} \cdot r_n \cdot (\gamma_n^*)^2 \cdot (2r+d)^{2L} \cdot B_n^{2L} \cdot a^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2
\end{aligned}$$

and (52) and (54) imply

$$\frac{1}{n} \sum_{s=1}^n (Y_s - f_{\mathbf{w}_2}(X_s))^2 \leq 4 \cdot r_n^2 \cdot (\gamma_n^*)^2.$$

And by Lemma 8 we can conclude for any $s \in \{1, \dots, n\}$

$$\sum_{i,j,l:(l,i,j) \in \mathcal{I} \text{ or } l \in \{0,L\}} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{i,j}^{(l)}}(X_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{i,j}^{(l)}}(X_s) \right)^2 \leq c_{38} \cdot r_n \cdot B_n^{4L} \cdot (\gamma_n^*)^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2.$$

Summarizing the above results we get the assertion. $\square$

5.4 Proof of Theorem 1

In the *first step of the proof* we show that if

$$|Y_i| \leq c_{39} \cdot \log n \quad (i = 1, \dots, n) \tag{55}$$

holds, then we have

$$F_n(\mathbf{w}^{(s+1)}) \leq F_n(\mathbf{w}^{(s)}) - \frac{\lambda_s}{2} \cdot \left\| \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)}) \right\|^2 \tag{56}$$

for $s = 0, 1, \dots, t_n - 1$.

To show this we have to demonstrate that if Algorithm 1 stops its inner loop because of $\lambda_s \leq \lambda_{\min}$, then (56) holds. Therefore it suffices to show for all $t \in \{0, 1, \dots, t_n - 1\}$ that if (56) holds for all $s \in \{0, 1, \dots, t - 1\}$ and if $\lambda \leq \lambda_{\min}$, then

$$\bar{\mathbf{w}} = \mathbf{w}^{(t)} - \lambda \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})$$

satisfies

$$F_n(\bar{\mathbf{w}}) \leq F_n(\mathbf{w}^{(t)}) - \frac{\lambda}{2} \cdot \left\| \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}) \right\|^2.$$

So assume that (56) holds for all $s \in \{0, 1, \dots, t-1\}$. Then we can conclude from Lemma 7 and (55) that we have

$$\|\mathbf{w}^{(s)} - \mathbf{w}^{(0)}\| \leq c_{40} \cdot \log n \quad (s = 1, \dots, t).$$

Using the bounds on $\mathbf{w}^{(0)}$ we can conclude from Lemma 9 (which we apply with $\gamma_n^* = B_n = c_{41} \cdot \log n$) that we have

$$\left\| \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}) \right\| \leq c_{42} \cdot (\log n)^{L+2} \cdot r_n^{3/2},$$

hence for any $\tau \in [0, 1]$ we have

$$\begin{aligned} \|\mathbf{w}^{(t)} + \tau \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)}) - \mathbf{w}^{(0)}\| &\leq \|\mathbf{w}^{(t)} - \mathbf{w}^{(0)}\| + \|\bar{\mathbf{w}} - \mathbf{w}^{(t)}\| \\ &\leq \|\mathbf{w}^{(t)} - \mathbf{w}^{(0)}\| + \lambda \cdot \left\| \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}) \right\| \\ &\leq 2 \cdot \log n + \lambda_{\min} \cdot \left\| \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}) \right\| \\ &\leq c_{43} \cdot \log n. \end{aligned}$$

Application of Lemma 10 (which we apply with $\gamma_n^* = B_n = c_{44} \cdot \log n$) yields

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)} + \tau \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)})) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)})\| \leq c_{45} \cdot r_n^{3/2} \cdot (\log n)^{2L+2} \cdot \|\tau \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)})\|.$$

For $\tau \in [0, 1]$ set

$$H(\tau) = F_n(\mathbf{w}^{(t)} + \tau \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)})).$$

Then the fundamental theorem of calculus, the chain rule and the Cauchy-Schwarz inequality imply

$$\begin{aligned} F_n(\bar{\mathbf{w}}) - F_n(\mathbf{w}^{(t)}) &= H(1) - H(0) = \int_0^1 H'(\tau) d\tau \\ &= \int_0^1 (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)} + \tau \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)})) \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)}) d\tau \\ &= \int_0^1 \left((\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)} + \tau \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)})) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)}) \right) \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)}) d\tau \\ &\quad + \int_0^1 (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)}) \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)}) d\tau \\ &\leq \int_0^1 \left\| (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)} + \tau \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)})) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)}) \right\| \cdot \|\bar{\mathbf{w}} - \mathbf{w}^{(t)}\| d\tau \\ &\quad + \int_0^1 (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)}) \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)}) d\tau \\ &\leq \int_0^1 c_{46} \cdot r_n^{3/2} \cdot (\log n)^{2L+2} \cdot \|\tau \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)})\| \cdot \|\bar{\mathbf{w}} - \mathbf{w}^{(t)}\| d\tau \end{aligned}$$

$$\begin{aligned}
& +(\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)}) \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)}) \\
& = \frac{c_{46} \cdot r_n^{3/2} \cdot (\log n)^{2L+2}}{2} \cdot \|\bar{\mathbf{w}} - \mathbf{w}^{(t)}\|^2 + (\nabla_{\mathbf{w}} F)(\mathbf{w}^{(t)}) \cdot (\bar{\mathbf{w}} - \mathbf{w}^{(t)}).
\end{aligned}$$

Using $\lambda \leq \lambda_{\min}$ and the definitions of $\lambda_{\min}$ and $\bar{\mathbf{w}}$ we get

$$\begin{aligned}
& F_n(\bar{\mathbf{w}}) - F_n(\mathbf{w}^{(t)}) \\
& \leq \frac{c_{46} \cdot r_n^{3/2} \cdot (\log n)^{2L+2}}{2} \cdot \lambda^2 \cdot \|(\nabla_{\mathbf{w}} F)(\mathbf{w}^{(t)})\|^2 - \lambda \|(\nabla_{\mathbf{w}} F)(\mathbf{w}^{(t)})\|^2 \\
& \leq \left(\frac{c_{46} \cdot r_n^{3/2} \cdot (\log n)^{2L+2}}{2} \cdot \lambda_{\min} - 1 \right) \cdot \lambda \cdot \|(\nabla_{\mathbf{w}} F)(\mathbf{w}^{(t)})\|^2 \\
& \leq -\frac{1}{2} \cdot \lambda \cdot \|(\nabla_{\mathbf{w}} F)(\mathbf{w}^{(t)})\|^2,
\end{aligned}$$

which follows from the fact that we can assume w.l.o.g. that n is so large that

$$c_{46} \cdot r_n^{3/2} \cdot (\log n)^{2L+2} \cdot \lambda_{\min} \leq 1$$

holds.

In the *second step of the proof* we show

$$\|\mathbf{w}^{(s)} - \mathbf{w}^{(0)}\| \leq c_{47} \cdot \log n \quad (s = 0, 1, \dots, t_n) \quad (57)$$

provided (55) holds. This follows directly from Lemma 7 and the first step of the proof.

In the *third step of the proof* we decompose the L_2 error of the estimate in a sum of several terms.

Let E_n be the event that $|Y_i|^2 \leq c_{48} \cdot \log n$ holds for $i = 1, \dots, n$ and set

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

Then we have

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

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

In the *fourth step of the proof* we show

$$\mathbf{E}\{T_{j,n}\} \leq c_{49} \cdot \frac{\log n}{n} \quad \text{for } j \in \{1, 3\}.$$

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

In the *fifth step of the proof* we show

$$\mathbf{E}\{T_{5,n}\} \leq c_{50} \cdot \frac{(\log n)^2}{n^2}.$$

W.l.o.g we can assume that n is so large that $\|m\|_\infty \leq \beta_n$ holds, and hence

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq \int |2 \cdot \beta_n|^2 \mathbf{P}_X(dx) \leq c_{51} \cdot (\log n)^2.$$

Consequently the assertion follows from

$$\begin{aligned}
\mathbf{P}(E_n^c) & \leq \mathbf{P}\{\max_{i=1,\dots,n} Y_i^2 > \beta_n\} \leq n \cdot \mathbf{P}\{Y^2 > \beta_n\} \\
& = n \cdot \mathbf{P}\{\exp(c_3 \cdot Y^2) > \exp(c_3 \cdot \beta_n)\} \leq n \cdot \frac{\mathbf{E}\{\exp(c_3 \cdot Y^2)\}}{\exp(c_3 \cdot c_2 \cdot \log n)} \\
& \leq \frac{c_{52}}{n^2},
\end{aligned}$$

where the last inequality followed from assumption $c_2 \cdot c_3 \geq 2$ and assumption (A3).

In the *sixth step of the proof* we show

$$\mathbf{E}\{T_{2,n}\} \leq c_{53} \cdot (\log n)^{4 \cdot d \cdot L + 3} \cdot n^{-\frac{2p}{2p+d}}.$$

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

$$|w_{1,k}^{(L)}| \leq c_{54} \cdot \log n \quad (k = 1, \dots, K_n),$$

$$|w_{i,j}^{(l)}| \leq c_{55} \cdot \log n \quad (l = 1, \dots, L-1),$$

$$|w_{i,j}^{(0)}| \leq c_{56} \cdot (\log n) \cdot n^{\frac{1}{2p+d}}$$

and

$$|\{j \in \{1, \dots, r_n\} : w_{i,j}^{(l)} \neq 0\}| \leq r \quad \text{for all } l \in \{1, \dots, L-1\}, i \in \{1, \dots, r_n\}.$$

By the second step of the proof and our assumptions on the initialization of the weights we know $\mathbf{w}^{(t_n)} \in \mathcal{W}_n$ on E_n , hence we have

$$m_n = f \quad \text{for some } f \in \mathcal{F}_n = \{T_{\beta_n} f_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}_n\}$$

on E_n . This implies for any $u > 0$

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

By Lemma 4 we get

$$\begin{aligned} & \mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n} \cdot f : f \in \mathcal{F}_n\right\}, x_1^n\right) \leq \mathcal{N}_1(\delta \cdot \beta_n, \mathcal{F}_n, x_1^n) \\ & \leq 3 \left(\frac{12e \cdot \beta_n}{\delta \cdot \beta_n}\right)^{2 \cdot (4ed)^d \cdot 2^{L \cdot d} \cdot (r+k)^{3 \cdot d \cdot L} \cdot \alpha^d \cdot ((\log n) \cdot n^{\frac{1}{2p+d}})^d \cdot (c_{57} \cdot \log n)^{(L-1) \cdot d} \cdot \left(\frac{rn \cdot c_{58} \cdot \log n}{\delta \cdot \beta_n}\right)^{\frac{d}{k}} + 2} \end{aligned}$$

for any $k \in \mathbb{N}$. Choosing

$$k = \lceil (c_7 + 2) \cdot \log n \rceil$$

we conclude from (15) for $\delta > 1/n^2$

$$\begin{aligned} & \mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n} \cdot f : f \in \mathcal{F}_n\right\}, x_1^n\right) \\ & \leq 3 (12e \cdot n)^{c_{59} \cdot (\log n)^{3 \cdot d \cdot L} \cdot (\log n)^d \cdot n^{\frac{d}{2p+d}} \cdot (\log n)^{(L-1) \cdot d} \cdot (n^{c_7+2})^{\frac{d}{(c_7+2) \cdot \log n}}} \\ & \leq 3 (12e \cdot n)^{c_{60} \cdot (\log n)^{4 \cdot d \cdot L} \cdot n^{\frac{d}{2p+d}}}. \end{aligned}$$

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

$$\mathbf{P}\{T_{2,n} > u\} \leq 14 \cdot c_{61} \cdot n^{c_{62} \cdot (\log n)^{4 \cdot d \cdot L} \cdot n^{\frac{d}{2p+d}}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot u\right).$$

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

$$\begin{aligned} \mathbf{E}\{T_{2,n}\} &\leq \epsilon_n + \int_{\epsilon_n}^{\infty} \mathbf{P}\{T_{2,n} > u\} du \\ &\leq \epsilon_n + 14 \cdot c_{61} \cdot n^{c_{62} \cdot (\log n)^{4 \cdot d \cdot L} \cdot n^{\frac{d}{2p+d}}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot \epsilon_n\right) \cdot \frac{5136 \cdot \beta_n^2}{n}. \end{aligned}$$

Setting

$$\epsilon_n = \frac{5136 \cdot \beta_n^2}{n} \cdot c_{62} \cdot (\log n)^{4 \cdot d \cdot L} \cdot n^{\frac{d}{2p+d}} \cdot \log n = \frac{5136 \cdot \beta_n^2}{n} \cdot \log\left(n^{c_{62} \cdot (\log n)^{4 \cdot d \cdot L} \cdot n^{\frac{d}{2p+d}}}\right)$$

yields the assertion of the sixth step of the proof.

In the *seventh step of the proof* we show

$$\mathbf{E}\{T_{4,n}\} \leq c_{63} \cdot n^{-\frac{2p}{2p+d}}.$$

Since $|T_{\beta_n} z - y| \leq |z - y|$ for $|y| \leq \beta_n$ we have

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

We will use Lemma 5 to derive an upper bound on $F_n(\mathbf{w}^{(t_n)})$ on E_n .

By the first step of the proof we know that (36) holds.

Next we choose $w_1^*, \dots, w_{r_n}^*$ as in Lemma 2 and define $\mathbf{w}^*$ by

$$(\mathbf{w}^*)_{i,j}^{(l)} = \begin{cases} w_j^* & \text{if } l = L, i = 1, \\ (\mathbf{w}^{(0)})_{i,j}^{(l)} & \text{if } l \in \{1, \dots, L-1\} \end{cases}$$

if E_n holds and set $\mathbf{w}^* = \mathbf{w}^{(0)}$ else. Then we have

$$\|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 \leq \sum_{k=1}^{r_n} |w_k^*|^2 \leq \frac{c_{64} \cdot n^{4L \cdot r^{L-1} \cdot (d+1)+4}}{r_n}.$$

By Lemma 8 we know that (37) holds with $C_n = c_{65} \cdot (\log n)^{2L+1} \cdot \sqrt{r_n}$, where we apply Lemma 8 with $B_n = \gamma_n^* = c_{66} \cdot \log n$, which is possible because of the result of the second step of the proof and our initial choice of $\mathbf{w}^{(0)}$. Here we have used that for any $\mathbf{w}$ with $\|\mathbf{w}^* - \mathbf{w}\| \leq \|\mathbf{w}^* - \mathbf{w}^{(0)}\|$ we have

$$\sum_{t=1}^{r_n} |w_{1,t}^{(L)}|^2 \leq 2 \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 + 2 \cdot \sum_{t=1}^{r_n} |(\mathbf{w}^*)_{1,t}^{(L)}|^2 \leq c_{67} \cdot (\log n)^2.$$

This implies also that (40) holds, and since $f_{\mathbf{w}^*}$ approximates m in supremum norm well on E_n and m is bounded, we can assume w.l.o.g. that n is so large that (39) holds.

So the assumptions of Lemma 5 are satisfied on E_n , and we can conclude

$$\begin{aligned} & \left[2 \cdot F_n(\mathbf{w}^{(t_n)}) - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right] \cdot 1_{E_n} \\ & \leq \left[2 \cdot F_n(\mathbf{w}^*) + 2 \cdot \left(5 \cdot \beta_n \cdot C_n + \frac{1}{2 \cdot 1} \right) \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 + 2 \cdot F_n(\mathbf{w}^{(0)}) \cdot \frac{\frac{1}{n}}{1} \right. \\ & \quad \left. - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right] \cdot 1_{E_n} \end{aligned}$$

This yields

$$\begin{aligned} T_{4,n} & \leq 2 \cdot \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^*}(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 + c_{68} \cdot \frac{\log n}{n} \\ & \quad + 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \cdot 1_{E_n^c}. \end{aligned}$$

Because of

$$\begin{aligned} & \mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^*}(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right\} \\ & = \mathbf{E} \int |f_{\mathbf{w}^*}(x) - m(x)|^2 \mathbf{P}_X(dx) \\ & \leq c_{69} \cdot n^{-\frac{2p}{2p+d}} + c_{70} \cdot n^{c_{71}} \cdot \exp(-(\log n)^2) \\ & \leq c_{72} \cdot n^{-\frac{2p}{2p+d}} \end{aligned}$$

and

$$\begin{aligned} & \mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \cdot 1_{E_n^c} \right\} \\ & \leq \frac{1}{n} \sum_{i=1}^n \sqrt{\mathbf{E} \{|m(X_i) - Y_i|^4\}} \cdot \sqrt{\mathbf{P}(E_n^c)} \leq \frac{c_{73}}{n} \end{aligned}$$

this implies the assertion. □

5.5 Proof of Theorem 2

In the proof we will need the following auxiliary results.

Lemma 11 *Let $p, C, \alpha > 0$ and assume that (A4) holds. Choose the parameters of the estimate as in Theorem 2. Let $\mathbf{w}^{(0)}$ be the randomly initialized and pruned weight vector of the network. Then outside of an event whose probability is bounded from above by*

$$n^{c_{74}} \cdot \exp(-(\log n)^2)$$

there exists (random) $w_1^, \dots, w_{r_n}^* \in \mathbb{R}$ such that*

$$\sup_{x \in [-\alpha, \alpha]^d} \left| \sum_{k=1}^{r_n} w_k^* \cdot f_{\mathbf{w}^{(0)}, k}^{(L)}(x) - m(x) \right| \leq c_{75} \cdot n^{-\frac{p}{2p+d^*}}$$

and

$$\sum_{k=1}^{r_n} |w_k^*|^2 \leq \frac{c_{76} \cdot n^{4L \cdot r^{L-1} \cdot (d+1) + 4}}{r_n}$$

Proof. The result follows by a straightforward modification of the proof of Lemma 2, therefore we only provide the outline of the proof.

W.l.o.g. we assume that n is so large that $A_n \geq B$ holds. Set $\tilde{K}_n = \lceil n^{\frac{1}{2p+d^*}} \rceil$ and

$$I_n = \lceil \frac{r_n}{n^{4L \cdot r^{L-1} \cdot (d+1) + 1}} \rceil.$$

In the *first step of the proof* we show that there exists a neural network

$$f_{\mathbf{w}}(x) = \sum_{j=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_j \cdot f_{\mathbf{w}_{j,1}}^{(L)}(x) \quad (58)$$

where $f_{\mathbf{w}_{j,1}}^{(L)}$ are recursively defined by (19) and (20) such that

$$\sup_{x \in [-\alpha, \alpha]^d} \left| \sum_{j=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_j \cdot f_{\mathbf{w}_{j,1}}^{(L)}(x) - m(x) \right| \leq c_{77} \cdot n^{-\frac{p}{2p+d^*}} \quad \text{and} \quad \sum_{j=1}^{I_n \cdot r \cdot \tilde{K}_n^d} w_j^2 \leq c_{78} \cdot \frac{n^3}{I_n}.$$

To show this we proceed as in the first step of the proof of Lemma 2. Observe that due to the different choice of $\tilde{K}_n$ here all inner weights are within the range of the uniform distributions used in Theorem 2 during the initialization of the weights.

Set

$$\epsilon_n = \frac{1}{n^3} \quad \text{and} \quad N_n = 2^{L \cdot r^{L-1} \cdot (d+1)} \cdot \left\lceil n^{4L \cdot r^{L-1} \cdot (d+1)} \cdot (\log n)^2 \right\rceil.$$

Next we observe that outside of an event whose probability is bounded from above by

$$I_n \cdot r \cdot \tilde{K}_n^d \cdot \exp \left(-N_n \cdot \frac{1}{2} \cdot \left(\frac{\epsilon_n}{A_n} \right)^{L \cdot r^{L-1} \cdot (d+1)} \right)$$

there exists (random) indices $j_1, \dots, j_{I_n \cdot r \cdot \tilde{K}_n^d} \in \{1, \dots, r_n\}$ such that

$$\sup_{x \in [-\alpha, \alpha]^d} |f_{\mathbf{w}_{k,k,1}}^{(L)}(x) - f_{\mathbf{w}^{(0)}, j_k}^{(L)}(x)| \leq (d+1) \cdot \alpha \cdot (2r+1)^{L-1} \cdot B^{L-1} \cdot \epsilon_n \quad (k = 1, \dots, I_n \cdot r \cdot \tilde{K}_n^d).$$

This follows from the second step of the proof of Lemma 2.

From this we get the assertion as in the third step of the proof of Lemma 2. $\square$

In order to prove an extension of our bound on the covering number of over-parametrized deep neural network to the case of manifold data, we need the following auxiliary result.

Lemma 12 *Let $\mathcal{M}$ be a d^* -dimensional Lipschitz-manifold. Let $h \in (0, 1]$ and let $\mathcal{P}$ be a partition of $\mathbb{R}^d$ into cubes with side length h . Then*

$$|\{C \in \mathcal{P} : C \cap \mathcal{M} \neq \emptyset\}| \leq c_{79} \cdot \left(\frac{1}{h}\right)^{d^*},$$

where $c_{79} = s \cdot (2 \cdot C_{\psi,2} \cdot \sqrt{d^*} + 4)^d$.

Proof. See Lemma 11 in Kohler, Krzyżak and Molinero Römer (2026). $\square$

Using the above lemma we are able to show the following modification of the bound on the covering number in Lemma 4.

Lemma 13 *Let $L, r_n, r \in \mathbb{N}$, let $\beta, A, B, C \geq 1$. Let σ be the logistic squasher and let $\mathcal{F}$ be the set of all neural networks $f_{\mathbf{w}}$ defined by (2)–(4) where the weight vector $\mathbf{w}$ satisfies (25)–(27). Let $0 < \epsilon \leq 1$, let $k \in \mathbb{N}$, let $\mathcal{M}$ be a d^* -dimensional Lipschitz-manifold, set*

$$\delta = \frac{1}{2} \cdot \left(\frac{k!}{2d^k} \cdot \frac{\epsilon}{(r+k)^{3 \cdot k \cdot L} \cdot (2^{k-1})^L \cdot C \cdot B^{(L-1) \cdot k} \cdot A^k} \right)^{1/k}$$

and let $\mathcal{M}_\delta$ be the δ -neighborhood of $\mathcal{M}$ introduced in Definition 3. Let $\alpha \geq 1$ be such that $\mathcal{M}_\delta \subseteq [-\alpha, \alpha]^d$ holds. Then we have for any $x_1, \dots, x_n \in \mathcal{M}_\delta$

$$\mathcal{N}(\epsilon, T_\beta \mathcal{F}, x_1^n) \leq 3 \cdot \left(\frac{12e \cdot \beta}{\epsilon} \right)^{c_{80} \cdot k^{3 \cdot L \cdot d^* + d} \cdot A^{d^*} \cdot B^{(L-1) \cdot d^*} \cdot \left(\frac{C}{\epsilon}\right)^{\frac{d^*}{k}}}$$

for some constant $c_{80} = c_{80}(d, r, L, C_{\psi,2}, s) > 0$.

Proof. Let $\mathcal{G}$ be the set of all polynomials of degree less than or equal to $k-1$, let Π be a partition of $[-\alpha, \alpha]^d$ into cubes of sidelength

$$h = \left(\frac{k!}{2d^k} \cdot \frac{\epsilon}{(r+k)^{3 \cdot k \cdot L} \cdot (2^{k-1})^L \cdot C \cdot B^{(L-1) \cdot k} \cdot A^k} \right)^{1/k} \leq 1,$$

and let $\mathcal{G} \circ \Pi$ be the class of all functions whose restriction on each cube in Π lies in $\mathcal{G}$ and which are zero outside of $[-\alpha, \alpha]^d$. The proof of Lemma 4 implies

$$\mathcal{N}_p(\epsilon, \{T_\beta f : f \in \mathcal{F}\}, x_1^n) \leq \mathcal{N}_p\left(\frac{\epsilon}{2}, T_\beta \mathcal{G} \circ \Pi, x_1^n\right),$$

so it suffices to derive an upper bound on the covering number on the right-hand side of the above inequality. Set $\Pi^* = \{I \in \Pi : I \cap x_1^n \neq \emptyset\}$. Obviously we have

$$\mathcal{N}_p\left(\frac{\epsilon}{2}, T_\beta \mathcal{G} \circ \Pi, x_1^n\right) = \mathcal{N}_p\left(\frac{\epsilon}{2}, T_\beta \mathcal{G} \circ \Pi^*, x_1^n\right),$$

where we set the functions in $T_\beta \mathcal{G} \circ \Pi^*$ to zero on all cubes in $I \in \Pi \setminus \Pi^*$. Since $x_1, \dots, x_n \in \mathcal{M}_\delta$, there exists $z_1, \dots, z_n \in \mathcal{M}$ such that $\|x_i - z_i\| \leq 2 \cdot \delta$ holds for $i = 1, \dots, n$. By Lemma 12 we know

$$|\{I \in \Pi : I \cap z_1^n \neq \emptyset\}| \leq |\{I \in \Pi : I \cap \mathcal{M} \neq \emptyset\}| \leq s \cdot (2 \cdot C_{\psi,2} \cdot \sqrt{d^*} + 4)^d \cdot \left(\frac{1}{h}\right)^{d^*}.$$

If $z_i \in I \in \Pi$, then $\|x_i - z_i\| \leq 2 \cdot \delta$ implies that x_i is contained in one of the $3^d - 1$ cubes of side length $h = 2 \cdot \delta$ which are adjacent to I . Hence

$$|\Pi^*| \leq 3^d \cdot |\{I \in \Pi : I \cap z_1^n \neq \emptyset\}| \leq 3^d \cdot s \cdot (2 \cdot C_{\psi,2} \cdot \sqrt{d^*} + 4)^d \cdot \left(\frac{1}{h}\right)^{d^*}.$$

Consequently, $\mathcal{G} \circ \Pi^*$ is a linear vector space of dimension less than or equal to

$$\begin{aligned} & k^d \cdot 3^d \cdot s \cdot (2 \cdot C_{\psi,2} \cdot \sqrt{d^*} + 4)^d \cdot \left(\frac{1}{h}\right)^{d^*} \\ & \leq k^d \cdot 3^d \cdot s \cdot (2 \cdot C_{\psi,2} \cdot \sqrt{d^*} + 4)^d \cdot d^{d^*} \cdot (r + k)^{3 \cdot L \cdot d^*} \cdot 2^{L \cdot d^*} \\ & \quad \cdot \left(\frac{C}{\epsilon}\right)^{d^*/k} \cdot B^{(L-1) \cdot d^*} \cdot A^{d^*}. \end{aligned}$$

As in the proof of Lemma 4 we conclude from Theorem 9.4 and Theorem 9.5 in Györfi et al. (2002),

$$\mathcal{N}\left(\frac{\epsilon}{2}, T_\beta \mathcal{G} \circ \Pi^*, x_1^n\right) \leq \cdot \left(\frac{12e \cdot \beta}{\epsilon}\right)^{c_{81} \cdot k^{3 \cdot L \cdot d^* + d} \cdot A^{d^*} \cdot B^{(L-1) \cdot d^*} \cdot \left(\frac{C}{\epsilon}\right)^{\frac{d^*}{k}}}$$

□

Proof of Theorem 2. We mimick the proof of Theorem 1.

In the *first step of the proof* we show that if (55) holds, then we have (56) for $s = 0, 1, \dots, t_n - 1$. This follows as in the first step of the proof of Theorem 1.

In the *second step of the proof* we show (57) provided (55) holds. This follows directly from Lemma 7 and the first step of the proof.

In the *third step of the proof* we decompose the L_2 error of the estimate in a sum of several terms.

Let E_n be the event that $|Y_i|^2 \leq c_{82} \cdot \log n$ holds for $i = 1, \dots, n$ and set

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

Then we have

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

In the *fourth step of the proof* we show

$$\mathbf{E}\{T_{j,n}\} \leq c_{83} \cdot \frac{\log n}{n} \quad \text{for } j \in \{1, 3\}.$$

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

In the *fifth step of the proof* we show

$$\mathbf{E}\{T_{5,n}\} \leq c_{84} \cdot \frac{(\log n)^2}{n^2}.$$

This mimics the fifth step of the proof of Theorem 1.

In the *sixth step of the proof* we show

$$\mathbf{E}\{T_{2,n}\} \leq c_{85} \cdot (\log n)^{4 \cdot d^* \cdot L + d + 3} \cdot n^{-\frac{2p}{2p+d^*}}.$$

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

$$|w_{1,k}^{(L)}| \leq c_{86} \cdot \log n =: C_n / r_n \quad (k = 1, \dots, K_n),$$

$$|w_{i,j}^{(l)}| \leq c_{87} \cdot \log n =: B_n \quad (l = 1, \dots, L-1),$$

$$|w_{i,j}^{(0)}| \leq c_{88} \cdot (\log n) \cdot n^{\frac{1}{2p+d^*}} =: A_n$$

and

$$|\{j \in \{1, \dots, r_n\} : w_{i,j}^{(l)} \neq 0\}| \leq r \quad \text{for all } l \in \{1, \dots, L-1\}, i \in \{1, \dots, r_n\}.$$

By the second step of the proof and our assumptions on the initialization of the weights we know $\mathbf{w}^{(t_n)} \in \mathcal{W}_n$ on E_n , hence we have

$$m_n = f \quad \text{for some } f \in \mathcal{F}_n = \{T_{\beta_n} f_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}_n\}$$

on E_n . This implies for any $u > 0$

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

Next we want to apply Lemma 13 with $x_1^n \in \text{supp}(\mathbf{P}_X)$, $k = \lceil (c_7 + 2) \cdot \log n \rceil$ and $\epsilon \geq 1/n$.

For this we have to show that $x_1^n \in \text{supp}(\mathbf{P}_X)$ are contained in $\mathcal{M}_{\delta_n}$, where

$$\delta_n = \frac{1}{2} \cdot \left(\frac{k!}{2d^k} \cdot \frac{\epsilon}{(r+k)^{3 \cdot k \cdot L} \cdot (2^{k-1})^L \cdot C_n \cdot B_n^{(L-1) \cdot k} \cdot A_n^k} \right)^{1/k}.$$

Since

$$\delta_n \geq c_{89} \cdot (\log n)^{-3L-L} \cdot n^{-\frac{1}{2p+d^*}},$$

this follows from the assumption on the distribution of X in Theorem 2. Application of Lemma 13 yields

$$\begin{aligned} & \mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n} \cdot f : f \in \mathcal{F}_n\right\}, x_1^n\right) \leq \mathcal{N}_1(\delta \cdot \beta_n, \mathcal{F}_n, x_1^n) \\ & \leq 3 \cdot \left(\frac{12e \cdot \beta_n}{\delta \cdot \beta_n}\right)^{c_{80} \cdot k^3 \cdot L \cdot d^* + d \cdot (c_{90} \cdot (\log n) \cdot n^{\frac{1}{2p+d^*}})^{d^*} \cdot (c_{91} \cdot \log n)^{(L-1) \cdot d^*} \cdot \left(\frac{r_n \cdot c_{92} \cdot \log n}{\delta \cdot \beta_n}\right)^{\frac{d^*}{k}}} \end{aligned}$$

for any $k \in \mathbb{N}$. Using

$$k = \lceil (c_7 + 2) \cdot \log n \rceil$$

we conclude from (15) for $\delta > 1/n^2$

$$\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n} \cdot f : f \in \mathcal{F}_n\right\}, x_1^n\right)$$

$$\begin{aligned}
&\leq 3(12e \cdot n)^{c_{93} \cdot (\log n)^{3 \cdot d^* \cdot L + d} \cdot (\log n)^{d^*} \cdot n^{\frac{d^*}{2p+d^*}} \cdot (\log n)^{(L-1) \cdot d^*} \cdot (n^{c_7+2})^{\frac{d^*}{(c_7+2) \cdot \log n}}} \\
&\leq 3(12e \cdot n)^{c_{94} \cdot (\log n)^{4 \cdot d^* \cdot L + d} \cdot n^{\frac{d^*}{2p+d^*}}}.
\end{aligned}$$

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

$$\mathbf{P}\{T_{2,n} > u\} \leq 14 \cdot c_{95} \cdot n^{c_{96} \cdot (\log n)^{4 \cdot d^* \cdot L + d} \cdot n^{\frac{d^*}{2p+d^*}}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot u\right).$$

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

$$\begin{aligned}
\mathbf{E}\{T_{2,n}\} &\leq \epsilon_n + \int_{\epsilon_n}^{\infty} \mathbf{P}\{T_{2,n} > u\} du \\
&\leq \epsilon_n + 14 \cdot c_{95} \cdot n^{c_{96} \cdot (\log n)^{4 \cdot d^* \cdot L + d} \cdot n^{\frac{d^*}{2p+d^*}}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot \epsilon_n\right) \cdot \frac{5136 \cdot \beta_n^2}{n}.
\end{aligned}$$

Setting

$$\epsilon_n = \frac{5136 \cdot \beta_n^2}{n} \cdot c_{96} \cdot (\log n)^{4 \cdot d^* \cdot L + d} \cdot n^{\frac{d^*}{2p+d^*}} \cdot \log n = \frac{5136 \cdot \beta_n^2}{n} \cdot \log\left(n^{c_{96} \cdot (\log n)^{4 \cdot d^* \cdot L + d} \cdot n^{\frac{d^*}{2p+d^*}}}\right)$$

yields the assertion of the sixth step of the proof.

In the *seventh step of the proof* we show

$$\mathbf{E}\{T_{4,n}\} \leq c_{97} \cdot n^{-\frac{2p}{2p+d^*}}.$$

Since $|T_{\beta_n} z - y| \leq |z - y|$ for $|y| \leq \beta_n$ we have

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

We will use Lemma 5 to derive an upper bound on $F_n(\mathbf{w}^{(t_n)})$ on E_n .

By the first step of the proof we know that (36) holds.

Next we choose $w_1^*, \dots, w_{r_n}^*$ as in Lemma 11 and define $\mathbf{w}^*$ by

$$(\mathbf{w}^*)_{i,j}^{(l)} = \begin{cases} w_j^* & \text{if } l = L, i = 1 \\ (\mathbf{w}^{(0)})_{i,j}^{(l)} & \text{if } l \in \{1, \dots, L-1\} \end{cases}$$

if E_n holds and set $\mathbf{w}^* = \mathbf{w}^{(0)}$ else. Then we have

$$\|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 \leq \sum_{k=1}^{r_n} |w_k^*|^2 \leq \frac{c_{98} \cdot n^{4L \cdot r^{L-1} \cdot (d+1) + 4}}{r_n}.$$

By Lemma 8 we know that (37) holds with $C_n = c_{99} \cdot \sqrt{r_n} \cdot (\log n)^{2L+1}$, where we apply Lemma 8 with $B_n = \gamma_n^* = c_{100} \cdot \log n$, which is possible because of the result of the second step of the proof and our initial choice of $\mathbf{w}^{(0)}$. Here we have used that for any $\mathbf{w}$ with $\|\mathbf{w}^* - \mathbf{w}\| \leq \|\mathbf{w}^* - \mathbf{w}^{(0)}\|$ we have

$$\sum_{t=1}^{r_n} |w_{1,t}^{(L)}|^2 \leq 2 \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 + 2 \cdot \sum_{t=1}^{r_n} |(\mathbf{w}^*)_{1,t}^{(L)}|^2 \leq c_{101} \cdot (\log n)^2.$$

This implies that also (40) holds, and since $f_{\mathbf{w}^*}$ approximates m in supremum norm well on E_n and m is bounded, we can assume w.l.o.g. that n is so large that (39) holds.

So the assumptions of Lemma 5 are satisfied, and we can conclude

$$\begin{aligned} & \left[2 \cdot F_n(\mathbf{w}^{(t_n)}) - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right] \cdot 1_{E_n} \\ & \leq \left[2 \cdot F_n(\mathbf{w}^*) + 2 \cdot \left(5 \cdot \beta_n \cdot C_n + \frac{1}{2 \cdot 1} \right) \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 + 2 \cdot F_n(\mathbf{w}^{(0)}) \cdot \frac{1}{1} \right. \\ & \quad \left. - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right] \cdot 1_{E_n} \end{aligned}$$

This yields

$$\begin{aligned} T_{4,n} & \leq 2 \cdot \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^*}(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 + c_{102} \cdot \frac{\log n}{n} \\ & \quad + 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \cdot 1_{E_n^c}. \end{aligned}$$

Because of

$$\begin{aligned} & \mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^*}(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right\} \\ & = \mathbf{E} \int |f_{\mathbf{w}^*}(x) - m(x)|^2 \mathbf{P}_X(dx) \\ & \leq c_{103} \cdot n^{-\frac{2p}{2p+d^*}} + c_{104} \cdot n^{c_{105}} \cdot \exp(-(\log n)^2) \\ & \leq c_{106} \cdot n^{-\frac{2p}{2p+d^*}} \end{aligned}$$

and

$$\begin{aligned} & \mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \cdot 1_{E_n^c} \right\} \\ & \leq \frac{1}{n} \sum_{i=1}^n \sqrt{\mathbf{E} \{ |m(X_i) - Y_i|^4 \}} \cdot \sqrt{\mathbf{P}(E_n^c)} \leq \frac{c_{107}}{n} \end{aligned}$$

this implies the assertion. $\square$

References

- [1] Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep learning via over-parameterization. *Proceedings of the 36th International Conference on Machine Learning (PMLR 2019)* **97**, pp. 242-252.
- [2] Armijo, L. (1966). Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics* **16**, pp. 1–3.
- [3] Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *Proceedings of the 36th International Conference on Machine Learning (PLMR 2019)*, pp. 322-332.
- [4] Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*, **14**, pp. 115-133.
- [5] 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.
- [6] Braun, A., Kohler, M., Langer, S., and Walk, H. (2024). Convergence rates for shallow neural networks learned by gradient descent. *Bernoulli*, **30**, pp. 475-502.
- [7] Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society* **46**, pp. 255–308.
- [8] Cheng, H., Zhang, M., and Shi, J. Q. (2024). A survey on deep neural network pruning: taxonomy, comparison, analysis, and recommendations. arXiv: 2308.06767.
- [9] Chizat, L., and Bach, F. (2016). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems* **31**, pp. 3036-3046.
- [10] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* **2**, pp. 303-314.
- [11] Drews, S. and Kohler, M. (2024). On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent. *Annals of the Institute of Statistical Mathematics* **76**, pp. 361–391.
- [12] Drews, S., and Kohler, M. (2026). Analysis of the expected L_2 error of an over-parametrized deep neural network estimate learned by gradient descent without regularization. *Electronic Journal of Statistics* **20**, pp. 163–214.
- [13] Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. *Proceedings of the 36th International Conference on Machine Learning (PMLR 2019)* **97**, pp. 1675-1685.

- [14] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer.
- [15] Hornik, K., Stinchcombe, M., and White, H. (1989). Multi-layer feedforward networks are universal approximators. *Neural Networks* **2**, pp. 359–366.
- [16] Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *Advances in neural information processing systems* **31**, pp. 8571-8580.
- [17] Jiao, Y., Shen, G., Lin, Y. and Huang, J. (2023). Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *Annals of Statistics* **51**, pp. 691–716.
- [18] Kenneweg, P., Kenneweg, T., and Hammer, B. (2024). Improving line search methods for large scale neural network training. arXiv: 2403.18519.
- [19] Kim, Y. (2014). Convolutional neural networks for sentence classification. Preprint, *arXiv: 1408.5882*.
- [20] Kohler, I. A., Gödicke, O., Kuder, T. A., Ladd, M. E., and Hesser, J. (2026). From whole-slide histology to ADC maps: Fast diffusion MRI simulation with neural operators. bioRxiv: 2026.07.22.740029.
- [21] Kohler, M. (2014). Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design. *Journal of Multivariate Analysis* **13**, pp. 197-208.
- [22] Kohler, M. (2026). On the rate of convergence of deep neural network regression estimates learned by gradient descent. *IEEE Transactions on Information Theory* **72(3)**, pp. 1777-1797.
- [23] Kohler, M. and Krzyżak, A. (2017). Nonparametric regression based on hierarchical interaction models. *IEEE Transaction on Information Theory* **63**, pp. 1620-1630.
- [24] Kohler, M. and Krzyżak, A. (2021). Over-parametrized deep neural networks minimizing the empirical risk do not generalize well. *Bernoulli*, **27**, pp. 2564-2597.
- [25] Kohler, M. and Krzyżak, A. (2025a). Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent. *IEEE Transactions on Information Theory* **71**, pp. 6165-6182.
- [26] Kohler, M. and Krzyżak, A. (2025b). Statistically guided deep learning. Preprint, *arXiv: 2504.08489v.1*.
- [27] Kohler, M., Krzyżak, A. and Molinero Römer, V. (2026). Estimation of a regression function from dependent data by over-parametrized deep neural networks learned by gradient descent. Preprint.

- [28] 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.
- [29] Kohler, M., Langer, L. and Reif, U. (2023). Estimation of a regression function on a manifold by fully connected deep neural networks. *Journal of Statistical Planning and Inference* **222**, pp. 160-181.
- [30] Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira et al. (Eds.), *Advances In Neural Information Processing Systems* Red Hook, NY: Curran. **25**, pp. 1097-1105.
- [31] Langer, S. (2020). Approximating smooth functions by deep neural networks with sigmoid activation function. *Journal of Multivariate Analysis* **182**, pp. 104696.
- [32] Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems* **32**, pp. 8572-8583.
- [33] Lu, J., Shen, Z., Yang, H. and Zhang, S. (2020). Deep network approximation for smooth functions. Preprint, *arXiv: 2001.03040*
- [34] Mei, S., Misiakiewicz, T., and Montanari, A. (2019). Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. *Proceedings of the 32nd Conference on Learning Theory (PLMR)* **99**, pp.2388–2464.
- [35] Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences* **115**, pp. E7665-E7671.
- [36] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X. and Gao, J. (2025). Large language models: a survey. Preprint, *arXiv: 2402.06196v3*.
- [37] Nguyen, M., and Mücke, N. (2024). How many neurons do we need? A refined analysis for shallow networks trained with gradient descent. *Journal of Statistical Planning and Inference* **233**, Article 106169.
- [38] Nguyen, P.-M., and Pham, H. T. (2023). A rigorous framework for the mean field limit of multilayer neural networks. *Mathematical Statistics and Learning* **6**, pp. 201-357.
- [39] Nitanda, A., and Suzuki, T. (2021). Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *arXiv: 2006.12297*.
- [40] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function (with discussion). *Annals of Statistics* **48**, pp. 1875–1897. Preprint, *arXiv:1708.06633v2*.

- [41] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Huber, T., et al. (2017). Mastering the game of go without human knowledge. *Nature* **550**, pp. 354-359.
- [42] Sirignano, J., and Spiliopoulos, K. (2020). Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics* **80(2)**, pp. 725-752.
- [43] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, pp. 1040-1053.
- [44] Vaswani, S., and Babanezhad, R. (2026). Armijo line-search can make (stochastic) gradient descent provably faster. arXiv: 2503.00229.
- [45] Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M., Macherey, W., Krikum, M., et al. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. Preprint, *arXiv: 1609.08144*.
- [46] Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, **94**, pp. 103-114.
- [47] Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. Preprint, *arXiv: 1802.03620*.
- [48] Yarotsky, D. and Zhevnerchuk, A. (2019). The phase diagram of approximation rates for deep neural networks. Preprint, *arXiv: 1906.09477*.