In Theory and Practice – On the Rate of Convergence of Implementable Neural Network Regression Estimates

Zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Dissertation von Alina Braun (M.Sc.) aus Frankfurt am Main
Tag der Einreichung: 21. Mai 2021, Tag der Prüfung: 25. Juni 2021

1. Gutachten: Prof. Dr. Michael Kohler
2. Gutachten: Prof. Dr. Volker Betz
Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

Mathematics Department
Research Group Probability and Statistics

In Theory and Practice – On the Rate of Convergence of Implementable Neural Network Regression Estimates

Doctoral thesis by Alina Braun (M.Sc.)

1. Review: Prof. Dr. Michael Kohler
2. Review: Prof. Dr. Volker Betz

Date of submission: 21. Mai 2021
Date of thesis defense: 25. Juni 2021

Darmstadt

Bitte zitieren Sie dieses Dokument als:
URN: urn:nbn:de:tuda-tuprints-190528
URL: http://tuprints.ulb.tu-darmstadt.de/19052

Dieses Dokument wird bereitgestellt von tuprints,
E-Publishing-Service der TU Darmstadt
http://tuprints.ulb.tu-darmstadt.de
tuprints@ulb.tu-darmstadt.de

Die Veröffentlichung steht unter folgender Creative Commons Lizenz:
Namensnennung - Nicht kommerziell - Keine Bearbeitungen 4.0 International
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de

Acknowledgements

I thank my thesis supervisor Prof. Dr. Michael Kohler for his support and for his expertise.

I would like to express my gratitude to Prof. Dr. Adam Krzyżak of the Concordia University and to Prof. Dr. Harro Walk of the Universität Stuttgart for the collaborations.

I thank Prof. Dr. Volker Betz for his review.

Thank you to my dear family, friends and colleagues.

Patrick, thank you.

Mama, thank you.

Abstract

In theory, recent results in nonparametric regression show that neural network estimates are able to achieve good rates of convergence provided suitable assumptions on the structure of the regression function are imposed. However, these theoretical analyses cannot explain the practical success of neural networks since the theoretically studied estimates are defined by minimizing the empirical L_2 risk over a class of neural networks and in practice, solving this kind of minimization problem is not feasible. Consequently, the neural networks examined in theory cannot be implemented as they are defined. This means that neural network in applications differ from the ones that are analyzed theoretically.

In this thesis we narrow the gap between theory and practice. We deal with neural network regression estimates for $(p, C)-$smooth regression functions m that satisfy a projection pursuit model. We construct three implementable neural network estimates and show that each of them achieve up to a logarithmic factor the optimal univariate rate of convergence.

Firstly, for univariate regression functions with $p \in [-\frac{1}{2}, 1]$ we construct a neural network estimate with one hidden layer where the weights are learned via gradient descent. The starting weights are randomly chosen from an interval independently of the data. The interval is large enough to guarantee that the estimate is close to a piecewise constant approximation.

Secondly, for multivariate regression functions with $p \in (0, 1]$ we construct a neural network estimate with one hidden layer where the weights are learned via gradient descent. The initial weights are chosen from specific intervals dependently on the data and the projection directions. This choice guarantees that the estimate is close to a piecewise constant approximation. The projection directions are repeatedly chosen randomly.

Lastly, for multivariate regression functions with $p > 0$ we construct a multilayer neural network estimate. The value of the inner weights are prescribed dependently on the projection directions by a new approximation result for a projection pursuit model by piecewise polynomials. The outer weights are chosen by solving a linear equation system. The projection directions are repeatedly chosen randomly.

Since we are able to show a rate of convergence that is independent of the dimension of the data our second and third estimates are able to circumvent the curse of dimensionality.

Zusammenfassung

Theoretische Resultate in der Nichtparametrischen Regressionsschätzung zeigen, dass unter geeigneten Annahmen an die Regressionsfunktion neuronale Netze Schätzer gute Konvergenzraten erreichen. Jedoch werden die dort untersuchten neuronalen Netze durch ein nicht praktikables Minimierungsproblem des empirischen L_2 Risikos über einer Klasse von neuronalen Netzen definiert. Folglich können diese theoretisch untersuchten neuronalen Netze nicht so implementiert werden, wie sie definiert werden. Also unterscheiden sich die in der Praxis verwendeten neuronalen Netze von den in der Theorie behandelten. In dieser Thesis verringern wir diese Kluft zwischen praktisch angewandten und theoretisch untersuchten neuronalen Netzen. Wir befassen uns mit neuronale Netze Schätzern für $(p, C)-$glatte Regressionsfunktionen m, die das Projection Pursuit Modell erfüllen. Wir konstruieren drei implementierbare neuronale Netze Schätzer und zeigen, dass diese bis auf einen logarithmischen Faktor die optimale univariate Konvergenzrate erreichen.
Zuerst konstruieren wir für univariate Regressionsfunktionen mit $p \in [-\frac{1}{2}, 1]$ einen neuronalen Netze Schätzer mit einer verdeckten Schicht, in dem die Gewichte durch das Gradientenabstiegsverfahren gelernt werden. Die Startgewichte werden zufällig aus einem Intervall gewählt, das groß genug ist, um zu garantieren, dass unser Schätzer nahe an einer stückweisen konstanten Approximation ist.
Danach konstruieren wir für multivariate Regressionsfunktionen mit $p \in (0, 1]$ einen neuronalen Netze Schätzer mit einer verdeckten Schicht, in dem die Gewichte durch das Gradientenabstiegsverfahren gelernt werden. Die Startgewichte werden aus speziellen Intervallen, abhängig von den Daten und der Projektionsrichtungen gewählt. Diese Wahl garantiert, dass unser Schätzer nahe an einer stückweisen konstanten Approximation ist. Die Projektionrichtungen werden wiederholt zufällig gewählt.
Zuletzt konstruieren wir für multivariate Regressionsfunktionen mit $p > 0$ einen neuronalen Netze Schätzer mit vielen verdeckten Schichten. Die inneren Gewichte werden durch ein neues Approximationsresultat für das Projection Pursuit Modell durch stückweise Polynome vorgegeben. Die äußeren Gewichte werden durch Lösen eines linearen Gleichungssystems bestimmt. Die Projektionrichtungen werden wiederholt zufällig gewählt. Da wir eine von der Dimension der Daten unabhängige Konvergenzrate zeigen, können unser zweiter und unser dritter Schätzer den Fluch der Dimensionalität umgehen.

Contents

Notation

In the following we provide a short summary of the notation we are going to use in this thesis.

Sets	
$\mathbb{N}$	the set of natural numbers
$\mathbb{N}_0$	the set of natural numbers including zero
$\mathbb{Z}$	the set of integer numbers
$\mathbb{R}_+$	the set of non-negative real numbers
$\mathbb{R}$	the set of real numbers

Operations on $z \in \mathbb{R}$	
$\lceil z \rceil$	the smallest integer number greater than or equal to z
$\lfloor z \rfloor$	the greatest integer number smaller than or equal to z
$\mathbf{1}_A(x)$	the indicator function of x on a set $A \subseteq \mathbb{R}$
$z_+ = \max\{z, 0\}$	z, if z is greater than zero, 0 otherwise
$T_\beta z = \max\{\min\{z, \beta\}, -\beta\}$	the truncation of z by $\beta > 0$

Balls around $x \in \mathbb{R}^d$	
$S_\epsilon(x)$	the $1-$dimensional closed ball around $x \in \mathbb{R}$ with radius $\epsilon > 0$
$S_\epsilon^{(d)}(x)$	the $d-$dimensional closed ball around $x \in \mathbb{R}^d$ with radius $\epsilon > 0$
$\partial S_\epsilon^{(d)}(x)$	the $(d-1)-$dimensional surface of the $d-$dimensional ball around $x \in \mathbb{R}^d$ with radius $\epsilon > 0$

Norms for $x = (x^{(1)}, \ldots, x^{(d)}) \in \mathbb{R}^d$

$\|x\|_1 = \sum_{i=1}^d \left\|x^{(i)}\right\|$	the Manhatten or 1-Norm of x	
$\|x\| = \|x\|_2 = \sqrt{\sum_{i=1}^d \left\|x^{(i)}\right\|^2}$	the Euclidean norm of x	
$\|x\|_\infty = \max_{i=1,\ldots,d} \left\|x^{(i)}\right\|$	the supremum norm of x	

Norms for a function $f : \mathbb{R}^d \to \mathbb{R}$

$\|f\|_\infty = \sup_{x\in\mathbb{R}^d} \|f(x)\|$	the supremum norm of f	
$\|f\|_{\infty,A} = \sup_{x\in A} \|f(x)\|$	the supremum norm of f on a set $A \subseteq \mathbb{R}^d$	
$\|f\|_{L_p(\nu)} = \left(\int \|f(x)\|^p \nu(dx)\right)^{\frac{1}{p}}$	the L_p-norm of f for $1 \leq p < \infty$ with respect to a probabilty measure ν on $\mathbb{R}^d$	

Operations on a class $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$

$T_\beta\mathcal{F} = \{T_\beta f : f \in \mathcal{F}\}$	the class of functions containing all functions in $\mathcal{F}$ truncated by β	

Complexity of a class $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$

$\mathcal{N}\left(\epsilon, \mathcal{F}, \|\cdot\|_{L_p(\nu)}\right)$	the $\epsilon-$covering number of $\mathcal{F}$ with respect to $\|\cdot\|_{L_p(\nu)}$ for $\epsilon > 0$	(👁)
$\mathcal{N}_p\left(\epsilon, \mathcal{F}, x_1^n\right)$	the $L_p - \epsilon-$covering number of $\mathcal{F}$ on x_1^n for $\epsilon > 0$	(👁)
$\mathcal{M}\left(\epsilon, \mathcal{F}, \|\cdot\|_{L_p(\nu)}\right)$	the $\epsilon-$packing number of $\mathcal{F}$ with respect to $\|\cdot\|_{L_p(\nu)}$ for $\epsilon > 0$	(👁)
$\mathcal{M}_p\left(\epsilon, \mathcal{F}, x_1^n\right)$	the $L_p - \epsilon-$packing number of $\mathcal{F}$ on x_1^n for $\epsilon > 0$	(👁)

Complexity of a class $\mathcal{A}$ of sets $A \subset \mathbb{R}^d$

$S(\mathcal{A}, n)$	the $n-$th shatter coefficient of $\mathcal{A}$	(👁)
$V_\mathcal{A}$	the Vapnik-Chervonenkis dimension of $\mathcal{A}$	(👁)

For a $d-$dimensional random variable X with distribution $\mathbf{P}_X$

$supp(X)$ or $supp(\mathbf{P}_X)$	the support of X with respect to $\mathbf{P}_X$	(👁)

(👁) : For a definiton see Supplement Section $A.1$.

1. Introduction and Overview of the Results

Ever since McCulloch and Pitts (1943) introduced their model of a neuron in the 1940's the advance of neural networks has been unstoppable. The notion of learning in a neural network was not part of the structure initially and came along later on, see e.g. Rummelhart et al. (1986). In particular, technical progress that led to the improvement of capacity and computing power in computers supported the development of neural networks. For a detailed overview of the history of neural networks see, e.g. Schmidhuber (2015). Today, neural networks are studied in numerous areas, for example the field of image classification (e.g. Krizhevsky, Sutskever and Hinton (2012)), the field of text classification (e.g. Kim (2014)), the field of machine translation(e.g. Wu et al. (2016)), the field of gaming (e.g. Silver et al. (2017), Tesauro (2012) and Yannakakis et al. (2004)), the field of virtual reality (e.g. Yang et al. (2018) and Weissmann and Salomon (2002)), the field of face reconstruction (e.g. Dou et al. (2017) and Richardson et al. (2017)), the field of autonomous driving (e.g. Tian et al. (2018)), the field of car plate recognition (e.g. Parisi et al. (2002)), the field of food production (e.g. Cotrim et al. (2020), Lamrini et al. (2012) and Sablani et al. (2002)), and in the medical field, such as the detection and classification of different types of cancer cells (e.g. Nasser and Abu-Naser (2019), Azar and El-Said (2012), Joshi et al. (2010)) or the detection of COVID-19 (e.g. Khan et al. (2020)). Naturally, this huge success in practical applications has motivated an increasing interest in the analysis of the theoretical background.
Formally, neural networks are functions of a specific structure. The smallest unit in a neural network is a neuron.

Definition 1.0.1. *A neuron is a function $f : \mathbb{R}^d \to \mathbb{R}$ defined as*

$$f(x) = \sigma\left(\sum_{i=1}^{d} w_i \cdot x^{(i)} + w_0\right),$$

where $w_0, w_1, \ldots, w_d$ are called the weights and

$$\sigma : \mathbb{R} \to \mathbb{R}$$

is called the activation function.

In this thesis we focus on the so-called sigmoid logistic squashing function or logistic squasher

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (x \in \mathbb{R}) \tag{1.1}$$

as the activation function. A bigger neural network is created by switching neurons together. This can be done in two ways: summing up neurons or concatenation of neurons. By the latter we create new layers in the network. Each concatenation corresponds to a new layer. By the former more neurons are added to a layer. Each summand corresponds to a new neuron. In visual representations we portray neurons by nodes and weights connecting the neurons by edges. Layers of the neural network are arranged horizontally. and neurons in the same layer are placed one below the other in a column. The input x is represented by d nodes (one for each component) and is called the input layer. The final output of the function is called the output layer. The layers between the input layer and the output layer are called hidden layers.

Definition 1.0.2. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function. Let $L \in \mathbb{N}$ and $\mathbf{k} = (k_1, \ldots, k_L) \in \mathbb{N}^L$. A multilayer feedforward neural network with L hidden layers and k_r neurons in the $r-$th hidden layer $(r = 1, \ldots, L)$ is a function $f : \mathbb{R}^d \to \mathbb{R}$ which is recursively defined as*

$$f(x) = \sum_{i=1}^{k_L} c_i^{(L)} \cdot f_i^{(L)}(x) + c_0^{(L)} \tag{1.2}$$

with output weights $c_0^{(L)}, c_1^{(L)}, \ldots, c_{k_L}^{(L)} \in \mathbb{R}$, and with hidden layers $r = 2, \ldots, L$

$$f_i^{(r)}(x) = \sigma\left(\sum_{j=1}^{k_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)}\right) \tag{1.3}$$

with inner weights $c_{i,0}^{(r-1)}, c_{i,1}^{(r-1)}, \ldots, c_{i,k_{r-1}}^{(r-1)} \in \mathbb{R}$ and lastly, with first hidden layer $r = 1$

$$f_i^{(1)} = \sigma\left(\sum_{j=1}^{d} c_{i,j}^{(0)} \cdot x^{(j)} + c_{i,0}^{(0)}\right) \tag{1.4}$$

for inner weights $c_{i,0}^{(0)}, c_{i,1}^{(0)}, \ldots, c_{i,d}^{(0)} \in \mathbb{R}$.

For a better understanding, a visualization is shown in Figure 1.1.

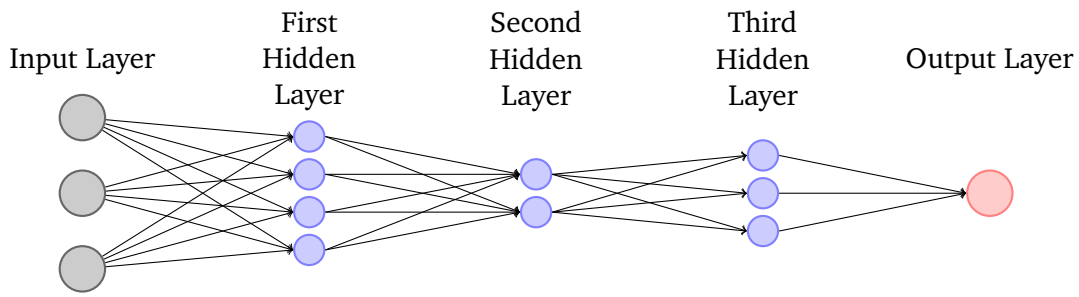

Figure 1.1.: Visualisation of a neural network. The function has a 3-dimensional input (black node) and a 1-dimensional output (red node). There are three hidden layers, the first consisting of 4 neurons, the second consisting of 2 neurons and the third consisting of 3 neurons (blue nodes).

We study neural networks in the context of nonparametric regression with random design. In this setting (X, Y) is an $\mathbb{R}^d \times \mathbb{R}-$valued random vector satisfying

$$\mathbf{E}\{Y^2\} < \infty.$$

The objective is to analyse the dependency of the so-called outcome variable Y on the so-called predictor variable X. For that we are given a sample set

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

of (X, Y) where $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed (in the following i.i.d.) random variables. The regression function $m : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$m(x) = \mathbf{E}\{Y \mid X = x\} \quad (x \in \mathbb{R}^d).$$

The goal is to construct an estimate

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

which, based on the given data set $\mathcal{D}_n$, approximates the regression function m such that the L_2-errror

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

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

Neural networks are considered to be one of the best approaches in nonparametric statistics, especially in multivariate statistical applications like pattern recognition and nonparametric regression, see, e.g., Hertz, Krogh and Palmer (1991), Devroye, Györfi and Lugosi (1996), Anthony and Bartlett (1999), Györfi et al. (2002), Haykin (2008) and Ripley (2008). In recent years, deep learning, where multilayer feedforward neural networks with many hidden layers are fitted to observed data, has also shifted into focus in applications, see, e.g., Schmidhuber (2015) and the literature cited therein.
By the so called slow-rate convergence result we know that in order to be able to derive non-trivial convergence rate results we need smoothness assumptions on the regression function (cf., e.g., Theorem 7.2 and Problem 7.2 in Devroye, Györfi and Lugosi (1996) and Section 3 in Devroye and Wagner (1980)). In this thesis we assume that the regression function m is $(p, C)-$smooth according to the following definition.

Definition 1.0.3. *Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $0 < s \leq 1$, where $\mathbb{N}_0$ is the set of nonnegative integers. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-smooth, if for every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = q$ the partial derivative $\frac{\partial^q f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ exists and satisfies*

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

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

It has been shown by Stone (1982) that the optimal minimax rate of convergence in nonparametric regression for (p, C)-smooth functions is

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

We see that with higher dimension d compared to the smoothness parameter p of the regression function this rate of convergence becomes increasingly slow. This is referred to as the so-called curse of dimensionality.

The natural questions to ask are:

> Do the neural network regression estimates achieve the optimal rate of convergence?
>
> Is it possible to circumvent the curse of dimensionality with the neural network regression estimates?

In order to circumvent the curse of dimensionality, a number of possible constraints can be imposed onto the structure of the regression function. For example, Stone (1985), assumed that the regression function is additive, meaning $m : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$m(x^{(1)}, \dots, x^{(d)}) = m_1(x^{(1)}) + \cdots + m_d(x^{(d)}) \quad (x^{(1)}, \dots, x^{(d)} \in \mathbb{R})$$

for some $(p, C)-$smooth univariate functions $m_1, \dots, m_d : \mathbb{R} \to \mathbb{R}$. He showed that, in this case, suitably defined spline estimates achieve a one-dimensional rate of convergence. Stone (1994) extended this result to interaction models. There, the regression function is assumed to be a sum of functions applied to at most $d^* < d$ components of x. In this case, he showed that suitably defined spline estimates achieve the d^*–dimensional rate of convergence.

Other constraints include the class of single index models, where

$$m(x) = g(\mathbf{a}^T x) \quad (x \in \mathbb{R}^d)$$

for some $\mathbf{a} \in \mathbb{R}^d$ and $g : \mathbb{R} \to \mathbb{R}$ see, e.g., Härdle and Stoker (1989), Härdle, Hall and Ichimura (1993), Yu and Ruppert (2002), Kong and Xia (2007) and Lepski and Serdyukova (2014) and, as we use it, the class projection pursuit models, where

$$m(x) = \sum_{s=1}^{r} g_s(\mathbf{a}_s^T x) \quad (x \in \mathbb{R}^d)$$

for some $r \in \mathbb{N}$, $\mathbf{a}_s \in \mathbb{R}^d$, and (p, C)-smooth functions $g_s : \mathbb{R} \to \mathbb{R}$ $(s = 1, \dots, r)$, see, e.g., Friedman and Stuetzle (1981) and Huber (1985). It has been shown that in this context that if the regression function is (p, C)-smooth then suitably defined (nonlinear) least squares estimates achieve the univariate rate of convergence $n^{-\frac{2p}{(2p+1)}}$ up to some logarithmic factor, see Section 22.3 in Györfi et al. (2002).

Horowitz and Mammen (2007) dealt with a generalization of projection pursuit, where the regression function satisfies

$$m(x) = g\left(\sum_{l_1=1}^{L_1} g_{l_1}\left(\sum_{l_2=1}^{L_2} g_{l_1,l_2}\left(\dots \sum_{l_r=1}^{L_r} g_{l_1,\dots,l_r}(x^{l_1,\dots,l_r})\right)\right)\right),$$

where $g, g_{l_1}, \dots, g_{l_1,\dots,l_r}$ are (p, C)-smooth univariate functions and $x^{l_1,\dots,l_r}$ are single components of $x \in \mathbb{R}^d$ (not necessarily different for two different indices $(l_1, \dots, l_r)$). They showed that a penalized least squares achieves the univariate rate of convergence $n^{-\frac{2p}{(2p+1)}}$.

The estimates considered in Section 22.3 in Györfi et al. (2002) for the projection pursuit model and in Horowitz and Mammen (2007) for its generalization are both nonlinear (penalized) least squares estimates. In practice it is unclear how these can be computed exactly. Friedman and Stuetzle (1981) described easily implementable estimates for projection pursuit, but in their definition a great deal of heuristic simplifications were used, which consequently makes it unclear whether it is possible to show any rate of convergence for their estimates or not.

With respect to the L_2 error of a neural network with one hidden layer Barron (1993, 1994) proved a rate of convergence $n^{-\frac{1}{2}}$ (up to some logarithmic factor) which is independent of the dimension, under the assumption that the Fourier transform has a finite first moment. Simply put, this requires that the function becomes smoother the higher the dimension d of X. McCaffrey and Gallant (1994) dealt with a certain cosine squasher as the activation function and showed a rate of convergence $n^{-\frac{2p}{2p+d+5}+\varepsilon}$ for the L_2 error of suitably defined neural network estimates with one hidden layer.

Various theoretical results are based on the derivation of new approximation results for piecewise polynomials by neural networks. In these works circumventing the curse of dimensionality is achieved by exploiting compository assumptions on the structure of the regression function through the network structure. Kohler and Krzyżak (2017) were the first to show that neural networks can achieve dimensionality reduction under the restriction that the regression function is a composition of (sums of) functions, where each of the function is a function of at most $d^* < d$ variables. There, it was shown that suitably defined multilayer neural networks achieve the rate of convergence $n^{-2p/(2p+d^*)}$ (up to some logarithmic factor) for $p \leq 1$. Bauer and Kohler (2019) extended this result to $p > 1$, provided the squashing function is suitably chosen. Kohler and Langer (2019) showed that these results also apply to very simply constructed fully connected feedforward neural networks. Schmidt-Hieber (2019) worked with ReLU activation function and obtained similar results as Bauer and Kohler (2019). For regression functions with low local dimensionality Kohler, Krzyżak and Langer (2020) showed that neural networks are able to circumvent the curse of dimensionality. Imaizumi and Fukamizu (2019) derived results concerning the estimation of piecewise polynomials with partitions with rather general smooth boundaries as regression functions.

The above results show that least squares neural network regression estimates are able to circumvent the curse of dimensionality under much more general assumptions than the projection pursuit model that will be assumed in this thesis. However, the key issue in all of the articles above is that they share the same definition of the neural network regression estimate as a nonlinear least squares estimate. Such an estimate is, for example, defined

as the miminum of the empirical L_2 risk

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

over a nonlinear class $\mathcal{F}_n$ of neural networks. Finding the global minimum of the empirical L_2 risk over a class of neural networks is practically impossible. Hence, the above mentioned estimates cannot be computed in practice. In practice, it is usually tried to find a local minimum instead using, for example, the gradient descent algorithm (so–called backpropagation).

Gradient descent is a topic of interest in many studies, see Poggio, Banburski and Liao (2020) for an overview and Karimi, Nutini and Schmidt (2018) and the literature cited therein. A standard reference is the monograph Luenberger and Ye (2016). As an early paper, Poljak (1981) should not go unmentioned where additionally, the case of noise corrupted function values was also considered. This case is also dealt with in stochastic approximation, see, e.g., the monograph Kushner and Yin (2003), where in a classic situation the constant factor at the gradient was replaced by a decreasing factor at a vector of divided differences (multidimensional Kiefer-Wolfowitz method). White (1989) and Fabian (1994) brought the two fields of stochastic approximation and neural network models together. In Dippon and Fabian (1994) and Dippon (1998) it is explained how gradient descent in stochastic approximation can be combined with a slowly convergent global optimizer in order to find not only a local but even a global minimum of a general function. The main difficulty of using such results to derive rate of convergence results for neural network regression estimates lies in the fact that for neural network regression estimates the networks adapts to the sample size and consists of more neurons the bigger the size. As a consequence, it is not sufficient to analyze gradient descent applied to a fixed function where the number of steps is tending to infinity. Instead the function is changing for increasing number of steps. Basically, this requires the ability to analyze the behaviour of gradient descent for a finite number of steps. As far as we know such results do not exist in the literature for general functions like the empirical L_2 risk of a neural network (which is not convex and has neither a global minimum nor an easily analysable Hessian matrix considered as a real-valued function of the weight vector).

There are a number of papers in computer science concerning the theoretical properties of neural network estimates learned by backpropagation. The most popular approach in this context is the so–called landscape approach. Choromanska et al. (2015) used random matrix theory to derive a heuristic argument showing that the risk of most of the local minima of the empirical L_2 risk is not much larger than the risk of the global minimum. It was possible to confirm this claim for neural networks with special activation

function, see, e.g., Arora et al. (2018), Kawaguchi (2016), and Du and Lee (2018), who have analyzed gradient descent for neural networks with linear or quadratic activation function. However, no good approximation results exist for such neural networks, making it impossible to derive rates of convergence results for these networks that are comparable to the ones above for the least squares neural network regression estimates.

Du et al. (2018) analyzed neural networks with one hidden layer learned by gradient descent in a setting where the input suffices a Gaussian distribution. They used the expected gradient instead of the gradient in their gradient descent routine. For that reason it is not possible to derive from their result a rate of convergence result for neural networks learned by gradient descent that is similar to the results for the least squares neural network estimates cited above. Liang et al. (2018) applied gradient descent to a modified loss function in classification, where it is assumed that the data can be interpolated by a neural network. Here, the last assumption is not satisfied in nonparametric regression and it is unclear whether the main idea (of simplifying the estimation by a modification of the loss function) can also be used in a regression setting. Several recent results including, e.g., Allen-Zhu, Li and Song (2019), Kawaguchi and Huang (2019) and the literature cited therein, showed that application of gradient descent to over-parameterized neural networks leads to small empirical L_2 risks. Here, it is also not possible to derive from their result a rate of convergence result for neural networks learned by gradient descent that is similar to the ones cited above for least squares neural network regression estimates, since it is not clear what the L_2 risk of the estimate is and exactly this term is critical in the derivation of such results. In particular, due to the fact that the networks are over-parameterized, a bound on the empirical L_2 risk might not be useful for bounding the L_2 risk. Kawaguchi and Huang (2019) presented a bound for the L_2 risk but they required that the weights in the network are small and it is not clear whether or not this condition can be guaranteed in an over-parameterized neural network learned by the gradient descent.

More results that analyse gradient descent in over-parameterized neural networks and show that for suitably chosen networks a global minimum of the empirical L_2 risk can be found include Allen-Zhu, Li and Liang (2019), Arora et al. (2019a, 2019b), Li and Liang (2018) and Zou et al. (2018). However, Kohler and Krzyżak (2019) presented a counterexample demonstrating that over-parameterized neural networks, which basically interpolate the training data, in general do not generalize well. In this counterexample the regression function is constant zero and hence satisfies the assumption on the regression function imposed in this thesis. In particular, this shows that results similar to the ones we will present in this thesis cannot be concluded from the papers cited above. Also, note the estimates considered in this thesis are not over-parameterized since the number of data-dependently chosen weights are much smaller than the sample size.

The generalization of neural networks can also be analyzed within the classical Vapnik Chervonenkis theory (cf., e.g., Györfi et al. (2002)). Here, the complexity of the underlying spaces of functions is measured by covering numbers, which can be bounded using the so–called Vapnik-Chervonenkis dimension (cf., e.g., Bartlett et al. (2019)). However, the resulting upper bounds on the generalization error might be too rough because during the gradient descent the neural network estimate does not attend all functions from the underlying function space. A similiar approach is to describe the complexity of the function space in which the estimate is contained during the gradient descent by the so–called Rademacher complexity (cf., Koltchinski (2004)). For neural networks with quadratic activation function this was already done successfully in Du and Lee (2018), but, unfortunately, such neural networks do not have good approximation properties and consequently it is not possible to derive any results that compare to those in this thesis.

We observe that, although theoretrical analyses of neural network learning show good results, the networks considered are not feasible and far away from those used in practice. Thus, the great perfomance of neural networks used in practice is a complete mystery from a theoretical point of view. Clearly, there is a gap between theoretical and practical results. In this thesis we narrow the gap between theoretical but unpractical and practical but theoretically foggy analysis of neural networks by constructing neural network regression estimates that are implementable and analyzing their rate of convergence theoretically. We present the following three results.

We start with a neural network estimate where the weights are chosen from a braod interval. For this we restrict our analysis to $1-$dimensional regression functions with smoothness factor $p \in [\frac{1}{2}, 1]$. The neural network regression estimate is inspired by approximation results for $(p, C)-$smooth functions with piecewise constant functions. We are able to analyze gradient descent in this network.

Result 1. We assume that the regression function

$$m : \mathbb{R} \to \mathbb{R}$$

is $(p, C)-$smooth and univariate. Consider networks with one hidden layer and K_n hidden neurons defined by

$$f_{net,\mathbf{w}}(x) = \sum_{j=1}^{K_n} w_{1,j}^{(1)} \cdot \sigma(w_{j,1}^{(0)} \cdot x + w_{j,0}^{(0)}) + w_{1,0}^{(1)}$$

$$= \sum_{j=1}^{K_n} \alpha_j \cdot \sigma(\beta_j \cdot x + \gamma_j) + \alpha_0 \tag{2.1}$$

where $K_n \in \mathbb{N}$, $\alpha_0, \alpha_i, \beta_i, \gamma_i \in \mathbb{R}$ $(i = 1, \ldots, K_n)$, the activation function is the logistic squasher

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

and

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

is the vector of weights. An example is shown in Figure 1.2. The construction procedure

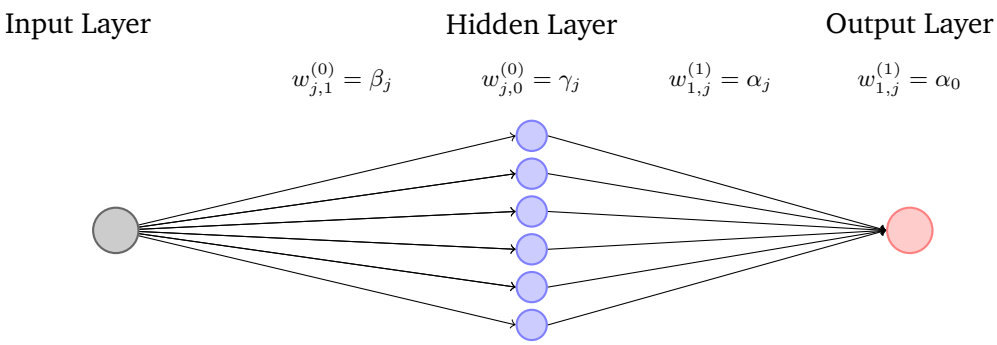

Figure 1.2.: Visualisation of an example of network estimate with parameter $K_n = 6$. The function has a 1-dimensional input (black node) as well as a 1-dimensional output (red node). There is one hidden layer consisting of 6 neurons (blue node).

for our neural network estimate has three steps:

1. Randomly choose our initial weights: Set

$$\mathbf{w}(0) = \mathbf{v} \tag{2.2}$$

where the initial weight vector

$$\mathbf{v} = (\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}) = (\alpha_0^{(0)}, \alpha_1^{(0)}, \ldots, \alpha_{K_n}^{(0)}, \beta_1^{(0)}, \ldots, \beta_{K_n}^{(0)}, \gamma_1^{(0)}, \ldots, \gamma_{K_n}^{(0)})$$

is chosen randomly such that

$$|\alpha_k^{(0)}| \leq \frac{c_2}{K_n} \quad (k = 1, \ldots, K_n)$$

and $\beta_1^{(0)}, \ldots, \beta_{K_n}^{(0)}, \gamma_1^{(0)}, \ldots, \gamma_{K_n}^{(0)}$ are independently uniformly distributed on the set $[-B_n, B_n]$. A choice of B_n is given in Theorem 2.2.1.

2. Apply gradient descent for $t = 0, 1, \ldots, t_n - 1$: Learn the weights by gradient descent. More precisely, we minimize the regularized least squares criterion

$$F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{net,\mathbf{w}}(X_i)|^2 + \frac{c_1}{K_n^{2p}} \cdot \sum_{k=0}^{K_n} \alpha_k^2, \tag{2.3}$$

where $c_1 > 0$ is an arbitrary constant by defining $\mathbf{w}(t+1)$ recursively by

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \lambda_n \cdot \nabla_{\mathbf{w}} F(\mathbf{w}(t)) \tag{2.4}$$

for $t = 0, 1, \ldots, t_n - 1$. We write

$$\mathbf{w}(t) = (\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) = (\alpha_0^{(t)}, \alpha_1^{(t)}, \ldots, \alpha_{K_n}^{(t)}, \beta_1^{(t)}, \ldots, \beta_{K_n}^{(t)}, \gamma_1^{(t)}, \ldots, \gamma_{K_n}^{(t)})$$

for $t = 0, 1, \ldots, t_n$. Here, $\lambda_n > 0$ denotes the stepsize and $t_n \in \mathbb{N}$ denotes the number of steps we perform in the gradient descent algorithm. A choice of λ_n and of t_n is given in Theorem 2.2.1. Note, that minimization of (2.3) with respect to $\mathbf{w}$ is a nonlinear least squares problem.

3. Set

$$\tilde{m}_n(\cdot) = f_{net,\mathbf{w}(t_n)}(\cdot) \tag{2.5}$$

and choose our estimate to be the by

$$\beta_n = c_3 \cdot \log n$$

truncated version

$$m_n(x) = T_{\beta_n} \tilde{m}_n(x). \tag{2.6}$$

For our neural network estimate we have the following up to a logarithmic factor optimal minimax rate of convergence.

Theorem 2.2.1. *Let (X, Y) be an $[0,1] \times \mathbb{R}$–valued random vector such that*

$$\mathbf{E}\left\{ e^{c_{10} \cdot Y^2} \right\} < \infty \tag{2.7}$$

hold for some constant $c_{10} > 0$ and assume that the corresponding regression function $m(x) = \mathbf{E}\{Y|X = x\}$ is (p, C)–smooth for some $p \in [\frac{1}{2}, 1]$ and $C > 0$. Let (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n) be independent and identically distributed. Set

$$K_n = \left\lceil c_5 \cdot n^{\frac{1}{2p+1}} \right\rceil, \quad B_n = 16 \cdot K_n \cdot (\log K_n)^2 \cdot n,$$

$$L_n = c_6 \cdot (\log n)^7 \cdot K_n^{3p+1}, \quad \lambda_n = \frac{1}{L_n}$$

and

$$t_n = \lceil K_n^{2p} \cdot (\log n)^2 \cdot L_n \rceil$$

and define the estimate m_n of m as described above. Then we have for n sufficiently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_7 \cdot (\log n)^{\max\{3,4p\}} \cdot n^{-2p/(2p+1)}$$

for some constant $c_7 > 0$ which does not depend on n.

Next, we analyze a neural network estimate for d-dimensional regression functions satisfying a projection pursuit model. Again, the unknown projection directions are chosen randomly and hence repeated initialization of the weights is needed. In contrast to the network presented in Result 1 the choice of the inital weights is more restricted. The inital outer weights are set to zero and the initial inner weights are chosen carefully from specific intervals dependent on $X_1, \ldots, X_n$ and on the projection directions. This means we are allowed less freedom in the choice of our weights but we trade it for analysis of high dimensional regression functions with $p \in (0, 1]$. Similarly to the network in Result 1, the neural network regression estimate is inspired by approximation results for $(p, C)-$smooth functions with piecewise constant functions. We are able to analyze gradient descent in this neural network. We show that our neural network estimate achieves the up to a logarithmic factor optimal univariate rate of convergence. Since this rate of convergence is independent of the dimension of the data, our second estimate can circumvent the curse of dimensionality.

Result 2. We deal with neural network regression in a projection pursuit model. This means, we assume that the regression function satisfies

$$m(x) = \sum_{s=1}^{r} g_s(\mathbf{c}_s^T x) \quad (x \in \mathbb{R}^d)$$

for some $r \in \mathbb{N}$, $\mathbf{c}_s \in \mathbb{R}^d$, where $\|\mathbf{c}_s\| = 1$ $(s = 1, \ldots, r)$, and (p, C)-smooth functions $g_s : \mathbb{R} \to \mathbb{R}$ $(s = 1, \ldots, r)$. We approximate m by networks with one hidden layer and $K \cdot r$ neurons in this hidden layer defined by

$$f_{net,(\mathbf{a},\mathbf{b})}(x) = \sum_{k=1}^{K \cdot r} a_k \cdot \sigma \left(\sum_{j=1}^{d} b_{k,j} \cdot x^{(j)} + b_{k,0} \right) + a_0, \tag{3.1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function and

$$a_k \in \mathbb{R} \quad (k = 0, \ldots, K \cdot r) \quad \text{and} \quad b_{k,j} \in \mathbb{R} \quad (k = 1, \ldots, K \cdot r, j = 0, \ldots, d)$$

are the weights. An example of the network is shown in Figure 1.3. The construction

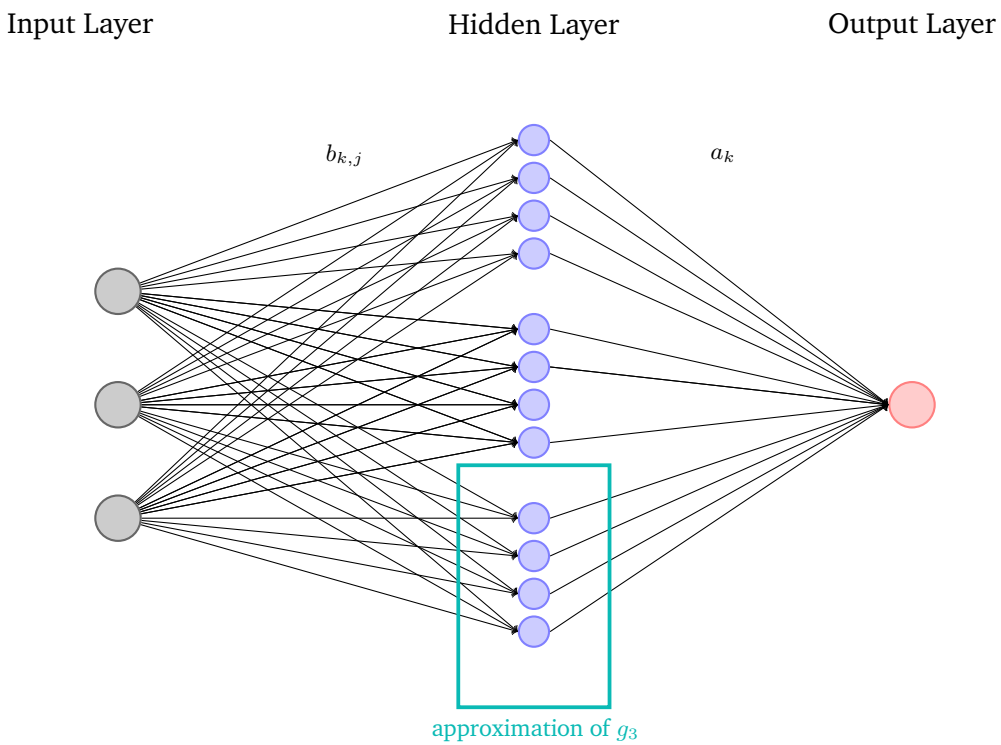

Figure 1.3.: Visualisation of an example of network estimate with parameters $d = 3$, $K = 4$ and $r = 3$. The function has a 3-dimensional input (black node) and a 1-dimensional output (red node). There is one hidden layer consisting of 12 neurons (blue node). The neurons in the turquoise box are neurons in the hidden layer of the example of the network approximation of g_3 (see Figure 3.2).

procedure for our neural network estimate has seven steps:

1. Randomly choose values

$$\mathbf{c}_1^*, \ldots, \mathbf{c}_r^* \in [-1, 1]^d$$

as an independent sample from a uniform distribution on $[-1,1]^d$ and set

$$\bar{\mathbf{c}}_s = \frac{\mathbf{c}_s^*}{\|\mathbf{c}_s^*\|} \quad (s = 1, \ldots, r).$$

These values approximate the direction of projection $\mathbf{c}_1, \ldots, \mathbf{c}_r$ in our projection pursuit model.

Note, that for $s = 1, \ldots, r$

$$\mathbf{P}\{\mathbf{c}_s^* = 0\} = 0,$$

which means we can assume w.l.o.g. $\|\mathbf{c}_s^*\| \neq 0$.

2. Prepare the definition of the initial inner weights $b_{(s-1)\cdot K+k,j}$ ($s = 1, \ldots, r$, $k = 1, \ldots, K$, $j = 0, \ldots, d$) according to $\bar{\mathbf{c}}_s$ and to $X_1, \ldots, X_n$: For each $s \in \{1, \ldots, r\}$ choose $\tilde{b}_{s,1}, \ldots, \tilde{b}_{s,K} \in \mathbb{R}$ such that $\tilde{b}_{s,1} < \tilde{b}_{s,2} < \cdots < \tilde{b}_{s,K}$ and

$$\tilde{b}_{s,1} \leq -A \cdot \sqrt{d} \quad \text{and} \quad \tilde{b}_{s,K} \geq A \cdot \sqrt{d} - \frac{4 \cdot \sqrt{d} \cdot A}{K-1},$$

$$\frac{\sqrt{d} \cdot A}{(n+1) \cdot (K-1)} \leq |\tilde{b}_{s,k+1} - \tilde{b}_{s,k}| \leq \frac{4 \cdot \sqrt{d} \cdot A}{K-1} \quad (k = 1, \ldots, K-1)$$

and

$$\min_{i=1,\ldots,n, k=1,\ldots,K} \left| \bar{\mathbf{c}}_s^T X_i - \tilde{b}_{s,k} \right| \geq \frac{\sqrt{d} \cdot A}{(n+1) \cdot (K-1)}.$$

Such a choice is always possible. For example, a viable way to do this is to set $\tilde{b}_{s,1} = -\sqrt{d} \cdot A - 2 \cdot \sqrt{d} \cdot A/((n+1) \cdot (K-1))$. Then, regarding $\tilde{b}_{s,k}$ ($k = 2, \ldots, K$), subdivide the interval

$$\left[-\sqrt{d} \cdot A + (k-2) \cdot \frac{2 \cdot \sqrt{d} \cdot A}{K-1}, -\sqrt{d} \cdot A + (k-1) \cdot \frac{2 \cdot \sqrt{d} \cdot A}{K-1} \right]$$

into $(n+1)$ subintervals of equal length $2 \cdot \sqrt{d} \cdot A/((K-1) \cdot (n+1))$ and choose $\tilde{b}_{s,k}$ as the midpoint of one of those intervals which does not contain any of the n values $\bar{\mathbf{c}}_s^T X_i$. Such an interval must always exist due to the impossibility of $n+1$ disjoint intervals containing at least one of the above n points each.

3. Define our initial inner weights

$$b_{(s-1)\cdot K+1,0}, \ldots, b_{(s-1)\cdot K+1,d}, \ldots, b_{s\cdot K,0}, \ldots, b_{s\cdot K,d}$$

for $s \in \{1, \dots, r\}$ such that we have

$$\sum_{j=1}^{d} b_{(s-1)\cdot K+k,j} \cdot x^{(j)} + b_{(s-1)\cdot K+k,0} = \rho_n \cdot (\bar{\mathbf{c}}_s^T x - \tilde{b}_{s,k}) \quad \text{for all } x \in \mathbb{R}^d,$$

for some $\rho_n > 0$. In other words, set

$$b_{(s-1)\cdot K+k,j} = \rho_n \cdot \bar{\mathbf{c}}_s^{(j)} \quad \text{and} \quad b_{(s-1)\cdot K+k,0} = -\rho_n \cdot \tilde{b}_{s,k}$$

($s = 1, \dots, r$, $k = 1, \dots, K$, $j = 1, \dots, d$) for some $\rho_n > 0$. A choice of ρ_n is given in Theorem 3.2.1.

4. Define the initial output weights by

$$a_l = 0 \quad \text{for all} \quad l \in \{0, \dots, K \cdot r\}.$$

5. Apply gradient descent for $t = 0, 1, \dots, t_n - 1$: Learn the weights by gradient descent. More precisely, we minimize the penalized empirical L_2 risk

$$F(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} |f_{net,(\mathbf{a},\mathbf{b})}(X_i) - Y_i|^2 + \frac{c_1}{n} \cdot \sum_{k=0}^{K \cdot r} a_k^2, \tag{3.2}$$

where $c_1 > 0$ is a constant by defining $(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)})$ recursively by

$$\begin{pmatrix} \mathbf{a}^{(t+1)} \\ \mathbf{b}^{(t+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{a}^{(t)} \\ \mathbf{b}^{(t)} \end{pmatrix} - \lambda_n \cdot (\nabla_{(\mathbf{a},\mathbf{b})} F)(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) \tag{3.3}$$

for some $\lambda_n > 0$ and $t = 0, 1, \dots, t_n - 1$. Here, $\lambda_n > 0$ denotes the step size and $t_n \in \mathbb{N}$ denotes the number of steps we perform in the gradient descent algorithm. A choice of λ_n and of t_n is given in Theorem 3.2.1. Note, that minimization of (3.2) with respect to $(\mathbf{a}, \mathbf{b})$ is a nonlinear least squares problem.

Repeat steps $1. - 5.$ I_n times. A choice of I_n is given in Theorem 3.2.1.

6. Choose the directions and the corresponding network which achieves the smallest penalized empirical L_2 error (3.2) among all the I_n networks as our neural network estimate $\tilde{m}_n$.

7. Set

$$\tilde{m}_n(\cdot) = f_{net,\mathbf{w}(t_n)}(\cdot) \tag{3.4}$$

and choose our estimate to be the by

$$\beta_n = c_3 \cdot \log n$$

truncated version

$$m_n(x) = T_{\beta_n} \tilde{m}_n(x). \tag{3.5}$$

For our neural network estimate we have the following up to a logarithmic factor optimal univariate rate of convergence.

Theorem 3.2.1. *Let $n \geq 2$, let $A \geq 1$ and let (X, Y), (X_1, Y_1), ..., (X_n, Y_n) be independent and identically distributed random variables with values in $[-A, A]^d \times \mathbb{R}$. Set $m(x) = \mathbf{E}\{Y|X = x\}$ and assume that (X, Y) satisfies*

$$\mathbf{E}\left(e^{c_2 \cdot |Y|^2}\right) < \infty \tag{3.6}$$

for some constant $c_2 > 0$, and that m satisfies

$$m(x) = \sum_{s=1}^{r} g_s(\mathbf{c}_s^T x) \quad (x \in \mathbb{R}^d)$$

for some $r \in \mathbb{N}$, $\mathbf{c}_s \in [-1, 1]^d$, where $\|\mathbf{c}_s\| = 1$, and $g_s : \mathbb{R} \to \mathbb{R}$ $(s = 1, \ldots, r)$. Assume that g_s is (p, C)-smooth for $s \in \{1, \ldots, r\}$, where $p \in (0, 1]$ and $C > 0$ are fixed. Define the regression estimate m_n as described above with

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

with parameter r as in the above projection pursuit model, and with the other parameters chosen by

$$K = K_n = \left\lceil \left(\frac{n}{(\log n)^3} \right)^{\frac{1}{2p+1}} \right\rceil,$$

$$\rho_n = n^2 \cdot K_n, \quad \lambda_n = \frac{1}{3 \cdot K_n \cdot r}, \quad t_n = K_n \cdot n \cdot (\log n)^2,$$

and

$$I_n = \left\lceil \left(\frac{n}{(\log n)^3} \right)^{\frac{r \cdot (d-1)}{2p+1}} \right\rceil.$$

Then m_n satisfies

$$\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_4 \cdot \left(\frac{(\log n)^3}{n}\right)^{\frac{2p}{2p+1}}$$

for some constant $c_4 > 0$ which does not depend on n.

Next, as in Result 2, we analyze a neural network estimate for d-dimensional regression functions satisfying a projection pursuit model. The unknown projection directions are chosen randomly and hence repeated initialization of the weights is needed. Our third estimate is different from the previously presented neural network estimates in several ways. This neural network regression estimate consists of many hidden layers where the structure and the value of the inner weights are prescribed by a new approximation result for a projection pursuit model by piecewise polynomials. The outer weights are determined by solving a linear equation system. This means there is no freedom in the choice of the weights but we trade it for analysis of high dimensional regression functions with $p > 0$ and analysis of multilayer neural networks. We show that our neural network estimate achieves the up to a logarithmic factor optimal univariate rate of convergence. Since this rate of convergence is independent of the dimension of the data, our third estimate can circumvent the curse of dimensionality.

Result 3. As in Result 2, we deal with neural network regression in a projection pursuit model. This means, we assume that the regression function satisfies

$$m(x) = \sum_{l=1}^{r} g_l\left(a_{(l-1)\cdot d+1} \cdot x^{(1)} + \cdots + a_{l\cdot d} \cdot x^{(d)}\right) \quad (x^{(1)}, \ldots, x^{(d)} \in \mathbb{R})$$

for some $r \in \mathbb{N}$, $\mathbf{a}_l = (a_{(l-1)\cdot d+1}, \ldots, a_{l\cdot d})^T \in \mathbb{R}^d$ where $\|\mathbf{a}_l\| = 1$ $(l = 1, \ldots, r)$, and (p, C)-smooth functions $g_s : \mathbb{R} \to \mathbb{R}$ $(s = 1, \ldots, r)$. Let $A \geq 1$, $M \in \mathbb{N}$, set

$$u_i = -\sqrt{d} \cdot A + i \cdot \frac{2 \cdot \sqrt{d} \cdot A}{M} \quad (i = 0, \ldots, M)$$

and set $\{i_1, \ldots, i_{M+1}\} = \{0, \ldots, M\}$. Denote by $g_l^{(j)}$ the j–th derivative of g_l $(l = 1, \ldots, r)$. We show that we can approximate m by a convex combination of Taylor polynomials of total degree q of the form

$$\sum_{l=1}^{r}\sum_{k=1}^{M+1}\left(\sum_{j=0}^{q}\frac{p_{l,j,i_k}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j\right) \cdot \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+$$

where p_{l,j,i_k} is the Taylor polynomial of $g_l^{(j)}$ of degree $q-j$ around u_{i_k}. Hence, the approximation function is contained in a class of functions defined by

$$\left\{ \sum_{l=1}^{r} \sum_{k=1}^{M+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,q\}, \\ j_1+\dots+j_d \le q}} a_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \cdot (x^{(1)})^{j_1} \cdot \dots \cdot (x^{(d)})^{j_d} \cdot \left(1 - \frac{M}{2\cdot\sqrt{d}\cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+ : a_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \in \mathbb{R} \right\}.$$

The key notion for the construction is to use smaller neural networks as building blocks to define a neural network $f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l}$ which approximates the functions

$$x \mapsto (x^{(1)})^{j_1} \cdots (x^{(d)})^{j_d} \cdot \left(1 - \frac{M}{2\cdot\sqrt{d}\cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+$$

and to choose the network architecture such that neural networks of the form

$$\sum_{l=1}^{r} \sum_{k=1}^{M+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,q\} \\ j_1+\dots+j_d \le q}} a_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \cdot f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l}(x) \quad (a_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \in \mathbb{R})$$

are contained in it. In order to do this, we introduce the following four neural networks. Choose $R \ge 1$. A choice of R is given in Theorem 4.2.1.

First, we approximate the function

$$f(x) = x$$

by the neural network

$$f_{id}(x) = 4R \cdot \sigma\left(\frac{x}{R}\right) - 2R. \tag{4.2}$$

Second, we approximate the function

$$f(x,y) = x \cdot y$$

by the neural network

$$f_{mult}(x,y) = \frac{R^2}{4} \cdot \frac{(1+e^{-1})^3}{e^{-2}-e^{-1}} \cdot \left(\sigma\left(\frac{2(x+y)}{R}+1\right) - 2\cdot\sigma\left(\frac{x+y}{R}+1\right)\right.$$

$$- \sigma \left(\frac{2(x-y)}{R} + 1 \right) + 2 \cdot \sigma \left(\frac{x-y}{R} + 1 \right) \Bigg). \tag{4.3}$$

Third, we approximate the function

$$f(x) = x_+$$

by the neural network

$$f_{ReLu}(x) = f_{mult}(f_{id}(x), \sigma(R \cdot x)). \tag{4.4}$$

Fourth, we approximate for fixed $y \in \mathbb{R}$ the function

$$f(x) = (1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |x-y|)_+$$

by the neural network

$$\begin{aligned} f_{hat,y}(x) = f_{ReLU} \left(\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (x-y) + 1 \right) - 2 \cdot f_{ReLU} \left(\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (x-y) \right) \\ + f_{ReLU} \left(\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (x-y) - 1 \right). \end{aligned} \tag{4.5}$$

The construction procedure for our neural network estimate has four steps:

1. Randomly choose values

   $$\mathbf{b}_1^*, \ldots, \mathbf{b}_r^* \in [-1,1]^d$$

   as an independent sample from a uniform distribution on $[-1,1]^d$ and set

   $$\mathbf{b}_l = \frac{\mathbf{b}_l^*}{\|\mathbf{b}_l^*\|} \quad (l = 1, \ldots, r).$$

   These values approximate the direction of projection $\mathbf{b}_1, \ldots, \mathbf{b}_r$ in our projection pursuit model.

   Note, that for $l = 1, \ldots, r$

   $$\mathbf{P}\{\mathbf{b}_l^* = 0\} = 0,$$

   which means we can assume w.l.o.g. $\|\mathbf{b}_l^*\| \neq 0$.

2. We recursively define the neural network $f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l}$ which approximates

$$x \mapsto (x^{(1)})^{j_1} \cdots (x^{(d)})^{j_d} \cdot \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+$$

with the building block networks. We choose $N \geq q$, set $s = \lceil \log_2(N+1) \rceil$ and define for $l \in \{1, \dots, r\}$, $j_1, \dots, j_d \in \{0, 1, \dots, N\}$ and $k \in \{1, \dots, M+1\}$

$$f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l}(x) = f_1^{(0)}(x), \tag{4.7}$$

where

$$f_k^{(t)}(x) = f_{mult}\left(f_{2k-1}^{(t+1)}(x), f_{2k}^{(t+1)}(x)\right) \tag{4.8}$$

for $k \in \{1, 2, \dots, 2^t\}$ and $t \in \{0, \dots, s-1\}$, and

$$f_k^{(s)}(x) = f_{id}(f_{id}(x^{(t)})) \tag{4.9}$$

for $j_1 + j_2 + \cdots + j_{t-1} + 1 \leq k \leq j_1 + j_2 + \cdots + j_t$ and $t = 1, \dots, d$,

$$f_{j_1+j_2+\cdots+j_d+1}^{(s)}(x) = f_{hat,u_{i_k}}(\mathbf{b}_l^T x), \tag{4.10}$$

and

$$f_k^{(s)}(x) = 1 \tag{4.11}$$

for $k = j_1 + j_2 + \cdots + j_d + 2, j_1 + j_2 + \cdots + j_d + 3, \dots, 2^s$. An example of the network is visualized in Figure 1.4.

3. We choose the output weights. The coefficients $a_{i_k,j_1,\dots,j_d,\mathbf{b}_l}$ are chosen by minimizing

$$\frac{1}{n}\sum_{i=1}^{n} |Y_i - \tilde{m}_n(X_i)|^2 + \frac{c_3}{n} \cdot \sum_{l=1}^{r} \sum_{k=1}^{M+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\cdots+j_d \leq N}} a_{i_k,j_1,\dots,j_d,\mathbf{b}_l}^2 \tag{4.16}$$

for some constant $c_3 > 0$. This regularized linear least squares estimate can be computed by solving the linear equation system

$$\left(\frac{1}{n}\mathbf{B}^T\mathbf{B} + \frac{c_3}{n} \cdot \mathbf{1}\right)\mathbf{a} = \frac{1}{n}\mathbf{B}^T\mathbf{Y} \tag{4.17}$$

where

$$\mathbf{B} = (B_j(X_i))_{1 \leq i \leq n, 1 \leq j \leq J} \quad \text{and} \quad \mathbf{Y} = (Y_i)_{i=1,\dots,n}$$

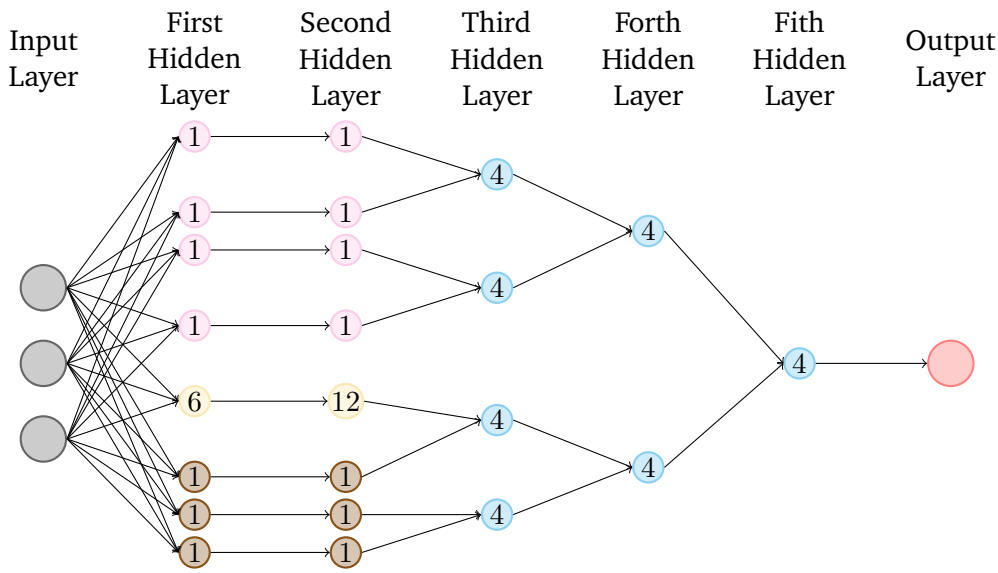

Figure 1.4.: Visualisation of an example of the neural network $f_{net,j_1,\ldots,j_d,i_k,\mathbf{b}_l}$ with parameters $d = 3$, $j_1 = 1, j_2 = 2, j_3 = 1$, $N = 5$ and $s = 3$. For readability, nodes that belong to the same building block network within a hidden layer are summarized into one node. The number inside the node indicates how many nodes it represents. An edge pointing to or leaving a node summarizes all the edges entering or leaving each node contained. The function has a 3-dimensional input (black node) and a 1-dimensional output (red node). There are five hidden layers. Nodes in the first and second hidden layer are produced by f_{id} (rose nodes), $f_{hat,x_{i_k}}(k = 1, \ldots, d)$ (yellow nodes) and constant function 1 (brown nodes). Nodes in the third, forth and fith layer are produced by f_{mult} (blue nodes).

with

$$J = r \cdot (M + 1) \cdot \binom{N + d}{d}$$

and

$$\begin{aligned}&\{B_j \,:\, j = 1, \ldots, J\}\\ &= \{f_{net,j_1,\ldots,j_d,i_k,\mathbf{b}_l}(x) \,:\, 1 \leq l \leq r,\, 1 \leq k \leq M + 1 \text{ and } 0 \leq j_1 + \cdots + j_d \leq N\}\,.\end{aligned}$$

Repeat steps $1. - 3.$ I_n times. A choice of I_n is given in Theorem $4.2.1$.

4. Choose the directions and the corresponding network which achieves the smallest penalized empirical L_2 error (4.16) among all the I_n networks as our neural network estimate $\tilde{m}_n$.

For our neural network estimate we have the following up to a logarithmic factor optimal univariate rate of convergence.

Theorem 4.2.1. *Assume that the distribution of (X, Y) satisfies*

$$\mathbf{E}\left(e^{c_4 \cdot |Y|^2}\right) < \infty \tag{4.19}$$

for some constant $c_4 > 0$ and that the distribution of X has bounded support $supp(X)$, and let $m(x) = \mathbf{E}\{Y|X = x\}$ be the corresponding regression function. Let $r \in \mathbb{N}$, $p > 0$ and $C > 0$, and assume that the regression function satisfies

$$m(x) = \sum_{l=1}^{r} g_l(\mathbf{a}_l^T x) \quad (x \in \mathbb{R}^d)$$

for some (p, C)–smooth functions $g_l : \mathbb{R} \to \mathbb{R}$ and some $\mathbf{a}_l \in \mathbb{R}^d$ with $\|\mathbf{a}_l\| = 1$ $(l = 1, \ldots, r)$.

Define the estimate $\tilde{m}_n$ as descriebed above, where σ is the logistic squasher

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

and the parameters are chosen as

$$I_n = \left\lceil \left(\frac{n}{(\log n)^2}\right)^{\frac{r\cdot(d-1)}{2p+1}} \right\rceil,$$

$$N \geq p, \quad M = M_n = \left\lceil c_{10} \cdot n^{\frac{1}{2p+1}} \right\rceil, \quad R = R_n = n^{\frac{9}{2}}$$

and

$$A = A_n = (\log n)^{\frac{1}{6(N+d)}}.$$

Set $\beta_n = c_6 \cdot \log(n)$ for some suitably large constant $c_6 > 0$ and define m_n by

$$m_n(x) = T_{\beta_n} \tilde{m}_n(x).$$

Then m_n satisfies for n suffciently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_{11} \cdot (\log n)^3 \cdot n^{-\frac{2p}{2p+1}},$$

where $c_{11} > 0$ does not depend on n.

Comparing our results we see that from our first result to our second result to our third result the choice of the weights becomes more and more restricted. However, at the same time we can consider more and more general regression functions m in the sense that we go from one-dimensional functions with $p \in [\frac{1}{2}, 1]$ in our first result to $d-$dimensional functions in our second and third result with $p \in (0, 1]$ and $p > 0$, respectively. Regarding the running time we see that in our first result we need one initialization and

$$t_n \leq c_{101} \cdot (\log n)^8 \cdot n^{\frac{5p+1}{2p+1}} \leq c_{101} \cdot n^{2+\epsilon}$$

many gradient descent steps for any $\epsilon > 0$ for a starting vector of size

$$3 \cdot K_n + 1 = 3 \cdot \left\lceil c_5 \cdot n^{\frac{1}{2p+1}} \right\rceil + 1,$$

while in our second result we need repeated initialization and so

$$I_n \cdot t_n \leq n^{2+r\cdot(d-1)}$$

many gradient descent steps for a starting vector of size

$$r \cdot (3 \cdot K_n + 1) = r \cdot \left(3 \cdot \left\lceil \left(\frac{n}{(\log n)^3} \right)^{\frac{1}{2p+1}} \right\rceil + 1 \right),$$

and in our third result we also need repeated initialization but we only have to solve a linear equation system with a quadratic $M_n \times M_n$ matrix which results in a computation time proportional to

$$I_n \cdot M_n^2 \approx n^{\frac{r\cdot(d-1)+2}{2p+1}}.$$

Clearly, our neural network estimate in result 3 is the most feasible which will also be reflected in our simulation.

The outline of this thesis is as follows: We present our first result in Chapter 2. We present our second result in Chapter 3 and we present our third result in Chapter 4. The sections can be read independently from each other.

These results are published in Braun, Kohler and Krzyżak (2019), Braun, Kohler and Walk (2019) and form the basis for Braun et al. (2021).

2. Neural Network Regression Estimates Learned by Gradient Descent Inspired by Approximation Results with Indicator Functions for Univariate Regression Functions

We assume that the regression function

$$m : \mathbb{R} \to \mathbb{R}$$

is $(p, C)-$smooth and univariate. We consider neural network regression estimates with one hidden layer where the weights are learned via gradient descent. The initial values for the inner weights $\beta_1^{(0)}, \ldots, \beta_{K_n}^{(0)}, \gamma_1^{(0)}, \ldots, \gamma_{K_n}^{(0)}$ are chosen independently of each other and of $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ from a uniform distribution on an interval $[-B_n, B_n]$ and the initial values for the outer weights $\alpha_0^{(0)}, \alpha_1^{(0)}, \ldots, \alpha_{K_n}^{(0)}$ are chosen from a smaller interval $[-\frac{c_2}{K_n}, \frac{c_2}{K_n}]$. We will see that the values in each step of the gradient descent algorithm do not leave some prespecified ball around the starting values of the parameters. We will see further that the choice of the value for B_n will guarantee with high probability that from the K_n inner weight vectors (β, γ) we can pick out $\tilde{K}_n$ many "good places" where our neural network estimate is close to a piecewise constant approximation of m. As a consequence, a single random initialization of the starting weights will be sufficient to find a "good" neural network estimate.

We define our neural network regression estimates in Section 2.1 and show the corresponding univariate rate of convergence result in Section 2.2. The finite sample size performance of our newly proposed estimate is illustrated in Section 2.3 by applying it to simulated data.

2.1. Constructing the Neural Network

Consider neural networks with one hidden layer and K_n hidden neurons defined by

$$\begin{aligned} f_{net,\mathbf{w}}(x) &= \sum_{j=1}^{K_n} w_{1,j}^{(1)} \cdot \sigma(w_{j,1}^{(0)} \cdot x + w_{j,0}^{(0)}) + w_{1,0}^{(1)} \\ &= \sum_{j=1}^{K_n} \alpha_j \cdot \sigma(\beta_j \cdot x + \gamma_j) + \alpha_0 \end{aligned} \tag{2.1}$$

where $K_n \in \mathbb{N}$, $\alpha_0, \alpha_i, \beta_i, \gamma_i \in \mathbb{R}$ $(i = 1, \ldots, K_n)$, the activation function is the logistic squasher

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

and

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

is the vector of weights. An example is shown in Figure 1.2.

The construction procedure for our neural network estimate has three steps:

1. Randomly choose our initial weights: Set

$$\mathbf{w}(0) = \mathbf{v} \tag{2.2}$$

   where the initial weight vector

$$\mathbf{v} = (\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}) = (\alpha_0^{(0)}, \alpha_1^{(0)}, \ldots, \alpha_{K_n}^{(0)}, \beta_1^{(0)}, \ldots, \beta_{K_n}^{(0)}, \gamma_1^{(0)}, \ldots, \gamma_{K_n}^{(0)})$$

   is chosen randomly such that

$$|\alpha_k^{(0)}| \leq \frac{c_2}{K_n} \quad (k = 1, \ldots, K_n)$$

   and $\beta_1^{(0)}, \ldots, \beta_{K_n}^{(0)}, \gamma_1^{(0)}, \ldots, \gamma_{K_n}^{(0)}$ are independently uniformly distributed on the set $[-B_n, B_n]$. A choice of B_n is given in Theorem 2.2.1.

2. Apply gradient descent for $t = 0, 1, \ldots, t_n - 1$: Learn the weights by gradient descent. More precisely, we minimize the regularized least squares criterion

$$F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{net,\mathbf{w}}(X_i)|^2 + \frac{c_1}{K_n^{2p}} \cdot \sum_{k=0}^{K_n} \alpha_k^2, \tag{2.3}$$

where $c_1 > 0$ is an arbitrary constant by defining $\mathbf{w}(t+1)$ recursively by

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \lambda_n \cdot \nabla_{\mathbf{w}} F(\mathbf{w}(t)) \tag{2.4}$$

for $t = 0, 1, \ldots, t_n - 1$. We write

$$\mathbf{w}(t) = (\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) = (\alpha_0^{(t)}, \alpha_1^{(t)}, \ldots, \alpha_{K_n}^{(t)}, \beta_1^{(t)}, \ldots, \beta_{K_n}^{(t)}, \gamma_1^{(t)}, \ldots, \gamma_{K_n}^{(t)})$$

for $t = 0, 1, \ldots, t_n$. Here, $\lambda_n > 0$ denotes the stepsize and $t_n \in \mathbb{N}$ denotes the number of steps we perform in the gradient descent algorithm. A choice of λ_n and of t_n is given in Theorem 2.2.1. Note, that minimization of (2.3) with respect to $\mathbf{w}$ is a nonlinear least squares problem.

3. Set

$$\tilde{m}_n(\cdot) = f_{net,\mathbf{w}(t_n)}(\cdot) \tag{2.5}$$

and choose our estimate to be the by

$$\beta_n = c_3 \cdot \log n$$

truncated version

$$m_n(x) = T_{\beta_n} \tilde{m}_n(x). \tag{2.6}$$

2.2. Rate of Convergence

For our neural network estimate we have the following up to a logarithmic factor optimal minimax rate of convergence.

Theorem 2.2.1. *Let (X, Y) be an $[0,1] \times \mathbb{R}$–valued random vector such that*

$$\mathbf{E}\left\{e^{c_{10} \cdot Y^2}\right\} < \infty \tag{2.7}$$

holds for some constant $c_{10} > 0$ and assume that the corresponding regression function $m(x) = \mathbf{E}\{Y|X = x\}$ is (p, C)–smooth for some $p \in [\frac{1}{2}, 1]$ and $C > 0$. Let (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n) be independent and identically distributed. Set

$$K_n = \left\lceil c_5 \cdot n^{\frac{1}{2p+1}} \right\rceil, \quad B_n = 16 \cdot K_n \cdot (\log K_n)^2 \cdot n,$$

$$L_n = c_6 \cdot (\log n)^6 \cdot K_n^{3p+1}, \quad \lambda_n = \frac{1}{L_n}$$

and

$$t_n = \lceil K_n^{2p} \cdot (\log n)^2 \cdot L_n \rceil$$

and define the estimate m_n of m as in Section 2.1. Then we have for n sufficiently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_7 \cdot (\log n)^{\max\{3,4p\}} \cdot n^{-2p/(2p+1)}$$

for some constant $c_7 > 0$ which does not depend on n.

Remark 2.2.2. *The computation of the estimate in Theorem 2.2.1 requires one initialization of the starting weights and*

$$t_n \leq c_{101} \cdot (\log n)^8 \cdot n^{\frac{5p+1}{2p+1}} \leq c_{101} \cdot n^{2+\epsilon}$$

gradient descent steps for any $\epsilon > 0$.

Remark 2.2.3. *The parameter K_n of the above algorithm is chosen data-dependently using the splitting of the sample technique as explained in Section 2.3.*

Remark 2.2.4. *It is possible to broaden the class of functions from which we choose our neural network estimate while maintaining the optimal rate of convergence up to a logratihmic factor by choosing all of the inital weights from the same broader interval $[-B_n, B_n]$ instead of choosing the outer weights from a smaller interval $[-\frac{c_2}{K_n}, \frac{c_2}{K_n}]$. The reason why this is possible in our context, is that Lemma 2.2.11 only the requires the inner weights to be containted in the interval $[-B_n, B_n]$. However, we pay for this broadness by an increase of computation time as the number of repititions of gradient descent steps increases significantly and is rather huge.*

More precisely, it is possible to choose the starting values for the gradient descent algorithm randomly such that $\alpha_1^{(0)}, \ldots, \alpha_{K_n}^{(0)}, \beta_1^{(0)}, \ldots, \beta_{K_n}^{(0)}, \gamma_1^{(0)}, \ldots, \gamma_{K_n}^{(0)}$ are independently uniformly distributed on the set $[-B_n, B_n]$. The function to be minimized is the regularized least squares criterion

$$F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{net,\mathbf{w}}(X_i)|^2 + \frac{c_1}{n} \cdot \sum_{k=1}^{K_n} \alpha_k^2$$

where $c_1 > 0$ is an arbitrary constant. In this setting, we can show in a completely analogous fashion, that under the same assumptions as in Theorem 2.2.1 and with the same parameter choice except for $L_n = n^3 \cdot (\log n)^6 \cdot K_n^{10} \cdot B_n^7$ and $t_n = \lceil n \cdot (\log n)^2 \cdot L_n \rceil$ we have for n sufficiently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_7' \cdot (\log n)^3 \cdot n^{-2p/(2p+1)}.$$

In this setting the computation of the estimate requires one initialization of the starting weights and

$$t_n \leq (\log n)^8 \cdot n^{13}$$

many gradient descent steps.

Since that huge computation time is not feasible, we focus on the smaller interval for the choice of the smaller interval for our inital output weights.

2.2.1. A Localization Lemma for Gradient Descent

We start with our key observation on gradient descent. Lemma 2.2.6 shows that, under suitable assumptions on the regularized least squares criterion function F and on the number of gradient descent steps t_n, the values in each step of the gradient descent algorithm do not leave some prespecified ball around the starting values of the parameters.

Lemma 2.2.5. *Let $F : \mathbb{R}^K \to \mathbb{R}$ be a differentiable function . Let $t \in \mathbb{N}$, $L > 0$, $\mathbf{a} \in \mathbb{R}^K$ and set*

$$\lambda = \frac{1}{L} \tag{2.8}$$

and

$$\bar{\mathbf{a}} = \mathbf{a} - \lambda \cdot (\nabla_{\mathbf{a}} F)(\mathbf{a}). \tag{2.9}$$

If

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}) - (\nabla_{\mathbf{a}} F)(\mathbf{b})\| \leq L \cdot \|\mathbf{a} - \mathbf{b}\| \tag{2.10}$$

holds for all

$$\mathbf{b} \in \{\mathbf{a} + s \cdot (\bar{\mathbf{a}} - \mathbf{a}) : s \in [0, 1]\}$$

then we have

$$F(\bar{\mathbf{a}}) - F(\mathbf{a}) \leq -\frac{1}{2 \cdot L} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2.$$

Proof. Lemma 2.2.5 follows from well-known bounds in the literature, see, e.g., Karimi, Nutini and Schmidt (2018). For the sake of completeness a complete proof is given here.

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

$$H(s) = F(\mathbf{a} + s \cdot (\bar{\mathbf{a}} - \mathbf{a})).$$

Then the fundamental theorem of calculus, the chain rule and assumption (2.10) imply

$$\begin{aligned} & F(\bar{\mathbf{a}}) - F(\mathbf{a}) \\ = \; & H(1) - H(0) \end{aligned}$$

$$
\begin{aligned}
&= \int_0^1 H'(s)\,ds \\
&= \int_0^1 (\nabla_{\mathbf{a}} F)(\mathbf{a} + s\cdot(\bar{\mathbf{a}} - \mathbf{a}))\cdot(\bar{\mathbf{a}} - \mathbf{a})\,ds \\
&= \int_0^1 ((\nabla_{\mathbf{a}} F)(\mathbf{a} + s\cdot(\bar{\mathbf{a}} - \mathbf{a})) - (\nabla_{\mathbf{a}} F)(\mathbf{a}))\cdot(\bar{\mathbf{a}} - \mathbf{a})\,ds \\
&\quad + \int_0^1 (\nabla_{\mathbf{a}} F)(\mathbf{a})\cdot(\bar{\mathbf{a}} - \mathbf{a})\,ds \\
&\le \int_0^1 L\cdot\|s\cdot(\bar{\mathbf{a}} - \mathbf{a})\|\cdot\|\bar{\mathbf{a}} - \mathbf{a}\|\,ds \\
&\quad + (\nabla_{\mathbf{a}} F)(\mathbf{a})\cdot(\bar{\mathbf{a}} - \mathbf{a}) \\
&= \frac{L}{2}\cdot\|\bar{\mathbf{a}} - \mathbf{a}\|^2 + (\nabla_{\mathbf{a}} F)(\mathbf{a})\cdot(\bar{\mathbf{a}} - \mathbf{a}).
\end{aligned}
$$

Using (2.9) and (2.8) we get

$$
\begin{aligned}
F(\bar{\mathbf{a}}) - F(\mathbf{a}) &\le \frac{L}{2}\cdot\lambda^2\cdot\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2 - \lambda\cdot\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2 \\
&= -\frac{1}{2\cdot L}\cdot\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2.
\end{aligned}
$$

□

Lemma 2.2.6. *Let $F : \mathbb{R}^K \to \mathbb{R}_+$ be a nonnegative differentiable function. Let $t \in \mathbb{N}$, $L > 0$, $\mathbf{a}_0 \in \mathbb{R}^K$ and set*

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

and

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

Assume

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\| \le \sqrt{2\cdot t\cdot L\cdot\max\{F(\mathbf{a}_0), 1\}} \tag{2.11}$$

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

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}) - (\nabla_{\mathbf{a}} F)(\mathbf{b})\| \le L\cdot\|\mathbf{a} - \mathbf{b}\| \tag{2.12}$$

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

$$\|\mathbf{a} - \mathbf{a}_0\| \le \sqrt{8\cdot\frac{t}{L}\cdot\max\{F(\mathbf{a}_0), 1\}} \quad \textit{and} \quad \|\mathbf{b} - \mathbf{a}_0\| \le \sqrt{8\cdot\frac{t}{L}\cdot\max\{F(\mathbf{a}_0), 1\}}. \tag{2.13}$$

Then we have

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

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

and

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

Proof. We show

$$\sum_{k=0}^{s-1} \|\mathbf{a}_{k+1} - \mathbf{a}_k\|^2 \leq \frac{2}{L} \cdot (F(\mathbf{a}_0) - F(\mathbf{a}_s)) \tag{2.14}$$

and

$$F(\mathbf{a}_k) \leq F(\mathbf{a}_{k-1}) \quad \text{for all } k \in \{1, \ldots, s\} \tag{2.15}$$

for all $s \in \{0, \ldots, t\}$ via induction on s.
Start of the induction: Trivially, (2.14) and (2.15) hold for $s = 0$.
Induction hypothesis: Assume now that

$$\sum_{k=0}^{s-1} \|\mathbf{a}_{k+1} - \mathbf{a}_k\|^2 \leq \frac{2}{L} \cdot (F(\mathbf{a}_0) - F(\mathbf{a}_s))$$

and

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

hold for some $s \in \{0, 1, \ldots, t-1\}$.
Induction step: We want to use (2.12) for $\mathbf{a} = \mathbf{a}_k$ and $\mathbf{b} \in \{\mathbf{a}_k + \gamma \cdot (\mathbf{a}_{k+1} - \mathbf{a}_k) : \gamma \in [0, 1]\}$ For that, we check (2.13) for those values. We verify (2.13) by (2.11). So, by the triangle inequality and the Cauchy-Schwarz inequality we can conclude from the induction hypothesis that for $k \in \{0, 1, \ldots, s\}$

$$\begin{aligned} \|\mathbf{a}_k - \mathbf{a}_0\| &\leq \sum_{j=0}^{k-1} \|\mathbf{a}_{j+1} - \mathbf{a}_j\| \\ &= \sqrt{\left(\sum_{j=0}^{k-1} 1 \cdot \|\mathbf{a}_{j+1} - \mathbf{a}_j\|\right)^2} \end{aligned}$$

$$
\begin{aligned}
&\leq \sqrt{\sum_{j=0}^{k-1} 1^2 \cdot \sum_{j=0}^{k-1} \|\mathbf{a}_{j+1} - \mathbf{a}_j\|^2} \\
&\leq \sqrt{s \cdot \sum_{j=0}^{k-1} \|\mathbf{a}_{j+1} - \mathbf{a}_j\|^2} \\
&\leq \sqrt{\frac{2 \cdot s}{L} \cdot (F(\mathbf{a}_0) - F(\mathbf{a}_k))} \\
&\leq \sqrt{\frac{2t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}}. \qquad (2.16)
\end{aligned}
$$

Hence, with (2.16) we conclude for $k \in \{0, 1, \ldots, s-1\}$ and for any $\gamma \in [0, 1]$

$$
\begin{aligned}
& \|\mathbf{a}_k + \gamma \cdot (\mathbf{a}_{k+1} - \mathbf{a}_k) - \mathbf{a}_0\| \\
= \; & \|\mathbf{a}_k - \gamma \cdot \mathbf{a}_k + \gamma \cdot \mathbf{a}_{k+1} - ((1-\gamma) + \gamma) \cdot \mathbf{a}_0\| \\
= \; & \|(1-\gamma) \cdot \mathbf{a}_k - (1-\gamma) \cdot \mathbf{a}_0 + \gamma \cdot \mathbf{a}_{k+1} - \gamma \cdot \mathbf{a}_0\| \\
\leq \; & (1-\gamma) \cdot \|\mathbf{a}_k - \mathbf{a}_0\| + \gamma \cdot \|\mathbf{a}_{k+1} - \mathbf{a}_0\| \\
\leq \; & (1-\gamma) \cdot \sqrt{\frac{2t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}} + \gamma \cdot \sqrt{\frac{2t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}} \\
= \; & \sqrt{\frac{2t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}}.
\end{aligned}
$$

Further, we conclude with (2.16) and with (2.11)

$$
\begin{aligned}
\|\mathbf{a}_{s+1} - \mathbf{a}_s\| &= \frac{1}{L} \cdot \|\nabla_{\mathbf{a}} F(\mathbf{a}_s)\| \\
&\leq \frac{1}{L} \cdot \sqrt{2 \cdot t \cdot L \cdot \max\{F(\mathbf{a}_0), 1\}} \\
&= \sqrt{\frac{2t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}}
\end{aligned}
$$

and consequently, for $k = s$ and for any $\gamma \in [0, 1]$

$$
\begin{aligned}
& \|\mathbf{a}_s + \gamma \cdot (\mathbf{a}_{s+1} - \mathbf{a}_s) - \mathbf{a}_0\| \\
\leq \; & \|\mathbf{a}_s - \mathbf{a}_0\| + \gamma \cdot \|\mathbf{a}_{s+1} - \mathbf{a}_s\| \\
\leq \; & \sqrt{\frac{2t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}} + \gamma \sqrt{\frac{2t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}}
\end{aligned}
$$

$$\leq \sqrt{\frac{8t}{L} \cdot \max\{F(\mathbf{a}_0), 1\}}.$$

Now, we can conlcude by (2.12) (with $\mathbf{a} = \mathbf{a}_k$ and $\mathbf{b} = \mathbf{a}_k + \gamma \cdot (\mathbf{a}_{k+1} - \mathbf{a}_k)$) for any $k \in \{0, 1, \ldots, s\}$ and for any $\gamma \in [0, 1]$ that

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

Next, we conclude by Lemma 2.2.5 that for $k \in \{0, 1, \ldots, s\}$

$$F(\mathbf{a}_{k+1}) - F(\mathbf{a}_k) \leq -\frac{1}{2 \cdot L} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}_k)\|^2.$$

This implies

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

and

$$\begin{aligned}
\sum_{k=0}^{s} \|\mathbf{a}_{k+1} - \mathbf{a}_k\|^2 &= \sum_{k=0}^{s} \lambda^2 \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}_k)\|^2 \\
&= \sum_{k=0}^{s} \frac{2}{L} \cdot \frac{1}{2 \cdot L} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}_k)\|^2 \\
&\leq \sum_{k=0}^{s} \frac{2}{L} \cdot (F(\mathbf{a}_k) - F(\mathbf{a}_{k+1})) \\
&= \frac{2}{L} \cdot (F(\mathbf{a}_0) - F(\mathbf{a}_{s+1})) .
\end{aligned}$$

This concludes the induction. Finally, by (2.14) we get

$$\begin{aligned}
\|\mathbf{a}_s - \mathbf{a}_0\| &\leq \sum_{k=0}^{s-1} \|\mathbf{a}_{k+1} - \mathbf{a}_k\| \\
&\leq \sqrt{s \cdot \sum_{k=0}^{s-1} \|\mathbf{a}_{k+1} - \mathbf{a}_k\|^2} \\
&\leq \sqrt{\frac{2 \cdot s}{L} \cdot (F(\mathbf{a}_0) - F(\mathbf{a}_s))},
\end{aligned}$$

which concludes the proof. □

2.2.2. Auxiliary Lemmas

We introduce Lemma 2.2.9 and 2.2.10 that will help us to verify the conditions in Lemma 2.2.6 in our setting. We will make use of them in the proof of Theorem 2.2.1. For the proof of these lemmas we need the auxiliary Lemma 2.2.7 and Lemma 2.2.8, which we present first.

The next lemma gives us a statement close to Lipschitz continuity with respect to the weights and $1-$norm for neural networks with one hidden layer and Lipschitz continuous activation function.

Lemma 2.2.7. *Let $\sigma : \mathbb{R}^d \to [0,1]$ be Lipschitz continuous with Lipschitz constant $C_{Lip} \geq 1$ and let $\mathbf{w}, \bar{\mathbf{w}} \in \mathbb{R}^{3K_n+1}$. Define*

$$f_{net,\mathbf{w}}(x) = \sum_{j=1}^{K_n} w_{1,j}^{(1)} \cdot \sigma(w_{j,1}^{(0)} \cdot x + w_{j,0}^{(0)}) + w_{1,0}^{(1)}$$

and

$$f_{net,\bar{\mathbf{w}}}(x) = \sum_{j=1}^{K_n} \bar{w}_{1,j}^{(1)} \cdot \sigma(\bar{w}_{j,1}^{(0)} \cdot x + \bar{w}_{j,0}^{(0)}) + \bar{w}_{1,0}^{(1)}.$$

Then we have for any $x \in \mathbb{R}$

$$\begin{aligned}
&|f_{net,\mathbf{w}}(x) - f_{net,\bar{\mathbf{w}}}(x)| \\
&\leq C_{Lip} \cdot \max\{|x|, 1\} \cdot \max_{k=1,\dots,K_n} \max\{|\bar{w}_{1,k}^{(1)}|, 1\} \\
&\qquad \cdot \left(\sum_{j=0}^{K_n} |w_{1,j}^{(1)} - \bar{w}_{1,j}^{(1)}| + \sum_{k=1}^{K_n} \left(|w_{k,0}^{(0)} - \bar{w}_{k,0}^{(0)}| + |w_{k,1}^{(0)} - \bar{w}_{k,1}^{(0)}| \right) \right).
\end{aligned}$$

Proof. By the Lipschitz continuity of σ we have

$$\begin{aligned}
&|f_{net,\mathbf{w}}(x) - f_{net,\bar{\mathbf{w}}}(x)| \\
= &\left| \sum_{j=1}^{K_n} w_{1,j}^{(1)} \cdot \sigma(w_{j,1}^{(0)} \cdot x + w_{j,0}^{(0)}) + w_{1,0}^{(1)} - \sum_{j=1}^{K_n} \bar{w}_{1,j}^{(1)} \cdot \sigma(\bar{w}_{j,1}^{(0)} \cdot x + \bar{w}_{j,0}^{(0)}) - \bar{w}_{1,0}^{(1)} \right| \\
\leq &\left| \sum_{j=1}^{K_n} w_{1,j}^{(1)} \cdot \sigma(w_{j,1}^{(0)} \cdot x + w_{j,0}^{(0)}) - \sum_{j=1}^{K_n} \bar{w}_{1,j}^{(1)} \cdot \sigma(\bar{w}_{j,1}^{(0)} \cdot x + \bar{w}_{j,0}^{(0)}) - \bar{w}_{1,0}^{(1)} \right| + \left| w_{1,0}^{(1)} - \bar{w}_{1,0}^{(1)} \right| \\
\leq &\left| \sum_{j=1}^{K_n} w_{1,j}^{(1)} \cdot \sigma(w_{j,1}^{(0)} \cdot x + w_{j,0}^{(0)}) - \sum_{j=1}^{K_n} \bar{w}_{1,j}^{(1)} \cdot \sigma(w_{j,1}^{(0)} \cdot x + w_{j,0}^{(0)}) \right|
\end{aligned}$$

$$
\begin{aligned}
&+\left|\sum_{j=1}^{K_n}\bar{w}_{1,j}^{(1)}\cdot\sigma(w_{j,1}^{(0)}\cdot x+w_{j,0}^{(0)})-\sum_{j=1}^{K_n}\bar{w}_{1,j}^{(1)}\cdot\sigma(\bar{w}_{j,1}^{(0)}\cdot x+\bar{w}_{j,0}^{(0)})\right| \\
&+\left|w_{1,0}^{(1)}-\bar{w}_{1,0}^{(1)}\right| \\
\leq\ & \sum_{j=1}^{K_n}\left|\left(w_{1,j}^{(1)}-\bar{w}_{1,j}^{(1)}\right)\cdot\sigma(w_{j,1}^{(0)}\cdot x+w_{j,0}^{(0)})\right| \\
&+\sum_{j=1}^{K_n}\left|\bar{w}_{1,j}^{(1)}\cdot\left(\sigma(w_{j,1}^{(0)}\cdot x+w_{j,0}^{(0)})-\sigma(\bar{w}_{j,1}^{(0)}\cdot x+\bar{w}_{j,0}^{(0)})\right)\right| \\
&+\left|w_{1,0}^{(1)}-\bar{w}_{1,0}^{(1)}\right| \\
\leq\ & \sum_{j=0}^{K_n}|w_{1,j}^{(1)}-\bar{w}_{1,j}^{(1)}|+\max_{k=1,\dots,K_n}\{|\bar{w}_{1,k}^{(1)}|,1\}\cdot\sum_{j=1}^{K_n}\left|\sigma(w_{j,1}^{(0)}\cdot x+w_{j,0}^{(0)})-\sigma(\bar{w}_{j,1}^{(0)}\cdot x+\bar{w}_{j,0}^{(0)})\right| \\
\leq\ & \sum_{j=0}^{K_n}|w_{1,j}^{(1)}-\bar{w}_{1,j}^{(1)}|+\max_{k=1,\dots,K_n}\{|\bar{w}_{1,k}^{(1)}|,1\}\cdot C_{Lip}\cdot\sum_{j=1}^{K_n}\left|w_{j,1}^{(0)}\cdot x+w_{j,0}^{(0)}-\bar{w}_{j,1}^{(0)}\cdot x-\bar{w}_{j,0}^{(0)}\right| \\
\leq\ & \sum_{j=0}^{K_n}|w_{1,j}^{(1)}-\bar{w}_{1,j}^{(1)}| \\
&+\max_{k=1,\dots,K_n}\{|\bar{w}_{1,k}^{(1)}|,1\}\cdot C_{Lip}\cdot\left(\sum_{j=1}^{K_n}\left|w_{j,1}^{(0)}-\bar{w}_{j,1}^{(0)}\right|\cdot|x|+\left|w_{j,0}^{(0)}-\bar{w}_{j,0}^{(0)}\right|\right) \\
\leq\ & C_{Lip}\cdot\max\{|x|,1\}\cdot\max_{k=1,\dots,K_n}\max\{|\bar{w}_{1,k}^{(1)}|,1\} \\
&\cdot\left(\sum_{j=0}^{K_n}|w_{1,j}^{(1)}-\bar{w}_{1,j}^{(1)}|+\sum_{k=1}^{K_n}\left(|w_{k,0}^{(0)}-\bar{w}_{k,0}^{(0)}|+|w_{k,1}^{(0)}-\bar{w}_{k,1}^{(0)}|\right)\right).
\end{aligned}
$$

□

Our next lemma gives us a statement close to Lipschitz continuity with respect to the weights for the gradient of our neural networks.

Lemma 2.2.8. *For a weight vector* $\mathbf{w}$ *define the network* $f_{net,\mathbf{w}}$ *by* (2.1). *Let*

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

be the logistic squasher. Let $\gamma_n \geq 1$ and assume

$$|\bar{w}_{1,j}^{(1)}| \leq \gamma_n$$

for $j \in \{1, \ldots, K_n\}$. Then we have for any $x \in [0,1]$:

$$\sum_{k,j,l} \left| \frac{\partial}{\partial w_{k,j}^{(l)}} f_{net,\mathbf{w}}(x) - \frac{\partial}{\partial w_{k,j}^{(l)}} f_{net,\bar{\mathbf{w}}}(x) \right|^2 \leq 34 \cdot \gamma_n^2 \cdot \|\mathbf{w} - \bar{\mathbf{w}}\|^2.$$

Proof. By definition we have for a weight vector $\mathbf{w}$

$$f_{net,\mathbf{w}}(x) = \sum_{k=1}^{K_n} w_{1,k}^{(1)} \cdot \sigma(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}) + w_{1,0}^{(1)},$$

where the logistic squasher σ satisfies

$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x)). \tag{2.17}$$

Obviously, this implies $|\sigma'(x)| \leq 1$ and consequently we know that σ is Lipschitz continuous with Lipschitz constant 1. Hence, we get the partial derivatives for the outer weights

$$\frac{\partial}{\partial w_{1,k}^{(1)}} f_{net,\mathbf{w}}(x) = \sigma(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)})$$

for $k = 1, \ldots, K_n$ and

$$\frac{\partial}{\partial w_{1,0}^{(1)}} f_{net,\mathbf{w}}(x) = 1$$

for $k = 0$. By the chain rule and by (2.17) we get the partial derivatives for the inner weights

$$\begin{aligned}
\frac{\partial}{\partial w_{k,1}^{(0)}} f_{net,\mathbf{w}}(x) &= w_{1,k}^{(1)} \cdot \frac{\partial}{\partial w_{k,1}^{(0)}} \sigma(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}) \\
&= w_{1,k}^{(1)} \cdot \sigma' \left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)} \right) \cdot x \\
&= w_{1,k}^{(1)} \cdot \sigma \left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)} \right) \cdot \left(1 - \sigma \left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)} \right) \right) \cdot x
\end{aligned}$$

and

$$\frac{\partial}{\partial w_{k,0}^{(0)}} f_{net,\mathbf{w}}(x) = w_{1,k}^{(1)} \cdot \frac{\partial}{\partial w_{k,0}^{(0)}} \sigma(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)})$$

$$
\begin{aligned}
&= w_{1,k}^{(1)} \cdot \sigma'\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) \cdot 1 \\
&= w_{1,k}^{(1)} \cdot \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right)\right) \cdot 1
\end{aligned}
$$

for $k = 1, \ldots, K_n$. Now, we look at

$$
\begin{aligned}
&\sum_{k,j,l} \left| \frac{\partial}{\partial w_{k,j}^{(l)}} f_{net,\mathbf{w}}(x) - \frac{\partial}{\partial w_{k,j}^{(l)}} f_{net,\bar{\mathbf{w}}}(x) \right|^2 \\
&= \sum_{k=0}^{K_n} \left| \frac{\partial}{\partial w_{1,k}^{(1)}} f_{net,\mathbf{w}}(x) - \frac{\partial}{\partial w_{1,k}^{(1)}} f_{net,\bar{\mathbf{w}}}(x) \right|^2 \\
&\quad + \sum_{k=1}^{K_n} \sum_{j=0}^{1} \left| \frac{\partial}{\partial w_{k,j}^{(0)}} f_{net,\mathbf{w}}(x) - \frac{\partial}{\partial w_{k,j}^{(0)}} f_{net,\bar{\mathbf{w}}}(x) \right|^2 .
\end{aligned}
$$

We bound the two terms on the right-hand side separately and we start with the first one. Since $x \in [0, 1]$ we get

$$
\begin{aligned}
&\sum_{k=0}^{K_n} \left| \frac{\partial}{\partial w_{1,k}^{(1)}} f_{net,\mathbf{w}}(x) - \frac{\partial}{\partial w_{1,k}^{(1)}} f_{net,\bar{\mathbf{w}}}(x) \right|^2 \\
= &\sum_{k=1}^{K_n} \left| \frac{\partial}{\partial w_{1,k}^{(1)}} f_{net,\mathbf{w}}(x) - \frac{\partial}{\partial w_{1,k}^{(1)}} f_{net,\bar{\mathbf{w}}}(x) \right|^2 + \left| \frac{\partial}{\partial w_{1,0}^{(1)}} f_{net,\mathbf{w}}(x) - \frac{\partial}{\partial w_{1,0}^{(1)}} f_{net,\bar{\mathbf{w}}}(x) \right|^2 \\
= &\sum_{k=1}^{K_n} \left| \sigma(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}) - \sigma(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}) \right|^2 + |1 - 1|^2 \\
\leq &\sum_{k=1}^{K_n} 1 \cdot \left| w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)} - \bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)} \right|^2 \\
\leq &\sum_{k=1}^{K_n} \left(\left| w_{k,1}^{(0)} - \bar{w}_{k,1}^{(0)} \right| \cdot |x| + \left| w_{k,0}^{(0)} - \bar{w}_{k,0}^{(0)} \right| \right)^2 \\
\leq &\sum_{k=1}^{K_n} \left(\sum_{j=0}^{1} |w_{k,j}^{(0)} - \bar{w}_{k,j}^{(0)}| \right)^2
\end{aligned}
$$

$$\leq \quad 2 \cdot \sum_{k=1}^{K_n} \sum_{j=0}^{1} |w_{k,j}^{(0)} - \bar{w}_{k,j}^{(0)}|^2.$$

Next, we look at the second term on the right-hand side. Since $|\sigma(x)| \leq 1$ and thus also $|1 - \sigma(x)| \leq 1$ we can conclude for $x \in [0,1]$ in the same manner as in the proof of Lemma 2.2.7 that for $k = 1, \ldots, K_n$

$$\begin{aligned}
& \left| \frac{\partial}{\partial w_{k,1}^{(0)}} f_{net,\mathbf{w}}(x) - \frac{\partial}{\partial w_{k,1}^{(0)}} f_{net,\bar{\mathbf{w}}}(x) \right| \\
= \quad & |w_{1,k}^{(1)} \cdot \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right)\right) \cdot x \\
& \quad -\bar{w}_{1,k}^{(1)} \cdot \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right)\right) \cdot x| \\
\leq \quad & |w_{1,k}^{(1)} \cdot \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right)\right) \cdot x \\
& \qquad -\bar{w}_{1,k}^{(1)} \cdot \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right)\right) \cdot x| \\
& \quad +|\bar{w}_{1,k}^{(1)} \cdot \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right)\right) \cdot x \\
& \qquad -\bar{w}_{1,k}^{(1)} \cdot \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right)\right) \cdot x| \\
& \quad +|\bar{w}_{1,k}^{(1)} \cdot \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right)\right) \cdot x \\
& \qquad -\bar{w}_{1,k}^{(1)} \cdot \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right)\right) \cdot x| \\
= \quad & |w_{1,k}^{(1)} - \bar{w}_{1,k}^{(1)}| \cdot |\sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right)\right)| \cdot |x| \\
& \quad +|\bar{w}_{1,k}^{(1)} \cdot \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) - \bar{w}_{1,k}^{(1)} \cdot \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right)| \\
& \qquad \cdot|\left(1 - \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right)\right)| \cdot |x| \\
& \quad +|\bar{w}_{1,k}^{(1)}| \cdot |\sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right)| \\
& \qquad \cdot|\left(1 - \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) - \left(1 - \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right)\right)\right)| \cdot |x| \\
\leq \quad & |w_{1,k}^{(1)} - \bar{w}_{1,k}^{(1)}| \\
& \quad +|\bar{w}_{1,k}^{(1)}| \cdot |\sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) - \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right)| \\
& \quad +|\bar{w}_{1,k}^{(1)}| \cdot |-\sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) + \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right)|
\end{aligned}$$

$$
\begin{aligned}
&\leq \quad |w_{1,k}^{(1)} - \bar{w}_{1,k}^{(1)}| + 2 \cdot |\bar{w}_{1,k}^{(1)}| \cdot 1 \cdot \max\{|x|, 1\} \cdot \left(|w_{k,0}^{(0)} - \bar{w}_{k,0}^{(0)}| + |w_{k,1}^{(0)} - \bar{w}_{k,1}^{(0)}|\right) \\
&\leq \quad |w_{1,k}^{(1)} - \bar{w}_{1,k}^{(1)}| + 2 \cdot \gamma_n \cdot \sum_{l=0}^{1} |w_{k,l}^{(0)} - \bar{w}_{k,l}^{(0)}|
\end{aligned}
$$

and analogously

$$
\begin{aligned}
&\left| \frac{\partial}{\partial w_{k,0}^{(0)}} f_{net,\mathbf{w}}(x) - \frac{\partial}{\partial w_{k,0}^{(0)}} f_{net,\bar{\mathbf{w}}}(x) \right| \\
= \quad & |w_{1,k}^{(1)} \cdot \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(w_{k,1}^{(0)} \cdot x + w_{k,0}^{(0)}\right)\right) \\
& \quad -\bar{w}_{1,k}^{(1)} \cdot \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right) \cdot \left(1 - \sigma\left(\bar{w}_{k,1}^{(0)} \cdot x + \bar{w}_{k,0}^{(0)}\right)\right) | \\
\leq \quad & |w_{1,k}^{(1)} - \bar{w}_{1,k}^{(1)}| + 2 \cdot \gamma_n \cdot \sum_{l=0}^{1} |w_{k,l}^{(0)} - \bar{w}_{k,l}^{(0)}|.
\end{aligned}
$$

This implies

$$
\begin{aligned}
& \sum_{k=1}^{K_n} \sum_{j=0}^{1} \left| \frac{\partial}{\partial w_{k,j}^{(0)}} f_{net,\mathbf{w}}(x) - \frac{\partial}{\partial w_{k,j}^{(0)}} f_{net,\bar{\mathbf{w}}}(x) \right|^2 \\
\leq \quad & \sum_{k=1}^{K_n} \sum_{j=0}^{1} \left| |w_{1,k}^{(1)} - \bar{w}_{1,k}^{(1)}| + 2 \cdot \gamma_n \cdot \sum_{l=0}^{1} |w_{k,l}^{(0)} - \bar{w}_{k,l}^{(0)}| \right|^2 \\
\leq \quad & \sum_{k=1}^{K_n} 2 \cdot 2 \cdot \left(|w_{1,k}^{(1)} - \bar{w}_{1,k}^{(1)}|^2 + 4 \cdot \gamma_n^2 \cdot \left(\sum_{l=0}^{1} |w_{k,l}^{(0)} - \bar{w}_{k,l}^{(0)}| \right)^2 \right) \\
\leq \quad & 4 \cdot \sum_{k=1}^{K_n} \left(|w_{1,k}^{(1)} - \bar{w}_{1,k}^{(1)}|^2 + 4 \cdot \gamma_n^2 \cdot \left(2 \cdot \sum_{l=0}^{1} |w_{k,l}^{(0)} - \bar{w}_{k,l}^{(0)}|^2 \right) \right) \\
\leq \quad & 4 \cdot \sum_{k=1}^{K_n} |w_{1,k}^{(1)} - \bar{w}_{1,k}^{(1)}|^2 + 32 \cdot \gamma_n^2 \cdot \sum_{k=1}^{K_n} \sum_{l=0}^{1} |w_{k,l}^{(0)} - \bar{w}_{k,l}^{(0)}|^2.
\end{aligned}
$$

Together, this gives us

$$
\sum_{k,j,l} \left| \frac{\partial}{\partial w_{k,j}^{(l)}} f_{net,\mathbf{w}}(x) - \frac{\partial}{\partial w_{k,j}^{(l)}} f_{net,\bar{\mathbf{w}}}(x) \right|^2
$$

$$
\begin{aligned}
&\leq 2 \cdot \sum_{k=1}^{K_n} \sum_{j=0}^{1} |w_{k,j}^{(0)} - \bar{w}_{k,j}^{(0)}|^2 + 4 \cdot \sum_{k=1}^{K_n} |w_{1,k}^{(1)} - \bar{w}_{1,k}^{(1)}|^2 + 32 \cdot \gamma_n^2 \cdot \sum_{k=1}^{K_n} \sum_{l=0}^{1} |w_{k,l}^{(0)} - \bar{w}_{k,l}^{(0)}|^2 \\
&\leq 34 \cdot \gamma_n^2 \cdot \left(\sum_{k=1}^{K_n} |w_{1,k}^{(1)} - \bar{w}_{1,k}^{(1)}|^2 + \sum_{k=1}^{K_n} \sum_{j=0}^{1} |w_{k,j}^{(0)} - \bar{w}_{k,j}^{(0)}|^2 \right) \\
&= 34 \cdot \gamma_n^2 \cdot \|\mathbf{w} - \bar{\mathbf{w}}\|^2.
\end{aligned}
$$

□

The following two lemmas will help us to verify the conditions in Lemma 2.2.6 in our setting.

Lemma 2.2.9. *Assume* $supp(X) \subseteq [0,1]$, $\gamma_n \geq 1$, $2 \cdot t_n \geq L_n$, $c_1^2 \leq K_n$ *and*

$$
|w_{1,k}^{(1)}| \leq \gamma_n \quad (k = 1, \ldots, K_n) \quad \textit{and} \quad \|\mathbf{w} - \mathbf{v}\|^2 \leq \frac{2t_n}{L_n} \cdot \max\{F(\mathbf{v}), 1\}. \tag{2.18}
$$

Then we have with probability one

$$
\|(\nabla_{\mathbf{w}} F)(\mathbf{w})\| \leq 25 \cdot \gamma_n^2 \cdot K_n \cdot \sqrt{\frac{t_n}{L_n} \cdot \max\{F(\mathbf{v}), 1\}}.
$$

Proof. Since

$$
\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x)) \in [0,1]
$$

we know that σ is Lipschitz continuous with Lipschitz constant 1. Moreover, we have

$$
\begin{aligned}
\|\mathbf{w} - \mathbf{v}\|_1^2 &= \left(\sum_{j,k,l} |w_{j,k}^{(l)} - v_{j,k}^{(l)}| \right)^2 \\
&\leq (3K_n + 1) \cdot \sum_{j,k,l} |w_{j,k}^{(l)} - v_{j,k}^{(l)}|^2 \\
&= (3K_n + 1) \cdot \|\mathbf{w} - \mathbf{v}\|^2 \\
&\leq 4K_n \cdot \frac{2t_n}{L_n} \cdot \max\{F(\mathbf{v}), 1\}.
\end{aligned}
$$

By the Cauchy-Schwarz inequality and by Lemma 2.2.7 we get

$$
\|(\nabla_{\mathbf{w}} F(\mathbf{w})\|^2
$$

$$
\begin{aligned}
&= \sum_{j,k,l}\left(\frac{2}{n}\sum_{i=1}^{n}(Y_i - f_{net,\mathbf{w}}(X_i))\cdot\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}}(X_i) + \frac{\partial}{\partial w_{j,k}^{(l)}}\left(\frac{c_1}{K_n}\cdot\sum_{l=0}^{K_n}(w_{1,l}^{(1)})^2\right)\right)^2 \\
&\leq \sum_{j,k,l} 2\cdot\left(\frac{2}{n}\sum_{i=1}^{n}(Y_i - f_{net,\mathbf{w}}(X_i))\cdot\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}}(X_i)\right)^2 \\
&\quad + \sum_{j,k,l} 2\cdot\left(\frac{\partial}{\partial w_{j,k}^{(l)}}\left(\frac{c_1}{K_n}\cdot\sum_{l=0}^{K_n}(w_{1,l}^{(1)})^2\right)\right)^2 \\
&\leq \sum_{j,k,l}\frac{8}{n}\sum_{i=1}^{n}(Y_i - f_{net,\mathbf{w}}(X_i))^2\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}}(X_i)\right)^2 \\
&\quad + \sum_{l=0}^{K_n} 2\cdot\left(\frac{c_1}{K_n}\cdot 2\cdot w_{1,l}^{(1)}\right)^2 \\
&= \sum_{j,k,l}\frac{8}{n}\sum_{i=1}^{n}(Y_i - f_{net,\mathbf{w}}(X_i))^2\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}}(X_i)\right)^2 \\
&\quad + \sum_{l=0}^{K_n}\frac{8\cdot c_1^2}{K_n^2}\cdot\left(w_{1,l}^{(1)}\right)^2 \\
&\leq \frac{8}{n}\sum_{i=1}^{n}\left((Y_i - f_{net,\mathbf{v}}(X_i)) + (f_{net,\mathbf{v}}(X_i) - f_{net,\mathbf{w}}(X_i))\right)^2 \\
&\qquad \cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{j,k,l}\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}}(X_i)\right)^2 + \frac{8\cdot c_1^2}{K_n^2}\cdot\sum_{l=0}^{K_n}\left(w_{1,l}^{(1)}\right)^2 \\
&\leq 16\cdot\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{net,\mathbf{v}}(X_i))^2 + \frac{1}{n}\sum_{i=1}^{n}(f_{net,\mathbf{v}}(X_i) - f_{net,\mathbf{w}}(X_i))^2\right) \\
&\qquad \cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{j,k,l}\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}}(X_i)\right)^2 + \frac{8\cdot c_1^2}{K_n^2}\cdot\sum_{l=0}^{K_n}\left(w_{1,l}^{(1)}\right)^2 \\
&\leq 16\cdot\left(F(\mathbf{v}) + 1^2\cdot 1^2\cdot\gamma_n^2\cdot\|\mathbf{v}-\mathbf{w}\|_1^2\right)\cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{j,k,l}\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}}(X_i)\right)^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{8\cdot c_1^2}{K_n^2}\cdot(K_n+1)\cdot\gamma_n^2\\
\leq\;& 16\cdot\left(F(\mathbf{v})+\gamma_n^2\cdot 4K_n\cdot\frac{2t_n}{L_n}\cdot\max\{F(\mathbf{v}),1\}\right)\cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{j,k,l}\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}}(X_i)\right)^2\\
&+16\cdot\frac{c_1^2}{K_n^2}\cdot K_n\cdot\gamma_n^2\\
\leq\;& 144\cdot\gamma_n^2\cdot K_n\cdot\frac{t_n}{L_n}\cdot\max\{F(\mathbf{v}),1\}\cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{j,k,l}\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}}(X_i)\right)^2\\
&+16\cdot\frac{c_1^2}{K_n}\cdot\gamma_n^2.
\end{aligned}
$$

We bound the right-hand side further. We have for any $i\in\{1,\ldots,n\}$

$$
\begin{aligned}
&\sum_{j,k,l}\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}}(X_i)\right)^2\\
=\;&\sum_{k=0}^{K_n}\left(\frac{\partial}{\partial w_{1,k}^{(1)}}f_{net,\mathbf{w}}(X_i)\right)^2+\sum_{k=1}^{K_n}\sum_{j=0}^{1}\left(\frac{\partial}{\partial w_{j,k}^{(0)}}f_{net,\mathbf{w}}(X_i)\right)^2\\
=\;&\sum_{k=0}^{K_n}\left(\frac{\partial}{\partial w_{1,k}^{(1)}}f_{net,\mathbf{w}}(X_i)\right)^2+\sum_{k=1}^{K_n}\left(\frac{\partial}{\partial w_{0,k}^{(0)}}f_{net,\mathbf{w}}(X_i)\right)^2+\sum_{k=1}^{K_n}\left(\frac{\partial}{\partial w_{1,k}^{(0)}}f_{net,\mathbf{w}}(X_i)\right)^2\\
=\;&1+\sum_{k=1}^{K_n}\left(\sigma\left(w_{k,1}^{(0)}\cdot X_i+w_{k,0}^{(0)}\right)\right)^2+\sum_{k=1}^{K_n}\left(w_{1,k}^{(1)}\cdot\sigma'\left(w_{k,1}^{(0)}\cdot X_i+w_{k,0}^{(0)}\right)\cdot 1\right)^2\\
&+\sum_{k=1}^{K_n}\left(w_{1,k}^{(1)}\cdot\sigma'\left(w_{k,1}^{(0)}\cdot X_i+w_{k,0}^{(0)}\right)\cdot X_i\right)^2\\
\leq\;&\sum_{k=0}^{K_n}1^2+\sum_{k=1}^{K_n}\left(w_{1,k}^{(1)}\right)^2+\sum_{k=1}^{K_n}\left(w_{1,k}^{(1)}\right)^2\\
\leq\;&(K_n+1)+K_n\cdot\gamma_n^2+K_n\cdot\gamma_n^2\\
\leq\;&4\cdot\gamma_n^2\cdot K_n.
\end{aligned}
$$

Together this gives us

$$\begin{aligned}
& \|(\nabla_{\mathbf{v}} F_{\mathbf{w}})(\mathbf{v})\|^2 \\
\leq\ & 144 \cdot \gamma_n^2 \cdot K_n \cdot \frac{t_n}{L_n} \cdot \max\{F(\mathbf{v}), 1\} \cdot 4 \cdot \gamma_n^2 \cdot K_n + 16 \cdot \frac{c_1^2}{K_n} \cdot \gamma_n^2 \\
\leq\ & 592 \cdot K_n^2 \cdot \frac{t_n}{L_n} \cdot \max\{F(\mathbf{v}), 1\} \cdot \gamma_n^4,
\end{aligned}$$

which implies the assertion. □

Lemma 2.2.10. *Assume* $supp(X) \subseteq [0,1]$, $\gamma_n \geq 1$, $t_n \geq L_n$ *and*

$$\max\left\{|(\mathbf{w}_2)_{1,k}^{(1)}|, |v_{1,k}^{(1)}|\right\} \leq \gamma_n \quad (k = 1, \ldots, K_n) \quad \textit{and} \quad \|\mathbf{w}_2 - \mathbf{v}\|^2 \leq 8 \cdot \frac{t_n}{L_n} \cdot \max\{F(\mathbf{v}), 1\}. \tag{2.19}$$

Then we have with probability one

$$\begin{aligned}
& \|(\nabla_{\mathbf{w}} F)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F)(\mathbf{w}_2)\| \\
\leq\ & 165 \cdot \max\{\sqrt{F(\mathbf{v})}, 1\} \cdot \max\{\frac{c_1}{K_n}, 1\} \cdot \gamma_n^2 \cdot K_n \cdot \sqrt{\frac{t_n}{L_n}} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|.
\end{aligned}$$

Proof. We have by the Cauchy-Schwarz inequality

$$\begin{aligned}
& \|(\nabla_{\mathbf{w}} F)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F)(\mathbf{w}_2)\|^2 \\
=\ & \sum_{j,k,l} \Bigg(\frac{2}{n} \sum_{i=1}^n (Y_i - f_{net,\mathbf{w}_1}(X_i)) \cdot \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_1}(X_i) \\
& \quad + \frac{\partial}{\partial w_{j,k}^{(l)}} \left(\frac{c_1}{K_n} \cdot \sum_{l=0}^{K_n} ((\mathbf{w}_1)_{1,l}^{(1)})^2 \right) \\
& \quad - \frac{2}{n} \sum_{i=1}^n (Y_i - f_{net,\mathbf{w}_2}(X_i)) \cdot \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_2}(X_i) \\
& \quad - \frac{\partial}{\partial w_{j,k}^{(l)}} \left(\frac{c_1}{K_n} \cdot \sum_{l=0}^{K_n} ((\mathbf{w}_2)_{1,l}^{(1)})^2 \right) \Bigg)^2 \\
\leq\ & 3 \cdot \sum_{j,k,l} \Bigg(\frac{2}{n} \sum_{i=1}^n (Y_i - f_{net,\mathbf{w}_1}(X_i)) \cdot \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_1}(X_i) \\
& \quad - \frac{2}{n} \sum_{i=1}^n (Y_i - f_{net,\mathbf{w}_2}(X_i)) \cdot \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_1}(X_i) \Bigg)^2
\end{aligned}$$

$$
\begin{aligned}
&+3\cdot\sum_{j,k,l}\left(\frac{2}{n}\sum_{i=1}^{n}(Y_i-f_{net,\mathbf{w}_2}(X_i))\cdot\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}_1}(X_i)\right.\\
&\quad\left.-\frac{2}{n}\sum_{i=1}^{n}(Y_i-f_{net,\mathbf{w}_2}(X_i))\cdot\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}_2}(X_i)\right)^2\\
&+3\cdot\sum_{l=0}^{K_n}\left(\frac{c_1}{K_n}\cdot 2\cdot(\mathbf{w}_1)_{1,l}^{(1)}-\frac{c_1}{K_n}\cdot 2\cdot(\mathbf{w}_2)_{1,l}^{(1)}\right)^2\\
=\;& 12\cdot\sum_{j,k,l}\left(\frac{1}{n}\sum_{i=1}^{n}(f_{net,\mathbf{w}_2}(X_i)-f_{net,\mathbf{w}_1}(X_i))\cdot\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}1}(X_i)\right)^2\\
&+12\cdot\sum_{j,k,l}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i-f_{net,\mathbf{w}_2}(X_i))\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}_1}(X_i)-\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}_2}(X_i)\right)\right)^2\\
&+12\cdot\frac{c_1^2}{K_n^2}\cdot\sum_{l=0}^{K_n}\left|(\mathbf{w}_1)_{1,l}^{(1)}-(\mathbf{w}_2)_{1,l}^{(1)}\right|^2\\
\leq\;& 12\cdot\frac{1}{n}\sum_{i=1}^{n}(f_{net,\mathbf{w}_1}(X_i)-f_{net,\mathbf{w}_2}(X_i))^2\cdot\sum_{j,k,l}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}_1}(X_i)\right)^2\\
&+12\cdot\frac{1}{n}\sum_{i=1}^{n}(Y_i-f_{net,\mathbf{w}_2}(X_i))^2\\
&\qquad\cdot\sum_{j,k,l}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}_1}(X_i)-\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}_2}(X_i)\right)^2\\
&+12\cdot\frac{c_1^2}{K_n^2}\cdot\sum_{l=0}^{K_n}\left|(\mathbf{w}_1)_{1,l}^{(1)}-(\mathbf{w}_2)_{1,l}^{(1)}\right|^2.
\end{aligned}
$$

We bound the three terms separately. Using Lemma 2.2.7 and the proof of Lemma 2.2.9 we get regarding the first term

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_{net,\mathbf{w}_1}(X_i)-f_{net,\mathbf{w}_2}(X_i))^2\cdot\sum_{j,k,l}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial w_{j,k}^{(l)}}f_{net,\mathbf{w}_1}(X_i)\right)^2\\
\leq\;& 1\cdot 1\cdot\gamma_n^2\cdot\|\mathbf{w}_1-\mathbf{w}_2\|_1^2\cdot 4\cdot\gamma_n^2\cdot K_n
\end{aligned}
$$

$$
\begin{aligned}
&\leq \gamma_n^2 \cdot 4 \cdot K_n \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2 \cdot 4 \cdot \gamma_n^2 \cdot K_n \\
&\leq 16 \cdot \gamma_n^4 \cdot K_n^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2.
\end{aligned}
$$

Again by Lemma 2.2.7 we get for the second term

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{net,\mathbf{w}_2}(X_i))^2 \\
\leq\ & \frac{2}{n}\sum_{i=1}^{n}(Y_i - f_{net,\mathbf{v}}(X_i))^2 + \frac{2}{n}\sum_{i=1}^{n}(f_{net,\mathbf{w}_2}(X_i) - f_{net,\mathbf{v}}(X_i))^2 \\
\leq\ & 2 \cdot F(\mathbf{v}) + 2 \cdot 1 \cdot 1 \cdot \gamma_n^2 \cdot \|\mathbf{w}_2 - \mathbf{v}\|_1^2 \\
\leq\ & 2 \cdot \max\{F(\mathbf{v}), 1\} + 2 \cdot \gamma_n^2 \cdot 4 \cdot K_n \cdot \|\mathbf{w}_2 - \mathbf{v}\|^2 \\
\leq\ & 2 \cdot \max\{F(\mathbf{v}), 1\} + 8 \cdot \gamma_n^2 \cdot K_n \cdot 8 \cdot \frac{t_n}{L_n} \cdot \max\{F(\mathbf{v}), 1\} \\
\leq\ & 66 \cdot \gamma_n^2 \cdot K_n \cdot \frac{t_n}{L_n} \cdot \max\{F(\mathbf{v}), 1\}
\end{aligned}
$$

and by Lemma 2.2.8 we get

$$
\begin{aligned}
&\sum_{j,k,l}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_1}(X_i) - \frac{\partial}{\partial w_{j,k}^{(l)}} f_{net,\mathbf{w}_2}(X_i)\right)^2 \\
&\leq 34 \cdot \gamma_n^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2.
\end{aligned}
$$

Summarizing the above results we get

$$
\begin{aligned}
&\|(\nabla_{\mathbf{w}} F)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F)(\mathbf{w}_2)\|^2 \\
\leq\ & 12 \cdot 16 \cdot \gamma_n^4 \cdot K_n^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2 \\
&\quad + 12 \cdot 66 \cdot \gamma_n^2 \cdot K_n \cdot \frac{t_n}{L_n} \cdot \max\{F(\mathbf{v}), 1\} \cdot 34 \cdot \gamma_n^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2 \\
&\quad + 12 \cdot \frac{c_1^2}{K_n^2} \cdot \sum_{l=0}^{K_n}\left|(\mathbf{w}_1)_{1,l}^{(1)} - (\mathbf{w}_2)_{1,l}^{(1)}\right|^2 \\
\leq\ & 27132 \cdot \gamma_n^4 \cdot K_n^2 \cdot \frac{t_n}{L_n} \cdot \max\{F(\mathbf{v}), 1\} \cdot \max\{\frac{c_1^2}{K_n^2}, 1\} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2,
\end{aligned}
$$

which implies the assertion. □

The following Lemma 2.2.11 guarantees the existence of "the right places" where our neural network estimate is close to a piecewise constant function approximation. We will use this in the proof of Theorem 2.2.1.

Lemma 2.2.11. *Let $K \in \mathbb{N}$, $\tilde{K} \in \{1, \dots, K-1\}$ and set*

$$I_k = \left[\frac{k-1}{\tilde{K}}, \frac{k}{\tilde{K}}\right) \quad \text{for } k \in \{0, \dots, \tilde{K}\}.$$

Let $x_1, \dots, x_n \in [0,1]$ and let β_1, γ_1, …, β_K, γ_K be independent and identically distributed random variables which are uniformly distributed on $[-B_n, B_n]$, where

$$B_n \geq 16 \cdot K \cdot (\log K)^2 \cdot n.$$

Then with probability at least $1 - (\tilde{K}+1) \cdot \exp(-K/(32 \cdot \tilde{K}))$ there exist $i_0, i_1, \dots, i_{\tilde{K}} \in \{1, \dots, K\}$ such that

$$\beta_{i_k} \geq \frac{B_n}{2}, \quad -\frac{\gamma_{i_k}}{\beta_{i_k}} \in I_k \quad \text{and} \quad \min_{i=1,\dots,n} |\beta_{i_k} \cdot x_i + \gamma_{i_k}| \geq 2 \cdot (\log K)^2$$

holds for every $k \in \{0, \dots, \tilde{K}\}$.

Proof. We prove the assertion by computing the counter-probability. We have

$$\begin{aligned}
&\mathbf{P}\Bigg\{\exists l \in \{0, \dots, \tilde{K}\}\, \forall k \in \{1, \dots, K\} : \\
&\qquad\qquad \beta_k < \frac{B_n}{2}, \quad -\frac{\gamma_k}{\beta_k} \notin I_l \quad \text{or} \quad \min_{i=1,\dots,n} |\beta_k \cdot x_i + \gamma_k| < 2 \cdot (\log K)^2\Bigg\} \\
&\leq (\tilde{K}+1) \cdot \max_{l=1,\dots,\tilde{K}} \prod_{k=1}^{K} \Bigg(1 - \\
&\qquad \mathbf{P}\left\{\beta_k \geq \frac{B_n}{2}, \quad -\frac{\gamma_k}{\beta_k} \in I_l \quad \text{and} \quad \min_{i=1,\dots,n} |\beta_k \cdot x_i + \gamma_k| \geq 2 \cdot (\log K)^2\right\}\Bigg).
\end{aligned}$$

In case $\beta_k > 0$ we have that $-\frac{\gamma_k}{\beta_k} \in I_l$ (for $l \in \{0, 1, \dots, \tilde{K}\}$) is equivalent to

$$\gamma_k \in \left(-l \cdot \frac{\beta_k}{\tilde{K}}, (-l+1) \cdot \frac{\beta_k}{\tilde{K}}\right].$$

Hence, in case $\beta_k > 0$ we have that $-\frac{\gamma_k}{\beta_k} \in I_l$ holds true if γ_k is contained in a subinterval of length

$$\frac{\beta_k}{\tilde{K}}$$

of $[-B_n, B_n]$. Furthermore, $\min_{i=1,\dots,n} |\beta_k \cdot x_i + \gamma_k| \geq 2 \cdot (\log K)^2$ holds if γ_k is not contained in a union of n intervals of length $4 \cdot (\log K)^2$. As a result, we know that in case that $\beta_k \geq B_n/2$ we have that γ_k satisfies

$$-\frac{\gamma_k}{\beta_k} \in I_l \quad \text{and} \quad \min_{i=1,\dots,n} |\beta_k \cdot x_i + \gamma_k| \geq 2 \cdot (\log K)^2$$

if γ_k is contained in a subset of $[-B_n, B_n]$ of Lebesgue measure at least

$$\frac{B_n}{2\tilde{K}} - 4 \cdot n \cdot (\log K)^2 \geq \frac{B_n}{2\tilde{K}} - \frac{B_n}{4K} \geq \frac{B_n}{2\tilde{K}} - \frac{B_n}{4\tilde{K}} = \frac{2B_n}{4\tilde{K}} - \frac{B_n}{4\tilde{K}} = \frac{B_n}{4 \cdot \tilde{K}},$$

where the first inequality holds since by definition of B_n

$$B_n \geq 4 \cdot K \cdot (4 \cdot n \cdot (\log K)^2).$$

This implies that the probability above is bounded from above by

$$\begin{aligned}(\tilde{K}+1) \cdot \prod_{k=1}^{K} \left(1 - \frac{1}{4} \cdot \frac{B_n/(4 \cdot \tilde{K})}{2 \cdot B_n}\right) &= (\tilde{K}+1) \cdot \left(1 - \frac{1}{32 \cdot \tilde{K}}\right)^K \\ &\leq (\tilde{K}+1) \cdot \exp\left(-\frac{1}{32} \cdot \frac{K}{\tilde{K}}\right).\end{aligned}$$

□

The following Lemma 2.2.12 will help us to analyze the development of the outer weights during gradient descent when only the outer weights are learned by gradient descent.

Lemma 2.2.12. *Let F be defined by*

$$F(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} |\sum_{k=1}^{K} a_k \cdot B_k(x_i) - y_i|^2 + \tau_n \cdot \|\mathbf{a}\|^2$$

where $\mathbf{a} = (a_1, \dots, a_K)^T \in \mathbb{R}^K$, $B_k(x) : \mathbb{R} \to \mathbb{R}$ and $\tau_n \in \mathbb{R}$, $\tau_n > 0$. Choose $\mathbf{a}_{opt}$ such that

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

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

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2 \geq 4 \cdot \tau_n \cdot (F(\mathbf{a}) - F(\mathbf{a}_{opt})).$$

Proof. Set

$$\mathbf{B} = (B_j(x_i))_{1\le i\le n, 1\le j\le K} \quad \text{and} \quad \mathbf{A} = \frac{1}{n}\cdot \mathbf{B}^T\cdot\mathbf{B} + \tau_n\cdot\mathbf{1},$$

where $\mathbf{1}$ is the unit matrix. Clearly, the matrix $\mathbf{A}$ is symmetric and positive definite and hence regular. As a consequence, we can write

$$\begin{aligned}
F(\mathbf{a}) &= \frac{1}{n}\cdot(\mathbf{B}\cdot\mathbf{a}-\mathbf{y})^T\cdot(\mathbf{B}\cdot\mathbf{a}-\mathbf{y}) + \tau_n\cdot\mathbf{a}^T\cdot\mathbf{a} \\
&= \frac{1}{n}(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{B}\mathbf{a}) + \mathbf{a}^T\left(\frac{1}{n}\mathbf{B}^T\mathbf{B} + \tau_n\cdot\mathbf{1}\right)\mathbf{a} \\
&= \mathbf{a}^T\mathbf{A}\mathbf{a} - 2\mathbf{y}^T\frac{1}{n}\mathbf{B}\mathbf{a} + \frac{1}{n}\mathbf{y}^T\mathbf{y} \\
&= \left(\mathbf{a} - \frac{1}{n}\mathbf{A}^{-1}\mathbf{B}^T\mathbf{y}\right)^T\mathbf{A}\left(\mathbf{a} - \frac{1}{n}\mathbf{A}^{-1}\mathbf{B}^T\mathbf{y}\right) + \frac{1}{n}\mathbf{y}^T\mathbf{y} - \frac{1}{n^2}\mathbf{y}^T\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T\mathbf{y}.
\end{aligned}$$

We observe that the last two terms are independent of $\mathbf{a}$ and hence the right-hand side is minimal for

$$\mathbf{a} - \frac{1}{n}\mathbf{A}^{-1}\mathbf{B}^T\mathbf{Y} = 0.$$

Thus,

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

This gives us

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

Since $\mathbf{A}$ is symmetric and positive definite we know by the Spectral theorem that $\mathbf{A}$ is diagonalizable with

$$\mathbf{A} = \mathbf{Q}^T\mathbf{D}\mathbf{Q}$$

where the orthogonal matrix $\mathbf{Q}$ has as column vectors the eigenvectors of $\mathbf{A}$ and the diagonal matrix $\mathbf{D}$ has on its diagonal the non-negative eigenvalues of $\mathbf{A}$. We denote by $\mathbf{D}^{1/2}$ the diagonal matrix where we take the square root of each entry of $\mathbf{D}$ and we denote by

$$\mathbf{A}^{1/2} = \mathbf{Q}^T\mathbf{D}^{1/2}\mathbf{Q}$$

the unique positive definite matrix satisfying

$$\mathbf{A}^{1/2}\cdot\mathbf{A}^{1/2} = \mathbf{A}.$$

Using

$$\mathbf{b}^T\mathbf{A}\mathbf{b} \geq \tau_n \cdot \mathbf{b}^T\mathbf{b}$$

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

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

where the last equality follows from

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

□

The following Lemma 2.2.13 will help us to show that our neural network is close to a piecewise constant approximation.

Lemma 2.2.13. *Let σ be the logistic squasher.*

a) For any $x \in \mathbb{R}$ we have

$$|\sigma(x) - 1_{[0,\infty)}(x)| \leq e^{-|x|}.$$

b) For any $b \in \mathbb{R}$, $c > 0$ and $x \in \mathbb{R}$ we have

$$|\sigma(c \cdot (x - b)) - 1_{[b,\infty)}(x)| \leq e^{-c \cdot |x-b|}.$$

Proof. a) For $x \geq 0$ we have

$$|\sigma(x) - 1_{[0,\infty)}(x)| = 1 - \frac{1}{1 + e^{-x}} = \frac{e^{-x}}{1 + e^{-x}} \leq e^{-x} = e^{-|x|}.$$

And for $x < 0$ we get

$$|\sigma(x) - 1_{[0,\infty)}(x)| = \frac{1}{1+e^{-x}} \leq e^{x} = e^{-|x|}.$$

b) From $c > 0$ and a) we get

$$|\sigma(c\cdot(x-b)) - 1_{[b,\infty)}(x)| = |\sigma(c\cdot(x-b)) - 1_{[0,\infty)}(c\cdot(x-b))| \leq e^{-|c\cdot(x-b)|} = e^{-c\cdot|x-b|}.$$

□

2.2.3. Auxiliary Lemmas from Empirical Process Theory

The next lemma is an auxiliary result from empirical process theory. We will need it again in in Section 3.2.3 and Section 4.2.1, where the idea will be adapted according to our needs.

Lemma 2.2.14. *Let*

- $\beta_n = c_3 \cdot \log(n)$ *for some suitably large constant* $c_3 > 0$
- $\mathcal{F}_n$ *be a set of functions* $f : \mathbb{R}^d \to \mathbb{R}$
- A_n *be an arbitrary event.*

Assume

- *the distribution of* (X, Y) *satisfies* (2.7) *for some constant* $c_{10} > 0$
- *the regression function* m *is bounded in absolute value*
- *the estimate* m_n *satisfies*

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

where on the event A_n

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

Then m_n *satisfies*

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

$$-2 \cdot \left(\frac{1}{n} \sum_{i=1}^{n} |\tilde{m}_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{\{|Y_j| \leq \beta_n \, (j \in \{1,\ldots,n\})\}} - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \right) \cdot \mathbf{1}_{A_n} \Bigg)$$
$$\leq \frac{c_7 \cdot (\log n)^2 \cdot \left(\log \left(\sup_{x_1^n} \mathcal{N}_1 \left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, x_1^n \right) \right) + 1 \right)}{n}$$

for $n > 1$ and some constant $c_7 > 0$, which does not depend on n, β_n or the parameters of the estimate.

Proof. This lemma follows in a straightforward way from the proof of Theorem 1 in Bagirov, Clausen and Kohler (2009). Since we have the characteristic function as a new term that needs to be carried through the proof, for the sake of completeness, a complete version of the proof is given in the Supplement in Section $A.2$. □

In order to bound the covering number $\mathcal{N}_1 \left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, x_1^n \right)$ we will use the following lemma. This lemma is embedded into our setting and close to the second auxiliary result in Section 4.2.1. Here, we consider a neural network with one hidden layer. Note that in this class of functions the inner weights do not necessarily need to be bounded.

Lemma 2.2.15. *Let $\max\{K, \beta_n, \gamma_n\} \leq n^{c_8}$ and define $\mathcal{F}$ by*

$$\mathcal{F} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \, : \, f(x) = \sum_{k=0}^{K} a_k \cdot \sigma \left(\sum_{j=1}^{d} b_{k,j} \cdot x^{(j)} + b_{k,0} \right) \quad (x \in \mathbb{R}^d) \right.$$
$$\left. \text{for some } a_k, b_{k,j} \in \mathbb{R} \text{ satisfying } \sum_{k=0}^{K} a_k^2 \leq \gamma_n \right\}.$$

Then we have for any $x_1^n \in (\mathbb{R}^d)^n$:

$$\log \left(\mathcal{N}_1 \left(\frac{1}{n \cdot \beta_n}, \mathcal{F}, x_1^n \right) \right) \leq c_9 \cdot \log n \cdot K.$$

Proof. Let

$$\mathcal{G} = \left\{ g : \mathbb{R}^d \to \mathbb{R} : g(x) = \sigma \left(\sum_{j=1}^{d} b_j \cdot x^{(j)} + b_0 \right) \quad (x \in \mathbb{R}^d) \right\}$$

for some $b_0, b_1, \ldots, b_d \in \mathbb{R}$. and let

$$\mathcal{F}' = \left\{ \sum_{i=1}^{K+1} w_i \cdot f_i : (w_1, \ldots, w_{K+1}) \in \mathbb{R}^{K+1} \text{ satisfying } \sum_{i=1}^{K+1} |w_i| \leq \sqrt{\gamma_n \cdot (K+1)} \right.$$

$$\text{and } f_i \in \mathcal{G} \text{ for } i = 1, \ldots, K+1 \Bigg\}.$$

Let

$$f(x) = \sum_{k=0}^{K} a_k \cdot \sigma \left(\sum_{j=1}^{d} b_{k,j} \cdot x^{(j)} + b_{k,0} \right) \in \mathcal{F}.$$

Then by the Cauchy-Schwarz inequality and by assumption of the lemma we have

$$\sum_{k=0}^{K} |a_k| = \sqrt{\left(\sum_{k=0}^{K} |a_k| \cdot 1 \right)^2} \leq \sqrt{\sum_{k=0}^{K} |a_k|^2} \cdot \sqrt{\sum_{k=0}^{K} 1^2} \leq \sqrt{\gamma_n} \cdot \sqrt{K+1}.$$

Hence, it also holds that $f \in \mathcal{F}'$ and consequently $\mathcal{F} \subseteq \mathcal{F}'$. Thus,

$$\mathcal{N}_1 \left(\frac{1}{n \cdot \beta_n}, \mathcal{F}, x_1^n \right) \leq \mathcal{N}_1 \left(\frac{1}{n \cdot \beta_n}, \mathcal{F}', x_1^n \right).$$

We continue by finding an upper bound for the covering number of $\mathcal{F}'$. For that, we apply Lemma 16.6 in Györfi et al. (2002) with $\mathcal{G}_i = \mathcal{G}$ for all $i = 1, \ldots, K$, $B = 1$, $b = \sqrt{\gamma_n \cdot (K+1)}$, $\eta + \delta = \frac{1}{n\beta_n}$ and $\eta = \delta = \frac{1}{n\beta_n} \cdot \frac{1}{2} > 0$ and receive

$$\begin{aligned}
& \mathcal{N}_1 \left(\frac{1}{n\beta_n}, \mathcal{F}', x_1^n \right) \\
\leq \; & \left(\frac{1 \cdot e \cdot \left(\frac{\sqrt{(K+1)\gamma_n} + 2 \cdot \frac{1}{2} \cdot \frac{1}{n\beta_n}}{1} \right)}{\frac{1}{2} \cdot \frac{1}{n\beta_n}} \right)^{K+1} \prod_{i=1}^{K+1} \mathcal{N}_1 \left(\frac{\frac{1}{2} \cdot \frac{1}{n\beta_n}}{\sqrt{(K+1)\gamma_n} + 2 \cdot \frac{1}{2} \cdot \frac{1}{n\beta_n}}, \mathcal{G}_i, x_1^n \right) \\
= \; & \left(\frac{e \cdot \left(\sqrt{(K+1)\gamma_n} + \frac{1}{n\beta_n} \right)}{\frac{1}{2n\beta_n}} \right)^{K+1} \cdot \left(\mathcal{N}_1 \left(\frac{\frac{1}{2n\beta_n}}{\sqrt{(K+1)\gamma_n} + \frac{1}{n\beta_n}}, \mathcal{G}, x_1^n \right) \right)^{K+1} \\
= \; & \left(e \cdot \left(\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2 \right) \cdot \mathcal{N}_1 \left(\frac{1}{\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2}, \mathcal{G}, x_1^n \right) \right)^{K+1}.
\end{aligned}$$

By Lemma 9.2 in Györfi et al. (2002) we know that we can bound the covering number by the packing number, i.e.

$$\mathcal{N}_1 \left(\frac{1}{\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2}, \mathcal{G}, x_1^n \right) \leq \mathcal{M}_1 \left(\frac{1}{\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2}, \mathcal{G}, x_1^n \right).$$

In order to bound the packing number on the right-hand side we will need a bound for the Vapnik-Chervonenkis dimension of the set of all subgraphs of functions of $\mathcal{G}$ defined as

$$\mathcal{G}^+ = \left\{ \{(z,t) \in \mathbb{R}^d \times \mathbb{R} : t \leq g(z)\} : g \in \mathcal{G} \right\}.$$

For that, let

$$\mathcal{H} = \left\{ \sum_{j=1}^{d} b_j \cdot x^{(j)} + b_0 : x \in \mathbb{R}^d \text{ and } b_0, \ldots, b_d \in \mathbb{R} \right\}.$$

Then we write

$$\mathcal{G} = \{\sigma \circ h : h \in \mathcal{H}\}$$

and since the logistic squasher σ is a non-decreasing function we know by Lemma 16.3 in Györfi et al. (2002) that

$$V_{\mathcal{G}^+} \leq V_{\mathcal{H}^+}.$$

Further, since

$$\begin{aligned}
\mathcal{H}^+ &= \left\{ \{(z,t) \in \mathbb{R}^d \times \mathbb{R} : t \leq h(z)\} : h \in \mathcal{H} \right\} \\
&= \left\{ \{(z,t) \in \mathbb{R}^d \times \mathbb{R} : (-1) \cdot t + h(z) \geq 0\} : h \in \mathcal{H} \right\} \\
&\subseteq \left\{ \{(z,t) \in \mathbb{R}^d \times \mathbb{R} : \alpha \cdot t + h(z) \geq 0\} : h \in \mathcal{H}, \alpha \in \mathbb{R} \right\} \\
&=: \mathcal{A}
\end{aligned}$$

we have

$$V_{\mathcal{H}^+} \leq V_{\mathcal{A}}.$$

We bound $V_{\mathcal{A}}$ by Theorem 9.5 in Györfi et al. (2002). For that we notice that

$$\{\alpha \cdot t + h(z) : h \in \mathcal{H}, \alpha \in \mathbb{R}\}$$

is an $(d+2)-$dimensional vector space of real functions on $\mathbb{R}^d$ as $\mathcal{H}$ is a linear vector space of dimension $(d+1)$. Consequently,

$$V_{\mathcal{A}} \leq d + 2.$$

Now, we bound the packing number by Theorem 9.4 in in Györfi et al. (2002) with $B = 1$, $p = 1$ and $\epsilon = 1/(\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2)$ which gives us

$$\mathcal{M}_1(\epsilon, \mathcal{G}, x_1^n)$$

$$
\begin{aligned}
&\leq 3 \cdot \left(\frac{2 \cdot e \cdot 1^1}{\left(\frac{1}{\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2}\right)^1} \cdot \log \left(\frac{3 \cdot e \cdot 1^1}{\left(\frac{1}{\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2}\right)^1} \right) \right)^{V_{\mathcal{G}^+}} \\
&= 3 \cdot \left(2 \cdot e \cdot \left(\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2\right) \cdot \log \left(3e \cdot (\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2)\right)\right)^{V_{\mathcal{G}^+}} \\
&\leq 3 \cdot \left(2 \cdot e \cdot \left(\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2\right) \cdot \log \left(3e \cdot (\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2)\right)\right)^{d+2}.
\end{aligned}
$$

Summarizing the above results we get

$$
\begin{aligned}
&\log \left(\mathcal{N}_1(\frac{1}{n\beta_n}, \mathcal{F}, x_1^n) \right) \\
\leq\ & \log \Bigg(\Bigg(e \cdot \left(\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2\right) \cdot 3 \\
&\qquad \cdot \Big(2 \cdot e \cdot \left(\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2\right) \\
&\qquad\qquad \cdot \log \left(3e \cdot (\sqrt{(K+1)\gamma_n} \cdot 2n\beta_n + 2)\right) \Big)^{d+2} \Bigg)^{K+1} \Bigg) \\
\leq\ & (K+1) \cdot \log \Bigg(e \cdot \left(\sqrt{(n^{c_8}+1) \cdot n^{c_8}} \cdot 2n \cdot n^{c_8} + 2\right) \cdot 3 \\
&\qquad \cdot \Big(2 \cdot e \cdot \left(\sqrt{(n^{c_8}+1) \cdot n^{c_8}} \cdot 2n \cdot n^{c_8} + 2\right) \\
&\qquad\qquad \cdot \log \left(3e \cdot (\sqrt{(n^{c_8}+1) \cdot n^{c_8}} \cdot 2n \cdot n^{c_8} + 2)\right) \Big)^{d+2} \Bigg) \\
\leq\ & K \cdot c_9 \cdot \log n
\end{aligned}
$$

for n large enough.

□

2.2.4. Proof of Theorem 2.2.1

The proof of Theorem 2.2.1 is quite long and technical. For a better understanding we present a brief and highly simplified outline of the proof before going into detail. The proof has ten steps.

Step 1: Preparations.

- W.l.o.g. we assume $\|m\|_\infty \le \beta_n$.
- Set $\tilde{K}_n = \lceil K_n/(\log K_n)^2 \rceil$.
- Define A_n to be the event where
  1. $|Y_i| \le \beta_n$ holds for all $i = 1, \dots, n$,
  2. there exist $i_0, i_1, \dots, i_{\tilde{K}_n} \in \{1, \dots, K_n\}$ such that

$$\beta_{i_k} \ge \frac{B_n}{2}, \quad -\frac{\gamma_{i_k}}{\beta_{i_k}} \in I_k \quad \text{and} \quad \min_{i=1,\dots,n} |\beta_{i_k} \cdot X_i + \gamma_{i_k}| \ge 2 \cdot (\log K_n)^2$$

  holds for every $k \in \{1, \dots, \tilde{K}_n\}$.

Step 2: Starting the proof. By adding zeros we can rewrite the left-hand side as the sum of three terms. We have

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) = ① + ② + ③$$

where

$$\begin{aligned} ① \quad = \quad \mathbf{E}\Bigg(\Bigg(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ -2 \cdot \Bigg(\frac{1}{n} \sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{\{|Y_j| \le \beta_n \ (j \in \{1,\dots,n\})\}} \\ -\frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \Bigg) \Bigg) \cdot \mathbf{1}_{A_n} \Bigg) \end{aligned}$$

is of the form of our auxiliary lemma from empirical process theory which is applicable due to bounds on the weights during gradient descent,

$$\begin{aligned} ② \quad = \quad 2 \cdot \mathbf{E}\Bigg(\Bigg(\frac{1}{n} \sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{\{|Y_j| \le \beta_n \ (j \in \{1,\dots,n\})\}} \\ -\frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \Bigg) \cdot \mathbf{1}_{A_n} \Bigg), \end{aligned}$$

for which we will show that "at the right places" our estimate is close to a piecewise constant function approximation and that there gradient descent changes the inner weights only slightly and finally,

$$③ \quad = \quad 4\beta_n^2 \cdot \mathbf{P}(A_n^c)$$

which is bounded by the choice of our weights.
Step 3: Bounding the first summand. We show that

$$① \ \leq c_{30} \cdot \frac{(\log n)^3 \cdot K_n}{n}.$$

Step 4: Bounding the third summand. We show that

$$③ \ \leq \frac{c_{43}}{n}.$$

Step 5: Looking at the second summand. We break down ② into four terms by adding zeros. For that we need to introduce new functions. We define on $[0,1]$ a piecewise constant approximation of m by

$$f(x) = \sum_{k=1}^{\tilde{K}_n} \alpha^*_{i_k} \cdot \mathbf{1}_{\left[-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}}, \infty\right)}(x) + \alpha^*_{i_0}$$

where

$$\alpha^*_{i_0} := m\left(-\frac{\gamma^{(0)}_{i_0}}{\beta^{(0)}_{i_0}}\right)$$

and

$$\alpha^*_{i_j} = m\left(-\frac{\gamma^{(0)}_{i_j}}{\beta^{(0)}_{i_j}}\right) - m\left(-\frac{\gamma^{(0)}_{i_j-1}}{\beta^{(0)}_{i_j-1}}\right) \quad (j = 1, \ldots, \tilde{K}_n).$$

We set

$$f^*(x) = \sum_{k=1}^{\tilde{K}_n} \alpha^*_{i_k} \cdot \sigma(\beta^{(0)}_{i_k} \cdot x + \gamma^{(0)}_{i_k}) + \alpha^*_{i_0}.$$

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

$$\text{pen}(g) = \frac{c_1}{K_n^{2p}} \cdot \sum_{k=1}^{K_n} \alpha_k^2.$$

We get

$$② \ \leq \ \clubsuit + \spadesuit + \heartsuit + \diamondsuit$$

where

$$\clubsuit = \mathbf{E}\Bigg(\Bigg(F(\alpha^{(t_n)}, \beta^{(t_n)}, \gamma^{(t_n)}) - \frac{1}{n}\sum_{i=1}^{n}|Y_i - f^*(X_i)|^2 - \text{ pen}(f^*)\Bigg) \cdot \mathbf{1}_{A_n}\Bigg),$$

$$\spadesuit = \mathbf{E}\Bigg(\Bigg(\frac{1}{n}\sum_{i=1}^{n}(|Y_i - f^*(X_i)|^2 - \frac{1}{n}\sum_{i=1}^{n}|Y_i - f(X_i)|^2\Bigg) \cdot \mathbf{1}_{A_n}\Bigg),$$

$$\heartsuit = \mathbf{E}\Bigg(\Bigg(\frac{1}{n}\sum_{i=1}^{n}|Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^{n}|m(X_i) - Y_i|^2\Bigg) \cdot \mathbf{1}_{A_n}\Bigg)$$

and

$$\diamondsuit = \mathbf{E}\Bigg(\text{pen}(f^*) \cdot \mathbf{1}_{A_n}\Bigg).$$

Step 6: Bounding the first summand of Step 5. We bound $\clubsuit$ in several steps. First, we show that on A_n we have for any $t \in \{1, \ldots, t_n - 1\}$

$$\begin{aligned} & F(\alpha^{(t+1)}, \beta^{(t+1)}, \gamma^{(t+1)}) - \frac{1}{n}\sum_{i=1}^{n}|Y_i - f^*(X_i)|^2 - \text{pen}(f^*) \\ \leq \quad & \left(1 - \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}}\right) \cdot \left(F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) - \frac{1}{n}\sum_{i=1}^{n}|Y_i - f^*(X_i)|^2 - \text{pen}(f^*)\right) \\ & + \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot \left(F(\alpha^*, \beta^{(0)}, \gamma^{(0)}) - F(\alpha^*, \beta^{(t)}, \gamma^{(t)})\right). \end{aligned}$$

Second, we show that

$$F(\alpha^*, \beta^{(0)}, \gamma^{(0)}) - F(\alpha^*, \beta^{(t)}, \gamma^{(t)}) \leq 2 \cdot \beta_n^2 \cdot \left(\sum_{k=1}^{\tilde{K}_n}|\beta_{i_k}^{(t)} - \beta_{i_k}^{(0)}| + |\gamma_{i_k}^{(t)} - \gamma_{i_k}^{(0)}|\right)^2.$$

Third, we show that

$$\left(\sum_{k=1}^{\tilde{K}_n}|\beta_{i_k}^{(t)} - \beta_{i_k}^{(0)}| + |\gamma_{i_k}^{(t)} - \gamma_{i_k}^{(0)}|\right)^2 \leq \frac{1}{n^8}.$$

Together, we get

$$\clubsuit \leq c_{62} \cdot \frac{K_n}{n}.$$

Step 7: Bounding the second summand of Step 5. We show

$$\spadesuit \leq \exp\left(-\frac{(\log K_n)^2}{4}\right).$$

Step 8: Bounding the third summand of Step 5. We show

$$\heartsuit \leq c_{80} \cdot \frac{(\log n)^{4p}}{K_n^{2p}}.$$

Step 9: Bounding the fourth summand of Step 5. We show

$$\diamondsuit \leq c_{90} \cdot \frac{(\log n)^{4p}}{K_n^{2p}}.$$

Step 10: Finishing the proof. Taking the previous steps together yields

$$\begin{aligned}
& \mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
\leq\ & c_{30} \cdot \frac{(\log n)^3 \cdot K_n}{n} + c_{100} \cdot \left(\frac{(\log n)^3 \cdot K_n}{n} + \frac{(\log n)^{4p}}{K_n^{2p}}\right) + \frac{c_{43}}{n} \\
\leq\ & c_7 \cdot (\log n)^{\max\{3,4p\}} \cdot n^{-\frac{2p}{2p+1}}.
\end{aligned}$$

Now we give the detailed proof.

Proof. Step 1: Preparations. Since by assumption X is an $[0,1]-$valued random variable and m is $(p,C)-$smooth we can assume w.l.o.g. that m is bounded in absolute value. So, we assume

$$\|m\|_\infty \leq \beta_n.$$

Further, set

$$\tilde{K}_n = \left\lceil \frac{K_n}{(\log K_n)^2} \right\rceil.$$

Moreover, let A_n be the event such that

1. $|Y_i| \leq \beta_n$ holds for all $i = 1, \ldots, n$,

2. there exist $i_0, i_1, \ldots, i_{\tilde{K}_n} \in \{1, \ldots, K_n\}$ such that

$$\beta_{i_k} \geq \frac{B_n}{2}, \quad -\frac{\gamma_{i_k}}{\beta_{i_k}} \in I_k \quad \text{and} \quad \min_{i=1,\ldots,n} |\beta_{i_k} \cdot X_i + \gamma_{i_k}| \geq 2 \cdot (\log K_n)^2$$

holds for every $k \in \{1, \ldots, \tilde{K}_n\}$.

Step 2: Starting the proof. We have

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

In the following we bound each of the terms $T_{i,n} (i = 1, 2, 3)$ separately.

Step 3: Bounding the first summand. We notice that $T_{1,n}$ is of the same form as the left-hand side in Lemma 2.2.14 where A_n is the event defined in Step 1. The only remaining condition we need to check in order to apply Lemma 2.2.14 is that on the event A_n

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

for some function class $\mathcal{F}_n$ we need to specify. For this we use Lemma 2.2.6. Consequently, we must first verify conditions (2.11) and (2.12) of Lemma 2.2.6. First, we check (2.11), i.e.

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

for all $\mathbf{w} \in \mathbb{R}^{3K_n+1}$ with

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

We verify this condition by Lemma 2.2.9 where $\mathbf{v} = \mathbf{w}(0)$. For this, we notice that on A_n

$$\begin{aligned}
F(\mathbf{w}(0)) &= \frac{1}{n}\sum_{i=1}^{n} |Y_i - f_{net,(\alpha^{(0)},\beta^{(0)},\gamma^{(0)})}(X_i)|^2 + \frac{c_1}{K_n^{2p}} \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left(2 \cdot |Y_i|^2 + 2 \cdot |f_{net,(\alpha^{(0)},\beta^{(0)},\gamma^{(0)})}(X_i)|^2\right) + \frac{c_1}{K_n^{2p}} \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 \\
&\leq 2 \cdot \beta_n^2 + 2 \cdot \max_{i=1,\ldots,n} \left(\sum_{j=1}^{K_n} \alpha_j^{(0)} \cdot \sigma(\beta_j^{(0)} \cdot X_i + \gamma_j^{(0)}) + \alpha_0^{(0)}\right)^2 \\
&\quad + \frac{c_1}{K_n^{2p}} \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 \\
&\leq 2 \cdot \beta_n^2 + 2 \cdot \left(\sum_{k=0}^{K_n} \alpha_k^{(0)}\right)^2 + \frac{c_1}{K_n^{2p}} \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 \\
&\leq 2 \cdot \beta_n^2 + 2 \cdot K_n \cdot \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 + \frac{c_1}{K_n^{2p}} \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 \\
&\leq 2 \cdot \beta_n^2 + 2 \cdot K_n \cdot (K_n + 1) \cdot \left(\frac{c_2}{K_n}\right)^2 + \frac{c_1}{K_n^{2p}} \cdot (K_n + 1) \cdot \left(\frac{c_2}{K_n}\right)^2 \\
&\leq 2 \cdot \beta_n^2 + 4 \cdot K_n^2 \cdot \frac{c_2^2}{K_n^2} + 2 \cdot c_1 \cdot K_n^{1-2p} \cdot \frac{c_2^2}{K_n^2}
\end{aligned}$$

$$
\begin{aligned}
&\leq 2 \cdot \beta_n^2 + 4 \cdot c_2^2 + 2 \cdot c_1 \cdot c_2^2 \\
&\leq c_{31} \cdot (\log n)^2 \qquad (2.20)
\end{aligned}
$$

and

$$
\frac{t_n}{L_n} = \frac{\lceil K_n^{2p} \cdot (\log n)^2 \cdot L_n \rceil}{L_n} \geq K_n^{2p} \cdot (\log n)^2. \qquad (2.21)
$$

So, obviously,

$$
t_n \geq 2 \cdot L_n
$$

and

$$
c_1^2 \leq K_n.
$$

By (2.20) and (2.21) we have

$$
\begin{aligned}
|w_{1,k}^{(1)}| &\leq \|\alpha^{(0)}\| + \|\mathbf{w} - \mathbf{w}(0)\| \\
&\leq \frac{c_2}{K_n} + \sqrt{\frac{2t_n}{L_n} \cdot \max\{F(\mathbf{w}(0)), 1\}} \\
&\leq \frac{c_2}{K_n} + \sqrt{2 \cdot 2 \cdot K_n^{2p} \cdot (\log n)^2 \cdot c_{31} \cdot (\log n)^2} \\
&\leq c_{32} \cdot K_n^p \cdot (\log n)^2. \qquad (2.22)
\end{aligned}
$$

Then,

$$
\gamma_n = c_{32} \cdot K_n^p \cdot (\log n)^2 \geq 1.
$$

Hence, Lemma 2.2.9 gives us

$$
\begin{aligned}
\|(\nabla_{\mathbf{w}} F)(\mathbf{w})\| &\leq 25 \cdot (c_{32} \cdot K_n^p \cdot (\log n)^2)^2 \cdot K_n \cdot \sqrt{\frac{t_n}{L_n} \cdot \max\{F(\mathbf{w}(0)), 1\}} \\
&\leq c_{33} \cdot \sqrt{2} \cdot \frac{c_6 \cdot (\log n)^7 \cdot K_n^{3p+1}}{(\log n)^3 \cdot K_n^p} \cdot \sqrt{\frac{t_n}{L_n} \cdot \max\{F(\mathbf{w}(0)), 1\}} \\
&= \frac{c_{33}}{(\log n)^3 \cdot K_n^p} \cdot \sqrt{2} \cdot L_n \cdot \sqrt{\frac{t_n}{L_n} \cdot \max\{F(\mathbf{w}(0)), 1\}} \\
&\leq \sqrt{2 \cdot L_n^2 \cdot \frac{t_n}{L_n} \cdot \max\{F(\mathbf{w}(0)), 1\}} \\
&= \sqrt{2 \cdot L_n \cdot t_n \cdot \max\{F(\mathbf{w}(0)), 1\}}
\end{aligned}
$$

for all $\mathbf{w} \in \mathbb{R}^{3K_n+1}$ with

$$
\|\mathbf{w} - \mathbf{w}(0)\|^2 \leq \frac{2t_n}{L_n} \cdot \max\{F(\mathbf{w}(0)), 1\}.
$$

Second, we check (2.12), i.e.

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

for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^{3K_n+1}$ satisfying

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

and

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

We verify this condition by Lemma 2.2.10. By (2.22) we have for $k = 1, \ldots, K_n$ and for $i = 1, 2$

$$\begin{aligned} \max\left\{|(\mathbf{w}_i)^{(1)}_{1,k}|, |\mathbf{w}(0)^{(1)}_{1,k}|\right\} &\leq \max\left\{c_{32} \cdot K_n^p \cdot (\log n)^2, \frac{c_2}{K_n}\right\} \\ &\leq c_{32} \cdot K_n^p \cdot (\log n)^2 \\ &= \gamma_n. \end{aligned}$$

Hence, Lemma 2.2.10 with (2.20) and (2.21) gives us

$$\begin{aligned} & \|(\nabla_{\mathbf{w}} F)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F)(\mathbf{w}_2)\| \\ \leq\ & 165 \cdot \max\{\sqrt{F(\mathbf{w}(0))}, 1\} \cdot \max\{\frac{c_1}{K_n}, 1\} \\ & \cdot \left(c_{32} \cdot K_n^p \cdot (\log n)^2\right)^2 \cdot K_n \cdot \sqrt{\frac{t_n}{L_n}} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\| \\ \leq\ & 165 \cdot \sqrt{c_{31}} \cdot \log n \cdot \max\{\frac{c_1}{K_n}, 1\} \\ & \cdot c_{32}^2 \cdot K_n^{2p} \cdot (\log n)^4 \cdot K_n \cdot \sqrt{2 \cdot K_n^{2p} \cdot (\log n)^2} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\| \\ \leq\ & c_6 \cdot (\log n)^6 \cdot K_n^{3p+1} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\| \\ =\ & L_n \cdot \|\mathbf{w}_1 - \mathbf{w}_2\| \end{aligned} \tag{2.23}$$

for all $\mathbf{w}_i \in \mathbb{R}^{3K_n+1}$ $(i = 1, 2)$ satisfying

$$\|\mathbf{w}_i - \mathbf{w}(0)\|^2 \leq 8 \cdot \frac{t_n}{L_n} \cdot \max\{F(\mathbf{w}(0)), 1\}.$$

By application of Lemma 2.2.6 together with (2.20) and (2.21) we can now conlude that on A_n

$$\tilde{m}_n(x) = \sum_{k=0}^{K_n} \alpha_k^{(t_n)} \cdot \sigma(\beta_k^{(t_n)} \cdot x + \gamma_k^{(t_n)}) \in \mathcal{F}$$

where the function class $\mathcal{F}$ is defined as in Lemma 2.2.15 (as we can write $m_n(x) = \sum_{j=1}^{K_n} \alpha_j^{(t_n)} \cdot \sigma(\beta_j^{(t_n)} \cdot x + \gamma_j^{(t_n)}) + 2 \cdot \alpha_0^{(t_n)} \sigma(0 \cdot x + 0)$) with

$$\begin{aligned}
& \sum_{k=0}^{K_n} (\alpha_k^{(t_n)})^2 \\
\leq\ & \sum_{k=0}^{K_n} (\alpha_k^{(t_n)} - \alpha_k^{(0)} + \alpha_k^{(0)})^2 + 2 \cdot \left(\sum_{k=1}^{K_n} (\beta_k^{(t_n)} - \beta_k^{(0)})^2 + (\gamma_k^{(t_n)} - \gamma_k^{(0)})^2 \right) \\
\leq\ & \sum_{k=0}^{K_n} 2 \cdot (\alpha_k^{(t_n)} - \alpha_k^{(0)})^2 + 2 \cdot (\alpha_k^{(0)})^2 + 2 \cdot \left(\sum_{k=1}^{K_n} (\beta_k^{(t_n)} - \beta_k^{(0)})^2 + (\gamma_k^{(t_n)} - \gamma_k^{(0)})^2 \right) \\
=\ & 2 \cdot \left(\sum_{k=0}^{K_n} (\alpha_k^{(t_n)} - \alpha_k^{(0)})^2 + \left(\sum_{k=1}^{K_n} (\beta_k^{(t_n)} - \beta_k^{(0)})^2 + (\gamma_k^{(t_n)} - \gamma_k^{(0)})^2 \right) \right) + 2 \cdot \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 \\
=\ & 2 \cdot \|\mathbf{w}(t_n) - \mathbf{w}(0)\|^2 + 2 \cdot \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 \\
\leq\ & 2 \cdot \left(\sqrt{2 \cdot \frac{t_n}{L_n} \cdot (F(\mathbf{w}(0)) - F(\mathbf{w}(t_n)))} \right)^2 + 2 \cdot \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 \\
\leq\ & 2 \cdot 2 \cdot \frac{t_n}{L_n} \cdot F(\mathbf{w}(0)) + 2 \cdot \sum_{k=0}^{K_n} (\frac{c_2}{K_n})^2 \\
\leq\ & 2 \cdot 2 \cdot 2 \cdot K_n^{2p} \cdot (\log n)^2 \cdot c_{31} \cdot (\log n)^2 + 2 \cdot (K_n + 1) \frac{c_2^2}{K_n^2} \\
\leq\ & 8 \cdot c_{31} \cdot K_n^{2p} \cdot (\log n)^4 + 4 \cdot \frac{c_2^2}{K_n} \\
\leq\ & c_{34} \cdot K_n^{2p} \cdot (\log n)^4.
\end{aligned}$$

As a result, application of Lemma 2.2.15 yields

$$\log \left(\mathcal{N}_1 \left(\frac{1}{n \cdot \beta_n}, \mathcal{F}, x_1^n \right) \right) \leq c_9 \cdot \log n \cdot K_n$$

and consequently by Lemma 2.2.14

$$\begin{aligned}
T_{1,n} &\leq \frac{c_7 \cdot (\log n)^2 \cdot \left(\log\left(\sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{n\cdot\beta_n}, \mathcal{F}_n, x_1^n\right)\right) + 1\right)}{n} \\
&\leq \frac{c_7 \cdot (\log n)^2 \cdot (c_9 \cdot \log n \cdot K_n + 1)}{n} \\
&\leq c_{30} \cdot \frac{(\log n)^3 \cdot K_n}{n}.
\end{aligned}$$

Step 4: Bounding the third summand. By Lemma 2.2.11, by Markov's inequality and by (2.7) we get

$$\begin{aligned}
& T_{3,n} \\
\leq\ & 4\beta_n^2 \cdot \Bigg(\mathbf{P}\Bigg\{\exists l \in \{0,\dots,\tilde{K}\}\, \forall k \in \{1,\dots,K\} : \\
& \qquad \beta_k < \frac{B_n}{2}, \quad -\frac{\gamma_k}{\beta_k} \notin I_l \quad \text{or} \quad \min_{i=1,\dots,n} |\beta_k \cdot x_i + \gamma_k| < 2 \cdot (\log K)^2\Bigg\} \\
& + \mathbf{P}\left\{|Y_i| > \beta_n \text{ for some } i \in \{1,\dots,n\}\right\}\Bigg) \\
\leq\ & 4\beta_n^2 \cdot \left((\tilde{K}_n + 1) \cdot \exp\left(-\frac{K_n}{32 \cdot \tilde{K}_n}\right) + n \cdot \mathbf{P}\left\{\exp(c_{10} \cdot Y^2) > \exp(c_{10} \cdot \beta_n^2)\right\}\right) \\
\leq\ & 4 \cdot (c_3 \cdot \log n)^2 \cdot \left((\tilde{K}_n + 1) \cdot \exp\left(-\frac{K_n}{32 \cdot \tilde{K}_n}\right) + n \cdot \frac{\mathbf{E}\left\{\exp(c_{10} \cdot Y^2)\right\}}{\exp\left(c_{10} \cdot (c_3 \cdot \log n)^2\right)}\right) \\
\leq\ & 4 \cdot c_3^2 \cdot (\log n)^2 \cdot 4 \cdot \frac{K_n}{(\log K_n)^2} \cdot \exp\left(-\frac{(\log K_n)^2}{64}\right) + 4 \cdot c_3^2 \cdot (\log n)^2 \cdot n \cdot \frac{c_{40}}{n^{c_{41}\cdot \log n}} \\
\leq\ & 16 \cdot c_3^2 \cdot (\log n)^2 \cdot \frac{K_n}{(c_{42} \cdot \log n)^2} \cdot \frac{1}{K_n^{\frac{\log K_n}{64}}} + 4 \cdot c_3^2 \cdot (\log n)^2 \cdot n \cdot \frac{c_{40}}{n^{c_{41}\cdot \log n}} \\
\leq\ & c_{43} \cdot \frac{1}{n}.
\end{aligned}$$

Step 5: Looking at the second summand. We break down $T_{2,n}$ into a sum of four terms. For that we need to introduce new functions. Let $i_0, i_1, \dots, i_{\tilde{K}_n}$ be the indices as in the

definition of the event A_n and define on $[0,1]$ a piecewise constant approximation of m by

$$f(x) = \sum_{k=1}^{\tilde{K}_n} \alpha^*_{i_k} \cdot \mathbf{1}_{\left[-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}},\infty\right)}(x) + \alpha^*_{i_0}$$

where

$$\alpha^*_{i_0} := m\left(-\frac{\gamma^{(0)}_{i_0}}{\beta^{(0)}_{i_0}}\right)$$

and

$$\alpha^*_{i_j} = m\left(-\frac{\gamma^{(0)}_{i_j}}{\beta^{(0)}_{i_j}}\right) - m\left(-\frac{\gamma^{(0)}_{i_j-1}}{\beta^{(0)}_{i_j-1}}\right) \quad (j = 1, \ldots, \tilde{K}_n).$$

We set

$$f^*(x) = \sum_{k=1}^{\tilde{K}_n} \alpha^*_{i_k} \cdot \sigma(\beta^{(0)}_{i_k} \cdot x + \gamma^{(0)}_{i_k}) + \alpha^*_{i_0}.$$

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

$$\text{pen}(g) = \frac{c_1}{K_n^{2p}} \cdot \sum_{k=1}^{K_n} \alpha_k^2.$$

We have

$$\begin{aligned}
& T_{2,n} \\
= \; & \mathbf{E}\left(\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{\{|Y_j| \le \beta_n \, (j \in \{1,\ldots,n\})\}} - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \cdot \mathbf{1}_{A_n}\right) \\
\le \; & \mathbf{E}\left(\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 + \frac{c_1}{n} \cdot \sum_{k=1}^{K_n} (\alpha_k^{(t_n)})^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \cdot \mathbf{1}_{A_n}\right) \\
= \; & \mathbf{E}\Bigg(\Bigg(F(\alpha^{(t_n)}, \beta^{(t_n)}, \gamma^{(t_n)}) - \frac{1}{n}\sum_{i=1}^n |Y_i - f^*(X_i)|^2 - \text{pen}(f^*) + \text{pen}(f^*) \\
& + \frac{1}{n}\sum_{i=1}^n |Y_i - f^*(X_i)|^2 - \frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|^2 \\
& + \frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\Bigg) \cdot \mathbf{1}_{A_n}\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}\Bigg(\Bigg(F(\alpha^{(t_n)},\beta^{(t_n)},\gamma^{(t_n)}) - \frac{1}{n}\sum_{i=1}^{n}|Y_i - f^*(X_i)|^2 - \text{pen}(f^*)\Bigg)\cdot \mathbf{1}_{A_n}\Bigg)\\
&\quad +\mathbf{E}\Bigg(\Bigg(\frac{1}{n}\sum_{i=1}^{n}(|Y_i - f^*(X_i)|^2 - \frac{1}{n}\sum_{i=1}^{n}|Y_i - f(X_i)|^2\Bigg)\cdot \mathbf{1}_{A_n}\Bigg)\\
&\quad +\mathbf{E}\Bigg(\Bigg(\frac{1}{n}\sum_{i=1}^{n}|Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^{n}|m(X_i) - Y_i|^2\Bigg)\cdot \mathbf{1}_{A_n}\Bigg)\\
&\quad +\mathbf{E}\Bigg(\text{pen}(f^*)\cdot \mathbf{1}_{A_n}\Bigg)\\
&=: T_{5,n} + T_{6,n} + T_{7,n} + T_{8,n}.
\end{aligned}
$$

In the remaining steps of the proof we will bound the terms $T_{i,n}(i = 5, 6, 7, 8)$ separately.

Step 6: Bounding the first summand of Step 5. All of the following considerations happen on A_n.

We bound $T_{5,n}$ in several steps. First, we will derive a recursive formula consisting of two terms where the first term will be the recursive term. Second, we reduce the second term to a difference of the inner weights at indices $i_0, i_1, \ldots, i_{\tilde{K}_n}$. Third, analysis of gradient descent will show that at those indices gradient descent affects these inner weights only slighty.

Let $\alpha^* \in \mathbb{R}^{K_n+1}$ with entries

$$\alpha^*_{i_k} \quad \text{for} \quad k \in \{0, 1, \ldots, \tilde{K}_n\}$$

and

$$\alpha^*_k = 0 \text{ for} \quad k \notin \{i_0, i_1, \ldots, i_{\tilde{K}_n}\}.$$

Then

$$T_{5,n} = \mathbf{E}\Bigg(\Bigg(F(\alpha^{(t+1)},\beta^{(t+1)},\gamma^{(t+1)}) - F(\alpha^*,\beta^{(0)},\gamma^{(0)})\Bigg)\cdot \mathbf{1}_{A_n}\Bigg).$$

By (2.23) we can apply Lemma 2.2.5 which gives us

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

Since

$$\|\nabla_{(\alpha,\beta,\gamma)}F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})\|^2$$

$$
\begin{aligned}
&= \sum_{k=1}^{K_n} \left| \frac{\partial}{\partial \alpha_k} F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) \right|^2 + \sum_{k=1}^{K_n} \left| \frac{\partial}{\partial \beta_k} F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) \right|^2 \\
&\quad + \sum_{k=1}^{K_n} \left| \frac{\partial}{\partial \gamma_k} F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) \right|^2 \\
&\geq \sum_{k=1}^{K_n} \left| \frac{\partial}{\partial \alpha_k} F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) \right|^2 \\
&= \|(\nabla_\alpha F)(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)})\|^2
\end{aligned}
$$

we get

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

we apply Lemma 2.2.12 (with $\tau_n = c_1/K_n^{2p}$) which yields

$$
\begin{aligned}
& \frac{1}{2L_n} \cdot \|(\nabla_\alpha F)(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)})\|^2 \\
\geq\ & \frac{1}{2L_n} \cdot \frac{4 \cdot c_1}{K_n^{2p}} \cdot (F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) - \min_{\mathbf{a} \in \mathbb{R}^{K_n+1}} F(\mathbf{a}, \beta^{(t)}, \gamma^{(t)})) \\
\geq\ & \frac{1}{2L_n} \cdot \frac{4 \cdot c_1}{K_n^{2p}} \cdot (F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) - F(\alpha^*, \beta^{(t)}, \gamma^{(t)})) \\
=\ & \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot (F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) - F(\alpha^*, \beta^{(0)}, \gamma^{(0)}) \\
& \qquad + F(\alpha^*, \beta^{(0)}, \gamma^{(0)}) - F(\alpha^*, \beta^{(t)}, \gamma^{(t)})) \\
=\ & \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot (F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) - F(\alpha^*, \beta^{(0)}, \gamma^{(0)})) \\
& + \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot (F(\alpha^*, \beta^{(0)}, \gamma^{(0)}) - F(\alpha^*, \beta^{(t)}, \gamma^{(t)})).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& F(\alpha^{(t+1)}, \beta^{(t+1)}, \gamma^{(t+1)}) - F(\alpha^*, \beta^{(0)}, \gamma^{(0)}) \\
\leq\ & F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) - \frac{1}{2L_n} \cdot \|(\nabla_\alpha F)(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)})\|^2 - F(\alpha^*, \beta^{(0)}, \gamma^{(0)}) \\
\leq\ & F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) - \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot (F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) - F(\alpha^*, \beta^{(0)}, \gamma^{(0)}))
\end{aligned}
$$

$$
\begin{aligned}
& -\frac{2\cdot c_1}{L_n\cdot K_n^{2p}}\cdot(F(\alpha^*,\beta^{(0)},\gamma^{(0)})-F(\alpha^*,\beta^{(t)},\gamma^{(t)})) \\
& -F(\alpha^*,\beta^{(0)},\gamma^{(0)}) \\
= & \left(1-\frac{2\cdot c_1}{L_n\cdot K_n^{2p}}\right)\cdot(F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})-F(\alpha^*,\beta^{(0)},\gamma^{(0)})) \\
& +\frac{2\cdot c_1}{L_n\cdot K_n^{2p}}\cdot(F(\alpha^*,\beta^{(t)},\gamma^{(t)})-F(\alpha^*,\beta^{(0)},\gamma^{(0)})) \\
= & \left(1-\frac{2\cdot c_1}{L_n\cdot K_n^{2p}}\right)\cdot(F(\alpha^{(t)},\beta^{(t)},\gamma^{(t)})-\frac{1}{n}\sum_{i=1}^{n}|Y_i-f^*(X_i)|^2-\text{pen}(f^*)) \\
& +\frac{2\cdot c_1}{L_n\cdot K_n^{2p}}\cdot(F(\alpha^*,\beta^{(t)},\gamma^{(t)})-F(\alpha^*,\beta^{(0)},\gamma^{(0)})).
\end{aligned}
$$

We observe that the first term on the right-hand side is the same term we started out with (up to a factor smaller than one) but one $t-$step down. Next, we have by the Cauchy-Schwarz inequality

$$
\begin{aligned}
& F(\alpha^*,\beta^{(t)},\gamma^{(t)})-F(\alpha^*,\beta^{(0)},\gamma^{(0)}) \\
= & \frac{1}{n}\sum_{i=1}^{n}\Bigg(|f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-y_i|^2+\frac{c_1}{n}\sum_{k=0}^{K}(\alpha_k^*)^2 \\
& \qquad -|f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i)-y_i|^2-\frac{c_1}{n}\sum_{k=0}^{K}(\alpha_k^*)^2\Bigg) \\
= & \frac{1}{n}\sum_{i=1}^{n}\left(|f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-y_i|^2-|f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i)-y_i|^2\right) \\
= & \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-y_i+(f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i)-y_i)) \\
& \qquad \cdot(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-y_i-(f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i)-y_i)) \\
= & \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)+f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i)-2y_i) \\
& \qquad \cdot(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i)) \\
= & \frac{1}{n}\sum_{i=1}^{n}(2f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i)-2y_i)\cdot(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i))
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i))^2\\
\leq\ & \sqrt{\frac{1}{n}\sum_{i=1}^{n}(2\cdot f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i)-2\cdot y_i)^2}\\
&\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i))^2}\\
&+\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i))^2\\
\leq\ & 2\cdot\sqrt{F(\alpha^*,\beta^{(0)},\gamma^{(0)})}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i))^2}\\
&+\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i))^2.
\end{aligned}
$$

We bound the term

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

further. For that we notice that by definition of the outer weights α^* both neural networks in the term are, in fact, neural networks with $\tilde{K}_n$ many hidden neurons. Applying Lemma 2.2.7 (with $\tilde{K}_n$) with Lipschitz continuity of σ (with Lipschitz constant 1) yields

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i)-f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i))^2\\
\leq\ & \Bigg(1\cdot 1\cdot\max_{k=1,\dots,\tilde{K}_n}\max\{|\alpha^*_{i_k}|,1\}\\
&\cdot\left(\sum_{j=0}^{\tilde{K}_n}|\alpha^*_{i_k}-\alpha^*_{i_k}|+\sum_{k=1}^{\tilde{K}_n}|\beta^{(t)}_{i_k}-\beta^{(0)}_{i_k}|+|\gamma^{(t)}_{i_k}-\gamma^{(0)}_{i_k}|\right)\Bigg)^2\\
=\ & \max_{k=1,\dots,\tilde{K}_n}\max\{|\alpha^*_{i_k}|^2,1\}\cdot\left(\sum_{k=1}^{\tilde{K}_n}|\beta^{(t)}_{i_k}-\beta^{(0)}_{i_k}|+|\gamma^{(t)}_{i_k}-\gamma^{(0)}_{i_k}|\right)^2.
\end{aligned}
$$

We will bound

$$\left(\sum_{k=1}^{\tilde{K}_n} |\beta_{i_k}^{(t)} - \beta_{i_k}^{(0)}| + |\gamma_{i_k}^{(t)} - \gamma_{i_k}^{(0)}|\right)^2$$

further. For that, we will show that application of gradient descent to $(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)})$ changes the inner weights $\beta^{(0)}, \gamma^{(0)}$ only slightly. In order to do so, we set

$$\frac{\delta_n}{2} = 2 \cdot (\log K_n)^2$$

and we show the following claim consisting of two inequalities by induction on s:

$$1. \quad \textstyle\sum_{k=1}^{\tilde{K}_n}(|\beta_{i_k}^{(s)} - \beta_{i_k}^{(0)}| + |\gamma_{i_k}^{(s)} - \gamma_{i_k}^{(0)}|) \quad \leq s \cdot c_{63} \cdot \frac{t_n}{L_n} \cdot \exp(-\delta_n/2) \tag{2.24}$$

$$2. \quad \min_{i=1,\ldots,n,k=0,\ldots,\tilde{K}} \left|\beta_{i_k}^{(s)} \cdot X_i + \gamma_{i_k}^{(s)}\right| \quad \geq \frac{\delta_n}{2} \tag{2.25}$$

for some constant $c_{63} > 0$ and for all $s \in \{0, \ldots, t_n\}$.
Start of the induction. For $s = 0$ we have:

1. For the first inequality, it trivially holds that

$$\sum_{k=1}^{\tilde{K}_n}(|\beta_{i_k}^{(0)} - \beta_{i_k}^{(0)}| + |\gamma_{i_k}^{(0)} - \gamma_{i_k}^{(0)}|) = 0 \leq s \cdot c_{63} \cdot \frac{t_n}{L_n} \cdot \exp(-\delta_n/2).$$

2. For the second inequality

$$\min_{i=1,\ldots,n,k=0,\ldots,\tilde{K}} \left|\beta_{i_k}^{(0)} \cdot X_i + \gamma_{i_k}^{(0)}\right| \geq \delta_n \geq \frac{\delta_n}{2}$$

holds by definition of A_n.
Induction hypothesis. Assume that (2.24) and (2.25) hold for some $s \in \{1, \ldots, t_n - 1\}$.
Induction step. We have:

1. Firstly, we observe that by the induction hypothesis

$$\begin{aligned} & \sum_{k=1}^{\tilde{K}_n}(|\beta_{i_k}^{(s+1)} - \beta_{i_k}^{(0)}| + |\gamma_{i_k}^{(s+1)} - \gamma_{i_k}^{(0)}|) \\ = & \sum_{k=1}^{\tilde{K}_n}(|\beta_{i_k}^{(s+1)} - \beta_{i_k}^{(s)} + \beta_{i_k}^{(s)} - \beta_{i_k}^{(0)}| + |\gamma_{i_k}^{(s+1)} - \gamma_{i_k}^{(s)} + \gamma_{i_k}^{(s)} - \gamma_{i_k}^{(0)}|) \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{\tilde{K}_n}(|\beta_{i_k}^{(s)}-\beta_{i_k}^{(0)}|+|\gamma_{i_k}^{(s)}-\gamma_{i_k}^{(0)}|)+\sum_{k=1}^{\tilde{K}_n}(|\beta_{i_k}^{(s+1)}-\beta_{i_k}^{(s)}|+|\gamma_{i_k}^{(s+1)}-\gamma_{i_k}^{(s)}|) \\
&\leq s\cdot c_{63}\cdot\frac{t_n}{L_n}\cdot\exp(-\delta_n/2)+\sum_{k=1}^{\tilde{K}_n}(|\beta_{i_k}^{(s+1)}-\beta_{i_k}^{(s)}|+|\gamma_{i_k}^{(s+1)}-\gamma_{i_k}^{(s)}|). \qquad (2.26)
\end{aligned}
$$

Secondly, we look at

$$
\sum_{k=1}^{\tilde{K}_n}(|\beta_{i_k}^{(s+1)}-\beta_{i_k}^{(s)}|+|\gamma_{i_k}^{(s+1)}-\gamma_{i_k}^{(s)}|).
$$

For that, we set

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

Using

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

(where the first inequality holds due to $\sigma(x)\in[0,1]$) we can conclude from Lemma 2.2.13 that

$$
\begin{aligned}
\max_{i=1,\ldots,n}\left|\sigma'\left(\beta_{i_k}\cdot X_i+\gamma_{i_k}\right)\right| &\leq \max_{i=1,\ldots,n}\exp\left(-\left|\beta_{i_k}\cdot X_i+\gamma_{i_k}\right|\right) \\
&= \exp\left(-\min_{i=1,\ldots,n}\left\{\left|\beta_{i_k}\cdot X_i+\gamma_{i_k}\right|\right\}\right). \qquad (2.27)
\end{aligned}
$$

We start with

$$
|\bar{\beta}_{i_k}-\beta_{i_k}|.
$$

As a consequence from (2.27), we get by the Cauchy-Schwarz inequality for $k\in\{1,\ldots,\tilde{K}_n\}$

$$
\begin{aligned}
&\left|\frac{\partial F}{\partial\beta_{i_k}}(\alpha,\beta,\gamma)\right| \\
=&\left|\frac{2}{n}\sum_{i=1}^{n}(f_{net,(\alpha,\beta,\gamma)}(X_i)-Y_i)\cdot\alpha_{i_k}\cdot\sigma'\left(\beta_{i_k}\cdot X_i+\gamma_{i_k}\right)\cdot X_i\right| \\
\leq\ & 2\cdot|\alpha_{i_k}|\cdot\frac{1}{n}\sum_{i=1}^{n}|f_{net,(\alpha,\beta,\gamma)}(X_i)-Y_i|\cdot|X_i|\cdot\left|\sigma'\left(\beta_{i_k}\cdot X_i+\gamma_{i_k}\right)\right| \\
\leq\ & 2\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}|f_{net,(\alpha,\beta,\gamma)}(X_i)-Y_i|^2\cdot(X_i)^2}\cdot|\alpha_{i_k}|\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\sigma'(\beta_{i_k}\cdot X_i+\gamma_{i_k})|^2}
\end{aligned}
$$

$$
\begin{aligned}
&\leq 2 \cdot \sqrt{F(\alpha,\beta,\gamma)} \cdot \max_{i=1,\dots,n} \{|X_i|\} \cdot |\alpha_{i_k}| \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\sigma'(\beta_{i_k} \cdot X_i + \gamma_{i_k})|^2} \\
&\leq 2 \cdot \sqrt{F(\alpha,\beta,\gamma)} \cdot \max_{i=1,\dots,n} \{|X_i|\} \cdot |\alpha_{i_k}| \cdot \exp\left(-\min_{i=1,\dots,n} \{|\beta_{i_k} \cdot X_i + \gamma_{i_k}|\}\right).
\end{aligned}
$$

Hence, we have shown

$$
\begin{aligned}
&|\bar{\beta}_{i_k} - \beta_{i_k}| \\
&= \lambda_n \cdot \left|\frac{\partial F}{\partial \beta_{i_k}}((\alpha,\beta,\gamma))\right| \\
&\leq \lambda_n \cdot 2 \cdot \sqrt{F((\alpha,\beta,\gamma))} \cdot \max_{i=1,\dots,n} \{|X_i|\} \cdot |\alpha_{i_k}| \\
&\qquad \cdot \exp\left(-\min_{i=1,\dots,n} \{|\beta_{i_k} \cdot X_i + \gamma_{i_k}|\}\right)
\end{aligned} \tag{2.28}
$$

for any $k \in \{1,\dots,\tilde{K}\}$. Next, we consider

$$
|\bar{\gamma}_{i_k} - \gamma_{i_k}|.
$$

Analogously to our previous consideration, we get by (2.27) and by the Cauchy-Schwarz inequality for $k \in \{1,\dots,\tilde{K}_n\}$

$$
\begin{aligned}
&\left|\frac{\partial F}{\partial \gamma_{i_k}}(\alpha,\beta,\gamma)\right| \\
&= \left|\frac{2}{n}\sum_{i=1}^{n} (f_{net,(\alpha,\beta,\gamma)}(X_i) - Y_i) \cdot \alpha_{i_k} \cdot \sigma'\left(\beta_{i_k} \cdot X_i + \gamma_{i_k}\right) \cdot 1\right| \\
&\leq 2 \cdot |a_{i_k}| \cdot \frac{1}{n}\sum_{i=1}^{n} |f_{net,(\alpha,\beta,\gamma)}(X_i) - Y_i| \cdot \left|\sigma'\left(\beta_{i_k} \cdot X_i + \gamma_{i_k}\right)\right| \\
&\leq 2 \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |f_{net,(\alpha,\beta,\gamma)}(X_i) - Y_i|^2} \cdot |a_{i_k}| \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\sigma'(\beta_{i_k} \cdot X_i + \gamma_{i_k})|^2} \\
&\leq 2 \cdot \sqrt{F(\alpha,\beta,\gamma)} \cdot |a_{i_k}| \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\sigma'(\beta_{i_k} \cdot X_i + \gamma_{i_k})|^2} \\
&\leq 2 \cdot \sqrt{F(\alpha,\beta,\gamma)} \cdot |a_{i_k}| \cdot \exp\left(-\min_{i=1,\dots,n} \{|\beta_{i_k} \cdot X_i + \gamma_{i_k}|\}\right).
\end{aligned}
$$

Hence, we have shown

$$\begin{aligned}
& |\bar{\gamma}_{i_k} - \gamma_{i_k}| \\
= \; & \lambda_n \cdot \left| \frac{\partial F}{\partial \gamma_{i_k}}(\alpha, \beta, \gamma) \right| \\
\leq \; & \lambda_n \cdot 2 \cdot \sqrt{F(\alpha, \beta, \gamma)} \cdot 1 \cdot |\alpha_{i_k}| \cdot \exp\left(- \min_{i=1,\ldots,n} \{|\beta_{i_k} \cdot X_i + \gamma_{i_k}|\} \right). \qquad (2.29)
\end{aligned}$$

for any $k \in \{1, \ldots, \tilde{K}\}$. Using (2.28) and (2.29) and the induction hypothesis we get

$$\begin{aligned}
& \sum_{k=1}^{\tilde{K}_n} (|\beta_{i_k}^{(s+1)} - \beta_{i_k}^{(s)}| + |\gamma_{i_k}^{(s+1)} - \gamma_{i_k}^{(s)}|) \\
\leq \; & \sum_{k=1}^{\tilde{K}_n} \Bigg(\lambda_n \cdot 2 \cdot \sqrt{F((\alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}))} \cdot \max_{i=1,\ldots,n} \{|X_i|\} \cdot |\alpha_{i_k}^{(s)}| \\
& \qquad \cdot \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \beta_{i_k}^{(s)} \cdot X_i + \gamma_{i_k}^{(s)} \right| \right\} \right) \\
& \qquad + \lambda_n \cdot 2 \cdot \sqrt{F((\alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}))} \cdot 1 \cdot |\alpha_{i_k}^{(s)}| \\
& \qquad \cdot \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \beta_{i_k}^{(s)} \cdot X_i + \gamma_{i_k}^{(s)} \right| \right\} \right) \Bigg) \\
\leq \; & \sum_{k=1}^{\tilde{K}_n} 4 \cdot \lambda_n \cdot \sqrt{F((\alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}))} \cdot |\alpha_{i_k}^{(s)}| \cdot \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \beta_{i_k}^{(s)} \cdot X_i + \gamma_{i_k}^{(s)} \right| \right\} \right) \\
\leq \; & \sum_{k=1}^{\tilde{K}_n} 4 \cdot \lambda_n \cdot \sqrt{F((\alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}))} \cdot |\alpha_{i_k}^{(s)}| \cdot \exp\left(- \frac{\delta_n}{2} \right) \\
= \; & 4 \cdot \lambda_n \cdot \sqrt{F((\alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}))} \cdot \|\alpha^{(s)}\|_1 \cdot \exp\left(- \frac{\delta_n}{2} \right).
\end{aligned}$$

Since

$$\begin{aligned}
& F(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}) \\
= \; & \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{net,(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)})}(X_i)|^2 + \frac{c_1}{K_n^{2p}} \cdot \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 \\
\leq \; & \frac{1}{n} \sum_{i=1}^{n} (2 \cdot |Y_i|^2 + 2 \cdot |f_{net,(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)})}(X_i)|^2) + \frac{c_1}{K_n^{2p}} \cdot \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2
\end{aligned}$$

$$
\begin{aligned}
&\leq 2 \cdot \beta_n^2 + 2 \cdot \left(\sum_{k=0}^{K_n} \alpha_k^{(0)} \right)^2 + \frac{c_1}{K_n^{2p}} \cdot \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 \\
&\leq 2 \cdot \beta_n^2 + 2 \cdot (K_n + 1) \cdot \sum_{k=0}^{K_n} (\alpha_k^{(0)})^2 + \frac{c_1}{K_n^{2p}} \cdot (K_n + 1) \cdot \left(\frac{c_2}{K_n} \right)^2 \\
&\leq 2 \cdot (c_3 \cdot \log n)^2 + 2 \cdot (K_n + 1)^2 \cdot \left(\frac{c_2}{K_n} \right)^2 + \frac{c_1}{K_n^{2p}} \cdot (K_n + 1) \cdot \left(\frac{c_2}{K_n} \right)^2 \\
&\leq 2 \cdot (c_3 \cdot \log n)^2 + 8 \cdot K_n^2 \cdot \frac{c_2^2}{K_n^2} + 2 \cdot \frac{c_1}{K_n^{2p}} \cdot K_n \cdot \frac{c_2^2}{K_n^2} \\
&\leq 2 \cdot c_3^2 \cdot (\log n)^2 + 8 \cdot c_2^2 + 2 \cdot \frac{c_1 \cdot c_2^2}{K_n^{2p+1}} \\
&\leq c_{65} \cdot (\log n)^2 \qquad (2.30)
\end{aligned}
$$

we conclude by Lemma 2.2.6

$$
F(\alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}) \leq F(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}) \leq c_{65} \cdot (\log n)^2
$$

and

$$
\begin{aligned}
& \|\alpha^{(s)}\|_1 \\
\leq\ & \sqrt{K_n + 1} \cdot \|\alpha^{(s)}\| \\
\leq\ & \sqrt{K_n + 1} \cdot (\|\alpha^{(s)} - \alpha^{(0)}\| + \|\alpha^{(0)}\|) \\
\leq\ & \sqrt{K_n + 1} \cdot (\|(\alpha^{(s)} - \alpha^{(0)}, \beta^{(s)} - \beta^{(0)}, \gamma^{(s)} - \gamma^{(0)})\| + \|\alpha^{(0)}\|) \\
=\ & \sqrt{K_n + 1} \cdot (\|(\alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}) - (\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)})\| + \|\alpha^{(0)}\|) \\
\leq\ & \sqrt{K_n + 1} \cdot \left(\sqrt{2 \cdot \frac{s}{L_n} \cdot (F(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}) - F(\alpha^{(s)}, \beta^{(s)}, \gamma^{(s)}))} + \|\alpha^{(0)}\| \right) \\
\leq\ & \sqrt{K_n + 1} \cdot \left(\sqrt{2 \cdot \frac{s}{L_n} \cdot F(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)})} + \|\alpha^{(0)}\| \right) \\
\leq\ & \sqrt{K_n + 1} \cdot \left(\sqrt{2 \cdot \frac{s}{L_n} \cdot F(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)})} + \sqrt{\sum_{k=0}^{K_n} (\alpha_k^{(0)})^2} \right) \\
\leq\ & \sqrt{K_n + 1} \cdot \left(\sqrt{2 \cdot \frac{s}{L_n} \cdot c_{65} \cdot (\log n)^2} + \sqrt{(K_n + 1) \cdot \frac{c_2^2}{K_n^2}} \right) \\
=\ & \sqrt{4 \cdot K_n \cdot \frac{s}{L_n} \cdot c_{65} \cdot (\log n)^2} + \sqrt{4 \cdot c_2^2}
\end{aligned}
$$

$$\leq \quad c_{64} \cdot \sqrt{K_n \cdot \frac{t_n}{L_n}} \cdot \log n.$$

As a result, we get

$$\begin{aligned}
& \sum_{k=1}^{\tilde{K}_n} (|\beta_{i_k}^{(s+1)} - \beta_{i_k}^{(s)}| + |\gamma_{i_k}^{(s+1)} - \gamma_{i_k}^{(s)}|) \\
\leq \quad & 4 \cdot \lambda_n \cdot \sqrt{c_{65} \cdot (\log n)^2} \cdot c_{64} \cdot \sqrt{K_n \cdot \frac{t_n}{L_n}} \cdot \log n \cdot \exp\left(-\frac{\delta_n}{2}\right) \\
= \quad & 4 \cdot \sqrt{c_{65}} \cdot c_{64} \cdot \frac{1}{L_n} \cdot (\log n)^2 \cdot \sqrt{K_n} \cdot \sqrt{\frac{t_n}{L_n}} \cdot \exp\left(-\frac{\delta_n}{2}\right) \\
\leq \quad & 4 \cdot \sqrt{c_{65}} \cdot c_{64} \cdot \frac{t_n}{L_n} \cdot \sqrt{\frac{K_n \cdot (\log n)^4}{t_n \cdot L_n}} \cdot \exp\left(-\frac{\delta_n}{2}\right) \\
\leq \quad & c_{63} \cdot \frac{t_n}{L_n} \cdot 1 \cdot \exp\left(-\frac{\delta_n}{2}\right).
\end{aligned}$$

Hence, with (2.26) we have

$$\begin{aligned}
& \sum_{k=1}^{\tilde{K}_n} (|\beta_{i_k}^{(s+1)} - \beta_{i_k}^{(0)}| + |\gamma_{i_k}^{(s+1)} - \gamma_{i_k}^{(0)}|) \\
\leq \quad & s \cdot c_{63} \cdot \frac{t_n}{L_n} \cdot \exp(-\delta_n/2) + c_{63} \cdot \frac{t_n}{L_n} \cdot 1 \cdot \exp\left(-\frac{\delta_n}{2}\right) \\
= \quad & (s+1) \cdot c_{63} \cdot \frac{t_n}{L_n} \cdot \exp(-\delta_n/2). \qquad (2.31)
\end{aligned}$$

2. From the induction hypothesis ($s = 0$) and from (2.31) we conclude

$$\begin{aligned}
& \min_{i=1,\ldots,n, k=0,\ldots,\tilde{K}} \left| \beta_{i_k}^{(s+1)} \cdot X_i + \gamma_{i_k}^{(s+1)} \right| \\
= \quad & \min_{i=1,\ldots,n, k=0,\ldots,\tilde{K}} \left| (\beta_{i_k}^{(s+1)} - \beta_{i_k}^{(0)}) \cdot X_i + \beta_{i_k}^{(0)} \cdot X_i + (\gamma_{i_k}^{(s+1)} - \gamma_{i_k}^{(0)}) + \gamma_{i_k}^{(0)} \right| \\
\geq \quad & \min_{i=1,\ldots,n, k=0,\ldots,\tilde{K}} \left| \beta_{i_k}^{(0)} \cdot X_i + \gamma_{i_k}^{(0)} \right| \\
& - \max_{i=1,\ldots,n, k=0,\ldots,\tilde{K}} \left(|\beta_{i_k}^{(s+1)} - \beta_{i_k}^{(0)}| \cdot |X_i| + |\gamma_{i_k}^{(s+1)} - \gamma_{i_k}^{(0)}| \right)
\end{aligned}$$

$$
\begin{aligned}
&\geq \delta_n - \max_{i=1,\ldots,n,k=0,\ldots,\tilde{K}} \left(|\beta_{i_k}^{(s+1)} - \beta_{i_k}^{(0)}| + |\gamma_{i_k}^{(s+1)} - \gamma_{i_k}^{(0)}|\right) \cdot \max\{1, \max_{i=1,\ldots,n}\{|X_i|\}\} \\
&\geq \delta_n - t_n \cdot c_{63} \cdot \frac{t_n}{L_n} \cdot \exp\left(-\delta_n/2\right) \\
&\geq \frac{\delta_n}{2}.
\end{aligned}
$$

This concludes the proof of the claim. From (2.24) we immediately get

$$
\begin{aligned}
\left(\sum_{k=1}^{\tilde{K}_n} |\beta_{i_k}^{(t)} - \beta_{i_k}^{(0)}| + |\gamma_{i_k}^{(t)} - \gamma_{i_k}^{(0)}|\right)^2 &\leq \left(t_n \cdot c_{63} \cdot \frac{t_n}{L_n} \cdot \exp(-\delta_n/2)\right)^2 \\
&= \left(c_{63} \cdot \frac{t_n^2}{L_n} \cdot \exp(-(\log K_n)^2)\right)^2 \\
&\leq \frac{1}{n^8}.
\end{aligned}
$$

This yields together with the definition of α^*

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i) - f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i))^2 \\
\leq\ & \max_{k=1,\ldots,\tilde{K}_n} \max\{|\alpha_{i_k}^*|^2, 1\} \cdot \frac{1}{n^8} \\
\leq\ & 2 \cdot \beta_n^2 \cdot \frac{1}{n^8} \\
=\ & 2 \cdot c_3 \cdot (\log n)^2 \cdot \frac{1}{n^8} \\
\leq\ & \frac{1}{n^6}.
\end{aligned}
$$

Since

$$
\frac{1}{n^6} \leq 1
$$

we have

$$
\sqrt{\frac{1}{n^6}} \geq \frac{1}{n^6}.
$$

Moreover, we have by definition of α^*

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

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}|Y_i - f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(X_i)|^2 + \frac{c_1}{K_n^{2p}}\cdot\sum_{k=0}^{K_n}(\alpha_k^*)^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}(2\cdot|Y_i|^2 + 2\cdot|f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(X_i)|^2) + \frac{c_1}{K_n^{2p}}\cdot\sum_{k=0}^{K_n}(\alpha_k^*)^2 \\
&\leq 2\cdot\beta_n^2 + 2\cdot\left(\sum_{k=0}^{K_n}\alpha_k^*\right)^2 + \frac{c_1}{K_n^{2p}}\cdot\sum_{k=0}^{K_n}(\alpha_k^*)^2 \\
&= 2\cdot\beta_n^2 + 2\cdot\left(\sum_{k=0}^{\tilde{K}_n}\alpha_{i_k}^*\right)^2 + \frac{c_1}{K_n^{2p}}\cdot\sum_{k=0}^{\tilde{K}_n}(\alpha_{i_k}^*)^2 \\
&\leq 2\cdot\beta_n^2 + 2\cdot(\tilde{K}_n+1)^2\cdot 4\cdot\beta_n^2 + \frac{c_1}{K_n^{2p}}\cdot(\tilde{K}_n+1)\cdot 4\cdot\beta_n^2 \\
&= \beta_n^2\cdot(2 + 32\cdot\tilde{K}_n^2 + \frac{8\cdot c_1}{K_n^{2p}}\cdot\tilde{K}_n) \\
&= c_{66}\cdot\beta_n^2\cdot\tilde{K}_n^2 \\
&= c_{67}\cdot\frac{(\log n)^2\cdot K_n^2}{(\log n)^2} \\
&= c_{67}\cdot K_n^2.
\end{aligned}
$$

Hence, plugging in yields

$$
\begin{aligned}
& F(\alpha^*,\beta^{(t)},\gamma^{(t)}) - F(\alpha^*,\beta^{(0)},\gamma^{(0)}) \\
\leq\ & 2\cdot\sqrt{F(\alpha^*,\beta^{(0)},\gamma^{(0)})}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i) - f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i))^2} \\
& + \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\alpha^*,\beta^{(t)},\gamma^{(t)})}(x_i) - f_{net,(\alpha^*,\beta^{(0)},\gamma^{(0)})}(x_i))^2 \\
\leq\ & 2\cdot\sqrt{c_{67}\cdot K_n^2}\cdot\sqrt{\frac{1}{n^6}} + \frac{1}{n^6} \\
\leq\ & 2\cdot\sqrt{c_{67}\cdot K_n^2}\cdot\sqrt{\frac{1}{n^6}} + \sqrt{\frac{1}{n^6}} \\
\leq\ & c_{68}\cdot\frac{K_n}{n^3}.
\end{aligned}
$$

Putting together our results above gives us

$$\begin{aligned}
& F(\alpha^{(t+1)}, \beta^{(t+1)}, \gamma^{(t+1)}) - F(\alpha^*, \beta^{(0)}, \gamma^{(0)}) \\
\leq \quad & \left(1 - \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}}\right) \cdot (F(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) - \frac{1}{n} \sum_{i=1}^{n} |Y_i - f^*(X_i)|^2 - \text{pen}(f^*)) \\
& + \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot c_{68} \cdot \frac{K_n}{n^3}.
\end{aligned}$$

By applying this relation recursively, we get with (2.30)

$$\begin{aligned}
& F(\alpha^{(t+1)}, \beta^{(t+1)}, \gamma^{(t+1)}) - F(\alpha^*, \beta^{(0)}, \gamma^{(0)}) \\
\leq \quad & \left(1 - \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}}\right)^{t_n} \cdot (F(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}) - \frac{1}{n} \sum_{i=1}^{n} |Y_i - f^*(X_i)|^2 - \text{pen}(f^*)) \\
& + t_n \cdot \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot c_{68} \cdot \frac{K_n}{n^3} \\
= \quad & \left(1 - \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}}\right)^{t_n} \cdot (F(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}) - F(\alpha^*, \beta^{(0)}, \gamma^{(0)})) \\
& + t_n \cdot \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot c_{68} \cdot \frac{K_n}{n^3} \\
\leq \quad & \left(1 - \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}}\right)^{t_n} \cdot F(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}) + t_n \cdot \frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot c_{68} \cdot \frac{K_n}{n^3} \\
\leq \quad & \exp\left(-\frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot t_n\right) \cdot F(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}) + (\log n)^2 \cdot c_1 \cdot c_{68} \cdot \frac{K_n}{n^3} \\
\leq \quad & \exp\left(-\frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot t_n\right) \cdot c_{65} \cdot (\log n)^2 + c_1 \cdot c_{68} \cdot \frac{K_n}{n} \\
\leq \quad & \exp\left(-\frac{2 \cdot c_1}{L_n \cdot K_n^{2p}} \cdot K_n^{2p} \cdot (\log n)^2 \cdot L_n\right) \cdot c_{65} \cdot (\log n)^2 + c_1 \cdot c_{68} \cdot \frac{K_n}{n} \\
= \quad & \frac{c_{65} \cdot (\log n)^2}{n^{2 \cdot c_1 \cdot \log n}} + c_1 \cdot c_{68} \cdot \frac{K_n}{n} \\
\leq \quad & c_{62} \cdot \frac{K_n}{n}.
\end{aligned}$$

Hence, we have

$$T_{5,n} \leq c_{62} \cdot \frac{K_n}{n}.$$

Step 7: Bounding the second summand of Step 5. We have

$$
\begin{aligned}
& T_{6,n} \\
= \; & \mathbf{E}\Bigg(\Bigg(\frac{1}{n} \sum_{i=1}^{n} (|Y_i - f^*(X_i)|^2 - \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|^2 \Bigg) \cdot \mathbf{1}_{\{|Y_j| \leq \beta_n \; (j \in \{1,\ldots,n\})\}} \cdot \mathbf{1}_{A_n} \Bigg) \\
\leq \; & \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\Bigg(|\, f(X_i) - f^*(X_i)| \cdot |2Y_i - f(X_i) - f^*(X_i)| \cdot \mathbf{1}_{\{|Y_j| \leq \beta_n \; (j \in \{1,\ldots,n\})\}} \cdot \mathbf{1}_{A_n} \Bigg) \\
\leq \; & \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\Bigg(|\, f(X_i) - f^*(X_i)| \cdot (2\beta_n + \tilde{K}_n \cdot c_{70}) \cdot \mathbf{1}_{\{|Y_j| \leq \beta_n \; (j \in \{1,\ldots,n\})\}} \cdot \mathbf{1}_{A_n} \Bigg).
\end{aligned}
$$

Then, by assumption of the event A_n Lemma 2.2.13 b) we get

$$
\begin{aligned}
& |\, f^*(X_i) - f(X_i)| \\
= \; & |\sum_{k=1}^{\tilde{K}_n} \alpha^*_{i_k} \cdot \sigma(\beta^{(0)}_{i_k} \cdot X_i + \gamma^{(0)}_{i_k}) + \alpha^*_{i_0} - \sum_{k=1}^{\tilde{K}_n} \alpha^*_{i_k} \cdot \mathbf{1}_{[-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}}, \infty)}(X_i) - \alpha^*_{i_0}| \\
= \; & |\sum_{k=1}^{\tilde{K}_n} \alpha^*_{i_k} \cdot \sigma(\beta^{(0)}_{i_k} \cdot (X_i - (-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}}))) - \sum_{k=1}^{\tilde{K}_n} \alpha^*_{i_k} \cdot \mathbf{1}_{[-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}}, \infty)}(X_i)| \\
\leq \; & \sum_{k=1}^{\tilde{K}_n} |\alpha^*_{i_k}| \cdot |\sigma(\beta^{(0)}_{i_k} \cdot (X_i - (-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}}))) - \mathbf{1}_{[-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}}, \infty)}(X_i)| \\
\leq \; & \sum_{k=1}^{\tilde{K}_n} |\alpha^*_{i_k}| \cdot \exp\left(-\beta^{(0)}_{i_k} \cdot |X_i - (-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}})| \right) \\
= \; & \sum_{k=1}^{\tilde{K}_n} |\alpha^*_{i_k}| \cdot \exp\left(-|\beta^{(0)}_{i_k} \cdot X_i + \gamma^{(0)}_{i_k})| \right) \\
\leq \; & \sum_{k=1}^{\tilde{K}_n} |\alpha^*_{i_k}| \cdot \exp\left(-2 \cdot (\log K_n)^2 \right) \\
\leq \; & \tilde{K}_n \cdot c_{71} \cdot \exp\left(-2 \cdot (\log K_n)^2 \right) \\
\leq \; & \exp\left(-(\log K_n)^2 \right).
\end{aligned}
$$

Hence, we have

$$T_{6,n} \leq \exp\left(-(\log K_n)^2\right) \cdot (2\beta_n + c_{70} \cdot \tilde{K}_n) \leq \exp\left(-\frac{(\log K_n)^2}{4}\right).$$

Step 8: Bounding the third summand of Step 5. Let $\tilde{A}_n$ be the event that there exists $i_0, \ldots, i_{\tilde{K}_n} \in \{1, \ldots, K\}$ such that

$$\beta_{i_k} \geq \frac{B_n}{2}, \quad -\frac{\gamma_{i_k}}{\beta_{i_k}} \in I_k \quad \text{and} \quad \min_{i=1,\ldots,n} |\beta_{i_k} \cdot X_i + \gamma_{i_k}| \geq 2 \cdot (\log K_n)^2$$

holds for every $k \in \{0, \ldots, \tilde{K}_n\}$. Set

$$W^{(0)}_{\tilde{K}_n} = \{\beta^{(0)}_1, \ldots, \beta^{(0)}_{\tilde{K}_n}, \gamma^{(0)}_1, \ldots, \gamma^{(0)}_{\tilde{K}_n}\}$$

which is independent of $\mathcal{D}_n$. By the Cauchy-Schwarz inequality we get

$$\begin{aligned}
& T_{7,n} \\
= \; & \mathbf{E}\left(\left(\frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \cdot (\mathbf{1}_{A_n} - \mathbf{1}_{\tilde{A}_n} + \mathbf{1}_{\tilde{A}_n})\right) \\
= \; & \mathbf{E}\left(\left(\frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \cdot \mathbf{1}_{\tilde{A}_n}\right) \\
& + \mathbf{E}\left(\left(\frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \cdot (\mathbf{1}_{A_n} - \mathbf{1}_{\tilde{A}_n})\right) \\
\leq \; & \mathbf{E}\left(\left(\frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \cdot \mathbf{1}_{\tilde{A}_n}\right) \\
& + \sqrt{\mathbf{E}\left(\left(\frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right)^2\right)} \cdot \sqrt{\mathbf{E}\{(\mathbf{1}_{A_n} - \mathbf{1}_{\tilde{A}_n})^2\}}
\end{aligned}$$

We bound the terms on the right-hand side separately. We start with

$$\begin{aligned}
& \mathbf{E}\left(\left(\frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \cdot \mathbf{1}_{\tilde{A}_n}\right) \\
= \; & \mathbf{E}\left(\mathbf{E}\left(\left(\frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \cdot \mathbf{1}_{\tilde{A}_n} \mid W^{(0)}_{\tilde{K}_n}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(\left(|Y_i - f(X_i)|^2 - |m(X_i) - Y_i|^2\right)\cdot \mathbf{1}_{\tilde{A}_n} \mid W_{\tilde{K}_n}^{(0)}\right)\right) \\
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(\left(|(Y_i - m(X_i)) + (m(X_i) - f(X_i))|^2 - |m(X_i) - Y_i|^2\right)\right.\right. \\
&\qquad\qquad \left.\left.\cdot \mathbf{1}_{\tilde{A}_n} \mid W_{\tilde{K}_n}^{(0)}\right)\right) \\
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(\left(|Y_i - m(X_i)|^2 + |m(X_i) - f(X_i)|^2 - |m(X_i) - Y_i|^2\right)\right.\right. \\
&\qquad\qquad \left.\left.\cdot \mathbf{1}_{\tilde{A}_n} \mid W_{\tilde{K}_n}^{(0)}\right)\right) \\
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(|m(X_i) - f(X_i)|^2 \cdot \mathbf{1}_{\tilde{A}_n} \mid W_{\tilde{K}_n}^{(0)}\right)\right) \\
&\leq \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(|m(X_i) - f(X_i)|^2 \cdot \mathbf{1}_{\{-\frac{\gamma_{i_k}^{(0)}}{\beta_{i_k}^{(0)}} \in I_k \ (k\in\{1,\ldots,\tilde{K}_n\})\}} \mid W_{\tilde{K}_n}^{(0)}\right)\right) \\
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(|m(X_i) - f(X_i)|^2 \mid W_{\tilde{K}_n}^{(0)}\right)\cdot \mathbf{1}_{\{-\frac{\gamma_{i_k}^{(0)}}{\beta_{i_k}^{(0)}} \in I_k \ (k\in\{1,\ldots,\tilde{K}_n\})\}}\right) \\
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\int |m(x) - f(x)|^2 \mathbf{P}_{X_i}(dx)\cdot \mathbf{1}_{\{-\frac{\gamma_{i_k}^{(0)}}{\beta_{i_k}^{(0)}} \in I_k \ (k\in\{1,\ldots,\tilde{K}_n\})\}}\right) \\
&= \mathbf{E}\left(\int |m(x) - f(x)|^2 \mathbf{P}_{X}(dx)\cdot \mathbf{1}_{\{-\frac{\gamma_{i_k}^{(0)}}{\beta_{i_k}^{(0)}} \in I_k \ (k\in\{1,\ldots,\tilde{K}_n\})\}}\right) \\
&\leq \mathbf{E}\left(\sup_{x\in[0,1]} |m(x) - f(x)|^2 \cdot \mathbf{1}_{\{-\frac{\gamma_{i_k}^{(0)}}{\beta_{i_k}^{(0)}} \in I_k \ (k\in\{1,\ldots,\tilde{K}_n\})\}}\right),
\end{aligned}
$$

where the fourth equality holds since

$$
\begin{aligned}
&\mathbf{E}\left(\left((Y_i - m(X_i))\cdot(m(X_i) - f(X_i))\cdot \mathbf{1}_{\tilde{A}_n} \mid W_{\tilde{K}_n}^{(0)}\right)\right. \\
= \ &\mathbf{E}\left(\mathbf{E}\left((Y_i - m(X_i))\cdot(m(X_i) - f(X_i))\cdot \mathbf{1}_{\tilde{A}_n} \mid \{X_1,\ldots,X_n\}\cup W_{\tilde{K}_n}^{(0)}\right) \mid W_{\tilde{K}_n}^{(0)}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbf{E}\Big((m(X_i) - f(X_i)) \cdot \mathbf{1}_{\tilde{A}_n} \cdot \mathbf{E}\Big(Y_i - m(X_i) \mid \{X_1, \ldots, X_n\} \cup W^{(0)}_{\tilde{K}_n} \Big) \mid W^{(0)}_{\tilde{K}_n} \Big) \\
&= \mathbf{E}\Big((m(X_i) - f(X_i)) \cdot \mathbf{1}_{\tilde{A}_n} \cdot \mathbf{E}\Big(Y_i - m(X_i) \mid X_i \Big) \mid W^{(0)}_{\tilde{K}_n} \Big) \\
&= \mathbf{E}\Big((m(X_i) - f(X_i)) \cdot \mathbf{1}_{\tilde{A}_n} \cdot \Big(\mathbf{E}(Y_i \mid X_i) - m(X_i) \Big) \mid W^{(0)}_{\tilde{K}_n} \Big) \\
&= 0.
\end{aligned}
$$

On the event $\left\{ -\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}} \in I_k \ (k \in \{1, \ldots, \tilde{K}_n\}) \right\}$ we have for $x \in [0,1]$ in case that $-\frac{\gamma^{(0)}_{i_l}}{\beta^{(0)}_{i_l}} \leq x < -\frac{\gamma^{(0)}_{i_l+1}}{\beta^{(0)}_{i_l+1}}$ $(l \in \{0, \ldots, \tilde{K}_n - 1\})$ by $(p, C)-$smoothness of m

$$
\begin{aligned}
|m(x) - f^*(x)|^2 &= |m(x) - \sum_{k=1}^{\tilde{K}_n} \alpha^*_{i_k} \cdot \mathbf{1}_{\left[-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}}, \infty\right)}(x) - \alpha^*_{i_0}|^2 \\
&= |m(x) - \sum_{k=1}^{l} \alpha^*_{i_k} \cdot \mathbf{1}_{\left[-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}}, \infty\right)}(x) - \alpha^*_{i_0}|^2 \\
&= \left| m(x) - \left(m\left(-\frac{\gamma^{(0)}_{i_0}}{\beta^{(0)}_{i_0}} \right) + \sum_{k=1}^{l} \left(m\left(-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}} \right) - m\left(-\frac{\gamma^{(0)}_{i_k-1}}{\beta^{(0)}_{i_k-1}} \right) \right) \right) \right|^2 \\
&= \left| m(x) - m\left(-\frac{\gamma^{(0)}_{i_l}}{\beta^{(0)}_{i_l}} \right) \right|^2 \\
&\leq C^2 \cdot \left| x - \left(-\frac{\gamma^{(0)}_{i_l}}{\beta^{(0)}_{i_l}} \right) \right|^{2p} \\
&\leq C^2 \cdot c_{82} \cdot 2^{2p} \left(\frac{1}{\tilde{K}_n} \right)^{2p} \\
&\leq C^2 \cdot c_{81} \cdot \frac{(\log K_n)^{4p}}{K_n^{2p}}
\end{aligned}
$$

and in case that $-\frac{\gamma^{(0)}_{\tilde{K}_n}}{\beta^{(0)}_{\tilde{K}_n}} \leq x \leq 1$ we get analogously

$$\begin{aligned}
|m(x) - f^*(x)|^2 &= |m(x) - \sum_{k=1}^{\tilde{K}_n} \alpha^*_{i_k} \cdot \mathbf{1}_{\left[-\frac{\gamma^{(0)}_{i_k}}{\beta^{(0)}_{i_k}}, \infty\right)}(x) - \alpha^*_{i_0}|^2 \\
&= \left| m(x) - m\left(-\frac{\gamma^{(0)}_{i_{\tilde{K}_n}}}{\beta^{(0)}_{i_{\tilde{K}_n}}}\right)\right|^2 \\
&\leq C^2 \cdot \left|1 - \left(-\frac{\gamma^{(0)}_{i_{\tilde{K}_n}}}{\beta^{(0)}_{i_{\tilde{K}_n}}}\right)\right|^{2p} \\
&\leq C^2 \cdot c_{81} \cdot \frac{(\log K_n)^{4p}}{K_n^{2p}}.
\end{aligned}$$

Hence, we get

$$\mathbf{E}\left(\left(\frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \cdot \mathbf{1}_{\tilde{A}_n}\right) \leq C \cdot c_{81} \cdot \frac{(\log K_n)^{4p}}{K_n^{2p}}.$$

Next, since

$$c_{10} \cdot |Y|^2 \leq \exp(c_{10} \cdot |Y|^2)$$

implies by (2.7)

$$\mathbf{E}(|Y|^2) \leq \frac{1}{c_{10}} \cdot \mathbf{E}(\exp(c_{10} \cdot |Y|^2))$$

and consequently

$$\begin{aligned}
\mathbf{E}(|Y|^4) &= \mathbf{E}(|Y|^2 \cdot |Y|^2) \\
&\leq \mathbf{E}\left(\frac{1}{c_{10}} \cdot \exp(c_{10} \cdot |Y|^2) \cdot \frac{1}{c_{10}} \cdot \exp(c_{10} \cdot |Y|^2)\right) \\
&= \frac{1}{c_{10}^2} \cdot \mathbf{E}(\exp(2 \cdot c_{10} \cdot |Y|^2)) \\
&\leq c_{82}
\end{aligned}$$

we conclude

$$\mathbf{E}\left(\left(\frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right)^2\right)$$

$$
\begin{aligned}
&= \mathbf{E}\left(\frac{1}{n^2}\left(\sum_{i=1}^n (|Y_i - f(X_i)|^2 - |m(X_i) - Y_i|^2)\right)^2\right) \\
&\leq \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n (|Y_i - f(X_i)|^2 - |m(X_i) - Y_i|^2)^2\right) \\
&\leq 4 \cdot \mathbf{E}\Big((|Y_i|^2 + |f(X_i)|^2 + |m(X_i)|^2 + |Y_i|^2)^2\Big) \\
&\leq 16 \cdot \mathbf{E}\Big(|Y_i|^4 + |f(X_i)|^4 + |m(X_i)|^4 + |Y_i|^4\Big) \\
&\leq 32 \cdot (c_{82} + \beta_n^4) \\
&\leq c_{83} \cdot (\log n)^4.
\end{aligned}
$$

Finally, as in Step 4 we have

$$
\begin{aligned}
& \mathbf{E}\{(\mathbf{1}_{A_n} - \mathbf{1}_{\tilde{A}_n})^2\} \\
\leq\ & \mathbf{E}\{(\mathbf{1}_{|Y_i|>\beta_n \text{ for some } i\in\{1,\dots,n\}})^2\} \\
=\ & \mathbf{E}\{\mathbf{1}_{|Y_i|>\beta_n \text{ for some } i\in\{1,\dots,n\}}\} \\
=\ & \mathbf{P}\left\{|Y_i| > \beta_n \text{ for some } i \in \{1,\dots,n\}\right\} \\
\leq\ & n \cdot \frac{c_{40}}{n^{c_{41}\cdot \log n}} \\
\leq\ & \frac{1}{n^3}.
\end{aligned}
$$

Taking the above together yields

$$
\begin{aligned}
T_{7,n} &\leq C \cdot c_{81} \cdot \frac{(\log K_n)^{4p}}{K_n^{2p}} + \sqrt{c_{83} \cdot (\log n)^4} \cdot \sqrt{\frac{1}{n^3}} \\
&= C \cdot c_{81} \cdot \frac{(\log K_n)^{4p}}{K_n^{2p}} + \sqrt{c_{83}} \cdot \frac{(\log n)^2}{n^{\frac{3}{2}}} \\
&\leq c_{80} \cdot \frac{(\log K_n)^{4p}}{K_n^{2p}}.
\end{aligned}
$$

Step 9: Bounding the fourth summand of Step 5. The (p, C)-smoothness of m implies for $j \in \{1, \dots, \tilde{K}_n\}$

$$
|\alpha_{i_j}^*| = \left| m\left(-\frac{\gamma_{i_j}^{(0)}}{\beta_{i_j}^{(0)}}\right) - m\left(-\frac{\gamma_{i_j-1}^{(0)}}{\beta_{i_j-1}^{(0)}}\right)\right| \leq C \cdot \left|\frac{\gamma_{i_j}^{(0)}}{\beta_{i_j}^{(0)}} - \frac{\gamma_{i_j-1}^{(0)}}{\beta_{i_j-1}^{(0)}}\right|^p.
$$

On A_n we know that $-\frac{\gamma_{i_j}^{(0)}}{\beta_{i_j}^{(0)}}$ and $-\frac{\gamma_{i_j-1}^{(0)}}{\beta_{i_j-1}^{(0)}}$ are contained in two adjacent intervals of length $1/\tilde{K}_n$. Using the assumption $p \in [\frac{1}{2}, 1]$ we get

$$\begin{aligned}
\text{pen}(f^*) \cdot \mathbf{1}_{A_n} &= \frac{c_1}{K_n^{2p}} \cdot \sum_{k=0}^{\tilde{K}_n} (\alpha_{i_k})^2 \\
&\leq \frac{c_1}{K_n^{2p}} \cdot \left(\|m\|_\infty^2 + \tilde{K}_n \cdot C^2 \cdot \frac{2^{2p}}{\tilde{K}_n^{2p}} \right) \\
&\leq \frac{c_1 \cdot c_3^2}{K_n^{2p}} \cdot (\log n)^2 + \frac{c_1}{K_n^{2p}} \cdot \tilde{K}_n \cdot C^2 \cdot \frac{2^{2p}}{\tilde{K}_n^{2p}} \\
&\leq c_{91} \cdot \frac{(\log n)^2}{K_n^{2p}} + c_{92} \cdot \frac{1}{K_n^{2p}} \cdot \frac{K_n}{(\log K_n)^2} \cdot \frac{(\log K_n)^{4p}}{K_n^{2p}} \\
&\leq c_{90} \cdot \frac{(\log K_n)^{4p}}{K_n^{2p}}
\end{aligned}$$

So,

$$T_{8,n} \leq c_{90} \cdot \frac{(\log K_n)^{4p}}{K_n^{2p}}.$$

Step 10: Finishing the proof. We put together the results from the previous steps. This gives us

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
\leq\ & c_{30} \cdot \frac{(\log n)^3 \cdot K_n}{n} \\
&+ c_{62} \cdot \frac{K_n}{n} + \exp\left(-\frac{(\log K_n)^2}{4} \right) + c_{80} \cdot \frac{(\log K_n)^{4p}}{K_n^{2p}} + c_{90} \cdot \frac{(\log K_n)^{4p}}{K_n^{2p}} \\
&+ c_{43} \cdot \frac{1}{n} \\
\leq\ & c_7 \cdot (\log n)^{\max\{3,4p\}} \cdot n^{-\frac{2p}{2p+1}}.
\end{aligned}$$

This concludes the proof.

□

2.3. Application to Simulated Data

We illustrate the finite sample size performance of our newly proposed estimate by applying it to simulated data using the software *MATLAB*.

For our simulation we choose the simulated data as follows: We choose X uniformly distributed on $[0,1]$, ϵ standard normal and independent of X, and we define Y by

$$Y = m_j(X) + \sigma \cdot \lambda_j \cdot \epsilon,$$

where $m_j : [-1,1]^d \to \mathbb{R}$ is described below, $\lambda_j > 0$ is a scaling value defined below and σ is chosen from $\{0.01, 0.05, 0.10, 0.20\}$ $(j \in \{1,2\})$. As regression functions we use

$$m_1(x) = \sin((-0.6x)^2) + \log\left(\frac{1}{(0.5x)^2} + 1\right)$$

and

$$m_2(x) = \frac{1}{1+\exp(-0.8317x)} + \sqrt{(2x)^2+1}.$$

λ_j is chosen approximately as IQR of a sample of size 100 of $m(X)$, and we use the values $\lambda_1 = 0.0293$ and $\lambda_2 = 0.5167$.

From this distribution we generate a sample of size $n = 100$ and apply our newly proposed neural network regression estimate and compare our results to that of six alternative regression estimates on the same data. Then we compute the L_2 errors of these estimates approximately by using the empirical L_2 error $\varepsilon_{L_2,\bar{N}}(\cdot)$ on an independent sample of X of size $\bar{N} = 10,000$. Since this error strongly depends on the behavior of the true function m_j, we set it in relation to the error of the simplest estimate for m_j we can think of, a completely constant function. The constant function estimate describes the average of the observed data according to the least squares approach. Thus, the scaled error measure we use for evaluation of the estimates is $\varepsilon_{L_2,\bar{N}}(m_{n,i})/\bar{\varepsilon}_{L_2,\bar{N}}(avg)$, where $\bar{\varepsilon}_{L_2,\bar{N}}(avg)$ is the median of 50 independent realizations of the value obtained if the average of n observations is plugged into $\varepsilon_{L_2,\bar{N}}(\cdot)$. To a certain extent, this quotient can be interpreted as the relative part of the error of the constant estimate that is still contained in the more sophisticated approaches. Of course, the resulting scaled errors depend on the random sample of (X,Y) and in order to still be able to compare these values we repeat the whole computation 50 times and report the median and the interquartile range of the 50 scaled errors for each of our estimates.

We choose the parameters for each of the estimates by splitting of the sample. Here we split our sample into a learning sample of size $n_l = 0.8 \cdot n$ and into a testing sample of size

$n_t = 0.2 \cdot n$. We compute the estimate for all parameter values from the sets described below using the learning sample. Then, we compute the corresponding empirical L_2 risk on the testing sample and choose the parameter value which leads to the minimal empirical L_2 risk on the testing sample.

Our first three estimates are built-in fully connected neural network estimates where the number of layers is fixed and the number of neurons per layer is chosen adaptively. The estimate *fc-neural-1* has one hidden layer, estimate *fc-neural-3* has three hidden layers, estimate *fc-neural-6* has six hidden layers and the number of neurons per layer is chosen from the set $\{5, 10, 25, 50, 75\}$, $\{3, 6, 9, 12, 15\}$, $\{2, 4, 6, 8, 10\}$, respectively.

Our fourth estimate *kernel* is the Nadaraya-Watson kernel estimate with so-called naive kernel where the bandwith is chosen from the set $\{2^k : k \in \{-5, -4, \dots, 5\}\}$.

Our fifth estimate *neighbor* is a nearest neighbor estimate where the number of nearest neighbors is chosen from the set $\{1, 2, 3\} \cup \{4, 8, 12, 16, \dots, 4 \cdot \lfloor \frac{n_l}{4} \rfloor\}$.

Our sixth estimate *RBF* is the interpoland with radial basis functions where the radial basis functions $\Phi(r) = (1 - r)_+^6 \cdot (35 \cdot r^2 + 18 \cdot r + 3)$ is used and the scaling radius is chosen adaptively.

Our seventh estimate *neural-1* is our newly proposed neural network estimate presented in this chapter. Here, the parameter K of the estimate is chosen from the set $\{5, 25, 50\}$ and we set the number of gradient descent steps to $t_n = 175000$.

The results are summarized in Table 2.1 and in Table 2.2. As we can see from the reported scaled errors, our newly proposed neural network estimate does not perform as well as the built-in fully connected neural networks. This is can be explained by the small set from which we choose the parameters of the network and the few number of performed gradient descent steps. Unfortunately, we do not have the capacities to run the network with greater parameter set and a greater number of gradient descent steps. While the built-in neural network estimates may show better performances than our neural network estimate, we would like to emphasize that the former have no theoretical background.

	m_1			
noise	1%	5%	10%	20%
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	0.0015	0.0015	0.0015	0.0015
approach	median (IQR)	median (IQR)	median (IQR)	median (IQR)
fc-neural-1	0.0011 (0.001)	**0.0006 (0.001)**	0.0019 (0.001)	0.014 (0.001)
fc-neural-3	**0.0005 (0.002)**	0.0011 (0.001)	**0.0012 (0.002)**	**0.0023 (0.001)**
fc-neural-6	0.0046 (0.002)	0.0048 (0.001)	0.0023 (0.001)	0.0079 (0.002)
kernel	0.0102 (0.012)	0.0104 (0.011)	0.0112 (0.011)	0.0141 (0.01)
neighbor	0.0020 (0.001)	0.0034 (0.001)	0.0055 (0.002)	0.0156 (0.016)
RBF	0.0094 (0.005)	0.247 (0.0121)	0.9279 (0.052)	3.6372 (0.222)
neural-1	0.0752 (0.078)	0.0751 (0.079)	0.0751 (0.081)	0.0752 (0.084)

Table 2.1.: Median and IQR of the scaled empirical L_2 error of estimates for m_1 for sample size $n = 100$. The smallest error values in each column is highlighted by bold letters.

	m_2			
noise	1%	5%	10%	20%
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	0.1954	0.1955	0.1956	0.1962
approach	median (IQR)	median (IQR)	median (IQR)	median (IQR)
fc-neural-1	**0.0001 (0.001)**	0.0006 (0.004)	**0.0013 (0.004)**	0.0036 (0.002)
fc-neural-3	0.0007 (0.002)	0.0021 (0.001)	0.0025 (0.001)	0.0037 (0.001)
fc-neural-6	**0.0001 (0.001)**	**0.0005 (0.001)**	**0.0013 (0.002)**	**0.0032 (0.001)**
kernel	0.0374 (0.069)	0.038 (0.068)	0.0409 (0.065)	0.0487 (0.006)
neighbor	0.0015 (0.001)	0.0035 (0.002)	0.0095 (0.010)	0.0118 (0.002)
RBF	0.0156 (0.015)	0.5344 (0.055)	2.528 (0.35)	8.4734 (0.844)
neural-1	0.0632 (0.002)	0.0632 (0.004)	0.0634 (0.006)	0.0643 (0.011)

Table 2.2.: Median and IQR of the scaled empirical L_2 error of estimates for m_2 for sample size $n = 100$. The smallest error values in each column is highlighted by bold letters.

3. Neural Network Regression Estimates Learned by Gradient Descent Inspired by Approximation Results with Indicator Functions for Projection Pursuit

We deal with neural network regression in a projection pursuit model. This means, we assume that the regression function satisfies

$$m(x) = \sum_{s=1}^{r} g_s(\mathbf{c}_s^T x) \quad (x \in \mathbb{R}^d)$$

for some $r \in \mathbb{N}$, $\mathbf{c}_s \in \mathbb{R}^d$, where $\|\mathbf{c}_s\| = 1$ $s \in \{1, \ldots, r\}$, and (p, C)-smooth functions $g_s : \mathbb{R} \to \mathbb{R}$ $(s = 1, \ldots, r)$. The constraint imposed upon the regression function has the effect that the $d-$dimensional input is reduced to a $1-$dimensional input for the functions g_s that make up m. The natural question to ask is whether we can now achieve a univariate rate of convergence. In this chapter we present an implementable neural network regression estimate with one hidden layer that achieves up to a logarithmic factor the univariate rate of convergence. We draw the inspiration for our neural network estimate from the observation that for large values of $\varphi \in \mathbb{R}$ the logistic squasher

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

is close to an indicator function. For graphic understanding, a visualization is shown in Figure 3.1. The idea is that we can build a neural network that is close to a piecewise constant approximation of m. We learn the weights of the network by gradient descent. In contrast to the network presented in Chapter 2, the choice of the inital weights is more restricted. The inital outer weights are set to zero and the initial inner weights are chosen carefully from specific intervals dependent on $X_1, \ldots, X_n$ and on the projection directions. This guarantees a good approximation of m and we will show that gradient descent changes the inner weights only slightly and finds the optimal outer weights. Since

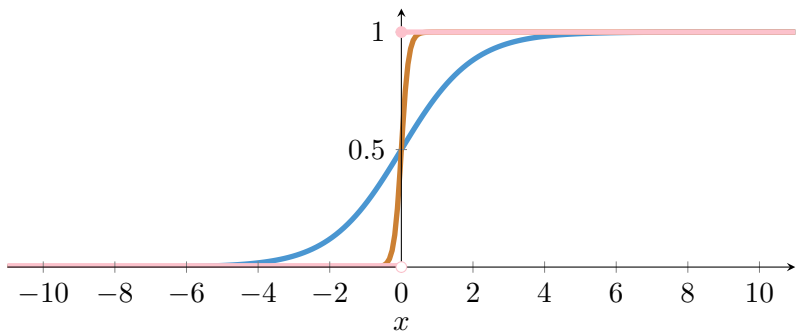

Figure 3.1.: Visualisation of the sigmoid logistic squashing function $\sigma(\varphi x)$ for $\varphi = 1$ (blue graph) and for $\varphi = 10$ (brown graph) and of the indicator function $1_{[0,\infty)}(x)$ (pink graph).

the projection directions are unknown, we will guess them repeatedly. This results in repeated initialization of the neural network from which we choose the one with minimal error. In comparison with the neural network regression esimates presented in in Chapter 2, we see that on the one hand there is less freedom in the choice of the weights as the intervals from which they are chosen is smaller but on the other hand we are able to analyze neural network estimates for multi-variate $(p, C)-$smooth functions where the smoothness factor $p \in (0, 1]$ is less restricted.
We construct our neural network regression estimate in Section 3.1 and show the corresponding univariate rate of convergence result in Section 3.2. The finite sample size performance of our newly proposed estimate is illustrated in Section 3.3 by applying it to simulated data.

3.1. Constructing the Neural Network

Let $A \geq 1$ and assume that the support of X is contained in the cube $[-A, A]^d$. The construction of our neural network regression estimate is motivated by an approximation in two steps:

1. Approximate each $g_s : \mathbb{R} \to \mathbb{R}$ by a piecewise constant function of the form

$$u \mapsto \sum_{l=1}^{K} a_{s,l} \cdot \mathbf{1}_{[b_l,\infty)} + a_{s,0}.$$

2. Approximate the pieceweise constant approximation by a neural network with logistic squasher of the form

$$u \mapsto \sum_{l=1}^{K} a_{s,l} \cdot \sigma(\rho_n \cdot (u - b_l)) + a_{s,0},$$

where $\rho_n > 0$ is a large constant. The error of this approximation will be small at all those points, where $\rho_n \cdot |u - b_l|$ is large. An example of the network is shown in Figure 3.2.

We will deal with this concept in Lemma 3.2.6. By replacing u with $\mathbf{c}_s^T x$ we see that we

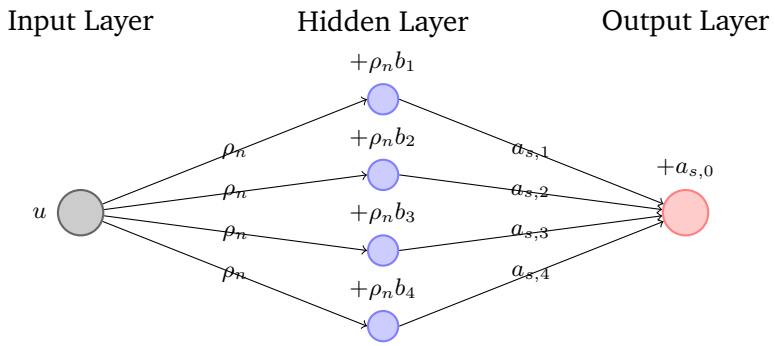

Figure 3.2.: Visualisation of an example of network estimate of g_s with parameters $K = 4$. The function has a 1-dimensional input (black node) as well as a 1-dimensional output (red node). There is one hidden layer consisting of 4 neurons (blue node).

can approximate m by networks with one hidden layer and $K \cdot r$ neurons in this hidden layer defined by

$$f_{net,(\mathbf{a},\mathbf{b})}(x) = \sum_{k=1}^{K \cdot r} a_k \cdot \sigma \left(\sum_{j=1}^{d} b_{k,j} \cdot x^{(j)} + b_{k,0} \right) + a_0, \tag{3.1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function and

$$a_k \in \mathbb{R} \quad (k = 0, \ldots, K \cdot r) \quad \text{and} \quad b_{k,j} \in \mathbb{R} \quad (k = 1, \ldots, K \cdot r, j = 0, \ldots, d)$$

are the weights. An example of the network is shown in Figure 1.3. We learn the weights $(\mathbf{a}, \mathbf{b})$ by gradient descent. The above condition that $\rho_n \cdot |u - b_l|$ is large in order to achieve

a small error at point u of the above neural network approximation of the piecewise constant function is replaced by the assumption that

$$\min_{i=1,\dots,n} |\sum_{j=1}^{d} b_{k,j} \cdot X_i^{(j)} + b_{k,0}|$$

is large, which will enable us to show that our approximation is good at all x-values of the data points. This condition will be ensured by a proper choice of the initial weights $(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})$ described below. We choose the projection directions by repeated random guessing.
In detail, the construction procedure for our neural network estimate has seven steps:

1. Randomly choose values
$$\mathbf{c}_1^*, \dots, \mathbf{c}_r^* \in [-1,1]^d$$
as an independent sample from a uniform distribution on $[-1,1]^d$ and set
$$\bar{\mathbf{c}}_s = \frac{\mathbf{c}_s^*}{\|\mathbf{c}_s^*\|} \quad (s = 1, \dots, r).$$
These values approximate the direction of projection $\mathbf{c}_1, \dots, \mathbf{c}_r$ in our projection pursuit model.
Note, that for $s = 1, \dots, r$
$$\mathbf{P}\{\mathbf{c}_s^* = 0\} = 0,$$
which means we can assume w.l.o.g. $\|\mathbf{c}_s^*\| \neq 0$.

2. Prepare the definition of the initial inner weights $b_{(s-1)\cdot K+k,j}$ ($s = 1, \dots, r$, $k = 1, \dots, K$, $j = 0, \dots, d$) according to $\bar{\mathbf{c}}_s$ and to $X_1, \dots, X_n$: For each $s \in \{1, \dots, r\}$ choose $\tilde{b}_{s,1}, \dots, \tilde{b}_{s,K} \in \mathbb{R}$ such that $\tilde{b}_{s,1} < \tilde{b}_{s,2} < \dots < \tilde{b}_{s,K}$ and
$$\tilde{b}_{s,1} \leq -A \cdot \sqrt{d} \quad \text{and} \quad \tilde{b}_{s,K} \geq A \cdot \sqrt{d} - \frac{4 \cdot \sqrt{d} \cdot A}{K-1},$$
$$\frac{\sqrt{d} \cdot A}{(n+1) \cdot (K-1)} \leq |\tilde{b}_{s,k+1} - \tilde{b}_{s,k}| \leq \frac{4 \cdot \sqrt{d} \cdot A}{K-1} \quad (k = 1, \dots, K-1)$$
and
$$\min_{i=1,\dots,n, k=1,\dots,K} \left| \bar{\mathbf{c}}_s^T X_i - \tilde{b}_{s,k} \right| \geq \frac{\sqrt{d} \cdot A}{(n+1) \cdot (K-1)}.$$

Such a choice is always possible. For example, a viable way to do this is to set $\tilde{b}_{s,1} = -\sqrt{d} \cdot A - 2 \cdot \sqrt{d} \cdot A/((n+1) \cdot (K-1))$. Then, regarding $\tilde{b}_{s,k}$ $(k = 2, \ldots, K)$, subdivide the interval

$$\left[-\sqrt{d} \cdot A + (k-2) \cdot \frac{2 \cdot \sqrt{d} \cdot A}{K-1}, -\sqrt{d} \cdot A + (k-1) \cdot \frac{2 \cdot \sqrt{d} \cdot A}{K-1}\right]$$

into $(n+1)$ subintervals of equal length $2 \cdot \sqrt{d} \cdot A/((K-1) \cdot (n+1))$ and choose $\tilde{b}_{s,k}$ as the midpoint of one of those intervals which does not contain any of the n values $\bar{\mathbf{c}}_s^T X_i$. Such an interval must always exist due to the impossibility of $n+1$ disjoint intervals containing at least one of the above n points each.

3. Define our initial inner weights

$$b_{(s-1)\cdot K+1,0}, \ldots, b_{(s-1)\cdot K+1,d}, \ldots, b_{s\cdot K,0}, \ldots, b_{s\cdot K,d}$$

for $s \in \{1, \ldots, r\}$ such that we have

$$\sum_{j=1}^{d} b_{(s-1)\cdot K+k,j} \cdot x^{(j)} + b_{(s-1)\cdot K+k,0} = \rho_n \cdot (\bar{\mathbf{c}}_s^T x - \tilde{b}_{s,k}) \quad \text{for all } x \in \mathbb{R}^d,$$

for some $\rho_n > 0$. In other words, set

$$b_{(s-1)\cdot K+k,j} = \rho_n \cdot \bar{\mathbf{c}}_s^{(j)} \quad \text{and} \quad b_{(s-1)\cdot K+k,0} = -\rho_n \cdot \tilde{b}_{s,k}$$

$(s = 1, \ldots, r,\ k = 1, \ldots, K,\ j = 1, \ldots, d)$ for some $\rho_n > 0$. A choice of ρ_n is given in Theorem 3.2.1.

4. Define the initial output weights by

$$a_l = 0 \quad \text{for all} \quad l \in \{0, \ldots, K \cdot r\}.$$

5. Apply gradient descent for $t = 0, 1, \ldots, t_n - 1$: Learn the weights by gradient descent. More precisely, we minimize the penalized empirical L_2 risk

$$F(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} |f_{net,(\mathbf{a},\mathbf{b})}(X_i) - Y_i|^2 + \frac{c_1}{n} \cdot \sum_{k=0}^{K \cdot r} a_k^2, \tag{3.2}$$

where $c_1 > 0$ is a constant by defining $(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)})$ recursively by

$$\begin{pmatrix} \mathbf{a}^{(t+1)} \\ \mathbf{b}^{(t+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{a}^{(t)} \\ \mathbf{b}^{(t)} \end{pmatrix} - \lambda_n \cdot (\nabla_{(\mathbf{a},\mathbf{b})} F)(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) \tag{3.3}$$

for some $\lambda_n > 0$ and $t = 0, 1, \ldots, t_n - 1$. Here, $\lambda_n > 0$ denotes the step size and $t_n \in \mathbb{N}$ denotes the number of steps we perform in the gradient descent algorithm. A choice of λ_n and of t_n is given in Theorem 3.2.1. Note, that minimization of (3.2) with respect to $(\mathbf{a}, \mathbf{b})$ is a nonlinear least squares problem.

Repeat steps $1. - 5.$ I_n times. A choice of I_n is given in Theorem 3.2.1.

6. Choose the directions and the corresponding network which achieves the smallest penalized empirical L_2 error (3.2) among all the I_n networks as our neural network estimate $\tilde{m}_n$.

7. Set
$$\tilde{m}_n(\cdot) = f_{net,\mathbf{w}(t_n)}(\cdot) \tag{3.4}$$
and choose our estimate to be the by
$$\beta_n = c_3 \cdot \log n$$
truncated version
$$m_n(x) = T_{\beta_n} \tilde{m}_n(x). \tag{3.5}$$

3.2. Rate of Convergence

Theorem 3.2.1 states that our neural network regression estimate constructed in Section 3.1 achieves up to a logarithmic factor the univariate rate of convergence.

Theorem 3.2.1. *Let $n \geq 2$, let $A \geq 1$ and let (X, Y), (X_1, Y_1), …, (X_n, Y_n) be independent and identically distributed random variables with values in $[-A, A]^d \times \mathbb{R}$. Set $m(x) = \mathbf{E}\{Y|X = x\}$ and assume that (X, Y) satisfies*
$$\mathbf{E}\left(e^{c_2 \cdot |Y|^2}\right) < \infty \tag{3.6}$$
for some constant $c_2 > 0$, and that m satisfies
$$m(x) = \sum_{s=1}^{r} g_s(\mathbf{c}_s^T x) \quad (x \in \mathbb{R}^d)$$
for some $r \in \mathbb{N}$, $\mathbf{c}_s \in [-1, 1]^d$, where $\|\mathbf{c}_s\| = 1$, and $g_s : \mathbb{R} \to \mathbb{R}$ $(s = 1, \ldots, r)$. Assume that g_s is (p, C)-smooth for $s \in \{1, \ldots, r\}$, where $p \in (0, 1]$ and $C > 0$ are fixed. Define the regression estimate m_n as in Section 3.1 with
$$\sigma(x) = \frac{1}{1 + e^{-x}},$$

with parameter r as in the above projection pursuit model, and with the other parameters chosen by

$$K = K_n = \left\lceil \left(\frac{n}{(\log n)^3} \right)^{\frac{1}{2p+1}} \right\rceil,$$

$$\rho_n = n^2 \cdot K_n, \quad \lambda_n = \frac{1}{3 \cdot K_n \cdot r}, \quad t_n = K_n \cdot n \cdot (\log n)^2,$$

and

$$I_n = \left\lceil \left(\frac{n}{(\log n)^3} \right)^{\frac{r \cdot (d-1)}{2p+1}} \right\rceil.$$

Then m_n satisfies

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_4 \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}}$$

for some constant $c_4 > 0$ which does not depend on n.

Remark 3.2.2. *Since the rate of convergence presented in Theorem 3.2.1 is independent of the dimension d of X, this means that our proposed computable neural network regression estimate for a regression function that satisfies the regression pursuit model is able to circumvent the curse of dimensionality. However, we can see that the dependence on the dimenion d has shifted into the necessary number of repititions I_n of the initial random choices of the directions $\bar{\mathbf{c}}_s$. As a consequence, we get with*

$$t_n \leq n^2$$

a repitition number of

$$I_n \cdot t_n \leq n^{2+r \cdot (d-1)}$$

which is rather huge.

Remark 3.2.3. *The parameters r and K_n, and also I_n of the above algorithm depend on the projection pursuit model and are hence unknown in any application. Using the splitting of the sample technique as explained in Section 3.3 it is possible to choose these parameters data-dependently.*

3.2.1. Learning of Linear Penalized Least Squares Estimates by Gradient Descent

We start with the consideration of a neural network where the inner weights are fixed and only the output weights are learned by mimization of $F(\mathbf{a}, \mathbf{b})$ as in (3.2) but with fixed $\mathbf{b}$ with respect to $\mathbf{a}$ by gradient descent. Essentially, this boils down to a linear least squares problem which is solved by gradient descent. We will need this observation in the proof of Lemma $3.2.10$.
So, let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, let $K \in \mathbb{N}$, let $B_1, \ldots, B_K : \mathbb{R}^d \to \mathbb{R}$ and let $c_1 > 0$. We consider the problem to minimize

$$F(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} |\sum_{k=1}^{K} a_k \cdot B_k(x_i) - y_i|^2 + \frac{c_1}{n} \cdot \|\mathbf{a}\|^2, \tag{3.7}$$

where $\mathbf{a} = (a_1, \ldots, a_K)^T$, by gradient descent. To do this, we choose $\mathbf{a}^{(0)} \in \mathbb{R}^K$ and set

$$\mathbf{a}^{(t+1)} = \mathbf{a}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)}) \tag{3.8}$$

for some properly chosen $\lambda_n > 0$ and $t = 0, \ldots, t_n - 1$.

Lemma 3.2.4. *Let $F : \mathbb{R}^K \to \mathbb{R}$ be a differentiable function and define $\mathbf{a}^{(t+1)}$ by (3.8), where*

$$\lambda_n = \frac{1}{L_n} \tag{3.9}$$

for some $L_n > 0$. Let $\mathbf{a}_{opt} \in \mathbb{R}^K$ be arbitrary.
If

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}_1) - (\nabla_{\mathbf{a}} F)(\mathbf{a}_2)\| \leq L_n \cdot \|\mathbf{a}_1 - \mathbf{a}_2\| \quad (\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^K) \tag{3.10}$$

and, in addition,

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2 \geq \rho_n \cdot (F(\mathbf{a}) - F(\mathbf{a}_{opt})) \quad (\mathbf{a} \in \mathbb{R}^K) \tag{3.11}$$

hold for some $\rho_n > 0$, then we have

$$F(\mathbf{a}^{(t+1)}) - F(\mathbf{a}_{opt}) \leq \left(1 - \frac{\rho_n}{2 \cdot L_n}\right) \cdot (F(\mathbf{a}^{(t)}) - F(\mathbf{a}_{opt})).$$

Proof. By (3.10) we can apply Lemma $2.2.5$ and with (3.11) we get

$$F(\mathbf{a}^{(t+1)}) - F(\mathbf{a}_{opt})$$

$$
\begin{aligned}
&\leq F(\mathbf{a}^{(t)}) - F(\mathbf{a}_{opt}) - \frac{1}{2 \cdot L_n} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)})\|^2 \\
&\leq F(\mathbf{a}^{(t)}) - F(\mathbf{a}_{opt}) - \frac{1}{2 \cdot L_n} \cdot \rho_n \cdot (F(\mathbf{a}^{(t)}) - F(\mathbf{a}_{opt})) \\
&= \left(1 - \frac{\rho_n}{2 \cdot L_n}\right) \cdot (F(\mathbf{a}^{(t)}) - F(\mathbf{a}_{opt})).
\end{aligned}
$$

□

The following lemma concerns the verification of the condition (3.10) of Lemma 3.2.4. Verification of the condition (3.11) of Lemma 3.2.4 will be done by Lemma 2.2.12. We will use this in the proof of Lemma 3.2.10.

Lemma 3.2.5. *Let F be defined by* (3.7). *Then we have for any $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^K$*

$$
\|(\nabla_{\mathbf{a}} F)(\mathbf{a}_1) - (\nabla_{\mathbf{a}} F)(\mathbf{a}_2)\| \leq \left(2 \cdot \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} B_k(x_i)^2 + \frac{2 \cdot c_1}{n}\right) \cdot \|\mathbf{a}_1 - \mathbf{a}_2\|.
$$

Proof. We have

$$
F(\mathbf{a}) = \frac{1}{n} \cdot (\mathbf{B} \cdot \mathbf{a} - \mathbf{y})^T \cdot (\mathbf{B} \cdot \mathbf{a} - \mathbf{y}) + \frac{c_1}{n} \cdot \mathbf{a}^T \cdot \mathbf{a}
$$

where

$$
\mathbf{B} = (B_j(x_i))_{1 \leq i \leq n, 1 \leq j \leq K} \quad \text{and} \quad \mathbf{y} = (y_1, \ldots, y_n)^T.
$$

Consequently,

$$
(\nabla_{\mathbf{a}} F)(\mathbf{a}) = \frac{2}{n} \cdot \left(\mathbf{B}^T \mathbf{B} \mathbf{a} - \mathbf{B}^T \mathbf{y}\right) + \frac{2 \cdot c_1}{n} \cdot \mathbf{a}
$$

and

$$
\|(\nabla_{\mathbf{a}} F)(\mathbf{a}_1) - (\nabla_{\mathbf{a}} F)(\mathbf{a}_2)\| \leq \|\frac{2}{n} \cdot \mathbf{B}^T \mathbf{B} \cdot (\mathbf{a}_1 - \mathbf{a}_2)\| + \frac{2 \cdot c_1}{n} \cdot \|\mathbf{a}_1 - \mathbf{a}_2\|.
$$

By applying twice the Cauchy-Schwarz inequality we get

$$
\begin{aligned}
\left\|\frac{2}{n} \cdot \mathbf{B}^T \mathbf{B} \cdot \mathbf{a}\right\|^2 &= \sum_{j=1}^{K} \left(\sum_{k=1}^{K} (\frac{2}{n} \sum_{i=1}^{n} B_j(x_i) \cdot B_k(x_i)) \cdot a_k\right)^2 \\
&\leq \sum_{j=1}^{K} \sum_{k=1}^{K} (\frac{2}{n} \sum_{i=1}^{n} B_j(x_i) \cdot B_k(x_i))^2 \cdot \|\mathbf{a}\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{j=1}^{K}\sum_{k=1}^{K} 4 \cdot \frac{1}{n}\sum_{i=1}^{n} B_j(x_i)^2 \cdot \frac{1}{n}\sum_{i=1}^{n} B_k(x_i))^2 \cdot \|\mathbf{a}\|^2 \\
&= \left(2 \cdot \sum_{k=1}^{K} \frac{1}{n}\sum_{i=1}^{n} B_k(x_i)^2\right)^2 \cdot \|\mathbf{a}\|^2,
\end{aligned}
$$

which implies the assertion.

□

3.2.2. Learning of Neural Networks Estimates with One Hidden Layer by Gradient Descent

We move on to the consideration of the scenario presented in Section 3.1 where both the inner and the outer weights of a neural network with one hidden layer are learned by gradient descent. We study neural networks with one hidden layer, which are defined by

$$
f_{net,(\mathbf{a},\mathbf{b})}(x) = \sum_{k=1}^{K} a_k \cdot \sigma\left(\sum_{j=1}^{d} b_{k,j} \cdot x^{(j)} + b_{k,0}\right) + a_0 \tag{3.12}
$$

(compare (3.1)), where $K \in \mathbb{N}$ is the number of neurons, the logistic squasher

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

is the activation function and where the weights

$$
a_k \quad (k = 0, \ldots, K) \quad \text{and} \quad b_{k,j} \in \mathbb{R} \quad (k = 1, \ldots, K, j = 0, \ldots, d)
$$

are learned by gradient descent. More precisely, we minimize

$$
F(\mathbf{a},\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n} |f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i|^2 + \frac{c_1}{n} \cdot \sum_{k=0}^{K} a_k^2 \tag{3.13}
$$

(compare (3.2)) by choosing an appropriate starting value $(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})$ and by setting

$$
\begin{pmatrix} \mathbf{a}^{(t+1)} \\ \mathbf{b}^{(t+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{a}^{(t)} \\ \mathbf{b}^{(t)} \end{pmatrix} - \lambda_n \cdot (\nabla_{(\mathbf{a},\mathbf{b})} F)(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) \tag{3.14}
$$

for some $\lambda_n > 0$ chosen below.

As described in Section 3.1, the concept for the construction of our neural network estimate is built on the observation that for appropriate inner weigths $b_{k,j}$ the neural network $f_{net,(\mathbf{a},\mathbf{b})}$ in (3.12) is close to a linear combination of indicator functions. We deal with this in Lemma 3.2.6. More precisely, we study the approximation of Hölder continuous functions by neural networks of the above form in the case of univariate functions and networks. To do this, we will need the auxiliary Lemma 2.2.13.

Further, we make the first key observation for our analysis, namely that in this scenario application of gradient descent affects the inner weights $b_{k,j}$ only slightly. We concern ourselves with this in Lemma 3.2.8 and in Lemma 3.2.9.

Finally, for the second key observation, we conclude from the above and from our results for linear least squares estimates in Section 3.2.1, that in this scenario application of gradient descent leads to a neural network estimate with optimally chosen outer weights a_k. We show this in Lemma 3.2.10.

Lemma 3.2.6. *Let σ be the logistic squasher. Let $\bar{\mathbf{c}} \in [-1,1]^d$ with $\|\bar{\mathbf{c}}\| = 1$ and let $g : \mathbb{R} \to \mathbb{R}$ be (p,C)-smooth for some $p \in (0,1]$ and $C > 0$. Let $\rho_n > 0$, $K \in \mathbb{N}$ and choose $b_1, b_2, \ldots, b_K \in \mathbb{R}$ such that $b_1 < b_2 < \cdots < b_K$ where*

$$b_1 \leq -A \cdot \sqrt{d} \quad \textit{and} \quad b_K \geq A \cdot \sqrt{d} - \frac{4 \cdot A \cdot \sqrt{d}}{K-1}$$

and

$$\frac{A \cdot \sqrt{d}}{(n+1) \cdot (K-1)} \leq |b_{k+1} - b_k| \leq \frac{4 \cdot A \cdot \sqrt{d}}{K-1} \quad (k = 1, \ldots, K-1).$$

Let

$$b_0 = b_1 - \frac{4 \cdot A \cdot \sqrt{d}}{K-1}$$

and set

$$a_0 = g(b_0) \quad \textit{and} \quad a_k = g(b_k) - g(b_{k-1}) \quad (k = 1, \ldots, K).$$

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

$$\sup_{x \in [-A,A]^d} \left| a_0 + \sum_{k=1}^{K} a_k \cdot \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - g(\bar{\mathbf{c}}^T x) \right|$$
$$\leq \frac{3 \cdot (4 \cdot A \cdot \sqrt{d})^p \cdot C}{(K-1)^p} + C \cdot (4 \cdot A \cdot \sqrt{d})^p \cdot (K-1)^{1-p} \cdot e^{-\frac{\rho_n \cdot (A \cdot \sqrt{d})}{(n+1) \cdot (K-1)}}.$$

Proof. Since for $x \in [-A, A]^d$ it holds that

$$|\bar{\mathbf{c}}^T x| = \sqrt{\left(\sum_{j=1}^{d} \bar{c}^{(j)} x^{(j)}\right)^2} \leq \sqrt{d \cdot A^2 \cdot \sum_{j=1}^{d} (\bar{c}^{(j)})^2} = \sqrt{d} \cdot A \cdot \|\bar{\mathbf{c}}\| = \sqrt{d} \cdot A$$

we have

$$\bar{\mathbf{c}}^T x \in [-\sqrt{d} \cdot A, \sqrt{d} \cdot A].$$

Further, we have

$$\begin{aligned} & \left| a_0 + \sum_{k=1}^{K} a_k \cdot \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - g(\bar{\mathbf{c}}^T x) \right| \\ \leq & \left| \sum_{k=1}^{K} a_k \cdot \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \sum_{k=1}^{K} a_k \cdot \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) \right| \\ & + \left| a_0 + \sum_{k=1}^{K} a_k \cdot \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) - g(\bar{\mathbf{c}}^T x) \right|. \end{aligned}$$

We bound the terms on the right-hand side separately. We start with the first term. In case that $b_j \leq \bar{\mathbf{c}}^T x < b_{j+1}$ ($j \in \{1, \ldots, K-1\}$) we conclude from the definition of a_k and from the (p, C)-smoothness of g that

$$\begin{aligned} \left| a_0 + \sum_{k=1}^{K} a_k \cdot \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) - g(\bar{\mathbf{c}}^T x) \right| &= \left| a_0 + \sum_{k=1}^{j} a_k - g(\bar{\mathbf{c}}^T x) \right| \\ &= |g(b_j) - g(\bar{\mathbf{c}}^T x)| \\ &\leq C \cdot |b_j - \bar{\mathbf{c}}^T x|^p \\ &\leq C \cdot |b_{j+1} - b_j|^p \\ &\leq \frac{C \cdot (4 \cdot A \cdot \sqrt{d})^p}{(K-1)^p}. \end{aligned}$$

In case that $b_K \leq \bar{\mathbf{c}}^T x \leq \sqrt{d} \cdot A$ we have by assumption

$$b_K \geq A \cdot \sqrt{d} - \frac{4 \cdot A \cdot \sqrt{d}}{K-1} \Leftrightarrow A \cdot \sqrt{d} - b_K \leq \frac{4 \cdot A \cdot \sqrt{d}}{K-1}$$

and hence, analogously to the previous case, we have

$$\left| a_0 + \sum_{k=1}^{K} a_k \cdot \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) - g(\bar{\mathbf{c}}^T x) \right| \leq C \cdot |b_K - \sqrt{d} \cdot A|^p$$

$$
\begin{aligned}
&\leq C \cdot |\sqrt{d} \cdot A - b_K|^p \\
&\leq \frac{C \cdot (4 \cdot A \cdot \sqrt{d})^p}{(K-1)^p}.
\end{aligned}
$$

So, we have shown

$$
\sup_{x \in [-A,A]^d} \left| a_0 + \sum_{k=1}^{K} a_k \cdot \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) - g(\bar{\mathbf{c}}^T x) \right| \leq \frac{C \cdot (4 \cdot A \cdot \sqrt{d})^p}{(K-1)^p}.
$$

Next, we consider the second term. Since

$$
\left| \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) \right| \in [0,1]
$$

we have in case that $b_j \leq \bar{\mathbf{c}}^T x \leq b_{j+1}$ $(j \in \{1, \ldots, K-1\})$

$$
\begin{aligned}
&\left| \sum_{k=1}^{K} a_k \cdot \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \sum_{k=1}^{K} a_k \cdot \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) \right| \\
\leq\ & \sum_{k=1}^{j-1} |a_k| \cdot \left| \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) \right| + |a_j| + |a_{j+1}| \\
&\quad + \sum_{k=j+2}^{K} |a_k| \cdot \left| \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) \right| \\
\leq\ & \max_{k=1,\ldots,K} |a_k| \\
&\quad \cdot \left(2 + (K-2) \cdot \max_{k \in \{1,2,\ldots,j-1,j+2,j+3,\ldots,K\}} \left| \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) \right| \right)
\end{aligned}
$$

and in case that $b_K \leq \bar{\mathbf{c}}^T x \leq \sqrt{d} \cdot A$

$$
\begin{aligned}
&\left| \sum_{k=1}^{K} a_k \cdot \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \sum_{k=1}^{K} a_k \cdot \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) \right| \\
\leq\ & \sum_{k=1}^{K-1} |a_k| \cdot \left| \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) \right| + |a_K| \\
\leq\ & \max_{k=1,\ldots,K} |a_k| \cdot \left(1 + (K-1) \cdot \max_{k \in \{1,2,\ldots,K-1\}} \left| \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x) \right| \right)
\end{aligned}
$$

$$\leq \max_{k=1,\dots,K} |a_k| \cdot \left(2 + (K-2) \cdot \max_{k\in\{1,2,\dots,K-1\}} \left|\sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x)\right|\right).$$

By definition of a_k and by the (p, C)-smoothness of g we have for $k = 1, \dots, K$

$$|a_k| = |g(b_k) - g(b_{k-1})| \leq C \cdot |b_k - b_{k-1}|^p \leq C \cdot \frac{(4 \cdot A \cdot \sqrt{d})^p}{(K-1)^p}.$$

Additionally, we know that in case that $b_j \leq \bar{\mathbf{c}}^T x \leq b_{j+1}$ $(j \in \{1, \dots, K-1\})$

$$|\bar{\mathbf{c}}^T x - b_k| \geq \frac{4 \cdot A \cdot \sqrt{d}}{(n+1) \cdot (K-1)}$$

for $k \in \{1, 2, \dots, j-1, j+2, j+3, \dots, K\}$ and in case that $b_K \leq \bar{\mathbf{c}}^T x \leq \sqrt{d} \cdot A$

$$|\bar{\mathbf{c}}^T x - b_k| \geq \frac{4 \cdot A \cdot \sqrt{d}}{(n+1) \cdot (K-1)}$$

for $k \in \{1, 2, \dots, K-1\}$. Together with Lemma 2.2.13 this implies in case that $b_j \leq \bar{\mathbf{c}}^T x \leq b_{j+1}$ $(j \in \{1, \dots, K-1\})$ it holds that

$$\begin{aligned}
&\left|\sum_{k=1}^{K} a_k \cdot \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \sum_{k=1}^{K} a_k \cdot \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x)\right| \\
\leq\ & C \cdot \frac{(4 \cdot A \cdot \sqrt{d})^p}{(K-1)^p} \cdot (2 + (K-2) \cdot \max_{k\in\{1,2,\dots,j-1,j+2,j+3,\dots,K\}} e^{-\rho_n \cdot |\bar{\mathbf{c}}^T x - b_k|}) \\
\leq\ & \frac{2 \cdot (4 \cdot A \cdot \sqrt{d})^p \cdot C}{(K-1)^p} + C \cdot (4 \cdot A \cdot \sqrt{d})^p \cdot (K-1)^{1-p} \cdot e^{-\frac{\rho_n \cdot (A \cdot \sqrt{d})}{(n+1) \cdot (K-1)}}
\end{aligned}$$

and in case that $b_K \leq \bar{\mathbf{c}}^T x \leq \sqrt{d} \cdot A$

$$\begin{aligned}
&\left|\sum_{k=1}^{K} a_k \cdot \sigma(\rho_n \cdot (\bar{\mathbf{c}}^T x - b_k)) - \sum_{k=1}^{K} a_k \cdot \mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x)\right| \\
\leq\ & C \cdot \frac{(4 \cdot A \cdot \sqrt{d})^p}{(K-1)^p} \cdot (2 + (K-2) \cdot \max_{k\in\{1,2,\dots K-1\}} e^{-\rho_n \cdot |\bar{\mathbf{c}}^T x - b_k|}) \\
\leq\ & \frac{2 \cdot (4 \cdot A \cdot \sqrt{d})^p \cdot C}{(K-1)^p} + C \cdot (4 \cdot A \cdot \sqrt{d})^p \cdot (K-1)^{1-p} \cdot e^{-\frac{\rho_n \cdot (A \cdot \sqrt{d})}{(n+1) \cdot (K-1)}}.
\end{aligned}$$

So, we have shown that for the second term

$$\begin{aligned}
&\sup_{x\in[-A,A]^d}\left|\sum_{k=1}^{K} a_k\cdot\sigma(\rho_n\cdot(\bar{\mathbf{c}}^T x-b_k))-\sum_{k=1}^{K} a_k\cdot\mathbf{1}_{[b_k,\infty)}(\bar{\mathbf{c}}^T x)\right| \\
\leq\ & \frac{2\cdot(4\cdot A\cdot\sqrt{d})^p\cdot C}{(K-1)^p}+C\cdot(4\cdot A\cdot\sqrt{d})^p\cdot(K-1)^{1-p}\cdot e^{-\frac{\rho_n\cdot(A\cdot\sqrt{d})}{(n+1)\cdot(K-1)}}.
\end{aligned}$$

Hence, we get

$$\begin{aligned}
&\left|a_0+\sum_{k=1}^{K} a_k\cdot\sigma(\rho_n\cdot(\bar{\mathbf{c}}^T x-b_k))-g(\bar{\mathbf{c}}^T x)\right| \\
\leq\ & \frac{C\cdot(4\cdot A\cdot\sqrt{d})^p}{(K-1)^p} \\
&+\frac{2\cdot(4\cdot A\cdot\sqrt{d})^p\cdot C}{(K-1)^p}+C\cdot(4\cdot A\cdot\sqrt{d})^p\cdot(K-1)^{1-p}\cdot e^{-\frac{\rho_n\cdot(A\cdot\sqrt{d})}{(n+1)\cdot(K-1)}} \\
=\ & \frac{3\cdot(4\cdot A\cdot\sqrt{d})^p\cdot C}{(K-1)^p}+C\cdot(4\cdot A\cdot\sqrt{d})^p\cdot(K-1)^{1-p}\cdot e^{-\frac{\rho_n\cdot(A\cdot\sqrt{d})}{(n+1)\cdot(K-1)}}
\end{aligned}$$

which concludes the proof.

□

Remark 3.2.7. *Note, that b_0 in Lemma 3.2.6 has nothing to do with our chosen grid points $b_1, b_2, \ldots, b_K$ but is needed for the defintion of $a_0, a_1, \ldots, a_K$. This does not pose a problem in our anaylsis later on, since we will show in Lemma 3.2.10 that the output weights are chosen optimally. We will use Lemma 3.2.6 in the proof of Theorem 3.2.1 to provide an upper bound. Naturally, $\min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(0)})\leq F(\mathbf{a}^*,\mathbf{b}^{(0)})$ for any $\mathbf{a}^*\in\mathbb{R}^{K+1}$, in particular, for the values in Lemma 3.2.6.*

In the next lemma we take a look at what happens to the inner weights $b_{k,j}$ $(k=1,\ldots,K, j=0,1,\ldots,d)$ after one gradient descent step applied to $F(\mathbf{a},\mathbf{b})$.

Lemma 3.2.8. *Let σ be the logistic squasher. Define F by (3.13) and set*

$$\bar{\mathbf{b}}=\mathbf{b}-\lambda_n\cdot(\nabla_{\mathbf{b}}F)(\mathbf{a},\mathbf{b})$$

for some $\lambda_n>0$, where

$$\mathbf{a}=(a_0,a_1,\ldots,a_K)^T\in\mathbb{R}^{K+1}$$

and

$$\mathbf{b} = (b_{1,0}, b_{1,1}, \ldots, b_{1,d}, \ldots, b_{K,0}, b_{K,1} \ldots, b_{K,d})^T \in \mathbb{R}^{K\cdot(d+1)}.$$

Then we have for any $k \in \{1, \ldots, K\}$ *and any* $j \in \{0, \ldots, d\}$

$$\begin{aligned} |\bar{b}_{k,j} - b_{k,j}| \quad &\leq \quad \lambda_n \cdot 2 \cdot \sqrt{F(\mathbf{a}, \mathbf{b})} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot |a_k| \\ &\qquad \cdot \exp\left(-\min_{i=1,\ldots,n}\left\{\left|\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right|\right\}\right). \end{aligned}$$

Proof. Using

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

(where the first inequality holds due to $\sigma(x) \in [0, 1]$) we can conclude from Lemma 2.2.13 that

$$\begin{aligned} \max_{i=1,\ldots,n}\left|\sigma'\left(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right)\right| \quad &\leq \quad \max_{i=1,\ldots,n} \exp\left(-\left|\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right|\right) \\ &= \quad \exp\left(-\min_{i=1,\ldots,n}\left\{\left|\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right|\right\}\right). \end{aligned}$$

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

$$\begin{aligned} &\left|\frac{\partial F}{\partial b_{k,j}}(\mathbf{a}, \mathbf{b})\right| \\ = \quad &\left|\frac{2}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i) \cdot a_k \cdot \sigma'\left(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right) \cdot x_i^{(j)}\right| \\ \leq \quad &2 \cdot |a_k| \cdot \frac{1}{n}\sum_{i=1}^{n}|f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i| \cdot |x_i^{(j)}| \cdot \left|\sigma'\left(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right)\right| \\ \leq \quad &2 \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}|f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i|^2 \cdot (x_i^{(j)})^2} \cdot |a_k| \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\sigma'(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0})|^2} \end{aligned}$$

$$
\begin{aligned}
&\leq 2 \cdot \sqrt{F(\mathbf{a},\mathbf{b})} \cdot \max_{i,l}\{|x_i^{(l)}|\} \cdot |a_k| \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\sigma'(\sum_{l=1}^{d} b_{k,l} \cdot x_i^{(l)} + b_{k,0})|^2} \\
&\leq 2 \cdot \sqrt{F(\mathbf{a},\mathbf{b})} \cdot \max_{i,l}\{|x_i^{(l)}|\} \cdot |a_k| \cdot \exp\left(-\min_{i=1,\ldots,n}\left\{\left|\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right|\right\}\right).
\end{aligned}
$$

Hence, we have shown

$$
\begin{aligned}
&|\bar{b}_{k,j} - b_{k,j}| \\
&= \lambda_n \cdot \left|\frac{\partial F}{\partial b_{k,j}}(\mathbf{a},\mathbf{b})\right| \\
&\leq \lambda_n \cdot 2 \cdot \sqrt{F(\mathbf{a},\mathbf{b})} \cdot \max_{i,l}\{|x_i^{(l)}|\} \cdot |a_k| \cdot \exp\left(-\min_{i=1,\ldots,n}\left\{\left|\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right|\right\}\right)
\end{aligned}
$$

for any $k \in \{1,\ldots,K\}$ and any $j \in \{1,\ldots,d\}$.

In case that $k \in \{1,\ldots,K\}$ and $j = 0$ we get in a similar fashion

$$
\begin{aligned}
&\left|\frac{\partial F}{\partial b_{k,0}}(\mathbf{a},\mathbf{b})\right| \\
&= \left|\frac{2}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i) \cdot a_k \cdot \sigma'\left(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right) \cdot 1\right| \\
&\leq 2 \cdot |a_k| \cdot \frac{1}{n}\sum_{i=1}^{n} |f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i| \cdot 1 \cdot \left|\sigma'\left(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right)\right| \\
&\leq 2 \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i|^2 \cdot 1^2} \cdot |a_k| \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\sigma'(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0})|^2} \\
&\leq 2 \cdot \sqrt{F(\mathbf{a},\mathbf{b})} \cdot 1 \cdot |a_k| \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\sigma'(\sum_{l=1}^{d} b_{k,l} \cdot x_i^{(l)} + b_{k,0})|^2} \\
&\leq 2 \cdot \sqrt{F(\mathbf{a},\mathbf{b})} \cdot 1 \cdot |a_k| \cdot \exp\left(-\min_{i=1,\ldots,n}\left\{\left|\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right|\right\}\right)
\end{aligned}
$$

and thus

$$\begin{aligned}|\bar{b}_{k,0} - b_{k,0}| &= \lambda_n \cdot \left|\frac{\partial F}{\partial b_{k,0}}(\mathbf{a},\mathbf{b})\right| \\ &\leq \lambda_n \cdot 2 \cdot \sqrt{F(\mathbf{a},\mathbf{b})} \cdot 1 \cdot |a_k| \cdot \exp\left(-\min_{i=1,\ldots,n}\left\{\left|\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right|\right\}\right),\end{aligned}$$

which implies the assertion. □

Next, we conclude from the previous lemma that if for the initially chosen inner weights

$$\min_{i=1,\ldots,n,k=1,\ldots,K}\left|\sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)}\right|$$

is large, the inner weights $b_{k,j}$ $(k = 1,\ldots,K, j = 0,1,\ldots,d)$ change only slightly in every gradient descent step.

Lemma 3.2.9. *Define F by* (3.13) *and define* $(\mathbf{a}^{(t)},\mathbf{b}^{(t)})$ *by* (3.14). *Assume that* $(\mathbf{a}^{(t)},\mathbf{b}^{(t)})$ *satisfy for* $t \in \{0,\ldots,t_n - 1\}$

$$F(\mathbf{a}^{(t)},\mathbf{b}^{(t)}) \leq c_{5,n} < \infty, \tag{3.15}$$

$$\|\mathbf{a}^{(t)}\|^2 \leq c_{6,n} \cdot n < \infty, \tag{3.16}$$

$$\min_{i=1,\ldots,n,k=1,\ldots,K}\left|\sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)}\right| \geq \delta_n > 0 \tag{3.17}$$

and

$$(d+1) \cdot t_n \cdot \lambda_n \cdot 2 \cdot \sqrt{c_{5,n}} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|^2\}\} \cdot \sqrt{c_{6,n} \cdot n} \cdot \exp(-\delta_n/2) \leq \frac{\delta_n}{2}. \tag{3.18}$$

Then we have for every $k \in \{1,\ldots,K\}$, any $j \in \{0,\ldots,d\}$ and any $t \in \{1,\ldots,t_n\}$

$$|b_{k,j}^{(t)} - b_{k,j}^{(t-1)}| \leq \lambda_n \cdot 2 \cdot \sqrt{c_{5,n}} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{c_{6,n} \cdot n} \cdot \exp(-\delta_n/2). \tag{3.19}$$

Proof. We show (3.19) by induction on t.
Start of the induction. For $t = 1$ we apply Lemma 3.2.8 together with (3.15)-(3.17). This yields for $k \in \{1, \dots, K\}$ and for $j \in \{0, \dots, d\}$

$$\begin{aligned} |b_{k,j}^{(1)} - b_{k,j}^{(0)}| &\leq \lambda_n \cdot 2 \cdot \sqrt{F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot |a_k^{(0)}| \\ &\qquad \cdot \exp\left(-\min_{i=1,\dots,n}\left\{\left|\sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)}\right|\right\}\right) \\ &\leq \lambda_n \cdot 2 \cdot \sqrt{c_{5,n}} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{c_{6,n} \cdot n} \cdot \exp\left(-\delta_n/2\right). \end{aligned}$$

Induction hypothesis. We assume that inequality (3.19) holds for all $t \in \{1, \dots, s\}$, where $s \in \{1, \dots, t_n - 1\}$.
Induction step. Application of Lemma 3.2.8 together with (3.15) and (3.16) gives us

$$\begin{aligned} |b_{k,j}^{(s+1)} - b_{k,j}^{(s)}| &\leq \lambda_n \cdot 2 \cdot \sqrt{F(\mathbf{a}^{(s)}, \mathbf{b}^{(s)})} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot |a_k^{(s)}| \\ &\qquad \cdot \exp\left(-\min_{i=1,\dots,n}\left\{\left|\sum_{j=1}^{d} b_{k,j}^{(s)} \cdot x_i^{(j)} + b_{k,0}^{(s)}\right|\right\}\right) \\ &\leq \lambda_n \cdot 2 \cdot \sqrt{c_{5,n}} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{c_{6,n} \cdot n} \\ &\qquad \cdot \exp\left(-\min_{i=1,\dots,n}\left\{\left|\sum_{j=1}^{d} b_{k,j}^{(s)} \cdot x_i^{(j)} + b_{k,0}^{(s)}\right|\right\}\right). \end{aligned} \tag{3.20}$$

We need to bound the sum in the exponential term. For this, we observe that with the induction hypothesis it holds that

$$\begin{aligned} |b_{k,j}^{(s)} - b_{k,j}^{(0)}| &\leq |b_{k,j}^{(s)} - b_{k,j}^{(s-1)}| + |b_{k,j}^{(s-1)} - b_{k,j}^{(s-2)}| + \dots + |b_{k,j}^{(1)} - b_{k,j}^{(0)}| \\ &\leq t_n \cdot \lambda_n \cdot 2 \cdot \sqrt{c_{5,n}} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{c_{6,n} \cdot n} \cdot \exp\left(-\delta_n/2\right). \end{aligned}$$

From this, together with the inverse triangle inequality and with assumption (3.17), we conclude that

$$\min_{i=1,\dots,n, k=1,\dots,K}\left|\sum_{j=1}^{d} b_{k,j}^{(s)} \cdot x_i^{(j)} + b_{k,0}^{(s)}\right|$$

$$
\begin{aligned}
&= \min_{i=1,\dots,n,k=1,\dots,K} \left| \left(\sum_{j=1}^{d} b_{k,j}^{(s)} \cdot x_i^{(j)} + b_{k,0}^{(s)} - \left(\sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right) \right) \right. \\
&\qquad \left. + \left(\sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right) \right| \\
&\geq \min_{i=1,\dots,n,k=1,\dots,K} \left| \left| \sum_{j=1}^{d} b_{k,j}^{(s)} \cdot x_i^{(j)} + b_{k,0}^{(s)} - \left(\sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right) \right| \right. \\
&\qquad \left. - \left| \sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right| \right| \\
&\geq \min_{i=1,\dots,n,k=1,\dots,K} \left(\left| \sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right| \right. \\
&\qquad \left. - \left| \sum_{j=1}^{d} b_{k,j}^{(s)} \cdot x_i^{(j)} + b_{k,0}^{(s)} - \left(\sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right) \right| \right) \\
&\geq \min_{i=1,\dots,n,k=1,\dots,K} \left| \sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right| \\
&\qquad - \max_{i=1,\dots,n,k=1,\dots,K} \left(\sum_{j=1}^{d} |b_{k,j}^{(s)} - b_{k,j}^{(0)}| \cdot |x_i^{(j)}| + |b_{k,0}^{(s)} - b_{k,0}^{(0)}| \right) \\
&\geq \delta_n - \max_{i=1,\dots,n,k=1,\dots,K} \left(\sum_{j=0}^{d} |b_{k,j}^{(s)} - b_{k,j}^{(0)}| \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \right) \\
&\geq \delta_n - (d+1) \cdot t_n \cdot \lambda_n \cdot 2 \cdot \sqrt{c_{5,n}} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|^2\}\} \cdot \sqrt{c_{6,n} \cdot n} \cdot \exp\left(-\delta_n/2\right) \\
&\geq \frac{\delta_n}{2}, \qquad (3.21)
\end{aligned}
$$

where the last inequality is implied by inequality (3.18). So, by (3.20) and by (3.21) we get

$$
|b_{k,j}^{(s+1)} - b_{k,j}^{(s)}| \leq \lambda_n \cdot 2 \cdot \sqrt{c_{5,n}} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{c_{6,n} \cdot n} \cdot \exp\left(-\delta_n/2\right).
$$

This concludes the proof. □

Finally, we conclude that if for the initially chosen inner weights

$$\min_{i=1,\dots,n,k=1,\dots,K} \left| \sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right|$$

is large and $F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})$ is bounded, the neural network learned by gradient descent finds the optimal output weights.

Lemma 3.2.10. *Define F by* (3.13)*, set*

$$\lambda_n = \frac{1}{3 \cdot K}$$

and define $(\mathbf{a}^{(t)}, \mathbf{b}^{(t)})$ *by* (3.14)*. Assume that* $(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})$ *is chosen such that*

$$F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) \leq c_{5,n} < \infty \tag{3.22}$$

and

$$\min_{i=1,\dots,n,k=1,\dots,K} \left| \sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right| \geq \delta_n \geq 1 \tag{3.23}$$

hold. Let $t_n \in \mathbb{N}$ and assume $2 \cdot c_1 \leq (K-2) \cdot n$,

$$\begin{aligned} 4 \cdot \max\{1, \frac{c_{5,n}}{c_1}\} \cdot \max\{1, \frac{1}{c_1^2}\} \cdot \lambda_n \cdot (d+1)^2 \cdot n^2 \cdot \max\{1, \max_{i,j} |x_i^{(j)}|^4\} \\ \cdot \left(1 + c_{5,n} + \frac{2}{n} \sum_{i=1}^{n} y_i^2\right)^4 \cdot t_n^2 \cdot \exp\left(-\delta_n/2\right) \leq 1 \end{aligned} \tag{3.24}$$

and

$$3 \cdot t_n \cdot \exp(-\delta_n/4) \leq 1. \tag{3.25}$$

Then for any $t \in \{0, 1, \dots, t_n - 1\}$ we have

$$\begin{aligned} & F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\ \leq \; & \left(1 - \frac{2 \cdot c_1}{3 \cdot K \cdot n}\right)^{t+1} \cdot \left(F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) + (2\sqrt{c_{5,n}} + 1) \cdot \exp\left(-\delta_n/4\right) \\ & + \frac{3 \cdot K \cdot n}{2 \cdot c_1} \cdot 3 \cdot \exp\left(-\delta_n/4\right). \end{aligned}$$

The proof of Lemma 3.2.10 is quite long and technical. For a better understanding we present a brief and highly simplified outline of the proof before going into detail. The proof has five steps.

Step 1: Starting the proof. By adding zeros we can rewrite the left-hand side as the sum of three terms. We have

$$F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) = \text{①} + \text{②} + \text{③}$$

with

$$\text{②} = F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})$$

where the terms differ only in the $\mathbf{a}$ component and with

$$\begin{aligned} \text{①} &= F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) \\ \text{③} &= \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}). \end{aligned}$$

Step 2: Looking at the second summand. In ② we are looking at the components $\mathbf{a}^{(t+1)}$ and $\mathbf{b}^{(t)}$ which we can interpret as making one gradient descent step only in the $\mathbf{a}$ component (outer weights) and leaving the $\mathbf{b}$ component (inner weights) untouched. This allows us to apply our results in Section 3.2.1 which gives us

$$F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) \leq \left(1 - \frac{4 \cdot c_1}{6 \cdot K \cdot n}\right) \cdot \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})\right).$$

Step 3: Bounding the third summand. In ③ we choose $\bar{\mathbf{a}}$ such that

$$F(\bar{\mathbf{a}}, \mathbf{b}^{(0)}) = \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})$$

which allows us to reduce the term to the difference of the $\mathbf{b}$ components. We get

$$\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \leq 2 \cdot \sqrt{F(\bar{\mathbf{a}}, \mathbf{b}^{(0)})} \cdot \sqrt{\text{✧}} + \text{✧}$$

with

$$\text{✧} = \frac{1}{n} \sum_{i=1}^{n} (f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(0)})}(x_i))^2$$

$$\leq \sum_{k=1}^{K} \bar{a}_k^2 \cdot \max\{1, \max_{i,j} |x_i^{(j)}|^2\} \cdot (d+1) \cdot \sum_{k=1}^{K} \sum_{j=0}^{d} |b_{k,j}^{(t)} - b_{k,j}^{(0)}|^2.$$

We bound this difference further by Lemma 3.2.9 which yields

$$\text{✧} \leq \exp\left(-\delta_n/2\right).$$

This gives us

$$③ \leq (2 \cdot \sqrt{c_{5,n}} + 1) \cdot \exp\left(-\delta_n/4\right) =: \beta_1.$$

Step 4: Showing that Lemma 3.2.9 is applicable in Step 3 (Bounding the first summand). We verify the conditions of Lemma 3.2.9. For that we show by induction on t the following claim consisting of two inequalities:

1. $F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \leq c_{5,n} + \frac{1}{n} \sum_{i=1}^{n} y_i^2 + 3 \cdot (t+1) \cdot \exp(-\delta_n/4)$
2. $F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \leq c_{5,n} + \frac{1}{n} \sum_{i=1}^{n} y_i^2 + 3 \cdot (t+1) \cdot \exp(-\delta_n/4)$

While showing the second inequality we derive a bound on ① in the process. We get

$$F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \leq ① + c_{5,n} + \frac{1}{n} \sum_{i=1}^{n} y_i^2 + 3t \cdot \exp(-\delta_n/4)$$

and

$$① \leq 3 \cdot \exp\left(-\delta_n/4\right) =: \beta_2.$$

Step 5: Finishing the proof. For simplicity we introduce the following notation

$$\gamma_t = \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) \quad \text{and} \quad \alpha = \frac{2 \cdot c_1}{3 \cdot K \cdot n}.$$

We apply the previous steps yielding

$$\begin{aligned} F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) &\leq (1-\alpha) \cdot \gamma_t + \alpha \cdot \beta_1 + \beta_2 \\ &\leq (1-\alpha)^{t+1} \cdot \gamma_0 + \beta_1 + \frac{\beta_2}{\alpha}. \end{aligned}$$

Now we give the detailed proof.

Proof. We prove the assertion in five steps.
Step 1: Starting the proof. We have for any $t \in \{0, 1, \ldots, t_n - 1\}$

$$\begin{aligned}
&F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
&= \left(F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})\right) + \left(F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})\right) \\
&\quad + \left(\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right).
\end{aligned}$$

Step 2: Looking at the second summand. We take a look at the second summand on the right-hand side of Step 1. We show that

$$F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) \leq \left(1 - \frac{4 \cdot c_1}{6 \cdot K \cdot n}\right) \cdot \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})\right).$$

The trick is that the extension in Step 1 results in a term where we look at components $\mathbf{a}^{(t+1)}$ and $\mathbf{b}^{(t)}$. We can interpret this as going from $(\mathbf{a}^{(t)}, \mathbf{b}^{(t)})$ to $(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})$ meaning we make one gradient descent step in the $\mathbf{a}$ component (outer weights) while leaving the $\mathbf{b}$ component (inner weights) untouched. This allows us to apply our results in Section 3.2.1. We bound the second summand by Lemma 3.2.4. In order to do so, we need to verify the conditions of Lemma 3.2.4. First, we check (3.10), i.e.

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}_1) - (\nabla_{\mathbf{a}} F)(\mathbf{a}_2)\| \leq L_n \cdot \|\mathbf{a}_1 - \mathbf{a}_2\| \quad (\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^{K+1}).$$

For this, we apply Lemma 3.2.5 (with shifted index) with

$$\begin{aligned}
F(\mathbf{a}, \mathbf{b}) &= \frac{1}{n} \sum_{i=1}^{n} |f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i|^2 + \frac{c_1}{n} \cdot \sum_{k=0}^{K} a_k^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} |\sum_{k=1}^{K} a_k \cdot \sigma\left(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right) + a_0 - y_i|^2 + \frac{c_1}{n} \cdot \sum_{k=0}^{K} a_k^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} |\sum_{k=0}^{K} a_k \cdot B_k(x_i) - y_i|^2 + \frac{c_1}{n} \cdot \|\mathbf{a}\|^2
\end{aligned}$$

where

$$B_0(x_i) = 1 \quad (i = 1, \ldots, n)$$

and for $k = 1, \ldots, K$

$$B_k(x_i) = \sigma\left(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0}\right) \leq 1 \quad (i = 1, \ldots, n)$$

which yields for $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^{K+1}$

$$\begin{aligned}
\left\|(\nabla_{\mathbf{a}} F)(\mathbf{a}_1, \mathbf{b}^{(t)}) - (\nabla_{\mathbf{a}} F)(\mathbf{a}_2, \mathbf{b}^{(t)})\right\| &\leq \left(2 \cdot \sum_{k=0}^{K} \frac{1}{n} \sum_{i=1}^{n} B_k(x_i)^2 + \frac{2 \cdot c_1}{n}\right) \cdot \|\mathbf{a}_1 - \mathbf{a}_2\| \\
&\leq \left(2 \cdot \sum_{k=0}^{K} \frac{1}{n} \sum_{i=1}^{n} 1 + \frac{2 \cdot c_1}{n}\right) \cdot \|\mathbf{a}_1 - \mathbf{a}_2\| \\
&\leq \left(2 \cdot (K+1) + \frac{2 \cdot c_1}{n}\right) \cdot \|\mathbf{a}_1 - \mathbf{a}_2\| \\
&\leq 3 \cdot K \cdot \|\mathbf{a}_1 - \mathbf{a}_2\|.
\end{aligned}$$

Second, we check (3.11), i.e.

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2 \geq \rho_n \cdot (F(\mathbf{a}) - F(\mathbf{a}_{opt})) \quad (\mathbf{a} \in \mathbb{R}^{K+1}).$$

For this, we apply Lemma 2.2.12 (with $\tau_n = c_1/n$) which yields for $\mathbf{a} \in \mathbb{R}^{K+1}$

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2 \geq \frac{4 \cdot c_1}{n} \cdot (F(\mathbf{a}) - \min_{\mathbf{a} \in \mathbb{R}^{K+1}} F(\mathbf{a})).$$

So, application of Lemma 3.2.4 gives us

$$\begin{aligned}
& F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) \\
\leq\ & \left(1 - \frac{\frac{4 \cdot c_1}{n}}{2 \cdot (3 \cdot K)}\right) \cdot \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})\right) \\
=\ & \left(1 - \frac{4 \cdot c_1}{6 \cdot K \cdot n}\right) \cdot \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})\right). \qquad (3.26)
\end{aligned}$$

Step 3: Bounding the third summand. We want to derive an upper bound β_1 on the third summand on the right-hand side of Step 1. We show that

$$\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \leq (2 \cdot \sqrt{c_{5,n}} + 1) \cdot \exp(-\delta_n/4) =: \beta_1.$$

For that, we choose $\bar{\mathbf{a}}$ such that

$$F(\bar{\mathbf{a}}, \mathbf{b}^{(0)}) = \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}).$$

Then, by assumption we get

$$\frac{c_1}{n}\sum_{k=0}^{K}\bar{a}_k^2 \leq \frac{c_1}{n}\sum_{k=0}^{K}\bar{a}_k^2 + \frac{1}{n}\sum_{i=1}^{n}|f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i) - y_i|^2 = F(\bar{\mathbf{a}}, \mathbf{b}^{(0)}) \leq F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) \leq c_{5,n}$$

and hence,

$$\sum_{k=0}^{K}\bar{a}_k^2 \leq \frac{c_{5,n}}{c_1}\cdot n. \tag{3.27}$$

We have by the Cauchy-Schwarz inequality

$$\begin{aligned}
& \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
= \; & \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - F(\bar{\mathbf{a}}, \mathbf{b}^{(0)}) \\
\leq \; & F(\bar{\mathbf{a}}, \mathbf{b}^{(t)}) - F(\bar{\mathbf{a}}, \mathbf{b}^{(0)}) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\left(|f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - y_i|^2 + \frac{c_1}{n}\sum_{k=0}^{K}\bar{a}_k^2 - |f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i) - y_i|^2 - \frac{c_1}{n}\sum_{k=0}^{K}\bar{a}_k^2\right) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\left(|f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - y_i|^2 - |f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i) - y_i|^2\right) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - y_i + (f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i) - y_i)) \\
& \qquad \cdot(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - y_i - (f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i) - y_i)) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) + f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i) - 2y_i)\cdot(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i)) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}(2f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i) - 2y_i)\cdot(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i)) \\
& + \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2 \\
\leq \; & \sqrt{\frac{1}{n}\sum_{i=1}^{n}(2\cdot f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i) - 2\cdot y_i)^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2}
\end{aligned}$$

$$
\begin{aligned}
& +\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2 \\
\leq\ & 2\cdot\sqrt{F(\bar{\mathbf{a}},\mathbf{b}^{(0)})}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2} \\
& +\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2.
\end{aligned}
$$

We bound the term

$$
\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2
$$

further. Applying the Cauchy-Schwarz inequality a second time and Lipschitz continuity of σ (with Lipschitz constant 1) yield

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2 \\
=\ & \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\bar{a}_k\cdot\sigma\left(\sum_{j=1}^{d}b_{k,j}^{(t)}\cdot x_i^{(j)} + b_{k,0}^{(t)}\right) + \bar{a}_0\right. \\
& \qquad\left. -\left(\sum_{k=1}^{K}\bar{a}_k\cdot\sigma\left(\sum_{j=1}^{d}b_{k,j}^{(0)}\cdot x_i^{(j)} + b_{k,0}^{(0)}\right) + \bar{a}_0\right)\right)^2 \\
=\ & \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\bar{a}_k\cdot\left(\sigma\left(\sum_{j=1}^{d}b_{k,j}^{(t)}\cdot x_i^{(j)} + b_{k,0}^{(t)}\right) - \sigma\left(\sum_{j=1}^{d}b_{k,j}^{(0)}\cdot x_i^{(j)} + b_{k,0}^{(0)}\right)\right)\right)^2 \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\bar{a}_k^2\cdot\sum_{k=1}^{K}\left(\sigma\left(\sum_{j=1}^{d}b_{k,j}^{(t)}\cdot x_i^{(j)} + b_{k,0}^{(t)}\right) - \sigma\left(\sum_{j=1}^{d}b_{k,j}^{(0)}\cdot x_i^{(j)} + b_{k,0}^{(0)}\right)\right)^2\right) \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\bar{a}_k^2\cdot\sum_{k=1}^{K}\left(\sum_{j=1}^{d}|b_{k,j}^{(t)}\cdot x_i^{(j)} + b_{k,0}^{(t)} - b_{k,j}^{(0)}\cdot x_i^{(j)} - b_{k,0}^{(0)}|\right)^2\right) \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\bar{a}_k^2\cdot\sum_{k=1}^{K}\left(\sum_{j=1}^{d}|b_{k,j}^{(t)} - b_{k,j}^{(0)}|\cdot x_i^{(j)} + |b_{k,0}^{(t)} - b_{k,0}^{(0)}|\right)^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\bar{a}_k^2\cdot\sum_{k=1}^{K}\left(\sum_{j=0}^{d}|b_{k,j}^{(t)}-b_{k,j}^{(0)}|\cdot\max\{1,\max_{i,j}|x_i^{(j)}|\}\right)^2\right)\\
&= \sum_{k=1}^{K}\bar{a}_k^2\cdot\sum_{k=1}^{K}\left(\max\{1,\max_{i,j}|x_i^{(j)}|\}\right)^2\cdot\left(\sum_{j=0}^{d}|b_{k,j}^{(t)}-b_{k,j}^{(0)}|\right)^2\\
&\leq \sum_{k=1}^{K}\bar{a}_k^2\cdot\sum_{k=1}^{K}\max\{1,\max_{i,j}|x_i^{(j)}|^2\}\cdot(d+1)\cdot\sum_{j=0}^{d}|b_{k,j}^{(t)}-b_{k,j}^{(0)}|^2\\
&= \sum_{k=1}^{K}\bar{a}_k^2\cdot\max\{1,\max_{i,j}|x_i^{(j)}|^2\}\cdot(d+1)\cdot\sum_{k=1}^{K}\sum_{j=0}^{d}|b_{k,j}^{(t)}-b_{k,j}^{(0)}|^2.
\end{aligned}
$$

Now, we consider

$$
|b_{k,j}^{(t)}-b_{k,j}^{(0)}|\leq|b_{k,j}^{(t)}-b_{k,j}^{(t-1)}|+|b_{k,j}^{(t-1)}-b_{k,j}^{(t-2)}|+\cdots+|b_{k,j}^{(1)}-b_{k,j}^{(0)}|
$$

and we apply to each term Lemma 3.2.9 with

$$
F(\mathbf{a}^{(t)},\mathbf{b}^{(t)})\leq 1+c_{5,n}+\frac{2}{n}\sum_{i=1}^{n}y_i^2
$$

and

$$
\|\mathbf{a}^{(t)}\|^2\leq\left(1+c_{5,n}+\frac{2}{n}\sum_{i=1}^{n}y_i^2\right)\cdot\frac{1}{c_1}\cdot n
$$

for $t\in\{0,\ldots,t_n-1\}$. We will show in Step 4 that Lemma 3.2.9 is, in fact, applicable here. This gives us

$$
\begin{aligned}
&|b_{k,j}^{(t)}-b_{k,j}^{(0)}|\\
\leq\ & t\cdot\lambda_n\cdot 2\cdot\sqrt{1+c_{5,n}+\frac{2}{n}\sum_{i=1}^{n}y_i^2}\cdot\max\{1,\max_{i,l}\{|x_i^{(l)}|\}\}\\
&\cdot\sqrt{\left(1+c_{5,n}+\frac{2}{n}\sum_{i=1}^{n}y_i^2\right)\cdot\frac{1}{c_1}\cdot n}\cdot\exp\left(-\delta_n/2\right)\\
\leq\ & t\cdot\lambda_n\cdot 2\cdot\sqrt{1+c_{5,n}+\frac{2}{n}\sum_{i=1}^{n}y_i^2}\cdot\max\{1,\max_{i,l}\{|x_i^{(l)}|\}\}
\end{aligned}
$$

$$
\begin{aligned}
& \cdot \sqrt{1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2} \cdot \sqrt{\frac{1}{c_1}} \cdot \sqrt{n} \cdot \exp\left(-\delta_n/2\right) \\
\leq \quad & t \cdot \lambda_n \cdot 2 \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \\
& \cdot \max\{1, \frac{1}{c_1}\} \cdot \sqrt{n} \cdot \exp\left(-\delta_n/2\right).
\end{aligned}
$$

From this we conclude with (3.27), with

$$
K \cdot \lambda_n^2 = K \cdot \frac{1}{3 \cdot K} \cdot \lambda_n \leq \lambda_n
$$

by definition of λ_n and with

$$
0 \leq F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) \leq c_{5,n}
$$

by definiton of $c_{5,n}$ that

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2 \\
\leq \quad & \sum_{k=1}^{K} \bar{a}_k^2 \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1) \cdot K \cdot (d+1) \\
& \cdot \left(t \cdot 2 \cdot \lambda_n \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \max\{1, \frac{1}{c_1}\}\right. \\
& \left. \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{n} \cdot \exp\left(-\delta_n/2\right)\right)^2 \\
\leq \quad & \frac{c_{5,n}}{c_1} \cdot n \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1)^2 \cdot K \\
& \cdot t^2 \cdot 4 \cdot \lambda_n^2 \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)^2 \cdot \max\{1, \frac{1}{c_1^2}\} \\
& \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|^2\}\} \cdot n \cdot \exp\left(-\delta_n\right) \\
\leq \quad & 4 \cdot t_n^2 \cdot \max\{1, \frac{c_{5,n}}{c_1}\} \cdot \max\{1, \frac{1}{c_1^2}\} \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)^4 \cdot n^2
\end{aligned}
$$

$$
\begin{aligned}
& \cdot \max\{1, \max_{i,j} |x_i^{(j)}|^4\} \cdot (d+1)^2 \cdot \lambda_n \cdot \exp(-\delta_n) \\
\leq \quad & \exp(-\delta_n/2),
\end{aligned}
$$

where the last inequality holds by (3.24). Since

$$
\exp(-\delta_n/2) \leq 1
$$

we have

$$
\sqrt{\exp(-\delta_n/2)} \geq \exp(-\delta_n/2).
$$

Hence, plugging in yields

$$
\begin{aligned}
& \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
\leq \quad & 2 \cdot \sqrt{F(\bar{\mathbf{a}}, \mathbf{b}^{(0)})} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} (f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2} \\
& + \frac{1}{n} \sum_{i=1}^{n} (f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2 \\
\leq \quad & 2 \cdot \sqrt{F(\bar{\mathbf{a}}, \mathbf{b}^{(0)})} \cdot \sqrt{\exp(-\delta_n/2)} + \exp(-\delta_n/2) \\
\leq \quad & 2 \cdot \sqrt{F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})} \cdot \sqrt{\exp(-\delta_n/2)} + \sqrt{\exp(-\delta_n/2)} \\
\leq \quad & (2 \cdot \sqrt{F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})} + 1) \cdot \sqrt{\exp(-\delta_n/2)} \\
\leq \quad & (2 \cdot \sqrt{c_{5,n}} + 1) \cdot \exp(-\delta_n/4) \\
=: \quad & \beta_1.
\end{aligned}
$$

Step 4: Showing that Lemma 3.2.9 *is applicable in Step 3 (Bounding the first summand).* We show that the application of Lemma 3.2.9 in Step 3 is justified. In the process we derive an upper bound β_2 for the first summand on the right-hand side of Step 1. We show that

$$
F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) \leq 3 \cdot \exp(-\delta_n/4) =: \beta_2.
$$

To avoid confusion and for readability we will denote by

$c_{5,n,Le9}$	the constant $c_{5,n}$ from Lemma 3.2.9
$c_{6,n,Le9}$	the constant $c_{6,n}$ from Lemma 3.2.9 .

We verify the four conditions of Lemma 3.2.9 for

$$c_{5,n,Le9} = 1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2 < \infty$$

and

$$c_{6,n,Le9} = \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \frac{1}{c_1} < \infty.$$

First, condition (3.17), i.e.

$$\min_{i=1,\dots,n,k=1,\dots,K} \left|\sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)}\right| \geq \delta_n > 0$$

is met by assumption (3.23). Second, we check condition (3.18), i.e.

$$(d+1) \cdot t_n \cdot \lambda_n \cdot 2 \cdot \sqrt{c_{5,n,Le9}} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|^2\}\} \cdot \sqrt{c_{6,n,Le9} \cdot n} \cdot \exp\left(-\delta_n/2\right) \leq \frac{\delta_n}{2}.$$

This is true since $0 \leq F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) \leq c_{5,n}$ by definition of $c_{5,n}$ and hence,

$$\begin{aligned}
& (d+1) \cdot t_n \cdot \lambda_n \cdot 2 \cdot \sqrt{c_{5,n,Le9}} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|^2\}\} \cdot \sqrt{c_{6,n,Le9} \cdot n} \cdot \exp\left(-\delta_n/2\right) \\
\leq\ & (d+1)^2 \cdot t_n^2 \cdot \lambda_n \cdot 2 \cdot \sqrt{1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|^4\}\} \\
& \cdot \sqrt{\left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \frac{1}{c_1}} \cdot n \cdot \exp\left(-\delta_n/2\right) \\
\leq\ & (d+1)^2 \cdot t_n^2 \cdot \lambda_n \cdot 2 \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)^4 \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|^4\}\} \\
& \cdot \max\{1, \left(\frac{1}{c_1}\right)^2\} \cdot n^2 \cdot \exp\left(-\delta_n/2\right) \cdot \max\{1, \frac{c_{5,n}}{c_1}\} \\
\leq\ & \frac{1}{2} \\
\leq\ & \frac{\delta_n}{2},
\end{aligned}$$

where the second to last inequality holds by (3.24) and the last inequality holds since $\delta_n \geq 1$ by assumption. There are two conditions remaining, namely (3.15) and (3.16), i.e.

$$F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) \leq c_{5,n,Le9}$$

and

$$\|\mathbf{a}^{(t)}\|^2 \leq c_{6,n,Le9}$$

for $t \in \{0, \ldots, t_n - 1\}$. For these we show the following claim for all $s \in \{0, 1, \ldots, t_n - 1\}$

$$\begin{aligned} & \max\left\{F(\mathbf{a}^{(s+1)}, \mathbf{b}^{(s)}), F(\mathbf{a}^{(s+1)}, \mathbf{b}^{(s+1)})\right\} - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\ \leq \quad & c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot (s+1) \cdot \exp(-\delta_n/4). \end{aligned} \tag{3.28}$$

In other words, we need to show two inequalities

1. $F(\mathbf{a}^{(s+1)}, \mathbf{b}^{(s)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \leq c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot (s+1) \cdot \exp(-\delta_n/4)$
2. $F(\mathbf{a}^{(s+1)}, \mathbf{b}^{(s+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \leq c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot (s+1) \cdot \exp(-\delta_n/4)$,

which we do by induction on s.
<u>Start of the induction.</u> For $s = 0$ we have:

1. For the first inequality of the claim we have by (3.26) and (3.22)

$$\begin{aligned} & F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\ \leq \quad & \left(1 - \frac{4 \cdot c_1}{6 \cdot K \cdot n}\right) \cdot \left(F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) \\ \leq \quad & 1 \cdot \left(F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) + \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) \\ \leq \quad & c_{5,n} + F(0, \mathbf{b}^{(0)}) \\ = \quad & c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2. \end{aligned} \tag{3.29}$$

2. For the second inequality of the claim we have

$$F(\mathbf{a}^{(1)}, \mathbf{b}^{(1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})$$

$$= \left(F(\mathbf{a}^{(1)}, \mathbf{b}^{(1)}) - F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)})\right) + \left(F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right).$$

The second term on the right-hand side is bounded by (3.29). It remains to bound the first term on the right-hand side. We have by the Cauchy-Schwarz inequality

$$\begin{aligned}
& F(\mathbf{a}^{(1)}, \mathbf{b}^{(1)}) - F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)}) \\
= \; & \frac{1}{n}\sum_{i=1}^{n} |f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - y_i|^2 + \frac{c_1}{n}\sum_{k=0}^{K}(a_k^{(1)})^2 \\
& -|f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i) - y_i|^2 - \frac{c_1}{n}\sum_{k=0}^{K}(a_k^{(1)})^2 \\
= \; & \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - y_i + (f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i) - y_i)) \\
& \qquad \cdot(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - y_i - (f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i) - y_i)) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) + f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i) - 2y_i) \\
& \qquad \cdot(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i)) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}(2f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i) - 2y_i) \cdot (f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i)) \\
& +\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i))^2 \\
\leq \; & \sqrt{\frac{1}{n}\sum_{i=1}^{n}(2 \cdot f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i) - 2 \cdot y_i)^2} \\
& \cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i))^2} \\
& +\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i))^2 \\
\leq \; & 2 \cdot \sqrt{F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)})} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i))^2}
\end{aligned}$$

$$+\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i))^2.$$

We bound the term

$$\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i))^2$$

further. Applying the Cauchy-Schwarz inequality a second time and Lipschitz continuity of σ (with Lipschitz constant 1) yield

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i))^2 \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K} a_k^{(1)}\cdot\sigma\left(\sum_{j=1}^{d} b_{k,j}^{(1)}\cdot x_i^{(j)} + b_{k,0}^{(1)}\right) + a_0^{(1)}\right. \\
&\qquad\left. -\left(\sum_{k=1}^{K} a_k^{(1)}\cdot\sigma\left(\sum_{j=1}^{d} b_{k,j}^{(0)}\cdot x_i^{(j)} + b_{k,0}^{(0)}\right) + a_0^{(1)}\right)\right)^2 \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K} a_k^{(1)}\cdot\left(\sigma\left(\sum_{j=1}^{d} b_{k,j}^{(1)}\cdot x_i^{(j)} + b_{k,0}^{(1)}\right) - \sigma\left(\sum_{j=1}^{d} b_{k,j}^{(0)}\cdot x_i^{(j)} + b_{k,0}^{(0)}\right)\right)\right)^2 \\
\leq \;& \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\left(a_k^{(1)}\right)^2\cdot\sum_{k=1}^{K}\left(\sigma\left(\sum_{j=1}^{d} b_{k,j}^{(1)}\cdot x_i^{(j)} + b_{k,0}^{(1)}\right)\right.\right. \\
&\qquad\left.\left. -\sigma\left(\sum_{j=1}^{d} b_{k,j}^{(0)}\cdot x_i^{(j)} + b_{k,0}^{(0)}\right)\right)^2\right) \\
\leq \;& \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\left(a_k^{(1)}\right)^2\cdot\sum_{k=1}^{K}\left(\sum_{j=1}^{d}|b_{k,j}^{(1)}\cdot x_i^{(j)} + b_{k,0}^{(1)} - b_{k,j}^{(0)}\cdot x_i^{(j)} - b_{k,0}^{(0)}|\right)^2\right) \\
\leq \;& \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\left(a_k^{(1)}\right)^2\cdot\sum_{k=1}^{K}\left(\sum_{j=1}^{d}|b_{k,j}^{(1)} - b_{k,j}^{(0)}|\cdot x_i^{(j)} + |b_{k,0}^{(1)} - b_{k,0}^{(0)}|\right)^2\right) \\
\leq \;& \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\left(a_k^{(1)}\right)^2\cdot\sum_{k=1}^{K}\left(\sum_{j=0}^{d}|b_{k,j}^{(1)} - b_{k,j}^{(0)}|\cdot\max\{1,\max_{i,j}|x_i^{(j)}|\}\right)^2\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{K}\left(a_k^{(1)}\right)^2 \cdot \sum_{k=1}^{K}\left(\max\{1, \max_{i,j}|x_i^{(j)}|\}\right)^2 \cdot \left(\sum_{j=0}^{d}|b_{k,j}^{(1)} - b_{k,j}^{(0)}|\right)^2 \\
&\leq \sum_{k=1}^{K}\left(a_k^{(1)}\right)^2 \cdot \sum_{k=1}^{K}\max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1) \cdot \sum_{j=0}^{d}|b_{k,j}^{(1)} - b_{k,j}^{(0)}|^2 \\
&= \sum_{k=1}^{K}\left(a_k^{(1)}\right)^2 \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1) \cdot \sum_{k=1}^{K}\sum_{j=0}^{d}|b_{k,j}^{(1)} - b_{k,j}^{(0)}|^2.
\end{aligned}
$$

We conclude from (3.29) that

$$
\begin{aligned}
& \frac{c_1}{n}\sum_{k=0}^{K}\left(a_k^{(1)}\right)^2 \\
\leq\ & \frac{c_1}{n}\sum_{k=0}^{K}\left(a_k^{(1)}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}|f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i) - y_i|^2 \\
\leq\ & F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)}) \\
\leq\ & \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) + c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 \\
\leq\ & F(0, \mathbf{b}^{(0)}) + c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 \\
\leq\ & 1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2
\end{aligned}
$$

and hence,

$$
\sum_{k=0}^{K}\left(a_k^{(1)}\right)^2 \leq \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \frac{1}{c_1} \cdot n. \tag{3.30}
$$

Now we consider

$$
|b_{k,j}^{(1)} - b_{k,j}^{(0)}|,
$$

to which we apply Lemma 3.2.9. This means we need to check (3.15) and (3.16) in this particular case of the start of the induction. We have

$$
F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) \leq c_{5,n} \leq 1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2 = c_{5,n,Le9}
$$

and

$$\begin{aligned}
\frac{c_1}{n}\sum_{k=1}^{K}\left(a_k^{(0)}\right)^2 &\leq \frac{c_1}{n}\sum_{k=1}^{K}\left(a_k^{(0)}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}|f_{net,(\mathbf{a}^{(0)},\mathbf{b}^{(0)})}(x_i)|^2 \\
&= F(\mathbf{a}^{(0)},\mathbf{b}^{(0)}) \\
&\leq 1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2
\end{aligned}$$

and hence,

$$\|\mathbf{a}^{(0)}\|^2 = \sum_{k=1}^{K}\left(a_k^{(0)}\right)^2 \leq \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)\cdot\frac{1}{c_1}\cdot n = c_{6,n,Le9}\cdot n.$$

So, Lemma 3.2.9 gives us

$$\begin{aligned}
&|b_{k,j}^{(1)} - b_{k,j}^{(0)}| \\
\leq\ & \lambda_n \cdot 2 \cdot \sqrt{1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2}\cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \\
&\cdot\sqrt{\left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)\cdot\frac{1}{c_1}\cdot n}\cdot \exp\left(-\delta_n/2\right) \\
=\ & \lambda_n \cdot 2 \cdot \sqrt{1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2}\cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \\
&\cdot\sqrt{1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2}\cdot\sqrt{\frac{1}{c_1}}\cdot\sqrt{n}\cdot \exp\left(-\delta_n/2\right) \\
\leq\ & \lambda_n \cdot 2 \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)\cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \\
&\cdot\frac{1}{\sqrt{c_1}}\cdot\sqrt{n}\cdot \exp\left(-\delta_n/2\right).
\end{aligned}$$

From this we conclude with (3.30), with

$$K\cdot\lambda_n^2 = K\cdot\frac{1}{3\cdot K}\cdot\lambda_n \leq \lambda_n$$

by definition of λ_n and with

$$0 \leq F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) \leq c_{5,n}$$

by definiton of $c_{5,n}$ that

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i))^2 \\
\leq \quad & \sum_{k=1}^{K}\left(a_k^{(1)}\right)^2 \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1) \cdot K \cdot (d+1) \\
& \cdot \left(2 \cdot \lambda_n \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \frac{1}{\sqrt{c_1}}\right. \\
& \left. \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{n} \cdot \exp\left(-\delta_n/2\right)\right)^2 \\
\leq \quad & \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \frac{1}{c_1} \cdot n \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1)^2 \cdot K \\
& \cdot 4 \cdot \lambda_n^2 \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)^2 \cdot \frac{1}{c_1} \\
& \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|^2\}\} \cdot n \cdot \exp\left(-\delta_n\right) \\
\leq \quad & 4 \cdot t_n^2 \cdot \max\{1, \frac{c_{5,n}}{c_1}\} \cdot \max\{1, \frac{1}{c_1^2}\} \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)^3 \cdot n^2 \\
& \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^4\} \cdot (d+1)^2 \cdot \lambda_n \cdot \exp\left(-\delta_n\right) \\
& \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \min\left\{1, \frac{1}{F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)})}\right\} \\
\leq \quad & \exp\left(-\delta_n/2\right) \cdot \min\left\{1, \frac{1}{F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)})}\right\},
\end{aligned}$$

where the last inequality holds by (3.24). Since

$$\min\left\{1, \frac{1}{F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)})}\right\} \cdot \exp\left(-\delta_n/2\right) \leq 1$$

we have

$$\sqrt{\min\left\{1, \frac{1}{F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)})}\right\} \cdot \exp\left(-\delta_n/2\right)} \geq \min\left\{1, \frac{1}{F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)})}\right\} \cdot \exp\left(-\delta_n/2\right).$$

Hence, plugging in yields

$$\begin{aligned}
& F(\mathbf{a}^{(1)}, \mathbf{b}^{(1)}) - \min_{\mathbf{a}} F(\mathbf{a}^{(1)}\mathbf{b}^{(0)}) \\
\leq\ & 2 \cdot \sqrt{F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)})} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i))^2} \\
& + \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(1)})}(x_i) - f_{net,(\mathbf{a}^{(1)},\mathbf{b}^{(0)})}(x_i))^2 \\
\leq\ & 2 \cdot \sqrt{F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)})} \cdot \sqrt{\min\left\{1, \frac{1}{F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)})}\right\} \cdot \exp\left(-\delta_n/2\right)} \\
& + \min\left\{1, \frac{1}{F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)})}\right\} \cdot \exp\left(-\delta_n/2\right) \\
\leq\ & 2 \cdot 1 \cdot \sqrt{\exp\left(-\delta_n/2\right)} + 1 \cdot \sqrt{\exp\left(-\delta_n/2\right)} \\
\leq\ & 3 \cdot \exp\left(-\delta_n/4\right).
\end{aligned}$$

Thus, we have

$$F(\mathbf{a}^{(1)}, \mathbf{b}^{(1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \leq 3 \cdot \exp\left(-\delta_n/4\right) + c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2.$$

This concludes the start of the induction.
Induction hypothesis. Assume that (3.28) holds for $s = t - 1$ for $t \in \{1, \ldots, t_n - 1\}$.
Induction step. We have:

1. For the first inequality of the claim we have by (3.26) and by the induction hypothesis

$$\begin{aligned}
& F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
=\ & \left(F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})\right) + \left(\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) \\
\leq\ & \left(1 - \frac{4 \cdot c_1}{6 \cdot K \cdot n}\right) \cdot \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})\right) \\
& + \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})
\end{aligned}$$

$$
\begin{aligned}
&= \left(1-\frac{4\cdot c_1}{6\cdot K\cdot n}\right)\cdot F(\mathbf{a}^{(t)},\mathbf{b}^{(t)}) - \left(\left(1-\frac{4\cdot c_1}{6\cdot K\cdot n}\right)+1\right)\cdot \min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(t)}) \\
&\quad - \left(\left(1-\frac{4\cdot c_1}{6\cdot K\cdot n}\right)+\frac{4\cdot c_1}{6\cdot K\cdot n}\right)\cdot \min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(0)}) \\
&= \left(1-\frac{4\cdot c_1}{6\cdot K\cdot n}\right)\cdot \left(F(\mathbf{a}^{(t)},\mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(0)})\right) \\
&\quad + \frac{4\cdot c_1}{6\cdot K\cdot n}\cdot \left(\min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(0)})\right) \\
&\leq \left(1-\frac{4\cdot c_1}{6\cdot K\cdot n}\right)\cdot \left(c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3\cdot t\cdot \exp\left(-\delta_n/4\right)\right) \\
&\quad + \frac{4\cdot c_1}{6\cdot K\cdot n}\cdot \left(\min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(0)})\right) \\
&\leq \left(1-\frac{4\cdot c_1}{6\cdot K\cdot n}\right)\cdot \left(c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3\cdot t\cdot \exp\left(-\delta_n/4\right)\right) + \frac{4\cdot c_1}{6\cdot K\cdot n}\cdot F(0,\mathbf{b}^{(t)}) \\
&= \left(1-\frac{4\cdot c_1}{6\cdot K\cdot n}\right)\cdot \left(c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3\cdot t\cdot \exp\left(-\delta_n/4\right)\right) + \frac{4\cdot c_1}{6\cdot K\cdot n}\cdot \frac{1}{n}\sum_{i=1}^{n} y_i^2 \\
&\leq c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3\cdot t\cdot \exp\left(-\delta_n/4\right). \qquad (3.31)
\end{aligned}
$$

2. For the second inequality of the claim we have proceed in the same fashion as in the start of the induction. We have

$$
\begin{aligned}
&F(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(0)}) \\
&= \left(F(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})\right) + \left(F(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(0)})\right).
\end{aligned}
$$

The second term on the right-hand side is bounded by (3.31). It remains to bound the first term on the right-hand side. We note that this remaining term is equal to the first summand on the right-hand side of Step 1. By the Cauchy-Schwarz inequality we get

$$
\begin{aligned}
&F(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)}) \\
&= \frac{1}{n}\sum_{i=1}^{n} |f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - y_i|^2 - |f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i) - y_i|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} (f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) + f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i) - 2y_i)
\end{aligned}
$$

$$
\begin{aligned}
&\cdot(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i)) \\
= \quad & \frac{1}{n}\sum_{i=1}^{n}(2f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i) - 2y_i)\cdot(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i)) \\
& + \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2 \\
\leq \quad & \sqrt{\frac{1}{n}\sum_{i=1}^{n}(2\cdot f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i) - 2\cdot y_i)^2} \\
& \cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2} \\
& + \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2 \\
\leq \quad & 2\cdot\sqrt{F(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2} \\
& + \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2.
\end{aligned}
$$

We bound the term

$$
\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2
$$

further. Applying the Cauchy-Schwarz inequality a second time and Lipschitz continuity of σ (with Lipschitz constant 1) yield

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2 \\
= \quad & \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}a_k^{(t+1)}\cdot\left(\sigma\left(\sum_{j=1}^{d}b_{k,j}^{(t+1)}\cdot x_i^{(j)} + b_{k,0}^{(t+1)}\right) - \sigma\left(\sum_{j=1}^{d}b_{k,j}^{(t)}\cdot x_i^{(j)} + b_{k,0}^{(t)}\right)\right)\right)^2 \\
\leq \quad & \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\left(a_k^{(t+1)}\right)^2\cdot\sum_{k=1}^{K}\left(\sigma\left(\sum_{j=1}^{d}b_{k,j}^{(t+1)}\cdot x_i^{(j)} + b_{k,0}^{(t+1)}\right)\right.\right.
\end{aligned}
$$

$$
\begin{aligned}
& \qquad\qquad -\sigma\left(\sum_{j=1}^{d} b_{k,j}^{(t)} \cdot x_i^{(j)} + b_{k,0}^{(t)}\right)\Bigg)^2\Bigg) \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\left(a_k^{(t+1)}\right)^2 \cdot \sum_{k=1}^{K}\left(\sum_{j=1}^{d}|b_{k,j}^{(t+1)} \cdot x_i^{(j)} + b_{k,0}^{(t+1)} - b_{k,j}^{(t)} \cdot x_i^{(j)} - b_{k,0}^{(t)}|\right)^2\right) \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\left(a_k^{(t+1)}\right)^2 \cdot \sum_{k=1}^{K}\left(\sum_{j=0}^{d}|b_{k,j}^{(t+1)} - b_{k,j}^{(t)}| \cdot \max\{1, \max_{i,j}|x_i^{(j)}|\}\right)^2\right) \\
=\ & \sum_{k=1}^{K}\left(a_k^{(t+1)}\right)^2 \cdot \sum_{k=1}^{K}\left(\max\{1, \max_{i,j}|x_i^{(j)}|\}\right)^2 \cdot \left(\sum_{j=0}^{d}|b_{k,j}^{(t+1)} - b_{k,j}^{(t)}|\right)^2 \\
\leq\ & \sum_{k=1}^{K}\left(a_k^{(t+1)}\right)^2 \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1) \cdot \sum_{k=1}^{K}\sum_{j=0}^{d}|b_{k,j}^{(t+1)} - b_{k,j}^{(t)}|^2.
\end{aligned}
$$

We conclude from (3.31) and (3.25) that

$$
\begin{aligned}
& \frac{c_1}{n}\sum_{k=0}^{K}\left(a_k^{(t+1)}\right)^2 \\
\leq\ & \frac{c_1}{n}\sum_{k=0}^{K}\left(a_k^{(t+1)}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}|f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(0)})}(x_i) - y_i|^2 \\
\leq\ & F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(0)}) \\
\leq\ & \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) + c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot t \cdot \exp\left(-\delta_n/4\right) \\
\leq\ & F(0, \mathbf{b}^{(0)}) + c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot t \cdot \exp\left(-\delta_n/4\right) \\
\leq\ & c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2 + 1
\end{aligned}
$$

and hence,

$$
\sum_{k=0}^{K}\left(a_k^{(t+1)}\right)^2 \leq \left(c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2 + 1\right) \cdot \frac{1}{c_1} \cdot n. \tag{3.32}
$$

Now we consider

$$|b_{k,j}^{(t+1)} - b_{k,j}^{(t)}|,$$

to which we apply Lemma 3.2.9. This means we need to check (3.15) and (3.16), which we do in the same way as in the start of the induction. We have by the induction hypothesis and by (3.25)

$$\begin{aligned} F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) &\leq c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot t \cdot \exp\left(-\delta_n/4\right) + \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\ &\leq c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 1 + F(0, \mathbf{b}^{(0)}) \\ &\leq c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2 + 1 \\ &= c_{5,n,Le9} \end{aligned}$$

from which we conclude that

$$\begin{aligned} \frac{c_1}{n}\sum_{k=1}^{K}\left(a_k^{(t)}\right)^2 &\leq \frac{c_1}{n}\sum_{k=1}^{K}\left(a_k^{(t)}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}|f_{net,(\mathbf{a}^{(t)},\mathbf{b}^{(t)})}(x_i)|^2 \\ &= F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) \\ &\leq 1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2 \end{aligned}$$

and hence,

$$\|\mathbf{a}^{(t)}\|^2 = \sum_{k=1}^{K}\left(a_k^{(t)}\right)^2 \leq \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \frac{1}{c_1} \cdot n = c_{6,n,Le9} \cdot n.$$

So, Lemma 3.2.9 gives us

$$\begin{aligned} & |b_{k,j}^{(t+1)} - b_{k,j}^{(t)}| \\ \leq\ & \lambda_n \cdot 2 \cdot \sqrt{1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \\ & \cdot \sqrt{\left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \frac{1}{c_1} \cdot n} \cdot \exp\left(-\delta_n/2\right) \end{aligned}$$

$$
\begin{aligned}
&\leq \quad \lambda_n \cdot 2 \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \\
&\qquad \cdot \max\{1, \frac{1}{c_1}\} \cdot \sqrt{n} \cdot \exp\left(-\delta_n/2\right).
\end{aligned}
$$

From this we conclude with (3.32), with

$$
K \cdot \lambda_n^2 = K \cdot \frac{1}{3 \cdot K} \cdot \lambda_n \leq \lambda_n
$$

by definition of λ_n and with

$$
0 \leq F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) \leq c_{5,n}
$$

by definiton of $c_{5,n}$ that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+11)},\mathbf{b}^{(t)})}(x_i))^2 \\
\leq \quad & \sum_{k=1}^{K}\left(a_k^{(t+1)}\right)^2 \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1) \cdot K \cdot (d+1) \\
& \cdot \left(2 \cdot \lambda_n \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \max\{1, \frac{1}{c_1}\}\right. \\
& \qquad \left. \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{n} \cdot \exp\left(-\delta_n/2\right)\right)^2 \\
\leq \quad & \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \frac{1}{c_1} \cdot n \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1)^2 \cdot K \\
& \cdot 4 \cdot \lambda_n^2 \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)^2 \cdot \max\{1, \frac{1}{c_1^2}\} \\
& \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|^2\}\} \cdot n \cdot \exp\left(-\delta_n\right) \\
\leq \quad & 4 \cdot t_n^2 \cdot \max\{1, \frac{c_{5,n}}{c_1}\} \cdot \max\{1, \frac{1}{c_1^2}\} \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)^3 \cdot n^2 \\
& \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^4\} \cdot (d+1)^2 \cdot \lambda_n \cdot \exp\left(-\delta_n\right)
\end{aligned}
$$

$$
\begin{aligned}
& \cdot \left(1 + c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \min\left\{1, \frac{1}{F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})}\right\} \\
\leq \quad & \exp\left(-\delta_n/2\right) \cdot \min\left\{1, \frac{1}{F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})}\right\},
\end{aligned}
$$

where the last inequality holds by (3.24). Since

$$
\min\left\{1, \frac{1}{F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})}\right\} \cdot \exp\left(-\delta_n/2\right) \leq 1
$$

we have

$$
\sqrt{\min\left\{1, \frac{1}{F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})}\right\} \cdot \exp\left(-\delta_n/2\right)} \geq \min\left\{1, \frac{1}{F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})}\right\} \cdot \exp\left(-\delta_n/2\right).
$$

Hence, plugging in yields

$$
\begin{aligned}
& F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) \\
\leq \quad & 2 \cdot \sqrt{F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2} \\
& + \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2 \\
\leq \quad & 2 \cdot \sqrt{F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})} \cdot \sqrt{\min\left\{1, \frac{1}{F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})}\right\} \cdot \exp\left(-\delta_n/2\right)} \\
& + \min\left\{1, \frac{1}{F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})}\right\} \cdot \exp\left(-\delta_n/2\right) \\
\leq \quad & 2 \cdot 1 \cdot \sqrt{\exp\left(-\delta_n/2\right)} + 1 \cdot \sqrt{\exp\left(-\delta_n/2\right)} \\
\leq \quad & 3 \cdot \exp\left(-\delta_n/4\right) \\
=: \quad & \beta_2.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
& F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
\leq \quad & 3 \cdot \exp\left(-\delta_n/4\right) + c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot t \cdot \exp\left(-\delta_n/4\right)
\end{aligned}
$$

$$= \quad c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3\cdot(t+1)\cdot\exp\left(-\delta_n/4\right).$$

This concludes the proof of the claim. With the claim we check the remaining two conditions (3.15) and (3.16) in the same manner as during the induction in order to apply Lemma 3.2.9. We conclude with the claim that for $s \in \{0,\dots,t_n-1\}$

$$\begin{aligned}
& F(\mathbf{a}^{(s)},\mathbf{b}^{(s)}) \\
\leq\ & c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3\cdot(s+1)\cdot\exp\left(-\delta_n/4\right) + \min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(0)}) \\
\leq\ & c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 1 + F(0,\mathbf{b}^{(0)}) \\
=\ & c_{5,n} + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 1 + \frac{1}{n}\sum_{i=1}^{n} y_i^2 \\
=\ & c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2 + 1 \\
=\ & c_{5,n,Le9}
\end{aligned}$$

which verifies (3.15) and

$$\begin{aligned}
\|\mathbf{a}^{(s)}\|^2 &= \frac{n}{c_1}\cdot\left(\frac{c_1}{n}\sum_{k=1}^{K}\left(a_k^{(s)}\right)^2\right) \\
&\leq \frac{n}{c_1}\cdot\left(\frac{c_1}{n}\sum_{k=1}^{K}\left(a_k^{(s)}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}|f_{net,(\mathbf{a}^{(s)},\mathbf{b}^{(s)})}(x_i)|^2\right) \\
&= \frac{n}{c_1}\cdot F(\mathbf{a}^{(s)},\mathbf{b}^{(s)}) \\
&\leq \frac{n}{c_1}\cdot\left(c_{5,n} + \frac{2}{n}\sum_{i=1}^{n} y_i^2 + 1\right) \\
&= c_{6,n,Le9}\cdot n
\end{aligned}$$

which verifies (3.16). Thus, all conditions of Lemma 3.2.9 are met.

Step 5: Finishing the proof. We put together the results from the previous steps. For simplicity, we introduce the following notation

$$\gamma_t = \left(F(\mathbf{a}^{(t)},\mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(0)})\right) \quad \text{and} \quad \alpha = \frac{2\cdot c_1}{3\cdot K\cdot n}.$$

As a consequence,

$$
\begin{aligned}
& \gamma_{t+1} \\
\leq\ & F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) + (1-\alpha) \cdot (F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})) \\
& \quad + \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
=\ & F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) + (1-\alpha) \cdot (F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})) \\
& \quad + \alpha \cdot (\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})) \\
=\ & (1-\alpha) \cdot \gamma_t + \alpha \cdot (\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})) \\
& \quad + F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) \\
\leq\ & (1-\alpha) \cdot \gamma_t + \alpha \cdot \beta_1 + \beta_2. \qquad (3.33)
\end{aligned}
$$

Applying the relation in (3.33) recursively we can bound the term by a geometric series and we get

$$
\begin{aligned}
\gamma_{t+1} &\leq (1-\alpha)^{t+1} \cdot \gamma_0 + \sum_{k=0}^{t} (1-\alpha)^k \cdot \alpha \cdot \beta_1 + \sum_{k=0}^{t} (1-\alpha)^k \cdot \beta_2 \\
&\leq (1-\alpha)^{t+1} \cdot \gamma_0 + \sum_{k=0}^{\infty} (1-\alpha)^k \cdot \alpha \cdot \beta_1 + \sum_{k=0}^{\infty} (1-\alpha)^k \cdot \beta_2 \\
&= (1-\alpha)^{t+1} \cdot \gamma_0 + \frac{\alpha \cdot \beta_1}{1-(1-\alpha)} + \frac{\beta_2}{1-(1-\alpha)} \\
&= (1-\alpha)^{t+1} \cdot \gamma_0 + \beta_1 + \frac{1}{\alpha} \cdot \beta_2.
\end{aligned}
$$

Plugging the values back in yields

$$
\begin{aligned}
& F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
\leq\ & \left(1 - \frac{2 \cdot c_1}{3 \cdot K \cdot n}\right)^{t+1} \cdot \left(F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) + (2 \cdot \sqrt{c_{5,n}} + 1) \cdot \exp\left(-\delta_n/4\right) \\
& + \frac{3 \cdot K \cdot n}{2 \cdot c_1} \cdot 3 \cdot \exp\left(-\delta_n/4\right),
\end{aligned}
$$

which concludes the proof. □

3.2.3. Auxiliary Lemmas from Empirical Process Theory

For the proof of Theorem 3.2.1 we will need the following auxiliary results from empirical process theory.

Lemma 3.2.11. *Let*

- $\beta_n = c_3 \cdot \log(n)$ *for some suitably large constant* $c_3 > 0$*,*
- $\mathcal{F}_n$ *be a set of functions* $f : \mathbb{R}^d \to \mathbb{R}$*.*

Assume

- *the distribution of* (X, Y) *satisfies (3.6) for some constant* $c_2 > 0$*,*
- *the regression function* m *is bounded in absolute value,*
- *the estimate* m_n *satisfies*
$$m_n = T_{\beta_n} \tilde{m}_n$$
with
$$\tilde{m}_n(\cdot) = \tilde{m}_n(\cdot, (X_1, Y_1), \ldots, (X_n, Y_n)) \in \mathcal{F}_n \tag{3.34}$$
- *the estimate* m_n *satisfies*
$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n |Y_i - \tilde{m}_n(X_i)|^2 \cdot \mathbf{1}_{\{|Y_i| \le \beta_n \text{ for all } i \in \{1,\ldots,n\}\}} \\ &\le \min_{l \in \Theta_n} \left(\frac{1}{n} \sum_{i=1}^n |Y_i - g_{n,l}(X_i)|^2 + pen_n(g_{n,l}) + \epsilon_{n,l} \right) \end{aligned} \tag{3.35}$$
for
  - *some nonempty set* Θ_n *of parameters,*
  - *some random functions* $g_{n,l} : \mathbb{R}^d \to \mathbb{R}$*, that only depend on the set*
$$\mathcal{D}_{n,r} = \{X_1, \ldots, X_n, \bar{\mathbf{c}}_1^{(1)}, \ldots, \bar{\mathbf{c}}_r^{(1)}, \ldots, \bar{\mathbf{c}}_1^{(I_n)}, \ldots, \bar{\mathbf{c}}_r^{(I_n)}\}$$
where
$$\bar{\mathbf{c}}_1^{(1)}, \ldots, \bar{\mathbf{c}}_r^{(1)}, \ldots, \bar{\mathbf{c}}_1^{(I_n)}, \ldots, \bar{\mathbf{c}}_r^{(I_n)}$$
are random variables independent of
$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

– *some deterministic penalty terms* $pen_n(g_{n,l}) \geq 0$*,*

– *some additional deterministic term* $\epsilon_{n,l} \geq 0$*.*

Then m_n *satisfies*

$$\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq \frac{c_7 \cdot (\log n)^2 \cdot \left(\log\left(\sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{n\cdot\beta_n}, \mathcal{F}_n, x_1^n\right)\right) + 1\right)}{n}$$
$$+ 2 \cdot \mathbf{E}\left(\min_{l\in\Theta_n} \frac{1}{n}\sum_{i=1}^n |g_{n,l}(X_i) - m(X_i)|^2 + pen_n(g_{n,l}) + \epsilon_{n,l}\right)$$

for $n > 1$ *and some constant* $c_7 > 0$*, which does not depend on* n, β_n *or the parameters of the estimate.*

Proof. Compared to Lemma 2.2.14, we see that (3.34) is assumed independently from any event A_n and (3.35) is an additional bound.
We apply Lemma 2.2.14 where the event A_n is the underlying set of our probability space. This gives us

$$\begin{aligned}
&\mathbf{E}\Bigg(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\quad -2 \cdot \left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{\{|Y_j| \leq \beta_n\ (j\in\{1,\dots,n\})\}} - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right)\Bigg) \\
&\leq \frac{c_7 \cdot (\log n)^2 \cdot \left(\log\left(\sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{n\cdot\beta_n}, \mathcal{F}_n, x_1^n\right)\right) + 1\right)}{n}
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\mathbf{E}\left(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)\right) \\
\leq\ & \frac{c_7 \cdot (\log n)^2 \cdot \left(\log\left(\sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{n\cdot\beta_n}, \mathcal{F}_n, x_1^n\right)\right) + 1\right)}{n} \\
&+2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{\{|Y_j| \leq \beta_n\ (j\in\{1,\dots,n\})\}} - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right).
\end{aligned}$$

We use (3.35) and bound the expected value on the right-hand side further by

$$2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{\{|Y_j| \leq \beta_n\ (j\in\{1,\dots,n\})\}} - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right)$$

$$\leq \quad 2 \cdot \mathbf{E}\left(\min_{l\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} |g_{n,l}(X_i) - Y_i|^2 + pen_n(g_{n,l}) + \epsilon_{n,l} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right).$$

Since $pen_n(g_{n,l})$ and $\epsilon_{n,l}$ are deterministic terms and and since $g_{n,l}$ are independent of Y_1, $\ldots$, Y_n given $\mathcal{D}_{n,r}$ we get that

$$\begin{aligned}
& \mathbf{E}\left(\min_{l\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} |g_{n,l}(X_i) - Y_i|^2 + pen_n(g_{n,l}) + \epsilon_{n,l} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right) \\
= \quad & \mathbf{E}\Bigg(\mathbf{E}\Bigg(\min_{l\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} |g_{n,l}(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2 \\
& \qquad + pen_n(g_{n,l}) + \epsilon_{n,l} \mid \mathcal{D}_{n,r}\Bigg)\Bigg) \\
\leq \quad & \mathbf{E}\Bigg(\min_{l\in\Theta_n} \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |g_{n,l}(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2 \mid \mathcal{D}_{n,r}\right) \\
& \qquad + pen_n(g_{n,l}) + \epsilon_{n,l}\Bigg) \\
= \quad & \mathbf{E}\Bigg(\min_{l\in\Theta_n} \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |g_{n,l}(X_i) - Y_i|^2 \mid \mathcal{D}_{n,r}\right) - \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2 \mid \mathcal{D}_{n,r}\right) \\
& \qquad + pen_n(g_{n,l}) + \epsilon_{n,l}\Bigg) \\
= \quad & \mathbf{E}\Bigg(\min_{l\in\Theta_n} \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |(g_{n,l}(X_i) - m(X_i)) + (m(X_i) - Y_i)|^2 \mid \mathcal{D}_{n,r}\right) \\
& \qquad - \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2 \mid \mathcal{D}_{n,r}\right) + pen_n(g_{n,l}) + \epsilon_{n,l}\Bigg) \\
= \quad & \mathbf{E}\Bigg(\min_{l\in\Theta_n} \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |g_{n,l}(X_i) - m(X_i)|^2 \mid \mathcal{D}_{n,r}\right) + \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2 \mid \mathcal{D}_{n,r}\right) \\
& \qquad - \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2 \mid \mathcal{D}_{n,r}\right) + pen_n(g_{n,l}) + \epsilon_{n,l}\Bigg) \\
= \quad & \mathbf{E}\left(\min_{l\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} |g_{n,l}(X_i) - m(X_i)|^2 + pen_n(g_{n,l}) + \epsilon_{n,l}\right)
\end{aligned}$$

where the fourth equality holds since the mixed term is

$$\begin{aligned}
& \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(g_{n,l}(X_i)-m(X_i))\cdot(m(X_i)-Y_i)\mid\mathcal{D}_{n,r}\right)\\
= & \frac{1}{n}\sum_{i=1}^{n}(g_{n,l}(X_i)-m(X_i))\cdot\mathbf{E}\left((m(X_i)-Y_i)\mid\mathcal{D}_{n,r}\right)\\
= & \frac{1}{n}\sum_{i=1}^{n}(g_{n,l}(X_i)-m(X_i))\cdot\mathbf{E}\left((m(X_i)-Y_i)\mid X_i\right)\\
= & 0.
\end{aligned}$$

This concludes the proof. □

We can bound the covering number $\mathcal{N}_1\left(\frac{1}{n\cdot\beta_n},\mathcal{F}_n,x_1^n\right)$ in Lemma 3.2.11 by Lemma 2.2.15. We will make use of this in the proof of Thereom 3.2.4.

3.2.4. Proof of Theorem 3.2.1

We give a proof for the rate of convergence presented in Theorem 3.2.1.

Proof. Since by assumption $supp(X)$ is bounded and m is $(p,C)-$ smooth we can assume w.l.o.g. that m is bounded in absolute value. So, we assume

$$\|m\|_\infty\leq\beta_n.$$

Let B_n be the event where

$$|Y_i|\leq\beta_n\quad(i=1,\ldots,n).$$

We get

$$\begin{aligned}
& \mathbf{E}\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\\
= & \mathbf{E}\left(\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\cdot(\mathbf{1}_{B_n}+\mathbf{1}_{B_n^c})\right)\\
\leq & \mathbf{E}\left(\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\cdot\mathbf{1}_{B_n}\right)+\mathbf{E}\left(\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\cdot\mathbf{1}_{B_n^c}\right)\\
\leq & \mathbf{E}\left(\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\cdot\mathbf{1}_{B_n}\right)
\end{aligned}$$

$$
\begin{aligned}
&+\mathbf{E}\left(\int 2\cdot|m_n(x)|^2+2\cdot|m(x)|^2\mathbf{P}_X(dx)\cdot\mathbf{1}_{B_n^c}\right)\\
\leq\ &\mathbf{E}\left(\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\cdot\mathbf{1}_{B_n}\right)+\mathbf{E}\left(\int 2\cdot\beta_n^2+2\cdot\beta_n^2\mathbf{P}_X(dx)\cdot\mathbf{1}_{B_n^c}\right)\\
=\ &\mathbf{E}\left(\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\cdot\mathbf{1}_{B_n}\right)+4\cdot\beta_n^2\cdot\mathbf{P}\{B_n^c\}.
\end{aligned}
$$

We bound the summands on the right-hand side separately. We start with the second term. Since we have by Markov's inequality and by (3.6)

$$
\begin{aligned}
\mathbf{P}\{B_n^c\} &= \mathbf{P}\left\{|Y_i|>\beta_n \text{ for some } i\in\{1,\dots,n\}\right\}\\
&\leq n\cdot\mathbf{P}\left\{\exp(c_2\cdot Y^2)>\exp(c_2\cdot\beta_n^2)\right\}\\
&\leq n\cdot\frac{\mathbf{E}\left(\exp(c_2\cdot Y^2)\right)}{\exp\left(c_2\cdot(c_3\cdot\log n)^2\right)}\\
&\leq n\cdot\frac{c_8}{n^{c_{29}\cdot\log n}}\\
&\leq c_{30}\cdot\frac{1}{n}
\end{aligned}
$$

we can conclude

$$
4\cdot\beta_n^2\cdot\mathbf{P}\{B_n^c\}\leq 4\cdot(\log n)^2\cdot c_{30}\cdot\frac{1}{n}\leq c_{11}\cdot\frac{(\log n)^2}{n}. \tag{3.36}
$$

Next, we bound the first summand further by Lemma 3.2.11. For that we consider an extension

$$
\bar{m}_n=\begin{cases} m_n=T_{\beta_n}\tilde{m}_n & \text{if } B_n\\ 0 & \text{if } B_n^c\end{cases}
$$

that coincides with m_n on the event B_n and has value 0 outside of the event. Then we have

$$
\begin{aligned}
\mathbf{E}\left(\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\cdot\mathbf{1}_{B_n}\right) &= \mathbf{E}\left(\int|\bar{m}_n(x)-m(x)|^2\mathbf{P}_X(dx)\cdot\mathbf{1}_{B_n}\right)\\
&\leq \mathbf{E}\left(\int|\bar{m}_n(x)-m(x)|^2\mathbf{P}_X(dx)\right).
\end{aligned}
$$

We apply Lemma 3.2.11 to the right-hand side of the above inequality. In order to apply Lemma 3.2.11 we need to check the conditions (3.34) and (3.35) for $\tilde{m}_n$ on B_n, since

obviously, the condtions (3.34) and (3.35) hold for 0. We do that with Lemma 3.2.10. In order to do that we first need to verify the conditions of Lemma 3.2.10. Condition (3.22) is satisfied since

$$\begin{aligned} F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) &= F(0, \mathbf{b}^{(0)}) \\ &= \frac{1}{n} \sum_{i=1}^{n} |f_{net,(0,\mathbf{b})}(X_i) - Y_i|^2 + \frac{c_1}{n} \cdot \sum_{k=0}^{K \cdot r} 0^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} |Y_i|^2 \\ &\leq (c_3 \cdot \log n)^2 \\ &= c_{5,n}. \end{aligned} \tag{3.37}$$

Set

$$\delta_n = \frac{\rho_n \cdot \sqrt{d} \cdot A}{(n+1) \cdot (K-1)}.$$

In particular, this means that

$$\delta_n \geq c_{12} \cdot \frac{n^2 \cdot K}{(n+1) \cdot (K-1)} \geq c_{13} \cdot n \geq 1.$$

Then condition (3.23) is satisfied since we have by construction of our network estimate for $s \in \{1, \ldots, r\}$

$$\begin{aligned} & \min_{i=1,\ldots,n, k=1,\ldots,K} \left| \sum_{j=1}^{d} b_{(s-1)\cdot K+k,j} \cdot X_i^{(j)} + b_{(s-1)\cdot K+k,0} \right| \\ = & \min_{i=1,\ldots,n, k=1,\ldots,K} \left| \rho_n \cdot (\bar{\mathbf{c}}_s^T X_i - \tilde{b}_{s,k}) \right| \\ = & \rho_n \cdot \min_{i=1,\ldots,n, k=1,\ldots,K} \left| \bar{\mathbf{c}}_s^T X_i - \tilde{b}_{s,k} \right| \\ \geq & \rho_n \cdot \frac{\sqrt{d} \cdot A}{(n+1) \cdot (K-1)} \\ = & \delta_n. \end{aligned}$$

Next, trivially we have

$$2 \cdot c_1 \leq (K-2) \cdot n.$$

Condition (3.24) is satisfied since we have with (3.37)

$$\begin{aligned}
& 4 \cdot \max\{1, \frac{c_{5,n}}{c_1}\} \cdot \max\{1, \frac{1}{c_1^2}\} \cdot \lambda_n \cdot (d+1)^2 \cdot n^2 \cdot \max\{1, \max_{i,j} |X_i^{(j)}|^4\} \\
& \qquad \cdot \left(1 + c_{5,n} + \frac{2}{n} \sum_{i=1}^n Y_i^2\right)^4 \cdot t_n^2 \cdot \exp\left(-\delta_n/2\right) \\
\leq \; & c_{14} \cdot (\log n)^2 \cdot \frac{1}{K} \cdot n^2 \cdot \left(1 + (\log n)^2 + 2 \cdot (c_3 \cdot \log n)^2\right)^4 \\
& \qquad \cdot K \cdot (\log n)^2 \cdot \exp\left(-\frac{\delta_n}{2}\right) \\
\leq \; & c_{15} \cdot \frac{(\log n)^9 \cdot n^3}{e^{\frac{n}{2}}} \\
\leq \; & 1.
\end{aligned}$$

Finally, condition (3.25) is satisfied since

$$\begin{aligned}
3 \cdot t_n \cdot \exp\left(-\frac{\delta_n}{4}\right) &= 3 \cdot K \cdot n \cdot (\log n)^2 \cdot \exp\left(-\frac{\delta_n}{4}\right) \\
&\leq c_{16} \cdot \frac{n^{\frac{2p+2}{2p+1}} \cdot \log n}{e^{\frac{n}{4}}} \\
&\leq 1.
\end{aligned}$$

Hence, we can apply Lemma 3.2.10 to check conditions (3.34) and (3.35). First, we verify (3.34). In the proof of Lemma 3.2.10 in (3.32) we have already shown that for any $t \in \{0, \ldots, t_n - 1\}$

$$\begin{aligned}
\|\mathbf{a}^{(t+1)}\|^2 &\leq \left(1 + c_{5,n} + \frac{2}{n} \sum_{i=1}^n Y_i^2\right) \cdot \frac{1}{c_1} \cdot n \\
&\leq \left(1 + (\log n)^2 + 2 \cdot (c_3 \cdot \log n)^2\right) \cdot \frac{1}{c_1} \cdot n \\
&\leq c_{17} \cdot (\log n)^2 \cdot n.
\end{aligned}$$

As a consequence, we know that on B_n

$$\tilde{m}_n(x) = \sum_{k=0}^{K \cdot r} a_k^{(t_n)} \cdot \sigma\left(\sum_{j=1}^d b_{k,j}^{(t_n)} \cdot x^{(j)} + b_{k,0}^{(t_n)}\right) \in \mathcal{F}$$

where the function class $\mathcal{F}$ is defined as in Lemma 2.2.15 (as we can write $\tilde{m}_n(x) = \sum_{k=1}^{K_n} a_k^{(t_n)} \cdot \sigma\left(\sum_{j=1}^d b_{k,j}^{(t_n)} \cdot x^{(j)} + b_{k,0}^{(t_n)}\right) + 2 \cdot a_0^{(t_n)} \cdot \sigma(0 \cdot x + 0)$) with

$$\gamma_n = c_{17} \cdot (\log n)^2 \cdot n.$$

Second, we verify (3.35). For the following consideration we make use of the construction of our neural network regression estimate. Denote for $l \in \{1, \dots, I_n\}$ and for $t \in \{0, \dots, t_n\}$ by

$$((\mathbf{a}^{(l)})^{(t)}, (\mathbf{b}^{(l)})^{(t)})$$

the weight vector in the $l-$th initialization after the $t-$gradient descent step. We use that we initialize the weights I_n times and after applying gradient descent t_n times we choose among those I_n possible weight vectors the vector $(\mathbf{a}^{(t_n)}, \mathbf{b}^{(t_n)})$ that has the smallest value in F, i.e.

$$F(\mathbf{a}^{(t_n)}, \mathbf{b}^{(t_n)}) = \min_{l=1,\dots,I_n} F((\mathbf{a}^{(l)})^{(t_n)}, (\mathbf{b}^{(l)})^{(t_n)}).$$

Since the outer weights are always initialized as zero, we have

$$(\mathbf{a}^{(l)})^{(0)} = \mathbf{a}^{(0)} = 0$$

for all $l \in \{1, \dots, I_n\}$ and consequently the bound in (3.37) for $t \in \{0, \dots, t_n\}$

$$F((\mathbf{a}^{(l)})^{(0)}, (\mathbf{b}^{(l)})^{(t)}) = F(\mathbf{a}^{(0)}, (\mathbf{b}^{(l)})^{(t)}) = F(0, (\mathbf{b}^{(l)})^{(t)}) \leq (c_3 \cdot \log n)^2$$

holds for all $l \in \{1, \dots, I_n\}$. Application of Lemma 3.2.10 gives us by construction of our neural network estimate

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n |Y_i - \tilde{m}_n(X_i)|^2 \cdot \mathbf{1}_{\{|Y_i| \leq \beta_n \text{ for all } i \in \{1,\dots,n\}\}} \\
\leq \; & \frac{1}{n} \sum_{i=1}^n |Y_i - \tilde{m}_n(X_i)|^2 \cdot \mathbf{1}_{\{|Y_i| \leq \beta_n \text{ for all } i \in \{1,\dots,n\}\}} + \frac{1}{n} \sum_{k=0}^{K \cdot r} (a^{(t_n)})_k^2 \\
= \; & \min_{l=1,\dots,I_n} \left(\frac{1}{n} \sum_{i=1}^n |Y_i - \tilde{m}_n(X_i)|^2 \cdot \mathbf{1}_{\{|Y_i| \leq \beta_n \text{ for all } i \in \{1,\dots,n\}\}} + \frac{1}{n} \sum_{k=0}^{K \cdot r} ((a^{(l)})_k^{(t_n)})^2 \right) \\
= \; & \min_{l=1,\dots,I_n} \left(F((\mathbf{a}^{(l)})^{(t_n)}, (\mathbf{b}^{(l)})^{(t_n)}) \right) \\
\leq \; & \min_{l=1,\dots,I_n} \left(\left(1 - \frac{2 \cdot c_1}{3 \cdot K \cdot r \cdot n}\right)^{t_n+1} \cdot \left(F(\mathbf{a}^{(0)}, (\mathbf{b}^{(l)})^{(0)}) - \min_{\mathbf{a}} F(\mathbf{a}, (\mathbf{b}^{(l)})^{(0)}) \right) \right.
\end{aligned}$$

$$
\begin{aligned}
& +(2\cdot\sqrt{c_{5,n}}+1)\cdot\exp\left(-\frac{\delta_n}{4}\right)+\frac{3\cdot K\cdot r\cdot n}{2\cdot c_1}\cdot 3\cdot\exp\left(-\frac{\delta_n}{4}\right) \\
& +\min_{\mathbf{a}} F(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})\Bigg) \\
\leq\ & \min_{l=1,\ldots,I_n}\Bigg(\left(1-\frac{2\cdot c_1}{3\cdot K\cdot r\cdot n}\right)^{t_n+1}\cdot(c_3\cdot\log n)^2+(2\cdot c_3\cdot\log n+1)\cdot\exp\left(-\frac{\delta_n}{4}\right) \\
& +\frac{3\cdot K\cdot r\cdot n}{2\cdot c_1}\cdot 3\cdot\exp\left(-\frac{\delta_n}{4}\right)+\min_{\mathbf{a}} F(\mathbf{a},\mathbf{b}^{(0)})\Bigg) \\
=\ & \min_{\mathbf{a}\in\mathbb{R}^{K+1},l=1,\ldots,I_n}\left(\frac{1}{n}\sum_{i=1}^{n}|Y_i-f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(X_i)|^2+\frac{c_1}{n}\sum_{k=0}^{K\cdot r}a_k^2+\epsilon_n\right)
\end{aligned}
$$

where

$$
\begin{aligned}
\epsilon_n\ =\ & \left(1-\frac{2\cdot c_1}{3\cdot K\cdot r\cdot n}\right)^{t_n+1}\cdot(c_3\cdot\log n)^2+(2\cdot c_3\cdot\log n+1)\cdot\exp\left(-\frac{\delta_n}{4}\right) \\
& +\frac{3\cdot K\cdot r\cdot n}{2\cdot c_1}\cdot 3\cdot\exp\left(-\frac{\delta_n}{4}\right).
\end{aligned}
$$

This verifies (3.35). Now, we can bound the first summand by Lemma 3.2.11 and Lemma 2.2.15 which yields

$$
\begin{aligned}
& \mathbf{E}\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx) \\
\leq\ & \frac{c_7\cdot(\log n)^2\cdot(c_9\cdot(\log n)\cdot K\cdot r)}{n} \\
& +2\cdot\mathbf{E}\left(\min_{\mathbf{a}\in\mathbb{R}^{K\cdot r+1},l=1,\ldots,I_n}\frac{1}{n}\sum_{i=1}^{n}|f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(X_i)-m(X_i)|^2+\frac{c_1}{n}\sum_{k=0}^{K\cdot r}a_k^2+\epsilon_n\right) \\
\leq\ & c_{10}\cdot\left(\frac{(\log n)^3}{n}\right)^{\frac{2p}{2p+1}} \\
& +2\cdot\mathbf{E}\left(\min_{\mathbf{a}\in\mathbb{R}^{K\cdot r+1},l=1,\ldots,I_n}\frac{1}{n}\sum_{i=1}^{n}|f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(X_i)-m(X_i)|^2+\frac{c_1}{n}\sum_{k=0}^{K\cdot r}a_k^2\right) \\
& +2\cdot\epsilon_n.
\end{aligned}
$$

We bound the second and the third summand on the right-hand side further. For the third

summand we have

$$\begin{aligned}
\epsilon_n &\leq \exp\left(-\frac{2\cdot c_1}{3\cdot r}\cdot(\log n)^2\right)\cdot(c_3\cdot\log n)^2+(2\cdot c_3\cdot\log n+1)\cdot\exp\left(-\frac{\sqrt{d}\cdot A\cdot n}{8}\right)\\
&\quad+\frac{3\cdot K\cdot r\cdot n}{2\cdot c_1}\cdot 3\cdot\exp\left(-\frac{\sqrt{d}\cdot A}{8}\cdot n\right)\\
&\leq c_{18}\cdot\frac{1}{n^{\frac{2\cdot c_1}{3\cdot r}\cdot\log n}}\cdot(c_3\cdot\log n)^2\\
&\leq \frac{1}{n}.
\end{aligned}$$

Now, we look at

$$\min_{\mathbf{a}\in\mathbb{R}^{K\cdot r+1},l=1,\ldots,I_n}\frac{1}{n}\sum_{i=1}^{n}|f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(x_i)-m(x_i)|^2+\frac{c_1}{n}\sum_{k=0}^{K\cdot r}a_k^2$$

for $x_i\in[-A,A]^d$ $(i=1,\ldots,n)$. We have

$$\begin{aligned}
&|f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(x_i)-m(x_i)|^2+\frac{c_1}{n}\sum_{k=0}^{K\cdot r}a_k^2\\
=\;&|f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(x_i)-\sum_{s=1}^{r}g_s((\bar{\mathbf{c}}_s^{(l)})^Tx_i)+\sum_{s=1}^{r}g_s((\bar{\mathbf{c}}_s^{(l)})^Tx_i)-m(x_i)|^2+\frac{c_1}{n}\sum_{k=0}^{K\cdot r}a_k^2\\
\leq\;&2\cdot|\sum_{s=1}^{r}g_s((\bar{\mathbf{c}}_s^{(l)})^Tx_i)-f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(x_i)|^2+\frac{c_1}{n}\sum_{k=0}^{K\cdot r}a_k^2\\
&+2\cdot|m(x_i)-\sum_{s=1}^{r}g_s((\bar{\mathbf{c}}_s^{(l)})^Tx_i)|^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\min_{\mathbf{a}\in\mathbb{R}^{K\cdot r+1},l=1,\ldots,I_n}\frac{1}{n}\sum_{i=1}^{n}|f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(x_i)-m(x_i)|^2+\frac{c_1}{n}\sum_{k=0}^{K\cdot r}a_k^2\\
\leq\;&2\cdot\min_{l=1,\ldots,I_n}\left(\min_{\mathbf{a}\in\mathbb{R}^{K\cdot r+1}}\left(|\sum_{s=1}^{r}g_s((\bar{\mathbf{c}}_s^{(l)})^Tx_i)-f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(x_i)|^2+\frac{c_1}{n}\sum_{k=0}^{K\cdot r}a_k^2\right)\right.
\end{aligned}$$

$$+|m(x_i) - \sum_{s=1}^{r} g_s((\bar{\mathbf{c}}_s^{(l)})^T x_i)|^2\Big)$$

We start with the first term on the right-hand side which we bound with Lemma 3.2.6. The conditions of Lemma 3.2.6 are satisfied by construction of our neural network estimate. We bound the term from above by plugging in specific values for $\mathbf{a}$. More precisely, we take the values for the outer weights from Lemma 3.2.6 for each g_s $(s = 1, \dots, r)$. In order to do this let for $s = 1, \dots, r$

$$\tilde{b}_{s,0} = \tilde{b}_{s,1} - \frac{4 \cdot A \cdot \sqrt{d}}{K-1}$$

and

$$\tilde{a}_{s,0} = g_s(\tilde{b}_{s,0}) \quad \text{and} \quad \tilde{a}_{s,k} = g_s(\tilde{b}_{s,k}) - g_s(\tilde{b}_{s,k-1}) \quad (k = 1, \dots, K).$$

Denote by

$$\tilde{\mathbf{a}} = (\tilde{a}_0, \tilde{a}_1, \dots, \tilde{a}_{K \cdot r})$$

with

$$\tilde{a}_0 = \sum_{s=1}^{r} \tilde{a}_{s,0} \quad \text{and} \quad \tilde{a}_{(s-1)\cdot K+k} = \tilde{a}_{s,k}$$

for $k = 1 \dots, K$ and $s = 1, \dots, r$. By Lemma 3.2.6 this gives us

$$\begin{aligned}
& \min_{\mathbf{a} \in \mathbb{R}^{K \cdot r+1}} |\sum_{s=1}^{r} g_s((\bar{\mathbf{c}}_s^{(l)})^T x_i) - f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(x_i)|^2 + \frac{c_1}{n} \sum_{k=0}^{K \cdot r} a_k^2 \\
\leq \; & |\sum_{s=1}^{r} g_s((\bar{\mathbf{c}}_s^{(l)})^T x_i) - f_{net,(\tilde{\mathbf{a}},(\mathbf{b}^{(l)})^{(0)})}(x_i)|^2 + \frac{c_1}{n} \sum_{k=0}^{K \cdot r} \tilde{a}_k^2 \\
= \; & |\sum_{s=1}^{r} g_s((\bar{\mathbf{c}}_s^{(l)})^T x_i) - \\
& \qquad \sum_{s=1}^{r} \sum_{k=1}^{K} \tilde{a}_{(s-1)\cdot K+k} \cdot \sigma \left(\sum_{j=1}^{d} (b^{(l)})^{(0)}_{(s-1)\cdot K+k,j} \cdot x_i^{(j)} + (b^{(l)})^{(0)}_{(s-1)\cdot K+k,0}) \right) - \tilde{a}_0|^2 \\
& + \frac{c_1}{n} \sum_{k=0}^{K \cdot r} \tilde{a}_k^2 \\
\leq \; & \left(\sum_{s=1}^{r} |g_s((\bar{\mathbf{c}}_s^{(l)})^T x_i) - \sum_{k=1}^{K} \tilde{a}_{s,k} \cdot \sigma \left(\rho_n \cdot ((\bar{\mathbf{c}}_s^{(l)})^T x_i - (\tilde{b}^{(l)})_{s,k}) \right) - \tilde{a}_{s,0}| \right)^2
\end{aligned}$$

$$
\begin{aligned}
& +\frac{c_1}{n} \cdot (K \cdot r + 1) \cdot \beta_n^2 \\
\leq \quad & \left(r \cdot 3 \cdot C \cdot \frac{(4 \cdot A \cdot \sqrt{d})^p}{(K-1)^p} + r \cdot C \cdot (4 \cdot A \cdot \sqrt{d})^p \cdot (K-1)^{1-p} \cdot e^{-\frac{\rho n \cdot A \cdot \sqrt{d}}{(n+1) \cdot (K-1)}} \right)^2 \\
& + c_{19} \cdot \frac{1}{n} \cdot \frac{n^{\frac{1}{2p+1}}}{(\log n)^{\frac{3}{2p+1}}} \cdot (\log n)^2 \\
\leq \quad & c_{20} \cdot \frac{1}{K^{2p}} + c_{21} \cdot \frac{\log n}{n^{\frac{2p}{2p+1}}}.
\end{aligned}
$$

Next, we deal with the second term on the right-hand side. For $i = 1, \ldots, n$ we have by $(p, C)-$smoothness of m and by the Cauchy-Schwarz inequality

$$
\begin{aligned}
& |m(x_i) - \sum_{s=1}^{r} g_s((\bar{\mathbf{c}}_s^{(l)})^T x_i)|^2 \\
= \quad & |\sum_{s=1}^{r} g_s(\mathbf{c}_s^T x_i) - \sum_{s=1}^{r} g_s((\bar{\mathbf{c}}_s^{(l)})^T x_i)|^2 \\
\leq \quad & \left(\sum_{s=1}^{r} |g_s(\mathbf{c}_s^T x_i) - g_s((\bar{\mathbf{c}}_s^{(l)})^T x_i)| \right)^2 \\
\leq \quad & \left(\sum_{s=1}^{r} C \cdot |\mathbf{c}_s^T x_i - (\bar{\mathbf{c}}_s^{(l)})^T x_i|^p \right)^2 \\
\leq \quad & C^2 \cdot \left(\sum_{s=1}^{r} |(\mathbf{c}_s^T - (\bar{\mathbf{c}}_s^{(l)})^T) \cdot x_i|^p \right)^2 \\
\leq \quad & C^2 \cdot \left(\sum_{s=1}^{r} \|\mathbf{c}_s - (\bar{\mathbf{c}}_s^{(l)})\|^p \cdot \|x_i\|^p \right)^2 \\
\leq \quad & C^2 \cdot r^2 \cdot \max_{s=1,\ldots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\|^{2p} \cdot (A \cdot \sqrt{d})^{2p} \\
= \quad & c_{31} \cdot \max_{s=1,\ldots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\|^{2p}.
\end{aligned}
$$

So, together we get for $x_i \in [-A, A]^d$ $(i = 1, \ldots, n)$

$$
\min_{\mathbf{a} \in \mathbb{R}^{K \cdot r + 1}, l = 1, \ldots, I_n} \frac{1}{n} \sum_{i=1}^{n} |f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(x_i) - m(x_i)|^2 + \frac{c_1}{n} \sum_{k=0}^{K \cdot r} a_k^2
$$

$$
\begin{aligned}
&\leq 2 \cdot \min_{l=1,\dots,I_n} \left(c_{20} \cdot \frac{1}{K^{2p}} + c_{21} \cdot \frac{\log n}{n^{\frac{2p}{2p+1}}} + c_{31} \cdot \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\|^{2p} \right) \\
&\leq c_{23} \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}} + c_{22} \cdot \min_{l=1,\dots,I_n} \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\|^{2p}.
\end{aligned}
$$

and hence,

$$
\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
\leq\ & c_{10} \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}} + c_{23} \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}} \\
& + c_{22} \cdot \mathbf{E} \left\{ \min_{l=1,\dots,I_n} \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\|^{2p} \right\} \\
=\ & c_{24} \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}} + c_{22} \cdot \mathbf{E} \left\{ \min_{l=1,\dots,I_n} \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\|^{2p} \right\}. \qquad (3.38)
\end{aligned}
$$

So, it remains to bound

$$
\mathbf{E} \left\{ \min_{l=1,\dots,I_n} \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\|^{2p} \right\}.
$$

By the random choice of the vectors $\bar{\mathbf{c}}_s^{(l)}$ $(s = 1, \dots, r)$ we know for any $u \in (0, 2]$

$$
\begin{aligned}
\mathbf{P} \left\{ \min_{l=1,\dots,I_n} \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\| > u \right\} &= \prod_{l=1}^{I_n} \left(1 - \mathbf{P} \left\{ \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\| \leq u \right\} \right) \\
&= \left(1 - \prod_{s=1}^{r} \mathbf{P} \left\{ \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(1)}\| \leq u \right\} \right)^{I_n}.
\end{aligned}
$$

By construction we choose the direction vector $\bar{\mathbf{c}}_s^{(1)}$ $(s = 1, \dots, r)$ by choosing a vector $(\mathbf{c}_s^*)^{(1)}$ $(s \in \{1, \dots, r\})$ from the $d-$dimensional unit cube and projecting it onto the surface of the $d-$dimensional Eucledian unit ball with center 0. Note that for $s = 1, \dots, r$

$$
\mathbf{P}\{\mathbf{c}_s^* = 0\} = 0,
$$

which means we can assume w.l.o.g. $\|\mathbf{c}_s^*\| \neq 0$. So, $\bar{\mathbf{c}}_s^{(1)}$ is chosen if and only if $(\mathbf{c}_s^*)^{(1)}$ lies on the line segment that starts in 0, passes through $\bar{\mathbf{c}}_s^{(1)}$ and continues until it hits the

boundary of the unit cube. Accordingly, the chosen direction lies in a neighborhood of $\bar{\mathbf{c}}_s^{(1)}$ if and only if $(\mathbf{c}_s^*)^{(1)}$ lies in the cone that has apex 0, extends to the boundary of the unit cube and intersects the surface of the unit ball at exactly the neighborhood. This means that $\bar{\mathbf{c}}_s^{(1)}$ is randomly distributed on $\partial S_1^{(d)}(0)$, the $(d-1)-$dimensional surface of the $d-$dimensional unit ball. Let λ_d be the Lebesgue measure on the $\sigma-$algebra of Borel sets in $\mathbb{R}^d$ and let λ' be the $(d-1)-$dimensional Lebesgue measure on the $\sigma-$algebra of Borel sets in $\partial S_1^{(d)}(0)$. We want to show that for any B in the $\sigma-$algebra of Borel sets in $\partial S_1^{(d)}(0)$ it holds that

$$c_{26} \cdot \lambda'(B) \leq \mathbf{P}_{\bar{\mathbf{c}}_s^{(1)}}(B) \leq c_{27} \cdot \lambda'(B) \tag{3.39}$$

for some constants $c_{26}, c_{27} > 0$. For that let

- A_1 be the cone inside $S_1^{(d)}(0)$ with apex 0 that intersects $\partial S_1^{(d)}(0)$ in B,
- A_2 be the cone inside the unit cube with apex 0 that intersects $\partial S_1^{(d)}(0)$ in B and
- A_3 be the cone inside $S_{\sqrt{d}}^{(d)}(0)$ with apex 0 that intersects $\partial S_1^{(d)}(0)$ in B.

For a better understanding, a visualisation for the case $d = 2$ is given in Figure 3.3. By

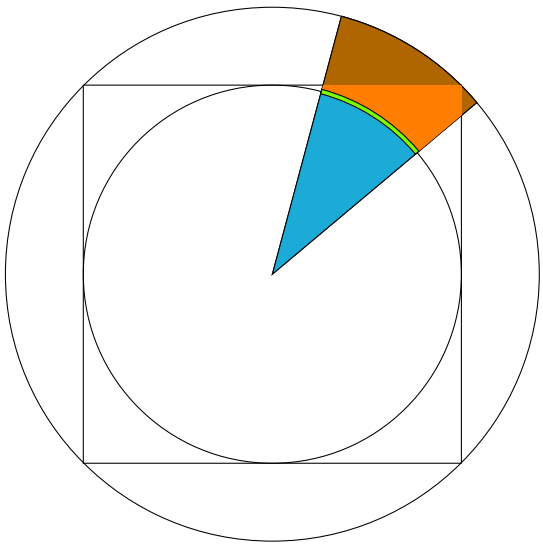

Figure 3.3.: Visualisation of the sets B (green), A_1 (blue area), A_2 (blue and orange area) and A_3 (blue, orange and brown area) for $d = 2$. The inner circle represents $S_1^{(2)}(0)$ and the outer circle represents $S_{\sqrt{2}}^{(2)}(0)$.

monotonicity of λ_d we have

$$\lambda_d(A_1) \leq \lambda_d(A_2) \leq \lambda_d(A_3)$$

and hence

$$\lambda'(B) = \frac{\lambda_d(A_1)}{\lambda_d(S_1^{(d)}(0))} \leq \frac{2^d \cdot \lambda_d(A_2)}{2^d \cdot \lambda_d(S_1^{(d)}(0))} = \frac{2^d}{\lambda_d(S_1^{(d)}(0))} \cdot \mathbf{P}_{\bar{\mathbf{c}}_s^{(1)}}(B)$$

and

$$\lambda'(B) = \frac{(\sqrt{d})^d \cdot \lambda_d(A_1)}{(\sqrt{d})^d \cdot \lambda_d(S_1^{(d)}(0))} = \frac{\lambda_d(A_3)}{\lambda_d(S_{\sqrt{d}}^{(d)}(0))} \geq \frac{2^d \cdot \lambda_d(A_2)}{2^d \cdot \lambda_d(S_{\sqrt{d}}^{(d)}(0))} = \frac{2^d}{\lambda_d(S_{\sqrt{d}}^{(d)}(0))} \cdot \mathbf{P}_{\bar{\mathbf{c}}_s^{(1)}}(B).$$

This proves (3.39). Note that (3.39) implies that $\mathbf{P}_{\bar{\mathbf{c}}_s^{(1)}}$ is absolutely continous with respect to λ' and by the Radon–Nikodym theorem we can conclude that $\bar{\mathbf{c}}_s^{(1)}$ has a density f with respect λ' on the $\sigma-$algebra of Borel sets in $\partial S_1^{(d)}(0)$ which is bounded away from zero.

For $d = 1$ we have $\bar{\mathbf{c}}_s^{(1)} \in \{-1, 1\}$. So, trivially, it holds that for $0 < u \leq 2$

$$\mathbf{P}\left\{\|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(1)}\| \leq u\right\} \geq \frac{1}{2}.$$

Next, we consider $d > 1$. By (3.39) we know that

$$\begin{aligned} \mathbf{P}\left\{\|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(1)}\| \leq u\right\} &= \mathbf{P}_{\bar{\mathbf{c}}_s^{(1)}}\left(\left\{\mathbf{w} \in \partial S_1^{(d)}(0) : \|\mathbf{c}_s - \mathbf{w}\| \leq u\right\}\right) \\ &\geq c_{26} \cdot \lambda'\left(\left\{\mathbf{w} \in \partial S_1^{(d)}(0) : \|\mathbf{c}_s - \mathbf{w}\| \leq u\right\}\right). \end{aligned}$$

By the rotational symmetry of λ' we can assume w.l.o.g. that

$$\mathbf{c}_s = (0, \ldots, 0, 1)^T.$$

In case that $u \geq \sqrt{2}$ the set $\left\{\mathbf{w} \in \partial S_1^{(d)}(0) : \|(0, \ldots, 0, 1)^T - \mathbf{w}\| \leq u\right\}$ contains the surface of the "upper hemisphere" $\left\{\mathbf{w} \in \partial S_1^{(d)}(0) : w^{(d)} \geq 0\right\}$. So, trivially, it holds that for $u \geq \sqrt{2}$

$$\begin{aligned} \lambda'\left(\left\{\mathbf{w} \in \partial S_1^{(d)}(0) : \|(0, \ldots, 0, 1)^T - \mathbf{w}\| \leq u\right\}\right) &\geq \lambda'\left(\left\{\mathbf{w} \in \partial S_1^{(d)}(0) : w^{(d)} \geq 0\right\}\right) \\ &\geq \frac{1}{2}. \end{aligned}$$

Hence, it suffices to consider $0 < u \leq \sqrt{2}$. We bound

$$\lambda'\left(\left\{\mathbf{w} \in \partial S_1^{(d)}(0) : \|(0,\ldots,0,1)^T - \mathbf{w}\| \leq u\right\}\right)$$

by exploiting monotonicity of λ'. We define a subset of $\left\{\mathbf{w} \in \partial S_1^{(d)}(0) : \|(0,\ldots,0,1)^T - \mathbf{w}\| \leq u\right\}$ for which the Lebesgue measure is easy to compute. For that, let

$$\mathbf{v} = \left(v^{(1)},\ldots,v^{(d)}\right)^T \in \partial S_1^{(d)}(0)$$

where

$$\|(v^{(1)}, v^{(2)}, \ldots, v^{(d-1)})\| \leq \frac{u}{\sqrt{2d}}$$

and

$$v^{(d)} > 0.$$

Since it must hold that $\|\mathbf{v}\| = 1$ we get

$$\begin{aligned}
1 &= \|\mathbf{v}\|^2 \\
&= \sum_{j=1}^{d-1}(v^{(j)})^2 + (v^{(d)})^2 \\
&= \|(v^{(1)}, v^{(2)}, \ldots, v^{(d-1)})\|^2 + (v^{(d)})^2 \\
&\leq \frac{u^2}{2d} + (v^{(d)})^2
\end{aligned}$$

meaning

$$v^{(d)} \geq \sqrt{1 - \frac{u^2}{2d}} \geq 1 - \frac{u^2}{2d}$$

where the last inequality holds because $1 - \frac{u^2}{2d} \leq 1$. Next, since

$$\begin{aligned}
\|(0,\ldots,0,1)^T - \mathbf{v}\|^2 &\leq \|(-v^{(1)}, -v^{(2)}, \ldots, -v^{(d-1)}, \left(1 - v^{(d)}\right))\|^2 \\
&= \frac{u^2}{2d} + \left(1 - v^{(d)}\right)^2 \\
&\leq \frac{u^2}{2d} + \left(1 - \left(1 - \frac{u^2}{2d}\right)\right)^2 \\
&\leq \frac{2 \cdot u^2}{2d}
\end{aligned}$$

$$\leq \quad u^2$$

where the second to last inequality holds because $\frac{u^2}{2d} \leq 1$, we know

$$\mathbf{v} \in \left\{ \mathbf{w} \in \partial S_1^{(d)}(0) : \|(0,\ldots,0,1)^T - \mathbf{w}\| \leq u \right\}.$$

Hence,

$$\begin{aligned} & \lambda' \left(\left\{ \mathbf{w} \in \partial S_1^{(d)}(0) : \|(0,\ldots,0,1)^T - \mathbf{w}\| \leq u \right\} \right) \\ \geq \quad & \lambda' \left(\left\{ \mathbf{v} \in \partial S_1^{(d)}(0) : \|(v^{(1)}, v^{(2)}, \ldots, v^{(d-1)})\| \leq \frac{u}{\sqrt{2d}}, v^{(d)} > 0 \right\} \right). \end{aligned}$$

By construction, the volume of the cone in the unit ball with apex 0 that intersects the unit ball at exactly

$$\left\{ \mathbf{v} \in \partial S_1^{(d)}(0) : \|(v^{(1)}, v^{(2)}, \ldots, v^{(d-1)})\| \leq \frac{u}{\sqrt{2d}}, v^{(d)} > 0 \right\}$$

is greater than the volume of the cone in the unit ball with apex 0 that has "flat" base

$$\left\{ \tilde{\mathbf{v}} \in \mathbb{R}^{d-1} : \|\tilde{\mathbf{v}}\| \leq \frac{u}{\sqrt{2d}} \right\}$$

and height

$$h = \sqrt{1 - \frac{u^2}{2d}} \geq \sqrt{1 - \frac{(\sqrt{2})^2}{2d}} \geq \sqrt{1 - \frac{1}{d}}.$$

For the volume of the latter cone we get

$$\frac{1}{d} \cdot \lambda_{d-1} \left(S_{\frac{u}{\sqrt{2d}}}^{(d-1)}(0) \right) \cdot h = \frac{1}{d} \cdot \left(\frac{u}{\sqrt{2d}} \right)^{d-1} \cdot \lambda_{d-1} \left(S_1^{(d-1)}(0) \right) \cdot h \geq c_{28} \cdot u^{d-1}.$$

Hence,

$$\begin{aligned} & \mathbf{P} \left\{ \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(1)}\| \leq u \right\} \\ \geq \quad & c_{26} \cdot \lambda' \left(\left\{ \mathbf{w} \in \partial S_1^{(d)}(0) : \|\mathbf{c}_s - \mathbf{w}\| \leq u \right\} \right) \\ \geq \quad & c_{26} \cdot \lambda' \left(\left\{ \mathbf{v} \in \partial S_1^{(d)}(0) : \|(v^{(1)}, v^{(2)}, \ldots, v^{(d-1)})\| \leq \frac{u}{\sqrt{2d}}, v^{(d)} > 0 \right\} \right) \end{aligned}$$

$$
\begin{aligned}
&\geq c_{26} \cdot \frac{c_{28} \cdot u^{d-1}}{\lambda_d \left(S_1^{(d)}(0) \right)} \\
&= c^* \cdot u^{d-1}. \qquad (3.40)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathbf{E}\left\{ \min_{l=1,\ldots,I_n} \max_{s=1,\ldots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\|^{2p} \right\} \\
&= \int_0^{2^{2p}} \mathbf{P}\left\{ \min_{l=1,\ldots,I_n} \max_{s=1,\ldots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\| > t^{\frac{1}{2p}} \right\} dt \\
&\leq \int_0^{2^{2p}} \left(1 - \left(c^* t^{\frac{d-1}{2p}} \right)^r \right)^{I_n} dt \\
&\leq \int_0^{2^{2p}} e^{-\left(c^* t^{\frac{d-1}{2p}} \right)^r I_n} dt \\
&\leq \frac{1}{(c^*)^{\frac{2p}{d-1}}} \cdot \frac{2p}{r(d-1)} \cdot \int_0^{\infty} e^{-v} v^{\frac{2p}{r(d-1)}-1} dv \cdot I_n^{-\frac{2p}{r(d-1)}} \\
&= \frac{1}{(c^*)^{\frac{2p}{d-1}}} \cdot \frac{2p}{r(d-1)} \cdot \Gamma\left(\frac{2p}{r(d-1)} \right) \cdot I_n^{-\frac{2p}{r(d-1)}} \\
&= c_{25} \cdot I_n^{-\frac{2p}{r(d-1)}} \\
&\leq c_{25} \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}},
\end{aligned}
$$

where the third inequality follows by substitution

$$
v = (c^*)^r t^{\frac{r(d-1)}{2p}} I_n \Leftrightarrow \left(\frac{v}{(c^*)^r \cdot I_n} \right)^{\frac{2p}{r(d-1)}} = t = \varphi(v)
$$

and

$$
\frac{d\varphi}{dv} = \frac{1}{(c^*)^{\frac{2p}{d-1}}} \cdot I_n^{-\frac{2p}{r(d-1)}} \cdot \frac{2p}{r(d-1)} \cdot v^{\frac{2p}{r(d-1)}-1}
$$

and where last equality follows since the the Gamma function Γ satisfies

$$
\Gamma\left(\frac{2p}{r(d-1)} \right) < \infty.
$$

Taking the above results together yields

$$\begin{aligned}
& \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
\leq \; & c_{24} \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}} + c_{22} \cdot c_{25} \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}} \\
= \; & c_4 \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}} .
\end{aligned}$$

This concludes the proof. □

3.3. Application to Simulated Data

We illustrate the finite sample size performance of our newly proposed estimate by applying it to simulated data using the software *MATLAB*.

For our simulation we choose the simulated data as follows: We choose X uniformly distributed on $[-1,1]^d$, where d is the dimension of the input, ϵ standard normal and independent of X, and we define Y by

$$Y = m_j(X) + \sigma \cdot \lambda_j \cdot \epsilon,$$

where $m_j : [-1,1]^d \to \mathbb{R}$ is described below, $\lambda_j > 0$ is a scaling value defined below and σ is chosen from $\{0.01, 0.05, 0.10, 0.20\}$ ($j \in \{1,2\}$). As regression functions we use

$$\begin{aligned}
& m_1(x^{(1)}, x^{(2)}, x^{(3)}) \\
= \; & \frac{1}{1 + \exp(-(0.8317x^{(1)}) - 0.0277x^{(2)} + 0.5545x^{(3)})} \\
& + \sqrt{(-0.6461x^{(1)} - 0.1412x^{(2)} + 0.7501x^{(3)})^2 + 1}
\end{aligned}$$

and

$$\begin{aligned}
& m_2(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}) \\
= \; & \frac{1}{1 + \exp(-(-0.4863x^{(1)} + 0.5976x^{(2)} - 0.0209x^{(3)} + 0.5949x^{(4)} - 0.2281x^{(5)})} \\
& + \log \left(\frac{1}{(-0.6236x^{(1)} + 0.1244x^{(2)} + 0.4735x^{(3)} + 0.1914x^{(4)} - 0.5786x^{(5)})^2 + 2} \right) .
\end{aligned}$$

λ_j is chosen approximately as IQR of a sample of size 100 of $m(X)$, and we use the values $\lambda_1 = 0.2444$ and $\lambda_2 = 0.2515$.

From this distribution we generate a sample of size $n = 100$ and apply our newly proposed neural network regression estimate and compare our results to that of six alternative regression estimates on the same data. Then we compute the L_2 errors of these estimates approximately by using the empirical L_2 error $\varepsilon_{L_2,\bar{N}}(\cdot)$ on an independent sample of X of size $\bar{N} = 10,000$. Since this error strongly depends on the behavior of the true function m_j, we set it in relation to the error of the simplest estimate for m_j we can think of, a completely constant function. The constant function estimate describes the average of the observed data according to the least squares approach. Thus, the scaled error measure we use for evaluation of the estimates is $\varepsilon_{L_2,\bar{N}}(m_{n,i})/\bar{\varepsilon}_{L_2,\bar{N}}(avg)$, where $\bar{\varepsilon}_{L_2,\bar{N}}(avg)$ is the median of 50 independent realizations of the value obtained if the average of n observations is plugged into $\varepsilon_{L_2,\bar{N}}(\cdot)$. To a certain extent, this quotient can be interpreted as the relative part of the error of the constant estimate that is still contained in the more sophisticated approaches. Of course, the resulting scaled errors depend on the random sample of (X, Y) and in order to still be able to compare these values we repeat the whole computation 50 times and report the median and the interquartile range of the 50 scaled errors for each of our estimates.

We choose the parameters for each of the estimates by splitting of the sample. Here we split our sample into a learning sample of size $n_l = 0.8 \cdot n$ and into a testing sample of size $n_t = 0.2 \cdot n$. We compute the estimate for all parameter values from the sets described below using the learning sample. Then, we compute the corresponding empirical L_2 risk on the testing sample and choose the parameter value which leads to the minimal empirical L_2 risk on the testing sample.

Our first three estimates are built-in fully connected neural network estimates where the number of layers is fixed and the number of neurons per layer is chosen adaptively. The estimate *fc-neural-1* has one hidden layer, estimate *fc-neural-3* has three hidden layers, estimate *fc-neural-6* has six hidden layers and the number of neurons per layer is chosen from the set $\{5, 10, 25, 50, 75\}$, $\{3, 6, 9, 12, 15\}$, $\{2, 4, 6, 8, 10\}$, respectively.

Our fourth estimate *kernel* is the Nadaraya-Watson kernel estimate with so-called naive kernel where the bandwith is chosen from the set $\{2^k : k \in \{-5, -4, \ldots, 5\}\}$.

Our fifth estimate *neighbor* is a nearest neighbor estimate where the number of nearest neighbors is chosen from the set $\{1, 2, 3\} \cup \{4, 8, 12, 16, \ldots, 4 \cdot \lfloor \frac{n_l}{4} \rfloor\}$.

Our sixth estimate *RBF* is the interpoland with radial basis functions where the radial basis functions $\Phi(r) = (1 - r)_+^6 \cdot (35 \cdot r^2 + 18 \cdot r + 3)$ is used and the scaling radius is chosen adaptively.

Our last estimate *neural-2* is our newly proposed neural network estimate presented in this chapter. Here, the following parameters of the estimate are fixed: N is set to 2, A

is set to 1, and R is set to 10^6, and r chosen from the set $\{2, 3, 4\}$. The parameter K of the estimate is chosen from the set $\{5, 10\}$. In order to accelerate the computation of this estimate we use only $I_n = 50$ random choices for the vectors of directions and $t_n = 1000$ gradient descent steps.

The results are summarized in Table 3.1 and in Table 3.2. As we can see from the reported scaled errors, our newly proposed neural network estimate does not perform as well as the built-in fully connected neural networks. This is can be explained by the small set from which we choose the parameters of the network and the few number of performed gradient descent steps. Unfortunately, we do not have the capacities to run the network with greater parameter set and a greater number of gradient descent steps. Still, our newly proposed neural network estimate even outperforms the other estimates in the case of 20% noise factor for m_2. While the built-in neural network estimates may show better performances than our neural network estimate, we would like to emphasize that the former have no theoretical background.

	m_1			
noise	1%	5%	10%	20%
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	0.0065	0.0065	0.0065	0.0066
approach	median (IQR)	median (IQR)	median (IQR)	median (IQR)
fc-neural-1	0.0358 (0.001)	0.0708 (0.001)	0.0581 (0.001)	0.1106 (0.001)
fc-neural-3	**0.0105 (0.002)**	0.0470 (0.002)	**0.0414 (0.001)**	0.1003 (0.002)
fc-neural-6	0.0209 (0.001)	**0.0361 (0.001)**	0.0497 (0.001)	**0.0867 (0.001)**
kernel	0.2478 (0.052)	0.2451 (0.067)	0.2436 (0.086)	0.246 (0.127)
neighbor	0.1168 (0.035)	0.1226 (0.046)	0.1815 (0.145)	0.2165 (0.121)
RBF	0.0117 (0.001)	0.2929 (0.012)	1.1759 (0.074)	5.9945 (3.003)
neural-2	0.1363 (0.116)	0.1421 (0.129)	0.1426 (0.121)	0.1665 (0.122)

Table 3.1.: Median and IQR of the scaled empirical L_2 error of estimates for m_1 for sample size $n = 100$. The smallest error values in each column is highlighted by bold letters.

	m_2			
noise	1%	5%	10%	20%
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	0.0073	0.0075	0.007	0.0073
approach	median (IQR)	median (IQR)	median (IQR)	median (IQR)
fc-neural-1	0.0278 (0.001)	0.0531 (0.004)	0.2241 (0.01)	0.5805 (0.006)
fc-neural-3	0.0567 (0.001)	0.0726 (0.001)	**0.0967 (0.002)**	1.2439 (0.005)
fc-neural-6	0.048 (0.002)	0.5121 (0.002)	0.4656 (0.002)	0.576 (0.005)
kernel	1.1081 (0.022)	1.1174 (0.013)	1.1386 (0.002)	1.2119 (0.040)
neighbor	0.3749 (0.158)	0.3978 (0.168)	0.4536 (0.195)	0.5734 (0.018)
RBF	**0.0038 (0.001)**	**0.0512 (0.013)**	0.1939 (0.039)	0.7595 (0.131)
neural-2	0.1624 (0.074)	0.1627 (0.083)	0.2618 (0.069)	**0.0561 (0.101)**

Table 3.2.: Median and IQR of the scaled empirical L_2 error of estimates for m_2 for sample size $n = 100$. The smallest error values in each column is highlighted by bold letters.

4. Neural Network Regression Estimates Inspired by Approximation Results with Piecewise Polynomials for Projection Pursuit

As in Chapter 3, we deal with neural network regression in a projection pursuit model. This means, we assume that the regression function satisfies

$$m(x) = \sum_{l=1}^{r} g_l \left(a_{(l-1)\cdot d+1} \cdot x^{(1)} + \cdots + a_{l\cdot d} \cdot x^{(d)} \right) \quad (x^{(1)}, \ldots, x^{(d)} \in \mathbb{R})$$

for some $r \in \mathbb{N}$, $\mathbf{a}_l = (a_{(l-1)\cdot d+1}, \ldots, a_{l\cdot d})^T \in \mathbb{R}^d$ where $\|\mathbf{a}_l\| = 1$ $(l = 1, \ldots, r)$, and (p, C)-smooth functions $g_s : \mathbb{R} \to \mathbb{R}$ $(s = 1, \ldots, r)$. The constraint imposed upon the regression function has the effect that the $d-$dimensional input is reduced to a $1-$dimensional input for the functions g_s that make up m. The natural question to ask is whether we can now achieve a univariate rate of convergence. In this chapter we present an implementable multilayer neural network regression estimate that achieves up to a logarithmic factor the univariate rate of convergence. We draw the inspiration for our neural network estimate from approximation of m by piecewise polynomials. Here, we present a new approximation result for a projection pursuit model by piecewise polynomials that is loosely related to an approximation result for $(p, C)-$smooth functions by a convex combination of Taylor polynomials by Schmidt-Hieber (2020). The idea is that we construct our neural network estimate by recreating this piecewise polynomial approximation using smaller neural networks with known approximation results as building blocks. In contrast to the neural network estimates presented in Chapter 2 and in Chapter 3, the inner weights are exactly prescribed and are thus chosen completely independently from the data set. The outer weights are chosen according to a regularized least squares criterion and we will show that they can be determined by solving a linear equation system. Since the projection directions are unknown, we will guess them repeatedly. This results in repeated initialization of the neural network from which we choose the one with minimal error. In comparison

with the neural network regression esimates presented in in Chapter 2 and in Chapter 3, we see that on the one hand there is no freedom in the choice of the weights but on the other hand we are able to analyze multilayer neural network estimates for multivariate $(p, C)-$smooth functions where the smoothness factor p has no restrictions.
We construct our neural network regression estimate in Section 4.1 and show the corresponding univariate rate of convergence result in Section 4.2. The finite sample size performance of our newly proposed estimate is illustrated in Section 4.3 by applying it to simulated data.

4.1. Constructing the Neural Network

The construction of our neural network estimate is motivated by an approximation in two steps:

1. Approximate the $(p, C)-$smooth regression function m in the projection pursuit model by a sum of convex combinations of polynomials.

2. Approximate the piecewise polynomials by neural networks.

In the following we will first present our approximation result for a projection pursuit model by piecewise polynomials in Section 4.1.1 that will set the underlying structure of our neural network regression estimate. We will then introduce the neural networks that we use to build our neural network estimate in Section 4.1.2. After that we will present the structure of our neural network estimate in Section 4.1.3 and we will choose the outer weights in Section 4.1.4. Lastly, we will explain the procedure of guessing the projection direction vectors which will also summarize the algorithm for the construction of our neural network estimate in Section 4.1.5.

4.1.1. Approximating a Projection Pursuit Model by Piecewise Polynomials

Let $A \geq 1$. We draw our inspiration for our neural network estimate from the approximation result of a $(p, C)-$smooth function $m : \mathbb{R}^d \to \mathbb{R}$ for $x \in [-A, A]^d$ by a local convex combination of polynomials. The following lemma gives an approximation result for a $(p, C)-$smooth function $m : \mathbb{R}^d \to \mathbb{R}$ in the projection pursuit model

$$m(x) = \sum_{l=1}^{r} g_l(\mathbf{a}_l^T x) \quad (x \in \mathbb{R}^d)$$

with $r \in \mathbb{N}$, (p, C)-smooth functions $g_l : \mathbb{R} \to \mathbb{R}$ $(l = 1, \ldots, r)$ and with projection directions $\mathbf{a}_l \in \mathbb{R}^d$ $(l = 1, \ldots, r)$.

Lemma 4.1.1. *Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $s \in (0, 1]$. Let $C > 0$, $r \in \mathbb{N}$, $g_l : \mathbb{R} \to \mathbb{R}$ (p, C)-smooth functions $(l = 1, \dots, r)$ and $\mathbf{a}_l \in \mathbb{R}^d$ $(l = 1, \dots, r)$. Set*

$$m(x) = \sum_{l=1}^{r} g_l(\mathbf{a}_l^T x) \quad (x \in \mathbb{R}^d).$$

For $\mathbf{b}_l \in \mathbb{R}^d$ $(l = 1, \dots, r)$ set

$$g(x) = \sum_{l=1}^{r} \sum_{j=0}^{q} \frac{g_l^{(j)}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j,$$

where $g_l^{(j)}$ denotes the j–th derivative of g_l. Then we have for any $x \in \mathbb{R}^d$

$$|m(x) - g(x)| \leq \frac{r \cdot d^p \cdot C}{q!} \cdot \|x\|_\infty^p \cdot \|\mathbf{a}_l - \mathbf{b}_l\|_\infty^p.$$

Proof. By the proof of Lemma 11.1 in Györfi et al. (2002) we have for any $z \in \mathbb{R}$

$$\left| g_l(u) - \sum_{j=0}^{q} \frac{g_l^{(j)}(z)}{j!} \cdot (u - z)^j \right| \leq \frac{1}{q!} \cdot C \cdot |u - z|^p \quad (u \in \mathbb{R}).$$

Applying this with $u = \mathbf{a}_l^T x$ and $z = \mathbf{b}_l^T x$ we get by the Cauchy-Schwarz inequality

$$\begin{aligned}
& |m(x) - g(x)| \\
\leq \; & \sum_{l=1}^{r} \left| g_l(\mathbf{a}_l^T x) - \sum_{j=0}^{q} \frac{g_l^{(j)}(\mathbf{b}_l^T x)}{j!} \cdot (\mathbf{a}_l^T x - \mathbf{b}_l^T x)^j \right| \\
\leq \; & \sum_{l=1}^{r} \frac{1}{q!} \cdot C \cdot |\mathbf{a}_l^T x - \mathbf{b}_l^T x|^p \\
\leq \; & \frac{r \cdot d^p \cdot C}{q!} \cdot \|x\|_\infty^p \cdot \|\mathbf{a}_l - \mathbf{b}_l\|_\infty^p.
\end{aligned}$$

□

Next, the following lemma gives an approximation result for the approximation in Lemma 4.1.1 by piecewise polynomials. For that we choose grid points from the interval $[-\sqrt{d} \cdot A, \sqrt{d} \cdot A]$ since we have for $x \in [-A, A]^d$ and for $l = 1, \dots, r$

$$|\mathbf{a}_l x| = \sqrt{\left(\sum_{j=1}^{d} a_l^{(j)} x^{(j)} \right)^2} \leq \sqrt{d \cdot A^2 \cdot \sum_{j=1}^{d} (a_l^{(j)})^2} = \sqrt{d} \cdot A \cdot \|\mathbf{a}_l\| = \sqrt{d} \cdot A.$$

Lemma 4.1.2. *Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $s \in (0,1]$. Let $C > 0$, $r \in \mathbb{N}$, $g_l : \mathbb{R} \to \mathbb{R}$ (p,C)-smooth functions $(l = 1,\dots,r)$ and $\mathbf{a}_l, \mathbf{b}_l \in \mathbb{R}^d$ with $\|\mathbf{a}_l\| = 1$ and $\|\mathbf{b}_l\| = 1$ $(l = 1,\dots,r)$. Let $A \geq 1$, $M \in \mathbb{N}$, set*

$$u_i = -\sqrt{d} \cdot A + i \cdot \frac{2 \cdot \sqrt{d} \cdot A}{M} \quad (i = 0,\dots,M)$$

and set $\{i_1,\dots,i_{M+1}\} = \{0,\dots,M\}$. Denote by $g_l^{(j)}$ the j–th derivative of g_l $(l = 1,\dots,r)$. Then there exist polynomials $p_{i_k,l} : \mathbb{R}^d \to \mathbb{R}$ of total degree q, which depend on $\mathbf{a}_l$ and $\mathbf{b}_l$ and where all coefficients are bounded in absolute value by

$$(p+1)^{d+2} \cdot 2^p \cdot d^{\frac{3p}{2}} \cdot A^p \cdot \max_{l \in \{1,\dots,r\}, j \in \{0,\dots,d\}} \|g_l^{(j)}\|_\infty \cdot \max\{\max_{l \in \{1,\dots,r\}} \|\mathbf{a}_l - \mathbf{b}_l\|_\infty^p, 1\}$$

such that we have for all $x \in [-A,A]^d$

$$\begin{aligned} &\Bigg| \sum_{l=1}^{r} \sum_{j=0}^{q} \frac{g_l^{(j)}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j \\ &\quad - \sum_{l=1}^{r} \sum_{k=1}^{M+1} p_{i_k,l}(x) \cdot \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+ \Bigg| \\ \leq \quad & r \cdot 2^p \cdot (p+1) \cdot C \cdot d^{\frac{3p}{2}} \cdot A^{2p} \cdot \left(\max\left\{\frac{1}{M}, \max_{l=1,\dots,r} \|\mathbf{a}_l - \mathbf{b}_l\|_\infty\right\}\right)^p. \end{aligned}$$

Proof. Let

$$p_{l,j,i_k} = \sum_{\tau=0}^{q-j} \frac{g_l^{(j+\tau)}(u_{i_k})}{\tau!} \cdot (\mathbf{b}_l^T x - u_{i_k})^\tau$$

be the Taylor polynomial of $g_l^{(j)}$ of degree $q - j$ around u_{i_k}. Since g_l is $(p,C)-$smooth we know that $g_l^{(j)}$ is $(p-j,C)-$smooth. So, by the proof of Lemma 11.1 in Györfi et al. (2002) we have that

$$\left| g_l^{(j)}(\mathbf{b}_l^T x) - p_{l,j,i_k}(\mathbf{b}_l^T x) \right| \leq \frac{1}{(q-j)!} \cdot C \cdot |\mathbf{b}_l^T x - u_{i_k}|^{(p-j)}.$$

From this we can conclude with the Cauchy-Schwarz inequality for $x \in [-A,A]^d$

$$\left| \frac{g_l^{(j)}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j - \frac{p_{l,j,i_k}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j \right|$$

$$
\begin{aligned}
&= \frac{1}{j!} \cdot \left| g_l^{(j)}(\mathbf{b}_l^T x) - p_{l,j,i_k}(\mathbf{b}_l^T x) \right| \cdot |(\mathbf{a}_l - \mathbf{b}_l)^T x|^j \\
&\leq \frac{1}{j!} \cdot \frac{1}{(q-j)!} \cdot C \cdot |\mathbf{b}_l^T x - u_{i_k}|^{(p-j)} \cdot (\|\mathbf{a}_l - \mathbf{b}_l\| \cdot \|x\|)^j \\
&\leq \frac{1}{(q-j)!} \cdot C \cdot |\mathbf{b}_l^T x - u_{i_k}|^{(p-j)} \cdot (\sqrt{d})^j \cdot \|\mathbf{a}_l - \mathbf{b}_l\|_\infty^j \cdot (\sqrt{d})^j \cdot \|x\|_\infty^j \\
&\leq \frac{1}{(q-j)!} \cdot C \cdot d^j \cdot A^j \cdot \left(\max\left\{|\mathbf{b}_l^T x - u_{i_k}|, \|\mathbf{a}_l - \mathbf{b}_l\|_\infty\right\}\right)^p . \qquad (4.1)
\end{aligned}
$$

By definition, in case that $u_{i_\theta} < \mathbf{b}_l^T x < u_{i_\theta+1}$ for $\theta \in \{1, \ldots, M+1\}$ we have for $x \in [-A, A]^d$

$$
\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_\theta}| = \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (\mathbf{b}_l^T x - u_{i_\theta}),
$$

$$
\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_\theta+1}| = \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (u_{i_\theta+1} - \mathbf{b}_l^T x)
$$

and for $i_k \neq i_\theta$, $i_k \neq (i_\theta + 1)$, $k \in \{1, \ldots, M+1\}$

$$
\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}| \geq \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot \frac{2 \cdot \sqrt{d} \cdot A}{M} = 1.
$$

Hence, for $u_{i_\theta} < \mathbf{b}_l^T x < u_{i_\theta+1}$ for $\theta \in \{1, \ldots, M+1\}$ and for $x \in [-A, A]^d$

$$
\begin{aligned}
&\sum_{k=1}^{M+1} \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+ \\
&= \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (\mathbf{b}_l^T x - u_{i_\theta})\right) + \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (u_{i_\theta+1} - \mathbf{b}_l^T x)\right) \\
&= 2 - \left(\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (u_{i_\theta+1} - u_{i_\theta})\right) \\
&= 1.
\end{aligned}
$$

In case that $\mathbf{b}_l^T x = u_{i_\theta}$ for $\theta \in \{1, \ldots, M+1\}$ we have for $x \in [-A, A]^d$

$$
\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_\theta}| = \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (\mathbf{b}_l^T x - u_{i_\theta}) = 0
$$

and for $i_k \neq i_\theta$, $k \in \{1, \ldots, M+1\}$

$$
\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}| \geq \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot \frac{2 \cdot \sqrt{d} \cdot A}{M} = 1.
$$

Hence, for $\mathbf{b}_l^T x = u_{i_\theta}$ for $\theta \in \{1, \dots, M+1\}$ and for $x \in [-A, A]^d$

$$\sum_{k=1}^{M+1} \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+ = 1.$$

As a consequence, with (4.1) and with $q < p$ we can conclude for $x \in [-A, A]^d$

$$\begin{aligned}
& \Bigg| \sum_{j=0}^{q} \frac{g_l^{(j)}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j \\
& \qquad - \sum_{j=0}^{q} \sum_{k=1}^{M+1} \frac{p_{l,j,i_k}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j \cdot \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+ \Bigg| \\
= & \Bigg| \sum_{j=0}^{q} \frac{g_l^{(j)}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j \cdot \sum_{k=1}^{M+1} \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+ \\
& \qquad - \sum_{j=0}^{q} \sum_{k=1}^{M+1} \frac{p_{l,j,i_k}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j \cdot \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+ \Bigg| \\
\leq & \sum_{j=0}^{q} \sum_{k=1}^{M+1} \left| \frac{g_l^{(j)}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j - \frac{p_{l,j,i_k}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j \right| \\
& \qquad \cdot \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+ \\
\leq & \sum_{j=0}^{q} \max_{\substack{i_k \in \{0, \dots, M\}, \\ |b_l^T x - u_{i_k}| \leq 2 \cdot \sqrt{d} \cdot A/M}} \left| \frac{g_l^{(j)}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j - \frac{p_{l,j,i_k}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j \right| \\
\leq & (q+1) \cdot C \cdot d^q \cdot A^q \cdot \left(\max \left\{ \frac{2 \cdot \sqrt{d} \cdot A}{M}, \|\mathbf{a}_l - \mathbf{b}_l\|_\infty \right\} \right)^p \\
\leq & (p+1) \cdot C \cdot d^p \cdot A^p \cdot (2 \cdot \sqrt{d} \cdot A)^p \cdot \left(\max \left\{ \frac{1}{M}, \|\mathbf{a}_l - \mathbf{b}_l\|_\infty \right\} \right)^p \\
= & (p+1) \cdot C \cdot d^{\frac{3p}{2}} \cdot 2^p \cdot A^{2p} \cdot \left(\max \left\{ \frac{1}{M}, \|\mathbf{a}_l - \mathbf{b}_l\|_\infty \right\} \right)^p .
\end{aligned}$$

With

$$p_{i_k,l}(x) = \sum_{j=0}^{q} \frac{p_{l,j,i_k}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j$$

we get the assertion. It remains to show the bounds on the coefficients. We have

$$
\begin{aligned}
& p_{i_k,l}(x) \\
= & \sum_{j=0}^{q}\sum_{\tau=0}^{q-j} \frac{g_l^{(j+\tau)}(u_{i_k})}{\tau!} \cdot (\mathbf{b}_l^T x - u_{i_k})^\tau \cdot \frac{1}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j \\
= & \sum_{j=0}^{q}\sum_{\tau=0}^{q-j} \frac{g_l^{(j+\tau)}(u_{i_k})}{\tau!} \cdot \sum_{\delta=0}^{\tau} \binom{\tau}{\delta} \cdot u_{i_k}^\delta \cdot (\mathbf{b}_l^T x)^{\tau-\delta} \cdot (-1)^{\tau-\delta} \\
& \qquad \cdot \frac{1}{j!} \sum_{\substack{\nu\in\mathbb{N}^d \\ |\nu|=j}} \binom{j}{\nu} \cdot (\mathbf{a}_l - \mathbf{b}_l)^\nu \cdot x^\nu \\
= & \sum_{j=0}^{q}\sum_{\tau=0}^{q-j} \frac{g_l^{(j+\tau)}(u_{i_k})}{\tau!} \cdot \sum_{\delta=0}^{\tau} \binom{\tau}{\delta} \cdot u_{i_k}^\delta \cdot \sum_{\substack{\nu'\in\mathbb{N}^d \\ |\nu'|=\tau-\delta}} \binom{\tau-\delta}{\nu'} \cdot \mathbf{b}_l^\nu \cdot x^{\nu'} \cdot (-1)^{\tau-\delta} \\
& \qquad \cdot \frac{1}{j!} \sum_{\substack{\nu\in\mathbb{N}^d \\ |\nu|=j}} \binom{j}{\nu} \cdot (\mathbf{a}_l - \mathbf{b}_l)^\nu \cdot x^\nu .
\end{aligned}
$$

We look at the term separately. We start with

$$
\sum_{\delta=0}^{\tau} \sum_{\substack{\nu'\in\mathbb{N}^d \\ |\nu'|=\tau-\delta}} (-1)^{\tau-\delta} \cdot \binom{\tau}{\delta} \cdot u_{i_k}^\delta \cdot \binom{\tau-\delta}{\nu'} \cdot \mathbf{b}_l^\nu \cdot x^{\nu'}
$$

and we see that, since $\|\mathbf{b}_l\| = 1$ implies $b_l^{(i)} \leq 1$ for all $i = 1, \ldots, d$,

$$
\left| (-1)^{\tau-\delta} \cdot \binom{\tau}{\delta} \cdot u_{i_k}^\delta \cdot \binom{\tau-\delta}{\nu'} \cdot \mathbf{b}_l^\nu \right| \leq 1 \cdot 2^\tau \cdot (\sqrt{d} \cdot A)^\delta \cdot d^{\tau-\delta} \cdot 1.
$$

Next, we look at

$$
\sum_{\substack{\nu\in\mathbb{N}^d \\ |\nu|=j}} \frac{g_l^{(j+\tau)}(u_{i_k})}{\tau!} \cdot \frac{1}{j!} \cdot \binom{j}{\nu} \cdot (\mathbf{a}_l - \mathbf{b}_l)^\nu \cdot x^\nu
$$

and we see that

$$
\frac{g_l^{(j+\tau)}(u_{i_k})}{\tau!} \cdot \frac{1}{j!} \cdot \binom{j}{\nu} \cdot (\mathbf{a}_l - \mathbf{b}_l)^\nu \leq \|g_l^{(j+\tau)}\|_\infty \cdot d^j \cdot \|\mathbf{a}_l - \mathbf{b}_l\|_\infty^j .
$$

Thus, we can conclude that the coefficients are bounded by

$$\begin{aligned} & (q+1)^2 \cdot q^d \cdot 2^q \cdot (\sqrt{d} \cdot A)^q \cdot d^q \cdot \max_{l \in \{1,\dots,r\}, j \in \{0,\dots,d\}} \|g_l^{(j)}\|_\infty \cdot \max_{l \in \{1,\dots,r\}} \|\mathbf{a}_l - \mathbf{b}_l\|_\infty^q \\ \leq \; & (p+1)^{d+2} \cdot 2^p \cdot d^{\frac{3p}{2}} \cdot A^p \cdot \max_{l \in \{1,\dots,r\}, j \in \{0,\dots,d\}} \|g_l^{(j)}\|_\infty \cdot \max\{\max_{l \in \{1,\dots,r\}} \|\mathbf{a}_l - \mathbf{b}_l\|_\infty^p, 1\}. \end{aligned}$$

This concludes the proof.

□

From the approximation results in Lemma 4.1.1 and in Lemma 4.1.2 we know that we can approximate m by a convex combination of polynomials of total degree q of the form

$$\sum_{l=1}^{r} \sum_{k=1}^{M+1} \left(\sum_{j=0}^{q} \frac{p_{l,j,i_k}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j \right) \cdot \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+$$

where p_{l,j,i_k} is the Taylor polynomial of $g_l^{(j)}$ of degree $q - j$ around u_{i_k}. Hence, the approximation function is contained in a class of functions defined by

$$\left\{ \sum_{l=1}^{r} \sum_{k=1}^{M+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,q\}, \\ j_1+\dots+j_d \leq q}} a_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \cdot (x^{(1)})^{j_1} \cdot \dots \cdot (x^{(d)})^{j_d} \cdot \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+ : a_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \in \mathbb{R} \right\}.$$

Note, $\mathbf{b}_l$ is our choice for the direction vector.

4.1.2. Building Blocks

The key concept for the construction is to use smaller neural networks as building blocks to define a neural network $f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l}$ which approximates

$$x \mapsto (x^{(1)})^{j_1} \cdots (x^{(d)})^{j_d} \cdot \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}_l^T x - u_{i_k}|\right)_+$$

and to choose the network architecture such that neural networks of the form

$$\sum_{l=1}^{r} \sum_{k=1}^{M+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,q\} \\ j_1+\dots+j_d \le q}} a_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \cdot f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l}(x) \quad (a_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \in \mathbb{R})$$

are contained in it. In order to do this, we introduce four neural networks. For the corresponding approximation results we need the following definition concerning the activation function of the neural network.

Definition 4.1.3. *Let $N \in \mathbb{N}_0$. A function $\sigma : \mathbb{R} \to [0,1]$ is called **N-admissible**, if it is nondecreasing and Lipschitz continuous and if, in addition, the following three conditions are satisfied:*

(i) The function σ is $N+1$ times continuously differentiable with bounded derivatives.

(ii) A point $t_\sigma \in \mathbb{R}$ exists, where all derivatives up to order N of σ are nonzero.

(iii) If $y > 0$, the relation $|\sigma(y) - 1| \le \frac{1}{y}$ holds. If $y < 0$, the relation $|\sigma(y)| \le \frac{1}{|y|}$ holds.

Remark 4.1.4. *It is easy to see that the logistic squasher*

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

is N–admissible for any $N \in \mathbb{N}$. A proof can be found in Bauer and Kohler (2019).

Choose $R \ge 1$.

First, we approximate the function

$$f(x) = x$$

by the neural network

$$f_{id}(x) = 4R \cdot \sigma\left(\frac{x}{R}\right) - 2R \tag{4.2}$$

which is depicted in Figure 4.1. We use the following approximation results for f_{id}.

Lemma 4.1.5. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a function, let $R, a > 0$. Assume that σ is two times continuously differentiable and let $t_{\sigma,id} \in \mathbb{R}$ be such that $\sigma'(t_{\sigma,id}) \neq 0$. Then*

$$f_{id}(x) = \frac{R}{\sigma'(t_{\sigma,id})} \cdot \left(\sigma\left(\frac{x}{R} + t_{\sigma,id}\right) - \sigma(t_{\sigma,id})\right)$$

satisfies for any $x \in [-a, a]$:

$$|f_{id}(x) - x| \le \frac{\|\sigma''\|_\infty \cdot a^2}{2 \cdot |\sigma'(t_{\sigma,id})|} \cdot \frac{1}{R}.$$

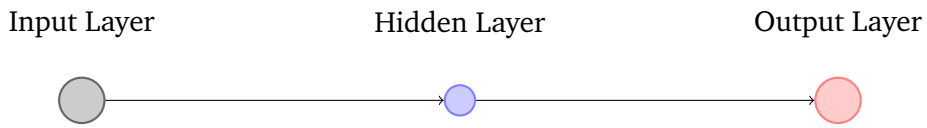

Figure 4.1.: Visualisation of the neural network f_{id}. The function has a 1-dimensional input (black node) as well as a 1-dimensional output (red node). There is one hidden layer consisting of 1 neuron (blue node).

Proof. The result follows in a straightforward way from the proof of Theorem 2 in Scarselli and Tsoi (1998), cf. Lemma 1 in Kohler, Krzyżak and Langer (2019). □

Remark 4.1.6. *Since the logistic squasher σ is $2-$admissible we apply Lemma* 4.1.5 *with* $t_{\sigma,id} = 0$*: With*

$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x)) \quad (x \in \mathbb{R})$$

we get

$$\sigma'(0) = \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) = \frac{1}{4}$$

and so the approximation results holds for

$$f_{id}(x) = \frac{R}{\frac{1}{4}} \cdot \left(\sigma\left(\frac{x}{R} + 0\right) - \sigma(0)\right) = 4R \cdot \left(\sigma\left(\frac{x}{R}\right) - \frac{1}{2}\right) = 4R \cdot \sigma\left(\frac{x}{R}\right) - 2R$$

which is (4.2).

Second, we approximate the function

$$f(x, y) = x \cdot y$$

by the neural network

$$\begin{aligned} f_{mult}(x, y) \quad = \quad & \frac{R^2}{4} \cdot \frac{(1 + e^{-1})^3}{e^{-2} - e^{-1}} \cdot \left(\sigma\left(\frac{2(x + y)}{R} + 1\right) - 2 \cdot \sigma\left(\frac{x + y}{R} + 1\right)\right. \\ & \left. -\sigma\left(\frac{2(x - y)}{R} + 1\right) + 2 \cdot \sigma\left(\frac{x - y}{R} + 1\right)\right), \end{aligned} \tag{4.3}$$

which is depicted in Figure 4.2. We use the following approximation results for f_{mult}.

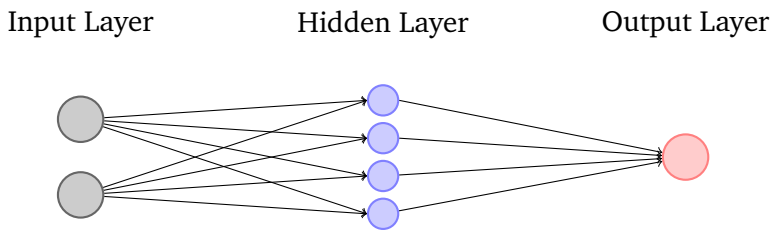

Figure 4.2.: Visualisation of the neural network f_{mult}. The function has a 2-dimensional input (black node) as well as a 1-dimensional output (red node). There is one hidden layer consisting of 4 neurons (blue node).

Lemma 4.1.7. *Let $\sigma : \mathbb{R} \to [0,1]$ be 2-admissible according to Definition 4.1.3. Then for any $R > 0$ and any $a > 0$ the neural network*

$$\begin{aligned} f_{mult}(x,y) &= \frac{R^2}{4 \cdot \sigma''(t_\sigma)} \cdot \left(\sigma\left(\frac{2 \cdot (x+y)}{R} + t_\sigma\right) - 2 \cdot \sigma\left(\frac{x+y}{R} + t_\sigma\right) \right. \\ &\qquad \left. -\sigma\left(\frac{2 \cdot (x-y)}{R} + t_\sigma\right) + 2 \cdot \sigma\left(\frac{x-y}{R} + t_\sigma\right) \right) \end{aligned}$$

satisfies for any $x \in [-a, a]$:

$$|f_{mult}(x,y) - x \cdot y| \leq \frac{20 \cdot \|\sigma'''\|_\infty \cdot a^3}{3 \cdot |\sigma''(t_\sigma)|} \cdot \frac{1}{R}.$$

Proof. See Lemma 2 in Kohler, Krzyżak and Langer (2019). □

Remark 4.1.8. *Since the logistic squasher σ is $2-$admissible we apply Lemma 4.1.7 with $t_\sigma = 1$. With*

$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x)) \quad (x \in \mathbb{R})$$

and

$$\sigma''(x) = \sigma(x) \cdot (1 - \sigma(x))^2 - \sigma(x)^2 \cdot (1 - \sigma(x))$$

we get

$$\sigma''(1) = \frac{1}{1+e^{-1}} \cdot \left(1 - \frac{1}{1+e^{-1}}\right)^2 - \left(\frac{1}{1+e^{-1}}\right)^2 \cdot \left(1 - \frac{1}{1+e^{-1}}\right) = \frac{e^{-2} - e^{-1}}{(1+e^{-1})^3}$$

and so the approximation results holds for

$$\begin{aligned} f_{mult}(x,y) &= \frac{R^2}{4\cdot\frac{e^{-2}-e^{-1}}{(1+e^{-1})^3}}\cdot\left(\sigma\left(\frac{2\cdot(x+y)}{R}+1\right)-2\cdot\sigma\left(\frac{x+y}{R}+1\right)\right. \\ &\qquad \left.-\sigma\left(\frac{2\cdot(x-y)}{R}+1\right)+2\cdot\sigma\left(\frac{x-y}{R}+1\right)\right) \end{aligned}$$

which is (4.3).

Third, we approximate the function

$$f(x) = x_+$$

by the neural network

$$f_{ReLu}(x) = f_{mult}(f_{id}(x), \sigma(R\cdot x)) \tag{4.4}$$

which is depicted in Figure 4.3 We use the following approximation results for f_{ReLu}.

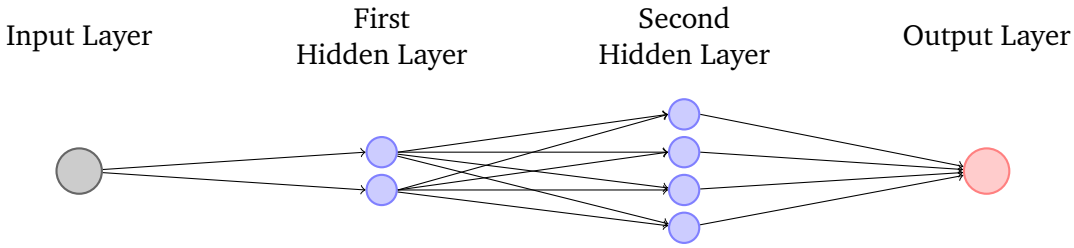

Figure 4.3.: Visualisation of the neural network f_{ReLu}. The function has a 1-dimensional input (black node) as well as a 1-dimensional output (red node). There are two hidden layer consisting of 2 neurons and 4 neurons, respectively (blue nodes).

Lemma 4.1.9. *Let $\sigma : \mathbb{R} \to [0,1]$ be 2-admissible according to Definition 4.1.3. Let f_{mult} be the neural network from Lemma 4.1.7 and let f_{id} be the network from Lemma 4.1.5. Assume*

$$a \geq 1 \quad \text{and} \quad R \geq \frac{\|\sigma''\|_\infty \cdot a}{2\cdot|\sigma'(t_{\sigma.id})|}.$$

Then the neural network

$$f_{ReLU}(x) = f_{mult}\left(f_{id}(x), \sigma\left(R\cdot x\right)\right)$$

$$= \sum_{k=1}^{4} d_k \cdot \sigma \left(\sum_{i=1}^{2} b_{k,i} \cdot \sigma(a_i \cdot x + t_\sigma) + b_{k,3} \cdot \sigma(a_3 \cdot x) + t_\sigma \right)$$

satisfies

$$|f_{ReLU}(x) - \max\{x, 0\}| \leq 56 \cdot \frac{\max\{\|\sigma''\|_\infty, \|\sigma'''\|_\infty, 1\}}{\min\{2 \cdot |\sigma'(t_{\sigma,id})|, |\sigma''(t_\sigma)|, 1\}} \cdot a^3 \cdot \frac{1}{R}$$

for all $x \in [-a, a]$.

Proof. See Lemma 3 in Kohler, Krzyżak and Langer (2019). □

Remark 4.1.10. *The approximation results in Lemma* 4.1.9 *hold for* f_{ReLu} *defined in* (4.4) *by Remark* 4.1.6 *and Remark* 4.1.8.

Fourth, we approximate for fixed $y \in \mathbb{R}$ the function

$$f(x) = (1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |x - y|)_+$$

by the neural network

$$\begin{aligned} f_{hat,y}(x) &= f_{ReLU}\left(\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (x - y) + 1\right) - 2 \cdot f_{ReLU}\left(\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (x - y)\right) \\ &\quad + f_{ReLU}\left(\frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot (x - y) - 1\right) \end{aligned} \tag{4.5}$$

which is depicted in Figure 4.4. We use the following approximation results for $f_{hat,y}(x)$.

Lemma 4.1.11. *Let* $M \in \mathbb{N}$ *and let* $\sigma : \mathbb{R} \to [0, 1]$ *be 2-admissible according to Definition 4.1.3. Let*

$$a > 0 \quad \text{and} \quad R \geq \frac{\|\sigma''\|_\infty \cdot (M + 1)}{2 \cdot |\sigma'(t_{\sigma.id})|},$$

let $y \in [-a, a]$ *and let* f_{ReLU} *be the neural network of Lemma 4.1.9. Then the network*

$$\begin{aligned} f_{hat,y}(x) &= f_{ReLU}\left(\frac{M}{2a} \cdot (x - y) + 1\right) - 2 \cdot f_{ReLU}\left(\frac{M}{2a} \cdot (x - y)\right) \\ &\quad + f_{ReLU}\left(\frac{M}{2a} \cdot (x - y) - 1\right) \end{aligned} \tag{4.6}$$

satisfies

$$\left| f_{hat,y}(x) - \left(1 - \frac{M}{2a} \cdot |x - y|\right)_+ \right| \leq 1792 \cdot \frac{\max\{\|\sigma''\|_\infty, \|\sigma'''\|_\infty, 1\}}{\min\{2 \cdot |\sigma'(t_{\sigma,id})|, |\sigma''(t_\sigma)|, 1\}} \cdot M^3 \cdot \frac{1}{R}$$

for all $x \in [-a, a]$.

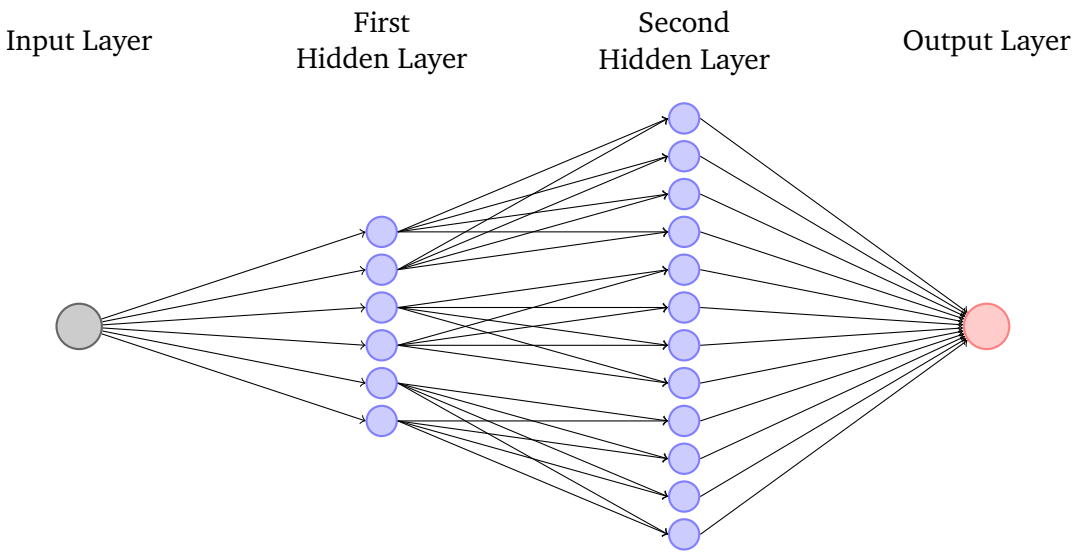

Figure 4.4.: Visualisation of the neural network $\bar{f}_{hat,y}(x)$. The function has a 1-dimensional input (black node) as well as a 1-dimensional output (red node). There are two hidden layers consisting of 6 neurons and 12 nodes, respectively (blue nodes). For readability, edges of weight 0 are omitted entirely.

Proof. Since

$$(1 - \frac{M}{2a} \cdot |x|)_+ = \max\{\frac{M}{2a} \cdot x + 1, 0\} - 2 \cdot \max\{\frac{M}{2a} \cdot x, 0\} + \max\{\frac{M}{2a} \cdot x - 1, 0\} \quad (x \in \mathbb{R})$$

and

$$x, y \in [-a, a] \Rightarrow \left(\frac{M}{2a} \cdot |x - y| + 1\right) \in [-(M+1), M+1],$$

we can apply Lemma 4.1.9 to each of the terms of $f_{hat,y}$ with $M + 1$ instead of a. Together with the triangle inequality we get

$$\begin{aligned} & \left| f_{hat,y}(x) - \left(1 - \frac{M}{2a} \cdot |x - y|\right)_+ \right| \\ \leq & \left| f_{ReLu}\left(\frac{M}{2a} \cdot (x - y) + 1\right) - \max\left\{\frac{M}{2a} \cdot (x - y) + 1, 0\right\} \right| \\ & +2 \cdot \left| f_{ReLu}\left(\frac{M}{2a} \cdot (x - y)\right) - \max\left\{\frac{M}{2a} \cdot (x - y), 0\right\} \right| \end{aligned}$$

$$
\begin{aligned}
&+\left| f_{ReLu}\left(\frac{M}{2a}\cdot(x-y)-1\right)-\max\left\{\frac{M}{2a}\cdot(x-y)-1,0\right\}\right| \\
\leq\ & (1+2+1)\cdot 56\cdot\frac{\max\{\|\sigma'''\|_\infty,\|\sigma''\|_\infty,1\}}{\min\{|\sigma''(t_\sigma)|,2\cdot|\sigma'(t_{\sigma,id})|,1\}}\cdot(M+1)^3\cdot\frac{1}{R} \\
\leq\ & 224\cdot\frac{\max\{\|\sigma'''\|_\infty,\|\sigma''\|_\infty,1\}}{\min\{|\sigma''(t_\sigma)|,2\cdot|\sigma'(t_{\sigma,id})|,1\}}\cdot(2M)^3\cdot\frac{1}{R} \\
=\ & 1792\cdot\frac{\max\{\|\sigma'''\|_\infty,\|\sigma''\|_\infty,1\}}{\min\{|\sigma''(t_\sigma)|,2\cdot|\sigma'(t_{\sigma,id})|,1\}}\cdot M^3\cdot\frac{1}{R}
\end{aligned}
$$

□

Remark 4.1.12. *The approximation results in Lemma* 4.1.11 *hold for* $f_{hat,y}(x)$ *defined in* (4.5) *by application of that lemma with* $a=\sqrt{d}\cdot A>0$ *together with Remark* 4.1.10.

4.1.3. The Network Architecture and the Inner Weights

With the building block networks presented Section 4.1.2 we will now recursively define the neural network $f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l}$ which approximates

$$
x\mapsto (x^{(1)})^{j_1}\cdots(x^{(d)})^{j_d}\cdot\left(1-\frac{M}{2\cdot\sqrt{d}\cdot A}\cdot|\mathbf{b}_l^T x-u_{i_k}|\right)_+ .
$$

First we give the definition of the neural network. We choose $N\geq q$, set $s=\lceil\log_2(N+1)\rceil$ and define for $l\in\{1,\dots,r\}$, $j_1,\dots,j_d\in\{0,1,\dots,N\}$ and $k\in\{1,\dots,M+1\}$

$$
f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l}(x)=f_1^{(0)}(x), \tag{4.7}
$$

where

$$
f_k^{(t)}(x)=f_{mult}\left(f_{2k-1}^{(t+1)}(x),f_{2k}^{(t+1)}(x)\right) \tag{4.8}
$$

for $k\in\{1,2,\dots,2^t\}$ and $t\in\{0,\dots,s-1\}$, and

$$
f_k^{(s)}(x)=f_{id}(f_{id}(x^{(t)})) \tag{4.9}
$$

for $j_1+j_2+\dots+j_{t-1}+1\leq k\leq j_1+j_2+\dots+j_t$ and $t=1,\dots,d$,

$$
f_{j_1+j_2+\dots+j_d+1}^{(s)}(x)=f_{hat,u_{i_k}}(\mathbf{b}_l^T x), \tag{4.10}
$$

and

$$
f_k^{(s)}(x)=1 \tag{4.11}
$$

for $k = j_1 + j_2 + \cdots + j_d + 2, j_1 + j_2 + \cdots + j_d + 3, \ldots, 2^s$. An example of the network is visualized in Figure 1.4.
It is easy to see that $f_{net,k,j_1,\ldots,j_d,\mathbf{b}_l}$ is a neural network with $s+2$ hidden layers and at most

$$6 \cdot 2^s,\ 12 \cdot 2^s,\ 2 \cdot 2^s, 2^s,\ \ldots, 8, 4$$

neurons in the layers $1, 2, \ldots, s+2$, respectively. Consequently, this network is contained in the class of all fully connected neural networks with $s+2$ hidden layers and $24 \cdot (N+1)$ neurons in each hidden layer. Furthermore, it is easy to see that all weights are bounded in absolute value by

$$c_8 \cdot \max\left\{1, \frac{M}{A}, R^2\right\}$$

for some constant $c_8 > 0$.
Second, we use the following approximation results for $f_{net,j_1,\ldots,j_d,i_k,\mathbf{b}_l}$.

Lemma 4.1.13. *Let $M \in \mathbb{N}$. Let $\sigma : \mathbb{R} \to [0,1]$ be 2-admissible according to Definition 4.1.3. Let $A \geq 1$, $\mathbf{b} \in \mathbb{R}^d$ with $\|\mathbf{b}\| = 1$. Let $N \in \mathbb{N}$ and let $j_1, \ldots, j_d \in \mathbb{N}_0$ such that $j_1 + \cdots + j_d \leq N$. Set $s = \lceil \log_2(N+d) \rceil$. Let*

$$\begin{aligned} R \quad \geq \quad \max\Bigg\{ & \frac{\|\sigma''\|_\infty \cdot (M+1)}{2 \cdot |\sigma'(t_{\sigma,id})|}, \frac{9 \cdot \|\sigma''\|_\infty \cdot A^2}{|\sigma'(t_{\sigma,id})|}, \frac{20 \cdot \|\sigma'''\|_\infty}{3 \cdot |\sigma''(t_\sigma)|} \cdot 3^{3 \cdot 3^s} \cdot A^{3 \cdot 2^s}, \\ & 1792 \cdot \frac{\max\{\|\sigma''\|_\infty, \|\sigma'''\|_\infty, 1\}}{\min\{2 \cdot |\sigma'(t_{\sigma,id})|, |\sigma''(t_\sigma)|, 1\}} \cdot d^{3/2} \cdot M^3 \Bigg\} \end{aligned} \tag{4.12}$$

and let $y \in [-A, A]$. Let f_{id}, f_{mult} and $f_{hat,z}$ (where $y = z$ in (4.5) for $z \in \mathbb{R}$) be the neural networks defined in Section 4.1.2. Define the network $f_{net,j_1,\ldots,j_d,y,\mathbf{b}}$ by

$$f_{net,j_1,\ldots,j_d,y,\mathbf{b}}(x) = f_1^{(0)}(x),$$

where $f_1^{(0)}$ is defined by backward recursion as follows:

$$f_k^{(l)}(x) = f_{mult}\left(f_{2k-1}^{(l+1)}(x), f_{2k}^{(l+1)}(x)\right)$$

for $k \in \{1, 2, \ldots, 2^l\}$ and $l \in \{0, \ldots, s-1\}$, and

$$f_k^{(s)}(x) = f_{id}(f_{id}(x^{(l)}))$$

for $j_1 + j_2 + \cdots + j_{l-1} + 1 \leq k \leq j_1 + j_2 + \cdots + j_l$ and $l = 1, \ldots, d$,

$$f_{j_1+j_2+\cdots+j_d+1}^{(s)}(x) = f_{hat,y}(\mathbf{b}^T x),$$

and

$$f_k^{(s)}(x) = 1$$

for $k = j_1 + j_2 + \cdots + j_d + 2, j_1 + j_2 + \cdots + j_d + 3, \ldots, 2^s$. *Then we have for any* $x \in [-A, A]^d$*:*

$$\left| f_{net,j_1,\ldots,j_d,y,\mathbf{b}}(x) - (x^{(1)})^{j_1} \cdots (x^{(d)})^{j_d} \cdot (1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}^T x - y|)_+ \right|$$
$$\leq c_{37} \cdot 3^{3 \cdot 3^s} \cdot A^{3 \cdot 2^s} \cdot M^3 \cdot \frac{1}{R}.$$

Proof. We define by backward recursion a function $g_1^{(0)}$ that has the same structure as our neural network. Each node of $g_1^{(0)}$ is assigned the function that the corresponding node in our neural network is supposed to approximate. So,

$$g_1^{(0)}(x) = (x^{(1)})^{j_1} \cdots (x^{(d)})^{j_d} \cdot (1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}^T x - y|)_+.$$

More precisely, we define

$$g_k^{(l)}(x) = g_{2k-1}^{(l+1)}(x) \cdot g_{2k}^{(l+1)}(x)$$

for $k \in \{1, 2, \ldots, 2^l\}$ and $l \in \{0, \ldots, s-1\}$, and

$$g_k^{(s)}(x) = x^{(l)}$$

for $j_1 + j_2 + \cdots + j_{l-1} + 1 \leq k \leq j_1 + j_2 + \cdots + j_l$ and $l = 1, \ldots, d$,

$$g_{j_1+j_2+\cdots+j_d+1}^{(s)}(x) = \left(1 - \frac{M}{2 \cdot \sqrt{d} \cdot A} \cdot |\mathbf{b}^T x - y|\right)_+,$$

and

$$g_k^{(s)}(x) = 1$$

for $k = j_1 + j_2 + \cdots + j_d + 2, j_1 + j_2 + \cdots + j_d + 3, \ldots, 2^s$.
We claim that we have for any $x \in [-A, A]^d$

$$|g_k^{(l)}(x)| \leq A^{2^{s-l}}. \tag{4.13}$$

We show (4.13) by induction on the layers of the network $g_1(0)$.
Start of the induction. For layer s we have for $j_1 + j_2 + \cdots + j_{l-1} + 1 \leq k \leq j_1 + j_2 + \cdots + j_l$ and $l = 1, \ldots, d$

$$|g_k^{(s)}(x)| = |x^{(l)}| \leq A = A^1 = A^{2^0}$$

and trivially

$$|g^{(s)}_{j_1+j_2+\cdots+j_d+1}(x)| = \left(1 - \frac{M}{2\cdot\sqrt{d}\cdot A}\cdot|\mathbf{b}^T x - y|\right)_+ \leq 1 \leq A = A^{2^0}$$

and for $k = j_1 + j_2 + \cdots + j_d + 2, j_1 + j_2 + \cdots + j_d + 3, \ldots, 2^s$

$$g^{(s)}_k(x) = 1 \leq A \leq A^{2^0}.$$

Induction hypothesis. Assume (4.13) holds for the $l-$th layer where $l \in \{0, 1, \ldots, s\}$.
Induction step. We consider the $(l-1)-$th layer. For $k \in \{1, 2, \ldots, 2^{l-1}\}$ and $(l-1) \in \{0, \ldots, s-1\}$ we have by the induction hypothesis

$$g^{(l-1)}_k(x) = g^{(l)}_{2k-1}(x)\cdot g^{(l)}_{2k}(x) \leq A^{2^{s-l}}\cdot A^{2^{s-l}} = A^{2\cdot 2^{s-l}} = A^{2^{s-(l-1)}}.$$

This concludes the proof of the claim (4.13). Next, we show that for $k \in \{1, \ldots, 2^l\}$ and $l \in \{0, \ldots, s\}$ it holds that

$$|f^{(l)}_k(x)| \leq 3^{3^{s-l}}\cdot A^{2^{s-l}}. \tag{4.14}$$

We show (4.14) by induction on the layers of the network $f^{(0)}_1$.
Start of the induction. For layer s we have by (4.12) for $j_1 + j_2 + \cdots + j_{l-1} + 1 \leq k \leq j_1 + j_2 + \cdots + j_l$ and $l = 1, \ldots, d$ we get by Lemma $4.1.5$ and

$$\begin{aligned}
|f^{(s)}_k(x)| &= |f_{id}(f_{id}(x^{(l)}))| \\
&= |f_{id}(f_{id}(x^{(l)})) - f_{id}(x^{(l)}) + f_{id}(x^{(l)}) - x^{(l)} + x^{(l)}| \\
&\leq |f_{id}(f_{id}(x^{(l)})) - f_{id}(x^{(l)})| + |f_{id}(x^{(l)}) - x^{(l)}| + |x^{(l)}| \\
&\leq \frac{\|\sigma''\|_\infty\cdot A^2}{2\cdot|\sigma'(t_{\sigma,id})|}\cdot\frac{1}{R} + \frac{\|\sigma''\|_\infty\cdot A^2}{2\cdot|\sigma'(t_{\sigma,id})|}\cdot\frac{1}{R} + A \\
&\leq \frac{1}{2} + \frac{1}{2} + A \\
&\leq 3\cdot A \\
&= 3^1\cdot A^1 \\
&= 3^{3^0}\cdot A^{2^0}
\end{aligned}$$

and by Lemma $4.1.11$

$$\begin{aligned}
& |f^{(s)}_{j_1+j_2+\cdots+j_d+1}(x)| \\
= \; & |f_{hat,y}(\mathbf{b}^T x)|
\end{aligned}$$

$$
\begin{aligned}
&\leq \left| f_{hat,y}(\mathbf{b}^T x) - \left(1 - \frac{M}{2\cdot\sqrt{d}A}\cdot|\mathbf{b}^T x - y|\right)_+ \right| + \left(1 - \frac{M}{2\cdot\sqrt{d}A}\cdot|\mathbf{b}^T x - y|\right)_+ \\
&\leq 1792 \cdot \frac{\max\{\|\sigma''\|_\infty, \|\sigma'''\|_\infty, 1\}}{\min\{2\cdot|\sigma'(t_{\sigma,id})|, |\sigma''(t_\sigma)|, 1\}} \cdot M^3 \cdot \frac{1}{R} + 1 \\
&\leq 1 + 1 \\
&\leq 2 \cdot A \\
&\leq 3^{3^0} \cdot A^{2^0}
\end{aligned}
$$

and for $k = j_1 + j_2 + \cdots + j_d + 2, j_1 + j_2 + \cdots + j_d + 3, \ldots, 2^s$

$$|f_k^{(s)}(x)| = 1 \leq A \leq 3^{3^0} \cdot A^{2^0}.$$

Induction hypothesis. Assume (4.14) holds for the $l-$th layer where $l \in \{0, 1, \ldots, s\}$.
Induction step. We consider the $(l-1)-$th layer. By the induction hypothesis we know that

$$f_{2k-1}^{(l)}(x), f_{2k}^{(l)}(x) \in [-3^{3^{s-l}} \cdot A^{2^{s-l}}, 3^{3^{s-l}} \cdot A^{2^{s-l}}]$$

and so for $k \in \{1, 2, \ldots, 2^{l-1}\}$ and $(l-1) \in \{0, \ldots, s-1\}$ we have by Lemma 4.1.7 and by (4.12)

$$
\begin{aligned}
\left|f_k^{(l-1)}(x)\right| &= \left|f_{mult}\left(f_{2k-1}^{(l)}(x), f_{2k}^{(l)}(x)\right)\right| \\
&\leq \left|f_{mult}\left(f_{2k-1}^{(l)}(x), f_{2k}^{(l)}(x)\right) - f_{2k-1}^{(l)}(x)\cdot f_{2k}^{(l)}(x)\right| + \left|f_{2k-1}^{(l)}(x)\cdot f_{2k}^{(l)}(x)\right| \\
&\leq \frac{20\cdot\|\sigma'''\|_\infty \cdot 3^{3\cdot 3^{s-l}} \cdot A^{3\cdot 2^{s-l}}}{3\cdot|\sigma''(t_\sigma)|} \cdot \frac{1}{R} + 3^{3^{s-l}} \cdot A^{2^{s-l}} \cdot 3^{3^{s-l}} \cdot A^{2^{s-l}} \\
&\leq 1 + 3^{2\cdot 3^{s-l}} \cdot A^{2\cdot 2^{s-l}} \\
&\leq 3^{3^{s-l}} \cdot 3^{2\cdot 3^{s-l}} \cdot A^{2\cdot 2^{s-l}} \\
&= 3^{3\cdot 3^{s-l}} \cdot A^{2\cdot 2^{s-l}} \\
&= 3^{3^{s-(l-1)}} \cdot A^{2^{s-(l-1)}}.
\end{aligned}
$$

This concludes the proof of the claim (4.13). Now, we prove the assertion of the lemma by showing

$$|f_k^{(l)}(x) - g_k^{(l)}(x)| \leq c_{37} \cdot 3^{3\cdot 3^{s-l}} \cdot A^{3\cdot 2^{s-l}} \cdot M^3 \cdot \frac{1}{R} \tag{4.15}$$

for $k \in \{1, \ldots, 2^l\}$ and $l \in \{0, \ldots, s\}$, where

$$c_{37} = \max\left\{\frac{20\cdot\|\sigma'''\|_\infty}{3\cdot|\sigma''(t_\sigma)|}, \frac{9\cdot\|\sigma''\|_\infty}{|\sigma'(t_{\sigma,id})|}, 1792\cdot\frac{\max\{\|\sigma''\|_\infty, \|\sigma'''\|_\infty, 1\}}{\min\{2\cdot|\sigma'(t_{\sigma.id})|, |\sigma''(t_\sigma)|, 1\}}\right\}$$

by induction on the layers.
Start of the induction. For layer s we have for $j_1+j_2+\cdots+j_{l-1}+1 \leq k \leq j_1+j_2+\cdots+j_l$ and $l=1,\ldots,d$ by Lemma 4.1.5

$$\begin{aligned}
|f_k^{(s)}(x)-g_k^{(s)}(x)| &= |f_{id}(f_{id}(x^{(l)}))-x^{(l)}| \\
&= |f_{id}(f_{id}(x^{(l)}))-f_{id}(x^{(l)})+f_{id}(x^{(t)})-x^{(l)}| \\
&\leq |f_{id}(f_{id}(x^{(l)}))-f_{id}(x^{(l)})|+|f_{id}(x^{(l)})-x^{(l)}| \\
&\leq \frac{9}{2}\cdot\frac{\|\sigma''\|_\infty\cdot A^2}{|\sigma'(t_{\sigma,id})|}\cdot\frac{1}{R}+\frac{9}{2}\cdot\frac{\|\sigma''\|_\infty\cdot A^2}{|\sigma'(t_{\sigma,id})|}\cdot\frac{1}{R} \\
&\leq 2\cdot\frac{1}{2}\cdot c_{37}\cdot A^2\cdot\frac{1}{R} \\
&\leq c_{37}\cdot 3^3\cdot A^3\cdot M^3\cdot\frac{1}{R} \\
&= c_{37}\cdot 3^{3\cdot 3^0}\cdot A^{3\cdot 2^0}\cdot M^3\cdot\frac{1}{R}
\end{aligned}$$

and by Lemma 4.1.11 for $k=j_1+j_2+\cdots+j_d+1$

$$\begin{aligned}
\left|f_k^{(s)}(x)-g_k^{(s)}(x)\right| &= \left|f_{hat,y}\left(\mathbf{b}^Tx\right)-\left(1-\frac{M}{2\cdot\sqrt{d}\cdot A}\cdot|\mathbf{b}^Tx-y|\right)_+\right| \\
&\leq 1792\cdot\frac{\max\{\|\sigma'''\|_\infty,\|\sigma''\|_\infty,1\}}{\min\{|\sigma''(t_\sigma)|,2\cdot|\sigma'(t_{\sigma,id})|,1\}}\cdot M^3\cdot\frac{1}{R} \\
&\leq c_{37}\cdot 3^3\cdot A^3\cdot M^3\cdot\frac{1}{R} \\
&\leq c_{37}\cdot 3^{3\cdot 3^0}\cdot A^{3\cdot 2^0}\cdot M^3\cdot\frac{1}{R}
\end{aligned}$$

and trivially for $k=j_1+j_2+\cdots+j_d+2, j_1+j_2+\cdots+j_d+3,\ldots,2^s$

$$\left|f_k^{(s)}(x)-g_k^{(s)}(x)\right| = |1-1| = 0 \leq c_{37}\cdot 3^{3\cdot 3^0}\cdot A^{3\cdot 2^0}\cdot M^3\cdot\frac{1}{R}.$$

Induction hypothesis. Assume (4.15) holds for the $l-$th layer where $l\in\{0,1,\ldots,s\}$.
Induction step. We consider the $(l-1)-$th layer. By Lemma 4.1.7, by (4.13), (4.14) and by the induction hypothesis we get for $k\in\{1,2,\ldots,2^{l-1}\}$ and $(l-1)\in\{0,\ldots,s-1\}$

$$\begin{aligned}
& |f_k^{(l-1)}(x)-g_k^{(l-1)}(x)| \\
= \; & |f_{mult}\left(f_{2k-1}^{(l)}(x),f_{2k}^{(l)}(x)\right)-g_{2k-1}^{(l)}(x)\cdot g_{2k}^{(l)}(x)|
\end{aligned}$$

$$
\begin{aligned}
&= |f_{mult}\left(f_{2k-1}^{(l)}(x), f_{2k}^{(l)}(x)\right) - f_{2k-1}^{(l)}(x) \cdot f_{2k}^{(l)}(x) + f_{2k-1}^{(l)}(x) \cdot f_{2k}^{(l)}(x) \\
&\quad -g_{2k-1}^{(l)}(x) \cdot f_{2k}^{(l)}(x) + g_{2k-1}^{(l)}(x) \cdot f_{2k}^{(l)}(x) - g_{2k-1}^{(l)}(x) \cdot g_{2k}^{(l)}(x)| \\
&\leq |f_{mult}\left(f_{2k-1}^{(l)}(x), f_{2k}^{(l)}(x)\right) - f_{2k-1}^{(l)}(x), f_{2k}^{(l)}(x)| \\
&\quad +|f_{2k-1}^{(l)}(x) \cdot f_{2k}^{(l)}(x) - g_{2k-1}^{(l)}(x) \cdot f_{2k}^{(l)}(x)| \\
&\quad +|g_{2k-1}^{(l)}(x) \cdot f_{2k}^{(l)}(x) - g_{2k-1}^{(l)}(x) \cdot g_{2k}^{(l)}(x)| \\
&= |f_{mult}\left(f_{2k-1}^{(l)}(x), f_{2k}^{(l)}(x)\right) - f_{2k-1}^{(l)}(x), f_{2k}^{(l)}(x)| + |f_{2k}^{(l)}(x)| \cdot |f_{2k-1}^{(l)}(x) - g_{2k-1}^{(l)}(x)| \\
&\quad +|g_{2k-1}^{(l)}(x)| \cdot |f_{2k}^{(l)}(x) - g_{2k}^{(l)}(x)| \\
&\leq \frac{20 \cdot \|\sigma'''\|_\infty \cdot 3^{3 \cdot 3^{s-l}} \cdot A^{3 \cdot 2^{s-l}}}{3 \cdot |\sigma''(t_\sigma)|} \cdot \frac{1}{R} \\
&\quad +3^{3^{s-1}} \cdot A^{2^{s-1}} \cdot c_{37} \cdot 3^{3 \cdot 3^{s-l}} \cdot A^{3 \cdot 2^{s-l}} \cdot M^3 \cdot \frac{1}{R} \\
&\quad +A^{2^{s-l}} c_{37} \cdot 3^{3 \cdot 3^{s-l}} \cdot A^{3 \cdot 2^{s-l}} \cdot M^3 \cdot \frac{1}{R} \\
&\leq c_{37} \cdot 3^{3 \cdot 3^{s-l}} \cdot A^{3 \cdot 2^{s-l}} \cdot \frac{1}{R} + c_{37} \cdot 3^{4 \cdot 3^{s-l}} \cdot A^{4 \cdot 2^{s-l}} \cdot M^3 \cdot \frac{1}{R} \\
&\quad +c_{37} \cdot 3^{3 \cdot 3^{s-l}} \cdot A^{4 \cdot 2^{s-l}} \cdot M^3 \cdot \frac{1}{R} \\
&\leq 3 \cdot c_{37} \cdot 3^{2 \cdot 3 \cdot 3^{s-l}} \cdot A^{3 \cdot 2 \cdot 2^{s-l}} \cdot M^3 \cdot \frac{1}{R} \\
&\leq c_{37} \cdot 3^{3 \cdot 3^{s-(l-1)}} \cdot A^{3 \cdot 2^{s-(l-1)}} \cdot M^3 \cdot \frac{1}{R}.
\end{aligned}
$$

This concludes the proof. □

Remark 4.1.14. *The approximation results in Lemma* 4.1.13 *hold with*

$$
y = u_{i_k} \in [-\sqrt{d} \cdot A, \sqrt{d} \cdot A]
$$

for the neural network $f_{net,j_1,\ldots,j_d,i_k,\mathbf{b}_l}$ *defined in* (4.7) − (4.11) *by Remark* 4.1.5*, Remark* 4.1.11 *and Remark* 4.1.7.

4.1.4. Definition of the Output Weights

Now, we choose the output weights for given directions $\mathbf{b}_l$ $(l = 1, \ldots, r)$. By construction of our neural network estimate in Section 4.1.3 the network is contained in the class of

functions

$$\left\{ \sum_{l=1}^{r} \sum_{k=1}^{M+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \le N}} a_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \cdot f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l}(x) : \quad a_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \in \mathbb{R} \right\}.$$

For given directions $\mathbf{b}_l$ $(l = 1, \dots, r)$ we define our neural network estimate $\tilde{m}_n(x)$ by

$$\tilde{m}_n(x) = \sum_{l=1,\dots,r} \sum_{k=1}^{M+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \le N}} a_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \cdot f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l}(x),$$

where the coefficients $a_{i_k,j_1,\dots,j_d,\mathbf{b}_l}$ are chosen by minimizing

$$\frac{1}{n} \sum_{i=1}^{n} |Y_i - \tilde{m}_n(X_i)|^2 + \frac{c_3}{n} \cdot \sum_{l=1}^{r} \sum_{k=1}^{M+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \le N}} a^2_{i_k,j_1,\dots,j_d,\mathbf{b}_l} \tag{4.16}$$

for some constant $c_3 > 0$. This regularized linear least squares estimate can be computed by solving a linear equation system. To see this, set

$$J = r \cdot (M+1) \cdot \binom{N+d}{d},$$

let

$$\begin{aligned} &\{B_j \,:\, j = 1, \dots, J\} \\ &= \{f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l} \,:\, 1 \le l \le r,\ 1 \le k \le M+1 \text{ and } 0 \le j_1 + \dots + j_d \le N\} \end{aligned}$$

and set

$$\mathbf{B} = (B_j(X_i))_{1 \le i \le n, 1 \le j \le J} \quad \text{and} \quad \mathbf{Y} = (Y_i)_{i=1,\dots,n}.$$

Then

$$\tilde{m}_n(x) = \sum_{k=1}^{J} a_k \cdot B_k(x).$$

Additionally, set

$$\mathbf{A} := \frac{1}{n} \mathbf{B}^T \mathbf{B} + \frac{c_3}{n} \cdot \mathbf{1}.$$

Clearly, the matrix $\mathbf{A}$ is symmetric and positive definite and hence regular. As a consequence, we can write (4.16) as

$$\begin{aligned}
& \frac{1}{n}(\mathbf{Y}-\mathbf{B}\mathbf{a})^T(\mathbf{Y}-\mathbf{B}\mathbf{a}) + \frac{c_3}{n}\mathbf{a}^T\mathbf{a} \\
= \; & \frac{1}{n}(\mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{B}\mathbf{a}) + \mathbf{a}^T\left(\frac{1}{n}\mathbf{B}^T\mathbf{B} + \frac{c_3}{n}\cdot\mathbf{1}\right)\mathbf{a} \\
= \; & \frac{1}{n}(\mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{B}\mathbf{a}) + \mathbf{a}^T\mathbf{A}\mathbf{a} \\
= \; & \left(\mathbf{a} - \frac{1}{n}\mathbf{A}^{-1}\mathbf{B}^T\mathbf{Y}\right)^T \mathbf{A}\left(\mathbf{a} - \frac{1}{n}\mathbf{A}^{-1}\mathbf{B}^T\mathbf{Y}\right) + \frac{1}{n}\mathbf{Y}^T\mathbf{Y} - \frac{1}{n^2}\mathbf{Y}^T\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T\mathbf{Y}.
\end{aligned}$$

We observe that the last two terms are independent of $\mathbf{a}$ and hence the right-hand side is minimal for

$$\mathbf{a} - \frac{1}{n}\mathbf{A}^{-1}\mathbf{B}^T\mathbf{Y} = 0 \quad \Leftrightarrow \quad \mathbf{A}\mathbf{a} = \frac{1}{n}\mathbf{B}^T\mathbf{Y}$$

So, the weight vector of our neural network estimate is the unique solution of the linear equation system

$$\left(\frac{1}{n}\mathbf{B}^T\mathbf{B} + \frac{c_3}{n}\cdot\mathbf{1}\right)\mathbf{a} = \frac{1}{n}\mathbf{B}^T\mathbf{Y}. \tag{4.17}$$

Since $\mathbf{a}$ minimizes (4.16), obviously, the value of this minimization problem will be also less than or equal to the value we get when we set all coefficients to zero. Hence, we have

$$\frac{1}{n}(\mathbf{Y}-\mathbf{B}\mathbf{a})^T(\mathbf{Y}-\mathbf{B}\mathbf{a}) + \frac{c_3}{n}\cdot\mathbf{a}^T\mathbf{a} \leq \frac{1}{n}(\mathbf{Y}-\mathbf{B}\mathbf{0})^T(\mathbf{Y}-\mathbf{B}\mathbf{0}) + \frac{c_3}{n}\cdot\mathbf{0}^T\mathbf{0} = \frac{1}{n}\mathbf{Y}^T\mathbf{Y}$$

and so

$$\mathbf{a}^T\mathbf{a} \leq \frac{1}{n}\mathbf{Y}^T\mathbf{Y}\cdot\frac{n}{c_3} \tag{4.18}$$

which will allow us to derive a bound on the maximal absolute value of our coefficients.

4.1.5. Choosing the Directions

In essence, we choose the projection directions by repeated random initialization of direction vectors. We then pick the neural network estimate with the smallest error. More precisely, the algorithm works as follows:

1. Randomly choose values
$$\mathbf{b}_1^*, \ldots, \mathbf{b}_r^* \in [-1,1]^d$$

as an independent sample from a uniform distribution on $[-1,1]^d$ and set

$$\mathbf{b}_l = \frac{\mathbf{b}_l^*}{\|\mathbf{b}_l^*\|} \quad (l = 1, \dots, r).$$

These values approximate the direction of projection $\mathbf{b}_1, \dots, \mathbf{b}_r$ in our projection pursuit model.

Note, that for $l = 1, \dots, r$

$$\mathbf{P}\{\mathbf{b}_l^* = 0\} = 0,$$

which means we can assume w.l.o.g. $\|\mathbf{b}_l^*\| \neq 0$.

2. For these vectors $\mathbf{b}_l$ $(l = 1, \dots, r)$ construct a neural network estimate as described in Section 4.1.3 and choose the outer weights as in Section 4.1.4.

Repeat steps 1. and 2. I_n times. A choice of I_n is given in Theorem 4.2.1.

3. Choose the directions and the corresponding network which achieves the smallest penalized empirical L_2 error (4.16) among all the I_n networks as our neural network estimate $\tilde{m}_n$.

4.2. Rate of Convergence

Theorem 4.2.1 states that our neural network regression estimate constructed in Section 4.1 achieves the univariate rate of convergence up to a logarithmic factor.

Theorem 4.2.1. *Assume that the distribution of (X, Y) satisfies*

$$\mathbf{E}\left(e^{c_4 \cdot |Y|^2}\right) < \infty \tag{4.19}$$

for some constant $c_4 > 0$ and that the distribution of X has bounded support $supp(X)$, and let $m(x) = \mathbf{E}\{Y|X = x\}$ be the corresponding regression function. Let $r \in \mathbb{N}$, $p > 0$ and $C > 0$, and assume that the regression function satisfies

$$m(x) = \sum_{l=1}^{r} g_l(\mathbf{a}_l^T x) \quad (x \in \mathbb{R}^d)$$

for some (p, C)–smooth functions $g_l : \mathbb{R} \to \mathbb{R}$ and some $\mathbf{a}_l \in \mathbb{R}^d$ with $\|\mathbf{a}_l\| = 1$ $(l = 1, \dots, r)$.

Define the estimate $\tilde{m}_n$ as in Section 4.1, where σ is the logistic squasher

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

and where the parameters are chosen as

$$I_n = \left\lceil \left(\frac{n}{(\log n)^2} \right)^{\frac{r\cdot(d-1)}{2p+1}} \right\rceil,$$

$$N \geq p, \quad M = M_n = \left\lceil c_{10} \cdot n^{\frac{1}{2p+1}} \right\rceil, \quad R = R_n = n^{\frac{9}{2}}$$

and

$$A = A_n = (\log n)^{\frac{1}{6(N+d)}}.$$

Set $\beta_n = c_6 \cdot \log(n)$ for some suitably large constant $c_6 > 0$ and define m_n by

$$m_n(x) = T_{\beta_n} \tilde{m}_n(x).$$

Then m_n satisfies for n suffciently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_{11} \cdot (\log n)^3 \cdot n^{-\frac{2p}{2p+1}},$$

where $c_{11} > 0$ does not depend on n.

Remark 4.2.2. *The rate of convergence in Theorem 4.2.1 is stated specifically for the logistic squasher. Nonetheless, the statement is equally true for all squashing functions that are Lipschitz continuous and $2-$admissible according to Definition 4.1.3 with f_{id} as in Lemma 4.1.5, f_{mult} as in Lemma 4.1.7, and f_{ReLU} as in Lemma 4.1.9.*

Remark 4.2.3. *Since the rate of convergence presented in Theorem 4.2.1 is independent of the dimension d of X, this means that our proposed neural network regression estimate for a regression function that satisfies the regression pursuit model is able to circumvent the curse of dimensionality. However, we can see that the dependence on the dimenion d has shifted into the necessary number of repitions I_n of the initial random choices of the directions $\bar{\mathbf{b}}_l$. Concerning the computation of our estimate, we need to solve a linear equation system with a quadratic $M_n \times M_n$ matrix I_n times. The computation time for this is proportional to*

$$I_n \cdot M_n^2 \approx n^{\frac{r\cdot(d-1)+2}{2p+1}}.$$

Hence in case

$$r \cdot (d-1) < 4 \cdot p$$

the computation time is $O(n^2)$. This means that if the number r of terms in the projection pursuit model and the dimension d of X are not too large our estimate can be computed in $O(n^2)$ time.

Remark 4.2.4. *The parameters N, r and M_n, and also I_n of the above algorithm depend on the projection pursuit model and are hence unknown in any application. Using the splitting of the sample technique as explained in Section 4.3 it is possible to choose these parameters data-dependently.*

4.2.1. Auxiliary Lemmas from Empirical Process Theory

For the proof of Theorem 4.2.1 we will need the following auxiliary results from empirical process theory.

Lemma 4.2.5. *Let*

- $\beta_n = c_6 \cdot \log(n)$ *for some suitably large constant* $c_6 > 0$*,*
- $\mathcal{F}_n$ *be a set of functions* $f : \mathbb{R}^d \to \mathbb{R}$*.*

Assume

- *the distribution of* (X, Y) *satisfies* (4.19) *for some constant* $c_4 > 0$*,*
- *the regression function* m *is bounded in absolute value,*
- *the estimate* m_n *satisfies*

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

  with

  $$\tilde{m}_n(\cdot) = \tilde{m}_n(\cdot, (X_1, Y_1), \ldots, (X_n, Y_n)) \in \mathcal{F}_n, \tag{4.20}$$

- *the estimate* m_n *satisfies*

  $$\frac{1}{n} \sum_{i=1}^{n} |Y_i - \tilde{m}_n(X_i)|^2 \leq \min_{l \in \Theta_n} \left(\frac{1}{n} \sum_{i=1}^{n} |Y_i - g_{n,l}(X_i)|^2 + pen_n(g_{n,l}) \right) \tag{4.21}$$

  for

  - *some nonempty set* Θ_n *of parameters,*

– *some random functions* $g_{n,l} : \mathbb{R}^d \to \mathbb{R}$*, that only depend on the set*

$$\mathcal{B}_{n,r} = \{\mathbf{b}_1^{(1)}, \dots, \mathbf{b}_r^{(1)}, \dots, \mathbf{b}_1^{(I_n)}, \dots, \mathbf{b}_r^{(I_n)}\}$$

where

$$\mathbf{b}_1^{(1)}, \dots, \mathbf{b}_r^{(1)}, \dots, \mathbf{b}_1^{(I_n)}, \dots, \mathbf{b}_r^{(I_n)}$$

are random variables independent of

$$(X_1, Y_1), \dots, (X_n, Y_n),$$

– *some deterministic penalty terms* $pen_n(g_{n,l}) \geq 0$.

Then m_n *satisfies*

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq \frac{c_{13} \cdot (\log n)^2 \cdot \left(\log\left(\sup_{x_1^n \in (supp(X))^n} \mathcal{N}_1\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, x_1^n\right)\right) + 1\right)}{n} \\
&\quad + 2 \cdot \mathbf{E}\left(\min_{l \in \Theta_n} \int |g_{n,l}(x) - m(x)|^2 \mathbf{P}_X(dx) + pen_n(g_{n,l})\right)
\end{aligned}$$

for $n > 1$ *and some constant* $c_{13} > 0$*, which does not depend on* n*.*

Proof. Compared to Lemma 2.2.14, we see that (4.20) is assumed independently from any event A_n and (4.21) is an additional bound.
As in the proof of Lemma 3.2.11 we apply Lemma 2.2.14 where the event A_n is the underlying set of our probability space. This gives us

$$\begin{aligned}
&\mathbf{E}\left(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)\right) \\
\leq\ & \frac{c_7 \cdot (\log n)^2 \cdot \left(\log\left(\sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, x_1^n\right)\right) + 1\right)}{n} \\
&+ 2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{\{|Y_j| \leq \beta_n\ (j \in \{1,\dots,n\})\}} - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right).
\end{aligned}$$

We use (4.21) and, since X_i is independent of $\mathbf{b}_1^{(1)}, \dots, \mathbf{b}_r^{(1)}, \dots, \mathbf{b}_1^{(I_n)}, \dots, \mathbf{b}_r^{(I_n)}$ for $i = 1, \dots, n$, we bound the expected value on the right-hand side further by

$$2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{\{|Y_j| \leq \beta_n\ (j \in \{1,\dots,n\})\}} - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right)$$

$$
\begin{aligned}
&\leq 2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}|\tilde{m}_n(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\right)\\
&\leq 2 \cdot \mathbf{E}\left(\min_{l\in\Theta_n}\frac{1}{n}\sum_{i=1}^{n}|g_{n,l}(X_i)-Y_i|^2+pen_n(g_{n,l})-\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\right)\\
&\leq 2 \cdot \mathbf{E}\left(\min_{l\in\Theta_n}\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}|g_{n,l}(X_i)-Y_i|^2+pen_n(g_{n,l})-\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\middle|\mathcal{B}_{n,r}\right)\right)\\
&= 2 \cdot \mathbf{E}\left(\min_{l\in\Theta_n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(|g_{n,l}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\middle|\mathcal{B}_{n,r}\right)+pen_n(g_{n,l})\right)\\
&= 2 \cdot \mathbf{E}\left(\min_{l\in\Theta_n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(|g_{n,l}(X_i)-m(X_i)|^2\middle|\mathcal{B}_{n,r}\right)+pen_n(g_{n,l})\right)\\
&= 2 \cdot \mathbf{E}\left(\min_{l\in\Theta_n}\frac{1}{n}\sum_{i=1}^{n}\int|g_{n,l}(x)-m(x)|^2\mathbf{P}_{X_i}(dx)+pen_n(g_{n,l})\right)\\
&= 2 \cdot \mathbf{E}\left(\min_{l\in\Theta_n}\int|g_{n,l}(x)-m(x)|^2\mathbf{P}_{X}(dx)+pen_n(g_{n,l})\right)
\end{aligned}
$$

where the second equality holds since

$$
\begin{aligned}
&\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(g_{n,l}(X_i)-m(X_i))\cdot(m(X_i)-Y_i)\mid\mathcal{B}_{n,r}\right)\\
&= \mathbf{E}\left(\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(g_{n,l}(X_i)-m(X_i))\cdot(m(X_i)-Y_i)\mid\{X_1,\ldots,X_n\}\cup\mathcal{B}_{n,r}\right)\mid\mathcal{B}_{n,r}\right)\\
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(g_{n,l}(X_i)-m(X_i))\cdot\mathbf{E}\left((m(X_i)-Y_i)\mid\{X_1,\ldots,X_n\}\cup\mathcal{B}_{n,r}\right)\mid\mathcal{B}_{n,r}\right)\\
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(g_{n,l}(X_i)-m(X_i))\cdot\mathbf{E}\left((m(X_i)-Y_i)\mid X_i\right)\mid\mathcal{B}_{n,r}\right)\\
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(g_{n,l}(X_i)-m(X_i))\cdot(m(X_i)-\mathbf{E}\left(Y_i\mid X_i\right))\mid\mathcal{B}_{n,r}\right)\\
&= 0.
\end{aligned}
$$

This concludes the proof. □

In order to bound the covering number $\mathcal{N}_1\left(\frac{1}{n\cdot\beta_n},\mathcal{F}_n,x_1^n\right)$ we will use the following lemma.

Lemma 4.2.6. *Let $a > 0$ and let $d, N, J_n \in \mathbb{N}$ be such that $J_n \leq n^{c_{14}}$ and set $\beta_n = c_6 \cdot \log n$. Let σ be 2–admissible according to Definition 4.1.3. Let $\mathcal{F}$ be the set of all functions defined by (1.2), (1.3) and (1.4) where $k_1 = k_2 = \cdots = k_L = 24 \cdot (N + d)$ and the weights are bounded in absolute value by $c_{15} \cdot n^{c_{16}}$. Set*

$$\mathcal{F}^{(J_n)} = \left\{ \sum_{j=1}^{J_n} a_j \cdot f_j \ : \ f_j \in \mathcal{F} \quad \text{and} \quad \sum_{j=1}^{J_n} a_j^2 \leq c_{17} \cdot n^{c_{18}} \right\}.$$

Then we have for $n > 1$

$$\log \left(supp_{x_1^n \in [-a,a]^{d \cdot n}} \mathcal{N}_1 \left(\frac{1}{n \cdot \beta_n}, \mathcal{F}^{(J_n)}, x_1^n \right) \right) \leq c_{19} \cdot \log n \cdot J_n$$

for some constant c_{19} which depends only on L, N, a and d.

Proof. The proof of this lemma has been thoroughly developed within the work of the author's Master Thesis. For that reason only a short sketch of the proof is given here. Since the networks in $\mathcal{F}^{(J_n)}$ are linear combinations of J_n fully connected neural networks with L hidden layers, a bounded number of neurons in each hidden layers and all weights bounded by a polynomial in n, the result follows by combining Lemma 16.6 in Györfi et al. (2002) with Lemma 7 in the Supplement of Bauer et al. (2019). □

4.2.2. Proof of Theorem 4.2.1

We give a proof for the rate of convergence presented in Theorem 4.2.1.

Proof. Since by assumption $supp(X)$ is bounded and m is $(p, C)-$ smooth we can assume w.l.o.g. that m is bounded in absolute value. So, we assume

$$\|m\|_\infty \leq \beta_n.$$

Let B_n be the event where

$$|Y_i| \leq \sqrt{n} \quad (i = 1, \ldots, n).$$

We get

$$\begin{aligned} & \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ = \ & \mathbf{E} \left(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot (\mathbf{1}_{B_n} + \mathbf{1}_{B_n^c}) \right) \end{aligned}$$

$$
\begin{aligned}
&\leq \mathbf{E}\left(\int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx)\cdot \mathbf{1}_{B_n}\right) + \mathbf{E}\left(\int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx)\cdot \mathbf{1}_{B_n^c}\right) \\
&\leq \mathbf{E}\left(\int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx)\cdot \mathbf{1}_{B_n}\right) \\
&\quad + \mathbf{E}\left(\int 2\cdot |m_n(x)|^2 + 2\cdot |m(x)|^2 \mathbf{P}_X(dx)\cdot \mathbf{1}_{B_n^c}\right) \\
&\leq \mathbf{E}\left(\int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx)\cdot \mathbf{1}_{B_n}\right) + \mathbf{E}\left(\int 2\cdot \beta_n^2 + 2\cdot \beta_n^2 \mathbf{P}_X(dx)\cdot \mathbf{1}_{B_n^c}\right) \\
&= \mathbf{E}\left(\int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx)\cdot \mathbf{1}_{B_n}\right) + \quad 4\cdot \beta_n^2 \cdot \mathbf{P}\{B_n^c\}.
\end{aligned}
$$

We bound the summands on the right-hand side separately. We start with the second term. Since we have by Markov's inequality and by (4.19)

$$
\begin{aligned}
\mathbf{P}\{B_n^c\} &= \mathbf{P}\left\{|Y_i| > \sqrt{n} \text{ for some } i \in \{1,\dots,n\}\right\} \\
&\leq n\cdot \mathbf{P}\left\{\exp(c_4\cdot Y^2) > \exp(c_4\cdot(\sqrt{n})^2)\right\} \\
&\leq n\cdot \frac{\mathbf{E}\left(\exp(c_4\cdot Y^2)\right)}{\exp\left(c_4\cdot n\right)} \\
&\leq n\cdot \frac{c_{21}}{\exp\left(c_4\cdot n\right)} \\
&\leq c_{23}\cdot \frac{1}{n^2},
\end{aligned}
$$

we can conclude

$$
4\cdot \beta_n^2 \cdot \mathbf{P}\{B_n^c\} \leq 4\cdot (c_6\cdot \log n)^2 \cdot c_{23}\cdot \frac{1}{n^2} \leq c_{11}\cdot \frac{(\log n)^2}{n^2}. \tag{4.22}
$$

Next, we bound the first summand further by Lemma 4.2.5. For that we define for given directions $\mathbf{b}_l$ $(l=1,\dots,r)$ a neural network $\hat{m}_n$ by

$$
\hat{m}_n(x) = \sum_{l=1,\dots,r} \sum_{k=1}^{M_n+1} \sum_{\substack{j_1,\dots,j_d\in\{0,\dots,N\} \\ j_1+\dots+j_d\leq N}} a_{i_k,j_1,\dots,j_d,\mathbf{b}_l}\cdot f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l}(x),
$$

where the coefficients $a_{k,j_1,\ldots,j_d,\mathbf{b}_l}$ are chosen from the set

$$\left\{ (a_{k,j_1,\ldots,j_d,\mathbf{b}_l})_{k,j_1,\ldots,j_d,l} \; : \; \sum_{k,j_1,\ldots,j_d,l} a^2_{k,j_1,\ldots,j_d,\mathbf{b}_l} \leq c_{27} \cdot n^2 \right\}$$

with

$$c_{27} = \max \left\{ \frac{1}{c_3}, 2 \cdot r \cdot \binom{N+d}{d} \cdot (q+1) \cdot 2^p \cdot d^{3p/2} \right\}$$

by minimizing

$$\frac{1}{n} \sum_{i=1}^n |Y_i - \hat{m}_n(X_i)|^2 + \frac{c_3}{n} \cdot \sum_{l=1}^r \sum_{k=0}^K \sum_{\substack{j_1,\ldots,j_d \in \{0,\ldots,N\} \\ j_1+\cdots+j_d \leq N}} a^2_{k,j_1,\ldots,j_d,\mathbf{b}_l}$$

for some constant $c_3 > 0$ and we let

$$\bar{m}_n = T_{\beta_n} \hat{m}_n$$

be the by β_n truncated version of $\hat{m}_n$. By (4.18) we know that the output weights of our neural network estimate $\tilde{m}_n$ satisfy on B_n

$$\begin{aligned} \sum_{k,j_1,\ldots,j_d,l} a^2_{k,j_1,\ldots,j_d,l} &\leq \frac{1}{n} \sum_{i=1}^n Y_i^2 \cdot \frac{n}{c_3} \\ &\leq \frac{1}{n} \cdot n \cdot (\sqrt{n})^2 \cdot \frac{n}{c_3} \\ &\leq c_{27} \cdot n^2. \end{aligned}$$

Thus, $\bar{m}_n$ coincides with m_n on the event B_n and we can write

$$\bar{m}_n = \begin{cases} m_n = T_{\beta_n} \tilde{m}_n & \text{if } B_n \\ T_{\beta_n} \hat{m}_n & \text{if } B_n^c. \end{cases}$$

So, we have

$$\begin{aligned} \mathbf{E}\left(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot \mathbf{1}_{B_n} \right) &= \mathbf{E}\left(\int |\bar{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot \mathbf{1}_{B_n} \right) \\ &\leq \mathbf{E}\left(\int |\bar{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \right). \end{aligned}$$

We apply Lemma 4.2.5 to the right-hand side of the above inequality. In order to apply Lemma 4.2.5 we need to check the conditions (4.20) and (4.21) for $\hat{m}_n$. By the definitions above, it suffices to consider $\tilde{m}_n$ under the additional assumption that

$$\sum_{k,j_1,\dots,j_d,l} a^2_{k,j_1,\dots,j_d,l} \leq c_{27} \cdot n^2. \tag{4.23}$$

Let $\mathcal{F}$ be the set of neural networks defined by (1.2), (1.3) and (1.4) with $s+2$ layers, i.e.

$$L = s + 2 = \lceil \log_2(N+d) \rceil + 2,$$

with $24 \cdot (N+d)$ nodes in each layer, i.e.

$$k_1 = k_2 = \dots = k_L = 24 \cdot (N+d)$$

and where the weights are bounded in absolute value by $n^{c_{20}}$. Let

$$J_n = r \cdot (M_n + 1) \cdot |\{(j_1, \dots, j_d) \,:\, j_1, \dots, j_d \in \{0, \dots, N\},\, j_1 + \dots + j_d \leq N\}|.$$

Then

$$J_n = r \cdot (M_n + 1) \binom{N+d}{d}.$$

Set

$$\mathcal{F}^{(J_n)} = \left\{ \sum_{j=1}^{J_n} a_j \cdot f_j \,:\, f_j \in \mathcal{F} \quad \text{and} \quad \sum_{j=1}^{J_n} a_j^2 \leq c_{27} \cdot n^2 \right\}$$

the function space defined in Lemma 4.2.6 with

$$c_{27} = \max\left\{ \frac{1}{c_3}, 2 \cdot r \cdot \binom{N+d}{d} \cdot (q+1) \cdot 2^p \cdot d^{\frac{3p}{2}} \right\}.$$

By construction and by 4.23, we immediately know that

$$\tilde{m}_n(x) \in \mathcal{F}^{(J_n)}$$

which verifies (4.20).
By Lemma 4.1.2 we know that for each $i \in \{1, \dots, I_n\}$ there exist coefficients

$$a^{(i)}_{k,j_1,\dots,j_d,l} \in [-c_{28} \cdot A_n^p, c_{28} \cdot A_n^p]$$

which depend on $\mathbf{a}_l$ and on $\mathbf{b}_l^{(i)}$, but which are independent of (X_1, Y_1), ..., (X_n, Y_n), such that we have for all $x \in [-A_n, A_n]^d$

$$\begin{aligned}
&\Bigg| \sum_{l=1}^{r} \sum_{j=0}^{q} \frac{g_l^{(j)}((\mathbf{b}_l^{(i)})^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l^{(i)})^T x)^j \\
&\qquad - \sum_{l=1}^{r} \sum_{k=1}^{M_n+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \le N}} a_{i_k,j_1,\dots,j_d,l}^{(i)} \cdot (x^{(1)})^{j_1} \cdots (x^{(d)})^{j_d} \\
&\qquad \cdot \left(1 - \frac{M_n}{2 \cdot \sqrt{d} \cdot A_n} \cdot |(\mathbf{b}_l^{(i)})^T x - u_{i_k}|\right)_+ \Bigg|^2 \\
\le\ & \Bigg(r \cdot 2^p \cdot (p+1) \cdot C \cdot d^{\frac{3p}{2}} \cdot A_n^{2p} \\
&\qquad \cdot \left(\max\left\{ \frac{1}{M_n}, \max_{l=1,\dots,r} \|\mathbf{a}_l - \mathbf{b}_l^{(i)}\|_\infty \right\} \right)^p \Bigg)^2 . \qquad (4.24)
\end{aligned}$$

From the definition of our estimate as the solution of a minimization problem we can conclude that for these coefficients it holds that

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} |Y_i - \tilde{m}_n(X_i)|^2 \\
\le\ & \min_{t=1,\dots,I_n} \Bigg\{ \frac{1}{n} \sum_{i=1}^{n} |Y_i - \sum_{l=1}^{r} \sum_{k=1}^{M_n+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \le N}} a_{i_k,j_1,\dots,j_d,l}^{(t)} \cdot f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l^{(t)}}(X_i)|^2 \\
&\qquad + \frac{c_3}{n} \cdot \sum_{l=1}^{r} \sum_{k=1}^{M_n+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \le N}} (a_{i_k,j_1,\dots,j_d,l}^{(t)})^2 \Bigg\} \\
\le\ & \min_{t=1,\dots,I_n} \Bigg\{ \frac{1}{n} \sum_{i=1}^{n} |Y_i - \sum_{l=1}^{r} \sum_{k=1}^{M_n+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \le N}} a_{i_k,j_1,\dots,j_d,l}^{(t)} \cdot f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l^{(t)}}(X_i)|^2 \\
&\qquad + c_{29} \cdot \frac{1}{n} \cdot r \cdot \binom{N+d}{d} \cdot M_n \cdot A_n^{2p} \Bigg\}
\end{aligned}$$

which verfies (4.21). Now, we can bound the first summand by Lemma 4.2.5 and Lemma

4.2.6 together with

$$A_n^{2p} \cdot \frac{M_n}{n} \leq (\log n)^{\frac{2p}{6(N+d)}} \cdot 2 \cdot c_{10} \cdot \frac{n^{\frac{1}{2p+1}}}{n} \leq 2 \cdot c_{10} \cdot \frac{(\log n)^{\frac{2p}{6p}}}{n^{\frac{2p}{2p+1}}} \leq 2 \cdot c_{10} \cdot \frac{\log n}{n^{\frac{2p}{2p+1}}}$$

which yields

$$\begin{aligned}
& \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
\leq \quad & c_{30} \cdot \frac{(\log n)^3 \cdot M_n}{n} \\
& +2 \cdot \mathbf{E} \Bigg(\min_{t=1,\dots,I_n} \int \Big| \sum_{l=1}^{r} \sum_{k=1}^{M_n+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \leq N}} a_{i_k,j_1,\dots,j_d,l}^{(t)} \cdot f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l^{(t)}}(x) \\
& \qquad\qquad -m(x) \Big|^2 \mathbf{P}_X(dx) \Bigg) \\
& +c_{31} \cdot (\log n) \cdot n^{-\frac{2p}{2p+1}}.
\end{aligned}$$

Now, we look at

$$\int \Big| \sum_{l=1}^{r} \sum_{k=1}^{M_n+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \leq N}} a_{i_k,j_1,\dots,j_d,l}^{(t)} \cdot f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l^{(t)}}(x) - m(x) \Big|^2 \mathbf{P}_X(dx).$$

Since $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$ $(a, b, c \in \mathbb{R})$ we have

$$\begin{aligned}
& \int |\sum_{l=1}^{r} \sum_{k=1}^{M_n+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \leq N}} a_{i_k,j_1,\dots,j_d,l}^{(t)} \cdot f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l^{(t)}}(x) - m(x)|^2 \mathbf{P}_X(dx) \\
\leq \quad & 3 \cdot \int \Bigg| \sum_{l=1}^{r} \sum_{k=1}^{M_n+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \leq N}} a_{i_k,j_1,\dots,j_d,l}^{(t)} \cdot f_{net,j_1,\dots,j_d,i_k,\mathbf{b}_l^{(t)}}(x) \\
& \qquad - \sum_{l=1}^{r} \sum_{k=1}^{M_n+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \leq N}} a_{i_k,j_1,\dots,j_d,l}^{(t)} \cdot (x^{(1)})^{j_1} \cdots (x^{(d)})^{j_d}
\end{aligned}$$

$$\cdot \left(1 - \frac{M_n}{2 \cdot \sqrt{d} \cdot A_n} \cdot |(\mathbf{b}_l^{(t)})^T x - u_{i_k}|\right)_+ \Bigg|^2 \mathbf{P}_X(dx)$$

$$+3 \cdot \int \Bigg| \sum_{l=1}^{r} \sum_{k=1}^{M_n+1} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \le N}} a_{i_k,j_1,\dots,j_d,l}^{(t)} \cdot (x^{(1)})^{j_1} \cdots (x^{(d)})^{j_d}$$

$$\cdot \left(1 - \frac{M_n}{2 \cdot \sqrt{d} \cdot A_n} \cdot |(\mathbf{b}_l^{(t)})^T x - u_{i_k}|\right)_+$$

$$- \sum_{l=1}^{r} \sum_{j=0}^{q} \frac{g_l^{(j)}((\mathbf{b}_l^{(t)})^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l^{(t)})^T x)^j \Bigg|^2 \mathbf{P}_X(dx)$$

$$+3 \cdot \int \Bigg| \sum_{l=1}^{r} \sum_{k=1}^{M_n+1} \frac{g_l^{(j)}(\mathbf{b}_l^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l^{(t)})^T x)^j - m(x) \Bigg|^2 \mathbf{P}_X(dx).$$

We bound the three summands on the right-hand side separately. Since

$$\max\left\{ n^{\frac{1}{2p+1}}, (\log n)^{\frac{2}{6(N+d)}}, (\log n)^{\frac{3(N+d)}{6(N+d)}}, n^{\frac{3}{2p+1}} \right\} \le n^{\frac{9}{2}} = R_n$$

holds, (4.12) is satisfied and for the first summand we have by Lemma 4.1.13 for all $x \in [-A_n, A_n]^d$

$$\Bigg| \sum_{l=1}^{r} \sum_{k=0}^{M_n} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \le N}} a_{k,j_1,\dots,j_d,l}^{(t)} \cdot f_{net,k,j_1,\dots,j_d,\mathbf{b}_l^{(t)}}(x)$$

$$- \sum_{l=1}^{r} \sum_{j=0}^{q} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \le N}} a_{k,j_1,\dots,j_d,l}^{(t)} \cdot (x^{(1)})^{j_1} \cdots (x^{(d)})^{j_d}$$

$$\cdot \left(1 - \frac{M_n}{2 \cdot \sqrt{d} \cdot A} \cdot |(\mathbf{b}_l^{(t)})^T x - u_k|\right)_+ \Bigg|^2$$

$$\le \Bigg(\sum_{l=1}^{r} \sum_{k=0}^{M_n} \sum_{\substack{j_1,\dots,j_d \in \{0,\dots,N\} \\ j_1+\dots+j_d \le N}} |a_{k,j_1,\dots,j_d,l}^{(t)}|$$

$$\cdot |f_{net,k,j_1,\dots,j_d,\mathbf{b}_l^{(t)}}(x) - (x^{(1)})^{j_1} \cdots (x^{(d)})^{j_d} \cdot \left(1 - \frac{M_n}{2 \cdot \sqrt{d} \cdot A} \cdot |(\mathbf{b}_l^{(t)})^T x - u_k|\right)_+ | \Bigg)^2$$

$$
\begin{aligned}
&\leq r^2 \cdot (M_n+1)^2 \cdot (N+d)^{2d} \cdot c_{28}^2 \cdot A_n^{2p} \cdot c_{37}^2 \cdot 3^{6\cdot 3^s} \cdot (\sqrt{d}\cdot A_n)^{6\cdot 2^s} \cdot M_n^6 \cdot \frac{1}{R_n^2} \\
&\leq c_{28} \cdot (\log n)^{\frac{2p}{6(N+d)}} \cdot (\log n)^{\frac{6(N+d)}{6(N+d)}} \cdot 2 \cdot c_{10} \cdot n^{\frac{8}{2p+1}} \cdot \frac{1}{n^9} \\
&\leq c_{32} \cdot \frac{(\log n)^2}{n}.
\end{aligned}
$$

The second summand is bounded by (4.24).
For the third summand application of Lemma 4.1.1 gives us for all $x \in [-A_n, A_n]^d$

$$
|\sum_{l=1}^{r}\sum_{j=0}^{q} \frac{g_l^{(j)}((\mathbf{b}_l^{(t)})^T x)}{j!} \cdot ((\mathbf{a}_l - \mathbf{b}_l)^T x)^j - m(x)|^2 \leq c_{33} \cdot A_n^{2p} \cdot \max_{l=1,\ldots,r} \|\mathbf{a}_l - \mathbf{b}_l^{(t)}\|_\infty^{2p}.
$$

Hence,

$$
\begin{aligned}
&\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
\leq\ & c_{30} \cdot \frac{(\log n)^3 \cdot M_n}{n} \\
&+2 \cdot \mathbf{E}\Bigg(\min_{t=1,\ldots,I_n} c_{32} \cdot \frac{(\log n)^2}{n} + \Bigg(r \cdot 2^p \cdot (p+1) \cdot C \cdot d^{\frac{3p}{2}} \cdot A_n^{2p} \\
&\qquad \cdot \left(\max\left\{ \frac{1}{M_n}, \max_{l=1,\ldots,r} \|\mathbf{a}_l - \mathbf{b}_l^{(t)}\|_\infty \right\} \right)^p \Bigg)^2 \\
&\qquad +c_{33} \cdot A_n^{2p} \cdot \max_{l=1,\ldots,r} \|\mathbf{a}_l - \mathbf{b}_l^{(t)}\|_\infty^{2p} \Bigg) \\
&+c_{31} \cdot (\log n) \cdot n^{-\frac{2p}{2p+1}}.
\end{aligned}
$$

In case that

$$
\max\left\{ \frac{1}{M_n}, \max_{l=1,\ldots,r} \|\mathbf{a}_l - \mathbf{b}_l^{(t)}\|_\infty \right\} = \frac{1}{M_n}
$$

we have

$$
\begin{aligned}
&\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
\leq\ & c_{30} \cdot \frac{(\log n)^3 \cdot M_n}{n} + c_{32} \cdot \frac{(\log n)^2}{n} + c_{39} \cdot A_n^{2p} \cdot \frac{1}{M_n^{2p}}
\end{aligned}
$$

$$
\begin{aligned}
& +\mathbf{E}\left(\min_{t=1,\dots,I_n} c_{33}\cdot A_n^{2p}\cdot \max_{l=1,\dots,r}\|\mathbf{a}_l-\mathbf{b}_l^{(t)}\|_\infty^{2p}\right)\\
& +c_{31}\cdot(\log n)\cdot n^{-\frac{2p}{2p+1}}\\
\leq\ & c_{40}\cdot\frac{(\log n)^3}{n^{\frac{2p}{2p+1}}}+c_{32}\cdot\frac{(\log n)^2}{n}+c_{42}\cdot(\log n)^{\frac{2p}{6(N+d)}}\cdot\frac{1}{n^{\frac{2p}{2p+1}}}\\
& +c_{43}\cdot(\log n)^{\frac{2p}{6(N+d)}}\cdot\mathbf{E}\left(\min_{t=1,\dots,I_n}\max_{l=1,\dots,r}\|\mathbf{a}_l-\mathbf{b}_l^{(t)}\|_\infty^{2p}\right)\\
& +c_{31}\cdot(\log n)\cdot n^{-\frac{2p}{2p+1}}\\
\leq\ & c_{44}\cdot(\log n)^3\cdot n^{-\frac{2p}{2p+1}}+c_{43}\cdot(\log n)^{\frac{2p}{6(N+d)}}\cdot\mathbf{E}\left(\min_{t=1,\dots,I_n}\max_{l=1,\dots,r}\|\mathbf{a}_l-\mathbf{b}_l^{(t)}\|_\infty^{2p}\right)
\end{aligned}
$$

and in case that

$$
\max\left\{\frac{1}{M_n},\max_{l=1,\dots,r}\|\mathbf{a}_l-\mathbf{b}_l^{(t)}\|_\infty\right\}=\max_{l=1,\dots,r}\|\mathbf{a}_l-\mathbf{b}_l^{(t)}\|_\infty
$$

we have

$$
\begin{aligned}
& \mathbf{E}\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\\
\leq\ & c_{30}\cdot\frac{(\log n)^3\cdot M_n}{n}+c_{32}\cdot\frac{(\log n)^2}{n}\\
& +\mathbf{E}\left(\min_{t=1,\dots,I_n} c_{45}\cdot A_n^{2p}\cdot\max_{l=1,\dots,r}\|\mathbf{a}_l-\mathbf{b}_l^{(t)}\|_\infty^{2p}+c_{33}\cdot A_n^{2p}\cdot\max_{l=1,\dots,r}\|\mathbf{a}_l-\mathbf{b}_l^{(t)}\|_\infty^{2p}\right)\\
& +c_{31}\cdot(\log n)\cdot n^{-\frac{2p}{2p+1}}\\
\leq\ & c_{46}\cdot(\log n)^3\cdot n^{-\frac{2p}{2p+1}}+c_{47}\cdot(\log n)^{\frac{2p}{6(N+d)}}\cdot\mathbf{E}\left(\min_{t=1,\dots,I_n}\max_{l=1,\dots,r}\|\mathbf{a}_l-\mathbf{b}_l^{(t)}\|_\infty^{2p}\right).
\end{aligned}
$$

So, it remains to bound

$$
\mathbf{E}\left\{\min_{t=1,\dots,I_n}\max_{l=1,\dots,r}\|\mathbf{a}_l-\mathbf{b}_l^{(t)}\|_\infty^{2p}\right\}.
$$

By the random choice of the vectors $\mathbf{b}_s^{(i)}$ $s\in\{1,\dots,r\}, i\in\{1,\dots,I_n\}$ we know for any $u\in(0,2]$

$$
\mathbf{P}\left\{\min_{i=1,\dots,I_n}\max_{s=1,\dots,r}\|\mathbf{b}_s^{(i)}-\mathbf{a}_s\|_\infty>u\right\}\ =\ \prod_{i=1}^{I_n}\left(1-\mathbf{P}\left\{\max_{s=1,\dots,r}\|\mathbf{b}_s^{(i)}-\mathbf{a}_s\|_\infty\leq u\right\}\right)
$$

$$= \left(1 - \prod_{s=1}^{r} \mathbf{P}\left\{\|\mathbf{b}_s^{(i)} - \mathbf{a}_s\|_\infty \le u\right\}\right)^{I_n}.$$

Further, by (3.40) we know that for any $u \in (0, 2]$

$$\mathbf{P}\left\{\|\mathbf{b}_s^{(i)} - \mathbf{a}_s\|_\infty \le u\right\} \ge \mathbf{P}\left\{\|\mathbf{b}_s^{(i)} - \mathbf{a}_s\| \le u\right\} \ge c^* \cdot u^{d-1}.$$

Thus,

$$\begin{aligned}
& \mathbf{E}\left\{\min_{i=1,\dots,I_n} \max_{s=1,\dots,r} \|\mathbf{b}_s^{(i)} - \mathbf{a}_s\|_\infty^{2p}\right\} \\
= & \int_0^{2^{2p}} \mathbf{P}\left\{\min_{i=1,\dots,I_n} \max_{s=1,\dots,r} \|\mathbf{b}_s^{(i)} - \mathbf{a}_s\|_\infty > t^{\frac{1}{2p}}\right\} dt \\
\le & \int_0^{2^{2p}} \left(1 - \left(c^* t^{\frac{d-1}{2p}}\right)^r\right)^{I_n} dt \\
\le & \int_0^{2^{2p}} e^{-\left(c^* t^{\frac{d-1}{2p}}\right)^r I_n} dt \\
\le & \frac{1}{(c^*)^{\frac{2p}{d-1}}} \cdot \frac{2p}{r(d-1)} \cdot \int_0^\infty e^{-v} v^{\frac{2p}{r(d-1)}-1} dv \cdot I_n^{-\frac{2p}{r(d-1)}} \\
= & \frac{1}{(c^*)^{\frac{2p}{d-1}}} \cdot \frac{2p}{r(d-1)} \cdot \Gamma\left(\frac{2p}{r(d-1)}\right) \cdot I_n^{-\frac{2p}{r(d-1)}} \\
= & c_{48} \cdot I_n^{-\frac{2p}{r(d-1)}} \\
\le & c_{48} \cdot \left(\frac{(\log n)^2}{n}\right)^{\frac{2p}{2p+1}},
\end{aligned}$$

where the third inequality follows by substitution

$$v = (c^*)^r t^{\frac{r(d-1)}{2p}} I_n \Leftrightarrow \left(\frac{v}{(c^*)^r \cdot I_n}\right)^{\frac{2p}{r(d-1)}} = t = \varphi(v)$$

and

$$\frac{d\varphi}{dv} = \frac{1}{(c^*)^{\frac{2p}{d-1}}} \cdot I_n^{-\frac{2p}{r(d-1)}} \cdot \frac{2p}{r(d-1)} \cdot v^{\frac{2p}{r(d-1)}-1}$$

and where last equality follows since the the Gamma function Γ satisfies

$$\Gamma\left(\frac{2p}{r(d-1)}\right) < \infty.$$

Taking the above results together yields

$$\begin{aligned}
& \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
\leq \; & c_{49} \cdot (\log n)^3 \cdot n^{-\frac{2p}{2p+1}} + c_{50} \cdot (\log n)^{\frac{2p}{6(N+d)}} \cdot c_{48} \cdot \left(\frac{(\log n)^2}{n} \right)^{\frac{2p}{2p+1}} \\
\leq \; & c_{11} \cdot (\log n)^3 \cdot \left(\frac{1}{n} \right)^{\frac{2p}{2p+1}} .
\end{aligned}$$

This concludes the proof. □

4.3. Application to Simulated Data

We illustrate the finite sample size performance of our newly proposed estimate by applying it to simulated data using the software *MATLAB*.

For our simulation we choose the simulated data as follows: We choose X uniformly distributed on $[-1,1]^d$, where d is the dimension of the input, ϵ standard normal and independent of X, and we define Y by

$$Y = m_j(X) + \sigma \cdot \lambda_j \cdot \epsilon,$$

where $m_j : [-1,1]^d \to \mathbb{R}$ is described below, $\lambda_j > 0$ is a scaling value defined below and σ is chosen from $\{0.01, 0.05, 0.10, 0.20\}$ ($j \in \{1,2\}$). For comparability, we choose as regression functions the same ones as in Chapter 3:

$$\begin{aligned}
& m_1(x^{(1)}, x^{(2)}, x^{(3)}) \\
= \; & \frac{1}{1 + \exp(-(0.8317x^{(1)}) - 0.0277x^{(2)} + 0.5545x^{(3)})} \\
& + \sqrt{(-0.6461x^{(1)} - 0.1412x^{(2)} + 0.7501x^{(3)})^2 + 1}
\end{aligned}$$

and

$$\begin{aligned}
& m_2(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}) \\
= \; & \frac{1}{1 + \exp(-(-0.4863x^{(1)} + 0.5976x^{(2)} - 0.0209x^{(3)} + 0.5949x^{(4)} - 0.2281x^{(5)})} \\
& + \log \left(\frac{1}{(-0.6236x^{(1)} + 0.1244x^{(2)} + 0.4735x^{(3)} + 0.1914x^{(4)} - 0.5786x^{(5)})^2 + 2} \right) .
\end{aligned}$$

λ_j is chosen approximately as IQR of a sample of size 100 of $m(X)$, and we use the values $\lambda_1 = 0.2444$ and $\lambda_2 = 0.2515$.

From this distribution we generate a sample of size $n = 100$ and apply our newly proposed neural network regression estimate and compare our results to that of six alternative regression estimates on the same data. Then we compute the L_2 errors of these estimates approximately by using the empirical L_2 error $\varepsilon_{L_2,\bar{N}}(\cdot)$ on an independent sample of X of size $\bar{N} = 10,000$. Since this error strongly depends on the behavior of the true function m_j, we set it in relation to the error of the simplest estimate for m_j we can think of, a completely constant function. The constant function estimate describes the average of the observed data according to the least squares approach. Thus, the scaled error measure we use for evaluation of the estimates is $\varepsilon_{L_2,\bar{N}}(m_{n,i})/\bar{\varepsilon}_{L_2,\bar{N}}(avg)$, where $\bar{\varepsilon}_{L_2,\bar{N}}(avg)$ is the median of 50 independent realizations of the value obtained if the average of n observations is plugged into $\varepsilon_{L_2,\bar{N}}(\cdot)$. To a certain extent, this quotient can be interpreted as the relative part of the error of the constant estimate that is still contained in the more sophisticated approaches. Of course, the resulting scaled errors depend on the random sample of (X, Y) and in order to still be able to compare these values we repeat the whole computation 50 times and report the median and the interquartile range of the 50 scaled errors for each of our estimates.

We choose the parameters for each of the estimates by splitting of the sample. Here we split our sample into a learning sample of size $n_l = 0.8 \cdot n$ and into a testing sample of size $n_t = 0.2 \cdot n$. We compute the estimate for all parameter values from the sets described below using the learning sample. Then, we compute the corresponding empirical L_2 risk on the testing sample and choose the parameter value which leads to the minimal empirical L_2 risk on the testing sample.

Our first three estimates are built-in fully connected neural network estimates where the number of layers is fixed and the number of neurons per layer is chosen adaptively. The estimate *fc-neural-1* has one hidden layer, estimate *fc-neural-3* has three hidden layers, estimate *fc-neural-6* has six hidden layers and the number of neurons per layer is chosen from the set $\{5, 10, 25, 50, 75\}$, $\{3, 6, 9, 12, 15\}$, $\{2, 4, 6, 8, 10\}$, respectively.

Our fourth estimate *kernel* is the Nadaraya-Watson kernel estimate with so-called naive kernel where the bandwith is chosen from the set $\{2^k : k \in \{-5, -4, \ldots, 5\}\}$.

Our fifth estimate *neighbor* is a nearest neighbor estimate where the number of nearest neighbors is chosen from the set $\{1, 2, 3\} \cup \{4, 8, 12, 16, \ldots, 4 \cdot \lfloor \frac{n_l}{4} \rfloor\}$.

Our sixth estimate *RBF* is the interpoland with radial basis functions where the radial basis functions $\Phi(r) = (1 - r)_+^6 \cdot (35 \cdot r^2 + 18 \cdot r + 3)$ is used and the scaling radius is chosen adaptively.

Our last estimate *neural-3* is our newly proposed neural network estimate presented in this chapter. Here, the following parameters of the estimate are fixed: N is set to 2, A is

set to 1, and R is set to 10^6, and r is set to 4. The parameter M of the estimate is chosen from the set $\{2, 3, 4\}$. In order to accelerate the computation of this estimate we use only $I_n = 50$ random choices for the vectors of directions.

The results are summarized in Table 4.1 and in Table 4.2. As we can see from the reported scaled errors, our newly proposed neural network estimate outperforms all other estimates in five out of eight cases. In the other settings our proposed neural network can easily compete with the other estimates, in the sense that the values of the former lie within a small range of the best error value.

	m_1			
noise	1%	5%	10%	20%
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	0.0065	0.0065	0.0065	0.0066
approach	median (IQR)	median (IQR)	median (IQR)	median (IQR)
fc-neural-1	0.0358 (0.001)	0.0708 (0.001)	0.0581 (0.001)	0.1106 (0.001)
fc-neural-3	**0.0105 (0.002)**	0.0470 (0.002)	0.0414 (0.001)	0.1003 (0.002)
fc-neural-6	0.0209 (0.001)	0.0361 (0.001)	0.0497 (0.001)	**0.0867 (0.001)**
kernel	0.2478 (0.052)	0.2451 (0.067)	0.2436 (0.086)	0.246 (0.127)
neighbor	0.1168 (0.035)	0.1226 (0.046)	0.1815 (0.145)	0.2165 (0.121)
RBF	0.0117 (0.001)	0.2929 (0.012)	1.1759 (0.074)	5.9945 (3.003)
neural-3	0.0139 (0.016)	**0.0187 (0.007)**	**0.0284 (0.011)**	0.1636 (0.071)

Table 4.1.: Median and IQR of the scaled empirical L_2 error of estimates for m_1 for sample size $n = 100$. The smallest error values in each column is highlighted by bold letters.

	m_2			
noise	1%	5%	10%	20%
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	0.0073	0.0075	0.007	0.0073
approach	median (IQR)	median (IQR)	median (IQR)	median (IQR)
fc-neural-1	0.0278 (0.001)	0.0531 (0.004)	0.2241 (0.01)	0.5805 (0.006)
fc-neural-3	0.0567 (0.001)	0.0726 (0.001)	**0.0967 (0.002)**	1.2439 (0.005)
fc-neural-6	0.048 (0.002)	0.5121 (0.002)	0.4656 (0.002)	0.576 (0.005)
kernel	1.1081 (0.022)	1.1174 (0.013)	1.1386 (0.002)	1.2119 (0.040)
neighbor	0.3749 (0.158)	0.3978 (0.168)	0.4536 (0.195)	0.5734 (0.018)
RBF	0.0038 (0.001)	0.0512 (0.013)	0.1939 (0.039)	0.7595 (0.131)
neural-3	**0.0035 (0.001)**	**0.024483 (0.010)**	0.1041 (0.007)	**0.2068 (0.029)**

Table 4.2.: Median and IQR of the scaled empirical L_2 error of estimates for m_2 for sample size $n = 100$. The smallest error values in each column is highlighted by bold letters.

Bibliography

[1] Allen-Zhu, Z., Li, Y., and Liang, Y. (2019). Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, pp. 6155-6166.

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

[3] Anthony, M., and Bartlett, P. L. (1999). *Neural Networks and Learning: Theoretical Foundations.* Cambridge University Press, Cambridge, UK.

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

[5] Arora, S., Du, S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. (2019a). On exact computation with an infinitely wide neural net. *33rd Conference on Neural Information Processing Systems (NIPS 2019)*. Vancouver, Canada.

[6] Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. (2019b). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *Proceedings of the 36th International Conference on Machine Learning (PMLR 2019)*, **97**, pp. 322-332. Long Beach, California.

[7] Azar, A. T., and El-Said, S. A. (2013). Probabilistic neural network for breast cancer classification. *Neural Computing and Applications*, **23**, pp. 1737-1751.

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

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

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

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

[12] Bauer, B., Heimrich, F., Kohler, M., and Krzyżak, A. (2019). On estimation of surrogate models for high-dimensional computer experiments. *Annals of the Institute of Statistical Mathematics* **71**, pp. 107–136.

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

[14] Braun, A., Kohler, M., and Krzyżak, A. (2019) Analysis of the rate of convergence of neural network regression estimates which are easy to implement. *arXiv:1912.05436*.

[15] Braun, A., Kohler, M., Langer, S. and Walk, H. (2021). The smoking gun: statistical theory improves neural network estimates. *submitted for publication*.

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

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

[18] Cotrim, W.d.S., Minim, V.P.R, Felix, L.B., and Minim L.A. (2020). Short convolutional neural networks applied to the recognition of the browning stages of bread crust. *Journal of Food Engineering*, **277**.

[19] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, US.

[20] Devroye, L., and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, **8**, pp. 231-239.

[21] Dippon, J. (1998). Globally convergent stochastic optimization with optimal asymptotic distribution. *Journal of Applied Probability* **35**, pp. 395-406.

[22] Dippon, J., and Fabian, V. (1994). Stochastic approximation of global minimum points. *Journal of Statistical Planning and Inference* **41**, pp. 327-347.

[23] Dou, P., Shah, S. K., Kakadiaris, I. A. (2017). End-To-End 3D Face Reconstruction With Deep Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017*, pp. 5908-5917.

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

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

[26] Fabian, V. (1994). Comment on White (1989). *Journal of the American Statistical Association*, **89**, p. 1571.

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

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

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

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

[31] Haykin, S. O. (2008). *Neural Networks and Learning Machines*. 3rd ed. Prentice-Hall, New York, US.

[32] Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, California, US.

[33] Horowitz, J. L., and Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Annals of Statistics*, **35**, pp. 2589-2619.

[34] Huber, P. J. (1985). Projection pursuit. *Annals of Statistics*, **13**, pp. 435-475.

[35] Imaizumi, M., and Fukamizu, K. (2018). Deep neural networks learn non-smooth functions effectively. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*. Naha, Okinawa, Japan.

[36] Joshi, D. M., Rana, N. K., and Misra, V. M. (2010). Classification of Brain Cancer using Artificial Neural Network. *2010 2nd International Conference on Electronic Computer Technology*. Kuala Lumpur, Malaysia.

[37] Kahn, A.I., Shah, J.L, and Bhat, M.M. (2020). CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. *Computer Methods and Programs in Biomedicine*, **196**.

[38] Karimi, H., Nutini, J., and Schmidt, M. (2018). Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2016)*, pp. 795-811.

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

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

[41] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *arXiv: 1408.5882*.

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

[43] Kohler, M., and Krzyżak, A. (2019). Over-parametrized deep neural networks do not generalize well. arXiv:1912.03925.

[44] Kohler, M., Krzyżak, A., and Langer, S. (2020). Estimation of a function of low local dimensionality by deep neural networks. Submitted for publication.

[45] Kohler, M., and Langer, S. (2019). On the rate of convergence of fully connected very deep neural network regression estimates. Submitted for publication.

[46] Koltchinskii, V. (2004). Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, **34**, pp. 2593-2656.

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

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

[49] Kushner, H., and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications (2nd ed.)*. Springer, New York, USA.

[50] Lamrini, B., Valle, G. D., Trelea, I.C., Perrot, N., and Trystram, G. (2012). A new method for dynamic modelling of bread dough kneading based on artificial neural network. *Food Control*, **26**, pp. 512-524.

[51] Lepski, O., and Serdyukova, O. (2014). Adaptive estimation under single-index constraint in a regression model. *Annals of Statistics*, **42**, pp. 1-28.

[52] Li, Y., and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018)*, pp. 8168- 8177. Montreal, Canada.

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

[54] Luenberger, D., and Ye, Y. (2016). *Linear and Nonlinear Programming (4th ed.)*. Springer Science + Business Media, New York, USA.

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

[56] McCulloch, W.S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, pp. 115-133.

[57] Nasser, I. M., Abu-Naser, S. S. (2019). Lung Cancer Detection Using Artificial Neural Network. *International Journal of Engineering and Information Systems*, **3**, pp. 17-23.

[58] Parisi, R., Di Claudio, E.D., Lucarelli, G., and Orlandi, G. (2002). Car plate recognition by neural networks and image processing. *1998 IEEE International Symposium on Circuits and Systems (ISCAS)*. Monterey, CA, USA.

[59] Poggio, T., Banburski, A. , and Liao,Q. (2020). Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences*, **117**, pp.30039-30045.

[60] Poljak, B. T. (1981). Iterative algorithms for singular minimization problems. *Nonlinear Programming*, **4**, pp. 147-166.

[61] Richardson, E., Sela, M., Or-El, R., and Kimme, R. (2017). Learning Detailed Face Reconstruction From a Single Image. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017*, pp. 1259-1268.

[62] Ripley, B. D. (2008). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK.

[63] Rummelhart, D.E., Hinton, G.E., an Williams, R.J. (1986). Learning internal representations by error propagation. In Rummelhart, D.E. and McClelland, J.L. *Parallel Distributed Processing*, **1**, pp. 318-362. MIT Press.

[64] Sablani, S.S., Baik, O., and Marcotte, M. (2002). Neural networks for predicting thermal conductivity of bakery products. *Journal of Food Engineering*, **52**, pp. 299-304.

[65] Scarselli, F., and Tsoi, A. C. (1998). Universal Approximation Using Feedforward Neural Networks: A Survey of Some Existing Methods, and Some New Results. *Neural Networks*, **11**, pp. 15-37.

[66] Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks*, **61**, pp. 85-117.

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

[68] 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.

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

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

[71] Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics*, **22**, pp. 118-184.

[72] Tesauro, G. (2012). Neurogammon: a neural-network backgammon program. *1990 IJCNN International Joint Conference on Neural Networks*. San Diego, CA, USA.

[73] Tian, Y., Pei, K., Jana, S., and Ray, B. (2018). Deep Test: automated testing of deep-neural-network-driven autonomous cars. *Proceedings of the 40th International Conference on Software Enginieering*, pp. 303-314. Gothenburg, Sweden.

[74] Weissmann, J., and Salomon, R. (2002). Gesture recognition for virtual reality applications using data gloves and neural networks. *International Joint Conference on Neural Networks*. Washington, DC, USA.

[75] White, H. (1989). Some asymptotic results for learning in single hidden-layer feed-forward network models. *Journal of the American Statistical association*, **84**, pp. 1003-1013. Correction ibid. 87, p. 1252.

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

[77] Yang, J., Liu, T., Jiang, B., Song, H., and Lu, W. (2018). 3D Panoramic Virtual Reality Video Quality Assessment Based on 3D Convolutional Neural Networks. *IEEE Access*, **6**, pp. 38669-38682.

[78] Yannakakis, G.N., Levine, J., and Hallam, J. (2004). An evolutionary approach for interactive computer games. *Porceedings of the 2004 Congress on Evolutionary Computation*. Portland, OR, USA.

[79] Yu, Y., and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, **97**, pp. 1042-1054.

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

A. Supplement

A.1. Definitions

Definition A.1.1. *Let $\epsilon > 0$ and let $1 \leq p < \infty$. Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$. Let ν be a probability measure on $\mathbb{R}^d$. For a function $f : \mathbb{R}^d \to \mathbb{R}$ set*

$$\|f\|_{L_p(\nu)} = \left(\int |f(x)|^p \nu(dx) \right)^{\frac{1}{p}}$$

a) *An $\epsilon-$**cover of $\mathcal{F}$ with respect to** $\|\cdot\|_{L_p(\nu)}$ is a collection of functions $g_1, \ldots, g_N : \mathbb{R}^d \to \mathbb{R}$ with $N < \infty$ such that for every $f \in \mathcal{F}$*

$$\min_{j=1,\ldots,N} \|f - g_j\|_{L_p(\nu)} < \epsilon$$

b) *An $\epsilon-$**covering number of $\mathcal{F}$ with respect to** $\|\cdot\|_{L_p(\nu)}$ is the size N of the smallest $\epsilon-$cover of $\mathcal{F}$. We denote the $\epsilon-$covering number of $\mathcal{F}$ with respect to $\|\cdot\|_{L_p(\nu)}$ by*

$$\mathcal{N}\left(\epsilon, \mathcal{F}, \|\cdot\|_{L_p(\nu)}\right).$$

If no finite $\epsilon-$cover exists, we set $\mathcal{N}\left(\epsilon, \mathcal{F}, \|\cdot\|_{L_p(\nu)}\right) = \infty$.

c) *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be fixed and let $x_1^n = (x_1, \ldots, x_n)$. Let ν_n be the corresponding empirical measure, i.e.,*

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(x_i) \quad (A \subseteq \mathbb{R}^d).$$

*Then the $L_p - \epsilon-$**covering number of $\mathcal{F}$ on** x_1^n is the minimal number $N \in \mathbb{N}_0$ so that there exists a collection of functions $g_1, \ldots, g_N : \mathbb{R}^d \to \mathbb{R}$, such that for every $f \in \mathcal{F}$ it holds that*

$$\min_{j=1,\ldots,N} \left(\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - g_j(x_i)|^p \right)^{1/p} < \epsilon.$$

We denote $L_p - \epsilon-$covering number of $\mathcal{F}$ on x_1^n by

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

Definition A.1.2. *Let $\epsilon > 0$ and let $1 \leq p < \infty$. Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$. Let ν be a probability measure on $\mathbb{R}^d$. For a function $f : \mathbb{R}^d \to \mathbb{R}$ set*

$$\|f\|_{L_p(\nu)} = \left(\int |f(x)|^p \nu(dx)\right)^{\frac{1}{p}}$$

a) *An $\epsilon-$**packing of $\mathcal{F}$ with respect to** $\|\cdot\|_{L_p(\nu)}$ is a collection of functions $g_1, \ldots, g_N : \mathbb{R}^d \to \mathbb{R}$ with $N < \infty$ such that for every $f \in \mathcal{F}$*

$$\min_{j=1,\ldots,N} \|f - g_j\|_{L_p(\nu)} \geq \epsilon$$

b) *An $\epsilon-$**packing number of $\mathcal{F}$ with respect to** $\|\cdot\|_{L_p(\nu)}$ is the size N of the largest $\epsilon-$packing of $\mathcal{F}$. We denote the $\epsilon-$packing number of $\mathcal{G}$ with respect to $\|\cdot\|_{L_p(\nu)}$ by*

$$\mathcal{M}(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}).$$

If there exists an $\epsilon-$packing of size N for every $N \in \mathbb{N}$, we set $\mathcal{M}(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}) = \infty$.

c) *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be fixed and let $x_1^n = (x_1, \ldots, x_n)$. Let ν_n be the corresponding empirical measure, i.e.,*

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1_A(x_i) \quad (A \subseteq \mathbb{R}^d).$$

*Then the $L_p - \epsilon-$**packing number of $\mathcal{G}$ on** x_1^n is the maximal number $N \in \mathbb{N}_0$ so that there exists a collection of functions $g_1, \ldots, g_N : \mathbb{R}^d \to \mathbb{R}$ with $g_1, \ldots, g_n \in \mathcal{G}$ with*

$$\left(\frac{1}{n}\sum_{i=1}^{n} |g_j(x_i) - g_k(x_i)|^p\right)^{1/p} \geq \epsilon$$

for all $1 \leq j < k \leq N$. We denote $L_p - \epsilon-$packing number of $\mathcal{G}$ on x_1^n by

$$\mathcal{M}_p\left(\epsilon, \mathcal{G}, x_1^n\right).$$

Definition A.1.3. *Let $\mathcal{A}$ be a class of subsets of $\mathbb{R}^d$ and let $n \in \mathbb{N}$.*

a) For $z_1, \ldots, z_n \in \mathbb{R}^d$ define

$$s(\mathcal{A}, \{z_1, \ldots, z_n\}) = |\{A \cap \{z_1, \ldots, z_n\} : A \in \mathcal{A}\}|.$$

*b) Let G be a subset of $\mathbb{R}^d$ of size n. We say that $\mathcal{A}$ **shatters** G if $s(\mathcal{A}, G) = 2^n$. This means that each subset of G can be represented in the form $A \cap G$ for some $A \in \mathcal{A}$.*

*c) The $n-$**th shatter coefficient** of $\mathcal{A}$ is*

$$S(\mathcal{A}, n) = \max_{\{z_1, \ldots, z_n\} \subseteq \mathbb{R}^d} s(\mathcal{A}, \{z_1, \ldots, z_n\}).$$

In other words, the shatter coefficient is the maximal number of different subsets of n points that can be picked out by sets of $\mathcal{A}$.

Definition A.1.4. *Let $\mathcal{A}$ be a class of subsets of $\mathbb{R}^d$ with $\mathcal{A} \neq \emptyset$. The **VC dimension** (or Vapnik-Chervonenkis dimension) $V_{\mathcal{A}}$ of $\mathcal{A}$ is defined by*

$$V_{\mathcal{A}} = \sup\{n \in \mathbb{N} : S(\mathcal{A}, n) = 2^n\}.$$

In other words, the VC dimension is the largest integer n such that there exists a set of n points in $\mathbb{R}^d$ which can be shattered by $\mathcal{A}$.

Definition A.1.5. *The **support of a d-dimensional random variable** X is given by*

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

The support is the smallest closed set with probabilty one according to $\mathbf{P}_X$.

A.2. Proof of Lemma 2.2.14

In the proof we use the following error decomposition:

$$\Bigg(\Bigg(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)$$
$$-2 \cdot \Bigg(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| \leq \beta_n \, (j \in \{1, \ldots, n\})\}}$$
$$-\frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\Bigg)\Bigg) \cdot 1_{A_n}\Bigg)$$

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

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

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

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

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

With the Cauchy-Schwarz inequality and

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

we conclude

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

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

$$|Y|^2 \leq \frac{2}{c_4} \cdot \exp\left(\frac{c_4}{2} \cdot |Y|^2\right) \tag{A.1}$$

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

$$\mathbf{E}\left(\frac{2}{c_4} \cdot \exp\left(c_4/2 \cdot |Y|^2\right) \cdot \exp(c_4/2 \cdot |Y|^2)\right) \leq \mathbf{E}\left(\frac{2}{c_4} \cdot \exp\left(c_4 \cdot |Y|^2\right)\right) \leq c_{38}$$

which is less than infinity by the assumptions of the lemma. Furthermore, in the same way the third term is bounded by $\sqrt{18\beta_n^2 + c_{39}}$ because

$$\mathbf{E}(|Y|^2) \leq \mathbf{E}(2/c_4 \cdot \exp(c_4 \cdot |Y|^2)) \leq c_{39} < \infty. \tag{A.2}$$

With $\beta_n = c_3 \cdot \log(n)$ we have for some constants $c_{40}, c_{41} > 0$ that

$$|T_{5,n}| \leq \sqrt{c_{38}} \cdot \exp\left(-c_{40} \cdot \log(n)^2\right) \cdot \sqrt{(18 \cdot c_3 \cdot \log(n)^2 + c_{39})} \leq c_{41} \cdot \frac{\log(n)}{n}.$$

From the Cauchy-Schwarz inequality we get

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

$$\cdot \sqrt{\mathbf{E}\left\{\left|m(X)+m_{\beta_n}(X)-Y-T_{\beta_n}Y\right|^2\right\}},$$

where we can bound the second factor on the right-hand side in the above inequality in the same way we have bounded the second factor from $T_{5,n}$, because by assumption $||m||_\infty$ is bounded and furthermore m_{β_n} is bounded by β_n. Thus, we get for some constant $c_{42} > 0$

$$\sqrt{\mathbf{E}\left\{\left|m(X)+m_{\beta_n}(X)-Y-T_{\beta_n}Y\right|^2\right\}} \leq c_{42} \cdot \log(n).$$

Next, we consider the first term. With Jensen's inequality we get that

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

Hence, we get

$$|T_{6,n}| \quad \leq \quad \sqrt{4 \cdot \mathbf{E}\left\{|Y-T_{\beta_n}Y|^2\right\}} \cdot c_{42} \cdot \log(n)$$

and therefore with the calculations from $T_{5,n}$ we have that $T_{6,n} \leq c_{43} \cdot \log(n)/n$ for some constant $c_{43} > 0$. Altogether we get

$$T_{1,n} \cdot 1_{A_n} \leq |T_{5,n}| + |T_{6,n}| \leq c_{44} \cdot \frac{\log(n)}{n}$$

for some constant $c_{44} > 0$.
Next, we consider $T_{2,n} \cdot 1_{A_n}$ and conclude for $t > 0$

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

where $T_{\beta_n,supp(X)}\mathcal{F}_n$ is defined as $\left\{T_{\beta_n} f \cdot 1_{supp(X)} : f \in \mathcal{F}_n\right\}$. Theorem 11.4 in Györfi et al. (2002) and the relation $\mathcal{N}\left(\delta, \left\{\frac{1}{\beta_n} g : g \in \mathcal{G}\right\}, \|\cdot\|_{\infty,supp(X)}\right) \leq \mathcal{N}\left(\delta \cdot \beta_n, \mathcal{G}, \|\cdot\|_{\infty,supp(X)}\right)$ for an arbitrary function space $\mathcal{G}$ and $\delta > 0$ lead to

$$\mathbf{P}\{T_{2,n} \cdot 1_{A_n} > t\} \leq 14 \cdot \mathcal{N}\left(\frac{t}{80 \cdot \beta_n}, \mathcal{F}_n, \|\cdot\|_{\infty,supp(X)}\right) \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot t\right).$$

Since the covering number and the exponential factor are decreasing in t, we can conclude for $\varepsilon_n \geq \frac{80}{n}$

$$\begin{aligned}
&\mathbf{E}(T_{2,n} \cdot 1_{A_n}) \\
&\leq \varepsilon_n + \int_{\varepsilon_n}^{\infty} \mathbf{P}\{T_{2,n} \cdot 1_{A_n} > t\} dt \\
&\leq \varepsilon_n + 14 \cdot \mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, \|\cdot\|_{\infty,supp(X)}\right) \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot \varepsilon_n\right) \cdot \frac{5136 \cdot \beta_n^2}{n}.
\end{aligned}$$

Choosing

$$\varepsilon_n = \frac{5136 \cdot \beta_n^2}{n} \cdot \log\left(14 \cdot \mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, \|\cdot\|_{\infty,supp(X)}\right)\right)$$

(which satisfies the necessary condition $\varepsilon_n \geq \frac{80}{n}$ if the constant c_5 in the definition of β_n is not too small) minimizes the right-hand side and implies

$$\mathbf{E}(T_{2,n} \cdot 1_{A_n}) \leq \frac{c_7 \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, \|\cdot\|_{\infty,supp(X)}\right)\right)}{n}.$$

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

$$\mathbf{E}(T_{3,n} \cdot 1_{A_n}) \quad \leq \quad c_{45} \cdot \frac{\log(n)}{n}$$

for some large enough constant $c_{45} > 0$ and hence we get in total

$$\mathbf{E}\left(\sum_{i=1}^{3} T_{i,n} \cdot 1_{A_n}\right) \quad \leq \quad \frac{c_{46} \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, \|\cdot\|_{\infty,supp(X)}\right)\right)}{n}$$

for some sufficient large constant $c_{46} > 0$.

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

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

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

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

$$
\begin{aligned}
&\mathbf{E}\{T_{4,n} \cdot 1_{A_n}\} \\
&\leq \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot 1_{\{|Y_j| > \beta_n \text{ for some } j \in \{1,\dots,n\}\}}\right\} \\
&= \mathbf{E}\left\{|m_n(X_1) - Y_1|^2 \cdot 1_{\{|Y_j| > \beta_n \text{ for some } j \in \{1,\dots,n\}\}}\right\} \\
&\leq \sqrt{\mathbf{E}\left\{|m_n(X_1) - Y_1|^4\right\}} \cdot \sqrt{\mathbf{P}\{|Y_j| > \beta_n \text{ for some } j \in \{1,\dots,n\}\}} \\
&\leq c_{104} \cdot \beta_n^2 \cdot \frac{1}{n},
\end{aligned}
$$

where the last inequality holds since by Markov's inequality

$$
\begin{aligned}
\mathbf{P}\left\{|Y_i| > \beta_n \text{ for some } i \in \{1,\dots,n\}\right\} &= n \cdot \mathbf{P}\left\{\exp(c_{10} \cdot Y^2) > \exp(c_{10} \cdot \beta_n^2)\right\} \\
&\leq n \cdot \frac{\mathbf{E}\left\{\exp(c_{10} \cdot Y^2)\right\}}{\exp\left(c_{10} \cdot (c_2 \cdot \log n)^2\right)} \\
&\leq 4 \cdot c_2^2 \cdot (\log n)^2 \cdot n \cdot \frac{c_{40}}{n^{c_{41} \cdot \log n}} \\
&\leq c_{103} \cdot \frac{1}{n}
\end{aligned}
$$

and (2.7) and $(A.1)$, which imply

$$
\begin{aligned}
\mathbf{E}\left\{|m_n(X_1) - Y_1|^4\right\} &\leq 16 \cdot (\beta_n^4 + \mathbf{E}\{Y^4\})) \leq 16 \cdot \left(\beta_n^4 + \frac{4}{c_4^1} \cdot \mathbf{E}\left\{\exp\left(c_4 \cdot Y^2\right)\right\}\right) \\
&\leq c_{105} \cdot \beta_n^4.
\end{aligned}
$$

In combination with the other considerations in the proof this implies the assertion of Lemma 2.2.14. □

Wissenschaftlicher Werdegang

Ausbildung

2012	**Abitur**, *Sankt Lioba Schule Bad Nauheim*, Bad Nauheim.
2012 - 2016	**Bachelor (B.Sc.)**, *Technische Universität Darmstadt*, Darmstadt. Mathematik mit Nebenfach Wirtschaftswissenschaften (bilingual). Thesis: Cover Times.
2016 - 2018	**Master (M.Sc.)**, *Technische Universität Darmstadt*, Darmstadt. Mathematik mit nicht-mathematischer Vertiefung Informatik, Nebenfach Psychologie. Thesis: A Linear Neural Network Regression Estimate.
2018 - 2021	**Wissenschaftliche Mitarbeiterin**, *Technische Universität Darmstadt*, Darmstadt. Fachbereich Mathematik, Arbeitsgruppe Stochastik.
2021	**Promotion**, *Technische Universität Darmstadt*, Darmstadt.

Publikationen

Braun, A., Kohler, M., and Krzyżak, A. (2019)
Analysis of the rate of convergence of neural network regression estimates which are easy to implement. *Zur Veröffentlichung eingereicht. Preprint, arXiv:1912.05436.*

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

Braun, A., Kohler, M., Langer, S. and Walk, H. (2021).
The smoking gun: statistical theory improves neural network estimates. *Zur Veröffentlichung eingereicht.*