Approximation Rates for Neural Networks with Encodable Weights in Smoothness Spaces

Ingo Gühring† Mones Raslan†

July 1, 2020

Abstract

We examine the necessary and sufficient complexity of neural networks to approximate functions from different smoothness spaces under the restriction of encodable network weights. Based on an entropy argument, we start by proving lower bounds for the number of nonzero encodable weights for neural network approximation in Besov spaces, Sobolev spaces and more. These results are valid for most practically used (and sufficiently smooth) activation functions. Afterwards, we derive almost optimal upper bounds for ELU-neural networks in Sobolev norms up to second-order. This work advances the theory of approximating solutions of partial differential equations by neural networks.

Keywords: Neural Networks, Expressivity, Approximation Rates, Smoothness Spaces, Exponential Linear Unit, Encodable Weights

MSC classification (2010): 35A35, 41A25, 41A46, 46E35, 68T05

1 Introduction

Deep learning algorithms have lately shown promising results for dealing with classical mathematical problems, such as the solution of partial differential equations (PDEs), see for instance [31, 13, 22, 23, 16, 12, 34, 5, 10, 9, 25, 24, 20, 29, 18, 30]. In this work, we investigate the necessary and sufficient number of non-zero, encodable weights for a vanilla feedforward neural network to approximate functions that are particularly relevant for the solution of PDEs. Notable works in this direction for neural networks with the ReLU (rectified linear unit) activation function are [21, 37]. Due to the limited regularity of the ReLU, one is only able to derive approximation rates with respect to first-order Sobolev norms. However, in order to appropriately approximate solutions of fourth-order PDEs, approximation rates with respect to second or higher-order Sobolev norms are required. As an example, consider the Dirichlet problem for the biharmonic operator $\Delta^2$ (see e.g. [8]) on some domain $\Omega \subset \mathbb{R}^d$, a typical fourth-order problem, which is given by

$$-\Delta^2 u = f, \quad \text{on } \Omega + \text{boundary conditions.} \quad (1.1)$$

In its weak formulation, this operator equation is uniquely solvable in some subspace $V$ (incorporating the boundary conditions) of the Sobolev space $W^{2,2}(\Omega)$. Additionally (see [8], Section 6), typical solutions $u$ of (1.1) are even in the Sobolev space $W^{n,2}(\Omega)$ for some $n \geq 3$. This motivates studying approximations of Sobolev-regular functions $f \in W^{n,p}(\Omega)$ by neural networks in higher-order Sobolev norms. In this paper, we make the following two contributions:

---

*Both authors contributed equally.
†Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany; E-Mail: {guehring, raslan}@math.tu-berlin.de
1i.e., representable by a bit-string of moderate length
I. General Lower Bounds based on Entropy Arguments

Let $\mathcal{C} \subset \mathcal{D}$ be two function spaces. We will lower bound the necessary number for nonzero, encodable weights of neural network approximations of functions from $\mathcal{C}$ with respect to the norm in $\mathcal{D}$. Our notion of a lower bound for the number of nonzero, encodable weights can be summarized as follows:

For some $\gamma > 0$ (depending on $\mathcal{C}$ and $\mathcal{D}$) we have: If for every $\varepsilon > 0$ there exists some $M_\varepsilon \in \mathbb{N}$ such that every $f \in \mathcal{C}$ can be $\varepsilon$-approximated by a neural network $\Phi_{\varepsilon,f}$ (i.e., $\|f - \Phi_{\varepsilon,f}\|_D \leq \varepsilon$) with $M_\varepsilon$ nonzero, encodable weights, then (up to a logarithmic factor and for some constant $C$) $M_\varepsilon \geq C \varepsilon^{-\gamma}$.

In [39], the concept of the $\varepsilon$-entropy $H_\varepsilon(\mathcal{C}, \mathcal{D})$ was used to derive lower bounds for $M_\varepsilon$ for specific choices of $\mathcal{C}$ and $\mathcal{D}$. In Theorem 3.3 we generalize that approach to a wide range of function spaces. In detail, we show that every lower bound on the $\varepsilon$-entropy $H_\varepsilon(\mathcal{C}, \mathcal{D})$ of the unit ball of $\mathcal{C}$ with respect to $\|\cdot\|_\mathcal{D}$ can directly be transferred to a lower bound on the number of nonzero, encodable weights of an approximating neural network. Concretely, if $H_\varepsilon(\mathcal{C}, \mathcal{D}) \geq C \varepsilon^{-\gamma}$, then $M_\varepsilon \geq C \varepsilon^{-\gamma} / \log_2(1/\varepsilon)$. This approach works with fairly general activation functions $\varphi$.

Since lower bounds on the $\varepsilon$-entropy are well-studied for a variety of classical function spaces $\mathcal{C}$ we give a nonexhaustive list of concrete lower complexity bounds in Corollary 3.4 for Sobolev and Besov spaces. Appositely to the upper bounds that we present below, we state the following special instance of these results: For $\mathcal{C} = W^{k,p}(\Omega)$ and $\mathcal{D} = W^{n,p}(\Omega)$ with $n, k \in \mathbb{N}, n > k$ and $1 \leq p \leq \infty$ we have $M_\varepsilon \geq C \varepsilon^{-d/(n-k)} / \log_2(1/\varepsilon)$.

II. Almost Optimal Upper Bounds for ELU-Neural Networks in second-order Sobolev Spaces

Here, we only consider neural networks with the ELU$_1$ activation function (see [9]) given by

$$\text{ELU}_\alpha : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \begin{cases} x, & x \geq 0, \\ \alpha(e^x - 1), & x < 0, \end{cases}$$

for $\alpha = 1$. We have chosen the ELU$_1$ as our object of study for three reasons: (a) its increasing popularity in applications; (b) sufficient smoothness to allow for approximations in Sobolev norms up to order two.

In detail, we have $\text{ELU}_1|_K \in W^{2,p}(K) \cap C^1(K)$ for all compact $K \subset \mathbb{R}$, where $p \in [1, \infty)$. (c) Our proof strategy heavily relies on the construction of a partition of unity with neural networks. This is at least approximately possible with the ELU. Other activation functions of interest are subject to future work.

We consider the function spaces $\mathcal{C} = W^{n,p}((0,1)^d)$ as well as $\mathcal{D} = W^{k,p}((0,1)^d)$, where $n \in \mathbb{N}_{\geq k+1}$, $k \in \{0,1,2\}$ and $1 \leq p \leq \infty$.

- For the case $k \in \{0,1\}$, we construct in Proposition 4.1 for every approximation accuracy $\varepsilon > 0$ and every $f$ in the unit ball of $W^{n,p}((0,1)^d)$ a neural network $\Phi_{\varepsilon,f}$ with at most $C \varepsilon^{-d/(n-k)}$ nonzero weights such that $\|f - \Phi_{\varepsilon,f}\|_{W^{k,p}} \leq \varepsilon$.

- For the case $k = 2$, we were not able to show the canonical upper bound $\varepsilon^{-d/(n-2)}$, but get arbitrarily close to it. In detail, for every $\mu > 0$ we construct a neural network, with at most $C \varepsilon^{-d/(n-2-\mu)}$ nonzero weights such that $\|f - \Phi_{\varepsilon,f}\|_{W^{2,p}} \leq \varepsilon$.

In both cases the depth of the constructed networks is constant (i.e. accuracy-independent) and greater than two. In Theorem 4.2 we additionally show that the weights of $\Phi_{\varepsilon,f}$ can be encoded by $C \log_2(1/\varepsilon)$ bits.

---

2 see for instance [30, 15]
3 and, for all $\alpha \geq 0, \alpha \neq 1$, there holds $\text{ELU}_\alpha \in W^{1,p}(K) \cap C(K)$
4 In [36, Theorem 4.1], for the case $k = 0$, $p = \infty$ the same upper bounds for the number of nonzero weights is shown, while the depth of the approximating networks grows logarithmically in $1/\varepsilon$. 

---
We note that, similarly to [51, 21, 39], the main proof strategy is based on the approximation of localized polynomials by neural networks. However, constructing localized bumps that together form a partition of unity with ELU-neural networks is highly-nontrivial and can only be done approximatively.

As already outlined in [21, Section 1.4], we observe in both, lower and upper bounds, a trade-off between the complexity of the approximating neural networks and the order of the approximation norm: A higher order of $k$ requires neural networks with asymptotically more nonzero weights. Additionally, up to a log-factor and up to $\mu > 0$ for $k = 2$, our upper bounds are tight if we only allow encodable weights.

**Related Work**

The universal approximation theorem [10, 24] is often regarded as the starting point of approximation theory for neural networks. It shows that every continuous function defined on a compact domain can be uniformly approximated by shallow neural networks with continuous, non-polynomial activation function. Extensions of this theorem (see [40, Section 4] and the references therein) also take derivatives into account. In more detail, it has been established that shallow neural networks with sufficiently regular activation function and unrestricted width are dense in the space $C^m$, where $m \in \mathbb{N}$. The necessary and sufficient complexity of neural network approximations for (piecewise) smooth functions has been studied, for example, in [51, 27, 2, 48, 43, 45] for ReLU-neural networks and in [3, 35, 6, 36, 32, 49, 33] for neural networks with activation functions that have higher regularity. The approximation error in all of these papers is measured with respect to $L^p$-norms. In particular, we mention [36, Theorem 4.1], which is applicable to ELU-neural networks and establishes bounds for Hölder functions in $L^\infty$. The papers [6, 39] consider the restriction of encodable weights.

In this paper we are primarily interested in the approximation of functions with respect to Sobolev norms. In this direction, we mention two works, which examine the approximation capabilities of ReLU-neural networks with respect to $W^{1,p}$ norms. The paper [21] derives lower complexity bounds based on a VC dimension argument for unrestricted neural network weights (similar to the one presented in [51]) and upper bounds based on the emulation of localized polynomials. In [37], approximation rates were derived by re-approximating finite elements. None of these papers examine neural networks with encodable weights.

**Outline**

After having introduced the necessary terminology for neural networks in Section 2, we start by proving general lower complexity bounds in Section 3. In Section 4, we derive almost optimal upper approximation rates for ELU-neural networks. The proofs of the two main results in this section, Proposition 4.1 and Theorem 4.2, can be found in Appendix D and Appendix E, respectively. To not interrupt the flow of reading, the notation section, basic facts about Sobolev spaces and basic operations one can perform with neural networks have been deferred to Appendices A-C, respectively.

### 2 Neural Networks with Encodable Weights: Terminology

We start by formally introducing neural networks closely sticking to the notions introduced in [39]. In the following, we will distinguish between a neural network as a structured set of weights and the associated function implemented by the network, called its realization. Towards this goal, let us fix numbers $L, d = N_0, N_1, \ldots, N_L \in \mathbb{N}$.

- A family $\Phi = (A_\ell, b_\ell)_{\ell=1}^L$ of matrix-vector tuples of the form $A_\ell \in \mathbb{R}^{N_{\ell-1} \times N_\ell}$ and $b_\ell \in \mathbb{R}^{N_\ell}$ is called a neural network.
• We refer to the entries of $A_\ell, b_\ell$ as the weights of $\Phi$ and call $M(\Phi) := \sum_{\ell=1}^L (\|A_\ell\|_0 + \|b_\ell\|_0)$ its number of nonzero weights, $L = L(\Phi)$ its number of layers and we call $N_\ell$ the number of neurons in layer $\ell$.
• We denote by $d := N_0$ the input dimension of $\Phi$ and by $N_L$ the output dimension.
• Moreover, we set $\|\Phi\|_{\text{max}} := \max_{\ell=1,\ldots,L} \max_{i=1,\ldots,N_\ell} \max_{j=1,\ldots,N_{\ell-1}} \{|(A_\ell)_{i,j}|, |(b_\ell)_i|\}$, which is the maximum absolute value of all weights.

5. Encodability

In the following, we study neural networks with encodable weights. This information-theoretic viewpoint has already been examined in [6, 39] and is motivated by the observation that on a computer only weights of limited complexity (w.r.t. their bit-length) can be stored. In this paper, we consider weights that can be encoded by bit-strings with length logarithmically growing in $1/\varepsilon$, where $\varepsilon$ is the approximation accuracy.

To make the notion of encodability more precise, we first introduce coding schemes (see [39]): A coding scheme (for real numbers) is a sequence $\mathcal{B} = (B_\ell)_{\ell \in \mathbb{N}}$ of maps $B_\ell : \{0,1\}^\ell \to \mathbb{R}$. Now we define sets of neural networks with weights encodable by a coding scheme. Given an arbitrary coding scheme $\mathcal{B} = (B_\ell)_{\ell \in \mathbb{N}}$, and $d \in \mathbb{N}, \varepsilon, M > 0$, we denote by

$$\mathcal{N}_d^{\mathcal{B}}_{M,[C_0 \log_2(1/\varepsilon)].d}$$

the set of all neural networks $\Phi$ with $d$-dimensional input, one-dimensional output and at most $M$ nonzero weights such that each nonzero weight of $\Phi$ is contained in $\text{Range}(B_{[C_0 \log_2(1/\varepsilon)].d})$.

3 Lower Bounds For Neural Networks with Encodable Weights and General Activation Functions

In this section, we derive lower bounds on the necessary number of nonzero, encodable weights of neural network approximations. The approximated function spaces include a wide variety of classical smoothness spaces and the accuracy is measured in rather general norms. For this, we only assume mild conditions on
the activation function of the neural network. We note that the proof of our result is essentially an abstract version of the proof of [39, Theorem 4.2]. After encouragement of one of the authors of [39] and after studying the paper more closely, we noticed that it is possible to consider the proof strategy of [39, Theorem 4.2] in a very general setting which we will outline below.

Throughout this section (unless stated otherwise) we fix some \( d \in \mathbb{N} \), some domain \( \Omega \subset \mathbb{R}^d \) and two normed spaces \( C, D \) of (equivalence classes of) functions defined on \( \Omega \) with values in \( \mathbb{R} \). Additionally, we assume that \( C \subset D \).

**Definition 3.1.** Let \( C_0 > 0 \) be fixed. Additionally, let \( f \in C \), and for some function \( \varrho : \mathbb{R} \to \mathbb{R} \) assume that \( \mathcal{N}_\varrho \subset D \). Finally, let \( \varepsilon > 0 \) and fix some coding scheme \( B \). Then, for \( C_0 > 0 \), we define the quantities:

\[
M^B_\varepsilon(f) := M^{B,\varrho,C_0,C,D}_\varepsilon(f) := \min \left\{ M \in \mathbb{N} : \exists \Phi \in \mathcal{N}_{M,\{C_0\log_2 \frac{1}{M}\},d} : \| f - R_\varrho(\Phi) \|_D \leq \varepsilon \right\},
\]

and

\[
M^B_\varepsilon(C, D) := M^{B,\varrho,C_0}_\varepsilon(C, D) := \sup_{f \in C, \| f \|_C \leq 1} M^{B,\varrho,C_0}_\varepsilon(C, D)(f).
\]

In other words, the quantity \( M^B_\varepsilon(f) \) denotes the required number of nonzero weights of a neural network \( \Phi \) to \( \varepsilon \)-approximate \( f \) with weights that can be encoded with \( \lceil C_0 \log_2 (1/\varepsilon) \rceil \) bits using the coding scheme \( B \). \( M^B_\varepsilon(C, D) \) gives a uniform bound of this quantity over the unit ball in \( C \). Next, we introduce the entropy \( H_\varepsilon(C,D) \), for which we show in Theorem 3.3 that it can directly be related to the approximation capabilities of neural networks.

**Definition 3.2 (28).** For \( \varepsilon > 0 \), define \( U_\varepsilon(C, D) \) as the minimal number of closed balls with radius \( \varepsilon \) in the space \( D \) needed to cover the unit ball in \( C \) embedded in \( D \). The \( \varepsilon \)-entropy of the unit ball of \( C \) embedded into \( D \) is then defined by

\[
H_\varepsilon(C, D) := \log_2 U_\varepsilon(C, D).
\]

Theorem 3.3 now states that if we can lower bound the \( \varepsilon \)-entropy, then we are also able to lower bound \( M^B_\varepsilon(C, D) \). Lower bounds on the \( \varepsilon \)-entropy (and hence for the quantity \( M^B_\varepsilon(C, D) \)) for specific, frequently used function spaces fulfilling the assumptions of the theorem will be given in Corollary 3.4.

**Theorem 3.3.** Let \( \varrho : \mathbb{R} \to \mathbb{R} \) with \( \varrho(0) = 0 \) and such that \( \mathcal{N}_\varrho \subset D \). Additionally, assume that \( H_\varepsilon(C,D) \geq C_1 \varepsilon^{-\gamma} \gamma \) for some \( \gamma = \gamma(C,D), C_1 = C_1(C,D) > 0 \) and all \( \varepsilon > 0 \). Then, for each \( C_0 > 0 \) there exists a constant \( C = C(\gamma,C,D,C_0) > 0 \), such that for each coding scheme of real numbers \( B \), and for all \( \varepsilon \in (0,\frac{1}{2}) \) we have

\[
M^B_\varepsilon(C_0,C,D) \geq C \cdot \varepsilon^{-\gamma} / \log_2 \left( \frac{1}{\varepsilon} \right).
\]

**Proof.** The proof of the theorem is along the lines of [39, Theorem 4.2] (see page 34 of the corresponding arXiv version), but in a more abstract setup. We shortly describe the idea. Afterwards we describe how one needs to adapt the quantities of the proof of [39, Theorem 4.2] in order to prove our result. First of all, we need the notion of the minimax code length \( L_\varepsilon(C,D) \) of \( C \) with respect to \( D \). The minimax code length describes the uniform description complexity of the set \( \{ f \in C : \| f \|_C \leq 1 \} \) in terms of the number of nonzero bits necessary to encode every \( f \) with distortion at most \( \varepsilon \) in \( D \). It is defined as follows (see also [39, Definition B.2]):

\[\text{We want to take the opportunity to thank Philipp Petersen for the fruitful suggestion.}\]

\[\text{We use the convention that min } \emptyset = \infty.\]
Let $\ell \in \mathbb{N}$. We denote by $\mathcal{E}^\ell := \{E : C \to \{0, 1\}^\ell\}$ the set of binary encoders mapping elements of $C$ to bit strings of length $\ell$, and by $\mathcal{D}^\ell := \{D : \{0, 1\}^\ell \to D\}$ the set of binary decoders mapping bit-strings of length $\ell$ into $D$. For $\varepsilon > 0$, we define the minimax code length by

$$L_\varepsilon(C, D) := \min \left\{ \ell \in \mathbb{N} : \exists (E^\ell, D^\ell) \in \mathcal{E}^\ell \times \mathcal{D}^\ell : \sup_{f \in C, \|f\|_C \leq 1} \|D^\ell(E^\ell(f)) - f\|_D \leq \varepsilon \right\}.$$ 

By [11] Remark 5.10 it holds that $L_\varepsilon(C, D) \geq H_\varepsilon(C, D)$ and, thus, since we assumed $H_\varepsilon(C, D) \geq C_1 \varepsilon^{-\gamma}$, we get $L_\varepsilon(C, D) \geq C_1 \varepsilon^{-\gamma}$.

Now, we prove the claim by contradiction: Assume that the lower bound (3.1) does not hold. Then, for every $f \in C$, $\|f\|_C \leq 1$ there exists some neural network $\Phi_{\varepsilon,f} \in \mathcal{N}^{B}_{M,[C_0 \log_2(1/\varepsilon)],d}$ such that (a) its $g$-realization approximates $f$ up to error $\varepsilon$ in $D$ and (b) it has $M \leq C_1 \varepsilon^{-\gamma}/\log_2(1/\varepsilon)$ nonzero weights. In [89] Lemma B.4 it is shown that there exists an injective map $\Gamma : \{R_\varepsilon(\Phi) : \Phi \in \mathcal{N}^{B}_{M,[C_0 \log_2(1/\varepsilon)],d}\} \to \{0, 1\}^\ell$ for some $\ell < C_1 \varepsilon^{-\gamma}$. We define an encoder and decoder as

$$E^\ell : \{f \in C : \|f\|_C \leq 1\} \to \{0, 1\}^\ell, \quad f \mapsto \Gamma(R_\varepsilon(\Phi_{\varepsilon,f})) \quad \text{and} \quad D^\ell : \{0, 1\}^\ell \to C, \quad c \mapsto \Gamma^{-1}(c),$$

where $\Gamma^{-1}$ is the left-inverse of $\Gamma$. We now get that

$$\sup_{f \in C, \|f\|_C \leq 1} \|f - D^\ell(E^\ell(f))\|_D \leq \varepsilon,$$

which is a contradiction to $L_\varepsilon(C, D) \geq C_1 \varepsilon^{-\gamma}$.

We close this proof outline by demonstrating how to replace the quantities in the proof of [89] Theorem 4.2 in order to prove our result.

$$\mathcal{H}_1 \times \mathcal{F}_{\alpha,d,B} \sim \{f \in C : \|f\|_C \leq 1\},$$

$L^p([-1/2, 1/2]^d) \sim \mathcal{D}$,

$[\mathcal{F}([-1/2, 1/2]^d] \sim \Omega,$

$p(d-1)/\beta \sim \gamma,$

$K_0, C_2 \sim$ equally defined,

$C \sim C := \min\{1, C_2/[2C_1(2 + C_0 + \gamma)]\} > 0,$

$M_0 \sim M_0 := [C \varepsilon^{-\gamma}/\log_2(1/\varepsilon)],$

$L_p(\varepsilon, \mathcal{H}_1 \times \mathcal{F}_{\alpha,d,B}) \sim L_\varepsilon(C, D).$

We proceed by listing a variety of lower bounds for specific examples for frequently used function spaces.

**Corollary 3.4.** Assume that $\Omega$ fulfills some regularity conditions\(^8\). Let $\varrho : \mathbb{R} \to \mathbb{R}$ be chosen such that $\varrho(0) = 0$ and $\mathcal{N}^{B}_{\alpha} \subset \mathcal{D}$ (where $\mathcal{D}$ is a function space on $\Omega$ specified below). Moreover, let $\mathcal{B}$ be an arbitrary coding scheme. Then, the following statements hold:

(i) **Besov spaces:** Let $s, t \in \mathbb{R}$ with $s < t$ as well as $p_1, p_2, q_1, q_2 \in (0, \infty]$ such that

$$t - s - d \max \left\{ \left( \frac{1}{p_1} - \frac{1}{p_2} \right), 0 \right\} > 0.$$

\(^8\)Many results estimating the $\varepsilon$-entropy are only formulated and proven for $C^\infty$-domains for simplicity of exposition. However, as has been described in [89] Section 4.10.3 and [15] Section 3.5, these results remain valid for function spaces on more general domains including cubes.
Moreover, let $\mathcal{C} = B_{p_1,q_1}^t(\Omega)$, and $\mathcal{D} = B_{p_2,q_2}^s(\Omega)$. Then, for some $C > 0$, we have

$$M_\epsilon^p(\mathcal{C}, \mathcal{D}) \geq C \epsilon^{-\frac{d}{2n}} \log_2 \left( \frac{1}{\epsilon} \right), \quad \text{for all } \epsilon \in (0, 1/2).$$

(ii) **Sobolev Spaces:** Let $s, t \in \mathbb{N}$ with $t > s$ and let $p \in (0, \infty]$. Then, for $\mathcal{C} = W^{t,p}(\Omega)$ and for $\mathcal{D} = W^{s,p}(\Omega)$ there exists some $C > 0$ with

$$M_\epsilon^p(\mathcal{C}, \mathcal{D}) \geq C \epsilon^{-\frac{d}{2n}} \log_2 \left( \frac{1}{\epsilon} \right), \quad \text{for all } \epsilon \in (0, 1/2).$$

**Proof.** (i) follows immediately from Theorem 3.3 in combination with Theorem [15, Section 3.5].

(ii) follows from Theorem 3.3 together with [14, Section 1.3], where we use the estimate on the approximation number $a_k(id)$ (cf. page 9) combined with the relation of $a_k(id)$ and the entropy. $\square$

**Discussion**

We conclude this section by a discussion putting our results into context.

(i) **Activation Functions:** Apart from having sufficient regularity, the only other requirement on the activation function present in our lower bounds is $\varrho(0) = 0$. Hence, we are in a position to conclude suitable lower bounds for many practically used activation functions such as the (leaky) ReLU, the ELU, the inverse (linear) square root unit, rectified power units, the tanh, the arctan, and more.

(ii) **Other Function Spaces:** As already mentioned above, one can deduce similar lower bounds for other choices of $\mathcal{C}, \mathcal{D}$. Notable examples include Hölder spaces, Triebel-Lizorkin, or Zygmund spaces (see for instance [50, 15] and the references therein for further examples).

(iii) **Relation to Lower and Upper Bounds With Non-Encodable Weights:** If one drops the restriction of encodable weights and considers the more general setting of arbitrary weights, a lesser number of weights is required in general. For this setting, we mention the results from [52, 21] which combined state:

For $\mathcal{C} = W^{n,\infty}((0,1)^d)$ and $\mathcal{D} = W^{k,\infty}((0,1)^d)$ with $k = 0, 1$, it holds for the necessary number of nonzero weights $M_\epsilon$ to achieve an $\epsilon$-approximation in $W^{k,\infty}$ norm that

$$M_\epsilon \geq C \epsilon^{-d/(2n-k)}.$$For $k = 0$, in [52] neural networks are constructed that achieve this approximation rate. In comparison, our entropy bounds show that under the assumption of encodable weights $M_\epsilon \geq C \epsilon^{-d/(n-k)}$ (suppressing the $\log_2(1/\epsilon)$ factor for simplicity of exposition).

4 **Upper Bounds for Function Approximation by ELU Neural Networks in $W^{k,p}$ for $k = 0, 1, 2$**

In this section, we show that for an arbitrary accuracy $\epsilon > 0$, every function from the unit ball of Sobolev space $W^{n,p}$ (with $n > k$)

$$\mathcal{F}_{n,d,p} := \{ f \in W^{n,p}((0,1)^d) : \|f\|_{W^{n,p}((0,1)^d)} \leq 1 \}$$
can be \( \varepsilon \)-approximated in weaker Sobolev norms \( W^{k,p} \) with order \( k = 0, 1, 2 \) by an ELU-neural network. For this, we explicitly construct approximating neural networks with constant depth (i.e., independent of \( \varepsilon \)) and give upper bounds for the number of nonzero, encodable weights (depending on \( \varepsilon \)), which in the light of the results of Section 2 are almost optimal. The ELU activation function is parametrized by \( \alpha > 0 \) which determines the function for \( x < 0 \). We have that

\[
\text{ELU}_\alpha(x) = x \cdot 1_{\{x \geq 0\}} + \alpha(e^x - 1) \cdot 1_{\{x < 0\}} \in \begin{cases} W^{2,p}(K), & \text{if } \alpha = 1, \\ W^{1,p}(K), & \text{if } \alpha \neq 1, \end{cases}
\]

for every compact subset \( K \) of \( \mathbb{R} \) and every \( p \in [1, \infty] \). Since the regularity of the neural network is determined by the regularity of the activation function, approximations in \( W^{2,p} \) norm can only be shown for the case \( \alpha = 1 \) (see Figure 1 for a visualization). We exclusively consider this case in our theorem and proofs. In the discussion at the end of this section, we shortly note that our results can be adapted to \( \alpha \in (0,1) \) and give an outlook how our proof strategy might be applied in a plug-and-play manner to different activation functions.

![Figure 1](image_url)

Figure 1: Depiction of the left: \( \text{ELU}_1 \), middle: \( \text{ELU}_{1/2} \), and right: \( \text{ReLU}(= \text{ELU}_0) \)- activation functions as well as their weak derivatives up to the highest possible order.

The proof of the main statement of this section can be roughly divided into two steps: First (see Proposi-
tions \([4,1]\), the approximating neural networks are constructed with weights whose absolute values are bounded polynomially in \(e^{-1}\). In Theorem \([4,2]\) the encodability of the weights is enforced. Note that for the case \(k = 2\) we can only get arbitrarily close to the best possible approximation rate. Towards this end, we introduce, for \(\mu > 0\), the notation

\[
\mu_{(k=2)} := \begin{cases} \mu, & \text{if } k = 2, \\ 0, & \text{else.} \end{cases}
\]

We start with Proposition \([4,1]\).

**Proposition 4.1.** Let \(d \in \mathbb{N}, k \in \{0, 1, 2\}, n \in \mathbb{N}_{> k+1}, 1 \leq p \leq \infty, \) and \(\mu > 0\). Then, there exist constants \(L, C, \theta, \xi\) depending on \(d, n, p, k, \mu\) with the following properties:

For every \(\varepsilon \in (0, \xi)\) and every \(f \in \mathcal{F}_{n,d,p}\), there is a neural network \(\Phi_{\varepsilon,f}\) with \(d\)-dimensional input and one-dimensional output, at most \(L\) layers and at most \(C\xi^{-d/(n-k-\mu_{(k=2)})}\) nonzero weights bounded in absolute value by \(C\xi^{-\theta}\) such that

\[
\|R_{\text{ELU}_1}(\Phi_{\varepsilon,f}) - f\|_{W^k,\ell_p(0,1)^d} \leq \varepsilon.
\]

**Proof.** The proof of this proposition is the subject of Appendix \([D]\). Here, we shortly outline the proof strategy. The main idea is based on the common strategy (see e.g. \([51,21,36]\)) of approximating \(f\) by localized polynomials which in turn are approximated by neural networks. For this, one needs to approximate bump functions that form a partition of unity as well as polynomials up to degree \(n - 1\) by neural networks.

- For a gridsize \(1/N\) (with \(N \in \mathbb{N}\)) we divide the domain \((0, 1)^d\) into \((N+1)^d\) equally large patches. In Appendix \([D.1]\) we construct for each patch (abbreviated by \(p\)) a bump function \(\phi_p \in W^{2,\infty}\) with rescaled shifts of ELU\(_1\). Deviating from usually used bump functions, our \(\phi_p\) are not compactly supported on the corresponding patch (but decay exponentially fast outside the patch for \(N \to \infty\)) and their sum only approximates \(1\) (in \(W^k,\ell_p(0,1)^d\) for \(N \to \infty\)).

- Using the Bramble-Hilbert Lemma \([B.4]\) and our bump functions from Appendix \([D.1]\) we show in Appendix \([D.2]\) that every function \(f \in \mathcal{F}_{n,d,p}\) can be approximated in \(\|\cdot\|_{W^k,\ell_p}\) (for \(k = 0, 1, 2\)) up to error \((1/N)^{d-k}\) by localized (averaged Taylor) polynomials of the form \(\sum \phi_p \cdot \text{poly}_p\).

- In Appendix \([D.3]\) we show that the localized polynomials \(\sum \phi_p \cdot \text{poly}_p\) can be approximated by ELU\(_1\)-neural networks (Lemma \([D.11]\)). Since the bumps \(\phi_p\) are by construction ELU\(_1\)-neural networks, we only need to approximate the polynomials (Proposition \([D.7]\)) and deal with the multiplication of \(\phi_p\) with \(\text{poly}_p\) (Propositions \([D.9]\) and \([D.10]\)).

- In Appendix \([D.4]\) we put everything together and choose \(N = N(\varepsilon)\) appropriately.

The main theorem now states that Proposition \([4,1]\) also holds with encodable weights, i.e. for each \(\varepsilon > 0\), every element of the set of weights \(W_\varepsilon = \bigcup_{\mu>0} W_{\varepsilon,f}\) (where \(W_{\varepsilon,f}\) denotes the weights of \(\Phi_{\varepsilon,f}\)) can be uniquely encoded by \([C \log_2(1/\varepsilon)]\) bits. To state this in a formal way, we use the notation introduced in Equation \([2.1]\).

**Theorem 4.2.** Let \(d \in \mathbb{N}, k \in \{0, 1, 2\}, n \in \mathbb{N}_{> k+1}, 1 \leq p \leq \infty, \) and \(\mu > 0\). Then, there exist constants \(L, C, \xi, \) and a coding scheme \(B = (B_\varepsilon)_{\varepsilon \in \mathbb{N}}\) depending on \(d, n, p, k, \mu\) with the following properties:

For every \(\varepsilon \in (0, \xi)\) and every \(f \in \mathcal{F}_{n,d,p}\), there is a neural network \(\Phi_{\varepsilon,f} \in \mathcal{N}_{M_\varepsilon}^d, [C \log_2(1/\varepsilon)]\) with \(d\)-dimensional input, one-dimensional output, at most \(L\) layers and at most \(M_\varepsilon = C \cdot \varepsilon^{-d/(n-k-\mu_{(k=2)})}\) nonzero weights, such that

\[
\|R_{\text{ELU}_1}(\Phi_{\varepsilon,f}) - f\|_{W^k,\ell_p(0,1)^d} \leq \varepsilon.
\]
Proof. We give a short outline of the proof here, the details can be found in Appendix E. Let \( \Phi_{\varepsilon,f} = ((A_1, b_1), \ldots, (A_{L-1}, b_{L-1}), (A_L, b_L)) \) be the network from Proposition 4.1 (where the main work has already been done). We denote the collection of entries of \( A_1, b_1, \ldots, A_{L-1}, b_{L-1} \) by \( W_{1,\ldots,L-1} \) and use that they are independent of \( f \) (i.e., they only depend on \( \varepsilon, n, d, p, k, \mu \)). Only the weights \( A_L, b_L \) in the last layer depend on \( f \).

- The number of independent weights \( |W_{1,\ldots,L-1}| \) is bounded by \( C \cdot \varepsilon^{-d/(n-k-\mu(k=2))} \) since the total number of nonzero weights is bounded by this quantity.

- We round the entries of \( A_L, b_L \) with a suitable precision \( \nu \) (see Lemma E.1) to the mesh \( [-\varepsilon^{-\theta}, \varepsilon^{-\theta}] \cap \varepsilon^{\nu}\mathbb{Z} \), where we also use the fact that the weights of \( \Phi_{\varepsilon,f} \) are bounded in absolute value by \( C\varepsilon^{-\theta} \).

Hence, the weights of the approximating neural networks can be chosen from a set \( W_{\varepsilon} \) with less than \( \varepsilon^{-s} \) real numbers, where \( s > 0 \) only depends on \( d, n, p, k, \mu \) and not on \( f \). Consequently, there exists a surjective mapping \( B_{\varepsilon} : \{0,1\}^\lceil s \log_2(1/\varepsilon) \rceil \to W_{\varepsilon} \). The collection of these maps constitutes the coding scheme.

Discussion

We discuss the tightness of our bounds, possible extensions, and differences to other works.

(i) Tightness of the Bounds: From Corollary 3.4(ii) it follows that our bounds for encodable neural network weights are tight up to a log factor for \( k = 0, 1 \). For \( k = 2 \) we get arbitrary close to the optimal bound (again up to a log factor) but were not able to reach it. If we allow for arbitrary weights, then this upper bound might be drastically improved (see also [iii] in the discussion of Section 3).

(ii) ELU\( \alpha \) for \( \alpha \in (0,1) \): It is possible to adapt the proofs of Proposition 4.1 and Theorem 4.2 (with some significant simplifications) to obtain analogous approximation rates for ELU\( \alpha \)-neural networks for \( \alpha \in (0,1) \) and \( k = 0, 1 \).

(iii) Difference to Other Works: As we have already remarked in the proof outline of Proposition 4.1 re-approximation of localized polynomials is a common strategy to obtain neural network approximation results (see e.g. [51, 21, 36]). Our work differs from these other works in three major aspects:
Our approximations include $W^{2,p}$ (instead of maximally $W^{1,p}$).

(b) Constructing a partition of unity with ELU-neural networks is tricky (contrary to ReLU networks) and can only be done approximately (see proof outline of Proposition 4.1 and Figure 2).

(c) ELU-neural networks admit approximations of polynomials with a depth independent of $\varepsilon$, which results in constant-depth approximations of $f$.

(iv) **Plug-and-Play**: We believe that our proof strategy can be used in a plug-and-play fashion with various activation functions. So far the concept of re-approximating localized polynomials (which goes back to [51]) has mainly been used for ReLU-like activation functions that allow the construction of an exact partition of unity. In our setting, an approximate partition of unity suffices. This might open the door for results involving e.g. the inverse square root unit, the softplus function, or sigmoidal functions. For instance, the approximative partition of unity introduced in [33] for $L^\infty$ could potentially be used as a starting point for deducing approximation rates in Sobolev spaces of sigmoidal neural networks. However, we believe that every approximation result with respect to norms in $W^{k,p}$ with $k \geq 3$ gets increasingly technical.

**Acknowledgments**

The authors would like to thank Philipp Petersen for fruitful discussions on the topic. I. Gühring acknowledges support from the Research Training Group “Differential Equation- and Data-driven Models in Life Sciences and Fluid Dynamics: An Interdisciplinary Research Training Group (DAEDALUS)” (GRK 2433) funded by the German Research Foundation (DFG).

**References**

[1] R. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.

[2] J. Almira, P. López-de-Teruel, D. Romero-López, and F. Voigtlaender. Some negative results for single layer and multilayer feedforward neural networks. *arXiv preprint arXiv:1810.10032*, 2018.

[3] A. Barron. Approximation and Estimation Bounds for Artificial Neural Networks. *Mach. Learn.*, 14(1):115–133, 1994.

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

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

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

[7] S. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer Science+Business Media, New York, 3rd edition, 2008.

[8] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. Society for Industrial and Applied Mathematics, 2002.
[9] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv preprint arXiv:1511.07289*, 2015.

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

[11] S. Dahlke, F. De Mari, P. Grohs, and D. Labate. *From Groups to Signals*. Harmonic and Applied Analysis. Birkhäuser, 2015.

[12] W. E, J. Han, and A. Jentzen. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Commun. Math. Stat.*, 5(4):349–380, 2017.

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

[14] D. E. Edmunds and W. D. Evans. *Hardy operators, function spaces and embeddings*. Springer Monographs in Mathematics. Springer, Berlin, 2004.

[15] D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge Tracts in Mathematics. Cambridge University Press, 1996.

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

[17] L. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 1999.

[18] M. Geist, P. Petersen, M. Raslan, R. Schneider, and G. Kutyniok. Numerical Solution of the Parametric Diffusion Equation by Deep Neural Networks. *arXiv preprint arXiv:2004.12131*, 2020.

[19] D. Gilbarg and N. Trudinger. *Elliptic Partial Differential Equations of Second Order*, volume 224 of *A Series of Comprehensive Studies in Mathematics*. Springer, Berlin, second edition, 1998.

[20] P. Grohs, F. Hornung, A. Jentzen, and P. von Wurstemberger. A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations. *arXiv preprint arXiv:1809.02362*, 2018.

[21] I. Gühring, G. Kutyniok, and P. Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Anal. Appl. (Singap.)*, pages 1–57, 2019.

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

[23] J. Han, M. Nica, and A. Stinchcombe. A Derivative-Free Method for Solving Elliptic Partial Differential Equations with Deep Neural Networks. *arXiv preprint arXiv:2001.06145*, 2020.

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

[25] A. Jentzen, D. Salimova, and T. Welti. A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of Kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients. *arXiv preprint arXiv:1809.07321*, 2018.
[26] H. Katsuura. Summations involving binomial coefficients. *The College Mathematics Journal, 40*(4):275–278, 2009.

[27] P. Kidger and T. Lyons. Universal Approximation with Deep Narrow Networks. *arXiv preprint arXiv:1905.08539*, 2019.

[28] A. Kolmogorov and V. Tihomirov. ε-entropy and ε-capacity of sets in function spaces. *Amer. Math. Soc. Translation*, 1961.

[29] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider. A Theoretical Analysis of Deep Neural Networks and Parametric PDEs. *arXiv preprint arXiv:1904.00377*, 2019.

[30] F. Laakmann and P. Petersen. Efficient approximation of solutions of parametric linear transport equations by ReLU DNNs. *arXiv preprint arXiv:2001.11441*, 2020.

[31] I. Lagaris, A. Likas, and D. Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Trans. Neural Netw. Learn. Syst.*, 9(5):987–1000, 1998.

[32] B. Li, S. Tang, and H. Yu. Better Approximations of High Dimensional Smooth Functions by Deep Neural Networks with Rectified Power Units. *Commun. in Comp. Phys.*, 27:379–411, 2020.

[33] S.-B. Lin. Generalization and Expressivity for Deep Nets. *IEEE T. Neur. Net. Lear.*, 30(5):1392 – 1406, 2019.

[34] L. Lu, X. Meng, Z. Mao, and G. Karniadakis. DeepXDE: A deep learning library for solving differential equations. *arXiv preprint arXiv:1907.04502*, 2019.

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

[36] I. Ohn and Y. Kim. Smooth function approximation by deep neural networks with general activation functions. *Entropy, 21*(7):627, 2019.

[37] J. A. A. Opschoor, P. C. Petersen, and C. Schwab. Deep ReLU networks and high-order finite element methods. *Anal. Appl., Online first:1–56*, 2020.

[38] P. Petersen, M. Raslan, and F. Voigtlaender. Topological Properties of the Set of Functions Generated by Neural Networks of Fixed Size. *Found. of Comp. Math.*, Online first:1–70, 2020.

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

[40] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numer.*, 8:143–195, 1999.

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

[42] T. Roubíček. *Nonlinear Partial Differential Equations with Applications*, volume 153 of *International Series of Numerical Mathematics*. Springer Science+Business Media, Basel, second edition, 2013.

[43] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *arXiv preprint arXiv:1708.06633*, 2017.

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

In this subsection, we depict the (mostly standard) notation used throughout this paper. We set $\mathbb{N} := \{1, 2, \ldots\}$ and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For $k \in \mathbb{N}_0$ we define $\mathbb{N}_{\geq k} := \{k, k+1, \ldots\}$. For a set $A$ we denote its cardinality by $|A| \in \mathbb{N} \cup \{\infty\}$ and by $1_A$ its indicator function of $A$. If $x \in \mathbb{R}$, then we write $[x] := \min\{k \in \mathbb{Z} : k \geq x\}$ where $\mathbb{Z}$ is the set of integers and $\lfloor x \rfloor := \max\{k \in \mathbb{Z} : k \leq x\}$.

If $d \in \mathbb{N}$ and $\|\| \|$ is a norm on $\mathbb{R}^d$, then we denote for $x \in \mathbb{R}^d$ and $r > 0$ by $B_r(\|\|)(x)$ the open ball around $x$ in $\mathbb{R}^d$ with radius $r$, where the distance is measured in $\|\|$. By $|x|$ we denote the euclidean norm of $x$ and by $\|x\|_{\infty}$ the maximum norm. We endow $\mathbb{R}^d$ with the standard topology and for $A \subset \mathbb{R}^d$ we denote by $\overline{A}$ the closure of $A$.

For $d_1, d_2 \in \mathbb{N}$ and a matrix $A \in \mathbb{R}^{d_1 \times d_2}$ the number of nonzero entries of $A$ is counted by $\|\|_0$, i.e.

$$\|A\|_0 := \{|(i, j) : A_{i, j} \neq 0\}|.$$ 

If $f : X \to Y$ and $g : Y \to Z$ are two functions, then we write $g \circ f : X \to Z$ for their composition. If additionally $U \subset X$, then $f|_U : U \to Y$ denotes the restriction of $f$ onto $U$. We use the usual multiindex notation, i.e. for $\alpha \in \mathbb{N}_0^d$ we write $|\alpha| := \alpha_1 + \ldots + \alpha_d$ and $\alpha! := \alpha_1! \ldots \alpha_d!$. Moreover, if $x \in \mathbb{R}^d$, then we have

$$x^\alpha := \prod_{i=1}^d x_i^{\alpha_i}.$$ 

Let from now on $\Omega \subset \mathbb{R}^d$ be open. For a function $f : \Omega \to \mathbb{R}$, we denote by

$$D^\alpha f := \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots \partial x_d^{\alpha_d}}.$$
its (weak or classical) derivative of order $\alpha$. For $n \in \mathbb{N}_0 \cup \{\infty\}$, we denote by $C^n(\Omega)$ the set of $n$ times continuously differentiable functions on $\Omega$. Additionally, if $\Omega$ is compact, we set, for $f \in C^n(\Omega)$

$$\|f\|_{C^n(\Omega)} := \max_{0 \leq |\alpha| \leq n} \sup_{x \in \Omega} |D^\alpha f(x)|,$$

We denote by $L^p(\Omega)$, $1 \leq p \leq \infty$ the standard Lebesgue spaces.

In the following, we will also make use of the following well-known fact stating that the exponential function decays faster than any polynomial.

**Proposition A.1.** Let $\alpha, \beta, c, c' > 0$. Then

$$\lim_{x \to \infty} c' x^\alpha e^{c \cdot x^\beta} = 0.$$

This implies that for all $\gamma > 0$ there exists some constant $C = C(\alpha, \beta, \gamma) > 0$ such that for all $x > 0$ there holds

$$\frac{c' x^\alpha}{e^{c \cdot x^\beta}} \leq C x^{-\gamma}.$$

**B Sobolev Spaces**

In this section, we introduce Sobolev spaces (see [1]) which constitute a crucial concept within the theory of PDEs (see e.g. [42, 17]).

**Definition B.1.** Given some domain $\Omega \subset \mathbb{R}^d$, $1 \leq p < \infty$, and $n \in \mathbb{N}$, the Sobolev space $W^{n,p}(\Omega)$ is defined as

$$W^{n,p}(\Omega) := \left\{ f : \Omega \to \mathbb{R} : \|D^\alpha f\|_{L^p(\Omega)}^p < \infty, \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \leq n \right\},$$

and is equipped with the norm

$$\|f\|_{W^{n,p}(\Omega)} := \left( \sum_{0 \leq |\alpha| \leq n} \|D^\alpha f\|_{L^p(\Omega)}^p \right)^{1/p}.$$

Additionally, we set

$$W^{n,\infty}(\Omega) := \left\{ f : \Omega \to \mathbb{R} : \|D^\alpha f\|_{L^\infty(\Omega)} < \infty, \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \leq n \right\},$$

and we equip this space with the norm $\|f\|_{W^{n,\infty}(\Omega)} := \max_{|\alpha| \leq n} \|D^\alpha f\|_{L^\infty(\Omega)}$. Moreover, for $0 \leq k \leq n$, on $W^{n,p}(\Omega)$ we introduce the family of semi-norms

$$|f|_{W^{k,p}(\Omega)} := \left( \sum_{|\alpha| = k} \|D^\alpha f\|_{L^p(\Omega)}^p \right)^{1/p}, \quad |f|_{W^{k,\infty}(\Omega)} := \max_{|\alpha| = k} \|D^\alpha f\|_{L^\infty(\Omega)},$$

respectively.

**Remark B.2.** If $\Omega$ is bounded and fulfills a local Lipschitz condition, arguments from [42] show that $W^{2,\infty}(\Omega)$ can be continuously embedded into $C^1(\overline{\Omega})$. This can be seen as follows: [42 Theorem 4.12] shows that $W^{2,p}(\Omega)$ can be continuously embedded into $C^1(\overline{\Omega})$ for $p > d$. Since also $W^{2,\infty}(\Omega)$ can be continuously embedded into $W^{2,p}(\Omega)$, the claim follows.
Lemma B.4 (Bramble-Hilbert).\footnote{been worked out in \cite[Section B.3 and Lemma C.4]{21}.} Again the overall statement follows easily.

Remark B.3. For purely technical reasons we sometimes make use of an extension operator. For this, let $E : W^{n,p}((0,1)^d) \to W^{n,p}(\mathbb{R}^d)$ be the extension operator from \cite[Theorem VI.3.1.5]{77} and set $f := Ef$. Note that for arbitrary $\Omega \subset \mathbb{R}^d$ and $0 \leq k \leq n$ it holds

$$
\|f\|_{W^{k,p}(\Omega)} \leq \|f\|_{W^{n,p}(\mathbb{R}^d)} \leq C\|f\|_{W^{n,p}((0,1)^d)},
$$

where $C = C(n,p,d)$ is the norm of the extension operator.

The following lemma which will be crucial for the proofs of our results can be stated in much more generality (see \cite[Chapter 4.1]{7}) and relies on the use of averaged Taylor polynomials. We only state a version tailored to our specific needs and will not give a proof since the details of this specific version have been worked out in \cite[Section B.3 and Lemma C.4]{21}.

Lemma B.5. \footnote{For purely technical reasons we sometimes make use of an extension operator. For this, let $E : W^{n,p}((0,1)^d) \to W^{n,p}(\mathbb{R}^d)$ be the extension operator from \cite[Theorem VI.3.1.5]{77} and set $f := Ef$. Note that for arbitrary $\Omega \subset \mathbb{R}^d$ and $0 \leq k \leq n$ it holds}

Then there exists a constant $C = C(n,d) > 0$ such that for all $f \in W^{n,p}(\mathbb{R}^d)$ and $m \in \{0, \ldots, N\}^d$ there is a polynomial $p_m(x) = \sum_{|\alpha| \leq n-1} c_\alpha x^\alpha$ such that

$$
\|f - p_m\|_{W^{k,p}(\Omega_{m,N})} \leq C \left( \frac{1}{N} \right)^{n-k} \|f\|_{W^{n,p}(\Omega_{m,N})}, \quad \text{for } k = 0, 1, \ldots, n
$$

and the coefficients $c_\alpha$ are bounded by $|c_\alpha| \leq CN^{d/p}\|f\|_{W^{n,p}(\Omega_{m,N})}$ for all $\alpha$ with $|\alpha| \leq n - 1$.

Now we turn our attention to a version of a product rule tailored to our needs.

Lemma B.6. \footnote{For purely technical reasons we sometimes make use of an extension operator. For this, let $E : W^{n,p}((0,1)^d) \to W^{n,p}(\mathbb{R}^d)$ be the extension operator from \cite[Theorem VI.3.1.5]{77} and set $f := Ef$. Note that for arbitrary $\Omega \subset \mathbb{R}^d$ and $0 \leq k \leq n$ it holds}

Now we turn our attention to a version of a product rule tailored to our needs.

Lemma B.5. Let $k \in \{0, 1, 2\}$, $f \in W^{k,\infty}(\Omega)$ and $g \in W^{k,p}(\Omega)$ with $1 \leq p \leq \infty$, then $fg \in W^{k,p}(\Omega)$ and there exists a constant $C = C(d, p) > 0$ such that

$$
\|fg\|_{W^{k,p}(\Omega)} \leq C \sum_{i=0}^k \|f\|_{W^{i,\infty}(\Omega)} \|g\|_{W^{k-i,p}(\Omega)},
$$

and, consequently

$$
\|fg\|_{W^{k,p}(\Omega)} \leq C\|f\|_{W^{k,\infty}(\Omega)}\|g\|_{W^{k,p}(\Omega)}.
$$

Proof. For $k = 0$ the statement is obvious.

For $k = 1$ we get from \cite[Lemma B.6]{21} that there exists a constant $C = C(d,p) > 0$ such that

$$
\|fg\|_{W^{1,p}(\Omega)} \leq C \left( \|f\|_{W^{1,\infty}(\Omega)} \|g\|_{L^p(\Omega)} + \|f\|_{L^\infty(\Omega)} \|g\|_{W^{1,p}(\Omega)} \right),
$$

from which the statement can easily be deduced.

For $k = 2$ it follows from \cite[Chap. 7.3]{19} that the usual product rule also holds for the second order derivatives such that we have

$$
\|fg\|_{W^{2,p}(\Omega)} \leq C \sum_{i,j=1}^d \left\| \frac{\partial^2}{\partial x_i \partial x_j} f g \right\|_{L^p(\Omega)} + \left\| \frac{\partial}{\partial x_i} f \frac{\partial}{\partial x_j} g \right\|_{L^p(\Omega)} + \left\| \frac{\partial}{\partial x_j} f \frac{\partial}{\partial x_i} g \right\|_{L^p(\Omega)} + \left\| \frac{\partial^2}{\partial x_i \partial x_j} g \right\|_{L^p(\Omega)}

\leq C \left( \|f\|_{W^{2,\infty}(\Omega)}\|g\|_{L^p(\Omega)} + \|f\|_{W^{1,\infty}(\Omega)}\|g\|_{W^{1,\infty}(\Omega)} + \|f\|_{L^\infty(\Omega)}\|g\|_{W^{2,\infty}(\Omega)} \right).
$$

Again the overall statement follows easily. \hfill \square
The following corollary establishes a chain rule estimate for $W^{2,\infty}$.

**Corollary B.6.** Let $d, m \in \mathbb{N}$ and $\Omega_1 \subset \mathbb{R}^d$, $\Omega_2 \subset \mathbb{R}^m$ both be open, bounded, and convex. Then, there exists a constant $C = C(d, m) > 0$ with the following property:

If $f \in W^{2,\infty}(\Omega_1; \mathbb{R}^m) \cap C^1(\Omega_1; \mathbb{R}^m)$ and $g \in W^{2,\infty}(\Omega_2) \cap C^1(\Omega_2)$ such that $\text{Range} f \subset \Omega_2$, then for the composition $g \circ f$ it holds that $g \circ f \in W^{2,\infty}(\Omega_1) \cap C^1(\Omega_1)$ and we have

$$|g \circ f|_{W^{1,\infty}(\Omega_1)} \leq C |g|_{W^{1,\infty}(\Omega_2)} |f|_{W^{1,\infty}(\Omega_1; \mathbb{R}^m)},$$

and

$$|g \circ f|_{W^{2,\infty}(\Omega_1)} \leq C \left(|g|_{W^{2,\infty}(\Omega_2)} |f|_{W^{1,\infty}(\Omega_1; \mathbb{R}^m)} + |g|_{W^{1,\infty}(\Omega_2)} |f|_{W^{2,\infty}(\Omega_1; \mathbb{R}^m)}\right).$$

**Proof.** The result can be shown by basic computations using the classical first derivative and [21, Corollary B.5, Lemma B.6].

---

**C Neural Network Calculus**

In this section, we introduce several operations one can perform with neural networks, namely the implementation of the identity function, the (sparse) concatenation and the parallelization.

A useful tool when constructing approximations of functions by neural networks is to be able to implement the identity function with a neural network. This is for example possible for ReLU-neural networks with an arbitrary number of layers and a moderate number of nonzero weights (see [39]). For arbitrary activation functions, however, this is in general not possible. One way to circumvent this problem is to use local approximations of the identity (see e.g. [38, Proposition B.3]). However in case of the ELU, we can use a simpler strategy. For non-negative $x$ the ELU is already defined as the identity. By shifting we can construct an ELU-neural network implementing the identity function on $[-B, \infty)^d$ for an arbitrary but fixed $B \geq 0$.

**Proposition C.1.** Let $\varrho : \mathbb{R} \rightarrow \mathbb{R}$ such that $\varrho(x) = x$ for all $x \geq 0$. Additionally, let $d, L \in \mathbb{N}$ and let $B \geq 0$. Then there exists a neural network $\Phi_B^{L,d}$ with $d$-dimensional input and

1. $R_\varrho(\Phi_B^{L,d})(x) = x$ for all $x \in [-B, \infty)^d$;
2. $L$ layers;
3. $M(\Phi_B^{L,d}) \leq (L + 2)d$ (for $B = 0$, only $M(\Phi_B^{L,d}) \leq Ld$);
4. $\|\Phi_B^{L,d}\|_{\max} \leq \max\{1, B\}$.

**Proof.** For $L = 1$ the claim is obvious. For $L \in \mathbb{N}_{\geq 2}$ the claim follows from the observation that

$$\varrho(\varrho(\ldots(\varrho(x + \underbrace{(B, \ldots, B)^T}_{\text{L-1 times}}))\ldots)) - \underbrace{(B, \ldots, B)^T}_{\in \mathbb{R}^d} = x + (B, \ldots, B)^T - (B, \ldots, B)^T = x$$

for all $x \in [-B, \infty)^d$.

Next, we consider the concatenation of two neural networks as given in [39].

**Definition C.2.** Let $\Phi^1 = ((A_1^1, b_1^1), \ldots, (A_{L_1}^1, b_{L_1}^1))$ and $\Phi^2 = ((A_1^2, b_1^2), \ldots, (A_{L_2}^2, b_{L_2}^2))$ be two neural networks such that the input dimension of $\Phi^1$ is equal to the output dimension of $\Phi^2$. Then the concatenation of $\Phi^1, \Phi^2$ is defined as the $L_1 + L_2 - 1$-layer neural network

$$\Phi^1 \bullet \Phi^2 := ((A_1^1, b_1^1), \ldots, (A_{L_2-1}^2, b_{L_2-1}^2), (A_1^1 A_{L_2}^2 + b_1^1), (A_2^1 b_{L_2}^2 + b_1^1), (A_1^1, b_2^1), \ldots, (A_{L_1}^1, b_{L_1}^1)).$$
It is easy to see that \( R_\varrho(\Phi^1 \cdot \Phi^2) = R_\varrho(\Phi^1) \circ R_\varrho(\Phi^2) \). However, it is not clear how the number of nonzero weights of \( \Phi^1 \cdot \Phi^2 \) relates to \( M(\Phi^1) \) and \( M(\Phi^2) \). This motivates the definition of a sparse concatenation, which is a slight adaption from [39] and makes use of the identity result from Proposition C.1.

**Lemma C.3.** Let \( \Phi^1, \Phi^2 \) be two neural networks such that the input dimension \( n \) of \( \Phi^1 \) is equal to the output dimension of \( \Phi^2 \). We denote by \( d \in \mathbb{N} \) the input dimension of \( \Phi^2 \). Let \( \varrho : \mathbb{R} \to \mathbb{R} \) such that \( \varrho(x) = x \) for all \( x \geq 0 \). Moreover, for some \( K \subset \mathbb{R}^d \) we assume that there exists \( B \geq 0 \) such that for all \( x \in K \) it holds \( R_\varrho(\Phi^2)(x) \in [-B, \infty)^n \). Then, the sparse concatenation of \( \Phi^1 \) and \( \Phi^2 \) is defined as

\[
\Phi^1 \circ \Phi^2 := \Phi^1 \cdot \Phi_B^{n,2} \cdot \Phi^2,
\]

where \( \Phi_B^{n,2} \) is the neural network as defined in Proposition C.1. We have

\[
(i) \quad R_\varrho(\Phi^1 \circ \Phi^2)(x) = R_\varrho(\Phi^1) \circ R_\varrho(\Phi^2)(x) \quad \text{for all } x \in K;
\]

\[
(ii) \quad L(\Phi^1 \circ \Phi^2) = L_1 + L_2;
\]

\[
(iii) \quad M(\Phi^1 \circ \Phi^2) \leq M(\Phi^1) + M(\Phi^2) + 2n;
\]

\[
(iv) \quad \|\Phi^1 \circ \Phi^2\|_{\text{max}} \leq 2 \max\{B, 1\} \cdot \max\{\|\Phi^1\|_{\text{max}}, \|\Phi^2\|_{\text{max}}\}.
\]

**Proof.** The proof of (i) and the fact that \( L(\Phi^1 \circ \Phi^2) = L_1 + L_2 \) immediately follows from the definition of the concatenation of two neural networks in combination with Proposition C.1. Moreover, it is not hard to see that

\[
\Phi^1 \circ \Phi^2 = 

(\sum_{i \in \mathbb{N}} (A^1_i, b^1_i), (A^2_{L_2}, b^2_{L_2}), (B, \ldots, B)^T, A^1_1, -A^1_1 \cdot (B, \ldots, B)^T + b^1_1), (A^1_{L_1}, b^1_{L_1}), \ldots, (A^1_{L_1}, b^1_{L_1})),
\]

from which the other statements follow.

In the next lemma we introduce the parallelization of neural networks with a potentially different number of layers. This result relies again on the identity implementation from Proposition C.1.

**Lemma C.4.** Let \( \varrho : \mathbb{R} \to \mathbb{R} \) such that \( \varrho(x) = x \) for all \( x \geq 0 \). Additionally, let \( \Phi^1, \ldots, \Phi^n \) be neural networks with \( d \)-dimensional input and \( L_1, \ldots, L_n \in \mathbb{N} \) layers, respectively. Then, for all \( B \geq 0 \) there exists a neural network \( P(\Phi^1, \ldots, \Phi^n) \) with \( d \)-dimensional input and

\[
(i) \quad R_\varrho(P(\Phi^1, \ldots, \Phi^n))(x) = (R_\varrho(\Phi^1)(x), \ldots, R_\varrho(\Phi^n)(x)) \quad \text{for all } x \in [-B, \infty)^d;
\]

\[
(ii) \quad \max\{L_1, \ldots, L_n\} \text{ layers};
\]

\[
(iii) \quad P(\Phi^1, \ldots, \Phi^n) \leq \sum_{i=1}^n M(\Phi^i) + nd \left( \max\{L(\Phi^1), \ldots, L(\Phi^n)\} + 4 \right)
\]

(for \( B = 0 \), only \( P(\Phi^1, \ldots, \Phi^n) \leq \sum_{i=1}^n M(\Phi^i) \));

\[
(iv) \quad ||P(\Phi^1, \ldots, \Phi^n)||_{\text{max}} \leq 2 \max\{B, 1\} \cdot \max\{\|\Phi^1\|_{\text{max}}, \ldots, \|\Phi^n\|_{\text{max}}\}.
\]

**Proof.** First of all we assume that \( L_1 = \cdots = L_n \). Then

\[
P(\Phi^1, \ldots, \Phi^n) := (A_1, b_1), \ldots, (A_L, b_L),
\]
with
\[
\tilde{A}_1 := \begin{pmatrix} A_1^1 \\ \vdots \\ A_1^n \end{pmatrix}, \quad \tilde{b}_1 := \begin{pmatrix} b_1^1 \\ \vdots \\ b_1^n \end{pmatrix} \quad \text{and} \quad \tilde{A}_\ell := \begin{pmatrix} A_1^1 & A_2^1 & \cdots & A_n^1 \\ A_1^2 & A_2^2 & \cdots & A_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ A_1^n & A_2^n & \cdots & A_n^n \end{pmatrix}, \quad \tilde{b}_\ell := \begin{pmatrix} b_1^1 \\ b_1^2 \\ \vdots \\ b_1^n \end{pmatrix}, \quad \text{for } 1 < \ell \leq L,
\]
fulfills all the desired properties. Now we assume that the number of layers \( L_1, \ldots, L_n \) are potentially different. Set \( L_{\text{max}} := \max\{L_1, \ldots, L_n\} \) and, for \( i = 1, \ldots, n \), define
\[
\Phi_i, L_{\text{max}} := \begin{cases} \Phi_i, & \text{if } L(\Phi_i) = L_{\text{max}}, \\ \Phi_i \circ \Phi_{B}^{L_{\text{max}}-L_d}, & \text{else}, \end{cases}
\]
where the networks \( \Phi_{B}^{L_{\text{max}}-L_d} \) are as in Proposition \( \ref{prop:approximate_partition_of_unity} \). Then, we define
\[
P(\Phi^1, \ldots, \Phi^n) := P(\Phi^1, L_{\text{max}}, \ldots, \Phi^n, L_{\text{max}}).
\]
It is easy to see that \( \text{(i)} \) and \( \text{(ii)} \) hold. Furthermore, we have
\[
M(\Phi_{i}, L_{\text{max}}) \leq n \sum_{i=1}^{n} M(\Phi^n) + (L_{\text{max}} + 4) nd
\]
and, finally,
\[
\|P(\Phi^1, \ldots, \Phi^n)\|_{\text{max}} \leq \max\{\|\Phi_1\|_{\text{max}}, \ldots, \|\Phi^n\|_{\text{max}} \cdot B, \|A_1^1\|_{\text{max}} B + \|b_1^1\|_{\text{max}} + \ldots, \|A_1^n\|_{\text{max}} B + \|b_1^n\|_{\text{max}} \}
\]
\[
\leq 2 \max\{B, 1\} \cdot \max\{\|\Phi_1\|_{\text{max}}, \ldots, \|\Phi^n\|_{\text{max}} \}. \]

\[\square\]

**D Proof of Proposition 4.1**

The goal of this section is the proof of Proposition 4.1. For an overview of the required steps see the proof outline below Proposition 4.1.

**D.1 Approximate Partition of Unity by ELU-neural networks**

We start with the proof of the existence and technical details of the approximate partition of unity. First of all, we give the definition of the involved bump functions whose sum approximates the 1-function on \((0, 1)^d\) (see Figure 3 for an illustration for the case \( d = 1 \)).

**Definition D.1.** For a scaling factor \( s \geq 1 \) we define the scaled ELU by
\[
\varphi^s : \mathbb{R} \to \mathbb{R}, \quad \varphi^s(x) := \frac{\text{ELU}_1(sx)}{s}.
\]

Moreover, we define a scaled one-dimensional bump function by
\[
\psi^s : \mathbb{R} \to \mathbb{R}, \quad \psi^s(x) := \varphi^s(x + 2) - \varphi^s(x + 1) - \varphi^s(x) + \varphi^s(x - 2),
\]
and, for \( d, N \in \mathbb{N} \) and \( m \in \{0, \ldots, N\}^d \) we define multi-dimensional bumps \( \phi^s_m : \mathbb{R}^d \to \mathbb{R} \) as a tensor product of scaled and shifted versions of \( \psi^s \). Concretely, we set
\[
\phi^s_m(x) := \prod_{i=1}^{d} \psi^s(3N \left(x_i - \frac{m_i}{N}\right)).
\]

Finally, the collection of bump functions is denoted by \( \Psi := \{\phi^s_m : m \in \{0, \ldots, N\}^d\} \).
We will use $\Psi$ as an approximate partition of unity, i.e., $\sum_{m \in \{0, \ldots, N\}^d} \phi^n_m$ approximates $1_{(0,1)^d}$ in $W^{2,\infty}$. Note that the $\phi^n_m$ do not have compact support but (as we will see in Proposition D.3) decay exponentially fast. Before we state the corresponding result, we first need a technical result for the one-dimensional functions $\psi^s$. For a reader more interested in the overall properties of the partition of unity we suggest to skip the lengthy computations of the following lemma and directly jump to Proposition D.3

**Lemma D.2.** Let $s \geq 1$, $N \in \mathbb{N}$, and $k \in \{0,1,2\}$. Then there exists some absolute constant $C > 0$ such that for all $m \in \{0, \ldots, N\}$ we have

$$\left| \psi^s \left( 3N \left( \cdot - \frac{m}{N} \right) \right) \right|_{W^{k,\infty}(\mathbb{R})} \leq C \cdot N^k \cdot s^{\max\{0,k-1\}}.$$ 

**Proof.** Let $x \in \mathbb{R}$. First of all, we note

$$\psi^s \left( 3N \left( \cdot - \frac{m}{N} \right) \right)(x)$$

\[
= \begin{cases} 
0, & \text{if } x \geq \frac{2}{3N} + \frac{m}{N}, \\
2 - 3Nx + 3m + \frac{1}{2} \left( e^{s(3Nx-3m-2)} - 1 \right), & \text{if } \frac{1}{3N} + \frac{m}{N} \leq x < \frac{2}{3N} + \frac{m}{N}, \\
1 + \frac{1}{2} \left( -e^{s(3Nx-3m-1)} + e^{s(3Nx-3m-2)} \right), & \text{if } \frac{1}{3N} + \frac{m}{N} \leq x < \frac{1}{3N} + \frac{m}{N}, \\
2 + 3Nx - 3m + \frac{1}{2} \left( 1 - e^{s(3Nx-3m+1)} - e^{s(3Nx-3m-1)} + e^{s(3Nx-3m-2)} \right), & \text{if } \frac{2}{3N} + \frac{m}{N} \leq x < \frac{1}{3N} + \frac{m}{N}, \\
\frac{1}{2} \left( e^{s(3Nx-3m+2)} - e^{s(3Nx-3m-1)} - e^{s(3Nx-3m-1)} + e^{s(3Nx-3m-2)} \right), & \text{if } x < \frac{2}{3N} + \frac{m}{N}.
\end{cases}
\]

Hence, we have for the first classical derivative

$$\left( \psi^s \left( 3N \left( \cdot - \frac{m}{N} \right) \right) \right)'(x)$$

\[
= \begin{cases} 
0, & \text{if } x \geq \frac{2}{3N} + \frac{m}{N}, \\
-3N + 3N \cdot e^{s(3Nx-3m-2)}, & \text{if } \frac{1}{3N} + \frac{m}{N} \leq x < \frac{2}{3N} + \frac{m}{N}, \\
3N \left( -e^{s(3Nx-3m-1)} + e^{s(3Nx-3m-2)} \right), & \text{if } \frac{1}{3N} + \frac{m}{N} \leq x < \frac{1}{3N} + \frac{m}{N}, \\
3N + 3N \left( -e^{s(3Nx-3m+1)} - e^{s(3Nx-3m-1)} + e^{s(3Nx-3m-2)} \right), & \text{if } \frac{2}{3N} + \frac{m}{N} \leq x < \frac{1}{3N} + \frac{m}{N}, \\
3N \left( e^{s(3Nx-3m+2)} - e^{s(3Nx-3m-1)} - e^{s(3Nx-3m-1)} + e^{s(3Nx-3m-2)} \right), & \text{if } x < \frac{2}{3N} + \frac{m}{N},
\end{cases}
\]

and for the second order weak derivative

$$\left( \psi^s \left( 3N \left( \cdot - \frac{m}{N} \right) \right) \right)''(x)$$

\[
= \begin{cases} 
0, & \text{if } x \geq \frac{2}{3N} + \frac{m}{N}, \\
9N^2s \cdot e^{s(3Nx-3m-2)}, & \text{if } \frac{1}{3N} + \frac{m}{N} \leq x < \frac{2}{3N} + \frac{m}{N}, \\
9N^2s \left( -e^{s(3Nx-3m-1)} + e^{s(3Nx-3m-2)} \right), & \text{if } \frac{1}{3N} + \frac{m}{N} \leq x < \frac{1}{3N} + \frac{m}{N}, \\
9N^2s \left( -e^{s(3Nx-3m+1)} - e^{s(3Nx-3m-1)} + e^{s(3Nx-3m-2)} \right), & \text{if } \frac{2}{3N} + \frac{m}{N} \leq x < \frac{1}{3N} + \frac{m}{N}, \\
9N^2s \left( e^{s(3Nx-3m+2)} - e^{s(3Nx-3m+1)} - e^{s(3Nx-3m-1)} + e^{s(3Nx-3m-2)} \right), & \text{if } x < \frac{2}{3N} + \frac{m}{N}.
\end{cases}
\]

We start with $k = 0$ and for this we estimate $\left| \psi^s \left( 3N \left( x - \frac{m}{N} \right) \right)(x) \right|$ for $x$ from different intervals. Keep in mind that $s \geq 1$.

**Case** $(x \geq \frac{2}{3N} + \frac{m}{N})$: We trivially have $\psi^s \left( 3N \left( x - \frac{m}{N} \right) \right)(x) = 0.$
Case \( \left( \frac{1}{3N} + \frac{m}{N} \leq x \leq \frac{2}{3N} + \frac{m}{N} \right) \): On the one hand we have

\[
\psi^s \left( 3N \left( x - \frac{m}{N} \right) \right) = 2 - 3Nx + 3m + \frac{1}{s} \left( e^{s(3Nx-3m-2)} - 1 \right)
\leq 2 - 3N \left( \frac{1}{3N} + \frac{m}{N} \right) + 3m + \frac{1}{s} \left( e^{s\left( \frac{3N}{N} + \frac{m}{N} \right) - 3m - 2} - 1 \right)
= 1 + \frac{1}{s} \left( e^{s(2+3m-3m-2)} - 1 \right) = 1.
\]

On the other hand we have

\[
\psi^s \left( 3N \left( x - \frac{m}{N} \right) \right) = 2 - 3Nx + 3m + \frac{1}{s} \left( e^{s(3Nx-3m-2)} - 1 \right)
\geq 2 - 3N \left( \frac{2}{3N} + \frac{m}{N} \right) + 3m + \frac{1}{s} \left( e^{s\left( \frac{3N}{N} + \frac{m}{N} \right) - 3m - 2} - 1 \right)
= \frac{1}{s} \left( e^{s(2+3m-3m-2)} - 1 \right) = \frac{1}{s} (e^{-s} - 1) \geq -\frac{1}{s}.
\]

This implies that we have \( |\psi^s (3N \left( x - \frac{m}{N} \right))| \leq \max \{ 1, \frac{1}{s} \} \).

Case \( \left( \frac{1}{3N} + \frac{m}{N} \leq x \leq \frac{1}{3N} + \frac{m}{N} \right) \): Then, on the one hand we have

\[
\psi^s \left( 3N \left( x - \frac{m}{N} \right) \right) (x) = 1 + \frac{1}{s} \left( e^{s(3Nx-3m)} + e^{s(3Nx-3m-2)} \right)
\leq 1 + \frac{1}{s} \left( e^{s(3N(-\frac{1}{3N} + \frac{m}{N}) - 3m)} + e^{s(3N(-\frac{1}{3N} + \frac{m}{N}) - 3m-2)} \right)
= 1 + \frac{1}{s} (e^{-s} - e^{-2s}) \leq 1 + \frac{e^{-s}}{s} \leq 1 + \frac{1}{s}.
\]

On the other hand we obtain that

\[
\psi^s \left( 3N \left( x - \frac{m}{N} \right) \right) (x) = 1 + \frac{1}{s} \left( e^{s(3Nx-3m)} + e^{s(3Nx-3m-2)} \right)
\geq 1 + \frac{1}{s} \left( e^{s(3N(-\frac{1}{3N} + \frac{m}{N}) - 3m-1)} + e^{s(3N(-\frac{1}{3N} + \frac{m}{N}) - 3m-2)} \right)
= 1 + \frac{1}{s} (e^{-3s} - e^{0}) \geq 1 - \frac{1}{s}.
\]

Together we get \( |\psi^s (3N \left( x - \frac{m}{N} \right))| \leq 1 + \frac{1}{s} \).

Case \( \left( \frac{3}{3N} + \frac{m}{N} \leq x \leq \frac{2}{3N} + \frac{m}{N} \right) \): We have

\[
\psi^s \left( 3N \left( x - \frac{m}{N} \right) \right) (x)
= 2 + 3Nx - 3m + \frac{1}{s} \left( -e^{s(3Nx-3m+1)} - e^{s(3Nx-3m-1)} + e^{s(3Nx-3m-2)} + 1 \right)
\leq 2 + 3N \left( -\frac{1}{3N} + \frac{m}{N} \right) - 3m
+ \frac{1}{s} \left( -e^{s(3N(-\frac{2}{3N} + \frac{m}{N}) - 3m+1)} - e^{s(3N(-\frac{2}{3N} + \frac{m}{N}) - 3m-1)} + e^{s(3N(-\frac{2}{3N} + \frac{m}{N}) - 3m-2)} + 1 \right)
= 1 + \frac{1}{s} \left( e^{-s} - e^{-3s} + e^{-3s} + 1 \right) = 1 + \frac{1}{s} \left( 1 - e^{-s} \right)
\leq 1 + \frac{1}{s}.
\]
On the other hand, we have
\[ \psi^s \left( 3N \left( x - \frac{m}{N} \right) \right) (x) \]
\[ = 2 + 3Nx - 3m + \frac{1}{s} \left( -e^{s(3Nx - 3m + 1)} - e^{s(3Nx - 3m - 1)} + e^{s(3Nx - 3m - 2)} + 1 \right) \]
\[ \geq 2 + 3N \left( -\frac{2}{3N} + \frac{m}{N} \right) - 3m \]
\[ + \frac{1}{s} \left( -e^{s(3N(-\frac{2}{3N} + \frac{m}{N}) - 3m + 1)} - e^{s(3N(-\frac{2}{3N} + \frac{m}{N}) - 3m - 1)} + e^{s(3N(-\frac{2}{3N} + \frac{m}{N}) - 3m - 2)} + 1 \right) \]
\[ = \frac{1}{s} \left( -e^0 - e^{-2s} + e^{-4s} + 1 \right) = \frac{1}{s} (-e^{-2s} + e^{-4s}) \]
\[ \geq e^{-4s} \geq 0. \]

Combining the two estimates we get \(|\psi^s \left( 3N \left( x - \frac{m}{N} \right) \right)\) \(\leq 1 + \frac{1}{s}\).

**Case** \((x \leq \frac{-2}{3N} + \frac{m}{N})\): Then
\[ \left| \psi^s \left( 3N \left( x - \frac{m}{N} \right) \right) \right| = \frac{1}{s} e^{s(3N x - 3m)} \left| e^2 - e - e^{-1} \right| \leq \frac{1}{s} e^{sN \left( \frac{2}{3N} + \frac{m}{N} \right) - 3m} . \]

In total, we obtain
\[ \left| \psi^s \left( 3N \left( x - \frac{m}{N} \right) \right) \right| \leq \begin{cases} 0, & \text{if } x \geq \frac{2}{3N} + \frac{m}{N}, \\ \max \left\{ 1, \frac{1}{s} \right\}, & \text{if } \frac{1}{3N} + \frac{m}{N} \leq x < \frac{2}{3N} + \frac{m}{N}, \\ 1 + \frac{1}{s}, & \text{if } \frac{-1}{3N} + \frac{m}{N} \leq x \leq \frac{1}{3N} + \frac{m}{N}, \\ 1 + \frac{1}{s}, & \text{if } \frac{-2}{3N} + \frac{m}{N} \leq x < \frac{1}{3N} + \frac{m}{N}, \\ \frac{5}{s}, & \text{if } x \leq \frac{-2}{3N} + \frac{m}{N}, \end{cases} \]

and, consequently, for \(s \geq 1\), we have
\[ \left| \psi^s \left( 3N \left( \cdot - \frac{m}{N} \right) \right) \right|_{W^{0,\infty}} \leq 5. \]

We proceed with estimating the first and second derivatives. First of all, it is clear that for all \(i \in \mathbb{Z}\) and \(x \leq \frac{i}{3N} + \frac{m}{N}\) and for all \(j \geq i\) we have
\[ e^{s(3Nx - 3m - j)} \leq e^{s(3N \left( \frac{2}{3N} + \frac{m}{N} \right) - 3m - j)} = e^{s(i + 3m - 3m - j)} \leq e^{s(i - 1)} = 1. \]

Hence, by using the triangle inequality on every of the involved summands as well as the properties of \(x\) with respect to the considered cases we obtain
\[ \left| \left( \psi^s \left( 3N \left( \cdot - \frac{m}{N} \right) \right) \right)' \right| \leq \begin{cases} 0, & \text{if } x \geq \frac{2}{3N} + \frac{m}{N}, \\ 6N, & \text{if } \frac{1}{3N} + \frac{m}{N} \leq x < \frac{2}{3N} + \frac{m}{N}, \\ 6N, & \text{if } \frac{-1}{3N} + \frac{m}{N} \leq x < \frac{1}{3N} + \frac{m}{N}, \\ 12N, & \text{if } \frac{-2}{3N} + \frac{m}{N} \leq x < \frac{-1}{3N} + \frac{m}{N}, \\ 12N, & \text{if } x < \frac{-2}{3N} + \frac{m}{N}. \end{cases} \]

22
This implies \( |\psi^s(3N \cdot - \frac{m}{N})|_{W^{1, \infty}(\mathbb{R})} \leq 12N \). In the same way we get

\[
\left| \left( \psi^s(3N \cdot - \frac{m}{N}) \right)''(x) \right| \leq \begin{cases} 
0, & \text{if } x > \frac{2}{3N} + \frac{m}{N}, \\
9N^2s, & \text{if } \frac{1}{3N} + \frac{m}{N} < x < \frac{2}{3N} + \frac{m}{N}, \\
18N^2s, & \text{if } \frac{1}{3N} + \frac{m}{N} < x < \frac{1}{3N} + \frac{m}{N}, \\
27N^2s, & \text{if } \frac{2}{3N} + \frac{m}{N} < x < \frac{1}{N} + \frac{m}{N}, \\
36N^2s, & \text{if } x < \frac{2}{3N} + \frac{m}{N},
\end{cases}
\]

which implies that \( |\psi^s(3N \cdot - \frac{m}{N})|_{W^{2, \infty}(\mathbb{R})} \leq 36N^2s. \)

The following result establishes the properties we need from our approximate partition of unity. In particular, we need to control the Sobolev norm of the bump functions; show that the bumps, which do not have compact support, decay sufficiently fast outside a ball containing their center; show that the sum of the bumps approximates \( 1_{(0, 1)^d} \) in \( W^{2, \infty} \); and, finally, that the bumps can be implemented by an ELU-neural network. The last property is not very surprising since was exactly the motivation for the construction of the bumps.

**Proposition D.3.** For any \( d, N \in \mathbb{N} \) and scaling parameter \( s > 1 \), the collection of functions

\[
\Psi = \{ \phi^s_m : m \in \{0, \ldots, N\}^d \}
\]

from Definition D.2 fulfills the following properties for all \( k \in \{0, 1, 2\} \) and a constant \( C = C(d) > 0 \):

(i) \( 0 \leq \phi^s_m(x) \leq 1 \) for every \( \phi^s_m \in \Psi \) and every \( x \in \mathbb{R}^d \);

(ii) \( \|\phi^s_m\|_{W^{k, \infty}(\mathbb{R}^d)} \leq CN^k \cdot s^{\max(0, k-1)} \) for every \( \phi^s_m \in \Psi \);

(iii) for \( \Omega^\text{exp}_m := \{x \in \mathbb{R}^d : \|x - \frac{m}{N}\|_{\infty} \geq \frac{5}{6N} \} \), we have \( \|\phi^s_m\|_{W^{k, \infty}(\Omega^\text{exp}_m)} \leq CN^k s^{k-1} e^{-\frac{1}{2}s} \) for every \( \phi^s_m \in \Psi \). In particular, if we choose \( s = N^\mu \) for an arbitrary \( \mu > 0 \), for all \( \gamma > 0 \) and some \( C' = C'(\mu, \gamma, k, d) > 0 \) there holds

\[
|\phi^s_m|_{W^{k, \infty}(\Omega^\text{exp}_m)} \leq C'N^{-\gamma};
\]
(iv) let \( \mu > 0 \), then there exists \( N_1 = N_1(\mu) \in \mathbb{N} \) such that

\[
\left\| I_{(0,1)^d} - \sum_{m=0}^{N} \phi_m^s \right\|_{W^{k,\infty}((0,1)^d)} \leq C N^{k} s^{k-1} e^{-s}
\]

for \( s = N^\mu \) and \( N \geq N_1 \);

(v) For each \( \phi_m^s \in \Psi \) there is a neural network \( \Phi_m^s \) with \( d \)-dimensional input and \( d \)-dimensional output, with two layers and \( C \) nonzero weights, that satisfies \( [R_\psi(\Phi_m^s)]_l = \psi^s \left( 3N \left( x_l - \frac{m}{N} \right) \right) \) for \( l = 1, \ldots, d \) such that

\[
\prod_{l=1}^{d} [R_\psi(\Phi_m^s)]_l = \phi_m^s,
\]

and \( \|R_\psi(\Phi_m^s)]_l\|_{W^{k,\infty}((0,1)^d)} \leq C N^k \cdot s^{\max\{0,k-1\}} \) for all \( l = 1, \ldots, d \) and \( k \in \{0,1,2\} \). Furthermore, for the weights of \( \Phi_m^s \) it holds that \( \|\Phi_m^s\|_{\text{max}} \leq CsN \).

**Proof.** We omit the proof of (v). It can be done by some very technical and unenlightening calculations which are of the flavor of those provided in the proof of Lemma \ref{lem:tensor}.4.

**ad (ii):** Since we will need it in the proof of (iii), we prove the following more general statement (Statement (i) follows by considering \( I = \{1, \ldots, d\} \)):

Let \( I \subset \{1, \ldots, d\} \) be arbitrary. Moreover, for \( m \in \{0, \ldots, N\}^{\|I\|} \) we define \( \phi_{m,I}^s \colon \mathbb{R}^{\|I\|} \to \mathbb{R}, x \mapsto \prod_{1 \leq i \leq \|I\|} \phi_{\psi}(3N (x_i - \frac{m_i}{N})) \) as well as \( \phi_{m,I}^s := \phi_{m,I}^s \), if \( I = \{1, \ldots, d\} \). Then for \( k \in \{0,1,2\} \) it holds that

\[
\left| \phi_{m,I}^s \right|_{W^{k,\infty}(\mathbb{R}^{\|I\|})} \leq C^{\|I\|} \cdot N^{k} \cdot s^{\max\{0,k-1\}}.
\]

It is clear that by the definition of \( \phi_{m,I}^s \) and by employing Lemma \ref{lem:tensor}.2 for \( k = 0 \) there holds

\[
\left| \phi_{m,I}^s \right|_{W^{0,\infty}(\mathbb{R}^{\|I\|})} \leq C^{\|I\|}. \tag{D.1}
\]

Now, let \( i \in I \) be arbitrary. Then, by using the tensor product structure of \( \phi_{m,I}^s \) in combination with Lemma \ref{lem:tensor}.2 for the case \( k = 1 \) and \ref{lem:tensor}.1 for \( I' := I \setminus \{i\} \) we obtain for an arbitrary \( x \in \mathbb{R}^{\|I\|} \)

\[
\left| \frac{\partial}{\partial x_i} \phi_{m,I}^s(x) \right| = \left| \phi_{m,I'}^s(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{\|I\|}) \right| \cdot \left| \left( \psi^s \left( 3N \left( -m_i/N \right) \right) \right)'(x_i) \right| 
\leq C^{\|I\| - 1} \cdot CN = C^{\|I\|} N
\]

which implies that \( \left| \phi_{m,I}^s \right|_{W^{1,\infty}(\mathbb{R}^{\|I\|})} \leq C^{\|I\|} N \).

Finally, let additionally be \( j \in I \) be arbitrary. If \( i = j \) then we have that (by using \ref{lem:tensor}.1 in combination with Lemma \ref{lem:tensor}.2 for \( k = 2 \)) that

\[
\left| \frac{\partial^2}{\partial x_i^2} \phi_{m,I}^s(x) \right| = \left| \phi_{m,I'}^s(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{\|I\|}) \right| \cdot \left| \left( \psi^s \left( 3N \left( -m_i/N \right) \right) \right)''(x_i) \right| 
\leq C^{\|I\| - 2} \cdot C N^2 = C^{\|I\|} N^2.
\]

Moreover, if \( i \neq j \), then, if we set \( I'' := I \setminus \{i, j\} \) we obtain with similar arguments as before that

\[
\left| \frac{\partial^2}{\partial x_i \partial x_j} \phi_{m,I}^s(x) \right| 
= \left| \phi_{m,I'}^s(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_j, \ldots, x_{\|I\|}) \right| 
\cdot \left| \left( \psi^s \left( 3N \left( -m_i/N \right) \right) \right)'(x_i) \right| \cdot \left| \left( \psi^s \left( 3N \left( -m_j/N \right) \right) \right)'(x_j) \right| 
\leq C^{\|I''\| - 2} \cdot C N \cdot C N = C^{\|I''\|} N^2,
\]

24
where we assumed w.l.o.g. that \( i < j \). This implies \( |\phi^*_{m,l}|_{W^{2,\infty}(\mathbb{R}^{|I|})} \leq C|l|N^2s \), which yields the claim.

\[ \text{ad (iii): For } x \in \mathbb{R}, \text{ we set} \]

\[ \Theta^s(x) := e^{s(3Nx-3m+2)} - e^{s(3Nx-3m+1)} - e^{s(3Nx-3m-1)} + e^{s(3Nx-3m-2)} = e^{3s(Nx-m)} \cdot (e^{2s} - e^{s} - e^{-s} + e^{-2s}) \]

and get

\[ \psi^s(3N(x - m/N)) = \begin{cases} 0, & \text{if } x \geq \frac{5}{6N} + \frac{m}{N}, \\ \frac{1}{2} \cdot \Theta^s(x), & \text{if } x \leq -\frac{5}{6N} + \frac{m}{N}. \end{cases} \]

Hence, we have

\[ (\psi^s(3N(-m/N)))' (x) = \begin{cases} 0, & \text{if } x \geq \frac{5}{6N} + \frac{m}{N}, \\ 3N \cdot \Theta^s(x), & \text{if } x \leq -\frac{5}{6N} + \frac{m}{N}, \end{cases} \]

and

\[ (\psi^s(3N(-m/N)))'' (x) = \begin{cases} 0, & \text{if } x \geq \frac{5}{6N} + \frac{m}{N}, \\ 9N^2s \cdot \Theta^s(x), & \text{if } x \leq -\frac{5}{6N} + \frac{m}{N}. \end{cases} \]

Furthermore, for every \( x \in \mathbb{R} \) such that \( x \leq -\frac{5}{6N} + \frac{m}{N} \) we have by the triangle inequality that

\[ |\Theta^s(x)| \leq 4e^{3s(Nx-m)} \cdot e^{2s} \leq 4e^{3s(N(-\frac{5}{6N} + \frac{m}{N})-m)+2s} = 4e^{-s}\frac{3}{6N}+2s = 4e^{-\frac{s}{2}}. \]

Trivially, we also have \( |\Theta^s(x)| \leq Ce^{-\frac{s}{2}} \) for all \( x \geq \frac{5}{6N} + \frac{m}{N} \). Hence, for \( k \in \{0, 1, 2\} \) we conclude directly that

\[ |\psi^s(3N(-m/N))|_{W^{k,\infty}((-\infty, -\frac{5}{6N}| \cup [\frac{5}{6N}, \infty)))} \leq C N^k s^{k-1} \cdot e^{-\frac{s}{2}}. \]  

(D.2)

Now, let \( x \in \Omega_{\text{exp}}^N \). Then there exists some \( l \in \{1, \ldots, d\} \) with \( |x_l - \frac{m}{N}| \geq \frac{5}{6N} \). This implies for \( I' = \{1, \ldots, d\} \setminus \{l\} \) by employing Equation (D.1) together with (D.2) that

\[ |\phi^*_{m}(x)| = |\phi^*_{m,l}(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_d)| \cdot |\psi^s(3N(x_l - m/N))| \leq C^{d-1} \cdot \frac{Ce^{-\frac{s}{2}}}{s}. \]

This shows that \( |\phi^*_{m}|_{W^{n,\infty}(\Omega_{\text{exp}}^N)} \leq C \frac{d}{s} e^{-\frac{s}{2}} \). By proceeding in a similar manner and with the same techniques as in the proof of (ii) and combining these with Equation (D.2), one can show the remaining Sobolev semi-norm estimates for the higher-order derivatives. The “in-particular” part then follows from Proposition A.1.

\[ \text{ad (iv): We start with an observation for the case } d = 1 \text{ that will be helpful later in the multi-dimensional case. For this, let } x \in \mathbb{R}. \text{ Then} \]

\[ \sum_{m=0}^{N} \phi^*_{m}(x) = \sum_{m=0}^{N} \phi^* (3N(x - m/N) + 2) - \phi^* (3N(x - m/N) + 1) - \phi^* (3N(x - m/N) - 1) + \phi^* (3N(x - m/N) - 2) \]

\[ = \phi^* (3Nx + 2) - \phi^* (3Nx + 1) - \phi^* (3N(x - 1) + 1) + \phi^* (3N(x - 1) - 2), \]

(D.3)
down to the same sum. We have for \( x \in \mathbb{R}^d \) that

\[
\sum_{m \in \{0, \ldots, N\}^d} \phi_m^s(x) = \sum_{m \in \{0, \ldots, N\}^d} \prod_{l=1}^d \psi^s \left( 3N \left( x_l - \frac{m_l}{N} \right) \right)
\]

\[
= \prod_{l=1}^d \sum_{m=0}^N \psi^s \left( 3N \left( x_l - \frac{m}{N} \right) \right)
\]

\[
= \prod_{l=1}^d \left( \varphi^s(3Nx_l + 2) - \varphi^s(3Nx_l + 1) - \varphi^s(3Nx_l - 1) + \varphi^s(3Nx_l - 2) \right),
\]

(D.4)

where we used the observation from Equation (D.3) in the last step.

Each of the first two terms of the sum (D.4) corresponds to the identity if \( x_l \geq -1/3N \). From this it follows that

\[
\varphi^s(3Nx_l + 2) - \varphi^s(3Nx_l + 1) = 3Nx_l + 2 - 3Nx_l - 1 = 1, \quad \text{for} \quad x_l \geq -\frac{1}{3N}.
\]

On the other hand, for the second two terms of the sum (D.4), we have

\[
-\varphi^s(3Nx_l - 1) + \varphi^s(3Nx_l - 2) = \frac{e^{s(3Nx_l - 1)} - e^{s(3Nx_l - 2)}}{s} =: \nu_l(x),
\]

for \( x_l \leq 1 + \frac{1}{3N} \) and where \( \nu_l : \mathbb{R}^d \to \mathbb{R} \). Combining this, we get

\[
\sum_{m \in \{0, \ldots, N\}^d} \phi_m^s(x) = \prod_{l=1}^d (1 + \nu_l(x)), \quad \text{for} \quad x \in (0, 1)^d \subset [-1/3N, 1 + 1/3N]^d.
\]

Now, we can estimate

\[
\left\| \mathbb{1}_{(0,1)^d} - \sum_{m \in \{0, \ldots, N\}^d} \phi_m^s \right\|_{W^{k, \infty}((0,1)^d)} = \left\| \sum_{i=1}^d \sum_{1 \leq l_1 < \ldots < l_i \leq d} \nu_{l_1}(x) \ldots \nu_{l_i}(x) \right\|_{W^{k, \infty}((0,1)^d)}
\]

\[
\leq d \max_{i=1, \ldots, d} \sum_{1 \leq l_1 < \ldots < l_i \leq d} \left\| \nu_{l_1}(x) \ldots \nu_{l_i}(x) \right\|_{W^{k, \infty}((0,1)^d)}
\]

(Product rule Lemma B.5) \leq C \max_{i=1, \ldots, d} \sum_{1 \leq l_1 < \ldots < l_i \leq d} \prod_{j=1}^i \left\| \nu_{l_j}(x) \right\|_{W^{k, \infty}((0,1)^d)}
\]

(D.5)

for a constant \( C = C(d) > 0 \). Next, we observe that

\[
\left\| \nu_{l_j}(x) \right\|_{W^{k, \infty}((0,1)^d)} = \max_{n=0, \ldots, k} (3sN)^n \left\| \frac{e^{s(3(N-1)l_j)} - e^{s(3(N-1)l_j-2)}}{s} \right\|_{L^\infty((0,1))}
\]

\[
= \max_{n=0, \ldots, k} (3N)^n s^{n-1} \frac{1}{e^s} \frac{1}{e^{2s}}
\]

\[
\leq \frac{C N^k s^{k-1}}{e^s} < 1 \quad \text{for} \quad s = N^\mu \quad \text{and} \quad N \geq N_1 = N_1(\mu),
\]

where we used \( s \geq 1 \) in the last step and \( C = C(d) > 0 \) is a constant. Combining this with Equation (D.5) finally yields (for \( s = N^\mu \) and \( N \geq N_1 \))

\[
\left\| \mathbb{1}_{(0,1)^d} - \sum_{m \in \{0, \ldots, N\}^d} \phi_m^s \right\|_{W^{k, \infty}((0,1)^d)} \leq C N^k s^{k-1} e^{-s}.
\]
Remark D.5. For all $N$ the coefficients of the polynomials $p$ and $f$ are such that for every $C$ is a constant $\Psi = \Psi(\cdot)$, with $d$-dimensional input and $d$-dimensional output, with two layers, $C$ nonzero weights, that satisfies

$$|R_\phi(\Phi_m^s)|_l = \psi^s\left(3N\left(x_l - \frac{m_l}{N}\right)\right)$$

for $l = 1, \ldots, d$. The largest weight of the network is $3sN$ such that $\|\Phi_m^s\|_{\max} \leq CsN$. The estimate for $\|R_\phi(\Phi_m^s)|_l\|_{W^{k,\infty}((0,1)^d)}$ follows from Lemma D.2.

D.2 Approximation by localized polynomials

With the help of the bump functions from Appendix D.1 and averaged Taylor polynomials, we can be in a position to approximate any $f \in \mathbb{W}^{n,p}((0,1)^d)$ by localized polynomials in higher-order Sobolev norms. The following proposition provides the details.

**Lemma D.4.** Let $k \in \{0, 1, 2\}$, $d, N \in \mathbb{N}$, $n \in \mathbb{N}_{>k+1}$, $1 \leq p \leq \infty$ and $\mu \in (0, 1)$. Set $s := N^\mu$ and let $\Phi = \Phi(d, N, \mu) = \left\{\phi_m^s : m \in \{0, \ldots, N\}^d\right\}$ be the partition of unity from Proposition D.3. Then there is a constant $C = C(d, n, p, \mu) > 0$ and $\tilde{N} = \tilde{N}(d, p, \mu) \in \mathbb{N}$ such that for every $f \in \mathbb{W}^{n,p}((0,1)^d)$ and every $m \in \{0, \ldots, N\}^d$, there exist polynomials $p_{f,m}(x) = \sum_{|\alpha| \leq n-1} c_{f,m,\alpha} x^\alpha$ for $m \in \{0, \ldots, N\}^d$ with the following properties:

Set $f_N := \sum_{m \in \{0, \ldots, N\}^d} \phi_m^s p_{f,m}$. Then, the operator $T_{\phi}^u : \mathbb{W}^{n,p}((0,1)^d) \to \mathbb{W}^{k,p}((0,1)^d)$ with $T_{\phi}^u f = f - f_N$ is linear and bounded with

$$\|T_{\phi}^u f\|_{W^{k,p}((0,1)^d)} \leq C \left(\frac{1}{N}\right)^{n-k} \|f\|_{\mathbb{W}^{n,p}((0,1)^d)}, \quad \text{for } k \in \{0, 1\},$$

and

$$\|T_{\phi}^u f\|_{W^{2,p}((0,1)^d)} \leq C \left(\frac{1}{N}\right)^{n-2-\mu} \|f\|_{\mathbb{W}^{n,p}((0,1)^d)},$$

for all $N \in \mathbb{N}$ with $N \geq \tilde{N}$.

**Remark D.5.** Since the polynomials utilized in Lemma D.4 are the averaged Taylor polynomials from the Bramble-Hilbert Lemma B.4, we get there is a constant $C = C(d, n) > 0$ such that for any $f \in \mathbb{W}^{n,p}((0,1)^d)$ the coefficients of the polynomials $p_{f,m}$ satisfy

$$|c_{f,m,\alpha}| \leq C \|\tilde{f}\|_{\mathbb{W}^{n,p}(\Omega_{m,N})} N^{d/p},$$

for all $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq n-1$, and for all $m \in \{0, \ldots, N\}^d$, where $\Omega_{m,N} := B_{\frac{1}{N}} \cap \|\cdot\|_{\infty} (\frac{m}{N})$ and $\tilde{f} \in \mathbb{W}^{n,p}(\mathbb{R}^d)$ is an extension of $f$.

Before we prove Lemma D.4 we state and prove an auxiliary result. The estimation will be very rough and can for sure be improved. This is, however, not necessary for our purpose.

**Lemma D.6.** Under the conditions of Lemma D.4 and with the notation from Remark D.3 we have for all $m, \tilde{m} \in \{0, \ldots, N\}^d$ the estimate

$$\|\tilde{f} - p_{f,m}\|_{W^{k,p}(\Omega_{m,N})} \leq C N^{d/p} \|f\|_{\mathbb{W}^{n,p}((0,1)^d)},$$

for a constant $C = C(n, d, p)$.  

27
Proof. We start with bounding the norm of the polynomial by using the triangle inequality. There holds
\[
\|p_{f,m}\|_{W^{k,p}(\Omega_m,N)} = \left\| \sum_{|\alpha|\leq n-1} c_{f,m,\alpha} x^\alpha \right\|_{W^{k,p}(\Omega_m,N;dx)} \leq \sum_{|\alpha|\leq n-1} |c_{f,m,\alpha}| \cdot \|x^\alpha\|_{W^{k,p}(\Omega_m,N;dx)}.
\]
Using that \(\Omega_m,N \subset B_{2,\|\cdot\|_\infty}\), we get
\[
\|x^\alpha\|_{W^{k,p}(\Omega_m,N;dx)} \leq (n-1)^2 2^{|\alpha|} \leq (n-1)^2 2^{n-1}.
\]
If we now combine Remark \[\ref{app:remark:extension-properties}\] with Equation \[\ref{eq:bramble-hilbert-estimate}\], we get
\[
\sum_{|\alpha|\leq n-1} |c_{f,m,\alpha}| \|x^\alpha\|_{W^{k,p}(\Omega_m,N;dx)} \leq C(n-1)^2 2^{n-1} \sum_{|\alpha|\leq n-1} N^{d/p} \|\tilde{f}\|_{W^{n,p}(\Omega_m,N)} \leq C N^{d/p} \|f\|_{W^{n,p}(0,1)^d)},
\]
where we have additionally used Remark \[\ref{app:remark:decay-of-bump-functions}\] in the last step. Finally, we can estimate, by the triangle inequality
\[
\|\tilde{f} - p_{f,m}\|_{W^{k,p}(\Omega_m,N)} \leq C \|f\|_{W^{k,p}(0,1)^d)} + C N^{d/p} \|f\|_{W^{n,p}(0,1)^d)} \leq C N^{d/p} \|f\|_{W^{n,p}(0,1)^d)},
\]
where we again used the extension property from Equation \[\ref{equation:extension-property}\] for the first step. \(\square\)

Proof of Lemma \[\ref{lemma:local-estimates-based-on-exponential-decay}\]. We use approximation properties of the polynomials from the Bramble-Hilbert Lemma \[\ref{lemma:bramble-hilbert-estimate}\] to derive local estimates and then combine them using an approximate partition of unity to obtain a global estimate. In order to use this strategy also near the boundary, we make use of an extension operator (see Remark \[\ref{app:remark:extension-properties}\]).

**Step 1 (Local estimates based on Bramble-Hilbert):** For each \(m \in \{0, \ldots, N\}^d\) we set
\[
\Omega_{m,N} := B_{\frac{1}{N}} \left( \tfrac{m}{N} \right)
\]
and denote by \(p_{m} = p_{f,m}\) the polynomial from Lemma \[\ref{lemma:bramble-hilbert-estimate}\] so that we can directly state the estimate
\[
\|\tilde{f} - p_{m}\|_{W^{k,p}(\Omega_{m,N})} \leq C \left( \frac{1}{N} \right)^{n-k} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})}.
\]
Furthermore, similarly to \[\text{[21] Lemma C.4}\], we obtain the estimate
\[
\|\phi_{m}^\alpha(\tilde{f} - p_{m})\|_{W^{k,p}(\Omega_{m,N})} \leq C \sum_{k=0}^{k} \|\phi_{m}^\alpha\|_{W^{n,\infty}(\Omega_{m,N})} \|\tilde{f} - p_{m}\|_{W^{k-p,\infty}(\Omega_{m,N})} 
\leq C \sum_{k=0}^{k} N^{n+k+\mu_{k=2}} \left( \frac{1}{N} \right)^{n-k+\mu_{k=2}} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})}
\leq C \left( \frac{1}{N} \right)^{n-k+\mu_{k=2}} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})},
\]
where we used the product rule from Lemma \[\ref{app:remark:product-rule}\] for the first step and the estimate of the derivative of \(\phi_{m}^\alpha\) from Proposition \[\ref{app:proposition:derivative-of-bump-functions}\] together with the Bramble-Hilbert estimate in Equation \[\ref{equation:bramble-hilbert-estimate}\] for the second step.

**Step 2 (Local estimates based on exponential decay):** Since our localizing bump functions \(\phi_{m}^\alpha\) do not have compact support on \(\Omega_{m,N}\), we also need to bound the influence of \(\phi_{m}^\alpha(\tilde{f} - p_{m})\) on patches \(\Omega_{\tilde{m},N}\) with \(\tilde{m} \neq m\) where we cannot use the Bramble-Hilbert lemma. Here, we will make use of the exponential decay of the bump functions \(\phi_{m}^\alpha\) outside a certain ball centered at \(m/N\) (see Proposition \[\ref{app:proposition:derivative-of-bump-functions}\][iii]).
This is possible for the case where $\Omega_{\tilde{m},N}$ is not a neighboring patch of $\Omega_{m,N}$, i.e. $\|\tilde{m} - m\|_{\infty} > 1$. Then 
$\Omega_{\tilde{m},N} \subset \Omega_{m,N}^{\text{exp}} := \{ x \in \mathbb{R}^d : \| x - \frac{m}{N} \|_{\infty} \geq \frac{3}{5N} \}$ and we have (by using Lemma B.5 in the first step), that

$$\left\| \phi_m^s (\tilde{f} - p_m) \right\|_{W^{k,p}(\Omega_{\tilde{m},N})} \leq C \|\phi_m^s\|_{W^{k,\infty}(\Omega_{\tilde{m},N})} \|\tilde{f} - p_m\|_{W^{k,p}(\Omega_{\tilde{m},N})}$$

(Proposition B.3) with $\Omega_{\tilde{m},N} \subset \Omega_{m,N}^{\text{exp}} \leq C N^k N^{\mu(k-1)} e^{-\frac{1}{4} N^k} \|\tilde{f} - p_m\|_{W^{k,p}(\Omega_{\tilde{m},N})}$ (Lemma B.6).

Then, by Proposition A.1 there exists $N_1 = N_1(\mu, d, p) \in \mathbb{N}$ such that $e^{-\frac{1}{4} N^k} \leq C \gamma(N)^{-1} \cdot (N + 1)^{-d - d/p} \cdot N^{-(n-k-\mu(k-2))}$ for all $N \geq N_1$. Consequently, we have

$$\|\phi_m^s (\tilde{f} - p_m)\|_{W^{k,p}(\Omega_{\tilde{m},N})} \leq C (N + 1)^{-d - d/p} N^{-(n-k-\mu(k-2))} \|f\|_{W^{k,p}(0,1)^d},$$

for all $N \geq N_1$.

**Step 3 (Mixed local estimates):** If $\Omega_{\tilde{m},N}$ is a neighboring patch of $\Omega_{m,N}$, i.e. $\|\tilde{m} - m\|_{\infty} = 1$, then we have to split into a region $\Omega_{m,N} \cap \Omega_{m,N}^{\text{exp}}$ where we have exponential decay of the bump function and a region $\Omega_{m,N} \setminus \Omega_{m,N}^{\text{exp}} \subset \Omega_{m,N}$ where we can make use of the Bramble-Hilbert Lemma. In detail, we have

$$\|\phi_m^s (\tilde{f} - p_m)\|_{W^{k,p}(\Omega_{\tilde{m},N})} \leq \|\phi_m^s (\tilde{f} - p_m)\|_{W^{k,p}(\Omega_{m,N} \cap \Omega_{m,N}^{\text{exp}})} + \|\phi_m^s (\tilde{f} - p_m)\|_{W^{k,p}(\Omega_{m,N} \setminus \Omega_{m,N}^{\text{exp}})} \leq C N^{-(n-k-\mu(k-2))} \left( \|f\|_{W^{k,p}(\Omega_{m,N})} + (N + 1)^{-d - d/p} \|f\|_{W^{k,p}(0,1)^d} \right),$$

for all $N \geq N_1$. Here we used Step 1 to bound the first term of the sum and Step 2 for the second.

**Step 4 (Global estimate):** Using that $\tilde{f}$ is an extension of $f$ on $(0,1)^d$ we can write

$$\left\| \tilde{f} - \sum_{m \in \{0,\ldots,N\}^d} \phi_m^s \right\|_{W^{k,p}(0,1)^d} \leq \left\| \tilde{f} - \sum_{m \in \{0,\ldots,N\}^d} \phi_m^s \tilde{f} \right\|_{W^{k,p}(0,1)^d} + \left\| \sum_{m \in \{0,\ldots,N\}^d} \phi_m^s (\tilde{f} - p_m) \right\|_{W^{k,p}(0,1)^d}$$

(D.8)

where the last step follows from $(0,1)^d \subset \bigcup_{m \in \{0,\ldots,N\}^d} \Omega_{\tilde{m},N}$.

**Step 4a (Partition of Unity):** For the first term in Equation (D.8), we get by the product rule from Lemma B.5

$$\left\| \tilde{f} (1_{(0,1)^d} - \sum_{m \in \{0,\ldots,N\}^d} \phi_m^s) \right\|_{W^{k,p}(0,1)^d} \leq C \left\| \tilde{f} \right\|_{W^{k,p}(0,1)^d} \left\| 1_{(0,1)^d} - \sum_{m \in \{0,\ldots,N\}^d} \phi_m^s \right\|_{W^{k,\infty}(0,1)^d}$$

(Property 2 from Proposition A.1) \leq C \|f\|_{W^{k,p}(0,1)^d} \cdot N^{-(n-k-\mu(k-2))},

(D.9)

for all $N \geq N_2 = N_2(\mu)$. For the second inequality we used the same trick as in Step 2 which is based on Proposition A.1.
Step 4b (Patches): Considering the second term from Equation (D.8), we obtain for each $m \in \{0, \ldots, N\}^d$

$$
\left\| \sum_{m \in \{0, \ldots, N\}^d} \phi_m^s(\tilde{f} - p_m) \right\|_{W^{k,p}(\Omega_{\tilde{m},N})} \leq \left\| \phi_m^s(\tilde{f} - p_{\tilde{m}}) \right\|_{W^{k,p}(\Omega_{\tilde{m},N})} + \sum_{m \in \{0, \ldots, N\}^d, \|m-\tilde{m}\|_\infty = 1} \left\| \phi_m^s(\tilde{f} - p_m) \right\|_{W^{k,p}(\Omega_{\tilde{m},N})} + \sum_{m \in \{0, \ldots, N\}^d, \|m-\tilde{m}\|_\infty > 1} \left\| \phi_m^s(\tilde{f} - p_m) \right\|_{W^{k,p}(\Omega_{\tilde{m},N})}.
$$

(\ast)

(\ast\ast)

(\ast\ast\ast)

(D.10)

The term (\ast) can be handled with Step 1, the term (\ast\ast) with Step 3 and the third one (\ast\ast\ast) with Step 2.

Since (\ast\ast) and (\ast\ast\ast) require a similar strategy we only demonstrate it for the third term. We get from Step 2

$$
\sum_{m \in \{0, \ldots, N\}^d, \|m-\tilde{m}\|_\infty > 1} \left\| \phi_m^s(\tilde{f} - p_m) \right\|_{W^{k,p}(\Omega_{\tilde{m},N})} \leq C N^{-(n-k-p(k=2))(N+1)^{-d/p}} \sum_{m \in \{0, \ldots, N\}^d, \|m-\tilde{m}\|_\infty > 1} \|f\|_{W^{n,p}(\Omega_{\tilde{m},N})^d}.
$$

D.10

We can now bound the sum from Equation (D.10) for each $\tilde{m} \in \{0, \ldots, N\}^d$ by

$$
\left\| \sum_{m \in \{0, \ldots, N\}^d} \phi_m^s(\tilde{f} - p_m) \right\|_{W^{k,p}(\Omega_{\tilde{m},N})} \leq C N^{-(n-k-p(k=2))} \left( 2(N+1)^{-d/p} \|f\|_{W^{n,p}(\Omega_{\tilde{m},N})^d} + \sum_{m \in \{0, \ldots, N\}^d, \|m-\tilde{m}\|_\infty \leq 1} \|\tilde{f}\|_{W^{n,p}(\Omega_{\tilde{m},N})} \right).
$$

(D.11)

Consequently, we get

$$
\sum_{\tilde{m} \in \{0, \ldots, N\}^d} \left\| \sum_{m \in \{0, \ldots, N\}^d} \phi_m^s(\tilde{f} - p_m)^p \right\|_{W^{k,p}(\Omega_{\tilde{m},N})}^p \leq C \left( \sum_{\tilde{m} \in \{0, \ldots, N\}^d} \left( 2(N+1)^{-d/p} \|f\|_{W^{n,p}(\Omega_{\tilde{m},N})^d} + \sum_{m \in \{0, \ldots, N\}^d, \|m-\tilde{m}\|_\infty \leq 1} \|\tilde{f}\|_{W^{n,p}(\Omega_{\tilde{m},N})} \right) \right)^p.
$$

$$
\leq C \left( \sum_{\tilde{m} \in \{0, \ldots, N\}^d} 2^p(N+1)^{-d/p} \|f\|_{W^{n,p}(\Omega_{\tilde{m},N})^d}^p + \sum_{\tilde{m} \in \{0, \ldots, N\}^d} \sum_{m \in \{0, \ldots, N\}^d, \|m-\tilde{m}\|_\infty \leq 1} \|\tilde{f}\|_{W^{n,p}(\Omega_{\tilde{m},N})}^p \right)
$$

$$
\leq C \left( \sum_{\tilde{m} \in \{0, \ldots, N\}^d} 2^p(N+1)^{-d/p} \|f\|_{W^{n,p}(\Omega_{\tilde{m},N})^d}^p + \sum_{\tilde{m} \in \{0, \ldots, N\}^d} \sum_{m \in \{0, \ldots, N\}^d, \|m-\tilde{m}\|_\infty \leq 1} \|\tilde{f}\|_{W^{n,p}(\Omega_{\tilde{m},N})}^p \right)^{p/q}
$$

30
where the first step follows from plugging in Equation (D.11), the second step follows from Hölder's inequality (with \( q := 1 - 1/p \)) and the last step follows from the definition of \( \Omega_{\tilde{m}, N} \). Moreover, we use in the second and the last step the fact that the number of neighbors of a particular patch is bounded by \( 3^d - 1 \). To conclude Step 4b we note that from the definition of \( \Omega_{\tilde{m}, N} \) it follows that there exist \( 2^d \) disjoint subsets \( \mathcal{M}_i \subset \{0, \ldots, N\}^d \) such that \( \bigcup_{i=1, \ldots, 2^d} \mathcal{M}_i = \{0, \ldots, N\}^d \) and \( \Omega_{m_1, N} \cap \Omega_{m_2, N} = \emptyset \) for all \( m_1, m_2 \in \mathcal{M}_i \) with \( m_1 \neq m_2 \) and all \( i = 1, \ldots, 2^d \). From this we get

\[
\sum_{\tilde{m} \in \{0, \ldots, N\}^d} \| f \|^p_{W^{n,p}((0,1)^d)} = \sum_{i=1, \ldots, 2^d} \sum_{\tilde{m} \in \mathcal{M}_i} \| \tilde{f}\|^p_{W^{n,p}((\Omega_{\tilde{m}, N})^d)} \leq 2^d \| \tilde{f}\|^p_{W^{n,p}((\bigcup_{m \in \{0, \ldots, N\}} \Omega_{\tilde{m}, N})^d)} \leq C \| f \|^p_{W^{n,p}((0,1)^d)}.
\]

(D.13)

and, finally, together with Remark B.3

\[
\sum_{\tilde{m} \in \{0, \ldots, N\}^d} \| f \|^p_{W^{n,p}((\Omega_{\tilde{m}, N})^d)} \leq 2^d \| f \|^p_{W^{n,p}((\bigcup_{m \in \{0, \ldots, N\}} \Omega_{\tilde{m}, N})^d)} \leq C \| f \|^p_{W^{n,p}((0,1)^d)}.
\]

(D.14)

Step 4c (Wrap it all up): Combining Equation (D.12) with Equation (D.14) from Step 4b and inserting it into Equation (D.8) together with the estimate in Equation (D.9) from Step 4a finally yields

\[
\| f - f_N \|^p_{W^{n,p}((0,1)^d)} \leq C N^{-(n-k-\mu(k+2))} \| f \|^p_{W^{n,p}((0,1)^d)}.
\]

for all \( N \geq \bar{N} := \max\{N_1, N_2\} \) and a constant \( C = C(n, d, p) > 0 \). The linearity of \( T_k^m \), \( k \in \{0,1,2\} \) is a consequence of the linearity of the averaged Taylor polynomial (cf. [21] Remark B.8). 

\[
\]

D.3 Approximation of Localized Polynomials by Neural Network Realizations

The goal of this section is now to approximate the localized polynomials from Appendix B.2 by neural networks.

We start by approximating monomials on \( \mathbb{R} \) by two-layered neural networks with (rather general) activation functions that have non-vanishing Taylor coefficients up to order \( n \in \mathbb{N} \). The construction is mainly based on finite backward differences in combination with ideas from [41]. We note that the statement can be simplified in case of the ELU-activation function. However, since we believe this result to be an interesting contribution in itself, we have formulated in this generality.

Proposition D.7. Let \( \varphi : \mathbb{R} \to \mathbb{R} \) be a function. Assume, that for some \( n \in \mathbb{N} \) there exists \( x_0 \in \mathbb{R} \) such that \( \varphi \) is \( n+1 \) times continuously differentiable in some open neighborhood \( U \) around \( x_0 \) and \( \varphi^{(m)}(x_0) \neq 0 \) for some \( m \in \{1, \ldots, n\} \). Then, for every \( \varepsilon \in (0,1) \), and every \( B > 0 \) there exists a constant \( C = C(B, \varphi, m, n) > 0 \) as well as a neural network \( \Phi_\varepsilon^m \) with \( R_\varphi(\Phi_\varepsilon^m) \subset C^{n+1}([-B,B]) \) and the following properties:

(i) \( \| R_\varphi(\Phi_\varepsilon^m) - x^m \|_{C^n([-B,B])} \leq \varepsilon; \)

(ii) \( | R_\varphi(\Phi_\varepsilon^m)|_{W^{k,\infty}([-B,B])} \leq C \frac{m!}{(m-k)!} B^{m-k} \) for \( k = 0, \ldots, m; \)

(iii) \( L(\Phi_\varepsilon^m) = 2 \), as well as \( M(\Phi_\varepsilon^m) \leq 3(m+1); \)

(iv) \( \| \Phi_\varepsilon^m \|_{\max} \leq C \varepsilon^{-m}. \)
Proof. Choose $C_0 > 1$ so that $[x_0 - \frac{nB}{C_0}, x_0 + \frac{nB}{C_0}] \subset U$. Moreover, let $\delta \geq C_0$ be arbitrary. Define the function

$$g_\delta^m : \mathbb{R} \to \mathbb{R}, \ x \mapsto \frac{\delta^m}{g^{(m)}(x_0)} \sum_{j=0}^{m} (-1)^j \binom{m}{j} \cdot g \left( x_0 - j \frac{x}{\delta} \right).$$

Then $g_\delta^m|_{[-B,B]} \in C^{m+1}([-B,B])$. Using the Taylor expansion and the following identity from [26]

$$\sum_{j=0}^{m} (-1)^j \binom{m}{j} = \begin{cases} 0, & \text{if } 1 \leq k < m, \\ (-1)^m m!, & \text{if } k = m, \end{cases}$$

it can easily be shown that $g_\delta^m(x) \approx x^m$ for $\delta > 0$ sufficiently large. In detail, we have by Taylor’s Theorem (where $\xi_j$ is between $x_0$ and $x_0 - j \frac{\delta}{m}$ for $j = 1, \ldots, m$) that

$$\sum_{j=0}^{m} (-1)^j \binom{m}{j} \cdot g \left( x_0 - j \frac{x}{\delta} \right) = g(x_0) + \sum_{k=0}^{m} \frac{(-x)}{k!} \frac{g^{(k)}(x_0)}{k!} \sum_{j=1}^{m} (-1)^j \binom{m}{j} j^k + \sum_{j=1}^{m} (-1)^j \binom{m}{j} \frac{g^{(m+1)}(\xi_j)}{(m+1)!} \left( \frac{-\frac{m+1}{x}x}{\delta} \right)^{m+1}$$

$$= g(x_0) \sum_{j=0}^{m} (-1)^j \binom{m}{j} + \sum_{k=1}^{m} \frac{(-x)}{k!} \frac{g^{(k)}(x_0)}{k!} \sum_{j=1}^{m} (-1)^j \binom{m}{j} j^k + r_\delta^m(x)$$

$$= \left( \frac{x}{\delta} \right)^m g^{(m)}(x_0) + r_\delta^m(x).$$

Hence, for every $k = 0, \ldots, n$ and every $x \in [-B,B]$, we have

$$\left| (g_\delta^m)^{(k)}(x) - (x^m)^{(k)} \right| \leq 2^n \left\| g^{(m+1)} \right\|_{C^{m+1}(U)} \left( \frac{\delta}{\min_{i=0,\ldots,n,|g^{(i)}(x_0)|}} \right)^{n+1} \left( \frac{\min_{i=0,\ldots,n,|g^{(i)}(x_0)|}}{\delta} \right) \leq n! \max\{B,1\}^{n+1}.$$

This implies that there exists some $C \geq \max\{C_0, C'(B,n, \delta)\}$ such that for every $\varepsilon \in (0,1)$ and the neural network $\Phi_\delta^m := ((A_1, b_1), (A_2, b_2))$ with

$$A_1 := \left( 0, -\frac{\varepsilon}{C}, \ldots, -\frac{m\varepsilon}{C} \right)^T \in \mathbb{R}^{m+1},$$

$$b_1 := (x_0, \ldots, x_0) \in \mathbb{R}^{m+1},$$

$$A_2 := \left( \frac{C^m}{\varepsilon^{m+1} g^{(m)}(x_0)} \right) \left( (-1)^0 \binom{m}{0}, (-1)^1 \binom{m}{1}, \ldots, (-1)^m \binom{m}{m} \right) \in \mathbb{R}^{1,m+1},$$

$$b_2 := 0 \in \mathbb{R},$$

32
fulfills
\[ \| R_\varrho (\Phi^m_\varepsilon) - x^m \|_{C^\infty([-B,B])} \leq \varepsilon. \]

Moreover, \( L(\Phi^m_\varepsilon) = 2 \) and \( M(\Phi^m_\varepsilon) \leq 3(m + 1) \).

Additionally, for every \( k = 0, \ldots, m \) and for every \( x \in [-B,B] \) we have
\[
\left\| (R_\varrho(\Phi^m_\varepsilon))^{(k)}(x) \right\| \leq \left\| (R_\varrho(\Phi^m_\varepsilon))^{(k)} - (x^m)^{(k)} \right\|_{C^\infty([-B,B])} + \left\| (x^m)^{(k)} \right\|_{C^\infty([-B,B])} \leq \varepsilon + \frac{n!}{(n-k)!} \max\{1,B\}^{m-k}.
\]

Finally, for all \( k = m + 1, \ldots, n \) we have that
\[
\left\| (R_\varrho(\Phi^m_\varepsilon))^{(k)}(x) \right\| \leq \left\| (R_\varrho(\Phi^m_\varepsilon))^{(k)} - (x^m)^{(k)} \right\|_{C^\infty([-B,B])} + \left\| (x^m)^{(k)} \right\| \leq \varepsilon + 0 = \varepsilon.
\]

This completes the proof. \( \square \)

**Remark D.8.** The assumptions of Proposition D.7 apply to the activation function ELU\(_\alpha\) for all \( \alpha \in \mathbb{R} \setminus \{0\} \) and for an arbitrary \( n \in \mathbb{N} \). One only needs to choose an arbitrary \( x_0 < 0 \) for which we have that ELU\(_{\alpha}^{(k)}(x_0) = \alpha x_0 \neq 0 \) for all \( k \in \mathbb{N}_{\geq 1} \).

With the help of Proposition D.7, we are now in a position to construct neural networks which implement an approximate multiplication.

**Proposition D.9.** Let \( \varrho : \mathbb{R} \to \mathbb{R} \) and \( x_0 \in \mathbb{R} \) such that \( \varrho \) is three times continuously differentiable in a neighborhood of \( x_0 \) and \( \varrho''(x_0) \neq 0 \). Let \( B > 0 \), then there exists a constant \( C = C(B, \varrho) > 0 \) such that for every \( \varepsilon \in (0,1/2) \), there is a neural network \( \widetilde{x} \) with two-dimensional input and one-dimensional output that satisfies the following properties:

\( i \) \( \| R_\varrho(\widetilde{x})(x,y) - xy \|_{W^{2,\infty}((-B,B)^2;dx,dy)} \leq \varepsilon; \)

\( ii \) \( |R_\varrho(\widetilde{x})|_{W^{k,\infty}((-B,B)^2)} \leq C \) for \( k \in \{0,1,2\} \);

\( iii \) \( L(\widetilde{x}) = 2 \) and \( M(\widetilde{x}) \leq C; \)

\( iv \) \( \| \widetilde{x} \|_{\max} \leq C\varepsilon^{-2}. \)

**Proof.** Let \( C \) be the constant from Corollary B.6 and set \( \bar{\varepsilon} := \varepsilon/2C \). Proposition D.7 yields that there exists a neural network \( \Phi^2_{\bar{\varepsilon}} \) with 2 layers and at most 9 nonzero weights such that for all \( k \in \{0,1,2\} \) we have
\[
\left| R_\varrho(\Phi^2_{\bar{\varepsilon}}) - x^2 \right|_{W^{k,\infty}([-2B,2B];dx)} \leq \bar{\varepsilon}.
\]

As in [51], we make use of the polarization identity
\[ xy = \frac{1}{4} \left( (x+y)^2 - (x-y)^2 \right) \quad \text{for } x, y \in \mathbb{R}. \]

In detail, we define the neural network
\[
\widetilde{x}_{\bar{\varepsilon}} := \left( \begin{pmatrix} 1/4, -1/4 \end{pmatrix}, 0 \right) \bullet \Phi^2_{\bar{\varepsilon}} \bullet \left( \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, 0 \right),
\]

which fulfills for all \((x,y) \in \mathbb{R}^2\) that
\[
R_\varrho(\widetilde{x}_{\bar{\varepsilon}})(x,y) = \frac{1}{4} \left( R_\varrho(\Phi^2_{\bar{\varepsilon}})(x+y) - R_\varrho(\Phi^2_{\bar{\varepsilon}})(x-y) \right).
\]
Now, setting $f : [-2B, 2B] \rightarrow \mathbb{R}, x \mapsto x^2$ as well as

$$u : [-B, B]^2 \rightarrow [-2B, 2B], \ (x, y) \mapsto x + y \quad \text{and} \quad v : [-B, B]^2 \rightarrow [-2B, 2B], \ (x, y) \mapsto x - y,$$

we see that for all $(x, y) \in [-B, B]^2$ there holds $xy = 1/4 (f \circ u(x, y) - f \circ v(x, y))$. We estimate

$$\| R_\varepsilon(\tilde{x}_\varepsilon)(x, y) - xy \|_{W^{k,\infty}([-B, B]^2;dxdy)} = \frac{1}{4} \| R_\varepsilon(\Phi_\varepsilon^2) \circ u - R_\varepsilon(\Phi_\varepsilon^2) \circ v - (f \circ u - f \circ v) \|_{W^{k,\infty}([-B, B]^2)}$$

and directly see for $k = 0$ that

$$\| R_\varepsilon(\tilde{x}_\varepsilon)(x, y) - xy \|_{W^{0,\infty}([-B, B]^2;dxdy)} \leq \frac{2}{4} \| R_\varepsilon(\Phi_\varepsilon^2) - x^2 \|_{L^{\infty}([-B, B]^2;dx)} \leq \frac{1}{2} \varepsilon \leq \varepsilon.$$

For the case $k = 0$ and $k = 1$ we first note that

$$|u|_{W^{0,\infty}([-B, B]^2)} = |v|_{W^{0,\infty}([-B, B]^2)} = 2B,$$

$$|u|_{W^{1,\infty}([-B, B]^2)} = |v|_{W^{1,\infty}([-B, B]^2)} = 1,$$

$$|u|_{W^{2,\infty}([-B, B]^2)} = |v|_{W^{2,\infty}([-B, B]^2)} = 0.$$

The composition rule from Corollary B.6 then yields that

$$\| R_\varepsilon(\tilde{x}_\varepsilon)(x, y) - xy \|_{W^{1,\infty}([-B, B]^2;dxdy)} \leq 2C \| R_\varepsilon(\Phi_\varepsilon^2) - x^2 \|_{W^{1,\infty}([-B, B]^2;dx)} \| u \|_{W^{1,\infty}([-B, B]^2)} \leq 2C \varepsilon = \varepsilon.$$

In the same way, we get

$$\| R_\varepsilon(\tilde{x}_\varepsilon)(x, y) - xy \|_{W^{2,\infty}([-B, B]^2;dxdy)} \leq 2C \| R_\varepsilon(\Phi_\varepsilon^2) - f \|_{W^{2,\infty}([-B, B]^2)} \| u \|_{W^{1,\infty}([-B, B]^2)}$$

and, thus, claim (i) is shown.

Finally, we have for $k \in \{0, 1, 2\}$

$$\| R_\varepsilon(\tilde{x}_\varepsilon) \|_{W^{k,\infty}([-B, B]^2)} \leq \| R_\varepsilon(\tilde{x}_\varepsilon) - xy \|_{W^{k,\infty}([-B, B]^2;dxdy)} + |xy|_{W^{k,\infty}([-B, B]^2;dxdy)} \leq C_1,$$

for a constant $C_1 = C_1(B) > 0$, yielding (ii). Claim (iii), (iv) immediately follow from the construction of $\tilde{x}_\varepsilon$ in combination with Proposition D.7.

Proposition D.9 is the foundation for the following results which implements a neural network that approximates the multiplication of multiple inputs:

**Lemma D.10.** Let $d, m, K \in \mathbb{N}$ and $N \geq 1$, $\mu, c > 0$ be arbitrary, and let $\varrho = \text{ELU}_1$. Then there are constants $C(d, m, c) > 0$ such that the following holds:

For any $\varepsilon \in (0, 1/2)$, and any neural network $\Phi$ with $d$-dimensional input and $m$-dimensional output and with number of layers and nonzero weights all bounded by $K$, such that

$$\| [R_\varepsilon(\Phi)]_l \|_{W^{k,\infty}((0,1)^d)} \leq cN^{k+\mu(k=2)} \quad \text{and} \quad 0 \leq [R_\varepsilon(\Phi)]_l(x) \leq 1,$$

for $k \in \{0, 1, 2\}$, $l = 1, \ldots, m$ and $x \in (0, 1)^d$ there exists a neural network $\Psi_{\varepsilon, \Phi}$ with $d$-dimensional input and one-dimensional output, and with
(i) number of layers and nonzero weights all bounded by $CK$;

(ii) $\|R_\varrho(\Psi_{\varepsilon,\Phi}) - \prod_{l=1}^{m}[R_\varrho(\Phi)]_l\|_{W^{k,\infty}((0,1)^d)} \leq CN^{k+\mu(k+2)}\varepsilon$;

(iii) $\|R_\varrho(\Psi_{\varepsilon,\Phi})\|_{W^{k,\infty}((0,1)^d)} \leq CN^{k+\mu(k+2)}$;

(iv) $\|\Psi_{\varepsilon,\Phi}\|_\text{max} \leq C\max\{\|\Phi\|_{\text{max}}, \varepsilon^{-2}\}$.

**Proof.** We show by induction over $m \in \mathbb{N}$ that the statement holds. To make the induction argument easier we will additionally show that the network $\Psi_{\varepsilon,\Phi}$ can be chosen such that the first $L(\Phi) - 1$ layers of $\Psi_{\varepsilon,\Phi}$ and $\Phi$ coincide.

If $m = 1$, then we can choose $\Psi_{\varepsilon,\Phi} = \Phi$ for any $\varepsilon \in (0, 1/2)$ and the claim holds.

Now, assume that the claim holds for an arbitrary, but fixed $m \in \mathbb{N}$. We show that it also holds for $m + 1$. For this, let $\varepsilon \in (0, 1/2)$ and let $\Phi = ((A_1, b_1), (A_2, b_2), \ldots, (A_L, b_L))$ be a neural network with $d$-dimensional input and $(m+1)$-dimensional output and with number of layers, and nonzero weights all bounded by $K$, where each $A_l$ is an $N_l \times N_{l-1}$ matrix, and $b_l \in \mathbb{R}^{N_l}$ for $l = 1, \ldots, L$.

**Step 1 (Invoking induction hypothesis):** We denote by $\Phi_m$ the neural network with $d$-dimensional input and $m$-dimensional output which results from $\Phi$ by removing the last output neuron and corresponding weights. In detail, we write

$$A_l = \begin{bmatrix} A^{(1,m)}_l \\ a^{(m+1)}_l \end{bmatrix} \quad \text{and} \quad b_l = \begin{bmatrix} b^{(1,m)}_l \\ b^{(m+1)}_l \end{bmatrix},$$

where $A^{(1,m)}_l$ is a $m \times N_{l-1}$ matrix and $a^{(m+1)}_l$ is a $1 \times N_{l-1}$ vector, and $b^{(1,m)}_l \in \mathbb{R}^m$ and $b^{(m+1)}_l \in \mathbb{R}^1$.

Now we set

$$\Phi_m := ((A_1, b_1), (A_2, b_2), \ldots, (A_{L-1}, b_{L-1}), (A^{(1,m)}_L, b^{(1,m)}_L)).$$

Using the induction hypothesis we get that there is a neural network

$$\Psi_{\varepsilon,\Phi_m} = ((A'_1, b'_1), (A'_2, b'_2), \ldots, (A'_{L'}, b'_{L'}))$$

with $d$-dimensional input and one-dimensional output, and at most $KC$ layers and nonzero weights such that

$$\left\| R_\varrho(\Psi_{\varepsilon,\Phi_m}) - \prod_{l=1}^{m}[R_\varrho(\Phi_m)]_l \right\|_{W^{k,\infty}((0,1)^d)} \leq CN^{k+\mu(k+2)}\varepsilon,$$

and $\|R_\varrho(\Psi_{\varepsilon,\Phi_m})\|_{W^{k,\infty}((0,1)^d)} \leq CN^{k+\mu(k+2)}$. Moreover, we have that $\|\Phi_m\|_\text{max} \leq \|\Phi\|_\text{max}$, so that there we can estimate $\|\Psi_{\varepsilon,\Phi_m}\|_\text{max} \leq C\max\{\|\Phi\|_\text{max}, \varepsilon^{-2}\}$. Furthermore, we can assume that the first $L - 1$ layers of $\Psi_{\varepsilon,\Phi_m}$ and $\Phi_m$ coincide and, thus, also the first $L - 1$ layers of $\Psi_{\varepsilon,\Phi_m}$ and $\Phi$, i.e. $A_l = A'_l$ for $l = 1, \ldots, L - 1$.

**Step 2 (Combining $\Psi_{\varepsilon,\Phi_m}$ and $[R_\varrho(\Phi)]_{m+1}$):** Now, we construct a network $\Psi_{\varepsilon,\Phi}$ where the first $L - 1$ layers of $\Psi_{\varepsilon,\Phi}$ and $\Psi_{\varepsilon,\Phi_m}$ (and, thus, also of $\Phi$) coincide (by definition), and $\Psi_{\varepsilon,\Phi}$ has two-dimensional output with $[R_\varrho(\Psi_{\varepsilon,\Phi})]_1 = R_\varrho(\Psi_{\varepsilon,\Phi_m})$ and $[R_\varrho(\Psi_{\varepsilon,\Phi})]_{L'} = [R_\varrho(\Phi)]_{m+1}$. For this, we add the formerly removed neuron with corresponding weights back to the $L$-th layer of $\Psi_{\varepsilon,\Phi_m}$ and recall that by Assumption (D.16) we have $[R_\varrho(\Phi)]_{m+1}(x) \geq 0$ for $x \in (0, 1)^d$. Thus, we can thus use the trick

$$\varrho(\ldots, \varrho(y) \ldots) = y, \quad \text{for} \quad y \in [0, \infty),$$

to pass the output through to the last last layer. For this, we set $e^n := [0, \ldots, 0, 1] \in \mathbb{R}^n$ and define

$$\Psi_{\varepsilon,\Phi} := \left( (A'_1, b'_1), \ldots, (A'_L, b'_L), \left( \begin{bmatrix} A^{(1,m+1)}_L \\ a^{(m+1)}_L \end{bmatrix}, \left( \begin{bmatrix} b^{(1,m+1)}_L \\ b^{(m+1)}_L \end{bmatrix} \right) \right), \ldots, \left( \begin{bmatrix} A^{(1,m)}_{L-1} \\ a^{(m+1)}_{L-1} \end{bmatrix}, \left( \begin{bmatrix} b^{(1,m)}_{L-1} \\ b^{(m+1)}_{L-1} \end{bmatrix} \right) \right) \right).$$
Counting the number of nonzero weights of \( \tilde{\Psi}_{\varepsilon, \Phi} \) we get

\[
M(\tilde{\Psi}_{\varepsilon, \Phi}) \leq M(\Psi_{\varepsilon, \Phi}) + M(\Phi) + (L' - (L + 1)) \leq CK + K + CK \leq CK,
\]

where we used in the second step the induction hypothesis twice together with the assumption on \( \Phi \). Similarly, we get the statement for \( N(\tilde{\Psi}_{\varepsilon, \Phi}) \) and \( L(\tilde{\Psi}_{\varepsilon, \Phi}) \). Furthermore, \( \| \tilde{\Psi}_{\varepsilon, \Phi} \|_{\text{max}} \leq C \max\{\|\Phi\|_{\text{max}}, \varepsilon^{-2}\} \).

Next, we want to apply the approximate multiplication network from Proposition \[ \text{D.9} \] to the output of \( \tilde{\Psi}_{\varepsilon, \Phi} \). For this, we need to find a bounding box for the range of \( R_{\varepsilon}(\tilde{\Psi}_{\varepsilon, \Phi}) \). We have

\[
\| R_{\varepsilon}(\Psi_{\varepsilon, \Phi}) \|_{L^\infty((0,1)^d)} \leq C \quad \text{and} \quad \| R_{\varepsilon}(\Phi) \|_{m+1} \|_{L^\infty((0,1)^d)} \leq 1,
\]

and get for \( B := \max\{C, 1\} \) that \( \text{Range} R_{\varepsilon}(\tilde{\Psi}_{\varepsilon, \Phi}) \subset [-B, B]^2 \). Now, we denote by \( \tilde{\times} \) the network from Proposition \[ \text{D.9} \] with \( B = B \) and accuracy \( \varepsilon \) and define

\[
\Psi_{\varepsilon, \Phi} := \tilde{\times} \circ \tilde{\Psi}_{\varepsilon, \Phi}.
\]

**Step 3 (\( \Psi_{\varepsilon, \Phi} \) fulfills induction hypothesis for \( m+1 \)):** ad (i): Clearly, \( \Psi_{\varepsilon, \Phi} \) has \( d \)-dimensional input, one-dimensional output and, combining Equation \[ \text{D.17} \] with \[ iii \] of Proposition \[ \text{D.9} \] and Lemma \[ C.3 \], at most \( CK + C' + 4 \leq CK \leq CK \) number of nonzero weights. Here, \( C' \) is the constant from Proposition \[ \text{D.9} \].

ad (ii): The first \( L - 1 \) layers of \( \Psi_{\varepsilon, \Phi} \) and \( \Phi \) coincide and for the approximation properties it holds that

\[
\left\| \left[ R_{\varepsilon}(\Psi_{\varepsilon, \Phi}) - \prod_{l=1}^{m+1} [R_{\varepsilon}(\Phi)]_l \right] \right\|_{W^k, \infty((0,1)^d)} = \left\| \left[ R_{\varepsilon}(\tilde{\times}) \circ R_{\varepsilon}(\tilde{\Psi}_{\varepsilon, \Phi}) - [R_{\varepsilon}(\Phi)]_{m+1} \prod_{l=1}^{m} [R_{\varepsilon}(\Phi)]_l \right] \right\|_{W^k, \infty((0,1)^d)} \leq \left\| \left[ R_{\varepsilon}(\tilde{\times}) \circ \left( R_{\varepsilon}(\Psi_{\varepsilon, \Phi, m}) \cdot [R_{\varepsilon}(\Phi)]_{m+1} - R_{\varepsilon}(\Psi_{\varepsilon, \Phi, m}) \cdot [R_{\varepsilon}(\Phi)]_{m+1} \right) \right] \right\|_{W^k, \infty((0,1)^d)} + \left\| \prod_{l=1}^{m} [R_{\varepsilon}(\Phi)]_l \right\|_{W^k, \infty((0,1)^d)}.
\]

We continue by considering the first term of the Inequality \[ \text{D.18} \] and bound the \( k \)-semi-norm of this term. Since the case \( k = 2 \) is the most complicated one and \( k \in \{0, 1\} \) can be shown similarly, we only state the case \( k = 2 \). We apply the chain rule from Corollary \[ \text{B.6} \] for \( g : \mathbb{R}^2 \to \mathbb{R} \) with \( g(x, y) = R_{\varepsilon}(\tilde{\times})(x, y) = x \cdot y \) and \( f : \mathbb{R}^d \to \mathbb{R}^2 \) with \( f = R_{\varepsilon}(\tilde{\Psi}_{\varepsilon, \Phi}) \). We get

\[
\left\| \left[ R_{\varepsilon}(\tilde{\times}) \circ \left( R_{\varepsilon}(\Psi_{\varepsilon, \Phi, m}) \cdot [R_{\varepsilon}(\Phi)]_{m+1} - R_{\varepsilon}(\Psi_{\varepsilon, \Phi, m}) \cdot [R_{\varepsilon}(\Phi)]_{m+1} \right) \right] \right\|_{W^{2, \infty}((0,1)^d)} \leq C \cdot \| R_{\varepsilon}(\tilde{\times})(x, y) - x \cdot y \|_{W^{2, \infty}((-B, B)^2; dx dy)} \right\|_{W^{2, \infty}((0,1)^d \times \mathbb{R}^2)}^2 + C \cdot \| R_{\varepsilon}(\tilde{\times})(x, y) - x \cdot y \|_{W^{1, \infty}((-B, B)^2; dx dy)} \right\|_{W^{2, \infty}((0,1)^d \times \mathbb{R}^2)} \leq C(\varepsilon CN^2 + \varepsilon CN^{2+2}) \leq C\varepsilon N^{2+2},
\]

where we used the induction hypothesis together with \( \| [R_{\varepsilon}(\Phi)]_{m+1} \|_{W^{1, \infty}((0,1)^d)} \leq cN^{k+\mu(x=2)} \) in the third step and assumed that \( c \leq C \). Similarly, we can show the bound \( C\varepsilon N^k \) for for the \( k \)-semi-norms with
\[ k \in \{0, 1\}. \text{ Combining the statements of the semi-norms then yields the required bound for the norm. To estimate the second term of } (D.18) \text{ we apply the product rule from Lemma B.5 and get}
\[
\left\| \left[ R_\varrho(\Phi) \right]_{m+1} \cdot \left( R_\varrho(\psi_{m}) - \prod_{l=1}^{m} [R_\varrho(\Phi)]_l \right) \right\|_{W^{k,\infty}(0,1)^d}
\leq \sum_{j=0}^{k} \left\| [R_\varrho(\Phi)]_{m+1} \right\|_{W^{j,\infty}(0,1)^d} \cdot \left\| R_\varrho(\psi_{m}) - \prod_{l=1}^{m} [R_\varrho(\Phi)]_l \right\|_{W^{k-j,\infty}(0,1)^d}
\leq \sum_{j=0}^{k} cN^{j+\mu(j+2)} \cdot CN^{k-j+\mu(k-j+2)} \varepsilon \leq 2cCN^{k+\mu(k+2)} \varepsilon,
\]
for \( k \in \{0, 1, 2\} \). For the second step, we used again the induction hypothesis together with
\[
\left\| [R_\varrho(\Phi)]_{m+1} \right\|_{W^{k,\infty}(0,1)^d} \leq cN^{k+\mu(k+2)}.
\]
Combining (D.18) with (D.19) and (D.20) yields
\[
\left\| R_\varrho(\psi_{m}) - \prod_{l=1}^{m} [R_\varrho(\Phi)]_l \right\|_{W^{k,\infty}(0,1)^d} \leq CN^{k+\mu(k+2)} \varepsilon.
\]

\textbf{ad (iii):} Next, we show that
\[
\left\| R_\varrho(\psi_{m}) \right\|_{W^{k,\infty}(0,1)^d} \leq CN^{k+\mu(k+2)}.
\]
Again, we only show the case \( k = 2 \). The other cases can be shown in the same way. Similarly as in (D.19) we have that
\[
\left\| R_\varrho(\psi_{m}) \right\|_{W^{2,\infty}(0,1)^d} = \left\| R_\varrho(\tilde{x} \circ \tilde{\psi}_{m}) \right\|_{W^{2,\infty}(0,1)^d}
\leq C \cdot \left\| R_\varrho(\tilde{x}) \right\|_{W^{2,\infty}((-B,B)^2)} \cdot \left\| R_\varrho(\tilde{\psi}_{m}) \right\|_{W^{1,\infty}(0,1)^d; \mathbb{R}^d}
+ C \cdot \left\| R_\varrho(\tilde{x}) \right\|_{W^{1,\infty}((-B,B)^2)} \cdot \left\| R_\varrho(\tilde{\psi}_{m}) \right\|_{W^{2,\infty}(0,1)^d; \mathbb{R}^d}
\leq CN^2 + CN^{2+\mu} \leq C \cdot N^{2+\mu},
\]
where Corollary B.6 was used for the second step and Proposition D.9 (ii), together with an argument as in (D.19), implies the third step.

\textbf{ad (iv):} Finally, we need to derive a bound for the absolute values of the weights. From the definition of \( \psi_{m, \varrho} \) and Lemma C.3 we get
\[
\left\| \psi_{m, \varrho} \right\|_{\max} = \left\| \tilde{x} \circ \tilde{\psi}_{m, \varrho} \right\|_{\max} \leq 2 \max\{B, 1\} \cdot \max\{\|\tilde{x}\|_{\max}, \|\tilde{\psi}_{m, \varrho}\|_{\max}\}.
\]
From \( \|\tilde{x}\|_{\max} \leq C \varepsilon^{-2} \) (see Proposition D.9 (iv)) and \( \|\tilde{\psi}_{m, \varrho}\|_{\max} \leq C \max\{\|\Phi\|_{\max}, \varepsilon^{-2}\} \) (see Step 2) it follows that \( \left\| \psi_{m, \varrho} \right\|_{\max} \leq C \max\{\|\Phi\|_{\max}, \varepsilon^{-2}\} \). This concludes the proof.

In the last part of this subsection, we are finally in a position to construct neural networks which approximate sums of localized polynomials.

\textbf{Lemma D.11.} \textit{Let } \( d, N \in \mathbb{N}, k \in \{0, 1, 2\}, n \in \mathbb{N}_{\geq k+1}, 1 \leq p \leq \infty, \text{ and } \mu > 0. \text{ Set } s := N^\mu \text{ and let } \Psi = \Psi(d, N, \mu) = \{ \phi_m^s : m \in \{0, \ldots, N\}^d \} \text{ be the partition of unity from Proposition D.9. Additionally, let } \varrho = \text{ELU}_1. \text{ Then, there is a constant } C = C(n, d, p) > 0 \text{ with the following properties:}
\]
Let $\varepsilon \in (0, 1/2)$, $f \in W^{n,p}(0, 1)^d$ and $p_m(x) := p_{f,m}(x) = \sum_{|\alpha| \leq n-1} c_{f,m,\alpha} x^{\alpha}$ for $m \in \{0, \ldots, N\}^d$ be the polynomials from Lemma \ref{lem:polynomial-prediction}. Then there is a neural network $\Phi_{P,\varepsilon} = \Phi_{P,\varepsilon}(f, d, n, N, \mu, \varepsilon)$ with $d$-dimensional input and one-dimensional output, with at most $C$ layers and $C(N+1)^d$ nonzero weights, such that

$$\left\| \sum_{m \in \{0, \ldots, N\}^d} \phi_m^* P_m - R_{\varepsilon}(\Phi_{P,\varepsilon}) \right\|_{W^{k,p}(0,1)^d} \leq C\|f\|_{W^{n,p}((0,1)^d)} \varepsilon,$$

and $\|\Phi_{P,\varepsilon}\|_{\text{max}} \leq C \max \{\|f\|_{W^{n,p}((0,1)^d)^d}, N^{d/p}, \varepsilon^{-2} N^{2(d/p+2d+\mu)}\}$.

**Proof.** **Step 1 (Approximating localized monomials $\phi_m^*(x) x^{\alpha}$):** Let $|\alpha| \leq n-1$ and $m \in \{0, \ldots, N\}^d$.

It is easy to see that there is a neural network $\Phi_\alpha$ with $d$-dimensional input and $|\alpha|$-dimensional output, with two layers, $2(n-1)$ nonzero weights bounded in absolute value by one such that

$$x^{\alpha} = \prod_{l=1}^{|\alpha|} |R_{\varepsilon}(\Phi_\alpha)|_l(x), \quad \text{for all } x \in (0,1)^d,$$

and

$$\| |R_{\varepsilon}(\Phi_\alpha)|_l \|_{W^{k,\infty}(0,1)^d} \leq 1, \quad \text{for all } l = 1, \ldots, |\alpha|. \quad (D.21)$$

Let now $\Phi_m$ be the neural network from Proposition \ref{prop:approximation-polynomials} (for $s = N^\mu$) and define the network

$$\Phi_{m,\alpha} := P(\Phi_m, \Phi_\alpha),$$

where the parallelization is provided by Lemma \ref{lem:parallelization} (here, $B = 0$). Consequently, $\Phi_{m,\alpha}$ has $2 \leq K_0$ layers and $C + 2(n-1) \leq K_0$ nonzero weights for a suitable constant $K_0 = K_0(n, d) \in \mathbb{N}$, $\|\Phi_{m,\alpha}\|_{\text{max}} \leq C N^{1+\mu}$ and $\prod_{l=1}^{|\alpha|+d} |R_{\varepsilon}(\Phi_{m,\alpha})|_l(x) = \phi_m^*(x) x^{\alpha}$ for all $x \in (0,1)^d$. Moreover, as a consequence of Proposition \ref{prop:approximation-polynomials} (v) together with Equation $(D.21)$ we have

$$\| |R_{\varepsilon}(\Phi_{m,\alpha})|_l \|_{W^{k,\infty}(0,1)^d} \leq C N^{k+\mu(k=2)}, \quad \text{for all } l = 1, \ldots, |\alpha| + d.$$

To construct an approximation of the localized monomials $\phi_m^*(x) x^{\alpha}$, set $\tilde{\varepsilon} := \varepsilon N^{-(d/p+2\mu)}$ and let $\Psi_{\tilde{\varepsilon},(m,\alpha)}$ be the neural network provided by Lemma \ref{lem:approximation-polynomials} (with $\Phi_{m,\alpha}$ instead of $\Phi$, $m = |\alpha| + d \in \mathbb{N}$, $K = K_0 \in \mathbb{N}$) for $m \in \{0, \ldots, N\}^d$ and $\alpha \in \mathbb{N}_0^d$, $|\alpha| \leq n-1$. Then $\Psi_{\tilde{\varepsilon},(m,\alpha)}$ has at most $C$ layers, number of nonzero weights and $\|\Psi_{\tilde{\varepsilon},(m,\alpha)}\|_{\text{max}} \leq C \max \{N^{1+\mu}, \varepsilon^{-2} N^{2(d/p+2\mu)}\}$. Moreover,

$$\| \phi_m^*(x) x^{\alpha} - R_{\varepsilon}(\Psi_{\tilde{\varepsilon},(m,\alpha)})(x) \|_{W^{k,\infty}(0,1)^d} \leq C N^{k+\mu(k=2)} \tilde{\varepsilon} \leq C \varepsilon N^{-d/p}.$$

**Step 2 (Constructing $\Phi_{P,\varepsilon}$):** We set

$$T := \{(m, \alpha) : m \in \{0, \ldots, N\}^d, \alpha \in \mathbb{N}_0^d, |\alpha| \leq n-1\}.$$

From Lemma \ref{lem:approximation-polynomials} (iii) we get there exists a constant $B = B(n, d)$ such that $\text{Range} \ R_{\varepsilon}(\Psi_{\tilde{\varepsilon},(m,\alpha)}) \subset [-B, B]$ for $m \in \{0, \ldots, N\}^d$ and $\alpha \in \mathbb{N}_0^d$, $|\alpha| \leq n-1$. We will now make use of the parallelization (Lemma \ref{lem:parallelization} for $B = 0$) of the localized polynomial approximations

$$P(\Psi_{\tilde{\varepsilon},(m,\alpha)} : m \in \{0, \ldots, N\}^d, \alpha \in \mathbb{N}_0^d, |\alpha| \leq n-1)$$

and note that the resulting network has at most $C$ layers and $CT$ nonzero weights bounded in absolute value by $C \max \{N^{1+\mu}, \varepsilon^{-2} N^{2(d/p+2\mu)}\} \leq C \varepsilon^{-2} N^{2(d/p+2\mu)}$. Next, we define the matrix $A_{\text{sum}} \in \mathbb{R}^{1,T}$ by $A_{\text{sum}} := |c_{f,m,\alpha} : m \in \{0, \ldots, N\}^d, \alpha \in \mathbb{N}_0^d, |\alpha| \leq n-1|$ and the neural network $\Phi_{\text{sum}} := (A_{\text{sum}}, 0)$. Finally, we set

$$\Phi_{P,\varepsilon} := \Phi_{\text{sum}} \circ P(\Psi_{\tilde{\varepsilon},(m,\alpha)} : m \in \{0, \ldots, N\}^d, \alpha \in \mathbb{N}_0^d, |\alpha| \leq n-1).$$
From Lemma [C.3](with \(B = B\)) we get \(\Phi_{P,\varepsilon}\) is a neural network with \(d\)-dimensional input and one-dimensional output, with at most \(1 + C\) layers and \(T + CT + 2M \leq C(N + 1)^d\) nonzero weights. For the absolute values of the weights it holds that \(\|\Phi_{P,\varepsilon}\|_{\text{max}} \leq C\max\{\|f\|_{W^{n,\infty}(0,1)^d}, N^{d/p}, \varepsilon^{-2(d/p+2+\mu)}\}\) where we used the bound for the coefficients \(c_{f,m,\alpha}\) from Remark [D.5](#D.5). Moreover, we have

\[
R_\varepsilon(\Phi_{P,\varepsilon}) = \sum_{m \in \{0, \ldots, N\}^d} \sum_{|\alpha| \leq n-1} c_{f,m,\alpha} \Psi_{\varepsilon,(m,\alpha)}.
\]

Note that the network \(\Phi_{P,\varepsilon}\) only depends on \(p_{f,m}\) (and thus on \(f\)) via the coefficients \(c_{f,m,\alpha}\).

**Step 3 (Estimating the approximation error in \(\|\cdot\|_{W^{k,p}}\))**: We get

\[
\left\| \sum_{m \in \{0, \ldots, N\}^d} \phi_m(x)p_m(x) - R_\varepsilon(\Phi_{P,\varepsilon})(x) \right\|_{W^{k,p}(0,1)^d;dx}
\]

\[
= \left\| \sum_{m \in \{0, \ldots, N\}^d} \sum_{|\alpha| \leq n-1} c_{f,m,\alpha} \left( \phi_m(x)x^\alpha - R_\varepsilon(\Psi_{\varepsilon,(m,\alpha)})(x) \right) \right\|_{W^{k,p}(0,1)^d;dx}
\]

\[
\leq \sum_{m \in \{0, \ldots, N\}^d} \sum_{|\alpha| \leq n-1} |c_{f,m,\alpha}| \left\| \phi_m(x)x^\alpha - R_\varepsilon(\Psi_{\varepsilon,(m,\alpha)})(x) \right\|_{W^{k,p}(0,1)^d;dx}
\]

\[
\leq \sum_{m \in \{0, \ldots, N\}^d} \sum_{|\alpha| \leq n-1} \|f\|_{W^{n,\infty}(\Omega_{m,N})} N^{d/p} C \varepsilon N^{-d/p},
\]

where we used again the bound for the coefficients \(c_{f,m,\alpha}\) together with \(\|\cdot\|_{W^{k,p}(0,1)^d} \leq C\|\cdot\|_{W^{k,\infty}(0,1)^d}\) in the last step. Similar as in Equation [D.13](#D.13) we finally have

\[
\left\| \sum_{m \in \{0, \ldots, N\}^d} \phi_m(x)p_m(x) - R_\varepsilon(\Phi_{P,\varepsilon})(x) \right\|_{W^{k,p}(0,1)^d;dx} \leq C \varepsilon \sum_{m \in \{0, \ldots, N\}^d} \|f\|_{W^{n,\infty}(\Omega_{m,N})} \leq C \|f\|_{W^{n,\infty}(0,1)^d}.
\]

This concludes the proof. □

**D.4 Putting Everything Together**

Now we conclude the proof of Proposition [4.1](#4.1).

**Proof of Proposition 4.1**. We divide the proof into two steps: First, we approximate the function \(f\) by a sum of localized polynomials. Afterwards, we proceed by approximating this sum by a neural network.

For the first step, we set

\[
N := \left( \frac{\varepsilon}{2C} \right)^{-1/(n-k-\mu(k=2))} \quad \text{and} \quad s := N^\mu,
\]

where \(\overline{C} = \overline{C}(n,d,p) > 0\) is the constant from Lemma [D.4](#D.4). Without loss of generality we may assume that \(\overline{C} \geq 1\). The same lemma yields that if \(\Psi = \Psi(d, N, \mu) = \{\phi_s^m : m \in \{0, \ldots, N\}^d\}\) is the partition of unity from Proposition [D.3](#D.3) and \(\overline{N} = \overline{N}(d, p, \mu)\) is the constant from Lemma [D.4](#D.4) then there exist polynomials \(p_m(x) = \sum_{|\alpha| \leq n-1} c_{f,m,\alpha} x^\alpha\) for \(m \in \{0, \ldots, N\}^d\) such that

\[
\left\| f - \sum_{m \in \{0, \ldots, N\}^d} \phi_s^m p_m \right\|_{W^{k,p}(0,1)^d} \leq \overline{C} \left( \frac{1}{N} \right)^{n-k-\mu(k=2)} \leq \overline{C} \frac{\varepsilon}{2C} = \frac{\varepsilon}{2}, \tag{D.22}
\]

where
for all $\varepsilon \in (0, \varepsilon_0)$, where $\varepsilon_0 = \varepsilon_0(d, p, \mu) > 0$ is chosen such that $N \geq \tilde{N}$.

For the second step, let $\tilde{C}' = \tilde{C}'(n, d, p)$ be the constant from Lemma D.11 and $\Phi_{P, \varepsilon}$ be the neural network provided by Lemma D.11 with $\varepsilon/(2\tilde{C}')$ instead of $\varepsilon$. Then $\Phi_{P, \varepsilon}$ has at most $\tilde{C}'$ layers and at most

$$\tilde{C}' \left( \frac{\varepsilon}{2\tilde{C}'} \right)^{-1/(n-k-\mu(k=2))} + 2 \leq \tilde{C}' \left( \frac{\varepsilon}{2\tilde{C}'} \right)^{-d/(n-k-\mu(k=2))} \leq C \varepsilon^{-d/(n-k-\mu(k=2))}$$

nonzero weights. In the first step we have used $(2\tilde{C}')/\varepsilon \geq 1$. The weights are bounded in absolute value by

$$\|\Phi_{P, \varepsilon}\|_{\max} \leq \tilde{C}' \varepsilon^{-2} N^{2(d/p+2+\mu)} \leq C \varepsilon^{-2-2(d/p+2+\mu)/(n-k-\mu(k=2))} = C \varepsilon^{-\theta},$$

for a suitable $\theta = \theta(d, p, k, n, \mu) > 0$. Additionally, there holds

$$\left\| \sum_{m \in \{0, \ldots, N\}^d} \phi_m p_m - R_\varepsilon(\Phi_{P, \varepsilon}) \right\|_{W^{k, p}(\Phi(0, 1)^d)} \leq \frac{\tilde{C}' \varepsilon}{2\tilde{C}'} \leq \frac{\varepsilon}{2}. \tag{D.23}$$

By applying the triangle inequality as well as Equations (D.22) and (D.23) we arrive at

$$\|f - R_\varepsilon(\Phi_{P, \varepsilon})\|_{W^{k, p}(\Omega)^d} \leq \left\| f - \sum_{m \in \{0, \ldots, N\}^d} \phi_m p_m \right\|_{W^{k, p}(\Omega)^d} + \left\| \sum_{m \in \{0, \ldots, N\}^d} \phi_m p_m - R_\varepsilon(\Phi_{P, \varepsilon}) \right\|_{W^{k, p}(\Omega)^d} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

thereby concluding the proof.

E Encodable Weights and Proof of Theorem 4.2

In this section, we investigate how rounding errors of the weights in the last layer affect the approximation error of a neural network. For this, the next lemma shows that under some conditions on the growth of the activation function and on the functions implemented by the first up to the next-to-last layer, the error resulting from rounding the weights in the last layer can be controlled in the $\| \cdot \|_{W^{k, \infty}(\Omega)}$ semi-norm for $k = 0, 1, 2$. The following lemma could be stated in more generality, in particular, the conditions on the activation function $\varphi$ could be weakened. However, to simplify the exposition, we took only care that the conditions are satisfied by the ELU.

Lemma E.1. Let $d \in \mathbb{N}$ and $\varepsilon \in (0, 1/2)$. Moreover, let $k \in \{0, 1, 2\}$, $s_1, s_2 \in \mathbb{N}_0$ and let $\varphi : \mathbb{R} \to \mathbb{R}$ be $k$-times weakly differentiable such that $\varphi$ is 1-Lipschitz and $\varphi(0) = 0$. We set $T := \lfloor \varepsilon^{-s_2} \rfloor$ and assume that for some function $f : ((0, 1)^d) \to \mathbb{R}$ we have $f_m \in W^{k, \infty}((0, 1)^d)$ for all $m = 1, \ldots, T$. In addition, let $A = (a_m)_{m=1}^T \in \mathbb{R}^{1, T}$ and $b \in \mathbb{R}$. Then we assume one of the following three setups:

(i) $k = 0$ and $\|f_m\|_{L^{\infty}((0, 1)^d)} \leq \varepsilon^{-s_1}$;

(ii) $k = 1$, $|\varphi|_{W^{1, \infty}(\mathbb{R})} \leq 1$ and $\|f_m\|_{W^{1, \infty}((0, 1)^d)} \leq \varepsilon^{-s_1}$;

(iii) $k = 2$, $|\varphi|_{W^{1, \infty}(\mathbb{R})}, |\varphi|_{W^{2, \infty}(\mathbb{R})} \leq 1$ and $\|f_m\|_{W^{2, \infty}((0, 1)^d)} \leq \varepsilon^{-s_1}$,
for $m = 1, \ldots, T$. In each case, there exists $\nu = \nu(k, s_1, s_2, \varrho) \in \mathbb{N}$ such that

$$
\left\| A \cdot \varphi(f) + b - \tilde{A} \cdot \varphi(f) - \tilde{b} \right\|_{W^{k, \infty}((0,1)^d)} = \left\| \sum_{m=1}^{T} a_m \varphi(f_m) + b - \sum_{m=1}^{T} \tilde{a}_m \varphi(f_m) - \tilde{b} \right\|_{W^{k, \infty}((0,1)^d)} \leq \varepsilon,
$$

where $\tilde{\cdot} : \mathbb{R} \to \varepsilon^\nu \mathbb{Z}$ is the rounding operator and $\varphi$ is applied componentwise.

**Proof.** (i): We set $\nu := s_1 + s_2 + 2$ and have

$$
\left\| \sum_{m=1}^{T} a_m \varphi(f_m) + b - \sum_{m=1}^{T} \tilde{a}_m \varphi(f_m) - \tilde{b} \right\|_{L^\infty((0,1)^d)} \leq \sum_{m=1}^{T} |a_m - \tilde{a}_m| \| \varphi \circ f_m \|_{L^\infty((0,1)^d)} + |b - \tilde{b}|
$$

(under the assumption $\varphi$ is 1-Lipschitz, $\varphi(0) = 0$)

$$
\leq \sum_{m=1}^{T} \varepsilon^\nu \| \varphi \circ f_m \|_{L^\infty((0,1)^d)} + \varepsilon^\nu \leq \varepsilon^\nu \varepsilon^{s_2} \varepsilon^{s_1} + 1 \leq 2 \varepsilon^\nu \varepsilon^{s_1} \varepsilon^{-s_2} \leq \varepsilon,
$$

where we have used in the last step that $\varepsilon \leq 1/2$.

(ii): We define $\nu := s_1 + s_2 + 1 + \log_2(C)$, where $C \geq 1$ is the constant from Corollary 4.6 and get in a similar manner as in (i) that

$$
\left\| \sum_{m=1}^{T} a_m \varphi(f_m) + b - \sum_{m=1}^{T} \tilde{a}_m \varphi(f_m) - \tilde{b} \right\|_{W^{1, \infty}((0,1)^d)} \leq \sum_{m=1}^{T} |a_m - \tilde{a}_m| \| \varphi \circ f_m \|_{W^{1, \infty}((0,1)^d)}
$$

(Corollary 4.6) 

$$
\leq \sum_{m=1}^{T} C \varepsilon^\nu \| \varphi \|_{W^{1, \infty}((0,1)^d)} \| f_m \|_{W^{1, \infty}((0,1)^d)} \leq \varepsilon^{-s_2} \varepsilon^{-\log_2(C)} \varepsilon^\nu \varepsilon^{s_1} \leq \varepsilon,
$$

where we used in the third step that $C = 2^{\log_2(C)} \leq \varepsilon^{-\log_2(C)}$. Combining this observation with (i) yields the claim.

(iii): Using the same steps as above we set $\nu := 2s_1 + 2s_2 + 1 + \log_2(C)$, where $C \geq 1$ is again the constant from Corollary 4.6 and get

$$
\left\| \sum_{m=1}^{T} a_m \varphi(f_m) + b - \sum_{m=1}^{T} \tilde{a}_m \varphi(f_m) - \tilde{b} \right\|_{W^{2, \infty}((0,1)^d)} \leq \sum_{m=1}^{T} |a_m - \tilde{a}_m| \| \varphi \circ f_m \|_{W^{2, \infty}((0,1)^d)}
$$

(Corollary 4.6) 

$$
\leq \sum_{m=1}^{T} \varepsilon^\nu C \left( \| \varphi \|_{W^{2, \infty}((0,1)^d)} \| f_m \|_{W^{2, \infty}((0,1)^d)}^2 + \| \varphi \|_{W^{1, \infty}((0,1)^d)} \| f_m \|_{W^{1, \infty}((0,1)^d)} \right) \leq \varepsilon^{-s_2} \varepsilon^\nu \varepsilon^{-\log_2(C)} \left( \varepsilon^{-2s_1} + \varepsilon^{-s_1} \right) \leq \varepsilon.
$$

Combining this observation with (i) and (ii) yields the claim. □
We are almost prepared to prove Theorem 4.2. As a preparation, the next remark is a collection of statements we did not include in Proposition 4.1 for improved readability, but are important to show the encodability of the weights.

**Remark E.2.** The neural networks $\Phi_{\varepsilon,f} = ((A_1,b_1),\ldots,(A_{L-1},b_{L-1}),(A_L,b_L))$ are constructed in such a way that only the weights $(A_L,b_L)$ in the last layer depend on the function $f$. In other words, the weights in the layers $1,\ldots,L-1$ are independent from $f$. They only depend on $\varepsilon,n,d,p,k,\mu$. This fact will play a crucial role in proving the encodability of the neural network weights in Theorem 4.2. Additionally, we will use the following fact:

Set $\Phi_{\varepsilon}^{1,\ldots,L-1} := ((A_1,b_1),\ldots,(A_{L-1},b_{L-1}))$, then Lemma D.10 in combination with Step 1 and 2 of the proof of Lemma D.11 show that for every $i = 1,\ldots,N_{L(\Phi_{\varepsilon,f})-1}$ we have, for $N := \lceil (\frac{\varepsilon}{C^\prime})^{-1/(n-k-\mu(k=2))} \rceil$, that

$$\|R_\varepsilon(\Phi_{\varepsilon}^{1,\ldots,L-1})|_{i}\|_{W^{k,\infty}(0,\infty)^d} \leq CN^{k+\mu(k=2)} \leq C\varepsilon^{-(k+\mu(k=2))/(n-k-\mu(k=2))}.$$ 

Finally, we use that the number of neurons $N_{L(\Phi_{\varepsilon,f})-1}$ in the $L-1$st layer is bounded by $C\varepsilon^{-d/(n-k-\mu(k=2))}$.

The proof of Theorem 4.2 now mainly consists of showing that the conditions from Lemma E.1 are fulfilled and that the weights from the layers $1,\ldots,L-1$, which are not rounded, can be encoded by a moderate number of bits.

**Proof of Theorem 4.2.** Let $C = C(d,n,p,\mu) > 0$, $\theta = \theta(d,n,p,k,\mu) > 0$ and $\bar{\varepsilon} = \bar{\varepsilon}(d,p,\mu,?) > 0$ be the constants from Proposition 4.1 and let $\varepsilon \in (0,\bar{\varepsilon})$. Moreover, for $f \in \mathcal{F}_{n,d,p}$, let $\Phi_{\varepsilon,f} = ((A_1,b_1)|_{\varepsilon,1}^{L-1})$ be the neural network from Proposition 4.1 with at most $L$ layers and $M(\Phi_{\varepsilon,f}) \leq C \cdot \varepsilon^{-d/(n-k-\mu(k=2))}$ nonzero weights bounded in absolute value by $C\varepsilon^{-\theta}$, such that

$$\|R_\varepsilon(\Phi_{\varepsilon,f}) - f\|_{W^{k,p},(0,1)^d} \leq \frac{\varepsilon}{2}.$$ 

**Step 1 (Rounding the weights in the last layer):** Set now $\Phi_{\varepsilon}^{1,\ldots,L-1} := ((A_1,b_1),\ldots,(A_{L-1},b_{L-1}))$ and denote by $N_{L-1}$ the number of neurons in the next-to-last layer of $\Phi_{\varepsilon,f}$ (which is the number of neurons in the last layer of $\Phi_{\varepsilon}^{1,\ldots,L-1}$). Moreover, let $C_1 = \max\{C',C''\}$ be the maximum of the constants from Remark E.2. We then get from this remark, that the weights of $\Phi_{\varepsilon}^{1,\ldots,L-1}$ do not depend on $f$ and that for $m = 1,\ldots,N_{L-1}$ and $k \in \{0,1,2\}$ we have

$$\|R_\varepsilon(\Phi_{\varepsilon}^{1,\ldots,L-1})|_{m}\|_{W^{k,\infty}(0,1)^d} \leq C_1 \left( \frac{\varepsilon}{2} \right)^{-\frac{\mu(k=2)}{n-k-\mu(k=2)}} \leq \left( \frac{\varepsilon}{2} \right)^{-s_1},$$

and

$$N_{L-1} \leq C_1 \left( \frac{\varepsilon}{2} \right)^{-\frac{d}{n-k-\mu(k=2)}} \leq \left( \frac{\varepsilon}{2} \right)^{-s_2},$$

where we set

$$s_1 := \frac{3}{n} + \log_2(C_1) > 0 \quad \text{and} \quad s_2 := \frac{d}{n} + \log_2(C_1) > 0$$

and used that $C_1 = 2\log_2(C_1) \leq e^{-\log_2(C_1)}$. We now apply Lemma E.1 and deduce that there exists a rounding precision $\nu = \nu(s_1,s_2,k)$ such that for the neural network

$$\tilde{\Phi}_{\varepsilon,f} = ((A_1,b_1),\ldots,(A_{L-1},b_{L-1}),(\tilde{A}_L,\tilde{b}_L))$$

where $\tilde{A}_L \in [\varepsilon^{-\theta},\varepsilon^{-\theta}]\cap C^{N_{L-1}}\mathbb{Z}$ and $\tilde{b}_L \in [-\varepsilon^{-\theta},\varepsilon^{-\theta}]\cap \mathbb{Z}$ are the rounded weight matrix $A_L \in \mathbb{R}^{1,N_{L-1}}$ and $b_L \in \mathbb{R}$, respectively, we have

$$\|\tilde{R}_\varepsilon(\Phi_{\varepsilon,f}) - R_\varepsilon(\tilde{\Phi}_{\varepsilon,f})\|_{W^{k,p},(0,1)^d} \leq \|A_L \cdot g(R_\varepsilon(\Phi_{\varepsilon}^{1,\ldots,L-1})) + b_L - \tilde{A}_L \cdot g(R_\varepsilon(\Phi_{\varepsilon}^{1,\ldots,L-1})) - \tilde{b}_L\|_{W^{k,\infty}(0,1)^d} \leq \varepsilon/2.$$

42
This implies (by the triangle inequality) that
\[ \| f - R_\theta(\tilde{\Phi}_{\varepsilon,f}) \|_{W^{k,p}((0,1)^d)} \leq \varepsilon. \]

**Step 2 (Construction of coding scheme):** We will now show that there is a constant \( C_2 = C_2(d,n,p,k,\mu) > 0 \) and a coding scheme \( B = (B_\ell)_{\ell \in \mathbb{N}} \) such that for each \( \varepsilon > 0 \) and each \( f \in F_{n,d,p} \) the nonzero weights of \( \tilde{\Phi}_{\varepsilon,f} \) are in Range \( B_{\lceil C_2 \log(1/\varepsilon) \rceil} \).

If we denote by \( W_1, \ldots, W_{L-1} \) the collection of nonzero weights of \( \Phi_{\varepsilon} \) (which are independent of \( f \)), then we have \( |W_1, \ldots, W_{L-1}| = M(\Phi_{\varepsilon}) \leq M(\Phi_{\varepsilon,f}) \leq \varepsilon^{-s_2} \) (with potentially a different constant in the definition of \( s_2 \)). Furthermore, we have \( |[-\varepsilon^{-\theta}, \varepsilon^{-\theta}] \cap \varepsilon^\nu \mathbb{Z}| = 2|\varepsilon^{-\theta-\nu}| + 1 \leq \varepsilon^{-s_3} \) with \( s_3 := \theta + \nu + 2 \), such that there exists a surjective mapping
\[
B_{\lceil (s_2 + s_3) \log_2(1/\varepsilon) \rceil} : \{0,1\}^{\lceil (s_2 + s_3) \log_2(1/\varepsilon) \rceil} \to W_1, \ldots, W_{L-1} \cup \left( [-\varepsilon^{-\theta}, \varepsilon^{-\theta}] \cap \varepsilon^\nu \mathbb{Z} \right).
\]
Setting \( C_2 := s_2 + s_3 \) yields the claim.