ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Full error analysis for the training of deep neural networks

Ch. Beck and A. Jentzen and B. Kuckuck

Research Report No. 2019-57
October 2019

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Full error analysis for the training of deep neural networks

Christian Beck[1], Arnulf Jentzen[2,3], and Benno Kuckuck[4,5]

[1] Department of Mathematics, ETH Zurich, Zürich, Switzerland, e-mail: christian.beck@math.ethz.ch

[2] Department of Mathematics, ETH Zurich, Zürich, Switzerland, e-mail: arnulf.jentzen@sam.math.ethz.ch

[3] Faculty of Mathematics and Computer Science, University of Münster, Münster, Germany, e-mail: ajentzen@uni-muenster.de

[4] Institute of Mathematics, University of Düsseldorf, Düsseldorf, Germany, e-mail: kuckuck@math.uni-duesseldorf.de

[5] Faculty of Mathematics and Computer Science, University of Münster, Münster, Germany, e-mail: benno.kuckuck@uni-muenster.de

October 3, 2019

Abstract

Deep learning algorithms have been applied very successfully in recent years to a range of problems out of reach for classical solution paradigms. Nevertheless, there is no completely rigorous mathematical error and convergence analysis which explains the success of deep learning algorithms. The error of a deep learning algorithm can in many situations be decomposed into three parts, the approximation error, the generalization error, and the optimization error. In this work we estimate for a certain deep learning algorithm each of these three errors and combine these three error estimates to obtain an overall error analysis for the deep learning algorithm under consideration. In particular, we thereby establish convergence with a suitable convergence speed for the overall error of the deep learning algorithm under consideration. Our convergence speed analysis is far from optimal and the convergence speed that we establish is rather slow, increases exponentially in the dimensions, and, in particular, suffers from the curse of dimensionality. The main contribution of this work is, instead, to provide a full error analysis (i) which covers each of the three different sources of errors usually emerging in deep learning algorithms and (ii) which merges these three sources of errors into one overall error estimate for the considered deep learning algorithm.

Contents

1 Introduction

In problems like image recognition, text analysis, speech recognition, or playing various games, to name a few, it is very hard and seems at the moment entirely impossible to provide a function or to hard-code a computer program which attaches to the input – be it a picture, a piece of text, an audio recording, or a certain game situation – a meaning or a recommended action. Nevertheless deep learning has been applied very successfully in recent years to such and related problems. The success of deep learning in applications is even more surprising as, to this day, the reasons for its performance are not entirely rigorously understood. In particular, there is no rigorous mathematical error and convergence analysis which explains the success of deep learning algorithms.

In contrast to traditional approaches, machine learning methods in general and deep learning methods in particular attempt to infer the unknown target function or at least a good enough approximation thereof from examples encountered during the training. Often a deep learning algorithm has three ingredients: (i) the *hypothesis class*, a parametrizable class of functions in which we try to find a reasonable approximation of the unknown target function, (ii) a *numerical approximation of the expected loss function* based on the training examples, and (iii) an *optimization algorithm* which tries to approximately calculate an element of the hypothesis class which minimizes the numerical approximation of the expected loss function from

(ii) given the training examples. Common approaches are to choose a set of suitable fully connected deep neural networks (DNNs) as hypothesis class in (i), empirical risks as approximations of the expected loss function in (ii), and stochastic gradient descent-type algorithms with random initializations as optimization algorithms in (iii). Each of these three ingredients contributes to the overall error of the considered approximation algorithm. The choice of the hypothesis class results in the so-called *approximation error* (cf., e.g., [3, 4, 19, 38, 40, 41] and the references mentioned at the beginning of Section 3), replacing the exact expected loss function by a numerical approximation leads to the so-called *generalization error* (cf., e.g., [5, 10, 18, 35, 51, 68, 71] and the references mentioned therein), and the employed optimization algorithm introduces the *optimization error* (cf., e.g., [2, 6, 9, 15, 20, 26, 43, 45] and the references mentioned therein).

In this work we estimate the approximation error, the generalization error, as well as the optimization error and we also combine these three errors to establish convergence with a suitable convergence speed for the overall error of the deep learning algorithm under consideration. Our convergence speed analysis is far from optimal and the convergence speed that we establish is rather slow, increases exponentially in the dimensions, and, in particular, suffers from the curse of dimensionality (cf., e.g., Bellman [8], Novak & Woźniakowski [56, Chapter 1], and Novak & Woźniakowski [57, Chapter 9]). The main contribution of this work is, instead, to provide a full error analysis (i) which covers each of the three different sources of errors usually emerging in deep learning algorithms and (ii) which merges these three sources of errors into one overall error estimate for the considered deep learning algorithm. In the next result, Theorem 1.1, we briefly illustrate the findings of this article in a special case and we refer to Section 4.2 below for the more general convergence results which we develop in this article.

Theorem 1.1. *Let $d \in \mathbb{N}$, $L, a, u \in \mathbb{R}$, $b \in (a, \infty)$, $v \in (u, \infty)$, $R \in [\max\{1, L, |a|, |b|, 2|u|, 2|v|\}, \infty)$, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $X_m \colon \Omega \to [a, b]^d$, $m \in \mathbb{N}$, be i.i.d. random variables, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be the standard norm on $\mathbb{R}^d$, let $\varphi \colon [a, b]^d \to [u, v]$ satisfy for all $x, y \in [a, b]^d$ that $|\varphi(x) - \varphi(y)| \leq L\|x - y\|$, for every $\mathfrak{d}, r, s \in \mathbb{N}$, $\delta \in \mathbb{N}_0$, $\theta = (\theta_1, \theta_2, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq \delta + rs + r$ let $\mathcal{A}^{\theta,\delta}_{r,s} \colon \mathbb{R}^s \to \mathbb{R}^r$ satisfy for all $x = (x_1, x_2, \ldots, x_s) \in \mathbb{R}^s$ that*

$$\mathcal{A}^{\theta,\delta}_{r,s}(x) = \left(\left[\sum_{i=1}^{s} x_i \theta_{\delta+i} \right] + \theta_{\delta+rs+1}, \left[\sum_{i=1}^{s} x_i \theta_{\delta+s+i} \right] + \theta_{\delta+rs+2}, \ldots, \left[\sum_{i=1}^{s} x_i \theta_{\delta+(r-1)s+i} \right] + \theta_{\delta+rs+r} \right), \quad (1)$$

let $\mathfrak{c} \colon \mathbb{R} \to [u, v]$ and $\mathfrak{R}_\tau \colon \mathbb{R}^\tau \to \mathbb{R}^\tau$, $\tau \in \mathbb{N}$, satisfy for all $\tau \in \mathbb{N}$, $x = (x_1, x_2, \ldots, x_\tau) \in \mathbb{R}^\tau$, $y \in \mathbb{R}$ that $\mathfrak{c}(y) = \min\{v, \max\{u, y\}\}$ and $\mathfrak{R}_\tau(x) = (\max\{x_1, 0\}, \max\{x_2, 0\}, \ldots, \max\{x_\tau, 0\})$, for every $\mathfrak{d}, \tau \in \{3, 4, \ldots\}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq \tau(d + 1) + (\tau - 3)\tau(\tau + 1) + \tau + 1$ let $\mathfrak{N}^{\theta,\tau} \colon \mathbb{R}^d \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}^d$ that

$$\big(\mathfrak{N}^{\theta,\tau}\big)(x) = \big(\mathfrak{c} \circ \mathcal{A}^{\theta,\tau(d+1)+(\tau-3)\tau(\tau+1)}_{1,\tau} \circ \mathfrak{R}_\tau \circ \mathcal{A}^{\theta,\tau(d+1)+(\tau-4)\tau(\tau+1)}_{\tau,\tau} \circ \mathfrak{R}_\tau \circ \ldots \circ \mathcal{A}^{\theta,\tau(d+1)}_{\tau,\tau} \circ \mathfrak{R}_\tau \circ \mathcal{A}^{\theta,0}_{\tau,d}\big)(x), \quad (2)$$

let $\mathfrak{E}_{\mathfrak{d},M,\tau} \colon [-R, R]^{\mathfrak{d}} \times \Omega \to [0, \infty)$, $\mathfrak{d}, M, \tau \in \mathbb{N}$, satisfy for all $\mathfrak{d}, M \in \mathbb{N}$, $\tau \in \{3, 4, \ldots\}$, $\theta \in [-R, R]^{\mathfrak{d}}$, $\omega \in \Omega$ with $\mathfrak{d} \geq \tau(d + 1) + (\tau - 3)\tau(\tau + 1) + \tau + 1$ that

$$\mathfrak{E}_{\mathfrak{d},M,\tau}(\theta, \omega) = \frac{1}{M} \left[\sum_{m=1}^{M} |\mathfrak{N}^{\theta,\tau}(X_m(\omega)) - \varphi(X_m(\omega))|^2 \right], \quad (3)$$

for every $\mathfrak{d} \in \mathbb{N}$ let $\Theta_{\mathfrak{d},k} \colon \Omega \to [-R, R]^{\mathfrak{d}}$, $k \in \mathbb{N}$, be i.i.d. random variables, assume for every $\mathfrak{d} \in \mathbb{N}$ that $\Theta_{\mathfrak{d},1}$ is continuous uniformly distributed on $[-R, R]^{\mathfrak{d}}$, assume for every $\mathfrak{d} \in \mathbb{N}$ that $(X_m)_{m\in\mathbb{N}}$ and $(\Theta_{\mathfrak{d},k})_{k\in\mathbb{N}}$ are independent, and let $\Xi_{\mathfrak{d},K,M,\tau} \colon \Omega \to [-R, R]^{\mathfrak{d}}$, $\mathfrak{d}, K, M, \tau \in \mathbb{N}$, satisfy for all $\mathfrak{d}, K, M, \tau \in \mathbb{N}$ that $\Xi_{\mathfrak{d},K,M,\tau} = \Theta_{\mathfrak{d},\min\{k\in\{1,2,\ldots,K\} \colon \mathfrak{E}_{\mathfrak{d},M,\tau}(\Theta_{\mathfrak{d},k})=\min_{l\in\{1,2,\ldots,K\}} \mathfrak{E}_{\mathfrak{d},M,\tau}(\Theta_{\mathfrak{d},l})\}}$. Then there exists $c \in (0, \infty)$ such that for all $\mathfrak{d}, K, M, \tau \in \mathbb{N}$, $\varepsilon \in (0, 1]$ with $[2d(2dL(v - u)\varepsilon^{-1} + 2)^d + 2]^3 \leq \tau^3 \leq \mathfrak{d}$ it holds that

$$\mathbb{P}\left(\int_{[a,b]^d} |\mathfrak{N}^{\Xi_{\mathfrak{d},K,M,\tau},\tau}(x) - \varphi(x)| \, \mathbb{P}_{X_1}(dx) > \varepsilon \right) \leq \exp\big(-K(c\tau)^{-\tau\mathfrak{d}}\varepsilon^{2\mathfrak{d}}\big) + 2\exp\big(\mathfrak{d}\ln\big((c\tau)^\tau\varepsilon^{-2}\big) - \tfrac{\varepsilon^4 M}{c}\big). \quad (4)$$

Theorem 1.1 is an immediate consequence of Corollary 4.8 in Section 4.2 below. Corollary 4.8 follows from Corollary 4.7 which, in turn, is implied by Theorem 4.5, the main result of this article. In the following we add some comments and explanations regarding the mathematical objects which appear in Theorem 1.1 above. For every $\mathfrak{d}, \tau \in \{3, 4, ...\}, \theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq \tau(d+1) + (\tau-3)\tau(\tau+1) + \tau + 1$ the function $\mathfrak{N}^{\theta,\tau} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ in (2) above describes the realization of a fully connected deep neural network with τ layers (1 input layer with d neurons [d dimensions], 1 output layer with 1 neuron [1 dimension], as well as $\tau - 2$ hidden layers with τ neurons on each hidden layer [τ dimensions in each hidden layer]). The vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ in (2) in Theorem 1.1 above stores the real parameters (the weights and the biases) for the concrete considered neural network. In particular, the architecture of the deep neural network in (2) is chosen so that we have $\tau d + (\tau-3)\tau^2 + \tau$ real parameters in the weight matrices and $(\tau-2)\tau+1$ real parameters in the bias vectors resulting in $[\tau d + (\tau-3)\tau^2 + \tau] + [(\tau-2)\tau+1] = \tau(d+1) + (\tau-3)\tau(\tau+1) + \tau + 1$ real parameters for the deep neural network overall. This explains why the dimension $\mathfrak{d}$ of the parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ must be larger or equal than the number of real parameters used to describe the deep neural network in (2) in the sense that $\mathfrak{d} \geq \tau(d+1) + (\tau-3)\tau(\tau+1) + \tau + 1$ (see above (2)). The affine linear transformations for the deep neural network, which appear just after the input layer and just after each hidden layer in (2), are specified in (1) above. The functions $\mathfrak{R}_\tau \colon \mathbb{R}^\tau \to \mathbb{R}$, $\tau \in \mathbb{N}$, describe the multi-dimensional rectifier functions which are employed as activation functions in (2). Realizations of the random variables $(X_m, Y_m) := (X_m, \varphi(X_m))$, $m \in \{1, \ldots, M\}$, act as training data and the neural network parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ should be chosen so that the empirical risk in (3) gets minimized. In Theorem 1.1 above, we use as an optimization algorithm just random initializations and perform no gradient descent steps. The inequality in (4) in Theorem 1.1 above provides a quantitative error estimate for the probability that the L^1-distance between the trained deep neural network approximation $\mathfrak{N}^{\Xi_{\mathfrak{d},K,M,\tau},\tau}(x)$, $x \in [a,b]^d$, and the function $\varphi(x)$, $x \in [a,b]^d$, which we actually want to learn, is larger than a possibly arbitrarily small real number $\varepsilon \in (0,1]$. In (4) in Theorem 1.1 above we measure the error between the deep neural network and the function $\varphi \colon [a,b]^d \to [u,v]$, which we intend to learn, in the L^1-distance. However, in the more general results in Section 4.2 below we measure the error in the L^2-distance and, just to keep the statement in Theorem 1.1 as easily accessible as possible, we restrict ourselves in Theorem 1.1 above to the L^1-distance. Observe that for every $\varepsilon \in (0,1]$ and every $\mathfrak{d}, \tau \in \{3, 4, \ldots\}$ with $\mathfrak{d} \geq \tau(d+1) + (\tau-3)\tau(\tau+1) + \tau + 1$ we have that the right hand side of (4) converges to zero as K and M tend to infinity. The right hand side of (4) also specifies a concrete speed of convergence and in this sense Theorem 1.1 provides a full error analysis for the deep learning algorithm under consideration. Our analysis is in parts inspired by Maggi [50], Berner et al. [10], Cucker & Smale [18], Beck et al. [6], and Fehrman et al. [26].

The remainder of this article is organized as follows. In Section 2 we present two elementary approaches how DNNs can be described in a mathematical fashion. Both approaches will be used in our error analyses in the later parts of this article. In Section 3 we separately analyze the approximation error, the generalization error, and the optimization error of the considered algorithm. In Section 4 we combine the separate error analyses in Section 3 to obtain an overall error analysis of the considered algorithm.

2 Deep neural networks (DNNs)

In this section we present two elementary approaches on how DNNs can be described in a mathematical fashion. More specifically, we present in Section 2.1 a vectorized description for DNNs and we present in Section 2.2 a structured description for DNNs. Both approaches will be used in our error analyses in the later parts of this article. Sections 2.1 and 2.2 are partially based on material in publications from the scientific literature such as Beck et al. [6, 7], Berner et al. [10], Goodfellow et al. [28], and Grohs et al. [31, 32]. In particular, Definition 2.1 is inspired by (25) in [7], Definition 2.2 is inspired by (26) in [7], Definition 2.3 is [31, Definition 2.2], Definitions 2.6, 2.7, and 2.8 are inspired by [10, Setting 2.3], Theorem 2.10 is a strengthened version of [10, Theorem 4.2], Definition 2.13 is [31, Definition 2.1], Definition 2.14 is [31, Definition 2.3], Definition 2.15 is [32, Definition 3.10], Definition 2.16 is [31, Definition 2.5], and Definition 2.17 is [31, Definition 2.17]. The proof of Theorem 2.10 is analogous to the proof of Theorem 4.2 in [10] and is therefore omitted.

2.1 Vectorized description of DNNs

2.1.1 Affine functions

Definition 2.1 (Affine function)**.** Let $d, r, s \in \mathbb{N}$, $\delta \in \mathbb{N}_0$, $\theta = (\theta_1, \theta_2, \dots, \theta_d) \in \mathbb{R}^d$ satisfy $d \geq \delta + rs + r$. Then we denote by $\mathcal{A}_{r,s}^{\theta,\delta} \colon \mathbb{R}^s \to \mathbb{R}^r$ the function which satisfies for all $x = (x_1, x_2, \dots, x_s) \in \mathbb{R}^s$ that

$$\begin{aligned} \mathcal{A}_{r,s}^{\theta,\delta}(x) &= \begin{pmatrix} \theta_{\delta+1} & \theta_{\delta+2} & \cdots & \theta_{\delta+s} \\ \theta_{\delta+s+1} & \theta_{\delta+s+2} & \cdots & \theta_{\delta+2s} \\ \theta_{\delta+2s+1} & \theta_{\delta+2s+2} & \cdots & \theta_{\delta+3s} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{\delta+(r-1)s+1} & \theta_{\delta+(r-1)s+2} & \cdots & \theta_{\delta+rs} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_s \end{pmatrix} + \begin{pmatrix} \theta_{\delta+rs+1} \\ \theta_{\delta+rs+2} \\ \theta_{\delta+rs+3} \\ \vdots \\ \theta_{\delta+rs+r} \end{pmatrix} \\ &= \Big(\Big[\textstyle\sum_{k=1}^{s} x_k \theta_{\delta+k} \Big] + \theta_{\delta+rs+1}, \Big[\textstyle\sum_{k=1}^{s} x_k \theta_{\delta+s+k} \Big] + \theta_{\delta+rs+2}, \dots, \Big[\textstyle\sum_{k=1}^{s} x_k \theta_{\delta+(r-1)s+k} \Big] + \theta_{\delta+rs+r} \Big). \end{aligned} \tag{5}$$

2.1.2 Vectorized description of DNNs

Definition 2.2. Let $d, L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $\delta \in \mathbb{N}_0$, $\theta \in \mathbb{R}^d$ satisfy

$$d \geq \delta + \sum_{k=1}^{L} l_k(l_{k-1} + 1) \tag{6}$$

and let $\Psi_k \colon \mathbb{R}^{l_k} \to \mathbb{R}^{l_k}$, $k \in \{1, 2, \dots, L\}$, be functions. Then we denote by $\mathcal{N}_{\Psi_1,\Psi_2,\dots,\Psi_L}^{\theta,\delta,l_0} \colon \mathbb{R}^{l_0} \to \mathbb{R}^{l_L}$ the function which satisfies for all $x \in \mathbb{R}^{l_0}$ that

$$\begin{aligned} \big(\mathcal{N}_{\Psi_1,\Psi_2,\dots,\Psi_L}^{\theta,\delta,l_0}\big)(x) = \big(\Psi_L \circ \mathcal{A}_{l_L,l_{L-1}}^{\theta,\delta+\sum_{k=1}^{L-1} l_k(l_{k-1}+1)} \circ \Psi_{L-1} \circ \mathcal{A}_{l_{L-1},l_{L-2}}^{\theta,\delta+\sum_{k=1}^{L-2} l_k(l_{k-1}+1)} \circ \dots \\ \dots \circ \Psi_2 \circ \mathcal{A}_{l_2,l_1}^{\theta,\delta+l_1(l_0+1)} \circ \Psi_1 \circ \mathcal{A}_{l_1,l_0}^{\theta,\delta}\big)(x) \end{aligned} \tag{7}$$

(cf. Definition 2.1).

2.1.3 Activation functions

Definition 2.3 (Multidimensional version)**.** Let $d \in \mathbb{N}$ and let $\psi \colon \mathbb{R} \to \mathbb{R}$ be a function. Then we denote by $\mathfrak{M}_{\psi,d} \colon \mathbb{R}^d \to \mathbb{R}^d$ the function which satisfies for all $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ that

$$\mathfrak{M}_{\psi,d}(x) = (\psi(x_1), \psi(x_2), \dots, \psi(x_d)). \tag{8}$$

Definition 2.4 (Rectifier function)**.** We denote by $\mathfrak{r} \colon \mathbb{R} \to \mathbb{R}$ the function which satisfies for all $x \in \mathbb{R}$ that

$$\mathfrak{r}(x) = \max\{x, 0\}. \tag{9}$$

Definition 2.5 (Multidimensional rectifier function)**.** Let $d \in \mathbb{N}$. Then we denote by $\mathfrak{R}_d \colon \mathbb{R}^d \to \mathbb{R}^d$ the function given by

$$\mathfrak{R}_d = \mathfrak{M}_{\mathfrak{r},d} \tag{10}$$

(cf. Definitions 2.3 and 2.4).

Definition 2.6 (Clipping function)**.** Let $u \in [-\infty, \infty)$, $v \in (u, \infty]$. Then we denote by $\mathfrak{c}_{u,v} \colon \mathbb{R} \to \mathbb{R}$ the function which satisfies for all $x \in \mathbb{R}$ that

$$\mathfrak{c}_{u,v}(x) = \max\{u, \min\{x, v\}\}. \tag{11}$$

Definition 2.7 (Multidimensional clipping function)**.** Let $d \in \mathbb{N}$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$. Then we denote by $\mathfrak{C}_{u,v,d} \colon \mathbb{R}^d \to \mathbb{R}^d$ the function given by

$$\mathfrak{C}_{u,v,d} = \mathfrak{M}_{\mathfrak{c}_{u,v},d} \tag{12}$$

(cf. Definitions 2.3 and 2.6).

2.1.4 Rectified DNNs

Definition 2.8 (Rectified clipped DNN)**.** Let $L, d \in \mathbb{N}$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$, $\mathbf{l} = (l_0, l_1, \ldots, l_L) \in \mathbb{N}^{L+1}$, $\theta \in \mathbb{R}^d$ satisfy

$$d \geq \sum_{k=1}^{L} l_k(l_{k-1} + 1). \tag{13}$$

Then we denote by $\mathscr{N}_{u,v}^{\theta,\mathbf{l}} \colon \mathbb{R}^{l_0} \to \mathbb{R}^{l_L}$ the function which satisfies for all $x \in \mathbb{R}^{l_0}$ that

$$\mathscr{N}_{u,v}^{\theta,\mathbf{l}}(x) = \begin{cases} \big(\mathcal{N}^{\theta,0,l_0}_{\mathfrak{C}_{u,v,l_L}}\big)(x) & : L = 1 \\ \big(\mathcal{N}^{\theta,0,l_0}_{\mathfrak{R}_{l_1},\mathfrak{R}_{l_2},\ldots,\mathfrak{R}_{l_{L-1}},\mathfrak{C}_{u,v,l_L}}\big)(x) & : L > 1 \end{cases} \tag{14}$$

(cf. Definitions 2.5, 2.7, and 2.2).

2.1.5 Local Lipschitz continuity of the parametrization function

Definition 2.9 (Maximum norm)**.** We denote by $|||\cdot||| \colon \big(\bigcup_{d=1}^{\infty} \mathbb{R}^d\big) \to [0, \infty)$ the function which satisfies for all $d \in \mathbb{N}$, $\theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \mathbb{R}^d$ that

$$|||\theta||| = \max_{i \in \{1,2,\ldots,d\}} |\theta_i|. \tag{15}$$

Theorem 2.10. *Let $a \in \mathbb{R}$, $b \in (a, \infty)$, $d, L \in \mathbb{N}$, $l = (l_0, l_1, \ldots, l_L) \in \mathbb{N}^{L+1}$ satisfy*

$$d \geq \sum_{k=1}^{L} l_k(l_{k-1} + 1). \tag{16}$$

Then it holds for all $\theta, \vartheta \in \mathbb{R}^d$ that

$$\sup_{x \in [a,b]^{l_0}} |||\mathscr{N}_{-\infty,\infty}^{\theta,l}(x) - \mathscr{N}_{-\infty,\infty}^{\vartheta,l}(x)||| \leq L \max\{1, |a|, |b|\} \left(|||l||| + 1\right)^L \left(\max\{|||\theta|||, |||\vartheta|||\}\right)^{L-1} |||\theta - \vartheta||| \tag{17}$$

(cf. Definition 2.8 and Definition 2.9).

Lemma 2.11. *Let $d \in \mathbb{N}$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$. Then it holds for all $x, y \in \mathbb{R}^d$ that*

$$|||\mathfrak{C}_{u,v,d}(x) - \mathfrak{C}_{u,v,d}(y)||| \leq |||x - y||| \tag{18}$$

(cf. Definitions 2.7 and 2.9).

Proof of Lemma 2.11. First, note that for all $x, y \in \mathbb{R}$ it holds that

$$|\mathfrak{c}_{u,v}(x) - \mathfrak{c}_{u,v}(y)| \leq |x - y| \tag{19}$$

(cf. Definition 2.6). Hence, we obtain that for all $x = (x_1, x_2, \ldots, x_d), y = (y_1, y_2, \ldots, y_d) \in \mathbb{R}^d$ it holds that

$$|||\mathfrak{C}_{u,v,d}(x) - \mathfrak{C}_{u,v,d}(y)||| = \max_{i \in \{1,2,\ldots,d\}} |\mathfrak{c}_{u,v}(x_i) - \mathfrak{c}_{u,v}(y_i)| \leq \max_{i \in \{1,2,\ldots,d\}} |x_i - y_i| = |||x - y||| \tag{20}$$

(cf. Definitions 2.7 and 2.9). This completes the proof of Lemma 2.11. ☐

Corollary 2.12. *Let $a \in \mathbb{R}$, $b \in (a, \infty)$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$, $d, L \in \mathbb{N}$, $l = (l_0, l_1, \ldots, l_L) \in \mathbb{N}^{L+1}$ satisfy*

$$d \geq \sum_{k=1}^{L} l_k(l_{k-1} + 1). \tag{21}$$

Then it holds for all $\theta, \vartheta \in \mathbb{R}^d$ that

$$\sup_{x \in [a,b]^{l_0}} |||\mathscr{N}_{u,v}^{\theta,l}(x) - \mathscr{N}_{u,v}^{\vartheta,l}(x)||| \leq L \max\{1, |a|, |b|\} \left(|||l||| + 1\right)^L \left(\max\{|||\theta|||, |||\vartheta|||\}\right)^{L-1} |||\theta - \vartheta||| \tag{22}$$

(cf. Definitions 2.8 and 2.9).

Proof of Corollary 2.12. Observe that Theorem 2.10 and Lemma 2.11 demonstrate that for all $\theta, \vartheta \in \mathbb{R}^d$ it holds that

$$\begin{aligned}\sup_{x\in[a,b]^{l_0}} |||\mathscr{N}^{\theta,l}_{u,v}(x) - \mathscr{N}^{\vartheta,l}_{u,v}(x)||| &= \sup_{x\in[a,b]^{l_0}} |||\mathfrak{C}_{u,v,l_L}(\mathscr{N}^{\theta,l}_{-\infty,\infty}(x)) - \mathfrak{C}_{u,v,l_L}(\mathscr{N}^{\vartheta,l}_{-\infty,\infty}(x))||| \\ &\le \sup_{x\in[a,b]^{l_0}} |||\mathscr{N}^{\theta,l}_{-\infty,\infty}(x) - \mathscr{N}^{\vartheta,l}_{-\infty,\infty}(x)||| \\ &\le L\max\{1,|a|,|b|\}\,(|||l|||+1)^L\,(\max\{|||\theta|||,|||\vartheta|||\})^{L-1}\,|||\theta-\vartheta|||\end{aligned} \tag{23}$$

(cf. Definitions 2.8, 2.9, and 2.7). This completes the proof of Corollary 2.12. □

2.2 Structured description of DNNs

2.2.1 Structured description of DNNs

Definition 2.13. We denote by $\mathbf{N}$ the set given by

$$\mathbf{N} = \textstyle\bigcup_{L\in\mathbb{N}} \bigcup_{(l_0,l_1,\ldots,l_L)\in\mathbb{N}^{L+1}} \left(\bigtimes_{k=1}^{L}(\mathbb{R}^{l_k\times l_{k-1}} \times \mathbb{R}^{l_k})\right) \tag{24}$$

and we denote by $\mathcal{P},\mathcal{L},\mathcal{I},\mathcal{O}\colon \mathbf{N}\to\mathbb{N}$, $\mathcal{H}\colon\mathbf{N}\to\mathbb{N}_0$, and $\mathcal{D}\colon\mathbf{N}\to\left(\bigcup_{L=2}^{\infty}\mathbb{N}^L\right)$ the functions which satisfy for all $L\in\mathbb{N}$, $l_0,l_1,\ldots,l_L\in\mathbb{N}$, $\Phi\in\left(\bigtimes_{k=1}^{L}(\mathbb{R}^{l_k\times l_{k-1}}\times\mathbb{R}^{l_k})\right)$ that $\mathcal{P}(\Phi)=\sum_{k=1}^{L} l_k(l_{k-1}+1)$, $\mathcal{L}(\Phi)=L$, $\mathcal{I}(\Phi)=l_0$, $\mathcal{O}(\Phi)=l_L$, $\mathcal{H}(\Phi)=L-1$, and $\mathcal{D}(\Phi)=(l_0,l_1,\ldots,l_L)$.

2.2.2 Realizations of DNNs

Definition 2.14 (Realization associated to a DNN)**.** Let $a\in C(\mathbb{R},\mathbb{R})$. Then we denote by $\mathcal{R}_a\colon\mathbf{N}\to\left(\bigcup_{k,l\in\mathbb{N}} C(\mathbb{R}^k,\mathbb{R}^l)\right)$ the function which satisfies for all $L\in\mathbb{N}$, $l_0,l_1,\ldots,l_L\in\mathbb{N}$, $\Phi=((W_1,B_1),(W_2,B_2),\ldots,(W_L,B_L))\in\left(\bigtimes_{k=1}^{L}(\mathbb{R}^{l_k\times l_{k-1}}\times\mathbb{R}^{l_k})\right)$, $x_0\in\mathbb{R}^{l_0}, x_1\in\mathbb{R}^{l_1},\ldots,x_{L-1}\in\mathbb{R}^{l_{L-1}}$ with $\forall\, k\in\mathbb{N}\cap(0,L)\colon x_k=\mathfrak{M}_{a,l_k}(W_k x_{k-1}+B_k)$ that

$$\mathcal{R}_a(\Phi)\in C(\mathbb{R}^{l_0},\mathbb{R}^{l_L}) \qquad\text{and}\qquad (\mathcal{R}_a(\Phi))(x_0)=W_L x_{L-1}+B_L \tag{25}$$

(cf. Definitions 2.13 and 2.3).

Definition 2.15 (Linear transformations as DNNs)**.** Let $m,n\in\mathbb{N}$, $W\in\mathbb{R}^{m\times n}$. Then we denote by $\mathfrak{N}_W\in\mathbb{R}^{m\times n}\times\mathbb{R}^m$ the pair given by $\mathfrak{N}_W=(W,0)$.

2.2.3 Compositions of DNNs

Definition 2.16 (Composition of DNNs)**.** We denote by $(\cdot)\bullet(\cdot)\colon\{(\Phi_1,\Phi_2)\in\mathbf{N}\times\mathbf{N}\colon\mathcal{I}(\Phi_1)=\mathcal{O}(\Phi_2)\}\to\mathbf{N}$ the function which satisfies for all $L,\mathfrak{L}\in\mathbb{N}$, $l_0,l_1,\ldots,l_L,\mathfrak{l}_0,\mathfrak{l}_1,\ldots,\mathfrak{l}_{\mathfrak{L}}\in\mathbb{N}$, $\Phi_1=((W_1,B_1),(W_2,B_2),\ldots,(W_L,B_L))\in\left(\bigtimes_{k=1}^{L}(\mathbb{R}^{l_k\times l_{k-1}}\times\mathbb{R}^{l_k})\right)$, $\Phi_2=((\mathfrak{W}_1,\mathfrak{B}_1),(\mathfrak{W}_2,\mathfrak{B}_2),\ldots,(\mathfrak{W}_{\mathfrak{L}},\mathfrak{B}_{\mathfrak{L}}))\in\left(\bigtimes_{k=1}^{\mathfrak{L}}(\mathbb{R}^{\mathfrak{l}_k\times\mathfrak{l}_{k-1}}\times\mathbb{R}^{\mathfrak{l}_k})\right)$ with $l_0=\mathcal{I}(\Phi_1)=\mathcal{O}(\Phi_2)=\mathfrak{l}_{\mathfrak{L}}$ that

$$\Phi_1\bullet\Phi_2 = \begin{cases}\begin{array}{r}\big((\mathfrak{W}_1,\mathfrak{B}_1),(\mathfrak{W}_2,\mathfrak{B}_2),\ldots,(\mathfrak{W}_{\mathfrak{L}-1},\mathfrak{B}_{\mathfrak{L}-1}),(W_1\mathfrak{W}_{\mathfrak{L}},W_1\mathfrak{B}_{\mathfrak{L}}+B_1),\\ (W_2,B_2),(W_3,B_3),\ldots,(W_L,B_L)\big)\end{array} & : L>1<\mathfrak{L}\\ \big((W_1\mathfrak{W}_1,W_1\mathfrak{B}_1+B_1),(W_2,B_2),(W_3,B_3),\ldots,(W_L,B_L)\big) & : L>1=\mathfrak{L}\\ \big((\mathfrak{W}_1,\mathfrak{B}_1),(\mathfrak{W}_2,\mathfrak{B}_2),\ldots,(\mathfrak{W}_{\mathfrak{L}-1},\mathfrak{B}_{\mathfrak{L}-1}),(W_1\mathfrak{W}_{\mathfrak{L}},W_1\mathfrak{B}_{\mathfrak{L}}+B_1)\big) & : L=1<\mathfrak{L}\\ \big((W_1\mathfrak{W}_1,W_1\mathfrak{B}_1+B_1)\big) & : L=1=\mathfrak{L}\end{cases} \tag{26}$$

(cf. Definition 2.13).

2.2.4 Parallelizations of DNNs

Definition 2.17 (Parallelization of DNNs)**.** Let $n \in \mathbb{N}$. Then we denote by

$$\mathbf{P}_n\colon \big\{(\Phi_1,\Phi_2,\dots,\Phi_n)\in\mathbf{N}^n\colon \mathcal{L}(\Phi_1)=\mathcal{L}(\Phi_2)=\ldots=\mathcal{L}(\Phi_n)\big\}\to\mathbf{N} \tag{27}$$

the function which satisfies for all $L\in\mathbb{N}$, $(l_{1,0},l_{1,1},\dots,l_{1,L}),(l_{2,0},l_{2,1},\dots,l_{2,L}),\dots,(l_{n,0},l_{n,1},\dots,l_{n,L})\in\mathbb{N}^{L+1}$, $\Phi_1=((W_{1,1},B_{1,1}),(W_{1,2},B_{1,2}),\dots,(W_{1,L},B_{1,L}))\in\big(\bigtimes_{k=1}^{L}(\mathbb{R}^{l_{1,k}\times l_{1,k-1}}\times\mathbb{R}^{l_{1,k}})\big)$, $\Phi_2=((W_{2,1},B_{2,1}),(W_{2,2},B_{2,2}),\dots,(W_{2,L},B_{2,L}))\in\big(\bigtimes_{k=1}^{L}(\mathbb{R}^{l_{2,k}\times l_{2,k-1}}\times\mathbb{R}^{l_{2,k}})\big)$, $\dots$, $\Phi_n=((W_{n,1},B_{n,1}),(W_{n,2},B_{n,2}),\dots,(W_{n,L},B_{n,L}))\in\big(\bigtimes_{k=1}^{L}(\mathbb{R}^{l_{n,k}\times l_{n,k-1}}\times\mathbb{R}^{l_{n,k}})\big)$ that

$$\begin{aligned}\mathbf{P}_n(\Phi_1,\Phi_2,\dots,\Phi_n)=\Bigg(&\left(\begin{pmatrix}W_{1,1}&0&0&\cdots&0\\0&W_{2,1}&0&\cdots&0\\0&0&W_{3,1}&\cdots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\cdots&W_{n,1}\end{pmatrix},\begin{pmatrix}B_{1,1}\\B_{2,1}\\B_{3,1}\\\vdots\\B_{n,1}\end{pmatrix}\right),\\&\left(\begin{pmatrix}W_{1,2}&0&0&\cdots&0\\0&W_{2,2}&0&\cdots&0\\0&0&W_{3,2}&\cdots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\cdots&W_{n,2}\end{pmatrix},\begin{pmatrix}B_{1,2}\\B_{2,2}\\B_{3,2}\\\vdots\\B_{n,2}\end{pmatrix}\right),\dots,\\&\left(\begin{pmatrix}W_{1,L}&0&0&\cdots&0\\0&W_{2,L}&0&\cdots&0\\0&0&W_{3,L}&\cdots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\cdots&W_{n,L}\end{pmatrix},\begin{pmatrix}B_{1,L}\\B_{2,L}\\B_{3,L}\\\vdots\\B_{n,L}\end{pmatrix}\right)\Bigg)\end{aligned} \tag{28}$$

(cf. Definition 2.13).

2.2.5 On the connection to the vectorized description of DNNs

Definition 2.18. We denote by $\mathcal{T}\colon\mathbf{N}\to\big(\bigcup_{d\in\mathbb{N}}\mathbb{R}^d\big)$ the function which satisfies for all $L,d\in\mathbb{N}$, $l_0,l_1,\dots,l_L\in\mathbb{N}$, $\Phi=((W_1,B_1),(W_2,B_2),\dots,(W_L,B_L))\in\big(\bigtimes_{m=1}^{L}(\mathbb{R}^{l_m\times l_{m-1}}\times\mathbb{R}^{l_m})\big)$, $\theta=(\theta_1,\theta_2,\dots,\theta_d)\in\mathbb{R}^d$, $k\in\{1,2,\dots,L\}$ with $\mathcal{T}(\Phi)=\theta$ that

$$d=\mathcal{P}(\Phi),\qquad B_k=\begin{pmatrix}\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+l_kl_{k-1}+1}\\\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+l_kl_{k-1}+2}\\\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+l_kl_{k-1}+3}\\\vdots\\\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+l_kl_{k-1}+l_k}\end{pmatrix},\qquad\text{and}$$

$$W_k=\begin{pmatrix}\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+1}&\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+2}&\cdots&\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+l_{k-1}}\\\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+l_{k-1}+1}&\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+l_{k-1}+2}&\cdots&\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+2l_{k-1}}\\\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+2l_{k-1}+1}&\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+2l_{k-1}+2}&\cdots&\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+3l_{k-1}}\\\vdots&\vdots&\ddots&\vdots\\\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+(l_k-1)l_{k-1}+1}&\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+(l_k-1)l_{k-1}+2}&\cdots&\theta_{(\sum_{i=1}^{k-1}l_i(l_{i-1}+1))+l_kl_{k-1}}\end{pmatrix}, \tag{29}$$

(cf. Definition 2.13).

Lemma 2.19. *Let $a, b \in \mathbb{N}$, $W = (W_{i,j})_{(i,j)\in\{1,2,\ldots,a\}\times\{1,2,\ldots,b\}} \in \mathbb{R}^{a\times b}$, $B = (B_i)_{i\in\{1,2,\ldots,b\}} \in \mathbb{R}^b$. Then*

$$\mathcal{T}\big(((W,B))\big) = \big(W_{1,1}, W_{1,2}, \ldots, W_{1,b}, W_{2,1}, W_{2,2}, \ldots, W_{2,b}, \ldots, W_{a,1}, W_{a,2}, \ldots, W_{a,b}, B_1, B_2, \ldots, B_b\big) \quad (30)$$

(cf. Definition 2.18).

Proof of Lemma 2.19. Observe that (29) establishes (30). The proof of Lemma 2.19 is thus completed. □

Lemma 2.20. *Let $L \in \mathbb{N}$, $l_0, l_1, \ldots, l_L \in \mathbb{N}$, let $W_k = (W_{k,i,j})_{(i,j)\in\{1,2,\ldots,l_k\}\times\{1,2,\ldots,l_{k-1}\}} \in \mathbb{R}^{l_k\times l_{k-1}}$, $k \in \{1, 2, \ldots, L\}$, and let $B_k = (B_{k,i})_{i\in\{1,2,\ldots,l_k\}} \in \mathbb{R}^{l_k}$, $k \in \{1, 2, \ldots, L\}$. Then*

(i) it holds for all $k \in \{1, 2, \ldots, L\}$ that

$$\mathcal{T}\big(((W_k, B_k))\big) = \big(W_{k,1,1}, W_{k,1,2}, \ldots, W_{k,1,l_{k-1}}, W_{k,2,1}, W_{k,2,2}, \ldots, W_{k,2,l_{k-1}}, \ldots, W_{k,l_k,1}, W_{k,l_k,2}, \ldots, W_{k,l_k,l_{k-1}}, B_{k,1}, B_{k,2}, \ldots, B_{k,l_k}\big) \quad (31)$$

and

(ii) it holds that

$$\begin{aligned}
&\mathcal{T}\Big(\big((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L)\big)\Big) \\
&= \Big(W_{1,1,1}, W_{1,1,2}, \ldots, W_{1,1,l_0}, \ldots, W_{1,l_1,1}, W_{1,l_1,2}, \ldots, W_{1,l_1,l_0}, B_{1,1}, B_{1,2}, \ldots, B_{1,l_1}, \\
&\quad W_{2,1,1}, W_{2,1,2}, \ldots, W_{2,1,l_1}, \ldots, W_{2,l_2,1}, W_{2,l_2,2}, \ldots, W_{2,l_2,l_1}, B_{2,1}, B_{2,2}, \ldots, B_{2,l_2}, \\
&\quad \ldots, \\
&\quad W_{L,1,1}, W_{L,1,2}, \ldots, W_{L,1,l_{L-1}}, \ldots W_{L,l_L,1}, W_{L,l_L,2}, \ldots, W_{L,l_L,l_{L-1}}, B_{L,1}, B_{L,2}, \ldots, B_{L,l_L}\Big)
\end{aligned} \quad (32)$$

(cf. Definition 2.18).

Proof of Lemma 2.20. Note that Lemma 2.19 proves item (i). Moreover, observe that (29) establishes item (ii). The proof of Lemma 2.20 is thus completed. □

Lemma 2.21. *Let $a \in C(\mathbb{R}, \mathbb{R})$, $\Phi \in \mathbf{N}$, $L \in \mathbb{N}$, $l_0, l_1, \ldots, l_L \in \mathbb{N}$ satisfy $\mathcal{D}(\Phi) = (l_0, l_1, \ldots, l_L)$ (cf. Definition 2.13). Then it holds for all $x \in \mathbb{R}^{l_0}$ that*

$$(\mathcal{R}_a(\Phi))(x) = \begin{cases} \big(\mathcal{N}^{\mathcal{T}(\Phi),0,l_0}_{\mathrm{id}_{\mathbb{R}^{l_L}}}\big)(x) & : L = 1 \\ \big(\mathcal{N}^{\mathcal{T}(\Phi),0,l_0}_{\mathfrak{M}_{a,l_1},\mathfrak{M}_{a,l_2},\ldots,\mathfrak{M}_{a,l_{L-1}},\mathrm{id}_{\mathbb{R}^{l_L}}}\big)(x) & : L > 1 \end{cases} \quad (33)$$

(cf. Definitions 2.14, 2.18, 2.3, and 2.2).

Proof of Lemma 2.21. Throughout this proof let $((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L)) \in \big(\times_{k=1}^{L}(\mathbb{R}^{l_k\times l_{k-1}} \times \mathbb{R}^{l_k})\big)$ satisfy $\Phi = ((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L))$. Note that (29) shows that for all $k \in \{1, 2, \ldots, L\}$, $x \in \mathbb{R}^{l_{k-1}}$ it holds that

$$W_k x + B_k = \big(\mathcal{A}^{\mathcal{T}(\Phi),\sum_{i=1}^{k-1} l_i(l_{i-1}+1)}_{l_{k-1},l_k}\big)(x) \quad (34)$$

(cf. Definitions 2.18 and 2.1). This demonstrates that for all $x_0 \in \mathbb{R}^{l_0}$, $x_1 \in \mathbb{R}^{l_1}, \ldots, x_{L-1} \in \mathbb{R}^{l_{L-1}}$ with $\forall\, k \in \mathbb{N} \cap (0, L)\colon x_k = \mathfrak{M}_{a,l_k}(W_k x_{k-1} + B_k)$ it holds that

$$x_{L-1} = \begin{cases} x_0 & : L = 1 \\ \big(\mathfrak{M}_{a,l_{L-1}} \circ \mathcal{A}^{\mathcal{T}(\Phi),\sum_{i=1}^{L-2} l_i(l_{i-1}+1)}_{l_{L-2},l_{L-1}} \circ \mathfrak{M}_{a,l_{L-2}} \circ \mathcal{A}^{\mathcal{T}(\Phi),\sum_{i=1}^{L-3} l_i(l_{i-1}+1)}_{l_{L-3},l_{L-2}} \circ \ldots \\ \quad \ldots \circ \mathfrak{M}_{a,l_1} \circ \mathcal{A}^{\mathcal{T}(\Phi),0}_{l_0,l_1}\big)(x_0) & : L > 1 \end{cases} \quad (35)$$

(cf. Definition 2.3). Combining this and (34) with (10), (7), and (25) proves that for all $x_0 \in \mathbb{R}^{l_0}$, $x_1 \in \mathbb{R}^{l_1}, \ldots, x_{L-1} \in \mathbb{R}^{l_{L-1}}$ with $\forall\, k \in \mathbb{N} \cap (0, L)\colon x_k = \mathfrak{M}_{a,l_k}(W_k x_{k-1} + B_k)$ it holds that

$$\begin{split}\big(\mathcal{R}_a(\Phi)\big)(x_0) = W_L x_{L-1} + B_L &= \big(\mathcal{A}_{l_{L-1},l_L}^{\mathcal{T}(\Phi),\sum_{i=1}^{L-1} l_i(l_{i-1}+1)}\big)(x_{L-1}) \\ &= \begin{cases} \big(\mathcal{N}_{\mathrm{id}_{\mathbb{R}^{l_L}}}^{\mathcal{T}(\Phi),0,l_0}\big)(x_0) & : L = 1 \\ \big(\mathcal{N}_{\mathfrak{M}_{a,l_1},\mathfrak{M}_{a,l_2},\ldots,\mathfrak{M}_{a,l_{L-1}},\mathrm{id}_{\mathbb{R}^{l_L}}}^{\mathcal{T}(\Phi),0,l_0}\big)(x_0) & : L > 1 \end{cases}\end{split} \tag{36}$$

(cf. Definitions 2.14, 2.3, and 2.2). The proof of Lemma 2.21 is thus completed. □

Corollary 2.22. *Let $\Phi \in \mathbf{N}$ (cf. Definition 2.13). Then it holds for all $x \in \mathbb{R}^{\mathcal{I}(\Phi)}$ that*

$$\big(\mathcal{N}_{-\infty,\infty}^{\mathcal{T}(\Phi),\mathcal{D}(\Phi)}\big)(x) = (\mathcal{R}_{\mathfrak{r}}(\Phi))(x) \tag{37}$$

(cf. Definitions 2.18, 2.8, and 2.14).

Proof of Corollary 2.22. Note that Lemma 2.21, (14), (10), and the fact that for all $d \in \mathbb{N}$ it holds that $\mathfrak{C}_{-\infty,\infty,d} = \mathrm{id}_{\mathbb{R}^d}$ establish (37). The proof of Corollary 2.22 is thus completed. □

2.2.6 Embedding DNNs in larger architectures

Lemma 2.23. *Let $a \in C(\mathbb{R},\mathbb{R})$, $L \in \mathbb{N}$, $l_0, l_1, \ldots, l_L, \mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_L \in \mathbb{N}$ satisfy for all $k \in \{1, 2, \ldots, L\}$ that $\mathfrak{l}_0 = l_0$, $\mathfrak{l}_L = l_L$, and $\mathfrak{l}_k \geq l_k$, for every $k \in \{1, 2, \ldots, L\}$ let $W_k = (W_{k,i,j})_{(i,j)\in\{1,2,\ldots,l_k\}\times\{1,2,\ldots,l_{k-1}\}} \in \mathbb{R}^{l_k \times l_{k-1}}$, $\mathfrak{W}_k = (\mathfrak{W}_{k,i,j})_{(i,j)\in\{1,2,\ldots,\mathfrak{l}_k\}\times\{1,2,\ldots,\mathfrak{l}_{k-1}\}} \in \mathbb{R}^{\mathfrak{l}_k \times \mathfrak{l}_{k-1}}$, $B_k = (B_{k,i})_{i\in\{1,2,\ldots,l_k\}} \in \mathbb{R}^{l_k}$, $\mathfrak{B}_k = (\mathfrak{B}_{k,i})_{i\in\{1,2,\ldots,\mathfrak{l}_k\}} \in \mathbb{R}^{\mathfrak{l}_k}$, assume for all $k \in \{1, 2, \ldots, L\}$, $i \in \{1, 2, \ldots, l_k\}$, $j \in \mathbb{N} \cap (0, l_{k-1}]$ that $\mathfrak{W}_{k,i,j} = W_{k,i,j}$ and $\mathfrak{B}_{k,i} = B_{k,i}$, and assume for all $k \in \{1, 2, \ldots, L\}$, $i \in \{1, 2, \ldots, l_k\}$, $j \in \mathbb{N} \cap (l_{k-1}, \mathfrak{l}_{k-1} + 1)$ that $\mathfrak{W}_{k,i,j} = 0$. Then*

$$\mathcal{R}_a\big(((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L))\big) = \mathcal{R}_a\big(((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \ldots, (\mathfrak{W}_L, \mathfrak{B}_L))\big) \tag{38}$$

(cf. Definition 2.14).

Proof of Lemma 2.23. Throughout this proof let $\pi_k\colon \mathbb{R}^{\mathfrak{l}_k} \to \mathbb{R}^{l_k}$, $k \in \{0, 1, \ldots, L\}$, satisfy for all $k \in \{0, 1, \ldots, L\}$, $x = (x_1, x_2, \ldots, x_{\mathfrak{l}_k})$ that

$$\pi_k(x) = (x_1, x_2, \ldots, x_{l_k}). \tag{39}$$

Observe that the hypothesis that $\mathfrak{l}_0 = l_0$ and $\mathfrak{l}_L = l_L$ shows that

$$\mathcal{R}_a\big(((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L))\big) \in C(\mathbb{R}^{\mathfrak{l}_0}, \mathbb{R}^{\mathfrak{l}_L}) \tag{40}$$

(cf. Definition 2.14). Furthermore, note that the hypothesis that for all $k \in \{1, 2, \ldots, l\}$, $i \in \{1, 2, \ldots, l_k\}$, $j \in \mathbb{N} \cap (l_{k-1}, \mathfrak{l}_{k-1}+1)$ it holds that $\mathfrak{W}_{k,i,j} = 0$ ensures that for all $k \in \{1, 2, \ldots, L\}$, $x = (x_1, x_2, \ldots, x_{\mathfrak{l}_{k-1}}) \in \mathbb{R}^{\mathfrak{l}_{k-1}}$ it holds that

$$\begin{split}\pi_k(\mathfrak{W}_k x + \mathfrak{B}_k) &= \left(\left[\sum_{i=1}^{\mathfrak{l}_{k-1}} \mathfrak{W}_{k,1,i} x_i\right] + \mathfrak{B}_{k,1}, \left[\sum_{i=1}^{\mathfrak{l}_{k-1}} \mathfrak{W}_{k,2,i} x_i\right] + \mathfrak{B}_{k,2}, \ldots, \left[\sum_{i=1}^{\mathfrak{l}_{k-1}} \mathfrak{W}_{k,l_k,i} x_i\right] + \mathfrak{B}_{k,l_k}\right) \\ &= \left(\left[\sum_{i=1}^{l_{k-1}} \mathfrak{W}_{k,1,i} x_i\right] + \mathfrak{B}_{k,1}, \left[\sum_{i=1}^{l_{k-1}} \mathfrak{W}_{k,2,i} x_i\right] + \mathfrak{B}_{k,2}, \ldots, \left[\sum_{i=1}^{l_{k-1}} \mathfrak{W}_{k,l_k,i} x_i\right] + \mathfrak{B}_{k,l_k}\right).\end{split} \tag{41}$$

Combining this with the hypothesis that for all $k \in \{1, 2, \ldots, L\}$, $i \in \{1, 2, \ldots, l_k\}$, $j \in \mathbb{N} \cap (0, l_{k-1}]$ that $\mathfrak{W}_{k,i,j} = W_{k,i,j}$ and $\mathfrak{B}_{k,i} = B_{k,i}$ shows that for all $k \in \{1, 2, \ldots, L\}$, $x = (x_1, x_2, \ldots, x_{\mathfrak{l}_{k-1}}) \in \mathbb{R}^{\mathfrak{l}_{k-1}}$ it holds that

$$\begin{split}\pi_k(\mathfrak{W}_k x + \mathfrak{B}_k) &= \left(\left[\sum_{i=1}^{l_{k-1}} W_{k,1,i} x_i\right] + B_{k,1}, \left[\sum_{i=1}^{l_{k-1}} W_{k,2,i} x_i\right] + B_{k,2}, \ldots, \left[\sum_{i=1}^{l_{k-1}} W_{k,l_k,i} x_i\right] + B_{k,l_k}\right) \\ &= W_k \pi_{k-1}(x) + B_k.\end{split} \tag{42}$$

Hence, we obtain that for all $x_0 \in \mathbb{R}^{\mathfrak{l}_0}$, $x_1 \in \mathbb{R}^{\mathfrak{l}_1}, \ldots, x_{L-1} \in \mathbb{R}^{\mathfrak{l}_{L-1}}$, $k \in \mathbb{N} \cap (0, L)$ with $\forall\, m \in \mathbb{N} \cap (0, L)\colon x_m = \mathfrak{M}_{a,\mathfrak{l}_m}(\mathfrak{W}_m x_{m-1} + \mathfrak{B}_m)$ it holds that

$$\pi_k(x_k) = \mathfrak{M}_{a,l_k}(\pi_k(\mathfrak{W}_k x_{k-1} + \mathfrak{B}_k)) = \mathfrak{M}_{a,l_k}(W_k \pi_{k-1}(x_{k-1}) + B_k) \tag{43}$$

(cf. Definition 2.3). Induction, the hypothesis that $l_0 = \mathfrak{l}_0$ and $l_L = \mathfrak{l}_L$, and (42) therefore prove that for all $x_0 \in \mathbb{R}^{\mathfrak{l}_0}$, $x_1 \in \mathbb{R}^{\mathfrak{l}_1}, \ldots, x_{L-1} \in \mathbb{R}^{\mathfrak{l}_{L-1}}$ with $\forall\, k \in \mathbb{N} \cap (0, L)\colon x_k = \mathfrak{M}_{a,\mathfrak{l}_k}(\mathfrak{W}_k x_{k-1} + \mathfrak{B}_k)$ it holds that

$$\begin{aligned}
\big(\mathcal{R}_a\big(((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L))\big)\big)(x_0) &= \big(\mathcal{R}_a\big(((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L))\big)\big)(\pi_0(x_0)) \\
&= W_L \pi_{L-1}(x_{L-1}) + B_L \\
&= \pi_L(\mathfrak{W}_L x_{L-1} + \mathfrak{B}_L) = \mathfrak{W}_L x_{L-1} + \mathfrak{B}_L \\
&= \big(\mathcal{R}_a\big(((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \ldots, (\mathfrak{W}_L, \mathfrak{B}_L))\big)\big)(x_0)
\end{aligned} \tag{44}$$

(cf. Definition 2.14). The proof of Lemma 2.23 is thus completed. □

Lemma 2.24. *Let $u \in [-\infty, \infty)$, $v \in (u, \infty]$, $L, d, \mathfrak{d} \in \mathbb{N}$, $l_0, l_1, \ldots, l_L, \mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_L \in \mathbb{N}$, $\theta \in \mathbb{R}^d$, assume for all $k \in \{1, 2, \ldots, L\}$ that $\mathfrak{l}_0 = l_0$, $\mathfrak{l}_L = l_L$, and $\mathfrak{l}_k \geq l_k$, and assume that*

$$d \geq \sum_{k=1}^{L} l_k(l_{k-1} + 1) \qquad \text{and} \qquad \mathfrak{d} \geq \sum_{k=1}^{L} \mathfrak{l}_k(\mathfrak{l}_{k-1} + 1). \tag{45}$$

Then there exists $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ such that $|||\vartheta||| = |||\theta|||$ and

$$\mathscr{N}_{u,v}^{\vartheta,(\mathfrak{l}_0,\mathfrak{l}_1,\ldots,\mathfrak{l}_L)} = \mathscr{N}_{u,v}^{\theta,(l_0,l_1,\ldots,l_L)} \tag{46}$$

(cf. Definitions 2.9 and 2.8).

Proof of Lemma 2.24. Throughout this proof let $B_k = (B_{k,i})_{i \in \{1,2,\ldots,l_k\}} \in \mathbb{R}^{l_k}$, $k \in \{1, 2, \ldots, L\}$, and $W_k = (W_{k,i,j})_{(i,j) \in \{1,2,\ldots,l_k\} \times \{1,2,\ldots,l_{k-1}\}} \in \mathbb{R}^{l_k \times l_{k-1}}$, $k \in \{1, 2, \ldots, L\}$, satisfy $\mathcal{T}\big(((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L))\big) = \theta$ and for every $k \in \{1, 2, \ldots, L\}$ let $\mathfrak{W}_k = (\mathfrak{W}_{k,i,j})_{(i,j) \in \{1,2,\ldots,\mathfrak{l}_k\} \times \{1,2,\ldots,\mathfrak{l}_{k-1}\}} \in \mathbb{R}^{\mathfrak{l}_k \times \mathfrak{l}_{k-1}}$, $\mathfrak{B}_k = (\mathfrak{B}_{k,i})_{i \in \{1,2,\ldots,\mathfrak{l}_k\}} \in \mathbb{R}^{\mathfrak{l}_k}$ satisfy for all $i \in \{1, 2, \ldots, \mathfrak{l}_k\}$, $j \in \{1, 2, \ldots, \mathfrak{l}_{k-1}\}$ that

$$\mathfrak{W}_{k,i,j} = \begin{cases} W_{k,i,j} & : (i \leq l_k) \wedge (j \leq l_{k-1}) \\ 0 & : (i > l_k) \vee (j > l_{k-1}) \end{cases} \qquad \text{and} \qquad \mathfrak{B}_{k,i} = \begin{cases} B_{k,i} & : i \leq l_k \\ 0 & : i > l_k \end{cases} \tag{47}$$

(cf. Definition 2.18). Observe that Lemma 2.23 shows that

$$\mathcal{R}_{\mathfrak{r}}\big(((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L))\big) = \mathcal{R}_{\mathfrak{r}}\big(((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \ldots, (\mathfrak{W}_L, \mathfrak{B}_L))\big) \tag{48}$$

(cf. Definitions 2.4 and 2.14). In the next step let $\vartheta = (\vartheta_1, \vartheta_2, \ldots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\eta \in \mathbb{R}^{\sum_{k=1}^{L} \mathfrak{l}_k(\mathfrak{l}_{k-1}+1)}$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$ with $\eta = \mathcal{T}\big(((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \ldots, (\mathfrak{W}_L, \mathfrak{B}_L))\big)$ that

$$\vartheta_i = \begin{cases} \eta_i & : i \leq \sum_{k=1}^{L} \mathfrak{l}_k(\mathfrak{l}_{k-1} + 1) \\ 0 & : i > \sum_{k=1}^{L} \mathfrak{l}_k(\mathfrak{l}_{k-1} + 1). \end{cases} \tag{49}$$

Note that (48) and Corollary 2.22 demonstrate that for all $x \in \mathbb{R}^{l_0}$ it holds that

$$\big(\mathscr{N}_{-\infty,\infty}^{\theta,(l_0,l_1,\ldots,l_L)}\big)(x) = \big(\mathscr{N}_{-\infty,\infty}^{\vartheta,(\mathfrak{l}_0,\mathfrak{l}_1,\ldots,\mathfrak{l}_L)}\big)(x) \tag{50}$$

(cf. Definition 2.8). Hence, we obtain that for all $x \in \mathbb{R}^{l_0}$ it holds that

$$\big(\mathscr{N}_{u,v}^{\theta,(l_0,l_1,\ldots,l_L)}\big)(x) = \mathfrak{C}_{u,v,l_L}\big(\big(\mathscr{N}_{-\infty,\infty}^{\theta,(l_0,l_1,\ldots,l_L)}\big)(x)\big) = \mathfrak{C}_{u,v,\mathfrak{l}_L}\big(\big(\mathscr{N}_{-\infty,\infty}^{\vartheta,(\mathfrak{l}_0,\mathfrak{l}_1,\ldots,\mathfrak{l}_L)}\big)(x)\big) = \big(\mathscr{N}_{u,v}^{\vartheta,(\mathfrak{l}_0,\mathfrak{l}_1,\ldots,\mathfrak{l}_L)}\big)(x) \tag{51}$$

(cf. Definition 2.7). Furthermore, observe that (47), (49), and Lemma 2.20 imply that

$$|||\vartheta||| = |||\theta||| \tag{52}$$

(cf. Definition 2.9). The proof of Lemma 2.24 is thus completed. □

3 Separate analyses of the error sources

In this section we study separately the approximation error (see Section 3.1 below), the generalization error (see Section 3.2 below), and the optimization error (see Section 3.3 below).

In particular, the main result in Section 3.1, Proposition 3.6 below, establishes an upper bound for the error in the approximation of a Lipschitz continuous function by DNNs. This approximation result is obtained by combining the essentially well-known approximation result in Lemma 3.1 with the DNN calculus in Section 2.2 above (cf., e.g., Grohs et al. [31, 32]). Some of the concepts and results in Section 3.1 are partially based on material in publications from the scientific literature. In particular, Definition 3.2 is [32, Definition 3.15] and the elementary results in Lemma 3.3 are basically well-known in the scientific literature. For further approximation results for DNNs we refer, e.g., to [1, 3, 4, 11, 12, 13, 14, 16, 17, 19, 21, 22, 23, 24, 25, 27, 29, 30, 31, 33, 34, 36, 38, 39, 40, 41, 42, 44, 47, 49, 52, 53, 54, 55, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 69, 70, 72, 73, 74] and the references mentioned therein.

In Lemmas 3.19 and 3.20 in Section 3.2 below we study the generalization error. Our analysis in Section 3.2 is in parts inspired by Berner et al. [10] and Cucker & Smale [18]. Proposition 3.11 in Section 3.2.1 is known as Hoeffding's inequality in the scientific literature and Proposition 3.11 is, e.g., proved as Theorem 2 in Hoeffding [37]. For further results on the generalization error we refer, e.g., to [5, 35, 51, 68, 71] and the references mentioned therein.

In the two elementary results in Section 3.3, Lemmas 3.21 and 3.22, we study the optimization error of the minimum Monte Carlo algorithm. A related result can, e.g., be found in [6, Lemma 3.5]. For further results on the optimization error we refer, e.g., to [2, 9, 15, 20, 26, 43, 45, 46, 48] and the references mentioned therein.

3.1 Analysis of the approximation error

3.1.1 Approximation for Lipschitz continuous functions

Lemma 3.1. *Let (E, δ) be a metric space, let $\mathcal{M} \subseteq E$ satisfy $\mathcal{M} \neq \emptyset$, let $L \in [0, \infty)$, let $f \colon E \to \mathbb{R}$ satisfy for all $x \in E$, $y \in \mathcal{M}$ that $|f(x) - f(y)| \leq L\delta(x, y)$, and let $F \colon E \to \mathbb{R} \cup \{\infty\}$ satisfy for all $x \in E$ that*

$$F(x) = \sup_{y \in \mathcal{M}} [f(z) - L\delta(x, y)]. \tag{53}$$

Then

(i) it holds for all $x \in E$ that $F(x) \leq f(x)$,

(ii) it holds for all $x \in \mathcal{M}$ that $F(x) = f(x)$,

(iii) it holds for all $x, y \in E$ that $|F(x) - F(y)| \leq L\delta(x, y)$, and

(iv) it holds for all $x \in E$ that

$$|F(x) - f(x)| \leq 2L \left[\inf_{y \in \mathcal{M}} \delta(x, y) \right]. \tag{54}$$

Proof of Lemma 3.1. First, observe that the hypothesis that for all $x \in E$, $y \in \mathcal{M}$ it holds that $|f(x) - f(y)| \leq L\delta(x, y)$ ensures that for all $x \in E$, $y \in \mathcal{M}$ it holds that

$$f(x) \geq f(y) - L\delta(x, y). \tag{55}$$

Hence, we obtain that for all $x \in E$ it holds that

$$f(x) \geq \sup_{y \in \mathcal{M}} [f(y) - L\delta(x, y)] = F(x). \tag{56}$$

This establishes item (i). Next observe that (53) implies that for all $x \in \mathcal{M}$ it holds that

$$F(x) \geq f(x) - L\delta(x, x) = f(x). \tag{57}$$

Combining this with item (i) establishes item (ii). In the next step we note that for all $x, y \in E$ it holds that

$$\begin{aligned} F(x) - F(y) &= \left[\sup_{v \in \mathcal{M}} (f(v) - L\delta(x, v))\right] - \left[\sup_{w \in \mathcal{M}} (f(w) - L\delta(y, w))\right] \\ &= \sup_{v \in \mathcal{M}} \left[f(v) - L\delta(x, v) - \sup_{w \in \mathcal{M}} (f(w) - L\delta(y, w))\right] \\ &\leq \sup_{v \in \mathcal{M}} \big[f(v) - L\delta(x, v) - (f(v) - L\delta(y, v))\big] \\ &= L\left[\sup_{v \in \mathcal{M}} (\delta(y, v) - \delta(x, v))\right] \\ &\leq L\left[\sup_{v \in \mathcal{M}} (\delta(y, x) + \delta(x, v) - \delta(x, v))\right] = L\delta(x, y). \end{aligned} \tag{58}$$

Combining this with the fact that for all $x, y \in E$ it holds that $\delta(x, y) = \delta(y, x)$ establishes item (iii). Observe that item (ii), the triangle inequality, item (iii), and the hypothesis that for all $x \in E$, $y \in \mathcal{M}$ it holds that $|f(x) - f(y)| \leq L\delta(x, y)$ ensure that for all $x \in E$ it holds that

$$\begin{aligned} |F(x) - f(x)| &= \inf_{y \in \mathcal{M}} |F(x) - F(y) + f(y) - f(x)| \\ &\leq \inf_{y \in \mathcal{M}} (|F(x) - F(y)| + |f(y) - f(x)|) \\ &\leq \inf_{y \in \mathcal{M}} (2L\delta(x, y)) = 2L\left[\inf_{y \in \mathcal{M}} \delta(x, y)\right]. \end{aligned} \tag{59}$$

This establishes item (iv). The proof of Lemma 3.1 is thus completed. □

3.1.2 DNN representations for maxima

Definition 3.2. We denote by $\mathfrak{I} = (\mathfrak{I}_d)_{d \in \mathbb{N}} \colon \mathbb{N} \to \mathbf{N}$ the function which satisfies for all $d \in \mathbb{N}$ that

$$\mathfrak{I}_1 = \left(\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right), \left(\begin{pmatrix} 1 & -1 \end{pmatrix}, 0\right)\right) \in \big((\mathbb{R}^{2\times 1} \times \mathbb{R}^2) \times (\mathbb{R}^{1\times 2} \times \mathbb{R}^1)\big) \tag{60}$$

and

$$\mathfrak{I}_d = \mathbf{P}_d(\mathfrak{I}_1, \mathfrak{I}_1, \ldots, \mathfrak{I}_1) \tag{61}$$

(cf. Definition 2.13 and Definition 2.17).

Lemma 3.3. *Let $\Phi \in \mathbf{N}$ satisfy*

$$\Phi = \left(\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\right), \left(\begin{pmatrix} 1 & 1 & -1 \end{pmatrix}, 0\right)\right) \in \big((\mathbb{R}^{3\times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1\times 3} \times \mathbb{R})\big) \tag{62}$$

(cf. Definition 2.13). Then

- *(i) it holds for all $k \in \mathbb{N}$ that $\mathcal{L}(\mathfrak{I}_k) = 2$,*
- *(ii) there exist unique $\phi_k \in \mathbf{N}$, $k \in \{2, 3, \ldots\}$, which satisfy for all $k \in \{2, 3, \ldots\}$ that $\phi_2 = \Phi$, $\mathcal{I}(\phi_k) = \mathcal{O}(\mathbf{P}_2(\Phi, \mathfrak{I}_{k-1}))$, and*

$$\phi_{k+1} = \phi_k \bullet \big(\mathbf{P}_2(\Phi, \mathfrak{I}_{k-1})\big), \tag{63}$$

(iii) it holds for all $k \in \{2, 3, \dots\}$ that $\mathcal{L}(\phi_k) = k$, and

(iv) it holds for all $k \in \{2, 3, \dots\}$ that $\mathcal{D}(\phi_k) = (k, 2k-1, 2k-3, \dots, 3, 1)$, and

(v) it holds for all $k \in \{2, 3, \dots\}$, $x = (x_1, x_2, \dots, x_k) \in \mathbb{R}^k$ that

$$\big(\mathcal{R}_{\mathfrak{r}}(\phi_k)\big)(x) = \max_{i \in \{1,2,\dots,k\}} x_i \tag{64}$$

(cf. Definitions 3.2, 2.17, 2.16, 2.4, and 2.14).

Proof of Lemma 3.3. First, note that $\mathcal{D}(\mathfrak{I}_1) = (1, 2, 1)$ (cf. Definition 3.2). Item (i) in [31, Proposition 2.20] hence shows that for all $k \in \mathbb{N}$ it holds that

$$\mathcal{D}(\mathfrak{I}_k) = (k, 2k, k). \tag{65}$$

This establishes item (i). Next note that

$$\mathcal{D}(\Phi) = (2, 3, 1). \tag{66}$$

Combining this and (65) with item (i) in [31, Proposition 2.20] shows that for all $k \in \mathbb{N}$ it holds that

$$\mathcal{D}(\mathbf{P}_2(\Phi, \mathfrak{I}_k)) = (k+2, 2k+3, k+1) \tag{67}$$

(cf. Definition 2.17). This implies that for all $k \in \{2, 3, \dots\}$ it holds that

$$\mathcal{D}(\mathbf{P}_2(\Phi, \mathfrak{I}_{k-1})) = (k+1, 2k+1, k). \tag{68}$$

Combining (67) and (68) ensures that for all $k \in \{2, 3, \dots\}$ it holds that

$$\mathcal{O}(\mathbf{P}_2(\Phi, \mathfrak{I}_k)) = k+1 = \mathcal{I}(\mathbf{P}_2(\Phi, \mathfrak{I}_{k-1})). \tag{69}$$

Moreover, note that (62) and (67) show that

$$\mathcal{I}(\Phi) = 2 = \mathcal{O}(\mathbf{P}_2(\Phi, \mathfrak{I}_1)). \tag{70}$$

Furthermore, observe that item (i) in [31, Proposition 2.6] shows that for all $k \in \{2, 3, \dots\}$, $\psi \in \mathbf{N}$ with $\mathcal{I}(\psi) = \mathcal{O}(\mathbf{P}_2(\Phi, \mathfrak{I}_{k-1}))$ it holds that

$$\mathcal{I}\big(\psi \bullet \big(\mathbf{P}_2(\Phi, \mathfrak{I}_{k-1})\big)\big) = \mathcal{I}(\mathbf{P}_2(\Phi, \mathfrak{I}_{k-1})) = k+1 = \mathcal{O}(\mathbf{P}_2(\Phi, \mathfrak{I}_k)) \tag{71}$$

(cf. Definition 2.16). Combining this and (70) with induction establishes item (ii). In the next step we note that (63) and item (ii) in [31, Proposition 2.6] imply that for all $k \in \{2, 3, \dots\}$ it holds that

$$\mathcal{L}(\phi_{k+1}) = \mathcal{L}(\phi_k) + \mathcal{L}(\mathbf{P}_2(\phi_2, \mathfrak{I}_{k-1})) - 1 = \mathcal{L}(\phi_k) + 1. \tag{72}$$

Combining this and the fact that $\mathcal{L}(\phi_2) = 2$ with induction establishes item (iii). Furthermore, observe that (68) and item (i) in [31, Proposition 2.6] demonstrate that for all $k \in \{2, 3, \dots\}$, $l_0, l_1, \dots, l_k \in \mathbb{N}$ with $\mathcal{D}(\phi_k) = (l_0, l_1, \dots, l_k)$ it holds that

$$\mathcal{D}(\phi_{k+1}) = \mathcal{D}\big(\phi_k \bullet \big(\mathbf{P}_2(\Phi, \mathfrak{I}_{k-1})\big)\big) = (k+1, 2k+1, l_1, l_2, \dots, l_k). \tag{73}$$

This, item (iii), the fact that $\mathcal{D}(\phi_2) = (2, 3, 1)$, and induction establish item (iv). In the next step we observe that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned}\big(\mathcal{R}_{\mathfrak{r}}(\mathfrak{I}_1)\big)(x) &= \begin{pmatrix} 1 & -1 \end{pmatrix} \mathfrak{M}_{\mathfrak{r},2}\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix} x + \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) + 0 = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \max\{x, 0\} \\ \max\{-x, 0\} \end{pmatrix} \\ &= \max\{x, 0\} + \max\{-x, 0\} = x\end{aligned} \tag{74}$$

(cf. Definitions 2.4, 2.14, and 2.3) Combining this with [31, Proposition 2.19] shows that for all $k \in \mathbb{N}$, $x = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$ it holds that

$$\big(\mathcal{R}_{\mathfrak{r}}(\mathfrak{I}_k)\big)(x) = \big(\big(\mathcal{R}_{\mathfrak{r}}(\mathfrak{I}_1)\big)(x_1), \big(\mathcal{R}_{\mathfrak{r}}(\mathfrak{I}_1)\big)(x_2), \ldots, \big(\mathcal{R}_{\mathfrak{r}}(\mathfrak{I}_1)\big)(x_k)\big) = (x_1, x_2, \ldots, x_k) = x. \tag{75}$$

Next note that for all $(x_1, x_2) \in \mathbb{R}^2$ it holds that

$$\begin{aligned}\big(\mathcal{R}_{\mathfrak{r}}(\Phi)\big)(x_1, x_2) &= \begin{pmatrix}1 & 1 & -1\end{pmatrix} \mathfrak{M}_{\mathfrak{r},3}\left(\begin{pmatrix}1 & -1\\ 0 & 1\\ 0 & -1\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} + \begin{pmatrix}0\\0\\0\end{pmatrix}\right) + 0\\ &= \begin{pmatrix}1 & 1 & -1\end{pmatrix}\begin{pmatrix}\max\{x_1 - x_2, 0\}\\ \max\{x_2, 0\}\\ \max\{-x_2, 0\}\end{pmatrix}\\ &= \max\{x_1 - x_2, 0\} + \max\{x_2, 0\} - \max\{-x_2, 0\}\\ &= \max\{x_1 - x_2, 0\} + x_2 = \max\{x_1, x_2\}.\end{aligned} \tag{76}$$

Combining this and (75) with [31, Proposition 2.19] shows that for all $k \in \{2, 3, \ldots\}$, $x = (x_1, x_2, \ldots, x_{k+1}) \in \mathbb{R}^{k+1}$ it holds that

$$\begin{aligned}\big(\mathcal{R}_{\mathfrak{r}}(\mathbf{P}_2(\Phi, \mathfrak{I}_{k-1}))\big)(x) &= \big(\big(\mathcal{R}_{\mathfrak{r}}(\Phi)\big)(x_1, x_2), \big(\mathcal{R}_{\mathfrak{r}}(\mathfrak{I}_{k-1})\big)(x_3, x_4, \ldots, x_{k+1})\big)\\ &= (\max\{x_1, x_2\}, x_3, x_4, \ldots, x_{k+1}).\end{aligned} \tag{77}$$

Item (v) in [31, Proposition 2.6] therefore demonstrates that for all $k \in \{2, 3, \ldots\}$, $x = (x_1, x_2, \ldots, x_{k+1}) \in \mathbb{R}^{k+1}$ it holds that

$$\begin{aligned}\big(\mathcal{R}_{\mathfrak{r}}(\phi_{k+1})\big)(x) &= \Big(\mathcal{R}_{\mathfrak{r}}\big(\phi_k \bullet \big(\mathbf{P}_2(\Phi, \mathfrak{I}_{k-1})\big)\big)\Big)(x) = \big(\mathcal{R}_{\mathfrak{r}}(\phi_k) \circ \mathcal{R}_{\mathfrak{r}}(\mathbf{P}_2(\Phi, \mathfrak{I}_{k-1}))\big)(x)\\ &= \big(\mathcal{R}_{\mathfrak{r}}(\phi_k)\big)(\max\{x_1, x_2\}, x_3, x_4, \ldots, x_{k+1}).\end{aligned} \tag{78}$$

This, the fact that $\phi_2 = \Phi$, (76), and induction establish item (v). The proof of Lemma 3.3 is thus completed. ☐

Lemma 3.4. *Let $A_k \in \mathbb{R}^{(2k-1)\times k}$, $k \in \{2, 3, \ldots\}$, and $C_k \in \mathbb{R}^{(k-1)\times(2k-1)}$, $k \in \{2, 3, \ldots\}$, satisfy for all $k \in \{2, 3, \ldots\}$ that*

$$A_k = \begin{pmatrix}1 & -1 & 0 & \cdots & 0\\ 0 & 1 & 0 & \cdots & 0\\ 0 & -1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ 0 & 0 & -1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\\ 0 & 0 & 0 & \cdots & -1\end{pmatrix} \quad \text{and} \quad C_k = \begin{pmatrix}1 & 1 & -1 & 0 & 0 & \cdots & 0 & 0\\ 0 & 0 & 0 & 1 & -1 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -1\end{pmatrix} \tag{79}$$

and let $\phi_k = ((W_{k,1}, B_{k,1}), (W_{k,2}, B_{k,2}), \ldots, (W_{k,k}, B_{k,k})) \in \mathbf{N}$, $k \in \{2, 3, \ldots\}$, satisfy for all $k \in \{2, 3, \ldots\}$ that $\mathcal{I}(\phi_k) = \mathcal{O}(\mathbf{P}_2(\phi_2, \mathfrak{I}_{k-1}))$, $\phi_{k+1} = \phi_k \bullet (\mathbf{P}_2(\phi_2, \mathfrak{I}_{k-1}))$, and

$$\phi_2 = \left(\left(\begin{pmatrix}1 & -1\\ 0 & 1\\ 0 & -1\end{pmatrix}, \begin{pmatrix}0\\0\\0\end{pmatrix}\right), \left(\begin{pmatrix}1 & 1 & -1\end{pmatrix}, 0\right)\right) \in \big((\mathbb{R}^{3\times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1\times 3} \times \mathbb{R})\big) \tag{80}$$

(cf. Definitions 2.13, 3.2, 2.16, and 2.17 and Lemma 3.3). Then

(i) it holds for all $k \in \{2, 3, \ldots\}$ that $W_{k,1} = A_k$,

(ii) it holds for all $k \in \{2, 3, \ldots\}$, $l \in \{1, 2, \ldots, k\}$ that $B_{k,l} = 0 \in \mathbb{R}^{2(k-l)+1}$,

(iii) it holds for all $k \in \{2, 3, \ldots\}$, $l \in \{3, 4, \ldots, k+1\}$ that $(W_{k+1,l}, B_{k+1,l}) = (W_{k,l-1}, B_{k,l-1})$,

(iv) it holds for all $k \in \{2, 3, \ldots\}$ that $W_{k+1,2} = W_{k,1}C_{k+1}$, and

(v) it holds for all $k \in \{2, 3, \ldots\}$ that $|||\mathcal{T}(\phi_k)||| \leq 1$

(cf. Definitions 2.18 and 2.9).

Proof of Lemma 3.4. First, note that for all $k \in \{2, 3, \ldots\}$ it holds that

$$\mathbf{P}_2(\phi_2, \mathfrak{I}_{k-1}) = (\mathfrak{N}_{A_{k+1}}, \mathfrak{N}_{C_{k+1}}) \tag{81}$$

(cf. Definition 2.15). This and (80) imply that for all $k \in \{2, 3, \ldots\}$ it holds that

$$\begin{aligned} \phi_{k+1} &= \phi_k \bullet \big(\mathbf{P}_2(\phi_2, \mathfrak{I}_{k-1})\big) \\ &= ((W_{k,1}, B_{k,1}), (W_{k,2}, B_{k,2}), \ldots, (W_{k,k}, B_{k,k})) \bullet (\mathfrak{N}_{A_{k+1}}, \mathfrak{N}_{C_{k+1}}) \\ &= (\mathfrak{N}_{A_{k+1}}, (W_{k,1}C_{k+1}, B_{k,1}), (W_{k,2}, B_{k,2}), \ldots, (W_{k,k}, B_{k,k})). \end{aligned} \tag{82}$$

This, (79), and (80) establish item (i). Combining (80) and (82) with induction and Lemma 3.3 proves item (ii). Moreover, note that (82) proves item (iii) and item (iv). In the next step we note that (82) and Lemma 2.20 ensure that for all $k \in \{2, 3, \ldots\}$ it holds that

$$\mathcal{T}(\phi_{k+1}) = \Big(\mathcal{T}\big((\mathfrak{N}_{A_{k+1}})\big), \mathcal{T}\big(((W_{k,1}C_{k+1}, B_{k,1}))\big), \mathcal{T}\big(((W_{k,2}, B_{k,2}))\big), \ldots, \mathcal{T}\big(((W_{k,k}, B_{k,k}))\big)\Big) \tag{83}$$

(cf. Definition 2.18). This implies that for all $k \in \{2, 3, \ldots\}$ it holds that

$$|||\mathcal{T}(\phi_{k+1})||| \leq \max\big\{|||\mathcal{T}\big((\mathfrak{N}_{A_{k+1}})\big)|||, |||\mathcal{T}\big(((W_{k,1}C_{k+1}, B_{k,1}))\big)|||, |||\mathcal{T}(\phi_k)|||\big\} \tag{84}$$

(cf. Definition 2.9). In addition, observe that item (i) proves that for all $k \in \{2, 3, \ldots\}$ it holds that

$$\begin{aligned} W_{k,1}C_{k+1} &= \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 & -1 & -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}. \end{aligned} \tag{85}$$

Item (ii) hence ensures that for all $k \in \{2, 3, \ldots\}$ it holds that $|||\mathcal{T}\big(((W_{k,1}C_{k+1}, B_{k,1}))\big)||| = 1$. Combining this and (79) with (84) shows that for all $k \in \{2, 3, \ldots\}$ it holds that

$$|||\mathcal{T}(\phi_{k+1})||| \leq \max\{1, |||\mathcal{T}(\phi_k)|||\}. \tag{86}$$

The fact that $|||\mathcal{T}(\phi_2)||| = 1$ and induction hence establish item (v). The proof of Lemma 3.4 is thus completed. ☐

3.1.3 Interpolation through DNNs

Lemma 3.5. *Let $d \in \mathbb{N}$, $L \in [0, \infty)$, let $\phi_k \in \mathbf{N}$, $k \in \{2, 3, \ldots\}$, satisfy for all $k \in \{2, 3, \ldots\}$ that $\mathcal{I}(\phi_k) = \mathcal{O}(\mathbf{P}_2(\phi_2, \mathfrak{I}_{k-1}))$, $\phi_{k+1} = \phi_k \bullet (\mathbf{P}_2(\phi_2, \mathfrak{I}_{k-1}))$, and*

$$\phi_2 = \left(\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right), \left(\begin{pmatrix} 1 & 1 & -1 \end{pmatrix}, 0 \right) \right) \in \left((\mathbb{R}^{3\times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1\times 3} \times \mathbb{R}) \right), \tag{87}$$

let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy $|\mathcal{M}| \in \{2, 3, \ldots\}$, let $m \colon \{1, 2, \ldots, |\mathcal{M}|\} \to \mathcal{M}$ be bijective, let $f \colon \mathcal{M} \to \mathbb{R}$ and $F \colon \mathbb{R}^d \to \mathbb{R}$ satisfy for all $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$ that

$$F(x) = \max_{y=(y_1,y_2\ldots,y_d)\in\mathcal{M}} \left[f(y) - L\left(\sum_{i=1}^{d} |x_i - y_i| \right) \right], \tag{88}$$

let $W_1 \in \mathbb{R}^{(2d)\times d}$, $W_2 \in \mathbb{R}^{1\times(2d)}$, let $B_1^{(z)} \in \mathbb{R}^{2d}$, $z \in \mathcal{M}$, assume for all $z = (z_1, z_2, \ldots, z_d) \in \mathcal{M}$ that

$$W_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & -1 \end{pmatrix}, \quad B_1^{(z)} = \begin{pmatrix} -z_1 \\ z_1 \\ -z_2 \\ z_2 \\ \vdots \\ -z_d \\ z_d \end{pmatrix}, \quad \text{and} \quad W_2 = \begin{pmatrix} -L & -L & \cdots & -L \end{pmatrix}, \tag{89}$$

let $\mathcal{W}_1 \in \mathbb{R}^{(2d|\mathcal{M}|)\times d}$, $\mathcal{B}_1 \in \mathbb{R}^{2d|\mathcal{M}|}$, $\mathcal{W}_2 \in \mathbb{R}^{|\mathcal{M}|\times(2d|\mathcal{M}|)}$, $\mathcal{B}_2 \in \mathbb{R}^{|\mathcal{M}|}$ satisfy

$$\mathcal{W}_1 = \begin{pmatrix} W_1 \\ W_1 \\ \vdots \\ W_1 \end{pmatrix}, \quad \mathcal{B}_1 = \begin{pmatrix} B_1^{(m(1))} \\ B_1^{(m(2))} \\ \vdots \\ B_1^{(m(|\mathcal{M}|))} \end{pmatrix}, \quad \mathcal{W}_2 = \begin{pmatrix} W_2 & 0 & \cdots & 0 \\ 0 & W_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_2 \end{pmatrix}, \text{ and } \mathcal{B}_2 = \begin{pmatrix} f(m(1)) \\ f(m(2)) \\ \vdots \\ f(m(|\mathcal{M}|)) \end{pmatrix}, \tag{90}$$

and let $\Phi \in \mathbf{N}$ satisfy $\Phi = \phi_{|\mathcal{M}|} \bullet ((\mathcal{W}_1, \mathcal{B}_1), (\mathcal{W}_2, \mathcal{B}_2))$ (cf. Definitions 2.13, 3.2, 2.17, and 2.16 and Lemma 3.3). Then

(i) it holds that $\mathcal{D}(\Phi) = (d, 2d|\mathcal{M}|, 2|\mathcal{M}| - 1, 2|\mathcal{M}| - 3, \ldots, 3, 1)$,

(ii) it holds that $\mathcal{L}(\Phi) = |\mathcal{M}| + 1$,

(iii) it holds that $|||\mathcal{T}(\Phi)||| \leq \max\{1, L, \sup_{z\in\mathcal{M}} |||z|||, 2[\sup_{z\in\mathcal{M}} |f(z)|]\}$, and

(iv) it holds that $F = \mathcal{R}_{\mathfrak{r}}(\Phi)$

(cf. Definitions 2.18, 2.9, 2.4, and 2.14).

Proof of Lemma 3.5. Throughout this proof let $\Psi \in \mathbf{N}$ satisfy $\Psi = ((\mathcal{W}_1, \mathcal{B}_1), (\mathcal{W}_2, \mathcal{B}_2))$, for every $k \in \{2, 3, \ldots\}$ let $A_k \in \mathbb{R}^{(2k-1)\times k}$ and $C_k \in \mathbb{R}^{(k-1)\times(2k-1)}$ satisfy that

$$A_k = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix} \quad \text{and} \quad C_k = \begin{pmatrix} 1 & 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}, \tag{91}$$

and let $\mathfrak{m}_{i,j} \in \mathbb{R}$, $i \in \{1,2,\dots,|\mathcal{M}|\}$, $j \in \{1,2\dots,d\}$, satisfy for all $i \in \{1,2,\dots,|\mathcal{M}|\}$, $j \in \{1,2\dots,d\}$ that $m(i) = (\mathfrak{m}_{i,1}, \mathfrak{m}_{i,2}, \dots, \mathfrak{m}_{i,d})$. Note that Lemma 3.3 establishes that there exists $((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \dots, (\mathfrak{W}_{|\mathcal{M}|}, \mathfrak{B}_{|\mathcal{M}|})) \in \mathbf{N}$ such that

$$\phi_{|\mathcal{M}|} = ((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \dots, (\mathfrak{W}_{|\mathcal{M}|}, \mathfrak{B}_{|\mathcal{M}|})). \tag{92}$$

Next observe that $\mathcal{L}(\Psi) = 2$ and

$$\mathcal{D}(\Psi) = (d, 2d|\mathcal{M}|, |\mathcal{M}|). \tag{93}$$

Moreover, note that item (iv) in Lemma 3.3 ensures that

$$\mathcal{D}(\phi_{|\mathcal{M}|}) = (\mathcal{M}, 2\mathcal{M} - 1, 2\mathcal{M} - 3, \dots, 3, 1). \tag{94}$$

This, the fact that $\Phi = \phi_{|\mathcal{M}|} \bullet \Psi$, (93), and item (i) in [31, Proposition 2.6] show that $\mathcal{L}(\Phi) = |\mathcal{M}| + 1$ and

$$\mathcal{D}(\Phi) = (d, 2d|\mathcal{M}|, 2|\mathcal{M}| - 1, 2|\mathcal{M}| - 3, \dots, 3, 1). \tag{95}$$

This establishes items (i) and (ii). In the next step we note that the hypothesis that $\Phi = \phi_{|\mathcal{M}|} \bullet ((\mathcal{W}_1, \mathcal{B}_1), (\mathcal{W}_2, \mathcal{B}_2))$ and (92) ensure that

$$\begin{aligned} \Phi &= ((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \dots, (\mathfrak{W}_{|\mathcal{M}|}, \mathfrak{B}_{|\mathcal{M}|})) \bullet ((\mathcal{W}_1, \mathcal{B}_1), (\mathcal{W}_2, \mathcal{B}_2)) \\ &= ((\mathcal{W}_1, \mathcal{B}_1), (\mathfrak{W}_1\mathcal{W}_2, \mathfrak{W}_1\mathcal{B}_2 + \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \dots, (\mathfrak{W}_{|\mathcal{M}|}, \mathfrak{B}_{|\mathcal{M}|})). \end{aligned} \tag{96}$$

Lemma 2.20 hence implies that

$$\mathcal{T}(\Phi) = \big(\mathcal{T}\big(((\mathcal{W}_1, \mathcal{B}_1))\big), \mathcal{T}\big(((\mathfrak{W}_1\mathcal{W}_2, \mathfrak{W}_1\mathcal{B}_2 + \mathfrak{B}_1))\big), \mathcal{T}\big(((\mathfrak{W}_2, \mathfrak{B}_2))\big), \dots, \mathcal{T}\big(((\mathfrak{W}_{|\mathcal{M}|}, \mathfrak{B}_{|\mathcal{M}|}))\big)\big) \tag{97}$$

(cf. Definition 2.18). Moreover, note that (90) and item (i) in Lemma 3.4 imply that

$$\mathfrak{W}_1\mathcal{W}_2 = \underbrace{\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}}_{\in \mathbb{R}^{(2|\mathcal{M}|-1)\times|\mathcal{M}|}} \mathcal{W}_2 = \begin{pmatrix} W_2 & -W_2 & 0 & \cdots & 0 \\ 0 & W_2 & 0 & \cdots & 0 \\ 0 & -W_2 & 0 & \cdots & 0 \\ 0 & 0 & W_2 & \cdots & 0 \\ 0 & 0 & -W_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & W_2 \\ 0 & 0 & 0 & \cdots & -W_2 \end{pmatrix}. \tag{98}$$

In addition, observe that (90) and items (i) and (ii) in Lemma 3.4 show that

$$\begin{aligned} \mathfrak{W}_1\mathcal{B}_2 + \mathfrak{B}_1 &= \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix} \mathcal{B}_2 + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix} \begin{pmatrix} f(m(1)) \\ f(m(2)) \\ \vdots \\ f(m(|\mathcal{M}|)) \end{pmatrix} = \begin{pmatrix} f(m(1)) - f(m(2)) \\ f(m(2)) \\ -f(m(2)) \\ f(m(3)) \\ -f(m(3)) \\ \vdots \\ f(m(|\mathcal{M}|)) \\ -f(m(|\mathcal{M}|)) \end{pmatrix}. \end{aligned} \tag{99}$$

This and (98) demonstrate that

$$\begin{aligned}&\left|\!\left|\!\left|\mathcal{T}\big(((\mathfrak{W}_1\mathcal{W}_2,\mathfrak{W}_1\mathcal{B}_2+\mathfrak{B}_1))\big)\right|\!\right|\!\right|\\&=\max\{L,|f(m(1))-f(m(2))|,|f(m(2))|,|f(m(3))|,\ldots,|f(m(|\mathcal{M}|))|\}\le\max\left\{L,2\left[\sup_{z\in\mathcal{M}}|f(z)|\right]\right\}\end{aligned}\tag{100}$$

(cf. Definition 2.9). Combining this, (90), and item (v) in Lemma 3.4 with (97) proves that

$$\begin{aligned}|||\mathcal{T}(\Phi)|||&\le\max\left\{\left|\!\left|\!\left|\mathcal{T}\big(((\mathcal{W}_1,\mathcal{B}_1))\big)\right|\!\right|\!\right|,\left|\!\left|\!\left|\mathcal{T}\big(((\mathfrak{W}_1\mathcal{W}_2,\mathfrak{W}_1\mathcal{B}_2+\mathfrak{B}_1))\big)\right|\!\right|\!\right|,|||\mathcal{T}(\phi_{|\mathcal{M}|})|||\right\}\\&\le\max\left\{1,\sup_{z\in\mathcal{M}}|||z|||,L,2\left[\sup_{z\in\mathcal{M}}|f(z)|\right]\right\}.\end{aligned}\tag{101}$$

This establishes item (iii). Observe that (89) ensures that for all $x=(x_1,x_2,\ldots,x_d)\in\mathbb{R}^d$, $z=(z_1,z_2,\ldots,z_d)\in\mathcal{M}$ it holds that

$$W_1x+B_1^{(z)}=\underbrace{\begin{pmatrix}1&0&\cdots&0\\-1&0&\cdots&0\\0&1&\cdots&0\\0&-1&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&1\\0&0&\cdots&-1\end{pmatrix}}_{\in\mathbb{R}^{(2d)\times d}}\begin{pmatrix}x_1\\x_2\\\vdots\\x_d\end{pmatrix}+\begin{pmatrix}-z_1\\z_1\\-z_2\\z_2\\\vdots\\-z_d\\z_d\end{pmatrix}=\begin{pmatrix}x_1\\-x_1\\x_2\\-x_2\\\vdots\\x_d\\-x_d\end{pmatrix}+\begin{pmatrix}-z_1\\z_1\\-z_2\\z_2\\\vdots\\-z_d\\z_d\end{pmatrix}=\begin{pmatrix}x_1-z_1\\-(x_1-z_1)\\x_2-z_2\\-(x_2-z_2)\\\vdots\\x_d-z_d\\-(x_d-z_d)\end{pmatrix}.\tag{102}$$

This and (90) prove that for all $x=(x_1,x_2,\ldots,x_d)\in\mathbb{R}^d$, $z=(z_1,z_2,\ldots,z_d)\in\mathcal{M}$ it holds that

$$\begin{aligned}W_2\mathfrak{R}_{2d}(W_1x+B_1^{(z)})&=\underbrace{\begin{pmatrix}-L&-L&\cdots&-L\end{pmatrix}}_{\in\mathbb{R}^{1\times(2d)}}\begin{pmatrix}\max\{x_1-z_1,0\}\\\max\{z_1-x_1,0\}\\\max\{x_2-z_2,0\}\\\max\{z_2-x_2,0\}\\\vdots\\\max\{x_d-z_d,0\}\\\max\{z_d-x_d,0\}\end{pmatrix}\\&=-L\left[\sum_{i=1}^d(\max\{x_i-z_i,0\}+\max\{z_i-x_i,0\})\right]=-L\left[\sum_{i=1}^d|x_i-z_i|\right]\end{aligned}\tag{103}$$

(cf. Definition 2.5). Moreover, note that (90) implies that for all $x\in\mathbb{R}^d$ it holds that

$$\mathcal{W}_1x+\mathcal{B}_1=\begin{pmatrix}W_1x+B_1^{(m(1))}\\W_1x+B_1^{(m(2))}\\\vdots\\W_1x+B_1^{(m(|\mathcal{M}|))}\end{pmatrix}.\tag{104}$$

Therefore, we obtain that for all $x\in\mathbb{R}^d$ it holds that

$$\mathfrak{R}_{2d|\mathcal{M}|}(\mathcal{W}_1x+\mathcal{B}_1)=\begin{pmatrix}\mathfrak{R}_{2d}(W_1x+B_1^{(m(1))})\\\mathfrak{R}_{2d}(W_1x+B_1^{(m(2))})\\\vdots\\\mathfrak{R}_{2d}(W_1x+B_1^{(m(|\mathcal{M}|))})\end{pmatrix}.\tag{105}$$

This, (90), and (103) imply that for all $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$ it holds that

$$
\begin{aligned}
\big(\mathcal{R}_{\mathfrak{r}}(\Psi)\big)(x) &= \mathcal{W}_2 \mathfrak{R}_{2d|\mathcal{M}|}(\mathcal{W}_1 x + \mathcal{B}_1) + \mathcal{B}_2 \\
&= \begin{pmatrix} W_2 & 0 & \cdots & 0 \\ 0 & W_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_2 \end{pmatrix} \begin{pmatrix} \mathfrak{R}_{2d}(W_1 x + B_1^{(m(1))}) \\ \mathfrak{R}_{2d}(W_1 x + B_1^{(m(2))}) \\ \vdots \\ \mathfrak{R}_{2d}(W_1 x + B_1^{(m(|\mathcal{M}|))}) \end{pmatrix} + \begin{pmatrix} f(m(1)) \\ f(m(2)) \\ \vdots \\ f(m(|\mathcal{M}|)) \end{pmatrix} \\
&= \begin{pmatrix} W_2 \mathfrak{R}_{2d}(W_1 x + B_1^{(m(1))}) \\ W_2 \mathfrak{R}_{2d}(W_1 x + B_1^{(m(2))}) \\ \vdots \\ W_2 \mathfrak{R}_{2d}(W_1 x + B_1^{(m(|\mathcal{M}|))}) \end{pmatrix} + \begin{pmatrix} f(m(1)) \\ f(m(2)) \\ \vdots \\ f(m(|\mathcal{M}|)) \end{pmatrix} \\
&= \begin{pmatrix} f(m(1)) - L\big[\sum_{i=1}^d |x_i - \mathfrak{m}_{1,i}|\big] \\ f(m(2)) - L\big[\sum_{i=1}^d |x_i - \mathfrak{m}_{2,i}|\big] \\ \vdots \\ f(m(|\mathcal{M}|)) - L\big[\sum_{i=1}^d |x_i - \mathfrak{m}_{|\mathcal{M}|,i}|\big] \end{pmatrix}
\end{aligned}
\tag{106}
$$

(cf. Definitions 2.4 and 2.14). This, the fact that $\Phi = \phi_{|\mathcal{M}|} \bullet \Psi$, item (v) in Lemma 3.3, and item (v) in [31, Proposition 2.6] ensure that for all $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$ it holds that

$$
\begin{aligned}
\big(\mathcal{R}_{\mathfrak{r}}(\Phi)\big)(x) = \big(\mathcal{R}_{\mathfrak{r}}(\phi_{|\mathcal{M}|}) \circ \mathcal{R}_{\mathfrak{r}}(\Psi)\big)(x) &= \max_{i \in \{1,2,\ldots,|\mathcal{M}|\}} \left[f(m(i)) - L\left(\sum_{j=1}^d |x_j - \mathfrak{m}_{i,j}| \right) \right] \\
&= \max_{z=(z_1,z_2\ldots,z_d) \in \mathcal{M}} \left[f(z) - L\left(\sum_{i=1}^d |x_i - z_i| \right) \right].
\end{aligned}
\tag{107}
$$

This establishes item (iv). The proof of Lemma 3.5 is thus completed. ☐

3.1.4 Explicit approximations through DNNs

Proposition 3.6. *Let $d \in \mathbb{N}$, $L \in [0, \infty)$, let $\phi_k \in \mathbf{N}$, $k \in \{2, 3, \ldots\}$, satisfy for all $k \in \{2, 3, \ldots\}$ that $\mathcal{I}(\phi_k) = \mathcal{O}(\mathbf{P}_2(\phi_2, \mathfrak{I}_{k-1}))$, $\phi_{k+1} = \phi_k \bullet (\mathbf{P}_2(\phi_2, \mathfrak{I}_{k-1}))$, and*

$$
\phi_2 = \left(\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right), \left(\begin{pmatrix} 1 & 1 & -1 \end{pmatrix}, 0 \right) \right) \in \big((\mathbb{R}^{3 \times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1 \times 3} \times \mathbb{R}) \big), \tag{108}
$$

let $A \subseteq \mathbb{R}^d$, let $f\colon A \to \mathbb{R}$ satisfy for all $x = (x_1, x_2, \ldots, x_d)$, $y = (y_1, y_2, \ldots, y_d) \in A$ that $|f(x) - f(y)| \leq L\big[\sum_{i=1}^d |x_i - y_i|\big]$, let $\mathcal{M} \subseteq A$ satisfy $|\mathcal{M}| \in \{2, 3, \ldots\}$, let $m\colon \{1, 2, \ldots, |\mathcal{M}|\} \to \mathcal{M}$ be bijective, let $W_1 \in \mathbb{R}^{(2d) \times d}$, $W_2 \in \mathbb{R}^{1 \times (2d)}$, let $B_1^{(z)} \in \mathbb{R}^{2d}$, $z \in \mathcal{M}$, assume for all $z = (z_1, z_2, \ldots, z_d) \in \mathcal{M}$ that

$$
W_1 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \\ \vdots \\ 1 \\ -1 \end{pmatrix}, \quad B_1^{(z)} = \begin{pmatrix} -z_1 \\ z_1 \\ -z_2 \\ z_2 \\ \vdots \\ -z_d \\ z_d \end{pmatrix}, \quad \text{and} \quad W_2 = \begin{pmatrix} -L & -L & \cdots & -L \end{pmatrix}, \tag{109}
$$

let $\mathcal{W}_1 \in \mathbb{R}^{(2d|\mathcal{M}|)\times d}$, $\mathcal{B}_1 \in \mathbb{R}^{(2d|\mathcal{M}|)}$, $\mathcal{W}_2 \in \mathbb{R}^{|\mathcal{M}|\times(2d|\mathcal{M}|)}$, $\mathcal{B}_2 \in \mathbb{R}^{|\mathcal{M}|}$ *satisfy*

$$\mathcal{W}_1 = \begin{pmatrix} W_1 \\ W_1 \\ \vdots \\ W_1 \end{pmatrix}, \quad \mathcal{B}_1 = \begin{pmatrix} B_1^{(m(1))} \\ B_1^{(m(2))} \\ \vdots \\ B_1^{(m(|\mathcal{M}|))} \end{pmatrix}, \quad \mathcal{W}_2 = \begin{pmatrix} W_2 & 0 & \cdots & 0 \\ 0 & W_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_2 \end{pmatrix}, \text{ and } \quad B_2 = \begin{pmatrix} f(m(1)) \\ f(m(2)) \\ \vdots \\ f(m(|\mathcal{M}|)) \end{pmatrix}, \tag{110}$$

and let $\Phi \in \mathbf{N}$ *satisfy* $\Phi = \phi_{|\mathcal{M}|} \bullet ((\mathcal{W}_1, \mathcal{B}_1), (\mathcal{W}_2, \mathcal{B}_2))$ *(cf. Definitions 2.13, 3.2, 2.17, and 2.16 and Lemma 3.3). Then*

(i) it holds that $\mathcal{D}(\Phi) = (d, 2d|\mathcal{M}|, 2|\mathcal{M}| - 1, 2|\mathcal{M}| - 3, \ldots, 3, 1)$,

(ii) it holds that $|||\mathcal{T}(\Phi)||| \leq \max\{1, L, \sup_{z\in\mathcal{M}}|||z|||, 2[\sup_{z\in\mathcal{M}} |f(z)|]\}$, *and*

(iii) it holds that

$$\left[\sup_{x\in A} \left|f(x) - \big(\mathcal{R}_{\mathfrak{r}}(\Phi)\big)(x)\right|\right] \leq 2L\left[\sup_{(x_1,x_2,\ldots,x_d)\in A} \left(\inf_{(z_1,z_2,\ldots,z_d)\in\mathcal{M}} \sum_{i=1}^{d} |x_i - z_i|\right)\right] \tag{111}$$

(cf. Definitions 2.18, 2.9, 2.4, and 2.14).

Proof of Proposition 3.6. Throughout this proof let $F\colon A \to \mathbb{R}$ satisfy for all $x = (x_1, x_2, \ldots, x_d) \in A$ that

$$F(x) = \max_{z=(z_1,z_2\ldots,z_d)\in\mathcal{M}} \left[f(z) - L\left(\sum_{i=1}^{d} |x_i - z_i|\right)\right]. \tag{112}$$

Observe that Lemma 3.5 establishes

(A) that $\mathcal{D}(\Phi) = (d, 2d|\mathcal{M}|, 2|\mathcal{M}| - 1, 2|\mathcal{M}| - 3, \ldots, 3, 1)$,

(B) that $|||\mathcal{T}(\Phi)||| \leq \max\{1, L, \sup_{z\in\mathcal{M}}|||z|||, 2[\sup_{z\in\mathcal{M}} |f(z)|]\}$, and

(C) that for all $x \in A$ it holds that $\big(\mathcal{R}_{\mathfrak{r}}(\Phi)\big)(x) = F(x)$

(cf. Definitions 2.18, 2.9, 2.4, and 2.14). Observe that item (A) and item (B) prove item (i) and item (ii). Next note that item (C) and Lemma 3.1 (with $E \leftarrow A$, $\delta \leftarrow (A \times A \ni ((x_1, x_2, \ldots, x_d), (y_1, y_2, \ldots, y_d)) \mapsto \sum_{i=1}^{d}|x_i - y_i| \in [0, \infty))$, $L \leftarrow L$, $\mathcal{M} \leftarrow \mathcal{M}$, $f \leftarrow f$, $F \leftarrow F$ in the notation of Lemma 3.1) ensure that

$$\sup_{x\in A} \left|f(x) - \big(\mathcal{R}_{\mathfrak{r}}(\Phi)\big)(x)\right| = \sup_{x\in A} |f(x) - F(x)| \leq 2L\left[\sup_{(x_1,x_2,\ldots,x_d)\in A} \left(\inf_{(z_1,z_2,\ldots,z_d)\in\mathcal{M}} \sum_{i=1}^{d} |x_i - z_i|\right)\right]. \tag{113}$$

The proof of Proposition 3.6 is thus completed. □

3.1.5 Implicit approximations through DNNs

Corollary 3.7. *Let* $d, \mathfrak{d} \in \mathbb{N}$, $L \in [0, \infty)$, *let* $A \subseteq \mathbb{R}^d$, *let* $f\colon A \to \mathbb{R}$ *satisfy for all* $x = (x_1, x_2, \ldots, x_d)$, $y = (y_1, y_2, \ldots, y_d) \in A$ *that* $|f(x) - f(y)| \leq L\big[\sum_{i=1}^{d} |x_i - y_i|\big]$, *let* $\mathcal{M} \subseteq A$ *satisfy* $|\mathcal{M}| \in \{2, 3, \ldots\}$, *let* $l_0, l_1, \ldots, l_{|\mathcal{M}|+1} \in \mathbb{N}$ *satisfy* $(l_0, l_1, \ldots, l_{|\mathcal{M}|+1}) = (d, 2d|\mathcal{M}|, 2|\mathcal{M}| - 1, 2|\mathcal{M}| - 3, \ldots, 3, 1)$, *and assume that* $\mathfrak{d} \geq \sum_{k=1}^{|\mathcal{M}|+1} l_k(l_{k-1}+1)$. *Then there exists* $\theta \in \mathbb{R}^{\mathfrak{d}}$ *such that* $|||\theta||| \leq \max\{1, L, \sup_{z\in\mathcal{M}}|||z|||, 2[\sup_{z\in\mathcal{M}} |f(z)|]\}$ *and*

$$\sup_{x\in A} \left|f(x) - \big(\mathscr{N}_{-\infty,\infty}^{\theta,(l_0,l_1,\ldots,l_{|\mathcal{M}|+1})}\big)(x)\right| \leq 2L\left[\sup_{(x_1,x_2,\ldots,x_d)\in A} \left(\inf_{(z_1,z_2,\ldots,z_d)\in\mathcal{M}} \sum_{i=1}^{d} |x_i - z_i|\right)\right] \tag{114}$$

(cf. Definitions 2.9 and 2.8).

Proof of Corollary 3.7. Observe that item (ii) in Lemma 3.3 and Proposition 3.6 ensure that there exists $\Phi \in \mathbf{N}$ which satisfies

(A) that $\mathcal{D}(\Phi) = (l_0, l_1, \ldots, l_{|\mathcal{M}|+1})$,

(B) that $|||\mathcal{T}(\Phi)||| \leq \max\{1, L, \sup_{z\in\mathcal{M}} |||z|||, 2[\sup_{z\in\mathcal{M}} |f(z)|]\}$, and

(C) that

$$\left[\sup_{x\in A} \left|f(x) - \big(\mathcal{R}_{\mathfrak{r}}(\Phi)\big)(x)\right|\right] \leq 2L\left[\sup_{(x_1,x_2,\ldots,x_d)\in A}\left(\inf_{(z_1,z_2,\ldots,z_d)\in\mathcal{M}} \sum_{i=1}^{d} |x_i - z_i|\right)\right] \tag{115}$$

(cf. Definitions 2.13, 2.18, 2.9, 2.4, and 2.14) Combining this with Corollary 2.22 establishes (114). The proof of Corollary 3.7 is thus completed. □

Corollary 3.8. *Let $d, \mathfrak{d} \in \mathbb{N}$, $L \in [0,\infty)$, $u \in [-\infty,\infty)$, $v \in (u,\infty]$, let $A \subseteq \mathbb{R}^d$, let $f\colon A \to \mathbb{R}$ satisfy for all $x = (x_1, x_2, \ldots, x_d)$, $y = (y_1, y_2, \ldots, y_d) \in A$ that $u \leq f(x) \leq v$ and $|f(x) - f(y)| \leq L\big[\sum_{i=1}^{d} |x_i - y_i|\big]$, let $\mathcal{M} \subseteq A$ satisfy $|\mathcal{M}| \in \{2, 3, \ldots\}$, let $l_0, l_1, \ldots, l_{|\mathcal{M}|+1} \in \mathbb{N}$ satisfy $(l_0, l_1, \ldots, l_{|\mathcal{M}|+1}) = (d, 2d|\mathcal{M}|, 2|\mathcal{M}| - 1, 2|\mathcal{M}| - 3, \ldots, 3, 1)$, and assume that $\mathfrak{d} \geq \sum_{k=1}^{|\mathcal{M}|+1} l_k(l_{k-1}+1)$. Then there exists $\theta \in \mathbb{R}^{\mathfrak{d}}$ such that $|||\theta||| \leq \max\{1, L, \sup_{z\in\mathcal{M}} |||z|||, 2[\sup_{z\in\mathcal{M}} |f(z)|]\}$ and*

$$\sup_{x\in A} \left|f(x) - \big(\mathscr{N}^{\theta,(l_0,l_1,\ldots,l_{|\mathcal{M}|+1})}_{u,v}\big)(x)\right| \leq 2L\left[\sup_{(x_1,x_2,\ldots,x_d)\in A}\left(\inf_{(z_1,z_2,\ldots,z_d)\in\mathcal{M}} \sum_{i=1}^{d} |x_i - z_i|\right)\right] \tag{116}$$

(cf. Definitions 2.9 and 2.8).

Proof of Corollary 3.8. First, observe that Corollary 3.7 (with $d \leftarrow d$, $\mathfrak{d} \leftarrow \mathfrak{d}$, $L \leftarrow L$, $A \leftarrow A$, $f \leftarrow (A \ni x \mapsto f(x) \in \mathbb{R})$, $\mathcal{M} \leftarrow \mathcal{M}$, $(l_0, l_1, \ldots, l_{|\mathcal{M}|+1}) \leftarrow (l_0, l_1, \ldots, l_{|\mathcal{M}|+1})$ in the notation of Corollary 3.7) ensures that there exists $\theta \in \mathbb{R}^{\mathfrak{d}}$ which satisfies $|||\theta||| \leq \max\{1, L, \sup_{z\in\mathcal{M}} |||z|||, 2[\sup_{z\in\mathcal{M}} |f(z)|]\}$ and

$$\sup_{x\in A} \left|f(x) - \big(\mathscr{N}^{\theta,(l_0,l_1,\ldots,l_{|\mathcal{M}|+1})}_{-\infty,\infty}\big)(x)\right| \leq 2L\left[\sup_{(x_1,x_2,\ldots,x_d)\in A}\left(\inf_{(z_1,z_2,\ldots,z_d)\in\mathcal{M}} \sum_{i=1}^{d} |x_i - z_i|\right)\right] \tag{117}$$

(cf. Definitions 2.9 and 2.8). The assumption that for all $x \in A$ it holds that $u \leq f(x) \leq v$ and Lemma 2.11 hence imply that

$$\begin{aligned}
&\sup_{x\in A} \left|f(x) - \big(\mathscr{N}^{\theta,(l_0,l_1,\ldots,l_{|\mathcal{M}|+1})}_{u,v}\big)(x)\right| = \sup_{x\in A} \left|\mathfrak{c}_{u,v}(f(x)) - \mathfrak{c}_{u,v}\big(\big(\mathscr{N}^{\theta,(l_0,l_1,\ldots,l_{|\mathcal{M}|+1})}_{-\infty,\infty}\big)(x)\big)\right| \\
&\leq \sup_{x\in A} \left|f(x) - \big(\mathscr{N}^{\theta,(l_0,l_1,\ldots,l_{|\mathcal{M}|+1})}_{-\infty,\infty}\big)(x)\right| \leq 2L\left[\sup_{(x_1,x_2,\ldots,x_d)\in A}\left(\inf_{(z_1,z_2,\ldots,z_d)\in\mathcal{M}} \sum_{i=1}^{d} |x_i - z_i|\right)\right]
\end{aligned} \tag{118}$$

(cf. Definition 2.6). This completes the proof of Corollary 3.8. □

Corollary 3.9. *Let $d, \mathfrak{d} \in \mathbb{N}$, $L \in [0,\infty)$, $u \in [-\infty,\infty)$, $v \in (u,\infty]$, let $A \subseteq \mathbb{R}^d$, let $f\colon A \to \mathbb{R}$ satisfy for all $x = (x_1, x_2, \ldots, x_d)$, $y = (y_1, y_2, \ldots, y_d) \in A$ that $u \leq f(x) \leq v$ and $|f(x) - f(y)| \leq L\big[\sum_{i=1}^{d} |x_i - y_i|\big]$, let $\mathcal{M} \subseteq A$ satisfy $|\mathcal{M}| \in \{2, 3, \ldots\}$, let $l_0, l_1, \ldots, l_{|\mathcal{M}|+1} \in \mathbb{N}$ satisfy for all $k \in \{2, 3, \ldots, |\mathcal{M}|\}$ that $l_0 = d$, $l_1 \geq 2d|\mathcal{M}|$, $l_k \geq 2|\mathcal{M}| - 2k + 3$, and $l_{|\mathcal{M}|+1} = 1$, and assume that $\mathfrak{d} \geq \sum_{k=1}^{|\mathcal{M}|+1} l_k(l_{k-1}+1)$. Then there exists $\theta \in \mathbb{R}^{\mathfrak{d}}$ such that $|||\theta||| \leq \max\{1, L, \sup_{z\in\mathcal{M}} |||z|||, 2[\sup_{z\in\mathcal{M}} |f(z)|]\}$ and*

$$\sup_{x\in A} \left|f(x) - \big(\mathscr{N}^{\theta,(l_0,l_1,\ldots,l_{|\mathcal{M}|+1})}_{u,v}\big)(x)\right| \leq 2L\left[\sup_{(x_1,x_2,\ldots,x_d)\in A}\left(\inf_{(z_1,z_2,\ldots,z_d)\in\mathcal{M}} \sum_{i=1}^{d} |x_i - z_i|\right)\right] \tag{119}$$

(cf. Definitions 2.9 and 2.8).

Proof of Corollary 3.9. Note that Corollary 3.8 and Lemma 2.24 establish (119). The proof of Corollary 3.9 is thus completed. $\square$

Corollary 3.10. *Let $d, \mathfrak{d}, N \in \mathbb{N}$, $L \in [0, \infty)$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$ satisfy $\mathfrak{d} \geq 2d^2(N+1)^d + 5d(N+1)^{2d} + \frac{4}{3}(N+1)^{3d}$, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be the standard norm, let $p = (p_1, p_2, \dots, p_d)$, $q = (q_1, q_2, \dots, q_d) \in \mathbb{R}^d$ satisfy for all $i \in \{1, 2, \dots, d\}$ that $p_i \leq q_i$ and $\max_{j \in \{1,2,\dots,d\}}(q_j - p_j) > 0$, let $A = \prod_{i=1}^d [p_i, q_i]$, let $\mathcal{M} \subseteq A$ satisfy*

$$\mathcal{M} = \left\{ (z_1, z_2, \dots, z_d) \in \mathbb{R}^d \colon \begin{pmatrix} \exists\, k_1, k_2, \dots, k_d \in \{0, 1, \dots, N\} \colon \\ \forall\, i \in \{1, 2, \dots, d\} \colon z_i = p_i + \frac{k_i}{N}(q_i - p_i) \end{pmatrix} \right\}, \tag{120}$$

and let $f \colon A \to \mathbb{R}$ satisfy for all $x, y \in A$ that $u \leq f(x) \leq v$ and $|f(x) - f(y)| \leq L\|x - y\|$. Then there exist $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\mathfrak{L} \in \mathbb{N}$, $l_0, l_1, \dots, l_{\mathfrak{L}} \in \mathbb{N}$ such that $|||\theta||| \leq \max\{1, L, |||p|||, |||q|||, 2[\sup_{z \in A} |f(z)|]\}$, $\mathfrak{d} \geq \sum_{k=1}^{\mathfrak{L}} l_k(l_{k-1} + 1)$, and

$$\sup_{x \in A} \left| f(x) - \mathscr{N}_{u,v}^{\theta,(l_0,l_1,\dots,l_{\mathfrak{L}})}(x) \right| \leq \frac{L}{N} \sum_{i=1}^d |q_i - p_i| \tag{121}$$

(cf. Definitions 2.9 and 2.8).

Proof of Corollary 3.10. Throughout this proof let $l_0, l_1, \dots, l_{|\mathcal{M}|+1} \in \mathbb{N}$ satisfy $(l_0, l_1, \dots, l_{|\mathcal{M}|+1}) = (d, 2d|\mathcal{M}|, 2|\mathcal{M}| - 1, 2|\mathcal{M}| - 3, \dots, 3, 1)$. Observe that the fact that $|\mathcal{M}| \leq (N+1)^d$ and the fact that for all $n \in \mathbb{N}$ it holds that $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6} \leq \frac{(n+1)^3}{3}$ ensure that

$$\begin{aligned} &\sum_{k=1}^{|\mathcal{M}|+1} l_k(l_{k-1} + 1) \\ &= \underbrace{d(2d|\mathcal{M}|) + 2d|\mathcal{M}|(2|\mathcal{M}| - 1) + \left[\sum_{i=1}^{|\mathcal{M}|-1} (2i-1)(2i+1) \right]}_{\text{number of weights}} + \underbrace{2d|\mathcal{M}| + \left[\sum_{i=1}^{|\mathcal{M}|} (2i-1) \right]}_{\text{number of biases}} \\ &= 2d^2|\mathcal{M}| + 4d|\mathcal{M}|^2 + 4\left[\sum_{i=1}^{|\mathcal{M}|-1} i^2 \right] - |\mathcal{M}| + 1 + |\mathcal{M}|^2 \\ &\leq 2d^2|\mathcal{M}| + 5d|\mathcal{M}|^2 + \frac{4}{3}|\mathcal{M}|^3 \leq 2d^2(N+1)^d + 5d(N+1)^{2d} + \frac{4}{3}(N+1)^{3d} \leq \mathfrak{d}. \end{aligned} \tag{122}$$

In addition, note that the hypothesis that for all $x, y \in A$ it holds that $|f(x) - f(y)| \leq L\|x - y\|$ implies that for all $x = (x_1, x_2, \dots, x_d), y = (y_1, y_2, \dots, y_d) \in A$ it holds that

$$|f(x) - f(y)| \leq L \left[\sum_{i=1}^d |x_i - y_i| \right]. \tag{123}$$

Furthermore, observe that the hypothesis that $\max_{j \in \{1,2,\dots,d\}}(q_j - p_j) > 0$ ensures that $|\mathcal{M}| \geq 2$. Combining this, (122), and (123) with Corollary 3.8 establishes that there exists $\theta \in \mathbb{R}^{\mathfrak{d}}$ such that $|||\theta||| \leq \max\{1, L, \sup_{z \in \mathcal{M}} |||z|||, 2[\sup_{z \in \mathcal{M}} |f(z)|]\}$ and

$$\sup_{x \in A} \left| f(x) - \left(\mathscr{N}_{u,v}^{\theta,(l_0,l_1,\dots,l_{|\mathcal{M}|+1})} \right)(x) \right| \leq 2L \left[\sup_{(x_1,x_2,\dots,x_d) \in A} \left(\inf_{(z_1,z_2,\dots,z_d) \in \mathcal{M}} \sum_{i=1}^d |x_i - z_i| \right) \right] \tag{124}$$

(cf. Definitions 2.9 and 2.8). Next note that the hypothesis that $\mathcal{M} \subseteq A = \prod_{i=1}^d [p_i, q_i]$ implies that for all $z \in \mathcal{M}$ it holds that

$$|||z||| \leq \max\{|||p|||, |||q|||\}. \tag{125}$$

Therefore, we obtain that

$$|||\theta||| \leq \max\left\{1, L, |||p|||, |||q|||, 2\left[\sup_{z\in\mathcal{M}} |f(z)|\right]\right\}. \tag{126}$$

In the next step we note that the fact that for all $N \in \mathbb{N}$, $r \in \mathbb{R}$, $s \in [r, \infty)$, $x \in [r, s]$ there exists $k \in \{0, 1, \ldots, N\}$ such that $|x - (r + \frac{k}{N}(s - r))| \leq \frac{s-r}{2N}$ ensures that for all $(x_1, x_2, \ldots, x_d) \in A$ there exists $(z_1, z_2, \ldots, z_d) \in \mathcal{M}$ such that

$$\sum_{i=1}^{d} |x_i - z_i| \leq \frac{1}{2N}\left[\sum_{i=1}^{d} |q_i - p_i|\right]. \tag{127}$$

Combining this, (122), (124), and (126) establishes (121). The proof of Corollary 3.10 is thus completed. □

3.2 Analysis of the generalization error

3.2.1 Hoeffding's concentration inequality

Proposition 3.11. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $N \in \mathbb{N}$, $\varepsilon, a_1, a_2, \ldots, a_N \in \mathbb{R}$, $b_1 \in [a_1, \infty)$, $b_2 \in [a_2, \infty)$, $\ldots$, $b_N \in [a_N, \infty)$, and let $X_n \colon \Omega \to [a_n, b_n]$, $n \in \{1, 2, \ldots, N\}$, be independent random variables. Then*

$$\mathbb{P}\left(\frac{1}{N}\left|\sum_{n=1}^{N} \big(X_n - \mathbb{E}[X_n]\big)\right| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2\varepsilon^2 N^2}{\sum_{n=1}^{N} (b_n - a_n)^2}\right). \tag{128}$$

3.2.2 Covering number estimates

Definition 3.12 (Covering number)**.** Let $(\mathcal{H}, d)$ be a metric space and let $r \in (0, \infty)$. Then we denote by $\mathcal{C}_{(\mathcal{H},d),r} \in \mathbb{N} \cup \{\infty\}$ (we denote by $\mathcal{C}_{\mathcal{H},r} \in \mathbb{N} \cup \{\infty\}$) the extended real number given by

$$\mathcal{C}_{(\mathcal{H},d),r} = \inf\Big(\Big\{n \in \mathbb{N} \colon \big(\exists\, v_1, v_2, \ldots, v_n \in \mathcal{H} \colon \big[\mathcal{H} \subseteq \textstyle\bigcup_{i=1}^{n} \{g \in \mathcal{H} \colon d(v_i, g) \leq r\}\big]\big)\Big\} \cup \{\infty\}\Big). \tag{129}$$

Proposition 3.13. *Let $(X, \|\cdot\|)$ be a finite-dimensional Banach space, let $R \in [1, \infty)$, $r \in (0, 1)$, $B = \{\theta \in X \colon \|\theta\| \leq R\}$, and let $d \colon B \times B \to [0, \infty)$ satisfy for all $\theta, \vartheta \in B$ that $d(\theta, \vartheta) = \|\theta - \vartheta\|$. Then $\ln(\mathcal{C}_{(B,d),r}) \leq \dim(X) \ln\big(\frac{4R}{r}\big)$ (cf. Definition 3.12).*

3.2.3 Measurability properties for suprema

Lemma 3.14. *Let $(X, \mathcal{X})$ be a topological space, let $Y \subseteq X$ be a countable set, assume Y is dense in X, let $(\Omega, \mathcal{F})$ be a measurable space, let $f_x \colon \Omega \to \mathbb{R}$, $x \in X$, be $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable functions, assume for all $\omega \in \Omega$ that $X \ni x \mapsto f_x(\omega) \in \mathbb{R}$ is a continuous function, and let $F \colon \Omega \to \mathbb{R} \cup \{\infty\}$ satisfy for all $\omega \in \Omega$ that $F(\omega) = \sup_{x\in X} f_x(\omega)$. Then*

(i) it holds for all $\omega \in \Omega$ that $F(\omega) = \sup_{x\in Y} f_x(\omega)$ and

(ii) it holds that F is an $\mathcal{F}/\mathcal{B}(\mathbb{R} \cup \{\infty\})$-measurable function.

Proof of Lemma 3.14. Note that the fact that Y is dense in X implies that for all $g \in C(X, \mathbb{R})$ it holds that

$$\sup_{x\in X} g(x) = \sup_{x\in Y} g(x). \tag{130}$$

This and the hypothesis that for all $\omega \in \Omega$ it holds that $X \ni x \mapsto f_x(\omega) \in \mathbb{R}$ is a continuous function show that for all $\omega \in \Omega$ it holds that

$$F(\omega) = \sup_{x\in X} f_x(\omega) = \sup_{x\in Y} f_x(\omega). \tag{131}$$

This establishes item (i). Combining (131) with the hypothesis that for all $x \in X$ it holds that $f_x \colon \Omega \to \mathbb{R}$ is an $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable function establishes item (ii). The proof of Lemma 3.14 is thus completed. □

3.2.4 Concentration inequalities for random fields

Lemma 3.15. *Let (X,d) be a separable metric space, let $\varepsilon, L \in (0,\infty)$, $N \in \mathbb{N}$, $z_1, z_2, \ldots, z_N \in X$ satisfy for all $x \in X$ that $\inf_{i\in\{1,2,\ldots,N\}} d(x,z_i) \leq \frac{\varepsilon}{4L}$, let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, and let $Z^{(x)}\colon \Omega \to \mathbb{R}$, $x \in X$, be random variables which satisfy for all $x,y \in X$ that $\mathbb{E}[|Z^{(x)}|] < \infty$ and $|Z^{(x)} - Z^{(y)}| \leq Ld(x,y)$. Then*

$$\mathbb{P}\big(\big[\sup_{x\in X}|Z^{(x)} - \mathbb{E}[Z^{(x)}]|\big] \geq \varepsilon\big) \leq \sum_{i=1}^{N} \mathbb{P}\big(\big[|Z^{(z_i)} - \mathbb{E}[Z^{(z_i)}]|\big] \geq \tfrac{\varepsilon}{2}\big) \tag{132}$$

(cf. Lemma 3.14).

Proof of Lemma 3.15. Throughout this proof let $B_1, B_2, \ldots, B_N \subseteq X$ satisfy for all $i \in \{1,2,\ldots,N\}$ that $B_i = \{x \in X\colon d(x,z_i) \leq \frac{\varepsilon}{4L}\}$. Observe that the triangle inequality and the hypothesis that for all $x,y \in X$ it holds that $|Z^{(x)} - Z^{(y)}| \leq Ld(x,y)$ show that for all $i \in \{1,2,\ldots,N\}$, $x \in B_i$ it holds that

$$\begin{aligned}|Z^{(x)} - \mathbb{E}[Z^{(x)}]| &\leq |Z^{(x)} - Z^{(z_i)}| + |Z^{(z_i)} - \mathbb{E}[Z^{(z_i)}]| + |\mathbb{E}[Z^{(z_i)}] - \mathbb{E}[Z^{(x)}]| \\ &\leq Ld(x,z_i) + |Z^{(z_i)} - \mathbb{E}[Z^{(z_i)}]| + \mathbb{E}[|Z^{(z_i)} - Z^{(x)}|] \\ &\leq |Z^{(z_i)} - \mathbb{E}[Z^{(z_i)}]| + 2Ld(x,z_i) \leq |Z^{(z_i)} - \mathbb{E}[Z^{(z_i)}]| + \tfrac{\varepsilon}{2}.\end{aligned} \tag{133}$$

Combining this with Lemma 3.14 proves that for all $i \in \{1,2,\ldots,N\}$ it holds that

$$\mathbb{P}\big(\big[\sup_{x\in B_i}|Z^{(x)} - \mathbb{E}[Z^{(x)}]|\big] \geq \varepsilon\big) \leq \mathbb{P}\big(|Z^{(z_i)} - \mathbb{E}[Z^{(z_i)}]| + \tfrac{\varepsilon}{2} \geq \varepsilon\big) = \mathbb{P}\big(|Z^{(z_i)} - \mathbb{E}[Z^{(z_i)}]| \geq \tfrac{\varepsilon}{2}\big). \tag{134}$$

(cf. Lemma 3.14). Next note that the hypothesis that for all $x \in X$ it holds that $\inf_{i\in\{1,2,\ldots,N\}} d(x,z_i) \leq \frac{\varepsilon}{4L}$ ensures that $\bigcup_{i=1}^N B_i = X$. Combining this and (134) with Lemma 3.14 establishes that

$$\begin{aligned}&\mathbb{P}\big(\big[\sup_{x\in X}|Z^{(x)} - \mathbb{E}[Z^{(x)}]|\big] \geq \varepsilon\big) = \mathbb{P}\Big(\Big[\sup_{x\in(\bigcup_{i=1}^N B_i)}|Z^{(x)} - \mathbb{E}[Z^{(x)}]|\Big] \geq \varepsilon\Big) \\ &= \mathbb{P}\Big(\textstyle\bigcup_{i=1}^N \big\{\big[\sup_{x\in B_i}|Z^{(x)} - \mathbb{E}[Z^{(x)}]|\big] \geq \varepsilon\big\}\Big) \\ &\leq \sum_{i=1}^{N} \mathbb{P}\big(\big[\sup_{x\in B_i}|Z^{(x)} - \mathbb{E}[Z^{(x)}]|\big] \geq \varepsilon\big) \leq \sum_{i=1}^{N} \mathbb{P}\big(\big[|Z^{(z_i)} - \mathbb{E}[Z^{(z_i)}]|\big] \geq \tfrac{\varepsilon}{2}\big).\end{aligned} \tag{135}$$

This completes the proof of Lemma 3.15. □

Lemma 3.16. *Let (X,d) be a separable metric space, let $\varepsilon, L \in (0,\infty)$, let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, and let $Z^{(x)}\colon \Omega \to \mathbb{R}$, $x \in X$, be random variables which satisfy for all $x,y \in X$ that $\mathbb{E}[|Z^{(x)}|] < \infty$ and $|Z^{(x)} - Z^{(y)}| \leq Ld(x,y)$. Then*

$$\mathbb{P}\big(\big[\sup_{x\in X}|Z^{(x)} - \mathbb{E}[Z^{(x)}]|\big] \geq \varepsilon\big) \leq \mathcal{C}_{(X,d),\frac{\varepsilon}{4L}}\big[\sup_{x\in X}\mathbb{P}(|Z^{(x)} - \mathbb{E}[Z^{(x)}]| \geq \tfrac{\varepsilon}{2})\big]. \tag{136}$$

(cf. Lemma 3.14 and Definition 3.12).

Proof of Lemma 3.16. Throughout this proof let $N \in \mathbb{N} \cup \{\infty\}$ satisfy $N = \mathcal{C}_{(X,d),\frac{\varepsilon}{4L}}$, assume without loss of generality that $N < \infty$, and let $z_1, z_2, \ldots, z_N \in X$ satisfy $X \subseteq \bigcup_{i=1}^N \{x \in X\colon d(x,z_i) \leq \frac{\varepsilon}{4L}\}$ (cf. Definition 3.12). Observe that Lemma 3.15 implies that

$$\begin{aligned}\mathbb{P}\big(\big[\sup_{x\in X}|Z^{(x)} - \mathbb{E}[Z^{(x)}]|\big] \geq \varepsilon\big) &\leq \sum_{i=1}^{N} \mathbb{P}\big(\big[|Z^{(z_i)} - \mathbb{E}[Z^{(z_i)}]|\big] \geq \tfrac{\varepsilon}{2}\big) \\ &\leq N\big[\sup_{x\in X}\mathbb{P}(|Z^{(x)} - \mathbb{E}[Z^{(x)}]| \geq \tfrac{\varepsilon}{2})\big].\end{aligned} \tag{137}$$

This completes the proof of Lemma 3.16. □

Lemma 3.17. *Let (X, d) be a separable metric space, let $M \in \mathbb{N}$, $\varepsilon, L, D \in (0, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $x \in X$ let $E_1^{(x)}, E_2^{(x)}, \ldots, E_M^{(x)} \colon \Omega \to [0, D]$ be independent random variables, let $Z^{(x)} \colon \Omega \to \mathbb{R}$ be random variables, and assume for all $x, y \in X$, $m \in \{1, 2, \ldots, M\}$ that $|E_m^{(x)} - E_m^{(y)}| \leq Ld(x, y)$ and $Z^{(x)} = \frac{1}{M}\big[\sum_{m=1}^M E_m^{(x)}\big]$. Then it holds for all $y \in X$ that $\mathbb{E}[|Z^{(y)}|] < \infty$ and*

$$\mathbb{P}\big(\big[\sup_{x\in X}|Z^{(x)} - \mathbb{E}[Z^{(x)}]|\big] \geq \varepsilon\big) \leq 2\mathcal{C}_{(X,d),\frac{\varepsilon}{4L}} \exp\!\left(\frac{-\varepsilon^2 M}{2D^2}\right) \tag{138}$$

(cf. Lemma 3.14 and Definition 3.12).

Proof of Lemma 3.17. First, observe that the triangle inequality and the hypothesis that for all $x, y \in X$, $m \in \{1, 2, \ldots, M\}$ it holds that $|E_m^{(x)} - E_m^{(y)}| \leq Ld(x, y)$ imply that for all $x, y \in X$ it holds that

$$\begin{aligned} |Z^{(x)} - Z^{(y)}| &= \left|\frac{1}{M}\left[\sum_{m=1}^M E_m^{(x)}\right] - \frac{1}{M}\left[\sum_{m=1}^M E_m^{(y)}\right]\right| = \left|\frac{1}{M}\left[\sum_{m=1}^M \big(E_m^{(x)} - E_m^{(y)}\big)\right]\right| \\ &\leq \frac{1}{M}\left[\sum_{m=1}^M \big|E_m^{(x)} - E_m^{(y)}\big|\right] \leq Ld(x, y). \end{aligned} \tag{139}$$

Next note that the hypothesis that for all $x \in X$, $m \in \{1, 2, \ldots, M\}$, $\omega \in \Omega$ it holds that $|E_m^{(x)}(\omega)| \leq D$ ensures that for all $x \in X$ it holds that

$$\mathbb{E}\big[|Z^{(x)}|\big] = \mathbb{E}\left[\frac{1}{M}\left|\sum_{m=1}^M E_m^{(x)}\right|\right] \leq \mathbb{E}\left[\frac{1}{M}\left[\sum_{m=1}^M \big|E_m^{(x)}\big|\right]\right] = \frac{1}{M}\left[\sum_{m=1}^M \mathbb{E}\big[\big|E_m^{(x)}\big|\big]\right] \leq D < \infty. \tag{140}$$

Hence, we obtain that for all $x \in X$ it holds that

$$|Z^{(x)} - \mathbb{E}[Z^{(x)}]| = \left|\frac{1}{M}\left[\sum_{m=1}^M E_m^{(x)}\right] - \mathbb{E}\left[\frac{1}{M}\left[\sum_{m=1}^M E_m^{(x)}\right]\right]\right| = \left|\frac{1}{M}\left[\sum_{m=1}^M \big(E_m^{(x)} - \mathbb{E}\big[E_m^{(x)}\big]\big)\right]\right|. \tag{141}$$

Combining this with Proposition 3.11 (with $(\Omega, \mathcal{F}, \mathbb{P}) \leftarrow (\Omega, \mathcal{F}, \mathbb{P})$, $N \leftarrow M$, $\varepsilon \leftarrow \frac{\varepsilon}{2}$, $(a_1, a_2, \ldots, a_N) \leftarrow (0, 0, \ldots, 0)$, $(b_1, b_2, \ldots, b_N) \leftarrow (D, D, \ldots, D)$, $(X_1, X_2, \ldots, X_N) \leftarrow (E_1^{(x)}, E_2^{(x)}, \ldots, E_M^{(x)})$ for $x \in X$ in the notation of Proposition 3.11) ensures that for all $x \in X$ it holds that

$$\mathbb{P}\big(|Z^{(x)} - \mathbb{E}[Z^{(x)}]| \geq \tfrac{\varepsilon}{2}\big) \leq 2\exp\!\left(\frac{-2\big[\frac{\varepsilon}{2}\big]^2 M^2}{MD^2}\right) = 2\exp\!\left(\frac{-\varepsilon^2 M}{2D^2}\right). \tag{142}$$

Combining this, (139), and (140) with Lemma 3.16 establishes (138). The proof of Lemma 3.17 is thus completed. □

3.2.5 Uniform estimates for the statistical learning error

Lemma 3.18. *Let (A, d) be a separable metric space, let $M \in \mathbb{N}$, $\varepsilon, L, D \in (0, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $x \in A$ let $(X_m^{(x)}, Y_m) \colon \Omega \to \mathbb{R} \times \mathbb{R}$, $m \in \{1, 2, \ldots, M\}$, be i.i.d. random variables, assume for all $x, y \in A$, $m \in \{1, 2, \ldots, M\}$ that $\big|X_m^{(x)} - X_m^{(y)}\big| \leq Ld(x, y)$ and $\big|X_m^{(x)} - Y_m\big| \leq D$, let $\mathfrak{E}^{(x)} \colon \Omega \to [0, \infty)$, $x \in A$, satisfy for all $x \in A$ that*

$$\mathfrak{E}^{(x)} = \frac{1}{M}\left[\sum_{m=1}^M |X_m^{(x)} - Y_m|^2\right], \tag{143}$$

and let $\mathcal{E}^{(x)} \in [0, \infty)$, $x \in A$, satisfy for all $x \in A$ that $\mathcal{E}^{(x)} = \mathbb{E}[|X_1^{(x)} - Y_1|^2]$. Then

$$\mathbb{P}\big(\big[\sup_{x\in A}|\mathfrak{E}^{(x)} - \mathcal{E}^{(x)}|\big] \geq \varepsilon\big) \leq 2\mathcal{C}_{(A,d),\frac{\varepsilon}{8LD}} \exp\!\left(\frac{-\varepsilon^2 M}{2D^4}\right) \tag{144}$$

(cf. Lemma 3.14 and Definition 3.12).

Proof of Lemma 3.18. Throughout this proof let $E_m^{(x)}\colon \Omega \to \mathbb{R}$, $x \in A$, $m \in \{1,2,\dots,M\}$, satisfy for all $x \in A$, $m \in \{1,2,\dots,M\}$ that

$$E_m^{(x)} = \big|X_m^{(x)} - Y_m\big|^2. \tag{145}$$

Observe that the hypothesis that for all $x \in A$, $m \in \{1,2,\dots,M\}$ it holds that $\big|X_m^{(x)} - Y_m\big| \le D$ ensures that for all $x \in A$, $m \in \{1,2,\dots,M\}$, $\omega \in \Omega$ it holds that

$$E_m^{(x)}(\omega) \in [0, D^2]. \tag{146}$$

Next note that for all $x_1, x_2, y \in \mathbb{R}$ it holds that

$$(x_1 - y)^2 - (x_2 - y)^2 = (x_1 - x_2)((x_1 - y) + (x_2 - y)). \tag{147}$$

This, the hypothesis that for all $x \in X$, $m \in \{1,2,\dots,M\}$ it holds that $\big|X_m^{(x)} - Y_m\big| \le D$, and the hypothesis that for all $x, y \in X$, $m \in \{1,2,\dots,M\}$ it holds that $\big|X_m^{(x)} - X_m^{(y)}\big| \le Ld(x,y)$ imply that for all $x, y \in A$, $m \in \{1,2,\dots,M\}$ it holds that

$$\begin{split}\big|E_m^{(x)} - E_m^{(y)}\big| &= \big|(X_m^{(x)} - Y_m)^2 - (X_m^{(y)} - Y_m)^2\big| = \big|X_m^{(x)} - X_m^{(y)}\big|\big|(X_m^{(x)} - Y_m) + (X_m^{(y)} - Y_m)\big| \\ &\le \big|X_m^{(x)} - X_m^{(y)}\big|\big(\big|X_m^{(x)} - Y_m\big| + \big|X_m^{(y)} - Y_m\big|\big) \le 2D\big|X_m^{(x)} - X_m^{(y)}\big| \le 2LDd(x,y).\end{split} \tag{148}$$

In addition, observe that for all $x \in A$ it holds that

$$\mathbb{E}\big[\mathfrak{E}^{(x)}\big] = \frac{1}{M}\left[\sum_{m=1}^{M} \mathbb{E}\Big[\big|X_m^{(x)} - Y_m\big|^2\Big]\right] = \frac{1}{M}\left[\sum_{m=1}^{M} \mathbb{E}\Big[\big|X_1^{(x)} - Y_1\big|^2\Big]\right] = \frac{1}{M}\left[\sum_{m=1}^{M} \mathcal{E}^{(x)}\right] = \mathcal{E}^{(x)}. \tag{149}$$

Furthermore, note that the hypothesis that for all $x \in A$ it holds that $(X_m^{(x)}, Y_m)\colon \Omega \to \mathbb{R} \times \mathbb{R}$, $m \in \{1,2,\dots,M\}$, are i.i.d. random variables ensures that for all $x \in A$ it holds that $E_m^{(x)}\colon \Omega \to \mathbb{R}$, $m \in \{1,2,\dots,M\}$, are i.i.d. random variables. Combining this, (146), (148), and (149) with Lemma 3.17 (with $(X,d) \leftarrow (A,d)$, $M \leftarrow M$, $\varepsilon \leftarrow \varepsilon$, $L \leftarrow 2LD$, $D \leftarrow D^2$, $(\Omega, \mathcal{F}, \mathbb{P}) \leftarrow (\Omega, \mathcal{F}, \mathbb{P})$, $((E_1^{(x)}, E_2^{(x)}, \dots, E_m^{(x)}))_{x\in A} \leftarrow ((E_1^{(x)}, E_2^{(x)}, \dots, E_m^{(x)}))_{x\in A}$, $(Z^{(x)})_{x\in A} = (\mathfrak{E}^{(x)})_{x\in A}$ in the notation of Lemma 3.17) establishes (144). The proof of Lemma 3.18 is thus completed. ☐

Lemma 3.19. *Let $d, \mathfrak{d}, M \in \mathbb{N}$, $R, L, \mathcal{R} \in [1,\infty)$, $\varepsilon \in (0,1)$, $B \subseteq \{\theta \in \mathbb{R}^{\mathfrak{d}}\colon |||\theta||| \le R\}$, let $D \subseteq \mathbb{R}^d$ be a compact set, let $H = (H_\theta)_{\theta\in B}\colon B \to C(D,\mathbb{R})$ satisfy for all $\theta, \vartheta \in B$, $x \in D$ that $|H_\theta(x) - H_\vartheta(x)| \le L|||\theta - \eta|||$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(X_m, Y_m)\colon \Omega \to D \times \mathbb{R}$, $m \in \{1,2,\dots,M\}$, be i.i.d. random variables which satisfy for all $\theta \in B$, $m \in \{1,2,\dots,M\}$ that $|H_\theta(X_m) - Y_m| \le \mathcal{R}$ and $\mathbb{E}[|Y_1|^2] < \infty$, let $\mathcal{E}\colon C(D,\mathbb{R}) \to [0,\infty)$ satisfy for all $f \in C(D,\mathbb{R})$ that $\mathcal{E}(f) = \mathbb{E}[|f(X_1) - Y_1|^2]$, and let $\mathfrak{E}\colon B \times \Omega \to [0,\infty)$ satisfy for all $\theta \in B$, $\omega \in \Omega$ that*

$$\mathfrak{E}(\theta,\omega) = \frac{1}{M}\left[\sum_{m=1}^{M} |H_\theta(X_m(\omega)) - Y_m(\omega)|^2\right] \tag{150}$$

(cf. Definition 2.9). Then

$$\mathbb{P}([\sup\nolimits_{\theta\in B}|\mathfrak{E}(\theta) - \mathcal{E}(H_\theta)|] \ge \varepsilon) \le 2\exp\!\left(\mathfrak{d}\ln\!\left(\frac{32LR\mathcal{R}}{\varepsilon}\right) - \frac{\varepsilon^2 M}{2\mathcal{R}^4}\right) \tag{151}$$

(cf. Lemma 3.14).

Proof of Lemma 3.19. Throughout this proof let $\delta\colon \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}} \to [0,\infty)$ satisfy for all $\theta, \vartheta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\delta(\theta,\vartheta) = |||\theta - \vartheta||| \tag{152}$$

and let $\mathcal{B} = \{\theta \in \mathbb{R}^{\mathfrak{d}} \colon |||\theta||| \le R\}$. Observe that the hypothesis that (X_m, Y_m), $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables and the hypothesis that for all $\theta \in B$ it holds that H_θ is a continuous function imply that for all $\theta \in B$ it holds that $(H_\theta(X_m), Y_m)$, $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables. Note that Lemma 3.18 (with $(A, d) \leftarrow (B, \delta|_{B\times B})$, $M \leftarrow M$, $\varepsilon \leftarrow \varepsilon$, $L \leftarrow L$, $D \leftarrow \mathcal{R}$, $(\Omega, \mathcal{F}, \mathbb{P}) \leftarrow (\Omega, \mathcal{F}, \mathbb{P})$, $(X_1^{(x)}, X_2^{(x)}, \dots, X_m^{(x)})_{x\in A} \leftarrow (H_\theta(X_1), H_\theta(X_2), \dots, H_\theta(X_m))_{\theta\in B}$ $(Y_1, Y_2, \dots, Y_m) \leftarrow (Y_1, Y_2, \dots, Y_m)$, $(\mathfrak{E}^{(x)})_{x\in A} \leftarrow (\Omega \ni \omega \mapsto \mathfrak{E}(\theta, \omega) \in [0, \infty))_{\theta\in B}$, $(\mathcal{E}^{(x)})_{x\in A} \leftarrow (\mathcal{E}(H_\theta))_{\theta\in B}$ in the notation of Lemma 3.18) establishes that

$$\mathbb{P}([\sup_{\theta\in B}|\mathfrak{E}(\theta) - \mathcal{E}(H_\theta)|] \ge \varepsilon) \le 2\mathcal{C}_{(B,\delta|_{B\times B}),\frac{\varepsilon}{8L\mathcal{R}}} \exp\left(\frac{-\varepsilon^2 M}{2\mathcal{R}^4}\right) \tag{153}$$

(cf. Definition 3.12). Furthermore, note that Proposition 3.13 (with $X \leftarrow \mathbb{R}^{\mathfrak{d}}$, $\|\cdot\| \leftarrow (\mathbb{R}^{\mathfrak{d}} \ni x \mapsto |||x||| \in [0, \infty))$, $R \leftarrow R$, $r \leftarrow \frac{\varepsilon}{8L\mathcal{R}}$, $B \leftarrow \mathcal{B}$, $d \leftarrow d|_{\mathcal{B}\times\mathcal{B}}$ in the notation of Proposition 3.13) shows that

$$\ln(\mathcal{C}_{(B,\delta|_{B\times B}),\frac{\varepsilon}{8L\mathcal{R}}}) \le \ln(\mathcal{C}_{(\mathcal{B},\delta|_{\mathcal{B}\times\mathcal{B}}),\frac{\varepsilon}{8L\mathcal{R}}}) \le \mathfrak{d} \ln\left(\frac{32LR\mathcal{R}}{\varepsilon}\right). \tag{154}$$

Combining this with (153) establishes (151). The proof of Lemma 3.19 is thus completed. ☐

Lemma 3.20. *Let $\mathfrak{d}, M \in \mathbb{N}$, $L \in \{2, 3, \dots, \}$, $\varepsilon \in (0, \infty)$, $u, a \in \mathbb{R}$, $v \in (u, \infty)$, $b \in (a, \infty)$, $R \in [1, \infty)$, $\mathbf{l} = (l_0, l_1, \dots, l_L) \in \mathbb{N}^{L+1}$, $B \subseteq \{\theta \in \mathbb{R}^{\mathfrak{d}} \colon |||\theta||| \le R\}$ satisfy $l_L = 1$ and $\mathfrak{d} \ge \sum_{k=1}^{L} l_k(l_{k-1} + 1)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(X_m, Y_m) \colon \Omega \to [a, b]^{l_0} \times [u, v]$, $m \in \{1, 2, \dots, M\}$, be i.i.d. random variables, let $\mathcal{E} \colon C([a, b]^{l_0}, \mathbb{R}) \to [0, \infty)$ satisfy for all $f \in C([a, b]^{l_0}, \mathbb{R})$ that $\mathcal{E}(f) = \mathbb{E}[|f(X_1) - Y_1|^2]$, and let $\mathfrak{E} \colon B \times \Omega \to [0, \infty)$ satisfy for all $\theta \in B$, $\omega \in \Omega$ that*

$$\mathfrak{E}(\theta, \omega) = \frac{1}{M}\left[\sum_{m=1}^{M} |\mathscr{N}_{u,v}^{\theta,\mathbf{l}}(X_m(\omega)) - Y_m(\omega)|^2\right] \tag{155}$$

(cf. Definitions 2.9 and 2.8). Then

$$\begin{aligned} &\mathbb{P}\big(\big[\sup_{\theta\in B}|\mathfrak{E}(\theta) - \mathcal{E}(\mathscr{N}_{u,v}^{\theta,\mathbf{l}})|\big] \ge \varepsilon\big) \\ &\le 2\exp\left(\mathfrak{d}\ln\left(\frac{32L\max\{1, |a|, |b|\}(|||l||| + 1)^L R^L (v - u)}{\varepsilon}\right) - \frac{\varepsilon^2 M}{2(v - u)^4}\right). \end{aligned} \tag{156}$$

Proof of Lemma 3.20. Throughout this proof let $\mathcal{L} \in \mathbb{R}$ satisfy

$$\mathcal{L} = L\max\{1, |a|, |b|\}\,(|||l||| + 1)^L R^{L-1}. \tag{157}$$

Observe that Corollary 2.12 (with $a \leftarrow a$, $b \leftarrow b$, $u \leftarrow u$, $v \leftarrow v$, $L \leftarrow L$, $d \leftarrow \mathfrak{d}$, $l \leftarrow \mathbf{l}$ in the notation of Theorem 2.10) shows that for all $\theta, \vartheta \in B$ it holds that

$$\begin{aligned} \sup_{x\in[a,b]^{l_0}} |||\mathscr{N}_{u,v}^{\theta,\mathbf{l}}(x) - \mathscr{N}_{u,v}^{\vartheta,\mathbf{l}}(x)||| &\le L\max\{1, |a|, |b|\}\,(|||l||| + 1)^L\,(\max\{|||\theta|||, |||\vartheta|||\})^{L-1}\,|||\theta - \vartheta||| \\ &\le L\max\{1, |a|, |b|\}\,(|||l||| + 1)^L R^{L-1}\,|||\theta - \vartheta|||. \end{aligned} \tag{158}$$

Furthermore, observe that for all $\theta \in B$, $m \in \{1, 2, \dots, M\}$ it holds that

$$|\mathscr{N}_{u,v}^{\theta,\mathbf{l}}(X_m) - Y_m| \le v - u. \tag{159}$$

Combining this and (158) with Lemma 3.19 (with $d \leftarrow l_0$, $\mathfrak{d} \leftarrow \mathfrak{d}$, $M \leftarrow M$, $R \leftarrow R$, $\varepsilon \leftarrow \varepsilon$, $L \leftarrow \mathcal{L}$, $\mathcal{R} \leftarrow v - u$, $B \leftarrow B$, $D \leftarrow [a, b]^{l_0}$, $H \leftarrow (B \ni \theta \mapsto \mathscr{N}_{u,v}^{\theta,\mathbf{l}} \in C([a, b]^{l_0}, \mathbb{R}))$, $(\Omega, \mathcal{F}, \mathbb{P}) \leftarrow (\Omega, \mathcal{F}, \mathbb{P})$, $((X_m, Y_m))_{m\in\{1,2,\dots,M\}} \leftarrow ((X_m, Y_m))_{m\in\{1,2,\dots,M\}}$, $\mathcal{E} \leftarrow \mathcal{E}$, $\mathfrak{E} \leftarrow \mathfrak{E}$ in the notation of Lemma 3.19) shows that

$$\mathbb{P}\big(\big[\sup_{\eta\in B}|\mathfrak{E}(\eta) - \mathcal{E}(\mathscr{N}_{u,v}^{\eta,\mathbf{l}})|\big] \ge \varepsilon\big) \le 2\exp\left(\mathfrak{d}\ln\left(\frac{32\mathcal{L}R(v - u)}{\varepsilon}\right) - \frac{\varepsilon^2 M}{2(v - u)^4}\right). \tag{160}$$

The proof of Lemma 3.20 is thus completed. ☐

3.3 Analysis of the optimization error

3.3.1 Convergence rates for the minimum Monte Carlo method

Lemma 3.21. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $\mathfrak{d}, N \in \mathbb{N}$, let $\|\cdot\| : \mathbb{R}^{\mathfrak{d}} \to [0,\infty)$ be a norm, let $\mathfrak{H} \subseteq \mathbb{R}^{\mathfrak{d}}$ be a set, let $\vartheta \in \mathfrak{H}$, $L, \varepsilon \in (0,\infty)$, let $\mathfrak{E}\colon \mathfrak{H} \times \Omega \to \mathbb{R}$ be $(\mathcal{B}(\mathfrak{H}) \otimes \mathcal{A})/\mathcal{B}(\mathbb{R})$-measurable, assume for all $x, y \in \mathfrak{H}$, $\omega \in \Omega$ that $|\mathfrak{E}(x,\omega) - \mathfrak{E}(y,\omega)| \leq L\|x-y\|$, and let $\Theta_n\colon \Omega \to \mathfrak{H}$, $n \in \{1,2,\ldots,N\}$, be i.i.d. random variables. Then*

$$\mathbb{P}\big(\big[\min\nolimits_{n\in\{1,2,\ldots,N\}} \mathfrak{E}(\Theta_n)\big] - \mathfrak{E}(\vartheta) > \varepsilon\big) \leq \big[\mathbb{P}\big(\|\Theta_1 - \vartheta\| > \tfrac{\varepsilon}{L}\big)\big]^N \leq \exp\big(-N\,\mathbb{P}\big(\|\Theta_1 - \vartheta\| \leq \tfrac{\varepsilon}{L}\big)\big). \tag{161}$$

Proof of Lemma 3.21. Note that

$$\begin{aligned}
&\big[\min\nolimits_{n\in\{1,2,\ldots,N\}} \mathfrak{E}(\Theta_n)\big] - \mathfrak{E}(\vartheta) = \min\nolimits_{n\in\{1,2,\ldots,N\}} [\mathfrak{E}(\Theta_n) - \mathfrak{E}(\vartheta)] \\
&\leq \min\nolimits_{n\in\{1,2,\ldots,N\}} |\mathfrak{E}(\Theta_n) - \mathfrak{E}(\vartheta)| \leq \min\nolimits_{n\in\{1,2,\ldots,n\}} \big[L\|\Theta_n - \vartheta\|\big] \\
&= L\big[\min\nolimits_{n\in\{1,2,\ldots,N\}} \|\Theta_n - \vartheta\|\big].
\end{aligned} \tag{162}$$

The hypothesis that Θ_n, $n \in \{1,2,\ldots,N\}$, are i.i.d. random variables and the fact that $\forall\, x \in \mathbb{R}\colon 1-x \leq e^x$ hence show that

$$\begin{aligned}
&\mathbb{P}\Big(\big[\min\nolimits_{n\in\{1,2,\ldots,N\}} \mathfrak{E}(\Theta_n)\big] - \mathfrak{E}(\theta) > \varepsilon\Big) \leq \mathbb{P}\Big(L\big[\min\nolimits_{n\in\{1,2,\ldots,N\}} \|\Theta_n - \vartheta\|\big] > \varepsilon\Big) \\
&= \mathbb{P}\Big(\min\nolimits_{n\in\{1,2,\ldots,N\}} \|\Theta_n - \vartheta\| > \frac{\varepsilon}{L}\Big) = \Big[\mathbb{P}\Big(\|\Theta_1 - \vartheta\| > \frac{\varepsilon}{L}\Big)\Big]^N \\
&= \Big[1 - \mathbb{P}\Big(\|\Theta_1 - \vartheta\| \leq \frac{\varepsilon}{L}\Big)\Big]^N \leq \exp\big(-N\,\mathbb{P}\big(\|\Theta_1 - \vartheta\| \leq \tfrac{\varepsilon}{L}\big)\big).
\end{aligned} \tag{163}$$

The proof of Lemma 3.21 is thus completed. □

3.3.2 Continuous uniformly distributed samples

Lemma 3.22. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $\mathfrak{d}, N \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a,\infty)$, $\vartheta \in [a,b]^{\mathfrak{d}}$, $L, \varepsilon \in (0,\infty)$, let $\mathfrak{E}\colon [a,b]^{\mathfrak{d}} \times \Omega \to \mathbb{R}$ be $\mathcal{B}([a,b]^{\mathfrak{d}})/\mathcal{B}(\mathbb{R})$-measurable, assume for all $x, y \in [a,b]^{\mathfrak{d}}$, $\omega \in \Omega$ that $|\mathfrak{E}(x,\omega) - \mathfrak{E}(y,\omega)| \leq L|||x-y|||$, let $\Theta_n\colon \Omega \to [a,b]^{\mathfrak{d}}$, $n \in \{1,2,\ldots,N\}$, be i.i.d. random variables, and assume that Θ_1 is continuous uniformly distributed on $[a,b]^{\mathfrak{d}}$ (cf. Definition 2.9). Then*

$$\mathbb{P}\big(\big[\min\nolimits_{n\in\{1,2,\ldots,N\}} \mathfrak{E}(\Theta_n)\big] - \mathfrak{E}(\vartheta) > \varepsilon\big) \leq \exp\bigg(-N \min\bigg\{1, \frac{\varepsilon^{\mathfrak{d}}}{L^{\mathfrak{d}}(b-a)^{\mathfrak{d}}}\bigg\}\bigg). \tag{164}$$

Proof of Lemma 3.22. Note that

$$\begin{aligned}
&\mathbb{P}\big(|||\Theta_1 - \vartheta||| \leq \tfrac{\varepsilon}{L}\big) \geq \mathbb{P}\big(|||\Theta_1 - (a,a,\ldots,a)||| \leq \tfrac{\varepsilon}{L}\big) = \mathbb{P}\big(|||\Theta_1 - (a,a,\ldots,a)||| \leq \min\{\tfrac{\varepsilon}{L}, b-a\}\big) \\
&= \bigg[\frac{\min\{\frac{\varepsilon}{L}, b-a\}}{(b-a)}\bigg]^{\mathfrak{d}} = \min\bigg\{1, \bigg[\frac{\varepsilon}{L\,(b-a)}\bigg]^{\mathfrak{d}}\bigg\}.
\end{aligned} \tag{165}$$

Combining this with Lemma 3.21 proves (164). The proof of Lemma 3.22 is thus completed. □

4 Overall error analysis

In this section we combine the separate error analyses of the approximation error, the generalization error, and the optimization error in Section 3 to obtain an overall analysis (cf. Theorem 4.5 below). In Lemma 4.1 below we present the well-known bias-variance decomposition. To formulate this bias-variance decomposition lemma we observe that for every probability space $(\Omega, \mathcal{A}, \mathbb{P})$, every measurable space $(S, \mathcal{S})$, every random variable $X\colon \Omega \to S$, and every $A \in \mathcal{S}$ it holds that $\mathbb{P}_X(A) = \mathbb{P}(X \in A)$. Moreover, note that for every probability space $(\Omega, \mathcal{A}, \mathbb{P})$, every measurable space $(S, \mathcal{S})$, every random variable $X\colon \Omega \to S$, and every $\mathcal{S}/\mathcal{B}(\mathbb{R})$-measurable function $f\colon S \to \mathbb{R}$ it holds that $\int_S |f|^2\, d\mathbb{P}_X = \int_S |f(x)|^2\, \mathbb{P}_X(dx) = \int_\Omega |f(X(\omega))|^2\, \mathbb{P}(d\omega) = \int_\Omega |f(X)|^2\, d\mathbb{P} = \mathbb{E}\big[|f(X)|^2\big]$. A result related to Lemmas 4.1 and 4.2 can, e.g., be found in Berner et al. [10, Lemma 2.8].

4.1 Bias-variance decomposition

Lemma 4.1 (Bias-variance decomposition)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $X \colon \Omega \to S$ and $Y \colon \Omega \to \mathbb{R}$ be random variables with $\mathbb{E}[|Y|^2] < \infty$, and let $\mathcal{E} \colon \mathcal{L}^2(\mathbb{P}_X; \mathbb{R}) \to [0, \infty)$ satisfy for all $f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ that $\mathcal{E}(f) = \mathbb{E}[|f(X) - Y|^2]$. Then*

(i) it holds for all $f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ that

$$\mathcal{E}(f) = \mathbb{E}\big[|f(X) - \mathbb{E}[Y|X]|^2\big] + \mathbb{E}\big[|Y - \mathbb{E}[Y|X]|^2\big], \tag{166}$$

(ii) it holds for all $f, g \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ that

$$\mathcal{E}(f) - \mathcal{E}(g) = \mathbb{E}\big[|f(X) - \mathbb{E}[Y|X]|^2\big] - \mathbb{E}\big[|g(X) - \mathbb{E}[Y|X]|^2\big], \tag{167}$$

and

(iii) it holds for all $f, g \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ that

$$\mathbb{E}\big[|f(X) - \mathbb{E}[Y|X]|^2\big] = \mathbb{E}\big[|g(X) - \mathbb{E}[Y|X]|^2\big] + (\mathcal{E}(f) - \mathcal{E}(g)). \tag{168}$$

Proof of Lemma 4.1. Note that for all $f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ it holds that

$$\begin{aligned}
\mathcal{E}(f) &= \mathbb{E}\big[|f(X) - Y|^2\big] = \mathbb{E}\big[|(f(X) - \mathbb{E}[Y|X]) + (\mathbb{E}[Y|X] - Y)|^2\big] \\
&= \mathbb{E}\big[|f(X) - \mathbb{E}[Y|X]|^2\big] + 2\,\mathbb{E}\big[\big(f(X) - \mathbb{E}[Y|X]\big)\big(\mathbb{E}[Y|X] - Y\big)\big] + \mathbb{E}\big[|\mathbb{E}[Y|X] - Y|^2\big] \\
&= \mathbb{E}\big[|f(X) - \mathbb{E}[Y|X]|^2\big] + 2\,\mathbb{E}\Big[\mathbb{E}\big[\big(f(X) - \mathbb{E}[Y|X]\big)\big(\mathbb{E}[Y|X] - Y\big)\big|X\big]\Big] + \mathbb{E}\big[|\mathbb{E}[Y|X] - Y|^2\big] \\
&= \mathbb{E}\big[|f(X) - \mathbb{E}[Y|X]|^2\big] + 2\,\mathbb{E}\Big[\big(f(X) - \mathbb{E}[Y|X]\big)\mathbb{E}\big[\big(\mathbb{E}[Y|X] - Y\big)\big|X\big]\Big] + \mathbb{E}\big[|\mathbb{E}[Y|X] - Y|^2\big] \\
&= \mathbb{E}\big[|f(X) - \mathbb{E}[Y|X]|^2\big] + 2\,\mathbb{E}\big[\big(f(X) - \mathbb{E}[Y|X]\big)\big(\mathbb{E}[Y|X] - \mathbb{E}[Y|X]\big)\big] + \mathbb{E}\big[|\mathbb{E}[Y|X] - Y|^2\big] \\
&= \mathbb{E}\big[|f(X) - \mathbb{E}[Y|X]|^2\big] + \mathbb{E}\big[|\mathbb{E}[Y|X] - Y|^2\big].
\end{aligned} \tag{169}$$

This implies that for all $f, g \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ it holds that

$$\mathcal{E}(f) - \mathcal{E}(g) = \mathbb{E}\big[|f(X) - \mathbb{E}[Y|X]|^2\big] - \mathbb{E}\big[|g(X) - \mathbb{E}[Y|X]|^2\big]. \tag{170}$$

Hence, we obtain that for all $f, g \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ it holds that

$$\mathbb{E}\big[|f(X) - \mathbb{E}[Y|X]|^2\big] = \mathbb{E}\big[|g(X) - \mathbb{E}[Y|X]|^2\big] + \mathcal{E}(f) - \mathcal{E}(g). \tag{171}$$

Combining (169)–(171) establishes items (i)–(iii). The proof of Lemma 4.1 is thus completed. □

4.2 Overall error decomposition

Lemma 4.2. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $d, M \in \mathbb{N}$, let $D \subseteq \mathbb{R}^d$ be a compact set, let $\mathcal{H} \subseteq C(D, \mathbb{R})$ be a set, let $(X_m, Y_m) = (X_{m,1}, \ldots, X_{m,d}, Y_m) \colon \Omega \to (D \times \mathbb{R})$, $m \in \{1, 2, \ldots, M\}$, be i.i.d. random variables with $\mathbb{E}\big[|Y_1|^2\big] < \infty$, let $\mathcal{E} \colon C(D, \mathbb{R}) \to [0, \infty)$ satisfy for all $f \in C(D, \mathbb{R})$ that $\mathcal{E}(f) = \mathbb{E}[|f(X_1) - Y_1|^2]$, let $\mathfrak{E} \colon C(D, \mathbb{R}) \times \Omega \to [0, \infty)$ satisfy for all $f \in C(D, \mathbb{R})$, $\omega \in \Omega$ that*

$$\mathfrak{E}(f, \omega) = \frac{1}{M}\left[\sum_{m=1}^{M} |f(X_m(\omega)) - Y_m(\omega)|^2\right], \tag{172}$$

and let $\phi \in \mathcal{H}$ satisfy $\mathcal{E}(\phi) = \inf_{f \in \mathcal{H}} \mathcal{E}(f)$. Then it holds for all $f \in \mathcal{H}$ that

$$\begin{aligned}
\mathbb{E}\big[|f(X_1) - \mathbb{E}[Y_1|X_1]|^2\big] &= \mathbb{E}\big[|\phi(X_1) - \mathbb{E}[Y_1|X_1]|^2\big] + \mathcal{E}(f) - \mathcal{E}(\phi) \\
&\leq \mathbb{E}\big[|\phi(X_1) - \mathbb{E}[Y_1|X_1]|^2\big] + \big[\mathfrak{E}(f) - \mathfrak{E}(\phi)\big] + 2\left[\sup_{v \in \mathcal{H}} |\mathfrak{E}(v) - \mathcal{E}(v)|\right].
\end{aligned} \tag{173}$$

Proof of Lemma 4.2. Note that Lemma 4.1 ensures that for all $f \in C(D, \mathbb{R})$ it holds that

$$
\begin{aligned}
&\mathbb{E}\big[|f(X_1) - \mathbb{E}[Y_1|X_1]|^2\big] \\
&= \mathbb{E}\big[|\phi(X_1) - \mathbb{E}[Y_1|X_1]|^2\big] + \mathcal{E}(f) - \mathcal{E}(\phi) \\
&= \mathbb{E}\big[|\phi(X_1) - \mathbb{E}[Y_1|X_1]|^2\big] + \mathcal{E}(f) - \mathfrak{E}(f) + \mathfrak{E}(f) - \mathfrak{E}(\phi) + \mathfrak{E}(\phi) - \mathcal{E}(\phi) \\
&= \underbrace{\mathbb{E}\big[|\phi(X_1) - \mathbb{E}[Y_1|X_1]|^2\big]}_{\text{Approximation error}} + \underbrace{\big[\big(\mathcal{E}(f) - \mathfrak{E}(f)\big) + \big(\mathfrak{E}(\phi) - \mathcal{E}(\phi)\big)\big]}_{\text{Statistical error}} + \underbrace{\big(\mathfrak{E}(f) - \mathfrak{E}(\phi)\big)}_{\text{Optimization error}} \\
&\leq \mathbb{E}\big[|\phi(X_1) - \mathbb{E}[Y_1|X_1]|^2\big] + \left[\sum_{v \in \{f, \phi\}} |\mathfrak{E}(v) - \mathcal{E}(v)|\right] + \big[\mathfrak{E}(f) - \mathfrak{E}(\phi)\big] \\
&\leq \mathbb{E}\big[|\phi(X_1) - \mathbb{E}[Y_1|X_1]|^2\big] + 2\left[\max_{v \in \{f, \phi\}} |\mathfrak{E}(v) - \mathcal{E}(v)|\right] + \big[\mathfrak{E}(f) - \mathfrak{E}(\phi)\big].
\end{aligned} \tag{174}
$$

The proof of Lemma 4.2 is thus completed. ☐

Lemma 4.3. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $d, \mathfrak{d}, M \in \mathbb{N}$, let $D \subseteq \mathbb{R}^d$ be a compact set, let $B \subseteq \mathbb{R}^{\mathfrak{d}}$ be a set, let $H \colon B \to C(D, \mathbb{R})$ be a function, let $(X_m, Y_m) = (X_{m,1}, \ldots, X_{m,d}, Y_m) \colon \Omega \to (D \times \mathbb{R})$, $m \in \{1, 2, \ldots, M\}$, be i.i.d. random variables with $\mathbb{E}[|Y_1|^2] < \infty$, let $\varphi \colon D \to \mathbb{R}$ be $\mathcal{B}(D)/\mathcal{B}(\mathbb{R})$-measurable, assume that it holds $\mathbb{P}$-a.s. that $\varphi(X_1) = \mathbb{E}[Y_1|X_1]$, let $\mathcal{E} \colon C(D, \mathbb{R}) \to [0, \infty)$ satisfy for all $f \in C(D, \mathbb{R})$ that $\mathcal{E}(f) = \mathbb{E}|f(X_1) - Y_1|^2$, let $\mathfrak{E} \colon B \times \Omega \to [0, \infty)$ satisfy for all $\theta \in B$, $\omega \in \Omega$ that*

$$
\mathfrak{E}(\theta, \omega) = \frac{1}{M}\left[\sum_{m=1}^{M} |H_\theta(X_m(\omega)) - Y_m(\omega)|^2\right], \tag{175}
$$

and let $\vartheta \in B$ satisfy $\mathcal{E}(H_\vartheta) = \inf_{\theta \in B} \mathcal{E}(H_\theta)$. Then it holds for all $\theta \in B$ that

$$
\begin{aligned}
&\int_D |H_\theta(x) - \varphi(x)|^2 \, \mathbb{P}_{X_1}(dx) = \int_D |H_\vartheta(x) - \varphi(x)|^2 \, \mathbb{P}_{X_1}(dx) + \mathcal{E}(H_\theta) - \mathcal{E}(H_\vartheta) \\
&\leq \int_D |H_\vartheta(x) - \varphi(x)|^2 \, \mathbb{P}_{X_1}(dx) + \mathfrak{E}(\theta) - \mathfrak{E}(\vartheta) + 2\left[\sup_{\eta \in B} |\mathfrak{E}(\eta) - \mathcal{E}(H_\eta)|\right].
\end{aligned} \tag{176}
$$

Proof of Lemma 4.3. First, observe that the assumption that it holds $\mathbb{P}$-a.s. that $\varphi(X_1) = \mathbb{E}[Y_1|X_1]$ ensures that for all $\theta \in B$ it holds that

$$
\int_D |H_\theta(x) - \varphi(x)|^2 \, \mathbb{P}_{X_1}(dx) = \mathbb{E}\big[|H_\theta(X_1) - \varphi(X_1)|^2\big] = \mathbb{E}\big[|H_\theta(X_1) - \mathbb{E}[Y_1|X_1]|^2\big]. \tag{177}
$$

Lemma 4.2 (with $(\Omega, \mathcal{A}, \mathbb{P}) = (\Omega, \mathcal{A}, \mathbb{P})$, $d = d$, $M = M$, $D = D$, $\mathcal{H} = \{H_\theta \colon \theta \in B\}$, $X_m = X_m$, $Y_m = Y_m$, $\mathcal{E} = \mathcal{E}$, $\mathfrak{E} = (C(D, \mathbb{R}) \times \Omega \ni (f, \omega) \mapsto \frac{1}{M}\sum_{m=1}^{M} |f(X_m(\omega)) - Y_m(\omega)|^2) \in [0, \infty))$, $\phi = H_\vartheta$ for $m \in \{1, 2, \ldots, M\}$ in the notation of Lemma 4.2) hence establishes (176). The proof of Lemma 4.3 is thus completed. ☐

4.3 Analysis of the convergence speed

4.3.1 Convergence rates for convergence in probability

Lemma 4.4. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $d, \mathfrak{d} \in \mathbb{N}$, $u \in \mathbb{R}$, $v \in (u, \infty)$, let $B \subseteq \mathbb{R}^{\mathfrak{d}}$ be a compact set, let $\mathfrak{l} \in \big(\bigcup_{s \in \mathbb{N} \cap [2, \infty)} \mathbb{N}^s\big)$, and let $(X, Y) \colon \Omega \to (\mathbb{R}^d \times [u, v])$ be a random variable. Then*

(i) it holds that $B \ni \theta \mapsto \mathbb{E}\big[|\mathscr{N}^{\theta, \mathfrak{l}}_{u,v}(X) - Y|^2\big] \in [0, \infty)$ is continuous and

(ii) there exists $\vartheta \in B$ such that $\mathbb{E}\big[|\mathscr{N}^{\vartheta, \mathfrak{l}}_{u,v}(X) - Y|^2\big] = \inf_{\theta \in B} \mathbb{E}\big[|\mathscr{N}^{\theta, \mathfrak{l}}_{u,v}(X) - Y|^2\big]$

(cf. Definition 2.8).

Proof of Lemma 4.4. First, observe that Corollary 2.12 ensures for all $\omega \in \Omega$ that $B \ni \theta \mapsto |\mathscr{N}^{\theta,\mathfrak{l}}_{u,v}(X(\omega)) - Y(\omega)|^2 \in [0,\infty)$ is continuous. Combining this with Lebesgue's dominated convergence theorem and the fact that for all $\omega \in \Omega$ it holds that $|\mathscr{N}^{\theta,\mathfrak{l}}_{u,v}(X(\omega)) - Y(\omega)|^2 \in [0,(v-u)^2]$ establishes item (i). Next note that item (i) and the assumption that $B \subseteq \mathbb{R}^{\mathfrak{d}}$ is compact establish item (ii). The proof of Lemma 4.4 is thus completed. ☐

Theorem 4.5. *Let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space, let $d,\mathfrak{d},K,l,M \in \mathbb{N}$, $\varepsilon \in (0,\infty)$, $L,R \in [0,\infty)$, $u \in \mathbb{R}$, $v \in (u,\infty)$, let $B = [-R,R]^{\mathfrak{d}}$, let $\|\cdot\| \colon \mathbb{R}^d \to [0,\infty)$ be the standard norm on $\mathbb{R}^d$, let $D \subseteq \mathbb{R}^d$ be a compact set, let $\mathcal{M} \subseteq D$ satisfy $|\mathcal{M}| \in [2,\infty)\cap\mathbb{N}$, assume that $\sup_{x=(x_1,\ldots,x_d)\in D}[\inf_{z=(z_1,\ldots,z_d)\in\mathcal{M}}(\sum_{i=1}^d |x_i - z_i|)] \leq \frac{\varepsilon}{4L}$, let $\mathfrak{l} = (\mathfrak{l}_0,\mathfrak{l}_1,\ldots,\mathfrak{l}_l) \in \mathbb{N}^{l+1}$, assume that $l \geq |\mathcal{M}|+1$, $\mathfrak{l}_0 = d$, $\mathfrak{l}_1 \geq 2d|\mathcal{M}|$, and $\mathfrak{l}_l = 1$, assume for all $i \in \{2,3,\ldots,|\mathcal{M}|\}$ that $\mathfrak{l}_i \geq 2(|\mathcal{M}|-i)+3$, assume for all $i \in \{|\mathcal{M}|+1,\ldots,l-1\}$ that $\mathfrak{l}_i \geq 2$, let $(X_m,Y_m) = (X_{m,1},\ldots,X_{m,d},Y_m) \colon \Omega \to D \times [u,v]$, $m \in \{1,2,\ldots,M\}$, be i.i.d. random variables, let $\mathfrak{E} \colon B \times \Omega \to [0,\infty)$ satisfy for all $\theta \in B$, $\omega \in \Omega$ that*

$$\mathfrak{E}(\theta,\omega) = \frac{1}{M}\left[\sum_{m=1}^{M} |\mathscr{N}^{\theta,\mathfrak{l}}_{u,v}(X_m(\omega)) - Y_m(\omega)|^2\right] \tag{178}$$

(cf. Definition 2.8) let $\varphi \colon D \to \mathbb{R}$ satisfy $\mathbb{P}$-a.s. that $\varphi(X_1) = \mathbb{E}[Y_1|X_1]$, assume for all $x,y \in D$ that $|\varphi(x)-\varphi(y)| \leq L\|x-y\|$, assume that $R \geq \max\{1,L,\sup_{z\in D}|||z|||, 2\sup_{z\in D}|\varphi(z)|\}$ and $\mathfrak{d} \geq \sum_{i=1}^l \mathfrak{l}_i(\mathfrak{l}_{i-1}+1)$, let $\Theta_k \colon \Omega \to B$, $k \in \{1,2,\ldots,K\}$, be i.i.d. random variables, assume that Θ_1 is continuous uniformly distributed on B, assume that $((X_m,Y_m))_{m\in\{1,\ldots,M\}}$ and $(\Theta_k)_{k\in\{1,\ldots,K\}}$ are independent, and let $\Xi \colon \Omega \to B$ satisfy $\Xi = \Theta_{\min\{k\in\{1,2,\ldots,K\}\colon \mathfrak{E}(\Theta_k)=\min_{l\in\{1,\ldots,K\}}\mathfrak{E}(\Theta_l)\}}$. Then

$$\begin{aligned}\mathbb{P}\bigg(\int_D |\mathscr{N}^{\Xi,\mathfrak{l}}_{u,v}(x)-\varphi(x)|^2\,\mathbb{P}_{X_1}(dx) > \varepsilon^2\bigg) &\leq \exp\bigg(-K\min\bigg\{1, \frac{\varepsilon^{2\mathfrak{d}}}{(16(v-u)l(|||\mathfrak{l}|||+1)^l R^{l+1})^{\mathfrak{d}}}\bigg\}\bigg)\\ &\quad + 2\exp\bigg(\mathfrak{d}\ln\bigg(\frac{128l(|||\mathfrak{l}|||+1)^l R^{l+1}(v-u)}{\varepsilon^2}\bigg) - \frac{\varepsilon^4 M}{32(v-u)^4}\bigg).\end{aligned} \tag{179}$$

Proof of Theorem 4.5. Throughout this proof let $\mathcal{E} \colon C(D,\mathbb{R}) \to [0,\infty)$ satisfy for all $f \in C(D,\mathbb{R})$ that $\mathcal{E}(f) = \mathbb{E}[|f(X_1)-Y_1|^2]$ and let $\vartheta \in B$ satisfy $\mathcal{E}(\mathscr{N}^{\vartheta,\mathfrak{l}}_{u,v}|_D) = \inf_{\theta\in B}\mathcal{E}(\mathscr{N}^{\theta,\mathfrak{l}}_{u,v}|_D)$ (cf. Lemma 4.4). Note that Lemma 4.3 (with $(\Omega,\mathcal{A},\mathbb{P}) \leftarrow (\Omega,\mathcal{A},\mathbb{P})$, $d \leftarrow d$, $\mathfrak{d} \leftarrow \mathfrak{d}$, $M \leftarrow M$, $D \leftarrow D$, $B \leftarrow B$, $H \leftarrow (B \ni \theta \mapsto (D \ni x \mapsto \mathscr{N}^{\theta,\mathfrak{l}}_{u,v}(x) \in \mathbb{R}) \in C(D,\mathbb{R}))$, $X_m \leftarrow X_m$, $Y_m \leftarrow Y_m$, $\varphi \leftarrow \varphi$, $\mathcal{E} \leftarrow \mathcal{E}$, $\mathfrak{E} \leftarrow \mathfrak{E}$, $\vartheta \leftarrow \vartheta$ for $m \in \{1,2,\ldots,M\}$ in the notation of Lemma 4.3) ensures for all $\omega \in \Omega$ that

$$\begin{aligned}&\int_D |\mathscr{N}^{\Xi(\omega),\mathfrak{l}}_{u,v}(x)-\varphi(x)|^2\,\mathbb{P}_{X_1}(dx)\\ &\leq \int_D \left|\mathscr{N}^{\vartheta,\mathfrak{l}}_{u,v}(x)-\varphi(x)\right|^2\mathbb{P}_{X_1}(dx) + \Big(\mathfrak{E}(\Xi(\omega)) - \mathfrak{E}(\vartheta)\Big) + 2\left[\sup_{\theta\in B}|\mathfrak{E}(\theta) - \mathcal{E}(\mathscr{N}^{\theta,\mathfrak{l}}_{u,v}|_D)|\right].\end{aligned} \tag{180}$$

Hence, we obtain that

$$\begin{aligned}&\mathbb{P}\bigg(\int_D |\mathscr{N}^{\Xi,\mathfrak{l}}_{u,v}(x)-\varphi(x)|^2\,\mathbb{P}_{X_1}(dx) > \varepsilon^2\bigg)\\ &\leq \mathbb{P}\bigg(\int_D \left|\mathscr{N}^{\vartheta,\mathfrak{l}}_{u,v}(x)-\varphi(x)\right|^2\mathbb{P}_{X_1}(dx) + \Big(\mathfrak{E}(\Xi)-\mathfrak{E}(\vartheta)\Big) + 2\left[\sup_{\theta\in B}|\mathfrak{E}(\theta)-\mathcal{E}(\mathscr{N}^{\theta,\mathfrak{l}}_{u,v}|_D)|\right] > \varepsilon^2\bigg)\\ &\leq \mathbb{P}\bigg(\int_D \left|\mathscr{N}^{\vartheta,\mathfrak{l}}_{u,v}(x)-\varphi(x)\right|^2\mathbb{P}_{X_1}(dx) > \tfrac{\varepsilon^2}{4}\bigg) + \mathbb{P}\Big(\mathfrak{E}(\Xi)-\mathfrak{E}(\vartheta) > \tfrac{\varepsilon^2}{4}\Big) + \mathbb{P}\bigg(\sup_{\theta\in B}|\mathfrak{E}(\theta)-\mathcal{E}(\mathscr{N}^{\theta,\mathfrak{l}}_{u,v}|_D)| > \tfrac{\varepsilon^2}{4}\bigg).\end{aligned} \tag{181}$$

Next observe that the assumption that for all $x,y \in D$ it holds that $|\varphi(x)-\varphi(y)| \leq L\|x-y\|$ implies that for all $x = (x_1,\ldots,x_d), y = (y_1,\ldots,y_d) \in D$ it holds that $|\varphi(x)-\varphi(y)| \leq L\big[\sum_{i=1}^d |x_i - y_i|\big]$. This, the assumption that $\mathfrak{d} \geq \sum_{i=1}^{l} \mathfrak{l}_i(\mathfrak{l}_{i-1}+1)$, the assumption that $R \geq \max\{1, L, \sup_{z\in D} |||z|||, 2\sup_{z\in D} |\varphi(z)|\}$, and Corollary 3.8 (with $d \leftarrow d$, $\mathfrak{d} \leftarrow \mathfrak{d}$, $a \leftarrow a$, $b \leftarrow b$, $L \leftarrow L$, $A \leftarrow D$, $f \leftarrow \varphi$, $\mathcal{M} \leftarrow \mathcal{M}$ in the notation of Corollary 3.8) ensure that there exists $\eta \in B$ which satisfies that

$$\sup_{x\in D} |\mathscr{N}_{u,v}^{\eta,\mathfrak{l}}(x) - \varphi(x)| \leq 2L \left[\sup_{x=(x_1,\dots,x_d)\in D} \left(\inf_{z=(z_1,\dots,z_d)\in\mathcal{M}} \sum_{i=1}^{d} |x_i - z_i| \right) \right] \leq 2L \frac{\varepsilon}{4L} = \frac{\varepsilon}{2}. \tag{182}$$

Lemma 4.3 (with $(\Omega, \mathcal{A}, \mathbb{P}) \leftarrow (\Omega, \mathcal{A}, \mathbb{P})$, $d \leftarrow d$, $\mathfrak{d} \leftarrow \mathfrak{d}$, $M \leftarrow M$, $D \leftarrow D$, $B \leftarrow B$, $H = (B \ni \theta \mapsto (D \ni x \mapsto \mathscr{N}_{u,v}^{\theta,\mathfrak{l}}(x) \in \mathbb{R}) \in C(D,\mathbb{R}))$, $X_m \leftarrow X_m$, $Y_m \leftarrow Y_m$, $\varphi \leftarrow \varphi$, $\mathcal{E} \leftarrow \mathcal{E}$, $\mathfrak{E} \leftarrow \mathfrak{E}$, $\vartheta \leftarrow \vartheta$ for $m \in \{1, 2, \dots, M\}$ in the notation of Lemma 4.3) hence ensures that

$$\begin{split} &\int_D \left|\mathscr{N}_{u,v}^{\vartheta,\mathfrak{l}}(x) - \varphi(x)\right|^2 \mathbb{P}_{X_1}(dx) = \int_D \left|\mathscr{N}_{u,v}^{\eta,\mathfrak{l}}(x) - \varphi\right|^2 \mathbb{P}_{X_1}(dx) + \underbrace{\mathcal{E}(\mathscr{N}_{u,v}^{\vartheta,\mathfrak{l}}|_D) - \mathcal{E}(\mathscr{N}_{u,v}^{\theta,\mathfrak{l}}|_D)}_{\leq 0} \\ &\leq \int_D \left|\mathscr{N}_{u,v}^{\eta,\mathfrak{l}}(x) - \varphi(x)\right|^2 \mathbb{P}_{X_1}(dx) \leq \sup_{x\in D} |\mathscr{N}_{u,v}^{\eta,\mathfrak{l}}(x) - \varphi(x)|^2 \leq \frac{\varepsilon^2}{4}. \end{split} \tag{183}$$

This implies that

$$\mathbb{P}\left(\int_D \left|\mathscr{N}_{u,v}^{\vartheta,\mathfrak{l}}(x) - \varphi(x)\right|^2 \mathbb{P}_{X_1}(dx) > \frac{\varepsilon^2}{4} \right) = 0. \tag{184}$$

Note that (181) therefore ensures that

$$\begin{split} &\mathbb{P}\left(\int_D |\mathscr{N}_{u,v}^{\Xi,\mathfrak{l}}(x) - \varphi(x)|^2 \, \mathbb{P}_{X_1}(dx) > \varepsilon^2 \right) \\ &\leq \mathbb{P}\left(\mathfrak{E}(\Xi) - \mathfrak{E}(\vartheta) > \frac{\varepsilon^2}{4} \right) + \mathbb{P}\left(\sup_{\theta\in B} \left|\mathfrak{E}(\theta) - \mathcal{E}(\mathscr{N}_{u,v}^{\theta,\mathfrak{l}}|_D)\right| > \frac{\varepsilon^2}{4} \right). \end{split} \tag{185}$$

Next observe that Corollary 2.12 (with $a \leftarrow -\sup\{|||z|||\colon z \in D\}$, $b \leftarrow \sup\{|||z|||\colon z \in D\}$, $u \leftarrow u$, $v \leftarrow v$, $L \leftarrow l$, $d \leftarrow \mathfrak{d}$, $l \leftarrow \mathfrak{l}$ in the notation of Corollary 2.12) ensures for all $\theta, \xi \in B$ that

$$\begin{split} &\sup_{x\in D} |||\mathscr{N}_{u,v}^{\theta,\mathfrak{l}}(x) - \mathscr{N}_{u,v}^{\xi,\mathfrak{l}}(x)||| \\ &\leq \sup\bigl\{|||\mathscr{N}_{u,v}^{\theta,\mathfrak{l}}(x) - \mathscr{N}_{u,v}^{\xi,\mathfrak{l}}(x)|||\colon x \in [-\sup\{|||z|||\colon z \in D\}, \sup\{|||z|||\colon z \in D\}]^d\bigr\} \\ &\leq l \max\{1, \sup\{|||z|||\colon z \in D\}\}(|||\mathfrak{l}||| + 1)^l (\max\{|||\theta|||, |||\xi|||\})^{l-1} |||\theta - \xi||| \\ &\leq l \max\{1, \sup\{|||z|||\colon z \in D\}\}(|||\mathfrak{l}||| + 1)^l R^{l-1} |||\theta - \xi||| \\ &\leq l(|||\mathfrak{l}||| + 1)^l R^l |||\theta - \xi|||. \end{split} \tag{186}$$

Combining this with the fact that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in D$ it holds that $\mathscr{N}_{u,v}^{\theta,\mathfrak{l}}(x) \in [u, v]$, with the assumption that $Y_1 \in [u, v]$, and with (178) ensures for all $\theta, \xi \in B$, $\omega \in \Omega$ that

$$\begin{split} &|\mathfrak{E}(\theta, \omega) - \mathfrak{E}(\xi, \omega)| \\ &= \left| \frac{1}{M} \left[\sum_{m=1}^{M} |\mathscr{N}_{u,v}^{\theta,\mathfrak{l}}(X_m(\omega)) - Y_m(\omega)|^2 \right] - \frac{1}{M} \left[\sum_{m=1}^{M} |\mathscr{N}_{u,v}^{\xi,\mathfrak{l}}(X_m(\omega)) - Y_m(\omega)|^2 \right] \right| \\ &\leq \frac{1}{M} \left[\sum_{m=1}^{M} \bigl(\left|\mathscr{N}_{u,v}^{\theta,\mathfrak{l}}(X_m(\omega)) - \mathscr{N}_{u,v}^{\xi,\mathfrak{l}}(X_m(\omega))\right| \left|\mathscr{N}_{u,v}^{\theta,\mathfrak{l}}(X_m(\omega)) + \mathscr{N}_{u,v}^{\xi,\mathfrak{l}}(X_m(\omega)) - 2Y_m(\omega)\right| \bigr) \right] \\ &\leq \frac{1}{M} \left[\sum_{m=1}^{M} \bigl(\left|\mathscr{N}_{u,v}^{\theta,\mathfrak{l}}(X_m(\omega)) - \mathscr{N}_{u,v}^{\xi,\mathfrak{l}}(X_m(\omega))\right| \underbrace{\bigl[\left|\mathscr{N}_{u,v}^{\theta,\mathfrak{l}}(X_m(\omega)) - Y_m(\omega)\right| + \left|\mathscr{N}_{u,v}^{\xi,\mathfrak{l}}(X_m(\omega)) - Y_m(\omega)\right| \bigr]}_{\leq 2(v-u)} \bigr) \right] \\ &\leq 2(v - u) l (|||\mathfrak{l}||| + 1)^l R^l |||\theta - \xi|||. \end{split} \tag{187}$$

Lemma 3.22 (with $(\Omega,\mathcal{A},\mathbb{P})\leftarrow(\Omega,\mathcal{A},\mathbb{P})$, $\mathfrak{d}\leftarrow\mathfrak{d}$, $N\leftarrow K$, $a\leftarrow -R$, $b\leftarrow R$, $L\leftarrow 2(v-u)l(|||\mathfrak{l}|||+1)^l R^l$, $\vartheta\leftarrow\vartheta$, $\mathfrak{E}\leftarrow\mathfrak{E}$, $\Theta_k\leftarrow\Theta_k$ for $k\in\{1,2,\dots,K\}$ in the notation of Lemma 3.22) therefore shows that

$$\begin{aligned}\mathbb{P}\bigg(\mathfrak{E}(\Xi)-\mathfrak{E}(\vartheta)>\frac{\varepsilon^2}{4}\bigg)&=\mathbb{P}\bigg(\Big[\min_{k\in\{1,2,\dots,K\}}\mathfrak{E}(\Theta_k)\Big]-\mathfrak{E}(\vartheta)>\frac{\varepsilon^2}{4}\bigg)\\&\le\exp\bigg(-K\min\bigg\{1,\frac{\varepsilon^{2\mathfrak{d}}}{(16(v-u)l(|||\mathfrak{l}|||+1)^lR^{l+1})^{\mathfrak{d}}}\bigg\}\bigg).\end{aligned}\tag{188}$$

Combining this with (185) ensures that

$$\begin{aligned}&\mathbb{P}\bigg(\int_D|\mathscr{N}^{\Xi,\mathfrak{l}}_{u,v}(x)-\varphi(x)|^2\,\mathbb{P}_{X_1}(dx)>\varepsilon^2\bigg)\\&\le\exp\bigg(-K\min\bigg\{1,\frac{\varepsilon^{2\mathfrak{d}}}{(16(v-u)l(|||\mathfrak{l}|||+1)^lR^{l+1})^{\mathfrak{d}}}\bigg\}\bigg)+\mathbb{P}\bigg(\sup_{\theta\in B}\big|\mathfrak{E}(\theta)-\mathcal{E}(\mathscr{N}^{\theta,\mathfrak{l}}_{u,v}|_D)\big|>\frac{\varepsilon^2}{4}\bigg).\end{aligned}\tag{189}$$

Next observe that Lemma 3.20 (with $\mathfrak{d}\leftarrow\mathfrak{d}$, $M\leftarrow M$, $L\leftarrow l$, $\varepsilon\leftarrow\frac{\varepsilon^2}{4}$, $u\leftarrow u$, $v\leftarrow v$, $a\leftarrow -\sup\{|||z|||\colon z\in D\}$, $b\leftarrow\sup\{|||z|||\colon z\in D\}$, $R\leftarrow R$, $\mathfrak{l}\leftarrow\mathfrak{l}$, $B\leftarrow B$, $(\Omega,\mathcal{F},\mathbb{P})\leftarrow(\Omega,\mathcal{A},\mathbb{P})$, $((X_m,Y_m))_{m\in\{1,2,\dots,M\}}\leftarrow((X_m,Y_m))_{m\in\{1,2,\dots,M\}}$, $\mathcal{E}\leftarrow\mathcal{E}$, $\mathfrak{E}\leftarrow\mathfrak{E}$ in the notation of Lemma 3.20) establishes that

$$\begin{aligned}&\mathbb{P}\bigg(\big[\sup\nolimits_{\theta\in B}|\mathfrak{E}(\theta)-\mathcal{E}(\mathscr{N}^{\theta,\mathfrak{l}}_{u,v}|_D)|\big]>\frac{\varepsilon^2}{4}\bigg)\\&\le 2\exp\bigg(\mathfrak{d}\ln\bigg(\frac{128l\max\{1,\sup\{|||z|||\colon z\in D\}\}(|||\mathfrak{l}|||+1)^lR^l(v-u)}{\varepsilon^2}\bigg)-\frac{\varepsilon^4M}{32(v-u)^4}\bigg)\\&\le 2\exp\bigg(\mathfrak{d}\ln\bigg(\frac{128l(|||\mathfrak{l}|||+1)^lR^{l+1}(v-u)}{\varepsilon^2}\bigg)-\frac{\varepsilon^4M}{32(v-u)^4}\bigg).\end{aligned}\tag{190}$$

Combining this with (189) proves that

$$\begin{aligned}\mathbb{P}\bigg(\int_D|\mathscr{N}^{\Xi,\mathfrak{l}}_{u,v}(x)-\varphi(x)|^2\,\mathbb{P}_{X_1}(dx)>\varepsilon^2\bigg)&\le\exp\bigg(-K\min\bigg\{1,\frac{\varepsilon^{2\mathfrak{d}}}{(16(v-u)l(|||\mathfrak{l}|||+1)^lR^{l+1})^{\mathfrak{d}}}\bigg\}\bigg)\\&\quad+2\exp\bigg(\mathfrak{d}\ln\bigg(\frac{128l(|||\mathfrak{l}|||+1)^lR^{l+1}(v-u)}{\varepsilon^2}\bigg)-\frac{\varepsilon^4M}{32(v-u)^4}\bigg).\end{aligned}\tag{191}$$

This establishes (179). The proof of Theorem 4.5 is thus completed. □

Corollary 4.6. *Let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space, let $d,\mathfrak{d},K,M,\tau\in\mathbb{N}$, $\varepsilon\in(0,\infty)$, $L,a,u\in\mathbb{R}$, $b\in(a,\infty)$, $v\in(u,\infty)$, $R\in[\max\{1,L,|a|,|b|,2|u|,2|v|\},\infty)$, let $B=[-R,R]^{\mathfrak{d}}$, let $\|\cdot\|\colon\mathbb{R}^d\to[0,\infty)$ be the standard norm on $\mathbb{R}^d$, assume that $[2d(2dL(b-a)\varepsilon^{-1}+2)^d+2]^3\le\tau^3\le\mathfrak{d}$, let $\mathfrak{l}\in\mathbb{N}^\tau$ satisfy $\mathfrak{l}=(d,\tau,\tau,\dots,\tau,1)$, let $X_m\colon\Omega\to[a,b]^d$, $m\in\{1,2,\dots,M\}$, be i.i.d. random variables, let $\varphi\colon[a,b]^d\to[u,v]$ satisfy for all $x,y\in[a,b]^d$ that $|\varphi(x)-\varphi(y)|\le L\|x-y\|$, let $\mathfrak{E}\colon B\times\Omega\to[0,\infty)$ satisfy for all $\theta\in B$, $\omega\in\Omega$ that*

$$\mathfrak{E}(\theta,\omega)=\frac{1}{M}\bigg[\sum_{m=1}^{M}|\mathscr{N}^{\theta,\mathfrak{l}}_{u,v}(X_m(\omega))-\varphi(X_m(\omega))|^2\bigg],\tag{192}$$

let $\Theta_k\colon\Omega\to B$, $k\in\{1,2,\dots,K\}$, be i.i.d. random variables, assume that Θ_1 is continuous uniformly distributed on B, assume that $(X_m)_{m\in\{1,2,\dots,M\}}$ and $(\Theta_k)_{k\in\{1,2,\dots,K\}}$ are independent, and let $\Xi\colon\Omega\to B$ satisfy $\Xi=\Theta_{\min\{k\in\{1,2,\dots,K\}\colon\mathfrak{E}(\Theta_k)=\min_{l\in\{1,2,\dots,K\}}\mathfrak{E}(\Theta_l)\}}$ (cf. Definition 2.8). Then

$$\begin{aligned}\mathbb{P}\bigg(\bigg[\int_{[a,b]^d}|\mathscr{N}^{\Xi,\mathfrak{l}}_{u,v}(x)-\varphi(x)|^2\,\mathbb{P}_{X_1}(dx)\bigg]^{1/2}>\varepsilon\bigg)&\le\exp\bigg(-K\min\bigg\{1,\frac{\varepsilon^{2\mathfrak{d}}}{(16(v-u)\tau(\tau+1)^\tau R^\tau)^{\mathfrak{d}}}\bigg\}\bigg)\\&\quad+2\exp\bigg(\mathfrak{d}\ln\bigg(\frac{128\tau(\tau+1)^\tau R^\tau(v-u)}{\varepsilon^2}\bigg)-\frac{\varepsilon^4M}{32(v-u)^4}\bigg)\end{aligned}\tag{193}$$

Proof of Corollary 4.6. Throughout this proof let $D = [a,b]^d$, let $N \in \mathbb{N}$ satisfy

$$N = \min\left\{k \in \mathbb{N} \colon k \geq \frac{2dL(b-a)}{\varepsilon}\right\}, \tag{194}$$

and let $\mathcal{M} = \{a, a+\frac{1}{N}(b-a), \ldots, a+\frac{N-1}{N}(b-a), b\}^d$. Note that $|\mathcal{M}| = (N+1)^d \in [2,\infty)$. Moreover, note that

$$\begin{split}\sum_{i=1}^{\tau-1} \mathfrak{l}_i(\mathfrak{l}_{i-1}+1) &= \mathfrak{l}_1(\mathfrak{l}_0+1) + \sum_{i=2}^{\tau-2} \mathfrak{l}_i(\mathfrak{l}_{i-1}+1) + \mathfrak{l}_l(\mathfrak{l}_{l-1}+1) \\ &= \tau(d+1) + \sum_{i=2}^{\tau-2} \tau(\tau+1) + \tau + 1 \leq (\tau-1)\tau(\tau+1) \leq \tau^3 \leq \mathfrak{d}.\end{split} \tag{195}$$

Furthermore, observe that for all $x = (x_1,\ldots,x_d) \in [u,v]^d$ there exists $z = (z_1,\ldots,z_d) \in \mathcal{M}$ which satisfies for all $i \in \{1,2,\ldots,d\}$ that $|x_i - z_i| \leq \frac{b-a}{2N}$. This implies that

$$\sup_{x=(x_1,\ldots,x_d)\in[a,b]^d} \left[\inf_{z=(z_1,\ldots,z_d)\in\mathcal{M}} \left(\sum_{i=1}^d |x_i - z_i|\right)\right] \leq \frac{d(b-a)}{2N} \leq \frac{\varepsilon}{4L}. \tag{196}$$

Next note that the assumption that $\tau^3 \geq [2d(\frac{2dL(b-a)}{\varepsilon}+2)^d+2]^3$ ensures that $\tau \geq 2d(N+1)^d+2 = 2d|\mathcal{M}|+2$. Moreover, observe that it holds that $\mathfrak{l}_0 = d$, $\mathfrak{l}_1 = \tau \geq 2d|\mathcal{M}|$, and $\mathfrak{l}_{\tau-1} = 1$. In addition, note that

(i) for all $i \in \{2,3,\ldots,|\mathcal{M}|\}$ it holds that $\mathfrak{l}_i = \tau \geq 2|\mathcal{M}| \geq 2(|\mathcal{M}|-i)+3$ and

(ii) for all $i \in \{|\mathcal{M}|+1,\ldots,\tau-1\}$ it holds that $\mathfrak{l}_i = \tau \geq 2$.

The fact that $\tau - 1 \geq |\mathcal{M}|+1$, (195), (196), and Theorem 4.5 (with $(\Omega,\mathcal{A},\mathbb{P}) \leftarrow (\Omega,\mathcal{A},\mathbb{P})$, $d \leftarrow d$, $\mathfrak{d} \leftarrow \mathfrak{d}$, $K \leftarrow K$, $l \leftarrow \tau-1$, $M \leftarrow M$, $\varepsilon \leftarrow \varepsilon$, $L \leftarrow L$, $R \leftarrow R$, $u \leftarrow u$, $v \leftarrow v$, $B \leftarrow B$, $D \leftarrow D$, $\mathcal{M} \leftarrow \mathcal{M}$, $\mathfrak{l} \leftarrow \mathfrak{l}$, $((X_m,Y_m))_{m\in\{1,2,\ldots,M\}} \leftarrow ((X_m,\varphi(X_m)))_{m\in\{1,2,\ldots,M\}}$, $\mathfrak{E} \leftarrow \mathfrak{E}$, $\varphi \leftarrow \varphi$, $(\Theta_k)_{k\in\{1,2,\ldots,K\}} \leftarrow (\Theta_k)_{k\in\{1,2,\ldots,K\}}$, $\Xi \leftarrow \Xi$ in the notation of Theorem 4.5) therefore ensure that

$$\begin{split}&\mathbb{P}\left(\left[\int_{[a,b]^d} |\mathscr{N}^{\Xi,\mathfrak{l}}_{u,v}(x) - \varphi(x)|^2 \,\mathbb{P}_{X_1}(dx)\right]^{1/2} > \varepsilon\right) \\ &\leq \exp\left(-K \min\left\{1, \frac{\varepsilon^{2\mathfrak{d}}}{(16(v-u)(\tau-1)(|||\mathfrak{l}|||+1)^{\tau-1}R^\tau)^{\mathfrak{d}}}\right\}\right) \\ &\quad + 2\exp\left(\mathfrak{d}\ln\left(\frac{128l(|||\mathfrak{l}|||+1)^{\tau-1}R^\tau(v-u)}{\varepsilon^2}\right) - \frac{\varepsilon^4 M}{32(v-u)^4}\right) \\ &= \exp\left(-K \min\left\{1, \frac{\varepsilon^{2\mathfrak{d}}}{(16(v-u)(\tau-1)(\tau+1)^{\tau-1}R^\tau)^{\mathfrak{d}}}\right\}\right) \\ &\quad + 2\exp\left(\mathfrak{d}\ln\left(\frac{128(\tau-1)(\tau+1)^{\tau-1}R^\tau(v-u)}{\varepsilon^2}\right) - \frac{\varepsilon^4 M}{32(v-u)^4}\right) \\ &\leq \exp\left(-K \min\left\{1, \frac{\varepsilon^{2\mathfrak{d}}}{(16(v-u)\tau(\tau+1)^{\tau}R^\tau)^{\mathfrak{d}}}\right\}\right) \\ &\quad + 2\exp\left(\mathfrak{d}\ln\left(\frac{128\tau(\tau+1)^{\tau}R^\tau(v-u)}{\varepsilon^2}\right) - \frac{\varepsilon^4 M}{32(v-u)^4}\right).\end{split} \tag{197}$$

This establishes (193). The proof of Corollary 4.6 is thus completed. ◻

Corollary 4.7. *Let $d \in \mathbb{N}$, $L, a, u \in \mathbb{R}$, $b \in (a,\infty)$, $v \in (u,\infty)$, $R \in [\max\{1, L, |a|, |b|, 2|u|, 2|v|\}, \infty)$, let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space, let $X_m \colon \Omega \to [a,b]^d$, $m \in \mathbb{N}$, be i.i.d. random variables, let $\|\cdot\| \colon \mathbb{R}^d \to$*

$[0,\infty)$ *be the standard norm on* $\mathbb{R}^d$*, let* $\varphi\colon [a,b]^d \to [u,v]$ *satisfy for all* $x,y \in [a,b]^d$ *that* $|\varphi(x)-\varphi(y)| \leq L\|x-y\|$*, let* $\mathfrak{l}_\tau \in \mathbb{N}^\tau$*,* $\tau \in \mathbb{N}$*, and* $\mathfrak{E}_{\mathfrak{d},M,\tau}\colon [-R,R]^{\mathfrak{d}} \times \Omega \to [0,\infty)$*,* $\mathfrak{d},M,\tau \in \mathbb{N}$*, satisfy for all* $\mathfrak{d},M \in \mathbb{N}$*,* $\tau \in \{3,4,\ldots\}$*,* $\theta \in [-R,R]^{\mathfrak{d}}$*,* $\omega \in \Omega$ *that* $\mathfrak{l}_\tau = (d,\tau,\tau,\ldots,\tau,1)$ *and*

$$\mathfrak{E}_{\mathfrak{d},M,\tau}(\theta,\omega) = \frac{1}{M}\left[\sum_{m=1}^{M} |\mathscr{N}_{u,v}^{\theta,\mathfrak{l}_\tau}(X_m(\omega)) - \varphi(X_m(\omega))|^2\right], \tag{198}$$

for every $\mathfrak{d} \in \mathbb{N}$ *let* $\Theta_{\mathfrak{d},k}\colon \Omega \to [-R,R]^{\mathfrak{d}}$*,* $k \in \mathbb{N}$*, be i.i.d. random variables, assume for every* $\mathfrak{d} \in \mathbb{N}$ *that* $\Theta_{\mathfrak{d},1}$ *is continuous uniformly distributed on* $[-R,R]^{\mathfrak{d}}$*, assume for every* $\mathfrak{d} \in \mathbb{N}$ *that* $(X_m)_{m\in\mathbb{N}}$ *and* $(\Theta_{\mathfrak{d},k})_{k\in\mathbb{N}}$ *are independent, and let* $\Xi_{\mathfrak{d},K,M,\tau}\colon \Omega \to [-R,R]^{\mathfrak{d}}$*,* $\mathfrak{d},K,M,\tau \in \mathbb{N}$*, satisfy for all* $\mathfrak{d},K,M,\tau \in \mathbb{N}$ *that* $\Xi_{\mathfrak{d},K,M,\tau} = \Theta_{\mathfrak{d},\min\{k\in\{1,2,\ldots,K\}\colon \mathfrak{E}_{\mathfrak{d},M,\tau}(\Theta_{\mathfrak{d},k}) = \min_{l\in\{1,\ldots,K\}} \mathfrak{E}_{\mathfrak{d},M,\tau}(\Theta_{\mathfrak{d},l})\}}$ *(cf. Definition 2.8). Then there exists* $c \in (0,\infty)$ *such that for all* $\mathfrak{d},K,M,\tau \in \mathbb{N}$*,* $\varepsilon \in (0,\infty)$ *with* $[2d(2dL(b-a)\varepsilon^{-1}+2)^d+2]^3 \leq \tau^3 \leq \mathfrak{d}$ *it holds that*

$$\begin{aligned}&\mathbb{P}\left(\left[\int_{[a,b]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau},\mathfrak{l}_\tau}(x) - \varphi(x)|^2 \,\mathbb{P}_{X_1}(dx)\right]^{1/2} > \varepsilon\right)\\ &\leq \exp\!\big(-K\min\big\{1,(c\tau)^{-\tau\mathfrak{d}}\varepsilon^{2\mathfrak{d}}\big\}\big) + 2\exp\!\Big(\mathfrak{d}\ln\!\big((c\tau)^\tau\varepsilon^{-2}\big) - \tfrac{\varepsilon^4 M}{c}\Big).\end{aligned} \tag{199}$$

Proof of Corollary 4.7. Throughout this proof let $c \in (0,\infty)$ satisfy

$$c = \max\{32(v-u)^4, 512(v-u+1)R\}. \tag{200}$$

Observe that Corollary 4.6 ensures that for all $\mathfrak{d},K,M,\tau \in \mathbb{N}$, $\varepsilon \in (0,\infty)$ with $[2d(\frac{2dL(b-a)}{\varepsilon}+2)^d+2]^3 \leq \tau^3 \leq \mathfrak{d}$ it holds that

$$\begin{aligned}\mathbb{P}\left(\left[\int_{[a,b]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau},\mathfrak{l}_\tau}(x) - \varphi(x)|^2 \,\mathbb{P}_{X_1}(dx)\right]^{1/2} > \varepsilon\right) &\leq \exp\!\left(-K\min\!\left\{1,\frac{\varepsilon^{2\mathfrak{d}}}{(16(v-u)\tau(\tau+1)^\tau R^\tau)^{\mathfrak{d}}}\right\}\right)\\ &\quad + 2\exp\!\left(\mathfrak{d}\ln\!\left(\frac{128\tau(\tau+1)^\tau R^\tau(v-u)}{\varepsilon^2}\right) - \frac{\varepsilon^4 M}{32(v-u)^4}\right).\end{aligned} \tag{201}$$

Note that (200) and the assumption that $\tau \geq 2$ ensure that

(i) $-\frac{1}{32(v-u)^4} \leq -\frac{1}{c}$,

(ii) $16(v-u)\tau(\tau+1)^\tau R^\tau \leq \tau(16(v-u+1)(\tau+1)R)^\tau \leq 2^\tau(32(v-u+1)R\tau)^\tau \leq (c\tau)^\tau$, and

(iii) $128\tau(\tau+1)^\tau R^\tau(v-u) \leq 128(2\tau)^\tau 2^\tau R^\tau(v-u) \leq (512R\tau(v-u+1))^\tau \leq (c\tau)^\tau$.

Combining this with (201) shows that for all $\mathfrak{d},K,M,\tau \in \mathbb{N}$, $\varepsilon \in (0,\infty)$ with $[2d(2dL(b-a)\varepsilon^{-1}+2)^d+2]^3 \leq \tau^3 \leq \mathfrak{d}$ it holds that

$$\begin{aligned}&\mathbb{P}\left(\left[\int_{[u,v]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau},\mathfrak{l}_\tau}(x) - \varphi(x)|^2 \,\mathbb{P}_{X_1}(dx)\right]^{1/2} > \varepsilon\right)\\ &\leq \exp\!\left(-K\min\!\left\{1,\frac{\varepsilon^{2\mathfrak{d}}}{(c\tau)^{\tau\mathfrak{d}}}\right\}\right) + 2\exp\!\left(\mathfrak{d}\ln\!\left(\frac{(c\tau)^\tau}{\varepsilon^2}\right) - \frac{\varepsilon^4 M}{c}\right).\end{aligned} \tag{202}$$

The proof of Corollary 4.7 is thus completed. ☐

Corollary 4.8. *Let* $d \in \mathbb{N}$*,* $L,a,u \in \mathbb{R}$*,* $b \in (a,\infty)$*,* $v \in (u,\infty)$*,* $R \in [\max\{1,L,|a|,|b|,2|u|,2|v|\},\infty)$*, let* $(\Omega,\mathcal{A},\mathbb{P})$ *be a probability space, let* $X_m\colon \Omega \to [a,b]^d$*,* $m \in \mathbb{N}$*, be i.i.d. random variables, let* $\|\cdot\|\colon \mathbb{R}^d \to [0,\infty)$ *be the standard norm on* $\mathbb{R}^d$*, let* $\varphi\colon [a,b]^d \to [u,v]$ *satisfy for all* $x,y \in [a,b]^d$ *that* $|\varphi(x)-\varphi(y)| \leq$

$L\|x-y\|$, *let* $\mathfrak{l}_\tau \in \mathbb{N}^\tau$, $\tau \in \mathbb{N}$, *and* $\mathfrak{E}_{\mathfrak{d},M,\tau} \colon [-R,R]^\mathfrak{d} \times \Omega \to [0,\infty)$, $\mathfrak{d}, M, \tau \in \mathbb{N}$, *satisfy for all* $\mathfrak{d}, M \in \mathbb{N}$, $\tau \in \{3,4,\ldots\}$, $\theta \in [-R,R]^\mathfrak{d}$, $\omega \in \Omega$ *that* $\mathfrak{l}_\tau = (d,\tau,\tau,\ldots,\tau,1)$ *and*

$$\mathfrak{E}_{\mathfrak{d},M,\tau}(\theta,\omega) = \frac{1}{M}\left[\sum_{m=1}^{M} |\mathscr{N}_{u,v}^{\theta,\mathfrak{l}_\tau}(X_m(\omega)) - \varphi(X_m(\omega))|^2\right], \tag{203}$$

for every $\mathfrak{d} \in \mathbb{N}$ *let* $\Theta_{\mathfrak{d},k} \colon \Omega \to [-R,R]^\mathfrak{d}$, $k \in \mathbb{N}$, *be i.i.d. random variables, assume for every* $\mathfrak{d} \in \mathbb{N}$ *that* $\Theta_{\mathfrak{d},1}$ *is continuous uniformly distributed on* $[-R,R]^\mathfrak{d}$, *assume for every* $\mathfrak{d} \in \mathbb{N}$ *that* $(X_m)_{m\in\mathbb{N}}$ *and* $(\Theta_{\mathfrak{d},k})_{k\in\mathbb{N}}$ *are independent, and let* $\Xi_{\mathfrak{d},K,M,\tau} \colon \Omega \to [-R,R]^\mathfrak{d}$, $\mathfrak{d}, K, M, \tau \in \mathbb{N}$, *satisfy for all* $\mathfrak{d}, K, M, \tau \in \mathbb{N}$ *that* $\Xi_{\mathfrak{d},K,M,\tau} = \Theta_{\mathfrak{d},\min\{k\in\{1,2,\ldots,K\} \colon \mathfrak{E}_{\mathfrak{d},M,\tau}(\Theta_{\mathfrak{d},k}) = \min_{l\in\{1,\ldots,K\}} \mathfrak{E}_{\mathfrak{d},M,\tau}(\Theta_{\mathfrak{d},l})\}}$ *(cf. Definition 2.8). Then there exists* $c \in (0,\infty)$ *such that for all* $\mathfrak{d}, K, M, \tau \in \mathbb{N}$, $\varepsilon \in (0,1]$ *with* $[2d(2dL(b-a)\varepsilon^{-1}+2)^d+2]^3 \le \tau^3 \le \mathfrak{d}$ *it holds that*

$$\begin{aligned}&\mathbb{P}\left(\int_{[a,b]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau},\mathfrak{l}_\tau}(x) - \varphi(x)|\,\mathbb{P}_{X_1}(dx) > \varepsilon\right)\\ &\le \exp\!\left(-K\min\left\{1,(c\tau)^{-\tau\mathfrak{d}}\varepsilon^{2\mathfrak{d}}\right\}\right) + 2\exp\!\left(\mathfrak{d}\ln\!\left((c\tau)^\tau\varepsilon^{-2}\right) - \tfrac{\varepsilon^4 M}{c}\right).\end{aligned} \tag{204}$$

Proof of Corollary 4.8. First, observe that Corollary 4.7 ensures that there exists $c \in (0,\infty)$ which satisfies that for all $\mathfrak{d}, K, M, \tau \in \mathbb{N}$, $\varepsilon \in (0,\infty)$ with $[2d(2dL(b-a)\varepsilon^{-1}+2)^d+2]^3 \le \tau^3 \le \mathfrak{d}$ it holds that

$$\begin{aligned}&\mathbb{P}\left(\left[\int_{[a,b]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau},\mathfrak{l}_\tau}(x) - \varphi(x)|^2\,\mathbb{P}_{X_1}(dx)\right]^{1/2} > \varepsilon\right)\\ &\le \exp\!\left(-K\min\left\{1,(c\tau)^{-\tau\mathfrak{d}}\varepsilon^{2\mathfrak{d}}\right\}\right) + 2\exp\!\left(\mathfrak{d}\ln\!\left((c\tau)^\tau\varepsilon^{-2}\right) - \tfrac{\varepsilon^4 M}{c}\right).\end{aligned} \tag{205}$$

Next let

$$\gamma = \max\{1, c\max\{1, v-u\}, c(v-u)^{2d}\}. \tag{206}$$

Hölder's inequality and (205) ensure that for all $\mathfrak{d}, K, M, \tau \in \mathbb{N}$, $\varepsilon \in (0,\infty)$ with $[2d(2dL(b-a)\varepsilon^{-1}+2)^d + 2]^3 \le \tau^3 \le \mathfrak{d}$ it holds that

$$\begin{aligned}&\mathbb{P}\left(\int_{[a,b]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau},\mathfrak{l}_\tau}(x) - \varphi(x)|\,\mathbb{P}_{X_1}(dx) > \varepsilon\right)\\ &\le \mathbb{P}\left(\left[\int_{[a,b]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau},\mathfrak{l}_\tau}(x) - \varphi(x)|^2\,\mathbb{P}_{X_1}(dx)\right]^{1/2} > \frac{\varepsilon}{(b-a)^{d/2}}\right)\\ &\le \exp\!\left(-K\min\left\{1,(c\tau)^{-\tau\mathfrak{d}}\varepsilon^{2\mathfrak{d}}(b-a)^{-d\mathfrak{d}}\right\}\right) + 2\exp\!\left(\mathfrak{d}\ln\!\left((c\tau)^\tau(b-a)^d\varepsilon^{-2}\right) - \tfrac{\varepsilon^4 M}{c(b-a)^{2d}}\right).\end{aligned} \tag{207}$$

This and (206) imply that for all $\mathfrak{d}, K, M, \tau \in \mathbb{N}$, $\varepsilon \in (0,1]$ with $[2d(2dL(b-a)\varepsilon^{-1}+2)^d+2]^3 \le \tau^3 \le \mathfrak{d}$ it holds that

$$\begin{aligned}&\mathbb{P}\left(\int_{[a,b]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau},\mathfrak{l}_\tau}(x) - \varphi(x)|\,\mathbb{P}_{X_1}(dx) > \varepsilon\right)\\ &\le \exp\!\left(-K\min\left\{1,(\gamma\tau)^{-\tau\mathfrak{d}}\varepsilon^{2\mathfrak{d}}\right\}\right) + \exp\!\left(\mathfrak{d}\ln\!\left((\gamma\tau)^\tau\varepsilon^{-2}\right) - \tfrac{\varepsilon^4 M}{\gamma}\right)\\ &\le \exp\!\left(-K(\gamma\tau)^{-\tau\mathfrak{d}}\varepsilon^{2\mathfrak{d}}\right) + 2\exp\!\left(\mathfrak{d}\ln\!\left((\gamma\tau)^\tau\varepsilon^{-2}\right) - \tfrac{\varepsilon^4 M}{\gamma}\right).\end{aligned} \tag{208}$$

This establishes (204). The proof of Corollary 4.8 is thus completed. □

4.3.2 Convergence rates for strong convergence

Lemma 4.9. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $c \in [0, \infty)$, and let $X : \Omega \to [-c, c]$ be a random variable. Then it holds for all $\varepsilon, p \in (0, \infty)$ that*

$$\mathbb{E}[|X|^p] \leq \varepsilon^p\, \mathbb{P}(|X| \leq \varepsilon) + c^p\, \mathbb{P}(|X| > \varepsilon) \leq \varepsilon^p + c^p\, \mathbb{P}(|X| > \varepsilon)\,. \tag{209}$$

Proof of Lemma 4.9. First, observe that for all $\varepsilon, p \in (0, \infty)$ it holds that

$$\mathbb{E}[|X|^p] = \mathbb{E}\big[|X|^p \mathbb{1}_{\{|X| \leq \varepsilon\}}\big] + \mathbb{E}\big[|X|^p \mathbb{1}_{\{|X| > \varepsilon\}}\big] \leq \varepsilon^p\, \mathbb{P}(|X| \leq \varepsilon) + c^p\, \mathbb{P}(|X| > \varepsilon)\,. \tag{210}$$

This establishes the first inequality in (209). The second inequality in (209) is immediate. The proof of Lemma 4.9 is thus completed. □

Corollary 4.10. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $d \in \mathbb{N}$, $L, a, u \in \mathbb{R}$, $b \in (a, \infty)$, $v \in (u, \infty)$, $R \in [\max\{1, L, |a|, |b|, 2|u|, 2|v|\}, \infty)$, let $\|\cdot\| : \mathbb{R}^d \to [0, \infty)$ be the standard norm on $\mathbb{R}^d$, let $(X_m) = (X_{m,1}, \ldots, X_{m,d}) : \Omega \to [a, b]^d$, $m \in \mathbb{N}$, be i.i.d. random variables, let $\varphi : [a, b]^d \to [u, v]$ satisfy for all $x, y \in [a, b]^d$ that $|\varphi(x) - \varphi(y)| \leq L\|x - y\|$, let $\mathfrak{l}_\tau \in \mathbb{N}^\tau$, $\tau \in \{3, 4, \ldots\}$, and $\mathfrak{E}_{\mathfrak{d},M,\tau} : [-R, R]^{\mathfrak{d}} \times \Omega \to [0, \infty)$, $M, \tau \in \mathbb{N}$, satisfy for all $\mathfrak{d}, M \in \mathbb{N}$, $\tau \in \{3, 4, \ldots\}$, $\theta \in [-R, R]^{\mathfrak{d}}$, $\omega \in \Omega$ that $\mathfrak{l}_\tau = (d, \tau, \tau, \ldots, \tau, 1)$ and*

$$\mathfrak{E}_{\mathfrak{d},M,\tau}(\theta, \omega) = \frac{1}{M}\left[\sum_{m=1}^{M} |\mathscr{N}_{u,v}^{\theta,\mathfrak{l}_\tau}(X_m(\omega)) - \varphi(X_m(\omega))|^2\right], \tag{211}$$

for every $\mathfrak{d}$ let $\Theta_{\mathfrak{d},k} : \Omega \to [-R, R]^{\mathfrak{d}}$, $k \in \mathbb{N}$, be i.i.d. random variables, assume for every $\mathfrak{d} \in \mathbb{N}$ that $\Theta_{\mathfrak{d},1}$ is continuous uniformly distributed on $[-R, R]^{\mathfrak{d}}$, assume for every $\mathfrak{d} \in \mathbb{N}$ that $(X_m)_{m\in\mathbb{N}}$ and $(\Theta_{\mathfrak{d},k})_{k\in\mathbb{N}}$ are independent, and let $\Xi_{\mathfrak{d},K,M,\tau} : \Omega \to [-R, R]^{\mathfrak{d}}$, $\mathfrak{d}, K, M \in \mathbb{N}$, $\tau \in \mathbb{N} \cap [3, \infty)$, satisfy for all $\mathfrak{d}, K, M, \tau \in \mathbb{N}$ that $\Xi_{\mathfrak{d},K,M,\tau} = \Theta_{\mathfrak{d},\min\{k\in\{1,2,\ldots,K\} : \mathfrak{E}_{\mathfrak{d},M,\tau}(\Theta_{\mathfrak{d},k}) = \min_{s\in\{1,\ldots,K\}} \mathfrak{E}_{\mathfrak{d},M,\tau}(\Theta_{\mathfrak{d},s})\}}$ (cf. Definition 2.8). Then there exists $c \in (0, \infty)$ such that for all $\mathfrak{d}, K, M, \tau \in \mathbb{N}$, $p \in [1, \infty)$, $\varepsilon \in (0, \infty)$ with $[2d(2dL(b - a)\varepsilon^{-1} + 2)^d + 2]^3 \leq \tau^3 \leq \mathfrak{d}$ it holds that

$$\begin{aligned}
&\left(\mathbb{E}\left[\left(\int_{[a,b]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau},\mathfrak{l}_\tau}(x) - \varphi(x)|^2\, \mathbb{P}_{X_1}(dx)\right)^{p/2}\right]\right)^{1/p} \\
&\leq (b - a)\left[\exp\!\big(-K \min\big\{1, (c\tau)^{-\tau\mathfrak{d}} \varepsilon^{2\mathfrak{d}}\big\}\big) + \exp\!\Big(\mathfrak{d} \ln\!\big((c\tau)^\tau \varepsilon^{-2}\big) - \tfrac{\varepsilon^4 M}{c}\Big)\right]^{1/p} + \varepsilon.
\end{aligned} \tag{212}$$

Proof of Corollary 4.10. First, observe that Corollary 4.7 ensures that there exists $c \in (0, \infty)$ which satisfies that for all $\mathfrak{d}, K, M, \tau \in \mathbb{N}$, $\varepsilon \in (0, \infty)$ with $[2d(2dL(b - a)\varepsilon^{-1} + 2)^d + 2]^3 \leq \tau^3 \leq \mathfrak{d}$ it holds that

$$\begin{aligned}
&\mathbb{P}\left(\left[\int_{[a,b]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau},\mathfrak{l}_\tau}(x) - \varphi(x)|^2\, \mathbb{P}_{X_1}(dx)\right]^{1/2} > \varepsilon\right) \\
&\leq \exp\!\big(-K \min\big\{1, (c\tau)^{-\tau\mathfrak{d}} \varepsilon^{2\mathfrak{d}}\big\}\big) + 2\exp\!\Big(\mathfrak{d} \ln\!\big((c\tau)^\tau \varepsilon^{-2}\big) - \tfrac{\varepsilon^4 M}{c}\Big).
\end{aligned} \tag{213}$$

Lemma 4.9 (with $(\Omega, \mathcal{A}, \mathbb{P}) \leftarrow (\Omega, \mathcal{A}, \mathbb{P})$, $c \leftarrow v - u$, $X \leftarrow (\Omega \ni \omega \mapsto [\int_{[a,b]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau}(\omega),\mathfrak{l}_\tau}(x) - \varphi(x)|^2\, \mathbb{P}_{X_1}(dx)]^{1/2} \in [u - v, v - u])$ in the notation of Lemma 4.9) hence ensures that for all $\mathfrak{d}, K, M, \tau \in \mathbb{N}$, $\varepsilon, p \in (0, \infty)$ with $[2d(2dL(b - a)\varepsilon^{-1} + 2)^d + 2]^3 \leq \tau^3 \leq \mathfrak{d}$ it holds that

$$\begin{aligned}
&\mathbb{E}\left[\left(\int_{[a,b]^d} |\mathscr{N}_{u,v}^{\Xi_{\mathfrak{d},K,M,\tau},\mathfrak{l}_\tau}(x) - \varphi(x)|^2\, \mathbb{P}_{X_1}(dx)\right)^{p/2}\right] \\
&\leq \varepsilon^p + (v - u)^p\left[\exp\!\big(-K \min\big\{1, (c\tau)^{-\tau\mathfrak{d}} \varepsilon^{2\mathfrak{d}}\big\}\big) + 2\exp\!\Big(\mathfrak{d} \ln\!\big((c\tau)^\tau \varepsilon^{-2}\big) - \tfrac{\varepsilon^4 M}{c}\Big)\right].
\end{aligned} \tag{214}$$

The fact that for all $p \in [1, \infty)$, $x, y \in [0, \infty)$ it holds that $(x + y)^{1/p} \leq x^{1/p} + y^{1/p}$ therefore establishes (212). The proof of Corollary 4.10 is thus completed. □

Acknowledgements

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2044 – 390685587, Mathematics Münster: Dynamics – Geometry – Structure.

References

[1] Bach, F. Breaking the curse of dimensionality with convex neutral networks. *J. Mach. Learn. Res. 18* (2017), 53 pages.

[2] Bach, F., and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in neural information processing systems* (2013), pp. 773–781.

[3] Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory 39*, 3 (1993), 930–945.

[4] Barron, A. R. Approximation and estimation bounds for artificial neural networks. *Machine learning 14*, 1 (1994), 115–133.

[5] Bartlett, P. L., Bousquet, O., and Mendelson, S. Local Rademacher complexities. *Ann. Statist. 33*, 4 (2005), 1497–1537.

[6] Beck, C., Becker, S., Grohs, P., Jaafari, N., and Jentzen, A. Solving stochastic differential equations and Kolmogorov equations by means of deep learning. *arXiv:1806.00421* (2018), 56 pages.

[7] Beck, C., E, W., and Jentzen, A. Machine Learning Approximation Algorithms for High-Dimensional Fully Nonlinear Partial Differential Equations and Second-order Backward Stochastic Differential Equations. *J. Nonlinear Sci. 29*, 4 (2019), 1563–1619.

[8] Bellman, R. *Dynamic programming*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 2010. Reprint of the 1957 edition, With a new introduction by Stuart Dreyfus.

[9] Bercu, B., and Fort, J.-C. Generic stochastic gradient methods. *Wiley Encyclopedia of Operations Research and Management Science* (2011), 1–8.

[10] Berner, J., Grohs, P., and Jentzen, A. Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. *arXiv:1809.03062* (2018), 35 pages.

[11] Blum, E. K., and Li, L. K. Approximation theory and feedforward networks. *Neural networks 4*, 4 (1991), 511–515.

[12] Bölcskei, H., Grohs, P., Kutyniok, G., and Petersen, P. Optimal approximation with sparsely connected deep neural networks. *SIAM J. Math. Data Sci. 1*, 1 (2019), 8–45.

[13] Burger, M., and Neubauer, A. Error bounds for approximation with neural networks. *J. Approx. Theory 112*, 2 (2001), 235–250.

[14] Candes, E. J. *Ridgelets: theory and applications*. PhD thesis, Stanford University Stanford, 1998.

[15] Chau, N. H., Moulines, É., Rásonyi, M., Sabanis, S., and Zhang, Y. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *arXiv:1905.13142* (2019), 27 pages.

[16] Chen, T., and Chen, H. Approximation capability to functions of several variables, nonlinear functionals, and operators by radial basis function neural networks. *IEEE Transactions on Neural Networks 6*, 4 (1995), 904–910.

[17] Chui, C. K., Li, X., and Mhaskar, H. N. Neural networks for localized approximation. *Math. Comp. 63*, 208 (1994), 607–623.

[18] Cucker, F., and Smale, S. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.) 39*, 1 (2002), 1–49.

[19] Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems 2*, 4 (1989), 303–314.

[20] Dereich, S., and Müller-Gronbach, T. General multilevel adaptations for stochastic approximation algorithms of Robbins-Monro and Polyak-Ruppert type. *Numer. Math. 142*, 2 (2019), 279–328.

[21] DeVore, R. A., Oskolkov, K. I., and Petrushev, P. P. Approximation by feed-forward neural networks. In *The heritage of P. L. Chebyshev: a Festschrift in honor of the 70th birthday of T. J. Rivlin*, vol. 4. Baltzer Science Publishers BV, Amsterdam, 1997, pp. 261–287.

[22] E, W., and Wang, Q. Exponential convergence of the deep neural network approximation for analytic functions. *arXiv:1807.00297* (2018), 7 pages.

[23] Elbrächter, D., Grohs, P., Jentzen, A., and Schwab, C. DNN expression rate analysis of high-dimensional PDEs: Application to option pricing. *arXiv:1809.07669* (2018), 50 pages.

[24] Eldan, R., and Shamir, O. The power of depth for feedforward neural networks. In *Conference on learning theory* (2016), pp. 907–940.

[25] Ellacott, S. W. Aspects of the numerical analysis of neural networks. In *Acta numerica, 1994*, Acta Numer. Cambridge Univ. Press, Cambridge, 1994, pp. 145–202.

[26] Fehrman, B., Gess, B., and Jentzen, A. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *arXiv:1904.01517* (2019), 59 pages.

[27] Funahashi, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural networks 2*, 3 (1989), 183–192.

[28] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2016.

[29] Gribonval, R., Kutyniok, G., Nielsen, M., and Voigtlaender, F. Approximation spaces of deep neural networks. *arXiv:1905.01208* (2019), 63 pages.

[30] Grohs, P., Hornung, F., Jentzen, A., and von Wurstemberger, P. A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. *arXiv:1809.02362* (2018), 124 pages. Revision requested from Memoirs of the AMS.

[31] Grohs, P., Hornung, F., Jentzen, A., and Zimmermann, P. Space-time error estimates for deep neural network approximations for differential equations. *arXiv:1908.03833* (2019), 86 pages.

[32] Grohs, P., Jentzen, A., and Salimova, D. Deep neural network approximations for Monte Carlo algorithms. *arXiv:1908.10828* (2019), 45 pages.

[33] Grohs, P., Perekrestenko, D., Elbrächter, D., and Bölcskei, H. Deep neural network approximation theory. *arXiv:1901.02220* (2019), 60 pages.

[34] Gühring, I., Kutyniok, G., and Petersen, P. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *arXiv:1902.07896* (2019), 42 pages.

[35] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. *A distribution-free theory of nonparametric regression.* Springer Series in Statistics. Springer-Verlag, New York, 2002.

[36] Hartman, E. J., Keeler, J. D., and Kowalski, J. M. Layered neural networks with Gaussian hidden units as universal approximations. *Neural computation 2*, 2 (1990), 210–215.

[37] Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 301 (1963), 13–30.

[38] Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural networks 4*, 2 (1991), 251–257.

[39] Hornik, K. Some new results on neural network approximation. *Neural networks 6*, 8 (1993), 1069–1072.

[40] Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks 2*, 5 (1989), 359–366.

[41] Hornik, K., Stinchcombe, M., and White, H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural networks 3*, 5 (1990), 551–560.

[42] Hutzenthaler, M., Jentzen, A., Kruse, T., and Nguyen, T. A. A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. *arXiv:1901.10854* (2019), 29 pages.

[43] Jentzen, A., Kuckuck, B., Neufeld, A., and von Wurstemberger, P. Strong error analysis for stochastic gradient descent optimization algorithms. *arXiv:1801.09324* (2018), 75 pages. Revision requested from IMA Journal of Numerical Analysis.

[44] Jentzen, A., Salimova, D., and Welti, T. A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of Kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients. *arXiv:1809.07321* (2018), 48 pages.

[45] Jentzen, A., and von Wurstemberger, P. Lower error bounds for the stochastic gradient descent optimization algorithm: Sharp convergence rates for slowly and fast decaying learning rates. *arXiv:1803.08600* (2018), 42 pages. To appear in Journal of Complexity.

[46] Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. Non-asymptotic Analysis of Biased Stochastic Approximation Scheme. *arXiv:1902.00629* (2019), 32 pages.

[47] Kutyniok, G., Petersen, P., Raslan, M., and Schneider, R. A theoretical analysis of deep neural networks and parametric PDEs. *arXiv:1904.00377* (2019), 43 pages.

[48] Lei, Y., Hu, T., Li, G., and Tang, K. Stochastic Gradient Descent for Nonconvex Learning without Bounded Gradient Assumptions. *arXiv:1902.00908* (2019), 6 pages.

[49] Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks 6*, 6 (1993), 861–867.

[50] Maggi, F. *Sets of finite perimeter and geometric variational problems*, vol. 135 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2012. An introduction to geometric measure theory.

[51] Massart, P. *Concentration inequalities and model selection*, vol. 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[52] Mhaskar, H. N. Neural networks for optimal approximation of smooth and analytic functions. *Neural computation 8*, 1 (1996), 164–177.

[53] Mhaskar, H. N., and Micchelli, C. A. Degree of approximation by neural and translation networks with a single hidden layer. *Adv. in Appl. Math. 16*, 2 (1995), 151–183.

[54] Mhaskar, H. N., and Poggio, T. Deep vs. shallow networks: an approximation theory perspective. *Anal. Appl. (Singap.) 14*, 6 (2016), 829–848.

[55] Nguyen-Thien, T., and Tran-Cong, T. Approximation of functions and their derivatives: A neural network implementation with applications. *Applied Mathematical Modelling 23*, 9 (1999), 687–704.

[56] Novak, E., and Woźniakowski, H. *Tractability of multivariate problems. Vol. 1: Linear information*, vol. 6 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2008.

[57] Novak, E., and Woźniakowski, H. *Tractability of multivariate problems. Volume II: Standard information for functionals*, vol. 12 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2010.

[58] Park, J., and Sandberg, I. W. Universal approximation using radial-basis-function networks. *Neural computation 3*, 2 (1991), 246–257.

[59] Perekrestenko, D., Grohs, P., Elbrächter, D., and Bölcskei, H. The universal approximation power of finite-width deep ReLU networks. *arXiv:1806.01528* (2018), 16 pages.

[60] Petersen, P., Raslan, M., and Voigtlaender, F. Topological properties of the set of functions generated by neural networks of fixed size. *arXiv:1806.08459* (2018), 56 pages.

[61] Petersen, P., and Voigtlaender, F. Equivalence of approximation by convolutional neural networks and fully-connected networks. *arXiv:1809.00973* (2018), 10 pages.

[62] Petersen, P., and Voigtlaender, F. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks 108* (2018), 296–330.

[63] Pinkus, A. Approximation theory of the MLP model in neural networks. In *Acta numerica, 1999*, vol. 8 of *Acta Numer.* Cambridge Univ. Press, Cambridge, 1999, pp. 143–195.

[64] Reisinger, C., and Zhang, Y. Rectified deep neural networks overcome the curse of dimensionality for nonsmooth value functions in zero-sum games of nonlinear stiff systems. *arXiv:1903.06652* (2019), 34 pages.

[65] Schmitt, M. Lower bounds on the complexity of approximating continuous functions by sigmoidal neural networks. In *Advances in neural information processing systems* (2000), pp. 328–334.

[66] Schwab, C., and Zech, J. Deep learning in high dimension: neural network expression rates for generalized polynomial chaos expansions in UQ. *Anal. Appl. (Singap.) 17*, 1 (2019), 19–55.

[67] Shaham, U., Cloninger, A., and Coifman, R. R. Provable approximation properties for deep neural networks. *Appl. Comput. Harmon. Anal. 44*, 3 (2018), 537–557.

[68] Shalev-Shwartz, S., and Ben-David, S. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[69] Shen, Z., Yang, H., and Zhang, S. Deep network approximation characterized by number of neurons. *arXiv:1906.05497* (2019), 36 pages.

[70] Shen, Z., Yang, H., and Zhang, S. Nonlinear approximation via compositions. *arXiv:1902.10170* (2019), 19 pages.

[71] van de Geer, S. A. *Applications of empirical process theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 2000.

[72] Voigtlaender, F., and Petersen, P. Approximation in $L^p(\mu)$ with deep ReLU neural networks. *arXiv:1904.04789* (2019), 4 pages.

[73] Yarotsky, D. Error bounds for approximations with deep relu networks. *Neural Networks 94* (2017), 103–114.

[74] Yarotsky, D. Universal approximations of invariant maps by neural networks. *arXiv:1804.10306* (2018), 64 pages.