On Sharpness of Error Bounds for Single Hidden Layer Feedforward Neural Networks

Steffen Goebbels

Fourth uploaded version, 05.11.2019

Abstract

A new non-linear variant of a quantitative extension of the uniform boundedness principle is used to show sharpness of error bounds for approximation by sums of sigmoid and ReLU functions. Single hidden layer feedforward neural networks perform such operations. Errors of best approximation can be expressed using moduli of smoothness of the function to be approximated (i.e., to be learned). In this context, the quantitative extension of the uniform boundedness principle indeed allows to construct counter examples that show approximation rates to be best possible. Approximation errors do not belong to the little-o class of given bounds. By choosing piecewise linear activation functions, the discussed problem becomes free knot spline approximation. Results of the present paper also hold for non-polynomial (and not piecewise defined) activation functions like inverse tangent. Based on Vapnik-Chervonenkis dimension, first results are shown for the logistic function.

Keywords: Neural Networks, Rates of Convergence, Sharpness of Error Bounds, Counter Examples, Uniform Boundedness Principle

AMS Subject Classification 2010: 41A25, 41A50, 62M45

1 Introduction

A feedforward neural network with an activation function $\sigma$, one input, one output node, and one hidden layer of $n$ neurons as shown in Figure 1 implements a real function $g$ of type

$$g(x) = \sum_{k=1}^{n} a_k \sigma(b_k x + c_k)$$

with weights $a_k, b_k, c_k \in \mathbb{R}$. Often, activation functions are sigmoid. A sigmoid function $\sigma : \mathbb{R} \to \mathbb{R}$ is a measurable function with

$$\lim_{x \to -\infty} \sigma(x) = 0 \text{ and } \lim_{x \to \infty} \sigma(x) = 1.$$
Fig. 1: One hidden layer neural network realizing $\sum_{k=1}^{n} a_k \sigma(b_k x + c_k)$

Sometimes also monotonicity, boundedness, continuity, or even differentiability may be prescribed. Deviant definitions are based on convexity and concavity. In case of differentiability, functions have a bell-shaped first derivative. Throughout this paper, approximation properties of following sigmoid functions are discussed:

$$\sigma_h(x) := \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \quad \text{(Heaviside function)},$$

$$\sigma_c(x) := \begin{cases} 0, & x < -\frac{1}{2} \\ x + \frac{1}{2}, & -\frac{1}{2} \leq x \leq \frac{1}{2} \\ 1, & x > \frac{1}{2} \end{cases} \quad \text{(cut function)},$$

$$\sigma_a(x) := \frac{1}{2} + \frac{1}{\pi} \arctan(x) \quad \text{(inverse tangent)},$$

$$\sigma_l(x) := \frac{1}{1 + e^{-x}} = \frac{1}{2} \left( 1 + \tanh \left( \frac{x}{2} \right) \right) \quad \text{(logistic function)}.$$ 

Although not a sigmoid function, the ReLU function (Rectified Linear Unit)

$$\sigma_r(x) := \max\{0, x\}$$

is often used as activation function for deep neural networks due to its computational simplicity. We also discuss the Exponential Linear Unit (ELU) activation function

$$\sigma_e(x) := \begin{cases} \alpha(e^x - 1), & x < 0 \\ x, & x \geq 0 \end{cases}$$

as a smoother variant of ReLU for $\alpha \neq 0$.

Qualitative approximation properties of neural networks have been studied extensively. For example, it is possible to choose an infinitely often differentiable, almost monotonous, sigmoid activation function $\sigma$ such that for each continuous function $f$, each compact interval and each bound $\varepsilon > 0$ weights $a_0, a_1, b_1, c_1 \in \mathbb{R}$ exist such that $f$ can be approximated by $a_0 + a_1 \sigma(b_1 x + c_1)$ pointwise on the interval within bound $\varepsilon$, see [23] and literature cited there. In this sense, a neural network with only one hidden neuron is capable of approximating every continuous function. However, a specific activation function is given in a typical scenario. In the late 1980s it was already known that, by increasing the number of neurons, all continuous functions can be approximated arbitrarily well with each non-constant, bounded, and monotone
increasing, continuous, sigmoid activation function, see [20] as well as [14] and [24], cf. [7] [12], for continuous discriminatory activation functions. The literature overviews in increasing, continuous, sigmoid activation function, see [20] as well as [14] and [24], cf. [7] [12], list a variety of such density propositions.

To approximate or interpolate a given but unknown function $f$, constants $a_k, b_k,$ and $c_k$ typically are obtained by learning based on sampled function values of $f$. The underlying optimization algorithm (like gradient descent with back propagation) might get stuck in a local but not in a global minimum. Thus, it might not find optimal constants to approximate $f$ best possible. This paper does not focus on learning but on general approximation properties of function spaces

$$\Phi_{n,\sigma} := \left\{ g : [0, 1] \to \mathbb{R} : g(x) = \sum_{k=1}^{n} a_k \sigma(b_k x + c_k) : a_k, b_k, c_k \in \mathbb{R} \right\}.$$ 

Thus, we discuss functions on the interval $[0, 1]$. Without loss of generality, it is used instead of an arbitrary compact interval $[a, b]$. In some papers, an additional constant function $a_0$ is allowed as summand in the definition of $\Phi_{n,\sigma}$. Please note that $a_k \sigma(0 \cdot x + b_k)$ already is a constant and that the definitions do not differ significantly.

For a function $f : [0, 1] \to \mathbb{R}$ let $\|f\|_{B[0,1]} := \sup\{ |f(x)| : x \in [0,1] \}$. By $C[0,1]$ we denote the Banach space of continuous functions on $[0,1]$ equipped with norm $\|f\|_{C[0,1]} := \|f\|_{B[0,1]}$. For Banach spaces $L^p[0,1], 1 \leq p < \infty$, of measurable functions we denote the norm by $\|f\|_{L^p[0,1]} := (\int_0^1 |f(x)|^p \, dx)^{1/p}$. To avoid case differentiation, we set $X^\infty[0,1] := C[0,1]$ with $\| \cdot \|_{X^\infty[0,1]} := \| \cdot \|_{B[0,1]}$, and $X^p[0,1] := L^p[0,1]$ with $\|f\|_{X^p[0,1]} := \|f\|_{L^p[0,1]}, 1 \leq p < \infty$.

The error of best approximation $E(\Phi_{n,\sigma}, f)_p, 1 \leq p \leq \infty$, is defined via

$$E(\Phi_{n,\sigma}, f)_p := \inf\{ \|f - g\|_{X^p[0,1]} : g \in \Phi_{n,\sigma} \}.$$ 

We use the abbreviation $E(\Phi_{n,\sigma}, f) := E(\Phi_{n,\sigma}, f)_\infty$ for $p = \infty$.

A trained network cannot approximate a function better than the error of best approximation. Therefore, it is an important measure of what can and what cannot be done with such a network.

The error of best approximation depends on the smoothness of $f$ that is measured in terms of moduli of smoothness (or moduli of continuity), a fundamental concept of Approximation Theory. In contrast to using derivatives, first and higher differences of $f$ obviously always exist. By applying a norm to such differences, moduli of smoothness measure a “degree of continuity” of $f$.

Let $f \in B[0,1]$, the set of bounded functions on the interval $[0,1]$. For a natural number $r \in \mathbb{N} = \{1, 2, \ldots \}$, the $r$th difference of $f$ at point $x \in [0, 1 - rh]$ with step size $h > 0$ is defined as

$$\Delta^r_h f(x) := f(x + h) - f(x), \quad \Delta^r_h f(x) := \Delta^r_h \Delta^{r-1}_h f(x), \quad r > 1,$$

or

$$\Delta^r_h f(x) := \sum_{j=0}^{r} (-1)^{r-j} \binom{r}{j} f(x + jh).$$

The $r$th uniform modulus of smoothness is the smallest upper bound of the absolute value of $r$th differences:

$$\omega_r(f, \delta) := \omega_r(f, \delta)_\infty := \sup \{ |\Delta^r_h f(x)| : x \in [0, 1 - rh], 0 < h \leq \delta \}.$$ 

With respect to $L^p$ spaces, $1 \leq p < \infty$, let

$$\omega_r(f, \delta)_p := \left( \int_0^{1-rh} |\Delta^r_h f(x)|^p \, dx \right)^{1/p} : 0 < h \leq \delta.$$
1 Introduction

Obviously, \( \omega_r(f, \delta)_p \leq 2^r \| f \|_{X_p[0,1]}, \) and for \( r \)-times continuously differentiable functions \( f \), there holds (cf. \[15\] p. 46)

\[
\omega_r(f, \delta)_p \leq \delta^r \| f^{(r)} \|_{X_p[0,1]}.
\] (1.1)

Barron applied Fourier methods in \[3\], cf. \[27\], to establish rates of convergence in an \( L^2 \)-norm. Makovoz \[29\] analyzed rates for uniform convergence. With respect to moduli of smoothness, Chen \[6\] proved a direct estimate that is here presented in a version of the textbook \[9\] p. 172ff. This estimate is independent of the choice of a bounded, sigmoid function \( \sigma \). Doctoral thesis \[10\], cf. \[11\], provides an overview of such direct estimates in Section 1.3.

Let function \( f : [0,1] \to \mathbb{R} \) be continuous on \([0,1]\), i.e., \( f \in C[0,1] \). Then, according to Chen

\[
E(\Phi_{n,\sigma}, f) \leq \left[ \sup_{x \in \mathbb{R}} |\sigma(x)| \right] \omega_1 \left( f, \frac{1}{n} \right).
\] (1.2)

This is the prototype estimate for which sharpness is discussed in this paper. In fact, the result of Chen for \( E(\Phi_{n,\sigma}, f) \) allows to additionally restrict weights such that \( b_k \in \mathbb{N} \) and \( c_k \in \mathbb{Z} \). The estimate has to hold true even for \( \sigma \) being a discontinuous Heaviside function. That is the reason why one can only expect an estimate against a first order modulus of smoothness. If the order of approximation of a continuous function \( f \) by such piecewise constant functions is \( o(1/n) \) then \( f \) itself is a constant, see \[15\] p. 366.

In fact, the idea behind Chen’s proof is that sigmoid functions can be asymptotically seen as Heaviside functions. One gets arbitrary step functions to approximate \( f \) by superposition of Heaviside functions. For quasi-interpolation operators based on the logistic activation function \( \sigma_l \), Chen and Zhao proved similar estimates in \[8\] (cf \[2,1\] for hyperbolic tangent). However, they only reach a convergence order of \( O(1/n^\alpha) \) for \( \alpha < 1 \). With respect to the error of best approximation, they prove

\[
E(\Phi_{n,\sigma_l}, f) \leq 80 \omega_1 \left( f, \frac{\exp \left( \frac{3}{2} \right)}{n} \right)
\]

by estimating against a polynomial of best approximation. Due to the different technique, constants are larger than in error bound (1.2).

If one takes additional properties of \( \sigma \) into account, higher convergence rates are possible. Continuous sigmoid cut function \( \sigma_c \) and ReLU function \( \sigma_r \) (as well as leaky ReLU) lead to spaces of continuous, piecewise linear functions. They consist of free knot spline functions of order one with at most \( 2n \) or \( n \) knots, cf. \[15\] Section 12.8. The error of best approximation can be estimated against the error of second order fixed simple knot spline approximation to improve convergence rates up to \( O(1/n^2) \), see \[15\] p. 225 for arbitrary orders. For \( f \in X_p[0,1], \ 1 \leq p \leq \infty \),

\[
E(\Phi_{n,\sigma_{c,r}}, f)_p \leq C \omega_2 \left( f, \frac{1}{n} \right)_p.
\] (1.3)

Section 2 deals with even higher order direct estimates. Similarly to \[1.3\], not only sup-norm bound (1.2) but also an \( L^p \)-bound, \( 1 \leq p \leq \infty \), for approximation with Heaviside function \( \sigma_h \), can be obtained from the corresponding bound of first order fixed simple knot spline approximation:

\[
E(\Phi_{n,\sigma_h}, f)_p \leq C \omega_1 \left( f, \frac{1}{n} \right)_p.
\] (1.4)
Lower error bounds are much harder to obtain than upper bounds, cf. \[31\] for some results with regard to multilayer feedforward perceptron networks. Often, lower bounds are given using a (non-linear) Kolmogorov n-width \( W_n \) (cf. \[30, 34\]),

\[
W_n := \inf_{b_1, \ldots, b_n, c_1, \ldots, c_n} \sup_{f \in X} \inf_{a_1, \ldots, a_n} \| f(\cdot) - \sum_{k=1}^{n} a_k \sigma(b_k \cdot + c_k) \|
\]

for a suitable function space \( X \) (of functions with a certain smoothness) and norm \( \| \cdot \| \). Thus, parameters \( b_k \) and \( c_k \) cannot be chosen individually for each function \( f \in X \). Higher rates of convergence might occur, if that becomes possible.

There are three somewhat different types of sharpness results that might be able to show that left sides of equations (1.2), (1.3), (1.4) or (2.2) in Section 2 do not converge faster to zero than the right sides.

The most far reaching results would provide lower estimates of errors of best approximation against moduli of smoothness. In connection with direct upper bounds against the same moduli, this would establish theorems similar to the equivalence between moduli of smoothness and K-functionals (cf. \[15, theorem of Johnen, p. 177\]) in which the error of best approximation replaces the K-functional. Let \( \sigma \) be \( r \)-times continuously differentiable like \( \sigma_a \) or \( \sigma_l \). Then for \( f \in C[0,1] \), a standard estimate based on (1.1) is

\[
\omega_r \left( f, \frac{1}{n} \right) \leq \inf_{g \in \Phi_n, \sigma} \left[ \omega_r \left( f - g, \frac{1}{n} \right) + \omega_r \left( g, \frac{1}{n} \right) \right] \leq \inf_{g \in \Phi_n, \sigma} \left[ 2^r \| f - g \|_{B[0,1]} + \frac{1}{n^r} \| g^{(n)} \|_{B[0,1]} \right].
\]

It unlikely that one can somehow bound \( \| g^{(n)} \|_{B[0,1]} \) by \( C \| f \|_{B[0,1]} \) to get

\[
\omega_r \left( f, \frac{1}{n} \right) \leq 2^r E(\phi_n, \sigma, f) + \frac{C}{n^r} \| f \|_{B[0,1]}.
\]

However, there are different attempts to prove such theorems in \[38, Remark 1, p. 620\], \[39, p. 101\] and \[40, p. 451\]. In the opinion of the author, the proofs contain difficulties. The lower bound estimates are based on an unproved assumption that two zero sequences converge to zero with the same order. With the same argument one could also improve the classical inverse theorem of trigonometric approximation, cf. \[15, p. 208\].

A second class of sharpness results consists of inverse and equivalence theorems. Inverse theorems provide upper bounds of moduli of smoothness in terms of weighted sums of approximation errors. If one adapts the mentioned inverse theorem for trigonometric approximation without considering effects related to interval endpoints in algebraic approximation then one gets a candidate inequality

\[
\omega_r \left( f, \frac{1}{n} \right) \leq \frac{C_r}{n^r} \sum_{k=1}^{n} k^{r-1} E(\phi_k, \sigma, f). \tag{1.5}
\]

Typically, the proof of an inverse theorem is based on a Bernstein-type inequality that is difficult to formulate for function spaces discussed here. The Bernstein inequality gives a bound for derivatives. If \( p_n \) is a trigonometric polynomial of degree at most \( n \) then \( \| p_n \|_{B[0,2\pi]} \leq n \| p_n \|_{B[0,\pi]} \), cf. \[15, p. 97\]. The problem here is that differentiating \( a \sigma(bx + c) \) leads to a factor \( b \) that cannot be bounded easily. Similar to
inverse theorems, equivalence theorems (like (1.6) below) describe equivalent behavior of expressions of moduli of smoothness and expressions of approximation errors. Both types of theorems allow to determine smoothness properties, typically membership to Lipschitz classes or Besov spaces, from convergence rates. Such a property is proved in [13] for max-product neural network operators activated by sigmoidal functions. For pseudo-interpolation operators based on piecewise linear activation functions and B-splines (but also not for errors of best approximation), [28] deals with an inverse estimate based on Bernstein polynomials. Inverse estimates for best approximation with activation functions of type (1.5) are stated in [40] and [38]. In the proofs, special sums of $n$ activation functions are constructed from polynomials of best approximation similar to the proof of Theorem 2.1 in Section 2 (see top of p. 623 in [38]). The obtained special sums do approximate not better than polynomials of best approximation. Then it is concluded without explanation that each possible sum of $n$ activation functions also does not approximate significantly better than the special sum. Thus, the proofs of these inverse theorems appear to be incomplete. A similar argument seems to be used in [39]. There, the lower bound does not hold for the error of best approximation because it is violated if one approximates the activation function itself with zero error. Since referenced equation (4.9) is missing in the paper, it appears that function $N_n f$ is a good approximation to Bernstein polynomial $B_i f$ instead of given $f$ in the proof of (5.7). However, there might be much better neural network approximations to $f$.

The relationship between order of convergence of best approximation and Besov spaces is well understood for approximation with free knot spline functions and rational functions, see [15, Section 12.8], cf. [26]. The Heaviside activation function leads to free knot splines of polynomial degree $r = 1$, cut and ReLU function correspond with degree $r = 2$. For $\sigma$ being one of these functions, and for $0 < \alpha < r$, $f \in L^p[0,1]$, $1 \leq p < \infty (p = \infty$ is excluded), $k := 1$ if $\alpha < 1$ and $k := 2$ otherwise, $q := \frac{1}{\alpha + 1/p}$, there holds the equivalence (see [16])

$$\int_0^\infty \frac{\omega_k(f, t)^q}{t^{1+\alpha q}} dt < \infty \iff \sum_{n=1}^\infty \left( \frac{E(\Phi_n, \sigma, f)}{n^{1-\alpha q}} \right)^q < \infty.$$  \hspace{1cm} (1.6)

However, such equivalence theorems might not be suited to obtain little-o results: Assume that $E(\Phi_n, \sigma, f) = \frac{1}{n^{1/(ln(n+1))^\beta}} = o \left( \frac{1}{n^\gamma} \right)$, then the right side of (1.6) converges exactly for the same smoothness parameters $0 < \alpha < \beta$ than if $E(\Phi_n, \sigma, f) = \frac{1}{n^\gamma} \neq o \left( \frac{1}{n^\gamma} \right)$.

The third type of sharpness results is based on counter examples. The present paper follows this approach to deal with little-o effects. Without further restrictions, counter examples show that convergence orders can not be faster than stated in (1.2), (1.3), (1.4) and the estimates in following Section 2 for some activation functions. To obtain such counter examples, a general theorem is introduced in Section 3. It is applied to neural network approximation in Section 4.

2 Direct estimates

In this section, two upper bounds in terms of higher order moduli of smoothness are derived from known results. Proofs are given for the sake of completeness. Under reasonable assumptions on $\sigma$, it is known that $E(\Phi_n, \sigma, p_{n-1}) = 0$ for all polynomials $p_{n-1}$ of degree at most $n-1$, i.e., $p_{n-1} \in \Pi^n := \{d_{n-1}x^{n-1} + d_{n-2}x^{n-2} + \cdots + d_0 :$
The Jackson estimate can be used to extend the proof given for the theorem of Jackson, the best approximation to an f such that for x in [0, 1], we see that there exists a constant C independent of f and n such that

\[
\inf\{\|f - p_n\|_{X^p[0,1]} : p_n \in \Pi^{n+1}\} \leq C\omega_r \left( \frac{f}{n}, \frac{1}{n} \right)_p.
\] (2.1)

Ritter proved an estimate against a first order modulus of smoothness for approximation with nearly exponential activation functions in [32]. Due to (2.1), Ritter’s proof can be extended in a straightforward manner to higher order moduli. The special case of estimating against a second order modulus is discussed in [33].

According to [32], a function \( \sigma : \mathbb{R} \to \mathbb{R} \) is called “nearly exponential” if for each \( \varepsilon > 0 \) there exist real numbers \( a, b, c, \) and \( d \) such that for all \( x \in (-\infty, 0] \)

\[
|a\sigma(bx + c) + d - e^x| < \varepsilon.
\]

The logistic function fulfills this condition with \( a = 1/\sigma_1(c) \), \( b = 1 \), \( d = 0 \), and \( c < \ln(\varepsilon) \) such that for \( x \leq 0 \) there is \( 0 < e^x \leq 1 \) and

\[
|a\sigma_1(bx + c) + d - e^x| = \left| \frac{\sigma_1(x + c)}{\sigma_1(c)} \right| e^x = \left| 1 + e^{-c} \right| e^x - e^x \leq e^x \left| \frac{1 - e^{-c}}{e^x + e^{-c}} \right| \leq \varepsilon < \varepsilon.
\]

**Theorem 2.1.** Let \( \sigma : \mathbb{R} \to \mathbb{R} \) be a nearly exponential function, \( f \in X^p[0,1], 1 \leq p \leq \infty \), and \( r \in \mathbb{N} \). Then, independently of \( n \geq \max\{r, 2\} \) and \( f \), a constant \( C_r \) exists such that

\[
E(\Phi_n, \sigma, f)_p \leq C_r\omega_r \left( \frac{f}{n}, \frac{1}{n} \right)_p.
\] (2.2)

**Proof.** Due to (2.1), for each \( \varepsilon > 0 \) there exists a polynomial \( p_n \in \Pi^{n+1} \) of degree at most \( n \) such that

\[
\|f - p_n\|_{X^p[0,1]} \leq C_r\omega_r \left( \frac{f}{n}, \frac{1}{n} \right)_p + \frac{\varepsilon}{3}.
\] (2.3)

The Jackson estimate can be used to extend the proof given for \( r = 1 \) in [32]: Auxiliary functions \( \alpha > 0 \)

\[
h_\alpha(x) := \alpha \left( 1 - \exp \left( -\frac{x}{\alpha} \right) \right)
\]

converge to \( x \) pointwise for \( \alpha \to \infty \) due to the theorem of L’Hospital. Since \( \frac{d}{dx}(h_\alpha(x) - x) = 0 \iff e^{-x/\alpha} = 1 \iff x = 0 \), the maximum of \( |h_\alpha(x) - x| \) on \([0, 1]\) is obtained at the endpoints 0 or 1, and convergence of \( h_\alpha(x) \) to \( x \) is uniform on \([0, 1]\) for \( \alpha \to \infty \). Thus \( \lim_{\alpha \to \infty} p_n(h_\alpha(x)) = p_n(x) \) uniformly on \([0, 1]\), and we can choose \( \alpha \) large enough to get

\[
\|p_n(\cdot) - p_n(h_\alpha(\cdot))\|_{X^p[0,1]} \leq \frac{\varepsilon}{3}.
\] (2.4)

Therefore, function \( f \) is approximated by an exponential sum of type

\[
p_n(h_\alpha(x)) = \gamma_0 + \sum_{k=1}^n \gamma_k \exp \left( -\frac{Kx}{\alpha} \right)
\]
within the bound \( C_r 2^{-r} \omega_r (f, n^{-1}) + 2 \varepsilon / 3 \). It remains to approximate the exponential sum by utilizing that \( \sigma \) is nearly exponential. For \( 1 \leq k \leq n \), \( \gamma_k \neq 0 \), there exist parameters \( a_k, b_k, c_k, d_k \) such that
\[
|a_k \sigma (b_k (x - \frac{kx}{\alpha}) + c_k) + d_k - \exp \left( -\frac{kx}{\alpha} \right)| < \frac{\varepsilon}{3n|\gamma_k|}
\]
for all \( x \in [0, 1] \). Also because \( \sigma \) is nearly exponential, there exists \( c \in \mathbb{R} \) with \( \sigma(c) \neq 0 \). Thus, a constant \( \gamma \) can be expressed via \( \frac{\gamma}{\sigma(c)} \sigma(0x + c) \). Therefore, there exists a function \( g_c \in \Phi_{n+1, \sigma} \),
\[
g_c(x) = \frac{\gamma_0}{\sigma(c)} \sigma(0x + c) + \sum_{k=1, \gamma_k \neq 0}^{n} \gamma_k \left[ a_k \sigma \left( b_k \left( -\frac{kx}{\alpha} \right) + c_k \right) + d_k \right]
\]
such that
\[
\|p_n(h_\alpha(\cdot)) - g_c\|_{B[0,1]} \leq \sum_{k=1, \gamma_k \neq 0}^{n} |\gamma_k| \left| a_k \sigma \left( b_k \left( -\frac{kx}{\alpha} \right) + c_k \right) + d_k - \exp \left( -\frac{kx}{\alpha} \right) \right|_{B[0,1]}
\]
\[
\leq \frac{n \varepsilon}{3n} = \frac{\varepsilon}{3}.
\]
By combining (2.3), (2.4) and (2.5), we get
\[
E(\Phi_{n+1, \sigma}, f)_p \leq \|f - p_n\|_{X_p[0,1]} + \|p_n - p_n(h_\alpha(\cdot))\|_{B[0,1]} + \|p_n(h_\alpha(\cdot)) - g_c\|_{B[0,1]} \leq \frac{C_r}{2^r} \omega_r \left( f, \frac{1}{n} \right)_p + \varepsilon.
\]
Since \( \varepsilon \) can be chosen arbitrarily, we obtain for \( n \geq 2 \)
\[
E(\Phi_{n, \sigma}, f)_p \leq \frac{C_r}{2^r} \omega_r \left( f, \frac{1}{n-1} \right)_p \leq \frac{C_r}{2^r} \omega_r \left( f, \frac{2}{n} \right)_p \leq C_r \omega_r \left( f, \frac{1}{n-1} \right)_p.
\]

By choosing \( a = 1/\alpha, b = 1, c = 0 \), and \( d = 1 \), the ELU activation function \( \sigma_e(x) \) obviously fulfills the condition to be nearly exponential. But its definition for \( x \geq 0 \) plays no role. Thus for ELU, Theorem 2.1 might not provide a sharp bound, cf. (4.7).

The “nearly exponential” property only fits with certain activation functions. A more general theorem is based on differentiability. If \( \sigma \) is arbitrarily often differentiable on some open interval and if \( \sigma \) is no polynomial on that interval then one can easily obtain an estimate against the \( r \)th modulus from the Jackson estimate (2.1) by considering that polynomials of degree at most \( n - 1 \) can be approximated arbitrarily well by functions in \( \Phi_{n, \sigma} \), see [31] Corollary 3.6, p. 157]. The idea is to approximate monomials by differential quotients of \( \sigma \). This is possible since derivative
\[
\frac{\partial^k}{\partial b^k} \sigma(bx + c) = x^k \sigma^{(k)}(bx + c)
\]
at \( b = 0 \) equals \( \sigma^{(k)}(c)x^k \). Because polynomials are excluded, constants \( \sigma^{(k)}(c) \neq 0 \) can be chosen. The following estimate extends [31] Theorem 6.8, p. 176] to moduli of smoothness.
3 A Uniform Boundedness Principle with Rates

Theorem 2.2. Let \( \sigma : \mathbb{R} \to \mathbb{R} \) be arbitrarily often differentiable on some open interval in \( \mathbb{R} \) and let \( \sigma \) be no polynomial on that interval, \( f, g \in X^p[0,1], 1 \leq p \leq \infty \), and \( r \in \mathbb{N} \). Then, independently of \( n \geq \max\{r, 2\} \) and \( f, g \), a constant \( C_r \) exists such that

\[
E(\Phi_{n,\sigma}, f)_p \leq C_r \omega_r \left( f, \frac{1}{n} \right)_p .
\]

This theorem can be applied to \( \sigma_1 \) but also to \( \sigma_n \) and \( \sigma_e \).

Proof. Let \( \varepsilon > 0 \). As in the previous proof, there exists a polynomial \( p_n \) of degree at most \( n \) such that (2.3) holds. Due to [31, p. 157] there exists a function \( g_e \in \Phi_{n+1,\sigma} \) such that \( \|g_e - p_n\|_{B[0,1]} < 2\varepsilon/3 \). This gives

\[
E(\Phi_{n+1,\sigma}, f)_p \leq \|f - p_n\|_{X^p[0,1]} + \|p_n - g_e\|_{B[0,1]} \leq \frac{C_r}{2} \omega_r \left( f, \frac{1}{n} \right)_p + \varepsilon.
\]

Since \( \varepsilon \) can be chosen arbitrarily, we get (2.8) via equation (2.9).

\( \square \)

3 A Uniform Boundedness Principle with Rates

In this paper, sharpness results are proved with a quantitative extension of the classical uniform boundedness principle of Functional Analysis. Dickmeis, Nessel and van Wickern developed several versions of such theorems. An overview of applications in Numerical Analysis can be found in [21 Section 6]. The given paper is based on [19 p. 108]. This and most other versions require error functionals to be sub-additive. Let \( X \) be a normed space. A functional \( T \) on \( X \), i.e., \( T \) maps \( X \) into \( \mathbb{R} \), is said to be (non-negative-valued) sub-linear and bounded, iff for all \( f, g \in X, c \in \mathbb{R} \)

\[
T(f) \geq 0, \quad T(f + g) \leq T(f) + T(g) \quad (\text{sub-additivity}),
\]

\[
T(cf) = cT(f) \quad (\text{absolute homogeneity}),
\]

\[
\|T\|_{X^\infty} := \sup\{T(f) : \|f\|_X \leq 1\} < \infty \quad (\text{bounded functional}).
\]

The set of non-negative-valued sub-linear bounded functionals \( T \) on \( X \) is denoted by \( X^\infty \). Typically, errors of best approximation are (non-negative-valued) sub-linear bounded functionals. Let \( U \subset X \) be a linear subspace. The best approximation of \( f \in X \) by elements \( u \in U \neq \emptyset \) is defined as \( E(f) := \inf\{\|f - u\|_X : u \in U\} \). Then \( E \) is sub-linear: \( E(f + g) \leq E(f) + E(g) \), \( E(cf) = |c|E(f) \) for all \( c \in \mathbb{R} \). Also, \( E \) is bounded: \( E(f) \leq \|f - 0\|_X = \|f\|_X \).

Unfortunately, function sets \( \Phi_{n,\sigma} \) are not linear spaces, cf. [31 p. 151]. In general, from \( f, g \in \Phi_{n,\sigma} \), one can only conclude \( f + g \in \Phi_{2n,\sigma} \) whereas \( cf \in \Phi_{n,\sigma} \), \( c \in \mathbb{R} \). Functionals of best approximation fulfill \( E(\Phi_{n,\sigma}, f)_p \leq \|f - 0\|_{X^p[0,1]} = \|f\|_{X^p[0,1]} \).

Absolute homogeneity \( E(\Phi_{n,\sigma}, cf)_p = |c|E(\Phi_{n,\sigma}, f)_p \) is obvious for \( c = 0 \). If \( c \neq 0 \),

\[
E(\Phi_{n,\sigma}, cf)_p = \inf \left\{ \left\| cf - \sum_{k=1}^{n} a_k \sigma(b_k x + c_k) \right\|_{X^p[0,1]} : a_k, b_k, c_k \in \mathbb{R} \right\}
\]

\[
= |c| \inf \left\{ \left\| f - \sum_{k=1}^{n} \frac{a_k}{c} \sigma(b_k x + c_k) \right\|_{X^p[0,1]} : \frac{a_k}{c}, b_k, c_k \in \mathbb{R} \right\}
\]
\[= |c| E(\Phi_n, f)_p.\]

But there is no sub-additivity. However, it is easy to prove a similar condition: For each \(\varepsilon > 0\) there exists elements \(u_{f,\varepsilon}, u_{g,\varepsilon} \in \Phi_n,\sigma\) that fulfill
\[
\|f - u_{f,\varepsilon}\|_{X_p[0,1]} \leq E(\Phi_n, f)_p + \frac{\varepsilon}{2}, \quad \|g - u_{g,\varepsilon}\|_{X_p[0,1]} \leq E(\Phi_n, g)_p + \frac{\varepsilon}{2}
\]
and \(u_{f,\varepsilon} + u_{g,\varepsilon} \in \Phi_n,\sigma\) such that
\[
E(\Phi_n, f + g)_p \leq \|f - u_{f,\varepsilon} + g - u_{g,\varepsilon}\|_{X_p[0,1]}
\leq \|f - u_{f,\varepsilon}\|_{X_p[0,1]} + \|g - u_{g,\varepsilon}\|_{X_p[0,1]} \leq E(\Phi_n, f)_p + E(\Phi_n, g)_p + \varepsilon,
\]
i.e.,
\[
E(\Phi_n, f + g)_p \leq E(\Phi_n, f)_p + E(\Phi_n, g)_p. \tag{3.1}
\]
Obviously, also \(E(\Phi_n, f)_p \geq E(\Phi_n, f)_p\) holds true.

In what follows, a quantitative extension of the uniform boundedness principle based on this condition is presented. The condition replaces sub-additivity. Another extension of the uniform boundedness principle to non-sub-linear functionals is proved in [18]. But this version of the theorem is stated for a family of error functionals with two parameters that has to fulfill a condition of quasi lower semi-continuity.

Functions \(S\) measuring smoothness also do not need to be sub-additive but have to fulfill a condition \(S_2(f + g) \leq B(S_2(f) + S_2(g))\) for a constant \(B \geq 1\). This theorem does not consider replacement \(\text{3.1}\) for sub-additivity.

Both rate of convergence and size of moduli of smoothness can be expressed by abstract moduli of smoothness, see [36, p. 96ff]. Such an abstract modulus of smoothness is a continuous, increasing function \(\omega\) on \([0, \infty)\) that has similar properties as \(\omega_r(f, \cdot)\), i.e., for \(0 < \delta_1, \delta_2\),
\[
0 = \omega(0) < \omega(\delta_1) \leq \omega(\delta_1 + \delta_2) \leq \omega(\delta_1) + \omega(\delta_2). \tag{3.2}
\]
Especially, for \(\lambda > 0\)
\[
\omega(\delta) \leq \omega(|\lambda| \delta) \leq |\lambda| \omega(\delta) \leq (\lambda + 1) \omega(\delta) \tag{3.3}
\]
and due to continuity \(\lim_{\delta \to 0^+} \omega(\delta) = 0\). For all \(0 < \delta_1 \leq \delta_2\), equation \(\text{3.3}\) also implies
\[
\frac{\omega(\delta_2)}{\delta_2} = \frac{\omega \left( \frac{\delta_2}{\delta_1} \delta_1 \right)}{\delta_2} \leq \left( 1 + \frac{\delta_1}{\delta_2} \right) \frac{\omega(\delta_1)}{\delta_2} = \frac{\delta_2 + \delta_1}{\delta_2} \frac{\omega(\delta_1)}{\delta_1} \leq 2 \frac{\omega(\delta_1)}{\delta_1}. \tag{3.4}
\]
Functions \(\omega(\delta) := \delta^\alpha, 0 < \alpha \leq 1\), are examples for abstract moduli of smoothness. They are used to define Lipschitz classes.

The aim is to discuss a sequence of remainders (that will be errors of best approximation) \((E_n)_{n=1}^\infty, E_n : X \to [0, \infty)\). These functionals do not have to be sub-linear but instead have to fulfill
\[
E_m \left( \sum_{k=1}^n f_k \right) \leq \sum_{k=1}^m E_n(f_k) \quad (\text{cf. 3.1}) \tag{3.5}
\]
\[
E_n(cf) = |c| E_n(f) \tag{3.6}
\]
\[
E_n(f) \leq D_n \|f\|_X \tag{3.7}
\]
\[
E_n(f) \geq E_{n+1}(f) \tag{3.8}
\]
for all \(m \in \mathbb{N}, f, f_1, f_2, \ldots, f_m \in X\), and constants \(c \in \mathbb{R}\). In the boundedness condition \(\text{3.7}\), \(D_n\) is a constant only depending on \(E_n\) but not on \(f\).
Theorem 3.1 (Adapted Uniform Boundedness Principle). Let $X$ be a (real) Banach space with norm $\| \cdot \|_X$. Also, a sequence $(E_n)_{n=1}^\infty$, $E_n : X \to [0, \infty)$ is given that fulfills conditions (3.5)–(3.8). To measure smoothness, sub-linear bounded functionals $S_\delta \in X^\sim$ are used for all $\delta > 0$.

Let $\mu(\delta) : (0, \infty) \to (0, \infty)$ be a positive function, and let $\varphi : [1, \infty) \to (0, \infty)$ be a strictly decreasing function with $\lim_{x \to \infty} \varphi(x) = 0$. An additional requirement is that for each $0 < \lambda < 1$ a point $X_\lambda = X_0(\lambda) \geq \lambda^{-1}$ and constant $C_\lambda > 0$ exist such that
\[
\varphi(\lambda x) \leq C_\lambda \varphi(x)
\] (3.9) for all $x > X_\lambda$.

If there exist test elements $h_n \in X$ such that for all $n \in \mathbb{N}$ with $n \geq n_0 \in \mathbb{N}$ and $\delta > 0$
\[
\|h_n\|_X \leq C_1, \tag{3.10}
\]
\[
S_\delta h_n \leq C_2 \min \left\{ 1, \frac{\mu(\delta)}{\varphi(n)} \right\}, \tag{3.11}
\]
\[
E_4 h_n \geq c_3 > 0, \tag{3.12}
\]
then for each abstract modulus of smoothness $\omega$ satisfying (3.2) and
\[
\lim_{\delta \to 0^+} \frac{\omega(\delta)}{\delta} = \infty \tag{3.13}
\]
there exists a counter example $f_\omega \in X$ such that $(\delta \to 0^+, n \to \infty)$
\[
S_\delta f_\omega = o(\omega(\mu(\delta))), \tag{3.14}
\]
\[
E_n f_\omega \neq o(\omega(\varphi(n))). \tag{3.15}
\]

For example, (3.9) is fulfilled for a standard choice $\varphi(x) = \frac{1}{x^\alpha}$.

The prerequisites of the theorem differ from the Theorems of Dickmeis, Nessel, and van Wickern in conditions (3.5)–(3.8) that replace $E_n \in X^\sim$. It also requires additional constraint (3.9). For convenience, resonance condition (3.12) replaces $E_n h_n \geq c_3$. Without much effort, (3.12) can be weakened to $\limsup_{n \to \infty} E_n h_n > 0$.

The proof is based on a gliding hump and follows the ideas of [19, Section 2.2] (cf. [17]) for sub-linear functionals and the literature cited there. For the sake of completeness, the whole proof is presented although changes were required only for estimates that are effected by missing sub-additivity.

Proof. The first part of the proof is not concerned with sub-additivity or its replacement. If a test element $h_j$ exists that already fulfills
\[
\limsup_{n \to \infty} \frac{E_n h_j}{\omega(\varphi(n))} > 0, \tag{3.16}
\]
then $f_\omega := h_j$ fulfills (3.10). To show that this $f_\omega$ also fulfills (3.11), one needs inequality
\[
\min \{1, \delta\} \leq A \omega(\delta) \tag{3.17}
\]
for all $\delta > 0$. This inequality follows from (3.3): If $0 < \delta < 1$ then $\omega(1)/1 \leq 2\omega(\delta)/\delta$, such that $\delta \leq 2\omega(\delta)/\omega(1)$. If $\delta > 1$ then $\omega(1) \leq \omega(\delta)$, i.e., $1 \leq \omega(\delta)/\omega(1)$. Thus, one can choose $A = 2/\omega(1)$.
Smoothness \(3.14\) of test elements \(h_j\) now follows from \(3.17\):
\[
S_k(h_j) \leq C_2 \min \left\{ \frac{1}{\varphi(j)} \right\} \leq AC_2 \omega \left( \frac{\mu(\delta)}{\varphi(j)} \right) \\
\leq AC_2 \left( 1 + \frac{1}{\varphi(j)} \right) \omega(\mu(\delta)).
\]

Under condition \(3.16\) function \(f_\omega := h_j\) indeed is a counter example. Thus, for the remaining proof one can assume that for all \(j \in \mathbb{N}, j \geq n_0:\)
\[
\lim_{n \to \infty} \frac{E_n(h_j)}{\omega(\varphi(n))} = 0.
\tag{3.18}
\]

The arguments of Dickmeis, Nessel and van Wickern have to be adjusted to missing sub-additivity in the next part of the proof. It has to be shown that for each fixed \(m \in \mathbb{N}\) a finite sum inherits limit \(3.15\). Let \((a_i)_{i=1}^{m} \subset \mathbb{R}\) and \(j_1, \ldots, j_m \geq n_0\) different indices. To prove
\[
\lim_{n \to \infty} \frac{E_n(\sum_{i=1}^{m} a_i h_{j_i})}{\omega(\varphi(n))} = 0,
\tag{3.19}
\]
one can apply \(3.5, 3.6,\) and \(3.3\) for \(n \geq 2m:\)
\[
0 \leq E_n(\sum_{i=1}^{m} a_i h_{j_i}) \leq E_n(\varphi(n)) \sum_{i=1}^{m} a_i h_{j_i} \leq \frac{\sum_{i=1}^{m} E_n(\varphi(n))}{\omega(\varphi(n))} \sum_{i=1}^{m} \frac{E_n(a_i h_{j_i})}{\omega(\varphi(n))}.
\tag{3.20}
\]

Since \(\varphi(x)\) is decreasing and \(\omega(\delta)\) is increasing, \(\omega(\varphi(x))\) is decreasing. Thus, \(3.9\) for \(\lambda := (2m)^{-1}\) and \(n > \max\{2m, X_0(\lambda)\}\) implies
\[
\omega(\varphi([n/m])) \leq \omega \left( \varphi \left( \frac{n-m}{m} \right) \right) \leq \omega \left( \varphi \left( \frac{n-n/2}{m} \right) \right) \leq \omega \left( C_{\lambda m} \varphi(n) \right) \leq C_{\lambda m} \omega(\varphi(n)).
\tag{3.21}
\]
With this inequality, estimate \(3.20\) becomes
\[
0 \leq E_n(\sum_{i=1}^{m} a_i h_{j_i}) \leq \frac{C_{\lambda m}}{\omega(\varphi(n))} \sum_{i=1}^{m} |a_i| E_n(a_i h_{j_i}) \omega(\varphi([n/m])).
\]
According to \(3.18\), this gives \(3.19\).

Now one can select a sequence \((n_k)_{k=1}^{\infty} \subset \mathbb{N}, n_0 \leq n_k < n_{k+1}\) for all \(k \in \mathbb{N},\) to construct a suitable counter example
\[
f_\omega := \sum_{k=1}^{\infty} \omega(\varphi(n_k)) h_{n_k}.
\tag{3.21}
\]

Let \(n_1 := n_0.\) If \(n_1, \ldots, n_k\) have already be chosen then select \(n_{k+1} \geq 2k\) large enough to fulfill following conditions:
\[
\omega(\varphi(n_{k+1})) \leq \frac{1}{2} \omega(\varphi(n_k)) \quad \left( \lim_{x \to \infty} \omega(\varphi(x)) = 0 \right)
\tag{3.22}
\]
Due to $4 \delta > C$, this implies $f \in L^p$. Thus, the Banach condition is fulfilled and counter example $k$ can be found.

With this estimate (and because of $\lim_{x \to \infty} \omega(f(x)) = 0$), this case of $(3.23)$ holds true because $\omega(f(n_k))$, for all $n_k$.

Using this index $k$, the last estimate holds true because $\omega(\varphi(n_k)) < \mu(\delta)$. The next part of the proof does not consider properties of $E_n$, see [19].

Function $f_\omega$ in $(3.24)$ is well-defined: For $j \geq k$, iterative application of $(3.22)$ leads to

$$\omega(\varphi(n_j)) \leq 2^{-j} \omega(\varphi(n_{j-1})) \leq \cdots \leq 2^{j-k} \omega(\varphi(n_k)).$$

This implies

$$\sum_{j=k}^{\infty} \omega(\varphi(n_j)) \leq \sum_{j=k}^{\infty} 2^{j-k} \omega(\varphi(n_k)) = 2 \omega(\varphi(n_k)).$$

With this estimate (and because of $\lim_{k \to \infty} \omega(\varphi(n_k)) = 0$), it is easy to see that $(g_m)_{m=1}^\infty \in X$ is a Cauchy sequence that converges to $f_\omega$ in Banach space $X$: For a given $\varepsilon > 0$, there exists a number $N_0(\varepsilon)$ such that $\omega(\varphi(n_k)) < \varepsilon/(2C_1)$ for all $k > N_0$. Then, due to $(3.11)$, for all $k > i > n_0$:

$$\|g_k - g_i\| \leq \sum_{j=i+1}^{k} \omega(\varphi(n_j)) \|h_{n_j}\| \leq 2C_1 \omega(\varphi(n_{i+1})) < \varepsilon.$$

Thus, the Banach condition is fulfilled and counter example $f_\omega \in X$ is well defined.

Smoothness condition $(3.11)$ is proved in two cases. The first case covers numbers $\delta > 0$ for which $\mu(\delta) \leq \varphi(n_1)$. Since $\lim_{x \to \infty} \varphi(x) = 0$, there exists $k \in \mathbb{N}$ such that in this case $\varphi(n_k) < \mu(\delta)$. Using this index $k$ in connection with the two bounds in $(3.11)$, one obtains for sub-linear functional $S_\delta$

$$S_\delta(f_\omega) \leq \sum_{j=1}^{k} \omega(\varphi(n_j)) S_\delta(h_{n_j}) + \sum_{j=k+1}^{\infty} \omega(\varphi(n_j)) S_\delta(h_{n_j})$$

$$\leq C_2 \left( \sum_{j=1}^{k-1} \omega(\varphi(n_j)) \frac{\mu(\delta)}{\varphi(n_j)} \right) + \omega(\varphi(n_k)) \frac{\mu(\delta)}{\varphi(n_k)} + C_2 \sum_{j=k+1}^{\infty} \omega(\varphi(n_j))$$

$$\leq 2C_2 \mu(\delta) \frac{\omega(\varphi(n_k))}{\varphi(n_k)} + 2C_2 \omega(\varphi(n_{k+1}))$$

$$\leq 2C_2 \mu(\delta) \frac{\omega(\varphi(n_k))}{\varphi(n_k)} + 2C_2 \omega(\mu(\delta)).$$

The last estimate holds true because

$$\omega(\varphi(n_{k+1}) \leq \omega(\mu(\delta))$$

due to $\varphi(n_{k+1}) < \mu(\delta)$. The first expression in $(3.27)$ can be estimated against $4C_2 \omega(\mu(\delta))$: Because $\mu(\delta) \leq \varphi(n_k)$, one can apply $(3.11)$ to obtain

$$\frac{\omega(\varphi(n_k))}{\varphi(n_k)} \leq 2 \frac{\omega(\mu(\delta))}{\mu(\delta)}.$$
Thus, \( S_6(f_\omega) \leq 6C_2\omega(\mu(\delta)) \).

The second case is \( \mu(\delta) > \varphi(n_1) \). In this situation, let \( k := 0 \). Then only the second sum in (3.27) has to be considered: \( S_6(f_\omega) \leq 2C_2\omega(\mu(\delta)) \).

The little-o condition remains to be proven without sub-additivity. From (3.5) one obtains \( E_{2n}(f) = E_{2n}(f + g) \leq E_n(f + g) + E_n(-g) \), i.e.,

\[
E_n(f + g) \geq E_{2n}(f) - E_n(-g) \geq E_{2n}(f) - E_n(g).
\]

The estimate can be used to show the desired lower bound based on resonance condition (3.12):

\[
E_{nk}(f_\omega) = E_{nk} \left( \omega(\varphi(n_k))h_{nk} + \sum_{j=1}^{k-1} \omega(\varphi(n_j))h_{nj} + \sum_{j=k+1}^{\infty} \omega(\varphi(n_j))h_{nj} \right)
\]

\[
\geq E_{2nk} \left( \omega(\varphi(n_k))h_{nk} + \sum_{j=k+1}^{\infty} \omega(\varphi(n_j))h_{nj} \right) - E_{nk} \left( \sum_{j=1}^{k-1} \omega(\varphi(n_j))h_{nj} \right)
\]

\[
\geq \omega(\varphi(n_k))c_3 - D_{2nk} \left( \sum_{j=k+1}^{\infty} \omega(\varphi(n_j)) \|h_{nj}\|_X \right) - \frac{\omega(\varphi(n_k))}{k}
\]

\[
\geq \omega(\varphi(n_k))c_3 - D_{2nk} C_1 \omega(\varphi(n_{k+1})) - \frac{\omega(\varphi(n_k))}{k}
\]

\[
\geq \left( c_3 - \frac{2C_1 + 1}{k} \right) \omega(\varphi(n_k)).
\]

Thus \( E_n(f_\omega) \neq o(\omega(\varphi(n))) \).

\[\square\]

4 Sharpness

Free knot spline function approximations by Heaviside, cut and ReLU functions are first examples for application of Theorem 3.1.

Let \( S_n^\omega \) be the space of functions \( f \) for which \( n + 1 \) intervals \( [x_k, x_{k+1}], 0 = x_0 < x_1 < \cdots < x_{n+1} = 1 \), exist such that \( f \) equals (potentially different) polynomials \( p \) of degree less than \( r \) on each of these intervals, i.e. \( p \in \Pi^r \). No additional smoothness conditions are required at knots.

**Corollary 4.1 (Free Knot Spline Approximation).** For \( r, \tilde{r} \in \mathbb{N}, 1 \leq p \leq \infty, \) and for each abstract modulus of smoothness \( \omega \) satisfying (3.2) and (3.13), there exists a counter example \( f_\omega \in X^r[0,1] \) such that

\[
\omega_r(f_\omega, \delta)_p = O(\omega(\delta^r)),
\]

\[
E(S_n^\omega, f_\omega)_p := \inf\{\|f_\omega - g\|_{X^p[0,1]} : g \in S_n^\omega \} \neq o\left(\omega\left(\frac{1}{n^{\tilde{r}}}\right)\right).
\]
Note that $r$ and $\tilde{r}$ can be chosen independently. This corresponds with Marchaud inequality for moduli of smoothness.

The following lemma helps in the proof of this and the next corollary. It is used to show the resonance condition of Theorem 3.3.

**Lemma 4.1.** Let $g : [0, 1] \to \mathbb{R}$, and 0 = $x_0 < x_1 < \cdots < x_{N+1} = 1$. Assume that for each interval $I_k := (x_k, x_{k+1})$, 0 ≤ $k$ ≤ $N$, either $g(x) \geq 0$ for all $x \in I_k$ or $g(x) \leq 0$ for all $x \in I_k$ holds. Then $g$ can change its sign only at points $x_k$. Let $h(x) := \sin(2N \cdot 2\pi \cdot x)$. Then there exists a constant $c > 0$ that is independent of $g$ and $N$ such that

$$\|h - g\|_{L^p[0, 1]} \geq c > 0.$$

**Proof.** We discuss $2N$ intervals $A_k := (k(2N)^{-1}, (k + 1/2)(2N)^{-1})$, 0 ≤ $k$ < $2N$. Function $g$ can change its sign at most in $N$ of these intervals. Let $J \subset \{0, 1, \ldots, 2N\}$ the set of indices $k$ of the at least $N$ intervals $A_k$ on which $g$ is non-negative or non-positive. On each of these intervals, $h$ maps to both its maximum 1 and its minimum −1. Thus $\|h - g\|_{L^p[0, 1]} \geq 1$. This shows the Lemma for $p = \infty$. Functions $h$ and $g$ have different sign on $(a, b)$ where $(a, b) = ((k+1/2)(2N)^{-1}, (k+1/2)(2N)^{-1})$, $1 \leq k \leq 2N - 1$.

Thus, for $1 \leq p < \infty$,

$$\|h - g\|_{L^p[0, 1]} \geq \left[ \sum_{k \in J} \int_{A_k} |h - g|^p \right]^\frac{1}{p} \geq \left[ \frac{N}{N4\pi} \int_0^{2\pi} \sin^p(2N \cdot 2\pi \cdot x) \, dx \right]^\frac{1}{p} \geq \frac{N}{N4\pi} \int_0^{2\pi} \sin^p((4n \cdot 2\pi \cdot (x - a)) \, du \right]^\frac{1}{p} =: c > 0.$$

by Corollary 4.3 Theorem 3.1 can be applied with following parameters. Let Banach-space $X = X^p[0, 1]$.

$$E_n(f) := E(S^n, f)_p, \quad S_\delta(f) := \omega_r(f, \delta)_p.$$

Whereas $S_\delta$ is a sub-linear, bounded functional, errors of best approximation $E_n$ fulfill conditions (3.5), (3.6), (3.7), and (3.8). Let $\|\cdot\|_{X^p[0, 1]}$ be defined as in (3.1), with $D_n = 1$. Let $\mu(\delta) := \delta^r$ and $\varphi(x) = 1/x^r$ such that condition (3.9) holds: $\varphi(x) = \varphi(x)/x^r$. Resonance elements

$$h_n(x) := \sin((4n + 1)2\pi x)$$

obviously satisfy condition (3.10): $\|h_n\|_{X^p[0, 1]} \leq 1 =: C_1$. One obtains (3.11) because of

$$S_\delta(h_n) = \omega_r(h_n, \delta)_p \leq 2^r \|h_n\|_{X^p[0, 1]} \leq 2^r \quad \text{and (see (3.11))}$$

$$S_\delta(h_n) = \omega_r(h_n, \delta)_p \leq \delta^r \|h_n\|_{X^p[0, 1]} \leq (\tilde{r}2\pi)^r \delta^r (4n + 1)^r \leq (\tilde{r}2\pi)^r \delta^r (5n)^r = (\tilde{r}10\pi)^r \frac{\mu(\delta)}{\varphi(n)}.$$

Let $g \in S^{l}_{\delta, n}$, then $g$ is composed from at most $4n + 1$ polynomials on $4n + 1$ intervals. On each of these intervals, $g \equiv 0$ or $g$ at most has $\tilde{r} - 1$ zeroes. Thus $g$ can change sign at $4n$ interval borders and at zeroes of polynomials, and $g$ fulfills the prerequisites.
of Lemma 4.1 with \( N := (4n + 1) \cdot \tilde{r} > 4n + (4n + 1) \cdot (\tilde{r} - 1) \). Due to the lemma, \( \| h_n - g \|_{X \rho_1,1} \geq c > 0 \) independent of \( n \) and \( g \). Since this holds true for all \( g \), (3.12) is shown for \( c_4 = c \). All preliminaries of Theorem 3.1 are fulfilled such that counter examples exist as stated.

Since \( \Phi_{n, \sigma_b} \subset S_n^1 \), Corollary 4.1 directly shows sharpness of (1.4) for the Heaviside activation function if one chooses \( r = \tilde{r} = 1 \). Sharpness of (1.3) for cut and ReLU function follows for \( r = \tilde{r} = 2 \) because \( \Phi_{n, \sigma_c} \subset S_n^2, \Phi_{n, \sigma_r} \subset S_n^2 \). However, the case \( \omega(\delta) = \delta \) of maximum non-saturated convergence order is excluded by condition (3.13).

We discuss this case for \( r = \tilde{r} \). Then a simple counter example is \( f_\infty(x) := x^r \). For each sequence of coefficients \( d_0, \ldots, d_{r-1} \in \mathbb{R} \) we can apply the fundamental theorem of algebra to find complex zeroes \( a_0, \ldots, a_{r-1} \in \mathbb{C} \) such that

\[
\left| x^r - \sum_{k=0}^{r-1} d_k x^k \right| = \prod_{k=0}^{r-1} |x - a_k|.
\]

There exists an interval \( I := ((j + 1) - 1, (j + 1) - 1) \subset [0, 1] \) such that real parts of complex numbers \( a_k \) are not in \( I \) for all \( 0 \leq k < r \). Let \( I_0 := ((j + 1) - 1, (j + 3/4) - 1) \subset I \). Then for all \( x \in I_0 \)

\[
\prod_{k=0}^{r-1} |x - a_k| \geq \left[ \frac{1}{4(r + 1)} \right]^r := c_\infty > 0.
\]

This lower bound is independent of coefficients \( d_k \) such that

\[
\inf \{ \| x^r - q(x) \|_{B_{[0,1]}^r} : q \in \Pi^r \} \geq c_\infty > 0.
\]

We also see that

\[
\int_0^1 \left| x^r - \sum_{k=0}^{r-1} d_k x^k \right|^p dx \geq \int_0^1 \frac{1}{4(r + 1)^{pr}} dx = \frac{1}{2(r + 1)^{pr}} := c_p^r > 0,
\]

\[
\inf \{ \| x^r - q(x) \|_{L_p^\infty} : q \in \Pi^r \} \geq c_p > 0.
\]

Each function \( g \in S_n^r \) is a polynomial of degree less than \( r \) on at least \( n \) intervals \( (j(2n) - 1, (j + 1)(2n) - 1), j \in J \subset \{ 0, 1, \ldots, n - 1 \} \). For \( j \in J \):

\[
\inf_{q \in \Pi^r} \| x^r - q(x) \|_{B_{[2n, 2n]}^{r+1}} = \inf_{q \in \Pi^r} \left\{ \left( \frac{x}{2n} + \frac{j}{2n} \right)^r - q(x) \right\}_{B_{[0,1]}^r} \geq \frac{c_\infty}{(2n)^r}.
\]

Thus, \( E(S_n^r, x^r) \neq o \left( \frac{1}{n^r} \right) \). In case of \( L^p \)-spaces, we similarly obtain with substitution \( u = 2nx - j \)

\[
\inf_{q \in \Pi^r} \int_{2n}^{2n+2} |x^r - q(x)|^p dx = \inf_{q \in \Pi^r} \frac{1}{2n} \int_0^1 \left( \frac{u}{2n} + \frac{j}{2n} \right)^r - q(u) \right|_{B_{[0,1]}^r} \geq \frac{c_p^r}{2n \cdot (2n)^{pr}}.
\]

Sharpness is demonstrated by combining lower estimates of all \( n \) subintervals:

\[
E(S_n^r, x^r)_p \geq n \frac{c_p^r}{(2n)^{pr+1}}, \quad E(S_n^r, x^r)_p \geq \frac{c_p}{2^{r+\frac{1}{r} + 1}} \neq o \left( \frac{1}{n^r} \right).
\]
Although our counter example is arbitrarily often differentiable, the convergence order is limited to $n^{-\alpha}$. Reason is the piecewise definition of the activation function by polynomials. There is no such limitation for activation functions that are arbitrarily often differentiable on an interval without being a polynomial, see Theorem \ref{thm:approximation}. Thus, neural networks based on smooth non-polynomial activation functions might approximate better if smooth functions have to be learned.

Theorem 3 in \cite{4} states for the Heaviside function that for each $n \in \mathbb{N}$ a function $f_n \in C[0,1]$ exits such that the error of best uniform approximation exactly equals $\omega_1 \left( f_n, \frac{1}{2^{(n+1)}} \right)$. This is used to show optimality of the constant. Functions $f_n$ might be different for different $n$. One does not get the condensed sharpness result of Corollary \ref{cor:sharpness}.

Another example for the application of Corollary \ref{cor:sharpness} is the square non-linearity $g$ is not constant then

\begin{align*}
\left| f \right|_r &= \frac{1}{r} \left( \left| f \right|_r \right)_r \\
&\leq 2. \because \\
\text{Because } g \text{ is not constant then}
\end{align*}

Thus, neural networks based on smooth non-polynomial activation functions might approximate better if smooth functions have to be learned.

Thus, error bound in terms of moduli of smoothness are not able to express the advantages of non-linear free knot spline approximation in contrast to fixed knot spline approximation (cf. \cite{5}). For an error measured in an $L^p$ norm with an order like $n^{-\alpha}$, smoothness only is required in $L^p$, $q := 1/((\alpha + 1)/p)$, see \cite{5} and \cite[p. 368]{5}.

**Corollary 4.2** (Inverse Tangent). Let $\sigma = \sigma_a$ be the sigmoid function based on the inverse tangent function, $r \in \mathbb{N}$, and $1 \leq p \leq \infty$. For each abstract modulus of smoothness $\omega$ satisfying \ref{eq:modulus} and \ref{eq:modulus2}, there exists a counter example $f_\omega \in X^p [0,1]$ such that

\begin{align*}
\omega_1(f_\omega, \delta)_p = O \left( \omega(\delta^a) \right) \quad \text{and} \quad E(\Phi_{n, \sigma_a}, f_\omega)_p \neq o \left( \omega \left( \frac{1}{n^r} \right) \right).
\end{align*}

The corollary shows sharpness of the error bound in Theorem \ref{thm:approximation} applied to the arbitrarily often differentiable function $\sigma_a$.

**Proof.** Similarly to the proof of Corollary \ref{cor:sharpness} we apply Theorem \ref{thm:approximation} with parameters $X = X^p[0,1]$, $E_n(f) := E(\Phi_{n, \sigma_a}, f)_p$, $S_\delta(f) := \omega_1(f, \delta)_p$, $\mu(\delta) := \delta^\alpha$, $\sigma(x) = 1/x^\alpha$, and $h_a(x) := \sin(16n \cdot 2\pi x)$ such that condition \ref{eq:condition} is obvious and \ref{eq:condition2} can be shown by estimating the modulus against the $r$th derivative of $h_n$ with \ref{eq:condition}. Let $g \in \Phi_{4n, \sigma_a}$,

\begin{align*}
g(x) &= \sum_{k=1}^{4n} a_k \left( \frac{1}{2} + \frac{1}{\pi} \arctan(b_k x + c_k) \right).
\end{align*}

Then

\begin{align*}
g'(x) &= \sum_{k=1}^{4n} a_k b_k \frac{1}{\pi} \frac{1}{1 + (b_k x + c_k)^2} - \frac{s(x)}{q(x)}
\end{align*}

where $s(x)$ is a polynomial of degree $2(4n-1)$, and $q(x)$ is a polynomial of degree $8n$. If $g$ is not constant then $g'$ at most has $8n - 2$ zeroes and $f$ at most has $8n - 1$ zeroes due to the mean value theorem (Rolle's theorem). In both cases, the requirements of Lemma \ref{lem:counterexample} are fulfilled with $N := 8n > 8n - 1$ such that $\| h_n - g \|_{X^p[0,1]} \geq c > 0$ independent of $n$ and $g$. Since $g$ can be chosen arbitrarily, \ref{thm:counterexample} is shown with $E_{4n} h_n \geq c > 0$. \qed
Whereas lower estimates for sums of $n$ inverse tangent functions are easily obtained by considering $O(n)$ zeroes of their derivatives, sums of $n$ logistic functions (or hyperbolic tangent functions) might have an exponential number of zeroes. To illustrate the problem in the context of Theorem 3.1, let

$$g(x) := \sum_{k=1}^{4n} \frac{a_k}{1 + e^{-c_k (c_k - b_k)} x} \in \Phi_{4n, \sigma}.$$  \hspace{1cm} (4.1)

Using a common denominator, the numerator is a sum of type $\sum_{k=1}^{m} \alpha_k \kappa_k^4$ for some $\kappa_k > 0$ and $m < 2^{1n}$. According to [37], such a function has at most $m - 1 < 16^n - 1$ zeroes, or it equals the zero function. Therefore, an interval $[k(16)^{-n}, (k + 1)(16)^{-n}]$ exists on which $g$ does not change its sign. Thus using a resonance sequence $h_n(x) := \sin (16^n \cdot 2\pi x)$, one gets $E(\Phi_{4n, \sigma}, h_n) \geq 1$. But factor $16^n$ is by far too large. One has to choose $\phi(x) := 1/16^n$ and $\mu(\delta) := \delta$ to obtain a “counter example” $f_\omega$ with

$$E(\Phi_{n, \sigma}, f_\omega) \in O \left( \omega \left( \frac{1}{n} \right) \right) \text{ and } E(\Phi_{n, \sigma}, f_\omega) \neq o \left( \omega \left( \frac{1}{16^n} \right) \right). \hspace{1cm} (4.2)$$

The gap between rates is obvious. The same difficulties do not only occur for the logistic function but also for other activation functions based on $\exp(x)$ like the softmax function $\sigma_m(x) := \log(\exp(x) + 1)$. Similar to (4.2),

$$\frac{\partial}{\partial c} \sigma_m(bx + c) = \frac{\exp(bx + c)}{\exp(bx + c) + 1} = \sigma_l(bx + c).$$

Thus, sums of $n$ logistic functions can be approximated uniformly and arbitrarily well by sums of differential quotients that can be written by $2n$ softmax functions. A lower bound for approximation with $\sigma_m$ would also imply a similar bound for $\sigma_l$ and upper bounds for approximation with $\sigma_l$ imply upper bounds for $\sigma_m$.

With respect to the logistic function, a better estimate than (4.2) is possible based on the Vapnik-Chervonenkis dimension (VC dimension) of related function spaces. The VC dimension $\text{VC-dim}(V)$ of a (non-linear) set $V$ of functions $g : X \to \mathbb{R}$ on a set $X \subset \mathbb{R}$ is defined as the largest number $m \in \mathbb{N}$ (if exists) for which $m$ points $x_1, \ldots, x_m \in X$ exist such that for each sign sequence $s_1, \ldots, s_m \in \{-1, 1\}$ a function $g \in V$ can be found that fulfills

$$\sigma_n(g(x_i)) = s_i, \hspace{1cm} 1 \leq i \leq m.$$  

As before, $\sigma_n$ denotes the Heaviside-function. The VC dimension is an indicator for the number of degrees of freedom in the construction of $V$.

**Corollary 4.3** (Sharpness due to VC Dimension). Let $(V_n)_{n=1}^{\infty}$ be a sequence of (non-linear) function spaces $V_n \subset B[0,1]$ such that

$$E_n(f) := \inf \{ ||f - g||_{B[0,1]} : g \in V_n \}$$

fulfills conditions (3.9)–(3.11). Let $\tau : \mathbb{N} \to \mathbb{N}$ and

$$G_n := \left\{ \frac{j}{\tau(n)} : j \in \{0, 1, \ldots, \tau(n)\} \right\}$$

be an equidistant grid on the interval $[0, 1]$. We restrict functions in $V_n$ to this grid:

$$V_{n, \tau(n)} := \{ h : G_n \to \mathbb{R} : h(x) = g(x) \text{ for a function } g \in V_n \}.$$
Let function \( \varphi(x) \) be defined as in Theorem 3.1 such that 3.3 holds true. If, for a constant \( C > 0 \), function \( \tau \) fulfills

\[
\begin{align*}
\text{VC-dim}(V_{n, \tau(n)}) & < \tau(n), \\
\tau(4n) & \leq \frac{C}{\varphi(n)},
\end{align*}
\]

for all \( n \geq n_0 \in \mathbb{N} \) then for \( r \in \mathbb{N} \) and each abstract modulus of smoothness \( \omega \) satisfying 3.3 and 3.13, a counter example \( f_\omega \in C[0,1] \) exists such that

\[ \omega_r(f_\omega, \delta) = O(\omega(\delta^r)) \text{ and } E_{\omega}(f_\omega) \neq o(\omega(\varphi(n)^r)). \]

**Proof.** Let \( n \geq n_0/4 \). Due to 4.3, a sign sequence \( s_0, \ldots, s_{\tau(4n)} \in \{0,1\} \) exists such that for each function \( g \in V_{4n} \) there is a point \( x_0 = \frac{\tau(4n)}{\tau(4n)} \in G_{4n} \) such that \( \sigma_h(g(x_0)) \neq s_i \).

We utilize this sign sequence to construct resonance elements. It is well known, that auxiliary function

\[ h(x) := \begin{cases} 
\exp\left(1 - \frac{1}{1+x^2}\right) & \text{for } |x| < 1, \\
0 & \text{for } |x| \geq 1,
\end{cases} \]

is arbitrarily often differentiable on the real axis, \( h(0) = 1, \|h\|_{1(\mathbb{R})} = 1 \). This function becomes the building block for the resonance elements:

\[ h_n(x) := \sum_{i=0}^{\tau(4n)} s_i \cdot h\left(2\tau(4n)\left(x - \frac{i}{\tau(4n)}\right)\right). \]

The interior of the support of summands is non-overlapping, i.e., \( \|h_n\|_{1[0,1]} \leq 1 \), and because of 4.3 norm \( \|h_{n,r}\|_{1[0,1]} \) of the \( r \)th derivative is in \( O(|\varphi(n)|^{-r}) \).

We apply Theorem 3.1 with \( X = C[0,1], S_h(f) := \omega_r(f, \delta), \mu(\delta) := \delta^r \), \( E_{\omega}(f) \), as defined in the theorem, and resonance elements \( h_n(x) \) that represent the existing sign sequence. Function \( [\varphi(x)]^r \) fulfills the requirements of function \( \varphi(x) \) in Theorem 3.1.

Then conditions 3.10 and 3.11 are fulfilled due to the norms of \( h_n \) and its derivatives, cf. 1.14. Due to the initial argument of the proof, for each \( g \in V_{4n} \) at least one point \( x_0 = \tau(4n)^{-1}, 0 \leq i \leq \tau(4n) \), exists such that we observe \( \sigma_h(g(x_0)) \neq \sigma_h(h_n(x_0)) \). Since \( |h_n(x_0)| = 1 \), we get \( \|h_n - g\|_{1[0,1]} \geq |h_n(x_0) - g(x_0)| \geq 1 \), and \( E_{\omega}(h_n) \geq 1 \).

**Corollary 4.4** (Logistic Function). Let \( \sigma = \sigma_l \) be the logistic function and \( r \in \mathbb{N} \). For each abstract modulus of smoothness \( \omega \) satisfying 3.3 and 3.13, a counter example \( f_\omega \in C[0,1] \) exists such that

\[ \omega_r(f_\omega, \delta) = O(\omega(\delta^r)) \text{ and } E(\Phi_{n, \sigma_l}, f_\omega) \neq o\left(\omega\left(\frac{1}{n(1 + \log_2(n))^r}\right)\right). \]

The corollary extends the Theorem of Maiorov and Meir for worst case approximation with sigmoid functions in the case \( p = \infty \) to Lipschitz classes, see [31] p. 176.
Proof. We apply Corollary 4.3 in connection with a result concerning the VC dimension of function space \( (D \in \mathbb{N}) \)
\[
\Phi^{D}_{n,\sigma_{l}} := \left\{ g : \{-D, -D+1, \ldots, D\} \to \mathbb{R} : g(x) = a_{0} + \sum_{k=1}^{n} a_{k}\sigma_{l}(b_{k}x + c_{k}) : a_{0}, a_{k}, b_{k}, c_{k} \in \mathbb{R} \right\},
\]
see paper [5] that is based on [22], cf. [25]. Functions are defined on a discrete set with \( 2D + 1 \) elements, and in contrast to the definition of \( \Phi^{n,\sigma} \), a constant function with coefficient \( a_{0} \) is added.

According to Theorem 2 in [5], the VC dimension of \( \Phi^{D}_{n,\sigma_{l}} \) is upper bounded by (\( n \) large enough)
\[
2 \cdot (3 \cdot n + 1) \cdot \log_{2}(24e(3 \cdot n + 1)D),
\]
i.e., there exists \( 2 \leq n_{0} \in \mathbb{N} \) and a constant \( C > 0 \) such that for all \( n \geq n_{0} \)
\[
\text{VC-dim}(\Phi^{D}_{n,\sigma_{l}}) \leq Cn[\log_{2}(n) + \log_{2}(D)].
\]
Since \( \lim_{E \to \infty} \frac{1 + \log_{2}(E)}{E} = 0 \), we can choose a constant \( E > 1 \) such that
\[
\frac{1 + \log_{2}(E)}{E} < \frac{1}{4C}.
\]
With this constant, we define \( D = D(n) := [En(1 + \log_{2}(n))] \) such that the VC dimension of \( \Phi^{D}_{n,\sigma_{l}} \) is less than \( D \) for \( n \geq n_{0} \):
\[
\text{VC-dim}(\Phi^{D}_{n,\sigma_{l}}) \leq Cn[\log_{2}(n) + \log_{2}(En(1 + \log_{2}(n)))]
\leq Cn[2\log_{2}(n) + \log_{2}(E) + \log_{2}(2\log_{2}(n))]
\leq Cn[3\log_{2}(n) + \log_{2}(E) + 1] \leq 4Cn \log_{2}(n)[1 + \log_{2}(E)]
\leq En \log_{2}(n) \leq [En(1 + \log_{2}(n))] = D.
\]

By applying an affine transform that maps interval \([-D, D]\) to \([0, 1]\) and by omitting constant function \( a_{0} \), we immediately see that for \( V_{n} := \Phi_{n,\sigma} \) and \( \tau(n) := 2D(n) \)
\[
\text{VC-dim}(V_{n,\tau(n)}) < \tau(n)
\]
such that (4.3) is fulfilled.

We define strictly decreasing
\[
\varphi(x) := \frac{1}{x[1 + \log_{2}(x)]}.
\]
Obviously, \( \lim_{x \to \infty} \varphi(x) = 0 \). Condition (3.9) holds: Let \( x > X_{0}(\lambda) := \lambda^{-2} \). Then \( \log_{2}(\lambda) > -\log_{2}(x)/2 \) and
\[
\varphi(\lambda x) = \frac{1}{\lambda x[1 + \log_{2}(x) + \log_{2}(\lambda)]} \leq \frac{1}{\lambda x[1 + \frac{1}{2}\log_{2}(x)]} < \frac{2}{\lambda} \frac{1}{x} = 2 \frac{1}{\varphi(x)}.
\]

Also, (4.4) is fulfilled:
\[
\tau(4n) = 2D(4n) = 2E \cdot 4n(1 + \log_{2}(4n)) < \frac{8E(1 + \log_{2}(4))}{\varphi(n)} = \frac{24E}{\varphi(n)}.
\]

Thus, all prerequisites of Corollary 4.3 are shown. \( \square \)
The corollary improves (1.2): There exists a counter example \( f_\omega \in C[0, 1] \) such that (see (2.2), (2.8))

\[
\omega_r(f_\omega, \delta) \in O(\omega(\delta^r)),
\]

\[
E(\Phi_{n, \sigma_1}, f_\omega) \in O\left(\omega\left(\frac{1}{n^r}\right)\right) \quad \text{and} \quad E(\Phi_{n, \sigma_1}, f_\omega) \neq o\left(\omega\left(\frac{1}{n^{r\left[1 + \log_2(n)r\right]}\right)\right). \tag{4.6}
\]

The preceding corollary is a prototype for proving sharpness based on known VC dimensions. Also at the price of a log-factor, the VC dimension estimate for radial basis functions in \([5]\) or \([35]\) can be used similarly in connection with Corollary 4.3 to construct counter examples. The sharpness results for Heaviside, cut, ReLU and inverse tangent activation functions shown above for \( p = \infty \) can also be obtained with Corollary 4.3 by proving that VC dimensions of corresponding function spaces \( \Phi_{n, \sigma} \) are in \( O(n) \). This in turn can be shown by estimating the maximum number of zeroes like in the proof of the next corollary and in the same manner as in the proofs of Corollaries 4.1 and 4.2.

The problem of different rates in upper and lower bounds arises because different scaling coefficients \( b_k \) are allowed. In the case of uniform scaling, i.e. all coefficients \( b_k \) in (2.1) have the same value \( b_k = B = B(n) \), the number of zeroes is bounded by \( 4n - 1 \) instead of \( 16^n - 2 \). Let

\[
\tilde{\Phi}_{n, \sigma_1} := \left\{ g : [0, 1] \to \mathbb{R} : g(x) = \sum_{k=1}^{n} a_k \sigma(Bx + c_k) : a_k, B, c_k \in \mathbb{R} \right\}
\]

be the non-linear function space generated by uniform scaling. Because the quasi-interpolation operators used in the proof of direct estimate (1.2) are defined using such uniform scaling, see [8, p. 172], the error bound

\[
E(\tilde{\Phi}_{n, \sigma_1}, f) := \inf\{\|f - g\|_{B[0, 1]} : g \in \tilde{\Phi}_{n, \sigma_1}\} \leq \omega_1\left(f, \frac{1}{n}\right)
\]

holds. This bound is sharp:

**Corollary 4.5 (Logistic Function with Restriction).** Let \( \sigma = \sigma_1 \) be the logistic function and \( r \in \mathbb{N} \). For each abstract modulus of smoothness \( \omega \) satisfying (2.8) and (6.8), there exists a counter example \( f_\omega \in C[0, 1] \) such that

\[
\omega_r(f_\omega, \delta) = O(\omega(\delta^r)) \quad \text{and} \quad E(\tilde{\Phi}_{n, \sigma_1}, f_\omega) \neq o\left(\omega\left(\frac{1}{n^r}\right)\right).
\]

To prove the corollary, we apply following lemma.

**Lemma 4.2.** \( V_n \subset C[0, 1], 2 \leq \tau(n) \in \mathbb{N}, G_n := \left\{ \frac{j}{\tau(n)} : j \in \{0, 1, \ldots, \tau(n)\} \right\}, \) and \( V_{n, \tau(n)} := \{ h : G_n \to \mathbb{R} : h(x) = g(x) \text{ for a function } g \in V_n \} \) are given as in Corollary 4.3. If VC-dim \( \left(V_{n, \tau(n)}\right) \geq \tau(n) \) then there exists a function \( g \in V_n, g \neq 0, \) with a set of at least \( \lfloor \tau(n)/2 \rfloor \) zero points in \( [0, 1] \) such that \( g \) has non-zero function values between each two consecutive points of this set.

**Proof.** Because of the VC dimension, a subset \( J \subset \{0, 1, \ldots, \tau(n)\} \) with \( \tau(n) \) elements \( j_1 < j_2 < \cdots < j_{\tau(n)} \) and a function \( g \in V_n \) exist such that

\[
\sigma_n\left(g, \frac{j_k}{\tau(n)}\right) = \frac{1 + (-1)^{k+1}}{2}, \quad 1 \leq k \leq \tau(n).
\]
Then we find a zero on each interval \([j_{2k}^{-1} \tau(n)^{-1}, j_{2k+1}^{-1} \tau(n)^{-1}]\), \(1 \leq k \leq \lfloor \tau(n)/2 \rfloor\): If \(g(j_{2k-1}^{-1} \tau(n)^{-1}) \neq 0\) then on the interval \((j_{2k-1}^{-1} \tau(n)^{-1}, j_{2k}^{-1} \tau(n)^{-1})\) continuous \(g\) has a zero. Thus, \(g\) has \(\lfloor \tau(n)/2 \rfloor\) zeroes on different sub-intervals. Between zeroes are non-zero, negative function values \(g(j_k^{-1} \tau(n)^{-1})\) for every \(k\) because \(\sigma_k(g(j_k^{-1} \tau(n)^{-1})) = 0\).

of Corollary \ref{cor:4.3} We apply Corollary \ref{cor:4.3} with \(V_n = \tilde{\Phi}_{n, \sigma_l}\) and \(E_n(f) = E(\tilde{\Phi}_{n, \sigma_l}, f)\) such that conditions \ref{eq:5.3}–\ref{eq:5.8} are fulfilled. Let \(\tau(n) := 2n\) and \(\varphi(n) = 1/n\) such that \ref{eq:4.3} holds true: \(\tau(4n) = 8n = 8/\varphi(n)\). Assume that \(\text{VC-dim}(V_{n, \tau(n)}) \geq \tau(n)\), then according to Lemma \ref{lem:4.2} there exists a function \(f \in \tilde{\Phi}_{n, \sigma_l}\) such that \(f \neq 0\) has \(\lfloor \tau(n)/2 \rfloor\) zeroes. However, we can write \(f\) as

\[
 f(x) = \sum_{k=1}^{n} \frac{a_k}{1 + e^{-\gamma k} (e^{-B})^x} = \frac{s(x)}{q(x)}
\]

Using a common denominator \(q(x)\), the numerator is a sum of type

\[
 s(x) = \sum_{k=0}^{n-1} a_k (e^{-kB})^x
\]

which has at most \(n - 1\) zeroes, see \ref{eq:7.4}. Because of this contradiction to \(n\) zeroes, \ref{eq:10.3} is fulfilled.

If one allows at most \(M\), independent of \(n\), coefficients \(b_k\) that are different from \(B\), the number of zeroes of \ref{eq:1.1} is bounded by \(2^M(4(n - M) - 1)\) for \(n > M\). Then sharpness can be established as before. However, if one applies the proof of Theorem \ref{thm:2.1} to the special case of the logistic function then a sum is constructed in which the number of coefficients \(b_k\) depends on \(n\).

A result similar to Corollary \ref{cor:5.3} can also be shown for the ELU function \(\sigma_e\). However, without the restriction to \(b_k\), piecewise superposition of exponential functions leads to \(O(n^2)\) zeroes of sums of ELU functions. Then in combination with direct estimates Theorems \ref{thm:2.1} and \ref{thm:2.2} i.e., \(E(\tilde{\Phi}_{n, \sigma_e}, f) \leq C_{r, \omega_r}(f, \omega, \frac{1}{n})\), only following result can be shown in a straightforward manner.

Corollary 4.6 (Coarse estimate for ELU activation). Let \(\sigma = \sigma_e\) be the ELU function and \(r \in \mathbb{N}, n \geq \max\{2, r\}\) (see Theorem \ref{thm:2.1}). For each abstract modulus of smoothness \(\omega\) satisfying \ref{eq:6.4} and \ref{eq:6.5}, there exists a counter example \(f_n \in C[0, 1]\) that fulfills

\[
 E(\tilde{\Phi}_{n, \sigma_e}, f) \leq C_{r, \omega_r}(f, \omega, \frac{1}{n}) \in O\left(\omega\left(\frac{1}{n^r}\right)\right)
\]

\[
 E(\tilde{\Phi}_{n, \sigma_e}, f) \neq o\left(\omega\left(\frac{1}{n^{r'}}\right)\right).
\]

Proof. To prove the existence of a function \(f_\omega\) with \(\omega_\tau(f_\omega, \delta) \in O(\omega(\delta^\tau))\) and \ref{eq:6.4}, we apply Corollary \ref{cor:4.3} with \(V_n = \tilde{\Phi}_{n, \sigma_l}\) and \(E_n(f) = E(\tilde{\Phi}_{n, \sigma_l}, f)\) such that conditions \ref{eq:6.4}–\ref{eq:6.8} are fulfilled. For each function \(g \in V_n\) the interval \([0, 1]\) can be divided into at most \(n + 1\) subintervals such that on the \(l\)th interval \(g\) equals a function \(g_l\) of type

\[
 g_l(x) = \gamma_l + \delta_l x + \sum_{k=1}^{n} \alpha_{l,k} \exp(\beta_{l,k} x).
\]
Derivative
\[ g'(x) = \delta_l \exp(0 \cdot x) + \sum_{k=1}^{n} \alpha_{l,k} \beta_{l,k} \exp(\beta_{l,k} x) \]

has at most \( n \) zeroes or equals the zero function according to [37]. Thus, due to the mean value theorem (or Rolle’s theorem), \( g_l \) has at most \( n + 1 \) zeroes or is the zero function. By concatenating functions \( g_l \) to \( g \), one observes that \( g \) has at most \( (n + 1)^2 \) different zeroes such that \( g \) does not vanish between such consecutive zero points.

Let \( \tau(n) := 8n^2 \) and \( \varphi(n) = 1/n^2 \) such that (4.4) holds true: \( \tau(4n) = 128n^2 = 128/\varphi(n) \). If \( \text{VC-dim}(V_n, \tau(n)) \geq \tau(n) \) then due to Lemma [12] and because \( n \geq 2 \) there exists a function in \( \Phi_{n, \sigma} \) with at least \( \lceil \tau(n)/2 \rceil = (2n)^2 > (n + 1)^2 \) zeroes such that between consecutive zeroes, the function is not the zero function. This contradicts the previously determined number of zeroes and (4.3) is fulfilled.

Sums of \( n \) softsign functions \( \varphi(x) = x/(1 + |x|) \) can be expressed piecewise by \( n + 1 \) rational functions that each have at most \( n \) zeroes. Thus, one also has to deal with \( O(n^2) \) zeroes. However, a theorem similar to Corollary [4.4] can be proved with Corollary [4.3] based on VC dimension estimates shown in [4].

5 Conclusions

Corollaries [4.1] and [4.2] can be seen in the context of Lipschitz classes. Let \( r := 1 \) for the Heaviside, \( r := 2 \) for cut and ReLU functions and \( r \in \mathbb{N} \) for the inverse tangent activation function \( \sigma = \sigma_a \). By choosing \( \omega(\delta) := \delta^\alpha \), a counter example \( f_n \in X^p[0,1] \) exists for each \( \alpha \in (0, r) \) such that
\[ \omega_r(f_n, \delta)_p \in O(\delta^\alpha), \quad E(\Phi_{n, \sigma}, f_n)_p \in O\left(\frac{1}{n^\alpha}\right), \quad \text{and} \quad E(\Phi_{n, \sigma}, f_n)_p \neq o\left(\frac{1}{n^\alpha}\right). \]

With respect to Corollary [4.4] for the logistic function, in general no higher convergence order \( \omega_r, \beta > \alpha \) can be expected for functions in the Lipschitz class that is defined via \( \text{Lip}^r(\alpha, C[0,1]) := \{ f \in C[0,1] : \omega_r(f, \delta) = O(\delta^\alpha) \} \).

In terms of (non-linear) Kolmogorov \( n \)-width, let \( X := \text{Lip}^r(\alpha, C[0,1]) \). Then, for example, condensed counter examples \( f_a \) for piecewise linear or inverse tangent activation functions and \( p = \infty \) imply
\[
W_n := \inf_{b_1, \ldots, b_n, c_1, \ldots, c_n} \sup_{f \in \text{Lip}^r(\alpha, C[0,1])} \frac{\alpha_{1, \ldots, n} f(\cdot) - \sum_{k=1}^{n} a_k \sigma(b_k \cdot + c_k)}{B[0,1]} \\
\geq \inf_{b_1, \ldots, b_n, c_1, \ldots, c_n} \frac{f_a(\cdot) - \sum_{k=1}^{n} a_k \sigma(b_k \cdot + c_k)}{B[0,1]} \\
= E(\Phi_{n, \sigma}, f_a) \neq o\left(\frac{1}{n^\alpha}\right).
\]

The restriction to the univariate case of a single input node was chosen because of compatibility with cited results and simple one-dimensional definitions of moduli of smoothness. Theorem [3.1] can be applied to similar multivariate problems.

Without additional restrictions, a lower estimate for approximation with logistic function \( \sigma_l \) could only be obtained with a log-factor in (4.6). Thus, either direct bounds [22] or sharpness result [4.6] can be improved. There is also a gap between upper and lower bounds in Corollary [4.3] for the ELU function.
Acknowledgment

I would like to thank Michael Gref, Christian Neumann, Peer Ueberholz, and Lorens Imhof for their valuable comments.

References

[1] Anastassiou, G.A.: Multivariate hyperbolic tangent neural network approximation. Computers & Mathematics with Applications 61(4), 809–821 (2011)
[2] Anastassiou, G.A.: Univariate hyperbolic tangent neural network quantitative approximation. In: Intelligent Systems: Approximation by Artificial Neural Networks, vol. 19, pp. 33–65. Springer, Berlin (2011)
[3] Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory 39(3), 930–945 (1993)
[4] Bartlett, P.L., Maiorov, V., Meir, R.: Almost linear VC-dimension bounds for piecewise polynomial networks. Neural Computation 10(8), 2159–2173 (1998)
[5] Bartlett, P.L., Williamson, R.C.: The VC dimension and pseudodimension of two-layer neural networks with discrete inputs. Neural Computation 8(3), 625–628 (1996)
[6] Chen, D.: Degree of approximation by superpositions of a sigmoid function. Approximation Theory and its Applications 9(3), 17–28 (1993)
[7] Chen, T., Chen, H., Liu, R.: A constructive proof and an extension of cybenkos approximation theorem. In: C. Page, R. LePage (eds.) Computing Science and Statistics, pp. 163–168. Springer, New York, NY (1992)
[8] Chen, Z., Cao, F.: The properties of logistic function and applications to neural network approximation. J. Computational Analysis and Applications 15(6), 1046–1056 (2013)
[9] Cheney, E.W., Light, W.A.: A course in Approximation Theory. AMS, Providence, RI (2000)
[10] Costarelli, D.: Sigmoidal Functions Approximation and Applications. Doctoral thesis, Roma Tre University, Rome (2014)
[11] Costarelli, D., Spigler, R.: Approximation results for neural network operators activated by sigmoidal functions. Neural Networks 44, 101–106 (2013)
[12] Costarelli, D., Spigler, R.: Constructive approximation by superposition of sigmoidal functions. Analysis in Theory and Applications 29, 169–196 (2013)
[13] Costarelli, D., Vinti, G.: Saturation classes for max-product neural network operators activated by sigmoidal functions. Results in Mathematics 72, 1555–1569 (2017)
[14] Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 2(4), 303–314 (1989)
[15] DeVore, R.A., Lorentz, G.G.: Constructive Approximation. Springer, Berlin (1993)
[16] DeVore, R.A., Popov, V.A.: Interpolation spaces and non-linear approximation. In: M. Cwikel, J. Peetre, Y. Sagher, H. Wallin (eds.) Function Spaces and Applications, pp. 191–205. Springer, Berlin, Heidelberg (1988)
[17] Dickmeis, W., Nessel, R.J., van Wickeren, E.: A general approach to counterexamples in numerical analysis. Numerische Mathematik 43(2), 249–263 (1984)
[18] Dickmeis, W., Nessel, R.J., van Wickeren, E.: On nonlinear condensation principles with rates. manuscripta mathematica 52(1), 1–20 (1985)
[19] Dickmeis, W., Nessel, R.J., van Wickeren, E.: Quantitative extensions of the uniform boundedness principle. Jahresber. Deutsch. Math.-Verein. 89, 105–134 (1987)
[20] Funahashi, K.I.: On the approximate realization of continuous mappings by neural networks. Neural Networks 2, 183–192 (1989)
[21] Goebbels, S.: A sharp error estimate for numerical fourier transform of band-limited functions based on windowed samples. Z. Anal. Anwendungen 32, 371–387 (2013)
[22] Goldberg, P.W., Jerrum, M.R.: Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. Machine Learning 18(2), 131–148 (1995)
[23] Guliyev, N.J., Ismailov, V.E.: A single hidden layer feedforward network with only one neuron in the hidden layer can approximate any univariate function. Neural Networks 28(7), 1289–1304 (2016)
[24] Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks 2(5), 359–366 (1989)
[25] Karpinski, M., Macintyre, A.: Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks. Journal of Computer and System Sciences 54(1), 169–176 (1997)
[26] Lei, Y., Ding, L.: Approximation and estimation bounds for free knot splines. Computers & Mathematics with Applications 65(7), 1006–1024 (2013)
[27] Lewicki, G., Marino, G.: Approximation of functions of finite variation by superpositions of a sigmoidal function. Applied Mathematics Letters 17(10), 1147–1152 (2004)
[28] Li, F.J.: Constructive function approximation by neural networks with optimized activation functions and fixed weights. Neural Computing and Applications 31(9), 4613–4628 (2019)
[29] Makovoz, Y.: Uniform approximation by neural networks. Journal of Approximation Theory 95(2), 215–228 (1998)
[30] Pinkus, A.: n-Width in Approximation Theory. Springer, Berlin (1985)
[31] Pinkus, A.: Approximation theory of the MLP model in neural networks. Acta Numerica 8, 143–195 (1999)
[32] Ritter, G.: Efficient estimation of neural weights by polynomial approximation. IEEE Transactions on Information Theory 45(5), 1541–1550 (1999)
[33] Sanguineti, M.: Universal approximation by ridge computational models and neural networks: A survey. The Open Applied Mathematics Journal 2, 31–58 (2008)
[34] Sanguineti, M., Hlaváčková-Schindler, K.: Some comparisons between linear approximation and approximation by neural networks. In: Artificial Neural Nets and Genetic Algorithms, pp. 172–177. Springer, Vienna (1999)
[35] Schmitt, M.: Radial basis function neural networks have superlinear VC dimension. In: Proceedings of the 14th Annual Conference on Computational Learning Theory COLT 2001 and 5th European Conference on Computational Learning Theory EuroCOLT 2001, volume 2111 of Lecture Notes in Artificial Intelligence, pp. 14–30. Springer, Berlin (2001)

[36] Timan, A.: Theory of Approximation of Functions of a Real Variable. Pergamon Press, New York, NY (1963)

[37] Tossavainen, T.: The lost cousin of the fundamental theorem of algebra. Mathematics Magazine 80(4), 290–294 (2007)

[38] Wang, J., Xu, Z.: New study on neural networks: the essential order of approximation. Neural Networks 23(5), 618 – 624 (2010)

[39] Xu, Z., Cao, F.: The essential order of approximation for neural networks. Science in China Series F: Information Sciences 47(1), 97–112 (2004)

[40] Xu, Z., Wang, J.J.: The essential order of approximation for neural networks. Science in China Series F: Information Sciences 49(4), 446–460 (2006)