Numerische Mathematik (2024) 156:319–344
https://doi.org/10.1007/s00211-023-01384-6

Numerische Mathematik

Two-layer networks with the $\mathbf{ReLU}^k$ activation function: Barron spaces and derivative approximation

Yuanyuan Li[1] · Shuai Lu[1] · Peter Mathé[2] · Sergei V. Pereverzev[3]

Received: 20 January 2023 / Revised: 11 August 2023 / Accepted: 30 October 2023 /
Published online: 23 November 2023

Abstract
We investigate the use of two-layer networks with the rectified power unit, which is called the ReLU^k activation function, for function and derivative approximation. By extending and calibrating the corresponding Barron space, we show that two-layer networks with the ReLU^k activation function are well-designed to simultaneously approximate an unknown function and its derivatives. When the measurement is noisy, we propose a Tikhonov type regularization method, and provide error bounds when the regularization parameter is chosen appropriately. Several numerical examples support the efficiency of the proposed approach.

Mathematics Subject Classification 41A25 · 42B35 · 42C40 · 65D15 · 65D25

1 Introduction

We investigate a classical problem of approximating a smooth function and its derivatives defined in a bounded domain $\Omega \subset \mathbb{R}^d$ when only noisy measurements

✉ Shuai Lu
slu@fudan.edu.cn

Yuanyuan Li
20110180025@fudan.edu.cn

Peter Mathé
peter.mathe@wias-berlin.de

Sergei V. Pereverzev
sergei.pereverzyev@oeaw.ac.at

1 School of Mathematical Sciences, Fudan University, No. 220 Handan Road, Shanghai 200433, China

2 Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, 10117 Berlin, Germany

3 Johann Radon Institute for Computational and Applied Mathematics, Altenbergerstrasse 69, 4040 Linz, Austria

$f^\delta \in L^2(\Omega)$ with an error bound

$$\|f - f^\delta\|_{L^2(\Omega)} \leq \delta \tag{1}$$

are known. As it is customary in regularization theory we assume that in principle we do have access to the noisy function f^δ as an element of $L^2(\Omega)$. This allows us to find reconstructions given in terms of *projection schemes*, see e.g. [10, 17]. In this study the sought for approximating function is given in the form of a two-layer network

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} a_i \sigma(b_i \cdot x + c_i), \quad a_i, c_i \in \mathbb{R}, \quad b_i \in \mathbb{R}^d, \tag{2}$$

where σ is the activation function, i.e. a continuous function of sigmoidal or tanh form, the rectified linear unit (ReLU) etc. Here $x \in \Omega$ denotes the input data and $(a_i, b_i, c_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$, denote the network coefficients. We focus on the ReLU^k activation function $\sigma(z) = [\max(0, z)]^k$ for an integer $k \geq 1$, and we also use $(z)_+^k := [\max(0, z)]^k$ for brevity.

When accessing the error $\|f - f_n\|_{L^2(\Omega)}$ of a two-layer network f_n we have two parameters, the (known) noise level δ, and the number n of neurons. As in regularization theory we are interested in error bounds under the asymptotics $\delta \to 0$, it is natural to incorporate the knowledge of the noise level into the network design by choosing the number of neurons n in such a way that the corresponding approximation rate becomes of the order of the noise level. Because there would be no gain in accuracy from the use of networks with a larger number of neurons, while the networks with a smaller number of neurons may not be able to utilize the whole information encoded in the noisy data. Details will be given in Sect. 3.

To approximate a function by its measurement has been one of the long-lasting tasks in numerical analysis, approximation theory, and plays an important role in supervised learning. Classical approaches use kernel based methods [2], spline based methods [25], piecewise polynomials and wavelets based methods [9], neural networks [11, 29] etc. In view of the approximation by neural networks, the seminal paper [4] develops a universal approximation theorem on function approximation by feedforward neural networks with sigmoidal activation functions. Systematic investigation on universality properties of two-layer networks with the ReLU activation function have been carried out in [19, 22] under different *a priori* knowledge of the unknown target function. Recent work in [30] verifies the universality of deep convolutional neural network with the ReLU activation function. There are only a few studies where the authors consider neural networks with ReLU^k activation function, and we mention [16] where the universality is investigated by classical polynomial approximation theory. Also, the recent study [1] deals with ReLU^k networks, called 'rectified power units' (RePU) there. It is worth to mention that the above authors do not consider noisy data.

At the same time, the approximation of derivatives of an unknown function is intrinsically ill-posed [10]. Noise in the measurement, as in (1), may yield unstable reconstruction unless some regularization schemes are implemented [10, 17]. Tikhonov regularization is viable, and can easily be adopted to our problem [12,

18]. Denote X_n be a generic solution set. Tikhonov regularization, with a smoothing penalty term, seeks for $f_n \in X_n$ as a minimizer of the following functional,

$$f \longrightarrow \|f - f^\delta\|^2_{L^2(\Omega)} + \lambda \|f - f_*\|^2_{H^m(\Omega)}, \tag{3}$$

provided that the minimum exists. This is known for weakly sequentially closed sets X_n. Above, the penalty term is the Sobolev space norm with an integer $m \geq 2$, λ is the regularization parameter and f_* is an initial guess. Another form of Tikhonov regularization is used in the case when X_n is not weakly closed. By choosing an arbitrary constant $\epsilon > 0$, we seek for $f_n \in X_n$ which obeys

$$\begin{aligned} &\|f_n - f^\delta\|^2_{L^2(\Omega)} + \lambda \|f_n - f_*\|^2_{H^m(\Omega)} \\ &\quad \leq \|g - f^\delta\|^2_{L^2(\Omega)} + \lambda \|g - f_*\|^2_{H^m(\Omega)} + \epsilon, \end{aligned} \tag{4}$$

for all $g \in X_n$. The arbitrary constant $\epsilon > 0$ in the functional is necessary to guarantee the existence of such f_n. For simplicity's sake, we skip a detailed description on the regularization scheme but refer to [10, 12, 18, 26] for abundant discussion and the realization on numerical differentiation, respectively. To fit our proposed approximation problem by two-layer networks, we will specify the penalty in Tikhonov regularization as e.g.

$$\min_{(a_i,b_i,c_i)\in(\mathbb{R}\times\mathbb{R}^d\times\mathbb{R})^n} \left\| \frac{1}{n} \sum_{i=1}^{n} a_i \sigma(b_i \cdot x + c_i) - f^\delta(x) \right\|^2_{L^2(\Omega)} + \lambda \mathcal{P}(a_i, b_i, c_i), \tag{5}$$

where $\mathcal{P}(a_i, b_i, c_i)$ is a penalty term associated to two-layer networks. This penalty shall fit the explicit form of the networks and avoid the calculation of the standard Sobolev norm, which is hardly accessible. Various forms of the Tikhonov regularization associating two-layer networks have been proposed and implemented for different inverse problems of derivative approximation in [5, 6, 21], or inverse boundary value problems in [3]. Nevertheless, little is known for its regularizing properties and theoretical error bounds. To the best of our knowledge, only the pioneering work in [6] considers these issues for activation functions in (2) that are smooth enough and when the unknown solution belongs to certain Sobolev spaces.

Here we are particularly interested in the simultaneous approximation of a function and its derivatives by two-layer networks with the ReLU^k activation function. For the standard ReLU activation function in (2) with $\sigma = \max(0, x)$ the second-order weak derivative of f_n does not exist as a function while it is a distribution, and we cannot use the approximating f_n in order to mimic high-order derivatives. As we shall see below, by choosing σ in (2) as ReLU^k activation functions, a simultaneous approximation of a function and its derivative can be achieved.

To this end, Barron spaces and the related Barron norms, induced by the ReLU^k activation function shall be carefully examined in the forthcoming Sect. 2. In Sect. 3 we show that the two-layer networks with the ReLU^k activation function are well-designed to simultaneously approximate an unknown function and its derivatives. The

major contribution in this study concerns the stable reconstruction by neural networks under noisy data. We propose a Tikhonov regularization with weight decay as in (5), where the penalty term $\mathcal{P}(a_i, b_i, c_i)$ will be specified there. Finally, several numerical examples in Sect. 4 shed light on the efficiency of the proposed approach.

2 Extended Barron spaces by ReLU^k activation functions

To carry out our theoretical analysis, we need to extend the Barron norm and Barron space induced by the ReLU activation function in [19] towards ReLU^k activation functions. Some connection between the Barron spaces and Sobolev spaces will also be explored in this section, which allows us to further consider the derivative approximation of an unknown function.

Remark 1 In the original paper [4], Barron focused on function approximation by two-layer networks with sigmoidal activation functions. The function to be approximated there shares a smoothness assumption in terms of its Fourier representation, c.f. [14]. Accordingly, these so-called spectral Barron spaces have been defined in [22, 28] based on the Fourier coefficients of an unknown function. Function and derivative approximation by two-layer networks with the ReLU^k activation functions and noise free measurement have been carried out there for the Barron spectral space. The Barron space that we have in mind may be viewed as a space of *infinite-width neural networks*, or "one-hidden-layer neural network with infinitely many units", as in the early study [15].

2.1 Barron spaces and Barron norms

The Barron spaces and norms that we consider here extend those in [19] which are designed specifically for the ReLU activation function. It has been observed there that this Barron space is well suited for two-layer networks and optimal approximation can be derived for functions in that space. Our focus is on derivative approximation. Therefore, it is a straightforward step to extend the definition of the Barron spaces and the Barron norms towards ReLU^k activation functions. To this end we consider functions f defined in a bounded domain $\Omega \subset \mathbb{R}^d$, satisfying the following representation

$$f(x) = \int_P a\,(b \cdot x + c)_+^k\, \rho(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c), \tag{6}$$

where $P = \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ and ρ is a probability distribution on (P, Σ_P) with Σ_P being a Borel σ-algebra on P. The above form (6) can be viewed as the continuous version of the two-layer network (2).

Every network with finitely many units, as in (2), allows for a representation (6) for the probability ρ_n, uniform on the point set $(a_i, b_i, c_i)_{i=1}^n$, a fact which shall be used below.

We fix the integer value k of a ReLU^k activation function. For $1 \le p < \infty$, we define a value associated to functions of the above representation (6), with given

probability distribution ρ by

$$\|f\|_{B^k_{p,\rho}} := \left(\int_P |a|^p \left(\|b\|_1 + |c|\right)^{kp} \rho(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c)\right)^{\frac{1}{p}}.$$

Denote α be a multi-index, $|\alpha| = \sum_{i=1}^d |\alpha_i|, b^\alpha = \prod_{i=1}^d b_i^{\alpha_i}$. Notice that for functions from (6) we have that

$$\begin{aligned}\partial^\alpha f(x) &= \int_P \frac{k!}{(k-|\alpha|)!} a b^\alpha \left(b \cdot x + c\right)_+^{k-|\alpha|} \rho(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c) \\ &= \int_P a \left(b \cdot x + c\right)_+^{k-|\alpha|} \rho_\alpha(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c),\end{aligned}$$

where ρ_α is a probability distribution on P and

$$\rho_\alpha(A) := \rho\left(\left\{(a,b,c) : \left(\frac{k!}{(k-|\alpha|)!} a b^\alpha, b, c\right) \in A\right\}\right). \tag{7}$$

In principle, we can view ρ_α as a measure induced by a continuous map $(a,b,c) \mapsto \left(\frac{k!}{(k-|\alpha|)!} a b^\alpha, b, c\right)$.

For any integer $m \in \{0, 1, 2, \ldots, k\}$ and any probability measures ρ with corresponding probabilities ρ_α, $|\alpha| \le m$, as in (7), we define

$$\|f\|_{B^{k,m}_{p,\rho}} := \left(\sum_{|\alpha| \le m} \left\|\partial^\alpha f\right\|^p_{B^{k-|\alpha|}_{p,\rho_\alpha}}\right)^{\frac{1}{p}}.$$

The extended Barron norm of the function f in (6) is then formally defined by

$$\|f\|_{B^{k,m}_p} := \inf_\rho \|f\|_{B^{k,m}_{p,\rho}} = \inf_\rho \left(\sum_{|\alpha| \le m} \left\|\partial^\alpha f\right\|^p_{B^{k-|\alpha|}_{p,\rho_\alpha}}\right)^{\frac{1}{p}}, \tag{8}$$

where the infimum is taken over all probability measures ρ which allow for a representation (6).

For $p = \infty$ we analogously define

$$\begin{aligned}\|f\|_{B^k_{\infty,\rho}} &:= \operatorname*{esssup}_{(a,b,c)\sim\rho} |a| \left(\|b\|_1 + |c|\right)^k, \\ \|f\|_{B^{k,m}_{\infty,\rho}} &:= \max_{|\alpha| \le m} \|\partial^\alpha f\|_{B^{k-|\alpha|}_{\infty,\rho_\alpha}},\end{aligned}$$

where esssup means essential supremum and the extended Barron norm is

$$\|f\|_{B^{k,m}_\infty} := \inf_\rho \|f\|_{B^{k,m}_{\infty,\rho}} = \inf_\rho \max_{|\alpha| \le m} \|\partial^\alpha f\|_{B^{k-|\alpha|}_{\infty,\rho_\alpha}}. \tag{9}$$

We thus define

Definition 1 (Extended Barron space, $B_p^{k,m}$) Given any integer $k \geq 0$ and $0 \leq m \leq k$, and a real value $1 \leq p \leq \infty$, the set of functions of the form (6) with finite Barron norm (8) for $1 \leq p < \infty$ or (9) for $p = \infty$, respectively, is called the extended Barron space denoted by $B_p^{k,m}$.

We shall emphasize that the above notion as the extended Barron norm is not rigorous; we can only verify it to be a norm for $p = 1$, as shown below. It is not clear whether a similar result holds true for $\|\cdot\|_{B_p^{k,m}}$ whenever $p > 1$.

Lemma 1 *The set $B_1^{k,m}$ equipped with the functional $f \mapsto \|f\|_{B_1^{k,m}}$ is a normed space.*

Proof Clearly, it suffices to check the validity of the triangle inequality

$$\|f_1 + f_2\|_{B_1^{k,m}} \leq \|f_1\|_{B_1^{k,m}} + \|f_2\|_{B_1^{k,m}}.$$

Thus, we let

$$f_i = \int_P a\,(b \cdot x + c)_+^k\, \rho_i\,(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c)\,, \quad i = 1, 2.$$

Consider the related probability distribution $\tilde{\rho}$ given as

$$\tilde{\rho}(A) := \frac{1}{2}\,(\rho_1(A) + \rho_2(A))\,, \quad A \in \Sigma_P.$$

Then

$$\begin{aligned} f_1 + f_2 &= \int_P 2a\,(b \cdot x + c)_+^k\, \tilde{\rho}(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c) \\ &= \int_P a\,(b \cdot x + c)_+^k\, \rho_3\,(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c)\,, \end{aligned}$$

where

$$\rho_3\,(A) = \tilde{\rho}\,(\{(a, b, c) : (2a, b, c) \in A\})\,.$$

We thus obtain

$$\begin{aligned} \|f_1 + f_2\|_{B_{p,\rho_3}^k} &= \left(\int_P |2a|^p\,(\|b\|_1 + |c|)^{kp}\,\tilde{\rho}(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c)\right)^{\frac{1}{p}} \\ &= 2\left(\int_P |a|^p\,(\|b\|_1 + |c|)^{kp} \right. \\ &\quad \left. \times \left(\frac{1}{2}\rho_1(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c) + \frac{1}{2}\rho_2(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c)\right)\right)^{\frac{1}{p}}. \end{aligned}$$

In case that $p = 1$, we have

$$\|f_1 + f_2\|_{B^k_{1,\rho_3}} = \|f_1\|_{B^k_{1,\rho_1}} + \|f_2\|_{B^k_{1,\rho_2}},$$

or consequently

$$\|f_1 + f_2\|_{B^{k,0}_1} \le \|f_1\|_{B^{k,0}_1} + \|f_2\|_{B^{k,0}_1}.$$

Similarly, we can prove

$$\|f_1 + f_2\|_{B^{k,m}_1} \le \|f_1\|_{B^{k,m}_1} + \|f_2\|_{B^{k,m}_1},$$

and the proof is complete. □

For sake of simplicity, we still call $\|\cdot\|_{B^{k,m}_p}$ the extended Barron norm for a general p. For technical reasons it will be convenient for us to have another representation of probabilities ρ in (6), namely restricted to the unit sphere $\mathbb{S}^d := \left\{(b, c) \in \mathbb{R}^d \times \mathbb{R} : \|b\|_1 + |c| = 1\right\}$. The following lemma, inspired by [19, Prop. 1], uses the key property $(au)^k_+ = |a|^k (u)^k_+,\ u \in \mathbb{R}$, of the ReLU^k activation function.

Lemma 2 *Suppose that a function $f \in B^{k,0}_1$ has a representation (6) for given a probability ρ. Let*

$$c_{k,\rho}(f) := \|f\|_{B^k_{1,\rho}} = \mathbb{E}_\rho \left[|a| \left(\|b\|_1 + |c|\right)^k\right].$$

Then there exists a probability $\bar{\rho}$ on $\{-1, 1\} \times \mathbb{S}^d$ such that

$$f(x) = c_{k,\rho}(f) \int_{\{-1,1\}\times\mathbb{S}^d} \bar{a} \left(\bar{b} \cdot x + \bar{c}\right)^k_+ \bar{\rho} \left(\mathrm{d}\bar{a}, \mathrm{d}\bar{b}, \mathrm{d}\bar{c}\right). \tag{10}$$

Proof The lemma holds true when $c_{k,\rho}(f) = 0$. Without loss of generality, we can thus assume that $c_{k,\rho}(f) > 0$, and that neither $|a|$ nor $(\|b\|_1 + |c|)^k$ are zero. Using the above feature of ReLU^k, we have

$$\begin{aligned} f(x) &= \int_P a \left(b \cdot x + c\right)^k_+ \rho \left(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c\right) \\ &= \int_P a \left(\|b\|_1 + |c|\right)^k \left(\frac{b}{\|b\|_1 + |c|} \cdot x + \frac{c}{\|b\|_1 + |c|}\right)^k_+ \rho \left(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c\right). \end{aligned}$$

Define $\mathbb{S}^d := \left\{(b, c) \in \mathbb{R}^d \times \mathbb{R} : \|b\|_1 + |c| = 1\right\}$, new parameters

$$\bar{a} := \frac{a}{|a|}, \quad \left(\bar{b}, \bar{c}\right) := \left(\frac{b}{\|b\|_1 + |c|}, \frac{c}{\|b\|_1 + |c|}\right),$$

and a new probability distribution $\bar{\rho}$ on $\{-1, 1\} \times \mathbb{S}^d$ given as

$$\bar{\rho}(A) := \frac{1}{c_{k,\rho}(f)} \int_{\{(a,b,c):(\bar{a},\bar{b},\bar{c})\in A\}} |a| \left(\|b\|_1 + |c|\right)^k \rho \left(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c\right),$$

for Borel sets A in $\{-1, 1\} \times \mathbb{S}^d$. For this newly defined probability $\bar{\rho}$ we then have the representation (10). □

The following consequence is important for our study.

Corollary 3 *Let ρ be a representing probability for f as in (6) with a constant $c_{k,\rho}(f)$. Then we have for all $1 \le p \le \infty$ that*

$$\|f\|_{B_p^{k,0}} \le c_{k,\rho}(f),$$

and consequently $\|f\|_{B_p^{k,0}} \le \|f\|_{B_1^{k,0}}$ for all $1 \le p \le \infty$.

Proof For a representing probability ρ of f we turn to its representation form (10). By Lemma 2 we could derive

$$\begin{aligned}\|f\|_{B_p^{k,0}} &\le c_{k,\rho}(f) \left(\int_{\{-1,1\}\times\mathbb{S}^d} |\bar{a}|^p \left(\|\bar{b}\|_1 + |\bar{c}|\right)^{pk} \bar{\rho}\left(\mathrm{d}\bar{a}, \mathrm{d}\bar{b}, \mathrm{d}\bar{c}\right)\right)^{1/p} \\ &\le c_{k,\rho}(f),\end{aligned}$$

which proves the first assertion. Taking the infimum of all representing ρ we see the second assertion. □

Clearly, by the definition of the norms we immediately observe that for all $0 \le m_1 < m_2 \le k$ and a suitable f there holds $\|f\|_{B_p^{k,m_1}} \le \|f\|_{B_p^{k,m_2}}$, and hence the space $B_p^{k,m_2} \subset B_p^{k,m_1}$ is continuously embedded. Then, having fixed integers m, k such that $m \le k$, as a consequence of Corollary 3, we see that the spaces $B_p^{k,m}$ for $1 \le p \le \infty$ coincide.

Corollary 4 *There holds $B_1^{k,0} \subset B_p^{k,m}$ for all $0 \le m \le k$ given a fixed integer k and a real number $p > 1$.*

Proof Based on (10), the derivatives of f can be represented by

$$\partial^\alpha f(x) = c_{k,\rho}(f) \int_{\{-1,1\}\times\mathbb{S}^d} \frac{k!}{(k-|\alpha|)!} \bar{a}\bar{b}^\alpha \left(\bar{b}\cdot x + \bar{c}\right)_+^{k-|\alpha|} \bar{\rho}\left(\mathrm{d}\bar{a}, \mathrm{d}\bar{b}, \mathrm{d}\bar{c}\right),$$

where we are allowed to bound

$$\begin{aligned}\|f\|_{B_p^{k,m}} &\leq \Bigg(\sum_{|\alpha|\leq m} \int_{\{-1,1\}\times\mathbb{S}^d} \Big(\frac{k!}{(k-|\alpha|)!} c_{k,\rho}(f)\, |\bar{a}|\, \left|\bar{b}^\alpha\right| \\ &\qquad \left(\|\bar{b}\|_1 + |c|\right)^{k-|\alpha|} \Big)^p \rho\left(\mathrm{d}\bar{a}, \mathrm{d}\bar{b}, \mathrm{d}\bar{c}\right) \Bigg)^{\frac{1}{p}} \\ &\leq \left(\sum_{|\alpha|\leq m} \left(\frac{k!}{(k-|\alpha|)!} \right)^p \right)^{\frac{1}{p}} c_{k,\rho}(f).\end{aligned}$$

Thus,

$$\|f\|_{B_p^{k,m}} \leq \left(\sum_{|\alpha|\leq m} \left(\frac{k!}{(k-|\alpha|)!} \right)^p \right)^{\frac{1}{p}} \|f\|_{B_1^{k,0}}$$

which ends the proof. □

The following corollary links different extended Barron spaces together.

Corollary 5 *There holds* $\|f\|_{B_1^{k,0}} \leq \|f\|_{B_p^{k,0}}$ *for all fixed integer* k *such that* $B_p^{k,0} \subset B_1^{k,0}$.

Proof Based on (6),

$$\begin{aligned}\|f\|_{B_{1,\rho}^{k,0}} &= \int_P |a|\left(\|b\|_1 + |c|\right)^k \rho\left(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c\right) \\ &\leq \left(\int_P \left(|a|\left(\|b\|_1 + |c|\right)^k \right)^p \rho\left(\mathrm{d}a, \mathrm{d}b, \mathrm{d}c\right) \right)^{\frac{1}{p}} = \|f\|_{B_{p,\rho}^{k,0}}.\end{aligned}$$

Thus, $\|f\|_{B_1^{k,0}} \leq \|f\|_{B_p^{k,0}}$ holds true by taking infimum from both sides. □

We conclude the above discussion as follows.

Theorem 6 *The space* $B_1^k := B_1^{k,0}$ *is a normed space. For all* $0 \leq m \leq k$ *and* $1 \leq p \leq \infty$ *we have that* $B_p^{k,m} = B_1^k$ *(as sets).*

Proof First, by Lemma 1 the space B_1^k is normed. Furthermore, by combining Corollaries 3–5 we can find that $B_1^{k,0} \subset B_p^{k,m} \subset B_p^{k,0} \subset B_1^{k,0}$, which completes the proof. □

For $k = 1$, the extended Barron space is just the Barron space in [19], and it has been proven in [24] that such a Barron space is equivalent to a convex hulls space $\mathscr{K}_1(\mathbb{P}_1^d)$ defined in [22]. This equivalence can be extended to the general case of $k \geq 1$ since Ω is a bounded domain. The relation between the Barron spaces (with $k = 1$) and other classical spaces is discussed for instance in [7, 27].

2.2 Approximation properties of extended Barron spaces

As we have highlighted in the introduction, our aim is to simultaneously approximate an unknown function and its derivatives. For this it is necessary to explore the connection between the Barron spaces as introduced before, and the standard Sobolev spaces. The following theorem shows that for any fixed k the Barron space B_1^k is indeed a subspace of a related standard Sobolev spaces $H^k(\Omega)$.

Theorem 7 *Given any fixed integer k, there holds $B_1^k \subset H^k(\Omega)$. Moreover, there is a constant $C(\Omega, d, k)$ (depending on Ω, d and k) such that for all $f \in B_1^k$ there holds*

$$\|f\|_{H^m(\Omega)} \leq C(\Omega, d, k)\|f\|_{B_1^k}, \quad 0 \leq m \leq k.$$

Proof We only need to prove the case when $m = k$ since $\|f\|_{H^m} \leq \|f\|_{H^k}$ when $m \leq k$. Given any $f \in B_1^k$, by Lemma 2 there is a probability $\bar{\rho}$ such that (10) holds. Then, for all $|\alpha| \leq k$, the weak derivative of f can be represented by

$$\partial^\alpha f(x) = c_{k,\rho}(f) \int_{\{-1,1\}\times\mathbb{S}^d} \frac{k!}{(k-|\alpha|)!} \bar{a}\bar{b}^\alpha \left(\bar{b}\cdot x + \bar{c}\right)_+^{k-|\alpha|} \bar{\rho}\left(\mathrm{d}\bar{a}, \mathrm{d}\bar{b}, \mathrm{d}\bar{c}\right).$$

We first bound

$$\begin{aligned}
&\left(\int_\Omega \left(\int_{\{-1,1\}\times\mathbb{S}^d} \bar{a}\bar{b}^\alpha \left(\bar{b}\cdot x + \bar{c}\right)_+^{k-|\alpha|} \, d\bar{\rho}\right)^2 \mathrm{d}x\right)^{1/2} \\
&\leq \left(\int_\Omega \left(\int_{\{-1,1\}\times\mathbb{S}^d} \left|\bar{a}\bar{b}^\alpha \left(\bar{b}\cdot x + \bar{c}\right)_+^{k-|\alpha|}\right| \, d\bar{\rho}\right)^2 dx\right)^{1/2} \\
&\leq \int_{\{-1,1\}\times\mathbb{S}^d} |\bar{b}^\alpha| \left(\int_\Omega \left|\left(\bar{b}\cdot x + \bar{c}\right)_+^{k-|\alpha|}\right|^2 \mathrm{d}x\right)^{1/2} d\bar{\rho}.
\end{aligned}$$

Clearly we have that

$$\begin{aligned}
\left|\bar{b}\cdot x + \bar{c}\right| &\leq \|\bar{b}\|_1 \|x\|_\infty + |\bar{c}| \\
&\leq \max\{1, \|x\|_\infty\}\left(\|\bar{b}\|_1 + |c|\right) = \max\{1, \|x\|_\infty\},
\end{aligned}$$

by construction of $\bar{b}$ and $\bar{c}$. Therefore,

$$\begin{aligned}
\left(\int_\Omega \left|\left(\bar{b}\cdot x + \bar{c}\right)_+^{k-|\alpha|}\right|^2 \mathrm{d}x\right)^{1/2} &\leq \left(\int_\Omega \max\{1, \|x\|_\infty\}^{2(k-|\alpha|)} \, \mathrm{d}x\right)^{1/2} \\
&\leq \left(\int_\Omega \max\{1, \|x\|_\infty\}^{2k} \, \mathrm{d}x\right)^{1/2} \\
&=: C_1(\Omega, d, k).
\end{aligned}$$

Also, since each $\left|\bar{b}_i\right| \leq 1$ we find that $\left|\bar{b}^\alpha\right| \leq 1$, which yields that

$$\left(\int_\Omega \left(\int_{\{-1,1\}\times\mathbb{S}^d} \bar{a}\bar{b}^\alpha \left(\bar{b}\cdot x+\bar{c}\right)_+^{k-|\alpha|}\, d\bar{\rho}\right)^2 \mathrm{d}x\right)^{1/2} \leq C_1(\Omega,d,k).$$

This results in

$$\|\partial^\alpha f\|_{L^2(\Omega)} \leq c_{k,\rho}(f)\frac{k!}{(k-|\alpha|)!}C_1(\Omega,d,k),$$

and finally that there is some constant $C(\Omega,d,k)$ for which

$$\|f\|_{H^k(\Omega)} \leq C(\Omega,d,k)c_{k,\rho}(f).$$

Since this holds true for each representing measure ρ in (6), we can take the infimum on both sides. Using that $\|f\|_{B_1^k} := \inf_\rho c_{k,\rho}(f)$ allows us to complete the proof. □

Remark 2 As the above Theorem 7 shows, any function in the Barron space B_1^k automatically belongs to the Sobolev space $H^k(\Omega)$. In order to stably approximate a derivative of an unknown function, we suggest considering an index $k \geq 2$ where the chosen ReLU^k activation function can be viewed as the a priori information. Such smoothness assumption is usually included in the regularization schemes in form of a Sobolev space norm, for instance [12, 26].

Functions f which allow for a representation (6) are particularly suited for approximation by two-layer networks (2). The following theorem has its origin in [14]. As stated here it is an extension of [19, Thm. 4] by replacing the ReLU activation function by the ReLU^k activation functions.

Theorem 8 *Let $f \in B_1^k$, and denote ρ be its representing probability which satisfies $c_{k,\rho}(f) \leq (1+\epsilon)\|f\|_{B_1^k}$ for some small $\epsilon > 0$. For all $0 \leq m \leq k$, and any $n \in \mathbb{N}$, there exist a probability ρ_n and a two-layer network f_n, with $\|f_n\|_{B_1^k} \leq c_{k,\rho_n}(f_n) \leq c_{k,\rho}(f)$ such that*

$$\|f-f_n\|_{H^m(\Omega)} \leq \frac{C(\Omega,d,k,m)}{\sqrt{n}}\|f\|_{B_1^k}.$$

Proof We define probability distributions ρ and $\bar{\rho}$ in the same way as in Lemma 2, such that (10) holds with a constant $c_{k,\rho}(f)$.

Assume that $\left\{\left(\bar{a}_i,\bar{b}_i,\bar{c}_i\right)\right\}_{i=1}^n$ are n *i.i.d.* random vectors according to the probability distribution $\bar{\rho}$, with corresponding functions

$$g_i(x) := c_{k,\rho}(f)\bar{a}_i\left(\bar{b}_i\cdot x+\bar{c}_i\right)_+^k, \quad i=1,\ldots,n.$$

By construction, each g_i is an unbiased estimate for f, and this extends to derivatives $\partial^\alpha g_i$ and $\partial^\alpha f$, by linearity for $|\alpha| \leq k$. We emphasize that $\|g_i\|_{B_1^k} \leq c_{k,\rho}(f)$, $i=1,\ldots,n$, because $|\bar{a}_i|\left(\|\bar{b}_i\|_1+|\bar{c}_i|\right)^k \leq 1$.

Let

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} g_i(x). \tag{11}$$

Recall that the Sobolev space $H^m(\Omega)$ constitutes a Hilbert space, and the independence of the functions g_i yields their orthogonality (in $H^m(\Omega)$). Therefore, we can bound the mean squared error as

$$\begin{aligned}
\mathbb{E}_{\bar\rho}\|f - f_n\|^2_{H^m(\Omega)} &= \mathbb{E}_{\bar\rho}\|\mathbb{E}_{\bar\rho} f_n - f_n\|^2_{H^m(\Omega)} \\
&= \mathbb{E}_{\bar\rho}\|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\bar\rho} g_i - g_i\right)\|^2_{H^m(\Omega)} \qquad (12)\\
&= \frac{1}{n}\mathbb{E}_{\bar\rho}\|\left(\mathbb{E}_{\bar\rho} g_1 - g_1\right)\|^2_{H^m(\Omega)} \qquad (13)\\
&\le \frac{1}{n}\mathbb{E}_{\bar\rho}\|g_1\|^2_{H^m(\Omega)} \\
&\le \frac{C^2(\Omega,d,k)}{n}\mathbb{E}_{\bar\rho}\|g_1\|^2_{B_1^k} \\
&\le \frac{C^2(\Omega,d,k)}{n} c^2_{k,\rho}(f), \qquad (14)
\end{aligned}$$

such that we obtain

$$\left(\mathbb{E}_{\bar\rho}\|f - f_n\|^2_{H^m(\Omega)}\right)^{1/2} \le \frac{C(\Omega,d,k)}{\sqrt{n}} c_{k,\rho}(f).$$

In the above analysis we used that the variance of the sum equals to the sum of the variances to turn from (12), that the variance is less than or equal to the norm to turn from (13), and we applied Theorem 7 in (14). Consequently, there is a realization $(a_i, b_i, c_i)_{i=1}^n$ with $\|f - f_n\|_{H^m(\Omega)} \le \frac{C(\Omega,d,k,m)}{\sqrt{n}}\|f\|_{B_1^k}$.

Finally, let ρ_n denote the uniform distribution on the above sample $\left\{\left(\bar a_i, \bar b_i, \bar c_i\right)\right\}$. Then we see that

$$f_n = \int c_{k,\rho}(f)\bar a \left(\bar b x + \bar c\right)_+^k \rho_n(d\bar a, d\bar b, d\bar c),$$

and therefore we find

$$c_{k,\rho_n}(f_n) \le c_{k,\rho}(f),$$

which completes the proof. □

Beyond this, by choosing more but finite neurons in the two-layer network, we can obtain an improved approximating rate when $m < k$ below. The proof of following theorem use stratified sampling. This was first used in [14], our approach is inspired by [28].

Theorem 9 *Let $f \in B_1^k$ and denote ρ by its representing probability which satisfies $c_{k,\rho}(f) \leq (1+\epsilon)\|f\|_{B_1^k}$ for some small $\epsilon > 0$. Then, for all $n \in \mathbb{N}$, there exist an integer $N \in [n, 2n]$, a probability ρ_N and a two-layer network f_N with $\|f_N\|_{B_1^k} \leq c_{k,\rho_N}(f_N) \leq c_{k,\rho}(f)$, such that*

$$\|f - f_N\|_{H^m(\Omega)} \leq C(\Omega, d, k, m)\|f\|_{B_1^k} N^{-\frac{1}{2}-\frac{1}{d}} \tag{15}$$

with $0 \leq m < k$.

Proof We define probability distributions ρ and $\bar{\rho}$ in the same way as in Lemma 2, such that (10) holds with a constant $c_{k,\rho}(f)$. Define a finite covering $\{P_1, P_2, \ldots, P_n\}$ of $\{-1, 1\} \times \mathbb{S}^d$ such that any $(\bar{a}, \bar{b}, \bar{c}), (\bar{a}', \bar{b}', \bar{c}') \in P_s$ and there holds

$$\bar{a} = \bar{a}', \quad \|\bar{b} - \bar{b}'\|_1 + \left|\bar{c} - \bar{c}'\right| \leq C_2 n^{-\frac{1}{d}},$$

where C_2 is a constant depending on d, for all $s \in \{1, 2, \ldots, n\}$.

Let $\mu_s = (\bar{\rho}(P_s))^{-1} \bar{\rho}\,|_{P_s}$, i.e. μ_s is the normalized probability distribution of $\bar{\rho}$ restricted on P_s. Then, (10) can be rewritten as

$$f(x) = c_{k,\rho}(f) \sum_{s=1}^{n} \bar{\rho}(P_s)\, \mathbb{E}_{\mu_s}\left[\bar{a}\left(\bar{b}\cdot x + \bar{c}\right)_+^k\right],$$

and

$$\partial^\alpha f(x) = c_{k,\rho}(f) \sum_{s=1}^{n} \bar{\rho}(P_s)\, \mathbb{E}_{\mu_s}\left[\frac{k!}{(k-|\alpha|)!}\bar{a}\bar{b}^\alpha\left(\bar{b}\cdot x + \bar{c}\right)_+^{k-|\alpha|}\right],$$

for all $|\alpha| \leq m$.

Now we are going to construct an auxiliary random function. Let $N_s = \lceil n\bar{\rho}(P_2)\rceil$ and $N = \sum_{s=1}^n N_s$. It is easy to see that $n \leq N \leq 2n$. We choose N independent random vectors $\left\{\left(\bar{a}_{s,t}, \bar{b}_{s,t}, \bar{c}_{s,t}\right)\right\}_{t=1,2,\ldots,N_s;s=1,2,\ldots,n}$ such that $\left(\bar{a}_{s,t}, \bar{b}_{s,t}, \bar{c}_{s,t}\right)$ are sampled from μ_s. Let

$$f_N(x) = c_{k,\rho}(f) \sum_{s=1}^{n} \bar{\rho}(P_s)\, \frac{1}{N_s}\sum_{t=1}^{N_s} \bar{a}_{s,t}\left(\bar{b}_{s,t}\cdot x + \bar{c}_{s,t}\right)_+^k,$$

and

$$\partial^\alpha f(x) = c_{k,\rho}(f) \sum_{s=1}^{n} \bar{\rho}(P_s)\, \frac{1}{N_s}\sum_{t=1}^{N_s} \frac{k!}{(k-|\alpha|)!}\bar{a}_{s,t}\bar{b}_{s,t}^\alpha\left(\bar{b}_{s,t}\cdot x + \bar{c}_{s,t}\right)_+^{k-|\alpha|},$$

for all $|\alpha| \leq m$.

Let $g_{s,t}(x) = \frac{k!}{(k-|\alpha|)!}\bar{a}_{s,t}\bar{b}_{s,t}^\alpha\left(\bar{b}_{s,t}\cdot x + \bar{c}_{s,t}\right)_+^{k-|\alpha|}$. Then

$$\mathbb{E}_{\prod_{s=1}^n \mu_s^{N_s}}\left[\left\|D^\alpha f - D^\alpha f_N\right\|_{L^2}^2\right]$$

$$= c_{k,\rho}^2(f) \sum_{s=1}^{n} \frac{\bar{\rho}(P_s)^2}{N_s} \mathbb{E}_{\mu_s} \left\| \mathbb{E}_{\mu_s} g_{s,1} - g_{s,1} \right\|_{L^2}^2 .$$

Notice that there is a constant C_3 depending on Ω, d, m and satisfies

$$\partial_{b_i} \left(\frac{k!}{(k-|\alpha|)!} b^\alpha \, (b \cdot x + c)_+^{k-|\alpha|} \right) < C_3,$$
$$\partial_c \left(\frac{k!}{(k-|\alpha|)!} b^\alpha \, (b \cdot x + c)_+^{k-|\alpha|} \right) < C_3,$$

for all $|\alpha| \leq m$, $(b, c) \in \mathbb{S}^d$ and $x \in \Omega$. Then, noticing $0 \leq m < k$, we derive

$$\begin{aligned}
&\mathrm{Var}_{\mu_s} \left(g_{s,1}(x) \right) \\
&\leq \operatorname*{esssup}_{(b,c),(b',c') \in \mathbb{S}^d} \left| \frac{k!}{(k-|\alpha|)!} \left(b^\alpha \, (b \cdot x + c)_+^{k-|\alpha|} - b'^\alpha \left(b' \cdot x + c' \right)_+^{k-|\alpha|} \right) \right|^2 \\
&\leq C_3^2 \left(\|b - b'\|_1 + \left| c - c' \right| \right)^2 \\
&\leq C_2^2 C_3^2 n^{-\frac{2}{d}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\prod_{s=1}^n \mu_s^{N_s}} \left[\| D^\alpha f - D^\alpha f_N \|_{L^2}^2 \right] &\leq (c_{k,\rho}(f))^2 C_2^2 C_3^2 n^{-\frac{2}{d}} \sum_{s=1}^{n} \frac{(\bar{\rho}(P_s))^2}{N_s} \\
&\leq (c_{k,\rho}(f))^2 C_2^2 C_3^2 n^{-\frac{2}{d}} \sum_{s=1}^{n} \frac{(\bar{\rho}(P_s))}{n} \\
&= (c_{k,\rho}(f))^2 C_2^2 C_3^2 n^{-\frac{2}{d}} \sum_{s=1}^{n} \frac{(\bar{\rho}(P_s))}{n} \\
&\leq (c_{k,\rho}(f))^2 C_2^2 C_3^2 n^{-1-\frac{2}{d}},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{\prod_{s=1}^n \mu_s^{N_s}} \left[\| f - f_N \|_{H^m}^2 \right] &\leq C_{m,d} (c_{k,\rho}(f))^2 C_2^2 C_3^2 n^{-1-\frac{2}{d}} \\
&\leq C^2(\Omega, d, k, m) \| f \|_{B_1^k}^2 N^{-1-\frac{2}{d}},
\end{aligned}$$

which completes the proof. □

Remark 3 We emphasize an important aspect of the above approximation rates. When using rates in Sobolev spaces $H^m(\Omega)$ for $m \geq 1$ then the rates are valid for the approximation of all derivatives up to order m. This is a specific advantage of the present approach, focusing on the simultaneous approximation of a function along with several derivatives.

Remark 4 Under additional assumptions on the domain Ω the above given rate can be improved, and we cite the following result from [23, Thm. 2]. For each n there is a two-layer ReLU^k network f_n with at most n neurons such that

$$\|f - f_n\|_{H^m} \lesssim n^{-\frac{1}{2}-\frac{k-m}{d}}.$$

In particular this assertion is shown to hold for cubes $[0, 1]^d$. However, the constant in front of the upper bound may depend on $\|f\|_{B_1^k}$ as well.

3 Regularization under noisy measurements

Our main focus is the derivative approximation, which is ill-posed in the sense that noisy measurements may provide unsatisfactory reconstructions [10, 12, 26]. Here we assume that the measurements f^δ satisfy the noise assumption (1), which corresponds to the standard assumption in numerical differentiation. We emphasize that random noise, in particular Gaussian white noise, cannot be treated as this is done below, because the noisy data f^δ in general do not belong to $L^2(\Omega)$, hence a Tikhonov functional, see (16) below, cannot be used. Under Gaussian white noise we only have access to a real-valued random variable $\langle f^\delta, y\rangle$, $y \in L^2(\Omega)$, and possible reconstructions must take this into account. We refer to a review paper [8] for extended discussion.

In order to obtain a stable approximation of the unknown function and its derivatives from the noisy measurement f^δ, we shall use a specific variant of Tikhonov regularization. Denote $\mathcal{F}_n$ be the set of all two-layer networks with n neurons, i.e., each $g \in \mathcal{F}_n$ has the form $g = \frac{1}{n}\sum_{i=1}^n a_i\,(b_i \cdot x + c_i)_+^k$. For such $g \in \mathcal{F}_n$ we assign the penalty

$$\mathcal{R}(g) := \left(\frac{1}{n}\sum_{i=1}^n |a_i|\,(\|b_i\|_1 + |c_i|)^k\right)^2 \left(= c^2_{k,\rho_n}(g)\right), \quad g \in \mathcal{F}_n.$$

The reconstruction is given as the minimizer of the Tikhonov functional

$$J_\lambda(g) := \left\|g - f^\delta\right\|^2_{L^2(\Omega)} + \lambda\mathcal{R}(g), \qquad g \in \mathcal{F}_n. \tag{16}$$

The regularization parameter λ shall be chosen appropriately.

Remark 5 In classical regularization theory, in order to stably approximate the derivatives of an unknown function, one shall impose high smoothness penalty referring to (4). Nevertheless, as we will show below, the above regularization scheme (16) with a B_1^k norm penalty term yields good approximation of the unknown function and its derivatives simultaneously.

We next prove that for the regularization scheme (16) a minimizer always exists.

Lemma 10 *For each $\lambda > 0$ there exists a minimizer of* (16).

Proof First we notice that there is some ambiguity in scaling due to the k-homogeneity of the ReLU^k function, such that we may and do assume that for each $i = 1, \dots, n$ we have $\|b_i\|_1 + |c_i| = 1$. After accordingly confining the search for the minimizer the penalty $\mathcal{R}(g)$ reduces to the squared l_1-norm for the coefficients $a_1, \dots, a_n$. With respect to those coefficients the Tikhonov functional is coercive, and we may further restrict to a bounded l_1-ball in $\mathbb{R}^n$, say with radius R. Thus, the minimization problem reduces to a search on the compact set $\left\{(a_i, b_1, c_i) \in (\mathbb{R} \times \mathbb{S}^d)^n,\ \sum_{i=1}^n |a_i| \le R,\ \|b_i\|_1 + |c_i| = 1\right\}$. The continuity of the Tikhonov functional now guarantees the existence of a minimizer. □

The following error bound for the above regularization scheme (16) is important.

Proposition 11 *Let f_n be chosen as in Theorem 8. Suppose that an approximation rate $r_n = r_n(f)$ at the element $f \in B_1^k$ is known, i.e., there holds $\|f - f_n\|_{L_2(\Omega)} \le r_n$. Then any minimizer*

$$f_{n,\lambda}^\delta(x) = \frac{1}{n} \sum_{i=1}^n a_{i,\lambda}^\delta \left(b_{i,\lambda}^\delta \cdot x + c_{i,\lambda}\right)_+^k \in \mathcal{F}_n$$

of the Tikhonov functional (16) satisfies

$$\|f - f_{n,\lambda}^\delta\|_{L_2(\Omega)} \le 2\delta + r_n + \sqrt{\lambda}\|f\|_{B_1^k}, \tag{17}$$

and

$$\|f_{n,\lambda}^\delta\|_{B_1^k} \le \frac{\delta + r_n}{\sqrt{\lambda}} + \|f\|_{B_1^k}. \tag{18}$$

Proof We shall compare the Tikhonov functional at the chosen minimizer $f_{n,\lambda}^\delta$ with its value at the auxiliary element f_n from Theorem 8. The latter is given in (11) and its parameters $\bar{a}_i, \bar{b}_i, \bar{c}_i$ obey $|\bar{a}_i| = c_{k,\rho}(f)$. Hence, the respective penalty term is given as $c_{k,\rho}^2(f_n) \le c_{k,\rho}^2(f)$, see Theorem 8. By the minimizing property of $f_{n,\lambda}^\delta$ this yields the estimate

$$\|f^\delta - f_{n,\lambda}^\delta\|_{L^2(\Omega)}^2 \le \|f^\delta - f_n\|_{L^2(\Omega)}^2 + \lambda c_{k,\rho}^2(f_n), \tag{19}$$

hence

$$\begin{aligned}\|f^\delta - f_{n,\lambda}^\delta\|_{L^2(\Omega)} &\le \|f^\delta - f_n\|_{L^2(\Omega)} + \sqrt{\lambda} c_{k,\rho}(f_n) \\ &\le \delta + \|f - f_n\|_{L^2(\Omega)} + \sqrt{\lambda} c_{k,\rho}(f) \\ &\le \delta + r_n + \sqrt{\lambda} c_{k,\rho}(f).\end{aligned}$$

Therefore we find

$$\|f - f_{n,\lambda}^\delta\|_{L_2(\Omega)} \le \|f^\delta - f_{n,\lambda}^\delta\|_{L_2(\Omega)} + \|f - f^\delta\|_{L_2(\Omega)} \le 2\delta + r_n + \sqrt{\lambda} c_{k,\rho}(f).$$

Since this holds true for every representing probability ρ the bound (17) is proved.

Similarly, we see that

$$\sqrt{\lambda} c_{k,\rho_n}(f_{n,\lambda}^{\delta}) = \sqrt{\lambda}\left(\mathcal{R}(f_{n,\lambda}^{\delta})\right)^{\frac{1}{2}} \leq \delta + r_n + \sqrt{\lambda} c_{k,\rho}(f).$$

Again, this yields the norm bound (18), and the proof is complete. □

It is clear from the error bounds, just established, that a suitable choice of the regularization parameter λ must obey

$$\delta + r_n \leq C\sqrt{\lambda}\|f\|_{B_1^k}, \tag{20}$$

for a constant C which does not depend on δ, and accordingly on r_n, as $\delta \to 0$. In practice the norm $\|f\|_{B_1^k}$ is usually not known to us. By choosing the regularization parameter λ of the order $\delta + r_n \asymp \sqrt{\lambda}$ one balances the terms from the error bound (17) up to coefficients that do not depend on the noise level δ.

This gives the main error bound via interpolation.

Theorem 12 *There is a constant C_1 such that under the assumptions of Proposition 11, for a regularization parameter choice of λ with the property (20), and for $0 \leq m \leq k$ that*

$$\|f - f_{n,\lambda}^{\delta}\|_{H^m(\Omega)} \leq C_1 \lambda^{\frac{k-m}{2k}} \|f\|_{B_1^k}. \tag{21}$$

Proof The bound (17) gives under the choice of λ that

$$\|f - f_{n,\lambda}^{\delta}\|_{L_2(\Omega)} \leq (2C+1)\sqrt{\lambda}\|f\|_{B_1^k}.$$

Moreover, under the regularization parameter choice rule (20) the following norm estimate holds

$$\|f_{n,\lambda}^{\delta}\|_{H^k(\Omega)} \leq \tilde{C}\|f_{n,\lambda}^{\delta}\|_{B_1^k} \leq \tilde{C}(C+1)\|f\|_{B_1^k}.$$

Then, interpolating between $L_2(\Omega)$ and $H^k(\Omega)$ with $0 \leq \theta := m/k \leq 1$ yields that

$$\|f - f_{n,\lambda}^{\delta}\|_{H^m(\Omega)} \leq \|f - f_{n,\lambda}^{\delta}\|_{L_2(\Omega)}^{\frac{k-m}{k}} \|f - f_{n,\lambda}^{\delta}\|_{H^k(\Omega)}^{\frac{m}{k}}.$$

Inserting the previously obtained bounds allows to complete the proof with a constant C_1 depending on C and $\tilde{C}$. □

Corollary 13 *There is a constant C_2 such that if the regularization parameter λ obeys $\delta + r_n \asymp \sqrt{\lambda}$ then*

$$\|f - f_{n,\lambda}^{\delta}\|_{H^m(\Omega)} \leq C_2 \left(\delta + r_n\right)^{\frac{k-m}{k}} \|f\|_{B_1^k}.$$

This gives rise to the following discussion. The best possible rate which can be seen from the above analysis is $\delta^{\frac{k-m}{k}}$. In general, for k-times differentiable functions this rate cannot be improved, as it can be seen, for example, from [20], where the case $d = 1$ has been analyzed. If the rate r_n dominates the noise level δ, then an accuracy of order $r_n^{\frac{k-m}{k}}$ can be ensured. In that case, the better the rate r_n is, the smaller n can be chosen to achieve a desired accuracy. In Theorem 8 a rate $r_n \asymp n^{-\frac{1}{2}}$ was established. That bound was improved in Theorem 9 to $n^{-\frac{1}{2}-\frac{1}{d}}$.

We recall, as discussed in Remark 4, if Ω is a sufficiently smooth manifold, then using the technique developed for the proof of [23, Theorem 2], and taking into account the smoothness of ReLU^k functions (see also (1.23) in [23]), one can improve the rate of r_n up to $\mathcal{O}(n^{-1/2-(k-m)/d})$, for $k \geq 1$. However, see again [23], the constants implicit in the $\mathcal{O}$ symbol may be quite large and accurately estimating them would require careful consideration of the structure of the manifold Ω. In our current framework, we further stress an important difference: For the improved rate as mentioned in Remark 4, the network needs to be trained anew for different values of m. While the more conservative choice suggested in Corollary 13 provides a uniform parameter choice for all m satisfying $0 \leq m \leq k$, it is computationally more affordable to rely on modest rate bounds such as the ones in Theorem 9. Numerical experiments reported in the next section demonstrate that the regularization parameters chosen according to Corollary 13 for the rate bounds from Theorem 9 lead to regularization procedures performing stably in different dimensions and orders m.

4 Numerical examples

In this section, we present several numerical examples verifying the theoretical results in previous sections. Particularly we are interested in performance of the Tikhonov regularization scheme (16) when the regularization parameter is chosen *a priori* as in (20) based on the theoretical analysis in Theorems 9 and 12.

4.1 Setup for the simulations

Throughout the whole section we consider as domain the d-dimensional cube $\Omega := [0, 1]^d$, and the dimension d will vary as specified below. The test target functions are randomly generated polynomials

$$f(x) = \sum_{j=1}^{T} \phi_j K_d(x, \ell^j), \quad x \in [0, 1]^d, \tag{22}$$

where

– K_d is a polynomial kernel

$$K_d(x, \ell) = \left(\langle x, \ell \rangle_{\mathbb{R}^d} + 1\right)^r, \quad x \in [0, 1]^d, \tag{23}$$

with degree r, that will vary in the simulations between $r = 3$ in Examples 1–3 and $r = 2, 4$ in Example 4.
- The nodes $\ell^j = (\ell_1^j, \ell_2^j, \ldots, \ell_d^j)$ and ϕ_j, $j = 1, \ldots, T$ are uniformly distributed random elements in the cube $[-1, 1]^d$ and $[-1, 1]$ respectively.
- The number T of summands is fixed to $T = 5$, throughout.
- Approximation is done by two-layer networks $f_n = \frac{1}{n} \sum_{i=1}^n a_i \, (b_i \cdot x + c_i)_+^k$, with fixed $k := 3$, throughout.

The variety of polynomial kernels actually generates a reproducing kernel Hilbert space of polynomials. Here this is convenient, because the amount of randomness, measured in terms of T, does not depend on the spatial dimension.

From each of these polynomials f we generate noisy measurements f^δ by

$$f^\delta(x) = f(x) + \delta\xi, \quad x \in [0, 1]^d,$$

where δ is the (known) noise level, and for each fixed x the noise ξ is independently uniformly distributed in $[-1, 1]$. Notice, that this choice yields a (weak) random function f^δ with

$$\|f^\delta - f\|_{L_2(\Omega)} \leq \|f^\delta - f\|_{L_\infty(\Omega)} \, |\Omega| \leq \delta,$$

since $|\Omega| = 1$, and hence the noise model (1) holds true.

For given number n of neurons and given noise level δ the regularization parameter λ is chosen according to

$$\lambda = \frac{1}{2} \left(n^{-1-\frac{2}{d}} + \delta^2 \right), \tag{24}$$

based on Theorems 9, 12 and Corollary 13.

In order to realize the numerical simulation we shall replace the Tikhonov functional (16), which uses the $L_2(\Omega)$-norm, by approximating this at a finite number of points by Monte-Carlo simulation, which results in

$$\begin{aligned} L_\lambda \left(f_n; \mathcal{D} \right) := & \frac{1}{\mathcal{D}} \sum_{v=1}^{\mathcal{D}} \left(f_n(x_v) - f^\delta(x_v) \right)^2 \\ & + \lambda \left[\frac{1}{n} \sum_{i=1}^n |a_i| \left(\|b_i\|_1 + |c_i| \right)^3 \right]^2 . \end{aligned} \tag{25}$$

Above, the points x_v, $v = 1, \ldots, \mathcal{D}$ belong to a dataset D, whose cardinality is $\mathcal{D}$, consisting of i.i.d. uniform random points in the unit cube. One might argue that in the light of the functional (25) in principle we only need discrete measurements in order to find the minimizer. In our simulations the random points x_v work well. Nevertheless, when requiring a discrete set of data $f^\delta(x_1), \ldots, f^\delta(x_\mathcal{D})$ from the very beginning, we would face the impact of the design of the given nodes. In particular, in

higher dimensions this would cause serious obstacles. In order to avoid these additional difficulties our assumption (1) is more appropriate.

Minimization of the functional (25) is performed by using the popular gradient descent algorithm `Adam`, see [13]. We choose different batch size V and a fixed epoch number $\mathcal{E} = 1000$ in all the examples.

The obtained results are reported below.

4.2 Numerical examples for $r = k$

Here we deal with the case when the target function f from (22) belongs to the Barron space, which is achieved by letting $r = k = 3$.

Example 1 In this example, we generate a data pool of cardinality $\mathcal{D} = 10{,}000$ and the batch size $V = 500$. We focus on the univariate case with $d = 1$, and $\Omega = [0, 1]$.

Figure 1 shows the relative H_0, H_1 and H_2 errors of our approximated neural networks f_n for 5 different functions in form of (22) by choosing random seeds 1768, 6526, 2729, 6888, 6440 in Pytorch (NumPy).

In all pictures of Fig. 1, the solid line denotes the relative errors by choosing $n = 10$. The vertical solid line shows a critical value of $\delta_0 := n^{-\frac{1}{2}-\frac{1}{d}}$ when $n = 10$ and the regularization parameter depends more on n noticing the *a priori* choice rule (24). In order to visualize the influence of different n, we use the same data pool and the dashed line in these pictures denotes relative errors of $n = 500$. The vertical dashed line shows the critical δ_0 when $n = 500$.

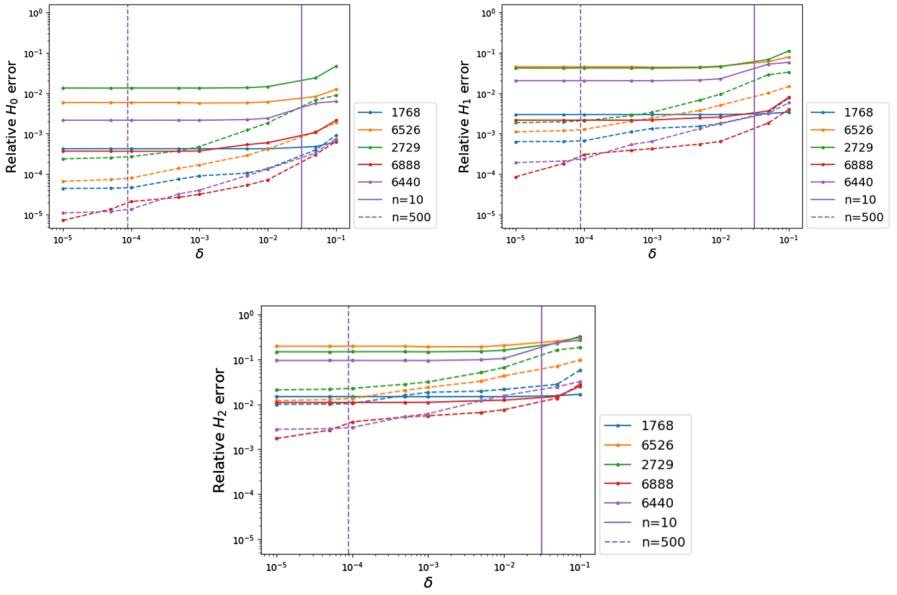

Fig. 1 Relative errors of five different experiments with $d = 1$ and $\mathcal{D} = 10000$. From top (left and right) to bottom: relative errors in H_0, H_1 and H_2 norm

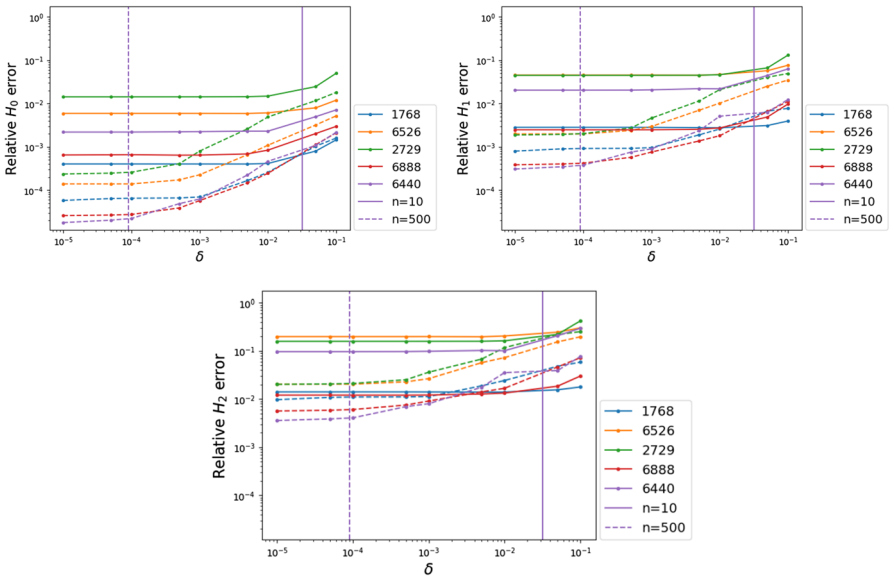

Fig. 2 Relative errors of five different experiments with $d = 1$ and $\mathcal{D} = 1000$. From top (left and right) to bottom: relative errors in H_0, H_1 and H_2 norm

As one can observe, when the noise level is large, the proposed algorithm yields a large relative error which decreases as the noise level decreases. If the noise level reaches a threshold value, i.e. that of the solid or dashed vertical line, the regularization parameter then depends more on the parameter n and relative error becomes flat when the noise vanishes. Though the random seeds provide different target functions the trend of all relative errors behaves similarly. In particular, when the noise level is small, we could observe a direct benefit of the large number n of neurons.

Example 2 Here we generate a smaller data pool with cardinality $\mathcal{D} = 1000$, and the batch size $V = 50$. The target functions are the same as in the previous example. The results are displayed in Fig. 2.

Though both the data pool and the batch size have been decreased essentially, we still observe a similar behavior as in the previous example. In particular, we can clearly observe the advantage of more neurons with the data pool of the same size under all norms. We shall emphasize that by decreasing the cardinality of the data pool we essentially decrease the computational cost.

Example 3 This example considers a high dimension $d = 5$ and hence the domain is $\Omega = [0, 1]^5$. In higher dimension the used data pool has cardinality $\mathcal{D} = 5000$, with a batch size $V = 250$. The results are displayed in Fig. 3. We shall mention that because of the increased dimensionality, the critical value of $\delta_0 := n^{-\frac{1}{2}-\frac{1}{d}}$ also increases as displayed by the corresponding solid and dashed vertical lines in Fig. 3. We observe a similar picture as in the previous two examples, and our algorithm performs better with more neurons in high-dimensional setting.

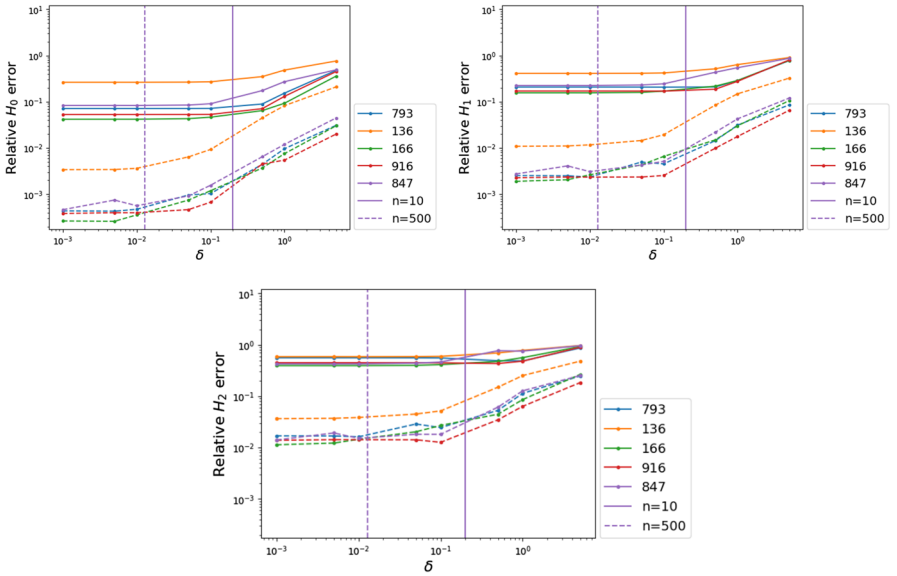

Fig. 3 Relative errors of five different experiments with $d = 5$ and $\mathcal{D} = 5000$. From top (left and right) to bottom: relative errors in H_0, H_1 and H_2 norm

4.3 Numerical examples for $r \neq k$

In general the smoothness of the target function is usually unknown, and we simulate this by letting $r \neq k$, still with $k = 3$. As illustration, we take $r = 2$ and $r = 4$ for the kernel function (23), respectively.

Example 4 In the final example, the data pool is chosen with the cardinality $\mathcal{D} = 1000$, and the batch size is again $V = 50$. By choosing different random seeds, we collect the relative errors in Fig. 4 ($r = 2$), and Fig. 5 ($r = 4$), respectively. It is not surprising that our proposed approach still works well for different choices of r. In case that $r < k$ the advantage of more neurons under high derivatives might be not so straight forward.

We emphasize the following observations.

1. Qualitatively, with measurement data size of the same level, the more neurons in the two-layer networks, the better approximate accuracy as shown in all examples. In high dimension, this conclusion is more solid.
2. In Figs. 1, 2, 3, 4 and 5, when the noise level vanishes, the regularization parameter will converge to $\frac{1}{2}n^{-1-\frac{2}{d}}$ as shown in (24). Then the relative error slopes become flat since the noise is vanishing and the regularization parameter is fixed. It can also be observed that by choosing $n = 10$ and $n = 500$ for very small noise, a (nearly) linear convergence can be achieved by comparing the relative errors of all tests.
3. The proposed approach is universal under different dimensions, which highlights the advantage of neural networks in high dimensional setting.

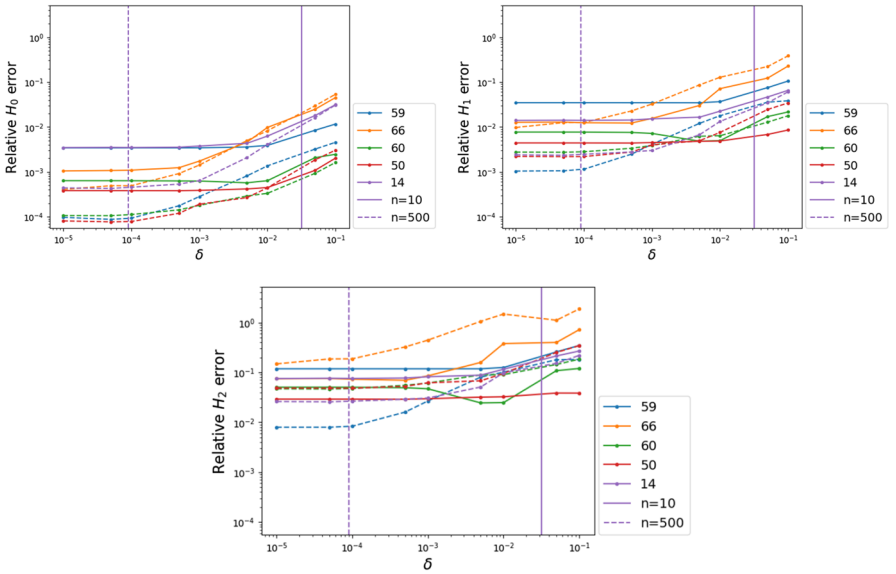

Fig. 4 $r = 2$: Relative errors of five different experiments with $d = 1$ and $\mathcal{D} = 1000$. From top (left and right) to bottom: relative errors in H_0, H_1 and H_2 norm

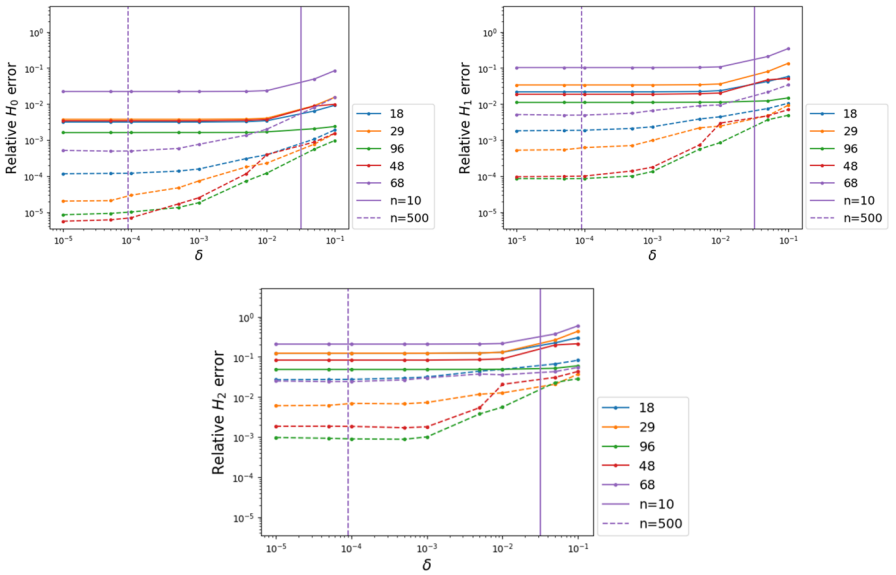

Fig. 5 $r = 4$: Relative errors of five different experiments when $d = 1$, and $\mathcal{D} = 1000$. From top (left and right) to bottom: relative errors in H_0, H_1 and H_2 norm

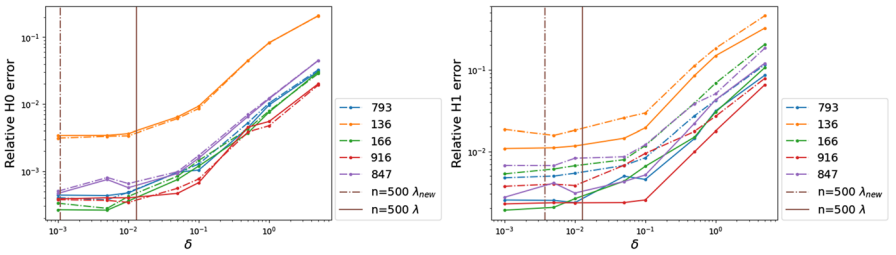

Fig. 6 Comparison of different parameter choice rules with high dimensionality and large number of neurons ($d = 5, n = 500$). **Left**: $m = 0$. **Right:** $m = 1$. The solid line is the error slope of the parameter choice λ (24) and the dashed-point line is that of the parameter choice λ_{new} (26)

4. We also perform simulations with

$$\lambda_{new} = \frac{1}{2}\left(n^{-1-\frac{2(k-m)}{d}} + \delta^2\right) \tag{26}$$

chosen on the basis of the improved approximation rate from Remark 4 for each $m = 0, 1, 2$ and $k = 3$, where some comparison is displayed in Fig. 6. Computationally we need to train the network for different m respectively and no significant improvements can be observed, provided that the number of neurons is large, such that the noise level dominates.

5 Concluding remarks

In the theoretical part we highlighted an intrinsic relation between the newly introduced Barron spaces and the commonly used Sobolev spaces, see e.g. Theorem 7. In spatial dimension one it is clear from the approximation theoretic bounds that $H^1(\Omega) \neq B_2^1$. Indeed, in one dimension the neural networks constitute splines. The best rate of approximation by splines with n nodes in $H^1(\Omega)$ is known to be n^{-1}, whereas the present Theorem 9 yields an order $n^{-3/2}$, and hence $B_1^1 = B_2^1 \subset H^1(\Omega)$ is a proper subspace. It would be interesting to see a similar result in higher spatial dimensions. However, the approximation by neural networks with ReLU^k functions is substantially different from tensor splines, which would serve as analog for splines in higher dimension. While tensor splines are supported on orthants, aligned with the coordinate axes, this does not hold true for neural networks. The support for a single function $x \to (b \cdot x + c)_+^k$ is a halfspace in general position. Thus, to approximate tensor splines by neural networks requires approximating orthants by such halfspaces. As far as we know this problem has not been considered. In the context of the present study this seems to be an important subject worth to be considered in the future. Meanwhile, whether similar convergence results are valid in a stronger norm, for instance the (extended) Barron norm, is also open for us.

Acknowledgements S. Lu is supported by NSFC (No. 11925104), and the Sino-German Mobility Programme (M-0187) by Sino-German Center for Research Promotion. S. Pereverzev is supported by the COMET Module S3AI managed by the Austrian Research Promotion Agency FFG. The authors thank two

anonymous referees for their careful reading of the manuscript and valuable remarks which greatly helped to improve the article.

References

1. Abdeljawad, A., Grohs, P.: Integral representations of shallow neural network with rectified power unit activation function. Neural Netw. **155**, 536–550 (2022)
2. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)
3. Bao, G., Ye, X., Zang, Y., Zhou, H.: Numerical solution of inverse problems by weak adversarial networks. Inverse Probl. **36**(11), 115003 (2020)
4. Barron, A.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inform. Theory **39**(3), 930–945 (1993)
5. Bishop, C.: Training with noise is equivalent to Tikhonov regularization. Neural Comput. **7**(1), 108–116 (1995)
6. Burger, M., Neubauer, A.: Analysis of Tikhonov regularization for function approximation by neural networks. Neural Netw. **16**(1), 79–90 (2003)
7. Caragea, A., Petersen, P., Voigtlaender, F.: Neural network approximation and estimation of classifiers with classification boundary in a Barron class (2022). Accessed: July 19, 2023. arXiv:2011.09363
8. Cavalier, L.: Ch.1 Inverse problems in statistics. In: P. Alquier et al. (eds.) Inverse Problems and High-Dimensional Estimation, Lecture Notes in Statistics, vol. 203. Springer, Berlin (2011)
9. DeVore, R.: Nonlinear approximation. Acta Numer. **7**, 51–150 (1998)
10. Engl, H., Hanke, M., Neubauer, A.: Regularization of inverse problems. In: Mathematics and its Applications, vol. 375. Kluwer Academic Publishers Group, Dordrecht (1996)
11. Gribonval, R., Kutyniok, G., Nielsen, M., Voigtlaender, F.: Approximation spaces of deep neural networks. Constr. Approx. **55**(1), 259–367 (2022)
12. Hanke, M., Scherzer, O.: Inverse problems light: numerical differentiation. Am. Math. Mon. **108**(6), 512–521 (2001)
13. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. ICLR 2015. arXiv:1412.6980 [cs] (2014)
14. Klusowski, J., Barron, A.: Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. IEEE Trans. Inform. Theory **64**(12), 7649–7656 (2018)
15. Kůrková, V.: Complexity estimates based on integral transforms induced by computational units. Neural Netw. **33**, 160–167 (2012)
16. Li, B., Tang, S., Yu, H.: Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. Commun. Comput. Phys. **27**(2), 379–411 (2020)
17. Lu, S., Pereverzev, S.V.: Regularization theory for ill-posed problems, volume 58 of Inverse and Ill-posed Problems Series. De Gruyter, Berlin. Selected topics (2013)
18. Lu, S., Pereverzev, S.V.: Numerical differentiation from a viewpoint of regularization theory. Math. Comput. **75**(256), 1853–1870 (2006)
19. Ma, C., Wu, L.: The Barron space and the flow-induced function spaces for neural network models. Constr. Approx. **55**(1), 369–406 (2022)
20. Magaril-Il'yaev, G.G., Osipenko, K.Y.: Optimal recovery of functions and their derivatives from inaccurate information about the spectrum and inequalities for derivatives. Funct. Anal. Appl. **37**, 203–214 (2003)
21. Moody, J.: The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In: Proceedings of the 4th International Conference on Neural Information Processing Systems, NIPS'91, pp. 847–854, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc (1991)
22. Siegel, J., Xu, J.: High-order approximation rates for shallow neural networks with cosine and ReLU^k activation functions. Appl. Comput. Harmon. Anal. **58**, 1–26 (2022)
23. Siegel, J., Xu, J.: Sharp bounds on the approximation rates, metric entropy, and n-widths of Shallow neural networks. Found. Comput. Math. (2022). https://doi.org/10.1007/s10208-022-09595-3
24. Siegel, J., Xu, J.: Characterization of the variation spaces corresponding to shallow neural networks. Constr. Approx. **57**, 1109–1132 (2023)

25. Wahba, G.: Spline Models for Observational Data, volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1990)
26. Wang, Y.B., Jia, X.Z., Cheng, J.: A numerical differentiation method and its application to reconstruction of discontinuity. Inverse Prob. **18**(6), 1461–1476 (2002)
27. Wojtowytsch, S.: Representation formulas and pointwise properties for Barron functions. Calc. Var. **61**(2), 1–37 (2022)
28. Xu, J.: Finite neuron method and convergence analysis. Commun. Comput. Phys. **28**(5), 1707–1745 (2020)
29. Yarotsky, D.: Error bounds for approximation with deep ReLU networks. Neural Netw. **94**, 103–114 (2017)
30. Zhou, D.: Universality of deep convolutional neural networks. Appl. Comput. Harmon. Anal. **48**(2), 787–794 (2020)