ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Deep ReLU Networks and High-Order Finite Element Methods

J. A. A. Opschoor and P. C. Petersen and Ch. Schwab

Research Report No. 2019-07
January 2019

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

Funding: P.P is supported by a DFG Research Fellowship ”Shearlet-based energy functionals for anisotropic phase-field methods”.

Deep ReLU Networks and High-Order Finite Element Methods

Joost A. A. Opschoor* Philipp C. Petersen† Christoph Schwab*

January 17, 2019

Abstract

Approximation rate bounds for expressions of real-valued functions on intervals by deep neural networks (DNNs for short) are established. The approximation results are given for DNNs based on ReLU activation functions, and the approximation error is measured with respect to Sobolev norms. It is shown that ReLU DNNs allow for essentially the same approximation rates as nonlinear, variable-order, free-knot (or so-called "*hp*-adaptive") spline approximations and spectral approximations, for a wide range of Sobolev and Besov spaces. In particular, exponential convergence rates in terms of the DNN size for piecewise Gevrey functions with point singularities are established. Combined with recent results on ReLU DNN expression of rational, oscillatory, and high-dimensional functions, this corroborates that ReLU DNNs match the approximation power of "best in class" schemes for a wide range of approximations.

Keywords: Deep neural networks, finite element methods, function approximation, adaptivity

Subject Classification: 41A25, 41A46, 65N30

1 Introduction

Recent years have seen a dramatic increase in the application of deep neural networks (DNNs for short) in a wide range of problems. We mention only machine learning, including applications from speech recognition to image classification [19]. In scientific computing, computational experiments with DNNs for the numerical solution of partial differential equations (PDEs for short) have been reported to be strikingly successful, in a wide range of applications (e.g. [3, 4, 11, 12, 16, 25, 36]). The present paper aims at contributing to a mathematical understanding of these observations. Specifically, we investigate DNN expression rates of concrete architectures of DNNs for a number of widely used approximation spaces in numerical analysis. We present DNN architectures emulating fixed- and free-knot spline approximations, spectral- and *hp*-approximations. Moreover, we will observe that the so-constructed DNNs yield the same approximation properties as the best available approximations with comparable numbers of degrees of freedom.

Early mathematical work on approximation by neural networks (NNs for short) focused on *universality results* (e.g. [1, 2, 18, 28] and the references there). In these references, universality was established already for so-called shallow NNs, thereby implying universality also for DNNs, for many activation functions. These early universality results parallel, in a sense, density results for polynomial approximations such as the Stone-Weierstrass theorem. Moreover, this universality of shallow NNs paradoxically led to the belief that depth in NN architectures would, in practice, be of little benefit. In recent years, dramatic empirical evidence fuelled by the ubiquitous availability of massive computing power and training data shattered this folklore [19]. At the same time, and in response, mathematical analysis started to address the interplay of depth and architecture of DNNs with specific function classes and it was shown that DNNs afford significant quantitative advantages over their shallow counterparts in terms of expression rates for a wide range of function spaces.

[1]SAM, ETH Zürich, ETH Zentrum, HG G57.1, CH8092 Zürich, Switzerland, email: {joost.opschoor, christoph.schwab}@sam.math.ethz.ch

[2]Mathematical Institute, University of Oxford, OX2 6GG, Oxford, UK, email: Philipp.Petersen@maths.ox.ac.uk

Among these are expression rate bounds for analytic functions (e.g. [10, 22]), (piecewise) differentiable functions (e.g.[26, 41]), high-dimensional approximation (e.g. [12, 34]), oscillatory functions (e.g. [13]), cartoon functions in image segmentation (e.g. [14, 13]), rational function approximations (e.g. [38]), high degree polynomials (e.g. [20, 30]), and linear finite elements [17]. The standard approach employed in all results above is to first demonstrate that DNNs are capable of efficiently emulating other (linear or nonlinear) approximation architectures such as B-splines (e.g. [21]), wavelets (e.g. [5, 35]), and high degree polynomials (e.g. [20, 41]). Then, the approximation results of these classical architectures are transfered to DNN approximation.

Most approximation theoretical results on DNNs assess approximation fidelity with respect to L^p norms, $p \in [1, \infty]$. However, in view of applications in numerical PDE approximation, it appears to be more natural to measure the accuracy of approximation with respect to Sobolev norms.

In this work, we study DNN approximations of functions $f \in W^{s,p}([-1,1])$, $p \in [1,\infty]$, $s \geq 1$ with respect to weaker Sobolev norms. Concretely, in one space dimension, we will present a range of results for piecewise polynomial functions f. Based on these results, we then show that DNNs can emulate high-order h–FEM on general partitions of a bounded interval, as well as high-order, spectral and so-called p– and hp–FEM. Specifically, we conclude that, in terms of the number of degrees of freedom and from an approximation theoretical point of view, DNNs perform as well as the best available finite element method. This observation explains, to some extent, the at times dramatic success that deep learning methodologies display in computational mathematics such as the aforementioned numerical approximations of PDEs.

The outline of this article is as follows: In Section 2, we start this exposition by presenting a formal definition of a neural network as well as a formal description of some basic operations on neural networks. In Section 3, we present—as a motivation—a simple connection between ReLU approximations and continuous, piecewise affine (free-knot) spline approximation. Section 4 provides the emulation of polynomials by ReLU networks as well as associated error estimates with respect to Sobolev norms. This construction is then the basis for the emulation of a range of FE spaces in Section 5. We conclude this article in Section 6.

Notation

Throughout this paper, C denotes a generic constant which may be different at each appearance, even within an equation. Dependence of C on parameters is indicated explicitly by $C(\cdot)$, e.g. $C(\eta, \theta)$.

When denoting the norm of a function, we will sometimes write the argument of the function explicitly. For example, we will write $\left\| mx^{m-1} - f(x) \right\|^2_{L^2(I)}$ for $m \in \mathbb{N}$, some bounded domain I and $f \in L^2(I)$. Here, $x \in I$ is the variable of integration.

For continuous, piecewise polynomial functions, we will use the following notation: Let $\mathcal{T}$ be a partition of the interval $I \coloneqq (0,1)$ with nodes $0 = x_0 < x_1 < \ldots < x_{N-1} < x_N = 1$, elements $I_i \coloneqq (x_{i-1}, x_i)$ and element sizes $h_i \coloneqq x_i - x_{i-1}$ for $i \in \{1, \ldots, N\}$. Let $h \coloneqq \max_{i \in \{1,\ldots,N\}} h_i$. For a polynomial degree distribution $\boldsymbol{p} = (p_i)_{i \in \{1,\ldots,N\}} \subset \mathbb{N}$ on $\mathcal{T}$, we define the maximal degree $p_{\max} \coloneqq \max_{i=1}^N p_i$ and the corresponding approximation space

$$S_{\boldsymbol{p}}(I, \mathcal{T}) \coloneqq \{ v \in H^1(I) : v|_{I_i} \in \mathbb{P}_{p_i}(I_i) \text{ for all } i \in \{1, \ldots, N\} \}.$$

For $N, p \in \mathbb{N}$, we define the space of free-knot splines with less than N interior knots on $I \coloneqq (0,1)$ which are continuous, piecewise polynomial functions of degree p by

$$S_p^N(I) \coloneqq \bigcup \{ S_{\boldsymbol{p}}(I, \mathcal{T}) \colon \mathcal{T} \text{ partition of } I \text{ with } N \text{ elements} \},$$

where $\boldsymbol{p} = (p, \ldots, p)$. These are often referred to as *free-knot splines of degree* $p+1$.

2 Neural Networks and ReLU Calculus

Following standard practice, we differentiate between a NN as a set of parameters and the so-called *realization* of the network. The realization is an associated function resulting from repeatedly applying affine linear transformations—defined through the parameters—and a so-called *activation function*, denoted generically by ϱ.

Definition 2.1. *Let $d, L \in \mathbb{N}$. A* neural network Φ with input dimension d and L layers *is a sequence of matrix-vector tuples*

$$\Phi = \big((A_1, b_1), (A_2, b_2), \dots, (A_L, b_L)\big),$$

where $N_0 \coloneqq d$ and $N_1, \dots, N_L \in \mathbb{N}$, and where $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{N_\ell}$ for $\ell = 1, ..., L$.

For a NN Φ and an activation function $\varrho : \mathbb{R} \to \mathbb{R}$, we define the associated realization of the NN Φ *as*

$$\mathrm{R}(\Phi) : \mathbb{R}^d \to \mathbb{R}^{N_L} : x \mapsto x_L \coloneqq \mathrm{R}(\Phi)(x),$$

where the output $x_L \in \mathbb{R}^{N_L}$ results from

$$\begin{aligned} &x_0 \coloneqq x, \\ &x_\ell \coloneqq \varrho(A_\ell\, x_{\ell-1} + b_\ell) \quad \text{for } \ell = 1, \dots, L-1, \\ &x_L \coloneqq A_L\, x_{L-1} + b_L. \end{aligned} \tag{2.1}$$

Here ϱ is understood to act component-wise on vector-valued inputs, i.e., for $y = (y^1, \dots, y^m) \in \mathbb{R}^m$, $\varrho(y) \coloneqq (\varrho(y^1), \dots, \varrho(y^m))$. We call $N(\Phi) \coloneqq d + \sum_{j=1}^{L} N_j$ the number of neurons of the NN Φ*, $L(\Phi) \coloneqq L$ the* number of layers *or* depth*, $M_j(\Phi) \coloneqq \|A_j\|_{\ell^0} + \|b_j\|_{\ell^0}$ the* number of weights in the j-th layer*, and $M(\Phi) \coloneqq \sum_{j=1}^{L} M_j(\Phi)$ the* number of weights of Φ*, also referred to as its* size*. The number of weights in the first layer is also denoted by $M_{\mathrm{fi}}(\Phi)$, the number of weights in the last layer by $M_{\mathrm{la}}(\Phi)$. We refer to N_L as the* dimension of the output layer of Φ*.*

In this work, the only activation function that we will consider is the so-called *rectified linear unit* (*ReLU* for short) defined by

$$\varrho : \mathbb{R} \to \mathbb{R} : x \mapsto \max\{0, x\}. \tag{2.2}$$

One fundamental ingredient of this work is to establish the approximation of piecewise polynomials by deep ReLU neural networks. Our results will imply, in view of classical results on approximation by continuous, piecewise polynomial functions, DNN expression rate bounds for functions in a collection of classical function spaces, in particular of Sobolev, Besov, and Hölder type. We will accomplish this construction of approximate piecewise polynomials by first demonstrating how to approximate certain universal building blocks by realizations of DNNs. Then, we invoke a so-called *calculus of ReLU NNs*, as introduced in [26]. This is a formal framework describing how to concatenate, parallelize, or extend DNNs. Using this framework, we can assemble complex functions from the fundamental building blocks.

Below, we recall three results of [26] which also serve as definitions of the associated procedures. We start with *concatenation of NNs*.

Proposition 2.2 (NN concatenation [26])**.** *Let $L_1, L_2 \in \mathbb{N}$, and let Φ^1, Φ^2 be two NNs of respective depths L_1 and L_2 such that $N_0^1 = N_{L_2}^2 =: d$, i.e., the input layer of Φ^1 has the same dimension as the output layer of Φ^2.*

Then, there exists a NN $\Phi^1 \odot \Phi^2$, called the sparse concatenation of Φ^1 and Φ^2*, such that $\Phi^1 \odot \Phi^2$ has $L_1 + L_2$ layers, $\mathrm{R}(\Phi^1 \odot \Phi^2) = \mathrm{R}(\Phi^1) \circ \mathrm{R}(\Phi^2)$,*

$$M_{\mathrm{fi}}(\Phi^1 \odot \Phi^2) \le \begin{cases} 2M_{\mathrm{fi}}(\Phi^2) & \text{if } L_2 = 1, \\ M_{\mathrm{fi}}(\Phi^2) & \text{else,} \end{cases} \qquad M_{\mathrm{la}}(\Phi^1 \odot \Phi^2) \le \begin{cases} 2M_{\mathrm{la}}(\Phi^1) & \text{if } L_1 = 1, \\ M_{\mathrm{la}}(\Phi^1) & \text{else,} \end{cases}$$

and

$$M\left(\Phi^1 \odot \Phi^2\right) \le M\left(\Phi^1\right) + M_{\mathrm{fi}}\left(\Phi^1\right) + M_{\mathrm{la}}\left(\Phi^2\right) + M\left(\Phi^2\right) \le 2M\left(\Phi^1\right) + 2M\left(\Phi^2\right).$$

The second fundamental operation on NNs is parallelization. This can be achieved with the following construction.

Proposition 2.3 (NN parallelization [26])**.** *Let $L, d \in \mathbb{N}$ and let Φ^1, Φ^2 be two NNs with L layers and with d-dimensional input each. Then there exists a network $\mathrm{P}(\Phi^1, \Phi^2)$ with d-dimensional input and L layers, which we call the* parallelization of Φ^1 and Φ^2*, such that*

$$\mathrm{R}\left(\mathrm{P}\left(\Phi^1, \Phi^2\right)\right)(x) = \left(\mathrm{R}\left(\Phi^1\right)(x), \mathrm{R}\left(\Phi^2\right)(x)\right), \text{ for all } x \in \mathbb{R}^d, \tag{2.3}$$

$M(\mathrm{P}(\Phi^1, \Phi^2)) = M(\Phi^1) + M(\Phi^2)$, $M_{\mathrm{fi}}(\mathrm{P}(\Phi^1, \Phi^2)) = M_{\mathrm{fi}}(\Phi^1) + M_{\mathrm{fi}}(\Phi^2)$ and $M_{\mathrm{la}}(\mathrm{P}(\Phi^1, \Phi^2)) = M_{\mathrm{la}}(\Phi^1) + M_{\mathrm{la}}(\Phi^2)$.

Proposition 2.3 only enables us to parallelize NNs of equal depth. To make two NNs have the same depth one can extend the shorter of the two by concatenating with a network that implements the identity. One possible construction of such a NN is presented next.

Proposition 2.4 (DNN emulation of Id [26]). *For every $d, L \in \mathbb{N}$ there exists a NN $\Phi^{\mathrm{Id}}_{d,L}$ with $L(\Phi^{\mathrm{Id}}_{d,L}) = L$, $M(\Phi^{\mathrm{Id}}_{d,L}) \leq 2dL$, $M_{\mathrm{fi}}(\Phi^{\mathrm{Id}}_{d,L}) \leq 2$ and $M_{\mathrm{la}}(\Phi^{\mathrm{Id}}_{d,L}) \leq 2$ such that $\mathrm{R}(\Phi^{\mathrm{Id}}_{d,L}) = \mathrm{Id}_{\mathbb{R}^d}$.*

Finally, we sometimes require a parallelization of NNs that do not share inputs.

Proposition 2.5 (Full parallelization of NNs with distinct inputs [12]). *Let $L \in \mathbb{N}$ and let*

$$\Phi^1 = \left(\left(A^1_1, b^1_1\right), \ldots, \left(A^1_L, b^1_L\right)\right), \quad \Phi^2 = \left(\left(A^2_1, b^2_1\right), \ldots, \left(A^2_L, b^2_L\right)\right)$$

be two NNs with L layers each and with input dimensions $N^1_0 = d_1$ and $N^2_0 = d_2$, respectively.

Then there exists a NN, denoted by $\mathrm{FP}(\Phi^1, \Phi^2)$, with $d = d_1 + d_2$-dimensional input and L layers, which we call the full parallelization of Φ^1 and Φ^2, *such that for all $x = (x_1, x_2) \in \mathbb{R}^d$ with $x_i \in \mathbb{R}^{d_i}, i = 1, 2$*

$$\mathrm{R}\left(\mathrm{FP}\left(\Phi^1, \Phi^2\right)\right)(x_1, x_2) = \left(\mathrm{R}\left(\Phi^1\right)(x_1), \mathrm{R}\left(\Phi^2\right)(x_2)\right),$$

$M(\mathrm{FP}(\Phi^1, \Phi^2)) = M(\Phi^1) + M(\Phi^2)$, $M_{\mathrm{fi}}(\mathrm{FP}(\Phi^1, \Phi^2)) = M_{\mathrm{fi}}(\Phi^1) + M_{\mathrm{fi}}(\Phi^2)$ *and* $M_{\mathrm{la}}(\mathrm{FP}(\Phi^1, \Phi^2)) = M_{\mathrm{la}}(\Phi^1) + M_{\mathrm{la}}(\Phi^2)$.

Proof. Set $\mathrm{FP}\left(\Phi^1, \Phi^2\right) \coloneqq \left(\left(A^3_1, b^3_1\right), \ldots, \left(A^3_L, b^3_L\right)\right)$ where, for $j = 1, \ldots, L$, we define

$$A^3_j \coloneqq \begin{pmatrix} A^1_j & 0 \\ 0 & A^2_j \end{pmatrix} \text{ and } b^3_j \coloneqq \begin{pmatrix} b^1_j \\ b^2_j \end{pmatrix}.$$

□

The four operations: concatenation, extension, parallelization with and without shared inputs; will be used to assemble more complex networks out of fundamental building blocks.

3 ReLU Network Approximation and Linear Splines

In this section, we analyze the connection between shallow ReLU networks and linear splines. The goal of this simple analysis is to identify the functional roles of the hidden parameters of a network. Concretely, we will see that approximation by shallow ReLU networks, where one is only varying the parameters in the output layer, corresponds to linear spline approximation with fixed nodes. On the other hand, an adaptive choice of the internal parameters of a network corresponds to free-knot linear spline approximation. This motivation highlights a first functional role of the hidden parameters. In Section 4 and after, we also identify further, more high-level roles of hidden parameters for deeper networks such as controlling the degree of the emulated polynomial approximation.

We begin by describing a network with exact emulation of continuous, piecewise affine-linear functions on arbitrary partitions of I.

Lemma 3.1. *For every partition $\mathcal{T}$ of $I = (0, 1)$ with N elements and every $v \in S_1(I, \mathcal{T})$ there exists a NN Φ^v such that*

$$\mathrm{R}\left(\Phi^v\right) = v, \quad L\left(\Phi^v\right) = 2, \quad M\left(\Phi^v\right) \leq 3N + 1, \quad M_{\mathrm{fi}}\left(\Phi^v\right) \leq 2N, \text{ and } M_{\mathrm{la}}\left(\Phi^v\right) \leq N + 1. \tag{3.1}$$

Proof. We set $\Phi^v \coloneqq ((A^v_1, b^v_1), (A^v_2, b^v_2))$ such that

$$A^v_1 \coloneqq [1, \ldots, 1]^T \in \mathbb{R}^{N \times 1}, \quad b^v_1 \coloneqq [-x_0, -x_1, \ldots, -x_{N-1}]^T \in \mathbb{R}^N, \quad b^v_2 \coloneqq v(x_0) \in \mathbb{R},$$

and, for $i \in \{1, \ldots, N\}$,

$$A^v_2 \in \mathbb{R}^{1 \times N}, \qquad (A^v_2)_{1,i} \coloneqq \begin{cases} \frac{v(x_i) - v(x_{i-1})}{x_i - x_{i-1}} - \frac{v(x_{i-1}) - v(x_{i-2})}{x_{i-1} - x_{i-2}} & \text{if } i > 1 \\ \frac{v(x_i) - v(x_{i-1})}{x_i - x_{i-1}} & \text{if } i = 1. \end{cases}$$

The claimed properties follow directly. □

We remark that the (simple) construction (3.1) contains both, *fixed-knot* spline approximations, as well as *free-knot* spline approximations. The former are obtained by constraining the NN parameters x_j in the hidden layer, the latter by allowing these hidden layer parameters to adapt during training of the NN. Then, the NN (3.1) is "h-adaptive", by design.

Lemma 3.1 can be combined with the following result on free-knot spline approximations.

Proposition 3.2 ([24, Theorem 3]). *Let $s < \max\{2, 1 + 1/q\}$, let $0 < q < q' \leq \infty$ and $0 < s' < \min\{1 + 1/q', s - 1/q + 1/q'\}$, and let $0 < t, t' \leq \infty$. Then for some $C \coloneqq C(q, q', t, t', s, s) > 0$, for every $N \in \mathbb{N}$ and every $f \in B^s_{q,t}(I)$ there exists $h^N \in S^N_1(I)$ such that*

$$\left\| f - h^N \right\|_{B^{s'}_{q',t'}(I)} \leq C N^{-(s-s')} \|f\|_{B^s_{q,t}(I)}.$$

For comparison, the approximation error for fixed-knot continuous, piecewise linear spline approximation on uniform partitions is of the order $\mathcal{O}(N^{-(s-s'-1/q+1/q')})$.

As a consequence of Proposition 3.2 and Lemma 3.1 we conclude the following corollary.

Corollary 3.3. *Let $s < \max\{2, 1 + 1/q\}$, let $0 < q < q' \leq \infty$ and $0 < s' < \min\{1 + 1/q', s - 1/q + 1/q'\}$, and let $0 < t, t' \leq \infty$. Then for some $C \coloneqq C(q, q', t, t', s, s) > 0$, for every $N \in \mathbb{N}$ and every $f \in B^s_{q,t}(I)$ there exists a NN Φ^N_f such that*

$$\left\| f - \mathrm{R}\left(\Phi^N_f\right) \right\|_{B^{s'}_{q',t'}(I)} \leq C \left(M\left(\Phi^N_f\right) \right)^{-(s-s')} \|f\|_{B^s_{q,t}(I)}.$$

Corollary 3.3 shows that ReLU NNs achieve the same convergence rate in terms of the network size as the convergence rate in terms of pieces N in Proposition 3.2.

The weights of the networks constructed in Lemma 3.1 have two types of freedom: First, the weights depend nonlinearly on the nodes $\{x_i\}_{i=0}^N$ of the partition $\mathcal{T}$. Second, the weights in the output layer depend linearly on the function values $\{v(x_i)\}_{i=0}^N$.

Fixing the weights in the first layer corresponds to fixing the partition, i.e. optimizing only the weights in the output layer corresponds to fixed-knot continuous, piecewise linear spline approximation. Exploiting the linearity of the output layer, the weights of the output layer can be determined by linear optimization.

4 Emulation of Polynomials by ReLU Networks

In this section, we present an emulation of polynomials of arbitrary degrees by ReLU NNs. Here, we analyze the approximation error with respect to Sobolev norms. In the sequel, it will prove to be important to have control of the emulated polynomials on the end points of the reference interval. Therefore, we present a construction of a polynomial emulation which is exact at the endpoints in Proposition 4.4.

The results below are based on a construction of DNNs emulating the multiplication function with two-dimensional input which has been derived in [41]. We recall here a version of this result and provide an estimate of the error with respect to the $W^{1,\infty}$ norm, from [34], as required in approximation rate bounds for PDEs.

Proposition 4.1 ([34, 41]). *There exist constants $C_L, C'_L, C_M, C'_M > 0$ such that, for every $\kappa > 0$ and $\delta \in (0, 1/2)$, there exists a NN $\widetilde{\times}_{\delta,\kappa}$ with two-dimensional input and such that*

$$\begin{aligned} &\sup_{|a|,|b| \leq \kappa} \left| ab - \mathrm{R}\left(\widetilde{\times}_{\delta,\kappa}\right)(a,b) \right| \leq \delta \text{ and} \\ &\operatorname*{esssup}_{|a|,|b| \leq \kappa} \max\left\{ \left| a - \frac{\mathrm{d}}{\mathrm{d}b} \mathrm{R}\left(\widetilde{\times}_{\delta,\kappa}\right)(a,b) \right|, \left| b - \frac{\mathrm{d}}{\mathrm{d}a} \mathrm{R}\left(\widetilde{\times}_{\delta,\kappa}\right)(a,b) \right| \right\} \leq \delta, \end{aligned} \tag{4.1}$$

where $\mathrm{d}/\mathrm{d}a$ and $\mathrm{d}/\mathrm{d}b$ denote weak derivatives. Furthermore, for every $\kappa > 0$ and for every $\delta \in (0, 1/2)$

$$M\left(\widetilde{\times}_{\delta,\kappa}\right) \leq C_M \left(\log_2\left(\frac{\max\{\kappa, 1\}}{\delta} \right) \right) + C'_M \text{ and } L\left(\widetilde{\times}_{\delta,\kappa}\right) \leq C_L \left(\log_2\left(\frac{\max\{\kappa, 1\}}{\delta} \right) \right) + C'_L.$$

Moreover, for all $a, b \in \mathbb{R}$,

$$\mathrm{R}\left(\widetilde{\times}_{\delta,\kappa}\right)(a, 0) = \mathrm{R}\left(\widetilde{\times}_{\delta,\kappa}\right)(0, b) = 0. \tag{4.2}$$

We now prove results on the approximation of polynomials on the reference interval $\hat{I} := (-1,1)$ by realizations of NNs, using the networks from Proposition 4.1.

Proposition 4.2. *For each $n \in \mathbb{N}_0$ and each polynomial $v \in \mathbb{P}_n([-1,1])$, such that $v(x) = \sum_{\ell=0}^n \tilde{v}_\ell x^\ell$, for all $x \in [-1,1]$ with $C_0 := \sum_{\ell=2}^n |\tilde{v}_\ell|$, there exist NNs $\{\Phi_\beta^v\}_{\beta\in(0,1)}$ with input dimension one and output dimension one which satisfy*

$$\begin{aligned}
\big\|v - \mathrm{R}(\Phi_\beta^v)\big\|_{W^{1,\infty}(\hat{I})} &\leq \beta,\\
\mathrm{R}(\Phi_\beta^v)(0) &= v(0),\\
L(\Phi_\beta^v) &\leq C_L(1+\log_2(n))\log_2(C_0/\beta) + \tfrac{1}{3}C_L(\log_2(n))^3 + C(1+\log_2(n))^2,\\
M(\Phi_\beta^v) &\leq 2C_M n\log_2(C_0/\beta) + 4C_M n\log_2(n) + 4C_L(1+\log_2(n))\log_2(C_0/\beta) + C(1+n),\\
M_{\mathrm{fi}}(\Phi_\beta^v) &\leq C_{\mathrm{fi}} + 4,\\
M_{\mathrm{la}}(\Phi_\beta^v) &\leq 2n+2
\end{aligned}$$

if $C_0 > \beta$. If $C_0 \leq \beta$ the same estimates hold, but with C_0 replaced by 2β.

To prove the proposition, we use the following technical lemma. For $k \in \mathbb{N}$ and a given polynomial v, this lemma produces a tree-structured network with $2^{k-1}+2$ outputs. The first $2^{k-1}+1$ of these correspond to high-order monomials of degree between 2^{k-1} and 2^k. The last output dimension contains an approximation to the partial sum of v of degree 2^{k-1}.

This network is constructed by repeatedly applying the product network introduced in Proposition 4.1.

Lemma 4.3. *Let $n \in \mathbb{N}$ and $v \in \mathbb{P}_n(\hat{I})$ such that $v(x) = \sum_{\ell=0}^n \tilde{v}_\ell x^\ell$, for all $x \in [-1,1]$. We define $\tilde{v}_\ell := 0$ for $\ell > n$. For every $k \in \mathbb{N}$ there exist NNs $\{\Psi_\delta^k\}_{\delta\in(0,1)}$ with input dimension one and output dimension $2^{k-1}+2$ such that with $\widetilde{X}_\delta^\ell := \mathrm{R}(\Psi_\delta^k)_{1+\ell-2^{k-1}}$ for $\ell \in \{2^{k-1},\dots,2^k\}$ and $\mathrm{psum}_{2^{k-1},\delta} := \mathrm{R}(\Psi_\delta^k)_{2^{k-1}+2}$ it holds that*

$$\begin{aligned}
\mathrm{R}(\Psi_\delta^k)(x) &= \big(\widetilde{X}_\delta^{2^{k-1}}(x),\dots,\widetilde{X}_\delta^{2^k}(x),\mathrm{psum}_{2^{k-1},\delta}(x)\big), \qquad x \in \hat{I},\\
\Big\|x^\ell - \widetilde{X}_\delta^\ell(x)\Big\|_{W^{1,\infty}(\hat{I})} &\leq \delta, \quad \ell \in \{2^{k-1},\dots,2^k\}, &(4.3)\\
\widetilde{X}_\delta^\ell(0) &= 0, \quad \ell \in \{2^{k-1},\dots,2^k\}, &(4.4)\\
\mathrm{psum}_{2^{k-1},\delta}(0) &= v(0), &(4.5)\\
\left\|\sum_{\ell=0}^{2^{k-1}} \tilde{v}_\ell x^\ell - \mathrm{psum}_{2^{k-1},\delta}(x)\right\|_{W^{1,\infty}(\hat{I})} &\leq \delta \sum_{\ell=2}^{2^{k-1}} |\tilde{v}_\ell|, &(4.6)\\
L(\Psi_\delta^k) &\leq C_L\big(\tfrac{1}{3}k^3 + 2k^2 + k\log_2(1/\delta)\big) + (4C_L + C_L' + 1)k, &(4.7)\\
C_1 &:= 7C_M + C_M' + C_{\mathrm{fi}} + \tfrac{1}{2}C_{\mathrm{la}} + 10,\\
C_2 &:= 4C_L' + 16C_L + 17,\\
M(\Psi_\delta^k) &\leq 2C_M k 2^k + C_M 2^k \log_2(1/\delta) + 4kC_L\log_2(1/\delta)\\
&\quad + C_1 2^k + \tfrac{4}{3}C_L k^3 + 6C_L k^2 + C_2 k, &(4.8)\\
M_{\mathrm{fi}}(\Psi_\delta^k) &\leq C_{\mathrm{fi}} + 4, &(4.9)\\
M_{\mathrm{la}}(\Psi_\delta^k) &\leq C_{\mathrm{la}} 2^{k-1} + 5. &(4.10)
\end{aligned}$$

Proof. We prove the lemma by induction over $k \in \mathbb{N}$.

Induction basis. For arbitrary $\delta \in (0,1)$ let $L_1 := L(\widetilde{\times}_{\delta/2,1})$, let $A := [1,1]^\top$ be a 2×1-matrix and let $\widetilde{\times}_{\delta/2,1} =: ((A_1,b_1),\dots,(A_{L_1},b_{L_1}))$ according to Proposition 4.1. Then we define

$$\Psi_\delta^1 := \mathrm{P}(\Phi_{1,L_1}^{\mathrm{Id}}, ((A_1A,b_1),\dots,(A_{L_1},b_{L_1})), \tilde{v}_1\Phi_{1,L_1}^{\mathrm{Id}} + \tilde{v}_0),$$

where the network $\tilde{v}_1\Phi_{1,L_1}^{\mathrm{Id}} + \tilde{v}_0$ can be constructed from $\Phi_{1,L_1}^{\mathrm{Id}}$ by adjusting the weights in the last layer. For all $x \in \hat{I}$ it holds that $\widetilde{X}_\delta^1(x) := [\mathrm{R}(\Psi_\delta^1)(x)]_1 = x$, $\widetilde{X}_\delta^2(x) := [\mathrm{R}(\Psi_\delta^1)(x)]_2 = \mathrm{R}(\widetilde{\times}_{\delta/2,1})(x,x)$ and $\mathrm{psum}_{1,\delta}(x) := [\mathrm{R}(\Psi_\delta^1)(x)]_3 = \tilde{v}_1 x + \tilde{v}_0$, which shows that Equations (4.4)–(4.6) hold for $k=1$.

We now estimate the depth and the size of Ψ_δ^1.

$$\begin{aligned}
L(\Psi_\delta^1) &= L_1 \leq C_L \log_2(2/\delta) + C_L', \\
M(\Psi_\delta^1) &= M(\Phi_{1,L_1}^{\mathrm{Id}}) + M(((A_1 A, b_1), \dots, (A_{L_1}, b_{L_1}))) + M(\tilde{v}_1 \Phi_{1,L_1}^{\mathrm{Id}} + \tilde{v}_0) \\
&\leq 2L_1 + \big(C_M \log_2(2/\delta) + C_M'\big) + (2L_1 + 1) \\
&\leq \big(4C_L + C_M\big) \log_2(2/\delta) + 4C_L' + C_M' + 1, \\
M_{\mathrm{fi}}(\Psi_\delta^1) &= M_{\mathrm{fi}}(\Phi_{1,L_1}^{\mathrm{Id}}) + M_{\mathrm{fi}}(((A_1 A, b_1), \dots, (A_{L_1}, b_{L_1}))) + M_{\mathrm{fi}}(\tilde{v}_1 \Phi_{1,L_1}^{\mathrm{Id}} + \tilde{v}_0) \\
&\leq 2 + C_{\mathrm{fi}} + 2 = C_{\mathrm{fi}} + 4, \\
M_{\mathrm{la}}(\Psi_\delta^1) &= M_{\mathrm{la}}(\Phi_{1,L_1}^{\mathrm{Id}}) + M_{\mathrm{la}}(((A_1 A, b_1), \dots, (A_{L_1}, b_{L_1}))) + M_{\mathrm{la}}(\tilde{v}_1 \Phi_{1,L_1}^{\mathrm{Id}} + \tilde{v}_0) \\
&\leq 2 + C_{\mathrm{la}} + 3 = C_{\mathrm{la}} + 5.
\end{aligned}$$

Finally, it follows from Proposition 4.1 that

$$\begin{aligned}
\Big| x^2 - \widetilde{X}_\delta^2(x) \Big|_{W^{1,\infty}(\hat{I})} &\leq \big\| 2x - [D\widetilde{\times}_{\delta/2,1}]_1(x,x) - [D\widetilde{\times}_{\delta/2,1}]_2(x,x) \big\|_{L^\infty(\hat{I})} \leq \tfrac{\delta}{2} + \tfrac{\delta}{2} = \delta, \\
\Big\| x^2 - \widetilde{X}_\delta^2(x) \Big\|_{W^{1,\infty}(\hat{I})} &\leq \delta,
\end{aligned}$$

where the last inequality follows from Poincaré's inequality and Equation (4.4). This shows that Equation (4.3) holds for $k = 1$. This finishes the proof of the induction basis.

Induction hypothesis (IH). For some $\delta \in (0,1)$ and some $k \in \mathbb{N}$ define $\theta \coloneqq 2^{-k-3}\delta$ and assume that there exists a network Ψ_θ^k for which Equations (4.3)–(4.10) hold with θ instead of δ.

Induction step. We show that Equations (4.3)–(4.10) hold with δ as in (IH) and with $k+1$ instead of k.

We note that, for all $\ell \in \{2^{k-1}, \dots, 2^k\}$,

$$\Big\| \widetilde{X}_\theta^\ell \Big\|_{L^\infty(\hat{I})} \leq \Big\| x^\ell \Big\|_{L^\infty(\hat{I})} + \Big\| x^\ell - \widetilde{X}_\theta^\ell(x) \Big\|_{W^{1,\infty}(\hat{I})} \leq 1 + \theta < 2. \tag{4.11}$$

Hence, we may use $\widetilde{X}_\theta^\ell(x)$ as input of $\widetilde{\times}_{\theta,2}$. For $\Phi^{1,k}$ and $\Phi_\delta^{2,k}$ introduced below, we define

$$\Psi_\delta^{k+1} \coloneqq \Phi_\delta^{2,k} \odot \Phi^{1,k} \odot \Psi_\theta^k. \tag{4.12}$$

Here, $\Phi^{1,k}$ is a NN of depth one which implements the linear map

$$\begin{aligned}
\mathbb{R}^{2^{k-1}+2} \to \mathbb{R}^{2^{k+1}+2} : \Big(z_1, \dots, z_{2^{k-1}+2}\Big) \mapsto (&z_{2^{k-1}+1}, z_1, z_2, z_2, z_2, z_2, z_3, z_3, z_3, z_3, z_4, z_4, z_4, \\
&\dots, z_{2^{k-1}}, z_{2^{k-1}+1}, z_{2^{k-1}+1}, z_{2^{k-1}+1}, z_{2^{k-1}+2} + \sum_{\ell=2^{k-1}+1}^{2^k} \tilde{v}_\ell z_{\ell+1-2^{k-1}}\Big).
\end{aligned}$$

The network $((A^{1,k}, b^{1,k})) \coloneqq \Phi^{1,k}$ satisfies $b^{1,k} = 0$ and

$$(A^{1,k})_{m,i} = \begin{cases}
1 & \text{if } m = 1, i = 2^{k-1} + 1, \\
1 & \text{if } m \in \{2, \dots, 2^{k+1} + 1\}, i = \lceil \tfrac{m+2}{4} \rceil, \\
\tilde{v}_{i-1+2^{k-1}} & \text{if } m = 2^{k+1} + 2, i \in \{2, \dots, 2^{k-1} + 1\}, \\
1 & \text{if } m = 2^{k+1} + 2, i = 2^{k-1} + 2, \\
0 & \text{else.}
\end{cases}$$

Moreover,

$$L(\Phi^{1,k}) = 1, \qquad M_{\mathrm{fi}}(\Phi^{1,k}) = M_{\mathrm{la}}(\Phi^{1,k}) = M(\Phi^{1,k}) \leq (1 + 2^{k+1} + 2^{k-1} + 1) = \tfrac{5}{2} 2^k + 2.$$

With $L_\theta \coloneqq L(\widetilde{\times}_{\theta,2})$ we define the network $\Phi_\delta^{2,k}$ as

$$\Phi_\delta^{2,k} \coloneqq \mathrm{FP}(\Phi_{1,L_\theta}^{\mathrm{Id}}, \widetilde{\times}_{\theta,2}, \dots, \widetilde{\times}_{\theta,2}, \Phi_{1,L_\theta}^{\mathrm{Id}}),$$

which contains 2^k $\widetilde{\times}_{\theta,2}$-networks. It holds that

$$\begin{aligned}
L(\Phi_\delta^{2,k}) &= L(\widetilde{\times}_{\theta,2}) \leq C_L\big(\log_2(2/\theta)\big) + C'_L \\
&= C_L\big(k+4+\log_2(1/\delta)\big) + C'_L, \\
M(\Phi_\delta^{2,k}) &\leq 2M(\Phi_{1,L_\theta}^{\mathrm{Id}}) + 2^k M(\widetilde{\times}_{\theta,2}) \\
&\leq 4L(\widetilde{\times}_{\theta,2}) + 2^k M(\widetilde{\times}_{\theta,2}) \\
&\leq (4C_L + C_M 2^k)\log_2(2/\theta) + 4C'_L + C'_M 2^k \\
&\leq (4C_L + C_M 2^k)(k+4+\log_2(1/\delta)) + 4C'_L + C'_M 2^k, \\
M_{\mathrm{fi}}(\Phi_\delta^{2,k}) &= 2M_{\mathrm{fi}}(\Phi_{1,L_\theta}^{\mathrm{Id}}) + 2^k M_{\mathrm{fi}}(\widetilde{\times}_{\theta,2}) \\
&\leq C_{\mathrm{fi}} 2^k + 4, \\
M_{\mathrm{la}}(\Phi_\delta^{2,k}) &= 2M_{\mathrm{la}}(\Phi_{1,L_\theta}^{\mathrm{Id}}) + 2^k M_{\mathrm{la}}(\widetilde{\times}_{\theta,2}) \\
&= C_{\mathrm{la}} 2^k + 4.
\end{aligned}$$

The realization of Ψ_δ^{k+1}, defined in Equation (4.12), is given by

$$[\mathrm{R}(\Psi_\delta^{k+1})(x)]_1 = \widetilde{X}_\theta^{2^k}(x), \qquad \text{for } x \in \hat{I}, \tag{4.13}$$

$$[\mathrm{R}(\Psi_\delta^{k+1})(x)]_{\ell+1-2^k} = \mathrm{R}(\widetilde{\times}_{\theta,2})\left(\widetilde{X}_\theta^{\lceil \ell/2 \rceil}(x), \widetilde{X}_\theta^{\lfloor \ell/2 \rfloor}(x)\right), \qquad \text{for } x \in \hat{I},\ \ell \in \{2^k+1,\dots,2^{k+1}\}, \tag{4.14}$$

$$[\mathrm{R}(\Psi_\delta^{k+1})(x)]_{2^k+2} = \mathrm{psum}_{2^{k-1},\theta}(x) + \sum_{\ell=2^{k-1}+1}^{2^k} \tilde{v}_\ell \widetilde{X}_\theta^\ell(x), \qquad \text{for } x \in \hat{I}. \tag{4.15}$$

We define, for $x \in \hat{I}$ and $\ell \in \{2^k+1,\dots,2^{k+1}\}$

$$\widetilde{X}_\delta^\ell(x) \coloneqq [\mathrm{R}(\Psi_\delta^{k+1})(x)]_{\ell+1-2^k} \text{ and } \mathrm{psum}_{2^k,\delta}(x) \coloneqq [\mathrm{R}(\Psi_\delta^{k+1})(x)]_{2^k+2}.$$

Equations (4.4)–(4.5) for $k+1$ follow from the induction hypothesis and Equation (4.2).

We will now give bounds on the depth and the size of Ψ_δ^{k+1}.

$$\begin{aligned}
L(\Psi_\delta^{k+1}) &= L(\Phi_\delta^{2,k}) + L(\Phi^{1,k}) + L(\Psi_\theta^k) \\
&\leq \big(C_L\big(k+4+\log_2(1/\delta)\big) + C'_L\big) + 1 \\
&\quad + \Big(C_L\big(\tfrac{1}{3}k^3 + 2k^2 + k\log_2(2^{k+3}/\delta)\big) + (4C_L + C'_L + 1)k\Big) \\
&\leq C_L\big(\tfrac{1}{3}(k+1)^3 + 2(k+1)^2 + (k+1)\log_2(1/\delta)\big) + (4C_L + C'_L + 1)(k+1), \\
M(\Psi_\delta^{k+1}) &\leq M(\Phi_\delta^{2,k}) + M_{\mathrm{fi}}(\Phi_\delta^{2,k}) + M_{\mathrm{la}}(\Phi^{1,k} \odot \Psi_\theta^k) + M(\Phi^{1,k} \odot \Psi_\theta^k) \\
&\leq M(\Phi_\delta^{2,k}) + M_{\mathrm{fi}}(\Phi_\delta^{2,k}) + 2M_{\mathrm{la}}(\Phi^{1,k}) + M(\Phi^{1,k}) + M_{\mathrm{fi}}(\Phi^{1,k}) + M_{\mathrm{la}}(\Psi_\theta^k) + M(\Psi_\theta^k) \\
&\leq \Big((4C_L + C_M 2^k)(k+4+\log_2(1/\delta)) + 4C'_L + C'_M 2^k\Big) + (C_{\mathrm{fi}} 2^k + 4) \\
&\quad + 2(\tfrac{5}{2}2^k + 2) + (\tfrac{5}{2}2^k + 2) + (\tfrac{5}{2}2^k + 2) + (C_{\mathrm{la}} 2^{k-1} + 5) \\
&\quad + \Big(2C_M k 2^k + C_M 2^k \log_2(2^{k+3}/\delta) + 4kC_L \log_2(2^{k+3}/\delta) + C_1 2^k + \tfrac{4}{3}k^3 C_L + 6k^2 C_L + C_2 k\Big) \\
&\leq 2C_M(k+1)2^{k+1} + C_M 2^{k+1}\log_2(1/\delta) + 4(k+1)C_L \log_2(1/\delta) \\
&\quad + C_1 2^{k+1} + \tfrac{4}{3}C_L(k+1)^3 + 6(k+1)^2 C_L + C_2(k+1), \\
C_1 &\coloneqq 7C_M + C'_M + C_{\mathrm{fi}} + \tfrac{1}{2}C_{\mathrm{la}} + 10, \\
C_2 &\coloneqq 4C'_L + 16C_L + 17, \\
M_{\mathrm{fi}}(\Psi_\delta^{k+1}) &= M_{\mathrm{fi}}(\Psi_\theta^k) \leq C_{\mathrm{fi}} + 4, \\
M_{\mathrm{la}}(\Psi_\delta^{k+1}) &= M_{\mathrm{la}}(\Phi_\delta^{2,k}) = C_{\mathrm{la}} 2^k + 4.
\end{aligned}$$

This finishes the proof of Equations (4.7)–(4.10) for $k+1$. We now estimate the NN expression error. Because $\theta < \delta$, it follows from the induction hypothesis that

$$\begin{aligned}\left\|\sum_{\ell=0}^{2^k} \tilde{v}_\ell x^\ell - \mathrm{psum}_{2^k,\delta}(x)\right\|_{W^{1,\infty}(\hat{I})} &\leq \left\|\sum_{\ell=0}^{2^{k-1}} \tilde{v}_\ell x^\ell - \mathrm{psum}_{2^{k-1},\theta}(x)\right\|_{W^{1,\infty}(\hat{I})} + \sum_{\ell=2^{k-1}+1}^{2^k} |\tilde{v}_\ell| \left\|x^\ell - \widetilde{X}_\theta^\ell(x)\right\|_{W^{1,\infty}(\hat{I})} \\ &\leq \delta \sum_{\ell=2}^{2^k} |\tilde{v}_\ell|.\end{aligned}$$

It follows from the induction hypothesis and Equation (4.13) that Equation (4.3) holds for $\ell = 2^{(k+1)-1}$. For $\ell \in \{2^k+1, \dots, 2^{k+1}\}$, with $\ell_0 := \lceil \ell/2 \rceil$, we use that, analogous to Equation (4.11), it holds that for $m \in \{\ell_0, \ell - \ell_0\}$

$$\begin{aligned}\left\|\widetilde{X}_\theta^m\right\|_{L^\infty(\hat{I})} &\leq 1 + \theta < 2, \\ \left\|\tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^m(x)\right\|_{L^\infty(\hat{I})} &\leq \left\|m x^{m-1}\right\|_{L^\infty(\hat{I})} + \left\|x^m - \widetilde{X}_\theta^m(x)\right\|_{W^{1,\infty}(\hat{I})} \leq m + \theta < m+1.\end{aligned}$$

We find

$$\begin{aligned}\left|x^\ell - \widetilde{X}_\delta^\ell(x)\right|_{W^{1,\infty}(\hat{I})} &\leq \left\|\ell_0 x^{\ell-1} - [D\mathrm{R}(\widetilde{\times}_{\theta,2})]_1\big(\widetilde{X}_\theta^{\ell_0}(x), \widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell_0}(x)\right\|_{L^\infty(\hat{I})} \\ &\quad + \left\|(\ell-\ell_0)x^{\ell-1} - [D\mathrm{R}(\widetilde{\times}_{\theta,2})]_2\big(\widetilde{X}_\theta^{\ell_0}(x),\ \widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell-\ell_0}(x)\right\|_{L^\infty(\hat{I})} \\ &\leq \left\|\ell_0 x^{\ell_0-1}\big(x^{\ell-\ell_0} - \widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\right\|_{L^\infty(\hat{I})} \\ &\quad + \left\|\widetilde{X}_\theta^{\ell-\ell_0}(x)\big(\ell_0 x^{\ell_0-1} - \tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell_0}(x)\big)\right\|_{L^\infty(\hat{I})} \\ &\quad + \left\|\big(\widetilde{X}_\theta^{\ell-\ell_0}(x) - [D\mathrm{R}(\widetilde{\times}_{\theta,2})]_1\big(\widetilde{X}_\theta^{\ell_0}(x), \widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\big)\tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell_0}(x)\right\|_{L^\infty(\hat{I})} \\ &\quad + \left\|(\ell-\ell_0)x^{\ell-\ell_0-1}\big(x^{\ell_0} - \widetilde{X}_\theta^{\ell_0}(x)\big)\right\|_{L^\infty(\hat{I})} \\ &\quad + \left\|\widetilde{X}_\theta^{\ell_0}(x)\big((\ell-\ell_0)x^{\ell-\ell_0-1} - \tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\right\|_{L^\infty(\hat{I})} \\ &\quad + \left\|\big(\widetilde{X}_\theta^{\ell_0}(x) - [D\mathrm{R}(\widetilde{\times}_{\theta,2})]_2\big(\widetilde{X}_\theta^{\ell_0}(x), \widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\big)\tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell-\ell_0}(x)\right\|_{L^\infty(\hat{I})} \\ &\overset{(4.11),(\mathrm{IH})}{\leq} \ell_0\theta + 2\theta + (\ell_0+1)\theta + (\ell-\ell_0)\theta + 2\theta + (\ell-\ell_0+1)\theta \\ &\leq (2\ell+6)\theta \leq \delta,\end{aligned}$$

where $[D\mathrm{R}(\widetilde{\times}_{\delta,2})]$ is the Jacobian and where we have used that $3 \leq \ell \leq 2^{k+1}$, which implies that $2\ell + 6 \leq 4\ell \leq 2^{k+3}$. Because $\widetilde{X}_\delta^\ell(0) = 0 = 0^\ell$ it follows with Poincaré's inequality that $\left\|x^\ell - \widetilde{X}_\delta^\ell(x)\right\|_{W^{1,\infty}(\hat{I})} \leq \delta$.

For k satisfying the induction hypothesis and arbitrary $\delta \in (0,1)$, we have constructed Ψ_δ^{k+1} and have shown that Equations (4.3)–(4.10) hold for $k+1$ instead of k and with δ as in (IH). This finishes the induction step. The lemma now follows by induction, as the induction basis shows the induction hypothesis for $k = 1$. □

Proof of Proposition 4.2. Below, we consider the case $C_0 > \beta$. The proof of the case $C_0 \leq \beta$ is analogous. The distinction is needed to ensure that we do not invoke Lemma 4.3 with $\delta \geq 1$.

In case $n \in \{0,1\}$, for all $\beta \in (0,1)$ we define $\Phi_\beta^v := ((A,b))$, where $A = \tilde{v}_1 \in \mathbb{R}^{1\times 1}$ and $b = \tilde{v}_0 \in \mathbb{R}^1$. It holds that $\left\|v - \mathrm{R}(\Phi_\beta^v)\right\|_{W^{1,\infty}(\hat{I})} = 0$, $\mathrm{R}(\Phi_\beta^v)(0) = \tilde{v}_0 = v(0)$, $L(\Phi_\beta^v) = 1$ and $M(\Phi_\beta^v) = M_{\mathrm{fi}}(\Phi_\beta^v) = M_{\mathrm{la}}(\Phi_\beta^v) \leq 2$.

In case $n \geq 2$ we define $k := \lceil \log_2(n) \rceil$ and $\delta := \beta/C_0$ and use Lemma 4.3. We define

$$\Phi_\beta^v := \Phi^{3,n} \odot \Psi_\delta^k,$$

where $\Phi^{3,n}$ is a NN which implements the linear map

$$\mathbb{R}^{2^{k-1}+2} \to \mathbb{R} : (z_1, \dots, z_{2^{k-1}+2}) \mapsto z_{2^{k-1}+2} + \sum_{\ell=2^{k-1}+1}^{2^k} \tilde{v}_\ell z_{\ell+1-2^{k-1}}.$$

It satisfies $L(\Phi^{3,n}) = 1$ and $M(\Phi^{3,n}) = M_{\mathrm{fi}}(\Phi^{3,n}) = M_{\mathrm{la}}(\Phi^{3,n}) \leq 2^{k-1} + 1$.

The realization of Φ_β^v is

$$\mathrm{R}(\Phi_\beta^v)(x) = \mathrm{psum}_{2^{k-1},\delta}(x) + \sum_{\ell=2^{k-1}+1}^{2^k} \tilde{v}_\ell \widetilde{X}_\delta^\ell(x), \qquad x \in \hat{I}.$$

From Equations (4.4) and (4.5) we conclude that $\mathrm{R}(\Phi_\beta^v)(0) = \tilde{v}_0 = v(0)$.

Using $2^k \leq 2n$, we can bound the depth and the size of Φ_β^v as follows:

$$\begin{aligned}
L\left(\Phi_\beta^v\right) &= L\left(\Phi^{3,n}\right) + L\left(\Psi_\delta^k\right) \\
&\leq 1 + \left(C_L\left(\tfrac{1}{3}k^3 + k\log_2\left(\tfrac{C_0}{\beta}\right)\right) + Ck^2\right) \\
&\leq C_L(1+\log_2(n))\log_2\left(\tfrac{C_0}{\beta}\right) + \tfrac{1}{3}C_L\log_2^3(n) + C\log_2^2(n), \\
M\left(\Phi_\beta^v\right) &\leq M\left(\Phi^{3,n}\right) + M_{\mathrm{fi}}\left(\Phi^{3,n}\right) + M_{\mathrm{la}}\left(\Psi_\delta^k\right) + M\left(\Psi_\delta^k\right) \\
&\leq \left(2^{k-1}+1\right) + \left(2^{k-1}+1\right) + \left(2^{k-1}C_{\mathrm{la}} + 5\right) \\
&\quad + \left(2C_M k 2^k + C_M 2^k \log_2\left(\frac{C_0}{\beta}\right) + 4kC_L\log_2\left(\frac{C_0}{\beta}\right) + C2^k\right) \\
&\leq 2C_M n\log_2\left(\frac{C_0}{\beta}\right) + 4C_M n\log_2(n) + 4C_L(1+\log_2(n))\log_2\left(\frac{C_0}{\beta}\right) + Cn, \\
M_{\mathrm{fi}}\left(\Phi_\beta^v\right) &= M_{\mathrm{fi}}\left(\Psi_\delta^k\right) = C_{\mathrm{fi}} + 4, \\
M_{\mathrm{la}}\left(\Phi_\beta^v\right) &= 2M_{\mathrm{la}}\left(\Phi^{3,n}\right) \leq 2n + 2.
\end{aligned}$$

Finally, we estimate the error.

$$\begin{aligned}
\left\|v - \mathrm{R}\left(\Phi_\beta^v\right)\right\|_{W^{1,\infty}(\hat{I})} &\leq \left\|\sum_{\ell=0}^{2^{k-1}} \tilde{v}_\ell x^\ell - \mathrm{psum}_{2^{k-1},\delta}(x)\right\|_{W^{1,\infty}(\hat{I})} + \sum_{\ell=2^{k-1}+1}^{2^k} |\tilde{v}_\ell| \left\|x^\ell - \widetilde{X}_\delta^\ell(x)\right\|_{W^{1,\infty}(\hat{I})} \\
&\leq \delta \sum_{\ell=2}^{2^{k-1}} |\tilde{v}_\ell| + \sum_{\ell=2^{k-1}+1}^{2^k} |\tilde{v}_\ell|\delta \leq \beta.
\end{aligned}$$

This finishes the proof of the proposition. □

Later, we will consider approximations of *piecewise* polynomial functions by realizations of NNs. For the results in Section 5, it is important that we can approximate polynomials on an interval with exactness in the endpoints. After subtracting an affine function, it suffices to approximate polynomials which vanish at the endpoints by NNs the realizations of which vanish at the endpoints. This is the aim of the following proposition.

In Section 5, we will mainly restrict our attention to estimates of the error in the H^1-norm. Therefore, the error estimates in the following proposition are L^2-based.

Proposition 4.4. *For all $q \in \mathbb{N}_{\geq 2}$ and all $w \in (\mathbb{P}_q \cap H_0^1)(\hat{I})$ there exist NNs $(\Phi_\varepsilon^{w,0})_{\varepsilon\in(0,1)}$ with input*

dimension one and output dimension one which satisfy $\mathrm{R}(\Phi_\varepsilon^{w,0})|_{\mathbb{R}\setminus\hat{I}} = 0$ *and for all* $1 \leq q' \leq \infty$

$$\begin{aligned}
\left\|w - \mathrm{R}(\Phi_\varepsilon^{w,0})\right\|_{W^{1,q'}(\hat{I})} &\leq 2^{\frac{2}{q'}-1}\varepsilon\,|w|_{H^1(\hat{I})}\,,\\
L\left(\Phi_\varepsilon^{w,0}\right) &\leq C_L(1+\log_2(q))\,(2q+\log_2(1/\varepsilon)) + C_L\log_2(1/\varepsilon) + C\log_2^3(q),\\
M\left(\Phi_\varepsilon^{w,0}\right) &\leq 2C_M\left(2q^2 + q\log_2(1/\varepsilon)\right) + (6C_L\,(1+\log_2(q)) + 2C_M)\log_2(1/\varepsilon)\\
&\quad + Cq\log_2(q),\\
M_{\mathrm{fi}}\left(\Phi_\varepsilon^{w,0}\right) &\leq C_{\mathrm{fi}} + 12,\\
M_{\mathrm{la}}(\Phi_\varepsilon^{w,0}) &= C_{\mathrm{la}}.
\end{aligned}$$

In particular, for $q' = 2$ *it holds that* $\left\|w - \mathrm{R}(\Phi_\varepsilon^{w,0})\right\|_{H^1(\hat{I})} \leq \varepsilon\,|w|_{H^1(\hat{I})}$.

Proof. The main observation in the proof is the fact that the polynomial w is divisible by ψ, known as *quadratic bubble function* and defined by $\psi(x) \coloneqq (1+x)(1-x) = 1-x^2$ for $x \in \hat{I}$ and $\psi(x) = 0$ for $x \in \mathbb{R}\setminus\hat{I}$. In addition, we use that ψ can be approximated with $W^{1,\infty}(\hat{I})$-error at most $\eta > 0$ by a NN Φ_η^ψ which satisfies $\mathrm{R}(\Phi_\varepsilon^\psi)|_{\mathbb{R}\setminus\hat{I}} = 0$ and we approximate $Q \coloneqq w/\psi \in \mathbb{P}_{q-2}(\hat{I})$ using Proposition 4.2. We use the product network from Proposition 4.1 to multiply the approximation of ψ with the approximation of Q. In order to apply Proposition 4.2 for the approximation of Q, we need to bound the sum of the absolute values of the Taylor coefficients of Q. In the first step of the proof we will derive such a bound. In the second step we construct networks which satisfy the desired properties.

Step 1. We first estimate the sum of the absolute values of the Taylor coefficients of the $L^2(\hat{I})$-normalized Legendre polynomials $\{L_j\}_{j\in\mathbb{N}_0}$. For $j \in \mathbb{N}_0$, it holds that $L_j(x) = \sum_{\ell=0}^j c_\ell^j x^\ell$ for $x \in \mathbb{R}$, where, for $\ell \in \mathbb{N}$ and $m \coloneqq (j-\ell)/2$,

$$c_\ell^j \coloneqq \begin{cases} 0 & \text{for } j-\ell \in \{0,\dots,j\} \cap 2\mathbb{Z}+1,\\ (-1)^m 2^{-j}\binom{j}{m}\binom{j+\ell}{j}\sqrt{j+\frac{1}{2}} & \text{for } j-\ell \in \{0,\dots,j\}\cap 2\mathbb{Z},\\ 0 & \text{for } \ell > j. \end{cases}$$

The sum of these coefficients can be estimated using the following inequalities (cf. [29]):

$$\forall n \in \mathbb{N}: \qquad \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}e^{\frac{1}{12n}}. \tag{4.16}$$

In addition, we will use that for all $j \in \mathbb{N}$ and all $m \in \{0,\dots,\lfloor j/2\rfloor\}$

$$\binom{2j-2m}{j} \leq \binom{2j-2m}{j}\prod_{i=0}^{2m-1}\frac{2j-i}{j-i} = \binom{2j}{j}.$$

It follows, from (4.16) that, for all $j \in \mathbb{N}$,

$$\binom{2j}{j} \leq \frac{\sqrt{2\pi}(2j)^{2j+\frac{1}{2}}e^{-2j}e^{\frac{1}{24j}}}{\sqrt{2\pi}j^{j+\frac{1}{2}}e^{-j}e^{\frac{1}{12j+1}}\sqrt{2\pi}j^{j+\frac{1}{2}}e^{-j}e^{\frac{1}{12j+1}}} \leq \frac{4^j}{\sqrt{\pi j}}\frac{e^{\frac{1}{24j}}}{e^{\frac{2}{12j+1}}} < \frac{4^j}{\sqrt{\pi j}}$$

and as a result that

$$\begin{aligned}
\sum_{\ell=0}^j |c_\ell^j| &= \sum_{m\in\{0,\dots,\lfloor j/2\rfloor\}} |c_{j-2m}^j|\\
&\leq 2^{-j}\left(\sum_{m=0}^j\binom{j}{m}\right)\max_{m=0}^{\lfloor j/2\rfloor}\binom{2j-2m}{j}\sqrt{j+\frac{1}{2}}\\
&\leq \sqrt{j+\frac{1}{2}}\binom{2j}{j} \leq \frac{4^j\sqrt{j+\frac{1}{2}}}{\sqrt{\pi j}} \leq 4^j.
\end{aligned} \tag{4.17}$$

We now consider a general polynomial $v \in \mathbb{P}_n$ of degree $n \in \mathbb{N}_0$. We denote the Legendre expansion of v by $v = \sum_{j=0}^n v_j L_j$. We find the following expression for the Taylor expansion of v at $x = 0$:

$$v(x) = \sum_{j=0}^n v_j L_j(x) = \sum_{j=0}^n v_j \sum_{\ell=0}^j c_\ell^j x^\ell = \sum_{\ell=0}^n \left(\sum_{j=0}^n v_j c_\ell^j \right) x^\ell =: \sum_{\ell=0}^n \tilde{v}_\ell x^\ell, \qquad x \in \hat{I}.$$

It follows that

$$\begin{aligned} \sum_{\ell=0}^n |\tilde{v}_\ell| = \sum_{\ell=0}^n \left| \sum_{j=0}^n v_j c_\ell^j \right| &\leq \left(\max_{j=0}^n |v_j| \right) \sum_{\ell=0}^n \sum_{j=0}^n |c_\ell^j| = \left(\max_{j=0}^n |v_j| \right) \sum_{j=0}^n \left(\sum_{\ell=0}^n |c_\ell^j| \right) \\ &\overset{(*)}{\leq} \|v\|_{L^2(\hat{I})} \sum_{j=0}^n 4^j \leq \tfrac{1}{3} 4^{n+1} \|v\|_{L^2(\hat{I})}. \end{aligned} \tag{4.18}$$

At (*) we used Equation (4.17) and

$$\max_{j=0}^n |v_j| \leq \left\| (v_j)_{j=0}^n \right\|_{\ell^2} = \|v\|_{L^2(\hat{I})}. \tag{4.19}$$

We now consider $w \in (\mathbb{P}_q \cap H_0^1)(\hat{I})$ of degree $q \geq 2$ and write $w = \psi Q$, where $Q \in \mathbb{P}_{q-2}(\hat{I})$. We recall Hardy's inequality: for all functions $g \in H^1((0,1))$ satisfying $g(0) = 0$, it holds that $\left\| \frac{g(x)}{x} \right\|_{L^2((0,1))} \leq 2 \|g'\|_{L^2((0,1))}$. It follows that

$$\begin{aligned} \|Q\|_{L^2(\hat{I})}^2 &= \left\| \frac{w(x)}{1-x^2} \right\|_{L^2(\hat{I})}^2 = \left\| \frac{w(x)}{1-x^2} \right\|_{L^2((-1,0))}^2 + \left\| \frac{w(x)}{1-x^2} \right\|_{L^2((0,1))}^2 \\ &\leq \left\| \frac{w(x)}{1+x} \right\|_{L^2((-1,0))}^2 + \left\| \frac{w(x)}{1-x} \right\|_{L^2((0,1))}^2 = \left\| \frac{w(y-1)}{y} \right\|_{L^2((0,1))}^2 + \left\| \frac{w(1-z)}{z} \right\|_{L^2((0,1))}^2 \\ &\leq 2^2 \left\| w'(y-1) \right\|_{L^2((0,1))}^2 + 2^2 \left\| w'(1-z) \right\|_{L^2((0,1))}^2 = 2^2 \left\| w' \right\|_{L^2((-1,0))}^2 + 2^2 \left\| w' \right\|_{L^2((0,1))}^2 \\ &= 2^2 |w|_{H^1(\hat{I})}^2. \end{aligned} \tag{4.20}$$

Writing $Q(x) = \sum_{\ell=0}^{q-2} \widetilde{Q}_\ell x^\ell$ for $x \in \hat{I}$, it follows from Equation (4.18) with $v = Q$, $\tilde{v}_\ell = \widetilde{Q}_\ell$ and $n = q-2$ that

$$\sum_{\ell=0}^{q-2} |\widetilde{Q}_\ell| \leq \tfrac{1}{6} 4^q |w|_{H^1(\hat{I})}. \tag{4.21}$$

We now estimate the $W^{1,\infty}(\hat{I})$-norm of Q. Writing $Q = \sum_{j=0}^{q-2} Q_j L_j$ it follows from Equation (4.19) for $v = Q$ and $v_j = Q_j$ and from Equation (4.20) that for all $j \in \{0, \dots, q-2\}$

$$|Q_j| \leq 2 |w|_{H^1(\hat{I})}.$$

Using that for all $j \in \mathbb{N}_0 : \|L_j\|_{L^\infty(\hat{I})} = \sqrt{j+1/2} \leq \sqrt{j+1} \leq 1 + j/2$, we find

$$\begin{aligned} \|Q\|_{L^\infty(\hat{I})} &\leq \sum_{j=0}^{q-2} |Q_j| \, \|L_j\|_{L^\infty(\hat{I})} \\ &\leq 2 |w|_{H^1(\hat{I})} \sum_{j=0}^{q-2} (1 + \tfrac{j}{2}) \\ &\leq 2 |w|_{H^1(\hat{I})} \left(q - 1 + \frac{(q-1)(q-2)}{4} \right) \\ &= \tfrac{1}{2} \left(q^2 + q - 2 \right) |w|_{H^1(\hat{I})} \leq (q^2 - 1) |w|_{H^1(\hat{I})}. \end{aligned}$$

By Markov's inequality (e.g. [9, Chapter 4, Theorem 1.4]), we get

$$|Q|_{W^{1,\infty}(\hat{I})} \leq (q-2)^2 \|Q\|_{L^\infty(\hat{I})} \leq (q-2)^2 \left(q^2 - 1 \right) |w|_{H^1(\hat{I})}$$

and hence, since $q \geq 2$, $\|Q\|_{W^{1,\infty}(\hat{I})} \leq (q^4 - 1)\,|w|_{H^1(\hat{I})}$.

Step 2. Let $\varepsilon \in (0,1)$. We first assume that $|w|_{H^1(\hat{I})} = 1$ and define $\beta \coloneqq \varepsilon/36$ and $\eta \coloneqq \varepsilon(12q^4)^{-1}$.

We write $w = \psi Q$ and approximate ψ by a NN whose realization is supported in $\hat{I}$. To approximate Q we use Φ_β^Q from Proposition 4.2, with $C_0 \leq \frac{1}{6}4^q$ according to Equation (4.21). We will use that

$$\left\|\mathrm{R}\left(\Phi_\beta^Q\right)\right\|_{W^{1,\infty}(\hat{I})} \leq \|Q\|_{W^{1,\infty}(\hat{I})} + \beta \leq (q^4-1)\,|w|_{H^1(\hat{I})} + \beta \leq q^4.$$

With the 2×1-matrix $A \coloneqq [1,-1]^\top$ and the vector $b \coloneqq (1,1)^\top$ we define $\Phi_\eta^\psi \coloneqq \widetilde{\times}_{\frac{\eta}{2},1} \odot ((A,b))$, which has realization $\mathrm{R}(\Phi_\eta^\psi)(x) = \mathrm{R}\left(\widetilde{\times}_{\frac{\eta}{2},1}\right)\big(\varrho(1+x), \varrho(1-x)\big)$ for $x \in \mathbb{R}$. By Equation (4.2), it follows that $\mathrm{R}(\Phi_\eta^\psi)|_{\mathbb{R}\setminus\hat{I}} = 0$. It holds that $L(\Phi_\eta^\psi) = C_L \log_2(2/\eta) + C_L' + 1$,

$$\begin{aligned} M(\Phi_\eta^\psi) &\leq M\left(\widetilde{\times}_{\frac{\eta}{2},1}\right) + M_{\mathrm{fi}}\left(\widetilde{\times}_{\frac{\eta}{2},1}\right) + M_{\mathrm{la}}(((A,b))) + M(((A,b))) \\ &\leq \left(C_M \log_2\left(\frac{2}{\eta}\right) + C_M'\right) + C_{\mathrm{fi}} + 4 + 4, \end{aligned}$$

$M_{\mathrm{fi}}(\Phi_\eta^\psi) \leq 2M_{\mathrm{fi}}(((A,b))) = 8$, and $M_{\mathrm{la}}(\Phi_\eta^\psi) = C_{\mathrm{la}}$. The error can be estimated as follows:

$$\begin{aligned} \left|\psi - \mathrm{R}(\Phi_\eta^\psi)\right|_{W^{1,\infty}(\hat{I})} &= \left\| \tfrac{\mathrm{d}}{\mathrm{d}x}\psi(x) - \tfrac{\mathrm{d}}{\mathrm{d}x}\left(\mathrm{R}\left(\widetilde{\times}_{\frac{\eta}{2},1}\right)(1+x,1-x)\right)\right\|_{L^\infty(\hat{I})} \\ &\leq \left\| \left((1-x) - \left[D\mathrm{R}\left(\widetilde{\times}_{\frac{\eta}{2},1}\right)\right]_1 (1+x,1-x)\right) \tfrac{\mathrm{d}}{\mathrm{d}x}(1+x)\right\|_{L^\infty(\hat{I})} \\ &\quad + \left\| \left((1+x) - \left[D\mathrm{R}\left(\widetilde{\times}_{\frac{\eta}{2},1}\right)\right]_2 (1+x,1-x)\right) \tfrac{\mathrm{d}}{\mathrm{d}x}(1-x)\right\|_{L^\infty(\hat{I})} \\ &\leq \tfrac{\eta}{2} + \tfrac{\eta}{2} = \eta. \end{aligned}$$

Because $\mathrm{R}(\Phi_\eta^\psi)(\pm 1) = 0 = \psi(\pm 1)$, it follows from Poincaré's inequality that $\left\|\psi - \mathrm{R}(\Phi_\eta^\psi)\right\|_{L^\infty(\hat{I})} \leq \eta$. As a result,

$$\left\|\mathrm{R}\left(\Phi_\eta^\psi\right)\right\|_{W^{1,\infty}(\hat{I})} \leq \|\psi\|_{W^{1,\infty}(\hat{I})} + \left\|\psi - \mathrm{R}\left(\Phi_\eta^\psi\right)\right\|_{W^{1,\infty}(\hat{I})} \leq 2 + \eta \leq 3.$$

We define

$$K \coloneqq \max\{2, \|Q\|_{L^\infty(\hat{I})} + \beta\} \leq \max\{2, (q^2-1)\,|w|_{H^1(\hat{I})} + \beta\} \leq \max\{2, q^2\} \leq q^2.$$

The last inequality holds because $q \geq 2$. The definition of K is such that $\left\|\mathrm{R}(\Phi_\eta^\psi)\right\|_{L^\infty(\hat{I})}, \left\|\mathrm{R}(\Phi_\beta^Q)\right\|_{L^\infty(\hat{I})} \leq K$. With $L_* \coloneqq L(\Phi_\beta^Q) - L(\Phi_\eta^\psi) \leq L(\Phi_\beta^Q)$, we define

$$\Phi_\varepsilon^{w,0} \coloneqq \begin{cases} \widetilde{\times}_{\eta,K} \odot \mathrm{P}\left(\Phi_\beta^Q, \Phi_{1,L_*}^{\mathrm{Id}} \odot \Phi_\eta^\psi\right), & \text{for } L_* > 0, \\ \widetilde{\times}_{\eta,K} \odot \mathrm{P}\left(\Phi_\beta^Q, \Phi_\eta^\psi\right), & \text{for } L_* = 0, \\ \widetilde{\times}_{\eta,K} \odot \mathrm{P}\left(\Phi_{1,-L_*}^{\mathrm{Id}} \odot \Phi_\beta^Q, \Phi_\eta^\psi\right), & \text{for } L_* < 0. \end{cases}$$

By Equation (4.2) and the fact that $\mathrm{R}(\Phi_\eta^\psi)|_{\mathbb{R}\setminus\hat{I}} = 0$, it follows that $\mathrm{R}(\Phi_\varepsilon^{w,0})|_{\mathbb{R}\setminus\hat{I}} = 0$.

For the estimate on the network depth and the network size, we only need to consider the case $L(\Phi_\beta^Q) > L(\Phi_\eta^\psi)$, for the following reason. We have two upper bounds: $L(\Phi_\eta^\psi) \leq 4C_L \log_2(q) + C_L \log_2(1/\varepsilon) + C$ and $L(\Phi_\beta^Q) \leq C_L q(1 + \log_2(q)) + C_L \log_2(q)\log_2(1/\varepsilon) + C(1 + \log_2^3(q))$. In addition, by Propositions 2.2 and 2.4, it follows that we can increase the depth of the network Φ_β^Q such that Φ_β^Q still satisfies the properties of Proposition 4.2, possibly with a larger universal constant in the estimate on the network size, and such that $L(\Phi_\beta^Q) \geq C(\log_2(q))^3$ for some $C > 0$. It then follows that $L(\Phi_\beta^Q) > L(\Phi_\eta^\psi)$ for sufficiently large $q \geq 2$. This implies that bounds on the size and the depth derived under the assumption that $L(\Phi_\beta^Q) > L(\Phi_\eta^\psi)$ also hold in case $L(\Phi_\beta^Q) \leq L(\Phi_\eta^\psi)$. The latter inequality only holds for finitely many q, and these cases can be covered by increasing the universal constants.

Assuming that $L(\Phi_\beta^Q) > L(\Phi_\eta^\psi)$, it follows that

$$\begin{aligned}
L\left(\Phi_\varepsilon^{w,0}\right) &= L\left(\widetilde{\times}_{\eta,K}\right) + L\left(\Phi_\beta^Q\right) \\
&\leq \left(C_L \log_2\left(\tfrac{K}{\eta}\right) + C_L'\right) + \left(C_L(1+\log_2(q))\left(2q + \log_2\left(\tfrac{1}{\beta}\right)\right) + C\log_2^3(q)\right) \\
&\leq C_L\left(6\log_2(q) + \log_2\left(\tfrac{12}{\varepsilon}\right)\right) + C_L\left(1+\log_2(q)\right)\left(2q + \log_2\left(\tfrac{36}{\varepsilon}\right)\right) + C\log_2^3(q) \\
&\leq C_L\left(1+\log_2(q)\right)\left(2q + \log_2\left(\tfrac{1}{\varepsilon}\right)\right) + C_L\log_2\left(\tfrac{1}{\varepsilon}\right) + C\log_2^3(q).
\end{aligned}$$

Moreover,

$$\begin{aligned}
M\left(\Phi_\varepsilon^{w,0}\right) &\leq M\left(\widetilde{\times}_{\eta,K}\right) + M_{\mathrm{fi}}\left(\widetilde{\times}_{\eta,K}\right) + M_{\mathrm{la}}\left(\Phi_\beta^Q\right) + M_{\mathrm{la}}\left(\Phi_{1,L_*}^{\mathrm{Id}} \odot \Phi_\eta^\psi\right) + M\left(\Phi_\beta^Q\right) + M\left(\Phi_{1,L_*}^{\mathrm{Id}} \odot \Phi_\eta^\psi\right) \\
&\leq M\left(\widetilde{\times}_{\eta,K}\right) + M_{\mathrm{fi}}\left(\widetilde{\times}_{\eta,K}\right) + M_{\mathrm{la}}\left(\Phi_\beta^Q\right) + 2M_{\mathrm{la}}\left(\Phi_{1,L_*}^{\mathrm{Id}}\right) \\
&\quad + M\left(\Phi_\beta^Q\right) + M\left(\Phi_{1,L_*}^{\mathrm{Id}}\right) + M_{\mathrm{fi}}\left(\Phi_{1,L_*}^{\mathrm{Id}}\right) + M_{\mathrm{la}}\left(\Phi_\eta^\psi\right) + M\left(\Phi_\eta^\psi\right) \\
&\leq \left(C_M\log_2\left(\tfrac{K}{\eta}\right) + C_M'\right) + C_{\mathrm{fi}} + (2q-2) + 4 \\
&\quad + \left(2C_M q\left(2q + \log_2\left(\tfrac{1}{\beta}\right)\right) + 4C_M q\log_2(q) + 4C_L(1+\log_2(q))\left(2q + \log_2\left(\tfrac{1}{\beta}\right)\right) + Cq\right) \\
&\quad + 2\left(C_L(1+\log_2(q))\left(2q + \log_2\left(\tfrac{1}{\beta}\right)\right) + C\log_2^3(q)\right) + 2 + C_{\mathrm{la}} \\
&\quad + \left(C_M\log_2\left(\tfrac{2}{\eta}\right) + C_M' + C_{\mathrm{fi}} + 8\right) \\
&\leq C_M\left(6\log_2(q) + \log_2\left(\tfrac{12}{\varepsilon}\right)\right) + 2C_M\left(2q^2 + q\log_2\left(\tfrac{36}{\varepsilon}\right)\right) + 6C_L(1+\log_2(q))\log_2\left(\tfrac{36}{\varepsilon}\right) \\
&\quad + C_M\left(4\log_2(q) + \log_2\left(\tfrac{24}{\varepsilon}\right)\right) + Cq\log_2(q) \\
&\leq 2C_M\left(2q^2 + q\log_2\left(\tfrac{1}{\varepsilon}\right)\right) + \left(6C_L(1+\log_2(q)) + 2C_M\right)\log_2\left(\tfrac{1}{\varepsilon}\right) + Cq\log_2(q), \\
M_{\mathrm{fi}}\left(\Phi_\varepsilon^{w,0}\right) &= M_{\mathrm{fi}}\left(\Phi_\beta^Q\right) + M_{\mathrm{fi}}\left(\Phi_\eta^\psi\right) = (C_{\mathrm{fi}}+4) + 8 = C_{\mathrm{fi}} + 12, \\
M_{\mathrm{la}}\left(\Phi_\varepsilon^{w,0}\right) &= M_{\mathrm{la}}\left(\widetilde{\times}_{\eta,K}\right) = C_{\mathrm{la}}.
\end{aligned}$$

The approximation error can be estimated by

$$\begin{aligned}
2\left|w - \mathrm{R}\left(\Phi_\varepsilon^{w,0}\right)\right|_{W^{1,\infty}(\hat{I})} &= 2\left\|\tfrac{\mathrm{d}}{\mathrm{d}x}\left(Q(x)\psi(x)\right) - \tfrac{\mathrm{d}}{\mathrm{d}x}\left(\mathrm{R}\left(\widetilde{\times}_{\eta,K}\right)\left(\mathrm{R}(\Phi_\beta^Q)(x), \mathrm{R}(\Phi_\eta^\psi)(x)\right)\right)\right\|_{L^\infty(\hat{I})} \\
&\leq 2\left\|\left(\psi(x) - \mathrm{R}\left(\Phi_\eta^\psi\right)(x)\right)\tfrac{\mathrm{d}}{\mathrm{d}x}Q(x)\right\|_{L^\infty(\hat{I})} \\
&\quad + 2\left\|\mathrm{R}\left(\Phi_\eta^\psi\right)(x)\tfrac{\mathrm{d}}{\mathrm{d}x}\left(Q(x) - \mathrm{R}\left(\Phi_\beta^Q\right)(x)\right)\right\|_{L^\infty(\hat{I})} \\
&\quad + 2\left\|\left(\mathrm{R}\left(\Phi_\eta^\psi\right)(x) - \left[D\mathrm{R}\left(\widetilde{\times}_{\eta,K}\right)\right]_1\left(\mathrm{R}\left(\Phi_\beta^Q\right)(x), \mathrm{R}\left(\Phi_\eta^\psi\right)(x)\right)\right)\tfrac{\mathrm{d}}{\mathrm{d}x}\mathrm{R}\left(\Phi_\beta^Q\right)(x)\right\|_{L^\infty(\hat{I})} \\
&\quad + 2\left\|Q(x)\tfrac{\mathrm{d}}{\mathrm{d}x}\left(\psi(x) - \mathrm{R}\left(\Phi_\eta^\psi\right)(x)\right)\right\|_{L^\infty(\hat{I})} \\
&\quad + 2\left\|\left(Q(x) - \mathrm{R}\left(\Phi_\beta^Q\right)(x)\right)\tfrac{\mathrm{d}}{\mathrm{d}x}\mathrm{R}\left(\Phi_\eta^\psi\right)(x)\right\|_{L^\infty(\hat{I})} \\
&\quad + 2\left\|\left(\mathrm{R}\left(\Phi_\beta^Q\right)(x) - \left[D\mathrm{R}\left(\widetilde{\times}_{\eta,K}\right)\right]_2\left(\mathrm{R}\left(\Phi_\beta^Q\right)(x), \mathrm{R}\left(\Phi_\eta^\psi\right)(x)\right)\right)\tfrac{\mathrm{d}}{\mathrm{d}x}\mathrm{R}\left(\Phi_\eta^\psi\right)(x)\right\|_{L^\infty(\hat{I})} \\
&\leq 2\eta\left|Q\right|_{W^{1,\infty}(\hat{I})} + 2\left\|\mathrm{R}\left(\Phi_\eta^\psi\right)\right\|_{L^\infty(\hat{I})}\beta + 2\eta\left|\mathrm{R}\left(\Phi_\beta^Q\right)\right|_{W^{1,\infty}(\hat{I})} \\
&\quad + 2\left\|Q\right\|_{L^\infty(\hat{I})}\eta + 2\beta\left|\mathrm{R}\left(\Phi_\eta^\psi\right)\right|_{W^{1,\infty}(\hat{I})} + 2\eta\left|\mathrm{R}\left(\Phi_\eta^\psi\right)\right|_{W^{1,\infty}(\hat{I})} \\
&\leq \tfrac{\varepsilon}{6} + \tfrac{\varepsilon}{6} + \tfrac{\varepsilon}{6} + \tfrac{\varepsilon}{6}q^{-2} + \tfrac{\varepsilon}{6} + \tfrac{\varepsilon}{2}q^{-4} \overset{(*)}{\leq} \varepsilon.
\end{aligned}$$

At (*) we used that $q \geq 2$. It follows from Poincaré's inequality and $\Phi_\varepsilon^{w,0}(\pm 1) = 0 = w(\pm 1)$ that $\varepsilon/2$

also bounds the $L^\infty(\hat{I})$-error. Finally, we get from Hölder's inequality for all $1 \leq q' < \infty$

$$\begin{aligned}\left\|w - \mathrm{R}\left(\Phi_\varepsilon^{w,0}\right)\right\|_{W^{1,q'}(\hat{I})}^{q'} &= \left\|w - \mathrm{R}\left(\Phi_\varepsilon^{w,0}\right)\right\|_{L^{q'}(\hat{I})}^{q'} + \left|w - \mathrm{R}\left(\Phi_\varepsilon^{w,0}\right)\right|_{W^{1,q'}(\hat{I})}^{q'} \\ &\leq 2|\hat{I}| \cdot \left\|w - \mathrm{R}\left(\Phi_\varepsilon^{w,0}\right)\right\|_{W^{1,\infty}(\hat{I})}^{q'} \\ &\leq 4\left(\tfrac{\varepsilon}{2}\right)^{q'} = 2^{2-q'}\varepsilon^{q'}.\end{aligned}$$

This finishes the proof in case $|w|_{H^1(\hat{I})} = 1$.

If $|w|_{H^1(\hat{I})} = 0$, then $w \in H_0^1(\hat{I})$ implies that $w \equiv 0$, which can be implemented exactly by a NN of depth 1 and size 0. If $|w|_{H^1(\hat{I})} > 0$ we can use the linearity of the output layer of NNs: we can approximate $w/|w|_{H^1(\hat{I})}$ as before, and multiply the weights in the output layer by $|w|_{H^1(\hat{I})}$, which gives the desired result. This finishes the proof of the proposition. □

Remark 4.5. *We note that by Hölder's inequality for all $2 \leq q' \leq \infty$*

$$|w|_{H^1(\hat{I})} \leq 2^{\frac{1}{2}-\frac{1}{q'}} |w|_{W^{1,q'}(\hat{I})}.$$

Because w' is a polynomial of degree $q-1$, it follows that for all $1 \leq q' \leq 2$

$$|w|_{H^1(\hat{I})} \leq \left((q'+1)(q-1)^2\right)^{\frac{1}{q'}-\frac{1}{2}} |w|_{W^{1,q'}(\hat{I})} \leq 2(q-1)\,|w|_{W^{1,q'}(\hat{I})}.$$

5 Finite Element Spaces

Based on the efficient approximation of polynomials of the previous section we can now present an emulation of higher-order spline approximations, approximations by Chebyshev polynomials, and *hp*-FEM approximations which correspond to so-called free-knot, variable-degree spline approximations ([31] and the references there).

5.1 Approximation of Piecewise Polynomials

We start by demonstrating how to emulate piecewise polynomial functions in general.

Proposition 5.1. *For all $\boldsymbol{p} = (p_i)_{i\in\{1,\dots,N\}} \subset \mathbb{N}$, all partitions $\mathcal{T}$ of $I = (0,1)$ with N elements and all $v \in S_{\boldsymbol{p}}(I,\mathcal{T})$, for $0 < \varepsilon < 1$ there exist NNs $\{\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\}_{\varepsilon\in(0,1)}$ such that for all $1 \leq q' \leq \infty$*

$$\begin{aligned}\left\|v - \mathrm{R}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right)\right\|_{W^{1,q'}(I)} &\leq \varepsilon\,|v|_{W^{1,q'}(I)}, \\ L\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right) &\leq C_L(1+\log_2(p_{\max}))\,(2p_{\max}+\log_2(1/\varepsilon)) + C_L\log_2(1/\varepsilon) + C\left(1+\log_2^3(p_{\max})\right), \\ M\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right) &\leq 4C_M\sum_{i=1}^N p_i^2 + 2C_M\log_2(1/\varepsilon)\sum_{i=1}^N p_i + \log_2(1/\varepsilon)\,C\left(1+\sum_{i=1}^N \log_2(p_i)\right) \\ &\quad + C\left(1+\sum_{i=1}^N p_i\log_2(p_i)\right) \\ &\quad + 2N\left(C_L(1+\log_2(p_{\max}))\,(2p_{\max}+\log_2(1/\varepsilon)) + C\left(1+\log_2^3(p_{\max})\right)\right), \\ M_{\mathrm{fi}}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right) &\leq 6N, \\ M_{\mathrm{la}}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right) &\leq 2N+2.\end{aligned}$$

In addition, it holds that $\mathrm{R}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right)(x_j) = v(x_j)$ for all $j \in \{0,\dots,N\}$, where $\{x_j\}_{j=0}^N$ are the nodes of $\mathcal{T}$.

Proof. We write v as the sum of its continuous, piecewise linear interpolant $\bar{v} \in S_{\mathbf{1}}(I, \mathcal{T})$ and a function $v - \bar{v} \in S_{\boldsymbol{p}}(I, \mathcal{T})$ which satisfies $(v - \bar{v})(x_j) = 0$ for $j \in \{0, \dots, N\}$. The network $\Phi^{\bar{v}}$, constructed in Lemma 3.1, satisfies

$$\mathrm{R}\left(\Phi^{\bar{v}}\right) = \bar{v}, \quad L\left(\Phi^{\bar{v}}\right) = 2, \quad M\left(\Phi^{\bar{v}}\right) \leq 3N + 1, \quad M_{\mathrm{fi}}\left(\Phi^{\bar{v}}\right) \leq 2N \text{ and } M_{\mathrm{la}}\left(\Phi^{\bar{v}}\right) \leq N + 1. \tag{5.1}$$

For all $i \in \{1, \dots, N\}$, we denote by $P_i : \mathbb{R} \to \mathbb{R} : x \mapsto \frac{2}{h_i}(x - \frac{x_{i-1}+x_i}{2})$ the affine transformation which satisfies $P_i(I_i) = \hat{I}$, $P_i(x_{i-1}) = -1$ and $P_i(x_i) = 1$.

Let

$$\gamma_i(q') \coloneqq \tfrac{\varepsilon}{2} 2^{1-\frac{2}{q'}} \begin{cases} 2^{\frac{1}{q'}-\frac{1}{2}} & \text{if } 2 \leq q' \leq \infty, \\ (2p_i)^{-1} & \text{if } 1 \leq q' < 2. \end{cases}$$

It follows that $\frac{1}{\gamma_i(q')} \leq \frac{2}{\varepsilon}$ for $2 \leq q' \leq \infty$ and $\frac{1}{\gamma_i(q')} \leq 8p_i\frac{1}{\varepsilon} =: \frac{1}{\varepsilon_i}$ for $1 \leq q' \leq 2$, hence $\frac{1}{\gamma_i(q')} \leq \frac{1}{\varepsilon_i}$ for $1 \leq q' \leq \infty$.

For $w_i \coloneqq (v - \bar{v})|_{I_i} \in (\mathbb{P}_{p_i} \cap H^1_0)(I_i)$, it holds that $\hat{w}_i \coloneqq w_i \circ P_i^{-1} \in (\mathbb{P}_{p_i} \cap H^1_0)(\hat{I})$, hence Proposition 4.4 shows the existence of a NN $\Phi_{\varepsilon_i}^{\hat{w}_i,0}$ such that $\mathrm{R}(\Phi_{\varepsilon_i}^{\hat{w}_i,0})|_{\mathbb{R}\setminus\hat{I}} = 0$ and

$$\begin{aligned}
L\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) &\leq C_L(1 + \log_2(p_i))\left(2p_i + \log_2\left(\tfrac{1}{\varepsilon_i}\right)\right) + C_L \log_2\left(\tfrac{1}{\varepsilon_i}\right) + C\left(1 + \log_2^3(p_i)\right), \\
&\leq C_L(1 + \log_2(p_i))\left(2p_i + \log_2\left(\tfrac{1}{\varepsilon}\right)\right) + C_L \log_2\left(\tfrac{1}{\varepsilon}\right) + C\left(1 + \log_2^3(p_i)\right), \\
M\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) &\leq 2C_M\left(2p_i^2 + p_i \log_2\left(\tfrac{1}{\varepsilon_i}\right)\right) + (6C_L(1 + \log_2(p_i)) + 2C_M)\log_2\left(\tfrac{1}{\varepsilon_i}\right) \\
&\quad + C\left(1 + p_i \log_2(p_i)\right), \\
&\leq 2C_M\left(2p_i^2 + p_i \log_2\left(\tfrac{1}{\varepsilon}\right)\right) + (6C_L(1 + \log_2(p_i)) + 2C_M)\log_2\left(\tfrac{1}{\varepsilon}\right) \\
&\quad + C\left(1 + p_i \log_2(p_i)\right), \\
M_{\mathrm{fi}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) &\leq C_{\mathrm{fi}} + 12, \\
M_{\mathrm{la}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) &= C_{\mathrm{la}}.
\end{aligned}$$

The affine transformation P_i can be implemented exactly by a NN Φ^{P_i} of depth 1 satisfying $M(\Phi^{P_i}) = M_{\mathrm{fi}}(\Phi^{P_i}) = M_{\mathrm{la}}(\Phi^{P_i}) = 2$. Now, the concatenation $\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}$ approximates w_i. It holds by Proposition 4.4 that $\mathrm{R}(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i})|_{\mathbb{R}\setminus I_i} = 0$ and that

$$\begin{aligned}
\left\| w_i - \mathrm{R}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right)\right\|_{W^{1,q'}(I_i)} &= \left(\tfrac{h_i}{2}\right)^{\frac{1}{q'}-1} \left\| \hat{w}_i - \mathrm{R}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right)\right\|_{W^{1,q'}(\hat{I})} \\
&\leq \left(\tfrac{h_i}{2}\right)^{\frac{1}{q'}-1} 2^{\frac{2}{q'}-1} \gamma_i(q') \left|\hat{w}_i\right|_{H^1(\hat{I})} \\
&\leq \left(\tfrac{h_i}{2}\right)^{\frac{1}{q'}-1} 2^{\frac{2}{q'}-1} \gamma_i(q') \left(\left|(v|_{I_i}) \circ P_i^{-1}\right|_{H^1(\hat{I})} + \left|(\bar{v}|_{I_i}) \circ P_i^{-1}\right|_{H^1(\hat{I})}\right) \\
&\overset{(*)}{\leq} \left(\tfrac{h_i}{2}\right)^{\frac{1}{q'}-1} 2^{\frac{2}{q'}-1} \gamma_i(q') 2 \left|(v|_{I_i}) \circ P_i^{-1}\right|_{H^1(\hat{I})} \\
&\overset{(**)}{\leq} \left(\tfrac{h_i}{2}\right)^{\frac{1}{q'}-1} \varepsilon \left|(v|_{I_i}) \circ P_i^{-1}\right|_{W^{1,q'}(\hat{I})} \\
&= \varepsilon \left|(v|_{I_i})\right|_{W^{1,q'}(I_i)}.
\end{aligned}$$

At (*) we used that $\left|(\bar{v}|_{I_i}) \circ P_i^{-1}\right|_{H^1(\hat{I})} \leq \left|(v|_{I_i}) \circ P_i^{-1}\right|_{H^1(\hat{I})}$, which follows e.g. from the fact that $\bar{v}'|_{I_i} \circ P_i^{-1}$ is a truncation of the Legendre expansion of $v'|_{I_i} \circ P_i^{-1}$. At (**) we used a result similar to

Remark 4.5, for $q = p_i$ and for $(v|_{I_i}) \circ P_i^{-1} \in \mathbb{P}_{p_i}(\hat{I})$ instead of w. In addition, it follows that

$$\begin{aligned}
L\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right) &\leq 1 + C_L(1+\log_2(p_i))\left(2p_i + \log_2\left(\tfrac{1}{\varepsilon}\right)\right) + C_L \log_2\left(\tfrac{1}{\varepsilon}\right) + C\left(1+\log_2^3(p_i)\right),\\
M\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right) &\leq M\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) + M_{\mathrm{fi}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) + M_{\mathrm{la}}\left(\Phi^{P_i}\right) + M\left(\Phi^{P_i}\right)\\
&\leq \Big(2C_M\left(2p_i^2 + p_i\log_2\left(\tfrac{1}{\varepsilon}\right)\right) + \left(6C_L(1+\log_2(p_i)) + 2C_M\right)\log_2\left(\tfrac{1}{\varepsilon}\right)\\
&\quad + C\left(1+p_i\log_2(p_i)\right)\Big) + (C_{\mathrm{fi}}+12) + 2 + 2\\
&\leq 2C_M\left(2p_i^2 + p_i\log_2\left(\tfrac{1}{\varepsilon}\right)\right) + \left(6C_L(1+\log_2(p_i)) + 2C_M\right)\log_2\left(\tfrac{1}{\varepsilon}\right)\\
&\quad + C\left(1+p_i\log_2(p_i)\right),\\
M_{\mathrm{fi}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right) &\leq 2M_{\mathrm{fi}}\left(\Phi^{P_i}\right) = 4,\\
M_{\mathrm{la}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right) &= M_{\mathrm{la}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) = C_{\mathrm{la}}.
\end{aligned}$$

Let $\{\ell_j\}_{j\in\{1,\dots,N+1\}} \subset \mathbb{N}$ be such that

$$\begin{aligned}
\ell_1 + L\left(\Phi^{\bar{v}}\right) &= \ell_2 + L\left(\Phi_{\varepsilon_1}^{\hat{w}_1,0} \odot \Phi^{P_1}\right) = \ldots = \ell_{N+1} + L\left(\Phi_{\varepsilon_N}^{\hat{w}_N,0} \odot \Phi^{P_N}\right)\\
&= 1 + \max\left\{L\left(\Phi^{\bar{v}}\right), \max_{i=1}^{N} L\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right)\right\}\\
&\leq 3 + \max_{i=1}^{N} L\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right),
\end{aligned}$$

where the inequality follows from $L(\Phi^{\bar{v}}) = 2$. In addition, we have

$$\begin{aligned}
\max_{j=1}^{N+1} \ell_j &\leq 3 + \max_{i=1}^{N} L\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right)\\
&\leq C_L(1+\log_2(p_{\max}))\left(2p_{\max} + \log_2\left(\tfrac{1}{\varepsilon}\right)\right) + C_L\log_2\left(\tfrac{1}{\varepsilon}\right) + C\left(1+\log_2^3(p_{\max})\right).
\end{aligned}$$

We define $\Phi_{N+1}^{\mathrm{Sum}} \coloneqq (([1,\dots,1],0))$, where $[1,\dots,1]$ is an $1\times(N+1)$-matrix. It holds that $L(\Phi_{N+1}^{\mathrm{Sum}}) = 1$ and $M(\Phi_{N+1}^{\mathrm{Sum}}) = M_{\mathrm{fi}}(\Phi_{N+1}^{\mathrm{Sum}}) = M_{\mathrm{la}}(\Phi_{N+1}^{\mathrm{Sum}}) = N+1$. We now define $\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}$ by

$$\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}} \coloneqq \Phi_{N+1}^{\mathrm{Sum}} \odot \mathrm{P}(\Phi_{1,\ell_1}^{\mathrm{Id}} \odot \Phi^{\bar{v}}, \Phi_{1,\ell_2}^{\mathrm{Id}} \odot \Phi_{\varepsilon_1}^{\hat{w}_1,0} \odot \Phi^{P_1}, \dots, \Phi_{1,\ell_{N+1}}^{\mathrm{Id}} \odot \Phi_{\varepsilon_N}^{\hat{w}_N,0} \odot \Phi^{P_N}).$$

Because the realisation of $\Phi^{\bar{v}}$ equals $\bar{v}$, it holds that $\mathrm{R}(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}})|_{I_i} = \bar{v}|_{I_i} + \mathrm{R}(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i})$ for all $i \in \{1,\dots,N\}$. The depth and the size of $\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}$ can be estimated as follows:

$$\begin{aligned}
L\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right) &\leq L\left(\Phi_{N+1}^{\mathrm{Sum}}\right) + \ell_1 + L\left(\Phi^{\bar{v}}\right) \leq 1 + \max_{j=1}^{N+1}\ell_j + 1\\
&\leq C_L(1+\log_2(p_{\max}))\left(2p_{\max} + \log_2\left(\tfrac{1}{\varepsilon}\right)\right) + C_L\log_2\left(\tfrac{1}{\varepsilon}\right) + C\left(1+\log_2^3(p_{\max})\right),\\
M\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right) &\leq M\left(\Phi_{N+1}^{\mathrm{Sum}}\right) + M_{\mathrm{fi}}\left(\Phi_{N+1}^{\mathrm{Sum}}\right) + M_{\mathrm{la}}\left(\Phi_{1,\ell_1}^{\mathrm{Id}} \odot \Phi^{\bar{v}}\right) + \sum_{i=1}^{N} M_{\mathrm{la}}\left(\Phi_{1,\ell_{i+1}}^{\mathrm{Id}} \odot \Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right)\\
&\quad + M\left(\Phi_{1,\ell_1}^{\mathrm{Id}} \odot \Phi^{\bar{v}}\right) + \sum_{i=1}^{N} M\left(\Phi_{1,\ell_{i+1}}^{\mathrm{Id}} \odot \Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right)\\
&\leq M\left(\Phi_{N+1}^{\mathrm{Sum}}\right) + M_{\mathrm{fi}}\left(\Phi_{N+1}^{\mathrm{Sum}}\right) + \sum_{i=0}^{N} 2M_{\mathrm{la}}\left(\Phi_{1,\ell_{i+1}}^{\mathrm{Id}}\right) + M\left(\Phi_{1,\ell_1}^{\mathrm{Id}}\right) + M_{\mathrm{fi}}\left(\Phi_{1,\ell_1}^{\mathrm{Id}}\right) + M_{\mathrm{la}}\left(\Phi^{\bar{v}}\right)\\
&\quad + M\left(\Phi^{\bar{v}}\right) + \sum_{i=1}^{N}\Big(M\left(\Phi_{1,\ell_{i+1}}^{\mathrm{Id}}\right) + M_{\mathrm{fi}}\left(\Phi_{1,\ell_{i+1}}^{\mathrm{Id}}\right) + M_{\mathrm{la}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) + M\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right)\\
&\quad + M_{\mathrm{fi}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) + M_{\mathrm{la}}\left(\Phi^{P_i}\right) + M\left(\Phi^{P_i}\right)\Big)\\
&\leq (N+1) + (N+1) + 4(N+1) + 2\max_{j=1}^{N+1} + 2 + 2N + (3N+1) + 2N\max_{j=1}^{N+1}\ell_j + 2N + C_{\mathrm{la}}N
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{N}\Big(2C_M\left(2p_i^2+p_i\log_2\left(\tfrac{1}{\varepsilon}\right)\right)+(6C_L(1+\log_2(p_i))+2C_M)\log_2\left(\tfrac{1}{\varepsilon}\right)\\
&+C\left(1+p_i\log_2(p_i)\right)\Big)+(C_{\mathrm{fi}}+12)N+2N+2N\\
\leq\ & 4C_M\sum_{i=1}^{N}p_i^2+2C_M\log_2\left(\tfrac{1}{\varepsilon}\right)\sum_{i=1}^{N}p_i+\log_2\left(\tfrac{1}{\varepsilon}\right)C\left(1+\sum_{i=1}^{N}\log_2(p_i)\right)+C\left(1+\sum_{i=1}^{N}p_i\log_2(p_i)\right)\\
&+2(N+1)\left(C_L(1+\log_2(p_{\max}))\left(2p_{\max}+\log_2\left(\tfrac{1}{\varepsilon}\right)\right)+C\left(1+\log_2^3(p_{\max})\right)\right),\\
M_{\mathrm{fi}}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right)\leq\ & M_{\mathrm{fi}}\left(\Phi^{\bar v}\right)+\sum_{i=1}^{N}2M_{\mathrm{fi}}\left(\Phi^{P_i}\right)\leq 2N+4N=6N,\\
M_{\mathrm{la}}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right)\leq\ & 2M_{\mathrm{la}}\left(\Phi_{N+1}^{\mathrm{Sum}}\right)=2N+2.
\end{aligned}
$$

To estimate the error we use that $\mathrm{R}(\Phi_{\varepsilon_i}^{\hat w_i,0}\odot\Phi^{P_i})|_{I_j}=0$ for $j\neq i$:

$$
\begin{aligned}
\left\|v-\mathrm{R}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right)\right\|_{W^{1,q'}(I)}^{q'}&=\left\|\sum_{i=1}^{N}w_i-\sum_{i=1}^{N}\mathrm{R}\left(\Phi_{\varepsilon_i}^{\hat w_i,0}\odot\Phi^{P_i}\right)\right\|_{W^{1,q'}(I)}^{q'}\\
&=\sum_{i=1}^{N}\left\|w_i-\mathrm{R}\left(\Phi_{\varepsilon_i}^{\hat w_i,0}\odot\Phi^{P_i}\right)\right\|_{W^{1,q'}(I_i)}^{q'}\\
&\leq\sum_{i=1}^{N}\varepsilon^{q'}\left|(v|_{I_i})\right|_{W^{1,q'}(I_i)}^{q'}=\varepsilon^{q'}\left|v\right|_{W^{1,q'}(I)}^{q'},
\end{aligned}
$$

where w_i is extended to I such that $w_i|_{I\setminus I_i}=0$. Finally, because $\mathrm{R}\left(\Phi_{\varepsilon_i}^{\hat w_i,0}\odot\Phi^{P_i}\right)(x_j)=0$ for all $i\in\{1,\dots,N\}$ and all $j\in\{0,\dots,N\}$, it follows that $\mathrm{R}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right)(x_j)=\mathrm{R}\left(\Phi^{\bar v}\right)(x_j)=v(x_j)$ for all $j\in\{0,\dots,N\}$. This finishes the proof. □

5.2 Free-knot Spline Approximation

The following classical result due to Petruchev [27] and Oswald [24] describes the rates of best approximation of Besov-regular functions by free-knot splines of fixed degree. This setting and the corresponding approximation rate bounds correspond to the so-called "h-adaptive FEM".

Theorem 5.2 ([24, Theorems 3 and 6])**.** *Let $q,q',t,t',s,s'\in(0,\infty]$, $p\in\mathbb{N}$, and*

$$
q<q',\quad s<p+1/q,\quad s'<s-1/q+1/q'.
$$

Then, there exists a $C_3:=C(q,q',t,t',s,s',p)>0$ and, for every $N\in\mathbb{N}$ and every f in $B^s_{q,t}(I)$, there exists $h^N\in S_p^N(I)$ such that

$$
\left\|f-h^N\right\|_{B^{s'}_{q',t'}(I)}\leq C_3N^{-(s-s')}\|f\|_{B^s_{q,t}(I)}. \tag{5.2}
$$

Moreover,

$$
\left\|h^N\right\|_{B^{s}_{q,t}(I)}\leq C_3\|f\|_{B^s_{q,t}(I)}. \tag{5.3}
$$

Equation (5.2) follows from [24, Theorem 3], where p,p',q,q' in their notation correspond with q,q',t,t' in our notation, where $k-1$ corresponds with p, where λ' corresponds with $p+1/q'$ ([24, Proposition 1]), where δ corresponds with $\max\{0,1/q-1\}$ ([24, Section 1]), where N equals 1 and where n corresponds with N. The assumptions in [24, Theorems 3] are that $0<q<q'\leq\infty$, that $0<t,t'\leq\infty$, that $0<s\leq(p+1)+\max\{0,1/q-1\}$ (equality only if $t=\infty$) and that $0<s'<\min\{p+1/q',s-1/q+1/q'\}$.

Under those assumptions Equation (5.3) follows from [24, Theorem 6], where λ corresponds with $p+1/q$ ([24, Proposition 1]) and under the additional assumption that $s<p+1/q$.

Remark 5.3 ([24])**.** *In fact, h^N is* of defect one *(or,* of minimal defect*), i.e. $h^N\in C^{p-1}(I)$.*

As a consequence of Theorem 5.2, we obtain the following result describing the approximation of Besov-regular functions by ReLU NNs.

Theorem 5.4. *Let $0 < q < q' \leq \infty$, $q' \geq 1$, $0 < t \leq \infty$. Let $p \in \mathbb{N}$, $0 < s' \leq 1 < s < p + 1/q$, $1-1/q' < s-1/q$ and $s' < 1$ if $p = 1$ and $q' = \infty$. Then, there exists a constant $C_4 \coloneqq C(q, q', t, s, s', p) > 0$ and, for every $N \in \mathbb{N}$ and every $f \in B^s_{q,t}(I)$, there exists a NN Φ_f^N such that*

$$\left\| f - \mathrm{R}\left(\Phi_f^N\right) \right\|_{W^{s',q'}(I)} \leq C_4 N^{-(s-s')} \|f\|_{B^s_{q,t}(I)} \tag{5.4}$$

and

$$L\left(\Phi_f^N\right) \leq C_L(1+\log_2(p))\left(2p + (s-s')\log_2(N)\right) + C_L(s-s')\log_2(N) + C\left(1+\log_2^3(p)\right), \tag{5.5}$$

$$\begin{aligned} M\left(\Phi_f^N\right) \leq\ & 4C_M N p^2 + 2C_M(s-s')N\log_2(N)p + C(s-s')\log_2(N)\left(1+N\log_2(p)\right) \\ & + C\left(1+Np\log_2(p)\right) \\ & + 2N\left(C_L(1+\log_2(p))\left(2p+(s-s')\log_2(N)\right) + C\left(1+\log_2^3(p)\right)\right), \end{aligned} \tag{5.6}$$

$$M_{\mathrm{fi}}\left(\Phi_f^N\right) \leq 6N, \tag{5.7}$$

$$M_{\mathrm{la}}\left(\Phi_f^N\right) \leq 2N + 2. \tag{5.8}$$

Proof. Let $p \in \mathbb{N}$, $s, s', q, q', t > 0$, and $f \in B^s_{q,t}(I)$ be as in the statement of the theorem.

The assumptions on p, s, s', q, q', and t allow us to apply Theorem 5.2 with $t' \coloneqq \min\{q', 2\}$. Hence there exists $C(q, q', t, s, s', p) > 0$ and $h^N \in S_p^N(I)$ such that

$$\left\| f - h^N \right\|_{B^{s'}_{q',\min\{q',2\}}(I)} \leq C(q, q', t, s, s', p) N^{-(s-s')} \|f\|_{B^s_{q,t}(I)} \tag{5.9}$$

and

$$\left\| h^N \right\|_{B^s_{q,t}(I)} \leq C(q, q', t, s, s', p) \|f\|_{B^s_{q,t}(I)}. \tag{5.10}$$

By [24, Equation 6] or [39, Equation (1.3.3/3)], $B^{s'}_{q',\min\{q',2\}}(I)$ is continuously embedded in $W^{s',q'}(I)$. Hence

$$\|u\|_{W^{s',q'}(I)} \leq C(s', q') \|u\|_{B^{s'}_{q',\min\{q',2\}}(I)} \quad \text{for all } u \in B^{s'}_{q',\min\{q',2\}}(I). \tag{5.11}$$

Applying Equation (5.11) to Equation (5.9) yields that

$$\left\| f - h^N \right\|_{W^{s',q'}(I)} \leq C(q, q', t, s, s', p) N^{-(s-s')} \|f\|_{B^s_{q,t}(I)}. \tag{5.12}$$

We invoke Proposition 5.1 with $\varepsilon = N^{-(s-s')}$, $v = h^N$ and polynomial degree distribution $\boldsymbol{p} = (p_i)_{i=1}^N$, where $p_i = p$. This yields a network Φ_f^N such that

$$\left\| h^N - \mathrm{R}\left(\Phi_f^N\right) \right\|_{W^{s',q'}(I)} \leq C(s', q') \left\| h^N - \mathrm{R}\left(\Phi_f^N\right) \right\|_{W^{1,q'}(I)} \leq C(s', q') N^{-(s-s')} \left\| h^N \right\|_{W^{1,q'}} \tag{5.13}$$

and Equations (5.5)–(5.8) hold. Invoking [24, Equation 6] or [39, Equation (1.3.3/3)] again, we obtain that

$$\left\| h^N \right\|_{W^{1,q'}(I)} \leq C(q') \left\| h^N \right\|_{B^1_{q',\min\{q',2\}}(I)} \leq C(q, q', s, t) \left\| h^N \right\|_{B^s_{q,t}(I)} \leq C(q, q', t, s, s', p) \|f\|_{B^s_{q,t}(I)}, \tag{5.14}$$

where the second estimate holds by [40, Section 3.3.1, Equation (7)] since $s - 1/q > 1 - 1/q'$ and the last estimate follows from Equation (5.10).

We have by the triangle inequality and by invoking Equations (5.12), (5.13), and (5.14) that

$$\begin{aligned} \left\| f - \mathrm{R}\left(\Phi_f^N\right) \right\|_{W^{s',q'}(I)} &\leq \left\| f - h^N \right\|_{W^{s',q'}(I)} + \left\| h^N - \mathrm{R}\left(\Phi_f^N\right) \right\|_{W^{s',q'}(I)} \\ &\leq C(q, q', t, s, s', p) N^{-(s-s')} \|f\|_{B^s_{q,t}(I)}. \end{aligned}$$

This yields Equation (5.4) and completes the proof. □

Remark 5.5. *Note that, if $s' = 1$, then we could also obtain the estimate of Equation (5.14) by applying the inverse triangle inequality to Equation (5.9). Hence, for $s' = 1$, Equation (5.3) is not required for the proof of Theorem 5.4. As is clear from the discussion after Theorem 5.2 the statement of that theorem holds without Equation (5.3) when replacing the assumption $s < p+1/q$ by the weaker $s \leq p+1+\max\{0, 1/q-1\}$. Hence, in the case $s' = 1$, Theorem 5.4 can be improved.*

Theorem 5.2 excludes the case $s' = 0$, which is treated separately in [24, Theorem 5]. Using that result, it is not hard to see that Theorem 5.4 can be extended to situations where $s' = 0$.

5.3 Spectral Methods

We now study ReLU NN emulations of spectral element approximations. We first show that on a given partition $\mathcal{T}$ of $I = (0,1)$ spectral FEM for $r \in \mathbb{N}_0$ and $u \in H^{r+1}(I)$ can be emulated by ReLU NNs. We will demonstrate that the H^1-error decreases algebraically with the network size. Concretely, this decay happens at least with rate $r/2$. This is half the convergence rate of spectral FEM in terms of degrees of freedom, which in Theorem 5.8 equals $Np + 1$. This reduction in the convergence rate is caused by the fact that the size of the networks constructed in Proposition 4.4 depends quadratically on the polynomial degree, whereas the the number of degrees of freedom depends linearly on the polynomial degree.

Theorem 5.6 ([33, Theorem 3.17])**.** *Let $\mathcal{T}$ be a partition of $I = (0,1)$ with N elements, let $r \in \mathbb{N}_0$, $u \in H^{r+1}(I)$ and $p \in \mathbb{N}$. Then for $\boldsymbol{p} \coloneqq (p, \dots, p)$ there exists a $v \in S_{\boldsymbol{p}}(I, \mathcal{T})$ such that for all $s \in \mathbb{N}_0$ satisfying $s \leq \min\{r, p\}$*

$$\|u - v\|_{H^1(I)} \leq C_5(r) \left(\tfrac{h}{p}\right)^s |u|_{H^{s+1}(I)}.$$

Remark 5.7. *Inspection of the proof of Theorem 5.6 reveals that $v'|_{I_i}$ is a truncation of the Legendre expansion of $u'|_{I_i}$ for all $i \in \{1, \dots, N\}$, which implies that $|v|_{H^1(I)} \leq |u|_{H^1(I)}$.*

Theorem 5.8. *Let $I = (0,1)$, $r \in \mathbb{N}_0$, $u \in H^{r+1}(I)$ and $p \in \mathbb{N}$. For all partitions $\mathcal{T}$ of I with N elements there exists a NN $\Phi^{u,\mathcal{T},p}$ such that for all $s \in \mathbb{N}_0$ satisfying $s \leq \min\{r, p\}$*

$$\begin{aligned}
\left\|u - \mathrm{R}(\Phi^{u,\mathcal{T},p})\right\|_{H^1(I)} &\leq \big(1 + C_5(r)\big) \left(\tfrac{h}{p}\right)^s \|u\|_{H^{s+1}(I)}, \\
L(\Phi^{u,\mathcal{T},p}) &\leq 2C_L p \log_2(p) + C_L r(2 + \log_2(p)) \log_2\left(\tfrac{p}{h}\right) + C(1 + \log_2(p))^3, \\
M(\Phi^{u,\mathcal{T},p}) &\leq N[4C_M p^2 + 2C_M rp \log_2\left(\tfrac{p}{h}\right) + r \log_2\left(\tfrac{p}{h}\right) C(1 + \log_2(p) + C(1 + p\log_2(p))], \\
M_{\mathrm{fi}}(\Phi^{u,\mathcal{T},p}) &\leq 6N, \\
M_{\mathrm{la}}(\Phi^{u,\mathcal{T},p}) &\leq 2N + 2.
\end{aligned}$$

Proof. For v as in Theorem 5.6 and for $\boldsymbol{p} = (p, \dots, p)$ we apply Proposition 5.1 and define $\Phi^{u,\mathcal{T},p} \coloneqq \Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}$ with $\varepsilon = \left(\tfrac{h}{p}\right)^r$. Using Remark 5.7, it follows that

$$\begin{aligned}
\left\|u - \mathrm{R}(\Phi^{u,\mathcal{T},p})\right\|_{H^1(I)} &\leq \|u - v\|_{H^1(I)} + \left\|v - \mathrm{R}(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}})\right\|_{H^1(I)} \\
&\leq C_5(r) \left(\tfrac{h}{p}\right)^s |u|_{H^{s+1}(I)} + \left(\tfrac{h}{p}\right)^r |v|_{H^1(I)} \\
&\leq \big(1 + C_5(r)\big) \left(\tfrac{h}{p}\right)^s \|u\|_{H^{s+1}(I)}, \\
L(\Phi^{u,\mathcal{T},p}) &\leq C_L(1 + \log_2(p))\big(2p + r \log_2\left(\tfrac{p}{h}\right)\big) + C_L r \log_2\left(\tfrac{p}{h}\right) + C(1 + \log_2(p))^3 \\
&\leq 2C_L p(1 + \log_2(p)) + C_L r(2 + \log_2(p)) \log_2\left(\tfrac{p}{h}\right) + C(1 + \log_2^3(p)), \\
M(\Phi^{u,\mathcal{T},p}) &\leq 4C_M N p^2 + 2C_M r \log_2\left(\tfrac{p}{h}\right) Np + r \log_2\left(\tfrac{p}{h}\right) NC(1 + \log_2(p)) \\
&\quad + NC(1 + p\log_2(p)) + 2N\left(C_L(1 + \log_2(p))\big(2p + r \log_2\left(\tfrac{p}{h}\right)\big) + C(1 + \log_2^3(p))\right) \\
&\leq N\left(4C_M p^2 + 2C_M rp \log_2\left(\tfrac{p}{h}\right) + r \log_2\left(\tfrac{p}{h}\right) C(1 + \log_2(p)) + C(1 + p\log_2(p))\right), \\
M_{\mathrm{fi}}(\Phi^{u,\mathcal{T},p}) &= M_{\mathrm{fi}}(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}) \leq 6N, \\
M_{\mathrm{la}}(\Phi^{u,\mathcal{T},p}) &= M_{\mathrm{la}}(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}) \leq 2N + 2.
\end{aligned}$$

This finishes the proof. □

We now study exponential expressive power bounds for deep ReLU NN emulation of spectral approximations of functions which are analytic on $\hat{I} = (-1,1)$ and admit a holomorphic continuation to the Bernstein ellipse $\mathcal{E}_r \subset \mathbb{C}$ for some $r > 1$. We recall that for $r > 1$ the Bernstein ellipse $\mathcal{E}_r \subset \mathbb{C}$ is defined as $\mathcal{E}_r \coloneqq \{\frac{z+z^{-1}}{2} \in \mathbb{C} : 1 \leq |z| \leq r\}$. For neural networks with certain smooth activation functions, this has been investigated in [22]. A similar result is given in [10], but under considerably stronger assumptions on the regularity of the function, namely that its Taylor series converges absolutely on $[-1,1]$, which implies that it admits a holomorphic continuation to the complex unit disk.

For a function u which is analytic on $\mathcal{E}_r$ we define $M(\ln r) \coloneqq \max_{z\in\mathcal{E}_r} |u(z)|$. Moreover, for $p \in \mathbb{N}$ let the Gauss-Lobatto Chebyshev nodes be defined as $x_i = -\cos(i\pi/p)$ for $i \in \{0,\dots,p\}$.

Theorem 5.9 ([37, Corollary 4.5])**.** *Let $u : [-1,1] \to \mathbb{R}$ be an analytic function which admits an analytic continuation to the Bernstein ellipse $\mathcal{E}_{r_0} \subset \mathbb{C}$ for some $r_0 > 1$. Define $\eta_0 \coloneqq \ln(r_0)$.*

Then for every $p \in \mathbb{N}$ the polynomial $v \in \mathbb{P}_p([-1,1])$ interpolating u in the Gauss-Lobatto Chebyshev nodes $\{x_i\}_{i\in\{0,\dots,p\}}$ satisfies for every $k \in \mathbb{N}_0$ and every $0 < \eta < \eta_0$

$$|u-v|^2_{H^k(\hat{I})} \leq \int_{-1}^{1} \frac{|D^k u(x) - D^k v(x)|^2}{(1-x^2)^{1/2}} dx \leq C_6(k) \frac{M(\eta)}{\sinh(\eta)} p^{2k} e^{-\eta p},$$

where $C_6(k)$ is independent of η and p. In particular, it holds that $v(\pm 1) = u(\pm 1)$.

Using this polynomial approximation result, we readily obtain DNN expression rate bounds.

Theorem 5.10. *Let $u : [-1,1] \to \mathbb{R}$ be an analytic function which admits an analytic continuation to the Bernstein ellipse $\mathcal{E}_{r_0} \subset \mathbb{C}$ for some $r_0 > 1$. Define $\eta_0 \coloneqq \ln(r_0)$.*

Then there exist NNs $\{\Phi^{u,p,0}\}_{p\in\mathbb{N}}$ such that $\mathrm{R}(\Phi^{u,p,0})(\pm 1) = u(\pm 1)$ for all $p \in \mathbb{N}$ and such that

$$\begin{aligned}
\left\|u - \mathrm{R}(\Phi^{u,p,0})\right\|_{H^1(\hat{I})} &\leq \left(C \left(\frac{M(\eta)}{\sinh(\eta)} \right)^{1/2} + |u|_{H^1(\hat{I})} \right) p e^{-\eta p/2}, \\
L(\Phi^{u,p,0}) &\leq (2+\eta/2) C_L p (1 + \log_2(p)) + C(\eta) p, \\
M(\Phi^{u,p,0}) &\leq (4+\eta) C_M p^2 + C(\eta) p \log_2(p), \\
M_{\mathrm{fi}}(\Phi^{u,p,0}) &\leq 4, \\
M_{\mathrm{la}}(\Phi^{u,p,0}) &\leq 4.
\end{aligned}$$

Proof. Let $p \in \mathbb{N}$. Let v be as given by Theorem 5.9. Let $I \coloneqq (0,1)$ and let $P : \mathbb{R} \to \mathbb{R} : x \mapsto 2x-1$ denote the affine transformation which satisfies $P(I) = \hat{I}$, $P(0) = -1$ and $P(1) = 1$. The affine transformation P^{-1} can be implemented exactly by a NN $\Phi^{P^{-1}}$ of depth 1 satisfying $M\left(\Phi^{P^{-1}}\right) = M_{\mathrm{fi}}\left(\Phi^{P^{-1}}\right) = M_{\mathrm{la}}\left(\Phi^{P^{-1}}\right) = 2$.

Note that $v \circ P \in \mathbb{P}_p(I)$. With $\varepsilon \coloneqq e^{-\eta p/2}$ and with $\Phi_\varepsilon^{v\circ P,\{I\},p}$ as constructed in Theorem 5.1 (with $N = 1$), we define $\Phi^{u,p,0} \coloneqq \Phi_\varepsilon^{v\circ P,\{I\},p} \odot \Phi^{P^{-1}}$. It follows that

$$\begin{aligned}
\left\|v - \mathrm{R}\left(\Phi^{u,p,0}\right)\right\|^2_{H^1(\hat{I})} &\leq 2 \left\| v\circ P - \mathrm{R}\left(\Phi_\varepsilon^{v\circ P,\{I\},p}\right)\right\|^2_{H^1(I)} \\
&\leq 2\varepsilon^2 |v\circ P|^2_{H^1(I)} \\
&= \varepsilon^2 |v|^2_{H^1(\hat{I})}.
\end{aligned}$$

By construction, it holds that $\mathrm{R}(\Phi^{u,p,0})(\pm 1) = v(\pm 1) = u(\pm 1)$.

It follows from Proposition 5.1 that for all $0 < \eta < \eta_0$

$$\begin{aligned}
\left\|u - \mathrm{R}(\Phi^{u,p,0})\right\|_{H^1(\hat{I})} &\leq \|u - v\|_{H^1(\hat{I})} + \left\|v - \mathrm{R}(\Phi^{u,p,0})\right\|_{H^1(\hat{I})} \\
&\leq \|u - v\|_{H^1(\hat{I})} + \varepsilon\, |v|_{H^1(\hat{I})} \\
&\leq 2\, \|u - v\|_{H^1(\hat{I})} + \varepsilon\, |u|_{H^1(\hat{I})} \\
&\leq C \left(\frac{M(\eta)}{\sinh(\eta)}\right)^{1/2} p e^{-\eta p/2} + |u|_{H^1(\hat{I})}\, e^{-\eta p/2} \\
&\leq \left(C \left(\frac{M(\eta)}{\sinh(\eta)}\right)^{1/2} + |u|_{H^1(\hat{I})}\right) p e^{-\eta p/2}, \\
L(\Phi^{u,p,0}) &= L\left(\Phi_\varepsilon^{v \circ P, \{I\}, p}\right) + L\left(\Phi^{P^{-1}}\right) \\
&\leq \left(C_L(1 + \log_2(p))\big(2p + \eta p/2\big) + C_L \eta p/2 + C(1 + \log_2(p))^3\right) + 1 \\
&\leq C_L(2 + \eta/2) p (1 + \log_2(p)) + C(\eta) p, \\
M(\Phi^{u,p,0}) &\leq M\left(\Phi_\varepsilon^{v \circ P, \{I\}, p}\right) + M_{\mathrm{fi}}\left(\Phi_\varepsilon^{v \circ P, \{I\}, p}\right) + M_{\mathrm{la}}\left(\Phi^{P^{-1}}\right) + M\left(\Phi^{P^{-1}}\right) \\
&\leq \Big(4 C_M p^2 + 2 C_M (\eta p/2) p + (\eta p/2) C (1 + \log_2(p)) + C(1 + p \log_2(p)) \\
&\quad + 2\left(C_L(1 + \log_2(p))\,(2p + \eta p/2) + C\left(1 + \log_2^3(p)\right)\right)\Big) + 6 + 2 + 2 \\
&\leq C_M(4 + \eta) p^2 + C(\eta) p \log_2(p), \\
M_{\mathrm{fi}}(\Phi^{u,p,0}) &\leq 2 M_{\mathrm{fi}}\left(\Phi^{P^{-1}}\right) \leq 4, \\
M_{\mathrm{la}}(\Phi^{u,p,0}) &\leq M_{\mathrm{la}}\left(\Phi_\varepsilon^{v \circ P, \{I\}, p}\right) \leq 4.
\end{aligned}$$

This finishes the proof. □

Theorem 5.10 shows that for any $0 < \eta < \eta_0$, any $\theta > 0$ and some $c_1(\eta, \theta) > 0$

$$\left\|u - \mathrm{R}(\Phi^{u,p,0})\right\|_{H^1(\hat{I})} \leq C(\eta, \theta, |u|_{H^1(\hat{I})}) \exp\big(-c_1 L(\Phi^{u,p,0})^{1/(1+\theta)}\big)$$

and that for any $0 < \eta < \eta_0$, any $\theta > 0$ and some $c_2(\eta, \theta) > 0$

$$\left\|u - \mathrm{R}(\Phi^{u,p,0})\right\|_{H^1(\hat{I})} \leq C(\eta, \theta, |u|_{H^1(\hat{I})}) \exp\big(-c_2 M(\Phi^{u,p,0})^{1/(2+\theta)}\big).$$

5.4 DNN Emulation of Piecewise Gevrey Functions

We now study expression rates for ReLU NN emulations of *hp*-approximations of functions on $I = (0, 1)$ which are singular at $x = 0$ and which belong to a Gevrey class. We refer to [6] and the references there for such spaces.

For any $\beta \in \mathbb{R}_{>0}$ we define $\psi_\beta : I \to \mathbb{R} : x \mapsto x^\beta$. For any $k, \ell \in \mathbb{N}_0$ we define a seminorm and a norm:

$$\begin{aligned}
|u|_{H^{k,\ell}_\beta(I)} &:= \left\|\psi_{\beta + k - \ell} D^k u\right\|_{L^2(I)}, \\
\|u\|^2_{H^{k,\ell}_\beta(I)} &:= \begin{cases} \sum_{k'=0}^{k} |u|^2_{H^{k',0}_\beta(I)}, & \text{if } \ell = 0, \\ \sum_{k'=\ell}^{k} |u|^2_{H^{k',\ell}_\beta(I)} + \|u\|^2_{H^{\ell-1}(I)}, & \text{if } \ell \in \mathbb{N}. \end{cases}
\end{aligned}$$

All functions for which this norm is finite form the space $H^{k,\ell}_\beta(I)$. In addition, for any $\delta \geq 1$ the Gevrey class $\mathcal{G}^{\ell,\delta}_\beta(I)$ is defined as the class of functions $u \in \bigcap_{k \geq \ell} H^{k,\ell}_\beta(I)$ for which there exist $C_*(u), d(u) > 0$ such that

$$\forall k \geq \ell : |u|_{H^{k,\ell}_\beta(I)} \leq C_* d^{k-\ell} ((k - \ell)!)^\delta. \tag{5.15}$$

For $N \in \mathbb{N}_0$ and $\sigma \in (0,1)$ the mesh $\mathcal{T}_{\sigma,N}$, which is geometrically graded towards $x = 0$, is defined as follows: let $x_0 \coloneqq 0$ and $x_i \coloneqq \sigma^{N-i}$ for $i \in \{1,\dots,N\}$. Let $\mathcal{T}_{\sigma,N}$ be the partition of I into intervals $\{I_{\sigma,i}\}_{i=1}^N$, where $I_{\sigma,i} \coloneqq (x_{i-1}, x_i)$.

The following theorem is a generalization of [33, Theorem 3.36] which, in turn, generalizes earlier results in [7, 31, 15] in the analytic case. The present analysis covers in particular the original results for the piecewise analytic case $\delta = 1$, i.e. functions in $\mathcal{G}_\beta^{\ell,1}(I)$ for $\ell \geq 2$, which are analytic on the interval $(0,1)$ and may have an algebraic singularity at the left endpoint $x = 0$. The proof for general $\delta \geq 1$ is very similar to the proof for $\delta = 1$. For convenience of the reader, it is provided in the appendix.

Theorem 5.11 (Generalization of [33, Theorem 3.36])**.** *Let $\sigma, \beta \in (0,1)$, $\lambda \coloneqq \sigma^{-1} - 1$, $\delta \geq 1$, $u \in \mathcal{G}_\beta^{2,\delta}(I)$ and $N \in \mathbb{N}$ be given. For $\mu_0 \coloneqq \mu_0(\sigma,\delta,d) \coloneqq \max\left\{1, \frac{d\lambda}{2}\left(\frac{e}{\sigma}\right)^{1-\delta}\right\}$ and for any $\mu > \mu_0$ let $\boldsymbol{p} = (p_i)_{i=1}^N \subset \mathbb{N}$ be defined as $p_1 \coloneqq 1$ and $p_i \coloneqq \lfloor \mu i^\delta \rfloor$ for $i \in \{2,\dots,N\}$.*

Then there exists a continuous, piecewise polynomial function $v \in S_{\boldsymbol{p}}(I, \mathcal{T}_{\sigma,N})$ such that $v(x_i) = u(x_i)$ for $i \in \{1,\dots,N\}$ and such that for a constant $C_7(\sigma,\beta,\delta,\mu,C_*,d) > 0$ (where $C_*(u)$ and $d(u)$ are as in Equation (5.15)) it holds that

$$\|u - v\|_{H^1(I)} \leq C_7 \exp\big(-(1-\beta)\log(1/\sigma)N\big) =: C_7 \exp(-cN).$$

As $N \to \infty$, $M = \dim(S_{\boldsymbol{p}}(I, \mathcal{T}_{\sigma,N})) = O(N^{1+\delta})$.

We present the proof of this assertion in Appendix A.

Remark 5.12. *Note that $v(0)$ need not equal $u(0)$. Besides that, it follows from the construction of v in the proof of Theorem 5.11 that $|v|_{H^1(I\setminus I_{1,\sigma})} \leq |u|_{H^1(I\setminus I_{1,\sigma})}$.*

Theorem 5.13. *For all $\delta \geq 1$, all $\beta, \sigma \in (0,1)$, all $\mu > \mu_0(\sigma,\delta,d)$ and all $u \in \mathcal{G}_\beta^{2,\delta}(I)$ there exist NNs $\{\Phi^{u,\sigma,N}\}_{N\in\mathbb{N}}$ such that*

$$\left\|u - \mathrm{R}(\Phi^{u,\sigma,N})\right\|_{H^1(I)} \leq C_8 \exp\big(-(1-\beta)\log(1/\sigma)N\big) = C_8 \exp(-cN),$$

where $C_8 \coloneqq C_8(\sigma,\beta,\delta,\mu,C_*(u),d(u),|u|_{H^1(I)}) > 0$, and such that

$$\begin{aligned}
L(\Phi^{u,\sigma,N}) &\leq C_L\delta\big(2\mu N^\delta \log_2(N) + cN\log_2(N)\big) + C(\sigma,\beta,\delta,\mu)N^\delta,\\
M(\Phi^{u,\sigma,N}) &\leq 2C_M\big(2\mu^2 N^{2\delta+1} + c\mu N^{\delta+2}\big) + C(\sigma,\beta,\delta,\mu)\big(1 + N^{\delta+1}\log_2(N)\big),\\
M_{\mathrm{fi}}(\Phi^{u,\sigma,N}) &\leq 6N,\\
M_{\mathrm{la}}(\Phi^{u,\sigma,N}) &\leq 2N + 2.
\end{aligned}$$

Proof. Let $v \in S_{\boldsymbol{p}}(I, \mathcal{T}_{\sigma,N})$ be as in Theorem 5.11, with $\boldsymbol{p} \subset \mathbb{N}$ defined by $p_1 = 1$ and $p_i = \lfloor \mu i^\delta \rfloor$ for $i \in \{2,\dots,N\}$. Let $\varepsilon \coloneqq \exp(-cN)$. We define $\Phi^{u,\sigma,N} \coloneqq \Phi_\varepsilon^{v,\mathcal{T}_{\sigma,N},\boldsymbol{p}}$, where $\Phi_\varepsilon^{v,\mathcal{T}_{\sigma,N},\boldsymbol{p}}$ is as constructed in Proposition 5.1.

Using that $\left\|v - \mathrm{R}(\Phi^{u,\sigma,N})\right\|_{H^1(I_{1,\sigma})} = 0$ because $p_1 = 1$ and using Remark 5.12 it follows that

$$\begin{aligned}
\left\|u - \mathrm{R}(\Phi^{u,\sigma,N})\right\|_{H^1(I)} &\leq \|u - v\|_{H^1(I)} + \left\|v - \mathrm{R}(\Phi_\varepsilon^{v,\mathcal{T}_{\sigma,M},\boldsymbol{p}})\right\|_{H^1(I)}\\
&\leq C_7\exp(-cN) + \exp(-cN)\,|v|_{H^1(I\setminus I_{\sigma,1})}\\
&\leq \big(C_7 + |u|_{H^1(I)}\big)\exp(-cN),\\
L(\Phi^{u,\sigma,N}) &\leq C_L(1+\log_2(\mu N^\delta))\big(2\mu N^\delta + cN\big) + C_L cN + C(1+\log_2^3(\mu N^\delta))\\
&\leq C_L\delta\big(2\mu N^\delta\log_2(N) + cN\log_2(N)\big) + C(\sigma,\beta,\delta,\mu)N^\delta,\\
M(\Phi^{u,\sigma,N}) &\leq 4C_M\sum_{i=1}^N(\mu i^\delta)^2 + 2C_M cN\sum_{i=1}^N(\mu i^\delta) + cNC\left(1 + \sum_{i=1}^N\log_2(\mu i^\delta)\right)\\
&\quad + C\left(1 + \sum_{i=1}^N \mu i^\delta\log_2(\mu i^\delta)\right)\\
&\quad + 2N\left(C_L(1+\log_2(\mu N^\delta))\big(2\mu N^\delta + cN\big) + C(1+\log_2^3(\mu N^\delta))\right)
\end{aligned}$$

$$\leq 2C_M\big(2\mu^2 N^{2\delta+1} + c\mu N^{\delta+2}\big) + C(\sigma,\beta,\delta,\mu)\big(1+N^{\delta+1}\log_2(N)\big),$$
$$M_{\mathrm{fi}}(\Phi^{u,\sigma,N}) \leq 6N,$$
$$M_{\mathrm{la}}(\Phi^{u,\sigma,N}) \leq 2N+2.$$

This finishes the proof. □

Theorem 5.13 shows that for any $\theta > 0$ and for $c_3(\beta,\sigma,\theta,\delta,\mu), C_9(\sigma,\beta,\delta,\mu,d,|u|_{H^1(I)},\theta) > 0$

$$\Big\|u - \mathrm{R}(\Phi^{u,\sigma,N})\Big\|_{H^1(I)} \leq C_9 \exp\big(-c_3 L(\Phi^{u,\sigma,N})^{1/(\delta+\theta)}\big),$$

and that for $c_4(\beta,\sigma,\delta,\mu), C_{10}(\sigma,\beta,\delta,\mu,d,|u|_{H^1(I)}) > 0$

$$\Big\|u - \mathrm{R}(\Phi^{u,\sigma,N})\Big\|_{H^1(I)} \leq C_{10} \exp\big(-c_4 M(\Phi^{u,\sigma,N})^{1/(2\delta+1)}\big).$$

Remark 5.14. *In Theorem 5.11, we proved exponential expression rate bounds for deep ReLU NNs in the Sobolev spaces $H^1(I)$ for classes of Gevrey δ-regular functions in $I=(0,1)$ which exhibit one algebraic singularity at the endpoint $x=0$ of I. It is straightforward to generalize this result to functions with a finite number of algebraic singularities at singular support sets $\mathcal{S}=\{x_1,...,x_J\}\subset\bar{I}$. Multivariate versions of Theorem 5.11 also hold [23].*

6 Conclusion

We established expression rate bound estimates for the expression by deep neural networks of univariate functions which belong to several types of function spaces. In particular, Sobolev and Besov spaces, and spaces of piecewise analytic and Gevrey-regular functions. We proved that ReLU DNNs can achieve in each of these function classes approximation rate bounds which are either identical to or closely match the best available approximation rates from classical approximation by piecewise polynomial spline functions. Notably, DNNs match the rates achieved by both, free-knot ("h-adaptive") and order-adaptive ("hp"-adaptive) approximations. They offer a partial explanation for the recent success of numerical solution strategies in using DNNs for the numerical approximation of PDEs.

The present results were established in the univariate case for ease of presentation and to keep mathematical technicalities at bay. We hasten to add, however, that corresponding results are valid also in several space dimensions. As the mathematical apparatus characterizing the analytic function classes is somewhat more involved (see, e.g., [32] and the references there), we will present these in [23].

Acknowledgements

P.P is supported by a DFG Research Fellowship "Shearlet-based energy functionals for anisotropic phase-field methods".

A Proof of Theorem 5.11

We note that [33, Lemma 3.41], which is formulated for any $\beta\in(0,1)$ and any $u\in\mathcal{G}^{2,1}_\beta(I) =: \mathcal{B}^2_\beta(I)$, also holds for any $\delta\geq 1$, any $\beta\in(0,1)$ and any $u\in\mathcal{G}^{2,\delta}_\beta(I)$.

Lemma A.1 ([33, Lemma 3.41])**.** *Let $I=(0,1)$, $\delta\geq 1$, $\beta\in(0,1)$ and $u\in\mathcal{G}^{2,\delta}_\beta(I)$. Let $\sigma\in(0,1)$, $\lambda := \sigma^{-1}-1$ and let $\boldsymbol{p}=(p_i)_{i=1}^N\subset\mathbb{N}$ be such that $p_1=1$ and such that $p_i\geq 2$ for $i\in\{2,\dots,N\}$.*

Then there exists a $v\in S_{\boldsymbol{p}}(I,\mathcal{T}_{\sigma,N})$ such that

$$\|u-v\|^2_{H^1(I)} \leq C\left[x_1^{2(1-\beta)}\,|u|^2_{H^{2,2}_\beta(I)} + \sum_{i=2}^N x_{i-1}^{2(1-\beta)}\,\frac{(p_i-s_i)!}{(p_i+s_i)!}\,(\tfrac{\lambda}{2})^{2s_i}\,|u|^2_{H^{s_i+1,2}_\beta(I)}\right],$$

where $s_i\in\{2,\dots,p_i\}$ for all $i\in\{2,\dots,N\}$.

Lemma A.2 ([8, Lemma 4.3.4]). *Let $N \in \mathbb{N}$, $\alpha > 0$ and $\mu_0 \coloneqq \max\{1, \alpha e^{1-\delta}\}$. For any $\mu > \mu_0$ let $\boldsymbol{p} = (p_i)_{i=1}^N \subset \mathbb{N}$ be defined by $p_i \coloneqq \lfloor \mu i^\delta \rfloor$ for all $i \in \{1, \dots, N\}$.*

Then it holds that

$$\sum_{i=1}^{N} \alpha^{2i} \frac{(p_i - i)!}{(p_i + i + 1)!} ((i+1)!)^{2\delta} \leq C(\alpha, \mu, \delta).$$

In particular, $C(\alpha, \mu, \delta)$ is independent of N.

Proof of Theorem 5.11. We use Lemma A.1 with $x_i = \sigma^{N-i}$ and $s_i = i + 1$ for all $i \in \{1, \dots, N\}$. Because $u \in \mathcal{G}_\beta^{2,\delta}(I)$, it holds that $|u|_{H_\beta^{i+2,2}(I)} \leq C d^i (i!)^\delta$ for all $i \in \{0, \dots, N-2\}$. With $\alpha \coloneqq \frac{d\lambda}{2\sigma^{1-\beta}}$, $\mu_0 = \max\left\{1, \frac{d\lambda}{2}\left(\frac{e}{\sigma}\right)^{1-\delta}\right\}$, and C_* as in Equation (5.15), it follows that

$$\begin{aligned}
\|u - v\|_{H^1(I)}^2 &\leq C \left[\sigma^{2(1-\beta)(N-1)} C_*^2 + \sum_{i=2}^{N} \sigma^{2(1-\beta)(N+1-i)} \frac{(p_i - i - 1)!}{(p_i + i + 1)!} (\tfrac{\lambda}{2})^{2i+2} C_*^2 d^{2i} (i!)^{2\delta} \right] \\
&\leq C C_*^2 \sigma^{2(1-\beta)N} \left[\sigma^{-2(1-\beta)} + (\sigma^{1-\beta} \tfrac{\lambda}{2})^2 \sum_{i=2}^{N} \left(\frac{d\lambda}{2\sigma^{1-\beta}} \right)^{2i} \frac{(p_i - i - 1)!}{(p_i + i + 1)!} (i!)^{2\delta} \right] \\
&\leq C C_*^2 \sigma^{2(1-\beta)N} \left[\sigma^{-2(1-\beta)} + (\sigma^{1-\beta} \tfrac{\lambda}{2})^2 C(\alpha, \mu, \delta) \right] \\
&\leq C_7^2 \sigma^{2(1-\beta)N},
\end{aligned}$$

where $C_7(\sigma, \beta, \delta, \mu, C_*, d) > 0$. □

References

[1] A. R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics (Spetses, 1990)*, volume 335 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pages 561–576. Kluwer Acad. Publ., Dordrecht, 1991.

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

[3] C. Beck, S. Becker, P. Grohs, N. Jaafari, and A. Jentzen. Solving stochastic differential equations and Kolmogorov equations by means of deep learning. Technical Report 1806.00421, arXiv, 2018.

[4] C. Beck, W. E, and A. Jentzen. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. Technical Report 2017-49, Seminar for Applied Mathematics, ETH Zürich, 2017.

[5] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM J. Math. Data Sci.*, 2018.

[6] A. Chernov, T. von Petersdorff, and C. Schwab. Exponential convergence of *hp* quadrature for integral operators with Gevrey kernels. *ESAIM Math. Mod. Num. Anal.*, 45:387–422, 2011.

[7] W. Dahmen and K. Scherer. Best approximation by piecewise polynomials with variable knots and degrees. *J. Approx. Theory*, 26(1):1–13, 1979.

[8] D. Devaud. *hp-approximation of linear parabolic evolution problems in $H^{1/2}$*. PhD thesis, ETH Zürich, 2017.

[9] R. A. DeVore and G. G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1993.

[10] W. E and Q. Wang. Exponential convergence of the deep neural network approximation for analytic functions. *Sci. China Math.*, 61(10):1733–1740, 2018.

[11] W. E and B. Yu. The Deep Ritz method: A deep learning-based numerical algorithm for solving variational problems. *Commun. Math. Stat.*, 6(1):1–12, 2018.

[12] D. Elbrächter, P. Grohs, A. Jentzen, and C. Schwab. DNN Expression Rate Analysis of High-dimensional PDEs: Application to Option Pricing. Technical Report 1809.07669, arXiv, 2018.

[13] P. Grohs, D. Perekrestenko, D. Elbrächter, and H. Bölcskei. Deep neural network approximation theory. Technical Report 1901.02220, arXiv, 2019.

[14] P. Grohs, T. Wiatowski, and H. H. Boelcskei. Deep convolutional neural networks on cartoon functions. Technical Report 2016-25, Seminar for Applied Mathematics, ETH Zürich, 2016.

[15] W. Gui and I. Babuška. The h, p and h-p versions of the finite element method in 1 dimension. II. The error analysis of the h- and h-p versions. *Numer. Math.*, 49(6):613–657, 1986.

[16] J. Han, A. Jentzen, and W. E. Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci. USA*, 115(34):8505–8510, 2018.

[17] J. He, L. Li, J. Xu, and C. Zheng. ReLU deep neural networks and linear finite elements. *arXiv:1807.03973*, 2018.

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

[19] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[20] S. Liang and R. Srikant. Why deep neural networks for function approximation? In *Proc. of ICLR 2017*, pages 1 – 17, 2017.

[21] H. N. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Adv. Comput. Math.*, 1(1):61–80, 1993.

[22] H. N. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput.*, 8:164–177, 1996.

[23] J. A. A. Opschoor, P. Petersen, and C. Schwab. Deep ReLU networks and multivariate high-order finite element methods, 2019. in preparation.

[24] P. Oswald. On the degree of nonlinear spline approximation in Besov-Sobolev spaces. *J. Approx. Theory*, 61(2):131–157, 1990.

[25] G. Pang, L. Lu, and G. E. Karniadakis. fPINNs: Fractional Physics-Informed Neural Networks. Technical Report 1811.08967, arXiv, 2018.

[26] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Netw.*, 108:296 – 330, 2018.

[27] P. P. Petrushev. Direct and converse theorems for spline and rational approximation and Besov spaces. In *Function spaces and applications*, pages 363–377. Springer, 1988.

[28] A. Pinkus. Approximation theory of the MLP model in neural networks. In *Acta Numerica*, volume 8, pages 143–195. Cambridge Univ. Press, Cambridge, 1999.

[29] H. Robbins. A Remark on Stirling's Formula. *Amer. Math. Monthly*, 62(1):26–29, 1955.

[30] D. Rolnik and M. Tegmark. The power of deeper networks for expressing natural functions. Technical Report 1705.05502v1, arXiv, 2017.

[31] K. Scherer. On optimal global error bounds obtained by scaled local error estimates. *Numer. Math.*, 36(2):151–176, 1980/81.

[32] D. Schötzau and C. Schwab. Exponential convergence of hp-FEM for elliptic problems in polyhedra: mixed boundary conditions and anisotropic polynomial degrees. *Found. Comput. Math.*, 18(3):595–660, 2018.

[33] C. Schwab. *p- and hp-finite element methods*. Numerical Mathematics and Scientific Computation. The Clarendon Press, Oxford University Press, New York, 1998.

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

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

[36] J. Sirignano and K. Spiliopoulos. DGM: A deep learning algorithm for solving partial differential equations. *J. Comput. Phys.*, 2018.

[37] E. Tadmor. The exponential accuracy of Fourier and Chebyshev differencing methods. *SIAM J. Numer. Anal.*, 23(1):1–10, 1986.

[38] M. Telgarsky. Neural networks and rational functions. *Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.*, 2017.

[39] H. Triebel. *Interpolation theory, function spaces, differential operators*. Johann Ambrosius Barth, Heidelberg, second edition, 1995.

[40] H. Triebel. *Theory of function spaces*. Modern Birkhäuser Classics. Birkhäuser/Springer Basel AG, Basel, 2010.

[41] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94:103 – 114, 2017.