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
Latest revision: December 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*

December 16, 2019

Abstract

Approximation rate bounds for emulations 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. 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 univariate, piecewise Gevrey functions with point singularities are established. Combined with recent results on ReLU DNN approximation of rational, oscillatory, and high-dimensional functions, this corroborates that continuous, piecewise affine ReLU DNNs afford algebraic and exponential convergence rate bounds which are comparable to "best in class" schemes for several important function classes of high and infinite smoothness. Using composition of DNNs, we also prove that radial-like functions obtained as compositions of the above with the euclidean norm and, possibly, anisotropic affine changes of co-ordinates can be emulated at exponential rate in terms of the DNN size and depth without the curse of dimensionality.

Keywords: Deep neural networks, finite element methods, exponential covergence, Gevrey regularity, singularities

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 [25]. 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, 16, 17, 22, 38, 50]). The present paper aims at contributing to a mathematical understanding of these observations. Specifically, we investigate DNN approximation rates of concrete architectures of DNNs for a number of widely used approximation spaces in numerical analysis. We present DNN architectures with ReLU activation which emulate a wide range of fixed- and free-knot spline approximations, spectral- and *hp*-approximations. Moreover, we will show that the so-constructed DNNs yield approximation properties (algebraic, (sub)exponential) comparable to the best available approximations with the same numbers of degrees of freedom. As (realizations of) ReLU DNNs are continuous, piecewise affine functions, the presently proved results not only shed light on the (exponential) expressive power of DNNs as compared to (possibly nonlinear) approximation rate bounds, but conversely indicate *exponential expressivity of iterated systems of classical (Courant-type) linear spline spaces* for piecewise analytic functions.

Early mathematical work on approximation by neural networks (NNs for short) focused on *universality results* (e.g. [1, 2, 24, 41] 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.

[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

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 [25]. 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 approximation rates for a wide range of function spaces.

Among these are approximation rate bounds for analytic functions (e.g. [15, 32, 33]), differentiable functions (e.g. [26, 55]), piecewise differentiable functions (e.g. [39]) and high-dimensional approximation (e.g. [17, 48, 34]), oscillatory functions (e.g. [19]), cartoon functions in image segmentation (e.g. [20, 19]), manifold approximation (e.g. [9, 49, 11]), rational function approximations (e.g. [51]), continuous, piecewise affine finite elements [23], radial functions (e.g. [29, 10]), spatially sparse functions (e.g. [27]) and multivariate functions which have a sparse polynomial expansion (e.g. [43, 48]). The standard approach employed in all the proofs of results above is to first demonstrate that DNNs are capable of efficiently emulating other existing (linear or nonlinear) approximation architectures such as B-splines (e.g. [32, 8]), wavelets (e.g. [5, 49]) and high degree polynomials (e.g. [26, 55, 43]). By this argument, approximation rate bounds for these classical architectures are then transferred to DNN approximation.

Main Results

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 is more appropriate to measure the accuracy of approximation with respect to Sobolev norms. Indeed, if the approximations of the solution of a PDE have a small L^p-error but remain very rough, then any attempt to learn these approximations to the solution of a PDE based on minimizing a derivative-based energy functional is futile because the approximants produce an excessive energy.

In the present work, we study DNN approximation rate bounds of functions $f \in W^{s,p}([-1, 1])$, $p \in [1, \infty]$, $s > 1$ with respect to stronger Besov and Sobolev norms. Specifically, we establish ReLU DNN approximation rate bounds in the strongest norms which are admissible by the ReLU activation. As it is an important special case of wide interest, in Section 3 we address the ReLU DNN approximation rates achieveable by emulation of continuous, piecewise linear functions.

For example, in Corollary 3.3 we establish ReLU approximation rate bounds in Besov norms $B^{s'}_{q',t'}(I)$ in the unit interval $I = (0, 1)$ with weak differentiation order $s' \in (1, 2)$ (the precise range depending on the summability and fine indices q, q', t, t' in the assumptions of these results). We emphasize that these bounds hold for ReLU activations despite the nondifferentiability of the ReLU activation.

To extend these results to higher polynomial degrees, in Section 4, we address the ReLU DNN emulation of polynomials with approximation rate bounds and NN size estimates which are explicit in the polynomial degree. As we work with Taylor representations of polynomials, the sizes of weights in our (constructive) NN emulations of polynomials tend to grow quickly with the polynomial degree, and the presently obtained bounds may not be best suited for quantization of NNs.

In Section 5, we address ReLU DNN approximation rates for the emulation of so-called *hp*-finite element spaces (corresponding to so-called "free-knot, variable degree" spline approximation). Here, we obtain ReLU approximation rate bounds for analytic or Gevrey-regular functions in $I = (0, 1)$ with possibly a point singularity in I (or a finite number thereof). In Section 5.5, we finally address rates of ReLU DNN emulation of so-called (exponential) "boundary layer functions" of the type $x \mapsto \exp(-x/\varepsilon)$ for $x \in I$ with the length scale $\varepsilon > 0$ determined by physical parameters of the phenomenon of interest. We show, based on the corresponding approximation result for finite elements in [47] in Theorem 5.14, that ReLU DNNs afford *exponential convergence rates* which are uniform with respect to the boundary layer scale parameter ε.

The present exposition is focussed on *univariate results*. Due to the closedness of ReLU NNs under composition, their scope is considerably wider. To illustrate this, we demonstrate in Section 6 how our univariate results imply straightforwardly ReLU NN approximation rates for multivariate, possibly radial functions. Not only do the results in Section 6 constitute novel high-dimensional approximation rate bounds, the proofs also outline a general recipe to relate all presently obtained approximation rate bounds in the univariate setting to anisotropic radial basis function systems with corresponding d-dimensional NN approximation rate results, with moderate (polylogarithmic in accuracy and quadratic in d, i.e. without the curse of dimensionality) NN size.

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. In terms of the NN size ("number of degrees of freedom" in finite element terminology) and from an approximation theoretical point of view, ReLU DNNs perform as well as the best available finite element approximation for a number of function classes which arise as solutions of elliptic (partial) differential equations. 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.

Outline

The structure 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. In Section 6, we extend these results to the multivariate setting of isotropic and anisotropic radial basis functions.

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. $C(\eta, \theta)$.

For $d \in \mathbb{N}$ and for $x, y \in \mathbb{R}^d$, $\langle x, y\rangle \in \mathbb{R}$ denotes the standard Euclidean inner product on $\mathbb{R}^d$. The Euclidean norm on $\mathbb{R}^d$ is denoted by $\|x\|_{2,\mathbb{R}^d}$.

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 $I := (0,1)$ with nodes $0 = x_0 < x_1 < \ldots < x_{N-1} < x_N = 1$, elements $I_i := (x_{i-1}, x_i)$ and element sizes $h_i := x_i - x_{i-1}$ for $i \in \{1, \ldots, N\}$. Let $h := \max_{i\in\{1,\ldots,N\}} h_i$. 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 $p_{\max} := \max_{i=1}^N p_i$ and the corresponding approximation space

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

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

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

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*

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

where $N_0 := d$ and $N_1, \ldots, N_L \in \mathbb{N}$, 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*

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

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

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

Here ϱ is understood to act component-wise on vector-valued inputs, i.e., for $y = (y^1, \dots, y^m) \in \mathbb{R}^m$, $\varrho(y) := (\varrho(y^1), \dots, \varrho(y^m))$. We call $N(\Phi) := d + \sum_{j=1}^{L} N_j$ the number of neurons of the NN Φ, $L(\Phi) := L$ *the* number of layers *or* depth, $M_j(\Phi) := \|A_j\|_{\ell^0} + \|b_j\|_{\ell^0}$ *the* number of weights in the j-th layer, *and* $M(\Phi) := \sum_{j=1}^{L} M_j(\Phi)$ *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

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

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 [39]. 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 [39] which also serve as definitions of the associated procedures. We provide bounds on the number of weights in the first layer and in the last layer. They can be derived from the definitions in [39]. We start with the *concatenation of NNs*.

Proposition 2.2 (NN concatenation [39, Remark 2.6]). *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,* $\mathrm{R}(\Phi^1 \odot \Phi^2) = \mathrm{R}(\Phi^1) \circ \mathrm{R}(\Phi^2)$,

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

and

$$M\left(\Phi^1 \odot \Phi^2\right) \leq M\left(\Phi^1\right) + M_{\mathrm{fi}}\left(\Phi^1\right) + M_{\mathrm{la}}\left(\Phi^2\right) + M\left(\Phi^2\right) \leq 2M\left(\Phi^1\right) + 2M\left(\Phi^2\right). \tag{2.1}$$

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

Proposition 2.3 (NN parallelization [39, Definition 2.7]). *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*

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

$M(\mathrm{P}(\Phi^1, \Phi^2)) = M(\Phi^1) + M(\Phi^2)$, $M_{\mathrm{fi}}(\mathrm{P}(\Phi^1, \Phi^2)) = M_{\mathrm{fi}}(\Phi^1) + M_{\mathrm{fi}}(\Phi^2)$ *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 [39, Remark 2.4]). *For every $d, L \in \mathbb{N}$ there exists a NN $\Phi^{\mathrm{Id}}_{d,L}$ with $L(\Phi^{\mathrm{Id}}_{d,L}) = L$, $M(\Phi^{\mathrm{Id}}_{d,L}) \le 2dL$, $M_{\mathrm{fi}}(\Phi^{\mathrm{Id}}_{d,L}) \le 2d$ and $M_{\mathrm{la}}(\Phi^{\mathrm{Id}}_{d,L}) \le 2d$ 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 [17, Setting 5.2]). *Let $L \in \mathbb{N}$ and let*

$$\Phi^1 = \left(\left(A_1^1, b_1^1\right), \dots, \left(A_L^1, b_L^1\right)\right), \quad \Phi^2 = \left(\left(A_1^2, b_1^2\right), \dots, \left(A_L^2, b_L^2\right)\right)$$

be two NNs with L layers each and with input dimensions $N_0^1 = d_1$ and $N_0^2 = d_2$, 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 $\mathrm{FP}\left(\Phi^1, \Phi^2\right) := \left(\left(A_1^3, b_1^3\right), \dots, \left(A_L^3, b_L^3\right)\right)$ where, for $j = 1, \dots, L$, we define

$$A_j^3 := \begin{pmatrix} A_j^1 & 0 \\ 0 & A_j^2 \end{pmatrix} \text{ and } b_j^3 := \begin{pmatrix} b_j^1 \\ b_j^2 \end{pmatrix}.$$

□

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. For previous work on ReLU network reapproximation of continuous, piecewise linear functions we refer to e.g. [55, Section 3.3], [56].

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 $v \in S_{\mathbf{1}}(I, \mathcal{T})$ there exists a NN Φ^v such that*

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

Proof. We set $\Phi^v := ((A_1^v, b_1^v), (A_2^v, b_2^v))$ such that

$$A_1^v := [1, \dots, 1]^T \in \mathbb{R}^{N \times 1}, \quad b_1^v := [-x_0, -x_1, \dots, -x_{N-1}]^T \in \mathbb{R}^N, \quad b_2^v := v(x_0) \in \mathbb{R},$$

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

$$A_2^v \in \mathbb{R}^{1 \times N}, \qquad (A_2^v)_{1,i} := \begin{cases} \frac{v(x_i) - v(x_{i-1})}{x_i - x_{i-1}} - \frac{v(x_{i-1}) - v(x_{i-2})}{x_{i-1} - x_{i-2}} & \text{if } i > 1 \\ \frac{v(x_i) - v(x_{i-1})}{x_i - x_{i-1}} & \text{if } i = 1. \end{cases}$$

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. For definitions and basic properties of Besov spaces we refer to [52, 53].

Proposition 3.2 ([37, Theorem 3])**.** *Let $s < \max\{2, 1 + 1/q\}$, let $0 < q < q' \le \infty$ and $0 < s' < \min\{1+1/q', s-1/q+1/q'\}$, and let $0 < t, t' \le \infty$. Then there exists a constant $C \coloneqq C(q, q', t, t', s, s') > 0$, such that 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*

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

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 $0 < q < q' \le \infty$ and $0 < s' < \min\{1 + 1/q', s - 1/q + 1/q'\}$, and let $0 < t, t' \le \infty$. 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*

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

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 the number N of the partition size in Proposition 3.2.

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

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 4.6.

The results below are based on a construction of DNNs emulating the multiplication function with two-dimensional input which has been derived in [55]. We recall here a version of this result and provide an estimate of the error with respect to the $W^{1,\infty}$ norm, from [48], as required in approximation rate bounds for PDEs.

Proposition 4.1 ([48, Proposition 3.1] [55, Proposition 3])**.** *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*

$$\sup_{|a|,|b|\le\kappa} \left| ab - \mathrm{R}\left(\widetilde{\times}_{\delta,\kappa}\right)(a, b)\right| \le \delta \text{ and}$$
$$\operatorname*{esssup}_{|a|,|b|\le\kappa} \max\left\{ \left| a - \frac{\mathrm{d}}{\mathrm{d}b}\mathrm{R}\left(\widetilde{\times}_{\delta,\kappa}\right)(a, b)\right|, \left| b - \frac{\mathrm{d}}{\mathrm{d}a}\mathrm{R}\left(\widetilde{\times}_{\delta,\kappa}\right)(a, b)\right| \right\} \le \delta,$$

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

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

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

$$\mathrm{R}\left(\widetilde{\times}_{\delta,\kappa}\right)(a, 0) = \mathrm{R}\left(\widetilde{\times}_{\delta,\kappa}\right)(0, b) = 0. \tag{4.1}$$

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 $\{\Phi^v_\beta\}_{\beta\in(0,1)}$ with input dimension one and output dimension one which satisfy*

$$
\begin{aligned}
\left\|v - \mathrm{R}(\Phi^v_\beta)\right\|_{W^{1,\infty}(\hat{I})} &\le \beta, \\
\mathrm{R}(\Phi^v_\beta)(0) &= v(0), \\
L(\Phi^v_\beta) &\le C_L(1+\log_2(n))\log_2(C_0/\beta) + \tfrac{1}{3}C_L(\log_2(n))^3 + C(1+\log_2(n))^2, \\
M(\Phi^v_\beta) &\le 4C_M n\log_2(C_0/\beta) + 8C_M n\log_2(n) + 4C_L(1+\log_2(n))^2\log_2(C_0/\beta) + C(1+n), \\
M_{\mathrm{fi}}(\Phi^v_\beta) &\le 4\log_2(n) + 4, \\
M_{\mathrm{la}}(\Phi^v_\beta) &\le 4n+2
\end{aligned}
$$

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

Remark 4.3. *As will become apparent in the proof, for given $n \in \mathbb{N}_0$ only the weights $\{\tilde{v}_\ell\}_{\ell=0}^n$ in the output layer of Φ^v_β depend on v (which are the Legendre coefficients of v, depending linearly on v). Due to the linearity of the output layer of NNs (cf. Definition 2.1), the approximation depends linearly on v. In particular, the network weights depend continuously on v with respect to the $L^2(\hat{I})$-norm, hence also with respect to stronger norms such as the $L^\infty(\hat{I})$-norm.*

Remark 4.4. *An alternative approach for the expression of polynomials by ReLU NNs, is to use networks of finite width, as proposed for example in [55, 15]. Both the networks constructed in these references and the network from Proposition 4.2 have network size bounds growing only logarithmically in the accuracy, and are of the order $O(n\log n)$ in terms of the polynomial degree n. For finite width networks, the network size is proportional to the depth. Here, by allowing varying widths, we obtain smaller bounds on the network depth.*

To prove the proposition, we use the following technical lemma. For $k \in \mathbb{N}$, this lemma produces a tree-structured network with $2^{k-1}+1$ outputs, which correspond to high-order monomials of degrees $2^{k-1}, ..., 2^k$. This network is constructed by repeatedly applying the product network introduced in Proposition 4.1.

Lemma 4.5. *For every $k \in \mathbb{N}$ there exist NNs $\{\Psi^k_\delta\}_{\delta\in(0,1)}$ with input dimension one and output dimension $2^{k-1}+1$ such that with $\widetilde{X}^\ell_\delta := \mathrm{R}(\Psi^k_\delta)_{1+\ell-2^{k-1}}$ for $\ell \in \{2^{k-1},\dots,2^k\}$ it holds that*

$$
\begin{aligned}
\mathrm{R}(\Psi^k_\delta)(x) &= \big(\widetilde{X}^{2^{k-1}}_\delta(x),\dots,\widetilde{X}^{2^k}_\delta(x)\big), \qquad x \in \hat{I}, \\
\left\|x^\ell - \widetilde{X}^\ell_\delta(x)\right\|_{W^{1,\infty}(\hat{I})} &\le \delta, \quad \ell \in \{2^{k-1},\dots,2^k\}, && (4.2) \\
\widetilde{X}^\ell_\delta(0) &= 0, \quad \ell \in \{2^{k-1},\dots,2^k\}, && (4.3) \\
L(\Psi^k_\delta) &\le C_L\big(\tfrac{1}{3}k^3 + 2k^2 + k\log_2(1/\delta)\big) + (4C_L + C'_L + 1)k, && (4.4) \\
C_1 &:= 7C_M + C'_M + C_{\mathrm{fi}} + \tfrac{1}{2}C_{\mathrm{la}} + 8, \\
C_2 &:= 2C'_L + 8C_L + 8, \\
M(\Psi^k_\delta) &\le 2C_M k 2^k + C_M 2^k\log_2(1/\delta) + 2kC_L\log_2(1/\delta) \\
&\quad + C_1 2^k + \tfrac{2}{3}C_L k^3 + 3C_L k^2 + C_2 k, && (4.5) \\
M_{\mathrm{fi}}(\Psi^k_\delta) &\le C_{\mathrm{fi}} + 2, && (4.6) \\
M_{\mathrm{la}}(\Psi^k_\delta) &\le C_{\mathrm{la}} 2^{k-1} + 2. && (4.7)
\end{aligned}
$$

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

$$
\Psi^1_\delta := \mathrm{P}\left(\Phi^{\mathrm{Id}}_{1,L_1}, ((A_1 A, b_1),\dots,(A_{L_1},b_{L_1}))\right).
$$

For all $x \in \hat{I}$ it holds that $\widetilde{X}_\delta^1(x) := [\mathrm{R}(\Psi_\delta^1)(x)]_1 = x$ and $\widetilde{X}_\delta^2(x) := [\mathrm{R}(\Psi_\delta^1)(x)]_2 = \mathrm{R}(\widetilde{\times}_{\delta/2,1})(x,x)$, which with (4.1) shows that Equation (4.3) holds for $k = 1$.

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

$$\begin{aligned}
L(\Psi_\delta^1) &= L_1 \le C_L \log_2(2/\delta) + C_L', \\
M(\Psi_\delta^1) &= M(\Phi^{\mathrm{Id}}_{1,L_1}) + M(((A_1 A, b_1), \dots, (A_{L_1}, b_{L_1}))) \\
&\le 2L_1 + \big(C_M \log_2(2/\delta) + C_M'\big) \\
&\le \big(2C_L + C_M\big)\log_2(2/\delta) + 2C_L' + C_M', \\
M_{\mathrm{fi}}(\Psi_\delta^1) &= M_{\mathrm{fi}}(\Phi^{\mathrm{Id}}_{1,L_1}) + M_{\mathrm{fi}}(((A_1 A, b_1), \dots, (A_{L_1}, b_{L_1}))) \\
&\le C_{\mathrm{fi}} + 2, \\
M_{\mathrm{la}}(\Psi_\delta^1) &= M_{\mathrm{la}}(\Phi^{\mathrm{Id}}_{1,L_1}) + M_{\mathrm{la}}(((A_1 A, b_1), \dots, (A_{L_1}, b_{L_1}))) \\
&\le C_{\mathrm{la}} + 2.
\end{aligned}$$

Finally, it follows from Proposition 4.1 that

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

where the last inequality follows from Poincaré's inequality and Equation (4.3). This shows that Equation (4.2) holds for $k = 1$. This finishes the proof of the induction hypothesis.

Induction hypothesis (IH). For some $\delta \in (0,1)$ and some $k \in \mathbb{N}$ define $\theta := 2^{-k-3}\delta$ and assume that there exists a network Ψ_θ^k for which Equations (4.2)–(4.7) hold with θ instead of δ.

Induction step. We show that Equations (4.2)–(4.7) hold with δ as in (IH) and with $k+1$ instead of k.

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

$$\Big\|\widetilde{X}_\theta^\ell\Big\|_{L^\infty(\hat{I})} \le \Big\|x^\ell\Big\|_{L^\infty(\hat{I})} + \Big\|x^\ell - \widetilde{X}_\theta^\ell(x)\Big\|_{W^{1,\infty}(\hat{I})} \le 1 + \theta < 2. \tag{4.8}$$

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

$$\Psi_\delta^{k+1} := \Phi_\delta^{2,k} \odot \Phi^{1,k} \odot \Psi_\theta^k. \tag{4.9}$$

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

$$\begin{aligned}
\mathbb{R}^{2^{k-1}+1} \to \mathbb{R}^{2^{k+1}+1} : \big(z_1, \dots, z_{2^{k-1}+1}\big) \mapsto \big(&z_{2^{k-1}+1}, z_1, z_2, z_2, z_2, z_2, z_3, z_3, z_3, z_3, z_4, z_4, z_4, \\
&\dots, z_{2^{k-1}}, z_{2^{k-1}+1}, z_{2^{k-1}+1}, z_{2^{k-1}+1}\big).
\end{aligned}$$

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

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

Moreover,

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

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

$$\Phi_\delta^{2,k} := \mathrm{FP}\left(\Phi^{\mathrm{Id}}_{1,L_\theta}, \widetilde{\times}_{\theta,2}, \dots, \widetilde{\times}_{\theta,2}\right),$$

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

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

The realization of Ψ^{k+1}_δ, defined in Equation (4.9), is given by

$$[\mathrm{R}(\Psi^{k+1}_\delta)(x)]_1 = \widetilde{X}^{2^k}_\theta(x), \qquad \text{for } x \in \hat{I}, \tag{4.10}$$

$$[\mathrm{R}(\Psi^{k+1}_\delta)(x)]_{\ell+1-2^k} = \mathrm{R}(\widetilde{\times}_{\theta,2})\left(\widetilde{X}^{\lceil \ell/2 \rceil}_\theta(x), \widetilde{X}^{\lfloor \ell/2 \rfloor}_\theta(x)\right), \qquad \text{for } x \in \hat{I},\ \ell \in \{2^k + 1, \ldots, 2^{k+1}\}. \tag{4.11}$$

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

$$\widetilde{X}^\ell_\delta(x) \coloneqq [\mathrm{R}(\Psi^{k+1}_\delta)(x)]_{\ell+1-2^k}.$$

Equation (4.3) for $k+1$ follows from the induction hypothesis and Equations (4.1) and (4.11).

We bound the depth and the size of Ψ^{k+1}_δ.

$$\begin{aligned}
L(\Psi^{k+1}_\delta) &= L(\Phi^{2,k}_\delta) + L(\Phi^{1,k}) + L(\Psi^k_\theta) \\
&\leq \big(C_L\big(k + 4 + \log_2(1/\delta)\big) + C'_L\big) + 1 \\
&\quad + \Big(C_L\big(\tfrac{1}{3}k^3 + 2k^2 + k\log_2(2^{k+3}/\delta)\big) + (4C_L + C'_L + 1)k\Big) \\
&\leq C_L\big(\tfrac{1}{3}(k+1)^3 + 2(k+1)^2 + (k+1)\log_2(1/\delta)\big) + (4C_L + C'_L + 1)(k+1), \\
M(\Psi^{k+1}_\delta) &\leq M(\Phi^{2,k}_\delta) + M_{\mathrm{fi}}(\Phi^{2,k}_\delta) + M_{\mathrm{la}}(\Phi^{1,k} \odot \Psi^k_\theta) + M(\Phi^{1,k} \odot \Psi^k_\theta) \\
&\leq M(\Phi^{2,k}_\delta) + M_{\mathrm{fi}}(\Phi^{2,k}_\delta) + 2M_{\mathrm{la}}(\Phi^{1,k}) + M(\Phi^{1,k}) + M_{\mathrm{fi}}(\Phi^{1,k}) + M_{\mathrm{la}}(\Psi^k_\theta) + M(\Psi^k_\theta) \\
&\leq \Big((2C_L + C_M 2^k)(k + 4 + \log_2(1/\delta)) + 2C'_L + C'_M 2^k\Big) + (C_{\mathrm{fi}} 2^k + 2) \\
&\quad + 2(1 + 2^{k+1}) + (1 + 2^{k+1}) + (1 + 2^{k+1}) + (C_{\mathrm{la}} 2^{k-1} + 2) \\
&\quad + \Big(2C_M k 2^k + C_M 2^k \log_2(2^{k+3}/\delta) + 2kC_L \log_2(2^{k+3}/\delta) + C_1 2^k + \tfrac{2}{3}k^3 C_L + 3k^2 C_L + C_2 k\Big) \\
&\leq 2C_M(k+1)2^{k+1} + C_M 2^{k+1}\log_2(1/\delta) + 2(k+1)C_L \log_2(1/\delta) \\
&\quad + C_1 2^{k+1} + \tfrac{2}{3}C_L(k+1)^3 + 3(k+1)^2 C_L + C_2(k+1), \\
C_1 &\coloneqq 7C_M + C'_M + C_{\mathrm{fi}} + \tfrac{1}{2}C_{\mathrm{la}} + 8, \\
C_2 &\coloneqq 2C'_L + 8C_L + 8, \\
M_{\mathrm{fi}}(\Psi^{k+1}_\delta) &= M_{\mathrm{fi}}(\Psi^k_\theta) \leq C_{\mathrm{fi}} + 2, \\
M_{\mathrm{la}}(\Psi^{k+1}_\delta) &= M_{\mathrm{la}}(\Phi^{2,k}_\delta) = C_{\mathrm{la}} 2^k + 2.
\end{aligned}$$

This finishes the proof of Equations (4.4)–(4.7) for $k+1$. We now estimate the NN expression error. Because $\theta < \delta$, it follows from the induction hypothesis and Equation (4.10) that Equation (4.2) holds for $\ell = 2^{(k+1)-1}$. For $\ell \in \{2^k + 1, \ldots, 2^{k+1}\}$, with $\ell_0 := \lceil \ell/2 \rceil$, we use that, analogous to Equation (4.8), it holds that for $m \in \{\ell_0, \ell-\ell_0\}$

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

We find

$$
\begin{aligned}
\left|x^\ell - \widetilde{X}_\delta^\ell(x)\right|_{W^{1,\infty}(\hat{I})} \le{}& \left\|\ell_0 x^{\ell-1} - [\mathrm{DR}(\widetilde{\times}_{\theta,2})]_1\big(\widetilde{X}_\theta^{\ell_0}(x), \widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell_0}(x)\right\|_{L^\infty(\hat{I})}\\
&+ \left\|(\ell-\ell_0)x^{\ell-1} - [\mathrm{DR}(\widetilde{\times}_{\theta,2})]_2\big(\widetilde{X}_\theta^{\ell_0}(x),\ \widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell-\ell_0}(x)\right\|_{L^\infty(\hat{I})}\\
\le{}& \left\|\ell_0 x^{\ell_0-1}\big(x^{\ell-\ell_0} - \widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\right\|_{L^\infty(\hat{I})}\\
&+ \left\|\widetilde{X}_\theta^{\ell-\ell_0}(x)\big(\ell_0 x^{\ell_0-1} - \tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell_0}(x)\big)\right\|_{L^\infty(\hat{I})}\\
&+ \left\|\big(\widetilde{X}_\theta^{\ell-\ell_0}(x) - [\mathrm{DR}(\widetilde{\times}_{\theta,2})]_1\big(\widetilde{X}_\theta^{\ell_0}(x), \widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\big)\tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell_0}(x)\right\|_{L^\infty(\hat{I})}\\
&+ \left\|(\ell-\ell_0)x^{\ell-\ell_0-1}\big(x^{\ell_0} - \widetilde{X}_\theta^{\ell_0}(x)\big)\right\|_{L^\infty(\hat{I})}\\
&+ \left\|\widetilde{X}_\theta^{\ell_0}(x)\big((\ell-\ell_0)x^{\ell-\ell_0-1} - \tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\right\|_{L^\infty(\hat{I})}\\
&+ \left\|\big(\widetilde{X}_\theta^{\ell_0}(x) - [\mathrm{DR}(\widetilde{\times}_{\theta,2})]_2\big(\widetilde{X}_\theta^{\ell_0}(x), \widetilde{X}_\theta^{\ell-\ell_0}(x)\big)\big)\tfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{X}_\theta^{\ell-\ell_0}(x)\right\|_{L^\infty(\hat{I})}\\
\overset{(4.8),(\mathrm{IH})}{\le}{}& \ell_0\theta + 2\theta + (\ell_0+1)\theta + (\ell-\ell_0)\theta + 2\theta + (\ell-\ell_0+1)\theta\\
\le{}& (2\ell+6)\theta \le \delta,
\end{aligned}
$$

where $[\mathrm{DR}(\widetilde{\times}_{\delta,2})]$ is the Jacobian and where we have used that $3 \le \ell \le 2^{k+1}$, which implies that $2\ell+6 \le 4\ell \le 2^{k+3}$. Because $\widetilde{X}_\delta^\ell(0) = 0 = 0^\ell$ it follows with Poincaré's inequality that $\left\|x^\ell - \widetilde{X}_\delta^\ell(x)\right\|_{W^{1,\infty}(\hat{I})} \le \delta$. In summary, for k satisfying the induction hypothesis and for arbitrary $\delta \in (0,1)$, we have constructed Ψ_δ^{k+1} and have shown that Equations (4.2)–(4.7) 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 $k=1$. ☐

Note that the number of consecutive concatenations in Ψ_δ^k depends on k. Therefore, we have to use the sharper bound in Equation (2.1) involving $M_{\mathrm{fi}}(\cdot)$ and $M_{\mathrm{la}}(\cdot)$. Using the second inequality in (2.1) instead would introduce factors of 2, resulting in an extra k-dependent factor in the bound on the network size.

Proof of Proposition 4.2. Below, we consider the case $C_0 > \beta$. The proof of the case $C_0 \le \beta$ is analogous. The distinction is needed to ensure that we do not invoke Lemma 4.5 with $\delta \ge 1$.

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 $M(\Phi_\beta^v) = M_{\mathrm{fi}}(\Phi_\beta^v) = M_{\mathrm{la}}(\Phi_\beta^v) \le 2$.

In case $n \ge 2$, we define $k := \lceil \log_2(n) \rceil$ and $\delta := \beta/C_0$ and use Lemma 4.5. Let $\{\ell_j\}_{j=1}^k \subset \mathbb{R}$ be such that $L\left(\Psi_\delta^k\right) + 1 = L\left(\Psi_\delta^j\right) + \ell_j$ for $j = 1,\dots,k$, hence it holds that $\ell_j \le L\left(\Psi_\delta^k\right)$. We define

$$
\Phi_\beta^v := \Phi^{3,n} \odot \mathrm{P}\left(\Psi_\delta^1 \odot \Phi_{1,\ell_1}^{\mathrm{Id}}, \dots, \Psi_\delta^k \odot \Phi_{k,\ell_k}^{\mathrm{Id}}\right),
$$

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

$$
\mathbb{R}^{2^k+k-1} \to \mathbb{R} : (z_1,\dots,z_{2^k+k-1}) \mapsto \tilde{v}_0 + \tilde{v}_1 z_1 + \tilde{v}_2 z_2 + \sum_{j=2}^{k} \sum_{\ell=2^{j-1}+1}^{2^j} \tilde{v}_\ell z_{\ell+j-1}.
$$

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

The realization of Φ^v_β is

$$\mathrm{R}(\Phi^v_\beta)(x) = \tilde{v}_0 + \sum_{\ell=1}^{2^k} \tilde{v}_\ell \widetilde{X}^\ell_\delta(x), \qquad x \in \hat{I}.$$

From Equation (4.3) we conclude that $\mathrm{R}(\Phi^v_\beta)(0) = \tilde{v}_0 = v(0)$.

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

$$\begin{aligned}
L\left(\Phi^v_\beta\right) &= L\left(\Phi^{3,n}\right) + L\left(\Psi^k_\delta\right) \\
&\leq 1 + \left(C_L\left(\tfrac{1}{3}k^3 + k\log_2\left(\tfrac{C_0}{\beta}\right)\right) + Ck^2\right) \\
&\leq C_L(1+\log_2(n))\log_2\left(\tfrac{C_0}{\beta}\right) + \tfrac{1}{3}C_L\log_2^3(n) + C\log_2^2(n), \\
M\left(\Phi^v_\beta\right) &\leq M\left(\Phi^{3,n}\right) + M_{\mathrm{fi}}\left(\Phi^{3,n}\right) + \sum_{j=1}^k M_{\mathrm{la}}\left(\Psi^j_\delta \odot \Phi^{\mathrm{Id}}_{1,\ell_j}\right) + \sum_{j=1}^k M\left(\Psi^j_\delta \odot \Phi^{\mathrm{Id}}_{1,\ell_j}\right) \\
&\leq M\left(\Phi^{3,n}\right) + M_{\mathrm{fi}}\left(\Phi^{3,n}\right) + \sum_{j=1}^k M_{\mathrm{la}}\left(\Psi^j_\delta\right) + \sum_{j=1}^k M\left(\Psi^j_\delta\right) + \sum_{j=1}^k M_{\mathrm{fi}}\left(\Psi^j_\delta\right) + \sum_{j=1}^k M_{\mathrm{la}}\left(\Phi^{\mathrm{Id}}_{1,\ell_j}\right) \\
&\quad + \sum_{j=1}^k M\left(\Phi^{\mathrm{Id}}_{1,\ell_j}\right) \\
&\leq \left(2^k+1\right) + \left(2^k+1\right) + \sum_{j=1}^k \left(2^{k-1}C_{\mathrm{la}} + 2\right) \\
&\quad + \sum_{j=1}^k \left(2C_M j 2^j + C_M 2^j \log_2\left(\frac{C_0}{\beta}\right) + 2jC_L\log_2\left(\frac{C_0}{\beta}\right) + C2^j\right) + \sum_{j=1}^k (C_{\mathrm{fi}} + 2) \\
&\quad + 2k + 2k\left(C_L\left(\tfrac{1}{3}k^3 + k\log_2\left(\tfrac{C_0}{\beta}\right)\right) + Ck^2\right) \\
&\leq 4C_M n\log_2\left(\frac{C_0}{\beta}\right) + 8C_M n\log_2(n) + 4C_L(1+\log_2(n))^2\log_2\left(\frac{C_0}{\beta}\right) + Cn, \\
M_{\mathrm{fi}}\left(\Phi^v_\beta\right) &= \sum_{j=1}^k M_{\mathrm{fi}}\left(\Psi^j_\delta \odot \Phi^{\mathrm{Id}}_{1,\ell_j}\right) \leq \sum_{j=1}^k 2M_{\mathrm{fi}}\left(\Phi^{\mathrm{Id}}_{1,\ell_j}\right) = 4k \leq 4\log_2(n) + 4, \\
M_{\mathrm{la}}\left(\Phi^v_\beta\right) &= 2M_{\mathrm{la}}\left(\Phi^{3,n}\right) \leq 4n + 2.
\end{aligned}$$

Finally, we estimate the error.

$$\left\|v - \mathrm{R}\left(\Phi^v_\beta\right)\right\|_{W^{1,\infty}(\hat{I})} \leq \sum_{\ell=1}^{2^k} |\tilde{v}_\ell| \left\|x^\ell - \widetilde{X}^\ell_\delta(x)\right\|_{W^{1,\infty}(\hat{I})} \leq \sum_{\ell=2}^{2^k} |\tilde{v}_\ell|\delta \leq \beta.$$

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. The error estimates in the following proposition are with respect to more general $W^{1,q'}(\hat{I})$ norms.

Proposition 4.6 (ReLU approximation rate bounds of polynomials of degree q in $\hat{I}$)**.** *For all $q \in \mathbb{N}_{\geq 2}$ and all $w \in (\mathbb{P}_q \cap H^1_0)(\hat{I})$ there exist NNs $(\Phi^{w,0}_\varepsilon)_{\varepsilon\in(0,1)}$ 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* $1 \le q' \le \infty$

$$\begin{aligned}
\left\|w - \mathrm{R}(\Phi_\varepsilon^{w,0})\right\|_{W^{1,q'}(\hat{I})} &\le 2^{\frac{2}{q'}-1}\varepsilon\, |w|_{H^1(\hat{I})}\,,\\
L\left(\Phi_\varepsilon^{w,0}\right) &\le C_L(1+\log_2(q))\,(2q+\log_2(1/\varepsilon)) + C_L\log_2(1/\varepsilon) + C\log_2^3(q),\\
M\left(\Phi_\varepsilon^{w,0}\right) &\le 4C_M\left(2q^2 + q\log_2(1/\varepsilon)\right) + \left(6C_L\left(1+\log_2(q)\right)^2 + 2C_M\right)\log_2(1/\varepsilon)\\
&\quad + Cq\log_2^2(q),\\
M_{\mathrm{fi}}\left(\Phi_\varepsilon^{w,0}\right) &\le 4\log_2(q) + 8,\\
M_{\mathrm{la}}(\Phi_\varepsilon^{w,0}) &= C_{\mathrm{la}}.
\end{aligned}$$

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

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 $\psi(x) := (1+x)(1-x) = 1-x^2$ 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 $\mathrm{R}(\Phi_\eta^\psi)|_{\mathbb{R}\setminus\hat{I}} = 0$ and we approximate $Q := w/\psi \in \mathbb{P}_{q-2}(\hat{I})$ 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 $\ell \in \mathbb{N}_0$ and $m := (j-\ell)/2$, (see e.g. [18, Section 10.10, Equation (16)])

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

The sum of these coefficients can be estimated using the following inequalities (cf. [42]):

$$\forall n \in \mathbb{N}: \qquad \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}e^{\frac{1}{12n}}. \tag{4.12}$$

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

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

It follows, from (4.12) that, for all $j \in \mathbb{N}$,

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

and as a result that

$$\begin{aligned}
\sum_{\ell=0}^j |c_\ell^j| &= \sum_{m\in\{0,\dots,\lfloor j/2\rfloor\}} |c_{j-2m}^j|\\
&\le 2^{-j}\left(\sum_{m=0}^j\binom{j}{m}\right)\max_{m=0}^{\lfloor j/2\rfloor}\binom{2j-2m}{j}\sqrt{j+\frac{1}{2}}\\
&\le \sqrt{j+\frac{1}{2}}\binom{2j}{j} \le \frac{4^j\sqrt{j+\frac{1}{2}}}{\sqrt{\pi j}} \le 4^j.
\end{aligned} \tag{4.13}$$

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

$$\begin{aligned}\sum_{\ell=0}^{n} |\tilde{v}_\ell| = \sum_{\ell=0}^{n} \left| \sum_{j=0}^{n} v_j c_\ell^j \right| &\le \left(\max_{j=0}^{n} |v_j| \right) \sum_{\ell=0}^{n} \sum_{j=0}^{n} |c_\ell^j| = \left(\max_{j=0}^{n} |v_j| \right) \sum_{j=0}^{n} \left(\sum_{\ell=0}^{n} |c_\ell^j| \right) \\ &\overset{(*)}{\le} \|v\|_{L^2(\hat{I})} \sum_{j=0}^{n} 4^j \le \tfrac{1}{3} 4^{n+1} \|v\|_{L^2(\hat{I})} . \end{aligned} \tag{4.14}$$

At (*) we used Equation (4.13) and

$$\max_{j=0}^{n} |v_j| \le \left\| (v_j)_{j=0}^{n} \right\|_{\ell^2} = \|v\|_{L^2(\hat{I})} . \tag{4.15}$$

We now consider $w \in (\mathbb{P}_q \cap H_0^1)(\hat{I})$ of degree $q \ge 2$ 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 $\left\| \frac{g(x)}{x} \right\|_{L^2((0,1))} \le 2 \|g'\|_{L^2((0,1))}$. It follows that

$$\begin{aligned}\|Q\|_{L^2(\hat{I})}^2 &= \left\| \frac{w(x)}{1-x^2} \right\|_{L^2(\hat{I})}^2 = \left\| \frac{w(x)}{1-x^2} \right\|_{L^2((-1,0))}^2 + \left\| \frac{w(x)}{1-x^2} \right\|_{L^2((0,1))}^2 \\ &\le \left\| \frac{w(x)}{1+x} \right\|_{L^2((-1,0))}^2 + \left\| \frac{w(x)}{1-x} \right\|_{L^2((0,1))}^2 = \left\| \frac{w(y-1)}{y} \right\|_{L^2((0,1))}^2 + \left\| \frac{w(1-z)}{z} \right\|_{L^2((0,1))}^2 \\ &\le 2^2 \left\| w'(y-1) \right\|_{L^2((0,1))}^2 + 2^2 \left\| w'(1-z) \right\|_{L^2((0,1))}^2 = 2^2 \left\| w' \right\|_{L^2((-1,0))}^2 + 2^2 \left\| w' \right\|_{L^2((0,1))}^2 \\ &= 2^2 |w|_{H^1(\hat{I})}^2 . \end{aligned} \tag{4.16}$$

Writing $Q(x) = \sum_{\ell=0}^{q-2} \widetilde{Q}_\ell x^\ell$ for $x \in \hat{I}$, it follows from Equation (4.14) with $v = Q$, $\tilde{v}_\ell = \widetilde{Q}_\ell$ and $n = q-2$ that

$$\sum_{\ell=0}^{q-2} |\widetilde{Q}_\ell| \le \tfrac{1}{6} 4^q |w|_{H^1(\hat{I})} . \tag{4.17}$$

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 (4.15) for $v = Q$ and $v_j = Q_j$ and from Equation (4.16) that for all $j \in \{0, \dots, q-2\}$

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

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

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

By Markov's inequality (e.g. [14, Chapter 4, Theorem 1.4]), we get

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

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

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

We write $w = \psi Q$ and approximate ψ by a NN whose realization is supported in $\hat{I}$. To approximate Q we use Φ^Q_β from Proposition 4.2, with $C_0 \leq \frac{1}{6} 4^q$ according to Equation (4.17). We will use that

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

With the 2×2 identity matrix Id_2, the vector $c := (0,0)^\top$, the 2×1-matrix $A := [1,-1]^\top$ and the vector $b := (1,1)^\top$ we define $\Phi^\psi_\eta := \widetilde{\times}_{\frac{\eta}{2},1} \odot ((\mathrm{Id}_2, c), (A, b))$, which has realization $\mathrm{R}(\Phi^\psi_\eta)(x) = \mathrm{R}\left(\widetilde{\times}_{\frac{\eta}{2},1}\right)\big(\varrho(1+x), \varrho(1-x)\big)$ for $x \in \mathbb{R}$. By Equation (4.1), it follows that $\mathrm{R}(\Phi^\psi_\eta)|_{\mathbb{R}\setminus\hat{I}} = 0$. It holds that $L(\Phi^\psi_\eta) = C_L \log_2(2/\eta) + C'_L + 2$,

$$\begin{aligned} M(\Phi^\psi_\eta) &\leq M\left(\widetilde{\times}_{\frac{\eta}{2},1}\right) + M_{\mathrm{fi}}\left(\widetilde{\times}_{\frac{\eta}{2},1}\right) + M_{\mathrm{la}}(((\mathrm{Id}_2, c), (A,b))) + M(((\mathrm{Id}_2, c), (A,b))) \\ &\leq \left(C_M \log_2\left(\frac{2}{\eta}\right) + C'_M\right) + C_{\mathrm{fi}} + 2 + 6, \end{aligned}$$

$M_{\mathrm{fi}}(\Phi^\psi_\eta) \leq M_{\mathrm{fi}}(((\mathrm{Id}_2, c), (A,b))) = 4$, and $M_{\mathrm{la}}(\Phi^\psi_\eta) = C_{\mathrm{la}}$. The error can be estimated as follows:

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

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

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

We define

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

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

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

By Equation (4.1) and the fact that $\mathrm{R}(\Phi^\psi_\eta)|_{\mathbb{R}\setminus\hat{I}} = 0$, it follows that $\mathrm{R}(\Phi^{w,0}_\varepsilon)|_{\mathbb{R}\setminus\hat{I}} = 0$.

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

Assuming that $L(\Phi^Q_\beta) > L(\Phi^\psi_\eta)$, it follows that

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

Moreover,

$$\begin{aligned}
M\left(\Phi^{w,0}_\varepsilon\right) &\leq M\left(\widetilde{\times}_{\eta,K}\right) + M_{\mathrm{fi}}\left(\widetilde{\times}_{\eta,K}\right) + M_{\mathrm{la}}\left(\Phi^Q_\beta\right) + M_{\mathrm{la}}\left(\Phi^{\mathrm{Id}}_{1,L_*}\odot\Phi^\psi_\eta\right) + M\left(\Phi^Q_\beta\right) + M\left(\Phi^{\mathrm{Id}}_{1,L_*}\odot\Phi^\psi_\eta\right) \\
&\leq M\left(\widetilde{\times}_{\eta,K}\right) + M_{\mathrm{fi}}\left(\widetilde{\times}_{\eta,K}\right) + M_{\mathrm{la}}\left(\Phi^Q_\beta\right) + 2M_{\mathrm{la}}\left(\Phi^{\mathrm{Id}}_{1,L_*}\right) \\
&\quad + M\left(\Phi^Q_\beta\right) + M\left(\Phi^{\mathrm{Id}}_{1,L_*}\right) + M_{\mathrm{fi}}\left(\Phi^{\mathrm{Id}}_{1,L_*}\right) + M_{\mathrm{la}}\left(\Phi^\psi_\eta\right) + M\left(\Phi^\psi_\eta\right) \\
&\leq \left(C_M\log_2\left(\tfrac{K}{\eta}\right) + C'_M\right) + C_{\mathrm{fi}} + (4q-6) + 4 \\
&\quad + \left(4C_M q\left(2q + \log_2\left(\tfrac{1}{\beta}\right)\right) + 8C_M q\log_2(q) + 4C_L(1+\log_2(q))^2\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) + 4C_M\left(2q^2 + q\log_2\left(\tfrac{36}{\varepsilon}\right)\right) + 6C_L(1+\log_2(q))^2\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_2^2(q) \\
&\leq 4C_M\left(2q^2 + q\log_2\left(\tfrac{1}{\varepsilon}\right)\right) + \left(6C_L(1+\log_2(q))^2 + 2C_M\right)\log_2\left(\tfrac{1}{\varepsilon}\right) + Cq\log_2^2(q), \\
M_{\mathrm{fi}}\left(\Phi^{w,0}_\varepsilon\right) &= M_{\mathrm{fi}}\left(\Phi^Q_\beta\right) + M_{\mathrm{fi}}\left(\Phi^\psi_\eta\right) \leq (4\log_2(q)+4) + 4 = 4\log_2(q) + 8, \\
M_{\mathrm{la}}\left(\Phi^{w,0}_\varepsilon\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^{w,0}_\varepsilon\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^Q_\beta)(x), \mathrm{R}(\Phi^\psi_\eta)(x)\right)\right)\right\|_{L^\infty(\hat{I})} \\
&\leq 2\left\|\left(\psi(x) - \mathrm{R}\left(\Phi^\psi_\eta\right)(x)\right)\tfrac{\mathrm{d}}{\mathrm{d}x}Q(x)\right\|_{L^\infty(\hat{I})} \\
&\quad + 2\left\|\mathrm{R}\left(\Phi^\psi_\eta\right)(x)\tfrac{\mathrm{d}}{\mathrm{d}x}\left(Q(x) - \mathrm{R}\left(\Phi^Q_\beta\right)(x)\right)\right\|_{L^\infty(\hat{I})} \\
&\quad + 2\left\|\left(\mathrm{R}\left(\Phi^\psi_\eta\right)(x) - \left[D\mathrm{R}\left(\widetilde{\times}_{\eta,K}\right)\right]_1\left(\mathrm{R}\left(\Phi^Q_\beta\right)(x), \mathrm{R}\left(\Phi^\psi_\eta\right)(x)\right)\right)\tfrac{\mathrm{d}}{\mathrm{d}x}\mathrm{R}\left(\Phi^Q_\beta\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^\psi_\eta\right)(x)\right)\right\|_{L^\infty(\hat{I})} \\
&\quad + 2\left\|\left(Q(x) - \mathrm{R}\left(\Phi^Q_\beta\right)(x)\right)\tfrac{\mathrm{d}}{\mathrm{d}x}\mathrm{R}\left(\Phi^\psi_\eta\right)(x)\right\|_{L^\infty(\hat{I})} \\
&\quad + 2\left\|\left(\mathrm{R}\left(\Phi^Q_\beta\right)(x) - \left[D\mathrm{R}\left(\widetilde{\times}_{\eta,K}\right)\right]_2\left(\mathrm{R}\left(\Phi^Q_\beta\right)(x), \mathrm{R}\left(\Phi^\psi_\eta\right)(x)\right)\right)\tfrac{\mathrm{d}}{\mathrm{d}x}\mathrm{R}\left(\Phi^\psi_\eta\right)(x)\right\|_{L^\infty(\hat{I})} \\
&\leq 2\eta\left|Q\right|_{W^{1,\infty}(\hat{I})} + 2\left\|\mathrm{R}\left(\Phi^\psi_\eta\right)\right\|_{L^\infty(\hat{I})}\beta + 2\eta\left|\mathrm{R}\left(\Phi^Q_\beta\right)\right|_{W^{1,\infty}(\hat{I})} \\
&\quad + 2\left\|Q\right\|_{L^\infty(\hat{I})}\eta + 2\beta\left|\mathrm{R}\left(\Phi^\psi_\eta\right)\right|_{W^{1,\infty}(\hat{I})} + 2\eta\left|\mathrm{R}\left(\Phi^\psi_\eta\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 $\Phi^{w,0}_\varepsilon(\pm 1) = 0 = w(\pm 1)$ that $\varepsilon/2$

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

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

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 4.7. *We note that by Hölder's inequality for all* $2 \le q' \le \infty$

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

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

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

5 ReLU emulation of *hp*-Finite Element Spaces

Based on the ReLU NN approximation rate bounds of univariate polynomials obtained in 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 ([44] and the references there). We discuss in detail several classes of functions whose relevance derives them appearing as solution components in a wide range of elliptic and parabolic PDEs. In particular, we study NN approximation of smooth functions with point singularities which appear in solutions of elliptic boundary value problems on polygonal and polyhedral domains. Moreover, we address NN approximation of (exponential) boundary layers which are ubiquitous in solutions of singular perturbation problems in fluid and solid mechanics.

5.1 Approximation of Piecewise Polynomials

We start by demonstrating how to emulate continuous, 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* $I = (0,1)$ *into* N *open, disjoint, connected subintervals and for all* $v \in S_{\boldsymbol{p}}(I,\mathcal{T})$*, for* $0 < \varepsilon < 1$ *exist NNs* $\{\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\}_{\varepsilon\in(0,1)}$ *such that for all* $1 \le q' \le \infty$ *holds*

$$\begin{aligned}\left\|v - \mathrm{R}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right)\right\|_{W^{1,q'}(I)} &\le \varepsilon\,|v|_{W^{1,q'}(I)}, \\ L\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right) &\le C_L(1+\log_2(p_{\max}))\left(2p_{\max}+\log_2(1/\varepsilon)\right) + C_L\log_2(1/\varepsilon) + C\left(1+\log_2^3(p_{\max})\right), \\ M\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right) &\le 8C_M\sum_{i=1}^N p_i^2 + 4C_M\log_2(1/\varepsilon)\sum_{i=1}^N p_i + \log_2(1/\varepsilon)\,C\left(1+\sum_{i=1}^N \log_2^2(p_i)\right) \\ &\quad + C\left(1+\sum_{i=1}^N p_i\log_2^2(p_i)\right) \\ &\quad + 2N\left(C_L(1+\log_2(p_{\max}))\left(2p_{\max}+\log_2(1/\varepsilon)\right) + C\left(1+\log_2^3(p_{\max})\right)\right), \\ M_{\mathrm{fi}}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right) &\le 6N, \\ M_{\mathrm{la}}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right) &\le 2N+2.\end{aligned}$$

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 affine 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 $j \in \{0, \ldots, N\}$. The network $\Phi^{\bar{v}}$, constructed in Lemma 3.1, satisfies

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

For all $i \in \{1, \ldots, N\}$, 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 $w_i \coloneqq (v - \bar{v})|_{I_i} \in (\mathbb{P}_{p_i} \cap H_0^1)(I_i)$, it holds that $\hat{w}_i \coloneqq w_i \circ P_i^{-1} \in (\mathbb{P}_{p_i} \cap H_0^1)(\hat{I})$, hence Proposition 4.6 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

$$\begin{aligned}
L\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) &\leq C_L(1 + \log_2(p_i))\left(2p_i + \log_2\left(\tfrac{1}{\varepsilon_i}\right)\right) + C_L \log_2\left(\tfrac{1}{\varepsilon_i}\right) + C\left(1 + \log_2^3(p_i)\right), \\
&\leq C_L(1 + \log_2(p_i))\left(2p_i + \log_2\left(\tfrac{1}{\varepsilon}\right)\right) + C_L \log_2\left(\tfrac{1}{\varepsilon}\right) + C\left(1 + \log_2^3(p_i)\right), \\
M\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) &\leq 4C_M\left(2p_i^2 + p_i \log_2\left(\tfrac{1}{\varepsilon_i}\right)\right) + \left(6C_L(1 + \log_2(p_i))^2 + 2C_M\right)\log_2\left(\tfrac{1}{\varepsilon_i}\right) \\
&\quad + C\left(1 + p_i \log_2^2(p_i)\right), \\
&\leq 4C_M\left(2p_i^2 + p_i \log_2\left(\tfrac{1}{\varepsilon}\right)\right) + \left(6C_L(1 + \log_2(p_i))^2 + 2C_M\right)\log_2\left(\tfrac{1}{\varepsilon}\right) \\
&\quad + C\left(1 + p_i \log_2^2(p_i)\right), \\
M_{\mathrm{fi}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) &\leq 4\log_2(p_i) + 8, \\
M_{\mathrm{la}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) &= C_{\mathrm{la}}.
\end{aligned}$$

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 4.6 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 4.7, 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

$$\begin{aligned}
L\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right) &\leq 1 + C_L(1+\log_2(p_i))\left(2p_i + \log_2\left(\tfrac{1}{\varepsilon}\right)\right) + C_L \log_2\left(\tfrac{1}{\varepsilon}\right) + C\left(1+\log_2^3(p_i)\right), \\
M\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right) &\leq M\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) + M_{\mathrm{fi}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) + M_{\mathrm{la}}\left(\Phi^{P_i}\right) + M\left(\Phi^{P_i}\right) \\
&\leq \Big(4C_M\left(2p_i^2 + p_i \log_2\left(\tfrac{1}{\varepsilon}\right)\right) + \left(6C_L(1+\log_2(p_i))^2 + 2C_M\right)\log_2\left(\tfrac{1}{\varepsilon}\right) \\
&\quad + C\left(1 + p_i \log_2^2(p_i)\right)\Big) + (4\log_2(p_i) + 8) + 2 + 2 \\
&\leq 4C_M\left(2p_i^2 + p_i \log_2\left(\tfrac{1}{\varepsilon}\right)\right) + \left(6C_L(1+\log_2(p_i))^2 + 2C_M\right)\log_2\left(\tfrac{1}{\varepsilon}\right) \\
&\quad + C\left(1 + p_i \log_2^2(p_i)\right), \\
M_{\mathrm{fi}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right) &\leq 2M_{\mathrm{fi}}\left(\Phi^{P_i}\right) = 4, \\
M_{\mathrm{la}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0} \odot \Phi^{P_i}\right) &= M_{\mathrm{la}}\left(\Phi_{\varepsilon_i}^{\hat{w}_i,0}\right) = C_{\mathrm{la}}.
\end{aligned}$$

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 $\Phi_{N+1}^{\mathrm{Sum}} := (([1,\dots,1],0))$, where $[1,\dots,1]$ is a $1\times(N+1)$-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

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

Because the realization 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:

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

$$
\begin{aligned}
&+\sum_{i=1}^{N}\Big(4C_M\left(2p_i^2+p_i\log_2\left(\tfrac{1}{\varepsilon}\right)\right)+\left(6C_L(1+\log_2(p_i))^2+2C_M\right)\log_2\left(\tfrac{1}{\varepsilon}\right)\\
&+C\left(1+p_i\log_2^2(p_i)\right)\Big)+N(4\log_2(p_i)+8)+2N+2N\\
&\le 8C_M\sum_{i=1}^{N}p_i^2+4C_M\log_2\left(\tfrac{1}{\varepsilon}\right)\sum_{i=1}^{N}p_i+\log_2\left(\tfrac{1}{\varepsilon}\right)C\left(1+\sum_{i=1}^{N}\log_2^2(p_i)\right)+C\left(1+\sum_{i=1}^{N}p_i\log_2^2(p_i)\right)\\
&+2(N+1)\left(C_L(1+\log_2(p_{\max}))\left(2p_{\max}+\log_2\left(\tfrac{1}{\varepsilon}\right)\right)+C\left(1+\log_2^3(p_{\max})\right)\right),
\end{aligned}
$$

$$
M_{\mathrm{fi}}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right)\le M_{\mathrm{fi}}\left(\Phi^{\bar v}\right)+\sum_{i=1}^{N}2M_{\mathrm{fi}}\left(\Phi^{P_i}\right)\le 2N+4N=6N,
$$

$$
M_{\mathrm{la}}\left(\Phi_\varepsilon^{v,\mathcal{T},\boldsymbol{p}}\right)\le 2M_{\mathrm{la}}\left(\Phi_{N+1}^{\mathrm{Sum}}\right)=2N+2.
$$

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

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

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 $i\in\{1,\ldots,N\}$ and all $j\in\{0,\ldots,N\}$, 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 $j\in\{0,\ldots,N\}$. This finishes the proof. □

5.2 Free-knot Spline Approximation

The following classical result due to Petruchev [40] and Oswald [37] describes the rates of best approximation of Besov-regular functions on $I:=(0,1)$ by free-knot splines of fixed degree. We refer to [52, 53] for definitions and basic properties of function spaces. This setting and the corresponding approximation rate bounds correspond to the so-called "h-adaptive FEM".

Theorem 5.2 ([37, 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 $B_{q,t}^s(I)$, there exists $h^N\in S_p^N(I)$ such that

$$
\left\|f-h^N\right\|_{B_{q',t'}^{s'}(I)}\le C_3N^{-(s-s')}\|f\|_{B_{q,t}^s(I)}. \tag{5.1}
$$

Moreover,

$$
\left\|h^N\right\|_{B_{q,t}^s(I)}\le C_3\|f\|_{B_{q,t}^s(I)}. \tag{5.2}
$$

Equation (5.1) follows from [37, Theorem 3]. We recall the assumptions in [37, Theorem 3], in our notation: $0<q<q'\le\infty$, $0<t,t'\le\infty$, $0<s\le(p+1)+\max\{0,1/q-1\}$ (with equality only if $t=\infty$) and $0<s'<\min\{p+1/q',s-1/q+1/q'\}$. For ease of the reader we compared our notation with that in [37, Theorem 3] in Table 1 below.

[37, Theorem 3]	p	p'	q	q'	$k-1$	λ'	δ	N	n
Theorem 5.2	q	q'	t	t'	p	$p+1/q'$	$\max\{0,1/q-1\}$	1	N

Table 1: Correspondence in notation between [37, Theorem 3] and Theorem 5.2. In [37], λ' and δ are defined in [37, Proposition 1] and [37, Section 1], respectively.

Under the assumptions in [37, Theorem 3], Equation (5.2) follows from [37, Theorem 6], where λ corresponds with $p+1/q$ ([37, Proposition 1]) and under the additional assumption that $s < p+1/q$.

Remark 5.3 ([37]). *The approximant h^N of Theorem 5.2 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 $0 < q < q' \le \infty$, $q' \ge 1$, $0 < t \le \infty$. Let $p \in \mathbb{N}$, $0 < s' \le 1 < s < p+1/q$, $1-1/q' < s-1/q$ and $s' < 1$ if $p=1$ and $q'=\infty$. Then, there exists a constant $C_4 := C(q,q',t,s,s',p) > 0$ and, for every $N \in \mathbb{N}$ and every $f \in B^s_{q,t}(I)$, there exists a NN Φ^N_f such that*

$$\left\| f - \mathrm{R}\left(\Phi^N_f\right)\right\|_{W^{s',q'}(I)} \le C_4 N^{-(s-s')} \|f\|_{B^s_{q,t}(I)} \tag{5.3}$$

and

$$L\left(\Phi^N_f\right) \le C_L(1+\log_2(p))\left(2p+(s-s')\log_2(N)\right) + C_L(s-s')\log_2(N) + C\left(1+\log_2^3(p)\right), \tag{5.4}$$

$$\begin{aligned} M\left(\Phi^N_f\right) \le\, & 8C_M N p^2 + 4C_M(s-s')N\log_2(N)p + C(s-s')\log_2(N)\left(1+N\log_2^2(p)\right) \\ & + C\left(1+Np\log_2^2(p)\right) \\ & + 2N\left(C_L(1+\log_2(p))\left(2p+(s-s')\log_2(N)\right) + C\left(1+\log_2^3(p)\right)\right), \end{aligned} \tag{5.5}$$

$$M_{\mathrm{fi}}\left(\Phi^N_f\right) \le 6N, \tag{5.6}$$

$$M_{\mathrm{la}}\left(\Phi^N_f\right) \le 2N+2. \tag{5.7}$$

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 $t' := \min\{q', 2\}$. Hence there exists $C(q,q',t,s,s',p) > 0$ and $h^N \in S^N_p(I)$ such that

$$\left\| f - h^N \right\|_{B^{s'}_{q',\min\{q',2\}}(I)} \le C(q,q',t,s,s',p) N^{-(s-s')} \|f\|_{B^s_{q,t}(I)} \tag{5.8}$$

and

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

By [37, Equation 6] or [52, Equation (1.3.3/3)], $B^{s'}_{q',\min\{q',2\}}(I)$ is continuously embedded in $W^{s',q'}(I)$. Hence

$$\|u\|_{W^{s',q'}(I)} \le C(s',q')\|u\|_{B^{s'}_{q',\min\{q',2\}}(I)} \quad \text{for all } u \in B^{s'}_{q',\min\{q',2\}}(I). \tag{5.10}$$

Applying Equation (5.10) to Equation (5.8) yields that

$$\left\| f - h^N \right\|_{W^{s',q'}(I)} \le C(q,q',t,s,s',p) N^{-(s-s')} \|f\|_{B^s_{q,t}(I)}. \tag{5.11}$$

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^N_f such that

$$\left\| h^N - \mathrm{R}\left(\Phi^N_f\right)\right\|_{W^{s',q'}(I)} \le C(s',q')\left\| h^N - \mathrm{R}\left(\Phi^N_f\right)\right\|_{W^{1,q'}(I)} \le C(s',q') N^{-(s-s')}\left\| h^N \right\|_{W^{1,q'}} \tag{5.12}$$

and Equations (5.4)–(5.7) hold. Invoking [37, Equation 6] or [52, Equation (1.3.3/3)] again, we obtain that

$$\left\|h^N\right\|_{W^{1,q'}(I)} \le C(q') \left\|h^N\right\|_{B^1_{q',\min\{q',2\}}(I)} \le C(q,q',s,t) \left\|h^N\right\|_{B^s_{q,t}(I)} \le C(q,q',t,s,s',p)\|f\|_{B^s_{q,t}(I)}, \tag{5.13}$$

where the second estimate holds by [53, Section 3.3.1, Equation (7)] since $s - 1/q > 1 - 1/q'$ and the last estimate follows from Equation (5.9).

We have by the triangle inequality and by invoking Equations (5.11), (5.12), and (5.13) that

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

This yields Equation (5.3) and completes the proof. □

Remark 5.5. *Note that, if $s' = 1$, then we could also obtain the estimate of Equation (5.13) by applying the inverse triangle inequality to Equation (5.8). Hence, for $s' = 1$, Equation (5.2) 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 (5.2) when replacing the assumption $s < p+1/q$ by the weaker $s \le p+1+\max\{0, 1/q-1\}$. 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 [37, 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 4.6 depends quadratically on the polynomial degree, whereas the the number of degrees of freedom depends linearly on the polynomial degree.

Theorem 5.6 ([46, 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 $\boldsymbol{p} := (p, \dots, p)$ there exists a $v \in S_{\boldsymbol{p}}(I, \mathcal{T})$ such that for all $s \in \mathbb{N}_0$ satisfying $s \le \min\{r, p\}$*

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

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 $|v|_{H^1(I)} \le |u|_{H^1(I)}$.*

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 $s \le \min\{r, p\}$*

$$\begin{aligned}\left\|u - \mathrm{R}(\Phi^{u,\mathcal{T},p})\right\|_{H^1(I)} &\le \left(1 + C_5(r)\right) \left(\tfrac{h}{p}\right)^s \|u\|_{H^{s+1}(I)}, \\ L(\Phi^{u,\mathcal{T},p}) &\le 2C_L p \log_2(p) + C_L r(2 + \log_2(p)) \log_2\left(\tfrac{p}{h}\right) + C(1 + \log_2(p))^3, \\ M(\Phi^{u,\mathcal{T},p}) &\le N[8C_M p^2 + 4C_M rp \log_2\left(\tfrac{p}{h}\right) + r \log_2\left(\tfrac{p}{h}\right) C(1 + \log_2(p))^2 + C(1 + p\log_2^2(p))], \\ M_{\mathrm{fi}}(\Phi^{u,\mathcal{T},p}) &\le 6N, \\ M_{\mathrm{la}}(\Phi^{u,\mathcal{T},p}) &\le 2N + 2.\end{aligned}$$

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

$$\begin{aligned}
\left\| u - \mathrm{R}(\Phi^{u,\mathcal{T},p}) \right\|_{H^1(I)} &\leq \|u - v\|_{H^1(I)} + \left\| v - \mathrm{R}(\Phi_{\varepsilon}^{v,\mathcal{T},\boldsymbol{p}}) \right\|_{H^1(I)} \\
&\leq C_5(r) \left(\tfrac{h}{p}\right)^s |u|_{H^{s+1}(I)} + \left(\tfrac{h}{p}\right)^r |v|_{H^1(I)} \\
&\leq \big(1 + C_5(r)\big) \left(\tfrac{h}{p}\right)^s \|u\|_{H^{s+1}(I)}, \\
L(\Phi^{u,\mathcal{T},p}) &\leq C_L(1 + \log_2(p))\big(2p + r\log_2\left(\tfrac{p}{h}\right)\big) + C_L r \log_2\left(\tfrac{p}{h}\right) + C(1 + \log_2(p))^3 \\
&\leq 2C_L p(1 + \log_2(p)) + C_L r(2 + \log_2(p)) \log_2\left(\tfrac{p}{h}\right) + C(1 + \log_2^3(p)), \\
M(\Phi^{u,\mathcal{T},p}) &\leq 8C_M N p^2 + 4C_M r \log_2\left(\tfrac{p}{h}\right) N p + r\log_2\left(\tfrac{p}{h}\right) NC(1 + \log_2(p))^2 \\
&\quad + NC(1 + p\log_2^2(p)) + 2N\left(C_L(1 + \log_2(p))\big(2p + r\log_2\left(\tfrac{p}{h}\right)\big) + C(1 + \log_2^3(p))\right) \\
&\leq N\left(8C_M p^2 + 4C_M r p \log_2\left(\tfrac{p}{h}\right) + r\log_2\left(\tfrac{p}{h}\right) C(1 + \log_2(p))^2 + C(1 + p\log_2^2(p))\right), \\
M_{\mathrm{fi}}(\Phi^{u,\mathcal{T},p}) &= M_{\mathrm{fi}}(\Phi_{\varepsilon}^{v,\mathcal{T},\boldsymbol{p}}) \leq 6N, \\
M_{\mathrm{la}}(\Phi^{u,\mathcal{T},p}) &= M_{\mathrm{la}}(\Phi_{\varepsilon}^{v,\mathcal{T},\boldsymbol{p}}) \leq 2N + 2.
\end{aligned}$$

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 $\mathcal{E}_r := \{\frac{z+z^{-1}}{2} \in \mathbb{C} : 1 \leq |z| \leq r\}$. For neural networks with certain smooth activation functions, this has been investigated in [32, 33]. We note that the results in [32] are proved even if the activation function is merely continuous. For the presently considered ReLU activations, the result is (a special case of) Theorem 3.7 in [36]. A similar result is given in [15], 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.

Theorem 5.9 ([36, Theorem 3.7])**.** *Assume that $u : [-1, 1] \to \mathbb{R}$ admits a holomorphic extension to $\mathcal{E}_\rho \subset \mathbb{C}$, for $\rho \in (1, \infty)$. Then, there exist constants $C(\rho), C_u(\rho, u) > 0$ and NNs $\{\Phi^{u,p}\}_{p \in \mathbb{N}}$ such that for all $p \in \mathbb{N}$*

$$M(\Phi^{u,p}) \leq Cp^2, \quad L(\Phi^{u,p}) \leq C(1 + p\log_2(p)), \quad \|u - \mathrm{R}(\Phi^{u,p})\|_{W^{1,\infty}(\hat{I})} \leq C_u \exp\left(-\log(\rho) p\right).$$

In addition, $M_{\mathrm{fi}}(\Phi^{u,p}) \leq C(\rho)$ and $M_{\mathrm{la}}(\Phi^{u,p}) \leq C(\rho)$.

Theorem 5.9 shows that for every $\theta > 0$ and some $c_1(\rho, \theta) > 0$

$$\|u - \mathrm{R}(\Phi^{u,p})\|_{W^{1,\infty}(\hat{I})} \leq C(\rho, \theta, u) \exp\big(-c_1 L(\Phi^{u,p})^{1/(1+\theta)}\big)$$

and that for every $\theta > 0$ and some $c_2(\rho, \theta) > 0$

$$\|u - \mathrm{R}(\Phi^{u,p})\|_{W^{1,\infty}(\hat{I})} \leq C(\rho, \theta, u) \exp\big(-c_2 M(\Phi^{u,p})^{1/2}\big).$$

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 [7] and the references there for such spaces.

For $\beta \in \mathbb{R}_{>0}$ we define $\psi_\beta : I \to \mathbb{R} : x \mapsto x^\beta$. For $k, \ell \in \mathbb{N}_0$ 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 $H_\beta^{k,\ell}(I)$. In addition, for $\delta \geq 1$, $\ell \in \mathbb{N}_0$ and $\beta \in (0,1)$ the Gevrey class $\mathcal{G}_\beta^{\ell,\delta}(I)$ is defined as the class of functions $u \in \bigcap_{k\geq\ell} H_\beta^{k,\ell}(I)$ for which there exist $C_*(u), d(u) > 0$ such that

$$\forall k \geq \ell : |u|_{H_\beta^{k,\ell}(I)} \leq C_* d^{k-\ell}((k-\ell)!)^\delta. \tag{5.14}$$

For $N \in \mathbb{N}$ and $\sigma \in (0,1)$, the mesh $\mathcal{T}_{\sigma,N}$ which is geometrically graded towards $x = 0$, is defined as follows: let $x_0 := 0$ and $x_i := \sigma^{N-i}$ 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 $I_{\sigma,i} := (x_{i-1}, x_i)$.

The following theorem is a generalization of [46, Theorem 3.36] which, in turn, generalizes earlier results in [12, 44, 21] 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 5.10 (Generalization of [46, Theorem 3.36]). *Let $\sigma, \beta \in (0,1)$, $\lambda := \sigma^{-1} - 1$, $\delta \geq 1$, $u \in \mathcal{G}_\beta^{2,\delta}(I)$ and $N \in \mathbb{N}$ be given. For $\mu_0 := \mu_0(\sigma,\beta,\delta,d) := \max\left\{1, \frac{d\lambda e^{1-\delta}}{2\sigma^{1-\beta}}\right\}$ and for $\mu > \mu_0$ let $\boldsymbol{p} = (p_i)_{i=1}^N \subset \mathbb{N}$ be defined as $p_1 := 1$ and $p_i := \lfloor \mu i^\delta \rfloor$ 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 (5.14)) 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 5.11. *Note that $v(0)$ need not equal $u(0)$. Besides, it follows from the construction of v in the proof of Theorem 5.10 that $|v|_{H^1(I\setminus I_{1,\sigma})} \leq |u|_{H^1(I\setminus I_{1,\sigma})}$.*

Theorem 5.12. *For all $\sigma, \beta \in (0,1)$, all $\delta \geq 1$, all $u \in \mathcal{G}_\beta^{2,\delta}(I)$ and all $\mu > \mu_0(\sigma,\beta,\delta,d(u))$ there exist NNs $\{\Phi^{u,\sigma,N}\}_{N\in\mathbb{N}}$ such that, for all $N \in \mathbb{N}$,*

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

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

$$\begin{aligned}
L(\Phi^{u,\sigma,N}) &\leq C_L \delta\big(2\mu N^\delta \log_2(N) + cN\log_2(N)\big) + C(\sigma,\beta,\delta,\mu)N^\delta, \\
M(\Phi^{u,\sigma,N}) &\leq 4C_M\big(2\mu^2 N^{2\delta+1} + c\mu N^{\delta+2}\big) + C(\sigma,\beta,\delta,\mu)\big(1 + N^{\delta+1}\log_2^2(N)\big), \\
M_{\mathrm{fi}}(\Phi^{u,\sigma,N}) &\leq 6N, \\
M_{\mathrm{la}}(\Phi^{u,\sigma,N}) &\leq 2N + 2.
\end{aligned}$$

Proof. Let $v \in S_{\boldsymbol{p}}(I, \mathcal{T}_{\sigma,N})$ be as in Theorem 5.10, 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 $\varepsilon := \exp(-cN)$. We define $\Phi^{u,\sigma,N} := \Phi_\varepsilon^{v,\mathcal{T}_{\sigma,N},\boldsymbol{p}}$, 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 5.11 it follows that

$$
\begin{aligned}
\left\|u - \mathrm{R}(\Phi^{u,\sigma,N})\right\|_{H^1(I)} &\leq \|u - v\|_{H^1(I)} + \left\|v - \mathrm{R}(\Phi_{\varepsilon}^{v,\mathcal{T}_{\sigma,M},\boldsymbol{p}})\right\|_{H^1(I)} \\
&\leq C_7 \exp(-cN) + \exp(-cN)\,|v|_{H^1(I\setminus I_{\sigma,1})} \\
&\leq \big(C_7 + |u|_{H^1(I)}\big)\exp(-cN), \\
L(\Phi^{u,\sigma,N}) &\leq C_L(1+\log_2(\mu N^\delta))\big(2\mu N^\delta + cN\big) + C_L cN + C(1+\log_2^3(\mu N^\delta)) \\
&\leq C_L\delta\big(2\mu N^\delta \log_2(N) + cN\log_2(N)\big) + C(\sigma,\beta,\delta,\mu)N^\delta, \\
M(\Phi^{u,\sigma,N}) &\leq 8C_M \sum_{i=1}^{N}(\mu i^\delta)^2 + 4C_M cN\sum_{i=1}^{N}(\mu i^\delta) + cNC\left(1+\sum_{i=1}^{N}\log_2^2(\mu i^\delta)\right) \\
&\quad + C\left(1+\sum_{i=1}^{N}\mu i^\delta \log_2^2(\mu i^\delta)\right) \\
&\quad + 2N\Big(C_L(1+\log_2(\mu N^\delta))\big(2\mu N^\delta + cN\big) + C(1+\log_2^3(\mu N^\delta))\Big) \\
&\leq 4C_M\big(2\mu^2 N^{2\delta+1} + c\mu N^{\delta+2}\big) + C(\sigma,\beta,\delta,\mu)\big(1+N^{\delta+1}\log_2^2(N)\big), \\
M_{\mathrm{fi}}(\Phi^{u,\sigma,N}) &\leq 6N, \\
M_{\mathrm{la}}(\Phi^{u,\sigma,N}) &\leq 2N+2.
\end{aligned}
$$

This finishes the proof. □

Theorem 5.12 shows that for $\theta > 0$ and for $c_3(\sigma,\beta,\delta,\mu,\theta), C_9(\sigma,\beta,\delta,\mu,C_*,d,|u|_{H^1(I)},\theta) > 0$

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

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

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

Remark 5.13. *In Theorem 5.12, we proved exponential expression rate bounds for deep ReLU NNs in the Sobolev space $H^1(I)$ 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 $\mathcal{S} = \{x_1, \ldots, x_J\} \subset \bar{I}$. Multivariate versions of Theorem 5.10 also hold [35].*

5.5 DNN Emulation of Boundary Layer Functions

Another class of functions for which variable order, free-knot spline approximations (or, "*rp*-approximations"[1]) achieve exponential convergence, are so-called *boundary layer functions*. Exponential boundary layer functions are ubiquitous solution components which arise from singularly perturbed elliptic and parabolic partial differential equations in several space dimensions. We refer to [31] for a regularity analysis in so-called "elliptic-elliptic" singular perturbation problems, and to [30] for corresponding approximation results. Importantly, the regularity results in [31] imply that the solutions of singular perturbation problems in one space dimension can be decomposed into a smooth (analytic) part, and into boundary layer functions, the prototypical example of which is the exponential boundary layer function $u_{1,\eta}(x) = \exp(-(x+1)/\eta)$ on the interval $\hat{I} := (-1,1)$ for the length scale parameter $\eta \in (0,1]$. The challenge is to approximate $u_{1,\eta}$ with error bounds which are uniform in the length scale parameter $\eta \in (0,1]$. Since $|u_{1,\eta}|^2_{H^1(\hat{I})} = \frac{1}{2\eta}(1-\exp(-4/\eta)) \to \infty$ for $\eta \to 0$, we do not obtain expression error bounds in $H^1(\hat{I})$ that are uniform with respect to the parameter η. Instead, we introduce the η-weighted

[1] The tag "*rp*" indicates knot (resp. node) *repositioning*, rather than knot (resp. node) *insertion*, as usually done in *hp* refinement. Node locations are NN parameters in the hidden layers. Therefore, node repositioning will affect neither NN architectures nor NN size.

norm $\|u\|_{\eta,\hat{I}} := \left(\|u\|^2_{L^2(\hat{I})} + \eta^2 |u|^2_{H^1(\hat{I})}\right)^{1/2}$ for $u \in H^1(\hat{I})$. The error in this parameter-dependent norm decreases exponentially, with constants bounded independently of η. In [47, 46], this was achieved with a p- and η-dependent mesh consisting of either one or two elements. We recall the following exponential approximation rate bounds for continuous, piecewise polynomial approximation of $u_{1,\eta}$.

Theorem 5.14 ([47, Theorem 5.1, Corollary 5.1], [46, Theorem 3.74, Corollary 3.77])**.** *For $p \in \mathbb{N}$, let the number of mesh elements N, the mesh $\mathcal{T}$ with nodes $\{x_i\}_{i=0}^N$ and the polynomial degree distribution $\boldsymbol{p}$ be as follows:*

$$\begin{cases} N = 2, \quad x_0 = -1, \quad x_1 = -1 + \kappa\tilde{p}\eta, \quad x_2 = 1, \quad \boldsymbol{p} = (p, 1), & \text{if } \kappa\tilde{p}\eta < 2, \\ N = 1, \quad x_0 = -1, \quad x_1 = 1, \quad \boldsymbol{p} = (p), & \text{if } \kappa\tilde{p}\eta \geq 2, \end{cases} \tag{5.15}$$

for $\tilde{p} := p + \frac{1}{2}$ and constants $0 < \kappa_0$ and $\kappa_0 \leq \kappa < 4/e$ which are independent of p and η.

Then, there exists $u_p \in S_{\boldsymbol{p}}(\hat{I}, \mathcal{T})$ with $u_p(\pm 1) = u_{1,\eta}(\pm 1)$ and

$$\begin{aligned} \|u_{1,\eta} - u_p\|_{L^2(\hat{I})} &\leq \eta^{1/2} C \alpha^{\tilde{p}}, \qquad |u_{1,\eta} - u_p|_{H^1(\hat{I})} \leq \eta^{-1/2} C \alpha^{\tilde{p}}, \\ \|u_{1,\eta} - u_p\|_{\eta,\hat{I}} &\leq \eta^{1/2} C \alpha^{\tilde{p}}, \quad \|u_{1,\eta} - u_p\|_{L^\infty(\hat{I})} \leq C \alpha^{\tilde{p}}, \end{aligned} \tag{5.16}$$

with

$$\alpha(p) := \left.\begin{cases} \max\{\kappa e/4, e^{-(\kappa-\epsilon)}\} & \text{if } \kappa\tilde{p}\eta < 2, \\ e/(2\tilde{p}\eta) & \text{if } \kappa\tilde{p}\eta \geq 2, \end{cases}\right\} < 1, \tag{5.17}$$

for arbitrary $\epsilon > \log(p)/(2p)$ and for C depending on κ_0 and α, but independent of p and η (except when $\kappa\tilde{p}\eta \geq 2$, then C has a factor $(1-\alpha^2)^{-1/2}$, cf. [46, Theorem 3.64] or [47, Theorem 4.1]).

We note that the η-dependence of the error bounds is natural, as it holds that

$$\|u_{1,\eta}\|_{L^2(\hat{I})} \leq (\eta/2)^{1/2}, \qquad |u_{1,\eta}|_{H^1(\hat{I})} \leq (2\eta)^{-1/2}, \qquad \|u_{1,\eta}\|_{\eta,\hat{I}} \leq \eta^{1/2}, \qquad \|u_{1,\eta}\|_{L^\infty(\hat{I})} \leq 1. \tag{5.18}$$

As a direct corollary of Theorem 5.14 and Proposition 5.1 we obtain:

Theorem 5.15. *For $0 < \kappa_0 \leq \kappa < 4/e$, $p \in \mathbb{N}$, $\eta \in (0,1]$, and $\tilde{p} := p + \frac{1}{2}$, let N, $\mathcal{T}$ and $\boldsymbol{p}$ be as in Equation* (5.15)*. Then, there exist neural networks $\{\Phi^{\eta,p}\}_{\eta\in(0,1],p\in\mathbb{N}}$, such that*

$$\begin{aligned} \|u_{1,\eta} - \mathrm{R}(\Phi^{\eta,p})\|_{L^2(\hat{I})} &\leq \eta^{1/2} C \alpha^{\tilde{p}}, \qquad |u_{1,\eta} - \mathrm{R}(\Phi^{\eta,p})|_{H^1(\hat{I})} \leq \eta^{-1/2} C \alpha^{\tilde{p}}, \\ \|u_{1,\eta} - \mathrm{R}(\Phi^{\eta,p})\|_{\eta,\hat{I}} &\leq \eta^{1/2} C \alpha^{\tilde{p}}, \quad \|u_{1,\eta} - \mathrm{R}(\Phi^{\eta,p})\|_{L^\infty(\hat{I})} \leq C \alpha^{\tilde{p}}, \end{aligned} \tag{5.19}$$

for α as in Equation (5.17) *with arbitrary $\epsilon > \log(p)/(2p)$ and for C depending on κ_0 and α, but independent of p and η (except when $\kappa\tilde{p}\eta \geq 2$, cf. Theorem 5.14). The network size and depth are bounded as follows:*

$$\begin{aligned} L(\Phi^{\eta,p}) &\leq C_L(1 + \log_2 p)(2p + p\log_2(1/\alpha) + \log_2(1/\eta)) + C(p + p\log_2(1/\alpha) + \log_2(1/\eta)), \\ M(\Phi^{\eta,p}) &\leq 4C_M p^2 (2 + \log_2(1/\alpha)) + 4C_M p \log_2(1/\eta) + C(1 + \log_2(p))^2 (p + p\log_2(1/\alpha) + \log_2(1/\eta)), \\ M_{\mathrm{fi}}(\Phi^{\eta,p}) &\leq 4, \\ M_{\mathrm{la}}(\Phi^{\eta,p}) &\leq 6. \end{aligned} \tag{5.20}$$

Proof. We define $I := (0,1)$. 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 $M\left(\Phi^{P^{-1}}\right) = M_{\mathrm{fi}}\left(\Phi^{P^{-1}}\right) = M_{\mathrm{la}}\left(\Phi^{P^{-1}}\right) = 2$.

For u_p from Theorem 5.14, we apply Proposition 5.1 to $u_p \circ P \in S_{\boldsymbol{p}}(I, \mathcal{T}')$, for $\mathcal{T}'$ obtained from $\mathcal{T}$ with the transformation P^{-1}. We choose $\varepsilon = \eta\alpha^{\tilde{p}}$ and obtain the NN $\Phi_\varepsilon^{u_p\circ P, \mathcal{T}', \boldsymbol{p}}$. Now, we define

$\Phi^{\eta,p} := \Phi_\varepsilon^{u_p\circ P, \mathcal{T}', \boldsymbol{p}} \odot \Phi^{P^{-1}}$. Using (5.18), it follows that

$$\begin{aligned}
\left\|u_p - \mathrm{R}\left(\Phi^{\eta,p}\right)\right\|_{H^1(\hat{I})} &\leq \sqrt{2}\left\|u_p \circ P - \mathrm{R}\left(\Phi_\varepsilon^{u_p\circ P, \mathcal{T}', \boldsymbol{p}}\right)\right\|_{H^1(I)} \\
&\leq \sqrt{2}\eta\alpha^{\tilde{p}}\left|u_p \circ P\right|_{H^1(I)} \\
&\leq \sqrt{2}\eta\alpha^{\tilde{p}}\sqrt{2}\left|u_p\right|_{H^1(\hat{I})} \\
&\leq 2\eta\alpha^{\tilde{p}}\left(\left|u_{1,\eta} - u_p\right|_{H^1(\hat{I})} + \left|u_{1,\eta}\right|_{H^1(\hat{I})}\right) \\
&\overset{(5.16)}{\underset{(5.18)}{\leq}} 2\eta\alpha^{\tilde{p}}\left(\eta^{-1/2}C\alpha^{\tilde{p}} + (2\eta)^{-1/2}\right) \\
&\leq C\eta^{1/2}\alpha^{\tilde{p}}.
\end{aligned}$$

This estimate, combined with Equation (5.16), shows Equation (5.19). By construction, it also holds that $\mathrm{R}(\Phi^{\eta,p})(\pm 1) = u_p(\pm 1) = u_{1,\eta}(\pm 1)$.

The bounds on the depth and the size are obtained as follows:

$$\begin{aligned}
L(\Phi^{\eta,p}) &= L\left(\Phi_\varepsilon^{u_p\circ P, \mathcal{T}', \boldsymbol{p}}\right) + L\left(\Phi^{P^{-1}}\right) \\
&\leq \Big(C_L(1+\log_2 p)\left(2p + p\log_2\left(1/\alpha\right) + \log_2(1/\eta)\right) + C\left(p\log_2\left(1/\alpha\right) + \log_2(1/\eta)\right) \\
&\quad + C(1+\log_2^3(p))\Big) + 1 \\
&\leq C_L(1+\log_2 p)\left(2p + p\log_2\left(1/\alpha\right) + \log_2(1/\eta)\right) + C\left(p + p\log_2\left(1/\alpha\right) + \log_2(1/\eta)\right), \\
M(\Phi^{\eta,p}) &\leq M\left(\Phi_\varepsilon^{u_p\circ P, \mathcal{T}', \boldsymbol{p}}\right) + M_{\mathrm{fi}}\left(\Phi_\varepsilon^{u_p\circ P, \mathcal{T}', \boldsymbol{p}}\right) + M_{\mathrm{la}}\left(\Phi^{P^{-1}}\right) + M\left(\Phi^{P^{-1}}\right) \\
&\leq \Big(8C_M(p^2+1) + 4C_M(p+1)\left(p\log_2\left(1/\alpha\right) + \log_2(1/\eta)\right) \\
&\quad + C\left(1+\log_2(p)\right)^2\left(p\log_2\left(1/\alpha\right) + \log_2(1/\eta)\right) + C(1+p\log_2^2(p))\Big) + 12 + 2 + 2 \\
&\leq 4C_M p^2\left(2+\log_2\left(1/\alpha\right)\right) + 4C_M p\log_2(1/\eta) + C(1+\log_2(p))^2\left(p + p\log_2\left(1/\alpha\right) + \log_2(1/\eta)\right), \\
M_{\mathrm{fi}}(\Phi^{\eta,p}) &\leq 2M_{\mathrm{fi}}\left(\Phi^{P^{-1}}\right) \leq 4, \\
M_{\mathrm{la}}(\Phi^{\eta,p}) &\leq M_{\mathrm{la}}\left(\Phi_\varepsilon^{u_p\circ P, \mathcal{T}', \boldsymbol{p}}\right) \leq 6.
\end{aligned}$$

□

Equations (5.20) and (5.19) can be combined to get, for some constants $C_{11}, C_{12} > 0$ depending only on α and κ_0, that

$$\begin{aligned}
M(\Phi^{\eta,p}) &\leq C_{11}\left(p\log_2\left(1/\alpha\right) + \tfrac{1}{2}\log_2(1/\eta)\right)^2, \\
\left\|u_{1,\eta} - \mathrm{R}\left(\Phi^{\eta,p}\right)\right\|_{L^2(\hat{I})}, \left\|u_{1,\eta} - \mathrm{R}\left(\Phi^{\eta,p}\right)\right\|_{\eta,\hat{I}} &\leq C_{12}\eta^{1/2}\alpha^p \\
&= C_{12}\exp\left(-C_{11}^{-1/2}\left(C_{11}\log_2^2\left(\eta^{1/2}\alpha^p\right)\right)^{1/2}\right) \\
&\leq C_{12}\exp\left(-C_{11}^{-1/2}M\left(\Phi^{\eta,p}\right)^{1/2}\right).
\end{aligned}$$

6 Multivariate Approximation: Radial Basis Functions

The preceding results on NN approximation adressed the univariate case only. One method to extend these results to the multivariate setting is through concatenation, which is naturally accommodated by DNNs. We illustrate this for the (widely used) class of *isotropic and anisotropic radial basis functions*, cf. [6] and [54] and the references therein. NN approximations of such functions have been considered in e.g. [28, 29, 10].

For dimension $d \in \mathbb{N}$, it was shown in [29] that the Euclidean norm $\mathbb{R}^d \ni x \mapsto \|x\|_{2,\mathbb{R}^d}$ can be approximated efficiently by NNs. To approximate a radially symmetric function $\mathbb{R}^d \ni x \mapsto g(\|x\|_{2,\mathbb{R}^d})$, with *radial profile function* g in any function class considered in Section 5, we concatenate the corresponding NN approximation from Section 5 with the NN approximation for the Euclidean norm $\mathbb{R}^d \ni x \mapsto \|x\|_{2,\mathbb{R}^d}$ based on [29], which we present in Section 6.1. As we shall show in Sections 6.2 and 6.3, the results from Section 5 then relate the smoothness of the profile function g to the size of the approximating NNs.

In addition, anisotropic radial-like functions can be approximated by concatenating a shallow NN emulating an affine transformation with a deep NN approximating an isotropic radially symmetric function.

Remark 6.1. *We point out that if instead of radial functions of the form $x \mapsto g(\|x\|_2)$ one only needs to express with NNs functions of the form $x \mapsto g(\|x\|_2^2)$ (i.e.,* even *functions $\zeta \mapsto g(\zeta)$ as arise in several kernels that are widely used in scattered data approximation, see [54]), the ensuing proofs could be simplified as one approximates*

$$x \mapsto \|x\|_2^2 = \sum_{k=1}^{d} x_i^2$$

through d Yarotsky products with a ReLU NN of size $O(d\log(d/\varepsilon))$ and depth $O(\log(d)\log(d/\varepsilon))$, i.e. near linear scaling with respect to the dimension d and logarithmic scaling with respect to the target accuracy $\varepsilon \in (0,1]$. The ensuing Proposition 6.2 achieves essentially this result for the euclidean norm $x \mapsto \|x\|_2$, albeit with a considerably more involved argument based on ideas in [29]. As it will allow us to cover NN approximation of general profile functions g, we opt to detail the NN emulation of the euclidean norm $x \mapsto \|x\|_2$.

As we will see in Section 6.3, combining results from Section 5.4 with the NN approximation of $x \mapsto \|x\|_2$ allows us to show exponential convergence in L^∞-norm of the approximation of $x \mapsto \|x\|_2^s$ for any $s > \frac{1}{2}$. Instead, combining Section 6.3 with the approximation of $x \mapsto \|x\|_2^2$ would only allow us to approximate $x \mapsto \|x\|_2^s$ for $s \geq 1$.

6.1 NN approximation of the euclidean norm $\mathbb{R}^d \ni x \mapsto \|x\|_{2,\mathbb{R}^d}$

We now recall the NN approximation of the Euclidean norm on $\mathbb{R}^d$ and derive a bound on the $W^{1,\infty}$-error. Then, we formulate our main result on the approximation of anisotropic radial-like functions.

Proposition 6.2 (cf. [29, Lemma 4])**.** *For all dimensions $d \geq 2$ and target accuracy $\delta > 0$, there exists a NN $\Phi^{\mathrm{Eucl}}_{d,\delta}$ with input dimension d and output dimension 1, such that $\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)$ is 1-Lipschitz continuous,*

$$\left|\|x\|_{2,\mathbb{R}^d} - \mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(x)\right| \leq \delta \|x\|_{2,\mathbb{R}^d}, \qquad \text{for all } x \in \mathbb{R}^d, \tag{6.1}$$

$$\left|\|\cdot\|_{2,\mathbb{R}^d} - \mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)\right|_{W^{1,\infty}(\mathbb{R}^d)} \leq \delta, \qquad \text{for a.e. } x \in \mathbb{R}^d, \tag{6.2}$$

and

$$L\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right) \leq \log_2(d)\log_2\left(10\pi\frac{d}{\delta}\right), \quad M\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right) \leq 16(d-1)\log_2\left(10\pi\frac{d}{\delta}\right).$$

Before we prove Proposition 6.2, we first discuss the proof of [29, Lemma 4]. To approximate the Euclidean norm on $\mathbb{R}^d$, in [29, Supplementary material, Equation (29)] it is observed that for $x \in \mathbb{R}^d$ with $n = \min\{2^k : k \in \mathbb{N}, 2^k \geq d\} = 2^{\lceil \log_2(d)\rceil}$ and with $x_k = 0$ for $k = d+1, \ldots, n$

$$\|x\|_{2,\mathbb{R}^d} = \sqrt{x_1^2 + x_2^2 + \ldots + x_d^2} = \sqrt{\sqrt{\sqrt{x_1^2+x_2^2}^2 + \sqrt{x_3^2+x_4^2}^2}^2 \ldots \sqrt{\ldots + \sqrt{x_{n-1}^2+x_n^2}^2}^2}.$$

The NN in the proof of [29, Lemma 4] consists of a binary tree of NNs from [29, Lemma 3], which approximate the Euclidean norm on $\mathbb{R}^2$. For a maximal input size $R > 0$ and an accuracy $\delta_1 > 0$, the NN constructed in [29, Lemma 3] is the $\lceil \log_2(R\pi/\delta_1)\rceil$-fold concatenation of folding networks from [29, Lemma 2] followed by a projection on the x_1-coordinate. We now recall [29, Lemma 2], providing bounds on the number of nonzero coefficients.

Lemma 6.3 ([29, Lemma 2])**.** *For every unit vector $l = (l_1, l_2)^\top \in \mathbb{R}^2$ there exists a* folding network *Φ^{l_1,l_2} with input dimension 2 and output dimension 2 such that*

$$\mathrm{R}\left(\Phi^{l_1,l_2}\right)(x) = \begin{cases} x & \text{if } \langle l, x\rangle \geq 0, \\ \begin{pmatrix} l_1^2 - l_2^2 & 2l_1l_2 \\ 2l_1l_2 & l_2^2 - l_1^2 \end{pmatrix} x & \text{otherwise.} \end{cases} \tag{6.3}$$

The NN realizes a folding transformation *about the line through the origin perpendicular to l, called* fold line*.*

It holds that

$$L\left(\Phi^{l_1,l_2}\right) = 2, \quad M\left(\Phi^{l_1,l_2}\right) = 16, \quad M_{\text{fi}}\left(\Phi^{l_1,l_2}\right) = 8, \quad M_{\text{la}}\left(\Phi^{l_1,l_2}\right) = 8,$$

and the hidden layer consists of 4 *neurons.*

Proof. Equation (6.3) is shown in the proof of [29, Lemma 2] in the supplementary material of [29]. All network weights are visualized in [29, Supplementary material, Figure 2], from which the bound on the number of nonzero weights directly follows. □

Remark 6.4 (the NN approximation of $x \mapsto \|x\|_{2,\mathbb{R}^2}$ is 1-Lipschitz continuous, almost everywhere the composition of $\lceil \log_2(R\pi/\delta_1) \rceil$ folding transformations is locally an isometry)**.** *Note that the folding transformation leaves the Euclidean norm invariant, if the fold line is a line through the origin. In addition, we note that if two points $x, y \in \mathbb{R}^2$ are on the same side of the fold line or if x or y lies on the fold line, then the folding transformation preserves their distance. If x, y are on opposite sides of the fold line, then the distance between the images of x, y under the folding transformation is less than the distance between x and y. We conclude that folding transformations are* 1*-Lipschitz continuous. In addition, also the projection onto the x_1-coordinate is* 1*-Lipschitz continuous. The realization of the NN from [29, Lemma 3], approximating $x \mapsto \|x\|_{2,\mathbb{R}^2}$, is the composition of $\lceil \log_2(R\pi/\delta_1) \rceil$ folding transformations and a projection onto the x_1 coordinate, thus* 1*-Lipschitz continuous.*

Moreover, considering a single folding transformation, for all $x \in \mathbb{R}^2$ not on the fold line, there exists an open neighborhood on which the folding transformation equals either the identity or the reflection in the fold line. In both cases, on this neighborhood the folding transformation is an isometry.

This argument extends to the composition of $\lceil \log_2(R\pi/\delta_1) \rceil$ folding transformations. Consider for each folding transformation the fold line, and the preimage of the fold line under the preceding folding transformations. Their union consists of $2^{\lceil \log_2(R\pi/\delta_1) \rceil - 1}$ lines through the origin. If $x \in \mathbb{R}^2$ is not element of any of those lines, then the composition of these folding transformations near x equals a composition of reflections and identity transformations, which is an isometry. Near almost every $x \in \mathbb{R}^2$ the composition of $\lceil \log_2(R\pi/\delta_1) \rceil$ folding transformations is an isometry, because the union of $2^{\lceil \log_2(R\pi/\delta_1) \rceil - 1}$ lines has zero Lebesgue measure.

Proof of Proposition 6.2. Similar to the proof of [29, Lemma 4] in the supplementary material of [29], for $\delta_1 := \frac{\delta}{10d} > 0$ we define the NN $\Phi^{\text{Eucl}}_{d,\delta}$ to be a binary tree of NNs $\Phi^{\text{Eucl}}_{2,\delta_1}$ approximating the Euclidean norm on $\mathbb{R}^2$.

Deviating from [29, Lemma 3], we define the NN $\Phi^{\text{Eucl}}_{2,\delta_1}$ to be the $f := \lceil \log_2(\pi/\delta_1) \rceil$-fold concatenation of folding networks from Lemma 6.3, followed by a projection onto the x_1-coordinate. We denote the composition of the f folding transformations by F_{δ_1}. Contrary to [29, Lemma 3], we do not consider a maximal input size for $\Phi^{\text{Eucl}}_{2,\delta_1}$. As a result, the error bound will scale linearly with the Euclidean norm of the input.

In polar coordinates, with each folding transformation the range of the angular coordinate is reduced by a factor 2, cf. [29, Figure 1a]. After f folding transformations, the angular coordinate of the output is contained in an interval of length $2^{-f}(2\pi)$. If we align the folding transformations such that the image $F_{\delta_1}(\mathbb{R}^2)$ is a cone at the origin symmetric around the positive x_1-axis, it holds that the angular coordinate of every point in $F_{\delta_1}(\mathbb{R}^2)$ is in $[-2^{-f}\pi, 2^{-f}\pi]$. To do so, for $i = 1, \dots, f$ we choose the folding directions as

$$l(i) = (l_{1,i}, l_{2,i})^\top, \quad \text{where} \quad l_{1,i} = \cos\left(\frac{\pi}{2^{i-1}} - \frac{\pi}{2} - \frac{\pi}{2^f}\right) \quad \text{and} \quad l_{2,i} = \sin\left(\frac{\pi}{2^{i-1}} - \frac{\pi}{2} - \frac{\pi}{2^f}\right).$$

Following the proof of [29, Lemma 3] in the supplementary material of [29], we define $\Phi^{\text{Eucl}}_{2,\delta_1}$ to be the concatenation (as in [39, Definition 2.2]) of f folding networks from Lemma 6.3 with previously described folding directions $(l(i))_{i=1,\dots,f}$. In addition, for the remainder of this proof we will use the following notation: for $k \in \mathbb{N}$ we inductively define

$$\Phi^k_{\delta_1} := \Phi^{\text{Eucl}}_{2,\delta_1} \odot \text{FP}\left(\Phi^{k-1}_{\delta_1}, \Phi^{k-1}_{\delta_1}\right),$$

and with $m := \lceil \log_2(d) \rceil$ we finally define $\Phi^{\text{Eucl}}_{d,\delta} := \Phi^m_{\delta_1}$.

Analogous to the error estimate in the proof of [29, Lemma 3], we find for the error of the approximation of the Euclidean norm on $\mathbb{R}^2$, for all $x \in \mathbb{R}^2$:

$$0 \le \|x\|_{2,\mathbb{R}^2} - \mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{2,\delta_1}\right)(x) \le \|x\|_{2,\mathbb{R}^2} - \|x\|_{2,\mathbb{R}^2}\cos(\pi 2^{-f}) \le \|x\|_{2,\mathbb{R}^2}\,\pi 2^{-f} \le \delta_1 \|x\|_{2,\mathbb{R}^2}. \tag{6.4}$$

The lower bound holds, because the fold operations preserve the Euclidean norm, and the projection onto the x_1-coordinate can only reduce it.

It follows by [29, Supplementary material, Equation (30)] for $n = 1$ and $i = \lceil \log_2(d) \rceil$ that, with $m := \lceil \log_2(d) \rceil$,

$$\|x\|_{2,\mathbb{R}^d} - \mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(x) \le \delta_1 \left((2^{\lfloor \frac{m+1}{2} \rfloor} - 1) + \sqrt{2}(2^{\lfloor \frac{m}{2} \rfloor} - 1) \right) \|x\|_{2,\mathbb{R}^d}, \qquad \text{for all } x \in \mathbb{R}^d. \tag{6.5}$$

The left-hand side is bounded from below by 0, as follows from (6.4) by induction.

With

$$\left((2^{\lfloor \frac{m+1}{2} \rfloor} - 1) + \sqrt{2}(2^{\lfloor \frac{m}{2} \rfloor} - 1) \right) \le (2\sqrt{2})2^{\frac{m}{2}} \le (2\sqrt{2})2^{\frac{\log_2(d)+1}{2}} = 4\sqrt{d}$$

and $\delta_1 = \frac{\delta}{10d}$ Equation (6.1) follows.

To see that $\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)$ is 1-Lipschitz continuous, we note that it is a binary tree of realizations $\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{2,\delta_1}\right)$. These are composed of 1-Lipschitz maps (folds and a projection, cf. Remark 6.4), and are thus 1-Lipschitz continuous. As a result, also $\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)$ is 1-Lipschitz continuous.

To bound the error in the first partial derivative, we note that for all $x \in \mathbb{R}^2$ $\partial_2 \|\cdot\|_{2,\mathbb{R}^2}(x) = \langle e_2\,,\, \widehat{r}(x) \rangle$, where e_2 denotes the unit vector in the x_2-direction and $\widehat{r}(x) = x/\|x\|$ is the unit vector in the radial direction. Denoting the angle between e_2 and $\widehat{r}(x)$ by $\theta_1(x)$, $\partial_2 \|\cdot\|_{2,\mathbb{R}^2}(x) = \cos(\theta_1(x))$.

We now determine the derivative of $\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{2,\delta_1}\right)(x)$, defined for almost every $x \in \mathbb{R}^2$. We denote the push-forward of the vectors e_2 and $\widehat{r}(x)$ at x under the transformation F_{δ_1} in the point $F_{\delta_1}(x)$ by $\mathrm{d}F_{\delta_1}(e_2)(F_{\delta_1}(x))$ and $\mathrm{d}F_{\delta_1}(\widehat{r})(F_{\delta_1}(x))$.

As for the NN from [29, Lemma 3] discussed in Remark 6.4, near almost every $x \in \mathbb{R}^2$ the map F_{δ_1} locally constitutes an isometry. Thus, for almost every $x \in \mathbb{R}^2$ it holds

$$\langle e_2\,,\, \widehat{r}(x) \rangle = \langle \mathrm{d}F_{\delta_1}(e_2)(F_{\delta_1}(x))\,,\, \mathrm{d}F_{\delta_1}(\widehat{r})(F_{\delta_1}(x)) \rangle.$$

Near those $x \in \mathbb{R}^2$, the map F_{δ_1} is a composition of reflections in a line through the origin and the identity. This means it preserves $\|\cdot\|_{2,\mathbb{R}^2}$, hence $\mathrm{d}F_{\delta_1}(\widehat{r})(F_{\delta_1}(x)) = \widehat{r}(F_{\delta_1}(x))$ and

$$\frac{\partial \|\cdot\|_{2,\mathbb{R}^2}}{\partial x_2}(x) = \cos(\theta_1(x)) = \langle e_2\,,\, \widehat{r}(x) \rangle = \langle \mathrm{d}F_{\delta_1}(e_2)(F_{\delta_1}(x))\,,\, \widehat{r}(F_{\delta_1}(x)) \rangle.$$

In addition, we note that

$$\frac{\partial \mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{2,\delta_1}\right)}{\partial x_2}(x) = \langle \mathrm{d}F_{\delta_1}(e_2)(F_{\delta_1}(x))\,,\, e_1 \rangle = \cos(\theta_2(x)),$$

where $\theta_2(x)$ denotes the angle between $\mathrm{d}F_{\delta_1}(e_2)(F_{\delta_1}(x))$ and e_1. By construction, for all $y \in \mathbb{R}^2$ the angular coordinate of $F_{\delta_1}(y)$ lies in the interval $[-\delta_1, \delta_1]$. This implies that the angle $\theta_3(x)$ between $\widehat{r}(F_{\delta_1}(x))$ and e_1 is at most δ_1. Also, note that $|\theta_1(x) - \theta_2(x)| = \theta_3(x) \le \delta_1$. With the 1-Lipschitz continuity of $\cos(\cdot)$ we obtain that for almost all $x \in \mathbb{R}^2$

$$\left| \frac{\partial \left(\|\cdot\|_{2,\mathbb{R}^2} - \mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{2,\delta_1}\right) \right)}{\partial x_2}(x) \right| \le |\cos(\theta_1(x)) - \cos(\theta_2(x))| \le |\theta_1(x) - \theta_2(x)| \le \delta_1. \tag{6.6}$$

The error in the derivative with respect to x_1 can be estimated in the same way.

Finally, with the chain rule it follows that

$$\left\| \nabla \|x\|_{2,\mathbb{R}^d} - \nabla \left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(x) \right) \right\|_{2,\mathbb{R}^d} \le 10d\delta_1, \qquad \text{for } a.e.\ x \in \mathbb{R}^d. \tag{6.7}$$

We show this by induction with respect to the dimension d. For $d = 2$, the assertion follows from Equation (6.6). To prove (6.7) for $d > 2$, we write $m = \lceil \log_2(d) \rceil$ and we show by induction that

$$\left\| \frac{\partial}{\partial x_\ell} \|\cdot\|_{2,\mathbb{R}^{2^i}} - \frac{\partial}{\partial x_\ell} \mathrm{R}\left(\Phi^i_{\delta_1}\right) \right\|_{L^\infty(\mathbb{R}^{2^i})} \leq \left(\sum_{j=1}^{i-1} \delta_j + i\delta_1 \right) \qquad \text{for all } \ell = 1, \dots, 2^i \text{ and all } i = 1, \dots, m, \tag{6.8}$$

where for $j = 1, \dots, m-1$

$$\delta_j := \left((2^{\lfloor \frac{j+1}{2} \rfloor} - 1) + \sqrt{2}(2^{\lfloor \frac{j}{2} \rfloor} - 1) \right).$$

We only consider $\ell = 1$, the bounds on the other derivatives are analogous. Assuming that for some $k \in \mathbb{N}$ Equation (6.8) holds for all $i < k$, we estimate for a.e. $x \in \mathbb{R}^{2^k}$

$$\begin{aligned}
&\left| \frac{\partial}{\partial x_1} \|x\|_{2,\mathbb{R}^{2^k}} - \frac{\partial}{\partial x_1} \left(\mathrm{R}\left(\Phi^k_{\delta_1}\right)(x) \right) \right| \\
&= \Bigg| \left(\frac{\partial}{\partial x_1} \|\cdot\|_{2,\mathbb{R}^2} \right) \left(\|(x_1, \dots, x_{2^{k-1}})\|_{2,\mathbb{R}^{2^{k-1}}}, \left\|(x_{2^{k-1}+1}, \dots, x_{2^k})\right\|_{2,\mathbb{R}^{2^{k-1}}} \right) \cdot \frac{\partial}{\partial x_1} \|(x_1, \dots, x_{2^{k-1}})\|_{2,\mathbb{R}^{2^{k-1}}} \\
&\quad - \left(\frac{\partial}{\partial x_1} \mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{2,\delta_1}\right) \right) \left(\mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_1, \dots, x_{2^{k-1}}), \mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_{2^{k-1}+1}, \dots, x_{2^k}) \right) \\
&\quad \cdot \frac{\partial}{\partial x_1} \left(\mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_1, \dots, x_{2^{k-1}}) \right) \Bigg| \\
&\leq \Bigg| \left(\frac{\partial}{\partial x_1} \|\cdot\|_{2,\mathbb{R}^2} \right) \left(\|(x_1, \dots, x_{2^{k-1}})\|_{2,\mathbb{R}^{2^{k-1}}}, \left\|(x_{2^{k-1}+1}, \dots, x_{2^k})\right\|_{2,\mathbb{R}^{2^{k-1}}} \right) \\
&\quad - \left(\frac{\partial}{\partial x_1} \|\cdot\|_{2,\mathbb{R}^2} \right) \left(\mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_1, \dots, x_{2^{k-1}}), \mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_{2^{k-1}+1}, \dots, x_{2^k}) \right) \Bigg| \\
&\quad \cdot \left| \frac{\partial}{\partial x_1} \|(x_1, \dots, x_{2^{k-1}})\|_{2,\mathbb{R}^{2^{k-1}}} \right| \\
&\quad + \left| \left(\frac{\partial}{\partial x_1} \|\cdot\|_{2,\mathbb{R}^2} - \frac{\partial}{\partial x_1} \mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{2,\delta_1}\right) \right) \left(\mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_1, \dots, x_{2^{k-1}}), \mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_{2^{k-1}+1}, \dots, x_{2^k}) \right) \right| \\
&\quad \cdot \left| \frac{\partial}{\partial x_1} \|(x_1, \dots, x_{2^{k-1}})\|_{2,\mathbb{R}^{2^{k-1}}} \right| \\
&\quad + \left| \left(\frac{\partial}{\partial x_1} \mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{2,\delta_1}\right) \right) \left(\mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_1, \dots, x_{2^{k-1}}), \mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_{2^{k-1}+1}, \dots, x_{2^k}) \right) \right| \\
&\quad \cdot \left| \frac{\partial}{\partial x_1} \|(x_1, \dots, x_{2^{k-1}})\|_{2,\mathbb{R}^{2^{k-1}}} - \frac{\partial}{\partial x_1} \left(\mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_1, \dots, x_{2^{k-1}}) \right) \right|.
\end{aligned} \tag{6.9}$$

Because the Euclidean norm is 1-Lipschitz continuous, it follows that $\left| \frac{\partial}{\partial x_1} \|(x_1, \dots, x_{2^{k-1}})\|_{2,\mathbb{R}^{2^{k-1}}} \right| \leq 1$. By Equation (6.6), the second term in (6.9) can be estimated by δ_1. To estimate the third term, we use the induction hypothesis and the fact that $\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{2,\delta_1}\right)$ is 1-Lipschitz continuous. As a result, the third term can be estimated by $\sum_{j=1}^{k-2} \delta_j + (k-1)\delta_1$.

To estimate the first term in (6.9), define

$$\begin{aligned}
a &:= \left(\|(x_1, \dots, x_{2^{k-1}})\|_{2,\mathbb{R}^{2^{k-1}}}, \left\|(x_{2^{k-1}+1}, \dots, x_{2^k})\right\|_{2,\mathbb{R}^{2^{k-1}}} \right) =: (a_1, a_2) \in \mathbb{R}^2, \\
b &:= \left(\mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_1, \dots, x_{2^{k-1}}), \mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_{2^{k-1}+1}, \dots, x_{2^k}) \right) =: (b_1, b_2) \in \mathbb{R}^2
\end{aligned}$$

and denote by θ_a, θ_b the angular coordinates of a and b, respectively. Then, with the 1-Lipschitz continuity of $\cos(\cdot)$ it follows that

$$\left| \left(\frac{\partial}{\partial x_1} \|\cdot\|_{2,\mathbb{R}^2} \right)(a) - \left(\frac{\partial}{\partial x_1} \|\cdot\|_{2,\mathbb{R}^2} \right)(b) \right| = |\cos\theta_a - \cos\theta_b| \leq |\theta_a - \theta_b|.$$

To estimate this difference, we use the following corollary of Equation (6.4):

$$0 \leq \|(x_1, \dots, x_{2^{k-1}})\|_{2,\mathbb{R}^{2^{k-1}}} - \mathrm{R}\left(\Phi^{k-1}_{\delta_1}\right)(x_1, \dots, x_{2^{k-1}}) = a_1 - b_1 \leq \delta_{k-1} \|(x_1, \dots, x_{2^{k-1}})\|_{2,\mathbb{R}^{2^{k-1}}} = \delta_{k-1} a_1,$$

where

$$\delta_{k-1} := \left((2^{\lfloor \frac{k-1+1}{2} \rfloor} - 1) + \sqrt{2}(2^{\lfloor \frac{k-1}{2} \rfloor} - 1) \right).$$

The upper bound follows from [29, Supplementary material, Equation (30)] with $n = 1$ and $i = k-1$, similar to (6.5). The lower bound follows from (6.4) by induction. Similarly, it holds that $0 \leq a_2 - b_2 \leq \delta_{k-1} a_2$. For arbitrary, fixed a and within the previously given bounds on b_1, b_2, the difference $|\theta_a - \theta_b|$ is largest when $b_1 = a_1(1 - \delta_{k-1})$ and $b_2 = a_2$ in case $|\frac{a_2}{a_1}| > 1$ or when $b_1 = a_1$ and $b_2 = a_2(1 - \delta_{k-1})$ in case $|\frac{a_2}{a_1}| \leq 1$. W.l.o.g. we only consider the case that $0 \leq \frac{a_2}{a_1} \leq 1$ with $a_1, a_2 \geq 0$. Then, $\tan(\theta_a) = \frac{a_2}{a_1}$ and $\tan(\theta_b) = \frac{b_2}{b_1} = \frac{a_2}{a_1}(1 - \delta_{k-1})$. Hence, using that $\frac{\mathrm{d}}{\mathrm{d}x} \tan(x) = (\cos(x))^{-2} \geq 1$ for all $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$, it follows that

$$|\theta_a - \theta_b| \leq |\tan(\theta_a) - \tan(\theta_a)| \leq \delta_{k-1} \tan(\theta_a) \leq \delta_{k-1}.$$

The same estimate $|\theta_a - \theta_b| \leq \delta_{k-1}$ can be derived in case $|\frac{a_2}{a_1}| > 1$, with the same argument, measuring angles with respect to the x_2-axis instead of the x_1-axis. This finishes the estimate of the first term in Equation (6.9), which is bounded by δ_{k-1}. Combining Equation (6.9) with the estimates for the three terms gives Equation (6.8) for $i = k$:

$$\left| \frac{\partial}{\partial x_1} \|x\|_{2,\mathbb{R}^{2^k}} - \frac{\partial}{\partial x_1} \left(\mathrm{R}\left(\Phi^k_{\delta_1} \right)(x) \right) \right| \leq \delta_{k-1} + \delta_1 + \left(\sum_{j=1}^{k-2} \delta_j + (k-1)\delta_1 \right) = \sum_{j=1}^{k-1} \delta_j + k\delta_1.$$

Next, we use (6.8) with $i = m \leq \log_2(d) + 1$, and use that $(\sqrt{2} - 1) \geq \frac{2}{5}$ and that $m > 1$ (because $d > 2$). We obtain

$$\begin{aligned} \sum_{j=1}^{m-1} \delta_j + m\delta_1 &\leq \left(\sum_{j=1}^{m-1} (2\sqrt{2}) 2^{\frac{j}{2}} - 2(m-1) \right) \delta_1 + m\delta_1 \\ &\leq \delta_1 (2\sqrt{2}) \frac{2^{\frac{m}{2}}}{\sqrt{2} - 1} \leq \delta_1 (5\sqrt{2}) 2^{\frac{\log_2(d)+1}{2}} = 10\sqrt{d}\delta_1. \end{aligned}$$

It follows that for a.e. $x \in \mathbb{R}^d$

$$\left\| \nabla \|x\|_{2,\mathbb{R}^d} - \nabla \left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta} \right)(x) \right) \right\|_{2,\mathbb{R}^d} \leq 10 d \delta_1.$$

This finishes the proof of Equation (6.2).

We now provide bounds on the network depth and size. Each NN $\Phi^{\mathrm{Eucl}}_{2,\delta_1}$ is fully connected with 4 neurons in each layer, hence there are at most 16 nonzero coefficients per layer. As a result, the number of nonzero weights in $\Phi^{\mathrm{Eucl}}_{d,\delta}$ is bounded by four times the number of neurons. Analogous to the proof of [29, Lemma 4], we obtain the following bounds on the depth and the number of neurons:

$$L\left(\Phi^{\mathrm{Eucl}}_{d,\delta} \right) \leq \log_2(d) \log_2\left(\frac{\pi}{\delta} 10d \right), \quad N\left(\Phi^{\mathrm{Eucl}}_{d,\delta} \right) \leq 4(d-1)\log_2\left(\frac{\pi}{\delta} 10d \right).$$

This gives the following bound on the network size:

$$M\left(\Phi^{\mathrm{Eucl}}_{d,\delta} \right) \leq 16(d-1)\log_2\left(\frac{\pi}{\delta} 10d \right).$$

□

6.2 Anisotropic Radial Functions in High Dimension

We now turn to the main result of this section, approximation rate bounds of deep NN aproximations for *anisotropic radial-like functions in high dimension.*

Theorem 6.5. *Let $d \in \mathbb{N}$, $R > 0$, $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$, and $D := \{x \in \mathbb{R}^d : \|Ax + b\|_{2,\mathbb{R}^d} \leq R\}$. Let $g \in W^{2,\infty}([0,R])$ be such that for all $\beta \in (0,1)$ the function g can be approximated by a NN $\Phi^g_{\beta,R}$ such that*

$$\left\| g - \mathrm{R}\left(\Phi^g_{\beta,R} \right) \right\|_{W^{1,\infty}([0,R])} \leq \beta \|g\|_{W^{1,\infty}([0,R])}, \tag{6.10}$$

$$L\left(\Phi^g_{\beta,R} \right) =: L_{\beta,R}, \quad M\left(\Phi^g_{\beta,R} \right) =: M_{\beta,R}.$$

Consider the anisotropic radial-like function

$$f : D \to \mathbb{R} : x \mapsto g(\|Ax + b\|_{2,\mathbb{R}^d}).$$

Then, for every $\varepsilon \in (0,1)$*, there exists a NN* $\Phi^f_{\varepsilon,D}$ *such that*

$$\left\| f - \mathrm{R}\left(\Phi^f_{\varepsilon,D}\right)\right\|_{L^\infty(D)} \leq \varepsilon \|g\|_{W^{1,\infty}([0,R])}, \tag{6.11}$$

$$\left\| f - \mathrm{R}\left(\Phi^f_{\varepsilon,D}\right)\right\|_{W^{1,\infty}(D)} \leq \varepsilon \|A\|_{2,\mathbb{R}^d} \|g\|_{W^{2,\infty}([0,R])}, \tag{6.12}$$

$$L\left(\Phi^f_{\varepsilon,D}\right) \leq L_{\beta,R} + \log_2(d)\log_2\left(30\pi d\sqrt{d}\max\{R,1\}/\varepsilon\right) + 1,$$

$$M\left(\Phi^f_{\varepsilon,D}\right) \leq 2M_{\beta,R} + 4d^2 + 64(d-1)\log_2\left(30\pi d\sqrt{d}\max\{R,1\}/\varepsilon\right) + 4d.$$

Remark 6.6. *For results in Section 5, if instead of* (6.10) *the right-hand side of the error bound for* $\Phi^g_{\beta,R}$ *depends linearly on a stronger norm of* g*, then a result similar to Theorem 6.5 holds, with the right-hand sides of* (6.11) *and* (6.12) *depending linearly on the stronger norm.*

Otherwise, an error bound for $\Phi^g_{\beta,R}$ *which scales linearly with* $\|g\|_{W^{1,\infty}([0,R])}$ *can be obtained as follows: If* $g \not\equiv 0$*, then* g *can be approximated by taking the approximation of* $g/\|g\|_{W^{1,\infty}([0,R])}$ *from Section 5, but with the weights of the output layer multiplied by* $\|g\|_{W^{1,\infty}([0,R])}$*. This NN satisfies an error bound linear in* $\|g\|_{W^{1,\infty}([0,R])}$ *as in* (6.10)*. Note that in all results in Section 5 the bounds on the network size are independent of* $\|g\|_{W^{1,\infty}([0,R])}$*, i.e. these bounds are not affected by this rescaling. On the other hand, if* $g \equiv 0$*, then it can be emulated exactly by a NN of depth* 1 *and size* 0*.*

Proof of Theorem 6.5. We note that the affine transformation $T : D \to \mathbb{R}^d : x \mapsto Ax + b$ can be emulated exactly by a NN $\Phi^{A,b}$ of depth 1 and size at most $d^2 + d$. For the approximation of $\|\cdot\|_{2,\mathbb{R}^d}$, we apply Proposition 6.2 with $\delta := \frac{\varepsilon}{3\sqrt{d}\max\{R,1\}}$. For the approximation of g, we use the assumption of the theorem with $\beta := \frac{\varepsilon}{3\sqrt{d}}$. We define

$$\Phi^f_{\varepsilon,D} := \Phi^g_{\beta,R} \odot \Phi^{\mathrm{Eucl}}_{d,\delta} \odot \Phi^{A,b}.$$

Its depth and its size can be bounded as follows:

$$\begin{aligned}
L\left(\Phi^f_{\varepsilon,D}\right) &\leq L\left(\Phi^g_{\beta,R}\right) + L\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right) + L\left(\Phi^{A,b}\right)\\
&\leq L_{\beta,R} + \log_2(d)\log_2\left(30\pi d\sqrt{d}\max\{R,1\}/\varepsilon\right) + 1,\\
M\left(\Phi^f_{\varepsilon,D}\right) &\leq 2M\left(\Phi^g_{\beta,R}\right) + 4M\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right) + 4M\left(\Phi^{A,b}\right)\\
&\leq 2M_{\beta,R} + 4\left(16(d-1)\log_2\left(30\pi d\sqrt{d}\max\{R,1\}/\varepsilon\right)\right) + 4(d^2+d)\\
&\leq 2M_{\beta,R} + 4d^2 + 64(d-1)\log_2\left(30\pi d\sqrt{d}\max\{R,1\}/\varepsilon\right) + 4d.
\end{aligned}$$

Using that the $W^{1,\infty}([0,R])$-seminorm equals the Lipschitz constant, we estimate the error as follows:

$$\begin{aligned}
&\left\| g(\|A\cdot+b\|_{2,\mathbb{R}^d}) - \mathrm{R}\left(\Phi^g_{\beta,R}\right)\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)}\\
\leq &\left\| g(\|A\cdot+b\|_{2,\mathbb{R}^d}) - g\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)}\\
&+ \left\|\left(g - \mathrm{R}\left(\Phi^g_{\beta,R}\right)\right)\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)}\\
\leq &\|g\|_{W^{1,\infty}([0,R])}\delta R + \beta\|g\|_{L^\infty([0,R])}\\
\leq &\varepsilon\|g\|_{W^{1,\infty}([0,R])}.
\end{aligned}$$

In order to estimate the $W^{1,\infty}(D)$-error, we use that $\Phi^{\mathrm{Eucl}}_{d,\delta}$ is 1-Lipschitz continuous. This implies $\left\|\frac{\partial}{\partial x_1}\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)} \leq \|A\|_{2,\mathbb{R}^d}$. It follows that

$$\begin{aligned}
&\left\|\frac{\partial}{\partial x_1}\left(g(\|A\cdot+b\|_{2,\mathbb{R}^d})\right)-\frac{\partial}{\partial x_1}\left(\mathrm{R}\left(\Phi^g_{\beta,R}\right)\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right)\right\|_{L^\infty(D)}\\
\leq&\left\|\left(\frac{\partial}{\partial x}g\right)\left(\|A\cdot+b\|_{2,\mathbb{R}^d}\right)\cdot\left(\frac{\partial}{\partial x_1}\|A\cdot+b\|_{2,\mathbb{R}^d}\right)\right.\\
&\left.-\left(\frac{\partial}{\partial x}g\right)\left(\|A\cdot+b\|_{2,\mathbb{R}^d}\right)\cdot\frac{\partial}{\partial x_1}\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)}\\
&+\left\|\left(\frac{\partial}{\partial x}g\right)\left(\|A\cdot+b\|_{2,\mathbb{R}^d}\right)\cdot\frac{\partial}{\partial x_1}\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right.\\
&\left.-\left(\frac{\partial}{\partial x}\mathrm{R}\left(\Phi^g_{\beta,R}\right)\right)\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\cdot\frac{\partial}{\partial x_1}\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)}\\
\leq&\left\|\left(\frac{\partial}{\partial x}g\right)\left(\|A\cdot+b\|_{2,\mathbb{R}^d}\right)\right\|_{L^\infty(D)}\left|\|\cdot\|_{2,\mathbb{R}^d}-\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)\right|_{W^{1,\infty}(\mathbb{R}^d)}\|A\|_2\\
&+\left\|\left(\frac{\partial}{\partial x}g\right)\left(\|A\cdot+b\|_{2,\mathbb{R}^d}\right)-\left(\frac{\partial}{\partial x}\mathrm{R}\left(\Phi^g_{\beta,R}\right)\right)\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)}\|A\|_2\\
\leq&\|A\|_2\left\|\left(\frac{\partial}{\partial x}g\right)\left(\|A\cdot+b\|_{2,\mathbb{R}^d}\right)\right\|_{L^\infty(D)}\left|\|\cdot\|_{2,\mathbb{R}^d}-\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)\right|_{W^{1,\infty}(\mathbb{R}^d)}\\
&+\|A\|_2\left\|\left(\frac{\partial}{\partial x}g\right)(\|A\cdot+b\|_{2,\mathbb{R}^d})-\left(\frac{\partial}{\partial x}g\right)\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)}\\
&+\|A\|_2\left\|\left(\frac{\partial}{\partial x}g-\frac{\partial}{\partial x}\mathrm{R}\left(\Phi^g_{\beta,R}\right)\right)\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)}\\
\leq&\|A\|_2\|g\|_{W^{1,\infty}([0,R])}\delta+\|A\|_2\|g\|_{W^{2,\infty}([0,R])}\delta R+\|A\|_2\,\beta\,\|g\|_{W^{1,\infty}([0,R])}\\
\leq&\,\varepsilon\|A\|_2\|g\|_{W^{2,\infty}([0,R])}/\sqrt{d},
\end{aligned}$$

and hence

$$\left\|\nabla\left(g(\|A\cdot+b\|_{2,\mathbb{R}^d})\right)-\nabla\left(\mathrm{R}\left(\Phi^g_{\beta,R}\right)\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right)\right\|_{L^\infty(D)}\leq\varepsilon\|A\|_2\|g\|_{W^{2,\infty}([0,R])}.$$

□

6.3 Singular Anisotropic Radial Functions in High Dimension

In this section, we derive a variation of Theorem 6.5, with weaker regularity assumptions on g: we only require $g \in W^{s,\infty}([0,R])$ for some $s > 0$. We show that deep NNs can emulate anisotropic, radial-like functions of d variables with pointwise accuracy $\varepsilon > 0$, with a NN size that scales polynomially in the dimension d. DNNs therefore break the curse of dimensionality in this case.

Theorem 6.7. *Let $d \in \mathbb{N}$, $s > 0$, $R > 0$, $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$, and $D := \{x \in \mathbb{R}^d : \|Ax+b\|_{2,\mathbb{R}^d} \leq R\}$. Moreover, let $g \in W^{s,\infty}([0,R])$ be such that for all $\beta \in (0,1)$ the function g can be approximated by a NN $\Phi^g_{\beta,R}$ such that*

$$\begin{aligned}
\left\|g-\mathrm{R}\left(\Phi^g_{\beta,R}\right)\right\|_{L^\infty([0,R])}&\leq\beta\|g\|_{L^\infty([0,R])},\\
L\left(\Phi^g_{\beta,R}\right)=:L_{\beta,R},\qquad M\left(\Phi^g_{\beta,R}\right)&=:M_{\beta,R}.
\end{aligned}$$

Consider the d-dimensional, anisotropic radial-like function

$$f : D \to \mathbb{R} : x \mapsto g(\|Ax+b\|_{2,\mathbb{R}^d}).$$

Then, for every $\epsilon \in (0,1)$, there exists a NN $\Phi^f_{\varepsilon,D}$ such that

$$\begin{aligned}\left\|f - \mathrm{R}\left(\Phi^f_{\varepsilon,D}\right)\right\|_{L^\infty(D)} &\leq \varepsilon \|g\|_{W^{s,\infty}([0,R])}, \\ L\left(\Phi^f_{\varepsilon,D}\right) &\leq L_{\beta,R} + \log_2(d)\log_2\left(10\pi d R(2/\varepsilon)^{1/s}\right) + 1, \\ M\left(\Phi^f_{\varepsilon,D}\right) &\leq 2M_{\beta,R} + 4d^2 + 64(d-1)\log_2\left(10\pi d R(2/\varepsilon)^{1/s}\right) + 4d.\end{aligned}$$

Proof of Theorem 6.7. We proceed as in the proof of Theorem 6.5. Again, let $\Phi^{A,b}$ be a NN of depth 1 and size at most d^2+d emulating exactly the affine transformation $T: D \to \mathbb{R}^d : x \mapsto Ax+b$. For the approximation of $\|\cdot\|_{2,\mathbb{R}^d}$, we apply Proposition 6.2 with $\delta := (\frac{\varepsilon}{2})^{1/s}R^{-1}$. For the approximation of g, we use the assumption of the theorem with $\beta := \frac{\varepsilon}{2}$. We define

$$\Phi^f_{\varepsilon,D} := \Phi^g_{\beta,R} \odot \Phi^{\mathrm{Eucl}}_{d,\delta} \odot \Phi^{A,b}.$$

The depth and size of $\Phi^f_{\varepsilon,D}$ can be bounded as follows:

$$\begin{aligned}L\left(\Phi^f_{\varepsilon,D}\right) &\leq L\left(\Phi^g_{\beta,R}\right) + L\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right) + L\left(\Phi^{A,b}\right) \\ &\leq L_{\beta,R} + \log_2(d)\log_2\left(10\pi d R(2/\varepsilon)^{1/s}\right) + 1, \\ M\left(\Phi^f_{\varepsilon,D}\right) &\leq 2M\left(\Phi^g_{\beta,R}\right) + 4M\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right) + 4M\left(\Phi^{A,b}\right) \\ &\leq 2M_{\beta,R} + 4\left(16(d-1)\log_2\left(10\pi d R(2/\varepsilon)^{1/s}\right)\right) + 4(d^2+d) \\ &\leq 2M_{\beta,R} + 4d^2 + 64(d-1)\log_2\left(10\pi d R(2/\varepsilon)^{1/s}\right) + 4d.\end{aligned}$$

Using that the value of the $W^{s,\infty}([0,R])$-seminorm equals the C^s-Hölder constant, we obtain that

$$\begin{aligned}&\left\|g(\|A\cdot+b\|_{2,\mathbb{R}^d}) - \mathrm{R}\left(\Phi^g_{\beta,R}\right)\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)} \\ \leq& \left\|g(\|A\cdot+b\|_{2,\mathbb{R}^d}) - g\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)} \\ &+ \left\|\left(g - \mathrm{R}\left(\Phi^g_{\beta,R}\right)\right)\left(\mathrm{R}\left(\Phi^{\mathrm{Eucl}}_{d,\delta}\right)(A\cdot+b)\right)\right\|_{L^\infty(D)} \\ \leq& \|g\|_{W^{s,\infty}([0,R])}(\delta R)^s + \beta\|g\|_{L^\infty([0,R])} \\ \leq& \varepsilon\|g\|_{W^{s,\infty}([0,R])}.\end{aligned}$$

□

Because $H^1((0,R)) \hookrightarrow W^{s,\infty}((0,R))$ for all $s \in [0,\frac{1}{2})$, this means that we can combine Theorem 6.7 with any of the results in Section 5. In particular, with $R=1$ we can apply the results of Section 5.4 concerning functions on $(0,1)$ with a singularity in $x=0$: for all $\beta \in (0,1)$ and all $\delta \geq 1$ it holds that $\mathcal{G}^{2,\delta}_\beta((0,1)) \subset H^1((0,1)) \hookrightarrow W^{s,\infty}((0,R))$, which means that we get exponential convergence of NN approximations of singular anisotropic radial-like functions, such as $x \mapsto \|x\|^s_{2,\mathbb{R}^d}$ for any $s > \frac{1}{2}$: we use that $x \mapsto x^s$ is an element of $\mathcal{G}^{2,1}_\beta((0,1))$ (e.g. for $\beta = \frac{5-2s}{4} \in (0,1)$).

As noted in Remark 6.1, if instead of Proposition 6.2 a NN approximation of $x \mapsto \sum_{i=1}^d x_i^2$ were combined with results from Section 5.4, then only $x \mapsto \|x\|^s_{2,\mathbb{R}^d}$ for $s \geq 1$ could be approximated.

7 Conclusions and Further Directions

We established bounds on approximation rates for the expression of univariate functions belonging to several types of function spaces by deep neural networks. The function spaces studied include, in particular, Sobolev and Besov spaces, as well as, spaces of piecewise analytic and Gevrey-regular functions. We proved that ReLU DNNs achieve for each of these function classes approximation rates which are either identical to or closely match the best available approximation rates by classical approximation schemes with piecewise polynomial spline functions. Notably, DNNs match the rates achieved by both, free-knot ("h-adaptive") and order-adaptive ("hp"-adaptive) approximations. These observations offer a partial explanation for the recent success of numerical solution strategies in using DNNs for the numerical approximation of PDEs as reported, e.g., in [50].

In addition to the univariate results, we demonstrated how these results imply bounds on approximation rates for isotropic and anisotropic radially symmetric functions in high dimension. Employing the closedness under composition of ReLU DNNs, we proved that radial-like functions obtained as compositions of univariate functions with the euclidean norm and, possibly, anisotropic affine changes of co-ordinates can be approximated to arbitrary accuracy with DNNs the size and depth of which scale logarithmically in the accuracy and polynomially in the input dimension. This implies that ReLU DNNs do not suffer from the curse of dimensionality when approximating radial functions.

Finally, we remark that it is certainly feasible to use alternative polynomial bases to the monomial basis and Legendre basis used in Proposition 4.2 and Proposition 4.6. We expect other polynomial bases to yield qualitatively similar results, however, with slight quantitative improvements of approximation rates and/or DNN size and depth bounds. Furthermore, emulating h-adaptive and hp-adaptive approximations by ReLU DNNs in space dimension two and higher dimensions is equally feasible, and builds on present results in the univariate case, combined with DNN tensorization as developed, e.g., in [17]. As the mathematical apparatus characterizing the analytic function classes is somewhat more involved (see, e.g., [45] and the references there), we present these in [35].

Acknowledgements

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

A Proof of Theorem 5.10

We note that [46, 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 ([46, Lemma 3.41])**.** *Let $I = (0,1)$, $\delta \geq 1$, $\beta \in (0,1)$ and $u \in \mathcal{G}^{2,\delta}_\beta(I)$. Let $\sigma \in (0,1)$, $N \in \mathbb{N}$, $\lambda := \sigma^{-1} - 1$ 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 $i \in \{2, \ldots, N\}$.*

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

We will use the following lemma to bound the right-hand side of the inequality in Lemma A.1.

Lemma A.2 ([13, Lemma 4.3.4])**.** *Let $N \in \mathbb{N}$, $\delta \geq 1$, $\alpha > 0$ and $\mu_0 := \max\{1, \alpha e^{1-\delta}\}$. For any $\mu > \mu_0$ let $\boldsymbol{p} = (p_i)_{i=1}^N \subset \mathbb{N}$ be defined by $p_i := \lfloor \mu i^\delta \rfloor$ for all $i \in \{1, \ldots, N\}$. 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 5.10. We use Lemma A.1 with $x_i = \sigma^{N-i}$ and $s_i = i + 1$ for all $i \in \{1, \ldots, N\}$. Because $u \in \mathcal{G}^{2,\delta}_\beta(I)$, it holds that $|u|_{H^{i+2,2}_\beta(I)} \leq C d^i (i!)^\delta$ for all $i \in \{0, \ldots, N-2\}$. With $\alpha := \frac{d\lambda}{2\sigma^{1-\beta}}$, $\mu_0 = \max\left\{1, \frac{d\lambda e^{1-\delta}}{2\sigma^{1-\beta}}\right\}$, and C_* as in Equation (5.14), it follows with Lemma A.2 that there exists a constant $C_7(\sigma,\beta,\delta,\mu,C_*,d) > 0$ such that for all N it holds that

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

This completes the proof. □

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. *Journal of Nonlinear Science*, 29(4):1563, 2019.

[5] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45, 2019.

[6] G. Cheng and V. Shcherbakov. Anisotropic radial basis function methods for continental size ice sheet simulations. *Journal of Computational Physics*, 372:161–177, 2018.

[7] 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.

[8] C. K. Chui, X. Li, and H. N. Mhaskar. Neural networks for localized approximation. *Mathematics of Computation*, 63(208):607–623, 1994.

[9] C. K. Chui, S.-B. Lin, and D.-X. Zhou. Construction of neural networks for realization of localized deep learning. *Frontiers in Applied Mathematics and Statistics*, 4, 2018.

[10] C. K. Chui, S.-B. Lin, and D.-X. Zhou. Deep neural networks for rotation-invariance approximation and learning. *Analysis and Applications*, 17(05), 2019.

[11] C. K. Chui and H. N. Mhaskar. Deep nets for local manifold learning. *Frontiers in Applied Mathematics and Statistics*, 4, 2018.

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

[13] D. Devaud. *hp-approximation of linear parabolic evolution problems in $H^{1/2}$*. PhD thesis, ETH Zürich, 2017. Dissertation 24393.

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

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

[16] 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.

[17] 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.

[18] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi. *Higher transcendental functions. Vol. II.* Robert E. Krieger Publishing Co., Inc., Melbourne, Fla., 1981. Based on notes left by Harry Bateman, Reprint of the 1953 original.

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

[20] 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.

[21] 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.

[22] 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.

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

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

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

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

[27] S.-B. Lin. Generalization and expressivity for deep nets. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5):1392–1406, 2019.

[28] B. McCane and L. Szymanski. Deep radial kernel networks: Approximating radially symmetric functions with deep networks. Technical Report 1703.03470, arXiv, 2017.

[29] B. McCane and L. Szymanski. Efficiency of deep networks for radially symmetric functions. *Neurocomputing*, 313:119 – 124, 2018.

[30] J. Melenk and C. Schwab. *hp* FEM for reaction-diffusion equations. I. Robust exponential convergence. *SIAM J. Numer. Anal.*, 35(4):1520–1557, 1998.

[31] J. Melenk and C. Schwab. Analytic regularity for a singularly perturbed problem. *SIAM J. Math. Anal.*, 30(2):379–400, 1999.

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

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

[34] H. Montanelli and Q. Du. New error bounds for deep ReLU networks using sparse grids. *SIAM J. Math. Data Sci.*, 1(1):78–92, 2019.

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

[36] J. A. A. Opschoor, C. Schwab, and J. Zech. Exponential relu dnn expression of holomorphic maps in high dimension. Technical Report 2019-35, Seminar for Applied Mathematics, ETH Zürich, Switzerland, 2019.

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

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

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

[40] 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.

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

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

[43] D. Rolnick and M. Tegmark. The power of deeper networks for expressing natural functions. In *International Conference on Learning Representations*, 2018.

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

[45] 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.

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

[47] C. Schwab and M. Suri. The *p* and *hp* versions of the finite element method for problems with boundary layers. *Math. Comp.*, 65(216):1403–1429, 1996.

[48] 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.

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

[50] J. Sirignano and K. Spiliopoulos. DGM: A deep learning algorithm for solving partial differential equations. *J. Comput. Phys.*, 375:1339–1364, 2018.

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

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

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

[54] H. Wendland. *Scattered data approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2005.

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

[56] D. Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 639–649. PMLR, 06–09 Jul 2018.